

Sequence Comparison: Significance of similarity scores

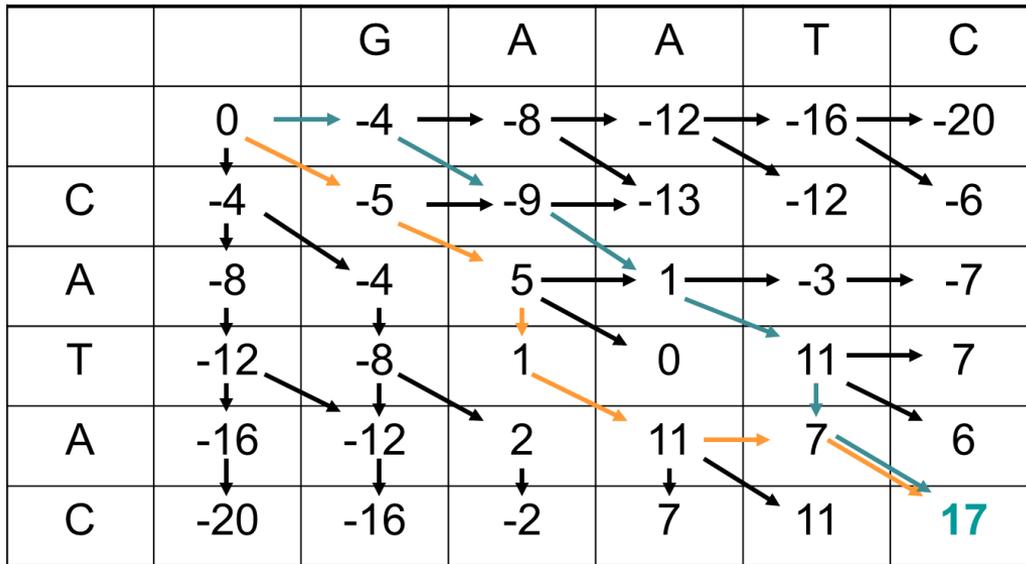
Genome 373
Genomic Informatics
Elhanan Borenstein

Summary

Global alignment algorithm:

Needleman-Wunsch.

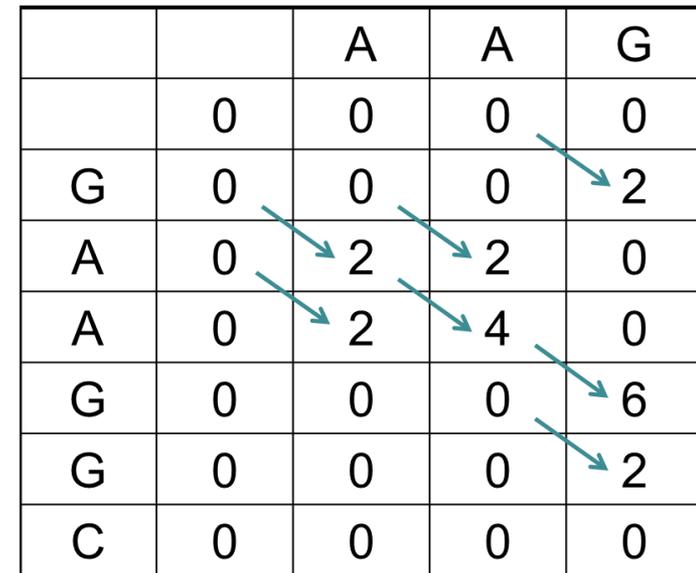
		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5	-9	-13	-12	-6
A	-8	-4	5	1	-3	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	17



Local alignment algorithm:

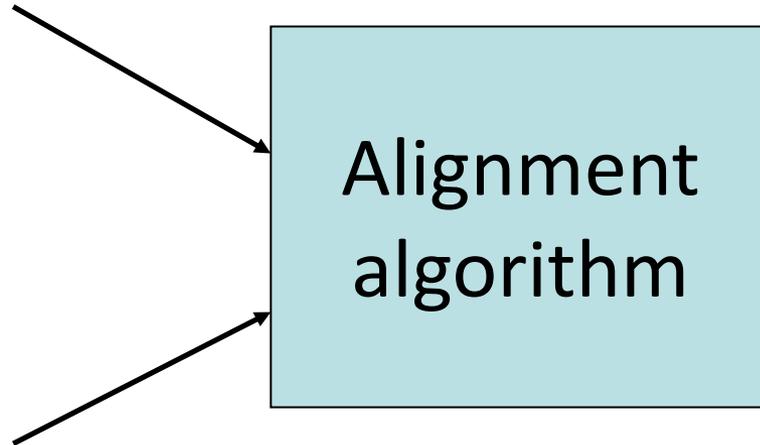
Smith-Waterman.

