Sequence comparison: Significance of similarity scores

Genome 559: Introduction to Statistical and Computational Genomics

Prof. James H. Thomas

Are these proteins related?

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

NO (score = 9)

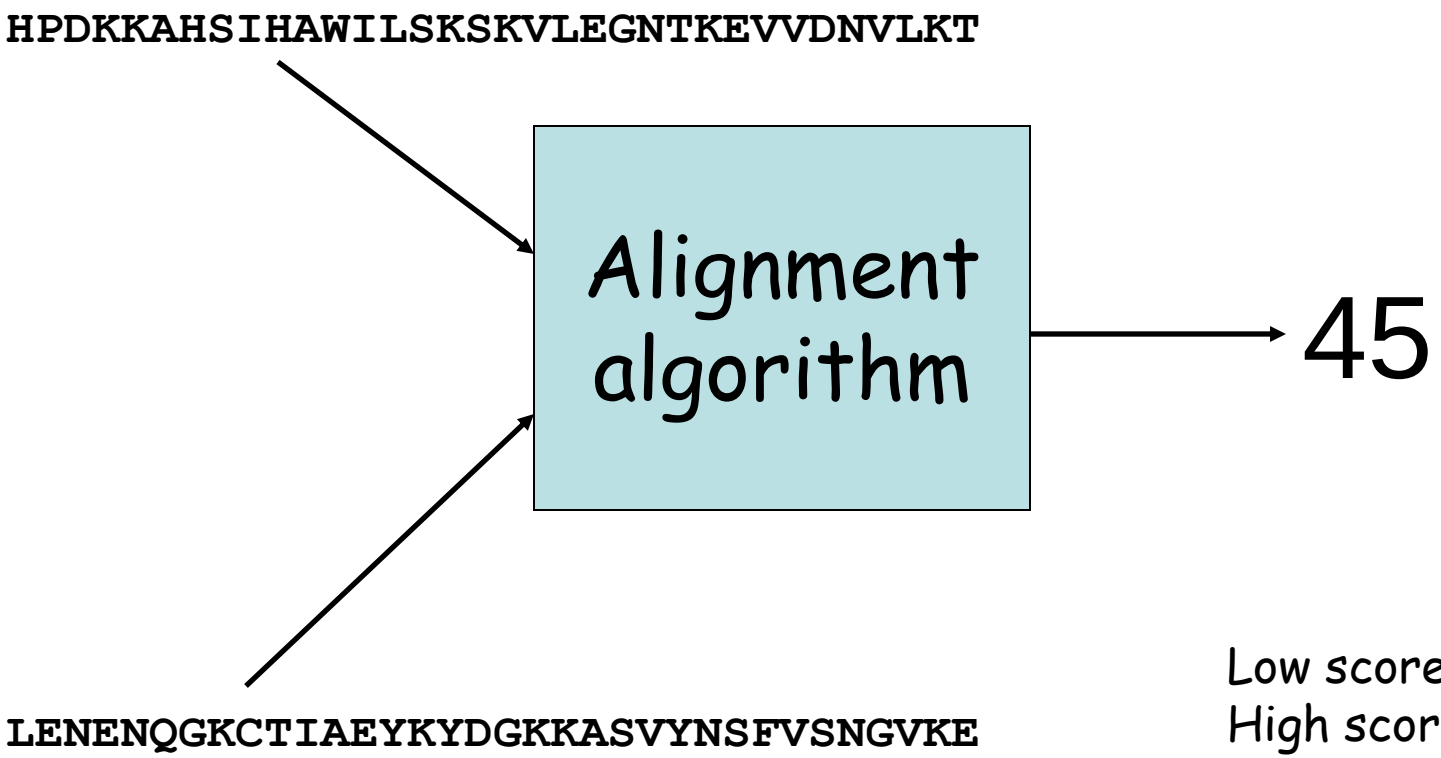
```
SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
           L P    W LDATYKNYA  Y C     L
SEQ 2: QFFPLMPPAPYWILDATYKNYALVYSCTTFFWLF
```

MAYBE (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 are interested in characterizing the distribution of scores from sequence comparisons.
- We measure how surprising a given score is, *assuming that the two sequences are not related.*
- The assumption is called the null hypothesis.

