BIOINFORMATICS

Vol. 1 no. 1 2003
Pages 1–9

CREME: A Framework for Identifying Cis-Regulatory Modules in Human-Mouse Conserved Segments

Roded Sharan[*†1], *Ivan Ovcharenko*[†2], *Asa Ben-Hur*[3] *and Richard M. Karp*[1]

[1]*International Computer Science Institute, 1947 Center St., Suite 600, Berkeley CA-94704. Email: {roded,karp}@icsi.berkeley.edu,* [2]*Genome Sciences Dept., Lawrence Berkeley National Laboratory, Berkeley CA-94720. Email: ivovcharenko@lbl.gov and* [3]*Dept. of Biochemistry, B400 Beckman Center, Stanford University CA-94305. Email: abenhur@stanford.edu.*

ABSTRACT

Motivation: The binding of transcription factors to specific regulatory sequence elements is a primary mechanism for controlling gene transcription. Recent findings suggest a modular organization of binding sites for transcription factors that cooperate in the regulation of genes. In this work we establish a framework for finding recurrent cis-regulatory modules in the promoters of a selected set of genes and scoring their statistical significance.

Results: Proceeding from a database of identified binding site motifs and their genomic locations we seek motifs whose frequency in the selected promoters is different than in a background promoter set. We present several statistical tests designed for this purpose. We provide a hashing algorithm for detecting combinations of these motifs that co-occur in clusters within the selected promoters. The significance of such co-occurrences is evaluated using novel statistical scores. Our methods are combined in CREME, a suite of software which includes a browser for viewing the pattern of occurrence of selected cis-regulatory modules. We applied our methodology to find modules within human-mouse conserved promoter segments, focusing on cell cycle regulated genes and stress response related genes. To validate the biological significance of the identified modules we tested whether the associated genes tended to be co-expressed or share similar function. In the cell cycle set five of the seven identified sets of genes were coherently expressed. On the stress response data four of the six detected sets fell predominantly into well-defined functional sub-categories.

Availability: http://icsi.berkeley.edu/~roded/creme.html.

Contact: roded@icsi.berkeley.edu.

Keywords: Cis-regulatory module, transcription factor binding site, motif cluster, statistical test.

*To whom correspondence should be addressed.
†These authors contributed equally to this work.

INTRODUCTION

The complex program of gene expression allows the living cell to cope with changing developmental and environmental conditions. A key mechanism for controlling protein abundance is the modulation of gene transcription by transcription factors (TFs) that bind specific regulatory sequence elements. Recent studies demonstrate combinatorial regulation of transcription in eukaryotes (Yuh et al., 1998): The expression level of a gene is determined by an interplay among several TFs, whose sites are organized in a modular fashion along the gene's promoter. Characterizing functional combinations of TF binding sites is key to understanding gene regulation and remains in large an unanswered computational challenge. The difficulty lies in the low specificity of the binding site motifs, which makes it hard to accurately detect them within long stretches of background sequence.

TF binding site motifs are commonly modeled using a position weight matrix (PWM). The most complete database of carefully evaluated binding sites is TRANSFAC (Wingender et al., 2000), which contains over 400 PWMs for vertebrate genomes. There is a vast literature on characterizing and finding TF binding sites (see (Stormo, 2000) and references thereof). Recently, computational and statistical methods were developed for identifying pairs of TFs that exhibit functional synergism, or tend to co-occur in close proximity in sequences of interest (Wasserman and Fickett, 1998; Pilpel et al., 2001; Sudarsanam et al., 2002; Hannenhalli and Levy, 2002; Thakurta and Stormo, 2001; Elkon et al., 2003).

Although much information can be gained from studying single TFs or interactions between pairs of TFs, the generalization to inferring whole regulatory mechanisms is a non-trivial task. A combination of binding sites for TFs that cooperate in the regulation of genes is termed a *cis-regulatory module (CRM)*. There are several studies on

the identification of occurrences of known CRMs, both in Drosophila (Berman et al., 2002; Rebeiz et al., 2002; Halfon et al., 2002) and in human (Krivan and Wasserman, 2001; Frith et al., 2001; Klingenhoff et al., 1998). Two previous papers addressed the statistical problem of scoring the occurrences of a given CRM, assuming that the binding sites of each participating TF occur according to independent Poisson processes (Wagner, 1999; Frith et al., 2002). Another study by Kel-Margoulis et al. (2002) used a genetic algorithm to search in the space of TF sets a densely occurring CRM.

