Sequence comparison: Significance of similarity scores

http://faculty.washington.edu/jht/GS559_2012/

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.
- Why gap scores should be affine.

Are these proteins related?

(intuitive answers)

```
SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
           L P      L   Y N    Y C     L
SEQ 2: QFFPLMPPAPYFILATDYENLPLVYSCTTFFWLF
```

NO (score = -1)

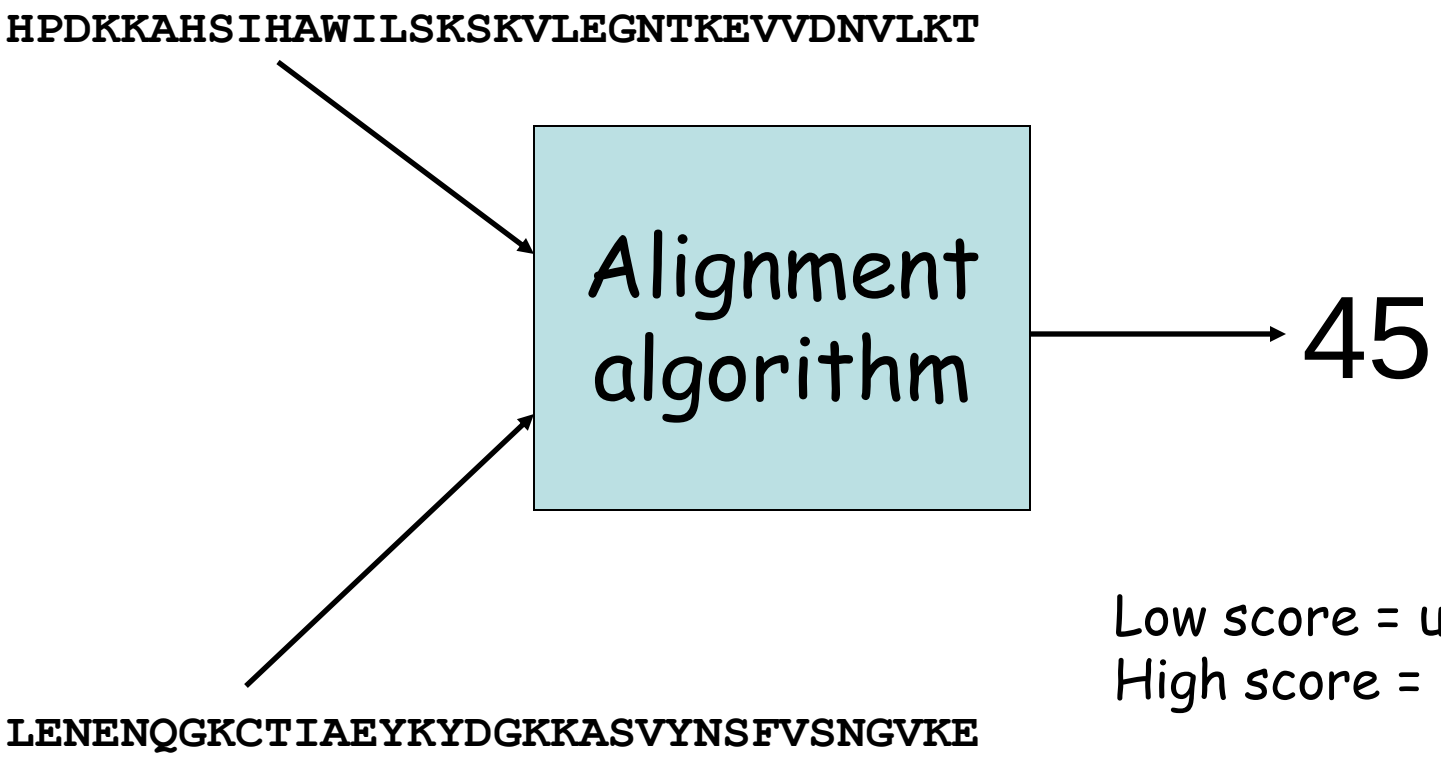
```
SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
           L P    W LDATYKNYA  Y C     L
SEQ 2: QFFPLMPPAPYWILDATYKNYALVYSCTTFFWLF
```

PROBABLY (score = 15)

```
SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
       RVV L PS   W LDATYKNYA  Y CDVTYKL
SEQ 2: RVVPLMPSAPYWILDATYKNYALVYSCDVTYKLF
```

YES (score = 24)

Significance of scores



Low score = unrelated
High score = related

How high is high enough?

