



Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm

Ö. Johansson¹, W. Alkema², W. W. Wasserman^{3,*} and J. Lagergren⁴

¹Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093-0114, USA, ²Center for Genomics and Bioinformatics, Karolinska Institutet, SE-171 77 Stockholm, Sweden, ³Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, University of British Columbia, Vancouver, B.C., V5Z 4H4, Canada and ⁴Stockholm Bioinformatics Center and Department of Numerical Analysis and Computer Science, KTH, SE-100 44 Stockholm, Sweden

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

Motivation: The identification of regulatory control regions within genomes is a major challenge. Studies have demonstrated that regulating regions can be described as locally dense clusters or modules of *cis*-acting transcription factor binding sites (TFBS). For well-described biological contexts, it is possible to train predictive algorithms to discern novel modules in genome sequences. However, utility of module detection methods has been severely limited by insufficient training data. For only a few tissues can one obtain sufficient numbers of literature-derived regulatory modules.

Results: We present a novel method, MSCAN, that circumvents the training data problem by measuring the statistical significance of any non-overlapping combination of TFBS in a window. Given a set of transcription factor binding profiles, a significance threshold, and a genomic sequence, MSCAN returns putative regulatory regions. We assess performance on two curated collections of regulatory regions; one each for tissue-specific expression in liver and skeletal muscle cells. The efficiency of MSCAN allows for predictive screens of entire genomes.

Availability: <http://tfscan.cgb.ki.se/cgi-bin/MSCAN>

Contact: wyeth@cmmt.ubc.ca

Keywords: transcription, gene networks, modules, motif, promoter.

INTRODUCTION

A complex regulatory network determines the properties of cells in multicellular organisms. These properties are defined primarily by the composition of the set

of expressed genes and the magnitude of expression of each active gene. One can consider the regulation of gene expression as a program directing the cell to exhibit specific characteristics and perform targeted tasks. While diverse biochemical mechanisms can contribute to the regulation of gene expression, the initial step of transcription appears to be the focal point of control in most cases (Gill, 2001). The transcription rate of genes is dictated primarily by interactions between DNA-binding transcription factors (TFs). The complexity of the regulatory network is enabled by the diverse range of cooperative interactions possible between members of the finite set of transcription factors (Davidson, 2001).

Biochemical and genetic studies have explored the biochemical mechanisms that drive cell-specific transcription, revealing key insights into the design of regulatory programs. While control of transcription rates can be exerted at levels ranging from the three-dimensional structure of chromatin (Labrador and Corces, 2002) to the initiation of transcription by the pol-II enzyme complex (Gill, 2001), the preponderance of available data addresses activation via *cis*-regulatory enhancers proximal to genes. Enhancers are segments of DNA demonstrated to control the rate of transcription initiation through a promoter sequence. Initial studies offering coarse suggestions of boundaries are occasionally followed by detailed analysis delineating minimal functional regions of between 30 and 300 basepairs. Genetic analyses of such regulatory regions indicate that they are generally composed of dense clusters of target binding sites (elements) for TFs (Davidson, 2001). While diverse characteristics have been observed, the clusters, commonly termed modules, typically function independently of sequence orientation

*To whom correspondence should be addressed.

and retain activity across species despite considerable sequence variation. Partially overlapping binding sites have been observed in some cases, although most reported sites occupy discrete segments within regulatory regions.

Computational approaches for the detection of regulatory modules are essential for the accelerated characterization of regulatory programs governing celltype-specific processes. While characterization of regulatory modules is an arduous process requiring amenable assays and extensive laboratory efforts, validation of predictions can be accomplished with relative ease. Detection of regulatory control sequences has emerged as one of the preeminent challenges in the sequence analysis of genomes. Considerable progress has been made in the computational detection of potential TF binding sites, but essentially all such putative elements are functionally inactive, as reviewed in Wasserman and Krivan (2003). The identification of segments of genomic sequence preferentially conserved over moderate periods of evolutionary time, phylogenetic footprinting, has received considerable attention as a means to delineate potential regulatory regions within genes (Duret and Bucher, 1997). However, such approaches offer little information as to the specific function of the conserved sequences. In order for a binding site to mediate a biological function, it must be situated in a suitable sequence context. Most successful efforts in bioinformatics to identify functional regulatory elements have been based on the module model. Early algorithms for the detection of modules employed statistical models that imposed biochemically unfounded restrictions. Such restrictions include rigid spacing rules between sites (Frech *et al.*, 1997) or limitations upon the number of contributing sites for each type of TF (Krivan and Wasserman, 2001; Wasserman and Fickett, 1998). The initial studies have been followed by more flexible discrimination algorithms (Frith *et al.*, 2001, 2002) and biochemical affirmation of the predictive value of module models (Berman *et al.*, 2002; Markstein *et al.*, 2002).