		A	A	G
	0	0	0	0
G	0	0	0	2
A	0	2	2	0
A	0	2	4	0
G	0	0	0	6
G	0	0	0	2
C	0	0	0	0



Significance of scores

HPDKKAHSIHAWILSKSKVLEGNTKEVVDNVLKT



45

Low score = unrelated
High score = related

LENENQGKCTIAEYKYDGKKASVYNSFVSNGVKE

But ... how high is high enough?

Subjective

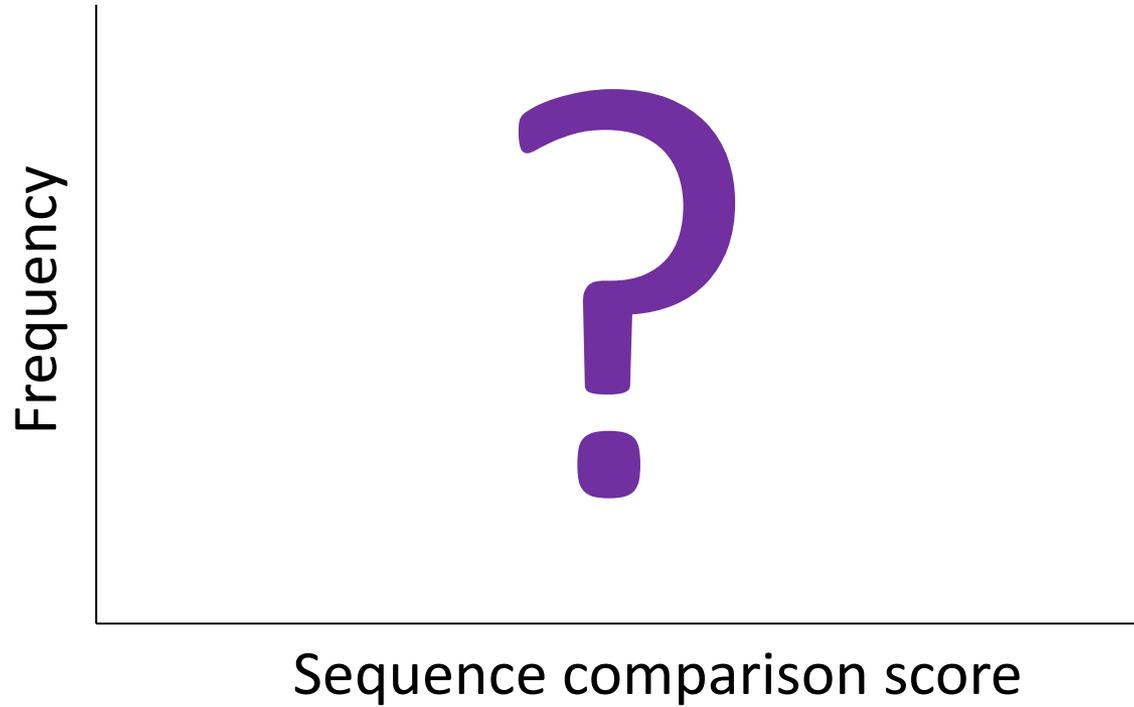
Problem specific

Parameter specific

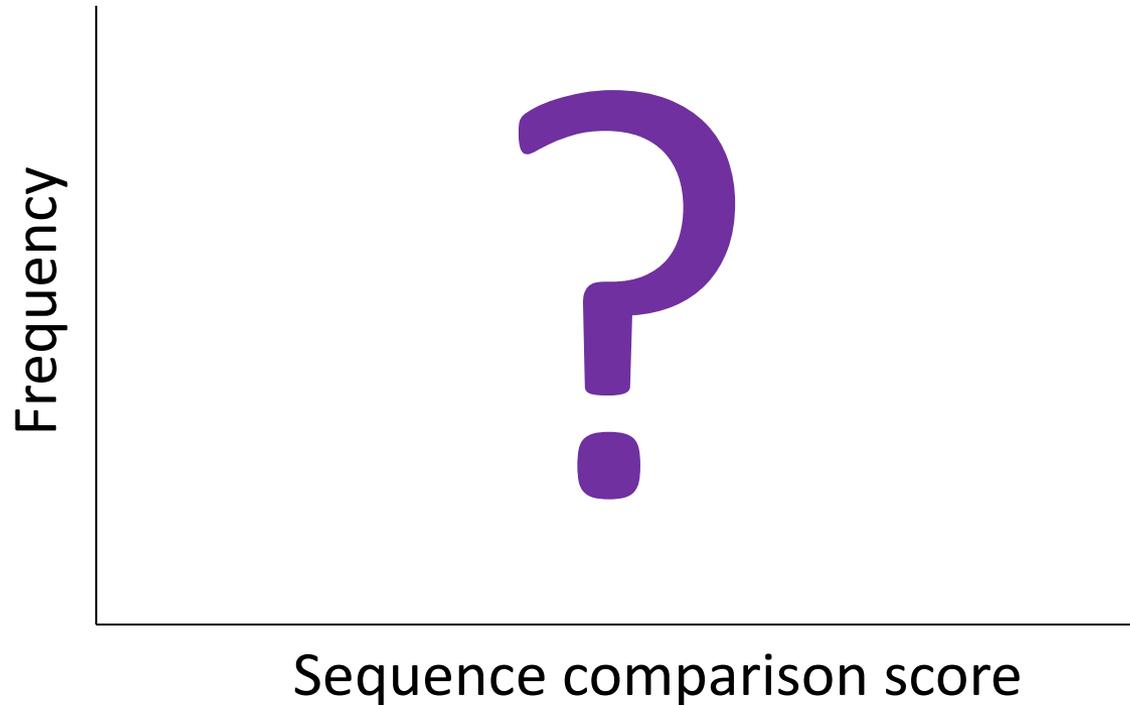
The null hypothesis

- We want to know 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.
- We want to characterize the distribution of scores from pairwise sequence alignments

