Session 3: BLAST algorithm
Learning Objectives

• Understand the principles of the BLAST algorithm
• Understand the different BLAST programs, parameters and their applications
• Be able to adjust the sensitivity and specificity of BLAST searches by adjusting parameters
• Understand and evaluate BLAST results
Learning Outcomes

• Select the correct BLAST algorithm and database for the appropriate biological question being asked, i.e. select between blastp, blastn, blastx etc. based on the question being addressed

• Understand the meaning of the various output metrics/results

• Comment on results based on the various output metrics
Why do a BLAST search?

• You can get important clues about the function of an as yet uncharacterised sequence.
• Identify homologous species for unknown species.
• Locate domains in protein sequence.
• Establish phylogeny.
• Mapping DNA from unknown location.
Principles of BLAST

• **Basic Local Alignment Search Tool** (Altschul et al., 1990) finds regions of local similarity between sequences.

• Compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches.

• Heuristic method for local alignment.

• Based on the assumption that good alignments contain short lengths of exact matches.
3 Parts of BLAST

1. Set-up

2. Find local alignments between the query sequence and a sequence in a database.

3. Produce p-values and a rank ordering of the local alignments according to the p-values.
BLAST - Part 1

1. Set-up
   - Read in query, search parameters and database
   - Check for low complexity or other repeats (optional)
Introduction to Bioinformatics online course: IBT

Query sequence

Select Database

Select Blast protocol

E-value

Turn on filter

Execute job

2. Find local alignments

   - Break query sequence into “words”
     • Typically, 3 for protein sequence and 11 for nucleotide sequence
   - Matches between the “words” in the “query” and “words” in the database sequences are found.
   - These can be exact matches (for nt-nt search) or matches that satisfy some positive-valued threshold score (for prot-prot search) as determined using a scoring matrix.
Sequences to “words”

MRRGLLEIALGFTVLLASYTSHGA

Break query sequence into “words”

Break database sequences into “words”
Find location of matching “words” in the database sequences

<table>
<thead>
<tr>
<th>Query “words”</th>
<th>Database sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRR</td>
<td>GRIAEVAARFTLDAMPGBKQMAIDADLNAGNIDETEAQRRGK</td>
</tr>
<tr>
<td>RRG</td>
<td>RRRRLCVLEIACVALWAPGAAGGQSKPPRGADEEPREE</td>
</tr>
<tr>
<td>RGR</td>
<td>VTKQTQSAKRRGGSVLRGPNVC</td>
</tr>
<tr>
<td>GRL</td>
<td>QLGRLVLGWAACWAALQAHHGQLDGLQASRFT</td>
</tr>
<tr>
<td>RLL</td>
<td></td>
</tr>
<tr>
<td>LLE</td>
<td></td>
</tr>
<tr>
<td>LEI</td>
<td></td>
</tr>
<tr>
<td>EIA</td>
<td></td>
</tr>
</tbody>
</table>

Introduction to Bioinformatics online course: IBT
Sequence Alignment | Sonal Henson
2. Find local alignments, continued (1)

- Each “word” match is scored

  • The score is computed by assigning a value to each aligned pair of letters and then summing these values over the length of the alignment.

  • For protein sequence alignments, scores for every possible amino acid letter pair are given in a “substitution matrix” where likely substitutions have positive values and unlikely substitutions have negative values. Default matrix used by BLASTP is BLOSUM62.

  • For nucleotide alignments, BLAST uses a reward of +2 for aligned pairs of identical letters and a penalty of –3 for each non-identical aligned pair.
• Words that do not score above a given threshold (T) are removed from the list.
  – E.g. When T is 13, AAI is removed.
• Matches with a score above a given threshold are used to “seed” an ungapped alignment.
Extend Hits

• The “words” are extended in both directions into gap-free extensions for as long as the cumulative alignment score increases or stays the same.

• Gap-free extensions with alignment score above a given threshold are used to seed gapped extensions.

