3572–3575 *Nucleic Acids Research, 2003, Vol. 31, No. 13*
DOI: 10.1093/nar/gkg511

SiteSeer: visualisation and analysis of transcription factor binding sites in nucleotide sequences

Paul E. Boardman, Stephen G. Oliver[1] and Simon J. Hubbard*

Department of Biomolecular Sciences, University of Manchester Institute of Science and Technology, PO Box 88, Manchester M60 1QD, UK and [1]School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Manchester M13 9PT, UK

Received February 13, 2003; Revised and Accepted March 11, 2003

ABSTRACT

The regulation of gene expression is a fundamental process within every living cell, which allows organisms to manage the precise levels of functional gene products with high sensitivity. It is well established that specific DNA sequences located upstream of the transcriptional start site are important in facilitating the binding of regulatory proteins that control the transcription of the gene. Indeed, microarray-based studies have successfully mined the upstream regions of co-expressed genes and discovered over-represented sequences corresponding to known promoter sites. Here we describe a tool for the visualisation of mapped transcription factor binding sites in the upstream regions of either single or grouped eukaryotic genes, which allows users to examine the positions of known and user-defined sites (http://rocky.bms.umist.ac.uk/SiteSeer/). SiteSeer allows the user to map different sections of the TRANSFAC and SCPD databases (or a set of user-defined sites) onto nucleotide sequences. Additionally, users may restrict the analysis by expectation values for certain DNA words as well as by known binding sites specific to a given organism. We believe this tool will prove particularly valuable for biologists who wish to examine sets of co-expressed or functionally-related genes and those who wish to visualise the positions of promoter sequences and generate displays for publications.

INTRODUCTION

The process by which transcription factor proteins bind DNA and regulate gene function on a genome-wide scale is still not well understood. The analysis of eukaryotic transcriptional regulation has concentrated mainly on activator proteins that have a positive effect on transcription when bound to DNA control elements (called enhancers or promoters). Many activators (or transcription factors) have been identified that are specific for genes or gene families and, typically, couple transcription to the physiological needs of the cell (1).

One way in which it is thought that activators exert a positive effect on transcription is by recruiting chromatin-modifying complexes that relieve nucleosomal repression via histone acetylation. The reverse mechanism has also been observed for a set of proteins called repressors, which have a negative effect on transcription by re-establishing nucleosome repression (2). The identification of these control sequences is therefore an important step in elucidating the mechanism of transcriptional regulation. Working from the assumption that the cell possesses a common control mechanism for genes of similar function, one would expect to find common control sequences in co-expressed genes. This allows the cell to initiate the expression of a whole range of genes that are required for a particular function, such as enzymes in a common metabolic pathway, and permits gene expression to be regulated in an efficient, concerted fashion. This concept has been exploited by many groups who have searched for common sequences in the upstream regions of co-expressed genes, typically defined by clustering microarray-based gene expression data (3–6). These sequences may be previously determined binding sites (3) or sequences that are predicted to be transcription factor binding sites due to the significance of their occurrence within the upstream regions of the co-expressed genes (4,7). In this way, potential common control sequences have been identified for many functionally related genes.

Motivation

As part of ongoing research in our laboratory, we wished to develop a tool, which allowed us to directly visualise the occurrence (or otherwise) of known transcription factor binding sites in the upstream region of related yeast genes. We were interested in the co-occurrence of such sites, not only in yeast genes that were co-expressed in microarray experiments, but also in genes that share some definable common functional properties. These include genes that are annotated with common functions or placed in common functional categories [such as those defined by the MIPS (8) and KEGG (9) databases], share common locations on the genome or related gene ontological descriptions, or represent

*To whom correspondence should be addressed. Tel: +44 1612008930; Fax: +44 1612360409; Email: simon.hubbard@umist.ac.uk

