Sequence Comparison: Significance of similarity scores

Genome 373

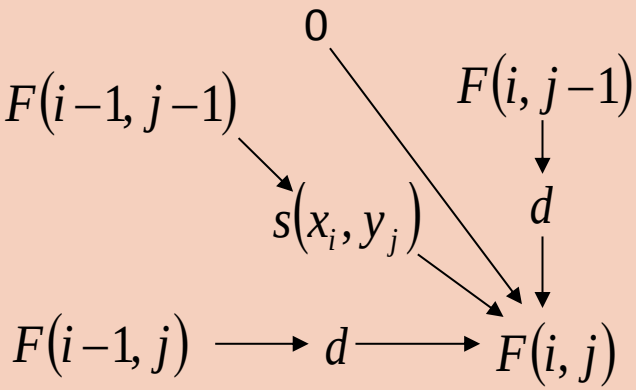
Genomic Informatics

Elhanan Borenstein

Quick review: Local alignment

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5



Find the optimal **local** alignment of AAG and GAAGGC.

Use a gap penalty of d = -5.

		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

Summary

Global alignment algorithm:

Needleman-Wunsch.

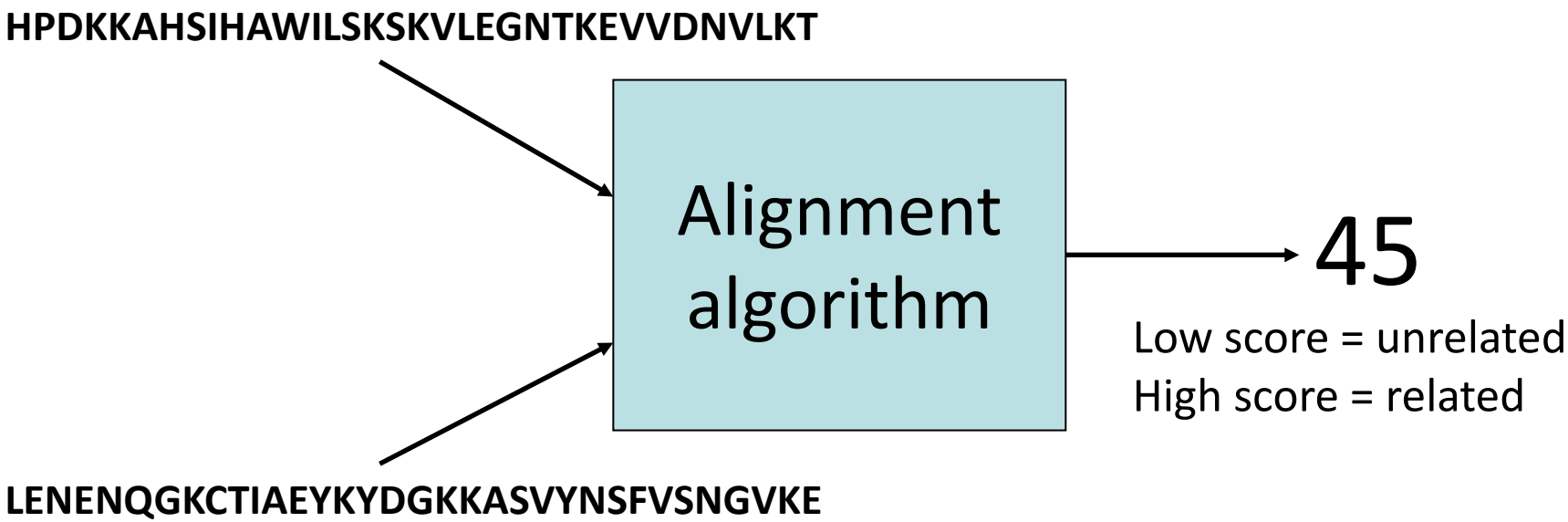
		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5	-9	-13	-12	-6
A	-8	-4	5	1	-3	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	**17**

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



But … how high is high enough?

