CS 466
Introduction to Bioinformatics

Instructor: Jian Peng
Teaching Assistant: Wesley Qian
Statistics: Question

• Given two random sequences of lengths m and n
• What is the probability that they will produce an MSP score of $\geq S$?
Statistics: Intuition

Frequency of aa occurring in nature

- Ala 0.1
- Val 0.3
- Trp 0.01
- ...

Random sequence 1

Random sequence 2

Real sequence 1

Real sequence 2

SCORE

SCORE
Approach

⇒ Evaluate the probability that a score between random or unrelated sequences will reach the score found between two real sequences of interest:

If that probability is very low, the alignment score between the real sequences is significant.

If \text{SCORE} > \text{SCORE} \Rightarrow \text{the alignment between the real sequences is significant}
How to compute the probability?
Simulation

1. Generate many random sequence pairs

- Frequency of aa occurring in nature
  - Ala 0.1
  - Val 0.3
  - Trp 0.01
  - ...

- Random sequence 1 → SCORE
- Random sequence 2

- Real sequence 1 → SCORE
  - Real sequence 2

2. Compute the distribution of the SCOREs
How to compute the p-value (probability)?
Statistical test

p-value = 0.45  p-value = 0.001
Is this efficient enough?
Another observation
Karlin and Altschul observed that in the framework of local alignments without gaps: the distribution of random sequence alignment scores follow an EVD.
Extreme value distribution

\[ Y = \lambda \exp[-\lambda(x - \mu) - e^{-\lambda(x - \mu)}] \]

\[ P(Z < x) = \exp[-e^{-\lambda(x - \mu)}] \]

\[ P(Z \geq x) = 1 - \exp[-e^{-\lambda(x - \mu)}] \]

**P-value** = the probability of obtaining a score equal or greater than \( x \) by chance
Compute a p-value

- The probability of observing a score $\geq 4$ is the area under the curve to the right of 4.
- For an *Unscaled EVD*:
  \[
P(S \geq x) = 1 - e^{(-e^{-x})}
  \]

\[
P(S \geq 4) = 1 - e^{(-e^{-4})}
  \]

\[
P(S \geq 4) = 0.018149
  \]
Parameters

\[ P(Z \geq x) = 1 - \exp[-e^{-\lambda(x-\mu)}] \] (1)

\( \mu, \lambda \) : parameters depend on the length and composition of the sequences and on the scoring system: \( \mu \) is the mode (highest point) of the distribution and \( \lambda \) is the decay parameter.

- They can be estimated by making many alignments of random or shuffled sequences.
Statistical test

p-value = 0.45  p-value = 0.001
Significance: P-value and E-value

In a database of size $N$: $P \times N = E$

- **P-value:**
  Probability that an alignment with this score occurs by chance in a database of size $N$.
  The closer the P-value is towards 0, the better the alignment.

- **E-value:**
  Number of matches with this score one can expect to find by chance in a database of size $N$.
  The closer the E-value is towards 0, the better the alignment.

  $\rightarrow$ Smaller E-value, more significant in statistics
  Bigger E-value, by chance

  $E[\# \text{ occurrences of a string of length } m \text{ in reference of length } L] \sim L/4^m$
Parameters

\[ P(Z \geq x) = 1 - \exp[-e^{-\lambda(x-\mu)}] \quad (1) \]

\( \mu, \lambda \): parameters depend on the length and composition of the sequences and on the scoring system: \( \mu \) is the mode (highest point) of the distribution and \( \lambda \) is the decay parameter. They can be estimated by making many alignments of random or shuffled sequences.

- For alignments without gaps they can be calculated from the scoring matrix and then:

\[ P(Z \geq x) = 1 - \exp[-Kmn e^{-\lambda x}] \quad (2) \]

\( K \): is a constant that depend on the scoring matrix values and the frequencies of the different residues in the sequences.