• Gapped extensions (HSPs) scoring above an empirically determined cut-off score, S, are retained.
**BLAST - Part 2**

- HSP is the fundamental unit of BLAST algorithm output.
- An HSP consists of two sequence fragments of arbitrary but equal length whose alignment is locally maximal and for which the alignment score meets or exceeds a cut-off score, S.

```
Query 1  MRRGRLLEIALGFTVLLASYTSHGA 25
  M  RGRL+  +A+G  +L+  +T  G
Sbjct 1  MCRGRLVRLAVGLVAVLSLWTEPGG 25
```

**High-Scoring Segment Pair (HSP)**
3. Determine the statistical significance of each HSP score (Raw score -> P-value -> E-value).
   – What is the probability (P-value) of the HSP alignment occurring by chance?
   – P-value is the probability of finding exactly a HSPs with score >=S
BLAST - Part 3

Problems:

• Longer sequences are more likely to find higher scoring pairs.
• Longer databases are more likely to result in higher scoring pairs.

Solution:

• Convert Probability values (P-values) to Expectation values (E-values)
BLAST – Part 3

• The “Expect Value” (E-value) is the number of times that an alignment as good or better than that found by BLAST would be expected to occur by chance, given the size of the database searched.

\[ E = \text{Length of Database} \times \text{Length of Sequence} \times \text{Probability} \]

• E-value is dependent on the length of the sequence and the length of the database

As the database grows the E-value will change.

• E-value < 1e-179 is written as 0.0
Bit Score

- Bit score ($S'$) is a normalised raw score that is independent of the search space.
  - Search space ($N$) = Query length ($n$) x Sum of length of sequences in the database ($m$)

- It measures sequence similarity independent of query sequence length and database size and is normalized based on the raw pairwise alignment score.
  - It is linearly related to the raw score.
  - The higher the Bit score, the more significant the match is.
Take a moment to reflect

1. How are the scores and E-value related? Will a high score give a high or low E-value?
2. Would a longer sequence give a higher or lower E-value?
3. How would an increase in database size affect the E-value?
4. What effect will adjusting the gap penalty have on the alignment?
### BLAST Output

**Job title: Protein Sequence (59 letters)**

<table>
<thead>
<tr>
<th>RID</th>
<th>HNMHTF3A014 (Expires on 05-17 18:19 pm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query ID</td>
<td>lcl</td>
</tr>
<tr>
<td>Description</td>
<td>None</td>
</tr>
<tr>
<td>Molecule type</td>
<td>amino acid</td>
</tr>
<tr>
<td>Query Length</td>
<td>59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Database Name</th>
<th>nr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects</td>
</tr>
</tbody>
</table>

**Program**

BLASTP 2.6.1+

**Other reports:**
- Search Summary
- Taxonomy reports
- Distance tree of results
- Multiple alignment
- MSA viewer

**Graphs**
- Graphic Summary
- Descriptions
- Alignments
BLAST Output

Graphic Summary

No putative conserved domains have been detected

Distribution of the top 100 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

