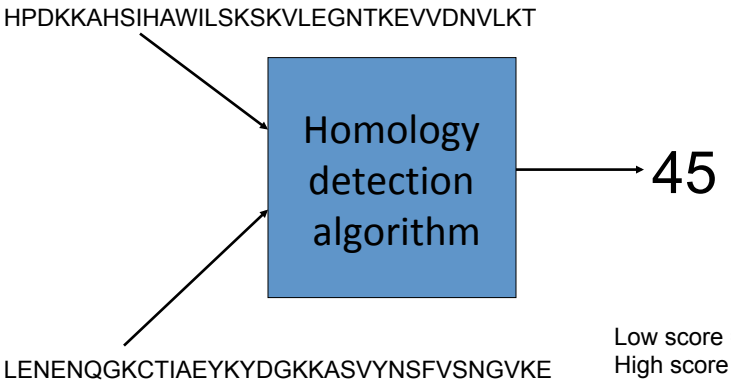
"Strength and growth come only through continuous effort and struggle."
-- Napoleon Hill

Significance of scores: Karlin Altshul Statistics



Low score = unrelated
High score = homologs

How high is high enough?

Significance of Scores ?

- Our scores are a log likelihood of H vs R
  - Score = log P[x,y | H] / P [x,y | R]
  - H == homology model
  - R == random sequence model

  Score > 1 => evidence for H

  Score < 1 => evidence for R

- Is a score of 2 convincing evidence of homology?
  - What about 5, 10, 15, or 20?
- We need some notion of "scale" for the score axis, some measure of confidence.

Lets first consider a Bayesian approach…

Are the two sequences, X and Y, homologous?

Want: P (H | x,y)

Are the two sequences, X and Y, homologous?
Want: P (H | x,y)

$$P(H \mid x,y) = P(x,y|H)\, P(H) / P(x,y)$$

Prior probability of homology model: P(H)

$$P(x,y) = P(x,y|H)P(H) + P(x,y|R)P(R)$$

Probability that random model is correct:

$$P(R) = 1 - P(H)$$

Similarity Score Significance

- Determining an appropriate prior log likelihood for the Bayesian analysis requires two pieces:
  - knowledge of homologies in database
  - model of non-homologies/random alignments
- Classical/frequentist approach:
  - Show that it is very unlikely to be random
  - Reject the null hypothesis...
    ...that random alignment is plausible

6

Alternative: Classical Statistics

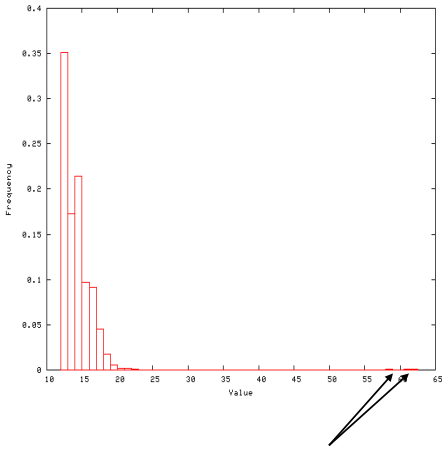
- We are interested in characterizing the distribution of scores from sequence comparison algorithms.
- We would like to measure how surprising a given score is, *assuming that the two sequences are not related.*
- The assumption is called the null hypothesis.
- The purpose of most statistical tests is to determine whether the observed results provide a reason to reject the hypothesis that they are merely a product of chance factors.

Sequence similarity score distribution

- Search a randomly generated database of DNA sequences using a randomly generated DNA query.
- What will be the form of the resulting distribution of pairwise sequence comparison scores?

Empirical score distribution

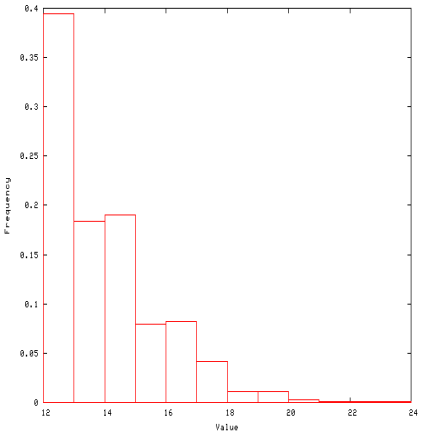
- The picture shows a distribution of scores from a real database search using BLAST.
- This distribution contains scores from non-homologous and homologous pairs.



