Sequence comparison: Significance of alignment scores

http://faculty.washington.edu/jht/GS559_2017/

Genome 559: Introduction to Statistical and Computational Genomics

Prof. James H. Thomas

Review

- How BLAST speeds up pair alignment
- How to interpret an E-value

Are these proteins related?

(intuitive answers)

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

NO (score = -1)

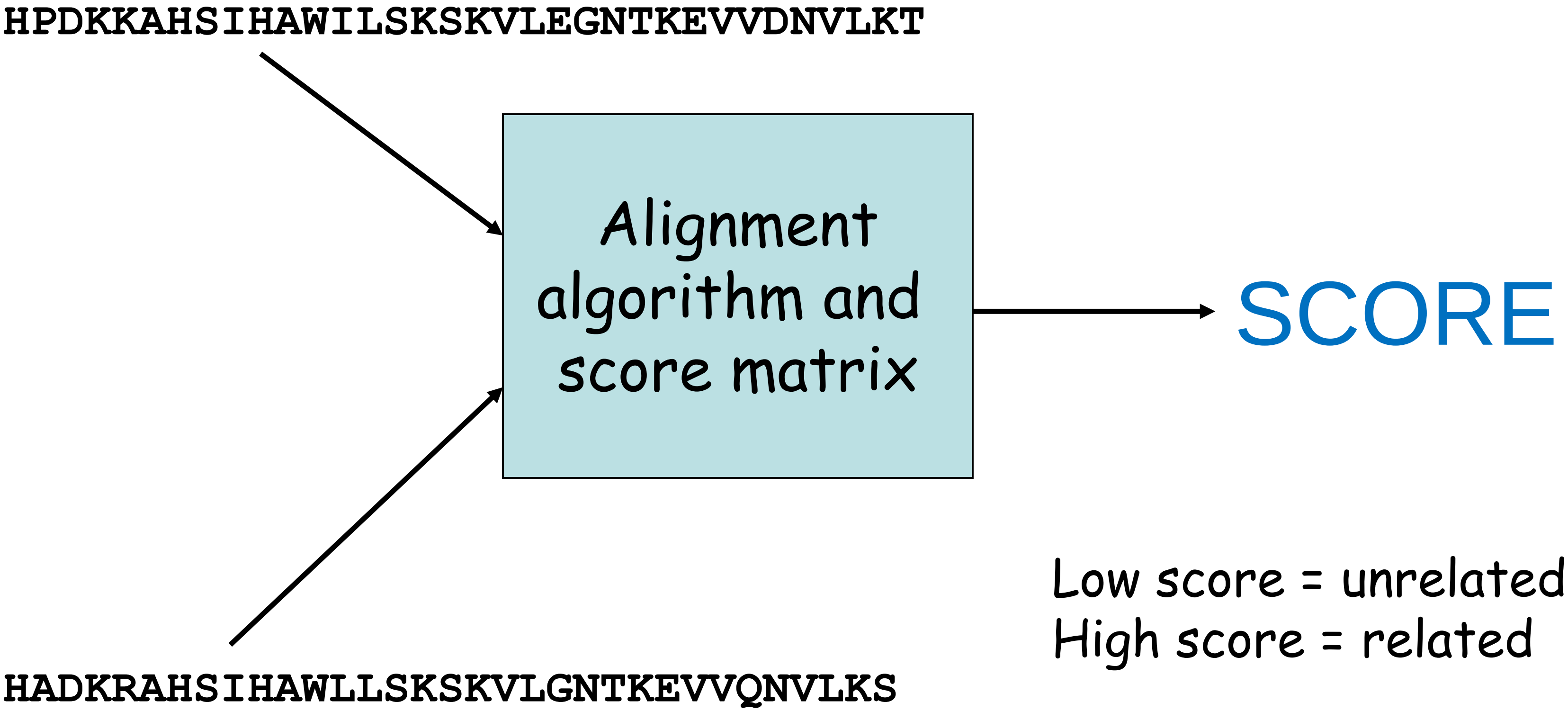
```
SEQ 1:       RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
                 L P    W LDATYKNYA  Y C     L
SEQ 2:       QFFPLMPPAPYWILDATYKNYALVYSCTTFFWLF
```

PROBABLY (score = 15)

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

YES (score = 24)

Significance of scores



HPDKKAHSIHAWILSKSKVLEGNTKEVVDNVLKT

Alignment algorithm and score matrix

SCORE

HADKRAHSIHAWLLSKSKVLGNTKEVVQNVLKS

Low score = unrelated
High score = related

How high is high enough?