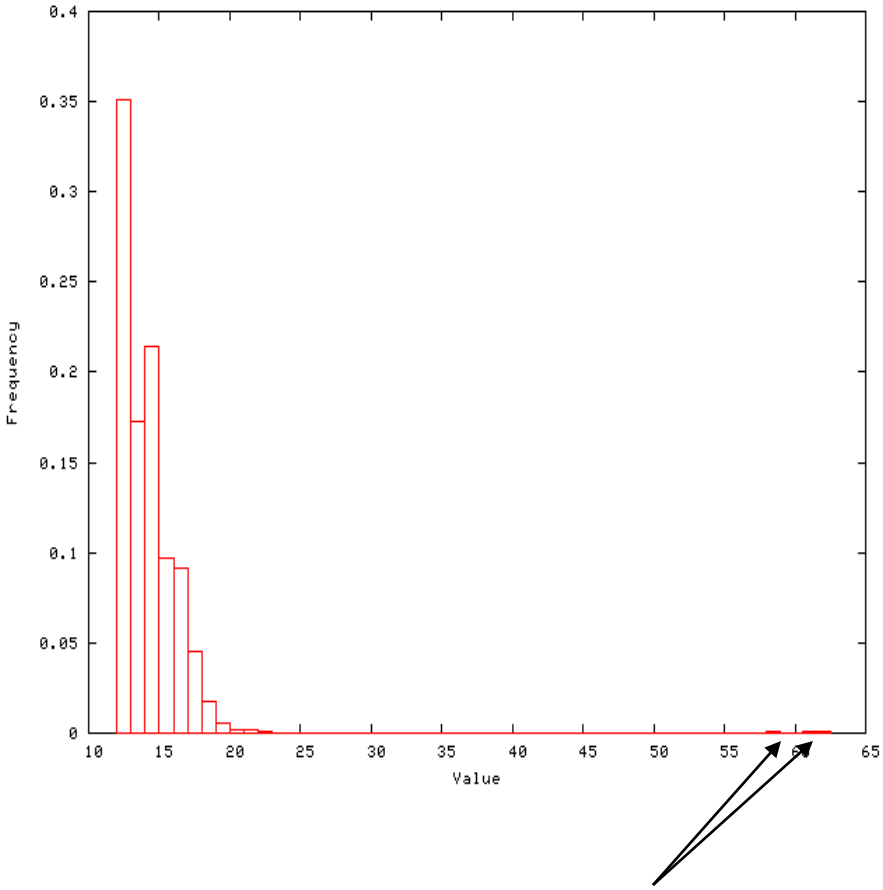
- The purpose of most statistical tests is to determine whether the observed results provide a reason to reject the hypothesis that they are merely a product of chance factors.