Color key for alignment scores:
- <40
- 40-50
- 50-80
- 80-200
- >=200

Query

1 10 20 30 40 50
### BLAST Output

#### Descriptions

<table>
<thead>
<tr>
<th>Description</th>
<th>Max score</th>
<th>Total score</th>
<th>Query cover</th>
<th>E value</th>
<th>Ident</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>fibrillin 1 (Marfan syndrome), isoform CRA_a [Homo sapiens]</td>
<td>126</td>
<td>126</td>
<td>100%</td>
<td>2e-32</td>
<td>100%</td>
<td>EAW77355.1</td>
</tr>
<tr>
<td>RecName: Full=Fibrillin-1; Contains: RecName: Full=Asprosin; Flags: Precursor</td>
<td>126</td>
<td>126</td>
<td>100%</td>
<td>2e-32</td>
<td>100%</td>
<td>P35555.3</td>
</tr>
<tr>
<td>fibrillin 1 [Homo sapiens]</td>
<td>126</td>
<td>126</td>
<td>100%</td>
<td>2e-32</td>
<td>100%</td>
<td>BAD16739.1</td>
</tr>
<tr>
<td>fibrillin [Homo sapiens]</td>
<td>126</td>
<td>126</td>
<td>100%</td>
<td>2e-32</td>
<td>100%</td>
<td>AAB02036.1</td>
</tr>
<tr>
<td>fibrillin-1 proprotein [Homo sapiens]</td>
<td>126</td>
<td>126</td>
<td>100%</td>
<td>2e-32</td>
<td>100%</td>
<td>NP_000129.3</td>
</tr>
<tr>
<td>fibrillin 1 [Homo sapiens]</td>
<td>125</td>
<td>125</td>
<td>100%</td>
<td>2e-32</td>
<td>100%</td>
<td>BAD16738.1</td>
</tr>
<tr>
<td>fibrillin [Homo sapiens]</td>
<td>124</td>
<td>124</td>
<td>100%</td>
<td>6e-32</td>
<td>100%</td>
<td>CAA45118.1</td>
</tr>
<tr>
<td>unnamed protein product [Homo sapiens]</td>
<td>124</td>
<td>124</td>
<td>100%</td>
<td>7e-32</td>
<td>100%</td>
<td>BAG65498.1</td>
</tr>
<tr>
<td>PREDICTED: fibrillin-1 [Pan paniscus]</td>
<td>124</td>
<td>124</td>
<td>100%</td>
<td>9e-32</td>
<td>98%</td>
<td>XP_003831532.1</td>
</tr>
<tr>
<td>PREDICTED: fibrillin-1-like [Pongo abelii]</td>
<td>115</td>
<td>115</td>
<td>100%</td>
<td>2e-31</td>
<td>97%</td>
<td>XP_009248093.1</td>
</tr>
<tr>
<td>PREDICTED: fibrillin-1 [Carlito syrichta]</td>
<td>123</td>
<td>123</td>
<td>100%</td>
<td>3e-31</td>
<td>98%</td>
<td>XP_008068093.1</td>
</tr>
<tr>
<td>PREDICTED: LOW QUALITY PROTEIN: fibrillin-1 [Roussetts aegyptiacus]</td>
<td>122</td>
<td>122</td>
<td>100%</td>
<td>3e-31</td>
<td>97%</td>
<td>XP_016018715.1</td>
</tr>
<tr>
<td>PREDICTED: fibrillin-1 isoform X3 [Ceratothorium simum simum]</td>
<td>122</td>
<td>122</td>
<td>100%</td>
<td>4e-31</td>
<td>97%</td>
<td>XP_014836182.1</td>
</tr>
<tr>
<td>PREDICTED: fibrillin-1 [Camelus bactrianus]</td>
<td>122</td>
<td>122</td>
<td>100%</td>
<td>4e-31</td>
<td>97%</td>
<td>XP_010853648.1</td>
</tr>
<tr>
<td>PREDICTED: fibrillin-1 [Camelus ferus]</td>
<td>122</td>
<td>122</td>
<td>100%</td>
<td>4e-31</td>
<td>97%</td>
<td>XP_006192393.1</td>
</tr>
<tr>
<td>PREDICTED: fibrillin-1 isoform X2 [Ceratothorium simum simum]</td>
<td>122</td>
<td>122</td>
<td>100%</td>
<td>5e-31</td>
<td>97%</td>
<td>XP_014838181.1</td>
</tr>
<tr>
<td>PREDICTED: fibrillin-1 [Chrysochloris asiatica]</td>
<td>122</td>
<td>122</td>
<td>100%</td>
<td>5e-31</td>
<td>97%</td>
<td>XP_006831741.1</td>
</tr>
<tr>
<td>PREDICTED: fibrillin-1-like [Physeter catodon]</td>
<td>122</td>
<td>122</td>
<td>100%</td>
<td>5e-31</td>
<td>97%</td>
<td>XP_007131588.1</td>
</tr>
<tr>
<td>PREDICTED: fibrillin-1 isoform X2 [Orcinus orca]</td>
<td>122</td>
<td>122</td>
<td>100%</td>
<td>5e-31</td>
<td>97%</td>
<td>XP_004281354.1</td>
</tr>
<tr>
<td>PREDICTED: fibrillin-1 isoform X1 [Ceratothorium simum simum]</td>
<td>122</td>
<td>122</td>
<td>100%</td>
<td>5e-31</td>
<td>97%</td>
<td>XP_004421808.1</td>
</tr>
<tr>
<td>PREDICTED: fibrillin-1 [Dasypus novemcinctus]</td>
<td>122</td>
<td>122</td>
<td>100%</td>
<td>5e-31</td>
<td>97%</td>
<td>XP_004448132.1</td>
</tr>
</tbody>
</table>
### BLAST Output

