UCRL-JRNL-202313

CREME: Cis-Regulatory Module Explorer for the Human Genome

R. Sharan, A. Ben-Hur, G. G. Loots, I.V. Ovcharenko

February 11, 2004

Nucleic Acid Research

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

CREME: Cis-Regulatory Module Explorer for the Human Genome

Roded Sharan[*] Asa Ben-Hur[†] Gabriela G. Loots[‡]

Ivan Ovcharenko[‡]

[*] International Computer Science Institute, 1947 Center St., Berkeley, CA 94704.

Email: roded@icsi.berkeley.edu.

[†] Dept. of Biochemistry, B400 Beckman Center, Stanford University, CA 94305.

Email: abenhur@stanford.edu.

[‡] EEBI and Genome Biology Divisions, L-441 Lawrence Livermore National Laboratory, 7000 East Ave., Livermore, CA 94550.

Email: {loots1,ovcharenko1}@llnl.gov.

Abstract

The binding of transcription factors to specific regulatory sequence elements is a primary mechanism for controlling gene transcription. Eukaryotic genes are often regulated by several transcription factors, whose binding sites are tightly clustered and form *cis*-regulatory modules. In this paper we present a web-server, CREME, for identifying and visualizing *cis*-regulatory modules in the promoter regions of a given set of potentially co-regulated genes. CREME relies on a database of putative transcription factor binding sites that have been annotated across the human genome using a library of position weight matrices and evolutionary conservation with the mouse and rat genomes. A search algorithm is applied to this dataset to identify combinations of transcription factors whose binding sites tend to co-occur in close proximity in the promoter regions of the input gene set. The identified *cis*-regulatory modules are statistically scored and significant combinations are reported and graphically visualized. Our web-server is available at `http://creme.dcode.org/`.

1 Introduction

Developmental and environmental factors constantly modulate the expression levels of genes in living cells. Gene transcription is primarily controlled by transcription factors (TFs) that physically interact with regulatory sequences located in promoter elements proximal to the transcription start site of a gene. In eukaryotes transciptional regulation is combinatorial in nature: The expression level of a gene is determined by an interplay among several TFs, whose binding sites are organized in a modular fashion along a gene's promoter [1, 2, 3].

A cis-*regulatory module (CRM)* is a sequence segment that contains several spatially-clustered transcription factor binding sites (TFBS), whose corresponding TFs cooperate in the regulation of a group of genes. The set of distinct TFBS that make up a *cis*-regulatory module is called a *TFBS module*. Previous work on identifying modules has followed two main directions. The first focuses on identifying pairs of transcription factors whose binding sites tend to co-occur in the promoter sequences of a group of related genes [4, 5]. However, these methods do not constrain the occurrences of binding sites in each combination to be close together within the given promoter regions. The second direction concerns the problem of identifying sequence segments that contain a known CRM [6, 7, 8, 9]. To date, only few studies have addressed the problem of identifying novel TFBS modules [10, 11, 12].

Recently, we have introduced a new method for identifying novel TFBS modules and assessing their statistical significance [11]. Our method relies on a database of putative TFBS across the promoters of known human genes which are highly conserved in orthologous genes from mouse and rat. We use evolutionary conservation to increase the reliability of TFBS predictions: it has been shown that the majority of computationally predicted TFBS ($> 95\%$) can be reliably eliminated based on evolutionary analysis [13]. Proceeding from this database of conserved TFBS, a search algorithm seeks all combinations of two or more TFs whose binding sites tend to co-occur in a selected set of promoters more frequently than expected by chance. Each TFBS module is statistically evaluated and significant modules are reported. We applied this strategy for the analysis of promoter regions of stress response genes and cell-cycle regulated genes. We identified several novel TFBS modules, most of which were shown to be associated with significantly co-expressed or functionally related groups of genes [11].