Sequence similarity score distribution

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

Empirical score distribution

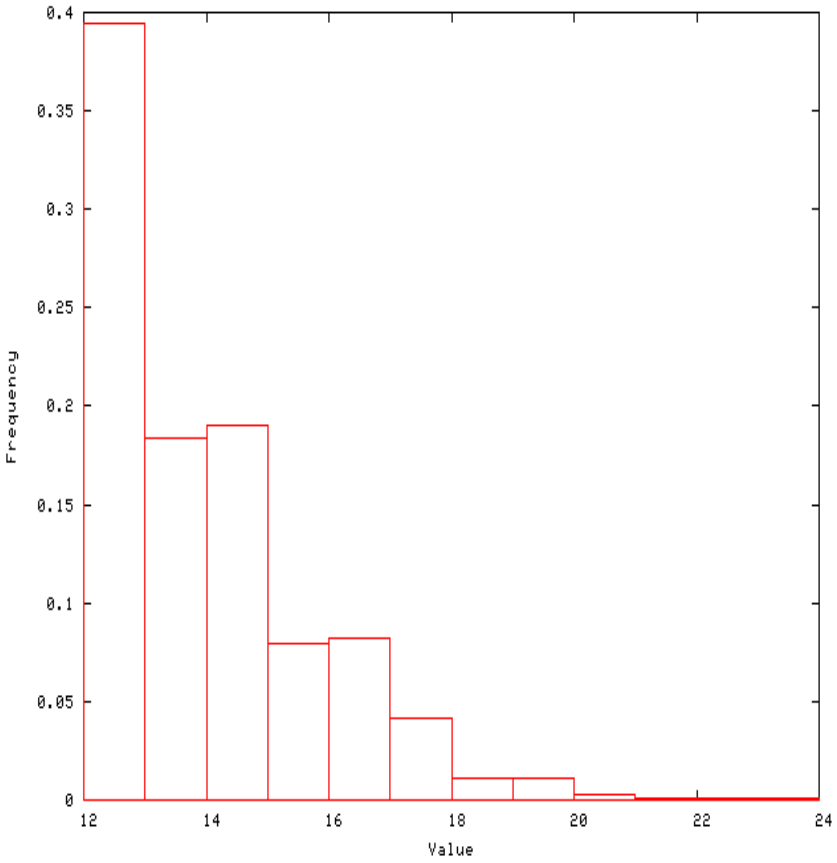
- This shows the distribution of scores from a real database search using BLAST.
- This distribution contains scores from unrelated and related pairs.



High scores from related sequences

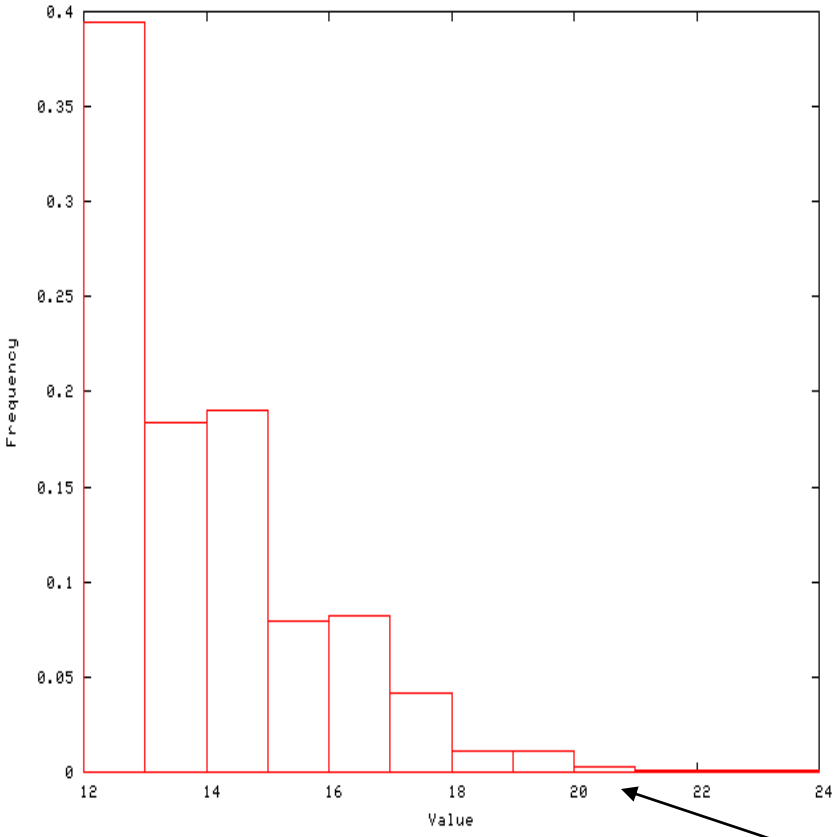
Empirical null score distribution

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



(notice the scale is shorter here)

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.

