BIOS 477/877
Bioinformatics and Molecular Evolution

Instructor: Etsuko Moriyama
(School of Biological Sciences)

Spring 2026 **Lecture 14**

1

Today's topics

- BLAST and FASTA statistics
- Multiple Alignment
  - Introduction
- Assignment 7

2

blastp statistics: bit score

glutamate synthase [NADPH] large chain-like [Nerophis lumbriciformis]
Sequence ID: XP_061781277.1 Length: 1377 Number of Matches: 1

Range 1: 755 to 1081 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
219 bits(557)	5e-58	Compositional matrix adjust.	134/330(41%)	183/330(55%)	9/330(2%)

```
Query  183   MSYGALSLNAHLSFAKAVKECGTFMGTGEGGLPKALY----PYADH----IITQVASGRFGV  236
             MS GALS  AH +   A+   G    +GEGG   A +   P  D+    I QVASGRFGV
Sbjct  755   MSLGALSPEAHKTLNVAMNRIGAKSDSGEGGEDPAHFVPEPNGDNPSAKIKQVASGRFGV  814

Query  237   NEEYLMKGSAIEIKIGQGAKPGIGGHLPGEKVTAEISATRMIPEGSDAISPAPHHDIYSI  296
               EYL    +EIK+ QGAKPG GG LPG KVT  I+  R   +G   ISP PHHDIYSI
Sbjct  815   TAEYLNHCEELEIKVAQGAKPGEGGQLPGIKVTDLIARLRHSTKGVTLISPPPHHDIYSI  874
```

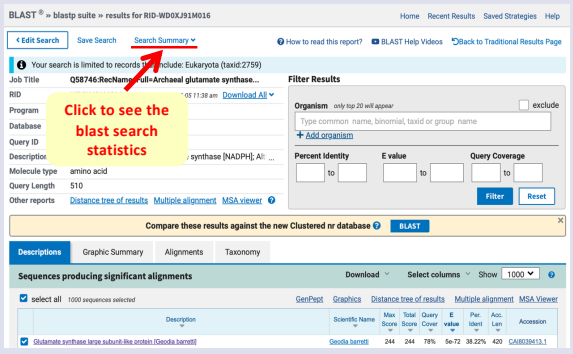
Raw Score (S): simply based on pairwise scores & gap penalties
Normalized Score or **Bit Score (S'_{bit}):**

$$S'_{bit} = (\lambda S - \log_e K) / \log_e 2, \quad [S'_{nat} = \lambda S - \log_e K]$$

λ and K are specific to the scoring system used (scoring matrix, gap penalties, etc.)
→ Where can we find the values for λ and K ?

3

blastp search summary



(Top portion of any blast output)

4

blastp search summary

Search Parameters	
Program	blastp
Word size	5
Expect value	0.05
Hitlist size	1000
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	0
Composition-based stats	2

Database	
Posted date	Feb 13, 2026 2:53 PM
Number of letters	34,352,650,388
Number of sequences	82,371,451
Entrez query	Includes: Eukaryota (taxid:2759)

Karlin-Altschul statistics		
Lambda	0.31987	0.267
K	0.137272	0.041
H	0.416015	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

Used to calculate bit scores and E-values for the alignments with gaps

5

blastp statistics: bit score

glutamate synthase [NADPH] large chain-like [Nerophis lumbriciformis]
Sequence ID: XP_061781277.1 Length: 1377 Number of Matches: 1

Range 1: 755 to 1081 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
219 bits(557)	5e-58	Compositional matrix adjust.	134/330(41%)	183/330(55%)	9/330(2%)

```
Query  183   MSYGALSLNAHLSFAKAVKECGTFMGTGEGGLPKALY----PYADH----IITQVASGRFGV  236
             MS GALS  AH +   A+   G    +GEGG   A +   P  D+    I QVASGRFGV
Sbjct  755   MSLGALSPEAHKTLNVAMNRIGAKSDSGEGGEDPAHFVPEPNGDNPSAKIKQVASGRFGV  814

Query  237   NEEYLMKGSAIEIKIGQGAKPGIGGHLPGEKVTAEISATRMIPEGSDAISPAPHHDIYSI  296
               EYL    +EIK+ QGAKPG GG LPG KVT  I+  R   +G   ISP PHHDIYSI
Sbjct  815   TAEYLNHCEELEIKVAQGAKPGEGGQLPGIKVTDLIARLRHSTKGVTLISPPPHHDIYSI  874
```