#### Alignments

**fibrillin 1** (Marfan syndrome), isoform CRA_a [Homo sapiens]
Sequence ID: **EAW77353.1**  Length: 2869  Number of Matches: 1

<table>
<thead>
<tr>
<th>Range: 1:1 to 59</th>
<th>GenPept</th>
<th>Graphics</th>
<th>Score</th>
<th>Expect</th>
<th>Method</th>
<th>Identities</th>
<th>Positives</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>126 bits(316)</td>
<td>2e-32</td>
<td>Composition-based stats.</td>
<td>59/59(100%)</td>
<td>59/59(100%)</td>
<td>0/59(0%)</td>
</tr>
</tbody>
</table>

Query 1
MRGRGLLEIALGFTLLASLYTHGADANLEAGNVKETRASRAKRGGGHDALKGPNCV 59

Subject 1
MRGRGLLEIALGFTLLASLYTHGADANLEAGNVKETRASRAKRGGGHDALKGPNCV 59

---

**RecName**: Full=Fibrillin-1; Contains: RecName: Full=Asprosin; Flags: Precursor
Sequence ID: **P35553.3**  Length: 2871  Number of Matches: 1

<table>
<thead>
<tr>
<th>Range: 1:1 to 59</th>
<th>GenPept</th>
<th>Graphics</th>
<th>Score</th>
<th>Expect</th>
<th>Method</th>
<th>Identities</th>
<th>Positives</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>126 bits(316)</td>
<td>2e-32</td>
<td>Composition-based stats.</td>
<td>59/59(100%)</td>
<td>59/59(100%)</td>
<td>0/59(0%)</td>
</tr>
</tbody>
</table>

Query 1
MRGRGLLEIALGFTLLASLYTHGADANLEAGNVKETRASRAKRGGGHDALKGPNCV 59

Subject 1
MRGRGLLEIALGFTLLASLYTHGADANLEAGNVKETRASRAKRGGGHDALKGPNCV 59

---

**PREDICTED**: fibrillin-1 [Myotis lucifugus]
Sequence ID: **XP_006086058.1**  Length: 2871  Number of Matches: 1

<table>
<thead>
<tr>
<th>Range: 1:1 to 59</th>
<th>GenPept</th>
<th>Graphics</th>
<th>Score</th>
<th>Expect</th>
<th>Method</th>
<th>Identities</th>
<th>Positives</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>115 bits(287)</td>
<td>2e-28</td>
<td>Composition-based stats.</td>
<td>53/59(90%)</td>
<td>55/59(93%)</td>
<td>0/59(0%)</td>
</tr>
</tbody>
</table>

Query 1
MRGRGLLEIALGFTLLASLYTHGADANLEAGNVKETRASRAKRGGGHDALKGPNCV 59
MRGGLLEVALLGFTLLASLYTHGADTRVAAGNAKETRASRAKRGGGHDALKGPNCV 59

Subject 1
MRGRGLLEIALGFTLLASLYTHGADANLEAGNVKETRASRAKRGGGHDALKGPNCV 59
### BLAST Output (nucleotide)

