BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 23 2005, pages 4304–4306
doi:10.1093/bioinformatics/bti691

Sequence analysis

CREDO: a web-based tool for computational detection of conserved sequence motifs in noncoding sequences

Tobias Hindemitt and Klaus F. X. Mayer*

MIPS/Institute for Bioinformatics, GSF Research Centre for Environment and Health, Ingolstaedter Landstrasse 1, 85758 Neuherberg, Germany

Received on May 13, 2005; revised on September 19, 2005; accepted on September 23, 2005
Advance Access publication October 4, 2005

ABSTRACT

Summary: CREDO is a user-friendly, web-based tool that integrates the analysis and results of different algorithms widely used for the computational detection of conserved sequence motifs in noncoding sequences. It enables easy comparison of the individual results. CREDO offers intuitive interfaces for easy and rapid configuration of the applied algorithms and convenient views on the results in graphical and tabular formats.

Availability: http://mips.gsf.de/proj/regulomips/credo.htm

Contact: kmayer@gsf.de

Supplementary information: A detailed help file for CREDO is available on http://mips.gsf.de/proj/regulomips/help.htm. Further supplementary material is available on *Bioinformatics* online.

In higher eukaryotes gene expression is regulated under a variety of constraints such as tissue specificity, developmental or environmental conditions. Understanding the mechanisms that orchestrate this tight regulation is a major challenge in modern biology. RNA-polymerase II-mediated transcription is activated or repressed by transcription factors (TFs). Transcription factor binding sites, also called *cis*-regulatory elements (CREs), constitute a gene's regulatory regions, in particular its promoter. CREs are typically short (6–12 bp) and often degenerate in sequence. Experimental detection and characterization of CREs are feasible but time-consuming and only few examples for experimental CRE detection on genome level have been reported (Lee *et al*., 2002).

Computational methods are powerful, cost effective and can support experimental approaches to detect CREs. The latest approaches are based on cross-species comparison of orthologous sequences (i.e. phylogenetic footprinting) (Gumucio *et al*., 1992; Duret and Bucher, 1997) and the comparative analysis of noncoding sequences from co-expressed genes (van Helden *et al*., 1998; Hughes *et al*., 2000). These approaches are based on the assumption that functional elements are conserved, which allows distinguishing them from non-functional regions in their vicinity. A number of computational tools have been developed to make use of this opportunity. Global alignment tools (e.g. DIALIGN; Morgenstern, 1999) represent the most widely used approach for phylogenetic footprinting. In addition, several motif detection procedures which do not rely on co-linearity have been published. They have either been specifically developed for phylogenetic footprinting (e.g. FootPrinter; Blanchette and Tompa, 2003) more generally for the detection of conserved motifs in the upstream regions of functionally related or co-expressed genes (e.g. AlignACE and MotifSampler; Hughes *et al*., 2000; Thijs *et al*., 2001) or for the identification of sequence conservation in biopolymers (e.g. MEME; Bailey and Elkan, 1994).

It has been suggested to compare and integrate results derived from different methods and to use complementary tools in combination rather than rely on a single method (Tompa *et al*., 2005). Such an approach is time-consuming and difficult owing to varying output formats and different graphical representation of the respective results. CREDO (*Cis*-Regulatory Element Detection Online) integrates, combines and visualizes the analyses of AlignACE, DIALIGN, FootPrinter, MEME and MotifSampler and therefore facilitates the comparison of their results. In contrast to AlignACE, MEME and MotifSampler—which are designed for a more general detection of conserved sequence motifs—FootPrinter is focused on the detection of conserved sequence motifs in noncoding sequences of orthologous genes and should not be used for applications like the identification of sequence conservation in the upstream regions of co-expressed genes. CREDO enables to run each of the algorithms simultaneously on a given dataset and summarizes the outputs of all programs graphically, in tables and within an XML file. In order to ensure complete platform independence and to avoid installation of additional software on the user side, CREDO interfaces are exclusively based on standard web technologies. Details on the integration of the different algorithms are provided as Supplementary material.

Input sequences can be pasted into the CREDO web form. Regions of special interest (e.g. known regulatory elements) can be indicated by capital letters. These regions will be highlighted in the graphical output (Fig. 1A).

Almost all parameters of the algorithms applied can be adapted. The CREDO web interface provides a structured parameter selection form. Parameters are subdivided into two groups: basic parameters (e.g. motif size or motif number) and advanced parameters. By default advanced parameters are hidden and preset values are being used. For expert users the opportunity to change these parameters and refine the analysis is provided. As a starting point, CREDO provides three different and widely applied presettings. The first presetting has been designed for users who aim to carry out phylogenetic footprinting with closely related species. The second presetting has been designed for phylogenetic footprinting with more distantly related species and finally the third presetting for users who set out to search for conserved sequence motifs in co-expressed

*To whom correspondence should be addressed.