Raw Score (S): simply based on pairwise scores & gap penalties
Normalized Score or **Bit Score (S'_{bit}):**

$$S'_{bit} = (\lambda S - \log_e K) / \log_e 2, \quad [S'_{nat} = \lambda S - \log_e K]$$

$\lambda = 0.267$, $K = 0.041$, $S'_{bit} = \{0.267 \times 557 - \log_e(0.041)\} / \log_e 2 = 219.2$

λ and K are scoring system specific

Karlin-Altschul statistics		
Lambda	0.31987	0.267
K	0.137272	0.041
H	0.416015	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

(for gapped alignments)

6

blastp statistics: bit score

glutamate synthase [NADPH] large chain-like [Nerophis lumbriciformis]
Sequence ID: XP_061781277.1 Length: 1377 Number of Matches: 1

Range 1: 755 to 1081 GenPept Graphics ▼ Next Match ▲ Previous

Score	Expect	Method	Identities	Positives	Gaps
219 bits(557)	5e-58	Compositional matrix adjust.	134/330(41%)	183/330(55%)	9/330(2%)

```
Query  143   MSYGALSLNAHLSFAKAVKECGTFMGTGEGGLPKALY----PYADH----IITQVASGRFGV  236
             MS GALS   AH +    A+    G    +GEGG     A +    P  D+     I  QVASGRFGV
Sbjct  755   MSLGALSPEAHKTLNVAMNRIGAKSDSGEGGEDPAHFVPEPNGDNPSAKIKQVASGRFGV  814

Query  237   NEEYLMKGSAIEIKIGDGAKPGIGGHLPGEKVTAEISATRMIPEGSDAISPAPHHDIYSI  296
               EYL     +EIK+ QGAKPG GG LPG KVT  I+  R    +G   ISP PHHDIYSI
Sbjct  815   TAEYLNHCEELEIKVAQGAKPGEGGQLPGIKVTDLIARLRHSTKGVTLISPPPHHDIYSI  874
```

Raw Score (S): simply based on pairwise scores & gap penalties

Normalized Score or Bit Score (S'_{bit}):

$$S'_{bit} = (\lambda S - \log_e K) / \log_e 2, \quad [S'_{nat} = \lambda S - \log_e K]$$

$\lambda = 0.267$, $K = 0.041$, $S'_{bit} = \{0.267 \times 557 - \log_e(0.041)\} / \log_e 2 = 219.2$

- Raw scores (S) depend on the scoring system; cannot be compared
- Bit scores (S'_{bit}) are normalized using λ and K
  → independent of scoring system; can be compared

7

Pairwise alignment vs. database searching

[For a pairwise alignment]

➢ Probability of getting the alignment score $S \geq x$ by chance

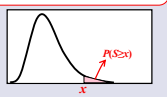
$P(S \geq x) = 1 - \exp[-Kmne^{-\lambda x}] \approx Kmne^{-\lambda x}$ ⟵ $e^a \approx 1+a$, where $a = -Kmne^{-\lambda x}$ — Taylor series approximation → works if a is small (<<1)

Karlin-Altschul equation (Karlin & Altschul, 1990)

[For database searching]

➢ Multiple pairwise alignments

→ Multiple testing problem



- $P(S \geq x)$: Probability of getting the alignment score (S) equal to or larger than x by chance from one pairwise alignment
  - If $P(S \geq x) = 0.05$, $P(S<x) = 1-P(S \geq x) = 0.95$
    → 0.95 is the prob. to have one pairwise alignment with $S<x$ by chance
  - For 10 alignments, $0.95^{10} \approx 0.60$ is the prob. to have all 10 alignments with $S<x$
    → 1-0.60 = 0.40 is the prob. to have at least one alignment with $S \geq x$ by chance
  - For 100 alignments, $0.95^{100} \approx 0.006$ is the prob. to have all 100- alignments with $S<x$
    → 1-0.006 ≈ 0.99 is the prob. to have at least one alignment with $S \geq x$ by chance

Pr=0.05 as the significance level is not good enough if many alignments need to be tested!

