Tutorial 4

BLAST – database similarity search
BLAST tutorial

• DB similarity search
• How to use BLAST
• The statistics behind a BLAST search: Score, E-value, P-value
• Statistical significance

• Cool story of the day:
From searching biological DBs to genetic engineering
DB similarity search

Goals:

• Find genes/proteins with possibly similar function
• Find the origin of a sequence (what organism it is taken form)
DB similarity search

• Computationally infeasible for large DBs

• How to define “best match”?
DB similarity search

1. Seed search - Reduces set of candidate DB sequences
2. Greedy seed extension
3. Full local alignment to calculate score
BLAST

What is BLAST?

- **Basic Local Alignment Search Tool**
- Set of similarity search programs for exploring sequence databases.
## BLAST Databases

<table>
<thead>
<tr>
<th>Query type</th>
<th>Database type</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastn</td>
<td>Genomic, Genomic</td>
</tr>
<tr>
<td>blastp</td>
<td>Proteomic, Proteomic</td>
</tr>
<tr>
<td>blastx</td>
<td>Translated genomic, Proteomic</td>
</tr>
<tr>
<td>tblastn</td>
<td>Proteomic, Translated genomic</td>
</tr>
<tr>
<td>tblastx</td>
<td>Translated genomic, Translated genomic</td>
</tr>
</tbody>
</table>

**Genomic:** A T G C  
**Proteomic:** G A S T C V L I M P F Y W D E N Q H K R

**Translated genomic:** The query is genomic, translated to protein using 6 possible reading frames

ATGCCGTTC -> **MPF**, **CR**, **AV**
BLAST is part of the NCBI genome browser

Query and DB parameters

Place Query

Job title – helpful when running multiple runs

Choose Database

In case you want to restrict to a specific organism

In case you want to eliminate specific sequences
How to choose the database?

- Depends on what you’re looking for...
- Why is it important to narrow the DB if possible?

A good place to start if you don’t know what you’re looking for

**nr/nt**: non-redundant nucleotide
Alignment parameters

Optimizes the parameters for the desired similarity level of the search
Alignment parameters

**General Parameters**
- Max target sequences: 100
- Short queries: Automatically adjust parameters for short input sequences
- Expect threshold: 10
- Word size: 28
- Max matches in a query range: 0

**Scoring Parameters**
- Match/Mismatch Scores: 1.2
- Gap Costs: Linear

**Filters and Masking**
- Low complexity regions
- Species-specific repeats for: Homo sapiens (Human)
- Mask for lookup table only
- Mask lower case letters

**BLAST**
- Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
- Show results in a new window

- **Threshold for results significance**
- **Primary word (Seed) match (16-64 nt)** Smaller length - more sensitive.
- **Scores of matching and mismatching bases**
- **Cost to create and extend a gap**
How to interpret BLAST results?
Search for homologous to chick “olfactory receptor 6” gene
Search results

<table>
<thead>
<tr>
<th>RID</th>
<th>2W07XWMZ01R (Expires on 11-19 02:11 am)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query ID</td>
<td>ID</td>
</tr>
<tr>
<td>Description</td>
<td>None</td>
</tr>
<tr>
<td>Molecule type</td>
<td>nucleic acid</td>
</tr>
<tr>
<td>Query Length</td>
<td>420</td>
</tr>
</tbody>
</table>

Other reports: [Search Summary] [Taxonomy reports] [Distance tree of results]

**Graphic Summary**

The image shows a diagram illustrating the distribution of 15 Blast Hits on the Query Sequence. The `Color key for alignment scores` is displayed with the following range:

- `<40`
- `40-50`
- `50-80`
- `80-200`
- `>=200`

The `Query sequence` is indicated by the horizontal line, and the `Matched sequences from DBs` are shown as vertical lines colored according to the alignment scores.
### Descriptions