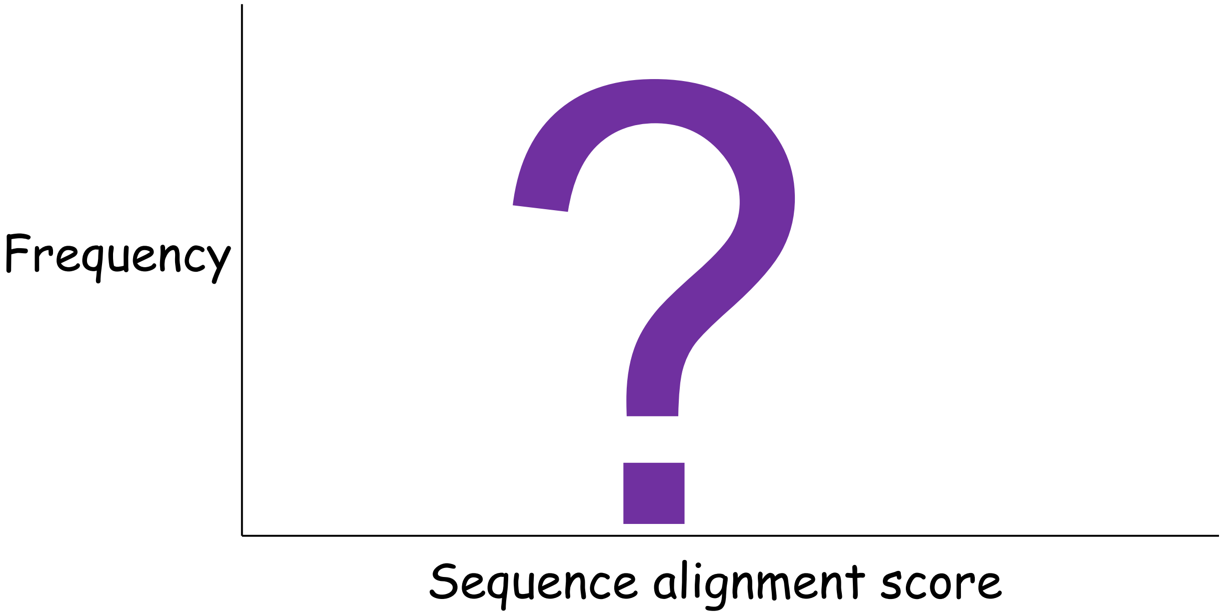
The null hypothesis

- First characterize distribution of alignment scores from sequences that are not related.

- This distribution constitutes a null hypothesis.

- The statistical test will determine whether the observed result provides a reason to reject the null hypothesis.