8

Bonferroni correction

➢ Multiple comparison correction

- Instead of using Prob = α as the threshold, use Prob = α/N (for N comparisons) as the threshold
  - For 10 alignments, use α' = 0.05/10 = 0.005 (instead of 0.05) as the threshold
    → For $P(S \geq x) = 0.005$, $(1-0.005)^{10} \approx 0.95$ is the probability to have all 10 alignments with $S<x$ by chance
    → 1-0.95 = 0.05 is the probability to have at least one alignment with $S \geq x$ by chance
  - For 100 alignments, use α' = 0.05/100 = 0.0005 (instead of 0.05) as the threshold
    → For $P(S \geq x) = 0.0005$, $(1-0.0005)^{100} \approx 0.95$ is the probability to have all 100 alignments with $S<x$ by chance
    → 1-0.95 = 0.05 is the probability to have at least one alignment with $S \geq x$ by chance

9

Bonferroni correction in database searching

➢ Multiple comparison correction

- Threshold without correction: $P = \alpha$
- Threshold with correction: $P = \alpha' = \alpha/N$ (for N comparisons)
- $E = N \times P$
  → For E-value, using $E = \alpha$ as the threshold is equivalent to using the threshold corrected for multiple comparisons
  → For database searching:
  N = the database size = the number of entries = the number of alignments

10

blastp statistics: E-value

➢ Karlin-Altschul equation (Karlin & Altschul, 1990)

[For a pairwise alignment]

$P = Kmne^{-\lambda S}$ (Lec 11 slide 3)

m, n: lengths of the sequences compared

→ $m \times n$: search space

[For database similarity searching]

Used by FASTA

$E = Kmne^{-\lambda S}$ (used in BLAST instead of $E=NxP$)

→ the expected number of HSPs with scores $\geq S$

- m: length of the query
- n: length of the database (total number of residues)

Consider a database as a single very long sequence

$P = 1 - e^{-E}$ ($P \approx E$ if $E < 0.01$)

→ the probability of having at least one HSP with its score $\geq S$

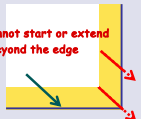
11

blastp statistics: E-value

➢ Karlin-Altschul equation (Karlin & Altschul, 1990) (See also Altschul & Gish, 1996)

$E = Km'n'e^{-\lambda S}$

- m': effective length of the query $m' = m - l$
- n': effective length of the database $n' = n - l$ x (number of sequences in the database)



- l: length adjustment → correction for edge effects
  → HSPs cannot occur too close to the search space edges.
  → Effective lengths of HSPs should be shorter than the actual lengths.

- For blastn and tblastx: $l = \ln(Kmn)/H$ is used
- For blastp, blastx, tblastn: adjusted m' and n' is calculated using the finite-size correction (FSC) (Park et al., 2012)

12

P-value, E-value, and database search

[FASTA]

- P-value for pairwise alignment: $1\text{-}\exp[-Kmne^{-\lambda S}] \approx Kmne^{-\lambda S}$
  - → Probability of getting the alignment score ≥ S from random pairwise comparison (m and n are the lengths of the two sequences compared)
- E-value = P x N, where N: database size (number of entries)

[BLAST]

- E-value = $Km'n'e^{-\lambda S}$
  - → Number of alignments with a score ≥ S expected by chance from a database search
  - m': effective length of the query
  - n': effective length of the database
- P-value for a database search: $P = 1 - e^{-E}$
  - → The probability of having at least one HSP with its score ≥ S

Effectively E=PxN if m' is the average length of queries and n' is the number of entries

Altschul *et al.* (1994)
See also BLAST Statistics @ NCBI BLAST website

13

blastp statistics: E-value

glutamate synthase [NADPH] large chain-like [Nerophis lumbriciformis]
Sequence ID: XP_061781277.1 Length: 1377 Number of Matches: 1

Range 1: 755 to 1081 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
219 bits(557)	5e-58	Compositional matrix adjust.	134/330(41%)	183/330(55%)	9/330(2%)

```
Query  199   MSYGALSLNAHLSFAKAVKECGTFMGTGEGGLPKALY---PYADH---IITQVASGRFGV  236
             MS GALS  AH +   A+   G    +GEGG    A +   P  D+     I QVASGRFGV
Sbjct  755   MSLGALSPEAHKTLNVAMNRIGAKSDSGEGGEDPAHFVPEPNGDNPSAKIKQVASGRFGV  814

Query  237   NEEYLMKGSAIEIKIGQGAKPGIGGHLPGEKVTAEISATRMIPEGSDAISPAPHHDIYSI  296
               EYL     +EIK+ QGAKPG GG LPG KVT  I+  R   +G   ISP PHHDIYSI
Sbjct  815   TAEYLNHCEELEIKVAQGAKPGEGGQLPGIKVTDLIARLRHSTKGVTLISPPPHHDIYSI  874
```

