Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 10

1

TODAY'S TOPICS

- Assignment 2 Review
- Amino Acid Substitution Matrix
  - Information Theory
- Statistical Significance of Alignment Scores
- Assignment 5

2

Log Odds Matrix

- **PAM matrix**

$S(i,j) = 10\log_{10}\{M(i,j)/f(i)\}$

$M(i,j)$: Mutation probability from AA_j to AA_i
$f(i)$: Frequency of AA_i (number of AA_i / total number of residues)
Probability to find AA_i by chance

- **BLOSUM matrix**

$S(i,j) = 2\log_2(q_{ij}/e_{ij})$

q_{ij}: Observed frequency of AA_i, AA_j pairs
e_{ij}: Expected frequencies of AA_i, AA_j pairs

- **General form**

$S(i,j) = 1/\lambda \log_2(q_{ij}/p_ip_j)$ [in bit unit]
$S(i,j) = 1/\lambda \log_e(q_{ij}/p_ip_j)$ [in nat unit]

3

Log Odds Matrix

	AA_1	AA_2	
AA_1	S_{11}	S_{21}	
AA_2	S_{12}	S_{22}	

- **Log odds (Lod) score**: general
  also called log odds ratio or log likelihood ratio

$S(i,j) = 1/\lambda \log_2(q_{ij}/p_ip_j)$ [in bit unit]
$S(i,j) = 1/\lambda \log_e(q_{ij}/p_ip_j)$ [in nat unit]

If λ=1/2, $S(i,j) = 2\log_2(q_{ij}/p_ip_j)$ [a half-bit unit]

q_{ij}: the frequency of the amino acid pair, AA_i and AA_j
p_i, p_j: the individual frequency of AA_i or AA_j
λ : a scaling factor

($1/\lambda$=2 is used with BLOSUM)

$$S(i,j)=1/\lambda\log\{\frac{\text{Observed freq. of amino acid pair } i \text{ and } j}{\text{Expected freq. of amino acid pair } i \text{ and } j}\}$$

* In the general format, substitutions does not have to be symmetrical. $S_{12} = S_{21}$ is not assumed.

4

Log Odds Matrix

	AA_1	AA_2	
AA_1	S_{11}	S_{21}	
AA_2	S_{12}	S_{22}	

- **Log odds (Lod) score**: general
  also called log odds ratio or log likelihood ratio

$S(i,j) = 1/\lambda \log_2(q_{ij}/p_ip_j)$ [in bit unit]
$S(i,j) = 1/\lambda \log_e(q_{ij}/p_ip_j)$ [in nat unit]

Target frequency (q_{ij})

$$S(i,j)=1/\lambda\log\{\frac{\text{Observed freq. of amino acid pair i,j}}{\text{Expected freq. of amino acid pair i,j}}\}$$

$[-<S(i,j)<+]$

Background frequency (p_ip_j)

H_1: Homologous hypothesis (residues i and j are related)
H_0: Random hypothesis (residues i and j are unrelated)

5

Log Odds Matrix

	AA_1	AA_2	
AA_1	S_{11}	S_{21}	
AA_2	S_{12}	S_{22}	

- **Log odds (Lod) score**: general
  also called log odds ratio or log likelihood ratio

$S(i,j) = 1/\lambda \log_2(q_{ij}/p_ip_j)$ [in bit unit]
$S(i,j) = 1/\lambda \log_e(q_{ij}/p_ip_j)$ [in nat unit]

$$\text{Likelihood ratio (LR)} = \frac{\text{Likelihood of } H_1}{\text{Likelihood of } H_0}$$

$$[0<LR<+\text{inf}] \quad = \frac{\text{Prob(an event}|H_1)}{\text{Prob(an event}|H_0)}$$

H_1: Hypothesis to be tested, H_0: Null hypothesis

6

Log Odds Matrix

	AA_1	AA_2
AA_1	s_{11}	s_{21}
AA_2	s_{12}	s_{22}

- **Log odds (Lod) score**: general
  also called log odds ratio or log likelihood ratio