Previous module prediction algorithms have consistently required two key resources: (1) quantitative models for the prediction of binding sites of TFs associated with the regulatory program under study and (2) well-defined sets of known modules for the training of discriminant functions. As molecular techniques can now provide accurate binding models for transcription factors (Liu *et al.*, 2002; Roulet *et al.*, 2002), the laboratory definition of regulatory modules to serve as training sequences for computational models is currently the primary limitation. Analysis of the training process for a liver module model indicates that the coefficients defined in a regression process can be defined with six to seven regulatory modules in the positive training set (Krivan and Wasserman, 2001). Laboratory data is too sparse for all but a few biologically interesting contexts to provide such positive training sets,

thus precluding the broad development and application of previously introduced predictive methods. In most cases, regulatory regions for only one or two genes have been extensively studied to identify the TFs associated with a context.

An alternative to the training of discriminant functions for the detection of regulatory modules is to develop methods based on the statistical significance of clusters of binding sites. While such an approach is unlikely to surpass the performance offered by methods incorporating positive training data, it offers the opportunity to address biologically important contexts as soon as the binding characteristics of contributing TFs are defined.

Here we present a novel method requiring no positive training to assess the significance of clusters of putative TF binding sites. Our algorithm, MSCAN, takes as input a set of TF binding models represented by position weight matrices (PWMs), a significance threshold and a genomic sequence. While creation of PWMs could be viewed as a training process, there are no direct adjustments to the algorithm for any set of user-selected TFs. Using dynamic programming, the most significant cluster of putative binding sites within a window of sequence is identified and those windows are reported for which the significance level is below the threshold.

ALGORITHM

Methods for the detection of transcriptional regulatory modules in genomic sequences must discern short segments containing high local concentrations of regulatory motifs. Within a window of sequence, MSCAN evaluates the combined statistical significance of sets of potential TF binding sites. When the significance level, or *p-value*, falls below a chosen threshold value, a prediction is output. In designing the MSCAN algorithm, there were three key issues to consider: (i) measuring the significance of individual patterns resembling TF binding sites ('hits'), (ii) calculating the combined significance for any given set of the putative TF binding sites ('multi-hits') and (iii) determining the optimal set of motifs for overlapping significant windows.

Computing the significance of putative TF binding sites

Due to the nature of TF protein-DNA interactions, the patterns of target sites are highly complex. Within the bioinformatics community, first order position weight matrices (PWMs) generated from alignments of known binding sites have emerged as a standard tool for quantitative representation of TF binding specificity (Stormo, 2000). While some reports have indicated that higher order models can marginally improve predictive performance (Bulyk *et al.*, 2002; Roulet *et al.*, 2002), binding site data is rarely

of sufficient depth to enable the performance of higher order models to exceed that of a first order PWM.

We use a previous method to compute p -value tables for single hits, this with respect to a model of random sequence where positions are independent and identically distributed (Claverie and Audic, 1996). Yet, genomic sequences can exhibit strongly biased local sequence characteristics. To reduce the number of false positives, we select the distribution for calculation of single hit p -values based on the nucleotide frequencies within the target window. As the p -value tables are somewhat costly to compute however, and as we have many windows to analyze, we first round the observed window nucleotide distribution to one out of a small set of potential distributions. The first time such a rounded distribution is encountered, its single-hit p -value tables are computed and saved in memory.

Single hits on either of the two DNA strands are equally interesting. We now give the details on how single hit p -values are bounded from above. Let f be a fixed nucleotide probability distribution and let S be a DNA sequence generated at random such that forward strand positions are independent and identically distributed according to f . Consider now a PWM M . The forward strand p -value of score x is the probability that M will produce score x or higher when applied at a predetermined position on the forward strand of S . Assuming that we have scaled and rounded the entries of M to integer values (which preserves the interesting features of the PWM provided that the scale factor is large enough), a table of this forward strand p -value for each possible score can be computed efficiently with dynamic programming (Claverie and Audic, 1996). Using the reverse strand probability distribution corresponding to f , we can similarly compute the reverse strand p -value table. Finally, by summing corresponding forward and reverse p -values, we get an upper bound for the ‘either-strand’ p -value for each score of M . Note again the dependency on the nucleotide probability distribution f .

