W128–W132 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* *Published online 2 May 2008*
doi:10.1093/nar/gkn195

ConTra: a promoter alignment analysis tool for identification of transcription factor binding sites across species

Bart Hooghe[1,2,3,*], Paco Hulpiau[1,2,3], Frans van Roy[2,3] and Pieter De Bleser[1,2,3]

[1]Bioinformatics Core Facility, VIB, [2]Department of Molecular Biology, Ghent University and [3]Department for Molecular Biomedical Research, VIB, B-9052 Ghent, Belgium

Received January 31, 2008; Revised March 28, 2008; Accepted April 2, 2008

ABSTRACT

Transcription factors (TFs) are key components in signaling pathways, and the presence of their binding sites in the promoter regions of DNA is essential for their regulation of the expression of the corresponding genes. Orthologous promoter sequences are commonly used to increase the specificity with which potentially functional transcription factor binding sites (TFBSs) are recognized and to detect possibly important similarities or differences between the different species. The ConTra (conserved TFBSs) web server provides the biologist at the bench with a user-friendly tool to interactively visualize TFBSs predicted using either TransFac (1) or JASPAR (2) position weight matrix libraries, on a promoter alignment of choice. The visualization can be preceded by a simple scoring analysis to explore which TFs are the most likely to bind to the promoter of interest. The ConTra web server is available at http://bioit.dmbr.ugent.be/ConTra/index.php.

INTRODUCTION

Nowadays, context-specific changes in gene expression levels can be easily monitored on a genome-wide scale by using microarray analysis and serial analysis of gene expression, but the molecular mechanisms and the specific transcription factors (TFs) that drive those specific changes remain unknown in most cases. Identification of the components and mechanisms of signaling pathways is a slow process that inevitably involves a strategy of trial-and-error. Therefore, *in silico* prediction of the components before and during the identification process is highly desirable.

In silico approaches estimate that there are about 2000 human TFs (3), of which about 800 have been characterized to varying degrees. For many of them, information on DNA-binding sites is available, allowing the modeling of binding characteristics to a reasonable extent. The most commonly used model for TF binding specificity is the position weight matrix (PWM), although it does not account for potential position dependencies within a transcription factor binding site (TFBS) (4).

When a PWM or even a more advanced model such as a hidden Markov model (HMM) is used to predict binding sites for a specific TF, the results include a very large proportion of false positives. The reason is that TFBSs are very short, often between 6 and 15 nt, and tolerate relatively high degrees of degeneracy in the sequence. The use of orthologous sequences to find conserved and, therefore, potentially functional TFBSs is called phylogenetic footprinting. This *in silico* technique is commonly and successfully used in combination with the PWM model to reduce its rate of false positive predictions. The main difficulties of this kind of approach lie in correct aligning regulatory elements in promoter sequences that might have diverged a lot during evolution (5).

Comparison of predicted TFBSs in one species with those of other species is not only used to reduce the number of false positive predictions, but also can be a goal in its own right. It is now widely accepted that many differences in animal morphology are due to specific changes in sequences that control gene expression, especially during development (6). Consequently, one expects to find important differences between species in the presence and position of TFBSs.

Conservation of a TFBS among several species observed in a multiple alignment is not proof that it is functional. Neither is the conservation of a TFBS required for functionality, because differences between species are at least as biologically important as the similarities. Furthermore, the apparent lack of conservation might not have biological reasons, but could result from 'incorrect' alignment. Thus, although systematic hard conclusions are extremely difficult to make, proper display of predicted

*To whom correspondence should be addressed. Tel: 003293313693; Email: bart.hooghe@dmbr.ugent.be

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

sites in several possible alignments would certainly be of help to the biologist seeking to generate or support a hypothesis.

Despite the availability of a number of web tools that offer phylogenetic footprinting together with some visualization interface, the biologist at the bench still lacks a compact and user-friendly tool that suggests answers to a regularly recurring question. ConTra, the web tool presented in this article, offers interactive visualization of all predicted sites for selected TFs on aligned sequences of orthologous promoters. ConTra works per alternative promoter to facilitate detection of their differences or similarities. Furthermore, a simple scoring analysis can be applied before visualization to identify the TFs that are most likely to bind the promoter(s) of interest.

