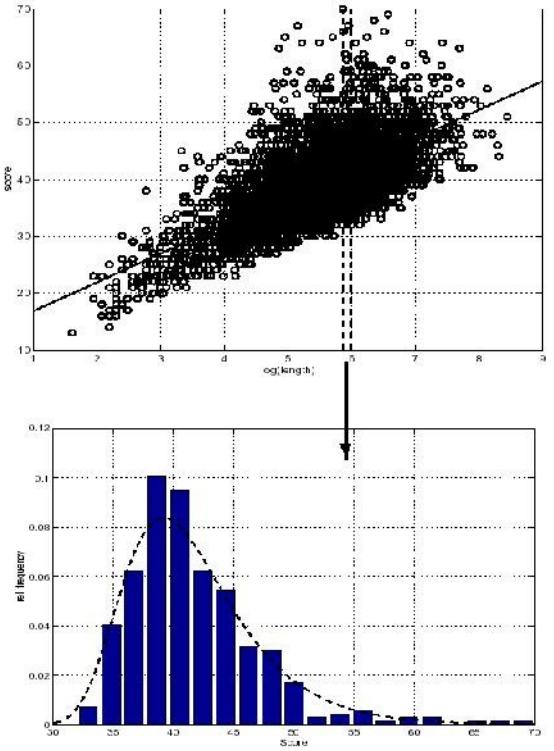
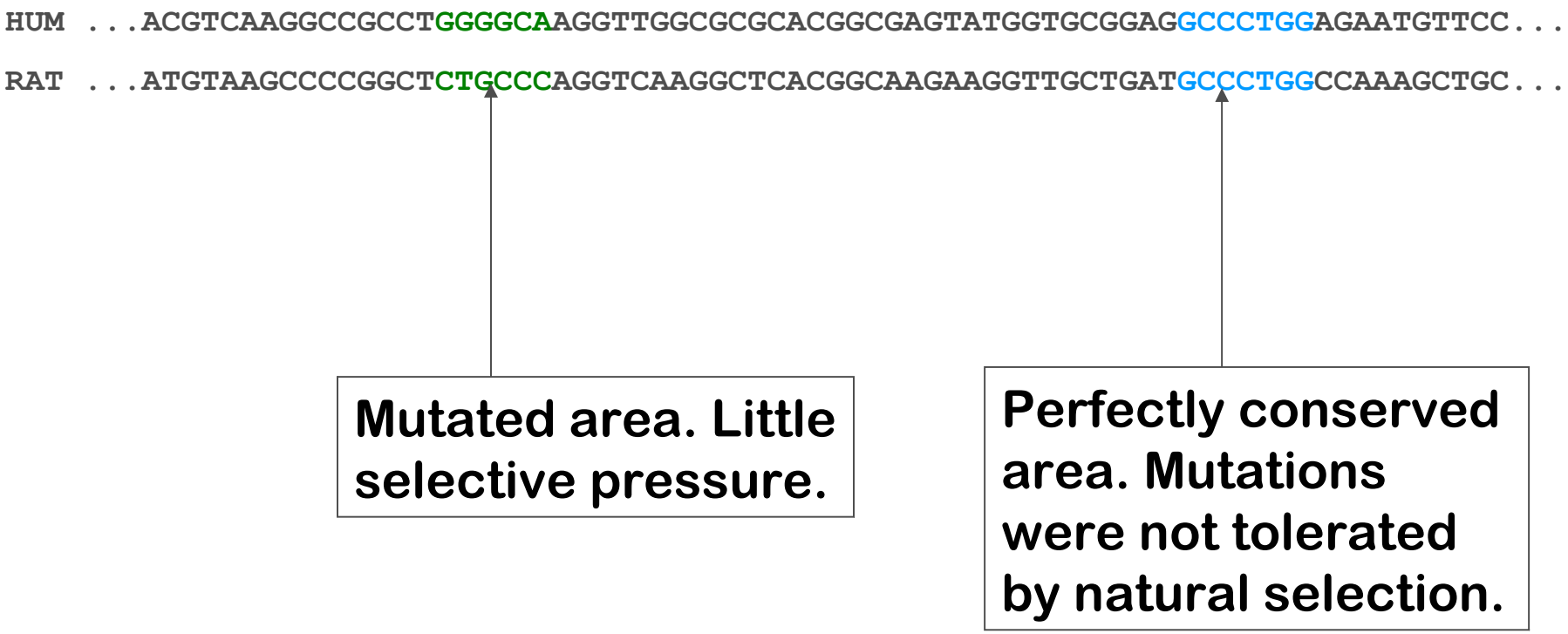
Random sequence similarities



Genomics and Bioinformatics

Chapter 10

Typically sequences are conserved only locally



Sequences need to be aligned locally to detect conserved segments

The conserved segments can be shifted towards each other.

There are substitutions and gaps inside conserved segments.

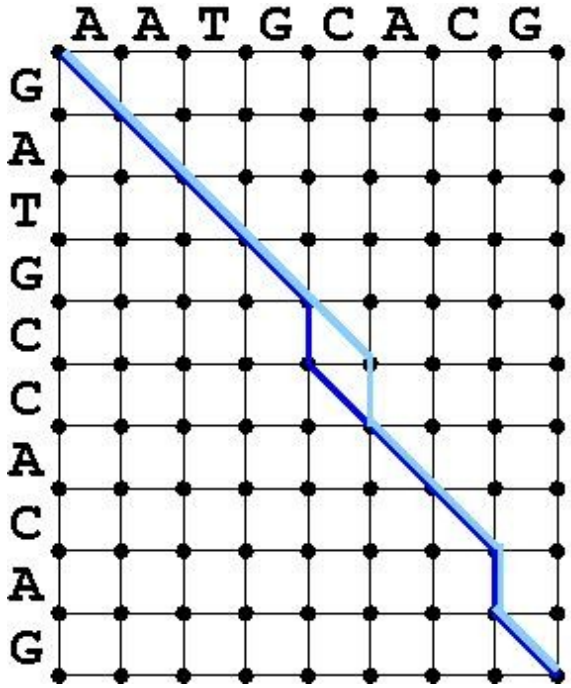
```
Cytochrome C human:   GDVEKGkkifimkcsqchtvekggkhktgpnlhglfgrkTGQAPG
                      YSYTAANKNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEE
                      RADLIAYLKKATNE

Cytochrome C550       MKWNPLTPFLLTAVLGTGLTFFLSVKGLDDSRETASGGFSKSAFK
Bacillus subtilis:    KDANASPeeiykanciachgenyegvsgpslkgvgdkkDVAEIKT
                      KIEKGGNGMPSGLVPADKLDDMAEWVSKIK
```

```
                                    10        20        30
Cytochrome C human:              ... KKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRK ...
                                     ..:.  .:  ::  .  :   .::.:.:.  .:
Cytochrome C550 Bacillus subtilis: ... EEIYKANCIACHGENYEG--VSGPSLKGVGDKK ...
                                           60        70        80
```

Difference between global and local alignment

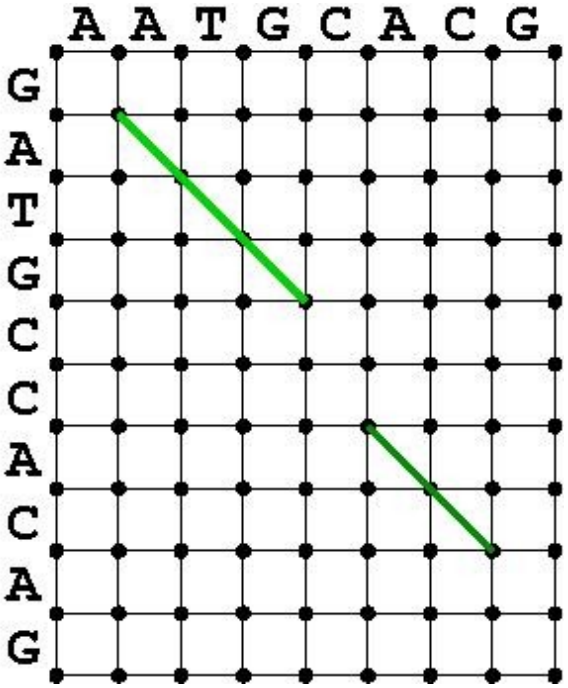
Global

```
AATG-CAC-G
|||  |||  |
GATGCCACAG
```



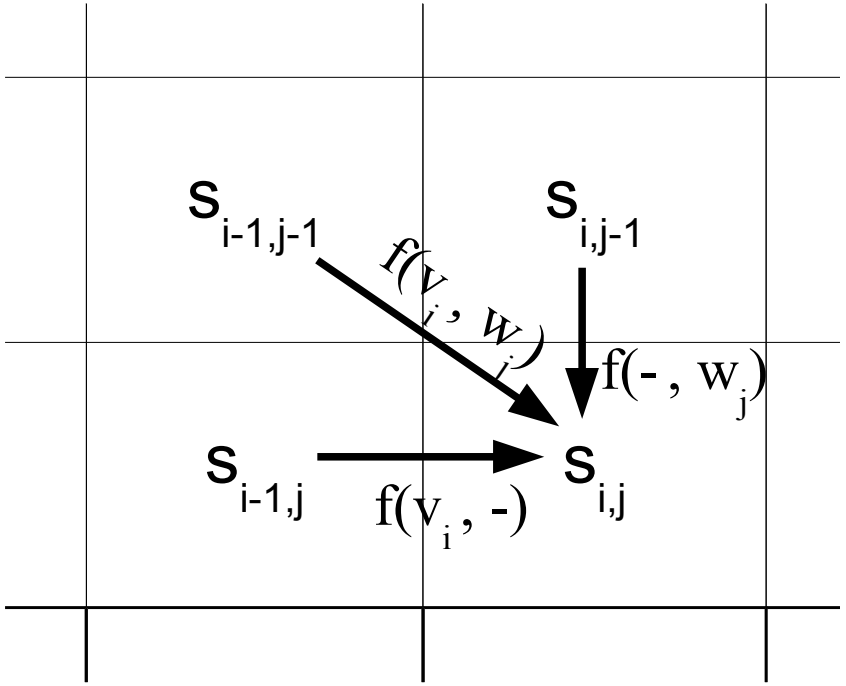
Local

```
AATGCACG
|||
GATGCCACAG
```



The Smith-Waterman algorithm can be used to calculate local alignments

$$s_{i,j} \leftarrow \max \begin{cases} 0 \\ s_{i-1,j} + f(v_i, -) \\ s_{i,j-1} + f(-, w_j) \\ s_{i-1,j-1} + f(v_i, w_j) \end{cases}$$



The start of a new alignment costs 0!