The p -values so obtained are with respect to a single PWM only. Suppose that we observe a hit for M with such a p -value p when M is just one of m different PWMs whose hits we are interested in. Either of these could have had a hit with this p -value or less at the position in question. Hence, mp is an upper bound for the p -value of this hit with respect to the whole PWM set.

Calculating the combined significance of multiple sites

The calculation of the combined significance of multiple motifs within biopolymer sequences remains a significant challenge within computational biology. Diverse approaches have been introduced previously. Several groups have reported methods based on Poisson distributions

of patterns (Wasserman and Fickett, 1998; Wagner, 1999). Below we present a new method for combining the p -values of non-overlapping binding sites, or *hits*, within a window of sequence. We call such a set of hits a *multi-hit*.

A k -hit is a collection of exactly k non-overlapping hits in a window. Let these have p -values p_1, \dots, p_k with respect to the whole set of PWMs. Intuitively, this k -hit will be interesting to us if all of these p -values are small. To capture this, we define the score of the k -hit as the maximum of p_1, \dots, p_k , and we define the k -hit score of a window as the minimum score among all its k -hits. We compute this using dynamic programming.

Notice that there is no ‘canonical’ mathematical method to derive a k -hit score from k individual p -values. On the contrary, this is strictly a modeling issue, and a suggested k -hit score function can only be assessed based upon its sensitivity and specificity with respect to biological data—only intuition and biological knowledge can assist us in the choice of this function. When such a choice has been made, p -values for k -hit scores are well-defined with respect to probabilistic models of random sequences. In our case, these p -values are hard to compute exactly and we are forced to compute upper bounds for them. Scores can then be classified as at least as unlikely as the upper bound.

Intuitively, the p -value of a k -hit with score p is the expected number of non-overlapping windows per sequence length that contain a k -hit with score at most p . To get an upper bound for this, notice that the expected number of positions per sequence length where we have a single hit with p -value p is just p ; then multiply this with the probability that such a hit will be followed by $k - 1$ others. Denote the window size by w , and let l be the minimum PWM width. The quantity we wish to compute is then less than p^k times the number of non-overlapping ways to place $k - 1$ identical hits of length l in a window of size $w - l$. We can write this as

$$p^k C_l(k - 1, w - l), \tag{1}$$

where C_l satisfies the following recursion, which immediately yields a dynamic programming algorithm for it:

$$C_l(k, w) = \begin{cases} w + 1 - l \\ C_l(k - 1, w - l) + C_l(k, w - 1) \\ 0 \end{cases}$$

corresponding to the cases

$$\begin{cases} k = 1 \text{ and } w \geq l \\ k \geq 2 \text{ and } w \geq l \\ w < l \end{cases}$$

This bound is good as long as p is small compared with $1/(kw)$, but gets worse when p is near or above this value.

In the latter case, a substantial number of k -hits with score p or less in a random sequence will in fact be embedded in windows with $k + 1$ or more hits with p -value p or less. Such windows have at least $k + 1$ different k -hits which all contribute to the bound. As one purpose of computing p -values for multi-hits is to allow us to choose a multi-hit of optimal size, we want good approximations of these p -values. Therefore, we only include hits with p -values less than or equal to a ‘use’ threshold t_u . We have found that a use threshold $t_u = 1/(5 * w)$ allows for acceptable approximation without noticeably restricting the set of true modules that we can point out (this with $w = 200$).

Assuming now that we are looking for multi-hits of size up to k_{\max} , we define the *score* of a window as the minimum of $p_1, \dots, p_{k_{\max}}$, where p_k in this case denotes the minimum p -value of a k -hit in the window. As this is a choice of optimal multi-hit size, we have to compute a p -value for the result. We define the p -value of a window score p as the expected number of non-overlapping windows per sequence length with score at most p . An upper bound for this is given by $\sum_{k=1}^{k_{\max}} \min(p, q_k)$, where $q_k = (t_u)^k C(k - 1, w - l)$, i.e. where q_k is the maximum possible p -value of a k -hit, given the hit use threshold t_u .