APPROACH AND FEATURES

ConTra enables easy and fast look-up of all known transcripts related to the human gene(s) or transcript(s) of interest, given by gene name, gene symbol, Ensembl gene id, Entrez gene id, RefSeq transcript id or Ensembl transcript id. The results are fully linked to NCBI (http://www.ncbi.nlm.nih.gov/), UCSC (http://genome.ucsc.edu/) and Ensembl (http://www.ensembl.org/).

Transcripts are grouped according to transcription start site (TSS), and each group can be analyzed separately. This important feature of ConTra differentiates it from most other web tools that provide only one promoter per gene for analysis. The potential importance of alternative promoter regulation is exemplified by an alternative promoter of the DICER1 gene. The TSS of the DICER1 transcript NM_030621, predominantly expressed in breast tissue (7), is positioned more than 16 kb upstream from the TSS of the transcript NM_177438, which has been reported to be predominantly expressed in several other tissues (8). It is very likely that some important differences in the spectrum of TFBSs between the two promoters are causing the observed transcript proportion differences in different tissues, and ConTra could help to start exploring these differences. The transcriptional regulation of the DICER1 transcripts is further discussed in the supplemental data document nr 1.

For every group of transcripts that has been selected, available qualitative pairwise and multiple alignments on man from Ensembl or UCSC are offered to choose from, and they can be retrieved by simple selection. Offered alignments include the multiz 17-way and 28-way multiple alignments from UCSC (9,10). The 28-way alignment has been produced recently and has been proven powerful for exploring vertebrate and mammalian evolution (11). Other offered alignments are the Pecan (http://www.ebi.ac.uk/~bjp/pecan/) 7-mammals and 10-amniota-vertebrates multiple alignments from Ensembl. The Pecan algorithm has been shown to be one of the best algorithms in terms of specificity and sensitivity (12). ConTra also offers most available pairwise blastz-net alignments on man from UCSC (13,14). The premade alignments offered by ConTra always have the human promoter sequence as the reference sequence because in our experience these alignments are the most frequently asked for. However, users can upload in fasta format their own alignment files with any other reference species. This upload feature also allows the use of alignment types other than those provided. We also plan to enable the upload of own PWMs in order to expand the series of TFs for which predicted binding sites can be visualized.

All potential TFBSs are determined independently for each orthologous promoter using 'vertebrate' PWMs from the most recent versions of TransFac (1) or JASPAR (2). We have chosen to visualize TFBSs predicted by the simple, often used PWM system as is. Restricting the predicted TFBSs to only those that are phylogenetically conserved or taking into account extra features such as clustering tendency (15) or distance from TSS (16) would produce less false positive predictions. However, these filters would also, respectively, create a bias of the true positive predictions towards conserved TFBSs or towards TFBSs that meet the theoretical assumptions of models developed with too little experimental data. Prediction of TFBSs must and can be improved a lot, but much more experimental data needs to be really available, not just dispersed throughout scientific literature. Recently a few databases were designed that are suitable to contain complex regulatory data, namely ORegAnno and Pazar (17,18), and biologists are strongly encouraged to deposit their regulatory findings in these databases.

The parameters that can be set are the length of upstream promoter sequences and the thresholds for PWMs that correspond to the stringency to be used when predicting TFBSs.

The visualization of predicted TFBSs in HTML allows Javascript user interaction that is similar to the interaction provided by Jalview, a freely downloadable Java alignment editor (19). The interaction is crucial to keep visualization compact and interpretable. It also facilitates observation of potential coincident binding of several TFs and hence possible coregulation. Files needed for customized Jalview visualization, which is suitable for publication purposes, are provided as well. The results also include an overview picture for every promoter alignment.

ConTra provides links to experimentally defined binding sites in the selected promoter region when these are available in ORegAnno (17).

A typical output of ConTra visualization is depicted in Figure 1. More ConTra visualizations of experimentally proven TFBSs are linked from the ConTra doc page at http://bioit.dmbr.ugent.be/ConTra/contradoc.php#examples. This collection of examples will be expanded continuously.