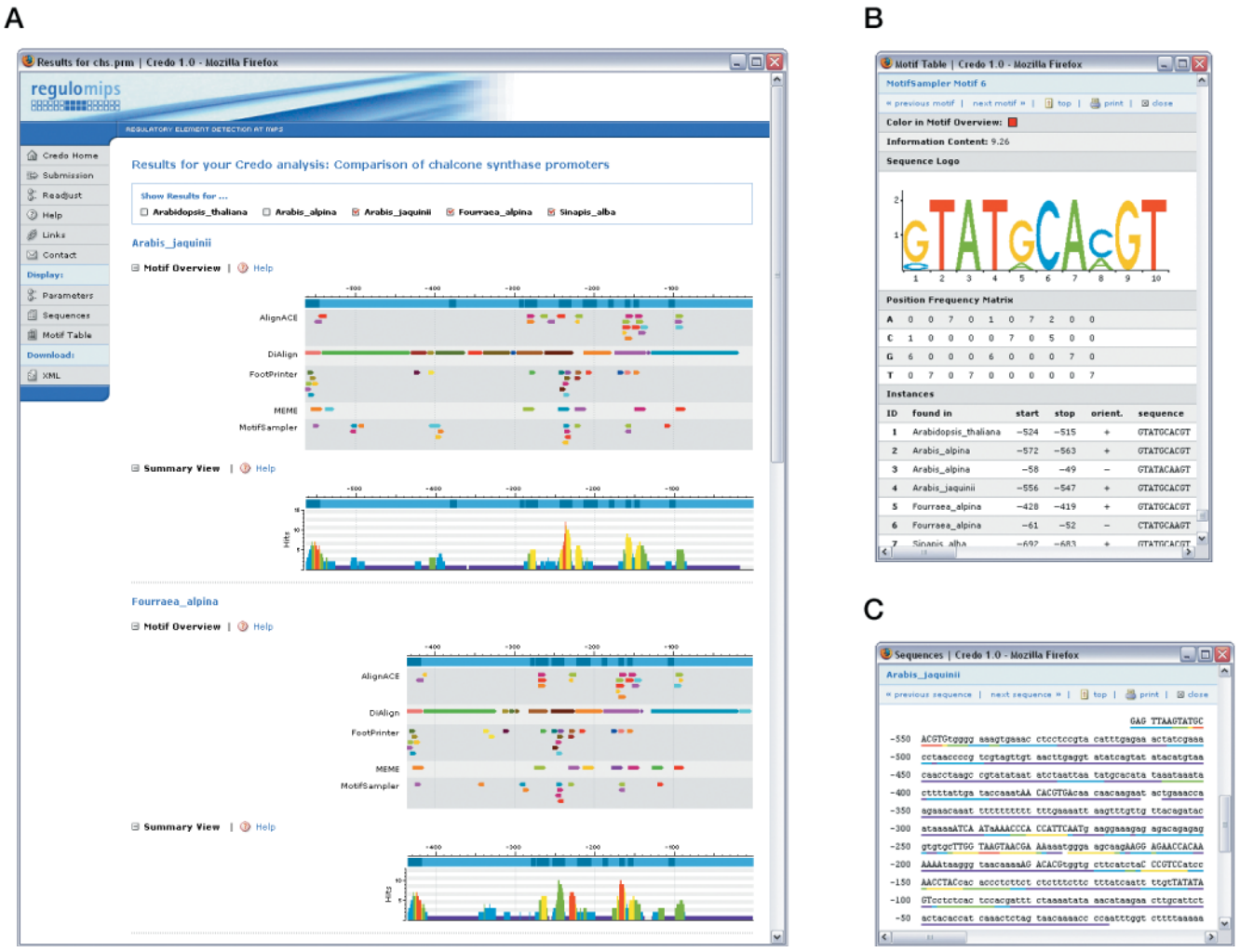

Fig. 1. CREDO result page. (**A**) Input sequences are represented as blue bars. Sequence regions marked within the sequence input are highlighted in darker blue. Within the motif overview occurrences of predicted motifs are depicted as coloured arrows and are linked to the respective motif data (see B). The summary view graphically summarizes the motifs found by all programs for each base pair. The number of total motif hits is indicated by the height of the bar and the number of different programs that report a motif is colour coded (purple, one; light blue, two; green, three; yellow, four; red, five). The panel can be hidden by clicking on the '−' symbol. (**B**) The motif table provides all relevant data and a sequence logo for each predicted motif. (**C**) Input sequences are displayed. Positions hit by one or several motif predictions are underlined. Colours correspond to those within the summary view (see A).

genes is provided. It is important to emphasize that these presettings provide only a starting point for the analysis. The parameter selection should be refined in subsequent analyse since optimized parametrization is a prerequisite for significant results.

After completion of the analysis the user is notified by e-mail and a graphical overview of the results is made available (Fig. 1A). For each input sequence the motifs detected by the individual algorithms along with a summary view are displayed. This view summarizes the motifs found by all programs and hence facilitates the identification of sequence regions where results of the different algorithms are coincident.

The graphical representations of motif occurrences are linked to underlying analytical data. The result pages include links to three pop-up windows that contain the table of found motif data, input sequences and chosen parameters, respectively. The motif table (Fig. 1B) provides all important motif data and includes a sequence logo (Schneider and Stephens, 1990; Lenhard and Wasserman, 2002)—a highly intuitive graphical representation—for each motif detected. To display motifs in their sequence environment, the respective positions are underlined in the pop-up window depicting the input sequences (Fig. 1C).

All relevant data, including input sequences and parameters used, as well as the analytical results can be downloaded as XML file. An example that illustrates the effectiveness of CREDO is provided within the Supplementary material and in more detail on the CREDO homepage (http://mips.gsf.de/proj/regulomips/credo.htm).

ACKNOWLEDGEMENTS

The authors would like to acknowledge Daniela Fölsl, Stefanie Maisel and Heiko Schoof for help during the web server implementation of CREDO and Thomas Rattei for providing additional

infrastructure. The authors would also like to thank Claudia Englbrecht for critical reading of the manuscript. This work has in part been founded by the GABI programme of German Ministry of Education and Research (BMBF).

Conflict of Interest: none declared.

REFERENCES

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Blanchette,M. and Tompa,M. (2003) FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res.*, **31**, 3840–3842.

Duret,L. and Bucher,P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, **7**, 399–406.

Gumucio,D.L. *et al.* (1992) Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Mol. Cell. Biol.*, **12**, 4919–4929.

Hughes,J.D. *et al.* (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.

Lee,T.I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.

Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Thijs,G. *et al.* (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113–1122.

Tompa,M. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

van Helden,J. *et al.* (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.