Assignment 1 review

Amino Acid Substitution Matrix
- Information Theory

Statistical Significance of Alignment Scores
- Assignment 5

TODAY’S TOPICS

Log Odds Score and Target Frequencies

\[ S(i,j) = \frac{1}{\lambda} \log(q_{ij}/p_i p_j) \]
\[ \lambda S(i,j) = \log(q_{ij}/p_i p_j) \]
\[ q_{ij} = p_i p_j e^{\lambda S(i,j)} \]

Target frequency  Expected frequency

\[ \sum_j q_{ij} = \sum_j p_i p_j e^{\lambda S(i,j)} = 1 \]
\[ \lambda \text{ can be estimated (matrix specific)} \]

Relative Entropy (H)

Expected Score (E)

\[ E = \sum_j p_j S(i,j) [p_i, p_j: \text{expected freq. of AA}_i, AA_j] \]

Relative Entropy (H)

\[ H = -\sum_j q_{ij} \lambda S(i,j) [q_{ij} \text{ is observed freq. of AA}_i, AA_j] \]

- the average information per residue pair
- summarizes the behavior of the scoring matrix
- the ability of the matrix to discriminate related from unrelated (nonrandom matching from random matching)
- \( H = 0 \) when target distribution equals to background distribution
- \( H \) increases when the two distributions become more distinguishable
- can be used to compare scoring matrices

Information Theory

Information: a decrease in uncertainty (unpredictability, a degree of surprise)

- If you are asking questions to somebody...
  - if you can guess every answer correctly
  - there is no surprise
  - you cannot gain any new information
  - but if you have no idea what answer you get
  - every answer is a surprise
  - you gain a lot of information

Information Theory Primer by Tom Schneider (also on Canvas):
http://schneider.ncifcrf.gov/paper/primer/

Introduction to Information Theory

Introduction to Information Theory

Information: a decrease in uncertainty (unpredictability, a degree of surprise)


Prob(A) = 1

Only one possible symbol

No surprise

No information
Introduction to Information Theory

- **Information:** a decrease in uncertainty (unpredictability, a degree of surprise)

<table>
<thead>
<tr>
<th>Device</th>
<th>A, B, A, B, A, B, A, B, ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob(A)</td>
<td>0.5</td>
</tr>
<tr>
<td>Prob(B)</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Two possible symbols
A little surprise
A small amount of information

- **Information:** a decrease in uncertainty (unpredictability, a degree of surprise)

<table>
<thead>
<tr>
<th>Device</th>
<th>A, B, C, A, B, C, A, B, ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob(A)</td>
<td>0.33</td>
</tr>
<tr>
<td>Prob(B)</td>
<td>0.33</td>
</tr>
<tr>
<td>Prob(C)</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Three possible symbols
More surprise, More information

- **Device**

<table>
<thead>
<tr>
<th>A, A, A, A, A, ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob(A) = 7/8</td>
</tr>
<tr>
<td>Prob(B) = 1/8</td>
</tr>
</tbody>
</table>

Two possible symbols
Big surprise! A lot more information...

But not much surprise in getting the symbol A’s

- **Bits:** the unit for values converted to base 2 logarithms (nats: the unit if base e is used)

  \[ H(p) = -\log_2 p \]

  - If an answer is highly unexpected (e.g., \( p = 0.1 \)),
    \[ H(0.1) = -\log_2 0.1 = 3.3 \text{ bits} \] (more information)
  - For a very much expected answer (e.g., \( p = 0.9 \)),
    \[ H(0.9) = -\log_2 0.9 = 0.15 \text{ bits} \] (less information)
  - If there is only one possible answer (symbol),
    \[ H(1.0) = -\log_2 1.0 = 0 \text{ bits} \] (no information)

- **Information is a decrease in uncertainty**

  - Surprising answers convey more information!
  - If each symbol is equally likely,
    \[ H(p) = \frac{-\log_2 p}{n} \]
  - The amount of information, or surprise of an answer, is inversely proportional to its probability.

- **Bits:** the unit for values converted to base 2 logarithms (nats: the unit if base e is used)

  \[ H(p) = -\log_2 p \]

  - If \( p_i \) is the same for all \( n \) symbols (equal probability)
    \[ H = -\sum p_i \log_2 p_i \]
  - For a very much unexpected answer (e.g., \( p = 0.1 \)),
    \[ H(0.1) = -\log_2 0.1 = 3.3 \text{ bits} \] (more information)
  - For a very much expected answer (e.g., \( p = 0.9 \)),
    \[ H(0.9) = -\log_2 0.9 = 0.15 \text{ bits} \] (less information)
  - If there is only one possible answer (symbol),
    \[ H(1.0) = -\log_2 1.0 = 0 \text{ bits} \] (no information)
Introduction to Information Theory

- Information can be represented by a series of symbols each with a certain probability:
  - Shannon Entropy: the average information per symbol
    \[ H = - \sum p_i \log_2(p_i) \]
  - For a random DNA sequence: ATGC (p = 0.25 for all)
    \[ H = -(0.25 \times 4) \times \log_2(0.25) = 2 \text{ bits} \]
  - For a AT-rich DNA sequence: pA = pT = 0.45 and pC = pG = 0.05
    \[ H = [-0.45 \times \log(0.45)] + [-0.45 \times \log(0.45)] + [-0.05 \times \log(0.05)] + [-0.05 \times \log(0.05)] = -0.45 \times (-1.15) + 2 \times (-0.05 \times -4.32) \times 2 = 1.47 \text{ bits} \]