In this paper we propose a framework for finding CRMs and scoring the statistical significance of their frequency of occurrence in a family of promoters. The biological significance of the overabundance or underabundance of a CRM within a family is not clear, but one might speculate that overabundance indicates a functional role in regulating genes within the family, whereas underabundance indicates that the CRM is not operative in such genes, or is operative only in a subfamily. For simplicity we orient our description of statistical methods towards the detection of overabundance (or enrichment), but the methods apply equally well to the detection of underabundance, and we report results in both directions. Our method is based on several components: (1) Restriction to conserved motif hits: We concentrate on promoter segments in the human genome that are conserved in mouse, and consider only sequence motifs that are aligned and locally conserved between human and mouse. This restriction allows us to reduce drastically the number of false positive hits of the PWMs, while preserving the number of true positives. (2) Statistical measures for the enrichment (or underrepresentation) of binding sites in a given set of genes compared to a background set. Using these scores we can pinpoint PWMs that are (statistically) relevant to the set of studied genes and carry further analysis on this set of motifs only. (3) A novel algorithm for finding CRMs in a set of promoters. The algorithm uses a hashing technique to go through all observed PWM combinations in the promoters of interest. (4) Significance measures for the co-occurrence of groups of motifs in the selected set relative to the frequencies of their constituent motifs. Our main measure explicitly treats the scoring problems resulting from similarity between different PWMs.

Our methods are embodied in CREME, a software suite which includes a browser for viewing the occurrences of selected cis-regulatory modules. We applied our methodology to find modules within human-mouse conserved promoter segments, focusing on cell cycle regulated genes and stress response genes. To validate the biological significance of the identified modules we tested whether the associated genes tended to be co-expressed or share similar function. On the cell cycle set we discovered seven putative CRMs with two to four component TFs. We tested the expression coherence of the genes containing each of the CRMs using the expression dataset of Whitfield et al. (2002); the coherence was highly significant in five cases and marginally significant in one more. For the stress response genes, we found six putative CRMs; in four cases the associated gene sets fell predominantly into well-defined functional subcategories.

PRELIMINARIES

A *motif* is a pattern of nucleotides, commonly modeled using a *position weight matrix*. A PWM for a transcription factor F of length l_F is a $4 \times l_F$ real matrix, recording for each relative position p and nucleotide N the probability of observing N at position p of a binding site for F. Using the PWM one can readily estimate the probability that a binding site for F occurs at a given position. Applying an appropriate threshold yields a set of positions in which the binding site is likely to occur along a given sequence. We call each such probable occurrence a *hit* and we associate a *position* with it. Full details on the computation of hits and their associated positions are given in the Results Section.

A *promoter* p is a sequence of nucleotides in the upstream region of a gene's transcription start site. We denote its length by $L(p)$. For a motif m of length l_m, $L_m(p) \equiv 2(L(p) - l_m + 1)$ is the number of positions in p at which m can occur when looking at both strands, and $h_m(p)$ is its number of hits in p.

The main subject of this work is identifying sets of TFs whose binding sites tend to co-occur in close proximity along a set of promoters of interest. This notion of co-occurrence is captured by the following definition: A *w-motif cluster* w.r.t. to a promoter p is a set of distinct motifs that occur at least once in p in an interval of length at most w. Biological examples suggest that the size of a motif cluster is typically bounded by a small number r of motifs (Krivan and Wasserman, 2001) giving rise to the notion of an *(r, w)-motif cluster*.

Let I be some instance of a w-motif cluster C. Define the *interval* of I as the interval $[l_I, r_I]$, where l_I is the starting position of I and r_I is the maximum position in $[l_I, l_I + w - 1]$ which contains a hit for some $m \in C$. We say that instance I_1 *dominates* an instance I_2 if the interval that corresponds to I_1 contains the interval that corresponds to I_2. The *count* $h_C(p)$ of a motif cluster C is obtained in the following way: We scan p in the direction of the positive strand. For each position that starts an instance I of C, we increment our count if no previous instance of C dominates I.

FINDING RECURRENT MOTIF CLUSTERS

In this section we treat the algorithmic question of identifying (r, w)-motif clusters in a given set of promoters $\mathcal{G}$ and calculating their counts. In the next section we study

the question of scoring the statistical significance of a cluster based on its count. For lack of space some algorithmic details are omitted.

Let n be the number of distinct motifs that we consider. Let $\mathcal{M}$ be the multiset of all motif hits in $\mathcal{G}$, ordered by their positions, and denote $M = |\mathcal{M}|$. Let q be the maximum number of distinct motifs that occur within an interval of length w in $\mathcal{G}$. Typical parameter values are $n = 30$, $r = 4$, $w = 100$, $q = 20$ and $M = 10000$. The count of a given (r, w)-motif cluster can be computed in $O(M)$ time. Thus, one can identify all such clusters in $\mathcal{G}$ and compute their counts in $O(\min\{n^r, q^r M\} \cdot M)$ time, by enumerating all possible sets of at most r motifs that occur in a window of length w and calculating their counts. This approach is too expensive for realistic values of the parameters. We next present a more efficient approach for the identification and counting of motif clusters.