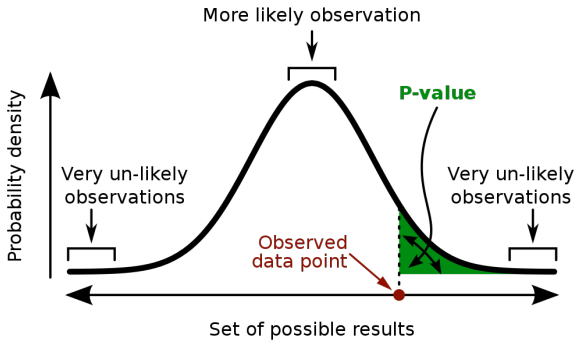
High scores likely from homology.

Empirical null score distribution

- This distribution is similar to the previous one, but generated **only** using a randomized sequence database.



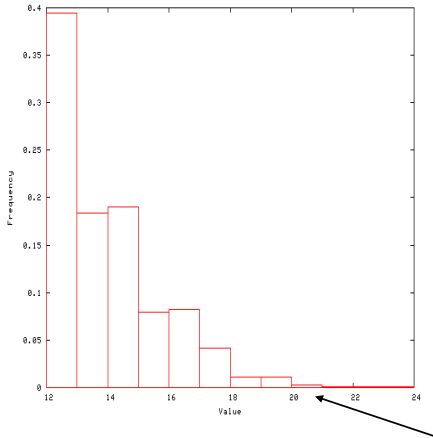
p-value or probability value is the probability for a given statistical model that, *when the null hypothesis is true*, the statistical summary would be the ***same*** as ***or of greater magnitude*** than the actual observed results.



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Note we're calculating P(X > x | R) here. This is NOT equivalent to P(H|x) and therefore using the p-value as a "score" is an egregious error!

Computing a p-value



- The probability of observing a score >X is the area under the curve to the right of X.
- This probability is called a p-value.

Out of 1685 scores, 28 receive a score of 20 or better. Thus, the empirical 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 distribution.
- Estimating a distribution with accurate tails is computationally *very expensive*.

A solution

- Solution: Characterize the form of the distribution mathematically.
- Fit the parameters of the distribution empirically, or compute them analytically.
- Use the resulting distribution to compute accurate p-values.

Karlin-Altschul theory

Assumes:

1. At least one alignment score is positive
2. Expected scores are negative
3. Characters of sequences are i.i.d.
4. No gaps

Then the expected number of alignments (e-value) with score at least S:

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

16

Follows Gumbel extreme value distribution

A link between scoring scheme and statistics.

$$E(s_{a,b}) = \Sigma_{a,b}\, p_a p_b s(x,y)$$

If the expected score is negative and s(x,y) contains at least one positive score, then the link between the scoring scheme and the log-odd ratio (i.e. Karlin-Altshul statistics) holds EVEN if the scoring scheme is chosen arbitrarily!

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

- Raw scores have little meaning without detailed knowledge of the scoring system used, or more simply its statistical parameters *K* and *lambda*.

Where λS_{ij} is a normalized score:

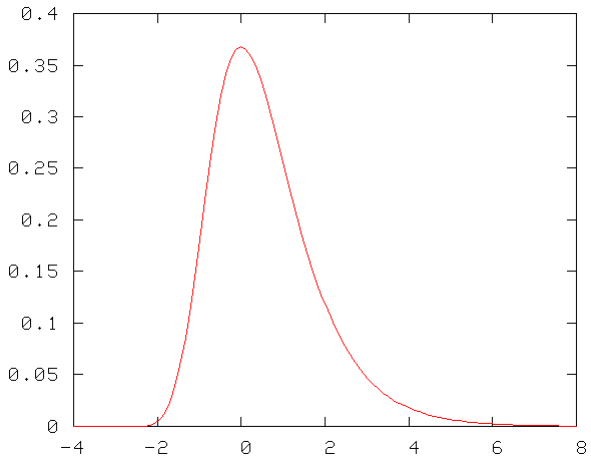
$$\lambda S_{ij} = \log (p_{ij} / p_i p_j)$$

K compensates for lack of independence of nearby local alignments, λ scales such that the expected score is 1.