$S(i,j) = 1/\lambda \log_2(q_{ij}/p_ip_j)$ [in bit unit]
$S(i,j) = 1/\lambda \log_e(q_{ij}/p_ip_j)$ [in nat unit]

$$\text{Log likelihood ratio} = \log\{\frac{\text{Likelihood of } H_1}{\text{Likelihood of } H_0}\}$$
$$= \log\{\text{Prob(an event}|H_1)\} - \log\{\text{Prob(an event}|H_0)\}$$
$$|- < \log(LR) < +|$$

H_1: Hypothesis to be tested, H_0: Null hypothesis

7

Log Odds Score and Target Frequencies

$$S(i,j) = 1/\lambda \log_e(q_{ij}/p_ip_j)$$
[or $S(i,j) = 1/\lambda \log_2(q_{ij}/p_ip_j)$ for BLOSUM]

$$\lambda S(i,j) = \log_e(q_{ij}/p_ip_j)$$
$$e^{\lambda S(i,j)} = q_{ij}/p_ip_j$$
$$q_{ij} = p_ip_je^{\lambda S(i,j)}$$

Target frequency ← q_{ij} → Expected (or background) frequency

$$\Sigma_i\Sigma_j\, q_{ij} = \Sigma_i\Sigma_j\, p_ip_j\, e^{\lambda S(i,j)} = 1$$
$$(i<j)$$

λ can be estimated (matrix specific)

8

Relative Entropy (H)

	AA_1	AA_2
AA_1	s_{11}	s_{21}
AA_2	s_{12}	s_{22}

- **Expected Score (E)**

$E = \Sigma_i\Sigma_j\, p_ip_jS(i,j)$ [p_i, p_j: expected freq. of AA_i, AA_j]

- **Relative Entropy (H)**

This is "ENTROPY" in Information Theory; Completely unrelated to "entropy" in thermodynamics!!

$H = \Sigma_i\Sigma_j\, q_{ij}\, \lambda S(i,j)$ [q_{ij} is observed freq. of AA_i, AA_j pair]

- the average information per residue pair
- summarizes the behavior of the scoring matrix
- the ability of the matrix to discriminate related from unrelated (nonrandom matching from random matching)
  - → $H = 0$ when target distribution equals to background distribution
    [If $q_{ij}=p_ip_j$, $S(i,j) = 1/\lambda\log_e(q_{ij}/p_ip_j) = 1/\lambda\log_e(1) = 0$]
  - → H increases when the two distributions become more distinguishable

→ can be used to compare scoring matrices

9

Introduction to Information Theory

- **Information:**
  a decrease in uncertainty (unpredictability, a degree of surprise)
  - If you are asking questions to somebody...
    - → if you can guess every answer correctly
    - → there is no surprise
    - → you cannot gain any new information
    - → but if you have no idea what answer you get
    - → every answer is a surprise
    - → you gain a lot of information

Information Theory Primer by Tom Schneider (also on Canvas):
http://users.fred.net/tds/lab/papers/primer/

10

Introduction to Information Theory

- **Information:**
  a decrease in uncertainty (unpredictability, a degree of surprise)

Device ⟹ A, A, A, A, A, A, A, A, ...

Prob(A) = 1
Only one possible symbol

No surprise
No information

11

Introduction to Information Theory

- **Information:**
  a decrease in uncertainty (unpredictability, a degree of surprise)

Device ⟹ A, B!, A, B!, A, B, A, B, ...

Prob(A) = 0.5
Prob(B) = 0.5

Two possible symbols
A little surprise
A small amount of information

12

Introduction to Information Theory

➢ **Information:**
a decrease in uncertainty (unpredictability, a degree of surprise)

Device ⟹ A, B!, C!!, A, B!, C!!, A, B, ...

Prob(A) = 0.33
Prob(B) = 0.33
Prob(C) = 0.33

Three possible symbols
More surprise, More information

13

Introduction to Information Theory

➢ **Information:**
a decrease in uncertainty (unpredictability, a degree of surprise)

Device ⟹ A, A, A, A, A, B!!!, A, A, ...

