W578–W583 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl193

PromAn: an integrated knowledge-based web server dedicated to promoter analysis

Aurélie Lardenois, Frédéric Chalmel[1], Laurent Bianchetti[2], José-Alain Sahel[3], Thierry Léveillard[3] and Olivier Poch*

Laboratoire de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP BP 163, 67404 Illkirch Cedex, France, [1]Biozentrum and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, CH-4056 Basel, Switzerland, [2]Plate-Forme de Bioinformatique de Strasbourg, Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP BP 163, 67404 Illkirch Cedex, France and [3]Laboratoire de Physiopathologie Cellulaire et Moléculaire de la Rétine, Inserm U592, Université Pierre et Marie Curie, 75571 Paris, France

Received February 14, 2006; Revised and Accepted March 21, 2006

ABSTRACT

PromAn is a modular web-based tool dedicated to promoter analysis that integrates distinct complementary databases, methods and programs. PromAn provides automatic analysis of a genomic region with minimal prior knowledge of the genomic sequence. Prediction programs and experimental databases are combined to locate the transcription start site (TSS) and the promoter region within a large genomic input sequence. Transcription factor binding sites (TFBSs) can be predicted using several public databases and user-defined motifs. Also, a phylogenetic footprinting strategy, combining multiple alignment of large genomic sequences and assignment of various scores reflecting the evolutionary selection pressure, allows for evaluation and ranking of TFBS predictions. PromAn results can be displayed in an interactive graphical user interface, PromAnGUI. It integrates all of this information to highlight active promoter regions, to identify among the huge number of TFBS predictions those which are the most likely to be potentially functional and to facilitate user refined analysis. Such an integrative approach is essential in the face of a growing number of tools dedicated to promoter analysis in order to propose hypotheses to direct further experimental validations. PromAn is publicly available at http://bips.u-strasbg.fr/PromAn.

INTRODUCTION

The functional genomics revolution has given rise to a huge number of transcriptomics studies. This, combined with the availability of numerous eukaryotic genome sequences, has lead to the current challenge of decoding the regulatory networks underlying gene expression. As a consequence, a rapid increase in the number of available databases, methods and programs dedicated to promoter analysis has emerged during the past few years. Transcription factor binding sites (TFBSs) are small, degenerative sequences, so a major problem in promoter analysis is the selection of the correct TFBS predictions in the resulting low signal-to-noise ratio environment. Cross-species comparison has been commonly used to filter TFBS predictions and to identify potentially active regulatory elements, as in ConSite (1) and rVista 2.0 (2). This approach is based on the assumption that gene regulatory regions and elements are often preferentially conserved during evolution but also suggests that selective pressure on orthologous genes must be similar in each respective organism. This phylogenetic footprinting (3,4) method is implemented in almost all web servers implicated in gene regulation studies, e.g. CONREAL (5) and Footer (6). However, the tools differ in their choice of TFBS databases, prediction programs and methods, as well as the genomic sequence alignment and statistical scoring methods implemented to estimate the TFBS conservation pressure.

In this context, integrative approaches become essential for in-depth promoter analysis. We thus developed PromAn, a web server aimed at integrating different publicly available databases, programs and methods dedicated to promoter analysis. PromAn integrates transcription start site (TSS) and TFBS databases, prediction programs, a phylogenetic

*To whom correspondence should be adressed. Tel: +33 388653294; Fax: +33 388653201; Email: poch@igbmc.u-strasbg.fr

footprinting approach, several statistical scoring methods as well as an interactive graphical user interface, PromAnGUI. PromAnGUI integrates all of this information and helps the biologist to refine results and further guide gene regulation hypotheses in parallel with experimental data and validations.

METHODS

Input genomic sequences

The PromAn web server requires as input a single genomic sequence that will be taken as the reference in subsequent analysis and a set of orthologous sequences in fasta format (Figure 1). PromAn provides the possibility of inputting large genomic sequences, which allows the user to begin analysis with minimal prior knowledge of proximal and distal active promoter regions. To limit the processing required and to allow for variation on the TSS position, the user can input genomic sequences up to 20 kb. According to the NCBI statistics of the human genome (http://www.ncbi.nih.gov/mapview/stats/BuildStats.cgi?taxid=9606&build=35), the average size of exons and introns are 231 and 5407 bp, respectively. Thus, large sequences up to 20 kb can allow for 5′ non-coding exons, TSS mis-location and coding exons anchoring the multiple alignment.

