

BIOS 477/877 Bioinformatics and Molecular Evolution

Instructor: Etsuko Moriyama
(School of Biological Sciences)

Spring 2026 Lecture 14

BIOS477/877 L14 - 1

1

Today's topics

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

BIOS477/877 L14 - 2

2

blast 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	1-33	MSYGALSLNAHLSFAKAVKEGTFMGTGEGGLPKALY---PYADH---IITQVASGRFGV	236
Sbjct	755	MS GALS AH + A+ G +GEGG A + P D+ I QVASGRFGV	814
Query	37	NEEYLMKGSATIEIKIGGAKPGIGGHLPEKVTAEISATRMIPEGSDAISAPPHDIYSI	296
Sbjct	815	EYL +EIK+ GGAKPG GG LPG KVT I+ R +G ISP PHHD IYSI	874

Raw Score (S): simply based on pairwise scores & gap penalties
Normalized Score or Bit Score (S'bit):
 $S'_{bit} = (\lambda S - \log_e K) / \log_e 2$, $[S'_{nat} = \lambda S - \log_e K]$

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 ?

BIOS477/877 L14 - 3

3

blast search summary

BLAST® - blastp suite - results for RID-WDDXJ91M016

Job Title: Q58746.Recham...-Archaean glutamate synthase...
 RID: 05/11/28 am

Click to see the blast search statistics

Sequences producing significant alignments

Description	Score	Score	Expect	Per	Acc
Glutamate synthase large subunit-like protein [Geobacillus baronii]	244	244	79%	5e-72	38.22%

(Top portion of any blast output)

BIOS477/877 L14 - 4

4

blast search summary

Search Parameters	
Program	blastp
Word size	5
Expect value	0.05
Hitlist size	1000
Gap costs	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
K	0.137272
H	0.416015
Alpha	0.7916
Alpha _v	4.96466
Sigma	43.6362

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

BIOS477/877 L14 - 5

5

blast 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	1-33	MSYGALSLNAHLSFAKAVKEGTFMGTGEGGLPKALY---PYADH---IITQVASGRFGV	236
Sbjct	755	MS GALS AH + A+ G +GEGG A + P D+ I QVASGRFGV	814
Query	37	NEEYLMKGSATIEIKIGGAKPGIGGHLPEKVTAEISATRMIPEGSDAISAPPHDIYSI	296
Sbjct	815	EYL +EIK+ GGAKPG GG LPG KVT I+ R +G ISP PHHD IYSI	874

Raw Score (S): simply based on pairwise scores & gap penalties
Normalized Score or Bit Score (S'bit):
 $S'_{bit} = (\lambda S - \log_e K) / \log_e 2$, $[S'_{nat} = \lambda S - \log_e K]$
 $\lambda = 0.267$, $K = 0.041$, $S'_{bit} = \{0.267 \times 557 - \log_e(0.041)\} / \log_e 2 = 219.2$

λ and K are scoring system specific

Karlin-Altschul statistics	
Lambda	0.31987
K	0.137272
H	0.416015
Alpha	0.7916
Alpha _v	4.96466
Sigma	43.6362

(for gapped alignments)

BIOS477/877 L14 - 6

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 [GenPart](#) [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 1:33 MSYGALSLNAHLSFAKAVKCEGTFMGTGEGGLPKALY---PYADH---IITQVASGRFGV 236
           MS  GALS  AH  +  A+  G  +GEGG  A  +  P  D+  I  DVASGRFGV
Sbjct 2:55 MSLGALSPFAHKTLLNVAMNRI GAKS DSGEGGDPAHFVPEPNGNPSAKIKQVASSGRFGV 814

Query 3:37 NEEYLMKGSATEIKIGGAKPGTGGHLGPEKVTAEISATRMIPGSDAISPAHPHDIVSI 296
           EYL  +EIK+ GGAKPG GG LPG KVT  I+  R  +G  ISP  PHHDIVSI
Sbjct 4:15 TAEYLNHCEELEIKVAQGAQKPGEGGOLPGIKVTDLIARLRHSTKGVTLISPPPHDIVSI 874
    