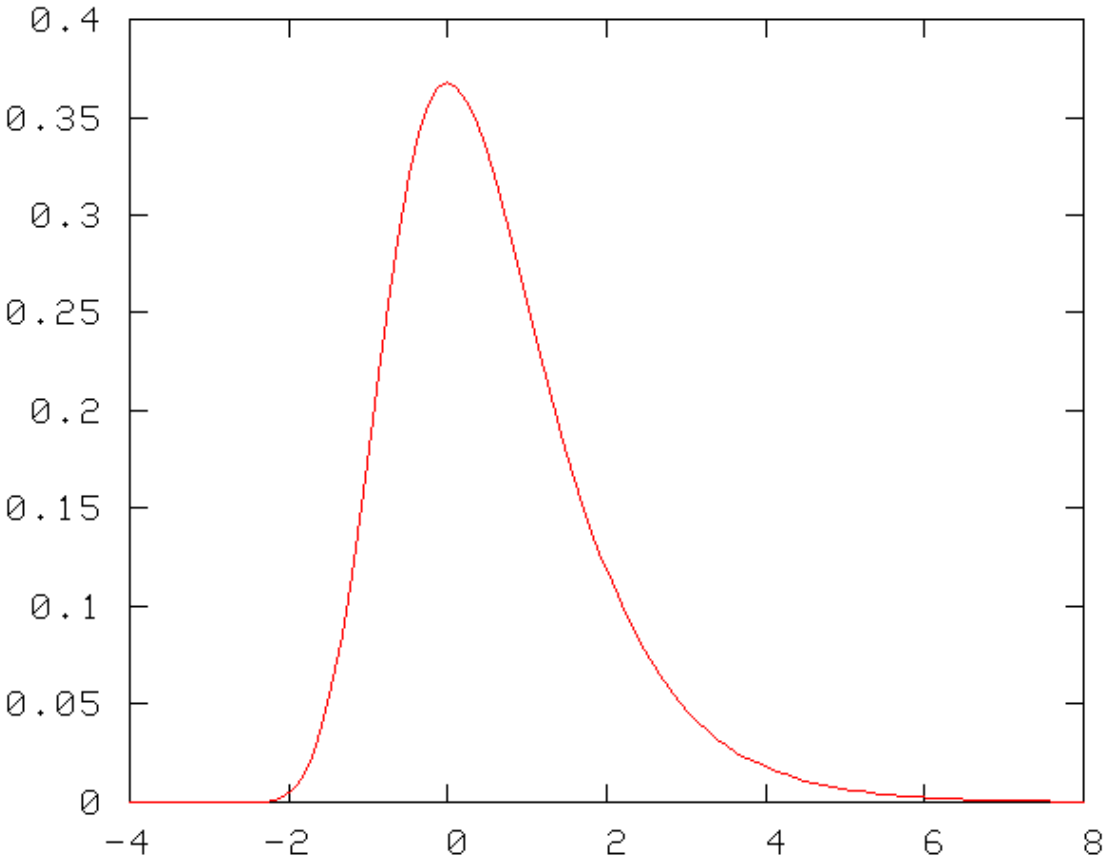
Problems with empirical distributions

- We are interested in very small probabilities.
- These are computed from the *tail* of the distribution.
- Estimating a distribution with an accurate tail is computationally very expensive.