In summary, we first compute the p -values for the different PWM scores at each position within the window. For each k from 1 up to k_{\max} , we then find an optimal k -hit. (The value of k_{\max} should be large enough to be non-critical; we have used the value 10.) Among the optimal multi-hits, we choose that with the lowest p -value, and we compute the window p -value from the result. If this is less than the selected significance threshold, a prediction is output, containing the window sequence, as well as the chosen multi-hit and the window p -value. When such windows overlap however, the combined sequence is presented together with the best (most significant) of these windows: its location, multi-hit, and p -value.

As already mentioned, we find the p -values of PWM scores using tables computed for a collection of fixed nucleotide probability distributions. We have tried two ways of defining these collections. In the more ideal of the two, we round the observed window nucleotide frequencies to the nearest number in a series a^{-2i} , where a is the accuracy, and i is an index between 0 and a suitable maximum value. We then scale the resulting numbers so that they sum to one. In experiments with random sequences generated using various fixed nucleotide distributions, we have seen that with this rounding scheme, MSCAN typically produces around 70 to 80 percent as many predictions as corresponds to the significance threshold when this is set to 10^{-5} . This underprediction stems from the fact that the p -values we compute are not exact but only upper bounds for

the true p -values with respect to the assumed nucleotide distribution.

An intrinsic problem with the above rounding scheme though is that it leads to a large collection of tables, thus requiring quite a lot of memory. In general therefore, we have instead rounded the combined AT- and CG-frequencies, scaled the result, and then divided evenly between A and T, and between C and G. A potential problem with this scheme is that in case a genomic sequence displays a marked bias in favor of C on one strand and G on the other, or similarly for A and T, the p -value tables (based on rounded unbiased distributions) are less relevant, and may lead to overprediction. Indeed, this is easy to see with artificially generated biased sequences. However, in our experience with real sequences (the *Fugu* genome, see below), the difference between the two rounding schemes has been small; for both of them we have seen approximately 8 times ‘too many’ predictions at the 10^{-5} significance level. This should not be too surprising though, as real genomes are not generated according to idealized nucleotide distributions.

As discussed above, it is not self-evident how to define the score (and hence the p -value) of a k -hit. A natural alternative to the approach we have described is to define the score of a k -hit as the product of the involved single hit p -values. The calculation of the p -value for this score will then be different. The derivation involves a multi-integral, and the result is that the factor p^k in (1) changes into

$$p \sum_{i=0}^{k-1} \frac{(-\ln(p/(t_u)^k))^i}{i!},$$

where p now denotes the product (the k -hit score) and t_u is the previously described ‘use threshold’ for single hit p -values. (A derivation in case of no such threshold can be found in Bailey and Gribskov (1998), and the expression above can easily be obtained by substitution in the integral formulation.) Although we initially did not see any improvement when using this alternative approach, Figure 2 shows that it might be at least as good.

IMPLEMENTATION

Predictive performance

The predictive performance of methods for the detection of regulatory modules is measured in terms of sensitivity and specificity. Specificity is measured as the rate of prediction in long genomic sequences, under the assumption that true positives are sufficiently rare that most predictions will be false. Sensitivity is commonly assessed by the percentage of positive examples identified for a richly studied tissue. For the vast majority of biological contexts, few genes have been studied extensively to characterize regulatory control sequences. The arduous biochemical procedures required to generate such data are