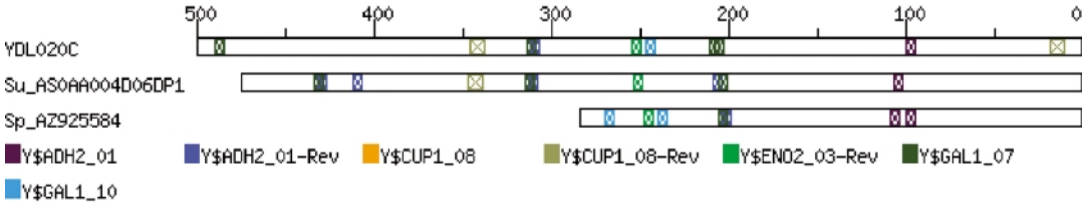

Figure 1. Visualisation of the upstream regions of the *Saccharomyces cerevisiae* gene YDL020C and two orthologues from the Génolevures database (http://cbi.labri.fr/Genolevures/). The output shows a high degree of conservation in binding site composition and position between these three species (*S.cerevisiae*, *Saccharomyces bayanus* var. *uvarum*, and *Saccharomyces paradoxus*, respectively).

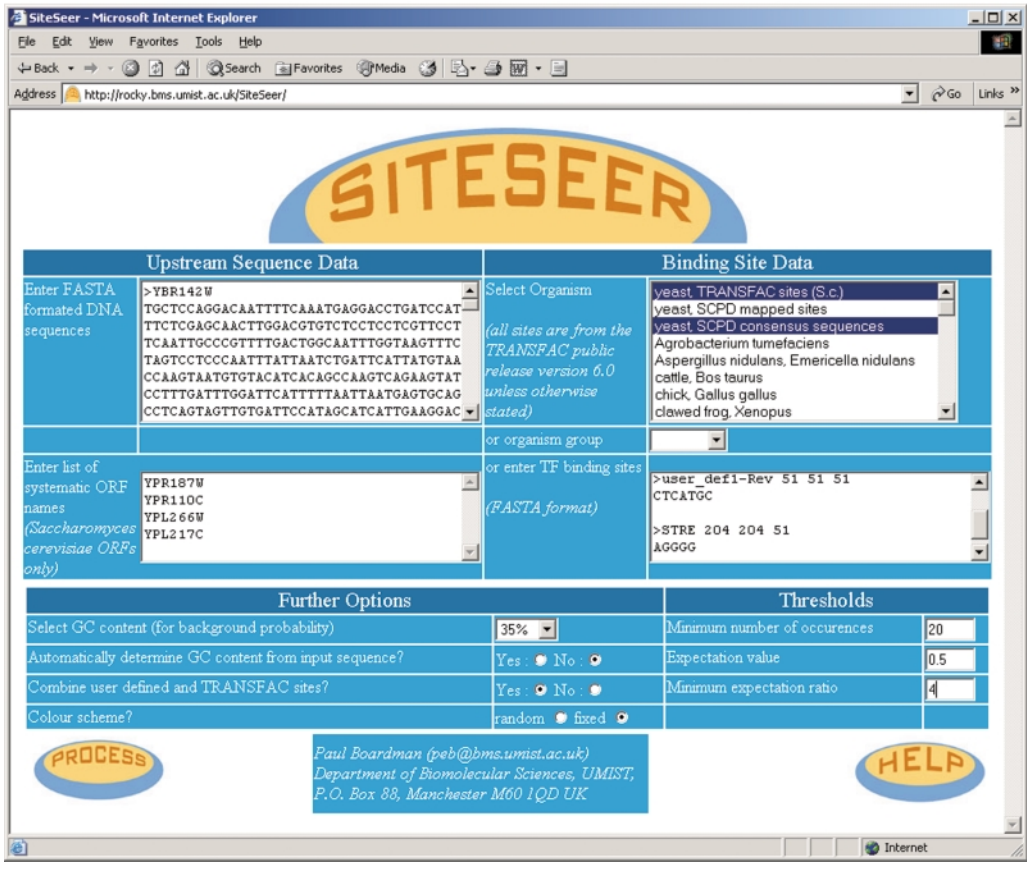

Figure 2. SiteSeer front page.

orthologous genes from different species (example shown in Fig. 1). To this end, we have developed a tool that allows the user to specify a set of related genes and to define the binding sites of interest [from a section of TRANSFAC (10), from the *Saccharomyces cerevisiae* Promoter Database SCPD (11), or a set of user-defined sites] and to investigate the occurrences of potential regulatory sites in the chosen gene sets.