#### Sequences producing significant alignments:

<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>Danio rerio myotubularin related protein 2, mRNA (cDNA clone MGC:192557 IMAGE:100)</td>
<td>776</td>
<td>776</td>
<td>100%</td>
<td>0.0</td>
<td>100%</td>
<td>BC165469.1</td>
</tr>
<tr>
<td>Danio rerio myotubularin related protein 2 (mtrr2), mRNA</td>
<td>776</td>
<td>776</td>
<td>100%</td>
<td>0.0</td>
<td>100%</td>
<td>NM_131371.1</td>
</tr>
<tr>
<td>PREDICTED: Sinocyclocheilus rhinoceros myotubularin-related protein 2 (LOC107747)</td>
<td>508</td>
<td>508</td>
<td>99%</td>
<td>3e-140</td>
<td>89%</td>
<td>XM_016562420.1</td>
</tr>
<tr>
<td>PREDICTED: Sinocyclocheilus anshuensis myotubularin-related protein 2 (LOC107680)</td>
<td>486</td>
<td>486</td>
<td>99%</td>
<td>1e-133</td>
<td>88%</td>
<td>XM_016476277.1</td>
</tr>
<tr>
<td>PREDICTED: Sinocyclocheilus grahami myotubularin-related protein 2-like (LOC107597)</td>
<td>486</td>
<td>486</td>
<td>99%</td>
<td>1e-133</td>
<td>88%</td>
<td>XM_016288179.1</td>
</tr>
<tr>
<td>PREDICTED: Sinocyclocheilus grahami myotubularin-related protein 2-like (LOC107597)</td>
<td>486</td>
<td>486</td>
<td>99%</td>
<td>1e-133</td>
<td>88%</td>
<td>XM_016288179.1</td>
</tr>
<tr>
<td>PREDICTED: Sinocyclocheilus rhinoceros myotubularin-related protein 2-like (LOC107597)</td>
<td>486</td>
<td>486</td>
<td>99%</td>
<td>1e-133</td>
<td>88%</td>
<td>XM_016288179.1</td>
</tr>
<tr>
<td>PREDICTED: Sinocyclocheilus rhinoceros myotubularin-related protein 2-like (LOC107597)</td>
<td>486</td>
<td>486</td>
<td>99%</td>
<td>1e-133</td>
<td>88%</td>
<td>XM_016288179.1</td>
</tr>
<tr>
<td>PREDICTED: Sinocyclocheilus grahami myotubularin-related protein 2-like (LOC107597)</td>
<td>486</td>
<td>486</td>
<td>99%</td>
<td>1e-133</td>
<td>88%</td>
<td>XM_016288179.1</td>
</tr>
<tr>
<td>PREDICTED: Sinocyclocheilus grahami myotubularin-related protein 2-like (LOC107597)</td>
<td>486</td>
<td>486</td>
<td>99%</td>
<td>1e-133</td>
<td>88%</td>
<td>XM_016288179.1</td>
</tr>
<tr>
<td>PREDICTED: Sinocyclocheilus anshuensis myotubularin-related protein 2-like (LOC107597)</td>
<td>486</td>
<td>486</td>
<td>99%</td>
<td>1e-133</td>
<td>88%</td>
<td>XM_016288179.1</td>
</tr>
<tr>
<td>Zebrafish DNA sequence from clone DKEY-110K5 in linkage group 5, complete sequence</td>
<td>193</td>
<td>747</td>
<td>94%</td>
<td>4e-45</td>
<td>99%</td>
<td>AL929305.4</td>
</tr>
</tbody>
</table>
Descriptions

Query covered=83%
Only 83% of the query is covered => ~348 bp

Identity=90%
Out of the 348 bp of alignment, 90% were matches

E-value=$e^{-122}$
This alignment is very significant
Alignments

DB sequence info
Alignment info
Alignment
It is possible to get multiple hits per sequence.
E-values and scores
Score vs. E-value

- The **score** is a measure of the **similarity** of the query to the a sequence from the database.

- The **E-value** is a measure of the **reliability** of the score.
Score

Score \( (S) = \sum (\text{identities} + \text{mismatches}) - \sum \text{gaps} \)

Bit Score \( (S') \): \[ S' = \frac{\lambda S - \ln K}{\ln 2} \]

- Unless the scoring system is understood, citing a raw score alone is like citing a distance without specifying feet, meters, or light years.
- By normalizing a raw score one attains a "bit score" \( S' \), which has a standard set of units.
- The parameters \( K \) and lambda can be thought of simply as natural scales for the search space size and the scoring system respectively.

E-value

The definition of the E-value is:
The number of expected alignments with observed score or higher due to chance.

\[ E = Kmne^{-\lambda S} \]

- \( E \)-values cannot be compared across different DBs, even if the score is the same.
The statistics behind scoring a DB match

Hypothesis testing

Experiment 1: 10 coin flips – 6 times “heads”.

Experiment 2: 100 coin flips – 60 times “heads”.

Experiment 2: 1000 coin flips – 600 times “heads”.

Questions:

• Is the coin fair?