Sequence similarity score distribution



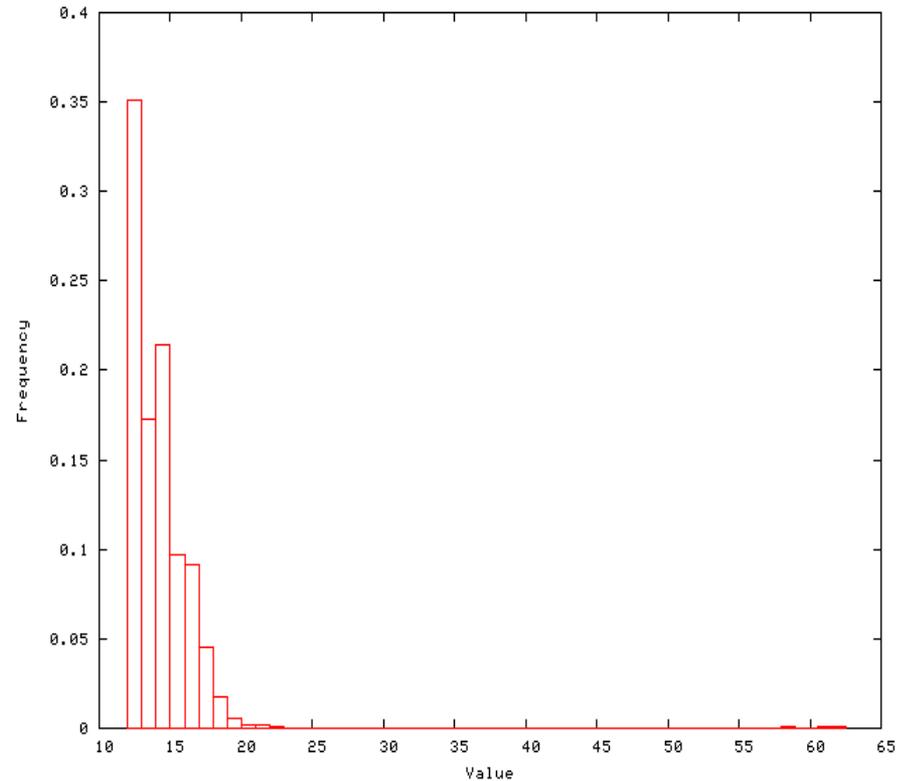
Sequence similarity score distribution



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

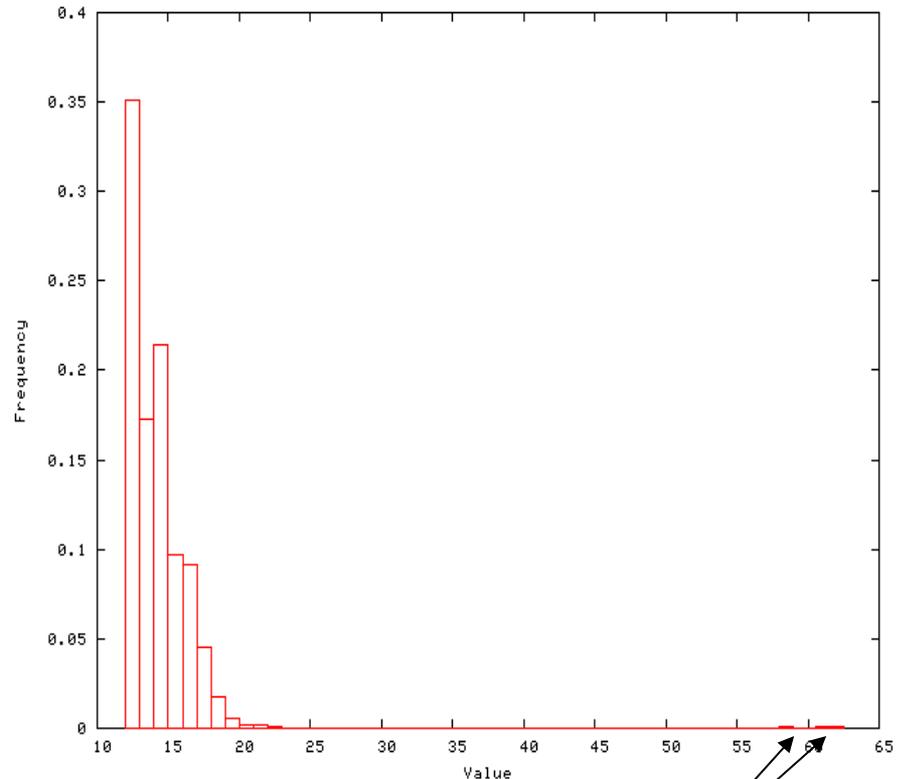
Empirical score distribution

- This shows the distribution of scores from a **real** database search using BLAST.



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.

