Sequence alignment

Statistical significance of alignment scores

Outline

- How do we determine whether the score of the best alignment is indicative of truly related sequences?
- Bayesian approach
- Classical approach

Bayesian approach

- Compute probability of Related model using Bayes rule
- Requires prior probability of R and U

$$\Pr(R \mid x, y)$$

$$= \frac{\Pr(x, y \mid R)\Pr(R)}{P(x, y)}$$

$$= \frac{\Pr(x, y \mid R)\Pr(R)}{\Pr(x, y \mid R)\Pr(R) + \Pr(x, y \mid U)\Pr(U)}$$

$$= \frac{\Pr(x, y \mid R)\Pr(R) / \Pr(x, y \mid U)\Pr(U)}{\Pr(x, y \mid R)\Pr(R) / \Pr(x, y \mid U)\Pr(U) + 1}$$

Needs to know Pr(R), Pr(U)

Classical approach

Determine how likely it is that such an alignment score would result from chance.

3 ways to calculate chance; look at alignment scores for

- real but non-homologous sequences
- real sequences shuffled to preserve compositional properties
- sequences generated randomly based upon a DNA/protein sequence model

Scores from Random Alignments

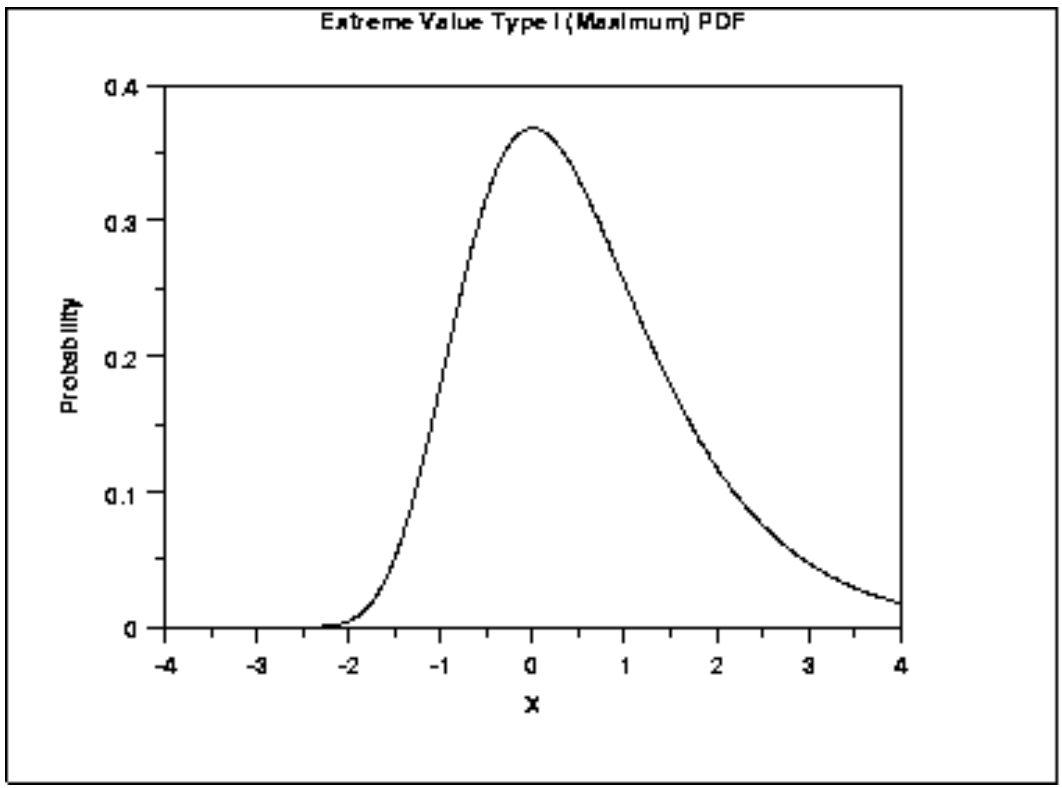
- suppose we assume
  - sequence lengths *m* and *n*
  - a particular substitution matrix and amino-acid frequencies

- and we consider generating random sequences of lengths *m* and *n* and finding the best alignment of these sequences

- this will give us a distribution over alignment scores for random pairs of sequences

Null Hypothesis

Statistics of Alignment Scores: The Extreme Value Distribution



- in particular, we get an *extreme value distribution*

Distribution of Scores

- the expected number of alignments, E, with score at least S is given by:

$$E(S) = Kmne^{-\lambda S}$$

- S is a given score threshold
- m and n are the lengths of the sequences under consideration
- K and λ are constants that can be calculated from
  - the substitution matrix
  - the frequencies of the individual amino acids

$$P(\text{score} > S) = 1 - e^{-E(S)}$$

Statistics of Alignment Scores

- to generalize this to searching a database, have *n* represent the summed length of the sequences in the DB (adjusting for edge effects)
- the NCBI BLAST server does just this
- theory for *gapped* alignments not as well developed
- computational experiments suggest this analysis holds for gapped alignments (but *K* and λ must be estimated from data)

Summary

- Statistical significance of optimal alignment scores
- Bayesian approach – requires priors
- Classical approach – uses extreme value distribution theory
  - Used in reporting BLAST search results