The null hypothesis

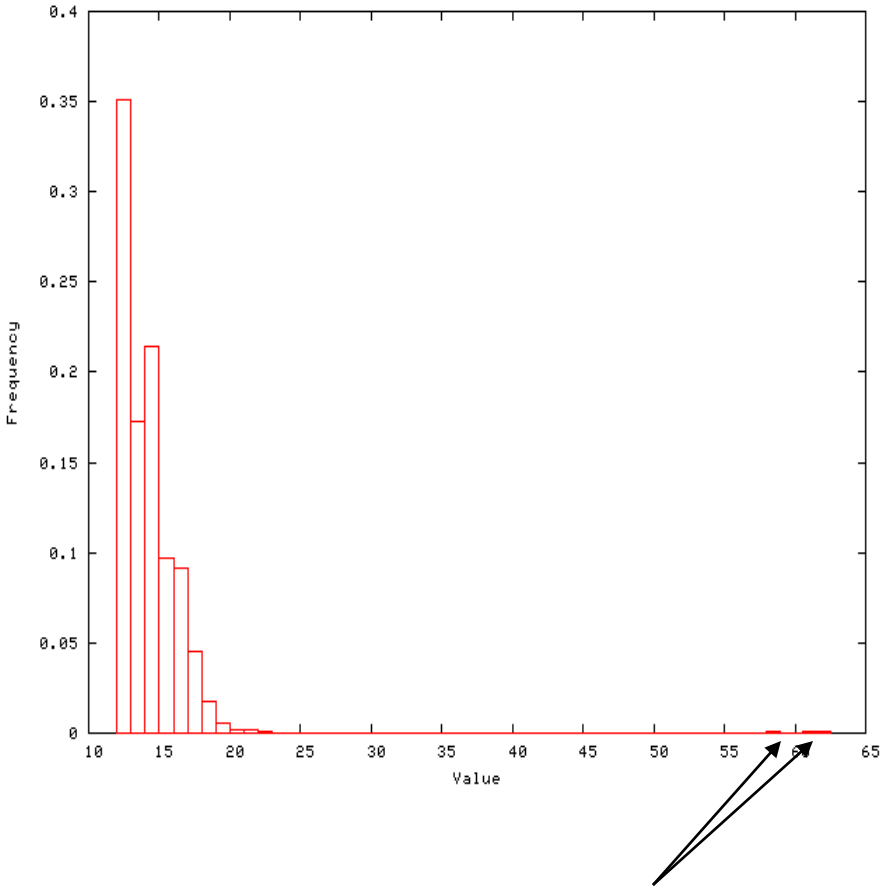
- We want to characterize the distribution of scores from pairwise sequence alignments.
- We measure how surprising a given score is, assuming that the two sequences are not related.
- This assumption is called the null hypothesis.
- The purpose of most statistical tests is to determine whether the observed result provides a reason to reject the null hypothesis.

Sequence similarity score distribution

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

Empirical score distribution

- This shows the distribution of scores from a real database search using BLAST.
- This distribution contains scores from a few related and lots of unrelated pairs.