Prob(A) = 7/8
Prob(B) = 1/8

Two possible symbols
Big surprise! A lot more information...?

But not much surprise in getting the symbol A's

14

Introduction to Information Theory

➢ Information is a decrease in uncertainty
- Surprising answers convey more information!

➜ If each symbol is equally likely,
➜ the amount of information increases with the number of different symbols.

➜ The amount of information, or surprise of an answer, is inversely proportional to its probability.

$I(p) = \log_2(1/p)$ or $I(p) = -\log_2 p$

I: information, p: probability

15

Introduction to Information Theory

➢ Bits: the unit for values converted to base 2 logarithms (nats: the unit if base e is used)

$$I(p) = -\log_2 p$$

- If an answer is highly unexpected (*e.g.*, $p = 0.1$),
➜ $I(0.1) = -\log_2 0.1 = 3.3$ bits (more information)
- For a very much expected answer (*e.g.*, $p = 0.9$),
➜ $I(0.9) = -\log_2 0.9 = 0.15$ bits (less information)
- If there is only one possible answer (symbol):
➜ $p = 1$, $I(p) =$?

16

Introduction to Information Theory

➢ Information can be represented by a series of symbols each with a certain probability:
- Shannon Entropy: the average information per symbol
$H = -\sum p_i(\log_2 p_i)$
- If all n symbols are equally possible (p_i is the same)
➜ $H = -\sum p(\log_2 p) = -(np \times \log_2 p)$
$= -\log_2 p$, since $np = 1$
$= -\log_2(1/n)$, since $p = 1/n$
$= \log_2(n)$

$H(1) = \log_2(1) = 0$ bit, $H(2) = \log_2(2) = 1$, $H(4) = \log_2(2^2) = 2$

18

Introduction to Information Theory

➢ Information can be represented by a series of symbols each with a certain probability:
- Shannon Entropy: the average information per symbol
$H = -\sum p_i(\log_2 p_i)$
- For a random DNA sequence: ATGC ($p = 0.25$ for all)
$H = -(0.25 \times 4) \times \log_2(0.25)$ or $\log_2(4) = 2$ bits
- For a AT-rich DNA sequence: $p_A = p_T = 0.45$ and $p_G = p_C = 0.05$
$H = \{-0.45 \times (\log_2 0.45)\} + \{-0.45 \times (\log_2 0.45)\} +$
$\{-0.05 \times (\log_2 0.05)\} + \{-0.05 \times (\log_2 0.05)\}$
$= \{-0.45 \times (-1.15)\} \times 2 + \{-0.05 \times (-4.32)\} \times 2 = 1.47$ bits

19

Relative Entropy (H)

- **Expected Score (E)**

  $E = \Sigma_i\Sigma_j\, p_i p_j S(i,j)$ [p_i, p_j: expected freq. of AA_i, AA_j]

- **Relative Entropy (H)**

  $H = \Sigma_i\Sigma_j\, q_{ij}\, \lambda S(i,j)$ [q_{ij} is observed freq. of AA_i, AA_j pair]

  Since $S(i,j) = 1/\lambda \log_2(q_{ij}/p_ip_j)$ or $1/\lambda \log_e(q_{ij}/p_ip_j)$

  $H = \Sigma_i\Sigma_j\, q_{ij} \log_2(q_{ij}/p_ip_j)$ or $\Sigma_i\Sigma_j\, q_{ij} \log_e(q_{ij}/p_ip_j)$

  $= \Sigma_i\Sigma_j\{q_{ij}\log_2(q_{ij}) - q_{ij}\log_2\{p_ip_j)\}$

  Connection to $H = -\Sigma p_i(\log_2 p_i)$

Note: Both Expected Score and Relative Entropy have their units in bit or nat.

