*Nucleic Acids Research, 2004, Vol. 32, Web Server issue* ***W195–W198***
*DOI: 10.1093/nar/gkh387*

# MSCAN: identification of functional clusters of transcription factor binding sites

**Wynand B.L. Alkema, Öjvind Johansson[1], Jens Lagergren[2] and Wyeth W. Wasserman[3,*]**

Center for Genomics and Bioinformatics, Karolinska Institutet, SE-17177 Stockholm, Sweden, [1]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093-0114, USA, [2]Stockholm Bioinformatics Center and Department of Numerical Analysis and Computer Science, KTH, SE-10044 Stockholm, Sweden and [3]Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, B.C. Children's Hospital, University of British Columbia, Vancouver, V5Z 4H4 Canada

Received February 12, 2004; Revised and Accepted March 24, 2004

## ABSTRACT

**Identification of functional transcription factor binding sites in genomic sequences is notoriously difficult. The critical problem is the low specificity of predictions, which directly reflects the low target specificity of DNA binding proteins. To overcome the noise produced in predictions of individual binding sites, a new generation of algorithms achieves better predictive specificity by focusing on locally dense clusters of binding sites. MSCAN is a leading method for binding site cluster detection that determines the significance of observed sites while correcting for local compositional bias of sequences. The algorithm is highly flexible, applying any set of input binding models to the analysis of a user-specified sequence. From the user's perspective, a key feature of the system is that no reference data sets of regulatory sequences from co-regulated genes are required to train the algorithm. The output from MSCAN consists of an ordered list of sequence segments that contain potential regulatory modules. We have chosen the features in MSCAN such that sequence and matrix retrieval is highly facilitated, resulting in a web server that is intuitive to use. MSCAN is available at http://mscan.cgb.ki.se/cgi-bin/MSCAN.**

## INTRODUCTION

Knowledge of the mechanisms of gene regulation can be of crucial importance in unraveling biological processes within cells. An important mechanism for the regulation of gene expression is the sequence-specific binding of transcription factors (TFs) to regulatory regions of genes, thereby repressing or activating transcription. The computational identification of functional TF binding sites is problematic. In general, methods that scan sequences for matches to a consensus binding site produce high false positive rates due to the low specificity of most of the profiles and the vast stretches of genomic sequences that have to be scanned (typically spanning at least several kilobases for human genes). A significant increase in predictive specificity has been achieved by methods that identify clusters of binding sites, rather than isolated motifs (1–7). These methods build on the biological evidence that many functional regulatory regions are composed of dense sets of TF binding sites (8).

MSCAN is an algorithm that identifies DNA segments that contain clusters of putative TF binding sites. The MSCAN algorithm for detection of site clusters has been validated on sets of genes that are expressed in muscle and liver cells (7). In short, MSCAN analyzes a subsequence (window) of user-specified length. The window is slid over a DNA sequence with a given window step size, and putative TF binding sites (hits) are identified within the specified window. For each window that contains multiple predicted binding sites, a score is calculated as follows. For each hit, the individual *P*-value is calculated based on the local nucleotide distribution of the sequence within the window. The *P*-value of an individual binding site is the expected frequency of equivalent or better sites in a random sequence that is similar in nucleotide composition to the sequence in the window. This *P*-value is adjusted for the number of TFs used in the search. The combined *P*-value for a window containing multiple hits is calculated using an intermediate score of the set of sites in the window, and then defined as the expected frequency of equivalent or better windows in a random sequence. Here, overlapping windows are only counted once. In principle, this means that for a random sequence (whose nucleotide content would be allowed to vary slowly), the algorithm should produce around one prediction per Mbase with a *P*-value less than $10^{-6}$. On real gene sequences, the number of

*To whom correspondence should be addressed. Tel: +1 604 875 3812; Fax: +1 604 875 3819; Email: wyeth@cmmt.ubc.ca

predictions is typically higher. When the probability of a cluster in a window is below a given *P*-value threshold, it is regarded as a regulatory region. Multiple overlapping windows that score below the *P*-value threshold are fused to create a larger regulatory region.

Here we describe a new web service that facilitates the application of the MSCAN algorithm to user-specified genes. While there are multiple web resources providing access to algorithms for the detection of modules of TF binding sites (3,4,6), each system includes a specific combination of features based in part on the underlying algorithms. The MSCAN server enables detailed analysis of a single sequence, as well as a mechanism for automated analysis of larger sequence sets. It is based on an open-access database of TF binding profiles, allowing all interested scientists unrestricted access.

