TODAY’S TOPICS

➤ Statistical Significance of Alignment Scores
➤ Similarity Search
  • FASTA and BLAST
➤ Assignment 6

Significance of Alignment Scores

➤ Hypothesis testing for sequence alignment
  ➤ Two hypotheses
    • Null-hypothesis
      \( H_0 \): Two sequences are not related (random)
    • Alternative hypothesis
      \( H_1 \): Two sequences are related
  ➤ Test statistic: alignment score \( S \)
  ➤ Significance level is chosen a priori (e.g., 0.05)
  ➤ \( P \)-value:
    \[ P = \frac{1}{N} \exp \left( -\frac{1}{\lambda} - \frac{1}{\mu} \right) \]
    
    \( \mu = \frac{\ln \left( \frac{Km}{n} \right)}{l} \)
    \( \lambda = \frac{1}{b} \)

  • Probability of getting the alignment score \( S \), even if the two sequences are not related but randomly matched
  • If \( P < \text{Significance level} \), reject \( H_0 \)
    (The score should not be obtained just by aligning unrelated sequences)

Significance of Alignment Scores

➤ Hypothesis testing for sequence alignment
  ➤ Two hypotheses
    • Null-hypothesis
      \( H_0 \): Two sequences are not related (random)
    • Alternative hypothesis
      \( H_1 \): Two sequences are related
  ➤ Test statistic: alignment score \( S \)
  ➤ Significance level is chosen a priori (e.g., 0.05)
  ➤ \( P \)-value:
    \[ P = \frac{1}{N} \exp \left( -\frac{1}{\lambda} - \frac{1}{\mu} \right) \]
    
    \( \mu = \frac{\ln \left( \frac{Km}{n} \right)}{l} \)
    \( \lambda = \frac{1}{b} \)

  • Probability of getting the alignment score \( S \), even if the two sequences are not related but randomly matched
  • If \( P < \text{Significance level} \), reject \( H_0 \)
    (The score should not be obtained just by aligning unrelated sequences)
**Significance of Alignment Scores**

How to calculate $K$ and $\lambda$ (in LALIGN and PRSS)

1. The second sequence is shuffled many times.
2. Smith-Waterman local alignment score is calculated from each alignment: $P(S \geq x | H_0)$
3. The distribution is fitted to an extreme value distribution to obtain estimates of $K$ and $\lambda$
4. $P$-value is estimated based on the $K$ and $\lambda$, and the original alignment score $x$: $P(S \geq x) \approx Kmn e^{-lx}$

**Simulation of Alignment Scores**

**RECA_ECOLI** (P0A7G6; 353 amino acids) vs. **RAD51_YEAST** (P25454; 400 amino acids)

Smith-Waterman local alignment score = 293 (BLOSUM50, gap opening: -10, gap extension: -1)

**FASTA Web server by William Pearson**

The origin of FASTA format

https://fasta.bioch.virginia.edu/
The second sequence is shuffled 1000 times to generate random sequences.

- Amino acid composition can be maintained.
- Window size can be set so that shuffling is done locally.

\[ P(S \geq x) \approx Knme^{-lx} \]

\[ E(10000) = 1.5 \times 10^{-11} \times 10000 = 1.5 \times 10^{-7} \]

From PRSS:

\[ P = E(1) = 1.5 \times 10^{-11} \]

Significance of Alignment Scores

- For BLOSUM:
  - \( a \): gap opening penalty
  - \( b \): gap extension penalty (affine gap penalty)
- Based on 10,000 random sequence pairs,
  - calculated optimal local alignment scores, \( S \),
  - \( K \) and \( l \) are estimated by fitting the distribution of \( S \) with \( P(S \geq x) = Knme^{-lx} \)

This is what BLAST does!
Significance of Alignment Scores
• Alignment scores cannot be compared directly because they depend on: scoring matrix, gap penalty, algorithms used
• Statistical significance of alignments can be tested
  ➢ Is the alignment possible just by chance?
  ➢ If the alignment score is statistically significant (not possible simply by chance), the alignment is meaningful.
  ➢ P-value or E-value ($E = NP$) can be compared regardless of the scoring systems used for alignments.

NOTE: E-values change depending on the number of data used (if $N$ is small, E-value becomes small)

0 < P-value < 1 vs. $0 < E$-value < $N$

Significance of Alignment Scores

Similarity Search
➢ Why do you want to perform similarity search?
• To find related genes in another organisms ➢ Homologue candidates
• To identify a possible function of a gene/protein
[From genomic sequence]
• To predict gene structure: against cDNA sequences ➢ exon-intron structure
• To predict gene locations

Query sequence (protein or nucleotide)
MVLSPA...

