W408–W411 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki486

CisMols Analyzer: identification of compositionally similar *cis*-element clusters in ortholog conserved regions of coordinately expressed genes

Anil G. Jegga, Ashima Gupta, Sivakumar Gowrisankar, Mrunal A. Deshmukh, Steven Connolly, Kevin Finley and Bruce J. Aronow*

Division of Biomedical Informatics, Children's Hospital Research Foundation, CCHMC, 3333 Burnet Avenue, Cincinnati, OH-45229, USA

Received February 14, 2005; Revised March 29, 2005; Accepted April 20, 2005

ABSTRACT

Combinatorial interactions of sequence-specific *trans*-acting factors with localized genomic *cis*-element clusters are the principal mechanism for regulating tissue-specific and developmental gene expression. With the emergence of expanding numbers of genome-wide expression analyses, the identification of the *cis*-elements responsible for specific patterns of transcriptional regulation represents a critical area of investigation. Computational methods for the identification of functional *cis*-regulatory modules are difficult to devise, principally because of the short length and degenerate nature of individual *cis*-element binding sites and the inherent complexity that is generated by combinatorial interactions within *cis*-clusters. Filtering candidate *cis*-element clusters based on phylogenetic conservation is helpful for an individual ortholog gene pair, but combining data from *cis*-conservation and coordinate expression across multiple genes is a more difficult problem. To approach this, we have extended an ortholog gene-pair database with additional analytical architecture to allow for the analysis and identification of maximal numbers of compositionally similar and phylogenetically conserved *cis*-regulatory element clusters from a list of user-selected genes. The system has been successfully tested with a series of functionally related and microarray profile-based co-expressed ortholog pairs of promoters and genes using known regulatory regions as training sets and co-expressed genes in the olfactory and immunohematologic systems as test sets. CisMols Analyzer is accessible via a Web interface at http://cismols.cchmc.org/.

INTRODUCTION

The integration of genomic sequences with transcription factor function and gene expression to decipher the gene regulatory networks underlying various developmental processes is a major challenge of the post-genomic era (1). Although the view that regulatory regions manifest as clusters of transcription factor binding sites (TFBSs) has been around for some time, it was the review by Arnone and Davidson (2) that clearly presented the case for emphasizing *cis*-clusters in both experimental and computational analyses. In fact, it is this paradigm shift that led to important advances in the detection of combinatorial occurrence of *cis*-elements and understanding transcriptional regulation (3). However, the availability of a number of completely sequenced eukaryotic genomes with an ever expanding volume of gene expression profile data has made computationally based strategies for deciphering genetic regulatory networks more viable. The methods range from sophisticated Gibbs sampling-based algorithms to more 'brute force' counting and analysis of fixed-length oligonucleotide words (*k*mer or *k*tuple word searching) (4,5). For a complete list of Internet resources and tools available to predict transcription regulatory clusters, refer to Ureta-Vidal *et al*. (6) and http://zlab.bu.edu/zlab/gene.shtml. Computational methods have focused primarily on trying to identify the co-occurrence of a set of TFBSs in a group of genes co-expressed or functionally related, and most of them have been restricted to the promoter or upstream regions. However, the basic problem of identifying the true positives from a list of combinatorial patterns remains. The problem becomes even more complicated and the results are difficult

*To whom correspondence should be addressed. Tel: +1 513 636 4865; Fax: +1 513 636 2056; Email: bruce.aronow@cchmc.org

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

to interpret when the entire stretch of non-coding regions comprising introns and upstream and downstream regions is considered. Adopting a phylogenetic approach allows substantial reduction in the number of false positives in the identification of regulatory regions of individual orthologous gene pairs (7–12). Although the need for experimental validation remains critical, at present, predicted *cis*-acting signature element searches can greatly focus experimental targets for validation studies.