λ=0.267, K=0.041, S=557, S'_{bit}={0.267 x 557 - ln(0.041)} / ln2 = 219.2

Expect (E) = $Km'n'e^{-\lambda S}$ or $m'n'e^{-S'_{nat}}$ or $m'n'2^{-S'_{bit}}$

$E = 0.041 \times m' \times n' \times e^{-0.267 \times 557}$ [from the raw score]
$E = m' \times n' \times 2^{-219}$ [from the bit score]

$m' \times n'$: Effective search space

14

blastp search summary

Query: Q58746.1 (510 amino acids) ← m = 510 (length of query)

Search Parameters	
Program	blastp
Word size	5
Expect value	0.05
Hitlist size	1000
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	0
Composition-based stats	2

Database	
Posted date	Feb 13, 2026 2:53 PM
Number of letters	34,352,650,388 ← n: length of database
Number of sequences	82,371,451
Entrez query	Includes: Eukaryota (taxid:2759)

Karlin-Altschul statistics		
Lambda	0.31987	0.267
K	0.137272	0.041
H	0.416015	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

15

blastp statistics: E-value

glutamate synthase [NADPH] large chain-like [Nerophis lumbriciformis]
Sequence ID: XP_061781277.1 Length: 1377 Number of Matches: 1

Range 1: 755 to 1081 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
219 bits(557)	5e-58	Compositional matrix adjust.	134/330(41%)	183/330(55%)	9/330(2%)

```
Query  199   MSYGALSLNAHLSFAKAVKECGTFMGTGEGGLPKALY---PYADH---IITQVASGRFGV  236
             MS GALS  AH +   A+   G    +GEGG    A +   P  D+     I QVASGRFGV
Sbjct  755   MSLGALSPEAHKTLNVAMNRIGAKSDSGEGGEDPAHFVPEPNGDNPSAKIKQVASGRFGV  814

Query  237   NEEYLMKGSAIEIKIGQGAKPGIGGHLPGEKVTAEISATRMIPEGSDAISPAPHHDIYSI  296
               EYL     +EIK+ QGAKPG GG LPG KVT  I+  R   +G   ISP PHHDIYSI
Sbjct  815   TAEYLNHCEELEIKVAQGAKPGEGGQLPGIKVTDLIARLRHSTKGVTLISPPPHHDIYSI  874
```

λ=0.267, K=0.041, S=557, S'_{bit}={0.267 x 557 - ln(0.041)} / ln2 = 219.2

Expect (E) = $Km'n'e^{-\lambda S}$ or $m'n'e^{-S'_{nat}}$ or $m'n'2^{-S'_{bit}}$

$E = 0.041 \times m' \times n' \times e^{-0.267 \times 557}$ [from the raw score]
$E = m' \times n' \times 2^{-219}$ [from the bit score]

W/O length adjustment: m = 510, n = 34,352,650,388
E = 0.041 x 510 x 34,352,650,388 x $e^{-0.267 \times 557}$ = 1.86e-53 = 1.86×10^{-53}
E = 510 x 34,352,650,388 x 2^{-219} = 2.08e-53 = 2.08×10^{-53} (> 5e-58)

(Without length adjustment, E-values are overestimated)

$P = 1 - e^{-E}$
$= 1 - \exp(-1.86 \times 10^{-53})$
≈ 0 ($P \approx E$ if $E < 0.01$)

16

blastp search summary

Query: Q58746.1 (510 amino acids)

Search Parameters		
Program	blastp	
Word size	5	← Word size (W)
Expect value	0.05	← E-value threshold
Hitlist size	1000	← Max target sequences
Gapcosts	11,1	← Scoring matrix & gap penalties
Matrix	BLOSUM62	
Filter string	F	
Genetic Code	1	
Window Size	40	← Length separating two HSPs to trigger extension (A: two-hit methods)
Threshold	0	← Neighborhood threshold (T) (no longer provided)
Composition-based stats	2	

Database	
Posted date	Feb 13, 2026 2:53 PM
Number of letters	34,352,650,388
Number of sequences	82,371,451
Entrez query	Includes: Eukaryota (taxid:2759)

Karlin-Altschul statistics		
Lambda	0.31987	0.267
K	0.137272	0.041
H	0.416015	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

λ, K, and H are pre-estimated for a combination of the scoring matrix and gap penalties

for gapped alignment

17

blastn search summary & statistics

Search Parameters	
Program	blastn
Word size	11
Expect value	0.05
Hitlist size	100
Match/Mismatch scores	2,-3
Gapcosts	5,2
Low Complexity Filter	Yes
Filter string	L;m;
Genetic Code	1

