Short String Search

1. Purpose

This document describes the use of a string matching method, to identify short linear sequences of amino acids in polypeptides. The blastP method is only designed to find sequence similarities for > 20-mers, so a custom algorithm for finding short stretches of identities is necessary.

2. Method Description

The method has several operational features.

1. GAP penalties are set HIGH. Fuzzy searches that return raw gapped alignment results are thus rejected.
2. Overlapping matches, such as the termini of oligopeptides, are not allowed. The target match has to be the same length as the query sequence, since introduction of an insert or delete is crucial in small peptides.
3. Oligos of 7-20 residues are allowed. Blast is capable of identifying larger string matches effectively.
4. ‘exact’ and ‘pattern’ searches call the seedtop program, and ‘fuzzy’ searches call the legacy program bl2seq, using the blosum62 similarity matrix.
5. A ticket number is assigned at runtime, and persists, so that search results can be retrieved later.
6. Non-alphabetic input characters are removed at the time of the job by an error-checking routine.
7. Sequences with Insertions and Deletions are not returned by fuzzy searches.
8. Extended IUPAC amino acid symbols can be used for pattern ambiguities, and include:

   B = "Asx";       Aspartic acid (R) or Asparagine (N)
   J = "Xle";       Leucine (L) or Isoleucine (I)
   U = "Sec";       Selenocysteine
   X = "Xxx";       Any, Unknown amino acid
   Z = "Glx";       Glutamic acid (E) or Glutamine (Q)

3. Input Data Preparation

The user selects a search type, and inputs an alphanumeric text string into an interface. The method allows input of a specified query amino acid ‘string’ of ASCII characters, written in IUPAC Amino Acid letter codes. Searches can be one of three separate types:

* EXACT matches are those for which 100% pairwise identity between target and query exist;
* FUZZY matches are those where there is a partial identity to the query, >50% total;
* PATTERN matches have an exact match to a Prosite Pattern, e.g. [LIT]-x(3)-[LT]-[NQK]-x-{G}.

If an exact or pattern search is specified, and program seedtop is called, a Prosite Pattern file (PA) is then written ‘on the fly’, and is stored. Input sequences are run through formatdb and are stored, and all other temp files (raw results) are stored at runtime. If fuzzy search is specified, the raw bl2seq results are parsed into objects by a BioPerl SeqIO method, and piped to the output.

User next selects a query set such as ‘ Arenaviridae all proteins’ or uses an upload file, and executes the search.
4. Output Data, Processing and Display

The initial output of the method is a tab-delimited table such as seen below:

Gene ID, match-string, start, stop, match score, sequence hit:

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>match-string</th>
<th>start</th>
<th>stop</th>
<th>match score</th>
<th>sequence hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>56707191</td>
<td>GVFSPLILAC</td>
<td>101</td>
<td>112</td>
<td>1.00</td>
<td>MQIFVGVFSPLILACNA...</td>
</tr>
<tr>
<td>56707193</td>
<td>GVFSPLILAC</td>
<td>6</td>
<td>17</td>
<td>1.00</td>
<td>MQIFVGVFSPLILACLI...</td>
</tr>
<tr>
<td>56707206</td>
<td>GVFSPLILAC</td>
<td>138</td>
<td>149</td>
<td>1.00</td>
<td>MQHIFVGVFSPLILACN...</td>
</tr>
<tr>
<td>56707208</td>
<td>GVFSPLILAC</td>
<td>165</td>
<td>176</td>
<td>1.00</td>
<td>MQIFVGVFSPLILACLS...</td>
</tr>
</tbody>
</table>

Results are then reformatted for the user interface, and presented as shown below:

Detailed prediction results are shown in the output. By clicking on the (i) symbol the user is redirected to the hit record, in this example the record for each virus isolate identified. The sequence of the hit is also available to the right in the colored link.

5. References