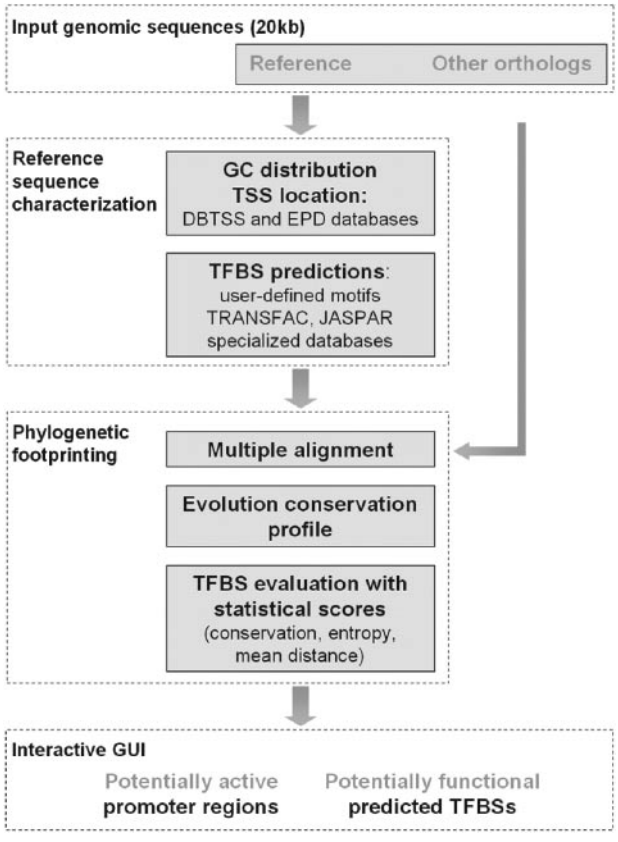

Figure 1. Flowchart of the PromAn integrated strategy. A single reference sequence and a set of orthologous genomic sequences are required as input. First, dinucleotide distribution, TSS and promoter location as well as TFBS predictions characterize the reference sequence. Next, a phylogenetic footprinting approach is used to determine the evolutionary conservation profile and TFBS evaluation with statistical scores. This integrated strategy is used to validate the TSS location, highlighting potentially active promoter regions and potentially functional TFBSs through the PromAn graphical user interface (GUI).

Reference sequence characterization

In order to locate the promoter region on the reference genomic sequence and validate the TSS location, PromAn integrates experimentally-based databases such as DBTSS (Database of Transcriptional Start Sites) (7) and EPD (Eukaryotic Promoter Database) (8) as well as promoter, first exon or exonic map prediction programs, such as First EF (9), Eponine (10) and GenScan (11). TSS location validation is an essential preliminary step in promoter analysis. PromAn determines the nucleotide distribution of the reference sequence to detect the presence of potential GC-rich regions within the given promoter of interest. On the reference sequence, PromAn also looks for TFBS predictions based on the following nucleotide matrices: public TRANSFAC (12) and JASPAR (13) databases, user-defined motifs and databases dedicated to a specific biological problems, such as retina or nuclear receptor databases. The Position Weight Matrices (PWM) scoring method (14,15) has been implemented to locate and score TFBS predictions on the reference promoter region of interest. The profiles are first converted to log-scale PWM to evaluate candidate TFBS predictions. A normalized matrix score, S, is assigned to each prediction where:

$$S = 100 \times \frac{\text{score} - \text{score}_{\text{min}}}{\text{score}_{\text{max}} - \text{score}_{\text{min}}}$$

A threshold is then applied to the scores to define the predictions that are considered to be candidate TFBS predictions.