Sequence alignment score distribution

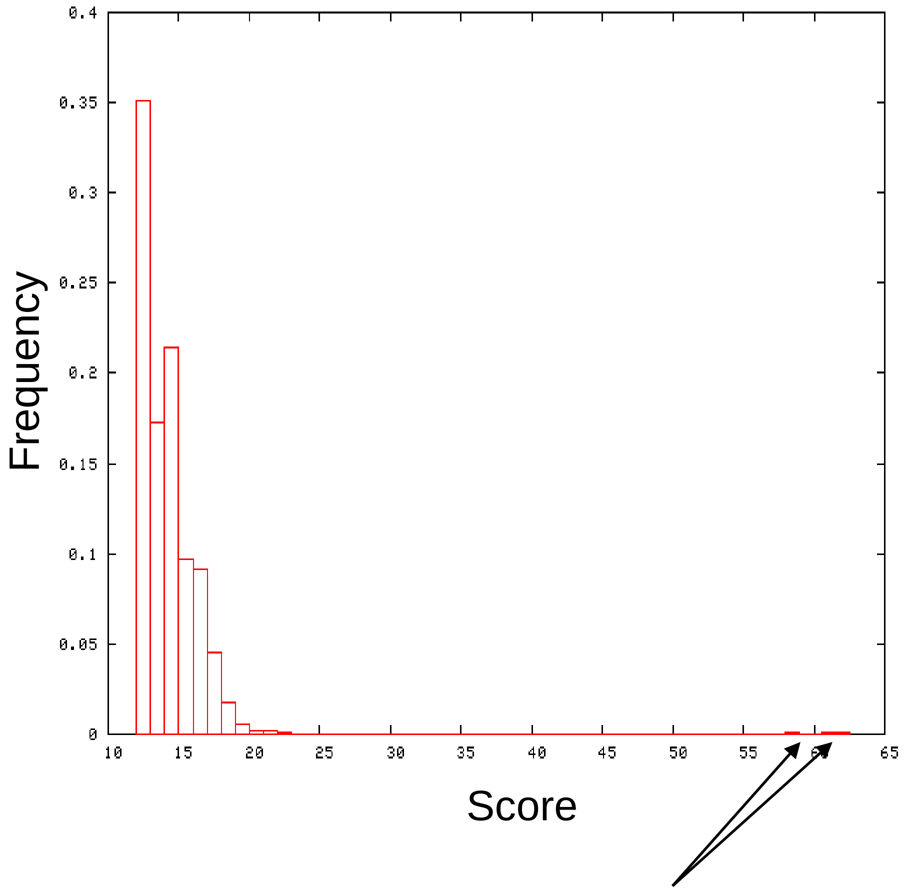


- Use BLAST to search a randomly generated database of sequences using a given query sequence.
- What will be the form of the resulting distribution of pairwise alignment scores?

Empirical score distribution

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

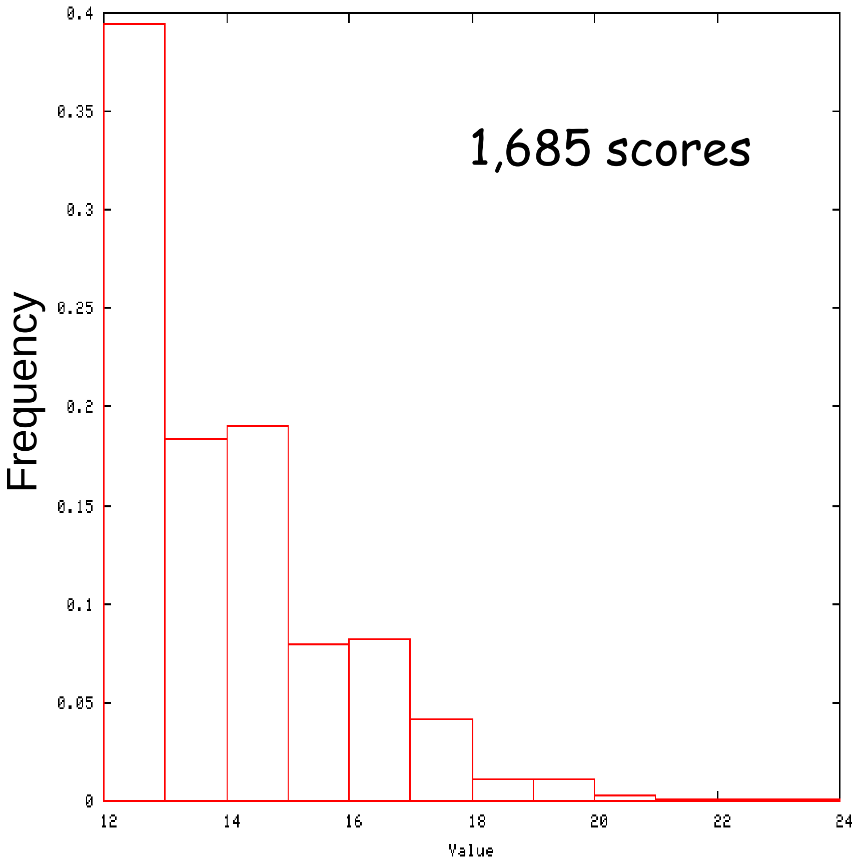
(note - there are lots of lower scoring alignments not reported)



High scores from related sequences

Empirical null score distribution

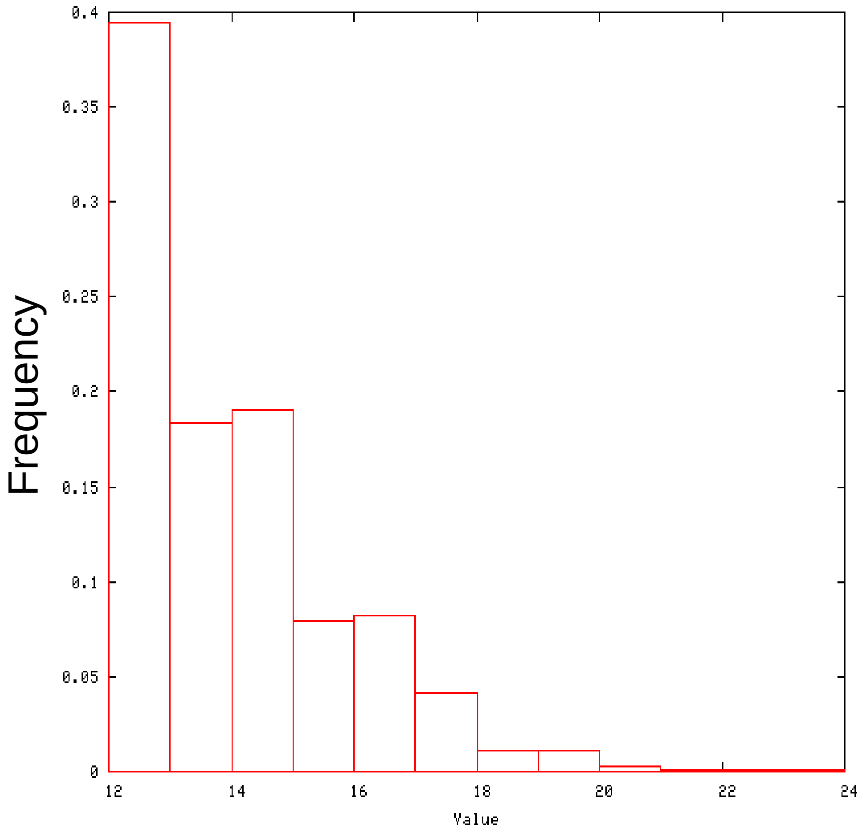
- This distribution is generated using a randomized sequence database (residue order in each sequence shuffled).



