Gene expression

TFBScluster: a resource for the characterization of transcriptional regulatory networks

Ian John Donaldson, Michael Chapman and Berthold Göttgens*

Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Hills Road, Cambridge CB2 2XY, UK

Received on January 20, 2005; revised on March 18, 2005; accepted on April 21, 2005
Advance Access publication April 26, 2005

ABSTRACT

Summary: One major challenge of the post-sequencing era of the human genome project will be the functional annotation of the non-coding portion of the genome, in particular gene regulatory sequences. We have developed a new web-based tool, TFBScluster, which performs genome-wide identification of transcription factor binding site clusters that are conserved in multiple mammalian genomes. Clusters representing candidate gene regulatory elements can be filtered further, based on the presence or absence of additional user-defined DNA sequence motifs or by constraining the orientation or order of binding sites. Comprehensive results files, returned by email, are designed to facilitate experimental validation of computationally identified candidate gene regulatory sequences. TFBScluster, therefore, has the potential to contribute to deciphering transcriptional networks that regulate a wide range of mammalian developmental processes.

Availability: http://hscl.cimr.cam.ac.uk/TFBScluster_genome_34.html
Contact: bg200@cam.ac.uk
Supplementary information: http://hscl.cimr.cam.ac.uk/sup_don05_app_api.html

A key goal of the post-genome sequencing era will be the identification of cis-regulatory sequences to reconstruct the transcriptional networks that lie at the heart of normal development and are frequently perturbed in human disease processes. Since transcription factor binding sites (TFBSs) are often short (4–6 nt) and degenerate, genome-wide computational identification of TFBSs must be able to discriminate functional (true positive) from the vast numbers of non-functional (false positive) sites. The proportion of functional TFBSs in such analyses may be increased by (1) concentrating on TFBSs conserved between species (Lenhard et al., 2003), (2) devising algorithms capable of predicting the likelihood that a given region of the genome is involved in regulating gene expression (Kolbe et al., 2004), and (3) exploiting the observation that most functional TFBSs form part of clusters of multiple sites (Krivan and Wasserman, 2001; Markstein et al., 2002; Wasserman and Fickett, 1998). Taking advantage of all three approaches, we have recently demonstrated for the first time that human gene regulatory sequences with predicted in vivo biological activity (when assayed under the most rigorous conditions in transgenic mice) can be identified through whole-genome computational analysis (Donaldson et al., 2005).

We have now built on this approach and developed a web-integrated suite of bioinformatics tools (TFBScluster) designed for the identification of candidate human gene regulatory sequences with predicted biological activity.

Briefly, TFBScluster utilizes genome-wide libraries of human TFBSs that are conserved in a range of other mammalian species. Binding site libraries currently implemented include libraries for 35 IUPAC consensus strings (conserved between human/mouse and human/mouse/rat) and 163 position weight matrices (conserved between human/mouse/rat). These libraries are used to identify TFBSs located in user-defined TFBS clusters. TFBScluster, therefore, significantly differs from existing tools that are specifically designed for the identification of TFBS clusters in lower model eukaryotes, such as Drosophila (Berman et al., 2002; Grad et al., 2004; Markstein et al., 2002; Sosinsky et al., 2003), limited to identifying TFBS clusters in relatively small sets of aligned sequences (Fridh et al., 2003; Johansson et al., 2003; Loots and Ovcharenko, 2004; Yu et al., 2004) or restricted to sequences flanking predicted transcriptional start sites (Dieterich et al., 2005; Jegga et al., 2002). Moreover, to increase the proportion of functional binding sites, TFBScluster search results can be further screened based on order, spacing and orientation of binding sites as well as presence or absence of user-defined motifs.