In this paper we summarize this strategy for identifying TFBS modules and describe a web-server, CREME (Cis-REgulatory Module Explorer), that provides an implementation of the method, controlled by a graphical user interface. The application requires as input a set of putatively co-regulated genes to initiate a search for abundant CRMs in the promoter regions of those genes. The interface allows users to customize the search and to visualize the identified CRMs. CREME is available online at `http://creme.dcode.org/`.

2 The Computational Pipeline

The CREME server is designed to identify combinations of TFBS that tend to co-occur in close proximity in the promoter regions of an input gene set. As a preprocessing step we prepare a database of (putative) TFBS across the promoter regions of known human genes. TFBS are commonly modeled using position weight matrices (PWMs). Each TFBS corresponds to a hit of a certain PWM that is conserved in the genomes of human, mouse and rat. The module search algorithm that uses this database, consists of four phases. In the first phase we identify non-redundant PWMs whose hits are enriched in the input promoters, compared to a background set of promoters. In the second phase we enumerate all combinations of these PWMs that occur within a short window in the input promoters. In the third phase these combinations are statistically evaluated. Last, significant combination are reported and visualized. We detail these steps below.

The database of conserved TFBS was prepared using PWMs cataloged in the TRANSFAC database [14], which contains over 500 vertebrate TF matrices that comprise about 400 TF families. We used evolutionary conservation to reduce the problem of false predictions of binding sites. To this end, we employed the rVista 2.0 tool [13] (`http://rvista.dcode.org/`) in combination with human-mouse and human-rat alignments obtained from the ECR Browser (*http://ecrbrowser.dcode.org/*). The rVista tool was applied to each of the alignments to find TFBS that are conserved between the two respective genomes. The conservation information from the two rVista computations was superimposed, to produce a list of TFBS that are conserved in the three genomes with TRANSFAC scores 0.8 and above. Each conserved hit is assigned its TRANSFAC score.

Genome alignments of human vs. mouse and human vs. rat were obtained from the ECR Browser, an interactive database consisting of pre-computed whole genome alignments from multiple vertebrate genomes (Ovcharenko I. et al., 2004). These alignments were generated using the most recent assemblies for the three genomes (hg16, mm4, and rn3; http://genome.ucsc.edu/cgi-bin/hgGateway/). By overlaying conservation profiles with RefSeq gene annotation mapped to the hg16 human genome assembly and extracting evolutionary conserved regions in the human genome, we determined that of the 16000 non-overlapping RefSeq transcripts, 46% of their promoters (1.5kb upstream of the transcriptional start site; 7307 genes) were highly conserved in both the mouse and rat genomes. In total, the CREME database contains 1.4 million TFBS corresponding to 487 different PWMs. These sites are highly conserved (conservation greater than 80% in over 20 bp) in human, mouse and rat, and are present in the promoter regions of 7307 human genes.

Given a database of TFBS we aim at identifying combinations of TFBS that tend to co-occur in the selected set of promoters. To reduce the large number of possible combinations we restrict attention to TFBS that are enriched in the given set of promoters compared to a background set. The enrichment scores are described in [11]. We search this filtered data using a hashing technique that identifies all the TFBS combinations that occur in the given promoters. The search is performed using a sliding sequence window of user-defined length. A window is considered to contain a module if it contains at least one binding site for each of the TFs that make up the module, where the maximum number of TFs per module is controlled by the user.

Abundant TFBS modules are then statistically scored, based on their number of occurrences, taking into account the frequency of occurrence of their constituent TFBS in a background set of promoters as well as the similarity between the PWM models for the different binding sites. The significance of the TF combination is weighted against the distribution of occurrence of the corresponding CRM under a null assumption that the binding sites for the different TFs occur independently. This is done via a permutation test, where randomized instances of the TFBS database are produced by permuting the identities of the different TFBS from the original database. The null distribution is approximately normal and its parameters are estimated based on these random datasets.

Significant combinations are further filtered to eliminate redundant modules, that is, modules that overlap considerably in the set of positions they occur in. The p-values of the resulting combinations are adjusted for multiple testing using the q-value method [15], which is specifically designed for adjusting the significance of a large number of observations. The reader is referred to [11] for further details on the algorithm and the statistical scoring method.

