The statistics of pairwise alignment

BMI/CS 576

www.biostat.wisc.edu/bmi576/

Irene Ong

irene.ong@wisc.edu

Fall 2017

Issues in scoring pairwise alignments

- How do we determine the substitution and gap scores for alignment?
- How do we determine whether the score of the best alignment is indicative of truly related sequences?
- These issues are related and addressed via statistical models

Circular logic in alignment and scoring

- How do we know what is the right distance without a good alignment?
- How do we construct a good alignment without knowing what substitutions were made previously?

Probabilistic Model of Alignments

- We'll focus on protein alignments without gaps
- given an alignment, we can consider two possibilities

  R: the sequences are related by evolution

  U: the sequences are unrelated

- How can we distinguish these possibilities?
- How is this view related to amino-acid substitution matrices?

Model for *Unrelated* Sequences

- We'll assume that each position in the alignment is sampled randomly from some distribution of amino acids
- We'll assume that amino acids at each position are **independent** of each other

- let q_a be the probability of amino acid *a*

- the probability of an *n*-character alignment of *x* and *y* is given by

$$\Pr(x, y \mid U) = \prod_{i=1}^{n} q_{x_i} \prod_{i=1}^{n} q_{y_i}$$

Model for *Related* Sequences

- We'll assume that each pair of aligned amino acids evolved from a common ancestor
- We'll assume each pair is **independent** of the other pairs

- let p_{ab} be the probability that evolution gave rise to amino acid *a* in one sequence and *b* in another sequence

- the probability of an alignment of *x* and *y* is given by

$$\Pr(x, y \mid R) = \prod_{i=1}^{n} p_{x_i y_i}$$

Probabilistic Model of Alignments

- How can we decide which possibility (U or R) is more likely?
- one principled way is to consider the relative likelihood of the two possibilities

$$\frac{\Pr(x, y \mid R)}{\Pr(x, y \mid U)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_i q_{y_i}} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} q_{y_i}}$$

- taking the log, we get

$$\log \frac{\Pr(x, y \mid R)}{\Pr(x, y \mid U)} = \sum_i \log \left(\frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \right)$$

- This is the *log-odds ratio* (or *log likelihood ratio)*

Probabilistic Model of Alignments

- If we let the substitution matrix score for the pair *a, b* be:

$$s(a,b) = \log\left(\frac{p_{ab}}{q_a q_b}\right)$$

- Then the score of an ungapped alignment is the log likelihood ratio:

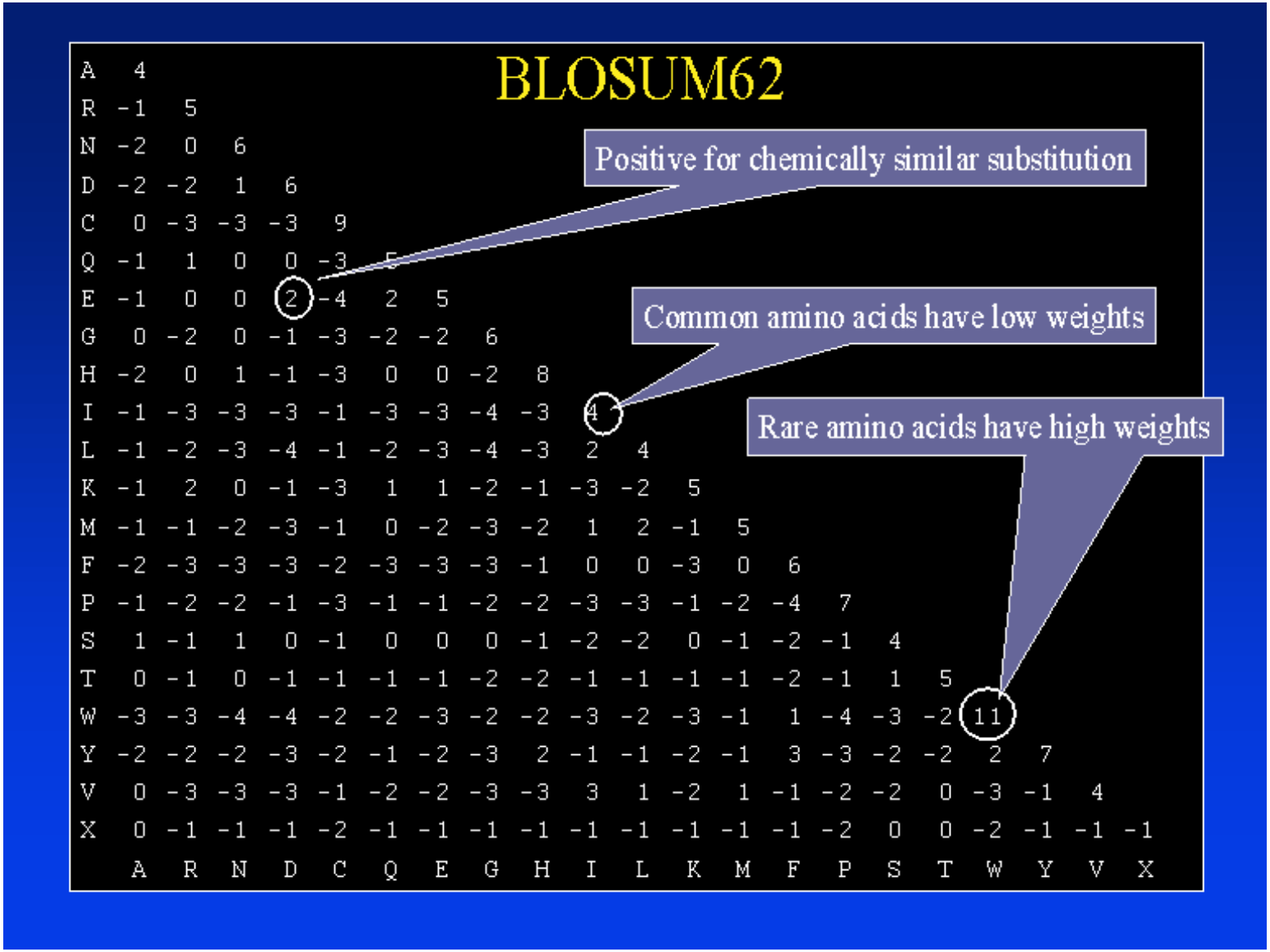
$$S = \sum_i s(x_i, y_i) = \log \frac{\Pr(x, y \mid R)}{\Pr(x, y \mid U)}$$

Substitution Matrices

- two popular sets of matrices for protein sequences
  - PAM matrices [Dayhoff *et al.,* 1978]
  - BLOSUM matrices [Henikoff & Henikoff, 1992]

- both try to capture the the relative substitutability of amino acid pairs in the context of evolution

Blosum 62 Matrix



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X
A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3	9																
Q	-1	1	0	0	-3	5															
E	-1	0	0	2	-4	2	5														
G	0	-2	0	-1	-3	-2	-2	6													
H	-2	0	1	-1	-3	0	0	-2	8												
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4											
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4										
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5									
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5				
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1

