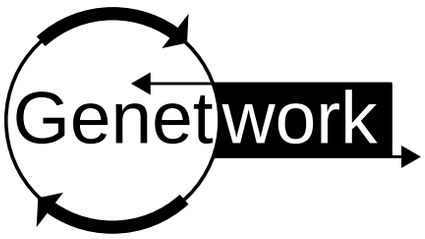


# Detection of potential target genes *in silico*?

Transcription factors (TFs) not only control a wide range of physiological processes, but are also responsible for a host of pathological phenomena in eukaryotic cells. These molecules specifically recognize and bind to regulatory sequences of target genes, whose transcription is up- or down-regulated as a consequence. Two phenomena that are typical of TFs in higher eukaryotes present serious obstacles to the analysis of their function by genetic means. These are pleiotropy (one TF might regulate many genes of sometimes apparently unrelated function) and genetic redundancy (several related TFs might regulate overlapping groups of genes). The latter phenomenon is known to hinder greatly interpretation of gene knock-out experiments in vertebrates[1].

Given the problems associated with a genetic approach, the direct identification of TF target genes is an attractive alternative for dissecting TF function. Several *in vitro* methods have been used for this purpose[2–7], with mixed success. An *in silico* analysis, aimed at identifying target genes via detection of potential TF-binding sites within general-purpose nucleotide sequence databases, such as GenBank, is a promising extension of *in vitro* methods. However, owing to the low-information content[8,9] of these often short and degenerate sites, many potential sites will be found occurring randomly almost anywhere in a genome. This is a major problem of almost any TF site-based approach that can be partially circumvented by (1) searching databases more specialized than GenBank, such as the Eukaryotic Promoter Database[10] (EPD), and (2) including context information into the binding-site search. A variety of methods and programs incorporating context information in promoter analysis has recently become available. A selection of them is reviewed here and contrasted to context-insensitive tools. Also, we put special emphasis on WWW-accessible tools because they will be most useful and accessible for most investigators. The surveyed methods are different from those addressing the related problem of promoter recognition, which have been reviewed recently[11] and will not be discussed further.

A number of well-established context-insensitive sequence analysis programs (e.g. FASTA[12], FindPatterns[19], ProfileSearch[19], PatScan[24], MatInspector[25], MatrixSearch[22], SignalScan[26]) can be employed for the purpose of screening whole nucleotide-sequence databases (or subsections) for matches to short sequences such as TF-binding sites. All of these programs can identify potential binding sites for a TF of interest, with major advantages on the side of programs using weight matrices as opposed to those using IUPAC consensus sequences or definite nucleotide sequences[13] (programs using weight matrices use the distribution of all four nucleotides at each single position of the matrix in order to calculate a quantitative score, which results in an enhanced specificity. IUPAC consensus searches use, instead, a majority rule, which results in a simple yes/no decision[13]). However, because these programs lack context-sensitivity, they will find spurious matches in many sequences that are not target genes. This high false-positive rate will obscure the real target genes also found in the search. The otherwise popular BLAST[14] program is even less-well suited for locating the limited similarities represented by TF-binding sites. In fact, it requires a minimum number of seven exactly matching bases, which is too stringent for the majority of TF-binding sites.

## WWW-accessible tools for context-sensitive sequence analysis

The functional context of a TF-binding site includes the following: local status of chromatin compaction; the position of the binding site, relative to the transcription start site; and the presence of other binding sites nearby. The computational methods discussed here (see Table 1 for URLs and references) try to include this context-sensitivity in different ways. None of these programs is capable of pinpointing real target genes specifically, but their output should be enriched in these genes owing to the enormous reduction in the number of spurious matches. The user is responsible for the definition of the type of context to be considered by these methods. This step is crucial for the quality of the results and, therefore, should receive special attention.

Programs like MatInspector and MatrixSearch come with a predefined library of carefully selected matrices, which are of immediate use to the researcher. MatInspector library is based on the TRANSFAC database[23], whereas MatrixSearch is based on the Information Matrix Database[22]. It is important to stress the need to use high-quality weight matrices that can contribute to a good search outcome even more than the chosen search-algorithm[13]. A good introduction to the general issues about the criteria that need to be met by a TF-binding site in order to be included in a high-quality weight matrix can be found on the documentation pages of the TRANSFAC Web site[15].

