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) and genetic redundancy (several related TFs might regulate overlapping groups of genes). The latter phenomenon is known to render greatly interpretation of genetic knock-out experiments in vertebrates.

Given the problems associated with a genetic approach, the direct identification of TF target genes is an attractive alternative for dissecting TF function. Several examples have been used for this purpose, for dissecting TF function. All of these programs can identity potential binding sites for a TF of interest, with major advantages on the side of using programs with weight matrices as opposed to those using IUPAC consensus sequences or define nucleotide sequences (programs using weight matrices use the distribution of all four nucleotides at each position of the matrix in order to calculate a quantitative measure, which results in an enhanced specificity. IUPAC consensus successes searches use instead, a majority rule, which results in a simple yes/no decision). However, because these programs lack context-sensitivity, they will find 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 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.

WWA-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 refined 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. MatrixInspector library is based on the TRANSFAC database, whereas MatrixSearch is based on the Information Matrix Database. 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. 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.

The NCBI server provides CosMoS, a yeast-specific tool that allows the detection of sequence similarities within positive promoter regions of the yeast genome (upstream of open reading frame start points), effectively covering a large number of positive promoter regions. The program Fasit exploits the fact that sites have been developed for the sequential order, between two different transcription factors in order to develop simple models of transcriptional units. The DNA sections, independent of a priori knowledge about the location of these units (e.g. using the TRANSFAC database), can be employed for the detection of such clusters in large genomic DNA regions. For example, the GenomeInspector tool can detect potential TFs between sequence elements (e.g. between ORFs and TF-binding sites) on megabases of nucleotide sequences. Another approach, not yet released as public-domain software, employs statistical tools to screen a genome for very closely spaced TF-binding sites.

In all, the approaches discussed here have the advantage of being independent of a priori knowledge about the location of these units (e.g. using the TRANSFAC database), can be employed for the detection of such clusters in large genomic DNA regions. For example, the GenomeInspector tool can detect potential TFs between sequence elements (e.g. between ORFs and TF-binding sites) on megabases of nucleotide sequences. Another approach, not yet released as public-domain software, employs statistical tools to screen a genome for very closely spaced TF-binding sites.

In all, the approaches discussed here have the advantage of being independent of a priori knowledge about the location of these units (e.g. using the TRANSFAC database), can be employed for the detection of such clusters in large genomic DNA regions. For example, the GenomeInspector tool can detect potential TFs between sequence elements (e.g. between ORFs and TF-binding sites) on megabases of nucleotide sequences. Another approach, not yet released as public-domain software, employs statistical tools to screen a genome for very closely spaced TF-binding sites.

In all, the approaches discussed here have the advantage of being independent of a priori knowledge about the location of these units (e.g. using the TRANSFAC database), can be employed for the detection of such clusters in large genomic DNA regions. For example, the GenomeInspector tool can detect potential TFs between sequence elements (e.g. between ORFs and TF-binding sites) on megabases of nucleotide sequences. Another approach, not yet released as public-domain software, employs statistical tools to screen a genome for very closely spaced TF-binding sites.

In all, the approaches discussed here have the advantage of being independent of a priori knowledge about the location of these units (e.g. using the TRANSFAC database), can be employed for the detection of such clusters in large genomic DNA regions. For example, the GenomeInspector tool can detect potential TFs between sequence elements (e.g. between ORFs and TF-binding sites) on megabases of nucleotide sequences. Another approach, not yet released as public-domain software, employs statistical tools to screen a genome for very closely spaced TF-binding sites.

In all, the approaches discussed here have the advantage of being independent of a priori knowledge about the location of these units (e.g. using the TRANSFAC database), can be employed for the detection of such clusters in large genomic DNA regions. For example, the GenomeInspector tool can detect potential TFs between sequence elements (e.g. between ORFs and TF-binding sites) on megabases of nucleotide sequences. Another approach, not yet released as public-domain software, employs statistical tools to screen a genome for very closely spaced TF-binding sites.

In all, the approaches discussed here have the advantage of being independent of a priori knowledge about the location of these units (e.g. using the TRANSFAC database), can be employed for the detection of such clusters in large genomic DNA regions. For example, the GenomeInspector tool can detect potential TFs between sequence elements (e.g. between ORFs and TF-binding sites) on megabases of nucleotide sequences. Another approach, not yet released as public-domain software, employs statistical tools to screen a genome for very closely spaced TF-binding sites.

In all, the approaches discussed here have the advantage of being independent of a priori knowledge about the location of these units (e.g. using the TRANSFAC database), can be employed for the detection of such clusters in large genomic DNA regions. For example, the GenomeInspector tool can detect potential TFs between sequence elements (e.g. between ORFs and TF-binding sites) on megabases of nucleotide sequences. Another approach, not yet released as public-domain software, employs statistical tools to screen a genome for very closely spaced TF-binding sites.
### References

18. Wagner, A. Genomics (in press)
19. Wisconsin Package Version 8.0, Genetics Computer Group (GCG), Madison

### Abbreviations

IMD: information matrix database; NDM: nucleotide distribution matrix; Regexps: regular expression search; TFs: transcription factors; TFD: transcription factor database; TRANSFAC: transcription factor database.

### Libraries of ready-to-use NDM and IUPAC consens provided with the programs.

- A 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. E. Bassett, Jr, M. Geraghty, S. J. Gould, P. Hieter and M. S. Boguski, pers. commun.

### Abbreviations

- IMD, information matrix database;
- NDM, nucleotide distribution matrix;
- Regexps, regular expression search;
- TFs, transcription factors;
- TFD, transcription factor database.

### Notes

- E. Boncinelli, pers. commun.
- F. Wagner, pers. commun.
- A. Guffanti and E. Boncinelli, pers. commun.
- G. Lavorgna, A. Guffanti and E. Boncinelli, pers. commun.

### Table 1. List of available resources for the purpose of searching transcription factor target genes in nucleotide sequence databases

<table>
<thead>
<tr>
<th>Program</th>
<th>Availability</th>
<th>Predefined library</th>
<th>Search</th>
<th>Heuristic</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PatScan</td>
<td>ftp://patscan.wustl.edu</td>
<td>None</td>
<td>A, T, G, C, N</td>
<td>IUPAC NDM Regexp</td>
<td>19</td>
</tr>
<tr>
<td>Cosmotools</td>
<td><a href="http://www.ncbi.nlm.nih.gov/STDFdb/">http://www.ncbi.nlm.nih.gov/STDFdb/</a></td>
<td>None</td>
<td>IUPAC NDM Regexp</td>
<td>Search confined to ORF upstream regions in yeast</td>
<td>16</td>
</tr>
</tbody>
</table>

### Notes

- TIG September 1998 Vol. 14 No. 9