3 User Interface

CREME requires as input a list of putatively co-regulated or functionally related human genes, provided as either Locus Link (LLID) or GenBank accession numbers (Figure 1, panel 1). The CREME web-site contains a sample input of 268 genes that were shown to be cell-cycle regulated [16].

The user has control over several search parameters: (1) Hit threshold; (2) module length; and (3) number of PWMs per module. The hit threshold parameter controls which TFBS in the database will be included in the analysis. Higher threshold values increase the specificity but decrease the sensitivity of the considered TFBS. The second parameter specifies the width of the sliding window to be used when searching for modules. Large window sizes increase the sensitivity of the search but may decrease the statistical significance associated with a module. The third parameter determines the maximum number of distinct TFBS that make up a module. The higher the value the more thorough the search at the expense of an increase in the processing time.

Upon submitting a query, CREME searches for sets of transcription factors whose binding sites tend to tightly cluster in the promoter regions of the input genes more frequently than expected by chance. The identified modules are reported and visualized (Figure 1, panel 2). For each identified module, CREME provides a graphical display that illustrates the promoter regions of the genes in the input set or, alternatively, of the genes that contain this module. For each promoter, shown are the occurrences of putative binding sites for the TFs comprising the module, where occurrences of the module are shaded (Figure 2).

To illustrate the application of the CREME server, we used it to analyze a set of 268 genes that have been previously shown to be cell-cycle regulated [16]. These genes are a subset of the original set of 651 genes with unique LLID, for which conserved promoter segments have been detected. We used the default running parameters (hit threshold of 0.85; 150bp-long modules; and at most 3 TFs per module) for the analysis. The computation resulted in 4 significant modules. The first two modules are detailed in Figure 1 (panel 2). The promoter regions of the genes that contain the second module are visualized in Figure 2.

4 Conclusions

We have introduced the CREME web-server for identifying *cis*-regulatory modules in a given group of genes. By combining transcription factor binding site motif searches, human/mouse/rat evolutionary conservation and statistical assessment of combinations of binding sites, the CREME server is able to identify *cis*-regulatory modules specific for the promoter regions of a set of functionally related genes. The server reports the modules shared by a subset of promoters from the original input gene set along with their p-values, and provides a graphical display of these co-occurrences. While there are several other available tools that provide methods for detecting given *cis*-regulatory modules (see, e.g., [7, 9]), CREME is able to identify novel *cis*-regulatory modules *de novo* and, thus, can potentially assist researchers in the discovery of transcription factors that synergistically activate genes and may therefore be responsible for their similar behavior. The identification of such combinatorial modules is critical for understanding how transcriptional regulatory elements are encoded in the human genome, as well as to help explain why certain factors induce genes to be turned on or off at the same time.

Acknowledgments

This research was supported in part by NSF ITR Grant CCR-0121555 and Livermore LDRD Grant 04-ERD-052. The work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory Contract No. W-7405-Eng-48.

References

[1] C.H. Yuh, H. Bolouri, and E.H. Davidson. Genomic cis-regulatory logic: Experimental and computational analysis of a sea urchin gene. *Science*, 279:1896–1902, 1998.

[2] M.Z. Ludwig, N.H. Patel, and M. Kreitman. Functional analysis of eve stripe 2 enhancer evolution in drosophila: rules governing conservation and change. *Development*, 125(5):949–58, 1998.

[3] W. Krivan and W.W. Wasserman. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, 11(9):1559–66, 2001.

[4] Y. Pilpel, P. Sudarsanam, and G.M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, 29(2):153–9, 2001.

[5] D.G. Thakurta and G.D. Stormo. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17(7):608–621, 2001.

[6] W.W. Wasserman and J.W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, 278(1):167–81, 1998.

[7] M.C. Frith, J.L. Spouge, U. Hansen, and Z. Weng. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.*, 30(14):3214–3224, 2002.