Phylogenetic footprinting

The TFBSs are small (6–20 bp) and degenerate sequences. Their inherent properties imply that there can at least be one TFBS prediction per genomic sequence base pair. Given the huge number of predictions, >90% are usually false positives. PromAn uses the multiz (16) multiple sequence alignment program to implement a phylogenetic footprinting method to take into account evolutionary selection pressure. Orthologous sequence alignment allows highlighting of regions conserved during evolution. Several studies have demonstrated that regulatory modules are under positive selection pressure, therefore regions of high conservation should correspond to potentially active promoter. Multiple sequence alignment also provides the basis for statistical scores estimating the significance of predicted TFBSs. As the number of orthologous input sequences is not limited in PromAn, use of at least three sequences provides a more precise multiple alignment that allows for more accurate statistical scores reflecting the conservation of TFBSs during evolution. The evolutionary distances between organisms should be taken into account, as sequence divergence between closely related organisms, such as human and chimpanzee or rat and mouse is usually insufficient to provide relevant evolutionary conservation information (1). PromAn implements three complementary statistical scores. The conservation score measures the identity of the orthologous sequences with respect to the reference sequence for a given region. It corresponds to the average percentage of nucleotides identical to the reference

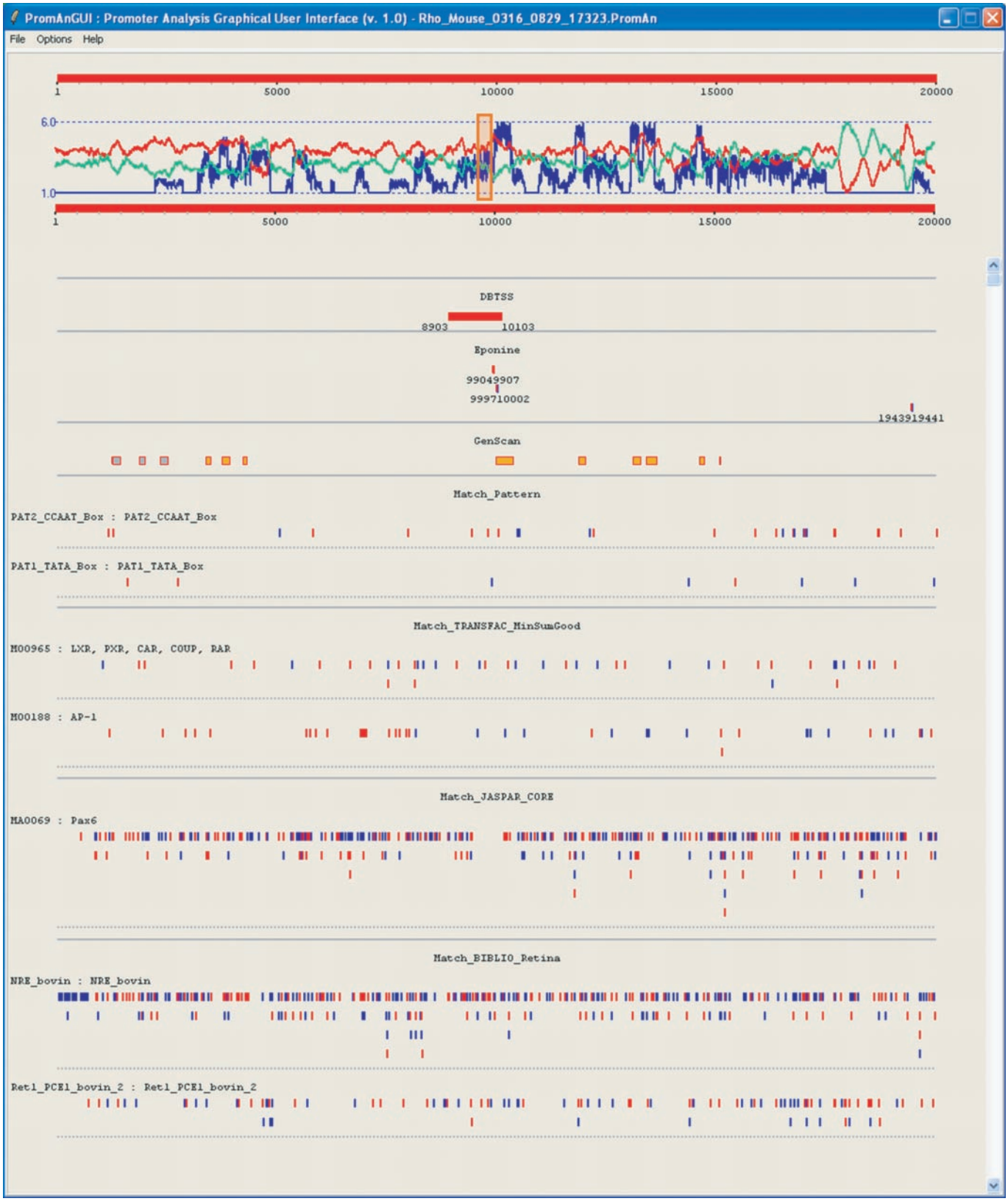

Figure 2. Example of a PromAn output. The upper frame displays the mouse rhodopsin genomic region (extracted from ±10 kb with respect to the start codon—red boxes), surrounding dinucleotide (AT in green and GC in red) and conservation (blue) profiles. The orange rectangle highlights the RPPR that is described in Figure 3. The lower frames depict the DBTSS, Eponine, GenScan, user-defined motifs (Match_Pattern), TRANSFAC (Match_TRANSFAC_MinSumGood), JASPAR (Match_JASPAR_CORE) and retina dedicated (Match_BIBLIO_Retina) predictions. Each prediction is displayed as a colored [gradient from low (grey) to high (red) Mean Distance score] box where the outline indicates the strand (blue for minus and red for plus).