Not all score functions yield localized alignments

```
Seq1   TACGGGTAT
         |||
Seq2  GGACGTACG
```

s(match) = 1

s(mismatch) = -5

s(gap) = -5

```
Seq1  T-ACGGGTAT-
       | |||   |||
Seq2  GGACG--TACG
```

s(match) = 1

s(mismatch) = 0.25

s(gap) = 0

Why is that?

Unrelated segments should not be aligned at all

```
Seq1   TACGGGTAT
         |||
Seq2  GGACGTACG
```

s(match) = 1

s(mismatch) = -5

s(gap) = -5

```
Seq1  T-ACGGGTAT-
      |  |||   |||
Seq2  GGACG--TACG
```

s(match) = 1

s(mismatch) = 0.25

s(gap) = 0

The score of unrelated segments should be negative.

The expected score of random sequences must be negative.

The expected score of random sequences should be negative

```
TACGCGTTCAATGCGTTAT . . .
GGTCATATACGGACGTACG . . .
          Score = 2a-4b
```

s(match) = a
s(mismatch) = - b
We ignore gaps for now.

This is just two random sequences written below each other without any alignment optimization done.

E(score) = a x P(match) -b x P(mismatch)

= a x 0.25 – b x 0.75

If E(S) < 0 alignment of random sequences is expected to result in a negative score and the local alignment algorithm should cut such segments out.

Random scores scatter around the expectation so they are not always negative

TACGCGTTCAATGCGTTAT...
GGTCATATACGGACGTACG...
Score = 2-4 = -2 < 0

s(match) = 1
s(mismatch) = -1

E(score) = a x P(match) -b x P(mismatch)

= 1 x 0.25 – 1 x 0.75 = - 0.5

TACGCGTTCAATGCGTTAT...
GGTCATATACGGACGTACG...
Score = 3-2 = 1 > 0

The red segments are random yet they are similar enough to generate a positive local score although the expected score for random sequences is negative.

The effect of random sequence similarities becomes worse if we allow for a shift between sequences

```
TACGCGTTCAATGCGTTAT...
GGTCATATACGGACGTACG...
Score = 2-4 = -2 < 0

TACGCGTTCAATGCGTTAT...
GGTCATATACGGACGTACG...
Score = 3-2 = 1 > 0

TACGCGTTCAATGCGTTAT...
GGTCATATACGGACGTACG...
Score = 4-0 = 4 > 0
```

s(match) = 1
s(mismatch) = -1

E(score) = a x P(match) -b x P(mismatch)

= 1 x 0.25 – 1 x 0.75 = - 0.5

Allowing for gaps will make the problem even worse.

The choice of the score function can ensure that alignments are local

```
TACGCGTTCCTTGCGTTAT...
GGATATATTACGTACGCCC...
```

conserved segments

```
TACGCGTTCCTTGCGTTAT...
GGATATATTACGTACGCCC...
```

enlarged local alignment due to random similarity

Positive scoring but unrelated segments can occur as a chance event but they are not frequent.

If random sequence similarity occurs next to a conserved segment it will enlarge the local alignment by a bit.

It is very unlikely that the alignment will be global unless the conservation is global.

However we can not avoid positive scoring local alignments that are caused by chance alone

```
TACGCGTTCAATGCGTTAT...
GGTCATACACGGATCCACG...
Score = +1

TACGCGTTCAATGCGTTAT...
GGTCATACACGGATCCACG...
Score = +2

TACGCGTTCAATGCGTTAT...
GGTCATACACGGATCCACG...
Score = +3
```

We essentially always get a local alignment with 100% sequence identity.
Just choose an A in the first sequence and one in the second sequence and "align" them.

This is a positive local random alignment score. But it is not very large.

We can get larger ones.

How large can local alignment scores become just by chance?

From which degree of similarity on should we consider a local sequence similarity to be a trace of common ancestry?

```
Cytochrome C human:   GDVEKGkkifimkcsqchtvekggkhktgpnlhglfgrkTGQAPG
                      YSYTAANKNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEE
                      RADLIAYLKKATNE

Cytochrome C550       MKWNPLTPFLLTAVLGTGLTFFLSVKGLDDSRETASGGFSKSAFK
Bacillus subtilis:    KDANASPeeiykanciachgenyegvsgpslkgvgdkkDVAEIKT
                      KIEKGGNGMPSGLVPADKLDDMAEWVSKIK
```

```
                                    10        20        30
Cytochrome C human:              ... KKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRK ...
                                     ...  .:  ::  .  :   .::.:.:.  .:
Cytochrome C550 Bacillus subtilis: ... EEIYKANCIACHGENYEG--VSGPSLKGVGDKK ...
                                           60        70          80
```

Is this very weak protein sequence similarity real or is it just random sequence similarity?

Random sequence similarity can occur between unrelated sequences.

The alignment process has many degrees of freedom:
- selecting the segments
- introducing gaps

We need to make sure that we do not “construct” similarity by the alignment.

To judge a local alignment, we need to know how strong local sequence similarities can become by chance alone.

We can formalize the problem

Consider two random sequences of lengths n and m.

Sequence X: $X_1,\ldots,X_n$ i.i.d. with $X_i \sim (1/4, 1/4, 1/4, 1/4)$

Sequence Y: $Y_1,\ldots,Y_m$ i.i.d. with $Y_i \sim (1/4, 1/4, 1/4, 1/4)$

All positions are independent of each other.

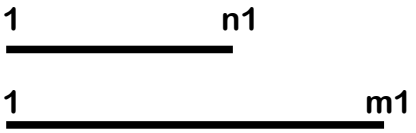
The two sequences are independent of each other.

Let H(X,Y) be the optimal local alignment score of X and Y.

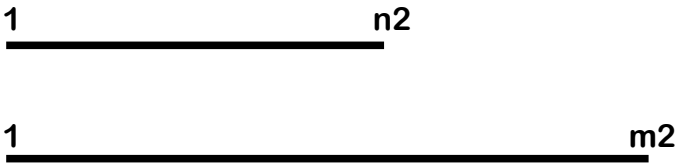
H(X,Y) is a random variable.

What is its distribution? $P(H<t) = ?$

The distribution of random alignment scores must depend on the length of the aligned sequences

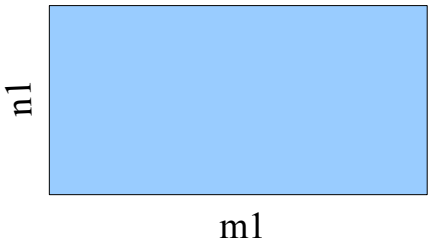


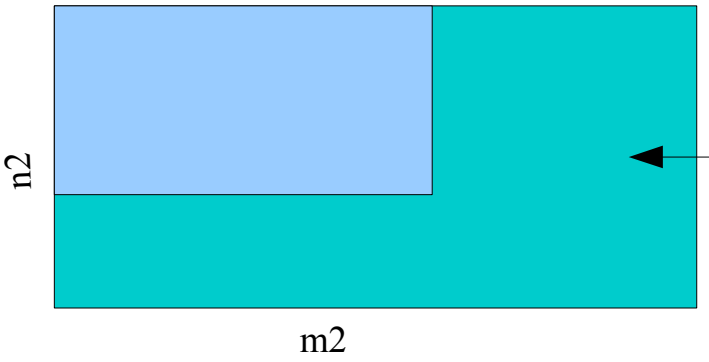
Pair 1



Pair 2

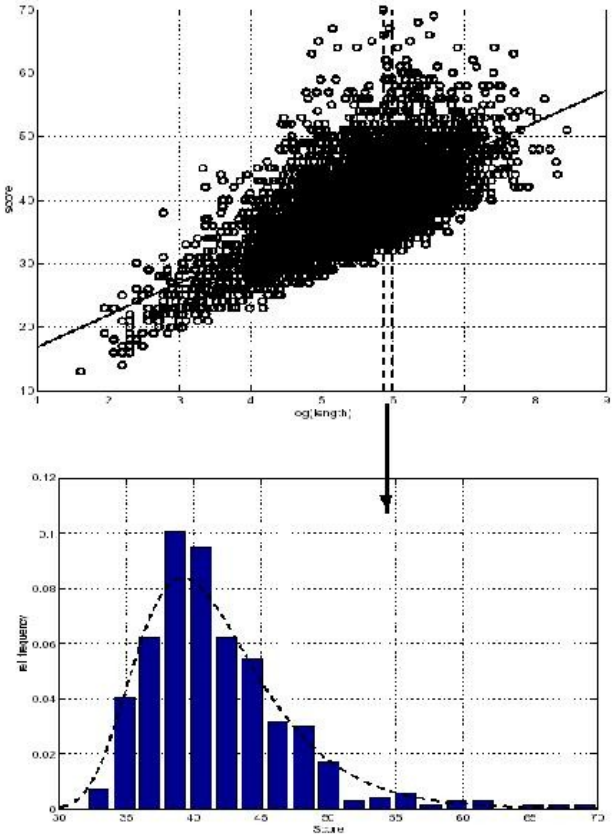
Alignment table





Additional possible end points for optimal local alignments

A simulation experiment shows that the expected random score grows with the log of the size of the search space

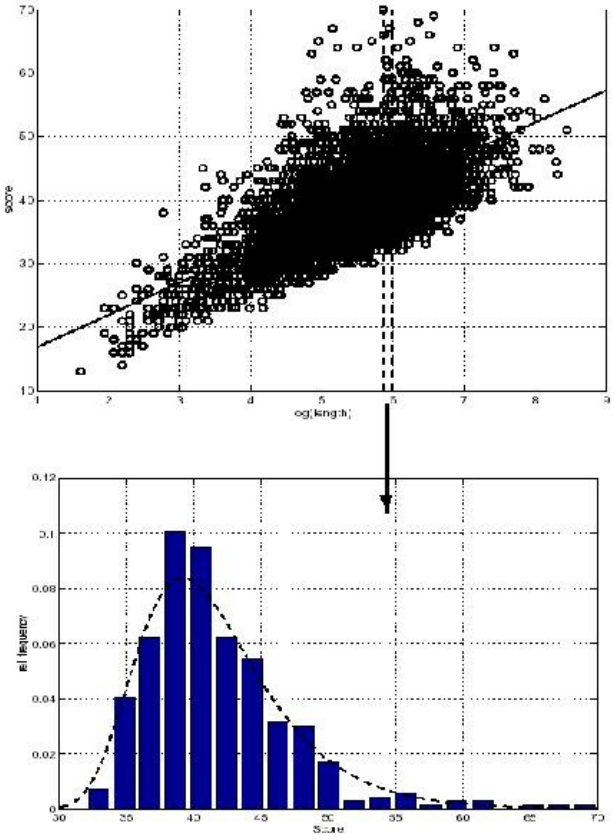


Generate many pairs of random sequences of different lengths n and m.

