Today’s Topics

- Assignment 4 Review
- BLASTP & BLASTN outputs
- BLAST & FASTA statistics

**BLASTP results**

- Proportion of the query aligned
- Bit score
- E-value

**Nucleotide Similarity Search**

- Default DB NR
  - megablast*: w=28 (16~256)
  - *This is the default search method
  - discontiguous megablast: w=11 (or 12)
  - allows some mismatches
  - blastn: w=11 (7~15)
  - w=7 for a short sequence

**Discontiguous megablast**

- If discontiguous megablast is chosen:
  - Word matching based on discontiguous pattern (template)
  - e.g., for coding: 110101011011011 (w=11, t=16)
  - mismatches are allowed for '0' positions

**BLASTN results**

- megablast (only 12 hits, all E<5e-153)
- blastn (125 hits, E<6.9)
- Discontiguous megablast (57 hits, E<6.9)
**BLASTN/BLASTX results**

- **[blastn]**
  - Query: Caenorhabditis briggsae clone CB051104, complete sequence
  - Database: C. elegans genomic scaffolds

- **[blastx]** (translated query vs. protein db)
  - Low complexity region is masked (shown in lower cases)
  - 6 possible frames

**BLASTP results**

- **[blastp]**
  - Positive (+) scoring AA pairs (similar AA pairs)

**BLAST results**

- Click to see the blast search statistics

**BLAST Statistics**

- **[blastp]**
  - Raw Score ($S$): simply based on pairwise scores & gap penalties
  - Normalized Score or Bit Score ($S'_{norm}$):
    
    \[
    S'_{norm} = \left( \frac{S - log_2 K}{\log_2(1 + L)} \right) / \log_2 2
    \]
  - $\lambda = 0.267$, $K = 0.041$, $S'_{norm} = (0.267 \times 465 - \log_2(0.041)) / \log_2 2 = 183.7$

- Used to calculate the scores for the alignments with gaps

- Raw scores ($S$) depend on the scoring system; cannot be compared
- Bit scores ($S'$) are normalized using $\lambda$ and $K$; independent of scoring system; can be compared
Pairwise alignment vs. database searching

[For a pairwise alignment]

- Karlin-Altschul equation (Karlin & Altschul, 1990)
  \[ P(S \geq x) = 1 - \exp[-Kmne^{-x}] \]
  Probability of getting the alignment score \( S \geq x \) by chance

- For the alignment score \( S \geq x \),
  \[ E = NP \] (\( N \): number of random alignments; used in PRSS and LALIGN)

[For database searching]

- Multiple pairwise alignments: multiple testing problem
  - If \( P(0 \leq x) = 0.05 \),
  - For 100 alignments, 0.95 is the probability of having \( S \geq x \) by chance for one alignment
  - For 10 alignments, 0.95 is the probability of having \( S \geq x \) by chance for all 10 alignments
  - If \( P(0 \leq x) = 0.05 \),
  - For 100 alignments, 0.95 is the probability of having \( S \geq x \) by chance for one alignment
  - For 10 alignments, 0.95 is the probability of having \( S \geq x \) by chance for all 10 alignments
  - If \( P(0 \leq x) = 0.05 \),
  - For 100 alignments, 0.95 is the probability of having \( S \geq x \) by chance for one alignment
  - For 10 alignments, 0.95 is the probability of having \( S \geq x \) by chance for all 10 alignments

Bonferroni correction

- Multiple comparison correction
  Instead of using \( P = \alpha \)
  Use \( P = \alpha/N \) (for \( N \) comparisons) as the threshold
  \[ \alpha = N \times \alpha' \]

BLAST Statistics

- Karlin-Altschul equation (Karlin & Altschul, 1990)
  \[ E = Kmne^{-xS} \]

- For database searching, \( E \) is the database size (the number of entries) \( \rightarrow \) the number of alignments

BLAST Statistics

- E-value threshold can be considered as a P-value threshold corrected for multiple comparisons in database searching

P-value, E-value, and database search

- P-value for pairwise alignment = \[ 1 - \exp[-Kmne^{-x}] \]
  Probability of getting the alignment score \( \geq x \) from random pairwise comparison (\( m \) and \( n \) are the lengths of the two sequences compared)

- E-value = \[ Kmne^{-xS} \]
  \( m \): length of the query (or effective length, \( m' \))
  \( n \): length of the database (or effective length, \( n' \))

- P-value for a database search (Bonferroni corrected)
  the probability of having at least one HSP with its score \( \geq S \)
  \[ P = 1 - e^{-xS} \] (\( P \approx E \) if \( E < 0.01 \))
  \( E \): the expected number of HSPs with scores \( \geq S \)
  \( E \approx E/N \) if \( E \) is used in FASTA; \( N \): database size

Note: Calculation methods for length adjustment (/) and \( m'/n' \) have been changed based on a new finite-size correction (FSC). See Park et al. (2012, BMC Research Note)
BLAST Statistics

**Normalized Score or Bit Score (S'bit):**

\[ S'_{\text{bit}} = \frac{LS - \log K}{\log 2}, \quad S'_{\text{nat}} = \frac{LS - \log K}{\log 2} \]

\( L = 0.267, K = 0.041, S = 465 \)

\[ S'_{\text{bit}} = \frac{0.267 \times 465 - \log 0.041}{\log 2} = 183.7 \]

**Raw Score (S):** simply based on pairwise scores & gap penalties

**Expect (E):**

\[ E = Kn' e^{-\frac{m'n' S'}{2}}, \quad E = \frac{m'n' e^{-S'}}{2} \]

\[ E = 0.041 \times 570 \times 53,472,480,294 \times e^{-183.7} \]

\[ E = 570 \times 53,472,480,294 \times 2^{-183} \]

**Probability (P):**

\[ P = 1 - e^{-E} \]

\[ P = 1 - e^{-0.041 \times 570 \times 53,472,480,294 \times e^{-183.7}} \]

\[ P = 1 - e^{-570 \times 53,472,480,294 \times 2^{-183}} \]

\[ P \approx 0 \] (if \( E \approx 0 \))

**BLAST search summary statistics**

- Word size (W)
- Scoring matrix & gap penalties
- Length separating two HSPs to trigger extension (A: two-hit methods)
- Alignment threshold (not user-definable)
- Neighborhood threshold (T)

**BLAST Statistics**

[blastp]

**BLAST Statistics**

[blastp HSP]

**BLAST Statistics**

[Footer in the old view in plain text format (from Formatting option)]
**BLAST Statistics**

[Footer in the old view in plain text format (from Formatting option)]

\( m' = m - l = 570 - 161 = 409 \)  
(Effective length of query)

\( n' = n - \text{(no. of sequences in the DB)} \times l = 53,472,480,294 - 145,899,600 \times 161 \)
= 29,982,644,694  
(Effective length of database)

\( m' \times n' = 409 \times 29,982,644,694 \)
= 12,262,901,679,846  
(Effective search space)

**BLAST search set vs. format option**

[Before search] Restrict a search against the selected organism or limit by Entrez Query

Search space will be limited
\( E \)-values become smaller

[After search] Restrict the result shown for a selected organism or limit by Entrez Query

Search space is not affected

**FASTA**

http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml  
(includes also SSEARCH)

http://www.ebi.ac.uk/Tools/sss/fasta/  
(includes also SSEARCH)

With graphic output
Results can be obtained through email

http://fasta.genome.jp/