A solution

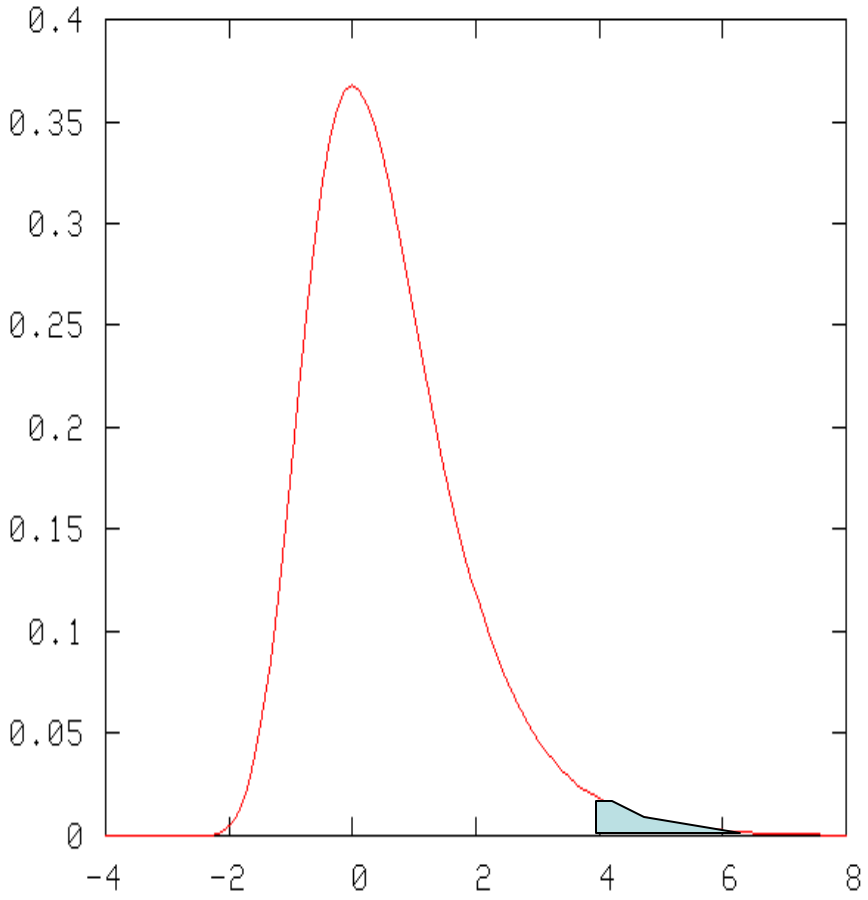
- Solution: Characterize the form of the distribution mathematically.
- Fit the parameters of the distribution empirically, or compute them analytically.
- Use the resulting distribution to compute accurate p-values.

Extreme value distribution



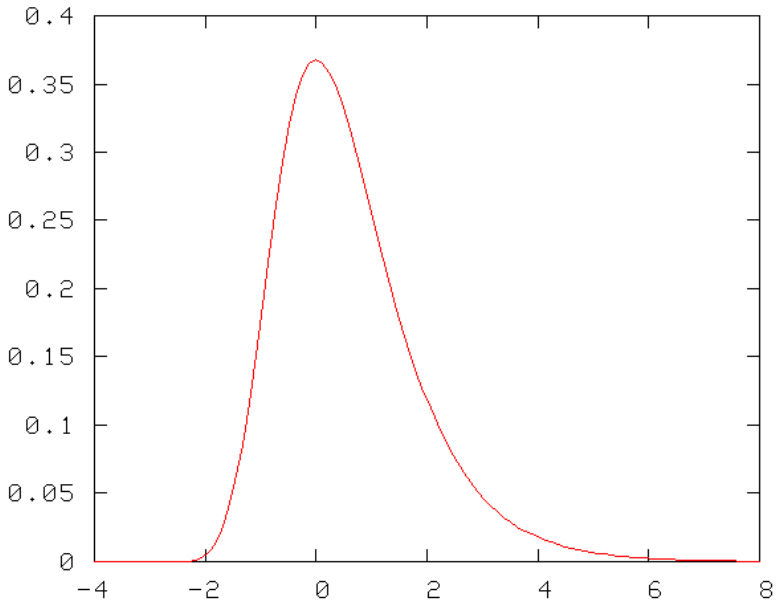
This distribution is roughly normal near the peak, but characterized by a larger tail on the right.

Computing a p-value



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

Extreme value distribution



Compute this value for x=4.