Query: DQ018115.1 (1533 bp) m = 1533

Database	
Posted date	Feb 18, 2026 1:15 AM
Number of letters	1,056,255,000,623
Number of sequences	121,169,851
Entrez query	None

n: Length of database
N: Number of sequences

Karlin-Altschul statistics		
Lambda	0.633731	0.625
K	0.408146	0.41
H	0.912438	0.78

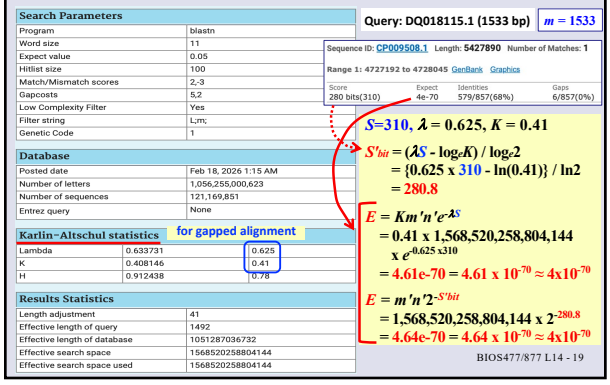
Results Statistics	
Length adjustment	41
Effective length of query	1492
Effective length of database	1051287036732
Effective search space	1568520258804144
Effective search space used	1568520258804144

For blastn, fixed values are used for effective lengths and shown in the search summary

l = 41
$m' = m - 41 = 1533 - 41 = 1492$
$n' = n - l \times N$
$= 1{,}056{,}255{,}000{,}623 - 41 \times 121{,}169{,}851$
$= 1{,}051{,}287{,}036{,}723$
$m' \times n' = 1492 \times 1{,}051{,}287{,}036{,}723$
$= 1{,}568{,}520{,}258{,}804{,}144$

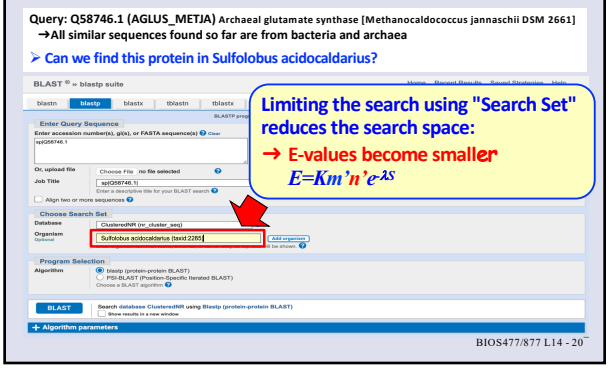
18

blastn search summary & statistics



19

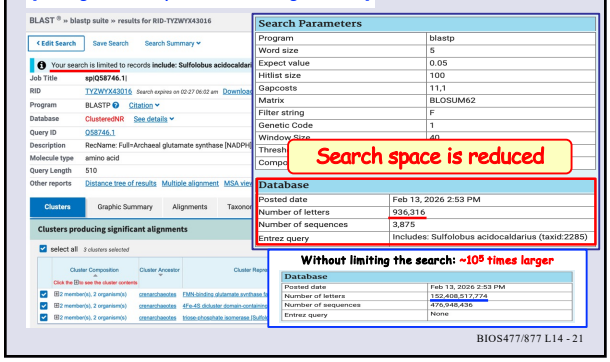
BLAST search set vs. format option



20

BLAST search set vs. format option

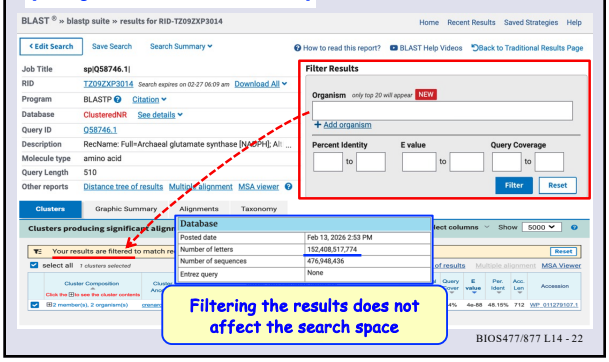
[Limiting the search space BEFORE starting the search]



21

BLAST search set vs. format option

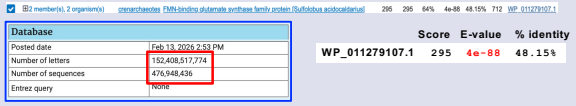
[Filtering the search result AFTER the search]



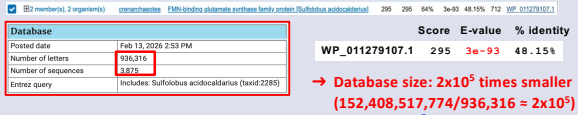
22

Database size and E-value

- Search is NOT limited; result is filtered



- Search is limited to Sulfolobus acidocaldarius



→ Database size: 2x10⁵ times smaller (152,408,517,774/936,316 ≈ 2x10⁵)
→ E-value: ~10⁵ times smaller (3e-93/4e-88 ≈ 10⁻⁵)

E=Km'n'e-λS
E-value is affected by the database size!

