doi:10.1093/nar/gkl879

PReMod: a database of genome-wide mammalian *cis*-regulatory module predictions

Vincent Ferretti, Christian Poitras[1], Dominique Bergeron[1], Benoit Coulombe[1], François Robert[1] and Mathieu Blanchette[2,*]

McGill University and Genome Quebec Innovation Center, 740 Dr Penfield, Montreal, Qc, Canada H3A 1A4, [1]Institut de Recherches Cliniques de Montréal, 110 Pine Avenue West, Montréal, Qc, Canada H2W 1R7 and [2]McGill Center for Bioinformatics. McGill University, 3775 University Street, room #332. Montréal, Qc, Canada H3A 2B4

Received August 15, 2006; Revised October 10, 2006; Accepted October 11, 2006

ABSTRACT

We describe PReMod, a new database of genome-wide *cis*-regulatory module (CRM) predictions for both the human and the mouse genomes. The prediction algorithm, described previously in Blanchette *et al*. (2006) *Genome Res*., 16, 656–668, exploits the fact that many known CRMs are made of clusters of phylogenetically conserved and repeated transcription factors (TF) binding sites. Contrary to other existing databases, PReMod is not restricted to modules located proximal to genes, but in fact mostly contains distal predicted CRMs (pCRMs). Through its web interface, PReMod allows users to (i) identify pCRMs around a gene of interest; (ii) identify pCRMs that have binding sites for a given TF (or a set of TFs) or (iii) download the entire dataset for local analyses. Queries can also be refined by filtering for specific chromosomal regions, for specific regions relative to genes or for the presence of CpG islands. The output includes information about the binding sites predicted within the selected pCRMs, and a graphical display of their distribution within the pCRMs. It also provides a visual depiction of the chromosomal context of the selected pCRMs in terms of neighboring pCRMs and genes, all of which are linked to the UCSC Genome Browser and the NCBI. PReMod: http://genomequebec.mcgill.ca/PReMod.

INTRODUCTION

The identification of DNA regulatory regions is one of the most important and challenging problems toward the functional annotation of genomes. In higher eukaryotes, transcription factor (TF) binding sites are often organized in clusters called *cis*-regulatory modules (CRM), which consists of DNA regions of up to a few hundred bases located in the (extended) neighborhood of the gene being regulated (1). While the prediction of individual TF-binding sites is a notoriously difficult problem, CRM predictions have proven to be more reliable and several algorithms have been developed in the last few years.

Most predictive methods rely on prior knowledge that has to be provided by the user. For instance, some methods will analyze the promoters of a set of (presumably) co-regulated genes obtained from some prior experiments in order to identify over-represented motif combinations (2–10). Other methods require a small set of TF position-weight matrices (PWMs) that are expected to co-occur in modules, and identify genomic regions densely populated in putative sites for these TFs (11–16). Because of the prior knowledge they require, none of these approaches are able to produce an unbiased, genome-wide survey of mammalian CRMs. Indeed, the only database of predicted *cis*-regulatory regions currently available for mammals, CisRed (17), is restricted to promoter regions.

In Blanchette *et al*. (18), we described a new sequence-based, genome-wide CRM identification method that exploits the observation that CRMs often contain several phylogenetically conserved binding sites for a few different TFs [see also a related approach by Philippakis and Bulyk (19)]. Applying this algorithm to the human and mouse genomes, we built the PReMod database, which contains the complete set of predicted CRMs (pCRMs) for those two genomes. Together with the recently published regulatory potential estimation from the Hardison group (20,21), our method represents the only computational approach that has been used for *de novo*, genome-wide prediction of CRMs.

