Cluster Analyzer for Transcription Sites (CATS): a C++-based program for identifying clustered transcription factor binding sites

Han Yu1,†, Andrew S. Yoo2,*,† and Iva Greenwald2,3

1Wave Systems Corp., 19925 Stevens Creek Blvd, Cupertino, CA 95014, USA, 2Integrated Program in Cellular, Molecular and Biophysical Studies and 3Department of Biochemistry and Molecular Biophysics and Howard Hughes Medical Institute, Columbia University, 701 West 168th Street, New York, NY 10032, USA

Received on August 28, 2003; revised on November 25, 2003; accepted on December 11, 2003
Advance Access publication February 5, 2004

Summary: We have developed a program, Cluster Analyzer for Transcription Sites (CATS), which identifies clusters of transcription factor binding sites in any genome sequence. The program searches for clusters of the consensus sequence for DNA binding within a window (length of DNA). The window size and the cluster size (number of consensus sequences within a given window) can be varied. CATS can be used for single or multiple transcription factors for which consensus sequences have been deduced based on biochemical and mutational analysis, or by comparative genomics. The use of CATS for clusters of different transcription factor binding sites may facilitate the identification of genes that are co-regulated in a cell type-specific or developmental stage-specific manner. CATS is simple to install and use on computers running any Windows NT-platforms.

Availability:  http://www.healthsciences.columbia.edu/dept/greenwaldlab/links.html

Contact: asy14@columbia.edu

ABSTRACT

Regulatory sequences present in the flanking and intronic regions of genes control their expression in space and time. During animal development, cell–cell interactions, mediated by ligands and receptor proteins, specify cell fate by modulating the activity of transcription factors that control the activation of specific target genes. The identification of such target genes is necessary to elucidate the molecular details of how they are regulated, and to explore how different cell types are created by the expression of different target genes.

There is increasing interest among developmental biologists in computational tools for identifying transcriptional targets of signaling pathways. For some transcription factors, the presence of clustered binding sites contributes to their ability to activate or repress target genes; in other cases, the occurrence of combinations of binding sites for different transcription factors can give cell type specificity (reviewed in Davidson, 2001). The search for clustered transcription factor binding sites is one way to approach the identification of target gene candidates in sequenced genomes (e.g. Berman et al., 2002; Rebeiz et al., 2002).

We have developed another tool to use for such a purpose, Cluster Analyzer for Transcription Sites (CATS). CATS differs from other tools currently in use (e.g. Berman et al., 2002; Halfon et al., 2002; Markstein et al., 2002; Rajewsky et al., 2002; Rebeiz et al., 2002) in that CATS has up to three scanning windows (as opposed to one), enables the user to define distances between different clusters and uses a local (as opposed to a remote) database that can be custom-tailored by importing selected FASTA files. We believe that CATS will be a generally useful adjunct to the other programs that are used in the search for transcription factor target genes. We have successfully used CATS to search for targets of the LIN-12/Notch signaling pathway in a specific Caenorhabditis elegans cell type, and found that 6 out of 10 genes identified in our search were bona fide targets, expressed in response to LIN-12 activation in that cell type (Yoo et al., 2004). The success rate using CATS, as for any of the currently available programs, will depend on the quality of the input data, i.e. how well the binding sites have been defined biochemically or computationally.

The user interface of CATS is shown in Figure 1. Input parameters include the file to be searched and the scanning window size(s) for up to 10 individual consensus sequences for each scanning window. Up to three scanning windows can be run simultaneously, and for each the minimum number of binding sites per scanning window may be specified. Additional parameters can be specified are the number of mismatches for each consensus sequence (default is set at 0) and the number of strands scanned (default is set at 2). For scanning windows 2 and 3, there is an additional parameter that may be

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
Fig. 1. User interface of CATS. Note help button in lower right corner. The first output window shows coordinate numbers, and the second output window shows highlighted matches to the consensus sequence.

set, the maximal distance in relation to window 1. The additional windows are useful for identifying candidate genes that are regulated by combinations of transcription factors, or for identifying ‘cis-regulatory modules’ (see Berman et al., 2002). Parameters are chosen based on existing biological data when possible.

Output is given in two different formats simultaneously. In the first format, the coordinates of the clusters are given. This output display is most useful when well-annotated genome sequence databases are used. In the second format, all sequences are shown in the scanning window, with a colored font used to show the sequence hits. The sequence information can be used in a BLAST search or any other manner to localize the information.

CATS is designed to be user-friendly. For each parameter, when the cursor is placed over the box, a help balloon is displayed so that everything is self-explanatory. For more detailed information, a ‘Help’ button may be clicked, also enabling all information about the program to be viewed in one file.

ACKNOWLEDGEMENTS

We thank Henry Bigelow for Web assistance. A.Y. is supported by a fellowship from the National Sciences and
Engineering Research Council of Canada. I.G. is an Investigator of the Howard Hughes Medical Institute.

REFERENCES


