Prediction of low complexity regions (LCRs) using the “seg” algorithm

1. Purpose

Low complexity regions (LCRs) are stretches of non-random, simplistic amino acid sequence order (‘compositionally-biased regions’). They are naturally abundant, and can be identified by ‘seg’, a legacy sequence analysis program from NIH. seg is embedded in InterProScan; it calculates a measure of sequence complexity. Although the related ‘blast’ program has incorporated some seg features in order to meaningfully score pairwise alignments, seg data is still useful on its own.

2. Input Data

Input protein data from public repositories are run against seg, parsed, and included in the ViPR and IRD as “value-added” content.

3. Output Data

Output is presented in detail pages as “Other Domains/Motifs”. The seg program will identify the predicted low complexity regions (start, finish) and their complexity scores. In the raw output, sequences are presented in fasta format. The whole query sequence is divided into low-complexity segments (left side, lower case) and high-complexity segments (right side, upper case). All sequence segments are read from left to right, top to bottom, and are shown by a central column. The data is parsed and presented in the IRD and ViPR front end, for use as below.

<table>
<thead>
<tr>
<th>Domain/Motif</th>
<th>Start</th>
<th>End</th>
<th>Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>low_complexity</td>
<td>320</td>
<td>337</td>
<td>seg</td>
</tr>
</tbody>
</table>

4. References

Standalone seg can be downloaded at:

Supporting documentation on InterProScan can be found at:
http://computing.bio.cam.ac.uk/local/doc/iprscan/README.html

More InterProScan component algorithms are described in the SOP:


