SIGNAL SCAN 3.0: new database and program features

Dan S. Prestridge and Gary Stormo

Abstract

SIGNAL SCAN is a program that utilizes a transcription factor database to find potential transcription factor binding sites in DNA sequences. The program is now in its third version. The SIGNAL SCAN transcription factor database format has changed and the program output format has been improved. New features allow the user to update the SIGNAL SCAN database automatically, to retrieve original journal citations and to develop user signal databases. The program now uses an indexing algorithm, improving scanning speed by a factor of 3. SIGNAL SCAN is now network compatible and is available for IBM-compatible PC, Unix and VMS platforms.

SIGNAL SCAN is a program developed to find putative transcription factor binding sites in DNA sequences (Prestridge, 1991). It was originally developed to use a small database of transcription factor binding sites reported in the literature to scan DNA sequences for these sites. The continually growing field of transcription factors (Wingender, 1990; Latchman, 1991) and the changing needs of investigators in the field have necessitated growth and changes in the SIGNAL SCAN database and program. The original database had only 100 transcription factor binding sites, while today >1800 binding sites are reported in the current version of David Ghosh’s Transcription Factor Database (TFD) (Ghosh, 1990, 1991). These data have mandated a complete revision of the SIGNAL SCAN database and program structure.

The SIGNAL SCAN database

In the present version (3.0) the SIGNAL SCAN database is entirely based upon the Ghosh TFD. The TFD is now updated on a monthly basis (D. Ghosh, personal communication) and the SIGNAL SCAN database is updated with each release. The information in the TFD is processed by a SIGNAL SCAN utility program into eight organism classes: mammal, bird, amphibian, insect, plant, yeast, prokaryote and miscellaneous classes. The prokaryote and miscellaneous classes are new with this release, the plant class was added in a previous release of SIGNAL SCAN. The current SIGNAL SCAN database has a total of 1887 sites: 1384 mammalian, 43 bird, 24 amphibian, 115 insect, 35 plant, 170 yeast, 97 prokaryote and 17 miscellaneous sites.

Both the binding factor name (if known) as well as the site name are now included in the database, and the SIGNAL SCAN program now reports the factor name rather than the TFD designated site name. This reduces confusion about the TFD site name nomenclature that is unique to the TFD. Only if a site is reported by the TFD to be bound by an unknown factor is the TFD site name used in the analysis.

Previously SIGNAL SCAN included only a few journal citations for the database sites while most sites (those obtained from the TFD) included only Medline database reference numbers, which are difficult to use by most investigators. Since the journal citations for each site are included in Ghosh’s TFD, the journal citations are now included in the SIGNAL SCAN database and are easily accessible through the program (discussed below).

In addition to the improvements in the SIGNAL SCAN database, the program now allows users to establish their own separate databases of binding sites. Previously, investigators had to re-enter their own signals into the program each time a scan was performed. With this release the investigator can create any number of separate databases of sites, save them, and use them in future sequence scans.

The SIGNAL SCAN program

Because of the rapidly changing nature of the TFD, it is imperative to include a feature in SIGNAL SCAN that allows the investigator to update the SIGNAL SCAN database frequently. This has now been largely automated. A new menu-selectable feature allows the user to rebuild the SIGNAL SCAN database using a current sites.dat file obtained from the TFD at the National Center for Biotechnology Information (NCBI).

Because of the increasing number of sites appearing in the literature and being incorporated into the TFD, there has been a need to increase the speed of scanning sequences to handle the increased data that must be stored in memory. Scanning speed has increased >8-fold over version 1.0, and 3-fold over version 2.0. The original increase in speed was accomplished by several improvements in the source code, but the algorithm remained unchanged (Prestridge, 1991). In this release, the speed increase is due to the incorporation of an indexing algorithm.

One disadvantage of using the raw TFD to scan sequences for transcriptional elements using other DNA analysis programs...
is that all sites in the TFD are searched for. If the investigator has a mammalian sequence, for example, a search using the raw TFD will result in the report of many yeast, plant and prokaryotic sites among others. This results in a very high number of irrelevant sites being reported and can be very confusing to researchers not familiar with using the TFD. To reduce the amount of irrelevant sites reported, SIGNAL SCAN incorporates the TFD sites data into the various organism classes. In past versions of SIGNAL SCAN, only one class at a time could be used to scan a sequence. With this release, multiple classes can be selected for the same scan, including the investigator’s own signal database. If all of the organism classes are selected for the scan, then the entire TFD is used.

The program output format has changed and includes more site information (Figure 1A). As mentioned earlier, the factor name is now given instead of the site name if the binding factor is known (Figure 1A, column 1). If the site name is used (the binding factor is unknown), then it is so designated in the second column (this column is blank if the factor name is given). The DNA strand on which the signal is located is now indicated by a (+) or (-) symbol. In addition, the TFD site number is given (last column), allowing the investigator to access the original journal citation for each site.

Original journal citations can now be accessed through the new journal citation look-up feature. When a sequence is scanned, each reported site has an associated TFD site number (Figure 1A, last column). By entering the TFD site number into the journal citation look-up feature, the original journal citation is reported and saved to a file (Figure 1B).

SIGNAL SCAN is now used at several large multi-user sites. Past versions have been modified by local on-site programmers to make the program network compatible. With this release, SIGNAL SCAN is now fully Unix network compatible. The program can reside in any location on the Unix network and be accessed by anyone on the network. It allows each investigator to keep and use their own signal databases, and the program remembers the last user signal database used by each investigator so that each investigator on the system does not have to tell the program each time to use their own unique signal databases.