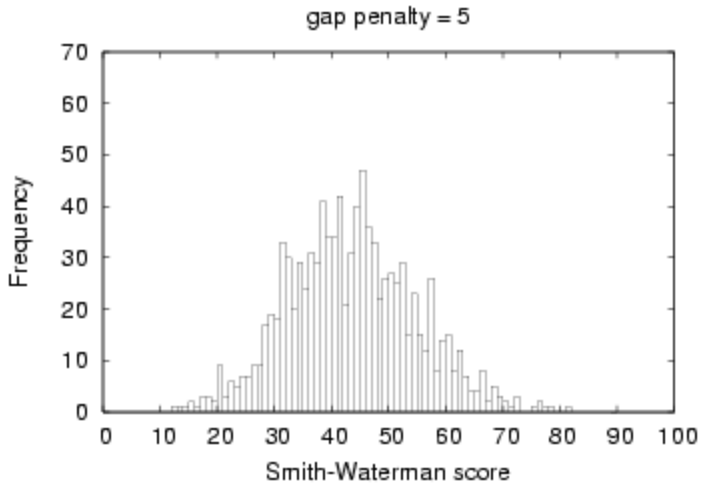
$$P\ S \geq x\ = 1 - e^{(-e^{-x})}$$

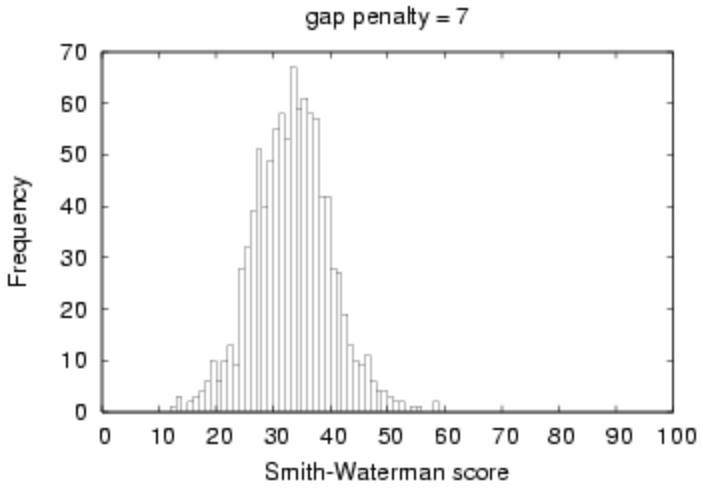
Computing a p-value

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

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

Scaling the EVD





- An EV distribution derived from, e.g., the Smith-Waterman algorithm with BLOSUM62 matrix has a characteristic mode μ and scale parameter λ.

$$P\ \ S \geq x\ \ = 1 - e^{(-e^{-x})}$$ scaled: $$P\ \ S \geq x\ \ = 1 - e^{(-e^{-\lambda(x-\mu)})}$$

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

$$
\begin{aligned}
P(S \geq 45) &= 1 - e^{\left(-e^{-0.693\,(45-25)}\right)} \\
&= 1 - e^{\left(-e^{-13.86}\right)} \\
&= 1 - e^{-9.565\times10^{-7}} \\
&= 1 - 0.999999043 \\
&= 9.565\times10^{-7}
\end{aligned}
$$

BLAST has precomputed values of μ and λ for all common matrices and gap penalties (and the run scales them for the size of the query and database)

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

Multiple testing

- Say that you perform a statistical test with a 0.05 threshold, but you repeat the test on twenty different observations (e.g. 20 different blast runs)

- 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?

Bonferroni correction

- Assume that individual tests are *independent*.
- Divide the desired p-value threshold by the number of tests performed.

Database searching

- Say that you search the non-redundant protein database at NCBI, containing roughly one million sequences (i.e. you are doing 10^6 pairwise tests). 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 $0.001 / 10^6 = 10^{-9}$.

E-values

- A p-value is the probability of making a mistake.
- An 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 1,000,000 = 1,000.