Subjective

Problem specific

Parameter specific

A statistical framework for interpreting sequence alignment scores

- The p-value is the probability that our hypothesis is false
- The p-value is the probability that the observed effects were produced by random chance
- P-value < 0.05 is significant
- The p-value indicates the size of the observed effect

Common misconceptions

- The p-value is the probability that our hypothesis is false
- The p-value is the probability that the observed effects were produced by random chance
- P-value < 0.05 is significant
- The p-value indicates the size of the observed effect

P Values Under Fire

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

Q: Why do so many colleges and grad schools teach $p = 0.05$?
A: Because that's still what the scientific community and journal editors use.
Q: Why do so many people still use $p = 0.05$?
A: Because that's what they were taught in college or grad school.

2014) and a statement on risk-limiting post-election audits (American Statistical Association 2010). However, these were truly policy-related statements. The VAM statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of VAMs as effective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific

SOCIAL SELECTION *Popular articles on social media*

Psychology journal bans *P* values

A controversial statistical test has met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values, because the values were too often used to support lower-quality research.

Authors are still free to submit papers to *BASP* with *P* values and other statistical measures that form part of 'null hypothesis significance testing' (NHST), but the numbers will be removed before publication. "Basic and Applied Social Psychology just went science rogue and banned NHST from their journal. Awesome," tweeted Nerisa Dozo, a PhD student in psychology at the University of Queensland in Brisbane, Australia. But Jan de Ruiter, a cognitive scientist at Bielefeld University in Germany, tweeted: "NHST is really problematic", adding that banning all inferential statistics is "throwing away the baby with the p-value".

Basic Appl. Soc. Psych. 37, **1–2 (2015)**

Based on data from altmetric.com. Altmetric is supported by Macmillan Science and Education, which owns Nature Publishing Group.

NATURE.COM
For more on popular papers:
go.nature.com/ynfi49

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Au

News & Comment > News > 2018 > April > Article

NATURE | NEWS

'One-size-fits-all' threshold for *P* values under fire

Scientists hit back at a proposal to make it tougher to call findings statistically significant.