Probability of match score greater than S:

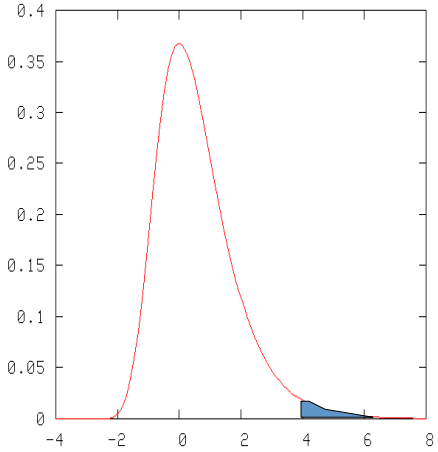
$$P(x > S) = 1 - e^{-E(S)}$$

Extreme value distribution



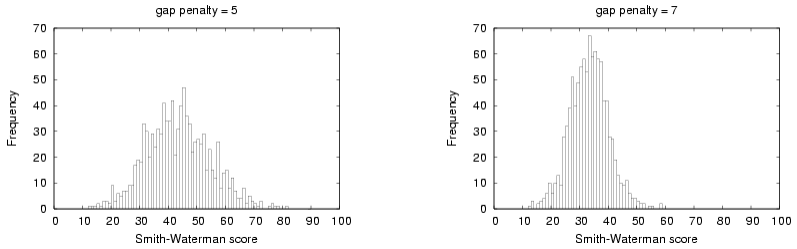
This distribution is characterized by a larger tail on the right.

Computing a p-value



- The probability of observing a score >4 is the area under the curve to the right of 4.
- This probability is the p-value!
- Exact same formulation as empirical!

Scaling the EVD



- An extreme value distribution derived from, e.g., the Smith-Waterman algorithm will have a characteristic mode μ (a scaled K!) and scale parameter λ.

$$P(S \geq x) = 1 - \exp\left[-e^{-\lambda(x-\mu)}\right]$$

- These parameters depend upon the size of the query, the size of the target database, the substitution matrix and the gap penalties.

An example

You run BLAST and get a score of 45. You then run BLAST on a shuffled version of the database, and fit an extreme value distribution to the resulting empirical distribution. The parameters of the EVD are μ = 25 and λ = 0.693. What is the p-value associated with 45?

$$P(S \geq x) = 1 - \exp\left[-e^{-\lambda(x-\mu)}\right]$$

An example

You run BLAST and get a score of 45. You then run BLAST on a shuffled version of the database, and fit an extreme value distribution to the resulting empirical distribution. The parameters of the EVD are μ = 25 and λ = 0.693. What is the p-value associated with 45?

$$\begin{aligned} P(S \geq 45) &= 1 - \exp\left[-e^{-0.693(45-25)}\right] \\ &= 1 - \exp\left[-e^{-13.86}\right] \\ &= 1 - \exp\left[-9.565 \times 10^{-7}\right] \\ &= 1 - 0.999999043 \\ &= 9.565 \times 10^{-7} \end{aligned}$$

Significance of Scores Summary

- Bayesian approach:
  - determine P [H | x,y]
  - need prior log likelihood for H vs R
- Frequentist approach:
  - determine P_R[max score > S]
  - need distribution for score function, e.g. EVD for P [max score > s]
- Significance of local ungapped alignment similarity score depends on:
  - Score matrix, length of query, size database

25

What p-value is significant?

$$Pr(\text{Reject } R \mid R) = Pr(p \leq \alpha \mid R) = \alpha$$

What p-value is significant?

- The most common thresholds are 0.01 and 0.05.
- A threshold of 0.05 means you are 95% sure that the result is significant.
- Is 95% enough? It depends upon the *cost* associated with making a mistake.
- Examples of costs:
  - Doing expensive wet lab validation.
  - Making clinical treatment decisions.
  - Misleading the scientific community.
- Most sequence analysis uses more stringent thresholds because the p-values are not very accurate.

Notes on relationship between substitution matrix and log-odds ratio:

$$s(x,y) = \log (q_{xy} / p_x p_y)$$

1. Our scoring scheme (substitution matrix) is additive, and hence S (the score for the entire alignment) is a measure of the relative likelihood of the whole alignment arising from homology compared to a random model.

2. However, a positive S (score for the total alignment) is not sufficient to test an alignment's significance. Just like a log-odds ratio > 1 wasn't sufficient for testing dinucleotide frequencies. We need a statistical test of significance!

3. We expect s(x,y) to contain positive and negative values, but this need NOT be the case!!

4. We can multiple all scores by a constant and still obtain the same relative ordering of global alignments. However, our constant multiplier *would* impact local alignments because they may not be of the same length!

Similarity score significance

- Karlin-Altschul doesn't extend to gapped scoring models...
  ...but empirical simulations suggest the same basic approach works.
- As with Bayesian approach, correct for "number of independent trials"
  – some fraction of nm (size of query and database).

29