\( m, n \): sequence lengths
Approximation:
if $x$ is very small, then $1 - \exp(-x)$ can be approximated by $x$

Therefore,

$$P(Z \geq x) \sim e^{-\lambda(x-\mu)} = Kmn e^{-\lambda x}$$

So E-value = DatabaseLength * p-value

$$E\text{-value} = KNme^{-\lambda x}$$

where $N$ is the database size (not the aligned length $n$)
Recap: Statistics: intuition

Given a binary 0/1 sequence and a query string of k consecutive ones

- Probability in a sequence of length k: \(\frac{1}{2^k}\)
- Probability in a sequence of length k+1?
  - \(1 - (1 - \frac{1}{2^k})^2\)
- How about the probability in a sequence of length k+n?
  - \(1 - (1 - \frac{1}{2^k})^{n+1}\)
- The longer the sequence, the more likely you are going to get k ones by chance!
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.

Web BLAST

Nucleotide BLAST
nucleotide ➤ nucleotide

blastx
translated nucleotide ➤ protein

tblastn
protein ➤ translated nucleotide

Protein BLAST
protein ➤ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

Human  Mouse  Rat  Microbes

Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an instance at a cloud provider
### Algorithm parameters

#### General Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max target sequences</td>
<td>100</td>
<td>Select the maximum number of aligned sequences to display</td>
</tr>
<tr>
<td>Short queries</td>
<td>✔</td>
<td>Automatically adjust parameters for short input sequences</td>
</tr>
<tr>
<td>Expect threshold</td>
<td>10</td>
<td>Expected number of chance matches in a random model.</td>
</tr>
<tr>
<td>Word size</td>
<td>6</td>
<td>The length of the seed that initiates an alignment.</td>
</tr>
<tr>
<td>Max matches in a query range</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

#### Scoring Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matrix</td>
<td>BLOSUM62</td>
<td></td>
</tr>
<tr>
<td>Gap Costs</td>
<td>Existence: 11 Extension: 1</td>
<td>Conditional compositional score matrix adjustment</td>
</tr>
</tbody>
</table>

#### Filters and Masking

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter</td>
<td>✔</td>
<td>Low complexity regions</td>
</tr>
<tr>
<td>Mask</td>
<td>✔</td>
<td>Mask lookup table only</td>
</tr>
<tr>
<td></td>
<td>✔</td>
<td>Mask lower case letters</td>
</tr>
</tbody>
</table>

---

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

- Show results in a new window
Understanding the output: graphics

The display can help you see that some matches do not extend over the entire length of your sequence => useful tool to discover domains.
Understanding the output: hit list

<table>
<thead>
<tr>
<th>Sequence ac number and name</th>
<th>Description</th>
<th>Score (bits)</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>P09505 RRPO BYDVP</td>
<td>Putative RNA-directed RNA polymerase (EC 2....</td>
<td>1652</td>
<td>0.0</td>
</tr>
<tr>
<td>P29045 RRPO BYDVR</td>
<td>Putative RNA-directed RNA polymerase (EC 2....</td>
<td>1635</td>
<td>0.0</td>
</tr>
<tr>
<td>P29044 RRPO BYDVI</td>
<td>Putative RNA-directed RNA polymerase (EC 2....</td>
<td>1625</td>
<td>0.0</td>
</tr>
<tr>
<td>P22956 RRPO RCNMV</td>
<td>Putative RNA-directed RNA polymerase (EC 2....</td>
<td>367</td>
<td>e-101</td>
</tr>
<tr>
<td>P17460 RRPO TCV</td>
<td>Probable RNA-directed RNA polymerase (EC 2.7....</td>
<td>286</td>
<td>1e-76</td>
</tr>
<tr>
<td>P22958 RRPO TTNVA</td>
<td>RNA-directed RNA polymerase (EC 2.7.7.48)</td>
<td>280</td>
<td>1e-74</td>
</tr>
</tbody>
</table>

- **Sequence ac number and name**: Hyperlink to the database entry: useful annotations
- **Description**: better to check the full annotation

- **Bit score (normalized score)**: A measure of the similarity between the two sequences: the higher the better (matches below 50 bits are very unreliable)