The other part of ConTra, the exploration part, predicts which TFs are most likely to bind to the given promoter sequence(s). This prediction is done by using a simple, intuitive but effective score that takes into account the number of predicted binding sites, the extent of phylogenetic conservation, the distance from the TSS, the proportion of conserved predicted TFBSs and the information content (IC) of the predicting PWM. This likelihood score for promoter regulation is calculated for each PWM from both TransFac and JASPAR (CORE and phyloFACTS). For every promoter sequence, the top

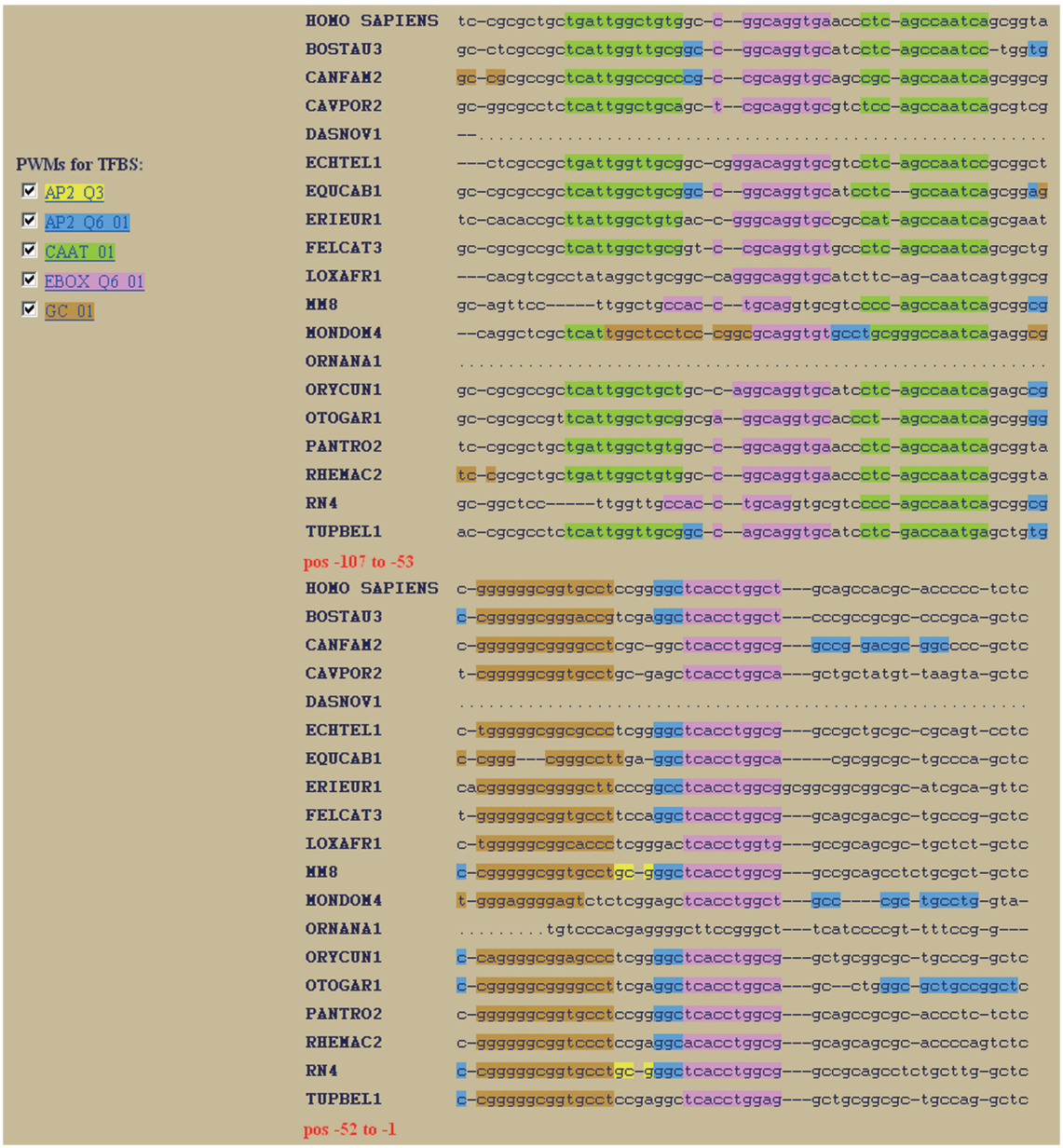

Figure 1. Visualization of the predicted TFBSs for TFs AP-2, CCAAT box, E-BOX and GC box in the multiz 28-way alignment of the promoter of the E-cadherin transcript NM_004360. The results are exactly as described by Comijn *et al.* (20).

