Basic Local Alignment Search Tool

A blast from the past...

<table>
<thead>
<tr>
<th>AGATCAC</th>
<th>A</th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

GATCA

GA–CA
Why BLAST?

Growth of GenBank
(1982 - 2005)

While you were sleeping...

LookUp Table

Database
BLAST Example

Query sequence
MLVFAHAYHESKWAHNQEILTPLV

LookUp Table

Database

BLAST Example

Query sequence
MLVFAHAYHESKWAHNQEILTPLV

Word List

LookUp Table

Database

Word List

LookUp Table

Database
BLAST Example

Query sequence
MLVFAHAYHESKWAAHNQEILTPLV

Word List
MLV AHN
LVF HNQ
VFA NQE
FAH QEI
AHY EIL
AYH ILY
AYH PLP
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
AHY EIL
AYH ILY
AYH PLP
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
MLVFAHAYHESKWAAHNQEILTPLV

- 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

BLAST (Basic Local Alignment Search Tool)

1. Enter Query Sequence
   - Enter accession number, gi, or FASTA sequence
   - Or, upload file
   - Job Title

2. Choose Search Set
   - Database
   - Organism
   - Optional
   - Enumer Query

3. Program Selection
   - Algorithm
     - Identity (protein-protein BLAST)
     - PiR (Position-Specific Iterated BLAST)
     - PSI-BLAST (Position-Specific Iterated BLAST)
     - Position Iterated BLAST (IP-PSI-BLAST)

4. Query
   - Enter accession number, gi, or FASTA sequence
   - Or, upload file
   - Job Title

5. Search
   - Database
   - Organism
   - Optional
   - Enumer Query

6. Program Selection
   - Algorithm
     - Identity (protein-protein BLAST)
     - PiR (Position-Specific Iterated BLAST)
     - PSI-BLAST (Position-Specific Iterated BLAST)
     - Position Iterated BLAST (IP-PSI-BLAST)

7. Result
   - Alignment
   - Similarity
   - Score

8. Result
   - Alignment
   - Similarity
   - Score
BLAST Output

**dbj|BAA29916|** (AP000003) 170aa long hypothetical protein [Pyrococcus horikoshii]

**Length = 170**

Score = 107 hits (264), Expect = 6e-21

Identities = 63/180 (39%), Positives = 97/160 (60%), Gaps = 7/160 (4%)

Query: 1 MSVMYKILYPTDFSETAEIAKRVFAFXTLKAEEVILLHVIDEREIKKKDDIFSLLGOVA 60

Sbjct: 1 HIFMFKRYLEPTDFEGAYEVEFKKVQEVLELVIDGLEE---LMDGYS 55

Query: 61 GLNKSVEPENELKXKLEEsXHREMENIKKELEDV---CFKVDIIIVQGPHENIVEKIAED 118

Sbjct: 56 FTYDNEELIDKXKLLEEsSRLQGEKAEEVKAFRAKNVRT1RFIPDEEKVAEE 115

Query: 119 BGVYIIEOSXKXTHLEEllLSSTVETNYKKNWPVIYVK 158

Sbjct: 116 ENYSLILPSRGLSLSHEFLGSTVWVRLRTKVPILIK 155

---

**Step 6. Statistical details of the search**

1. **Database:** All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples
2. **Posted date:** Feb 29, 2008 6:04 PM
3. **Number of letters in database:** 2,144,987,218
   **Number of sequences in database:** 6,276,778
4. **Lambda**  
   K  
   H  
   0.314  0.135  0.352
   **Gapped Lambda**  
   K  
   H  
   0.267  0.0410  0.140
5. **Matrix:** BLOSUM62
6. **Gap Penalties:** Existence: 11, Extension: 1
The heights of women are normally distributed, with a mean of 65.5 inches and a standard deviation of 2.5 inches.
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 of two 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.
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.

![E-values depend on sequences and scoring](image-url)