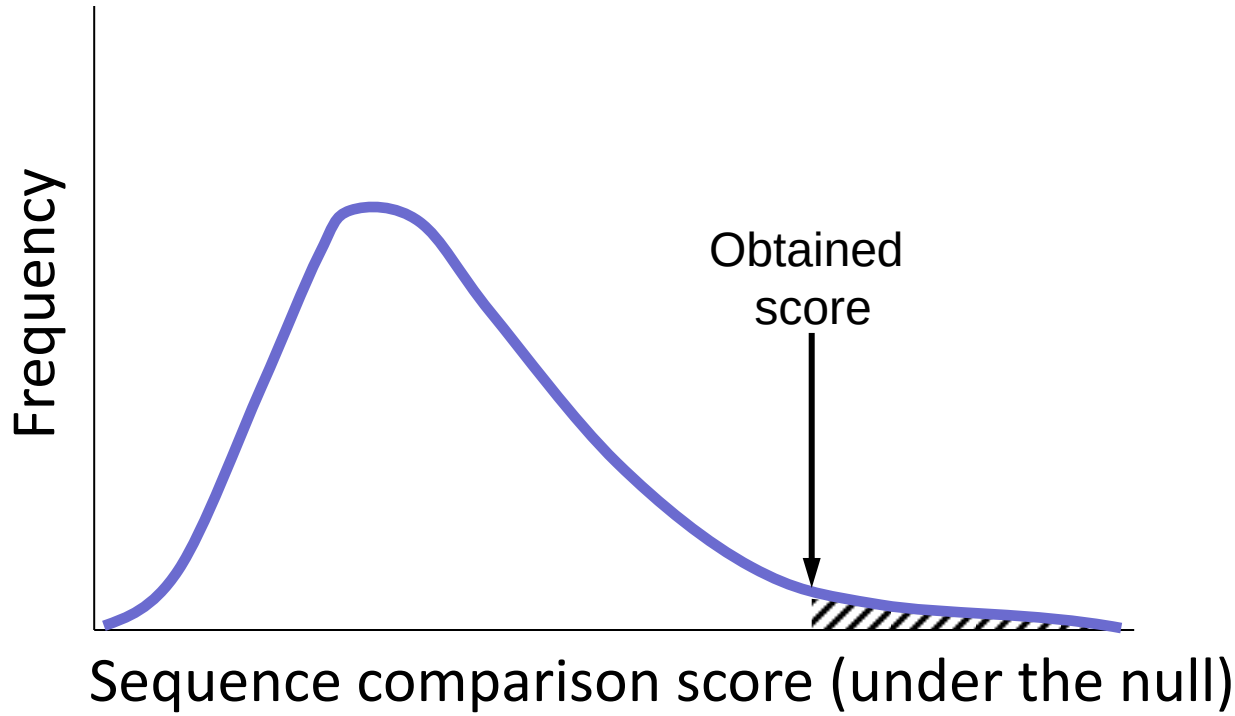
Dalmeet Singh Chawla

19 September 2017

Statistical hypothesis testing

- We want to know how **surprising** a given score is, …
  assuming that the two sequences are <u>not</u> 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.
- Put differently, we want to determine how likely is it to obtain a specific score (or higher) under the null hypothesis.

P-values as a representation of surprise



- 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)

Sequence similarity score distribution

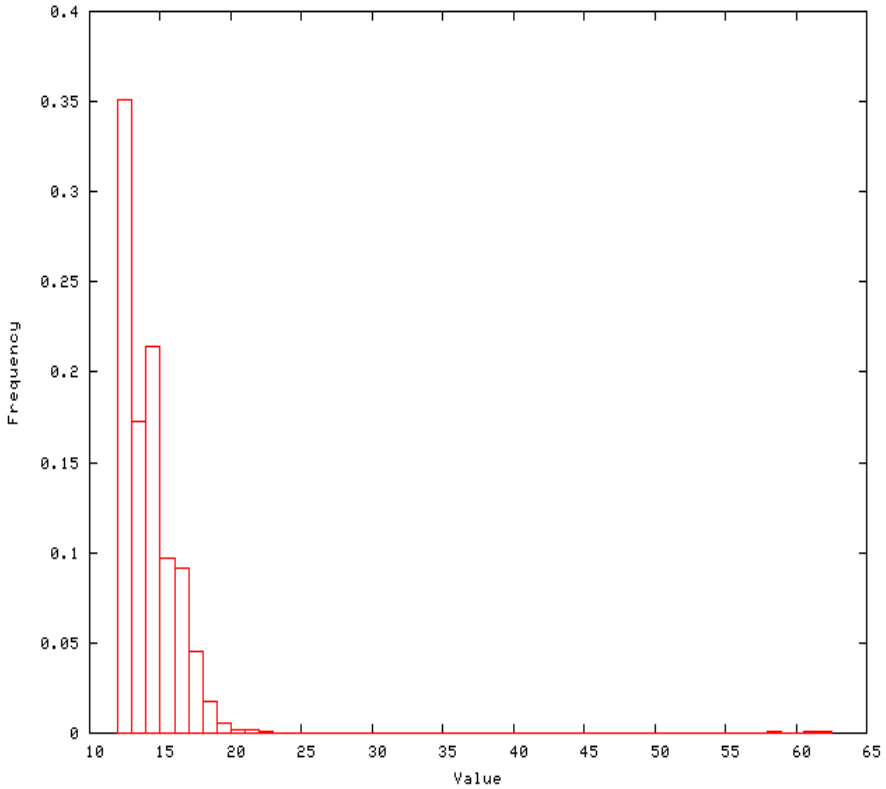
Approach 1:

Search a database of **unrelated** sequences using a given query sequence

(Empirical null score distribution)

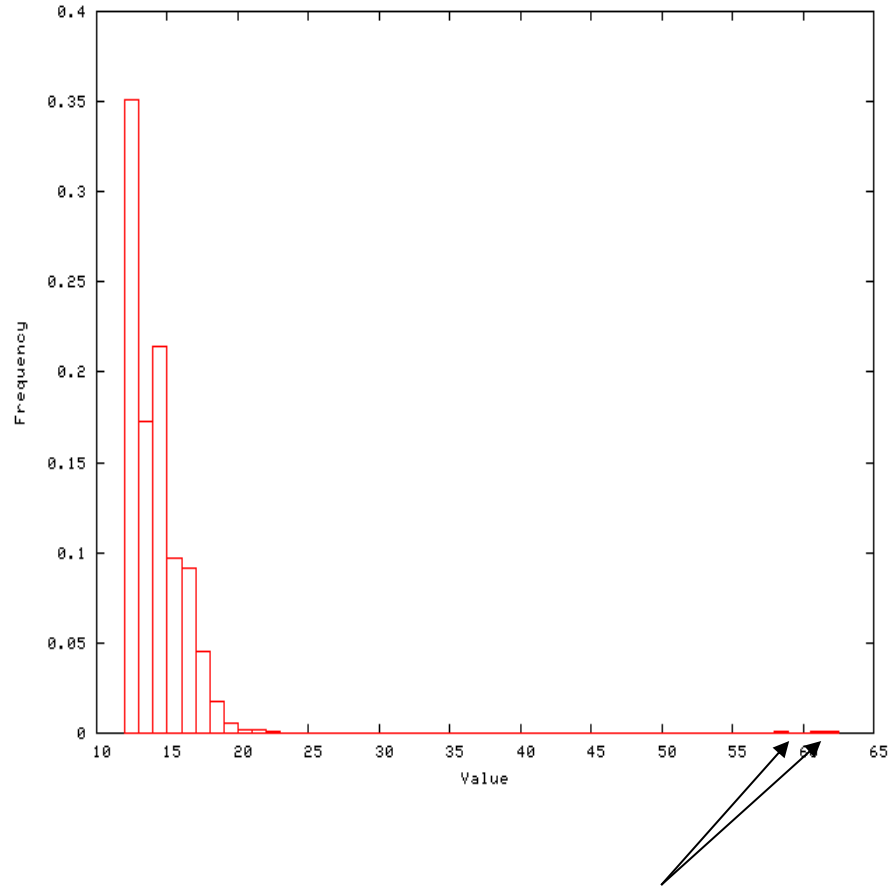
Empirical null score distribution

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



Empirical null score distribution

- This shows the distribution of scores from a real database search using BLAST.
- **<u>Problem</u>**: This distribution contains scores many unrelated sequences (but also from a few related sequences).