PReMod will be useful for several types of investigations. First, researchers interested in the regulation of a specific

*To whom correspondence should be addressed. Tel: 514 398 5209; Fax: 514 398 3387; Email: blanchem@mcb.mcgill.ca

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors gene can use PReMod to identify putative CRMs in the vicinity of that gene. The PReMod information is complementary to other types of data like inter-species conservation, CpG islands, regulatory potential, etc. However, it provides a richer annotation, as it predicts the TFs likely to be involved. Second, researchers interested in identifying the targets of a particular TF or TF family will find PReMod useful as it provides a ranked list of putative targets for all TFs for which PWMs are available in Transfac. Modules are ranked by their total binding site concentration for that factor. The list of pCRMs associated to a particular TF can then be used to validate experimentally some of the predictions. For example, Blanchette *et al*. (18) used the modules predicted to be bound by E2F4 and estrogen receptor (ER) to build a DNA microarray for chromatin immunoprecipation (ChIP) -chip. A total of 55 and 433 modules were thus validated for ER and E2F4, respectively. While this corresponds to a relatively low fraction of the total number of modules tested (17% for E2F4 and 3% for ER), it is expected that testing binding under different experimental conditions will validate a much larger number of pCRMs since TFs (and in particular ER) are known to regulate different genes in different cellular contexts (22,23). Predicted CRMs can also be tested for function using lower-throughput approaches, such as reporter assays [e.g. Woofle *et al*. (24) and the Vista Enhancer Database (http://enhancer.lbl.gov)], and their predicted binding sites can be confirmed via gel shifts or mutagenesis. Finally, PReMod can be used as a data source for data mining efforts to understand the relationship between TFs (e.g. through co-occurrence of binding sites) or between TFs and genes of a particular function or expression pattern [e.g. see Ref. (18)]. By providing TF target predictions that are more accurate than individual binding site predictions, PReMod affords the researchers a better dataset from which subtle patterns can emerge. For example, using PReMod, Blanchette *et al*. (18) highlighted a surprising enrichment of pCRMs near the 3′ end of genes; a results that is corroborated by a growing number of experimental evidence (25,26).

Users need to keep in mind that the different types of predictions contained within PReMod are associated with different expected specificity. We first clarify that PReMod is not meant to be an exhaustive list of CRMs, and that CRMs that would not fit the signature described above would go undetected. Among all the predictions contained in PReMod, those of individual TF-binding sites have the lowest expected accuracy. More accurate are the predictions of the interaction between a TF (or a family of TFs) and a particular module (but without specifying the exact position of the binding sites). Finally, the most accurate predictions of the location of the pCRMs themselves, although the precise boundaries of the modules remain difficult to establish.

METHODS

The pCRMs contained in PReMod were computed using the method described by Blanchette *et al*. (18). We only provide a short overview of the method, and refer the interested reader to that article for more details. At the base of PReMod is a set of individual binding site predictions for TFs whose binding preferences are described by PWMs from the Transfac 7.2 database (27). Putative human binding sites are scored based on how well the human site and its orthologs in mouse and rat match the matrix [orthology is based on Multiz genome-wide alignments (28)]. Putative mouse sites are computed based on an alignment to the human and dog genomes. More precisely, a binding site's score is a weighted sum of the log-likelihood ratio scores in the three species. The score of the modules reported in PReMod reflect the presence, in a region of 100–1000 bp, of a surprisingly large number of binding sites (or, more precisely, a surprisingly large sum of their individual scores), for a few different PWMs. Specifically, to assign a score to a given genomic region, each PWM is first assigned a 'matrixScore', which reflects the surprise associated with the density and quality of predicted sites in that region. This surprise (*P*-value) depends, among other things, on the length and GC-content of the region and the genome-wide number and scores of predicted sites for the same PWM. The PWM with the highest matrixScore is chosen as first 'tag' for the region. Its occurrences are then masked, and the process is repeated, selecting a second tag. Up to five tags can be selected for a given module. In the end, the region is assigned a 'moduleScore', which reflects the surprise associated with the combined scores of the tags. Depending on which number of tags gives the most significant result, the lower-scoring tags may be rejected. It is important to mention here that although PWMs chosen as tags for a module are likely to be of interest, other PWMs that were not selected could also correspond to factors binding the module. This is particularly true in the case where two or more different PWMs represent binding sites for factors of the same family (e.g. STAT1 and STAT3). Because factors from the same family tend to have similar PWMs, it is very difficult to distinguish between their binding sites. Since their predicted sites will heavily overlap, only one member of the family will be reported as tag. However, this should not be interpreted as an indication that this member is significantly more likely than its homologs to bind the module. Instead, the user should refer to the 'matrixScore' to assess the binding potential of a particular TF.