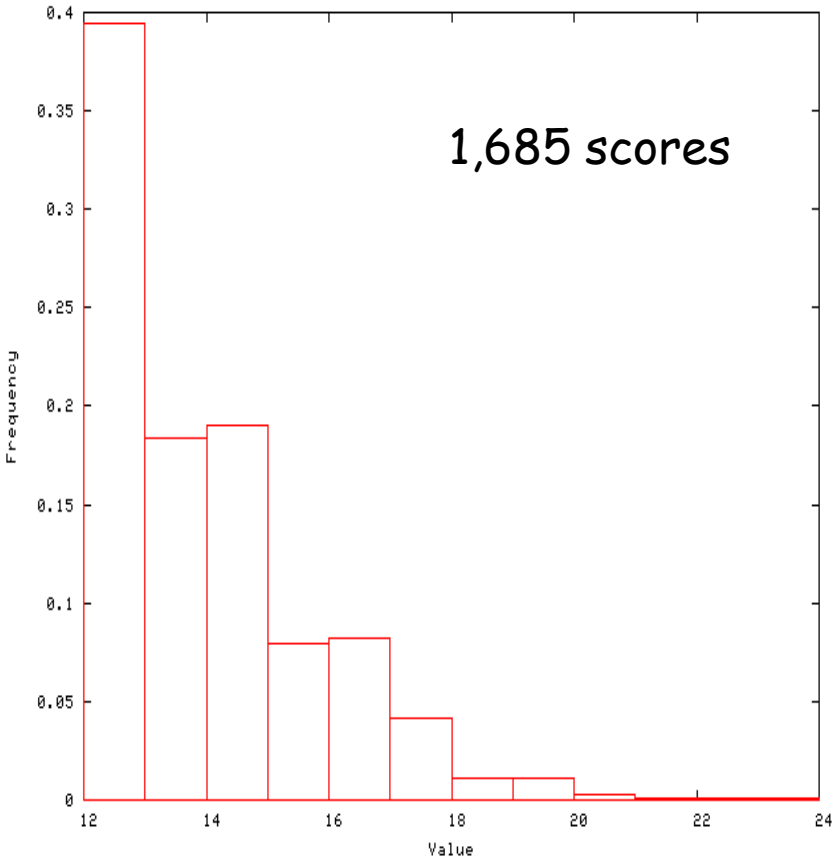


High scores from related sequences

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

Empirical null score distribution

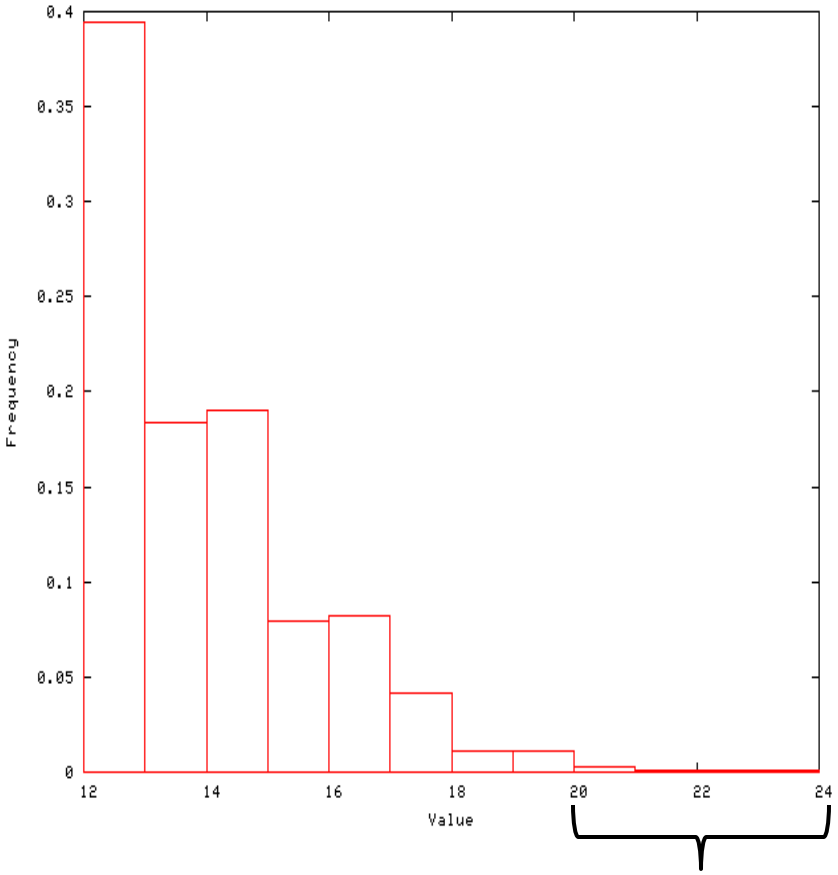
- This distribution is similar to the previous one, but generated using a randomized sequence database (each sequence shuffled).