## INPUT

### Transcription factor binding site models and sequences

The input for the MSCAN algorithm is a set of one or more TF binding site models and a set of sequences of interest (Figure 1). The binding site models are given as position frequency matrices (PFMs). Each PFM is a matrix consisting of four rows and a number of columns that is equal to the length of the binding site. Each column lists the frequency for each of the four nucleotides observed in an alignment of known binding sites at the indicated position in the profile.

Users may directly input or upload their own matrices, or a selection can be made from publicly available databases. Matrices can be selected directly from the JASPAR database (9), an open-access TF binding profile database that contains over a hundred binding profiles for eukaryotic transcription factors. Alternatively, files with matrix identifiers referring to the JASPAR database or the publicly available portion of the TRANSFAC database (10) can be uploaded. MSCAN reports only the highest-scoring regulatory module in a given sequence and is therefore ideally suited to explore regulatory modules in sequences for which some biological knowledge about important transcription factors exists. Extensions of the approach that are less stringently dependent upon prior knowledge are under development.

The gene sequences can be input directly, uploaded or specified to specific chromosome positions by the user. In order to facilitate sequence input, options are provided to

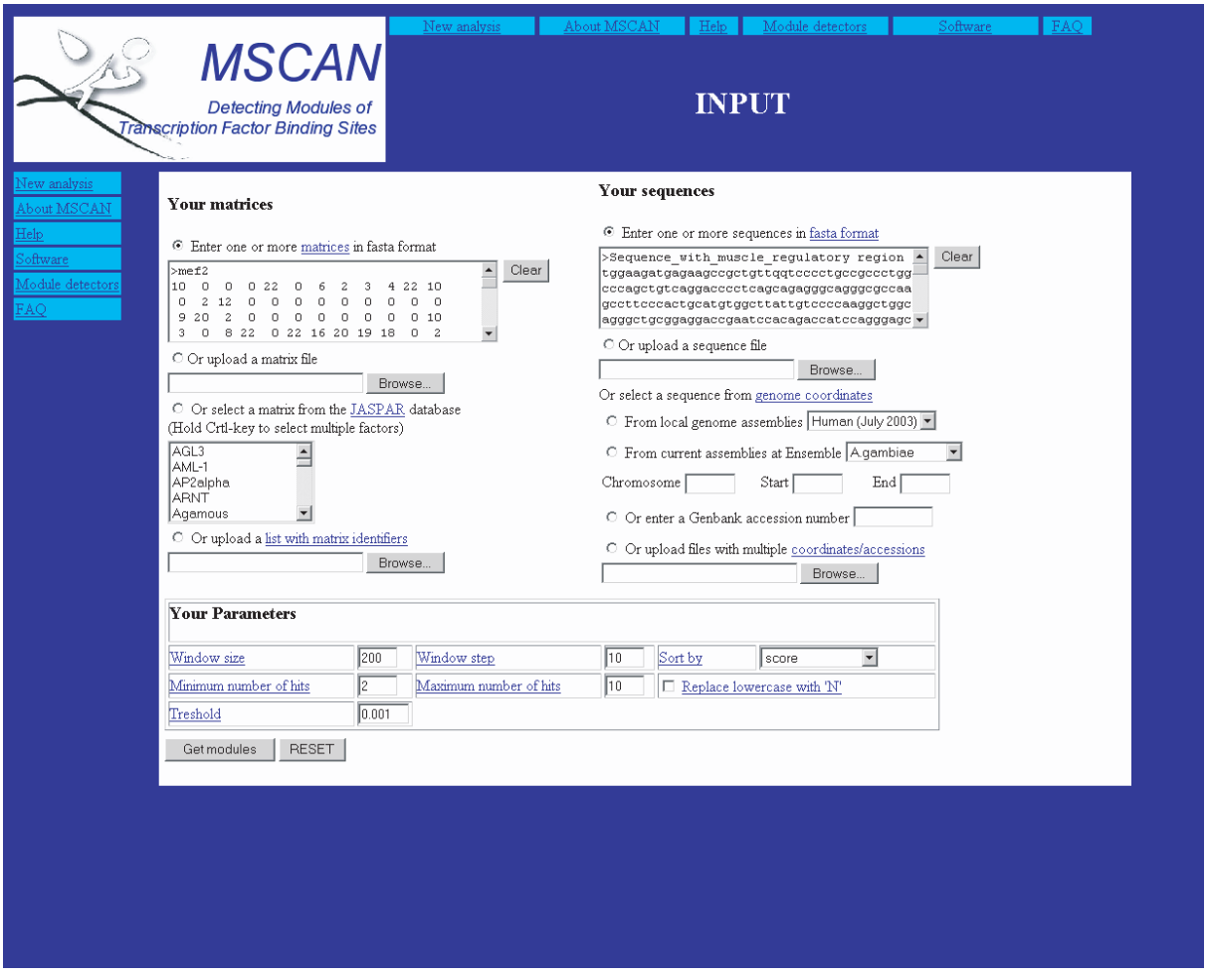


**Figure 1.** Screenshot of the MSCAN data input page. The initial input form contains fields for the provision of TF binding profiles (matrices), sequences and parameters. Links to the help file provide users with information about formats and description of the parameters.