Align them using the Smith-Waterman algorithm.

Draw the scores against log(nm).

The random scores scatter around the expectation

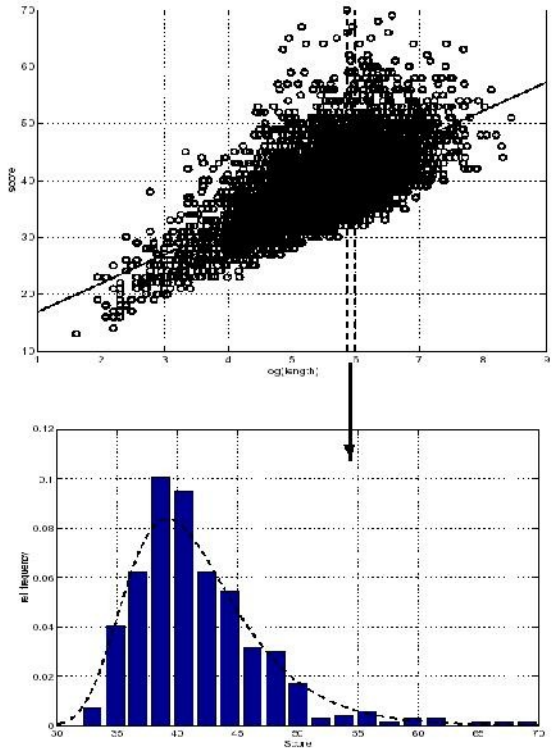


$$H \sim \alpha + \theta \log(m\,n) + \theta\, G$$

The regression line (deterministic)

The random part (residual)

The shape of the residual distribution is not symmetric around the mean



$$H \sim \alpha + \theta \log(m\,n) + \theta\, G$$

The regression line (deterministic)

The random part (residual)

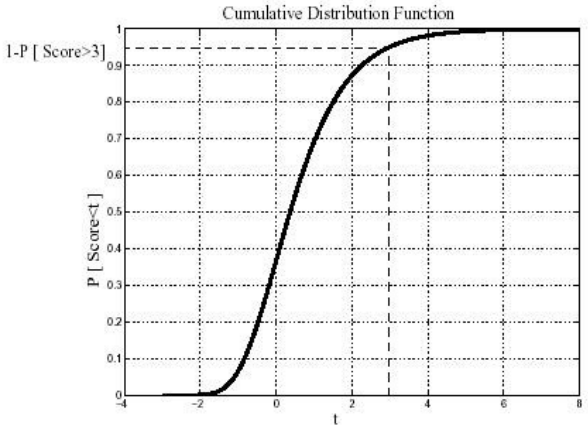
Large deviations towards higher scores are more frequent .

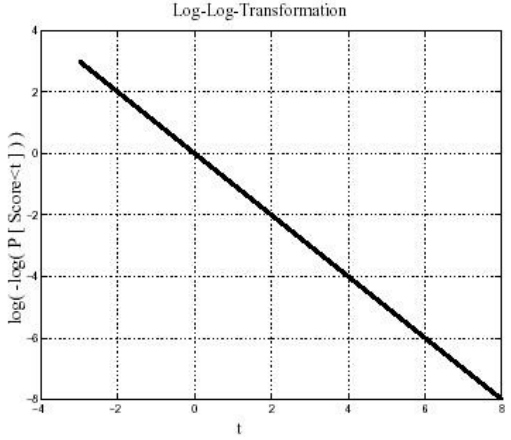
The random scores follow an extreme value distribution

A continuous random variable G with

$$P(G < t) = e^{-e^{-t}}$$

is called a standard extreme value distributed variable.





The extreme value distributions build a family of distributions

If G is standard extreme value distributed, the shifted and rescaled variable

X = θ G + ξ

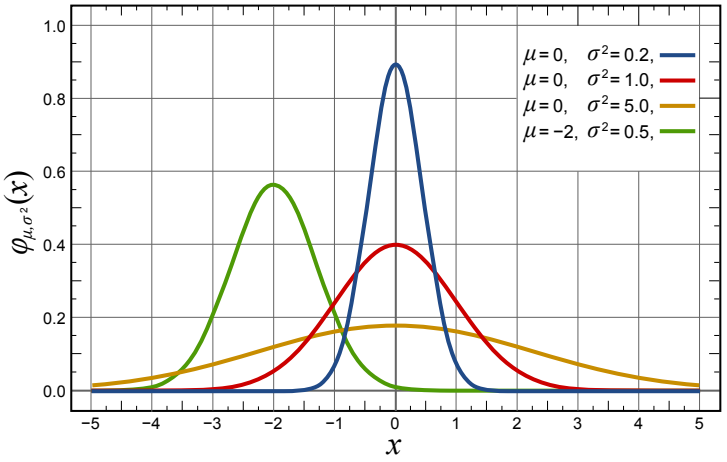
has the distribution function:

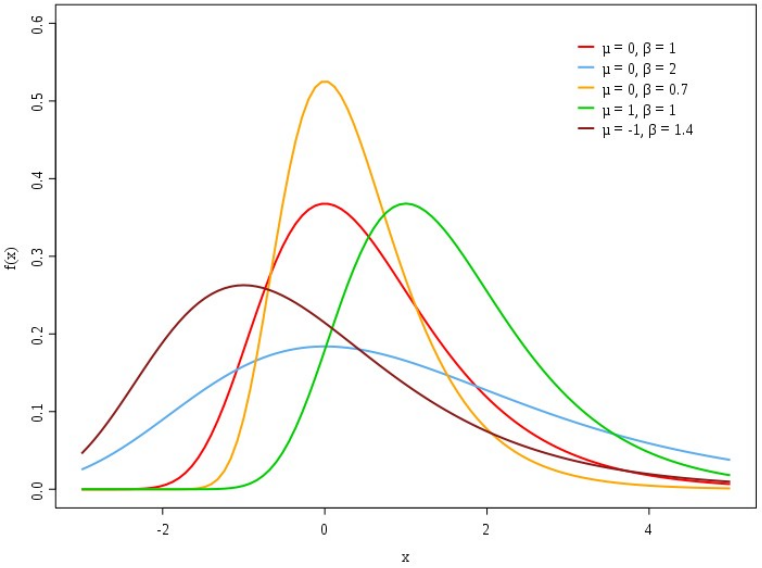
$$\begin{aligned} P[X < t] &= P\left[G < \frac{t-\xi}{\theta}\right] \\ &= \exp\left(-e^{-\frac{t-\xi}{\theta}}\right) \\ &= \exp\left(-e^{\frac{\xi}{\theta}} e^{-\frac{t}{\theta}}\right) \end{aligned}$$

X is extreme value distributed with scale parameter θ and location parameter ξ.

In short: X ~ G(ξ,θ)

Recap: normal distribution vs. extreme value distribution





Probability density function:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$f(x) = \frac{1}{\beta} \mathrm{e}^{-\frac{1}{\beta}(x-\mu)} \mathrm{e}^{-\mathrm{e}^{-\frac{1}{\beta}(x-\mu)}}, \quad x \in \mathbb{R}$$

Local alignments with score functions with sufficiently high mismatch and gap costs follow an extreme value distribution

$$P[H \geq t] \approx 1 - \exp\left(-\gamma\, n\, m\, e^{-\frac{t}{\theta}}\right)$$

For alignment without gaps there are explicit formulas for γ and θ (scale parameter).

For local alignments with gaps, we can estimate them by simulations.

The parameters of the extreme value distribution can be estimated by random simulation

Simulate 1000 pairs of random sequence and align them using the Smith-Waterman algorithm.

This gives you 1000 random alignment scores.

Calculate the mean and variance of your simulated scores.

You get the scale and location parameter by:

$$\theta = \frac{\sqrt{6}}{\pi} SD_X^2$$

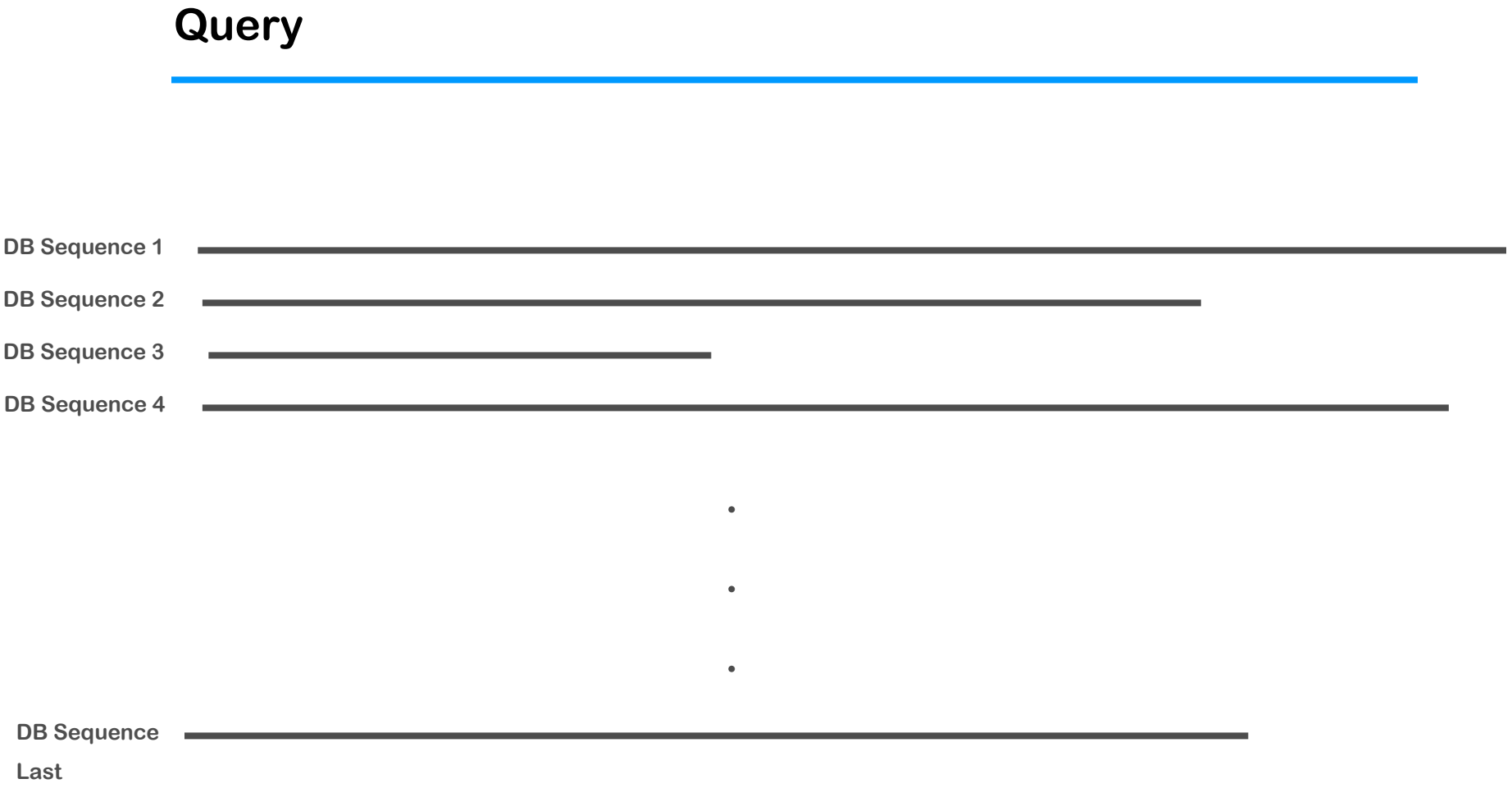
$$\xi = \bar{X} - c\theta$$

This formula tells you when you can conclude from a sequence similarity to common ancestry and hence to similar function of a gene

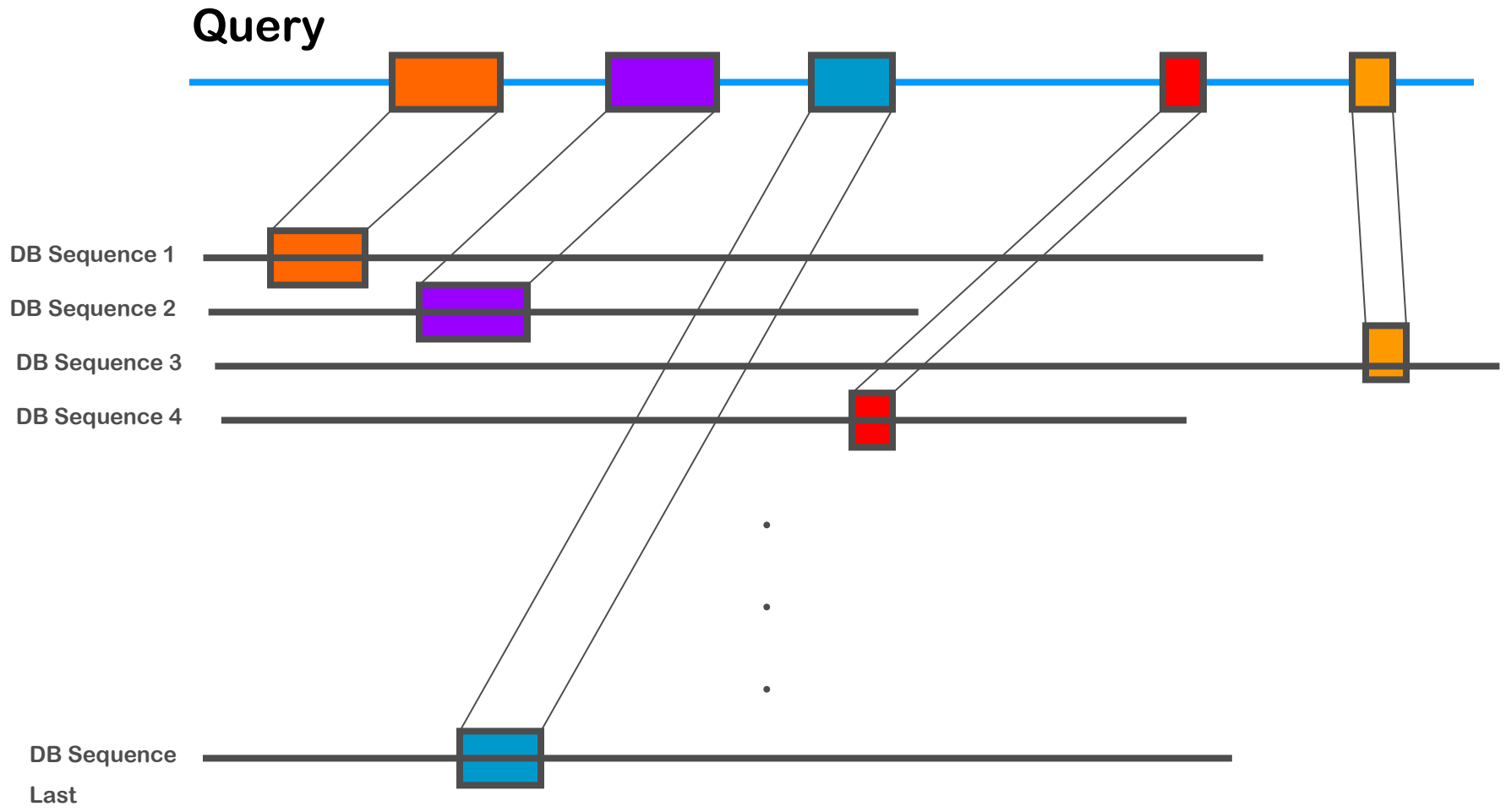
$$P[H \geq t] \approx 1 - \exp\left(-\gamma\, n\, m\, e^{-\frac{t}{\theta}}\right)$$

And it will greatly enhance the power of molecular database searches...

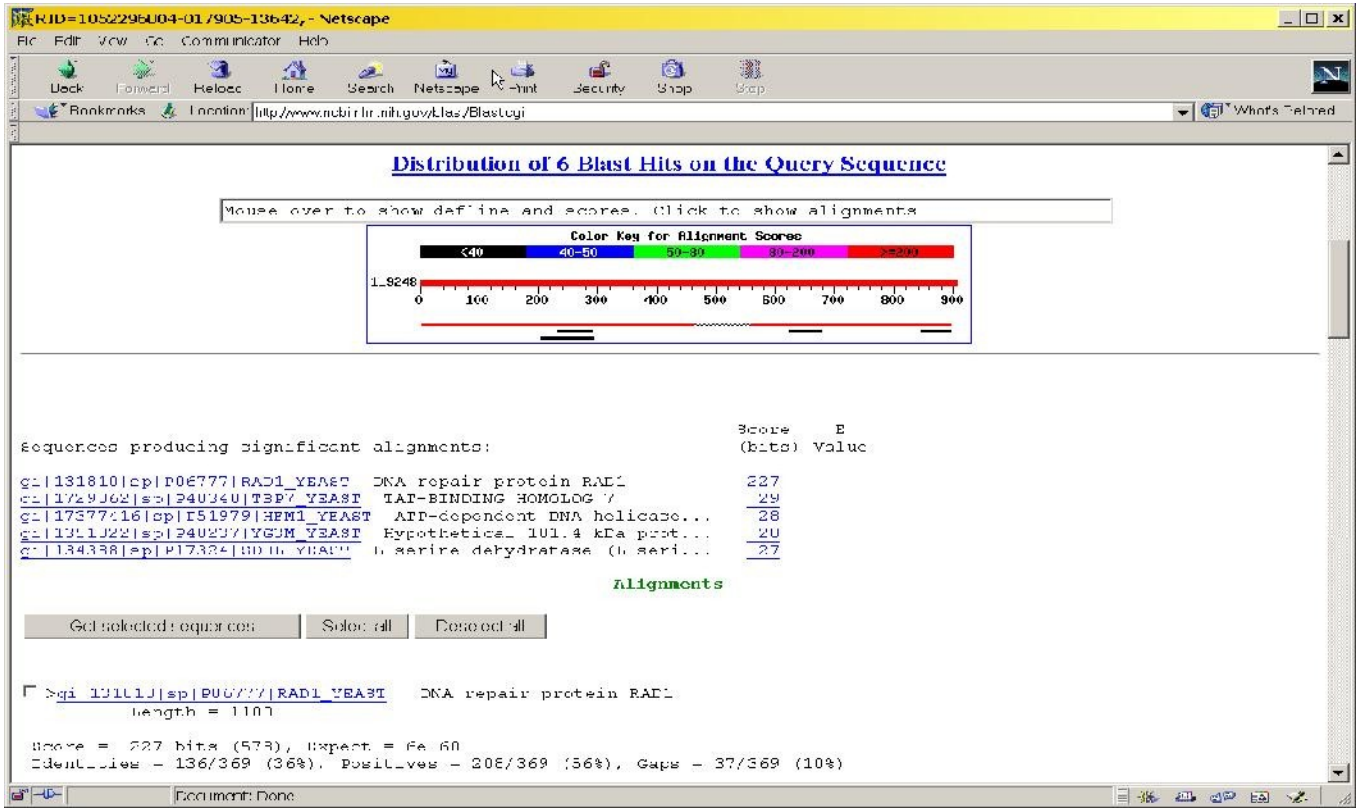
In a genomic database search a query sequence is compared to hundreds of thousands of database sequences



We typically search for local sequence similarities



We can rank the database entries by the local alignment scores



Do the scores of 29, 28, 27, … still reflect true homologies?

Molecular database searches are large scale random experiments

The vast majority of sequences in the database are not related to the query.

Local similarities occur just by chance.

Of course we hope that at least some database sequences are related to the query and we hope they show stronger than random similarity.

For the following slides we just assume all sequences are unrelated and all scores are random.

Our formula gives us a p-value for a score

Null Hypothesis (Model):
The score results from random sequence similarity

$$P[H \geq t] \approx 1 - \exp\left(-\gamma\, n\, m\, e^{-\frac{t}{\theta}}\right)$$

The formula gives us the probability of observing a score of t or higher (e.g. t=29) under the assumption that the sequences are not related.

The probability of an observation under the assumption that it is just a random fluctuation is called a p-value.

We can rewrite the distribution of H given the sequence lengths n and m as a regression equation

$$P[H \geq t] \approx 1 - \exp\left(-\gamma\, n\, m\, e^{-\frac{t}{\theta}}\right)$$

The location but not the scale of the score distribution depends on the sequence lengths.