Genomic regions obtaining significant moduleScores (*P*-value below e^{-10}) are reported in PReMod. We should however emphasize that the prediction algorithm is not very good at identifying the correct boundaries of the CRMs, and that one pCRM may sometimes actually contain two functionally distinct modules, or one module may be split between two CRMs. We encourage the user to consider all types of evidence, (e.g. regions of inter-species conservation) to decide on the correct CRM boundaries.

THE PREMOD DATABASE

Content

Table 1 reports the key statistics for the human and mouse versions of PReMod. The human version contains more than 123 000 predicted modules, slightly more than the 91 000 modules of mouse version. The difference is largely due to the fact that the mouse binding site predictions use the dog genome for comparison, resulting in more stringent predictions. Approximately 1.9% of the human genome

Table 1. Key statistics on the PReMod 1.0 database

	PreMod1.0 human	PReMod1.0 mouse
Genome assembly	Build 34—May 2004 (hg17)	Build 34—March 2005 (mm6)
Transfac version	7.2	7.2
Number of pCRMs	123 510	91 412
Fraction of genome contained in pCRMs	1.93%	1.68%
Average module length	481 bp	479 bp
Fraction of modules that are		
proximal (<2 kb from TSS)	10.8%	8.4%
distal (2–10 kb from TSS)	7.7%	8.3%
long-range (>10 kb from TSS)	81.6%	83.4%
Average number of tags per module	3.3	3.5
Average number of different PWMs per module	30.6	26.0
Average number of predicted sites per module	75.6	58.7
Average number of module containing sites for a given TF	7842	5395

(and 1.7% of the mouse genome) is covered by pCRMs, consistent with the hypothesis that a large fraction of the non-coding functional regions has a regulatory function (29). Note that the set of human modules in PreMod is based on a newer assembly than the dataset originally reported in (18). The large number of predicted sites per module is due in part to the fact that sites are predicted separately for each PWM, even though several matrices often represent the same or related TF. Thus, a single DNA location can be predicted as a binding site for more than one PWM.

Web interface

A Java web-based application allows users to browse, query or download the database. To address the various needs of the users, PReMod can be queried in a number of ways using an advanced search form (Figure 1A). First, users can request regulatory modules related to a given gene. Although there is currently no way to confidently assigning pCRMs to the gene they regulate, PReMod assumes that the gene whose transcription start site (TSS) is the closest to the module is the most likely target. However, this association is likely to often be incorrect, in particular for very-long-range regulators.

The second type of PReMod queries is TF-centric. Specifically, the user can request to see all the pCRMs containing predicted binding sites for one or more TFs (only TFs with Transfac matrices can be sought). By default, all modules containing predicted sites for the specified TF will be reported, although the search can also be restricted to only the PWMs used as tags for the pCRMs. An example of this type of query is shown in Figure 1A. Here, the user wants to identify the pCRMs containing tags for two nuclear receptor TFs, ER (M00191) and androgen receptor (M00447). All queries can be refined further by restricting the search to some chromosomal regions, to pCRMs that have a particular moduleScore, to located modules around specific genes, or to modules overlapping CpG islands.