retrieve sequences from GenBank or from specified coordinates in the current assemblies available via the Ensembl service (11). For fast access, sequences from the mouse and human genomes are available via a database on the MSCAN server. For the simultaneous analysis of sequences of multiple organisms, the user can upload a file with genome coordinates or a list of GenBank identifiers. The user has also an option to replace lowercase characters by 'N's. This option may be useful when repeat masked sequences with low-complexity regions marked as lowercase characters are sent to the server.

In order to acquaint users with the input formats and a typical MSCAN output, the input page is preloaded with matrices for muscle-specific transcription factors, Myf, Sp1, Mef2, Srf and Tef, together with experimentally verified sequences that contain muscle-specific regulatory modules. Users working on their own sequences must first clear the sample sequences.

### Parameters

MSCAN does not require positive and negative training sets, as the significance of putative modules is calculated based only on the significance of the observed motifs corrected for local nucleotide composition. The number of modules reported by MSCAN is primarily dependent upon the user-defined score threshold for the window. Other adjustable parameters include window size, the window step size, and the minimum and maximum number of non-overlapping hits that are considered for calculating the *P*-value of the window. These later parameters can be tuned, but they should not be modified without specific biological motivation.

## OUTPUT

The output generated by MSCAN is a set of putative regulatory modules (Figure 2). These can be sorted according either to score or to position within the genomic sequence. For each module, the location and sequence data are reported, as well as the location and score of the most significant window and the individual binding sites therein. This information can be displayed either graphically or in a tabular format.

When MSCAN is applied to sequences which have not been masked for low-complexity repeats, such repeats may result in the generation of numerous predicted modules of significance. Typically, these repeat-containing modules are characterized by a high number of motif matches for the same transcription factor, separated by a consistent distance. In order to discriminate between such spurious modules and modules more likely to serve a biological function, each transcription factor binding site in the highest-scoring window is assessed for its presence within a repeat. To this end, the base pairs flanking