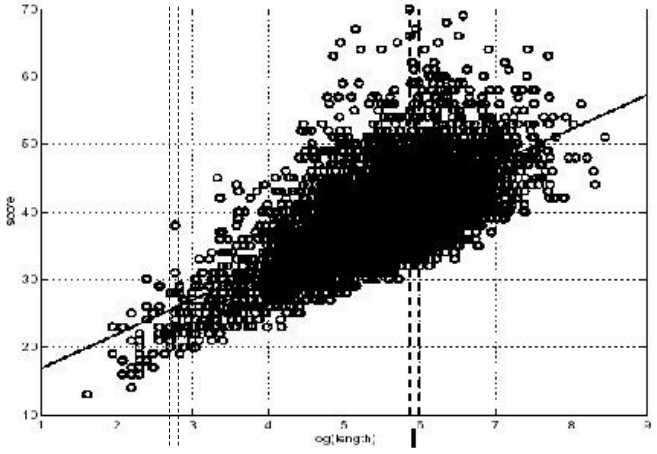
We can make this dependency more transparent by rewriting the above formula as:

$$H \sim \alpha + \theta \log(m\, n) + \theta\, G$$

H is distributed like a standard extreme value distribution rescaled by θ and shifted by $\alpha+\theta \log(mn)$.

According to the formula the variance of the residual is constant across all sequence lengths

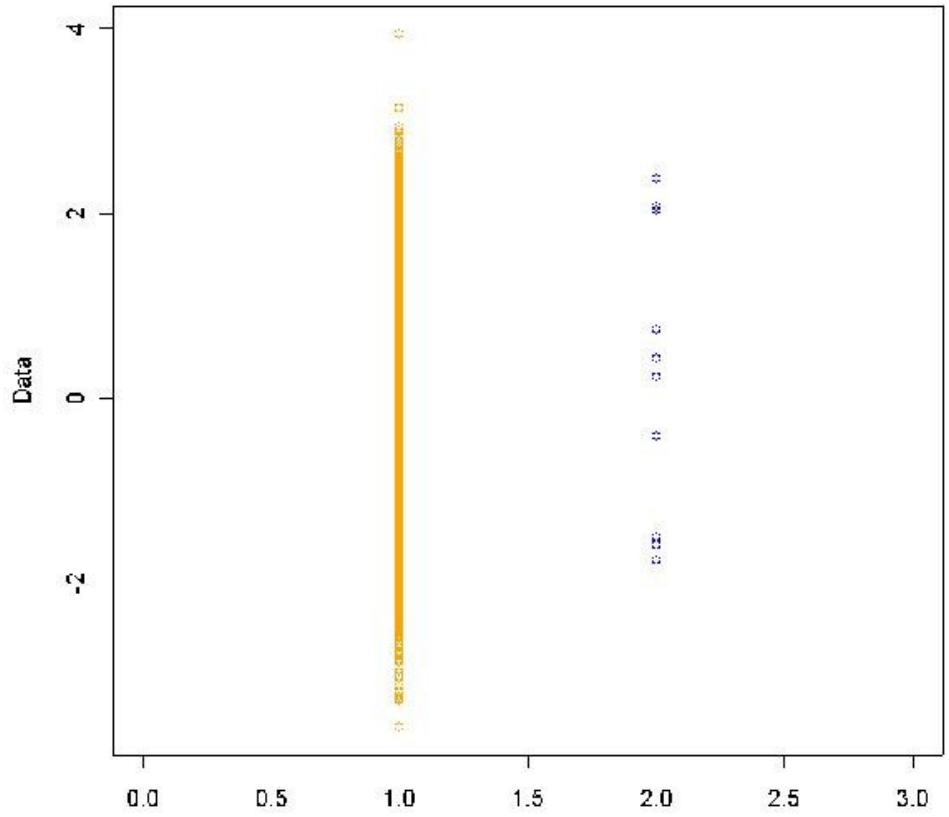
$$H \sim \alpha + \theta \log(m\,n) + \theta\, G$$



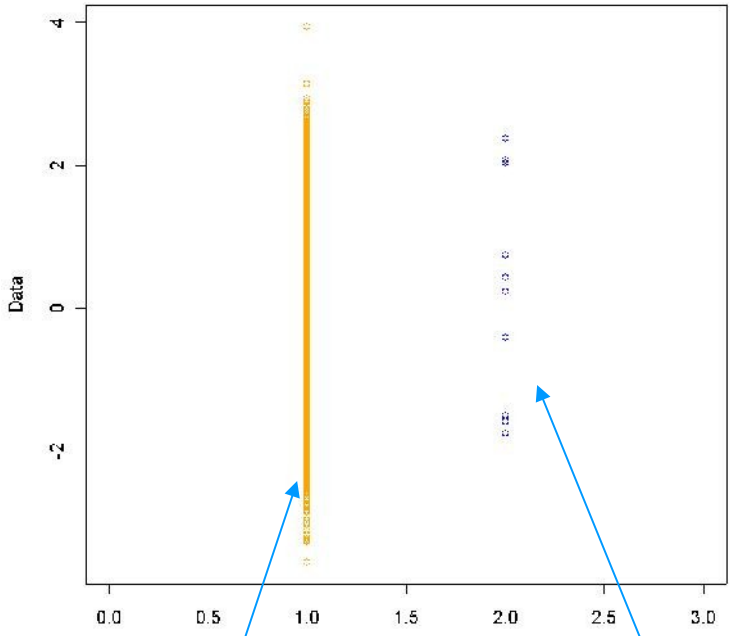
Really?

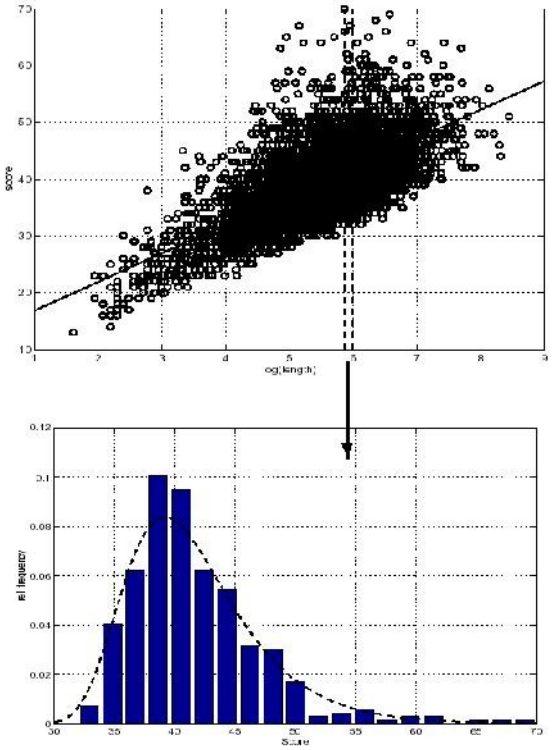
Is the variance of the data in both stripes the same?

Which data set has the higher variance?



The eye confuses the variance with the range

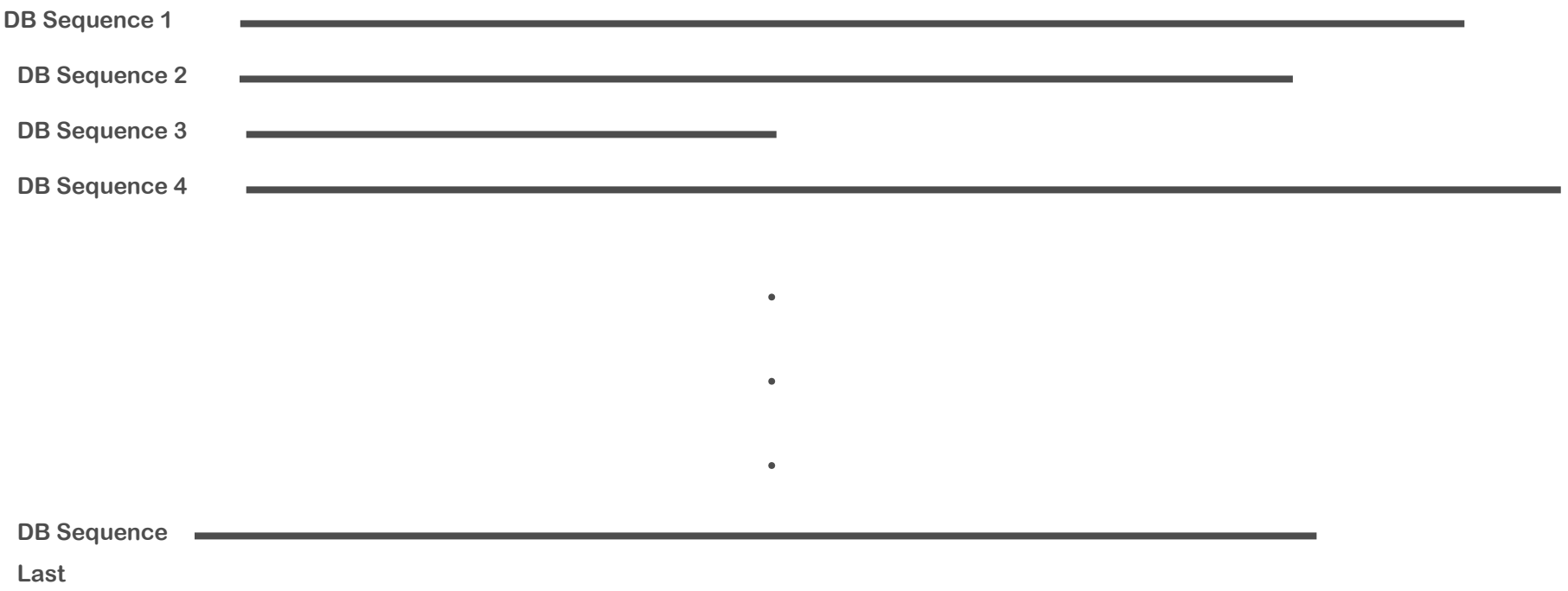




5000 N(0,1) points

10 N(0,1.5) points

In a database search every comparison is dealing with different sequence lengths



We need to adopt the null model to the sequence length for every sequence comparison in a database search

The length of the query is always the same, but the database entries have different lengths.

H_i: **score from comparing the query to DB sequence** i.