OVERVIEW

The software, SiteSeer, searches input nucleotide sequences for transcription factor binding sites and creates a graphical representation of the results. Binding site sequences have been extracted from TRANSFAC public release version 6.0 (available at http://www.gene-regulation.com) and SCPD (as of 12 February 2003) (11). The user may select to search with sites from a single organism, from a taxonomic group (yeasts, plants, viruses, animals or insects) or with user-defined sites. An expectation value and expectation ratio are calculated for each site. These two metrics provide an estimate of the significance of occurrence of the site and can be used as a filter to remove low-significance sites from the final output.

PROGRAM USAGE

Sequence input

Figure 2 shows a screenshot of the SiteSeer input page. Sets of FASTA-formatted nucleotide sequences can be 'cut-

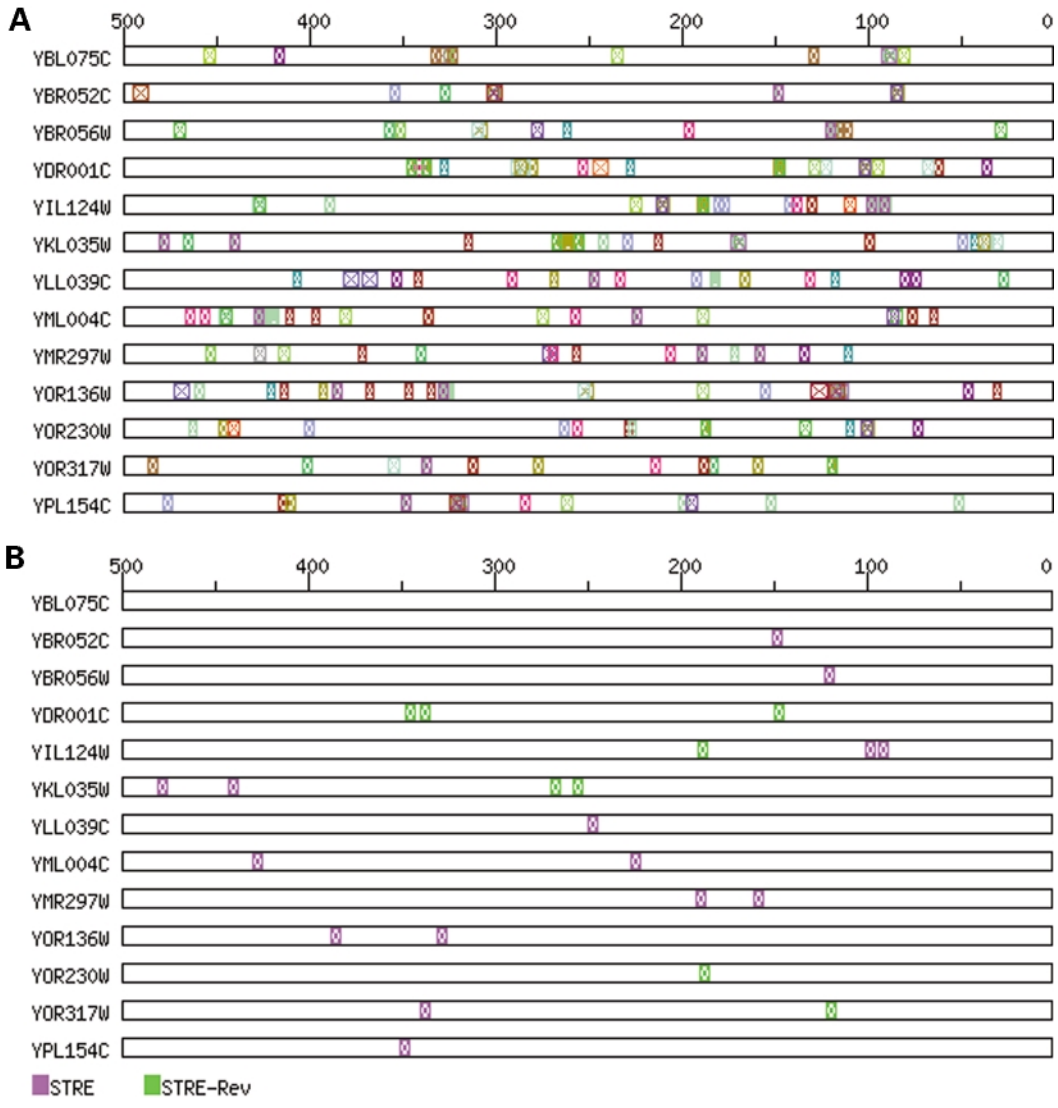

Figure 3. Visualisation of the upstream regions of clustered yeast genes. Genes were clustered using a self-organising map with the expression data from Gasch *et al.* (13). (**A**) Data shown with default thresholds and a user-defined binding-site (STRE). (**B**) The same data, but using threshold values to reduce the complexity of the display and to highlight potentially important sites (thresholds used were: minimum occurrence, 8; maximum expectation value, 0.45; and minimum expectation ratio, 4).

and-pasted' into the sequence input box (top left of the figure). This tool was originally designed with the yeast community in mind and, as a reflection of this, there is an input box for *Saccharomyces cerevisiae* systematic ORF names. If these are supplied, a sequence encompassing 800 bp upstream of the translational start site is automatically retrieved for each ORF specified. These input methods may be used in conjunction with each other to provide the set of sequence data to be analysed.

