Today's Topics

Amino Acid Substitution Matrix

• Information Theory

Statistical Significance of Alignment Scores

Log Odds Matrix

PAM matrix

\[ S(i,j) = 10 \log_{10} \left( \frac{M(i,j)/f(i)}{f(j)} \right) \]

- \( M(i,j) \): Mutation probability from \( AA_i \) to \( AA_j \)
- \( f(i) \): Frequency of \( AA_i \) (number of \( AA_i \) / total number of residues)
- Probability to find \( AA_i \) by chance

BLOSUM matrix

\[ S(i,j) = 2 \log_2 \left( \frac{q_{ij}}{e_{ij}} \right) \]

- \( q_{ij} \): Observed frequency of \( AA_i, AA_j \) pairs
- \( e_{ij} \): Expected frequencies of \( AA_i, AA_j \) pairs

General form

\[ S(i,j) = \frac{1}{\ln} \log_2 \left( \frac{q_{ij}}{p_i p_j} \right) \]

[bit unit]

\[ S(i,j) = \frac{1}{\ln} \log_e \left( \frac{q_{ij}}{p_i p_j} \right) \]

[nat unit]

Relative Entropy (H)

Expected Score (E)

\[ E = \Sigma_i \Sigma_j p_i p_j S(i,j) \] [\( p_i, p_j \): expected freq. of \( AA_i, AA_j \)]

Relative Entropy (H)

\[ H = \Sigma_i \Sigma_j q_{ij} \lambda S(i,j) \] [\( q_{ij} \) is observed freq. of \( AA_i, AA_j \) pair]

- 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

Introduction to 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://users.fred.net/tds/lab/papers/primer/
**Introduction to Information Theory**

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

- **Device:** A, B, A, B, A, B, B, ...

  - Prob(A) = 0.5
  - Prob(B) = 0.5

  - Two possible symbols
  - A little surprise
  - A small amount of information

- **Device:** A, B, C, A, B, C, A, B, ...

  - Prob(A) = 0.33
  - Prob(B) = 0.33
  - Prob(C) = 0.33

  - Three possible symbols
  - More surprise, More information

- **Device:** A, A, A, A, A, B, A, A, ...

  - Prob(A) = 7/8
  - Prob(B) = 1/8

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

  - But not much surprise in getting the symbol A’s

**Introduction to Information Theory**

- Information is a decrease in uncertainty

  - Surprising answers convey more information!

  - If each symbol is equally likely,
    - The amount of information increases with the number of different symbols.

  - The amount of information, or surprise of an answer, is inversely proportional to its probability.

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

  - I: information, p: probability

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

  - If all n symbols are equally possible (p, is the same)
    \[ H = -\sum p \log_2(p) = -(np \times \log_2(p)) \]

    - = -log_2(n), since np = 1
    - = -log_2(1/n), since p = 1/n

    \[ H(1) = \log_2(1) = 0 \text{ bit, } H(2) = \log_2(2) = 1, \text{ and } H(4) = \log_2(4) = 2 \]
**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
    \[ H = -\sum p \log p \]
  - For a random DNA sequence: ATGC (\( p = 0.25 \) for all)
    \[ H = -(0.25 \times 4 \times \log_2(0.25) + \log_2(4)) = 2 \text{ bits} \]
  - For an AT-rich DNA sequence, \( p_A = 0.65 \) and \( p_T = p_C = p_G = 0.15 \)
    \[ H = -0.65 \times (\log_2(0.65)) - 0.15 \times (\log_2(0.15)) = -0.45 \times (\log_2(2)) = 0.9 \text{ bits} \]

**Comparing Scoring Matrices**

- **Relative Entropy (\( H \)) of a scoring matrix**
  \[ H = \sum_{ij} p_{ij} \log \frac{q_{ij}}{p_{ij}} \]
  - 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
    \[ H = \sum_{ij} \log \frac{q_{ij}}{p_{ij}} \] when the two distributions become more distinguishable
    \[ H \] can be used to compare scoring matrices

**Relative Entropy (\( H \))**

- **Expected Score (\( E \))**
  \[ E = \sum_{ij} p_{ij} S_{ij} \]
  - \( p_{ij} \): expected freq. of \( AA, \, AA \) pair

- **Relative Entropy (\( H \))**
  \[ H = \sum_{ij} q_{ij} \log \frac{q_{ij}}{p_{ij}} \]
  - \( q_{ij} \): observed freq. of \( AA, \, AA \) pair

**Comparing Scoring Matrices**

- **Relative Entropy (\( H \))**
  \[ H = \sum_{ij} q_{ij} \log \frac{q_{ij}}{p_{ij}} \]
  - Higher BLOSUM is generated using more data (fewer information is eliminated), thus has a higher \( H \)

- **BLOSUM and PAM matrices** (default in BLAST)
  - BLOSUM80
  - BLOSUM62
  - BLOSUM45
  - PAM120
  - PAM160
  - PAM250

- **Expected Score**
  \( E = \sum_{ij} p_{ij} S_{ij} \)
  - **Relative Entropy**
  \( H = \sum_{ij} q_{ij} \log \frac{q_{ij}}{p_{ij}} \)

**Note:** Both Expected Score and Relative Entropy have their units in bit or nat.
Short alignments require shallow scoring matrices.