BIOS477/877 L14 - 23

23

FASTA Web servers

FASTA server @ UVA
(Pearson's original site; including also SSEARCH)

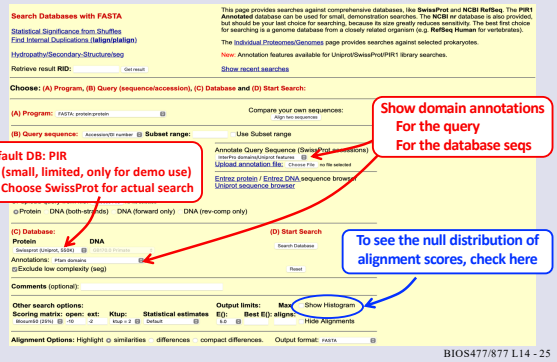
FASTA server @ EBI (including also SSEARCH)
- With graphic output
- Results can be obtained through email

FASTA server @ genome.jp (KEGG)
(search can be done against KEGG databases)

BIOS477/877 L14 - 24

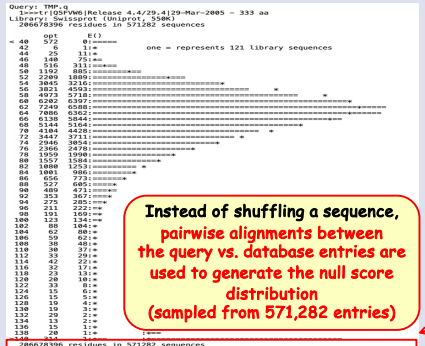
24

FASTA similarity search



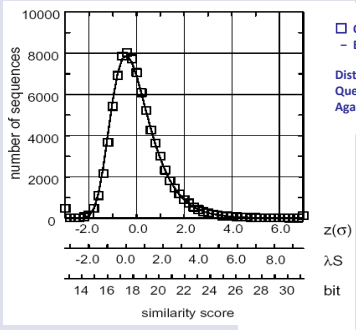
25

FASTA statistics



26

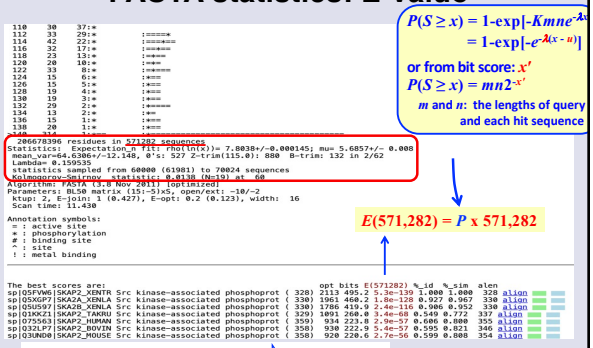
FASTA statistics



Pearson (1998, 2013)