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

(notice the x scale is shorter here)

Computing an empirical p-value



- 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)

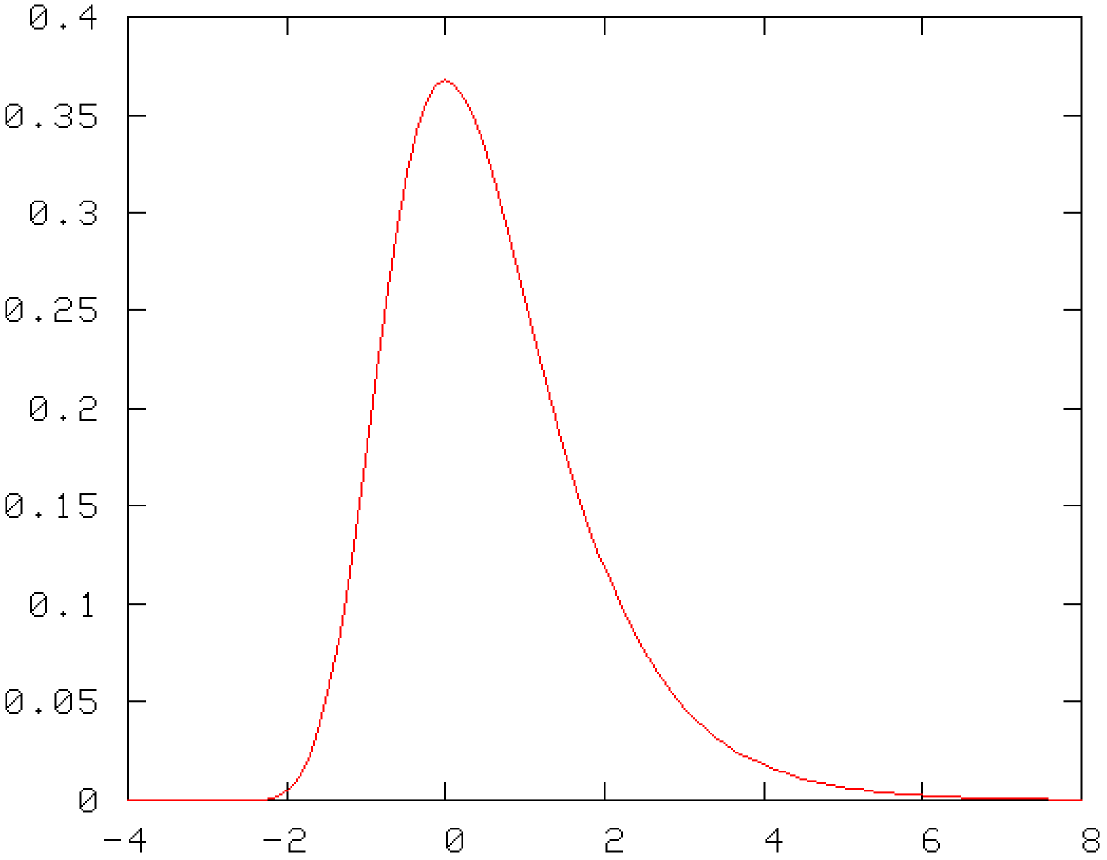
Problems with empirical distributions

- We are interested in very small probabilities (high scoring matches).
- These are computed from the *tail* of the null distribution.
- Estimating a distribution with an accurate tail is computationally expensive - it requires a very large number of alignments.