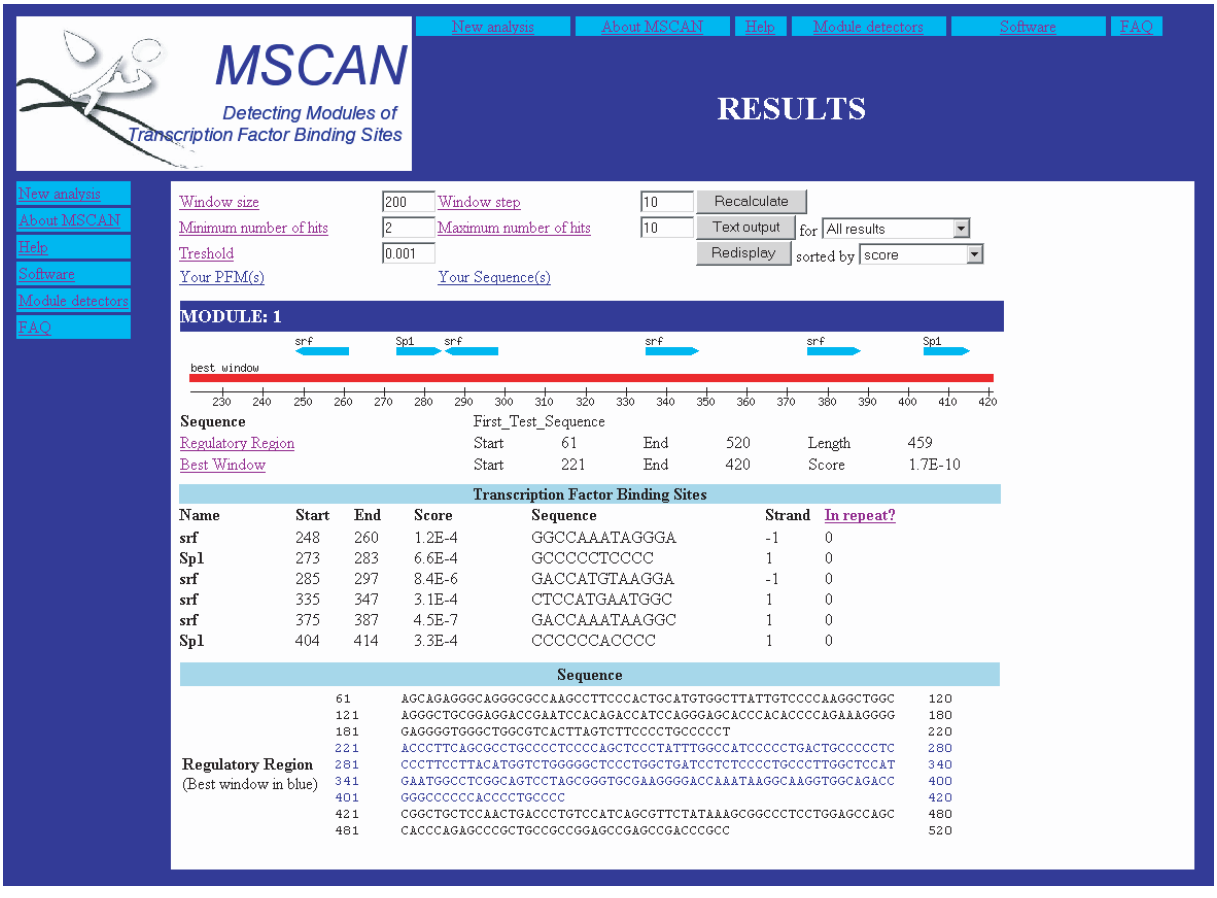


**Figure 2.** Screenshot of the output page. Shown is a page with a predicted module in the regulatory region of a gene known to be selectively expressed in a context linked to the selected TFs. Every category label on the output page is hyperlinked to the help page.

each site are compared with each other. Identical sites with identical flanking sequence which are located on the same strand are labeled as likely to be associated with repetitive sequence.

On the results page, users may select from options based on the initial results. A refined search with adjusted parameter settings can be launched or results can be displayed with different sorting criteria. Links to the actual PFMs and sequences that were used in the analysis are available. This information is useful when matrices and sequences are retrieved from databases instead of directly input by the user. Furthermore, several options are available to redisplay the results in different formats. These include FASTA-formatted sequences for the regulatory regions or the best windows in the regulatory regions, and text files containing the results.

## SOFTWARE AND AUTOMATED ACCESS

The MSCAN algorithm is written in JAVA. A typical computation on the MSCAN server, searching 1 Mb of sequence with five different matrices and default parameter settings takes ~10 s. Currently MSCAN accepts sequences up to 10 Mb for analysis. Analyses of entire genome sequences or sequences of pre-compiled sets of human promoters will be offered in a later stage or alternatively may be performed by users with the standalone version of MSCAN. Perl modules for remote access to the web server and parsing of the MSCAN results into Bioperl (12) and TFBS (13) compatible objects are available from the authors upon request.

## ACKNOWLEDGEMENTS

This work was supported by research funding from the Canadian Institutes of Health Research (W.W.W.) and a Marie Curie Fellowship no. MCFI-2002-01638 from the European Commission (W.A.).

## REFERENCES

1. Rebeiz,M., Reeves,N.L. and Posakony,J.W. (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**, 9888–9893.
2. Bailey,T.L. and Noble,W.S. (2003) Searching for statistically significant regulatory modules. *Bioinformatics*, **19** (Suppl. 2), II16–II25.
3. Frith,M.C., Li,M.C. and Weng,Z. (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, **31**, 3666–3668.
4. Sharan,R., Ovcharenko,I., Ben-Hur,A. and Karp,R.M. (2003) CREME: a framework for identifying *cis*-regulatory modules in human–mouse conserved segments. *Bioinformatics*, **19** (Suppl. 1), I283–I291.
5. Aerts,S., Van Loo,P., Thijs,G., Moreau,Y. and De Moor,B. (2003) Computational detection of *cis*-regulatory modules. *Bioinformatics*, **19** (Suppl 2), II5–II14.
6. Sosinsky,A., Bonin,C.P., Mann,R.S. and Honig,B. (2003) Target Explorer: an automated tool for the identification of new target genes for a specified set of transcription factors. *Nucleic Acids Res.*, **31**, 3589–3592.
7. Johansson,Ö., Alkema,W., Wasserman,W.W. and Lagergren,J. (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, **19** (Suppl. 1), I169–I176.
8. Davidson,E.H. (2001) *Genomic Regulatory Systems: Development and Evolution*. Academic Press, San Diego.
9. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32** (Database issue), D91–D94.
10. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
11. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al*. (2004) Ensembl 2004. *Nucleic Acids Res.*, **32** (Database issue), D468–D470.
12. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al*. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
13. Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.