Upon submitting a query, the user receives the list of modules satisfying the given constraints. All outputs can be viewed as HTML files or exported to an Excel spreadsheet. For each module reported, the module identifier, genomic position, length and score are given. Also given are the genes with the closest TSS upstream or downstream of the pCRM. Finally, the list of Transfac matrices selected as tags for the module is shown.

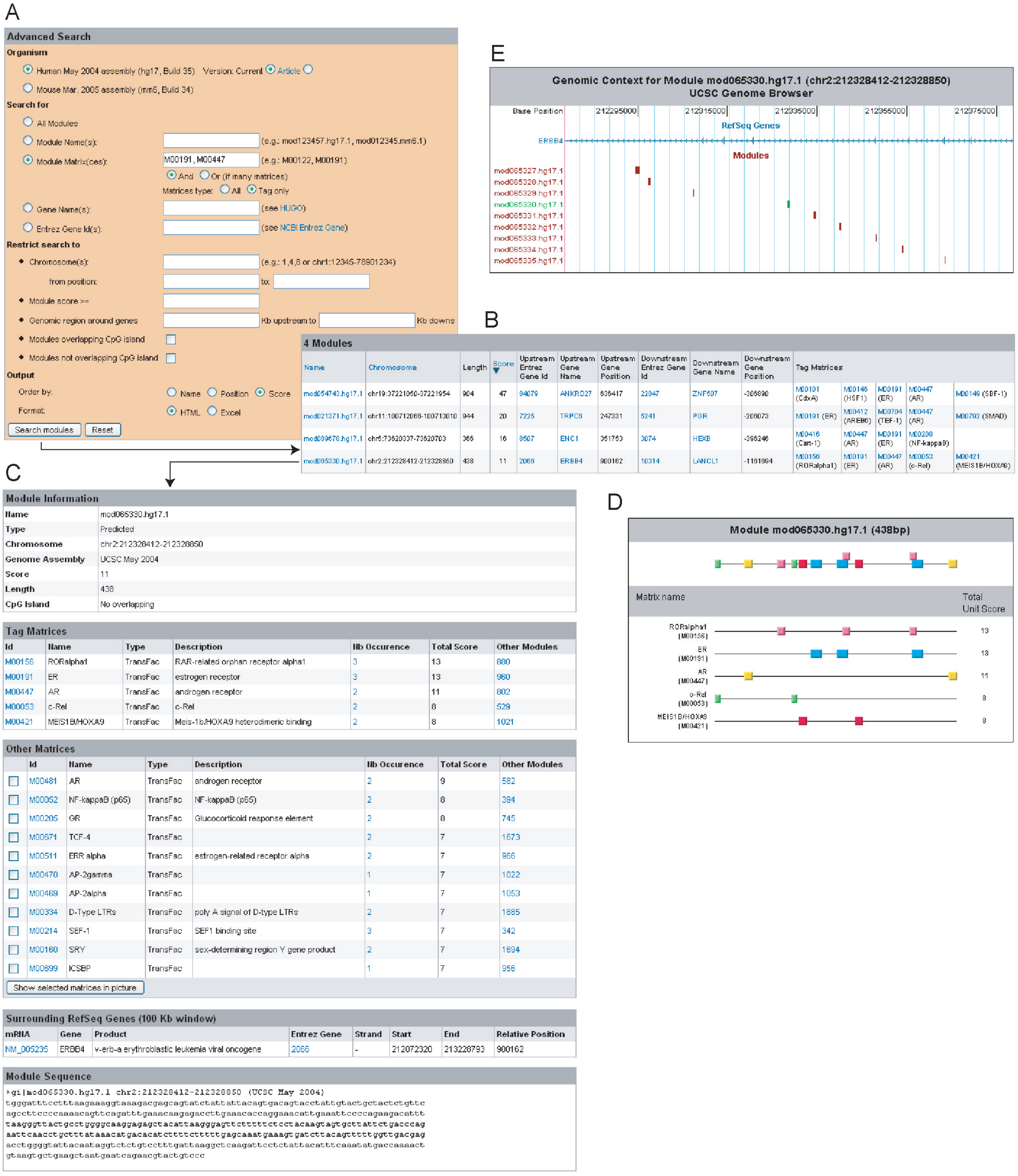
For example, given the query described above, the list of four modules produced as output is shown in Figure 1B. The second is a module located next to the progesterone receptor gene, which we showed to be bound by ER (18). However, we focus here on the fourth module reported, which is interesting for a number of reasons. First, not only are the estrogen and androgen receptors predicted to bind this module, but a third nuclear receptor, RORalpha1, which was not included in our query, is selected as first tag for this module. Given that different nuclear receptors are known to cooperatively (or antagonistically) bind regulatory regions (30), this association is promising. Second, this region is located within an intron of the *ERBB4* gene (v-erb-a erythroblastic leukemia viral oncogene), a key growth factor receptor tyrosine kinase inducing cell differentiation (31). The *ERBB* family is a key player in hormone-dependent breast (and other types of) cancer.