High scores from related sequences

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

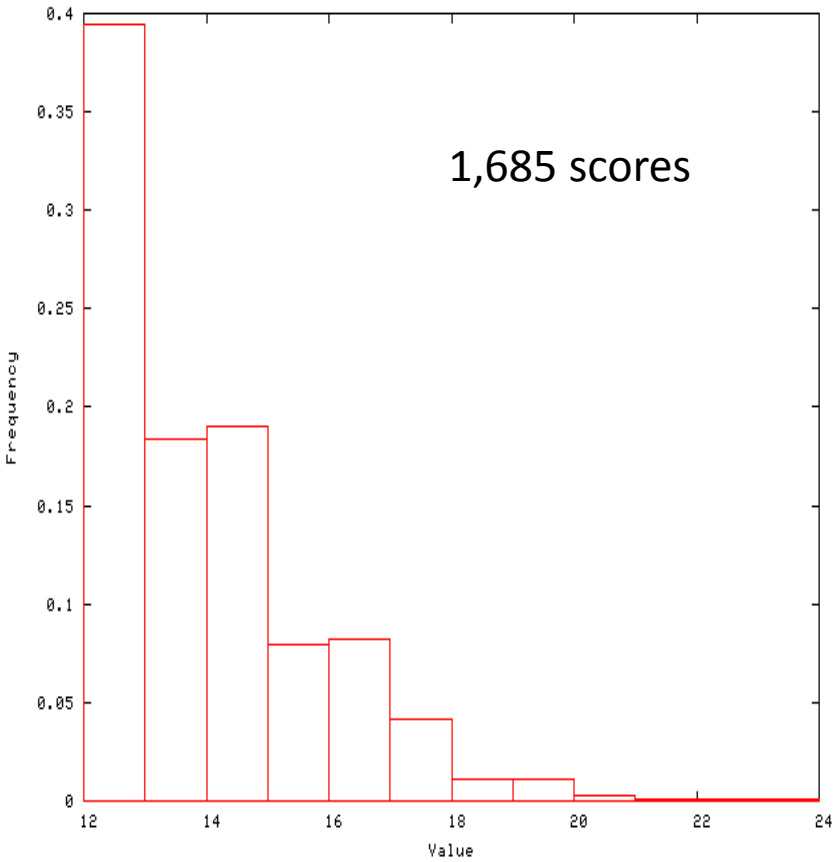
Approach 2:

Search a database of **random** sequences using a given query sequence

(Empirical null score distribution)

Empirical null score distribution

- The distribution of scores obtained from aligning a given sequence to a database of random sequences

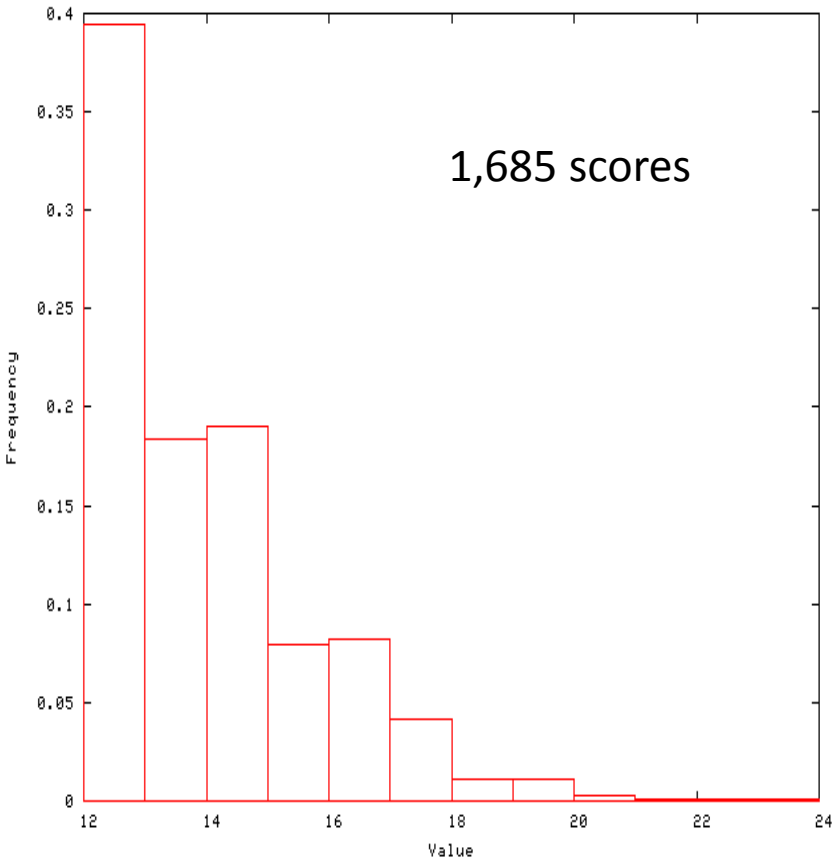


(note - there are lots of lower scoring alignments not reported)