- Default matrices (e.g., BLOSUM62) are good for identifying <25% identity.
- Deeper scoring matrices (e.g., BLOSUM, PAM250) require long sequence alignment to achieve significant scores (e.g., >90 bit).
- They are more likely to extend alignments outside of homologous region.

Alignment scores are used:

- Alignment scores cannot be compared directly
- They are more likely to extend alignments outside of homologous region.

**Substitution matrices for specific proteins**

**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
- Alignment scores are used:
  - For searching optimal alignments from the alignment matrix
  - For a given pair of sequences based on a given scoring system
Alignment Strategy
- **Protein alignment is easier** than DNA alignment
  - DNA has only 4 nucleotide types (they can match just by chance more easily)
  - Protein sequences evolve more slowly than DNA sequences (genetic code is redundant; nonsynonymous substitutions are less frequent than synonymous substitutions)
- If DNA sequences are from coding regions:
  - Translate them, and **align at the protein level** first
  - This ensures gaps inserted between codons (prevents insertion of frame-shifted gaps)
- Do not blindly rely on the default parameter set
  - Try various scoring matrices, gap penalties, etc.

DNA alignment at protein level
1. Translate DNA sequences to amino acid sequences
2. Align them at the protein level
3. Reverse translate the protein alignment to DNA alignment
   - **TranslatorX** http://translatorx.co.uk/
   - **RevTrans 2.0** https://services.healthtech.dtu.dk/service.php?RevTrans-2.0
   - **PAL2NAL** http://www.bork.embl.de/pal2nal/
   - **tranalign** https://www.bioinformatics.nl/cgi-bin/emboss/tranalign
     (included in EMBoss servers; see the course/Link page.)

Significance of Alignment Scores
- **Hypothesis testing (General)**
  - Two hypotheses
    - Null-hypothesis
    - Alternative hypothesis
  - Test statistic
  - Significance level is chosen a priori (e.g., 0.05)
  - \( P\)-value: \( P(S|H_0) \) is true
  - If \( P < \text{Significance level} \), 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 \( H_0 \)

- Central Limit Theorem:
  - If the sample size is large enough, the sampling distribution of the mean of any independent, random variables will be normal or nearly normal.

  **Experiment**: 1000 coin tossing
  - Count the number of heads
  - Repeat 1000 experiments

  **(Expected to see 500 heads/experiment)
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 Extreme Value Distribution (Gumbel distribution)**
  - The probability distribution of highest values in an experiment (e.g., optimal alignment scores)

**Karlin–Altschul equation** (Karlin and Altschul 1990)

$$P(S \geq x) = 1 - \exp[-Kmne^{-x}] \approx Kmne^{-x}$$

- $\lambda$ and $\mu$ calculated from the empirical distribution of $S$ based on a given scoring matrix and amino acid composition
- $m$ and $n$: lengths of sequences aligned
- Solved for ungapped local alignments
- Can be applied for gapped local alignments

- **E-value** = $P(S \geq x)$ × $N$
  - where $N$ is the number of sequences in the dataset
  - Expected number of sequences in the dataset to have a score $\geq x$

- **E-value = P-value**

- **How to calculate $\lambda$ and $\mu$ (in LALIGN and PRSS)**
  - Estimated from an empirical probability distribution.
  1) The second sequence is shuffled many times. (Simulates random sequences)
  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 $\lambda$ and $\mu$
  4) P-value is estimated based on the $\lambda$ and $\mu$ and the original alignment score $x$: $P(S \geq x) \approx Kmne^{-x}$
Simulation of Alignment Scores

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

Smith-Waterman local alignment score = 127 (BLOSUM50, gap opening: -12, gap extension: -5)

127

Random alignment score distribution

From 1000 random alignments

FASTA Web server by William Pearson

FASTA Sequence Comparison at the U. of Virginia

- UVa FASTA Server
- University of Virginia
- National Center for Biotechnology Information (NCBI)
- EMBL-EBI
- GenBank
- Swiss-Prot
- Protein Data Bank (PDB)
- National Library of Medicine (NLM)
- National Institutes of Health (NIH)

Original FASTA package was released on 1988 (earlier than BLAST)

The origin of FASTA format