Substitution Matrices

- the substitution matrix score for the pair *a*, *b* is given by:

$$s(a,b) = \log\left(\frac{p_{ab}}{q_a q_b}\right)$$

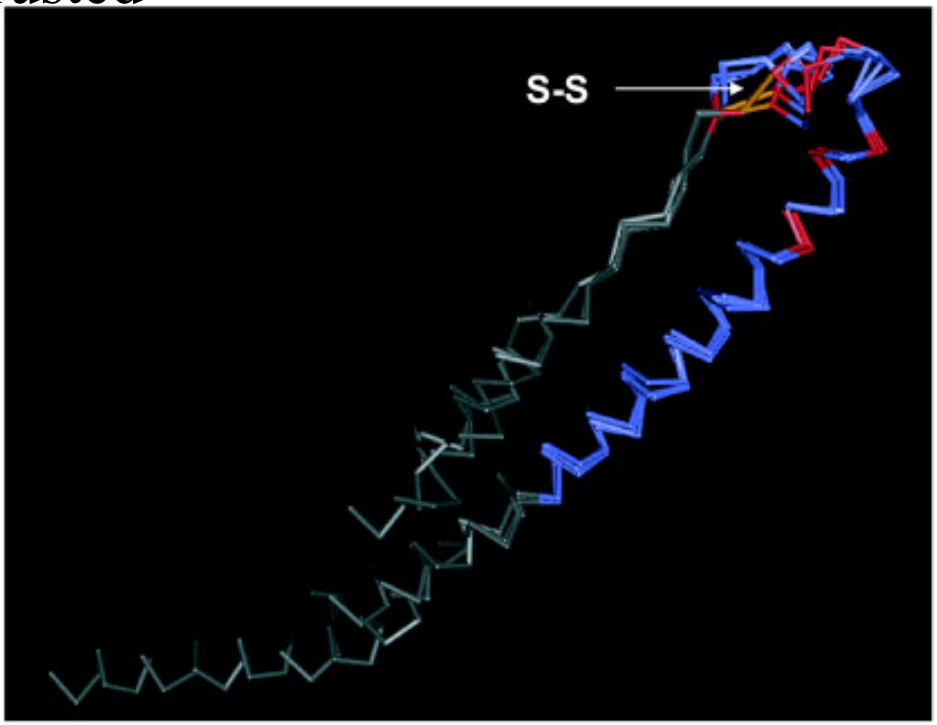
- but how do we get values for p_{ab} (probability of *a* and *b* given that they are derived from a common ancestor)?
- it depends on how long ago sequences diverged

  diverged recently: $p_{ab} \approx 0$ for $a \neq b$

  diverged long ago: $p_{ab} \approx q_a q_b$

Substitution Matrices

- key idea: trusted alignments of related sequences provide information about biologically permissible mutations
- protein structure similarity provides the gold standard for which alignments are trusted




```
D2/CEL/OR  sdvqAISSTIQDLQDQVDSLAEVVLQNRRGLDLLTAEQGGICLALQEKCcfyank
MMLV       ddlrEVEKSISNLEKSLTSLSEVVLQNRRGLDLLFLKEGGLCAALKEECafyad~
HTLV-1     kdisQLTQAIVKNHKNLLKIAQYAAQNRRGLDLLFWEQGGLCKALQEQCcflnit
Ebola      qlanETTQALQLFLRATTELRTFSILNRKAIDFLLQRWGGTCHILGPDCriephd
```

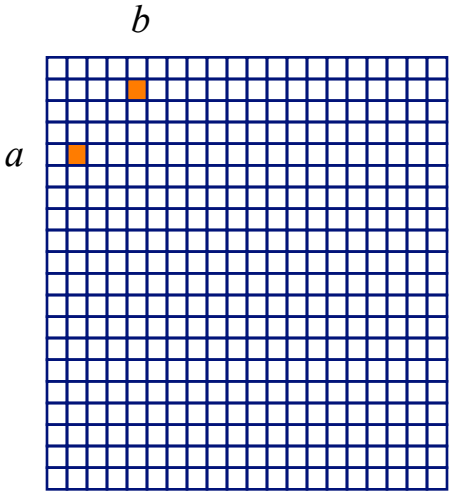
BLOSUM Matrices

- [Henikoff & Henikoff, *PNAS* 1992]
- probabilities estimated from "blocks" of sequence fragments that represent *structurally* conserved regions in proteins
- transition frequencies observed directly by counting pairs of characters between clusters in the blocks. Sequences within blocks are clustered at various levels:
  - 45% identical (BLOSUM-45)
  - 50% identical (BLOSUM-50)
  - 62% identical (BLOSUM-62)
  - etc.

BLOSUM Matrices

- given: a set of sequences in a block
- fill in matrix *A* with number of observed substitutions (we won't worry about details of some normalization that happens here)

a paired with *b* →



$$p_{ab} = \frac{A_{ab}}{\sum_{c,d} A_{cd}} \qquad q_a = \frac{\sum_b A_{ab}}{\sum_{c,d} A_{cd}}$$