A *consecutive instance* of a w-motif cluster C is an interval of length at most w in some $p \in \mathcal{G}$, that contains at least one hit for every $m \in C$, and no hit for any other motif. Our empirical results suggest that a large fraction of the recurrent motif clusters have at least one consecutive occurrence. An algorithm for identifying all w-motif clusters in $\mathcal{G}$ with an least one consecutive occurrence is given in Figure 1.

Algorithm ConsecID(·) runs in $O(Mr)$ time, since the size of the list of active clusters, C_{open}, is bounded by r. Using this algorithm we can identify and count motif clusters with at least one consecutive occurrence in $O(M^2)$ time, since there are at most $O(M)$ such motif clusters and counting each of them takes $O(M)$ time. The algorithm can be generalized in a straightforward way to identify and count motif clusters that do not necessarily have a consecutive occurrence, at the expense of increasing the complexity to $O(q^r M)$. We omit the details.

THE STATISTICAL FRAMEWORK

In this section we describe statistical scores for assessing the significance of the occurrences of a single motif or a motif cluster in a set of promoters. We denote by $\mathcal{G}$ the set of promoters of interest and by $\mathcal{B}$ a set of background promoters, which does not intersect $\mathcal{G}$. Typically, $\mathcal{G}$ consists of several hundred promoters and $\mathcal{B}$ is larger by an order of magnitude. We call $\mathcal{G}$ the *gene set* and $\mathcal{B}$ the *background set*.

Single Motif Abundance Tests

We present several scores for the frequency of a motif m in $\mathcal{G}$ compared to the background set $\mathcal{B}$. In the following we denote by n_b and n_g the number of promoters in the background set and the gene set, respectively. We denote by h_b and h_g the number of promoters that contain m in $\mathcal{B}$ and $\mathcal{G}$, respectively.

For any promoter p, let $X(p, m)$ be the event that promoter p contains at least one occurrence of m. Let us take the null hypothesis that the events $X(p, m)$ are independent and identically distributed; in particular, the probability of $X(p, m)$ does not depend on whether promoter p lies in $\mathcal{B}$ or in $\mathcal{G}$. Following (Tavazoie et al., 1999) the probability of observing h_g or more promoters in $\mathcal{G}$ that contain m, given that $h_g + h_b$ promoters in $\mathcal{G} \cup \mathcal{B}$ contain m, is the tail of a hypergeometric distribution. A small tail probability would indicate that occurrences of m are enriched in $\mathcal{G}$ relative to $\mathcal{B}$; a tail probability close to 1 would indicate the opposite. Note that this score does not take into account the number of motif hits in each promoter.

The above assumption of equal probability is reasonable when all examined promoters have similar lengths. However, promoters may vary in length, and our restriction to evolutionarily conserved segments creates additional variation. In order to take into account the length of the promoters we propose two approaches. The first approach is based on binning the promoters according to their length. For each bin b we use the background set to estimate the expectation $E(b)$ and variance $V(b)$ of the number of hits for m in a promoter belonging to b, and we take the null hypothesis that, for each bin b, the number of hits for m in promoters from $\mathcal{G}$ also have this expectation and variance. By the Central Limit Theorem (for independent, non-identically distributed random variables), the total number of hits for m in $\mathcal{G}$ is approximately normally distributed, where the expectation E and variance V of this distribution can be estimated from the bin information. Specifically, for a promoter p let $b(p)$ denote its bin. Then $E = \sum_{p \in \mathcal{G}} E(b(p))$ and $V = \sum_{p \in \mathcal{G}} V(b(p))$. These estimates allow us to derive a *normal-based* p-value for m.