100 best ranked PWMs are given, a selection of which can be directly forwarded to the visualization part. Predicting which TFs regulate the gene of interest is an extremely difficult task. The exploration part is mainly intended to give an idea of which TFs are more likely to bind to the promoter and thus to indicate the PWMs for which visualization of predicted TFBSs could be interesting. In the supplemental data document nr 1 we show that the exploration results seem to be biologically meaningful. We start with the extensively described promoter of the IL2 gene, encoding interleukin-2. Most experimentally defined TFBSs described in the literature are ranked at the top of the full list delivered by the ConTra exploration. The second example uses the promoter of MX1 (myxovirus resistance 1), which has two interferon-stimulated response element (ISRE) sites known to be crucial for its expression. The PWMs corresponding to TFs that bind to ISRE sites appear in the top of the resulting list. The third example considers the exploration of the promoter of the DICER1 gene, for which, as far as we know, no transcription regulation experiments have been described in the literature. Those results are intriguing in that they might be correlated with recent findings showing that miRNAs involved in cancer are regulated by TFs already known to play a role in cancerous processes (21).

Several other web tools provide information about TFBSs predicted by PWMs (or HMMs) in the context of

(multiple) promoter alignments. The supplemental data document nr 2 lists those web tools with their features. We think ConTra competes well with the other tools in this list as it is a compact and user-friendly web tool that provides the biologist at the bench with useful visualization of predicted TFBSs in a cross-species alignment context. The alignments are automatically fetched and contain up to 28 species. ConTra works per alternative promoter and is flexible with respect to promoter length, alignment type and PWM prediction stringency. Also important are the up-to-date PWM libraries of TransFac and JASPAR.

IMPLEMENTATION

Making input user-friendly was accomplished by the integration of resources from HGNC (22), UCSC and Ensembl. The alignment retrieval feature was implemented by perl scripts using data from the 'golden path' of UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/) and the program axtAndBed from the UCSC genome browser source code, or by perl scripts using the Ensembl Compara perl API.

The PWM libraries used by ConTra contain 101 'vertebrate' matrices from the latest JASPAR CORE database, 174 matrices from JASPAR phyloFACTS database and a nonredundant selection of 214 matrices from one of the latest TransFac database versions (11.4).

Jalview (19) is used to create an overview picture of each promoter alignment, whereas the dynamic view of predicted TFBSs in the HTML-embedded promoter alignments is accomplished by Javascript changing CSS properties.

The likelihood score for promoter regulation of each PWM in the exploration part is obtained by an accumulation of the weights of its predicted TFBSs on the reference sequence. The weight of a predicted TFBS depends mainly on the extent of phylogenetic conservation, which is determined by the number of species with a predicted TFBS for the same PWM at about the same position and by the conservation extent of that position. This simply represents the basic concept behind phylogenetic footprinting, i.e. cross-species conserved TFBSs are more likely to be functional compared to nonconserved ones. We do not require that the TFBS is conserved at exactly the same place. The score even rises if TFBSs predicted by the same PWM are near each other, because of the frequently observed presence of homotypic clusters of functional sites and weak 'shadow' sites around them (23). Another factor influencing the weight of a predicted TFBS is the distance to the TSS. This is supported by findings of ref. (16), which prove that functional TFBSs are mainly situated in the first 200 nt upstream of the TSS. Continuous high ranking of PWMs with a rather bad quality, i.e. predicting many false positives, is avoided by having the IC of the predicting PWM influence the weight of each predicted TFBS. For the same reason, the accumulated amount of weights is divided by a factor proportional to the number of nonconserved predicted TFBSs. The scoring formula is given as pseudocode in the supplemental data document nr 3.