Assessing significance of the alignment score

- There are two ways to do this
  - Bayesian approach
  - Classical approach

Bayesian approach

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

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

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

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

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

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

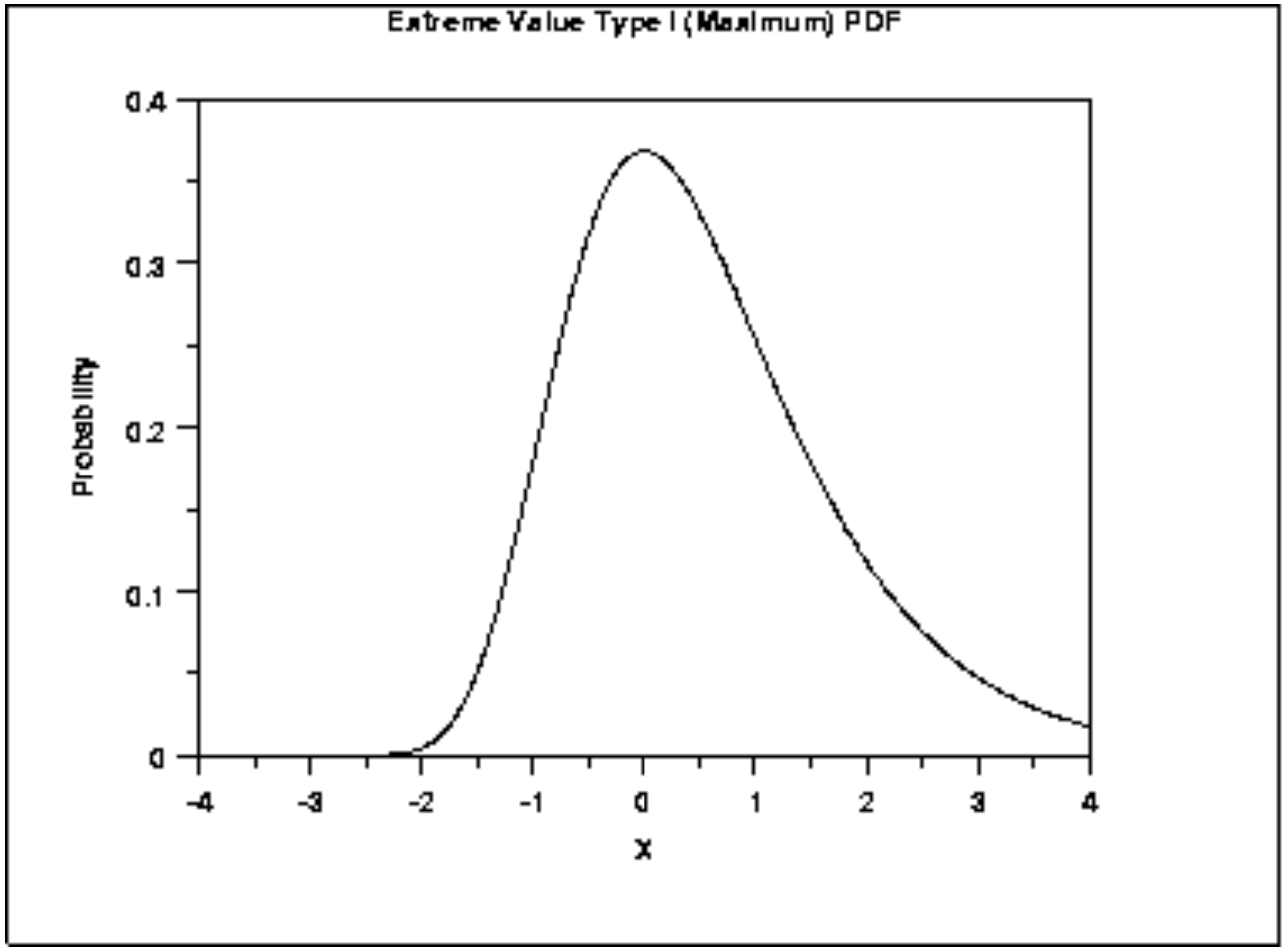
Scores from Random Alignments

- 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

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

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)

Scoring Matrices

- Differences between PAM and BLOSUM
  - PAM based on predictions of mutations when proteins diverged from common ancestor – explicit evolutionary model
  - BLOSUM based on common regions (BLOCKS) in protein families
- BLOSUM is better designed to find conserved domains
- Much larger data set was used than for PAM
- BLOSUM matrices with small percentage correspond to PAM with large evolutionary distances
  - BLOSUM64 ~ PAM120

Are these proteins homologs?

```
SEQ 1:RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
        | |     | |    | |     | |      |
SEQ 2:QFFPLMPPAPYWILATDYENLPLVYSCTTFFWLF
```

NO (score = 69)

```
SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
          | |     | |||||||||   | |      |
SEQ 2: QFFPLMPPAPYWILDATYKNYALVYSCTTFFWLF
```

MAYBE (score = 104)

```
SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
       |||  | ||    | |||||||||   | |||||||
SEQ 2: RVVPLMPSAPYWILDATYKNYALVYSCDVTYKLF
```

YES (score = 153)

Slides courtesy of William Noble

An empirical histogram

- Roll a 20-sided die 1000 times.
- Keep a tally of the number of times each value is observed.

Slides courtesy of William Noble

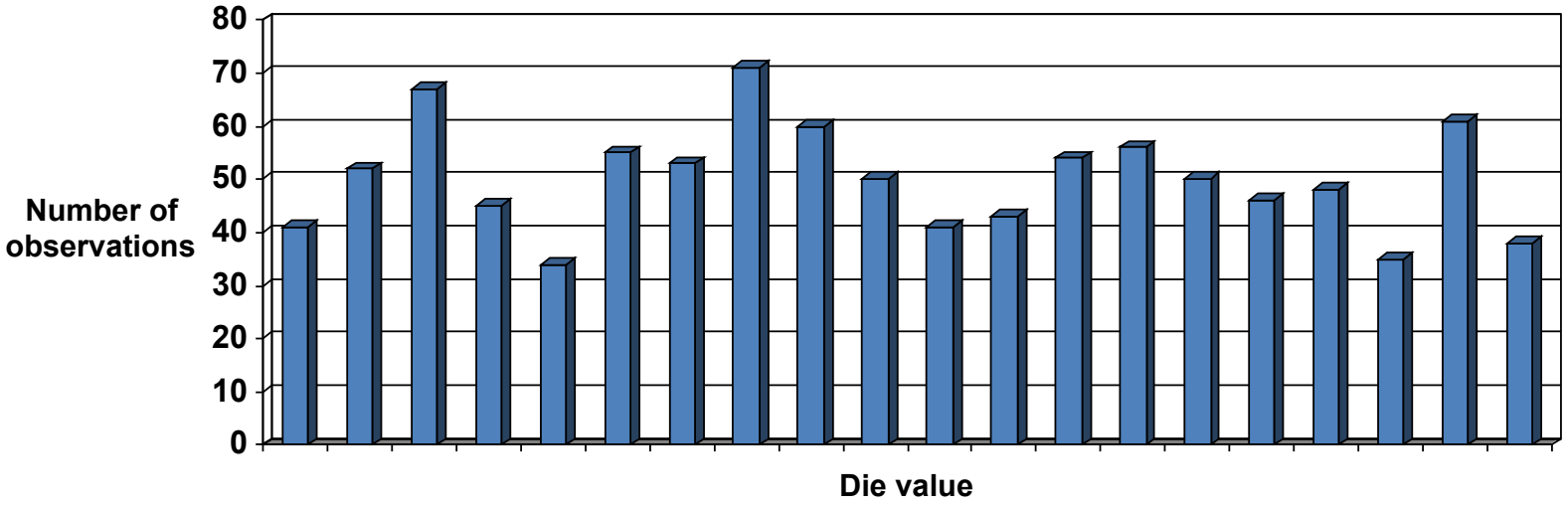
A table of results

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
41	52	67	45	34	55	53	71	60	50	41	43	54	56	50	46	48	35	61	38