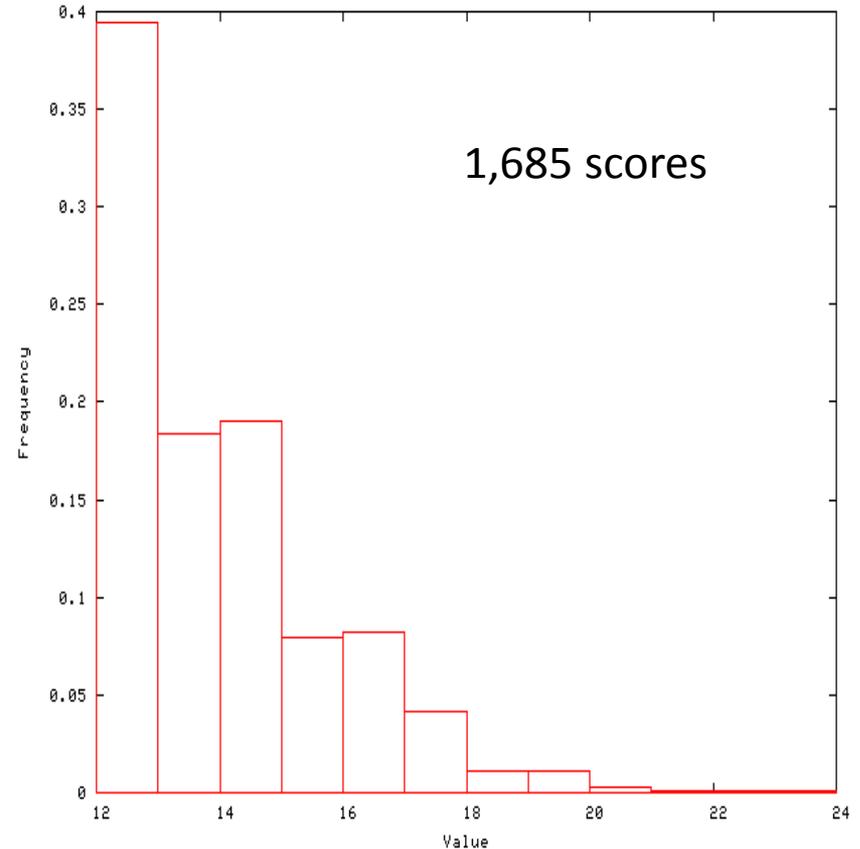


High scores from related sequences

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

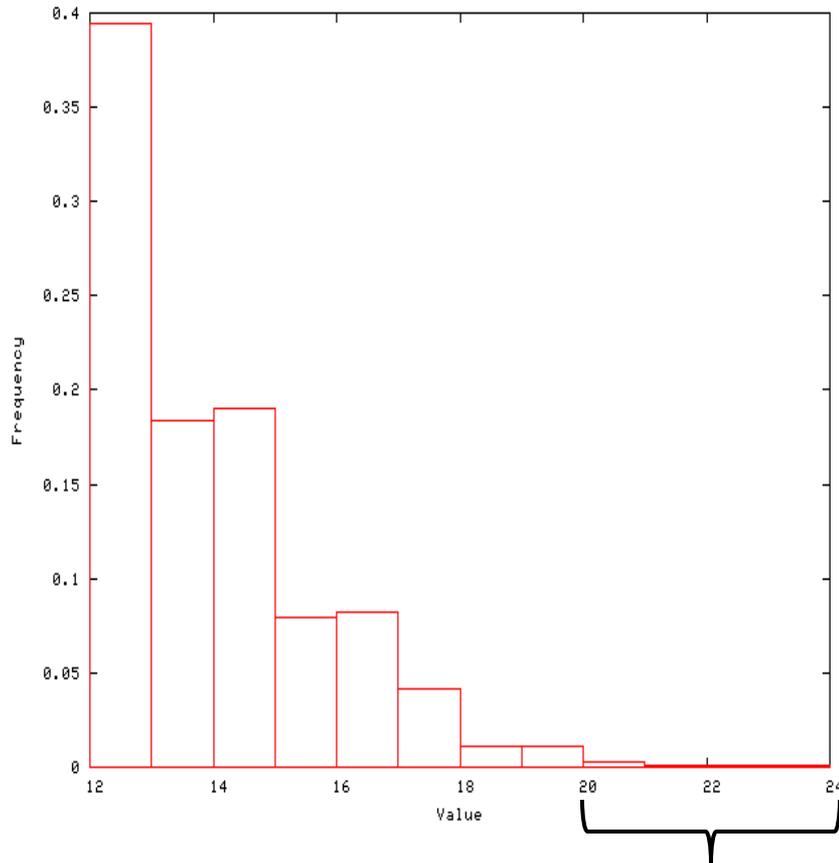
Empirical null score distribution

- The distribution of scores obtained from aligning a given sequence to a database of **randomized** sequences (e.g., each sequence was shuffled)



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

Computing an empirical p-value



- The probability of observing a score $\geq X$ is the area under the curve to the right of X .
- This probability is called a p-value.
- **p-value = $\Pr(\text{data} | \text{null})$**
(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 $\sim 28/1685 = 0.0166$.