20

Comparing Scoring Matrices

- **Relative Entropy (H) of a scoring matrix**

  $H = \Sigma_i\Sigma_j q_{ij}\lambda S(i,j)$ [q_{ij} is observed freq. of AA_i, AA_j pair]

  - the average information per residue pair for a scoring matrix
  - summarizes the behavior of the scoring matrix
  - the ability of the matrix to discriminate related from unrelated (nonrandom matching from random matching)
    - → H=0 when target distribution equals to background distribution [If $q_{ij}=p_ip_j$, $S(i,j) = 1/\lambda\log_e(q_{ij}/p_ip_j) = 1/\lambda\log_e(1) = 0$]
    - → H increases when the two distributions become more distinguishable

→ can be used to compare scoring matrices

21

Comparing Scoring Matrices

- **Relative Entropy (H)**

  H(PAM1) = 4.17 bits
  H(PAM50) = 2.00
  H(PAM120) = 0.98
  H(PAM160) = 0.70 ←
  H(PAM250) = 0.36
  from Altschul (1991)

H decreases with increasing PAM;
H increases with increasing BLOSUM

Higher BLOSUM is generated including sequences that are more similar to one another
↓
More amino acid pairs are used

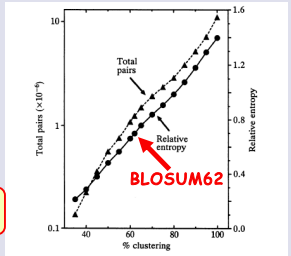


FIG. 1. Relationship between percentage clustering and total amino acid pair counts plotted on a logarithmic scale and relative entropy.

from Henikoff and Henikoff (1992)

22

Comparing Scoring Matrices

- **Relative Entropy (H) of a scoring matrix**

  $H = \Sigma_i\Sigma_j q_{ij}\lambda S(i,j)$

  - The average information per residue pair for a scoring matrix
  - Decreases with increasing PAM: H(PAM1) > H(PAM120)
    - → All PAMn is extrapolated from PAM1
    - → Higher PAM is less specific, contains less information, thus has a lower H
  - Increases with increasing BLOSUM: H(BLOSUM45) < H(BLOSUM80)
    - → Higher BLOSUM is generated using more data (fewer information is eliminated), thus has a higher H

[e.g., BLOSUM100 is generated using the threshold 100%; only identical sequences are down-weighted for calculation]

23

BLOSUM and PAM matrices

(default in BLAST)

BLOSUM80	BLOSUM62	BLOSUM45
Less divergent	←	More divergent
PAM120	PAM160	PAM250
H = 0.98 bits	0.7 bits	0.36 bits

24

BLOSUM62

```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =   0.6979, Expected =  -0.5209
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3  . . .
```

$S(i,j) = 2\log_2(q_{ij}/e_{ij})$

https://ftp.ncbi.nlm.nih.gov/blast/matrices/

25

Selecting the Right Similarity-Scoring Matrix

William R. Pearson[1] *Current Protocols in Bioinformatics* (2013)

Relative entropy (*H*)

Table 3.5.1 Scoring Matrix Target Identity, Information Content, and Alignment Length[a]

Matrix	Gap penalty[b]	% Identity	Bits/ position	Random alignment length	50-bit length
SSEARCH version 36.3.6					
BLOSUM50[c]	10/2	25.3	0.21	160	238
BLOSUM62	11/1	28.9	0.40	86	125
VTML 160[c,d]	12/2	23.9	0.25	139	200
VTML 140	10/1	28.4	0.44	82	114
VTML 120	11/1	32.1	0.54	62	93
VTML 80	10/1	40.5	0.74	47	68
VTML 40	13/1	64.7	1.92	18	26
VTML 20	15/2	86.1	3.30	11	15
VTML 10	16/2	90.9	3.87	9	13
BLAST version 2.2.27+					
BLOSUM50[c]	13/2	29.4	0.39	85	128
BLOSUM62	11/1	29.6	0.41	82	122
BLOSUM80	10/1	32.0	0.48	69	104
PAM70	10/1	33.9	0.58	56	86
PAM30	9/1	45.9	0.90	34	56

(VTML: PAM type)

- Default matrices (*e.g.*, BLOSUM62) are good for identifying <25% identity.
- Deep scoring matrices (*e.g.*, BLOSUM62, PAM250) require long sequence alignment to achieve significant scores (*e.g.*, >50 bits).
- They are more likely to extend alignments outside of homologous region.