The detection of a particular known *cis*-acting element in all or many of the genes in a particular expression cluster does not necessarily mean that the genes are regulated via that element. The likelihood of this prediction is greater if each of these shared clusters is also conserved in the corresponding inter-species ortholog. CisMols Analyzer is built based on these two hypotheses and is designed to identify significant *cis*-regulatory elements from sets of co-expressed or related groups of genes for elements that are also ortholog-conserved. To do this, ortholog-conserved *cis*-clusters for each individual gene pair are identified and stored in the database. Next, a gene list is compiled based on various criteria such as coordinate regulation and then the ortholog-conserved *cis*-clusters for each of the genes in the list are compared to identify occurrence of common *cis*-clusters. Since the existence of gene regulatory regions in intronic and downstream regions is well proven, our method to identify these sites is not confined to upstream regions alone, but is extended to intronic and 5′ and 3′ gene-flanking regions. We have successfully validated our algorithm on several data sets comprising skeletal muscle-specific genes, liver-specific genes, pancreas overexpressed genes, olfactory genes (13) and immune system genes (14).

Genomic regions of orthologous genes are retrieved from UCSC Golden Path, along with the exon annotations. Putative regulatory regions are identified either by using our earlier developed Trafac server (12) or by searching against the potential regulatory regions stored in the GenomeTrafac database (http://genometrafac.cchmc.org; Jegga *et al*., manuscript submitted). The conserved *cis*-element dense regions for each of the ortholog gene pairs are compared to identify the common binding sites in a group of genes. The web application is available at http://cismols.cchmc.org. Researchers can automatically (i) create gene groups and identify shared ortholog-conserved putative regulatory regions and individual binding sites, (ii) search genes for known *cis*-regulatory modules and (iii) identify potential novel gene targets for known *cis*-regulatory modules or novel clusters of individual binding sites.

INPUT

Creating and submitting gene groups for analysis

CisMols Analyzer is designed to analyze a list of genes—typically co-expressed or related genes—for *cis*-element clusters that are shared by genes in the list. In contrast, GenomeTrafac is a whole-genome repository of individual genes with specified gene orthologs that have been analyzed, in batch form over the entire genome, for phylogenetically conserved *cis*-elements (Jegga *et al*., manuscript submitted). Trafac is similarly single-gene oriented, but it allows for the entry of human-curated ortholog gene pairs (12). CisMols Analyzer operates on genes in either database by allowing lists of these genes to be formed and then subjecting the lists to shared *cis*-element analysis. It is possible to analyze existing gene groups that have been assembled by the system administrators and by other users. However, to create new groups and perform a clustering analysis to detect modules that contain shared *cis*-elements, a login account is needed. Options are also provided to select the genomic regions for a single gene or a group of genes that need to be searched for the occurrence of common TFBSs. By default, CisMols Analyzer searches for clusters in the genomic region comprising the 5′- and 3′-flanking 10 kb and also the intronic regions. After submitting the genes for analysis, the user will be notified by email of the availability of the results when the clustering is finished.

Searching for *cis*-clusters

Users can customize the search criteria using Boolean logic to restrict the search to known validated *cis*-regulatory modules and/or a combination of individual binding sites. The minimum number of binding sites that must appear in each cluster, or the minimum number of genes in which each cluster must appear, can also be specified. The validated known *cis*-regulatory modules are provided along with a PubMed citation. Users also have the option of saving their search parameters or queries.

OUTPUT

The output generated is a set of putative regulatory modules occurring within an ortholog conserved region of two or more genes. CisMolGram (Figure 1), a graphical representation of ortholog-conserved *cis*-clusters shared by a group of genes, depicts the location of clusters within gene regions. The shared *cis*-clusters (two or more than two binding sites) are represented as variously colored boxes on the gene. Each of these shared clusters is linked to its respective ortholog conserved regions and can be visualized as Trafacgrams or regulograms (described in the legend of Figure 1). The Trafac and regulogram image pages have the option to download the sequences (corresponding to that cluster window) in fasta format. Links are also provided to the human and mouse UCSC browser, in which the user can see a CisMols-identified cluster in the context of other annotations. Users can also download the UCSC browser-compatible GFF files (currently GFF3) for each sequence from the regulogram image page. These can be uploaded directly to UCSC Golden Path, enabling visualization of the CisMols cluster region in the context of all other features available as aligned annotation tracks (known genes, predicted genes, ESTs, mRNAs, CpG islands, assembly gaps and coverage, chromosomal bands, other species homologies and more). The table view or the legend (Figure 1D) displays a summary of the results. It indicates the details of each cluster (the individual constituent *cis*-elements and their total frequency) and the frequency of the clusters occurring in each of the genes compared and involved in cluster analysis. Users also have the option for flipping to the ortholog view. The default base sequence is the human gene, and it is mapped to the corresponding mouse ortholog. However, the program can be used