### SIGNAL SCAN analysis

The analysis of sequences using a database of transcriptional elements has resulted in finding previously unknown functional transcriptional elements in DNA sequences (Rupp et al., 1990). However, there are many reported putative transcriptional elements in any sequence scan, most of which are probably irrelevant (Wingender et al., 1991; D.S.Prestridge and C.Burks, submitted). To be able, in the future, to predict functional sites,

---

#### Table A

<table>
<thead>
<tr>
<th>Factor or Site name</th>
<th>Loc.Str.</th>
<th>Signal sequence</th>
<th>TFD S#</th>
</tr>
</thead>
<tbody>
<tr>
<td>heat-indf</td>
<td>6 (+)</td>
<td>CTGGAATATTCCCG</td>
<td>S00774</td>
</tr>
<tr>
<td>HSTF</td>
<td>6 (+)</td>
<td>CTGGAATATTCCCG</td>
<td>S01248</td>
</tr>
<tr>
<td>CAP-site</td>
<td>26 (+)</td>
<td>CANYY</td>
<td>S00089</td>
</tr>
<tr>
<td>T-Ag</td>
<td>28 (-)</td>
<td>GAGGC</td>
<td>S00973</td>
</tr>
<tr>
<td>CPI</td>
<td>41 (-)</td>
<td>YNNNNNRRCCAATCANYK</td>
<td>S01306</td>
</tr>
<tr>
<td>GATA-1</td>
<td>43 (-)</td>
<td>MYWATCWY</td>
<td>S00477</td>
</tr>
<tr>
<td>CTF/CBP</td>
<td>45 (+)</td>
<td>GATTGG</td>
<td>S00777</td>
</tr>
<tr>
<td>NFI</td>
<td>45 (-)</td>
<td>GCTCACT</td>
<td>S00281</td>
</tr>
</tbody>
</table>

#### Table B

<table>
<thead>
<tr>
<th>ID #</th>
<th>FACTOR</th>
<th>SITE</th>
<th>JOURNAL REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>S00973</td>
<td>T-Ag-SV40.2</td>
<td>T-Ag-SV40.2</td>
<td>J. Virol., 46, 143–150 (1983)</td>
</tr>
<tr>
<td>S01306</td>
<td>CPI</td>
<td>CCAAT.4</td>
<td>Cell, 55, 11–24 (1988)</td>
</tr>
</tbody>
</table>

**Fig. 1.** (A) Partial SIGNAL SCAN output using the linear sequence scan option. The human 70 kd heat shock (hsp 70) gene promoter region sequence was used as the input sequence for SIGNAL SCAN (GenBank entry HUMHSP, Bilofsky and Burks, 1988). The sequence was scanned using the mammalian database. **Loc.** is the location of the signal in the sequence (relative to the beginning of the sequence in the GenBank entry file). **Str.** indicates the DNA strand. **TFD S#** is the TFD site number. Only the first 45 bp of the analysis is shown. (B) SIGNAL SCAN reference look-up feature output. The output includes the TFD site number (ID #), the factor and TFD site names, followed by the journal citation. Note that some journal citations are cut short (but are still usable) due to a shorter journal field length in SIGNAL SCAN than in the TFD.
the putative transcriptional element sites found by SIGNAL SCAN will probably have to be considered along with other elements found nearby in the sequence. The building of computer programs to analyze patterns of transcriptional elements is currently underway.

Until such time that pattern analysis may allow us to predict functional sites with a relatively high degree of confidence, the best uses of SIGNAL SCAN are:

(i) If the investigator has a known site bound by an as yet unidentified protein, SIGNAL SCAN can be used to find candidate binding proteins.

(ii) If the investigator has a known promoter sequence, SIGNAL SCAN can be used to identify possible regulatory elements.

(iii) If the investigator has some evidence of the regulatory nature of a sequence but has not located the appropriate elements, SIGNAL SCAN can be used to scan for those elements.

The advantages that SIGNAL SCAN currently has over most other more general site recognition programs are:

(i) The database of TFD sites is broken down into organism classes.

(ii) Easy access to journal citations for each reported site.

(iii) The program comes with the TFD already integrated, and it is easy to update. This requires no knowledge of programming or of relational databases by the investigator. The program comes ready to run and requires essentially no setup.

(iv) The results can be output in three different formats (Prestridge, 1991), allowing the investigator to select the format that is most relevant to their research.

(v) There is an on-line help facility.

(vi) The investigator can create, maintain, and use their own database of sequence elements.

SIGNAL SCAN is available in IBM-compatible PC, Unix and VAX/VMS versions. The PC version is available by surface mail or anonymous ftp, the Unix version is available by an e-mail server, and the VMS version is available by anonymous ftp. For more information on SIGNAL SCAN or how to obtain a copy of the program, send requests to DANP@MOLBIO. UMN.EDU.

Acknowledgements

I would like to thank Gary Williams of the UK Human Genome Mapping Centre, and Alec Dunn and Alex Riesner of the Sydney University Sequence Analysis Interface for their suggestions and source code for network compatibility. I would also like to thank the various SIGNAL SCAN users for their comments and suggestions. D.P. is currently supported by NIH HG00249 (G.Stormo, PI) and a grant from the Genetics Computer Group (GCG).

References