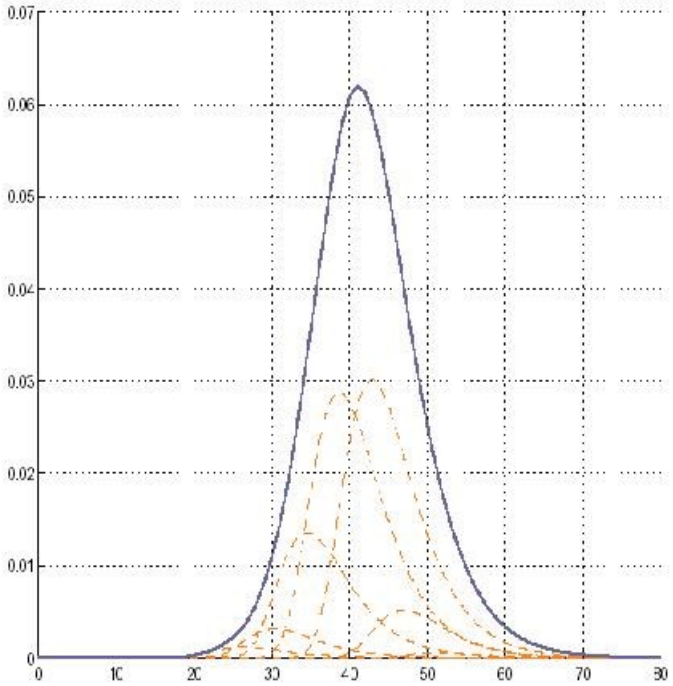
$$P[H_i \geq t] \approx 1 - \exp\left(-\hat{\gamma}\, m_i\, e^{-\frac{t}{\theta}}\right)$$

$$\hat{\gamma} = \gamma\, n$$

A Database search of 100.000 sequences produces 100.000 scores H_i.

Do the H_i follow an extreme value distribution?

The scores do not follow an extreme value distribution due to different sequence lengths



Every single H_i is drawn from an extreme value distribution.

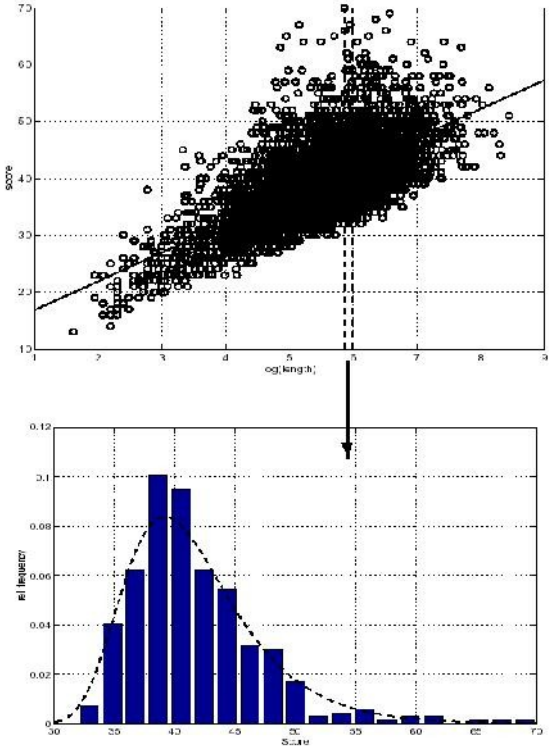
But these are different (shifted) distributions from the family of extreme value distributions.

The mixture is no longer extreme value distributed.

We can adopt the regression form of the alignment statistics formula to database searches

$$P[H_i \geq t] \approx 1 - \exp\left(-\hat{\gamma}\, m_i\, e^{-\frac{t}{\theta}}\right)$$

$$H_i \sim \hat{\alpha} + \theta \log(m_i) + \theta\, G$$



Longer database entries generate in average higher random alignment scores

$$H_i \sim \hat{\alpha} + \theta \log(m_i) + \theta G$$

If the database sequence is long:

- There are more segments that can be similar

- The Smith-Waterman dynamic programming tables are larger

- The search space is larger

- We take the maximum over more local random alignments

We can calculate lengths adjusted scores by shifting them

Define lengths adjusted scores by:

$$A_i := H_i - \theta \log(m_i)$$

Their distribution is given by:

$$A_i \sim \hat{\alpha} + \theta\, G$$

Which no longer depends on the lengths of database sequences.

All lengths adjusted scores follow the same extreme value distribution.

There are two ways to rank sequences in a database search

We can rank them by alignment scores.

We can rank them by lengths adjusted alignment scores.

What is better ?

Why?

Biological irrelevant experiments can still be instructive

Accession	Description	Max score
XP_821204.1	fumarate hydratase [Trypanosoma cruzi strain CL Brener] >gb|EAN99353.1| fumarate hydratas	34.7
BAG41970.1	alkaline phosphatase [Bombyx mori]	38.9
XP_501340.1	YALI0C02057p [Yarrowia lipolytica] >emb|CAG81648.1| YALI0C02057p [Yarrowia lipolytica]	34.7
NP_001024029.1	hypothetical protein K04F1.14 [Caenorhabditis elegans] >gb|AAQ62448.1| Hypothetical protein	35.4
AAN61918.1	IQGAP-like protein [Eremothecium gossypii]	36.2
NP_985400.2	AFL150Cp [Ashbya gossypii ATCC 10895] >gb|AAS53224.2| AFL150Cp [Ashbya gossypii ATCC	36.2
YP_641269.1	hypothetical protein Mmcs_4128 [Mycobacterium sp. MCS] >ref|YP_940185.1| hypothetical pr	35.4
NP_752388.1	RTX family exoprotein A gene [Escherichia coli CFT073] >gb|AAN78844.1|AE016756_27 Putativ	38.9
ZP_07446310.1	hypothetical protein ECNC101_09364 [Escherichia coli NC101] >gb|EFM54588.1| hypothetical p	38.5
ZP_07103045.1	conserved hypothetical protein [Escherichia coli MS 185-1] >ref|ZP_07511011.1| hypothetical	38.5
YP_663571.1	FHA domain-containing protein [Pseudoalteromonas atlantica T6c] >gb|ABG42517.1| FHA dom	35.0
EEC82183.1	hypothetical protein OsI_26313 [Oryza sativa Indica Group]	35.8
EEE67316.1	hypothetical protein OsJ_24561 [Oryza sativa Japonica Group]	35.8
BAC79582.1	putative receptor-like protein kinase 4 [Oryza sativa Japonica Group] >dbj|BAD32134.1| putati	35.0
ZP_07622560.1	hypothetical protein EcolH2_04186 [Escherichia coli H299]	37.4
ZP_05081555.1	cyclopropane-fatty-acyl-phospholipid synthase [beta proteobacterium KB13] >gb|EDZ64242.1	34.7
XP_002844715.1	chitinase [Arthroderma otae CBS 113480] >gb|EEQ33860.1| chitinase [Arthroderma otae CBS	35.4
XP_002266469.1	PREDICTED: hypothetical protein [Vitis vinifera]	35.0
XP_002930992.1	PREDICTED: zinc finger protein 275-like [Ailuropoda melanoleuca]	34.3
YP_002862685.1	hypothetical protein CLJ_B1902 [Clostridium botulinum Ba4 str. 657] >gb|ACQ53271.1| conserv	36.2
YP_002803843.1	hypothetical protein CLM_1653 [Clostridium botulinum A2 str. Kyoto] >gb|ACO84863.1| conser	36.2
ZP_02618644.1	Xaa-His dipeptidase [Clostridium botulinum Bf] >gb|EDT84912.1| Xaa-His dipeptidase [Clostridi	36.2
ZP_02994779.1	hypothetical protein CLOSPO_01898 [Clostridium sporogenes ATCC 15579] >gb|EDU35733.1| h	36.2
XP_002568183.1	Pc21g11520 [Penicillium chrysogenum Wisconsin 54-1255] >emb|CAP96049.1| Pc21g11520 [Pe	35.8
ZP_07005208.1	2,3-diketo-L-gulonate reductase [Eubacterium saburreum DSM 3986] >gb|EFU75819.1| 2,3-dik	34.3
ZP_01075968.1	Lysine exporter protein (LYSE/YGGA) [Marinomonas sp. MED121] >gb|EAQ66043.1| Lysine exp	35.0
ZP_06793519.1	ATP synthase beta subunit/transcription termination factor Rho [Brucella sp. NVSL 07-0026] >|	34.7
NP_539404.1	ATP synthase beta subunit/transription termination factor Rho [Brucella melitensis bv. 1 str. 16	34.7
BAA06425.1	TPA: TPA_inf: occludin-related Y protein [Drosophila ananassae]	34.3
ZP_07035819.1	ComEC/Rec2 family protein [Prevotella oris C735] >gb|EFI48606.1| ComEC/Rec2 family protein	37.0
ZP_06257090.1	ComEC/Rec2 family protein [Prevotella oris F0302] >gb|EFB30584.1| ComEC/Rec2 family protei	37.0
YP_001311441.1	YD repeat-containing protein [Clostridium beijerinckii NCIMB 8052] >gb|ABR36485.1| YD repeat	34.7
Q19319.3	RecName: Full=Cadherin-4; Flags: Precursor >emb|CAA84339.2| C. elegans protein F25F2.2a,	35.4
NP_497917.1	CaDHenn family member (cdh-4) [Caenorhabditis elegans]	35.4
XP_003105982.1	CRE-CDH-4 protein [Caenorhabditis remanei] >gb|EFO98508.1| CRE-CDH-4 protein [Caenorhab	35.0
XP_002160126.1	PREDICTED: similar to F59H6.5 [Hydra magnipapillata]	34.3
ZP_06191060.1	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase [Serratia odorifera	34.3
ZP_06987924.1	transcriptional regulator [Bacteroides sp. 3_1_19] >gb|EFI06718.1| transcriptional regulator [B	35.0
YP_003914852.1	formate dehydrogenase alpha subunit [Ferrimonas balearica DSM 9799] >gb|ADN77778.1| form	34.3
XP_002107718.1	hypothetical protein TRIADDRAFT_51513 [Trichoplax adhaerens] >gb|EDV28516.1| hypothetica	34.3
ZP_03729449.1	transcriptional activator domain protein [Dethiobacter alkaliphilus AHT 1] >gb|EEG78005.1| tra	35.0
ZP_07324283.1	ComEC/Rec2-like protein [Prevotella disiens F8035-09AN] >gb|EFL45210.1| ComEC/Rec2-like p	34.7
XP_002120492.1	PREDICTED: similar to HELZ protein [Ciona intestinalis]	35.8
EFQ26893.1	chromo domain-containing protein [Glomerella graminicola M1.001]	35.8

We have searched similarities to a randomly generated sequence of 1000 aa in a huge protein database (all known proteins 2010).

The list is ranked by scores.