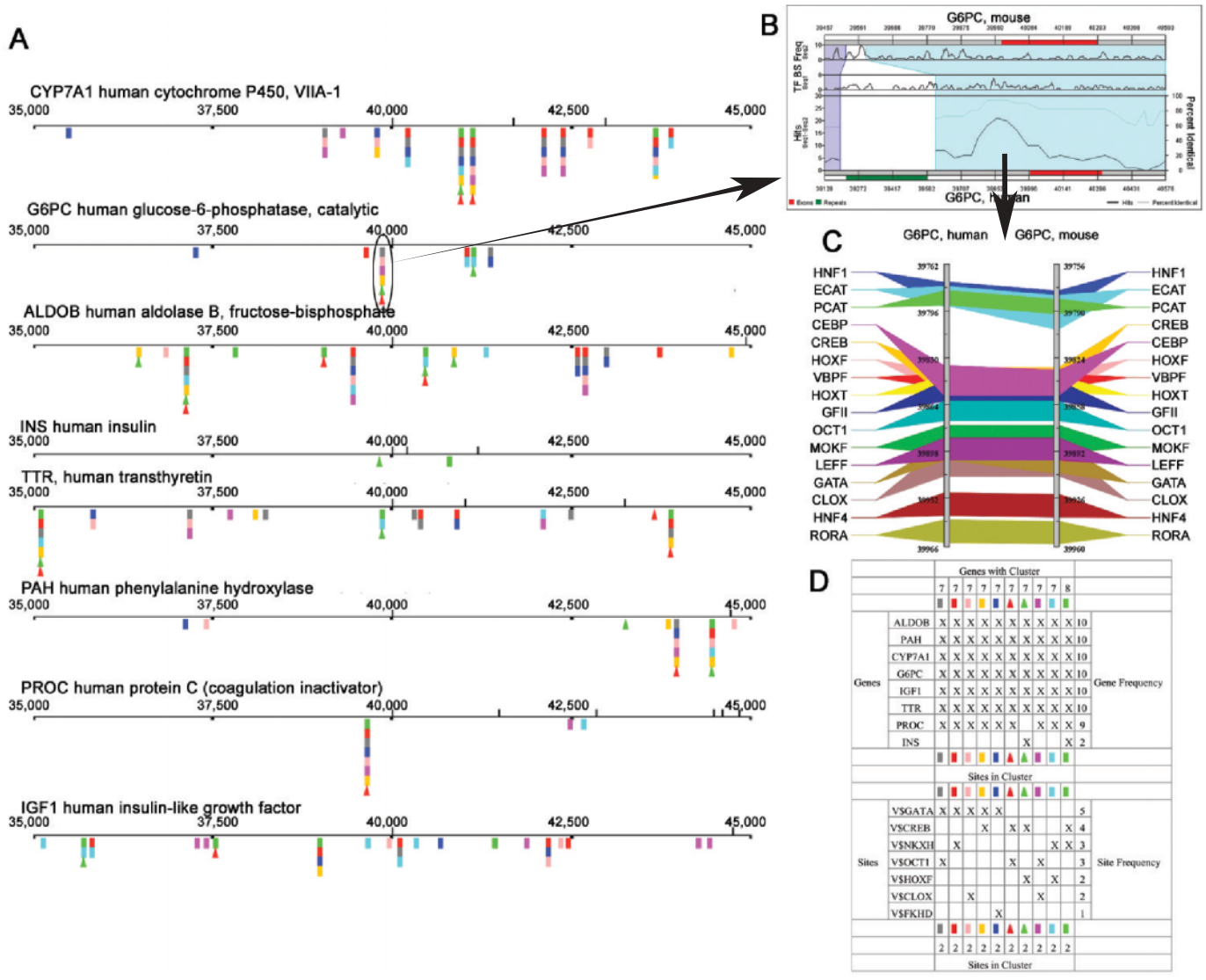

Figure 1. Detection and visualization of compositionally similar *cis*-regulatory modules shared by multiples genes. (**A**) CisMolGram provides a depiction of *cis*-element modules that are shared across two or more genes and contain two or more binding sites. Each module is represented by variously colored boxes located relative to the promoter of each gene in the analyzed group. Genes gathered from the GenomeTrafac database generally have their first exon locations at position 40 000. Most of the locations of *cis*-module clusters contain multiple modules. The locations of the clusters can be selected to be shown over either set of gene orthologs. By clicking on a cluster module (circled), a link is given to view either (**B**) the density of conserved elements over the regions (regulogram) or (**C**) the conserved *cis*-element arrangement at that location (Trafac image). The Regulogram shows the two sequences, mouse and human, represented as horizontal bars. The red-colored segments on these bars are exons. The green-colored bars shown parallel to the genomic sequences represent the repeat regions. The regions of sequence alignment are represented as differently colored quadrilaterals that relate one sequence to another. Within each shaded block, the percentage sequence similarity and the number of TFBSs are represented as two separate line graphs in the lower half. The frequencies of individual binding sites occurring in each of the sequences separately are shown as two running graphs in the top half of the pane. The percentage similarity is the average sequence conservation as determined by the BlastZ algorithm and the shared *cis*-element hits are determined by an algorithm that uses a 200 bp moving window to look through the *cis*-elements present within the conserved sequence block. Numbers are nucleotide positions. The regulogram can be clicked to zoom in or to view the TFBSs that are common to the two sequences at the click-point coordinate. In the Trafac image, the two gray vertical bars are the two genes (human and mouse) compared. The numbers