Basic Local Alignment Search Tool

A blast from the past...

<table>
<thead>
<tr>
<th>AGATCAC</th>
<th>C</th>
<th>G</th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGATCAC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CGACAG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

Why BLAST?

Sequenced Bases in GenBank

Why BLAST?

Sequenced Bases in GenBank
How Does BLAST Work?

Query sequence
MLVFAHAYHESKWAAHNQEILTPLV

BLAST Example

Query sequence
MLVFAHAYHESKWAAHNQEILTPLV

Database

Word List
MLV  AHN
LVF  HNQ
VFA  NQE
FAH  QEI
AHA  EIL
HAY  ILT
AYH  LTP
YHE  TPL
HES  PLV
ESK
SKW
KWA
WAA
AAH

LookUp Table
AAA
AAC
AAD
AAE
AAF
AAG
AAH
AAI

Database

BLAST Example

Query sequence
MLVFAHAYHESKWAAHNQEILTPLV

Word List
MLV  AHN
LVF  HNQ
VFA  NQE
FAH  QEI
AHA  EIL
HAY  ILT
AYH  LTP
YHE  TPL
HES  PLV
ESK
SKW
KWA
WAA
AAH

LookUp Table
AAA
AAC
AAD
AAE
AAF
AAG
AAH
AAI

Database

BLAST Example

Query sequence
MLVFAHAYHESKWAAHNQEILTPLV

Word List
MLV  AHN
LVF  HNQ
VFA  NQE
FAH  QEI
AHA  EIL
HAY  ILT
AYH  LTP
YHE  TPL
HES  PLV
ESK
SKW
KWA
WAA
AAH

LookUp Table
AAA
AAC
AAD
AAE
AAF
AAG
AAH
AAI

Database
**BLAST In a Nutshell**

Query sequence
MLVFAHAYHESKWAHNQEILTPLV

- Create "word list" from query sequence
- Locate words in database via "lookup table"
- Determine similarity of query sequence to each word-match sequence in database

**BLAST Program**

**LookUp Table**

<table>
<thead>
<tr>
<th>AAA</th>
<th>AAC</th>
<th>AAD</th>
<th>AAF</th>
<th>AAG</th>
<th>AAH</th>
<th>AAI</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>YYY</td>
<td>YYW</td>
<td>YYY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Database**

**BLAST Output**

Universal stress protein [Pyrococcus horikoshii]
Sequence ID: WP_048053230.1 Length: 167 Number of Matches: 1

Range 1:1 to 152

**Search Parameters**

<table>
<thead>
<tr>
<th>Program</th>
<th>blastn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word size</td>
<td>28</td>
</tr>
<tr>
<td>Expect value</td>
<td>0.06</td>
</tr>
<tr>
<td>Hitlist size</td>
<td>100</td>
</tr>
<tr>
<td>Gapopen</td>
<td>11.1</td>
</tr>
<tr>
<td>Matrix</td>
<td>BLOSUM62</td>
</tr>
<tr>
<td>Filter string</td>
<td>F</td>
</tr>
<tr>
<td>Genetic Code</td>
<td>1</td>
</tr>
</tbody>
</table>

**Database**

- **Query**
  - Id: WP_048053230.1
  - Length: 167

**Results Statistics**

- Lambda: 1.93271
- K: 2.62056
- H: 5.77265
- Length of sequence: 152
- Effective length of query: 1429
- Effective length of database: 673
- Effective search space: 9995
- Effective search space used: 929321

**BLAST Output**

**Search Parameters**

<table>
<thead>
<tr>
<th>Program</th>
<th>blastp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word size</td>
<td>8</td>
</tr>
<tr>
<td>Expect value</td>
<td>0.06</td>
</tr>
<tr>
<td>Hitlist size</td>
<td>100</td>
</tr>
<tr>
<td>Gapopen</td>
<td>11.1</td>
</tr>
<tr>
<td>Matrix</td>
<td>BLOSUM62</td>
</tr>
<tr>
<td>Filter string</td>
<td>F</td>
</tr>
<tr>
<td>Genetic Code</td>
<td>1</td>
</tr>
</tbody>
</table>

**Database**

- **Query**
  - Id: WP_048053230.1
  - Length: 167

**Results Statistics**

- Lambda: 0.18634
- K: 0.10658
- H: 0.70145
- Alpha: 0.0515
- Alpha_v: 0.49600
- Sigma: 45.8028
Normal Distributions

The widths of zebra stripes are normally distributed, with a mean of 7.3 centimeters and a standard deviation of 0.9 centimeters.

Extreme Value Distributions

Scores of optimal local alignments correspond to extreme value distributions.

Statistical Significance

Suppose we align two sequences, a query sequence and a target sequence, and we determine that their optimal local alignment score is $S = 60$.

Are the sequences similar? In other words, is a score of $S = 60$ significant? How likely is it that we would observe an alignment score of $S = 60$ by chance?

The $p$-value of an optimal local alignment score, $S$, is the likelihood that two random sequences* would have an optimal local alignment score greater than or equal to $S$.

* of the same lengths and compositions as the query and target sequences
**p-values for pairs of sequences**

What is the probability that the optimal local alignment score for two random sequences will be at least 60?

Solution 1: Count up all of the alignment scores greater than or equal to 60 and divide by the total number of alignment scores, i.e., 10,000.

Solution 2: Plug $x = 60$ into the following expression, where $\mu = 34.2$ and $\beta = 6.1$

$$1.0 - e^{-\frac{x-\mu}{\beta}}$$

**p-values for databases**

When searching a large database with many target sequences, our previous definition of the $p$-value is problematic because we can expect some small $p$-values by chance. For example, if we align a query sequence to 6,000,000 target sequences in a database, we can expect 60,000 scores with a $p$-value less than 0.01.

When we BLAST a query sequence against a database of many target sequences, the $p$-value of one of the alignment scores, $S$, indicates the likelihood that we would see a score of at least $S$ when BLASTing the query sequence against a comparable random database.

**E-values**

Instead of $p$-values, BLAST reports E-values. If the alignment score of a query sequence and some target sequence in the database is $S$, the E-value is the expected number of alignments with score $S$ or higher in a random database.

<table>
<thead>
<tr>
<th>Score</th>
<th>Expect</th>
<th>Method</th>
<th>Identities</th>
<th>Positives</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>40.0 bits(92)</td>
<td>1.4</td>
<td>Compositional matrix adjust.</td>
<td>27/88(31%)</td>
<td>42/88(47%)</td>
<td>3/88(3%)</td>
</tr>
</tbody>
</table>

Query 9
MKAPAWKGRGIVSLRKLGETTSSVAQSTOPSYSD|---KLOTQP8NINAYDSIL65|MKX F G RGI ++ +A ++Q+ +S ++LT F+ ++ ++S
Sbjct 127
MKSAEPVESGERIIINTASIAAFSEQIQGQAYAKSAGVHSLTLPAARELHGRIRMS 186

Query 66
MDGSGNGPPLSNGGDVVQGQAGAVKF|93| + GTI+ SG DDVQ+ AA F
Sbjct 187
IAQGLPFTPMLSLPPFDVQDSILAANTPF 214

**E-values depend on sequences and scoring**
If $n$ is the length of the query sequence and $m$ is the length of the target database...

What is the runtime of computing a pairwise alignment?

What is the runtime of BLAST?

How can we improve this runtime?