For each module, a details page is obtained by clicking the module name, as is exemplified in Figure 1C–E, for the module described above. This page contains all the binding site information about the selected module, starting with a visual representation of the position of the predicted binding sites for the TFs selected as tags (Figure 1D). The page gives the complete list of matrices with predicted sites in the module, and the position of these predicted sites can be visualized in the graphical display (Figure 1C). We emphasize again that, although the selection of TF tags for the modules is a necessary step algorithmically, the fact that a TF was not selected as a tag should not necessarily be interpreted as the TF being of less interest. Therefore, we recommend that users consider the matrix's total score as an indication of the binding potential.

Each module can be visualized in its genomic context, together with the genes nearby and the other surrounding modules (Figure 1E). By clicking the other modules in the image, one can explore their binding site content and properties. Quite often, interesting patterns will emerge by considering together several neighboring modules. For example, the module located upstream of the module described above is also predicted to be bound by several nuclear receptors. Finally, to explore the selected module in the context other types of annotations, a link to the UCSC genome browser is provided, where the pCRMs are displayed using a custom track.

Future developments

Several features will be added to PReMod in the near future. Predicted CRMs will soon be made available for other mammalian species (rat, dog, etc.), and orthologous pCRMs from different species will be linked to each other, allowing easy jumps from one species to the other. As new genome assemblies come out, new versions of the database will be released. To simplify querying the database, a BLAST server will be made available, allowing the quick identification of pCRMs homologous to a given query sequence. Finally, the



Advanced Search

Organism

● Human May 2004 assembly (hg17, Build 35) Version: Current ● Article ○

○ Mouse Mar. 2005 assembly (mm6, Build 34)

Search for

○ All Modules

○ Module Name(s): (e.g.: mod123457.hg17.1, mod012345.mm6.1)

● Module Matrix(ces): M00191, M00447 (e.g.: M00122, M00191)

● And ○ Or (if many matrices)

Matrices type: ○ All ● Tag only

○ Gene Name(s): (see HUGO)

○ Entrez Gene Id(s): (see NCBI Entrez Gene)

Restrict search to

- Chromosome(s): (e.g.: 1,4,8 or chr1:12345-78901234)

  from position: to:

- Module score >=
- Genomic region around genes Kb upstream to Kb downs
- Modules overlapping CpG island □
- Modules not overlapping CpG island □