represent the nucleotide positions with respect to the sequences used. The TFBSs occurring in both the genes are highlighted as variously colored bars drawn across the two genes. (**D**) The table view or the CisMolGram legend displays a summary of the results. It indicates the details of each cluster (the constituent individual *cis*-elements and their total frequency) and the frequency of the clusters occurring in each of the genes compared and involved in cluster analysis.

for any ortholog gene pair after uploading their alignments and binding sites to the Trafac database. Options are also provided to modify the viewable regions to focus on *cis*-clusters of interest. The CisMolGram can be saved or exported as images (SVG, TIFF, PDF, PNG or JPG format). The search parameters can also be saved for future generation of CisMolGrams.

SOFTWARE AND ACCESS

The CisMols Analyzer algorithm is implemented in Java. The time taken to analyze a group of genes depends upon factors such as number of genes in the group, gene lengths, percentage conservation between the orthologous pair of genes, and ortholog-conserved *cis*-element clusters. CisMols Analyzer looks for conserved *cis*-element clusters within conserved

regions [BlastZ-aligned genomic regions of ⩾70% sequence similarity (15)]. MatInspector (16) is used to identify the potential binding sites in each of the genomic sequences. The analysis parameters for identification of TFBSs were set to 0.85 for the core similarity and optimal for the matrix similarity. A typical analysis on the CisMols server—for, say, 50 ortholog gene pairs with default parameter settings of 10 kb upstream and downstream for each gene—takes approximately 1 h. The current upper limit of processing capability is about 242 ortholog gene pairs (a total sequence length of ∼32 Mb, of which about 7 Mb are BlastZ-aligned with ⩾70% sequence similarity). We intend to work on accommodating the analysis of larger sets in the future. Currently we are also working on providing the statistical significance and comparison between *cis*-clusters identified for two discrete gene groups.