The second approach assumes that motif occurrences are generated by a Poisson process with the same rate across all promoters. This assumption can be verified for each PWM separately, e.g., using a chi-square test: One counts the number of hits for the PWM on each of the promoters (more precisely, on equal-length segments of the promoters) and tests the goodness of fit of these counts to the Poisson assumption. Let $L_g = \sum_{p \in \mathcal{G}} L_m(p)$ and $L_b = \sum_{p \in \mathcal{B}} L_m(p)$. Then, assuming that hits for motif m occur according to a constant-rate Poisson process, the probability of observing h_g or more hits in $\mathcal{G}$, given that there are $h_b + h_g$ hits overall, is the tail of a binomial distribution with parameters $(h_b + h_g, L_g/L_b)$. This is true regardless of the (unknown) rate of the Poisson process. This tail probability can be used to detect whether hits of m are overrepresented or underrepresented in $\mathcal{G}$ compared

ConsecID($\mathcal{M}$):
$\mathcal{C} \leftarrow \emptyset$ # A hash of motif clusters whose keys are motif sets.
$C_{open} \leftarrow \emptyset$ # A hash of active clusters and their starting positions.
For $i = 1$ to $|\mathcal{M}|$ **do**:
    Let h be the i-th hit in $\mathcal{M}$ occurring at position $pos(h)$.
    For every $(C, start) \in C_{open}$ **do**:
        If ($pos(h) - start \geq w$ or $h \notin C$) **then** Insert($\mathcal{C}$,C); Delete(C_{open},C).
        If ($h \notin C$ and $|C| < r$) **then** Insert(C_{open},$(C \cup \{h\}, start)$).
        If $C = \{h\}$ **then** $start \leftarrow pos(h)$.
    If $\{h\} \notin C_{open}$ **then** Insert(C_{open},$(\{h\}, pos(h))$).
For every $C \in C_{open}$ **do**: Insert($\mathcal{C}$,C) # Add remaining active clusters.
Output $\mathcal{C}$.

Fig. 1. An algorithm for identifying all motif clusters with at least one consecutive instance in a given sequence. Procedures Insert(H,e) and Delete(H,e) insert/delete an element from a hash table H.

to $\mathcal{B}$. We refer the reader to (Waterman, 1995) for a thorough presentation of statistical tests for pattern counts.

Scoring a Cluster Count

In this section we restrict our attention to clusters of motifs and present methods for measuring the significance of the counts of such clusters within $\mathcal{G}$. Two motivating questions for these methods suggest themselves: (1) Which clusters occur more frequently in $\mathcal{G}$ than would be expected from their frequencies in $\mathcal{B}$? (2) Which clusters occur more frequently in $\mathcal{G}$ than would be expected from the frequencies in $\mathcal{G}$ of their component motifs? The first question could be attacked using the hypergeometric or normal tests described already for single motifs. However, a cluster could be overrepresented in $\mathcal{G}$ relative to $\mathcal{B}$ merely because its component motifs are overrepresented in $\mathcal{G}$. Therefore we have chosen to focus on the second question, which concerns whether these motif hits tend to co-occur in proximity within $\mathcal{G}$, leading to a higher cluster frequency than we would expect from the frequencies of their constituent motifs.

A complicating factor in addressing this question is the tendency of certain pairs of motifs to have frequent overlapping occurrences merely because certain positions within them have similar nucleotide distributions. To avoid giving weight to this effect we restrict attention to cluster occurrences that are *spaced*, meaning that, in addition to the usual requirement that all the motif hits comprising the cluster occurrence lie in a window of length w, the start positions of each such pair of hits must differ by at least h, where h is a parameter which is set to 4 in our computations.

Let us say that a set of motif hits within a promoter is *independent* if every pair of start positions among these hits differs by at least h. In order to focus on spaced cluster occurrences we replace each promoter by a set of t *spaced promoters*, each of which contains a randomly chosen maximal independent subset of the motif hits on the original promoter. These maximal independent subsets are generated by t executions of a randomized greedy algorithm. The idea is that these spaced promoters will contain a good sampling of the spaced occurrences of any cluster, while eliminating from consideration all cluster occurrences that are not spaced.

Having replaced the original promoters by t times as many spaced promoters, we wish to determine whether a given cluster occurs on these spaced promoters significantly more frequently than would be expected by chance, given the positions of the motif hits, the number of hits for each motif, and the number of (spaced) promoters containing each given motif. We attack this question by a Monte Carlo method. Consider each motif hit within the set of spaced promoters as having a location and a label, where the label is the name of the corresponding motif. Call a permutation of the labels *conservative* if for each motif, the number of promoters containing it is unchanged. We run k Monte Carlo calculations, each of which generates a random conservative permutation of the labels. Each Monte Carlo simulation starts with the original (true) labels and performs a long series of random conservative interchanges, each of which either interchanges the labels of two motif hits, or replaces all occurrences of some label A on one promoter by a different label B, and all occurrences of B on a second promoter by label A (where the requirement that the interchange is conservative implies that number of changes of A to B is equal to the number of changes of B to A).