Binding-site selection

There are three ways of defining the set(s) of binding-site sequences to investigate. We have provided sites from the TRANSFAC public release (version 6.0) separated by organism and also into larger taxonomic groups (yeasts, plants, insects, animals or viruses). We have also included both mapped sites and consensus sequences from SCPD. Users may also enter additional binding-site sequence data, in FASTA format, in the input box provided (an example is shown in Fig. 1). The binding sites may either be entered as exact matches (e.g. CAGATA) or via a simple regular expression grammar, using standard IUPAC codes (e.g. the pattern A-[any nucleotide]-[A or T]-G is represented as the string ANWG).

Thresholds

It was not our intention to develop a novel over-representation statistic or promoter-discovery tool, rather we planned to allow known or user-defined sites to be easily visualised. However, we have included a simple statistical measure for over-

representation derived from the input sequences themselves, and based on GC content and binding-site probabilities (using a simple zero-order Markov Model). We define two metrics: the expectation value and the expectation ratio. The expectation value of a site in a given sequence is a measure of the likelihood of the site occurring in the upstream region. Let l_q be the length of the query sequence, l_s be the length of the binding-site sequence and $p(x_i)$ be the probability of finding nucleotide x at position i in the binding site. Then, we define the expectation value of a site as:

$$(l_q - l_s + 1) \prod_{i=1}^{l_s} p(x_i) \quad (1)$$

The expectation ratio builds on this measure by taking into account the actual number of sites detected in a sequence. This is calculated by dividing the number of occurrences of a specific binding site in an upstream sequence by the expectation value of this site.