The top ranking sequences display quite some local similarity to the random query sequence

>[redacted] RTX family exoprotein A gene [Escherichia coli CFT073]
gb|AAN78844.1|AE016756_27 Putative RTX family exoprotein A gene [Escherichia coli CFT073]
Length=1610

[redacted] | RTX family exoprotein A gene [Escherichia coli CFT073]
(10 or fewer PubMed links)

Score = 38.9 bits (89), Expect = 4.1, Method: Compositional matrix adjust.
Identities = 26/99 (27%), Positives = 39/99 (40%), Gaps = 9/99 (9%)

```
Query  279  SWVMEDSQSGWTSHTPVHVYDMENTR--MHADTFCQKQKTWWAHYPPM----DPQV--QH  330
            +  +E   +GWTS  P ++ D    +  + AD     +       P     DP
Sbjct  874  TVTLEKGDNGWTSSDPTLIPDSTGDKATIPADNVKDNSEVTGVAKDPSGNESDPSTVTSK  933

Query  331  TDFLACTCLKYWCNKTTYNGNGFHSRVTVDGWTIMRRPA  369
            TD L    +     T  NG+GF    +V+G T+M  PA
Sbjct  934  TDVLPTVSISVETTSTDVNGDGFTGIASVNG-TVMDVPA  971
```

>dbj|BAG[redacted]970.1| alkaline phosphatase [Bombyx mori]
Length=550

[redacted] | membrane-bound alkaline phosphatase [Bombyx mori]
(10 or fewer PubMed links)

Score = 38.9 bits (89), Expect = 4.6, Method: Compositional matrix adjust.
Identities = 52/191 (28%), Positives = 79/191 (42%), Gaps = 32/191 (16%)

```
Query  552  DTCRCPD----YFSMKNGNAY--VFNCRRSG-------PEKGQLG--SDSVENPECIFGD  596
            D  RCPD        M  GN +  +F   R           E+G  G  +D     E    D
Sbjct  223  DVNRCPDIAHQLIKMAPGNKFKVIFGGGRREFLPTTQVDEEGTRGLRTDGRNLIEEWQND  282

Query  597  SKVRQ-DYKMVLYQQKRCRGFNVPCEYVYGVACNNHHSLHMEIHPRGD--MRHTCLEPTQ  653
             + ++  YK +  +Q+  +  + P +Y+ G+    +H   H+E    GD     T  E T
Sbjct  283  KESQKVSYKYLWNRQELLKLGSSPPDYLLGLFEGSHLQYHLE----GDESTEPTLAELTD  338

Query  654  HCCNVLCRQHFCYFI---RGQVHHS-----PHLAADPTTFI-RA-KIVTEEVDLRRSSVF  703
                 VLCR    +F+   RG++ H+       HLA D T   + RA K+ T+ +    S V
Sbjct  339  VAIRVLCRNERGFFLFVERGRIDHAHHDNYAHLALDETIEMDRAVKVATDALKEDESLVV  398

Query  704  YKMDKCLPMAF  714
                D    M+F
Sbjct  399  VTADHTHVMSF  409
```

>ref|ZP_[redacted]7446310.1| hypothetical protein ECNC101_09364 [Escherichia coli NC101]
gb|EFM5456[redacted].1| hypothetical protein ECNC101_09364 [Escherichia coli NC101]
Length=1274

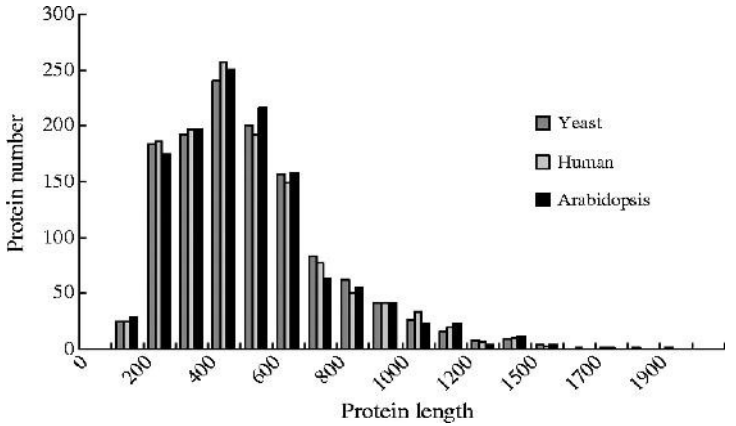
[redacted] 38.5 bits (88), Expect = 4.8, Method: Compositional matrix adjust.
Identities = 26/99 (27%), Positives = 39/99 (40%), Gaps = 9/99 (9%)

```
Query  279  SWVMEDSQSGWTSHTPVHVYDMENTR--MHADTFCQKQKTWWAHYPPM----DPQV--QH  330
            +  +E   +GWTS  P ++ D    +  + AD     +       P     DP
Sbjct  539  TVTLEKGDNGWTSSDPTLIPDSTGDKATIPADNVKDNSEVTGVAKDPSGNESDPSTVTSK  598

Query  331  TDFLACTCLKYWCNKTTYNGNGFHSRVTVDGWTIMRRPA  369
            TD L    +     T  NG+GF    +V+G T+M  PA
Sbjct  599  TDVLPTVSISVETTSTDVNGDGFTGIASVNG-TVMDVPA  636
```

>ref|ZP_07193045.1| conserved hypothetical protein [Escherichia coli MS 185-1]
ref|ZP_07511011.1| hypothetical protein EcolTA_00495 [Escherichia coli TA206]
gb|EFJ56[redacted]3.1| conserved hypothetical protein [Escherichia coli MS 185-1]
gb|ADN449[redacted].1| hypothetical protein ECABU_c03190 [Escherichia coli ABU 83972]
Length=1275

A look on the lengths of the top ranking sequences is instructive:

They are among the largest proteins.



If we only have unrelated sequences in the database the highest alignment scores in a database search result from long sequences

Long sequences produce more alignment noise.

They tend to rank higher than weak but real homologies.

This effect is compensated when ranking by adjusted scores or by p-values.

Note that the ranking of adjusted scores and p-values is identical.

The main benefit of the p-values in a genomic database search is not in judging significance but in improving the ranking

On average:

The real hits rank higher in the p-value driven ranking than in the score driven ranking.

The performance of the database search improved.

Long sequences go down in the ranking.
Shorter ones come up.

Most database search programs report an E-value instead of a p-value

Assume a sequence s reaches a score of t.

The E-value describes how many scores of t or higher one would expect just due to random sequence similarity.

P-value: $p(s) = P(H \geq t)$

The database has N sequences.

We perform N "random experiments" in a database search.

The expected value of the number of scores of t or above is:

$E(S) = N \times p(s)$

But the sequences in a database do not have identical lengths

… and hence the N random experiments have different probabilities p(i) i=1,…, N to reach a score of t.

Idea: Lets "make" the database homogeneous in lengths.

Let m_i be the lengths of sequence i.

And $L=\sum m_i$ the lengths of the database in base pairs.

Let $D(s)=L/m_s$ the number of times a sequence of the same length as s fits into the database.

Database search engines report the E-value:

$$E(s) = D(s) \times p(s)$$

What about the distribution of the score H^{max} of the highest ranking sequence?

We still assume all sequences, query and database, are random and independent from each other:

$H^{max}=\max(H_1,\ldots,H_N)$

What is the distribution of H^{max}? $P(H^{max}<t) = ?$

The event $H^{max}<t$ says that no sequence in the database reaches a score of t:

$$P[H^{max} < t] = P\left[\bigcap_{1\le i\le N} \{H_i < t\}\right]$$
$$= \prod_{1\le i\le N} P[H_i < t]$$

H^{max} follows an extreme value distribution, too

$$P[H^{max} < t] = P\left[\bigcap_{1 \le i \le N} \{H_i < t\}\right]$$
$$= \prod_{1 \le i \le N} P[H_i < t]$$

Plugging in the distributions of the H_i and taking the log yields:

$$\log\left(P[H^{max} < t]\right) = \sum_{i=1}^{N} -\hat{\gamma}\, m_i\, e^{-\frac{t}{\theta}}$$
$$= -\hat{\gamma}\, L\, e^{-\frac{t}{\theta}}$$

Where $L := \sum_i m_i$ **is the length of the database.**

The longer the database the higher the expected value of H^{max}

$$H^{max} \sim \hat{\alpha} + \theta \log(L) + \theta G$$

We can do the same calculations for the maximal length adjusted score A^{max}

$A^{max} = \max(A_1,\ldots,A_N)$

$$A^{max} \sim \hat{\alpha}_1 + \theta \log(N) + \theta G$$

The distribution of A^{max} also depends on the size of the database that we search but this time measured by the number of sequences N in the database and not the number of bases (or aa) L.

Random similarity noise growths with the databases

Genomic databases grow rapidly due to the numerous genome projects.

This seems to come for a price.

Real similarities must be stronger every year to be dissected from random similarities.

The database noise becomes louder every year.

We can simulate the database noise

Two distantly related sequences were hidden in a database of many random sequences. We search for similarities to HBB_HUMAN.

	Sequence	Score
1.	HBB_HUMAN	725
2.	GLB2_TYLHE	72
3.	random.6532	65
4.	random.2117	65
5.	random.9620	62
6.	random.1147	62
7.	random.549	61
8.	random.1661	61
9.	random.3562	61
10.	random.2800	60
11.	random.5711	59
12.	random.5170	59

10.000 *comparisons*

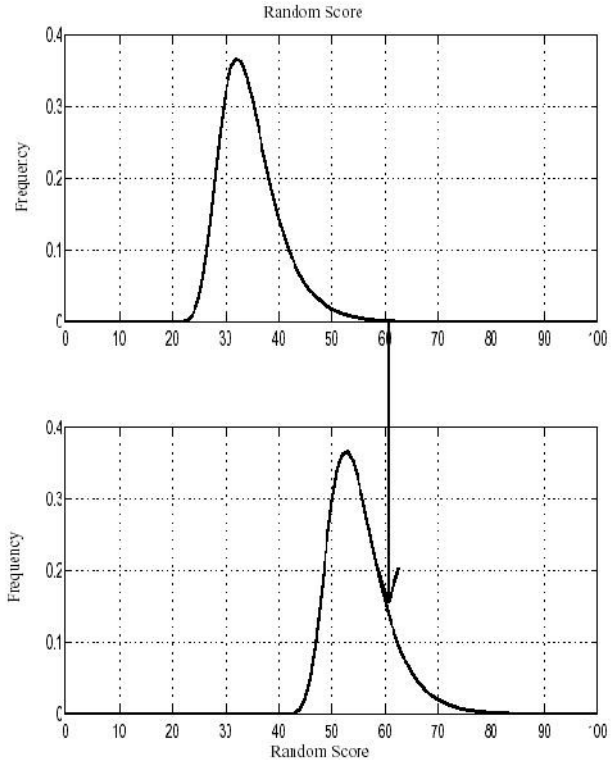
	Sequence	Score
1.	HBB_HUMAN	725
2.	random.20673	82
3.	random.29959	81
4.	random.95385	78
5.	random.77503	77
6.	random.60158	75
7.	random.57179	75
8.	random.46083	73
9.	GLB2_TYLHE	72
10.	random.68600	72
11.	random.87038	71
12.	random.40156	70

100.000 *comparisons*

	Sequence	Score
1.	HBB_HUMAN	725
2.	random.413406	100
3.	random.874986	97
4.	random.401601	90
5.	random.862697	89
6.	random.461280	85
7.	random.520651	84
8.	random.20673	82
9.	random.304933	82
10.	random.739210	81
11.	random.29959	81
12.	random.374090	81
	...	
46.	GLB2_TYLHE	72

1 *mio comparisons*

Database noise becomes louder



Database with 500 sequences of length 300

Database with 50.000 sequences of length 300

A score of 60 stands out of the noise in the small database but not in the large one.