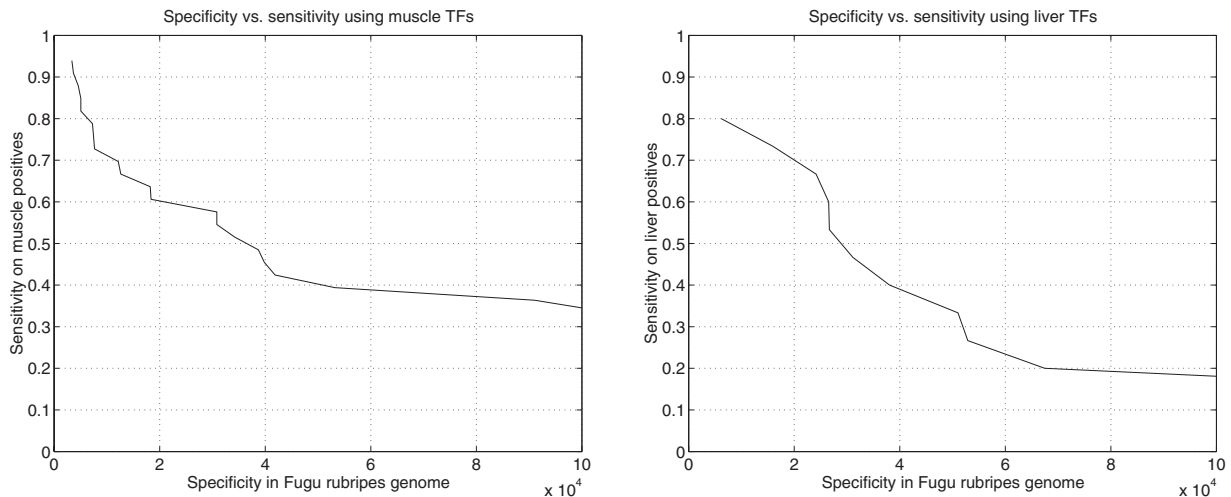


Fig. 1. MSCAN sensitivity versus specificity for the muscle and liver PWM sets. The sensitivity axis measures the fraction of known regulatory regions detected in the muscle and liver sequence collections. The specificity axis shows the average distance between predictions in the unmasked *Fugu rubripes* genome (v2.0, 332Mbp) at significance thresholds leading to the indicated sensitivities.

prohibitive, limiting assessments to the few exceptional cases.

Skeletal muscle An elegant tissue culture model for the differentiation of skeletal muscle cells has enabled extensive biochemical annotation of regulatory control sequences for a relatively large number of genes. An initial database of regulatory regions has emerged as a reference set for cluster detection algorithms (Wasserman and Fickett, 1998). The regulatory modules primarily contain binding sites for five critical TFs: myoD-family/myf, mef2, SRF, Tef and Sp1. While several of the TFs participate in the regulation of transcription in multiple contexts, certain combinations and arrangements of binding sites for these factors are sufficient to direct transcription specifically to mature skeletal muscle. Binding models for the TFs were taken from the literature (Wasserman and Fickett, 1998). The MSCAN algorithm was applied to the module collection and the *Fugu rubripes* genome sequence. This pufferfish genome is an attractive target for cluster analysis as the regulatory controls are in many cases similar to those for humans while the sequence is only 10 percent of the length. Specificity rates of 1 cluster detected every 12 and 15 kbp respectively in the *Fugu* unmasked and repeat-masked genome were observed at a module p -value threshold sufficient to identify 66 percent of true positives. See Figure 1.

Liver hepatocytes Hepatocytes are one of the few classes of human cells which can be effectively grown and manipulated in primary cell culture experiments. Furthermore, a widely studied tumor cell line (HepG2) retains many of the

characteristics of primary hepatocytes. These experimental resources have been effectively used to identify regulatory regions sufficient to direct gene expression specifically in hepatocytes. A set of four key TFs have been identified that primarily contribute to hepatocyte-specific gene expression: HNF-1, HNF-3, HNF-4 and C/EBP. Additional TFs have been linked to the regulation of some genes, but the above four appear to be the only broadly important factors.

It should be stressed that, although the muscle data set was analyzed periodically during the design of the MSCAN algorithm, MSCAN requires no training. In the final implementation, the algorithm requires only PWMs to define the TF binding sites. For instance, the liver regulatory modules were not screened until the final MSCAN algorithm was implemented. The results, consistent with those observed for skeletal muscle, indicates that the method functions well in the absence of training data. The observed prediction rates with the liver PWMs is comparable to the muscle results; 1 cluster detected every 22 and 23 kbp respectively in the *Fugu* unmasked and repeat-masked genome at a module p -value threshold correctly classifying 66 percent of true positives. See Figure 1.

A method to account for direct repeats in genome sequences

In reviewing the putative modules detected in the pufferfish genome with the skeletal muscle-linked TF binding models, it was immediately apparent that the significance values were inconsistent with the number of modules observed. Many modules contained uniformly spaced hits

Table 1. MSCAN compared with other methods

Method	Data	Sensitivity	Specificity	Reference
LRA	muscle	66 %	1 per 32 kb	Wasserman and Fickett (1998)
LRA	liver	62 %	1 per 35 kb	Krivan and Wasserman (2001)
Cister	muscle	59 %	1 per 35 kb	Frith <i>et al.</i> (2001)
COMET	muscle	59 %	1 per 29 kb	Frith <i>et al.</i> (2002)
MSCAN	muscle	60 %	1 per 18 kb	
MSCAN	liver	60 %	1 per 26 kb	