27

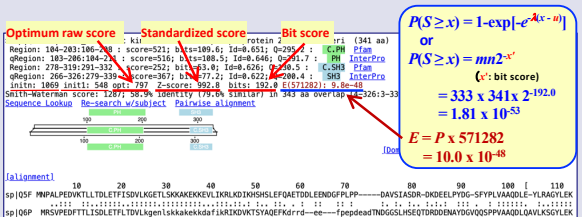
FASTA statistics: E-value



NOTE: FASTA calculates P-value for each pairwise alignment → $E = NxP$

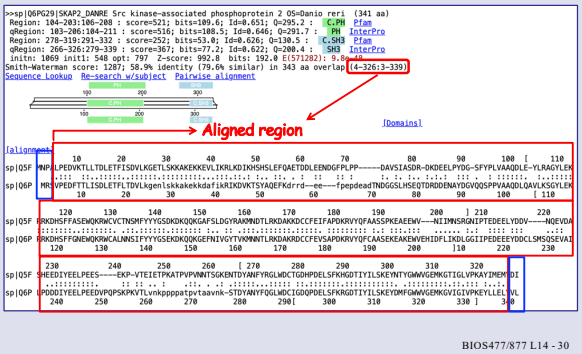
28

FASTA statistics: E-value



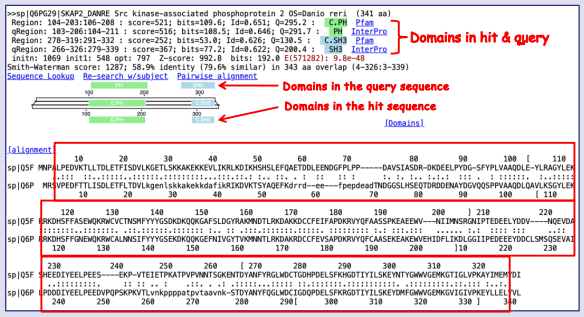
29

FASTA alignment



30

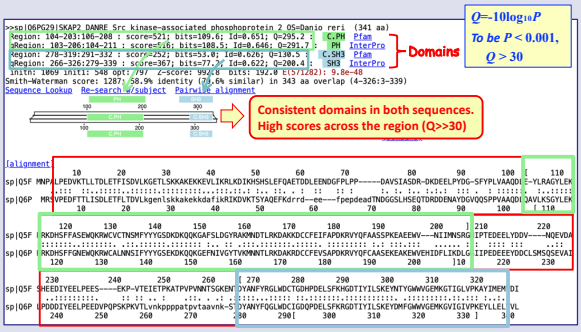
FASTA alignment



BIOS477/877 L14 - 31

31

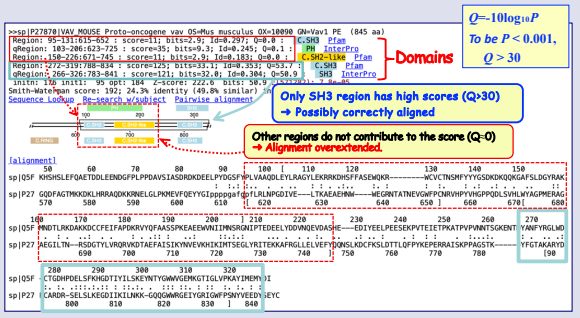
FASTA alingment



BIOS477/877 L14 - 32

32

FASTA alignment



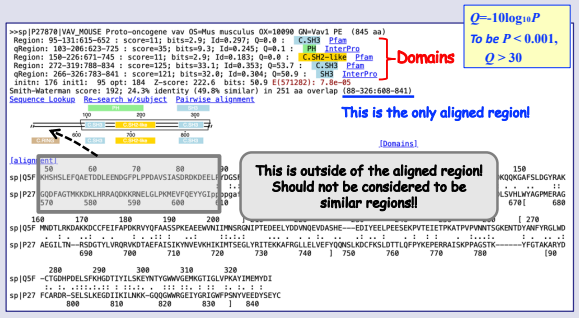
FASTA uses domain information to indicate possible incorrect alignment

See Mills and Pearson (2013)

BIOS477/877 L14 - 33

33

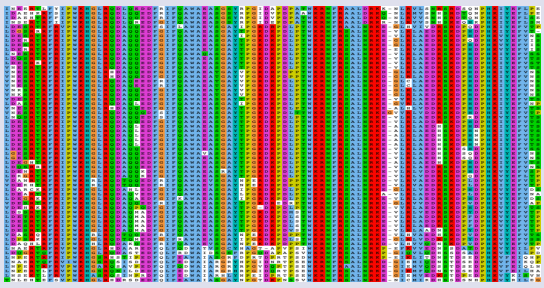
FASTA alignment



BIOS477/877 L14 - 34

34

Multiple sequence alignment (MSA)



BIOS477/877 L14 - 35

35

Why multiple alignment?

- To examine evolutionary relationships between sequences
  - → To reconstruct phylogenetic trees
- To predict protein functions
  (conserved regions, functional or structural domains)
- For homology modeling (structural prediction)
- To design PCR primers etc. etc. ...