CONCLUSION

The identification of signature clusters for a specific group of genes is still difficult. Most of the time, the *cis*-element clusters responsible for tissue specificity tend to be scored relatively low. For example, searching for ortholog-conserved shared *cis*-clusters in a group of pancreas overexpressed genes without any *cis*-element filter resulted in the identification of non-specific clusters. However, when the search was performed again restricting the results to only those clusters that have at least one Pdx binding site, the resulting shared clusters coincided with the validated regulatory regions of each of the individual genes (data not shown). The top hits, or the *cis*-element clusters shared by the most genes, tend to be more general, and, although they are important for gene expression, very little knowledge can be extracted from these about conferring tissue specificity for a group of genes. Using a control or a negative control does, however, improve understanding of the importance of the shared clusters—for instance, comparing the most shared *cis*-clusters in a group of genes with overexpression in the liver with genes overexpressed in the cerebellum. Clearly, there is a paradox in the phylogenetic footprinting approach. To allow the recognition of conserved (regulatory) elements, there should be enough evolutionary distance, but, at the same time, this evolutionary distance makes it difficult to recognize TFBSs—the short conserved elements. Nevertheless, the significance of a predicted shared *cis*-regulatory module for a group of co-expressed genes or a functionally related group of genes will be higher if the shared clusters are additionally conserved in both gene orthologs.

ACKNOWLEDGEMENTS

This work was supported by grants NCI UO1 CA84291-07 (Mouse Models of Human Cancer Consortium), NIH R24 DK 064403 (Digestive Diseases Research Development Center—DDRDC), NIEHS ES-00-005 (Comparative Mouse Genome Centers Consortium) and NIEHS P30-ES06096 (Center for Environmental Genetics). Funding to pay the Open Access publication charges for this article was provided by Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA.

Conflict of interest statement. None declared.

REFERENCES

1. Davidson,E.H., Rast,J.P., Oliveri,P., Ransick,A., Calestani,C., Yuh,C.H., Minokawa,T., Amore,G., Hinman,V., Arenas-Mena,C. *et al.* (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.
2. Arnone,M.I. and Davidson,E.H. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development*, **124**, 1851–1864.
3. Michelson,A.M. (2002) Deciphering genetic regulatory codes: a challenge for functional genomics. *Proc. Natl Acad. Sci. USA*, **99**, 546–548.
4. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
5. van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.
6. Ureta-Vidal,A., Ettwiller,L. and Birney,E. (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nature Rev. Genet.*, **4**, 251–262.
7. 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.
8. Gumucio,D.L., Heilstedt-Williamson,H., Gray,T.A., Tarle,S.A., Shelton,D.A., Tagle,D.A., Slightom,J.L., Goodman,M. and Collins,F.S. (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.
9. Hardison,R.C., Oeltjen,J. and Miller,W. (1997) Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.*, **7**, 959–966.
10. Loots,G.G., Locksley,R.M., Blankespoor,C.M., Wang,Z.E., Miller,W., Rubin,E.M. and Frazer,K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288**, 136–140.
11. 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.
12. Jegga,A.G., Sherwood,S.P., Carman,J.W., Pinski,A.T., Phillips,J.L., Pestian,J.P. and Aronow,B.J. (2002) Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res.*, **12**, 1408–1417.
13. Genter,M.B., Van Veldhoven,P.P., Jegga,A.G., Sakthivel,B., Kong,S., Stanley,K., Witte,D.P., Ebert,C.L. and Aronow,B.J. (2003) Microarray-based discovery of highly expressed olfactory mucosal genes: potential roles in the various functions of the olfactory system. *Physiol. Genomics*, **16**, 67–81.
14. Hutton,J.J., Jegga,A.G., Kong,S., Gupta,A., Ebert,C., Williams,S., Katz,J.D. and Aronow,B.J. (2004) Microarray and comparative genomics-based identification of genes and gene regulatory regions of the mouse immune system. *BMC Genomics*, **5**, 82.
15. 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.
16. Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.