Pairwise alignment (local)

Search result
high similarity
Sequence28
Sequence5
Sequence11
Sequence1
Sequence73
Sequence65
Sequence33
...
low similarity

Smith-Waterman Local Alignment
(used in SSEARCH)

Searches the entire alignment space (X multiple comparisons)
For a large database, it requires lots of time

Heuristic algorithm (FASTA/BLAST)

Minimizes the search space
Faster search speed
But search is not thorough

Similarity Search = Multiple Pairwise Alignments

Pairwise alignment search space

X multiple comparisons
**Speed and Sensitivity**

- Faster
- Slower
- Less sensitive
- More sensitive

**FASTA/BLAST Algorithm**

To reduce the search space:

```
TTAGACGCGTA
A
C
A
G
A
G
C
T
A
```

Break the query into **words**

**e.g., word length (k-tuple) = 2**

Larger the word size, more rapid, but less thorough (less sensitive)

```
TTAGACGCGTA
A
C
A
G
A
G
C
T
A
```

Word length (k-tuple) = 2
**FASTA/BLAST Algorithm**

- **Word length (k-tuple) = 2**
- **Word length (k-tuple) = 3**

**FASTA Algorithm**

1. Find identities using k-tuples
2. Join diagonals without gaps
3. Choose top 10 diagonals using a scoring matrix (e.g., BLOSUM62)

**FASTA Algorithm**

- **Ranking**
  - Database sequences are ranked based on z-values or OPT scores
  - \( z \)-value = the number of standard deviations from the mean (standardized score)
  - The high scored sequence pairs are aligned using the full Smith-Waterman dynamic programming algorithm
- **FASTA/SSSEARCH website**
  - [https://fastx.bioch.virginia.edu/fasta www2/fasta_list2.shtml](https://fastx.bioch.virginia.edu/fasta www2/fasta_list2.shtml)
  - [https://www.ebi.ac.uk/Tools/sss/fasta/](https://www.ebi.ac.uk/Tools/sss/fasta/)
- **[FASTA guide](https://fastx.bioch.virginia.edu/wrpearson/fastx_guide.pdf)**
  - [William R. Pearson's website](https://fastx.bioch.virginia.edu/wrpearson/)

**FASTA/BLAST Algorithm**

- **Word length (k-tuple) = 2**
- **Word length (k-tuple) = 3**
BLAST Similarity Search

BLAST resources

- BLAST
  [Guide to BLAST home and search pages]

- BLAST Report Description

- BLAST Statistics

- BLAST Command Line User Manual
  https://www.ncbi.nlm.nih.gov/books/NBK279690/

- BLAST YouTube Tutorials
  (Link is available from NCBI Help page or from Canvas)

BLAST Algorithm: seeding

Using words reduce the search space

Neighborhood increases the sensitivity

Word hits

Neighborhood Threshold (T) = 1

[minimum score allowed to be the neighborhood]
BLAST Algorithm: seeding

BLAST Algorithm

Search Space is Seeded

Extension of each alignment

BLAST Algorithm: extension

Extension of diagonal lines

BLAST Algorithm: extension

Extension is done until the score drops significantly

BLAST Algorithm: extension

>[Extension is allowed until its score drops $\geq X$ from the last maximum]

BLAST Algorithm: extension

>[Extension is allowed until its score drops $\geq X$ from the last maximum]
**BLAST Algorithm: extension**

- Extension is allowed until its score drops \( \geq X \) from the last maximum.

- Extension is trimmed back to the nearest maximum point.

**BLAST Algorithm: evaluation**

- Alignment threshold (\( S_1 \)) is used to choose only HSP (High-Scoring Segment Pairs).

- Consistent alignments are grouped for E-value calculation.

- Gapped extension is triggered after high score ungapped alignments are found.
**BLAST Algorithm: Gapped extension**

- Another threshold for gapped alignment (S2) is used to choose the final set of HSPs.

**Score ≥ S2**

W=3, Y=11, X=16, A=40, S1=41, S2=70

---

**BLAST Algorithm**

- Word-matching size (W)
  - longer words: faster but less sensitive
- Neighborhood threshold (T)
  - lower T: detects weaker similarities
  - slower but more sensitive
- Extension
  - Drop-off score (X)
  - Two-hit method
    - (A: distance b/w 2 hits)
- HSP selection
  - ungapped alignment threshold: S1
  - Gapped HSP extension
    - (gapped alignment threshold: S2)

---

**BLAST Similarity Search**

- Basic Local Alignment Search Tool
- BLAST finds regions of similarity between biological sequences.
- The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

- Blat
- Nucleotide BLAST
- Protein BLAST