26

Selecting the Right Similarity-Scoring Matrix

William R. Pearson[1] *Current Protocols in Bioinformatics* (2013)

VTML 20

	A	R	N	D	C	Q	E
A	7						
R	-7	8					
N	-6	-5	8				
D	-6	-12	-1	8			
C	-3	-7	-8	-14	12		
Q	-5	-2	-4	-4	-13	9	
E	-5	-10	-5	-1	-14	-1	7

BLOSUM62

	A	R	N	D	C	Q	E
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

Figure 3.5.2 Comparison of a "shallow" (VTML 20) and "deep" (BLOSUM62) scoring matrix. Both matrices are scaled in 1/2-bits. For the small part of the matrices shown here, the VTML20 matrix produces an average 2.80 half-bit identity score, and an average −0.59 nonidentical score (weighted by amino-acid abundance). In contrast, BLOSUM62 produces 1.86 for identities but only −0.06 for nonidentities. Thus, VTML20 targets shorter, higher-identity alignments, because it penalizes nonidentities much more strongly.

- Short alignments require shallow scoring matrices.
- Shallower scoring matrices (*e.g.*, PAM20) are more effective when searching over shorter evolutionary distances.

27

CORRESPONDENCE VOLUME 26 NUMBER 3 MARCH 2008 NATURE BIOTECHNOLOGY

BLOSUM62 miscalculations improve search performance

RBLOSUM

Mark P Styczynski[1,5,6], Kyle L Jensen[2,3,6], Isidore Rigoutsos[4] & Gregory Stephanopoulos[1]

Hess *et al. BMC Bioinformatics* (2016) 17:189
DOI 10.1186/s12859-016-1060-3

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Addressing inaccuracies in BLOSUM computation improves homology search performance

CorBLOSUM

Martin Hess[1,2†], Frank Keul[2*†], Michael Goesele[1] and Kay Hamacher[2]

Govindarajan *et al. BMC Res Notes* (2018) 11:328
https://doi.org/10.1186/s13104-018-3415-5

BMC Research Notes

RESEARCH NOTE Open Access

RBLOSUM performs better than CorBLOSUM with lesser error per query

Renganayaki Govindarajan*, Biji Christopher Leela and Achuthsankar S. Nair

28

Substitution matrices for specific proteins

Keul *et al. BMC Bioinformatics* (2017) 18:293
DOI 10.1186/s12859-017-1703-z

BMC Bioinformatics

RESEARCH ARTICLE Open Access

PFASUM: a substitution matrix from Pfam structural alignments

Frank Keul[1†], Martin Hess[2*†], Michael Goesele[2] and Kay Hamacher[1]

BIOINFORMATICS Vol. 27 ISMB 2011, pages i15–i23
doi:10.1093/bioinformatics/btr230

Environment specific substitution tables improve membrane protein alignment

Jamie R. Hill[1], Sebastian Kelm[1], Jiye Shi[2,3] and Charlotte M. Deane[1,*]

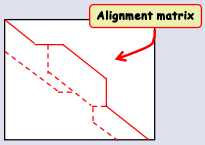
More substitution matrices reviewed in Trivedi & Nagarajaram (2020)

29

Pairwise alignment summary

- **Alignment score** depends on:
  - → **Scoring matrix** (match, mismatch, Ts/Tv, BLOSUM, PAM, *etc.*)
  - → **Gap penalty**
  - → **Alignment method** (*e.g.*, global or local)
- **Alignment scores cannot be compared directly**
  - → if the scoring systems used are different
  - → if sequences compared are different