The NCBI server provides CosMoS, a yeast-specific tool that allows the detection of user-defined patterns within putative promoter regions of the yeast genome (upstream of open reading frame start points), effectively focusing the search on known promoter regions. The program FastM exploits the spatial connection and, optionally, sequential order, between two different transcription factors in order to develop simple models of transcriptional regulatory DNA sections, independent of *a priori* knowledge about the location of promoters. FastM employs MatInspector and its matrix library, thus, greatly facilitating the user-selection of TF weight matrices or consensi. Another tool, TargetFinder, also uses MatInspector and a predefined TF library to search for TF sites in databases. The program takes advantage of annotated features present in GenBank entries to restrict matches to relevant gene sub-regions, significantly reducing the background usually associated with these searches. TargetFinder allows the inclusion of sequence annotation (e.g. TATA box, transcription start site, annotated promoters) that cannot be included in FastM models. The Transcription Factor Combination Discoverer[16] (TFCD) finds and analyzes combinations of transcription factor binding sites in the yeast genome and in upstream regions in particular.

## Other approaches for context-sensitive sequence analysis

Often, binding sites for one or more TFs are closely spaced in a regulatory region, indicating cooperativity in transcriptional regulation. Recently, statistical and heuristic techniques have been developed for the detection of such clusters in large genomic DNA regions[16–18]. For example, the GenomeInspector tool can detect distant correlations between sequence elements (e.g. between ORFs and TF-binding sites) on megabases of nucleotide sequences[17]. Another approach, not yet released as public-domain software, employs statistical tests to screen a genome for very closely spaced TF-binding sites[18]. Its application to the genome of yeast detects genes known to be regulated by particular TFs, and genes that are not known to be regulated by the TFs, but that act in the same cellular process (e.g. cell-cycle) as the studied TFs (Ref. 18).

Although the search for TF target genes in sequence databases is still a difficult task, the tools discussed above can help to reduce the signal-to-noise ratio considerably. Importantly, these individual approaches can be used complementarily and their incorporation into one integrated analysis tool is highly desirable. However, even the current improvements in specificity, achieved by incorporating very simple biological principles into database searches, demonstrate that context-sensitive approaches hold great promise.



 0168-9525/98/$19.00
PII: S0168-9525(98)01551-0

**TABLE 1. List of available resources for the purpose of searching transcription factor target-genes in nucleotide sequence databases**

| Program | Availability | Predefined library[a] | Search[b] | Heuristic[c] | Ref. |
|---|---|---|---|---|---|
| FASTA | ftp://ftp.virginia.edu/pub/fasta | None | A, T, G, C, N | None | 12 |
| PatScan | ftp://info.mcs.anl.gov/pub/overbeek/PatScan/scan_for_matches.tar.Z http://www.mcs.anl.gov/home/overbeek/PatScan/HTML/patscan.html | None | IUPAC NDM Regexps | None | 24 |
| FindPatterns | GCG Wisconsin Package | ftp://ncbi.nlm.nih.gov//pub/repository/TFD//datasets/sitesdata.gcg | IUPAC Regexps | None | 19 |
| ProfileSearch | GCG Wisconsin Package | None | NDM | None | 19 |
| MatInspector | ftp://ariane.gsf.de/pub/unix/matind_2.1.tar.Z http://www.gsf.de/cgi-bin/matsearch.pl | MatLibrary TRANSFAC | NDM IUPAC | None | 25 |
| CosMos | http://www.ncbi.nlm.nih.gov/XREFdb/CoSMoS/index.html | None | IUPAC Regexps | Search confined to ORF upstream regions in yeast | d |
| FastM | http://www.gsf.de/cgi-bin/fastm.pl | TRANSFAC | NDM IUPAC | Search for TF interaction | e |
| TargetFinder | http://gcg.tigem.it/TargetFinder.html | TRANSFAC | IUPAC NDM | Search confined to putative gene regulatory regions | f |
| TFCD | http://www.cs.helsinki.fi/~vilo/Yeast | IMD TRANSFAC | IUPAC NDM Regexps | Search for TF interaction in yeast promoters | 16 |