Output

Order by: ○ Name ○ Position ● Score

Format: ● HTML ○ Excel

Search modules | Reset

4 Modules

Name	Chromosome	Length	Score	Upstream Entrez Gene Id	Upstream Gene Name	Upstream Gene Position	Downstream Entrez Gene Id	Downstream Gene Name	Downstream Gene Position	Tag Matrices				
mod054743.hg17.1	chr19:37221050-37221954	904	47	84079	ANKRD27	636417	22847	ZNF507	-306890	M00101 (CdxA)	M00146 (HSF1)	M00191 (ER)	M00447 (AR)	M00149 (SBF-1)
mod021371.hg17.1	chr11:100712066-100713010	944	20	7225	TRPC6	247331	5241	PGR	-206073	M00191 (ER)	M00412 (AREB6)	M00704 (TEF-1)	M00447 (AR)	M00792 (SMAD)
mod089678.hg17.1	chr5:73620337-73620703	366	16	8507	ENC1	351753	3074	HEXB	-396246	M00416 (Cart-1)	M00447 (AR)	M00191 (ER)	M00208 (NF-kappaB)	
mod065330.hg17.1	chr2:212328412-212328850	438	11	2066	ERBB4	900162	10314	LANCL1	-1161694	M00156 (RORalpha1)	M00191 (ER)	M00447 (AR)	M00053 (c-Rel)	M00421 (MEIS1B/HOXA9)

Module Information

Name	mod065330.hg17.1
Type	Predicted
Chromosome	chr2:212328412-212328850
Genome Assembly	UCSC May 2004
Score	11
Length	438
CpG Island	No overlapping

Tag Matrices

Id	Name	Type	Description	Nb Occurence	Total Score	Other Modules
M00156	RORalpha1	TransFac	RAR-related orphan receptor alpha1	3	13	880
M00191	ER	TransFac	estrogen receptor	3	13	960
M00447	AR	TransFac	androgen receptor	2	11	802
M00053	c-Rel	TransFac	c-Rel	2	8	529
M00421	MEIS1B/HOXA9	TransFac	Meis-1b/HOXA9 heterodimeric binding	2	8	1021

Other Matrices

	Id	Name	Type	Description	Nb Occurence	Total Score	Other Modules
□	M00481	AR	TransFac	androgen receptor	2	9	582
□	M00052	NF-kappaB (p65)	TransFac	NF-kappaB (p65)	2	8	394
□	M00205	GR	TransFac	Glucocorticoid response element	2	8	745
□	M00671	TCF-4	TransFac		2	7	1673
□	M00511	ERR alpha	TransFac	estrogen-related receptor alpha	2	7	966
□	M00470	AP-2gamma	TransFac		1	7	1022
□	M00469	AP-2alpha	TransFac		1	7	1053
□	M00334	D-Type LTRs	TransFac	poly A signal of D-type LTRs	2	7	1885
□	M00214	SEF-1	TransFac	SEF1 binding site	3	7	342
□	M00160	SRY	TransFac	sex-determining region Y gene product	2	7	1694
□	M00699	ICSBP	TransFac		1	7	956

Show selected matrices in picture

Surrounding RefSeq Genes (100 Kb window)

mRNA	Gene	Product	Entrez Gene	Strand	Start	End	Relative Position
NM_005235	ERBB4	v-erb-a erythroblastic leukemia viral oncogene	2066	-	212072320	213228793	900162

Module Sequence

```
>gi|mod065330.hg17.1 chr2:212328412-212328850 (UCSC May 2004)
tgggatttcctttaagaaaggtaaagacgagcagtatctattattacagtgacagtacctattgtactgctactctgttc
agccttccccaaaacagttcagatttgaaacaagagaccttgaaacaccaggaaacattgaaattccccagaagacattt
taagggttactgcctggggcaaggagagctacattaagggagttctttttctcctacaagtagtgcttattctgacccag
aattcaacctgctttataaacatgacacatctttttcttttgagcaaatgaaagtgatcttacagttttggttgacgag
acctggggtattacaataggtctctgtcctttgattaaggctcaagattcctctattacatttcaaatatgaccaaaact
gtaagtgctgaagctaatgaatcagaacgtactgtccc
```

Module mod065330.hg17.1 (438bp)

Matrix name	Total Unit Score
RORalpha1 (M00156)	13
ER (M00191)	13
AR (M00447)	11
c-Rel (M00053)	8
MEIS1B/HOXA9 (M00421)	8