Slides courtesy of William Noble

A table of results

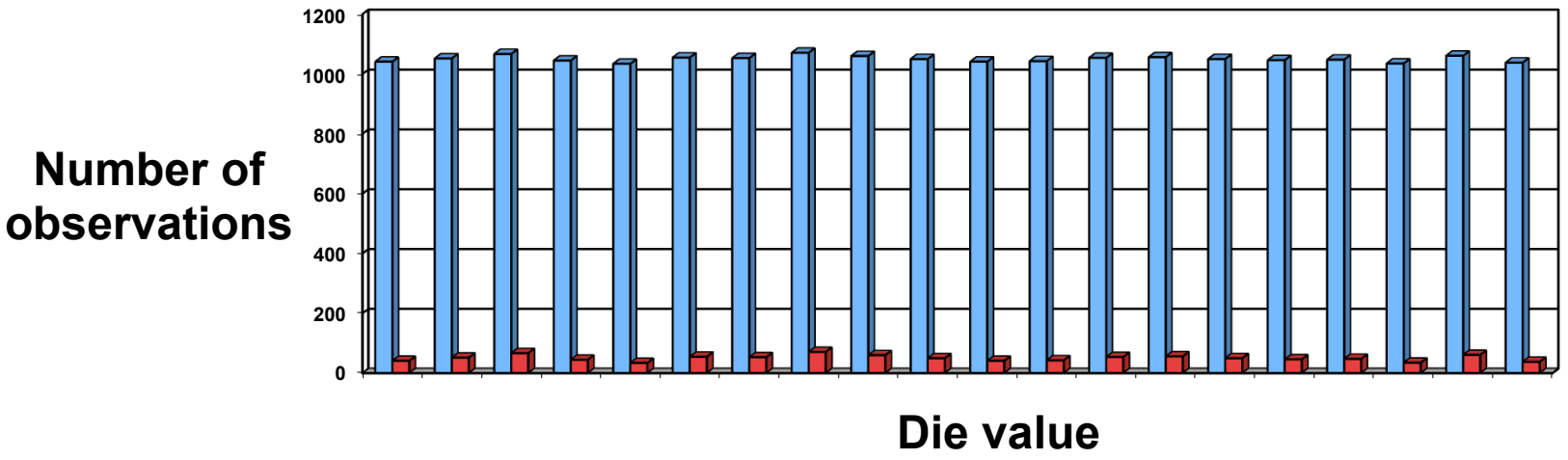
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
41	52	67	45	34	55	53	71	60	50	41	43	54	56	50	46	48	35	61	38



A *histogram* plots the number of times or the frequency with which each value of a given variable (e.g., the die value) is observed.

Slides courtesy of William Noble

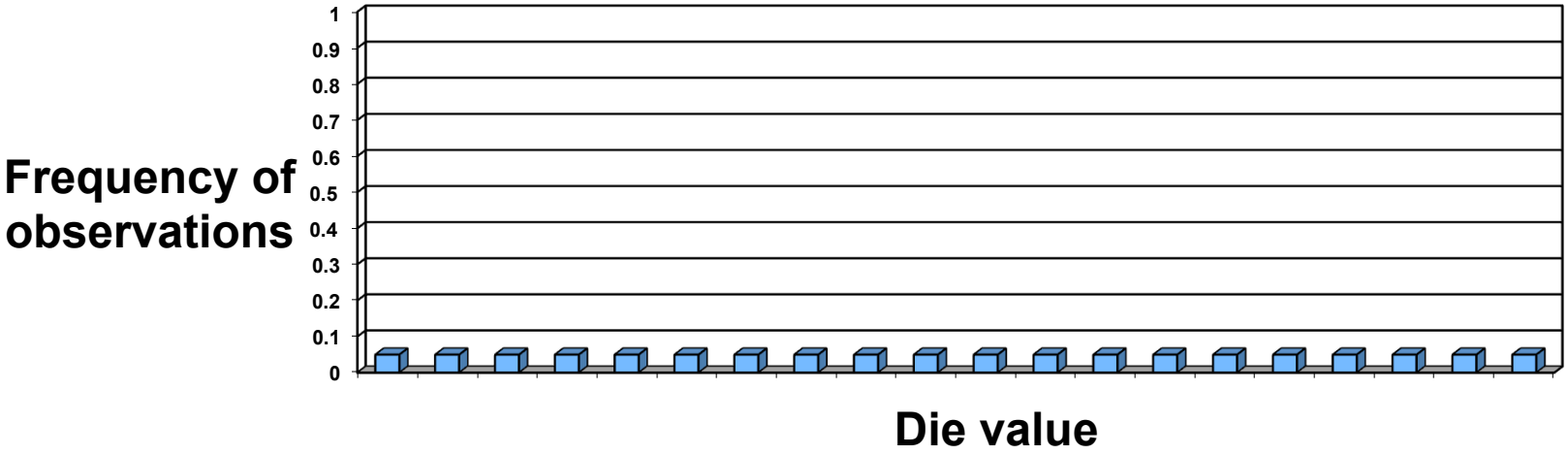
After 21,000 rolls



As the number of observations increases, the histogram becomes smoother.

Slides courtesy of William Noble

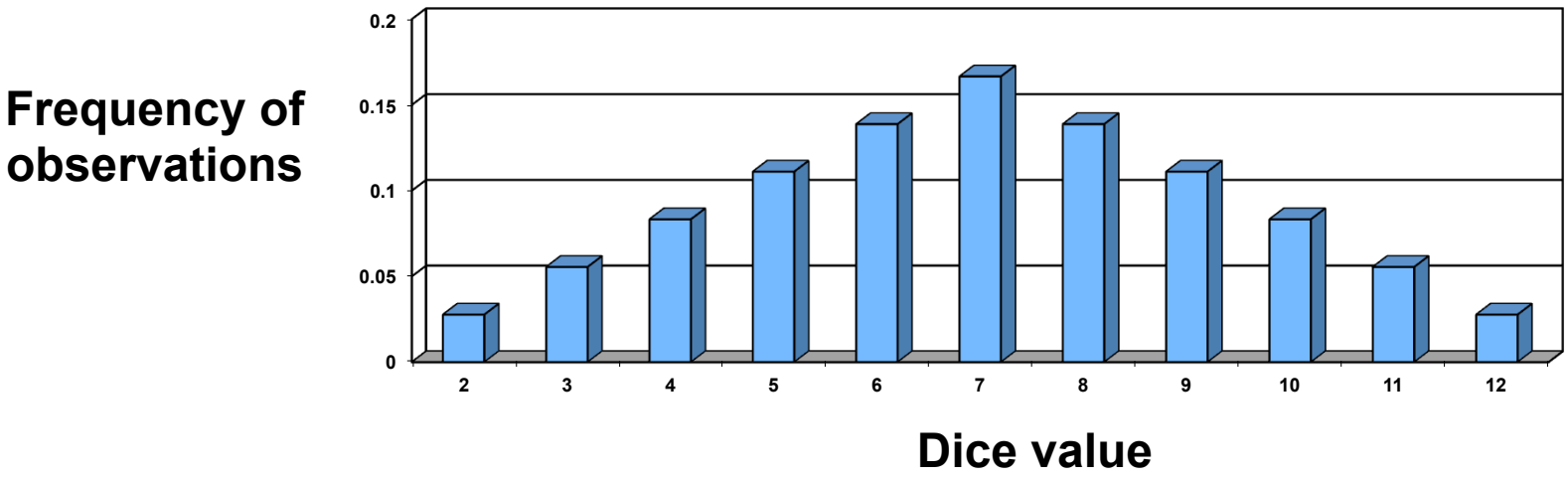
Frequency histogram = distribution



- Divide through by the total number of observations to get frequencies.
- The resulting histogram is called a distribution.
- The sum of the bars in the distribution is 1.
- This particular distribution is “uniform.”