A solution

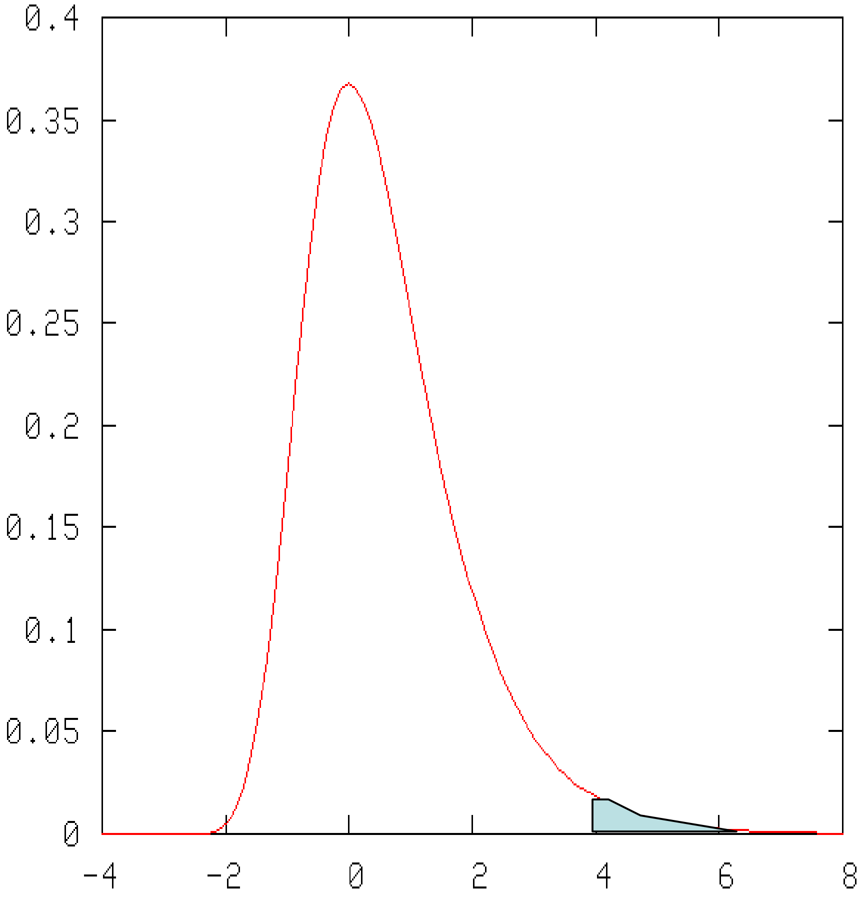
- Solution: characterize the form of the score distribution mathematically.
- Use the resulting distribution to compute accurate p-values.
- First solved by Karlin and Altschul.

Extreme value distribution (EVD) (aka Gumbel Distribution)



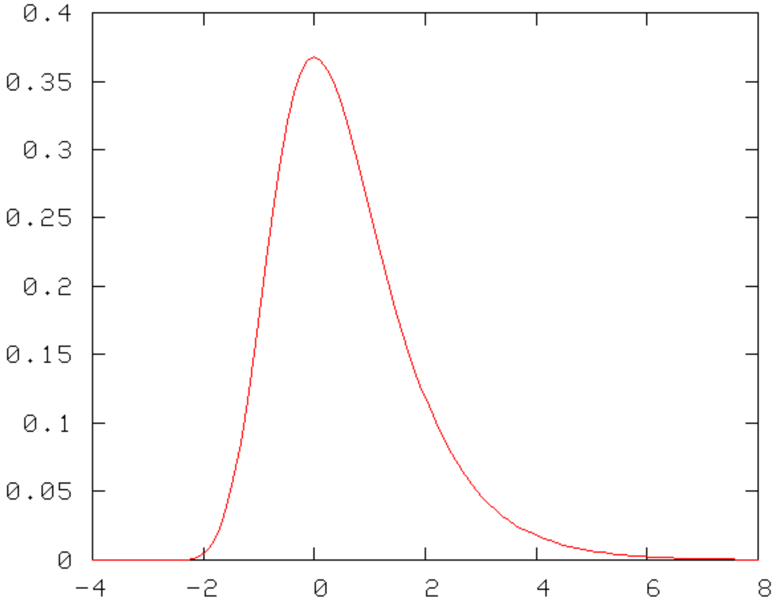
This distribution is roughly normal near the peak, but has a longer tail on the right.

Computing a p-value



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

Unscaled EVD equation (null)



Compute this value for x=4.

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

S is data score, x is test score

Computing a p-value

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

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

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 (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 - CDF**.

- for the unscaled extreme value distribution:

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

- and we want to compute **1 - CDF**:

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