sequence at a given position in the multiple sequence alignment. The entropy score is based on the ScoreCons program (17). It gives the degree of nucleic acid variability to quantify residue conservation in a multiple alignment. The mean distance score is based on the ClustalX conservation profile (18) this means on the mean pairwise distance between sequences in a continuous sequence space. The combined scores allow for an evaluation and ranking of TFBS predictions in order to estimate their biological relevance.

PromAn results display and analysis

PromAn results are sent to the user by e-mail. PromAnGUI gives the user the possibility to visualize and refine results as often as needed in parallel with expert biological knowledge and experimental validations, which are indispensable to complete and further guide gene regulation hypotheses. A help section relative to the PromAnGUI graphical user interface is available on the PromAn home page.

As an example, we consider the analysis of the mouse rhodopsin promoter region. *cis*-regulatory elements of the bovin rhodopsin proximal promoter region (RPPR) have been characterized in several studies (19,20). PromAn allows us to determine whether these elements are retrieved in the mouse. In other words, are these biologically active regions conserved during evolution? Input genomic sequences have been extracted from the HomoloGene Downloader tool (http://www.ncbi.nlm.nih.gov/HomoloGene) and the UCSC Genome Browser web server (http://genome.ucsc.edu). Mouse and a set of five orthologous sequences (human, cow, dog, chicken and *Xenopus tropicalis*) extracted from −10 kb to +10 kb with respect to the start codon have been used as input in this PromAn analysis.

Figure 2 illustrates the display of this analysis in the PromAnGUI. Results are always located with respect to the reference sequence (mouse), which is depicted as two red boxes surrounding different profiles in the upper frame. Both AT (green) and GC (red) profiles can be depicted to easily and immediately identify regions enriched in these dinucleotides. The DBTSS database allows us to locate an experimentally validated TSS at the 9903th bp of the mouse reference sequence. The dinucleotide profiles in Figure 2 show that the RPPR contains a GC-rich region around the TSS. The blue curve represents the conservation profile of