[8] S. Sinha, E. van Nimwegen, and E. D. Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19, Suppl. 1:I292–I301, 2003.

[9] O. Johansson, W. Alkema, W. W. Wasserman, and J. Lagergren. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, 19, Suppl. 1:I169–I176, 2003.

[10] O.V. Kel-Margoulis, T.G. Ivanova, E. Wingender, and A.E. Kel. Automatic annotation of genomic regulatory sequences by searching for composite clusters. *Pacific Symposium on Biocomputing*, pages 187–198, 2002.

[11] R. Sharan, I. Ovcharenko, A. Ben-Hur, and R.M. Karp. CREME: A framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, 19, Suppl. 1:I283–I291, 2003.

[12] E. Segal and R. Sharan. A discriminative model for identifying spatial cis-regulatory modules. *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology*, 2004.

[13] G.G. Loots, I. Ovcharenko, L. Pachter, et al. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.*, 12(5):832–9, 2002.

[14] E. Wingender, X. Chen, R. Hehl, et al. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, 28(1):316–9, 2000.

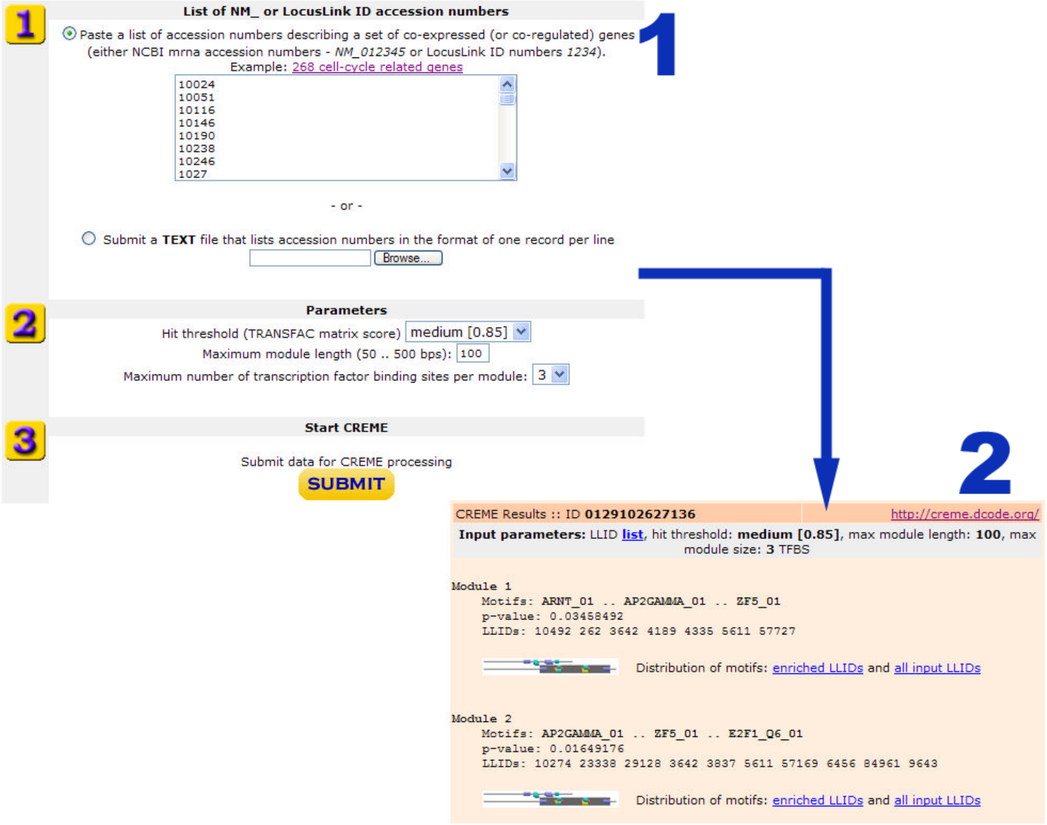
[15] J.D. Storey and R. Tibshirani. Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, 100:9440–9445, 2003.