```

Raw Score (S): simply based on pairwise scores & gap penalties
Normalized Score or Bit Score (S'_{bit}):
 $S'_{bit} = (\lambda S - \log_2 K) / \log_2 e$, $[S'_{nat} = \lambda S - \log_2 K]$
 $\lambda = 0.267$, $K = 0.041$, $S'_{bit} = \{0.267 \times 557 - \log_2(0.041)\} / \log_2 e = 219.2$

- 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

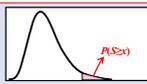
Pairwise alignment vs. database searching

[For a pairwise alignment]
 ➤ Probability of getting the alignment score $S \geq x$ by chance
 $P(S \geq x) = 1 - \exp[-Kmn e^{-\lambda x}] \approx Kmn e^{-\lambda x}$ ← Taylor series approximation → works if a is small ($\ll 1$)
 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, $0.95^{10} = 0.60$ is the prob. to have **all 10 alignments with $S < x$ by chance**
 → $1 - 0.60 = 0.40$ is the prob. to have **at least one alignment with $S \geq x$ by chance**
- For 100 alignments, $0.95^{100} = 0.006$ is the prob. to have **all 100-alignments with $S < x$ by chance**
 → $1 - 0.006 = 0.99$ 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 $\alpha' = 0.05/10 = 0.005$ (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 $\alpha' = 0.05/100 = 0.0005$ (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]
 $P = Kmn e^{-\lambda S}$ (Lec 11 slide 3)
 m, n : lengths of the sequences compared
 → $m \times n$: search space



[For database similarity searching] (Used by FASTA)
 $E = Kmn e^{-\lambda S}$ (used in BLAST instead of $E = N \times P$)
 → the expected number of HSPs with scores $\geq S$

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

$P = 1 - e^{-E}$ ($P \approx E$ 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
 $n' = n - l \times$ (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: $1 - \exp(-Kmn e^{-\lambda S}) \approx Kmn e^{-\lambda S}$
 - Probability of getting the alignment score $\geq S$ from random pairwise comparison (m and n are the lengths of the two sequences compared)
- > E-value = $P \times N$, where N: database size (number of entries)

[BLAST]

- > E-value = $Km'n'e^{-\lambda S}$
 - Number of alignments with a score $\geq S$ 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 $\geq S$

Altschul et al. (1994)
See also [BLAST Statistics @ NCBI BLAST website](#)

BIOS477/877 L14 - 13

13

blastp statistics: E-value

glutamate synthase [NADPH] large chain-like [Nerophis lumbliciformis]

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 155	MSYGALSLNAHLSFAKAVKCGTFMGTGEGGLPKALY---PYADH---IITQVASGRFGV	236			
Sbjct 755	MS GALS AH + A+ G +GEGG A + P D+ I QVASGRFGV	814			
Query 237	NEEYLMKGSATIEIKIGGAKPGIGGLHPEKVTAEISATRMIPGSDAISPAHPHDIIYSI	296			
Sbjct 815	EYL +EIK+ GGAKPG GG LPG KVT I+ R +G ISP PHHDIIYSI	874			

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

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

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

$$E = m' \times n' \times 2^{-219} \text{ [from the bit score]}$$

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

BIOS477/877 L14 - 14

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
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

BIOS477/877 L14 - 15

15

blastp statistics: E-value

glutamate synthase [NADPH] large chain-like [Nerophis lumbliciformis]

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 155	MSYGALSLNAHLSFAKAVKCGTFMGTGEGGLPKALY---PYADH---IITQVASGRFGV	236			
Sbjct 755	MS GALS AH + A+ G +GEGG A + P D+ I QVASGRFGV	814			