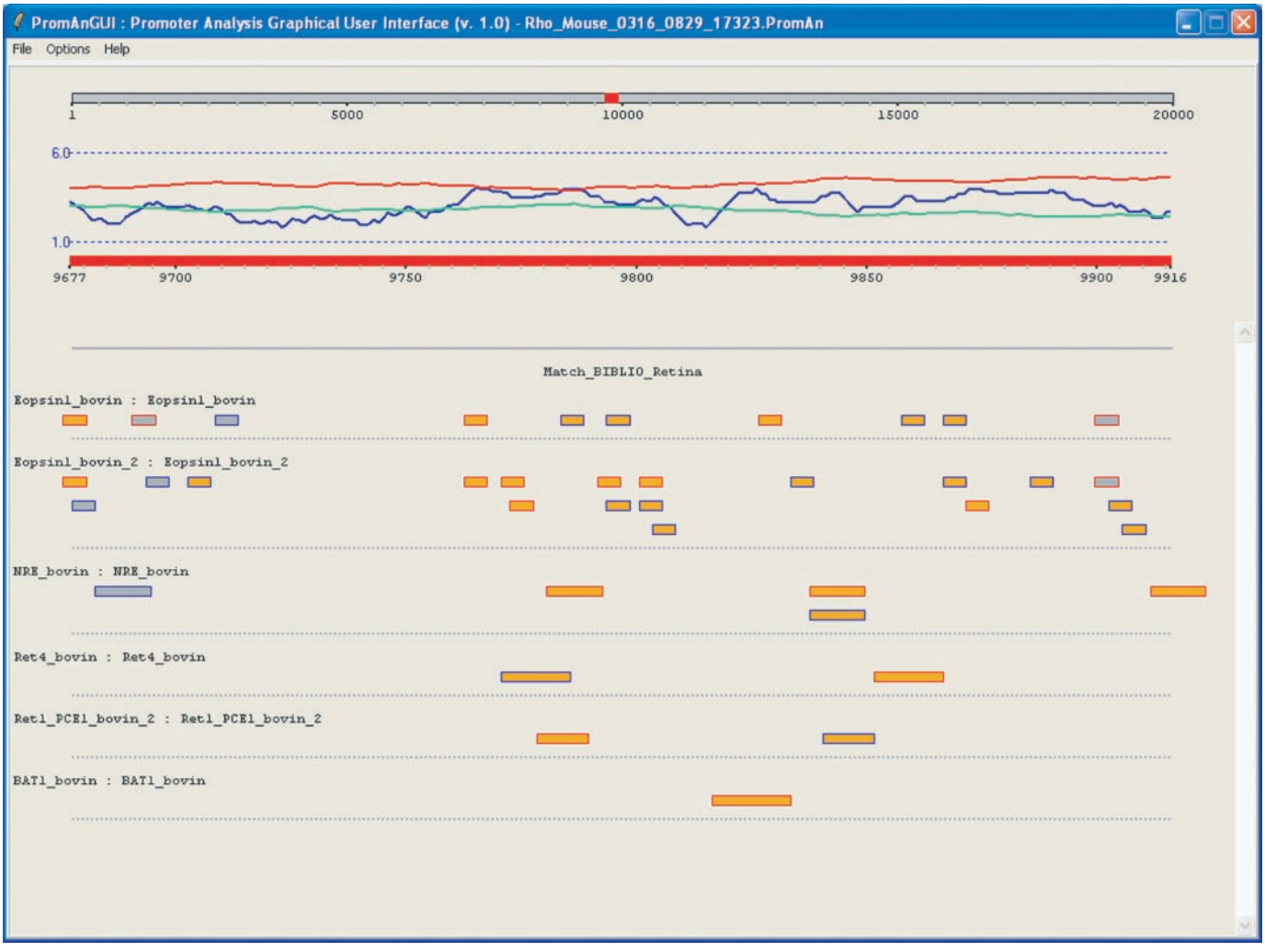

Figure 3. TFBS analysis of the RPPR responsible for photoreceptor specificity. User-defined motifs, TRANSFAC, JASPAR and retina specific predictions are depicted. They highlight the conservation of the Ret-1/PCE-1, BAT-1, NRE and Ret-4 elements among mammals.

the reference sequence based on a multiple alignment of the orthologous genomic sequences. The higher peaks correspond to coding exons present in the mouse sequence and conserved in the five other orthologs. The DBTSS, Eponine, GenScan and TFBS (user-defined TATA and CCAAT boxes, TRANSFAC, JASPAR and retina dedicated databases) predictions are depicted in the lower frames by boxes colored according to the mean distance [gradient from low (grey) to high (red)] score. The conservation profile shows a proximal region of about 1 kb long conserved in all four mammals. Within this region, a small region (from 9677 to 9916—depicted in orange) has been described as being responsible for the photoreceptor specificity and is named the RPPR.

Figure 3 presents a zoom-in on this conserved RPPR region. The full genomic reference sequence is depicted above the profiles and the zoomed-in region is displayed in red below the profiles. PWM and conservation score cut-offs of 0.6 were used to select TFBS predictions. These results highlight the conservation of the Ret-1/PCE-1, BAT-1, NRE and Ret-4 elements among mammals. Many Eopsin-1 binding sites are predicted because the motifs are only 6 bp long.

PromAn program implementation and future improvements

The PromAn web server and PromAnGUI visualization tool are written in Tcl/Tk 8.4, HTML and JavaScript. Both tools are modular and organized in order to allow for easily upgrade by the addition of supplementary genomic sequence alignment tools, promoter or TFBS databases and prediction programs.

