Conslnspector (Frech et al., 1993) is a program to scan nucleic acid sequences for matches to a pre-compiled library of transcription factor binding sites. The program carries out an extensive examination of binding site candidates; the real sequence is compared with randomly shuffled versions and sequence regions surrounding the conserved binding site are included in the analysis (default 40 bp upstream and 40 bp downstream of the highly conserved core sequence). This feature distinguishes the program from other methods available for the identification of transcription factor binding sites which are restricted to the binding sites: SIGNAL SCAN (Prestridge, 1991, 1996; Prestridge and Stormo, 1993), MATRIX SEARCH (Chen et al., 1995) and MatInspector (Quandt et al., 1995a). Recently, we showed the quality scores (C$_5$-scores) assigned by Conslnspector to correlate to some extent with biological functionality (Quandt et al., 1995b). Release 3.0 of Conslnspector, with enhanced performance and a considerably extended library of consensus profiles, is available now at ftp://ariane.gsf.de/pub/ or http://www.gsf.de/biodv/.

The program Consln (Frech et al., 1993) has been used to compile the library of consensus profiles. The library now encompasses 37 consensus profiles (Release 1.0: 12, Release 2.1: 17 consensus profiles) and is separated into four groups (Table 1). The extended weight matrices were deduced from experimentally confirmed binding sequences selected from the TRANSFAC database (Wingender et al., 1996) or directly from the literature. Most consensus profiles of the original library have been improved by the inclusion of additional sequences. Consensus profiles have been compiled from a minimum of nine sequences (Table 1).

The analysis of DNA sequences for transcription factor binding sites with Conslnspector has improved since Release 1.0:

1. Conslnspector now accepts EMBL-formatted sequences in addition to IG- and GCG-formatted sequences.
2. Sequences of unlimited length are accepted.
3. Scanning of both strands is possible.

4. Conslnspector compares the real sequence with randomly shuffled sequences of the same nucleotide composition. Results from individual program runs differ, since the quality rating (C$_5$-score) depends on random numbers. In order to restrict the variability of the results, each putative core position can now be examined multiple times (parameter). The final C$_5$-score is the average score of multiple test runs.

5. The scoring algorithm of Conslnspector was extended in order to enhance performance by pre-analysis of the conserved binding region. Conslnspector first identifies possible cores (sequence positions that reach a user-defined similarity with the consensus core; for definitions, see Quandt et al., 1995a). The sequence region around each core (default ± 40 bp) is aligned to the consensus profile and the similarity with the conserved binding region is calculated (comparable to matrix similarity of MatInspector; Quandt et al., 1995a). Analysis of the whole alignment region (including analysis of shuffled sequences) is only carried out for cores that reach the user-defined threshold for the region similarity. We have shown, for several example sequences and consensus profiles, that Conslnspector predicts the same possible binding sites as before in about half the CPU time.

6. Conslnspector is now available for DOS and Macintosh in addition to UNIX and VAX/VMS.

We used Conslnspector extensively for analysis of yeast (Quandt et al., 1996a,b) and multiple copies of the program have been downloaded from our ftp server. The extended consensus library, as well as the new program features, will further facilitate the use of Conslnspector and widen the range of its applicability. The consensus library will also be used to extend the matrix libraries of TRANSFAC (Wingender et al., 1996) and MatInspector (Quandt et al., 1995a). However, profiles of multi-component elements, especially if they have variable spacing regions, can only be used with Conslnspector, because they cannot be converted to a single matrix description. Conslnspector does a more extensive analysis of the region surrounding the binding site than MatInspector, e.g. it performs an alignment which allows insertions and deletions. This results in better quality of the predictions, but takes more time for the analysis. Therefore, Conslnspector should be a useful tool for the detailed analysis of a sequence.
Table I. Consensus library

<table>
<thead>
<tr>
<th>Group</th>
<th>No. of consensi</th>
<th>Minimum no. of sequences</th>
<th>Maximum no. of sequences</th>
<th>Average no. of sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prokaryotes</td>
<td>2</td>
<td>13</td>
<td>16</td>
<td>14.5</td>
</tr>
<tr>
<td>Vertebrates</td>
<td>16</td>
<td>9</td>
<td>67</td>
<td>21.4</td>
</tr>
<tr>
<td>Viruses</td>
<td>10</td>
<td>15</td>
<td>24</td>
<td>17.8</td>
</tr>
<tr>
<td>Yeast</td>
<td>9</td>
<td>9</td>
<td>22</td>
<td>16.7</td>
</tr>
</tbody>
</table>

*Minimum, maximum and average number of individual sequences from which the consensus profiles have been deduced.

with selected transcription factor binding sites. Conslnspector is a stand-alone application, but was also integrated in higher-level methods for the definition and recognition of regulatory regions (Frech et al., 1996; Frech et al., in preparation; Quandt et al., 1996a).

Conslnspector 3.0 is written in ANSI C. It is available as source code for UNIX and VAX/VMS with a menu-driven user interface. DOS and Macintosh versions (client modules for all Macintosh computers, compute-server for Power Macintosh only) with a graphical user interface are available as executables. Conslnspector 3.0 including the consensus library can be obtained via anonymous ftp from ariane.gsf.de or from our WWW-site http://www.gsf.de/biodv/.

Acknowledgements

We want to thank Thomas Taubenberger for programming the graphical user interface for DOS, and Ulrich Rückert for development of the Macintosh user interface of Conslnspector. This work was supported by the BMBF Verbundprojekt GENUS 413-4001-01 IB 306 D (Förderschwerpunkt Bioinformatik).

References


Received on September 2, 1996; revised on September 27, 1996; accepted on October 7, 1996.