Slides courtesy of William Noble

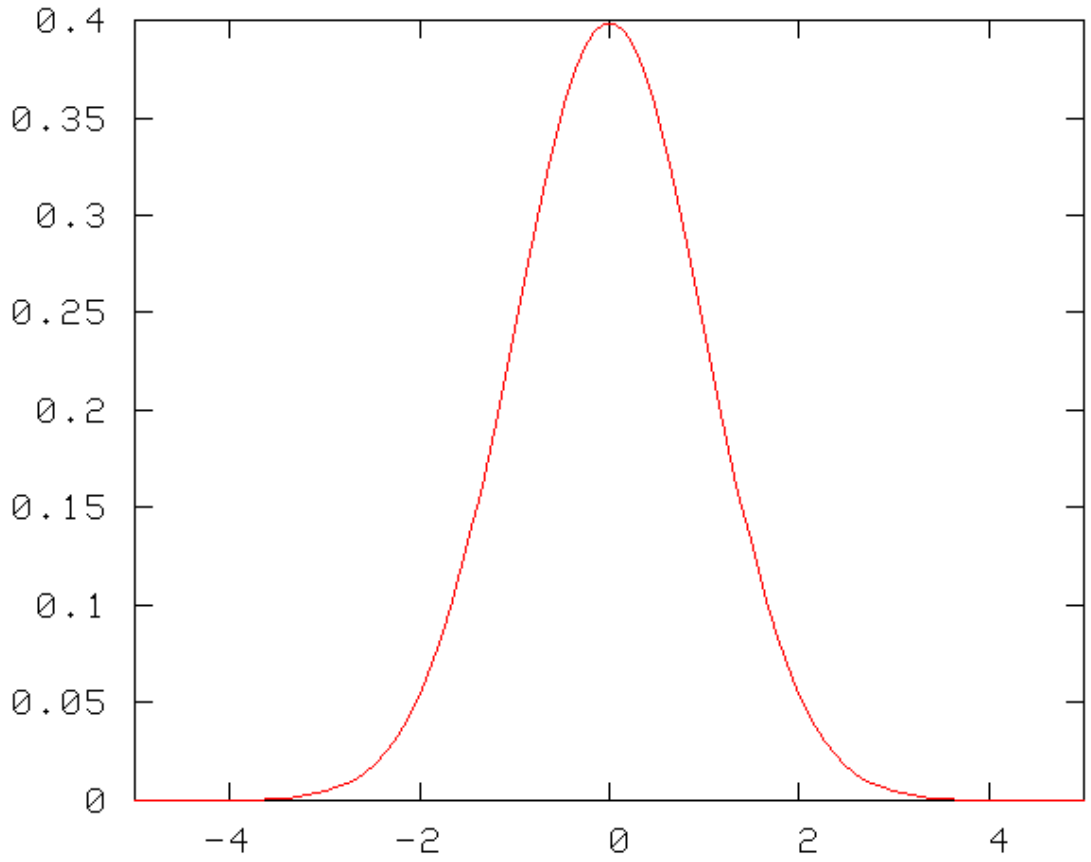
Distribution from two dice



- There is only one way to get a score of 2.
- How many ways can you get a 7 from two six-sided dice?
- This distribution is approximately normal (Gaussian).

Slides courtesy of William Noble

Normal distribution



This is probably the most common distribution in science.
The area under a continuous distribution curve is 1.

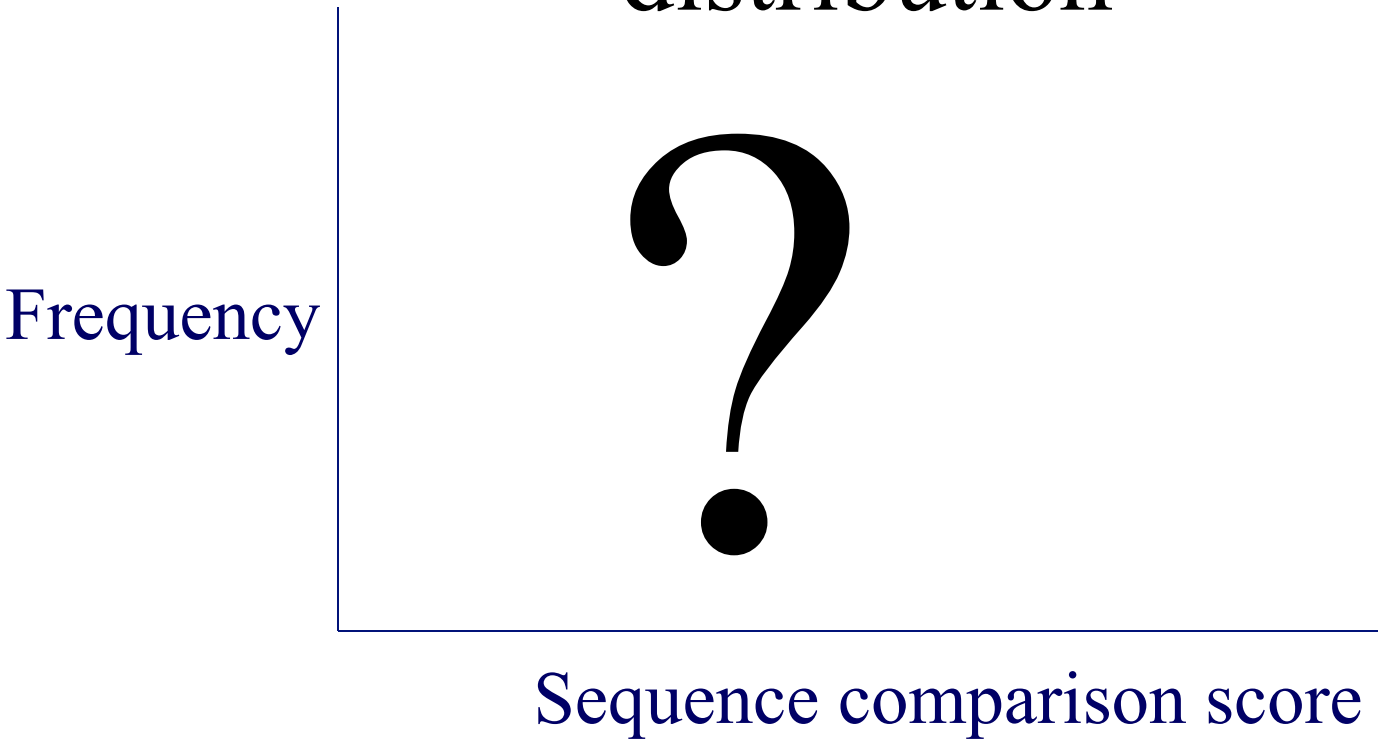
Slides courtesy of William Noble

The null hypothesis

- 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 the results are merely a product of chance factors.