For $j = 1, \ldots, t$, let p_j be the j-th spaced promoter derived from promoter $p \in \mathcal{G}$ and for $s = 1, \ldots, k$, let $C_s(p, j)$ be the number of occurrences of cluster C

in p_j after the s-th Monte Carlo relabeling of the motif hits on p_j. From these empirical values we can estimate $E(C(p,j))$ and $V(C(p,j))$, which are, respectively, the expectation and variance of the number of occurrences of cluster C on spaced promoter p_j after a random conservative relabeling. By the Central Limit Theorem, the total number of occurrences of C on spaced promoters is approximately normal with mean $\sum_{p,j} E(C(p,j))$ and variance bounded above by $t \sum_{p,j} V(C(p,j))$ (this upper bound accounts for the positive correlation between $C(p,j)$ for different j-s). This yields an upper bound on the p-value for the number of occurrences of C on spaced promoters under the original labeling. In this way we can determine whether occurrences of cluster C are significantly higher or lower than would be expected by chance.

An alternative approach for scoring a cluster $C = \{m_1, \ldots, m_l\}$ uses the *Poisson clump heuristic* (Waterman, 1995). It is based on the assumption that each of its motifs occurs independently according to a Poisson process. First, we evaluate the process rate λ_{m_i} for each motif m_i based on its hits in $\mathcal{G}$. The occurrences of C are grouped in *clumps*, i.e., maximal segments of overlapping intervals of instances of C. We also empirically evaluate the expected clump size μ_C. We then approximate the distribution of the number of clumps using a Poisson distribution and compute the quality of the approximation, similar to the approach in (Waterman, 1995): Denote $L = \sum_{p \in \mathcal{G}} L(p)$ and

$$p_C = \sum_{i=1}^{n} \lambda_{m_i} \prod_{1 \le j \le n, j \ne i} (1 - e^{-w\lambda_{m_j}}) .$$

Ignoring boundary effects, the probability density of a clump start is $\frac{p_C}{\mu_C}$ and the expected number of clumps is $\lambda = \frac{L p_C}{\mu_C}$. The event X_i, denoting a clump start at position i, is independent of any X_j for $\{j : |j - i| \ge w\}$. Applying Chen-Stein approximation bounds, as in (Waterman, 1995), we find that the deviation of the distribution of the number of clumps from the Poisson distribution (with parameter λ) is bounded by $\lambda^2(1 + \mu_C)\frac{4w}{L}$.

THE CREME FRAMEWORK

In this section we collect the methods developed above into a framework for motif cluster discovery, which we call CREME (Cis-REgulatory Module Explorer). CREME embodies an algorithm for finding and scoring motif clusters, as well as software for visualizing and evaluating the resulting clusters.

The cluster finding process consists of several stages: First, we find significantly enriched motifs in $\mathcal{G}$ by scanning the whole set of promoters and identifying PWMs that occur significantly more (less) frequently in the gene set compared to the background set. The enrichment p-values are computed using the normal-based score described in the Statistical Framework Section. We verify these p-values empirically, by simulating random gene sets of size $|\mathcal{G}|$, computing the p-values for m on these random sets and ranking the real p-value among the simulated ones. We retain all motifs whose p-value is smaller than 0.01 and was empirically verified, collecting both overrepresented and underrepresented motifs. This step reduces the size of the initial motif set by an order of magnitude and allows the subsequent stages to focus on motifs that are statistically relevant to the genes of interest, saving computation time and reducing the number of false discoveries.

In the next stage we filter similar PWMs – the output of the previous stage may include PWMs that are very similar, e.g., similar matrices that correspond to the same transcription factor. The goal of this stage is to produce a non-redundant list of PWMs. We say that two PWMs are *redundant* if at least 50% of the occurrences of one PWM take place in a window of length 7 around an occurrence of the other. For the filtering we construct a graph, in which each vertex corresponds to a PWM m and is assigned a weight $-\log p_m$, where p_m is its p-value as computed in the previous stage. We connect by an edge every pair of vertices that correspond to redundant PWMs. We now use a greedy algorithm to find a high weight independent set in this graph. The resulting set of motifs is passed to the next stage.

Next, we search for significant motif clusters. Concentrating on the set of enriched, non-redundant motifs we use our hashing algorithm to look for combinations of these motifs that tend to co-occur in $\mathcal{G}$. We implemented the general version of the algorithm, which does not assume consecutive occurrences of clusters. Each identified motif cluster is scored using the Monte Carlo approach described in the Statistical Framework Section. The p-values of all clusters are Bonferroni adjusted for multiple testing, and only those clusters that pass the 0.05 significance level (after adjustment) are reported.

Last, we filter similar clusters – the output of the previous stage may contain redundant clusters that share some of their motifs and occur at overlapping positions. To address this problem we filter the list of clusters using the same elimination procedure as in the single motif case: Two clusters have an edge in the conflict graph if they share at least two motifs and at least 75% of the occurrences of one cluster overlap with occurrences of the other cluster.