The chief reason for the inclusion of these thresholds is to mask low-complexity 'uninteresting' sites, such as the many small sites present in TRANSFAC and the ubiquitous TATA sites. Thus, users can remove potential noise from their sequence representations by choosing an appropriate threshold value, on any of these parameters, below which binding-sites are not drawn. In the example shown in Figure 3, the importance of the user-defined site STRE (the stress response element, 'AGGGG') is masked by many low-complexity sites. By applying threshold limits, the less significant sites are removed from the display, leaving only the most significant sites (in this case, the STRE elements). We believe that the expectation ratio, in particular, provides a simple, intuitive, threshold measure for users to reduce the complexity of the final display by masking low significance sites.

Colour scheme

There are two options for colouring the binding sites. By default, binding sites are randomly allocated one of the 215, non-white, web-safe colours (http://www.imswebtips.com/issue56top1.htm). The user may also choose the fixed colour scheme where the colour allocated is dependent on the name of the binding site. This ensures consistency between images, but can result in an image that is difficult to interpret as sites with similar names often have similar colours. Finally, colours can be individually assigned for user-defined sites by including an RGB value after the site name. For example, inputting the definition line '>STRE 0 0 0' would ensure the colour black is used when drawing all STRE sites (for more information see the SiteSeer help page).

OUTPUT

SiteSeer creates an image in Portable Network Graphics (PNG) format to represent the results (see Figs 1 and 3 for examples). Each input upstream sequence is displayed as an empty rectangle whose width (in pixels) is identical to the length of the original sequence (in nucleotides). A scale is provided at the top of the image to aid in interpretation. Individual sites are represented as a coloured cross, bounded by a rectangle of the same colour. The width of this rectangle is proportional to the length of the binding site. The image is part of a client-side image map, which allows further exploration of the data through interaction with the map. Placing the mouse over any mapped site provides brief details of the site and the number of occurrences, whilst clicking on a binding-site brings up further, more detailed, information about the site. This includes expectation scores, accession number, database and sequence data.

COMPUTATIONAL DETAILS

SiteSeer is freely available at http://rocky.bms.umist.ac.uk/SiteSeer/. The tool is written in the Perl programming language (12). A MySQL (http://www.mysql.com/) database is used to store the *S.cerevisiae* upstream sequences and the binding-site data. SiteSeer currently runs on a LINUX server with dual 1 GHz INTEL Pentium III processors.

ACKNOWLEDGEMENTS

We would like to thank Lindsey Jones for web-design assistance, Claire Wilson for beta testing and Fajar Restuhadi for providing data. P.B. was supported by an MRC Bioinformatics studentship.

REFERENCES

1. Kornberg,R.D. (1999) Eukaryotic transcription control. *Trends Genet.,* **15**, 46–49.
2. Kadonga,J.T. (1998) Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. *Cell*, **77**, 599–608.
3. Schuldiner,O., Yanover,C. and Benvenisty,N. (1998) Computer analysis of the entire budding yeast genome for putative targets of the GCN4 transcription factor. *Curr. Genet.*, **33**, 16–20.
4. Wolfsberg,T.G., Gabrielian,A.E., Campbell,M.J., Cho,R.J., Spouge,J.L. and Landsman,D. (1999) Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*. *Genome Res.*, **9**, 775–792.
5. Zhang,M.Q. (1999) Large-scale gene expression data analysis: a new challenge to computational biologists. *Genome Res.*, **9**, 681–688.
6. Fujibuchi,W., Anderson,J.S.J. and Landsman,D. (2001) PROSPECT improves *cis*-acting regulatory element prediction by integrating expression profile data with consensus pattern searches. *Nucleic Acids Res.*, **29**, 3988–3996.
7. Van Helden,J., André,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.
8. Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.
9. Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG database at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.
10. Matys,V., Fricke,E., Geffers,R., Gobling,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. Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae. Bioinformatics*, **15**, 607–611.
12. Wall,L., Christiansen,T. and Orwant,J. (2000) Programming Perl 3rd Edn. O'Reilly and Associates, Inc., Sebastopol, CA.
13. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell.*, **11**, 4241–4257.