Sequence comparison:
Significance of similarity scores

http://faculty.washington.edu/jht/GS559_2013/

Genome 559: Introduction to Statistical and Computational Genomics
Prof. James H. Thomas
Review

• How to compute and use a score matrix.

• log-odds of sum-of-pair counts vs. expected counts in aligned blocks.

• Why gap scores should be affine.
Are these proteins related?

SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
identities-> LP LYNYC L
SEQ 2: QFFPLMPPAPYFILATDYENLPLVYSCTTFFWLF

SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
     LP WLDATYKNYA YC L
SEQ 2: QFFPLMPPAPYWILDATYKNYALVYSCTTFFWLF

SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
     RVV LPS WLDATYKNYA YCDVTYKL
SEQ 2: RVVPLMPSAPYWILDATYKNYALVYSCDVTYKLF

(intuitive answers)

NO (score = -1)

PROBABLY (score = 15)

YES (score = 24)
Significance of scores

Alignment algorithm and score matrix

Low score = unrelated
High score = related

How high is high enough?

HPDKKAHSIHAWILSKSKVLEGNTKEVVDNVLKT

HADKRAHSIHAWLLSKSKVLGNTKEVVQNVNLKS

45
The null hypothesis

• We first characterize the distribution of scores expected from sequences that are not related.
• This assumption is called the null hypothesis.
• The statistical test will be to determine whether the observed result provides a reason to reject the null hypothesis.
• Use BLAST to search a randomly generated database of sequences using a given query sequence (recall that BLAST searches use DP local alignment).

• What will be the form of the resulting distribution of pairwise alignment scores?
Empirical score distribution

• Distribution of scores from a **real** database search using BLAST.
• This distribution contains scores from a few related and lots of unrelated pairs.

(note - there are lots of lower scoring alignments not reported)
Empirical null score distribution

• This distribution is similar to the previous one, but generated using a randomized sequence database (each sequence shuffled).

(note - there are lots of lower scoring alignments not reported)

(notice the x scale is shorter here)
Computing an empirical p-value

- The probability of observing a score $\geq X$ is the area under the 'curve' to the right of $X$.
- This probability is called a p-value.
- $p$-value = $\Pr(\text{data}|\text{null})$
  
  (read as probability of data given a null hypothesis)

e.g. out of 1,685 scores, 28 received a score of 20 or better. Thus, the p-value associated with a score of 20 is approximately $28/1685 = 0.0166$. 

Problems with empirical distributions

- We are interested in very small probabilities.
- These are computed from the tail of the null distribution.
- Estimating a distribution with an accurate tail is feasible but computationally very expensive because we have to make a very large number of alignments.
A solution

- Solution: characterize the form of the score distribution mathematically.
- Fit the parameters of the distribution empirically (or compute them analytically if possible).
- Use the resulting distribution to compute accurate p-values.
- First solved by Karlin and Altschul.
Extreme value distribution (EVD) (aka Gumbel Distribution)

This distribution is roughly normal near the peak, but has a longer tail on the right.
Computing a p-value

- The probability of observing a score $\geq 4$ is the area under the curve to the right of 4.
- $p$-value = $\Pr(\text{data} | \text{null})$
Unscaled EVD equation

Compute this value for $x=4$.

$$P \quad S \geq x \quad = 1 - e^{(-e^{-x})}$$

S is data score, x is test score
Computing a p-value

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

\[ P(S \geq 4) = 0.018149 \]
Other comments on probability distributions (FYI)

• the **PDF** (probability density function) is the equation that generates the probability curve.

• the **CDF** (cumulative distribution function) is the equation that describes the total area under the probability curve up to some point (intuitively the "area so far").

• for alignment scores we are interested in the area above some point. But since the total area under the curve is exactly 1, this is just $1 - \text{CDF}$.

• for the unscaled extreme value distribution (Gumbel):

$$\text{CDF} = e^{(-e^{-x})} \quad \text{PDF} = e^{-x} e^{(-e^{-x})}$$

• and we want to compute $1 - \text{CDF}$:

$$P \ S \geq x = 1 - e^{(-e^{-x})}$$