• How certain can we be in the answer?
The statistics behind scoring a DB match

Hypothesis testing

- **Null hypothesis** $H_0$ - the conservative hypothesis on which we want to defend
  (The coin is fair $\Rightarrow p = 0.5$)

- **Alternative hypothesis** $H_1$ - a new hypothesis that we want to check
  (The coin is biased $\Rightarrow p \neq 0.5$)

- We assume $H_0$ is true until we decide otherwise!
- We can only reject $H_0$
- We can not verify an hypothesis, only fail to reject it
The statistics behind scoring a DB match

Hypothesis testing

- **Test statistic**
  - Can be calculated from the sample
    \[ S = (\#"heads") \]
  - We know it’s distribution under \( H_0 \)
    \[ S \sim Binomial(n, p = 0.5) \]

- **p-value**
  - **Under** \( H_0 \), What is the probability to get a test statistic which is “more extreme” than the observed.
    \[ P(S \geq 6) + P(S \leq 4) \] when \( S \sim Bin(10, 0.5) \)
  - If the **probability is low** than \( H_0 \) is rejected
  - If the **probability is high** than \( H_0 \) cannot be rejected
The statistics behind scoring a DB match

**Experiment 1:** 10 coin flips – 6 times “heads”.
**Experiment 2:** 100 coin flips – 60 times “heads”.
**Experiment 3:** 1000 coin flips – 600 times “heads”.

\( H_0: \text{the coin is fair } p = 0.5; H_1: p \neq 0.5 \)

Exp1: \( S = 6, \ n = 10, \ S \sim Bin(10,0.5) \)
\( p\text{-value} = P(S \geq 6) + P(S \leq 4) = 0.75 \)

Exp2: \( S = 60, \ n = 100, \ S \sim Bin(100,0.5) \)
\( p\text{-value} = P(S \geq 60) + P(S \leq 40) = 0.057 (5.7\%) \)

Exp3: \( S = 600, \ n = 10, \ S \sim Bin(1000,0.5) \)
\( p\text{-value} = P(S \geq 600) + P(S \leq 400) < 10^{-5} \)
The statistics behind scoring a DB match

In a DB search

- We found the best match with score $S$
- $H_0$: There is no “real” match. The score is simply due to randomness in the sequences.
- $H_1$: otherwise

- Test statistic: $S$
- $p$-value:

  $P(\text{getting a Score } S \text{ or better for a random search})$
  - If the probability is low than $H_0$ is rejected (significant match)
  - If the probability is high than $H_0$ cannot be rejected
E-value and P-value

**E-value:** The number of random sequences in the DB that will align with the query with score $\geq S$.

**P-value:** The probability to find a sequence in the DB with score $\geq S$. 
Cool Story of the day

How searching biological DBs brought about the development of state of the art genetic engineering tool.
Genetic engineering
EDITING GENES WITH CRISPR

CRISPR is a tool used by scientists to precisely edit genes inside cells. It's comprised of two parts...

- **Cas9**
  - An enzyme that cuts DNA

  ![Cas9 Enzyme Cutting DNA](http://scienceblog.cancerresearchuk.org/-rpsirc/2016/02/01/eht-ni-tolb-ro-hcraeser-recnac-ni-retpahc-wen-gnitide-eneg/koobypoc-lacihte)

  + They stick together and seek out the target DNA

  ![Cas9 Enzyme and Guide RNA Binding](http://scienceblog.cancerresearchuk.org/-rpsirc/2016/02/01/eht-ni-tolb-ro-hcraeser-recnac-ni-retpahc-wen-gnitide-eneg/koobypoc-lacihte)

  - Guide RNA
    - Directs the Cas9 to the target DNA
    - Contains a sequence that matches the target DNA

  ![Guide RNA and Cas9 Cutting DNA](http://scienceblog.cancerresearchuk.org/-rpsirc/2016/02/01/eht-ni-tolb-ro-hcraeser-recnac-ni-retpahc-wen-gnitide-eneg/koobypoc-lacihte)

- **Cas9** unzips the target DNA and the guide RNA matches up...

  ![Cas9 Unzipping DNA and Guide RNA](http://scienceblog.cancerresearchuk.org/-rpsirc/2016/02/01/eht-ni-tolb-ro-hcraeser-recnac-ni-retpahc-wen-gnitide-eneg/koobypoc-lacihte)

  - Once cut, the DNA can be disabled or altered

  ![DNA Cut by Cas9](http://scienceblog.cancerresearchuk.org/-rpsirc/2016/02/01/eht-ni-tolb-ro-hcraeser-recnac-ni-retpahc-wen-gnitide-eneg/koobypoc-lacihte)

  - Scientists can now edit the gene how they wish

  ![DNA Editing with Cas9](http://scienceblog.cancerresearchuk.org/-rpsirc/2016/02/01/eht-ni-tolb-ro-hcraeser-recnac-ni-retpahc-wen-gnitide-eneg/koobypoc-lacihte)
But how was it developed?

It was first *discovered* in bacteria
The repeat sequences are found in 90% of Archaea and 40% of bacteria!

The “spacers” were found to be phage DNA...