ACKNOWLEDGEMENTS

This research was supported by VIB. We acknowledge Dr Amin Bredan for English proofreading of the manuscript and Dr Karl Vandepoele for testing feedback. Funding to pay the Open Access publication charges for this article was provided by Ghent University and VIB.

Conflict of interest statement. None declared.

REFERENCES

1. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
2. Bryne,J.C., Valen,E., Tang,M.H., Marstrand,T., Winther,O., da Piedade,I., Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.
3. Messina,D.N., Glasscock,J., Gish,W. and Lovett,M. (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.*, **14**, 2041–2047.
4. Tomovic,A. and Oakeley,E.J. (2007) Position dependencies in transcription factor binding sites. *Bioinformatics*, **23**, 933–941.
5. Fang,F. and Blanchette,M. (2006) FootPrinter3: phylogenetic footprinting in partially alignable sequences. *Nucleic Acids Res.*, **34**, W617–W620.
6. Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.
7. Irvin-Wilson,C.V. and Chaudhuri,G. (2005) Alternative initiation and splicing in dicer gene expression in human breast cells. *Breast Cancer Res.*, **7**, R563–R569.
8. Matsuda,S., Ichigotani,Y., Okuda,T., Irimura,T., Nakatsugawa,S. and Hamaguchi,M. (2000) Molecular cloning and characterization of a novel human gene (HERNA) which encodes a putative RNA-helicase. *Biochim. Biophys. Acta*, **1490**, 163–169.
9. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
10. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
11. Miller,W., Rosenbloom,K., Hardison,R.C., Hou,M., Taylor,J., Raney,B., Burhans,R., King,D.C., Baertsch,R., Blankenberg,D. *et al.* (2007) 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.*, **17**, 1797–1808.
12. Margulies,E.H., Cooper,G.M., Asimenos,G., Thomas,D.J., Dewey,C.N., Siepel,A., Birney,E., Keefe,D., Schwartz,A.S., Hou,M. *et al.* (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.*, **17**, 760–774.
13. Kent,W.J., Baertsch,R., Hinrichs,A., Miller,W. and Haussler,D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA*, **100**, 11484–11489.
14. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
15. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.
16. Tabach,Y., Brosh,R., Buganim,Y., Reiner,A., Zuk,O., Yitzhaky,A., Koudritsky,M., Rotter,V. and Domany,E. (2007) Wide-scale analysis of human functional transcription factor binding

reveals a strong bias towards the transcription start site. *PLoS ONE*, **2**, e807.
17. Griffith,O.L., Montgomery,S.B., Bernier,B., Chu,B., Kasaian,K., Aerts,S., Mahony,S., Sleumer,M.C., Bilenky,M., Haeussler,M. *et al.* (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.*, **36**, D107–D113.
18. Portales-Casamar,E., Kirov,S., Lim,J., Lithwick,S., Swanson,M.I., Ticoll,A., Snoddy,J. and Wasserman,W.W. (2007) PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biol.*, **8**, R207.
19. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
20. Comijn,J., Berx,G., Vermassen,P., Verschueren,K., van Grunsven,L., Bruyneel,E., Mareel,M., Huylebroeck,D. and van Roy,F. (2001) The two-handed E box binding zinc finger protein SIP1 downregulates E-cadherin and induces invasion. *Mol. Cell*, **7**, 1267–1278.
21. Corney,D.C., Flesken-Nikitin,A., Godwin,A.K., Wang,W. and Nikitin,A.Y. (2007) MicroRNA-34b and MicroRNA-34c are targets of p53 and cooperate in control of cell proliferation and adhesion-independent growth. *Cancer Res.*, **67**, 8433–8438.
22. Eyre,T.A., Ducluzeau,F., Sneddon,T.P., Povey,S., Bruford,E.A. and Lush,M.J. (2006) The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res.*, **34**, D319–D321.
23. Papatsenko,D.A., Makeev,V.J., Lifanov,A.P., Regnier,M., Nazina,A.G. and Desplan,C. (2002) Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers. *Genome Res.*, **12**, 470–481.