- **E-value**: The lower the E-value, the better. Sequences identical to the query have an E-value of 0. Matches above 0.001 are often close to the twilight zone. As a rule-of-thumb an E-value above 10^-4 (0.0001) is not necessarily interesting. If you want to be certain of the homology, your E-value must be lower than 10^-4
A good alignment should not contain too many gaps and should have a few patches of high similarity, rather than isolated identical residues spread here and there.
Very similar sequences

Query:  HBA_HUMAN Hemoglobin alpha subunit
Sbjct:  HBB_HUMAN Hemoglobin beta subunit

Score = 114 bits (285),  Expect = 1e-26
Identities = 61/145 (42%),  Positives = 86/145 (59%),  Gaps = 8/145 (5%)

Query 2  LSPADKTNVKAAGKAVGKYAGEYGAELERMFLSFPTTKFYPKF---DLSHGSAQV 55
          +K+ V A WGV + E G E A L R+ + P T+ + F  F  D  G+ + V
Sbjct 3  LTPKKEKSLAVTGKVL--NVDEVGGEALGRLLVLVVYPWTQRFSEFGDLSTPDAVMGNPKV 60

Query 56  KGHGKKVADALTNAHAVDDMPNALSALSLSLHAKLRLVDPVNFKLSSLHCLLTVTAAHLP 115
          K HGKKV A ++ +AH+D++ + LS+LH KL VDP NF+LL + L+ LA H
Sbjct 61  KAHGKKVLGAFSDGLAHLDNLKGTATLSELHCDEKHLVDPENFRLLGNVLCVLASHHGFK 120

Query 116  EFTPBAVHSLDKFLASVSTVLTSKY 140
          EFTP V A+ K +A V+ L K Y
Sbjct 121  EFTPPVQAAYQKVAGVANALAHKY 145
Similar sequences

Query: HBA_HUMAN Hemoglobin alpha subunit
Sbjct: MYG_HUMAN Myoglobin

Score = 51.2 bits (121), Expect = 1e-07,
Identities = 38/146 (26%), Positives = 58/146 (39%), Gaps = 6/146 (4%)

Query    2  LSPADKTNVKAAWGVGAHAGYGAEEALERLMFLSFTTTKTREFPHF------DLSSHGSAQV  55
       LS  +  V  WGKV  A  +G  E  L  R+F  P  T  F  F  D  S +
Sbjct    3  LSDGEWQLVLNVTGKEADIPGHGQEVLIRLFKGFHTLEKFDFKIKLHKLSEDKEKASDL  62

Query    56  KGHGKKVADALTNAHAVDHDDTPNALSALSDLHAKLRVDPVNFKLLSHCLLVTLLAHLPA  115
       K  HG  V  AL  +  +  L+  HA  K  +  +  +S  C++  L  +  P
Sbjct    63  KKHGATVLTLGILKKGHHGAEIKPLAQSHATKHKIPVKYLEFISECIIQVQSHKHP  122

Query    116  EFTPAVHASLDKFLASVSTVLTSKRY  141
             +F  +++++K  L  +  S  Y +
Sbjct     123  DFGADAQGAMNKALELFKDMASNYK  148
Not so similar sequences

Query: HBA_HUMAN Hemoglobin alpha subunit
Sbjct: SPAC869.02c [Schizosaccharomyces pombe]

Score = 33.1 bits (74), Expect = 0.24
Identities = 27/95 (28%), Positives = 50/95 (52%), Gaps = 10/95 (10%)

Query  30  ERMFLSFPTTKTYFPHFDLSHGSQAQVKGKHGKVADALTNAVAHVDDMPNALSALSDLH
++M ++P  P+F+  +H  +  +  +A  A  A  N  ++DD+  +LSA  D
Sbjct  59  QKMLGNYPEV---LYFNKAHQLSL---SQPRILAFALLNYAKNIDDL-TSLSAFMDQIVV 112

Query  90  K---LRVDPVNFKLSSHCLLVTLLAHLPQAEF-TPA 120
  K  L++  ++  ++  HCL  T+  LP++  TPA
Sbjct  113  KHVGLQIKAEHYPVGHCLLSTMQELLPSDVATPA 147