Slides courtesy of William Noble

Sequence similarity score distribution

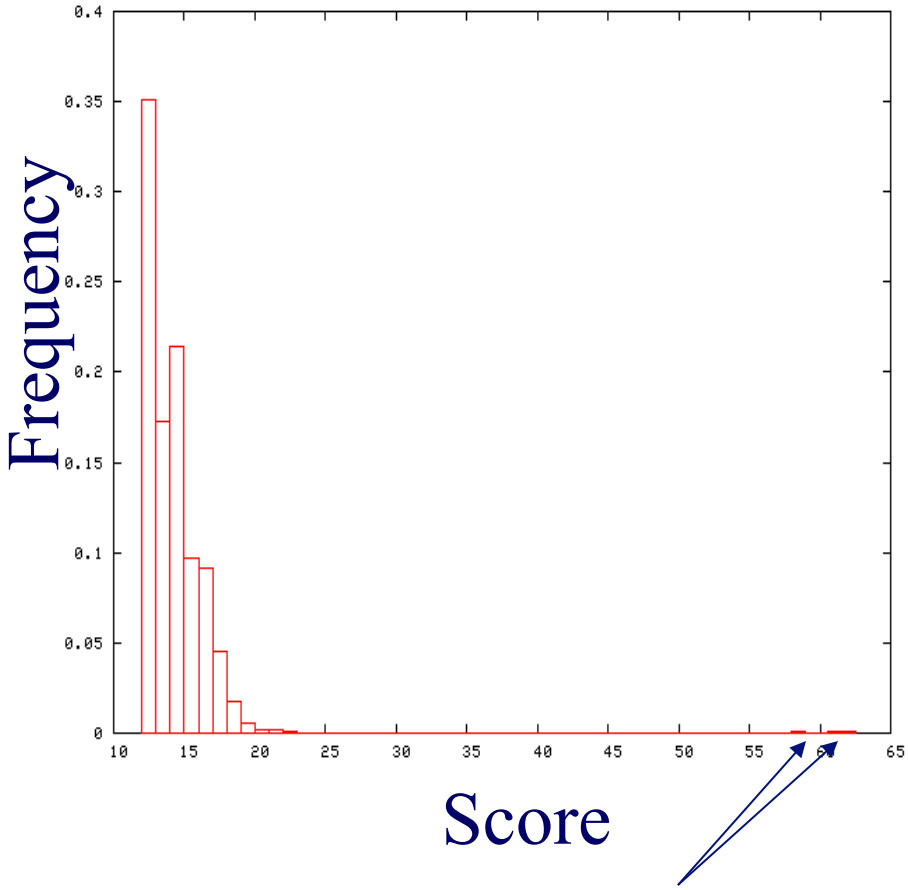


- Search a randomly generated database of DNA sequences using a randomly generated DNA query.
- What will be the form of the resulting distribution of pairwise sequence comparison scores?

Slides courtesy of William Noble

Empirical score distribution

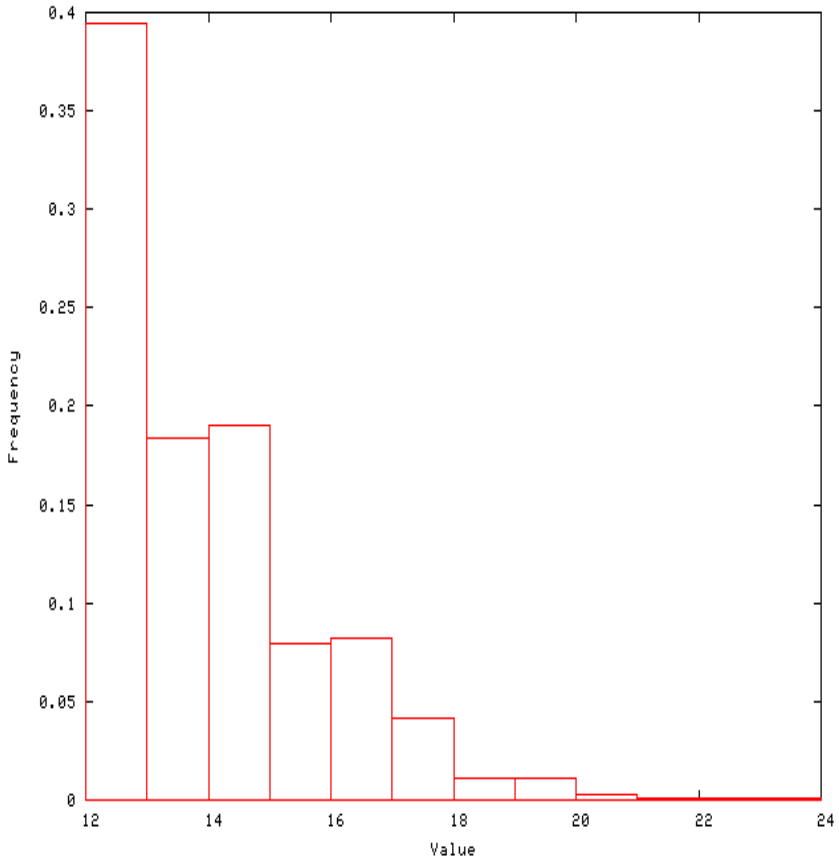
- The picture shows a distribution of scores from a real database search using BLAST.
- This distribution contains scores from non-homologous and homologous pairs.



Slides courtesy of William Noble

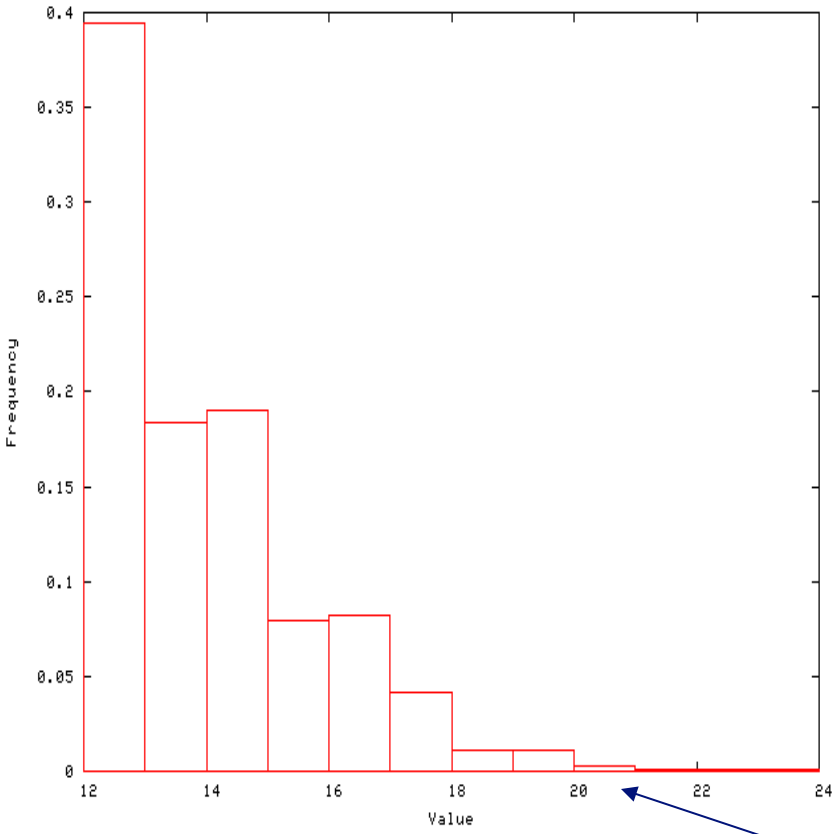
Empirical null score distribution

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



Slides courtesy of William Noble

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.
- p-value = Pr(data|null)

Out of 1685 scores, 28 receive a score of 20 or better. Thus, the p-value associated with a score of 20 is approximately 28/1685 = 0.0166.