Genomic Context for Module mod065330.hg17.1 (chr2:212328412-212328850)
UCSC Genome Browser

Figure 1. Sample screenshots of a query input and its related outputs generated by PReMod. (**A**) The Advanced Search page. By clicking on 'Search Predicted Modules', users access a page that allows searching pCRMs by module name, matrix name, gene name or Entrez gene Id. In the example shown here, the user wants to identify the pCRMs containing tags for both the ER (M00191) and the androgen receptor (M00447). Output can be ordered by name, position or score, and can be displayed in HTML or exported as an Excel file. (**B**) Query output page. Upon submitting a query, the list of modules satisfying the given constraints is displayed. 'B' shows the result of the query shown in 'A'. For each module reported, the module identifier, genomic position, length, and score are given. Also given are the genes with the closest TSS upstream or downstream of the pCRM. Finally, the list of Transfac matrices selected as tags for the module is shown. (**C**) Module Information page. By clicking on a module name in the Query output page (B), a details page is obtained that contains all the binding site information about the selected module. The page includes the list of all the matrices that were used to calculate the moduleScore (Tag Matrices) as well as all the other matrices found in that pCRM (Other Matrices). The page also includes a list of the surrounding genes (within a 100 kb window) and the DNA sequence of the module. (**D**) The Module view. The Module Information page also contains a graphical representation of the position of the predicted binding sites for the TFs selected as tags. Any matrices present in that module (those listed in Tag Matrices and in Other Matrices) can be added to (or removed from) the display by a simple click. (**E**) The Genomic Context view. The genomic context of the selected module can also be visualized within the Module Information page. In this display, the selected pCRM, together with any other pCRMs and genes present within a 100 kb window are shown. By clicking on any module in that image the user is sent to the appropriate Module Information page.

module prediction algorithm is under constant refinement and new releases of PReMod will likely improve the specificity of the predictions.

ACKNOWLEDGEMENTS

The authors thank Diane Bourque, Johanne Duhaime, Nathalie Edmond and Martin Leboeuf for their technical support, as well as the UCSC genome browser group for their support. The authors also thank Vincent Giguère for his advice about the nuclear receptor analysis. This work was funded by grants from Génome Québec and Génome Canada (M.B., B.C. and F.R.). F.R. holds a new investigator award from the CIHR. Funding to pay the Open Access publication charges for this article was provided by NSERC and CIHR.

Conflict of interest statement. None declared.