  (*e.g.*, longer alignments tend to have higher scores)
- **Alignment scores are used:**
  for searching optimal alignments
  from the alignment matrix
  for a given pair of sequences
  based on a given scoring system



30

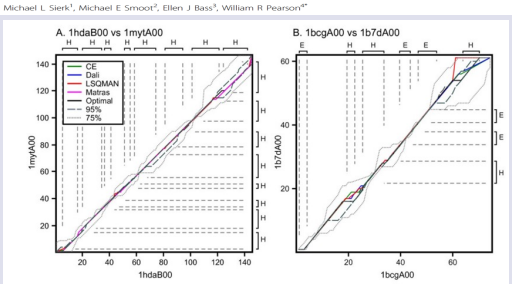
Sierk *et al. BMC Bioinformatics* 2010, **11**:146
http://www.biomedcentral.com/1471-2105/11/146

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments

Michael L Sierk[1], Michael E Smoot[2], Ellen J Bass[3], William R Pearson[4*]



Alignment paths of structure-based, and optimal and suboptimal sequence alignments. Two

31

Pairwise alignment summary

- Optimal alignments and biologically meaningful alignments may not be the same
- Depending on the scoring system, unreasonable alignments can become optimal

→ We need to choose a better (biologically reasonable) scoring system: level of divergence (scoring matrices), gap penalty (affine, *etc.*), algorithm (local, global, or semi-global)

→ Manual adjustment may be necessary

→ Test statistical significance of the alignment (is the alignment possible just by chance?)

32

Significance of Alignment Scores

➢ Hypothesis testing (General)

→ Two hypotheses

- Null-hypothesis
  H_0: The previous (original) belief is true
- Alternative hypothesis
  H_1: The previous (original) belief is false; the new theory is true

→ S: Test statistic

→ Significance level is chosen a priori (*e.g.*, 0.05)

→ P-value: $P(S|H_0 \text{ is true})$ Probability of getting S if H_0 is true

→ If $P <$ Significance level, reject H_0

33

Significance of Alignment Scores

➢ P-value: $P(S|H_0 \text{ is true})$

→ Need to be calculated from the test statistic S

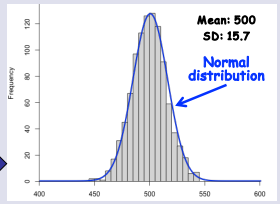
→ Need to know the probability distribution of the test statistic S under H_0

Central Limit Theorem:
If the sample size is large enough, the sampling distribution of the mean of any independent, random variables will be normal or nearly normal.

(Example)

Experiment: 1000 coin tossing
- Count the number of heads
- Repeat 1000 experiments

(Expect to see 500 heads/experiment)



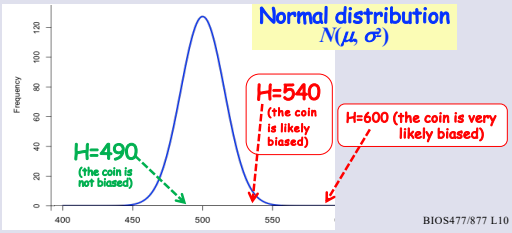
34

Significance of Alignment Scores

➢ P-value: $P(S|H_0 \text{ is true})$

→ Need to be calculated from the test statistic S

→ Need to know the probability distribution of the test statistic S under H_0



35

Significance of Alignment Scores

➢ Hypothesis testing for sequence alignment

→ Two hypotheses

- Null-hypothesis
  H_0: Two sequences are not related (random)
- Alternative hypothesis
  H_1: Two sequences are related

→ Test statistic: alignment score (S)

→ Significance level is chosen a priori (*e.g.*, 0.05)

→ P-value: $P(S|H_0 \text{ is true})$
Probability of getting the alignment score S, even if the two sequences are not related but randomly matched

→ If $P <$ Significance level, reject H_0
(The score should not be obtained just by aligning unrelated sequences)