Problems with empirical distributions

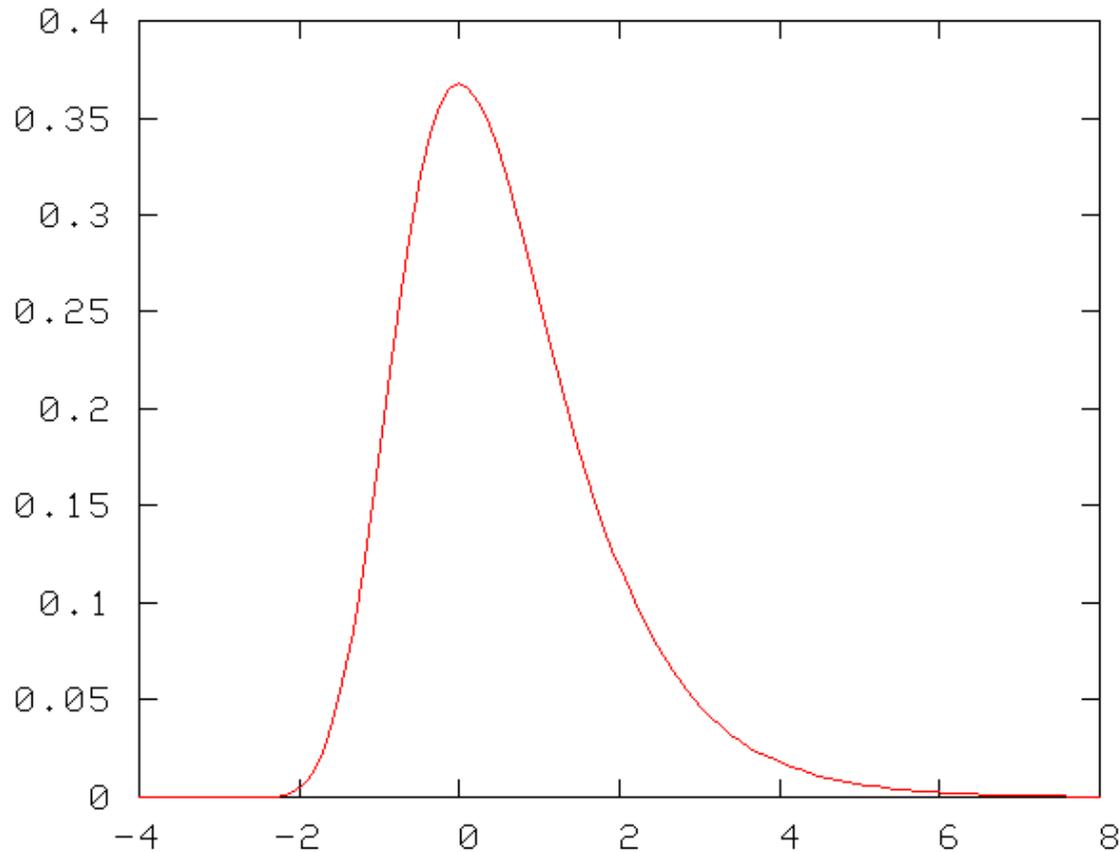
- We are interested in very small probabilities.
- These are computed from the *tail* of the null distribution.
- Estimating a distribution with an accurate tail is feasible but computationally very expensive because we have to make a very large number of alignments.

A solution

- 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)