In order to visualize the resulting motif clusters we created a web-based visualization tool available at http://icsi.berkeley.edu/~roded/creme.html. The tool allows the user to specify a list of motif clusters, a set of genes

of interest, a window length and other parameters. Full details are available on the web page.

RESULTS

Data Preparation

We extracted promoter regions for all RefSeq genes in the human genome, restricting attention to promoter segments that are conserved between human and mouse, and motif hits that are included in such segments and are locally conserved. This was shown to greatly reduce false positive hits, while preserving a high fraction of true positives (Loots et al., 2002). In order to obtain conserved promoter segments for the genes, we utilized the browser at http://nemo.lbl.gov/ecrBrowser/, which provides online access to the complete alignment of the human and mouse genomes. For every gene, we examined the overlap of the 1200bp upstream region of the transcription start site with the set of conserved segments. If an overlap was detected, the overlapping segment closest to the gene's transcription start site was extracted for the analysis, provided that its length exceeded 80. In total, 7749 unique human-mouse conserved segments were extracted. Henceforth, we call this set the *promoter set*. A histogram of the lengths of the conserved segments is shown in Figure 2.

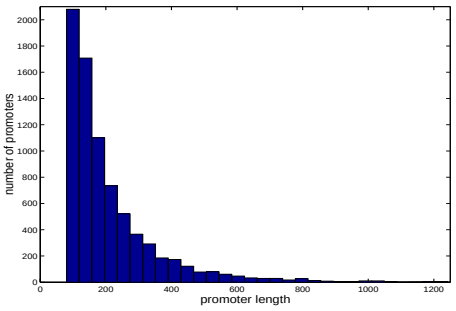

Fig. 2. Histogram of the lengths of 7749 human-mouse conserved segments used in this study.

Next, we searched the promoter set for hits for all 414 vertebrate specific PWMs available in the 6.2 version of TRANSFAC (Wingender et al., 2000). The search was performed using the rVista tool (Loots et al., 2002), which identifies PWM hits that are locally highly conserved, using the parameters 0.75 and 0.85 for matrix and core PWM similarity scores, respectively. 389 of the PWMs had a hit in our promoter set, and about $2,600,000$ hits in total. The position of each hit was set to be the center position of the PWM's core.

Enriched Clusters in Cell-Cycle Regulated Genes

We applied CREME to a set of genes that were shown to be cell-cycle regulated in (Whitfield et al., 2002). This dataset contains the expression profiles of synchronized HeLa cells in five independent experiments. Whitfield et al. (2002) identified in the data 874 genes that are cell-cycle regulated. The analysis was carried out on 336 of these genes that contained a conserved segment in their upstream regions.

Our single motif tests identified 47 significant PWMs. This list included E2F, a key TF in cell-cycle regulation, as well as other TFs (CREB and NFY) that were suggested to be related to cell cycle in Elkon et al. (2003). Further filtering of similar PWMs resulted in a set of 16 enriched (or underrepresented), non-redundant PWMs.

We searched for motif clusters over this set of motifs, using a window of length $w = 100$ and limiting the search to clusters with at most $r = 4$ distinct motifs. The algorithm discovered a total of 1089 motif clusters that had at least 10 hits in the gene set. About 90% of the clusters had at least one consecutive occurrence. These clusters were then scored using the Monte Carlo approach with two spaced instances for each promoter. 13 of the clusters were found to be significant at the 5% level (after Bonferroni adjustment for multiple testing), and further filtering resulted in 7 significant, non-redundant clusters. A list of these motif clusters along with their statistical scores is given in Table 1.

We validated our findings by checking the coherence of every set of genes containing a given significant cluster with expression data. The expression dataset of Whitfield et al. (2002) was obtained using two kinds of arrays. We chose to focus on three experiments (76 conditions) that used the larger array, which contained 265 of our genes, and applied the same data processing methods as described in (Whitfield et al., 2002). Using the approach of Pilpel et al. (2001), we define the *coherence* of a set of genes as the median pairwise similarity in this set excluding self-similarities, where the Pearson correlation coefficient is used as a similarity measure. Let S be a set of genes containing a given cluster, and let U denote the whole set of 336 genes. To score the coherence of S we randomly select 10000 subsets of $|S|$ genes from U, and compute the coherence of each subset. The p-value we assign to S is $\frac{k}{10000}$, where k is the number of subsets whose coherence was higher than that of S. Out of the seven reported clusters, five were found to have a significant coherence score, using a significance level of $\alpha = 0.05$ (Table 1). The coherence is further illustrated in Figure 3, which shows the distribution of pairwise similarities in cluster 673 compared to the background distribution in the whole set of 336 genes. A visualization of the promoters of this cluster is given in Figure 4.