<table>
<thead>
<tr>
<th>Score</th>
<th>Expect</th>
<th>Identities</th>
<th>Gaps</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>366 bits(198)</td>
<td>1e-94</td>
<td>274/311(88%)</td>
<td>4/311(1%)</td>
<td>Plus/Minus</td>
</tr>
<tr>
<td><strong>Query</strong></td>
<td><strong>Sbjct</strong></td>
<td><strong>Query</strong></td>
<td><strong>Sbjct</strong></td>
<td><strong>Query</strong></td>
</tr>
<tr>
<td>48468796</td>
<td>11303</td>
<td>48468856</td>
<td>11243</td>
<td>48468916</td>
</tr>
<tr>
<td>CGGTGGCTCACGCTTGTAATCCCAAGACAAGTTGGGAGGCCAGGCAGGACAGGTCCAAGT</td>
<td>CGGTGGCTCACGCTTGTAATCCCAAGACAAGTTGGGAGGCCAGGCAGGACAGGTCCAAGT</td>
<td>CAGGAGATCGAGACCATCTGGCTAACACAGTGAAAACCTCATCTCTACTAAAAATACAAA</td>
<td>CAGGAGATCGAGACCATCTGGCTAACACAGTGAAAACCTCATCTCTACTAAAAATACAAA</td>
<td>AAATTAGCCAGGCTGTTGGTGGGCTGTAGTCCACGCTACTGCGGAGGCTGAGGCAG</td>
</tr>
<tr>
<td>48468855</td>
<td>11244</td>
<td>48468915</td>
<td>11184</td>
<td>48468975</td>
</tr>
<tr>
<td>GAGAATGGCAGTGAACCTGGAGGCGGAGCTGGAGTCCAGTGACGCAAGATGTTGAGACTGACTGACT</td>
<td>GAGAATGGCAGTGAACCTGGAGGCGGAGCTGGAGTCCAGTGACGCAAGATGTTGAGACTGACT</td>
<td>CCAAGCTGGGCCAGCAGCGAGACTCCGTCTCaaaaaa---aaa---aaaatatatatatatata</td>
<td>CCAAGCTGGGCCAGCAGCGAGACTCCGTCTCaaaaaa---aaa---aaaatatatatatatata</td>
<td>48469035</td>
</tr>
<tr>
<td>11123</td>
<td>TACAATAAAAAT</td>
<td>10994</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Low Complexity Region

• “Low complexity region” is a region of sequence composed of few kinds of elements (low compositional complexity).

• They might give high scores that confuse the program to find the actual significant sequences in the database.

• Marked with X (protein sequences) or N (nucleotide sequences) and ignored in the BLAST program.

• SEG (proteins) or DUST (nucleotides)
Flavours of BLAST

- **Nucleotide BLAST**
  - nucleotide → nucleotide

- **Protein BLAST**
  - protein → protein

- **blastx**
  - translated nucleotide → protein

- **tblastn**
  - protein → translated nucleotide

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Can align two or more sequences in all flavours

Introduction to Bioinformatics online course: IBT
Sequence Alignment | Sonal Henson
PSI-BLAST

• Used to find distant relatives of a protein.
• Does an iterative search.
  – First search a regular BlastP
  – Generates a multiple alignment of the HSPs above an E-value threshold and calculates a profile or a Position Specific Scoring Matrix (PSSM)
    • PSSM captures the conservation pattern in the alignment and records it as a scoring matrix
  – In the next iteration this profile is used instead of the original substitution matrix to detect sequences that match the conservation pattern specified by the PSSM
  – After every iteration new sequences above the threshold are added to the the PSSM
• In this way, PSI-BLAST allows detection of distant sequence similarities.
Specialized Searches

- **SmartBLAST**: Find proteins highly similar to your query.
- **Primer-BLAST**: Design primers specific to your PCR template.
- **Global Align**: Compare two sequences across their entire span (Needleman-Wunsch).
- **CD-search**: Find conserved domains in your sequence.
- **GEO**: Find matches to gene expression profiles.
- **IgBLAST**: Search immunoglobulins and T cell receptor sequences.
- **VecScreen**: Search sequences for vector contamination.
- **CDART**: Find sequences with similar conserved domain architecture.
- **Targeted Loci**: Search markers for phylogenetic analysis.
- **Multiple Alignment**: Align sequences using domain and protein constraints.
- **BioAssay**: Search protein or nucleotide targets in PubChem BioAssay.
- **MOLE-BLAST**: Establish taxonomy for uncultured or environmental sequences.
Summary

• The algorithm underlying a BLAST search is complex.
• By understanding it you can adjust the sensitivity and specificity of the search.
• Higher the Bit score the lower the E-value
  – E-value depends on the size of the database and the length of the query.
  – A “significant” E-value of a result will depend on your sequence and goal of the search.
• How to choose amongst the different flavours of BLAST for proteins and nucleotides
Useful Resources

- **BLAST QuickStart**
- **BLAST Handbook**
- **PSI-BLAST**
Introduction to Bioinformatics online course: IBT
Sequence Alignment | Sonal Henson