Slides courtesy of William Noble

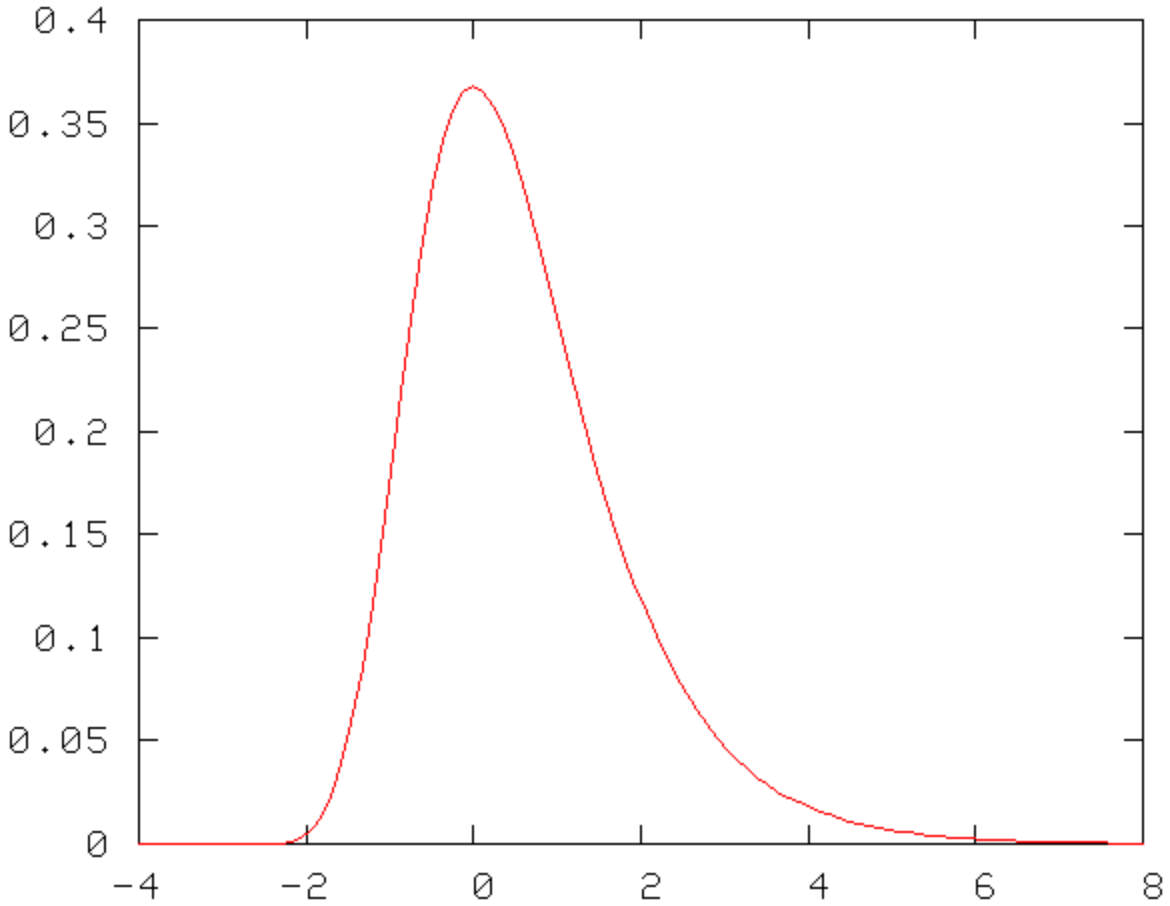
Problems with empirical distributions

- 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.

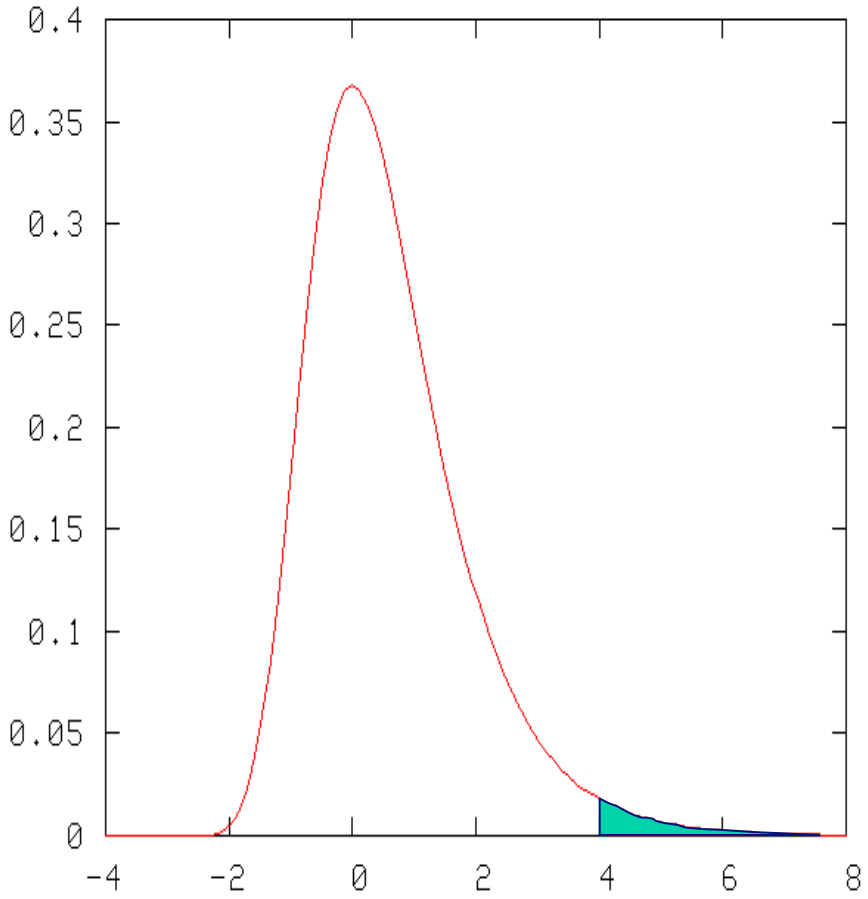
Extreme value distribution



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

Slides courtesy of William Noble

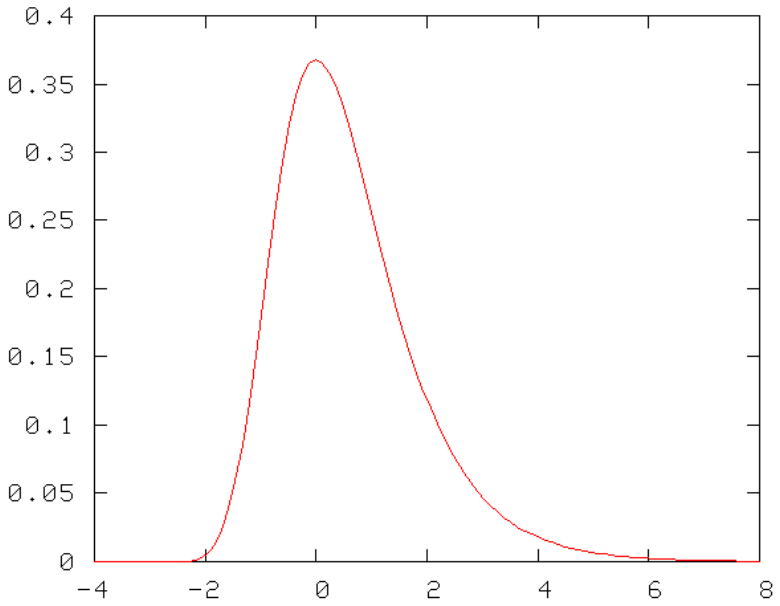
Computing a p-value



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

Slides courtesy of William Noble

Extreme value distribution



Compute this value for x=4.