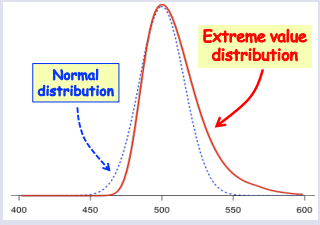
36

Significance of Alignment Scores

➢ P-value: $P(S|H_0 \text{ is true})$

→ Need to be calculated from the test statistic S

→ Need to know the probability distribution of the test statistic S under H_0

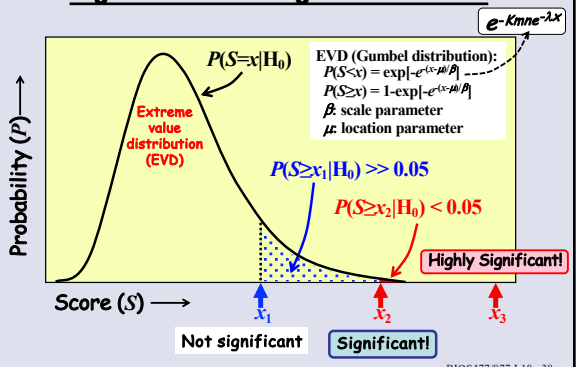


Distribution of alignment scores follow Extreme Value Distribution (Gumbel distribution)

The probability distribution of highest values in an experiment (*e.g.*, optimal alignment scores)

37

Significance of Alignment Scores



38

Significance of Alignment Scores

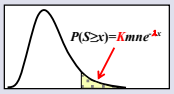
- $P(S\geq x|H_0)$: Probability of getting the alignment score $S \geq x$

Karlin-Altschul equation (Karlin and Altschul 1990)

$P(S\geq x) = 1 - \exp[-Kmne^{-\lambda x}] \approx Kmne^{-\lambda x}$

EVD (Gumbel distribution):
$P(S\geq x) = 1-\exp[-e^{-(x-\mu)/\beta}]$
$\lambda = 1/\beta$, $\mu = (\ln Kmn)/\lambda$

K and λ: calculated from the empirical distribution of S based on a given scoring matrix and amino acid composition
m and n: lengths of sequences aligned



→ Solved for ungapped local alignments
→ Can be applied for gapped local alignments

- E-value = $P(S\geq x|H_0) \times N$,
where N is the number of sequences in the dataset
→ Expected number of sequences in the dataset to have a score $\geq x$

E-value ≠ P-value

39

Significance of Alignment Scores

- How to calculate K and λ (in LALIGN and PRSS)
→ estimated from an empirical probability distribution.

1) The second sequence is shuffled many times.
(simulates random sequences)

2) Smith-Waterman local alignment score is calculated from each alignment: $P(S\geq x|H_0)$

3) The distribution is fitted to an extreme value distribution to obtain estimates of K and λ

4) P-value is estimated based on the K and λ, and the original alignment score x: $P(S\geq x) \approx Kmne^{-\lambda x}$

40

Simulation of Alignment Scores

- RECA_ECOLI (P0A7G6; 353 amino acids)
- RAD51_YEAST (P25454; 400 amino acids)