BIOS477/877 L14 - 36

36

Multiple alignment as an extension of pairwise alignment

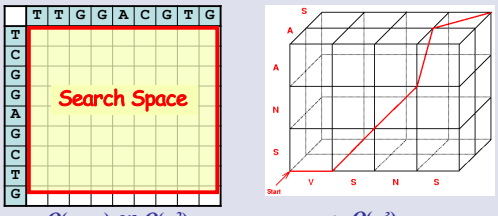
- Dynamic programming algorithm
  - → Guarantees to find the optimal alignment based on the scoring system
- Optimal alignments are searched based on alignment score
  - → Match/mismatch (S_{ij}) and gap penalties

37

Multiple alignment: complexity

- Dynamic programming algorithm

Pairwise alignment | Multiple alignment (3x)



$O(mxn)$ or $O(n^2)$ ← → $O(n^3)$

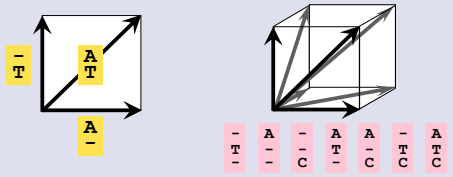
Complexity can be expressed with big-O notation

38

Multiple alignment: complexity

- Dynamic programming algorithm

Pairwise alignment (3 ways to align) | Multiple alignment for 3 (7 ways to align)



$O(mxn)$ or $O(n^2)$ ← → $O(n^3)$

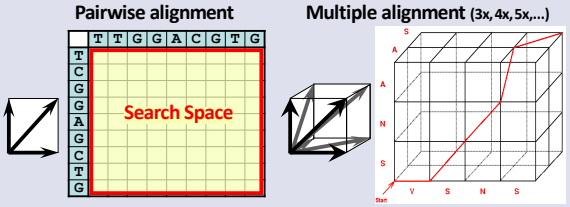
Complexity can be expressed with big-O notation

39

Multiple alignment: complexity

- Dynamic programming algorithm

Pairwise alignment | Multiple alignment (3x, 4x, 5x,...)



$O(n^2)$ $O(n^3)$, $O(n^4)$, $O(n^5)$ ... $O(n^x)$

Impossible for more than 5 - 6 sequences!

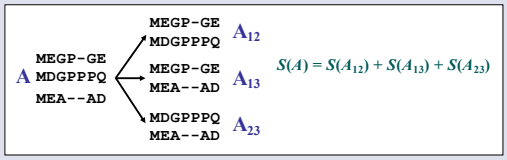
40

How to score multiple alignment

- Sum of pairs score

$S(A) = \Sigma_{i,j} S(A_{ij})$

A_{ij}: the score of the pairwise alignment between i and j



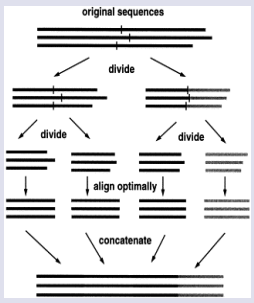
→ S(A) has no statistical justification

There is no single good method that can measure the overall quality of multiple alignments!

41

Multiple alignment: Divide and Conquer (DCA)

https://bibiserv.cebitec.uni-bielefeld.de/dca



Stoye (1998)

Divide and Conquer algorithm

i) Breaks down a problem into small sub-problems,

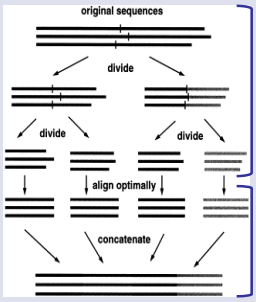
ii) Solves each sub-problem independently,

iii) Combines the solutions to the sub-problems to give a solution to the original large problem.

42

Multiple alignment: Divide and Conquer (DCA)

https://bibiserv.cebitec.uni-bielefeld.de/dca



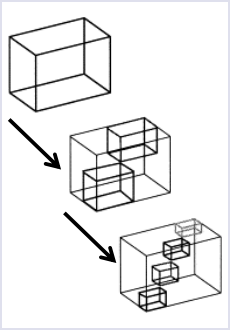
For multiple alignment:

1. Each sequence is cut into two.
2. Each group of subsequences is recursively cut into two until they are short enough.
3. Each group of short subsequence is aligned optimally (dynamic programming algorithm)
4. Short alignments are concatenated, yielding a solution of the original multiple alignment problem.

Stoye (1998)

43

Multiple alignment: Divide and Conquer (DCA)



DCA:

- Uses (almost) exact multiple alignment method on shorter fragments
- How can the suitable cutting positions be found?

DIALIGN (local fragment alignment) can be used to find cutting positions (based on highly conserved fragments)

Sammeth *et al.* (2003) *Bioinformatics* 19: ii189-ii195
[Not implemented.]

https://bibiserv.cebitec.uni-bielefeld.de/dialign/
https://dialign.gobics.de

44