The sequence similarity has not changed. Only the size of the database in which it was found has changed.

We can study this effect on real databases that grow over the years

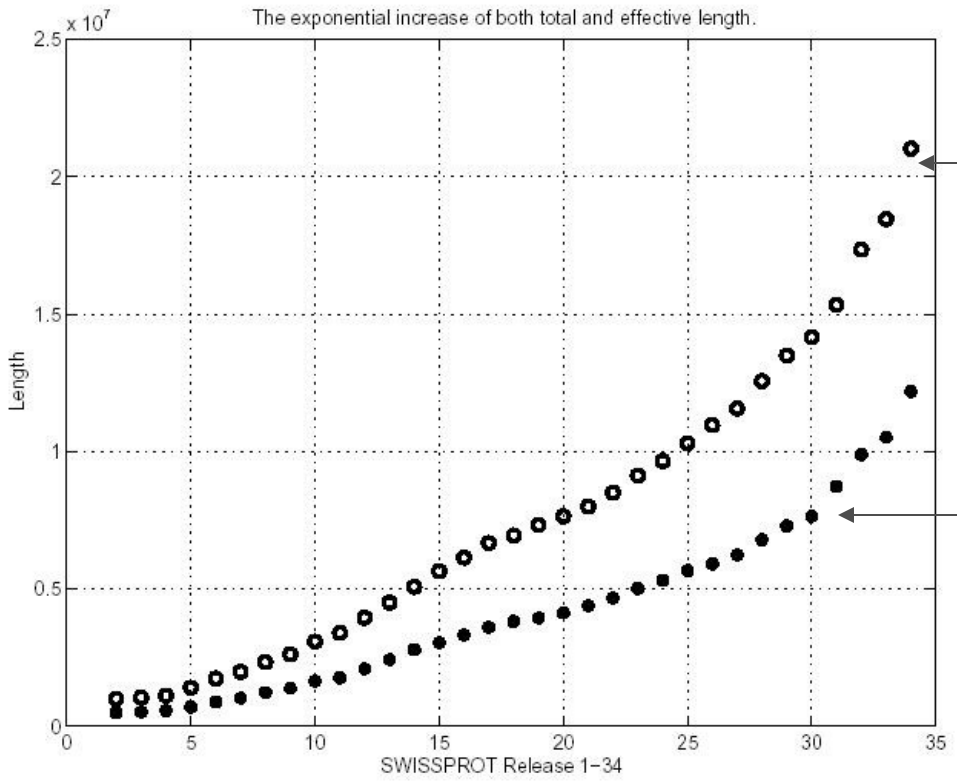
Search 1000 randomly generated query sequences against releases 1-34 of the protein database SWISSPROT.

For each release we obtain a sample of 1000 extreme value distributed scores.

We estimate the location parameter for each release.

We expect it to grow.

Increase of noise in the SWISSPROT database between 1985 and 1996



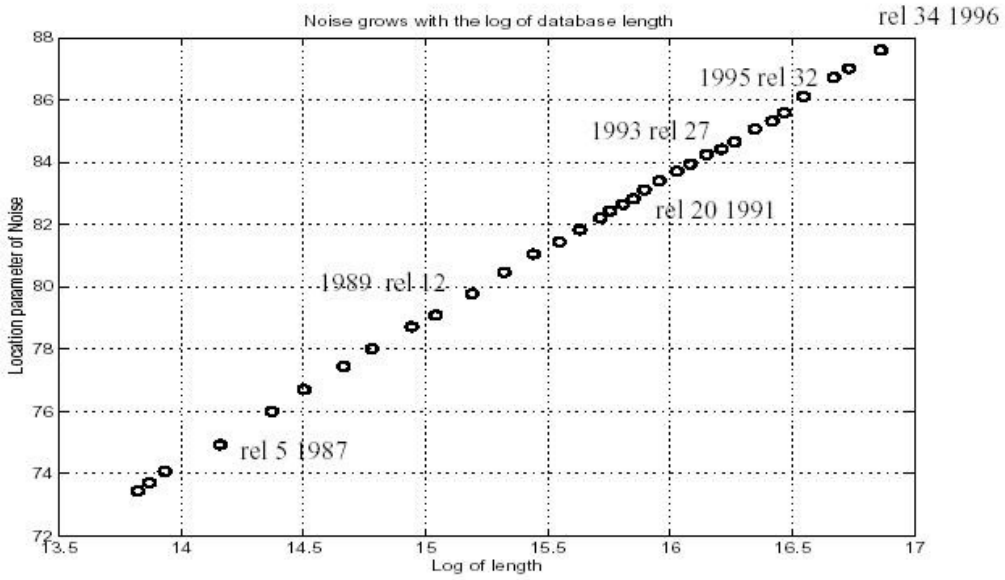
Length of SWISSPROT

Location parameter of the database noise

Our regression equation suggests that the noise grows with the log of the lengths of the database releases

$$H^{max} \sim \alpha_0 + \theta \log(L) + \theta G$$

And it does perfectly:



Summary

- **Biological sequences often show local rather than global similarities.**
- **In two sequences, local similarities can occur by chance.**
- **We need to distinguish between biologically relevant and chance (random) similarities.**
- **Looking at the distribution of local alignment scores of random sequences helps judging the relevance of an alignment.**
- **Scores from random sequences are extreme value distributed.**
- **Scores depend on sequence length, we can correct for this.**
- **Scores depend on database size, we can correct for this.**
- **The E-value reports the expected number of scores equal or higher than t by chance.**

End of Chapter 10

Appendix: probability density function and distribution function

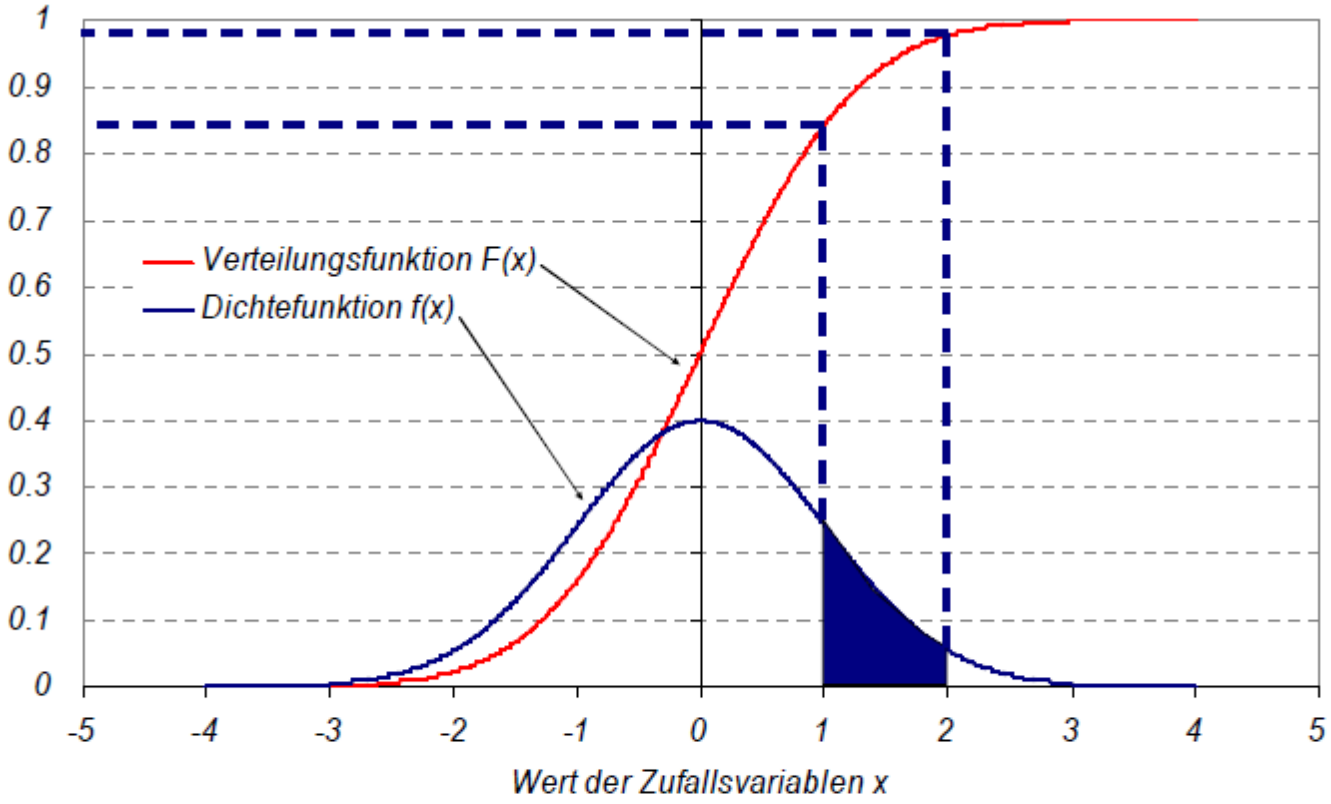


Abb. 1-2 Ableitung von Aussagen zur Auftretenswahrscheinlichkeit aus Dichte- und Verteilungsfunktion am Beispiel der Normalverteilung mit dem Mittelwert $\mu=0$ und der Standardabweichung $\sigma=1$