Empirical null score distribution

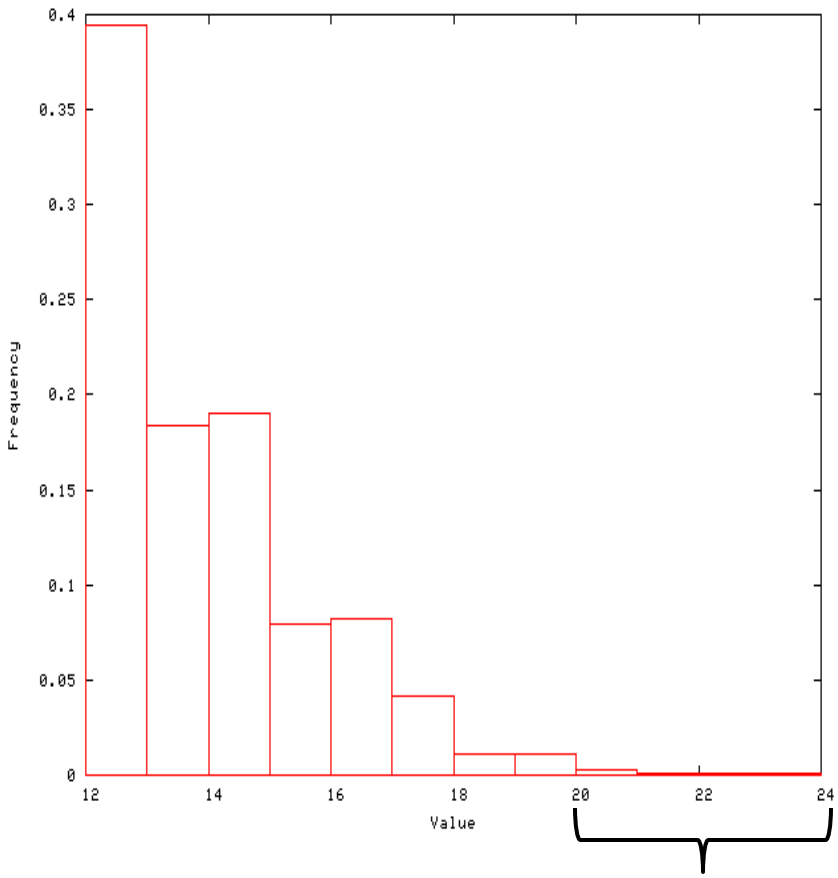
- The distribution of scores obtained from aligning a given sequence to a database of random sequences

- **Challenge: How will we generate a database of random sequences??**



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

Computing an empirical p-value



- P-value = The probability of observing a score >=X is the area under the curve to the right of X.

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 ~28/1685 = 0.0166.