The genome-wide positions of 35 IUPAC consensus strings were identified using a revised version of our PERL program TFBSsearch.pl (Chapman et al., 2004) in human/mouse and human/mouse/rat aligned genomes available from Genome Bioinformatics at the UCSC (http://genome.ucsc.edu/downloads.html). Background and relevant references for the consensus binding site sequences used can be found at http://hscl.cimr.cam.ac.uk/TFBScluster_genome_34_background_iupac.html. For each consensus string, five library files were prepared using increasing levels of sequence conservation. The first file contains all ‘non-exact’ core positions where degenerate IUPAC positions may differ between aligned sequences. The second file contains ‘exact’ positions of only those sites with identical sequence across the alignment. In the remaining three files, the IUPAC code ‘N’ (any nucleotide) is added to the start and end of the core search string, thereby extending the conserved site beyond the original core string to ensure that the core binding sites are situated within extended blocks of high sequence conservation. Search strings were used with one, two and three conserved flanking nucleotides to identify TFBSs located in blocks of increasing sequence conservation. The genome-wide positions of 163 position weight matrices.
Identification of conserved TFBS clusters

conserved in human/mouse/rat whole-genome alignments were obtained from the UCSC genome browser (http://genome.ucsc.edu/). Position weight matrices used are from the Transfac database v4.0 (Heinemeyer et al., 1999). A list of 163 matrices is provided as background information (http://bscl.cimr.cam.ac.uk/TFBScluster_genome_34_background_pwm.html) and includes hyperlinks to database records for the individual matrices maintained on the TESS website (http://www.chib.upenn.edu/tes). To increase the likelihood that TFBSs represent functional sites, TFBScluster can be restricted to only analyse sites located in regions of regulatory potential (Kolbe et al., 2004) with scores ≥0.0002, based on the threshold suggested by UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg16/regPotential/).

Using the web interface of TFBScluster, TFBS libraries are screened for clusters of TFBSs in the human genome that contain a set number of sites in a defined window of sequence. In TFBS-cluster, cluster sizes are defined by the start of the first (left-most) TFBS and the end of the last (right-most) TFBS, and overlapping clusters are merged, with the result that some clusters may be larger than the user-defined window size. By default there is no restriction to the distance between individual TFBSs. However, since some binding sites may overlap, the web interface includes an option that only counts the number of non-overlapping TFBSs when determining whether the minimum number of each TFBS for the user-defined cluster has been met. This ensures that the minimum set of TFBSs is free to bind proteins simultaneously.

The user may choose to receive a simple list of candidate clusters in the Sanger GFF format (http://www.sanger.ac.uk/Software/formats/GFF/) with the start and end chromosomal coordinates of each cluster relative to the human genome (‘short’ analysis). This output is directly portable into many downstream sequence analysis programs and also easily displayed in genome browsers. The TFBScluster interface also provides a more comprehensive analysis of candidate TFBS clusters (‘long’ analysis).

The results files for ‘long’ analysis include information for the genes that have been localised to such clusters of TFBSs, and are therefore, potentially regulated by such clusters. Comprehensive information about potentially regulated genes is retrieved via the Ensembl API (Stabenau et al., 2004), and web-links to the UCSC genome browser are provided (see http://bscl.cimr.cam.ac.uk/sup_don05_app_api/TFBScluster_ex_file.html for examples of result files). Consequently, lists of candidate TFBS clusters can be converted into lists of ‘localised’ genes with the relevant gene identifiers (Swiss-Prot/Locuslink accessions) to perform downstream analysis based, for example, on gene function (Gene Ontology). Moreover, ‘long’ analysis generates comprehensive information on the local sequence surrounding each cluster. An alignment is stored for the predicted cluster sequence as well as an extended sequence containing 350 nt on either side of the cluster. The TFBScluster web interface provides the option to search these alignments in order to retain or exclude cluster candidates, based on the presence of a user-supplied DNA sequence motif (any user-defined IUPAC string) or a pattern of motifs.

TFBScluster.pl and the associated programs are written in PERL and are accessible via a PERL CGI interface on a web server, hosted by the University of Cambridge. The user HTML input pages contain default parameters where appropriate. Through its ease of use and streamlined provision of information required for functional validation of candidate gene regulatory regions, TFBScluster has the potential to contribute to deciphering the transcriptional networks controlling a wide range of human developmental processes. Finally, TFBScluster can be easily adapted to analyse additional genomes as long as they have been annotated in Ensembl. The development of a mouse-centric resource is currently underway.

ACKNOWLEDGEMENTS

We are grateful for many stimulating discussions with Tony Green and Josette-Renee Landry (Cambridge University) and Noel Buckley, Alex Bruce and Michael Sadowski (Leeds University). Work in the authors’ laboratory is funded by the Leukaemia Research Fund, BBSRC, Cambridge MIT Institute and an IBM SUR equipment grant.

REFERENCES