Two data collections have been analyzed by cluster detection tools. The muscle set (29 sequences) from Wasserman and Fickett (1998) and the liver set (16 sequences) from Krivan and Wasserman (2001) refer to experimentally defined regulatory clusters mapped to regions of less than 200 bp length. LRA, Cister and COMET specificities refer to repeat-masked portions of the human genome whereas MSCAN specificities refer to the unmasked *Fugu* genome

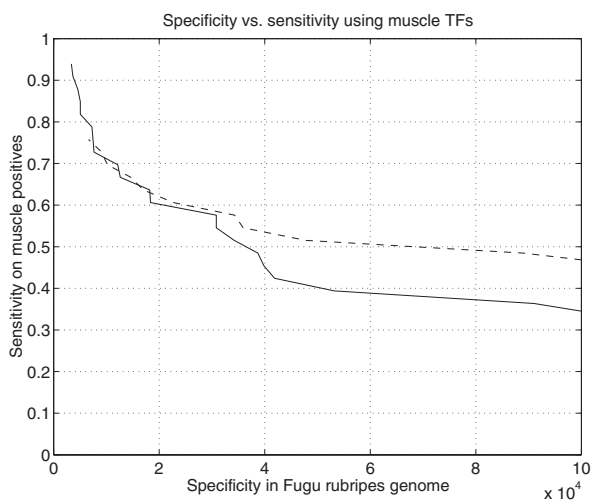


Fig. 2. Sensitivity versus specificity using two different ways to define the score and p -value of a k -hit. The solid line illustrates our main approach, in which the score of the k -hit is defined as the maximum p -value among the k involved single hits. The dotted line illustrates the alternative approach in which the score is defined as the product of the k single hit p -values.

of exactly the same p -value for only one of the TFs. Upon further review, these TF-specific regions were composed of short, direct repeats which resemble TF binding sites. As some of these regions may be of biological relevance, it was decided to leave the results for the interpretation of researchers. However, a post-analysis step was added to mark those regions containing an overabundance of hits with identical sequences that also extend a few base pairs upstream and downstream of what is 'covered' by the PWM itself.

Comparison to previously reported methods

Diverse approaches have been developed for the discrimination of regulatory modules. The LRA model analyzed only a single site for each of five TFs within a

fixed 200 bp window. A more flexible approach based on hidden Markov models (HMMs) was introduced to detect any combination of binding sites. Even though the MSCAN algorithm involves no training, it is comparable in performance to the previously published methods. See Table 1.

MSCAN server

To enable broad access to MSCAN, a web-server has been implemented. Researchers may scan a given genomic sequence with any specified set of TF binding models. Users can also elect to activate the screening function, which will mark multi-hits that are likely to result from short, locally repeated sequences. The program is available at <http://tfscan.cgb.ki.se/cgi-bin/MSCAN>.

DISCUSSION AND CONCLUSIONS

The flexible MSCAN algorithm enables the identification of significant clusters of binding sites for any combination of transcription factors in the absence of training data. This advance in the analysis of regulatory modules enables computational approaches for the identification of sets of co-regulated genes to be implemented at the earliest possible stage—when the set of mediating TFs are initially defined. By circumventing the requirement for extensive training data, the MSCAN algorithm promises to accelerate characterization of the regulatory control regions within genomes.

Of course, the choice of PWMs will have a significant impact upon the performance of the method. A biologically ill-conceived set of TFs will detect significant clusters, but their relevance would be difficult to assess. Furthermore, PWMs that fail to adequately represent the binding characteristics of a TF will limit the overall quality of predictions. Yet this does not diminish the fact that no parameters are adjusted in applying MSCAN. We stress that PWMs are requirements for the application of MSCAN, and we point to several emerging methods for accelerated generation of profiles.

We have introduced a novel method to assess the significance of clusters of putative transcription factor binding sites within a genomic sequence. The underlying advance is a process to assess the significance of combinations of TF binding sites. While the statistics for assessing the significance of multiple motifs within a biopolymer sequence remains an area of active research, we have introduced a robust means to establish an upper bound for the p -value of any combination of motifs within a fixed length window. This enables the selection of the most significant combination of patterns. There is no overlap here with Lenhard and Wasserman (2002), which introduces a set of Perl-based programming tools for open-source development.