Enriched Clusters Related to Stress Response

As another test of our methodology we looked for statistically significant clusters in a set of genes with

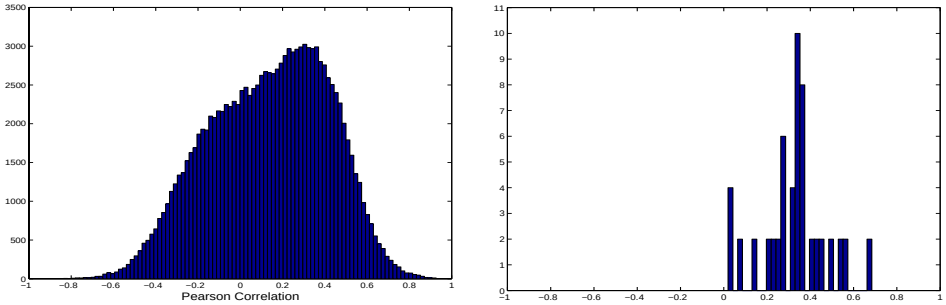

Fig. 3. Left: Histogram of similarity values for all 336 cell cycle regulated genes. Right: Histogram of similarity values for the genes containing cluster 673. The median similarity of genes in the cluster is significantly higher than in the whole set ($p = 0.04$).

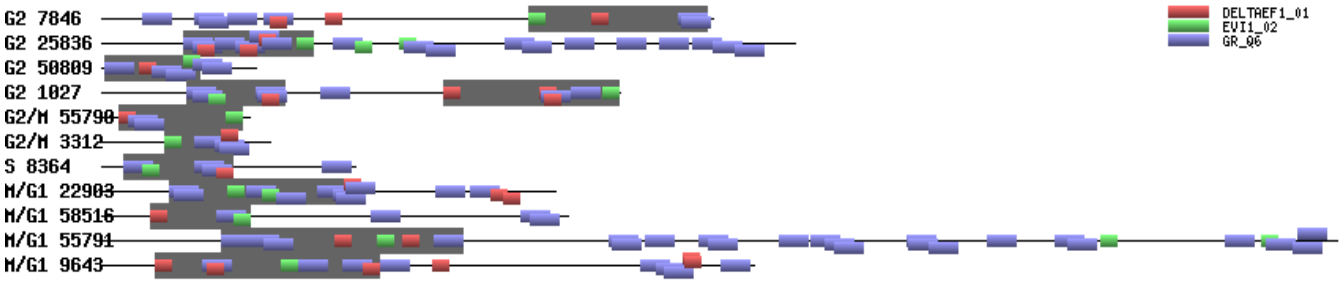

Fig. 4. Color visualization of motif cluster 673. Shown are all the promoters containing this cluster. A LocusLink ID and cell cycle phase assignment are indicated for each promoter. Each motif in the cluster is represented by a colored square with a distinct color. Motif hits are shown only for the motifs that appear in the cluster. Cluster instances are shaded.

similar function, focusing on stress response related genes. We used the following scheme: We applied CREME to genes that are annotated as stress response in the Gene Ontology (GO) database (The Gene Ontology Consortium, 2000), and then scored the resulting clusters based on the enrichment of their genes in sub-categories of this class. Note that the sub-category information is not used in the cluster discovery process.

The stress response class contains 253 genes for which we have promoter data. These genes belong to several sub-categories: Pathogen response (GO:0009613, 159 genes), response to wounding (GO:0009611, 188 genes), inflammatory response (GO:0006954, 83 genes), and humoral response (GO:00006959, 62 genes). Applying CREME to these 253 genes yielded 6 non-redundant clusters with p-values below 0.05 (after Bonferroni adjustment). We checked the enrichment of the genes containing each cluster w.r.t. the four sub-categories, using a standard hypergeometric score. For each gene we took the lowest p-value obtained. Four of the clusters had enrichment p-values below 0.05. These clusters are listed in Table 1. Markedly, one of the clusters (number 1938) obtained a very significant enrichment score ($p = 0.002$) and 17 of its 18 associated genes were GO annotated as related to pathogen response.

CONCLUSION

In this paper we presented a framework for discovering regulatory modules – clusters of transcription factor binding sites. Our framework includes statistical scores for assessing the significance of the discovered clusters, as well as enrichment tests for the motifs that make up these clusters. We use an efficient enumeration approach that can discover all combinations of a bounded number of motifs that appear in a window of certain size. We demonstrate the effectiveness of our approach on two real datasets, discovering modules whose associated sets of genes are coherently expressed or functionally related.