(notice the scale is shorter here)

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

Computing an empirical 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.
- p-value = Pr(data|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.

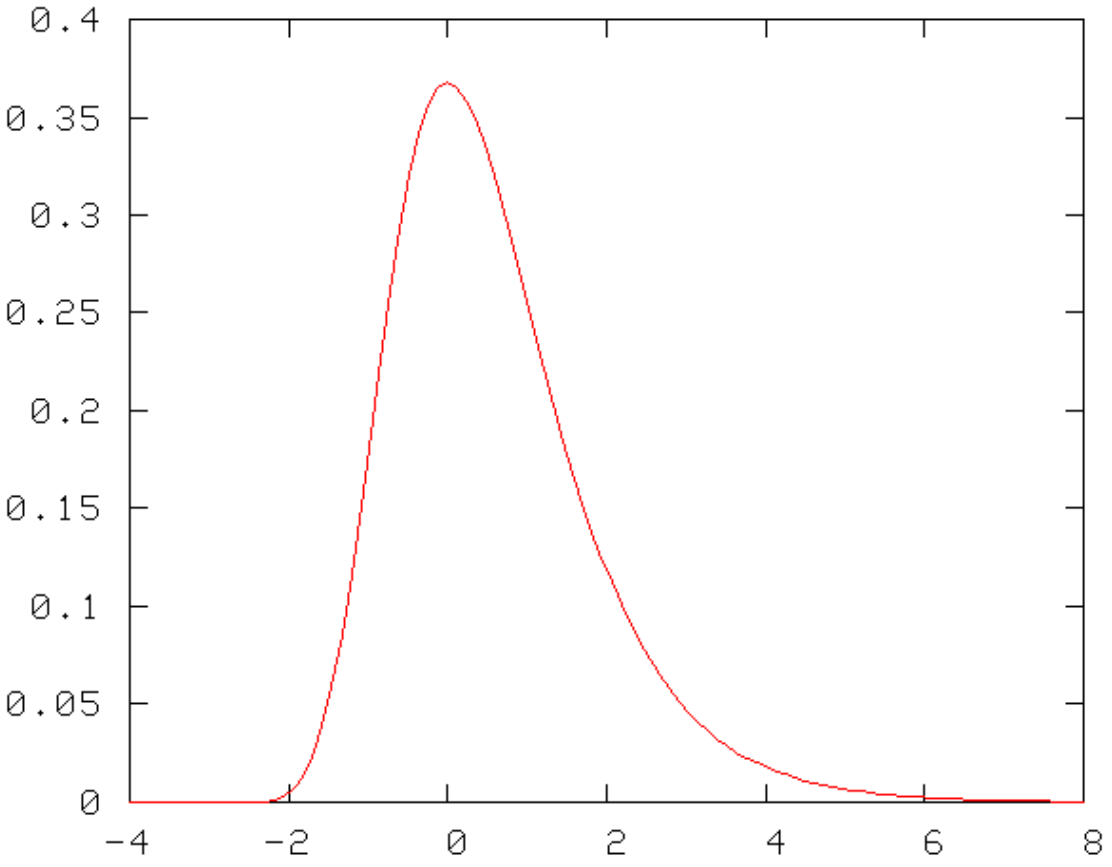
Problems with empirical distributions

- 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

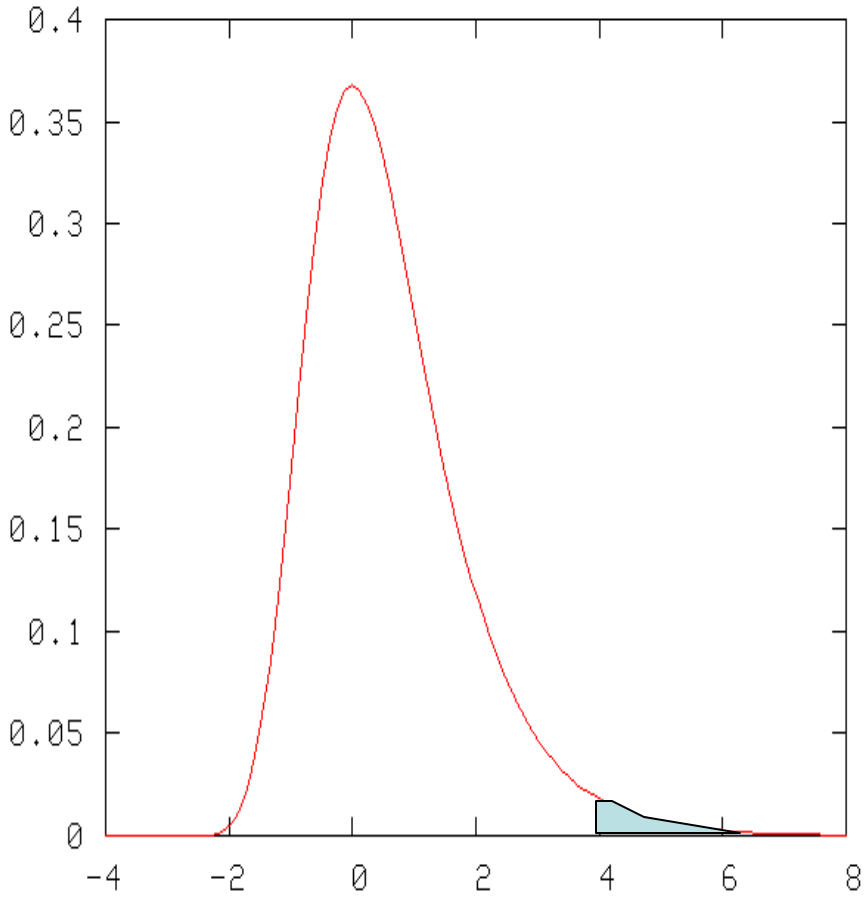
- Solution: Characterize the form of the score distribution mathematically.
- Fit the parameters of the distribution empirically, or compute them analytically.
- Use the resulting distribution to compute accurate p-values.