[a]Libraries of ready-to-use NDM and IUPAC consensi provided with the programs.
[b]Type of nucleotide pattern data that can be searched for by the programs.
[c]Search strategy adopted by the programs in order to implement context-sensitive sequence analysis. The expressions can include several non-search characters that are used to specify OR and NOT matching, begin and end constraints and repeat counts. A, T, G, C, N, IUPAC string searches restricted to the four bases and to a completely unspecified position (N).
[d]D.E. Bassett, Jr, M. Geraghty, S.J. Gould, P. Hieter and M.S. Boguski, pers. commun.
[e]K. Frech, K. Quandt and T. Werner, pers. commun.
[f]G. Lavorgna, A. Guffanti and E. Boncinelli, pers. commun.
Abbreviations: IMD, information matrix database[21]; NDM, nucleotide distribution matrix; Regexps, regular expression search; TFs, transcription factors; TFD, transcription factor database[20]; TRANSFAC, transcription factor database[22].

## References


1 Jacobson, D. and Anagnostopoulos, A. (1996) *Trends Genet.* 12, 117–118
2 Gould, A.P. *et al.* (1990) *Nature* 348, 308–312
3 Orlando, V., Strutt, H. and Paro, R. (1997) *Methods* 11, 205–214
4 Kinzler, K. and Vogelstein, B. (1989) *Nucleic Acids Res.* 17, 3645–3653
5 Joulin, V. and Richard-Foy, H. (1995) *Eur. J. Biochem.* 232, 620–626
6 Caubin, J. *et al.* (1994) *Nucleic Acids Res.* 11, 4132–4138
7 Liang, P. and Pardee, A.B. (1992) *Science* 257, 967–971
8 Schneider, T.D., Stormo, G.D., Gold, L. and Ehrenfeucht, A. (1986) *J. Mol. Biol.* 188, 415–431
9 Stormo, G.D. and Fields, D.S. (1998) *Trends Biochem. Sci.* 23, 109–113
10 Cavin Perier, R., Junier, T. and Bucher, P. (1998) *Nucleic Acids Res.* 26, 353–357
11 Fickett, J.W. and Hatzigeorgiou, A. (1997) *Genome Res.* 7, 861–878
12 Pearson, W.R. (1994) *Methods Mol. Biol.* 124, 307–331
13 Frech, K., Quandt, K. and Werner, T. (1997) *Trends Biochem. Sci.* 22, 103–104
14 Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
15 http://transfac.gbf.de/TRANSFAC/doc34/site3.html
16 Brazma, A., Vilo, J., Ukkonen, E. and Valtonen, K. (1997) *ISMB* 5, 65–74
17 Quandt, K., Grote, K. and Werner, T. (1996) *Genomics* 33, 301–304
18 Wagner, A. *Genomics* (in press)
19 Wisconsin Package Version 9.0, Genetics Computer Group (GCG), Madison
20 Audic, S. and Claverie, J.M. (1998) *Trends Genet.* 14, 10–11
21 Ghosh, D. (1991) *Trends Biochem. Sci.* 16, 445–447
22 Chen, Q.K., Hertz, J.K. and Stormo, G.D. (1995) *Comp. Appl. Biosci.* 11, 563–566
23 Heinemeyer, T. *et al.* (1998) *Nucleic Acids Res.* 26, 364–370
24 Dsouza, M., Larsen, N. and Overbeek, R. (1997) *Trends Genet.* 13, 497–498
25 Quandt, K. *et al.* (1995) *Nucleic Acids Res.* 23, 4878–4884
26 Prestridge, D.S. (1991) *Comp. Appl. Biosci.* 7, 203–206



**Giovanni Lavorgna**
giovanni.lavorgna@hsr.it

**Edoardo Boncinelli**
edoardo.boncinelli@hsr.it

*DIBIT, Istituto Scientifico H.S. Raffaele, Via Olgettina 60, 20132 Milano, Italy.*

**Andreas Wagner**
aw@santafe.edu

*The University of New Mexico, Department of Biology and The Santa Fe Institute University of New Mexico, Department of Biology, 167A Castetter Hall, Albuquerque, NM 87131-1091, USA.*

**Thomas Werner**
werner@gsf.de

*GSF-National Research Center for Environment and Health, Institute of Mammalian Genetics, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany.*