We are currently integrating additional multiple alignment programs dedicated to large genomic sequences, such as TBA (16) and Multi-LAGAN (21). A future version of PromAn will include an option to give the user the possibility to predict TFBSs on orthologous sequences. PromAn will thus integrate statistics similar to the ones available in Footer (6). Identification of *cis*-regulatory modules will be implemented in a future version of the graphical user interface. PromAnGUI will also integrate tissue-specific, transcriptomic and interactomic data relative to transcription factors to add new filtration dimensions aimed at improving TFBS predictions. Thus, the user will be able to select regulatory motifs or modules according to co-expressed transcription factors interacting together.

CONCLUSION

The PromAn web server provides a number of advantages over many existing systems for promoter analysis. First, minimal prior knowledge of the genomic region of interest is necessary. Second, PromAn provides the possibility of performing multiple alignment using more than two orthologous sequences, allowing refinement of the evolutionary conserved regions. The PromAnGUI graphical user interface is a powerful tool used to integrate and visualize results and to filter out false positive TFBS predictions with matrix, conservation scores and biological knowledge of the user. Therefore, PromAn facilitates the construction of hypotheses in terms of potentially regulatory regions and elements in order to direct further experimental validations.

ACKNOWLEDGEMENTS

The authors thank L. Moulinier, J. Muller, F. Plewniak, R. Ripp and S. Uge for stimulating discussions. The authors are also grateful to L. Hermida and J. D. Thompson for interesting critical comments and suggestions on the manuscript. The work was funded by the Institut National de le Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique, the Université Louis Pasteur de Strasbourg, the Réseau National des Génopoles de Strasbourg and the EVI-GENORET project (number LSHG-CT-2005-512036). Funding to pay the Open Access publication charges for this article was provided by GIE CERBM.

Conflict of interest statement. None declared.

REFERENCES

1. Sandelin,A., Wasserman,W.W. and Lenhard,B. (2004) ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.*, **32**, W249–W252.
2. Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W217–W221.
3. Tagle,D.A., Koop,B.F., Goodman,M., Slightom,J.L., Hess,D.L. and Jones,R.T. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.*, **203**, 439–455.
4. Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.*, **26**, 225–228.
5. Berezikov,E., Guryev,V. and Cuppen,E. (2005) CONREAL web server: identification and visualization of conserved transcription factor binding sites. *Nucleic Acids Res.*, **33**, W447–W450.
6. Corcoran,D.L., Feingold,E. and Benos,P.V. (2005) FOOTER: a web tool for finding mammalian DNA regulatory regions using phylogenetic footprinting. *Nucleic Acids Res.*, **33**, W442–W446.
7. Yamashita,R., Suzuki,Y., Wakaguri,H., Tsuritani,K., Nakai,K. and Sugano,S. (2006) DBTSS: database of human transcription start sites, progress report 2006. *Nucleic Acids Res.*, **34**, D86–D89.
8. Schmid,C.D., Perier,R., Praz,V. and Bucher,P. (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res.*, **34**, D82–D85.
9. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nature Genet.*, **29**, 412–417.
10. Down,T.A. and Hubbard,T.J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, **12**, 458–461.
11. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
12. 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.
13. Vlieghe,D., Sandelin,A., De Bleser,P.J., Vleminckx,K., Wasserman,W.W., van Roy,F. and Lenhard,B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.*, **34**, D95–D97.
14. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
15. Lenhard,B., Sandelin,A., Mendoza,L., Engstrom,P., Jareborg,N. and Wasserman,W.W. (2003) Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.*, **2**, 13.
16. 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.

17. Valdar,W.S. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.
18. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*., **25**, 4876–4882.
19. Mitton,K.P., Swain,P.K., Chen,S., Xu,S., Zack,D.J. and Swaroop,A. (2000) The leucine zipper of NRL interacts with the CRX homeodomain. A possible mechanism of transcriptional synergy in rhodopsin regulation. *J. Biol. Chem*., **275**, 29794–29799.
20. Chen,S., Wang,Q.L., Nie,Z., Sun,H., Lennon,G., Copeland,N.G., Gilbert,D.J., Jenkins,N.A. and Zack,D.J. (1997) Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron*, **19**, 1017–1030.
21. Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A. and Batzoglou,S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*., **13**, 721–731.