- First solved by Karlin and Altschul.

Extreme value distribution



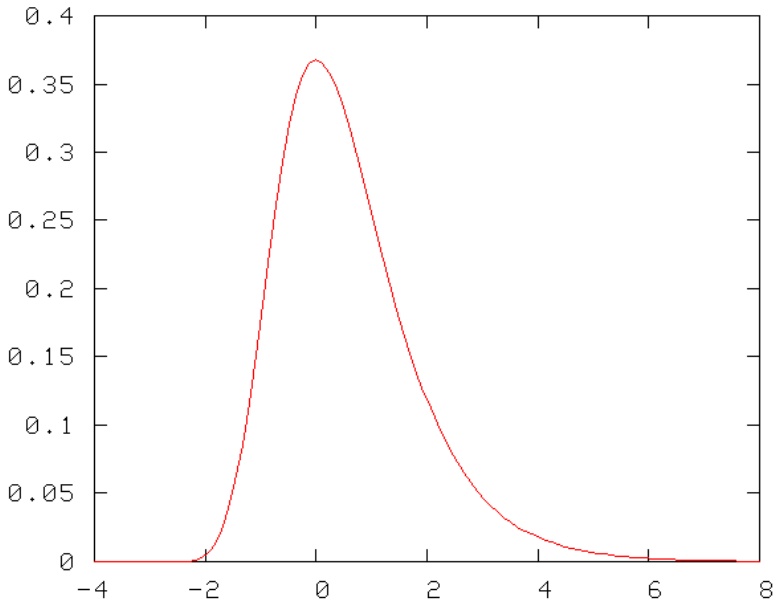
This distribution is roughly normal near the peak, but characterized by a larger tail on the right.

Computing a p-value



- The probability of observing a score >=4 is the area under the curve to the right of 4.
- p-value = Pr(data|null)

Unscaled EVD equation



Compute this value for x=4.

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

S is data score, x is test score

Computing a p-value

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

$$P(S \geq 4) = 0.018149$$