REFERENCES

1. Levine,M. and Davidson,E.H. (2005) Gene regulatory networks for development. *Proc. Natl Acad. Sci. USA*, **102**, 4936–4942.
2. Aerts,S., Van Loo,P., Thijs,G., Moreau,Y. and De Moor,B. (2003) Computational detection of cis-regulatory modules. *Bioinformatics*, **19**, II5–II14.
3. Aerts,S., Van Loo,P., Moreau,Y. and De Moor,B. (2004) A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics*, **20**, 1974–1976.
4. Gupta,M. and Liu,J.S. (2005) *De novo* cis-regulatory module elicitation for eukaryotic genomes. *Proc. Natl Acad. Sci. USA*, **102**, 7079–7084.
5. Krivan,W. and Wasserman,W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, **11**, 1559–1566.
6. Segal,E. and Sharan,R. (2005) A discriminative model for identifying spatial *cis*-regulatory modules. *J. Comput. Biol.*, **12**, 822–834.
7. Sharan,R., Ben Hur,A., Loots,G.G. and Ovcharenko,I. (2004) CREME: *cis*-regulatory module explorer for the human genome. *Nucleic Acids Res.*, **32**, W253–W256.
8. Thompson,W., Palumbo,M.J., Wasserman,W.W., Liu,J.S. and Lawrence,C.E. (2004) Decoding human regulatory circuits. *Genome Res.*, **14**, 1967–1974.
9. Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.
10. Zhou,Q. and Wong,W.H. (2004) CisModule: *de novo* discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl Acad. Sci. USA*, **101**, 12114–12119.
11. Alkema,W.B., Johansson,O., Lagergren,J. and Wasserman,W.W. (2004) MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W195–W198.
12. Bailey,T.L. and Noble,W.S. (2003) Searching for statistically significant regulatory modules. *Bioinformatics*, **19**, II16–II25.
13. Frith,M.C., Li,M.C. and Weng,Z. (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, **31**, 3666–3668.
14. Johansson,O., Alkema,W., Wasserman,W.W. and Lagergren,J. (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, **19**, i169–i176.
15. Sinha,S., Schroeder,M.D., Unnerstall,U., Gaul,U. and Siggia,E.D. (2004) Cross-species comparison significantly improves genome-wide prediction of *cis*-regulatory modules in *Drosophila*. *BMC Bioinformatics*, **5**, 129.
16. Sinha,S., van Nimwegen,E. and Siggia,E.D. (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19**, i292–i301.
17. Robertson,G., Bilenky,M., Lin,K., He,A., Yuen,W., Dagpinar,M., Varhol,R., Teague,K., Griffith,O.L., Zhang,X. *et al.* (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res.*, **34**, D68–D73.
18. Blanchette,M., Bataille,A.R., Chen,X., Poitras,C., Laganiere,J., Lefebvre,C., Deblois,G., Giguere,V., Ferretti,V., Bergeron,D. *et al.* (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.*, **16**, 656–668.
19. Bulyk,M.L. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol.*, **5**, 201.
20. Kolbe,D., Taylor,J., Elnitski,L., Eswara,P., Li,J., Miller,W., Hardison,R. and Chiaromonte,F. (2004) Regulatory potential scores from genome-wide three-way alignments of human, mouse and rat. *Genome Res.*, **14**, 700–707.
21. King,D.C., Taylor,J., Elnitski,L., Chiaromonte,F., Miller,W. and Hardison,R.C. (2005) Evaluation of regulatory potential and conservation scores for detecting *cis*-regulatory modules in aligned mammalian genome sequences. *Genome Res.*, **15**, 1051–1060.
22. Hartman,S.E., Bertone,P., Nath,A.K., Royce,T.E., Gerstein,M., Weissman,S. and Snyder,M. (2005) Global changes in STAT target selection and transcription regulation upon interferon treatments. *Genes Dev.*, **19**, 2953–2968.
23. Zeitlinger,J., Simon,I., Harbison,C.T., Hannett,N.M., Volkert,T.L., Fink,G.R. and Young,R.A. (2003) Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell*, **113**, 395–404.
24. Woolfe,A., Goodson,M., Goode,D.K., Snell,P., McEwen,G.K., Vavouri,T., Smith,S.F., North,P., Callaway,H., Kelly,K. *et al.* (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, **3**, e7.
25. Cawley,S., Bekiranov,S., Ng,H.H., Kapranov,P., Sekinger,E.A., Kampa,D., Piccolboni,A., Sementchenko,V., Cheng,J., Williams,A.J. *et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.
26. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engstrom,P.G., Frith,M.C. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.*, **38**, 626–635.
27. 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.
28. 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.
29. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
30. Laganiere,J., Deblois,G., Lefebvre,C., Bataille,A.R., Robert,F. and Giguere,V. (2005) From the cover: location analysis of estrogen receptor alpha target promoters reveals that FOXA1 defines a domain of the estrogen response. *Proc. Natl Acad. Sci. USA*, **102**, 11651–11656.
31. Grant,S., Qiao,L. and Dent,P. (2002) Roles of ERBB family receptor tyrosine kinases, and downstream signaling pathways, in the control of cell growth and survival. *Front Biosci.*, **7**, d376–d389.