Data	ID	Motifs	Size	p-value	Coherence/Enrichment
CC	11	ZF5, USF2	137	0.01	0.03
CC	355	DELTAEF1, GR, HAND1E47, LMO2COM	30	0.002	0.21
CC	421	GR, ZF5	84	0.003	0.02
CC	489	EVI1, GR, HAND1E47, LMO2COM	12	0.02	0.04
CC	673	DELTAEF1, EVI1, GR	11	0.01	0.04
CC	937	NFY, WHN, ZF5, E2F1	18	0.002	0.09
CC	948	HAND1E47, ZF5	90	0.01	0.02
STRESS	288	HSF1, OCT1, NFKB, XVENT1	12	0.02	0.01 (GO:0006954)
STRESS	1726	NFKAPPAB65, CHOP, STAT5B, TCF1P	12	0.004	0.03 (GO:0009613)
STRESS	1938	OCT1, CEBP, STAT, XVENT1	18	0.02	0.002 (GO:0009613)
STRESS	2035	STAT, CHOP, XVENT1, STAT5B	14	0.0003	0.01 (GO:0009613)

Table 1. Clusters discovered in the cell cycle and stress response datasets. The columns are: dataset, cluster ID, component motifs, size and the Monte Carlo based p-value of its count (corrected for multiple testing). The last column contains the expression coherence p-value for the cell cycle (CC) data and enrichment p-value and GO sub-category for the stress data. Clusters 11, 421 and 948 were significantly underrepresented; all other clusters were overrepresented.

Our methods can be refined in several ways: First, focusing on enriched, non-redundant motifs greatly reduces the number of false positive PWMs, but may also eliminate true positive motifs. To address this concern one can incorporate existing knowledge on regulation, and add biologically relevant PWMs to the list motifs considered. Second, the definition of a motif cluster can be refined by considering the order of motif hits in a cluster and the multiplicity of hits for each motif. This requires designing new statistical measures for motif cluster occurrences. Third, the statistical tests proposed here can be refined by taking into account the scores of the different motif hits. Finally, although our method could be applied under different window lengths (or even under other definitions of a cluster instance e.g., bounding the gap between adjacent hits of its motifs), a general score that measures the significance of the proximity of motif hits within a cluster, allowing a flexible definition of a cluster instance, is yet to be formulated. Some results in this direction were reported by Wagner (1999) and Frith et al. (2002).

ACKNOWLEDGMENTS

We thank Rani Elkon and Chaim Linhart for many fruitful suggestions. R.S. was supported by a Fulbright grant. This work was supported by NIH and performed under Department of Energy Contract DE-AC0376SF00098, University of California, Berkeley.

REFERENCES

Berman, B., Y. Nibu, B. Pfeiffer, et al. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome. *Proc. Natl. Acad. Sci. USA 99(2)*, 757–762.

Elkon, R., C. Linhart, R. Sharan, R. Shamir, and Y. Shiloh (2003). Genome-wide in-silico determination of transcriptional regulation mechanisms controlling cell cycle in human cells. *Genome Res.*. To appear.

Frith, M., U. Hansen, and Z. Weng (2001). Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics 17*(10), 878–889.

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

Halfon, M., Y. Grad, G. Church, and A. Michelson (2002). Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res. 12(7)*, 1019–28.

Hannenhalli, S. and S. Levy (2002). Predicting transcription factor synergism. *Nucleic Acids Res. 30(19)*, 4278–84.

Kel-Margoulis, O., T. Ivanova, E. Wingender, and A. Kel (2002). Automatic annotation of genomic regulatory sequences by searching for composite clusters. In *Pac. Symp. Biocomput.*, pp. 187–98.

Klingenhoff, A., K. Frech, K. Quandt, and T. Werner (1998). Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics 15*(3), 180–186.

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

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

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

Rebeiz, M., N. Reeves, and J. Posakony (2002). SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc. Natl. Acad. Sci. USA 99(15)*, 9888–93.

Stormo, G. (2000). DNA binding sites: representation and discovery. *Bioinformatics 16(1)*, 16–23.

Sudarsanam, P., Y. Pilpel, and G. Church (2002). Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in *saccharomyces cerevisiae*. *Genome Res. 12(11)*, 1723–31.

Tavazoie, S., J. Hughes, M. Campbell, et al. (1999). Systematic determination of genetic network architecture. *Nat. Genet. 22*, 281–285.

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

The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nature Genet. 25*, 25–29.

Wagner, A. (1999). Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics 15*(10), 776–784.

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

Waterman, M. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman and Hall.

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

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

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