Relative Entropy (H)

- Expected Score (E)
  \[ E = \sum p_i S(i,j) \quad \text{(p: expected freq. of AA, AA)} \]
- Relative Entropy (H)
  \[ H = - \sum q_{ij} S(i,j) \quad \text{([q is observed freq. of AA, AA pair])} \]

Since \( S(i,j) = 1/\lambda \log_2(q_{ij}/p_{ij}) \) or \( 1/\lambda \log_2(q_{ij}/p_{ij}) \)

\[ H = - \sum q_{ij} \log_2(q_{ij}/p_{ij}) - \sum q_{ij} \log_2(q_{ij}/p_{ij}) \]

\[ = -\sum q_{ij} \log_2(q_{ij}) - q_{ij} \log_2(p_{ij}) \]

Connection to \( H = \sum H_i \log_2(p_i) \)

Note: Both Expected Score and Relative Entropy have their units in bit or nat.

Comparing Scoring Matrices

- Relative Entropy (H) of a scoring matrix
  \[ H = -\sum q_{ij} S(i,j) \quad \text{([q is observed freq. of AA, AA pair])} \]
  - the average information per residue pair for a scoring matrix
  - summarizes the behavior of the scoring matrix
  - the ability of the matrix to discriminate related from unrelated (nonrandom matching from random matching)
    - \( H = 0 \) when target distribution equals to background distribution
    - \( H > 0 \) when the two distributions become more distinguishable
  - can be used to compare scoring matrices

BLOSUM and PAM matrices

(default in BLAST)

- BLOSUM80
- BLOSUM62
- BLOSUM45

Less divergent

Less divergent

More divergent

<table>
<thead>
<tr>
<th>PAM120</th>
<th>PAM160</th>
<th>PAM250</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.98 bits</td>
<td>0.7 bits</td>
<td>0.36 bits</td>
</tr>
</tbody>
</table>

From Henikoff and Henikoff (1992)

More amino acid pairs are used

Higher BLOSUM is generated including sequences that are more similar to one another.
Pairwise alignment summary

- Alignment score depends on:
  - Scoring matrix (match, mismatch, Ts/Tv, BLOSUM, PAM, etc.)
  - Gap penalty
  - Alignment method (e.g., global or local)
- Alignment scores cannot be compared directly
  - If the scoring systems used are different
  - If sequences compared are different (e.g., longer alignments tend to have higher scores)
- Alignment scores are used:
  - for searching optimal alignments from the alignment matrix
  - for a given pair of sequences based on a given scoring system

Pairwise alignment summary

- Optimal alignments and biologically meaningful alignments may not be the same
- Depending on the scoring system, unreasonable alignments can become optimal
- We need to choose better (biologically reasonable) scoring system: level of divergence (scoring matrices), gap penalty (affine, etc.), algorithm (local or global)
- Manual adjustment may be necessary
- Test statistical significance of the alignment (is the alignment possible just by chance?)

Significance of Alignment Scores

- Hypothesis testing (General)
  - Two hypotheses
    - Null-hypothesis
      - $H_0$: The previous (original) belief is true
    - Alternative hypothesis
      - $H_1$: The previous (original) belief is false; the new theory is true
  - Test statistic: alignment score ($S$)
  - Significance level is chosen a priori (e.g., 0.05)
  - P-value: $P(S|H_0)$ is true; Probability of getting $S$ if $H_0$ is true
  - If $P <$ Significance level, reject $H_0$

Significance of Alignment Scores

- P-value: $P(S|H_0)$ is true
  - Need to be calculated from the test statistic $S$
  - Need to know the probability distribution of the test statistic $S$ under $H_0$

Distribution of alignment scores follow the Extreme Value Distribution

The probability distribution of highest values in an experiment (e.g., optimal alignment scores)
Significance of Alignment Scores

- Score
- Probability

$P(S \geq x | H_0) >> 0.05$

$x_1$

$x_2$

$x_3$

Not significant

Significant!

Highly Significant!

Significance of Alignment Scores

$P(S \geq x | H_0)$: Probability of getting the alignment score $S \geq x$

Karlin-Altschul equation

$P(S \geq x) = 1 - \exp[-K \cdot m \cdot n] \approx Kmne^{-l}$

$K$ and $\lambda$: calculated from the empirical distribution of $S$ based on a given scoring matrix and amino acid composition

$\alpha$ and $\beta$: lengths of sequences aligned

(Karlin and Altschul 1990)

- Solved for ungapped local alignments
- Can be applied for gapped local alignments

$E$-value $= P(S \geq x | H_0) \times N$

where $N$ is the number of sequences

- Expected number of sequences in the data to have a score $\geq S$

$E$-value $\neq P$-value

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

- Estimated from an empirical probability distribution.

1) The second sequence is shuffled many times.
   (simulates random sequences)

2) Smith-Waterman 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 $x$: $P(S \geq x) \approx Kmne^{-l}$

LALIGN and PRSS

[Servers including both LALIGN and PRSS]

- http://www.ch.embnet.org/software/LALIGN_form.html
- http://www.ch.embnet.org/software/PRSS_form.html

- Read the description for PRSS!

[LALIGN only]

- http://www.ebi.ac.uk/Tools/psa/lalign/

[LALIGN server below does not provide E-values]

- http://xylian.igh.cnrs.fr/bin/lalign-guess.cgi