Problems with empirical distributions

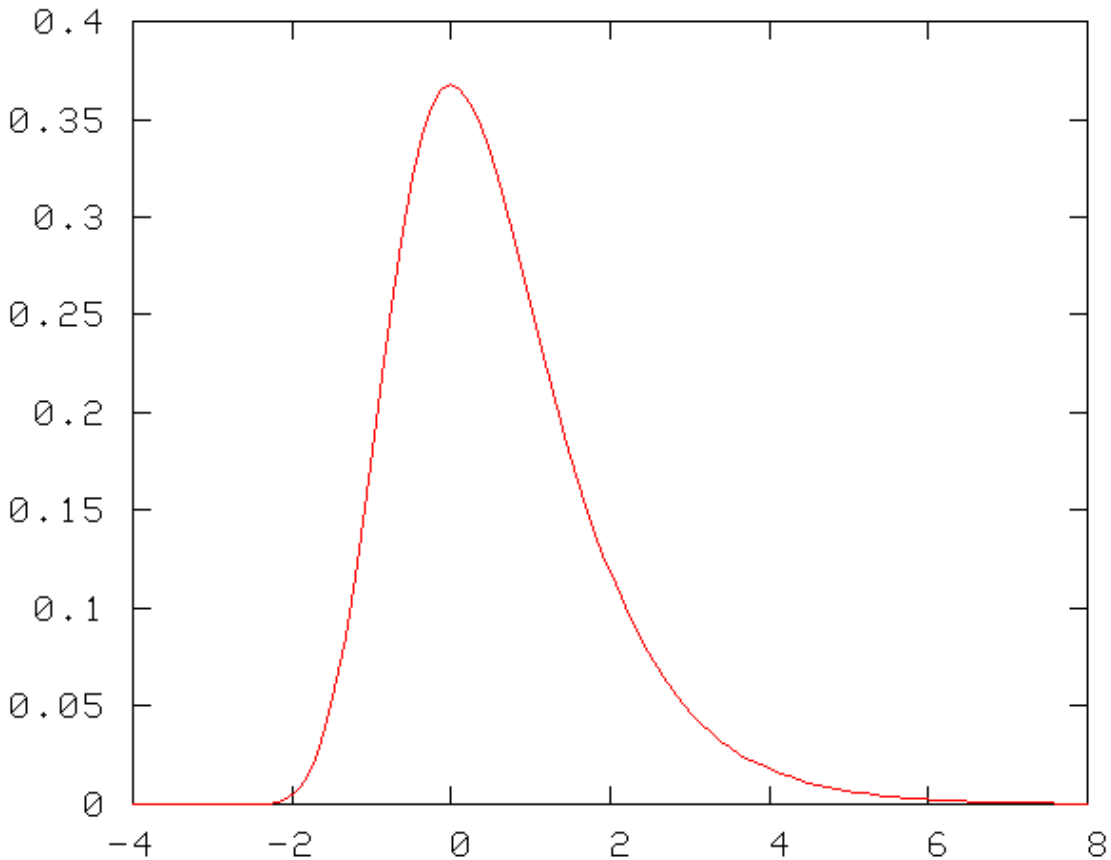
- We are interested in very small probabilities.
- These are computed from the *tail* of the null distribution.
- Estimating a distribution with an accurate tail is feasible but computationally very expensive because we have to make a very large number of alignments.

Approach 3:

- 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)

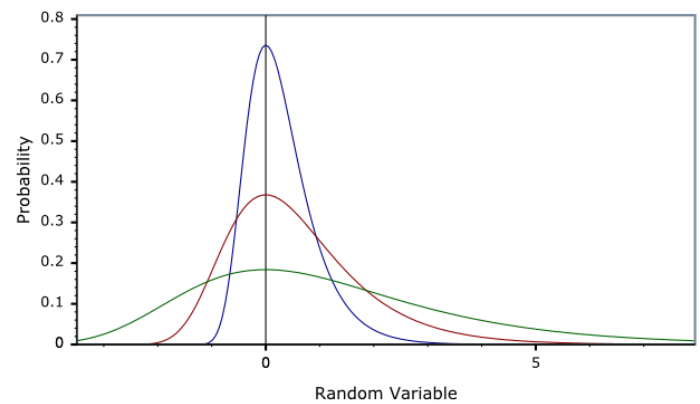
Extreme value distribution



- <u>For an *Unscaled EVD:*</u>

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

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



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.
- Why 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 you align your sequence to a candidate gene …
- And assume that the null hypothesis is correct (i.e., your sequence is not related to this gene)
- What is the chance that you get a p-value < 0.05?

Multiple testing

- Now, say you align your sequence to 20 different candidate genes …
- And still assume that the null hypothesis is correct (i.e., your sequence is not related to this gene)
- What is the chance that at least one of these tests will get a p-value < 0.05?

Multiple testing

- Now, say you align your sequence to 20 different candidate genes …
- And still assume that the null hypothesis is correct (i.e., your sequence is not related to this gene)
- What is the chance that at least one of these tests will get 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 × 1,000,000 = 10,000.

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

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

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.