Smith-Waterman local alignment score = 293
(BLOSUM50, gap opening: -10, gap extension: -1)

```
RECA_ECOLI      3 IDENK-QRALAAALGQIEKQFGKGSIMRLGEDRSMDVETISTGSLSLDIA   51
                  |.:.|  |.|.....|.....|.....|.....|.....|||.:||.||
RAD51_YEAST   124 ISEARADKLLNEAARLVPMGFPTAADFHM---RRSELICLYTGSEWLDTL   170

RECA_ECOLI     52 LGAGGLPMGRIVEIYGPESSGKTTL------TLQV-IAAAQREGKTCAFI    94
                  || ||:..|.|.|::|...:||:.|      |.|: :.....||| |.:|
RAD51_YEAST   171 LG-GGVETGSITELFGEFRTGKSQLCHTLAVTCQIPLDIGGGEGK-CLYI   218

RECA_ECOLI     95 DAEHALDPI----YARKLGVDID-----NLLCSQPDTGEQALEICDALAR-   135
                  |.|...|:    .|:::|:|.|     |:..::|....::|.:|.|.|:
RAD51_YEAST   219 DTEGTFRPVRLVSIAQRFGLDPDDALNNVAYARAYNADHQLRLLDAAAQM   268

RECA_ECOLI    136 --SGAVDVIVVDSVAALTPKAEI--EGEIGDSHMGLAARMMSQAMRKLAG   181
                  .......|||||.||. :.:.   .||:....|.||..|  :|:::||
RAD51_YEAST   269 MSERPSLIVVDSVMALY-RTDFSGRGELSARQMHLAKFM--RALQRLA-   314

RECA_ECOLI    182 NLKQSNTLLIFINQIRMKI--GVMFG-NPETTTGGNALKFYASVRLDIRR   228
                  .|....:..||:...:   |:.|. :|:...|||.:.....::.||
RAD51_YEAST   315 --DQFGVAVVVTNQVVAQVDGGMAFNPDPKKPIGGNIMAHSSTTRL----   358

RECA_ECOLI    229 IGAVKEGENVVGSETRVKVVKNKIAAPFKQAEFQILYGEGI    269
                  ..|:|:   |....|||.:....|....|.| |.:|:
RAD51_YEAST   359 --GFEKGK---GCQRLCKVV-DSPCLPEAECVFAI-YEDGV    392
```

https://www.ebi.ac.uk/idispatcher/psa/emboss_water

41

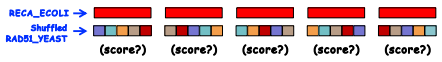
Simulation of Alignment Scores

- RECA_ECOLI (P0A7G6; 353 amino acids)
- RAD51_YEAST (P25454; 400 amino acids)

Smith-Waterman local alignment score = 293
(BLOSUM50, gap opening: -10, gap extension: -1)

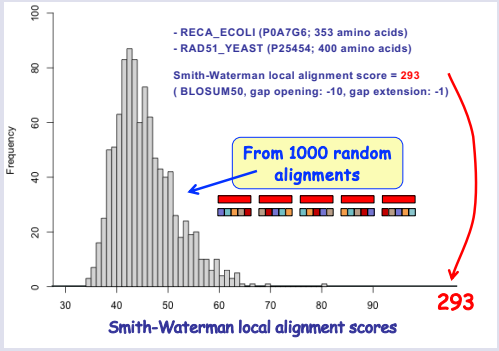
Shuffle RAD51_YEAST 1000 times
(generate 1000 random sequences)

Align with RECA_ECOLI
(generate 1000 random local alignments)



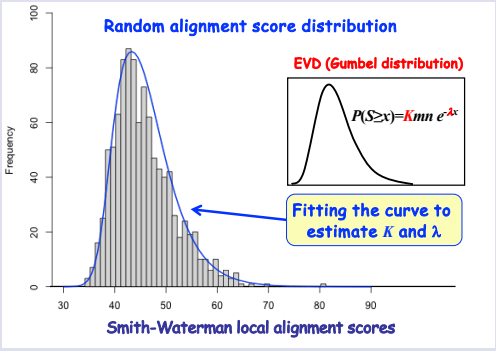
42

Simulation of Alignment Scores



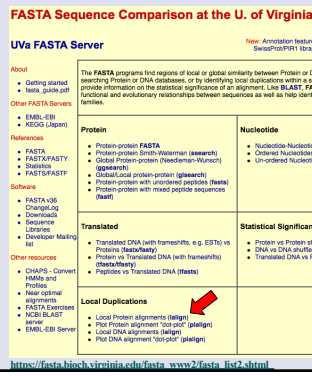
43

Simulation of Alignment Scores



44

FASTA Web server by William Pearson



https://fasta.bioch.virginia.edu/wrpearson/

Original FASTA package was released on 1988 (earlier than BLAST)

The origin of FASTA format

45