While the theory is the critical advance, the application of the method on a genome scale has indicated a somewhat unanticipated challenge. Short, local repeats of sequences can result in a region scoring as highly significant. While such repeats may be biologically significant, it is also possible that they represent an aberration of genome sequencing methods. We therefore introduced a process to identify site clusters associated with such regions. The MSCAN method and the repeat filtering function have been provided as a web service for community access.

Methods for the discrimination of regulatory modules have utilized diverse procedures. Early methods employed restrictive methods requiring either highly specific rules for the spacing between binding sites (Fickett, 1996; Frech *et al.*, 1997) or limitations on the number of binding sites for any single factor (Krivan and Wasserman, 2001; Wasserman and Fickett, 1998). A hidden Markov model-based approach demonstrated improved predictive specificity, but still required rich data resources for the training process (Frith *et al.*, 2001). A new generation of methods have been reported, but they appear particularly oriented towards clusters of sites for a single factor (Berman *et al.*, 2002; Wagner, 1999). A recently described method allows for the detection of site clusters with reduced training requirements (Frith *et al.*, 2002), but it does not account for the significance of the clusters. This precludes the comparison of modules predicted with slightly different combinations of TF binding profiles.

The MSCAN algorithm lays the foundation for rapid advances in the characterization of regulatory sequences. To achieve this goal, there is a dramatic need for data describing the binding specificities of transcription factors. The challenge of gene-specific studies suggests that high-throughput genomic approaches are required to characterize the binding properties of TFs. Both SELEX-based methods (Roulet *et al.*, 2002) and chromatin immunoprecipitation techniques (Lee *et al.*, 2002) can be broadly applied to surpass this remaining limitation on the characterization of regulatory control sequences. Given that most TFs fall within a relative small set of structural classes and,

with some key exceptions, most members of each structural class bind to similar target sites, it should be possible to characterize the *in vitro* binding characteristics of most TFs in short order.

The breadth of importance of regulatory modules remains to be established. While studies in frogs, flies and man have demonstrated the utility of computational approaches based on TF binding site clusters, there remains an open question regarding how far the model may extend. Most examples have been drawn from tissue-specific expression, while inducible expression by environmental or physiological signals may act through regulatory sites with less modular structure. It may be in some specific cases that the three dimensional local structure of DNA in the nucleus (chromatin) is the principal factor in gene expression and modulating regulatory modules play little or no role. The role of site clusters in diverse expression context will require continued study.

The analysis of regulatory modules offers an important means to link sets of genes which participate in linked biological processes. MSCAN will accelerate these studies by removing the previous limitation for large sets of characterized regulatory modules for the training of statistical models. By determining the significance of clusters of sites, it becomes possible to apply different combinations of factors.

REFERENCES

- Bailey, T.L. and Gribskov, M. (1998) Combining evidence using p -values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
- Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M. and Eisen, M.B. (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**, 757–762.
- Bulyk, M.L., Johnson, P.L. and Church, G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
- Claverie, J.-M. and Audic, S. (1996) The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.*, **12**, 431–439.
- Davidson, E.H. (2001) *Genomic Regulatory Systems: Development and Evolution*, San Diego, Academic Press.
- Duret, L. and Bucher, P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, **7**, 399–406.
- Fickett, J.W. (1996) Coordinate positioning of MEF2 and myogenin binding sites. *Gene*, **172**, GC19–32.
- Frech, K., Danescu-Mayer, J. and Werner, T. (1997) A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *J. Mol. Biol.*, **270**, 674–687.
- Frith, M.C., Hansen, U. and Weng, Z. (2001) Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, **17**, 878–889.

- Frith,M.C., Spouge,J.L., Hansen,U. and Weng,Z. (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.*, **30**, 3214–3224.
- Gill,G. (2001) Regulation of the initiation of eukaryotic transcription. *Essays Biochem.*, **37**, 33–43.
- Krivan,W. and Wasserman,W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, **11**, 1559–1566.
- Labrador,M. and Corces,V.G. (2002) Setting the boundaries of chromatin domains and nuclear organization. *Cell*, **111**, 151–154.
- Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
- Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.
- Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.
- Markstein,M., Markstein,P., Markstein,V. and Levine,M.S. (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl Acad. Sci. USA*, **99**, 763–768.
- Roulet,E., Busso,S., Camargo,A.A., Simpson,A.J., Mermod,N. and Bucher,P. (2002) High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.*, **20**, 831–835.
- Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
- Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.
- Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.
- Wasserman,W.W. and Krivan,W. (2003) Regulation. *Naturwissenschaften*, in press.