(BLAST actually calculates E-values in a more complex way, but they mean the same thing)

Search

```
>104K_THEPA 104 KD MICRONEME-RHOPTRY ANTIGEN
MKFLILLFNILCLFPVLAADNHGVGPQGASGVDPITFDINSNQTGPAFLTAVEMAGVKYLQ
HRLVEGNVVIWENASTPLYTGAIVTNNDGPYMAYVEVLGDPNLQFFIKSGDAWVTLSEHEY
AVHIESVFSLNMAFQLENNKYEVETHAKNGANMVTFIPRNGHICKMVYHKNVRIYKATGND
RGLRLLLINVFSIDDNGMMSNRYFQHVDDKYVPISQKNYETGIVKLKDYKHAYHPVDLDIK
```

Set subsequence From: To:

Choose database nr

Do CD-Search

Now: BLAST! or Reset query Reset all

Sequences producing significant alignments:	Score (bits)	E Value
gi|112670|sp|P15711|104K_THEPA 104 KD MICRONEME-RHOPTRY ANT...	1352	0.0
gi|14268530|gb|AAK56556.1| 104 kDa microneme-rhoptry antige...	243	1e-62
gi|14268528|gb|AAK56555.1| 104 kDa microneme-rhoptry antige...	242	4e-62
gi|14268526|gb|AAK56554.1| 104 kDa microneme-rhoptry antige...	238	7e-61
gi|31210185|ref|XP_314059.1| ENSANGP00000015608 [Anopheles ...	37	2.1
gi|22971724|ref|ZP_00018655.1| hypothetical protein [Chloro...	35	9.7
gi|32403566|ref|XP_322396.1| hypothetical protein [Neurospo...	35	12
gi|24639766|ref|NP_572189.1| CG2861-PA [Drosophila melanoga...	34	17
gi|30348569|emb|CAC84361.1| hypothetical protein [Saimiriin...	34	19
gi|6492132|gb|AAF14193.1| spherical body protein 3 [Babesia...	34	20
gi|9629342|ref|NP_044542.1| virion protein [Human herpesvir...	34	21
gi|24639768|ref|NP_726958.1| CG2861-PB [Drosophila melanoga...	34	21
gi|4757118|emb|CAB42096.1| TashAT2 protein [Theileria annul...	34	22
gi|17534529|ref|NP_495288.1| putative protein (2G676) [Caen...	34	22
gi|15241089|ref|NP_195809.1| leucine-rich repeat transmembr...	33	23
gi|43489677|gb|EAD99646.1| unknown [environmental sequence]	33	23
gi|44419062|gb|EAJ13596.1| unknown [environmental sequence]	33	25
gi|43969222|gb|EAG41329.1| unknown [environmental sequence]	33	29
gi|15792145|ref|NP_281968.1| putative oxidoreductase [Campy...	33	34
gi|43926327|gb|EAG18073.1| unknown [environmental sequence]	33	37
gi|39595869|emb|CAE67372.1| Hypothetical protein CBG12848 [...	33	38
gi|30020082|ref|NP_831713.1| Glycosyltransferase [Bacillus ...	33	40
gi|43723946|gb|EAF16931.1| unknown [environmental sequence]	33	41
gi|11545212|gb|AAG37800.1| hypothetical telomeric SfiI frag...	33	44
gi|40788024|emb|CAE47751.1| ubiquitin specific proteinase 5...	32	51
gi|42656951|ref|XP_052597.6| ubiquitin specific protease 53...	32	51
gi|32698642|ref|NP_872557.1| DNA-ligase [Adoxophyes orana g...	32	52
gi|12840300|dbj|BAB24814.1| unnamed protein product [Mus mu...	32	54
gi|28899333|ref|NP_798938.1| 4-diphosphocytidyl-2C-methyl-D...	32	55
gi|7243081|dbj|BAA92588.1| KIAA1350 protein [Homo sapiens]	32	62

Summary

- A distribution plots the frequencies of types 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 long 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.