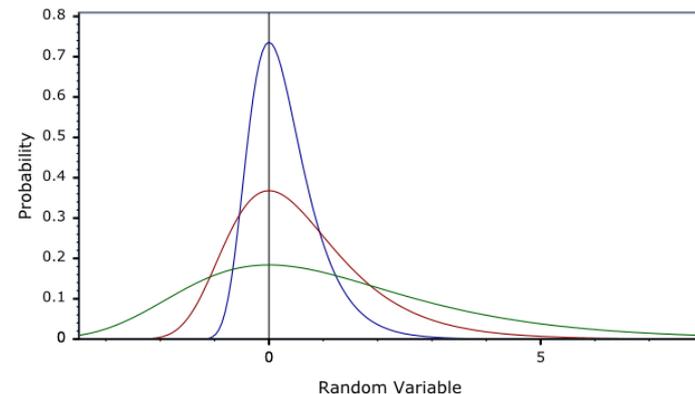
Extreme value distribution



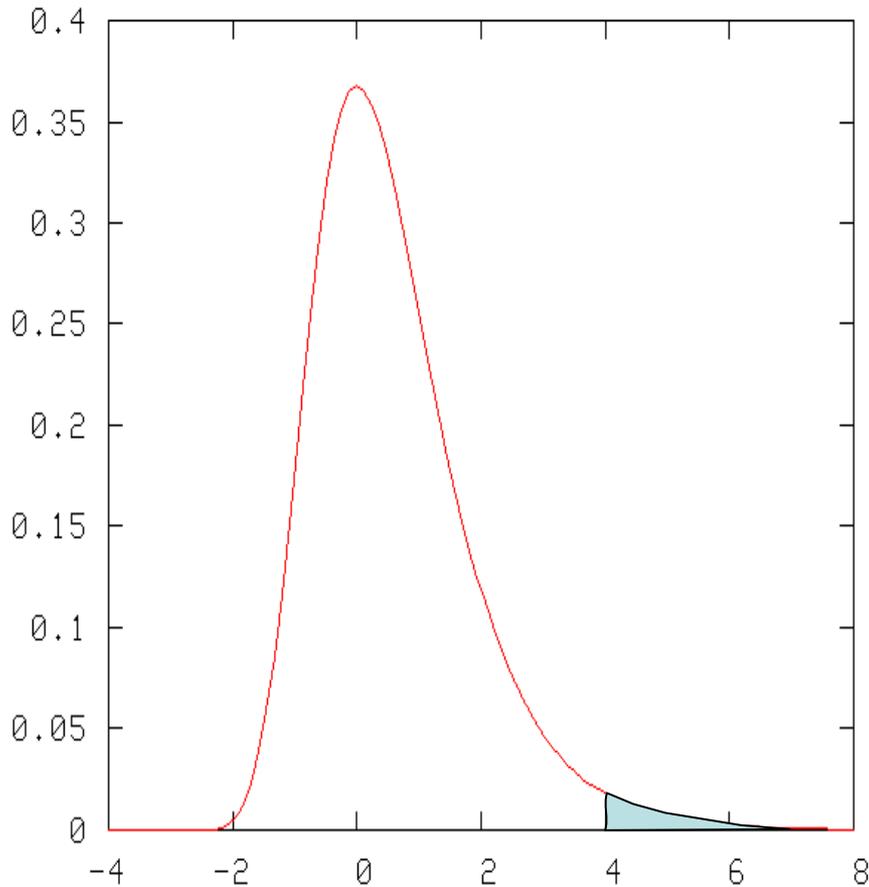
This distribution is roughly normal near the peak, but characterized by a larger tail on the right.

- For an Unscaled EVD:

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



Computing a p-value



- The probability of observing a score ≥ 4 is the area under the curve to the right of 4.

- For an *Unscaled EVD*:

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

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

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

What p-value is significant?

What p-value is significant?

- The most common thresholds are 0.01 and 0.05.
- A threshold of 0.05 means that even if the null hypothesis is correct you will still get such score (or higher) in 5% of cases.
- Is 0.05? It depends upon the cost associated with making a mistake.
- Examples of costs:
 - Doing extensive wet lab validation (expensive)
 - Making clinical treatment decisions (very expensive)
 - Misleading the scientific community (very expensive)
 - Doing further simple computational tests (cheap)
 - Telling your grandmother (very cheap)

Multiple testing

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 alignments)
- 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 < 0.05?

$$1 - 0.95^{20} = 0.6415$$

Bonferroni correction

- Assume that individual tests are *independent*.
- Divide the desired p-value threshold by the number of tests performed.
- In the example about, a Bonferroni correction would suggest using a p-value threshold of $0.05 / 20 = 0.0025$.

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).
- and ... you want to use a p-value of 0.01.
- Recall that you would observe such a p-value by chance approximately every 100 times in a random database.
- That is, without correcting for multiple testing you will get **~10,000 false positives!!!**
- A Bonferroni correction would suggest using a p-value threshold of $0.01 / 10^6 = 10^{-8}$.

E-values

- 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 0.01 and a database of 1,000,000 sequences, the corresponding E-value is $0.01 \times 1,000,000 = 100$.

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

Score E
(bits) Value

Sequences producing significant alignments:

Accession	Description	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

Take home message

- A [distribution](#) plots the frequencies of types of observation.
- The area under the distribution curve is 1.
- Most statistical tests compare observed data to the expected result according to a [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 a given score would appear in a random database of the given size.