[16] M.L. Whitfield, G. Sherlock, A. Saldanha, et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, 13:1977–2000, 2002.

5 Figure Legends

Figure 1. The CREME user interface. Upon submission of LLID (or GenBank) accession numbers as pasted text or as a file, and selection of search parameters, the user initiates a CREME process (panel 1) that automatically redirects to the results page (panel 2) when the computation is completed. The results page lists all the detected modules of TFBS whose enrichment in the promoters of the input genes is statistically significant, as well as a link to the list of PWMs that were enriched in the input set of genes. For each module, listed are its consitutent PWMs, its *p*-value and the LLIDs that contain it. In addition, two links to visualization are given for each module (see Figure 2). These show the hits of the TFs that appear in the module, either in the submitted set of genes, or only in the genes in which the module occurs.

Figure 2. CREME visualization of a detected module. Shown are the occurrences of the module and its constituent PWMs in the promoters of the selected gene set. In the example here, the module contains three TFs: AP2GAMMA, ZF5 and E2F1. 10 promoters that contain this module are visualized. AP2GAMMA sites are in red, ZF5 sites in green, and E2F1 sites in blue, and module occurrences are shaded in gray. The horizontal scale depicts the distance from each gene's transcription start site.



List of NM_ or LocusLink ID accession numbers

● Paste a list of accession numbers describing a set of co-expressed (or co-regulated) genes (either NCBI mrna accession numbers - *NM_012345* or LocusLink ID numbers *1234*).
Example: 268 cell-cycle related genes

```
10024
10051
10116
10146
10190
10238
10246
1027
```

- or -

○ Submit a **TEXT** file that lists accession numbers in the format of one record per line

Parameters

Hit threshold (TRANSFAC matrix score) medium [0.85]
Maximum module length (50 .. 500 bps): 100
Maximum number of transcription factor binding sites per module: 3

Start CREME

Submit data for CREME processing

CREME Results :: ID **0129102627136** http://creme.dcode.org/

Input parameters: LLID list, hit threshold: **medium [0.85]**, max module length: **100**, max module size: **3** TFBS

```
Module 1
    Motifs: ARNT_01 .. AP2GAMMA_01 .. ZF5_01
    p-value: 0.03458492
    LLIDs: 10492 262 3642 4189 4335 5611 57727
```

Distribution of motifs: enriched LLIDs and all input LLIDs

```
Module 2
    Motifs: AP2GAMMA_01 .. ZF5_01 .. E2F1_Q6_01
    p-value: 0.01649176
    LLIDs: 10274 23338 29128 3642 3837 5611 57169 6456 84961 9643
```

Distribution of motifs: enriched LLIDs and all input LLIDs

Figure 1:

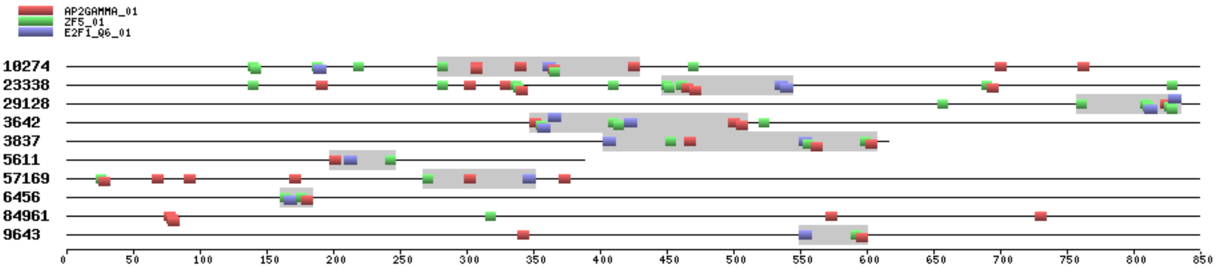


Figure 2: