# Significance of Alignments

COMP 571
Luay Nakhleh, Rice University

# Hypothesis Testing for Sequence Homology

* When a best local alignment is found, the next task is to assess its biological relevance
* This is most often done based on hypothesis testing

# Hypothesis Testing for Sequence Homology

1. A null hypothesis $H_0$, the validity of which we will test, is given
2. An alternative hypothesis, $H_1$, is also given
3. Perform a relevant experiment for testing $H_0$, and record the result
4. Find the probability, p, for the result, given that $H_0$ is valid.
5. If p is less than a given threshold (e.g., .05), reject $H_0$ and accept $H_1$

# Hypothesis Testing for Sequence Homology

1. $H_0$: the two sequences are not homologous
2. $H_1$: the two sequences are homologous
2. Determine the experiment: find the segment pair from the two sequences with the highest score
3. Determine the probability of the result, given $H_0$ (**details: next slide**)
4. Determine the rejection threshold for $H_0$ (e.g., $0.5 \times 10^{-5}$)
5. Perform the experiment chosen in (2): find the segment pair with the highest score and record the result
6. Determine the probability of achieving the result or higher, given $H_0$ (use the probability distribution found above), and compare with the rejection level for $H_0$

# Probability of the Result, Given $H_0$

* This is often done by finding the probability distribution for the highest-scoring segment pairs in randomly generated sequences (details: next slide)
* A large number of such sequences are generated, and compared with one of the two sequences being aligned, and the scores of these comparisons are the basis for the probability distribution of the scores

# Random Generation of Sequences

* A frequency distribution of the occurrences of the amino acids has to be used
* The amino acid of the random sequence is drawn using this distribution, often independent of the position and which amino acids are in the other positions

# Example

Assume an alphabet of four symbols {A, C, D, E} and the sequences

$$q = \mathtt{ACADAEA} \quad \text{and} \quad d = \mathtt{ECAEDACECE},$$

and we find the best local alignment to be

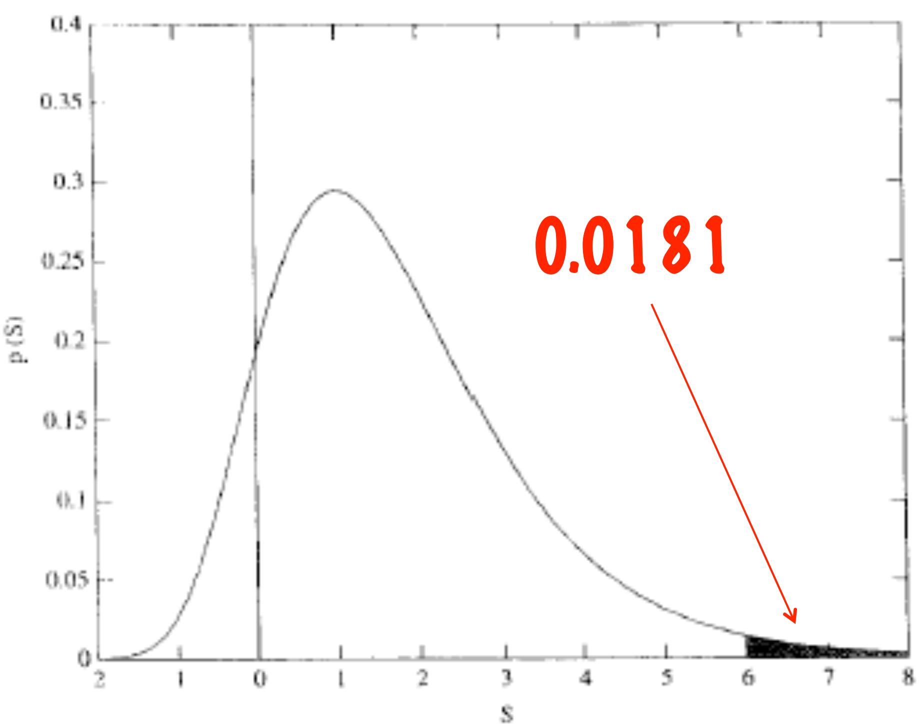
```
CA-DA
CAEDA
```

with score $S = 6$.

We then make random sequences, with the same amino acid distribution and length as $d$. The frequency distribution to use when generating random sequences is then $\{f_{\mathrm{A}} = 0.2,\ f_{\mathrm{C}} = 0.3,\ f_{\mathrm{D}} = 0.1,\ f_{\mathrm{E}} = 0.4\}$. E will be drawn with a probability twice the probability of drawing A.

# Example

Assume we make a lot (e.g. 10 000) of random sequences, and for each we make a local alignment to $q$ and note the score of the highest-scoring local alignment. We then find the distribution of these scores




The probability of finding a local alignment with score at least 6, when q is aligned with a random sequence with the same amino acid distribution and length as *d*

# Deriving the Amino Acid Frequency Distribution

* Universal
  * Over all known sequences
* Global
  * Super-family
* Local
  * From the sequences (q,d) themselves

# What Frequency Distribution to Use

* The distribution of the amino acids in d should be used to assess the score obtained by comparing q with d

# Using Z Scores to Estimate Statistical Significance

1. Compare the two sequences, and record $S'$, the score of the highest-scoring segment pair.

2. Make a distribution of the scores for random sequences, as explained above, or use a known distribution. Let $\mu$ be the mean of the random scores, and $\sigma$ the standard deviation. The standard deviation is a measure of variance, and for discrete values defined as

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \mu)^2},$$

   where $y_i$ are the scores obtained for randomized sequences.

3. Find the $Z$ value of the highest-scoring segment pair, with scoring $S'$. The $Z$ value is the number of standard deviations that the score $S'$ is above the mean value, calculated as

$$Z(S') = \frac{S' - \mu}{\sigma}.$$

   A threshold, $Z_t$, is determined, such that a score higher than $Z_t\sigma + \mu$ should indicate statistical significance, hence that the sequences are biologically related (homologous). Experiments have shown that $Z_t = 7$ is appropriate for protein sequence comparison and protein database search.

# Using Z Values to Estimate Statistical Significance: Example

Assume we have found a score $S' = 17.2$. We must then calculate the $Z$ value for $S'$. Let the probability distribution for the scores from random sequences have mean and standard deviation as $\mu = 4.2$, $\sigma = 3.4$. By use of the formula above, we find the $Z$ value for $S'$ to be 3.8. This is much less than 7, hence not convincing enough to indicate homology between the two sequences. $\triangle$

# Questions?