$$Y_{ev} = \exp\left[-x - e^{-x}\right]$$

Slides courtesy of William Noble

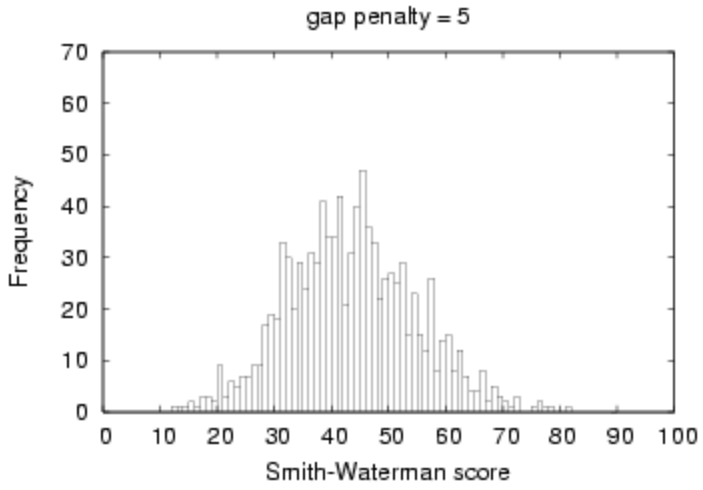
Computing a p-value

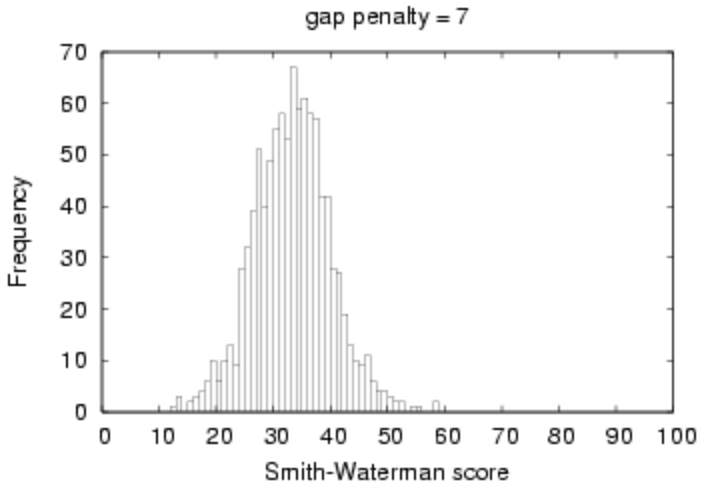
$$P(S \geq x) = 1 - \exp\left[-e^{-4}\right]$$

- Calculator keys: `4, +/-, inv, ln, +/-, inv, ln, +/-, +, 1, =`
- Solution: 0.018149

Slides courtesy of William Noble

Scaling the EVD





- An extreme value distribution derived from, e.g., the Smith-Waterman algorithm will have a characteristic mode μ 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.

Slides courtesy of William Noble

An example

You run the Smith-Waterman algorithm and get a score of 45. You then run Smith-Waterman with the same query but using a shuffled version of the database, and you fit an extreme value distribution to the resulting empirical distribution. The parameters of the EVD are $\mu = 25$ and $\lambda = 0.693$. What is the p-value associated with 45?

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

Slides courtesy of William Noble

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 $\mu = 25$ and $\lambda = 0.693$. What is the p-value associated with 45?

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

Slides courtesy of William Noble

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.

Slides courtesy of William Noble

Multiple testing

- Say that you perform a statistical test with a 0.05 threshold, but you repeat the test on twenty different observations.
- Assume that all of the observations are explainable by the null hypothesis.
- What is the chance that at least one of the observations will receive a p-value less than 0.05?

Slides courtesy of William Noble

Multiple testing

- Say that you perform a statistical test with a 0.05 threshold, but you repeat the test on twenty different observations. Assuming that all of the observations are explainable by the null hypothesis, what is the chance that at least one of the observations will receive a p-value less than 0.05?
- Pr(making a mistake) = 0.05
- Pr(not making a mistake) = 0.95
- Pr(not making any mistake) = 0.95^{20} = 0.358
- Pr(making at least one mistake) = 1 - 0.358 = 0.642
- There is a 64.2% chance of making at least one mistake.

Slides courtesy of William Noble

Bonferroni correction

- Assume that individual tests are *independent*.
- Divide the desired p-value threshold by the number of tests performed.
- For the previous example, 0.05 / 20 = 0.0025.
- Pr(making a mistake) = 0.0025
- Pr(not making a mistake) = 0.9975
- Pr(not making any mistake) = 0.9975^{20} = 0.9512
- Pr(making at least one mistake) = 1 - 0.9512 = 0.0488

Slides courtesy of William Noble

Database searching

- Say that you search the non-redundant protein database at NCBI, containing roughly one million sequences. What p-value threshold should you use?

Database searching

- Say that you search the non-redundant protein database at NCBI, containing roughly one million sequences. What p-value threshold should you use?
- Say that you want to use a conservative p-value of 0.001.
- Recall that you would observe such a p-value by chance approximately every 1000 times in a random database.
- A Bonferroni correction would suggest using a p-value threshold of 0.001 / 1,000,000 = 0.000000001 = 10^{-9}.

Slides courtesy of William Noble

E-values

- A p-value is the probability of making a mistake.
- The E-value is the expected number of times that the given score would appear in a random database of the given size.
- One simple way to compute the E-value is to multiply the p-value times the size of the database.
- Thus, for a p-value of 0.001 and a database of 1,000,000 sequences, the corresponding E-value is $0.001 \times 1{,}000{,}000 = 1{,}000$.

BLAST actually calculates E-values in a more complex way.

Slides courtesy of William Noble

Summary

- A distribution plots the frequency of a given type of observation.
- The area under the distribution is 1.
- Most statistical tests compare observed data to the expected result according to the null hypothesis.
- Sequence similarity scores follow an extreme value distribution, which is characterized by a larger tail.
- The p-value associated with a score is the area under the curve to the right of that score.
- Selecting a significance threshold requires evaluating the cost of making a mistake.
- Bonferroni correction: Divide the desired p-value threshold by the number of statistical tests performed.
- The E-value is the expected number of times that the given score would appear in a random database of the given size.

Slides courtesy of William Noble