Query 237	NEEYLMKGSATIEIKIGGAKPGIGGLHPEKVTAEISATRMIPGSDAISPAHPHDIIYSI	296			
Sbjct 815	EYL +EIK+ GGAKPG GG LPG KVT I+ R +G ISP PHHDIIYSI	874			

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

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

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

$$E = m' \times n' \times 2^{-219} \text{ [from the bit score]}$$

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

(Without length adjustment, E-values are overestimated)

$$P = 1 - e^{-E}$$

$$= 1 - \exp(-1.86 \times 10^{-53})$$

$$\approx 0 (P \approx E \text{ if } E < 0.01)$$

BIOS477/877 L14 - 16

16

blastp search summary

Query: Q58746.1 (510 amino acids)

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

- Word size (W)
- E-value threshold
- Max target sequences
- Scoring matrix & gap penalties
- Length separating two HSPs to trigger extension (A: two-hit methods)
- Neighborhood threshold (T) (no longer provided)
- $\lambda, K,$ and H are pre-estimated for a combination of the scoring matrix and gap penalties
- for gapped alignment

BIOS477/877 L14 - 17

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	Lm;
Genetic Code	1

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

Karlin-Altschul statistics		
Lambda	0.633731	0.625
K	0.488146	0.45
H	0.912438	0.6

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

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

n : Length of database

N : Number of sequences

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

$$l = 41$$

$$m' = m - l = 1533 - 41 = 1492$$

$$n' = n - l \times N$$

$$= 1,056,255,000,623 - 41 \times 121,169,851$$

$$= 1,051,287,036,723$$

$$m' \times n' = 1492 \times 1,051,287,036,723 = 1,568,520,258,804,144$$

BIOS477/877 L14 - 18

18

blastn search summary & statistics

Search Parameters		Query: DQ018115.1 (1533 bp) $m = 1533$
Program	blastn	Sequence ID: CP009508.1 Length: 5427890 Number of Matches: 1
Word size	11	Range 1: 4727192 to 4728045 GenBank Graphics
Expect value	0.05	Score 280 bits(310) Expect 4e-70 Identities 579/857(68%) Gaps 6/857(0%)
Hittail size	100	
Match/Mismatch scores	2,-3	
Gap costs	5,2	
Low complexity filter	Yes	
Filter string	Ltrc	
Genetic Code	1	

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

Karlin-Altschul statistics for gapped alignment	
Lambda	0.633731
K	0.408146
H	0.912438

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

$S'_{bit} = (\lambda S - \log_e K) / \log_e 2$
 $= \{0.625 \times 310 - \ln(0.41)\} / \ln 2$
 $= 280.8$

$E = Km'n'e^{-\lambda S}$
 $= 0.41 \times 1,568,520,258,804,144 \times e^{-0.625 \times 310}$
 $= 4.61e-70 = 4.61 \times 10^{-70} \approx 4 \times 10^{-70}$

$E = m'n'2^{-S'_{bit}}$
 $= 1,568,520,258,804,144 \times 2^{-280.8}$
 $= 4.64e-70 = 4.64 \times 10^{-70} \approx 4 \times 10^{-70}$

BIOS477/877 L14 - 19

19

BLAST search set vs. format option

Query: Q58746.1 (AGLUS_METJA) Archaeal glutamate synthase [Methanocaldococcus jannaschii DSM 2661]
 → All similar sequences found so far are from bacteria and archaea

Can we find this protein in *Sulfolobus acidocaldarius*?

Limiting the search using "Search Set" reduces the search space:
 → E-values become smaller
 $E = Km'n'e^{-\lambda S}$

BIOS477/877 L14 - 20

20

BLAST search set vs. format option

[Limiting the search space BEFORE starting the search]

Search space is reduced

Without limiting the search: ~10⁵ times larger

BIOS477/877 L14 - 21

21

BLAST search set vs. format option

[Filtering the search result AFTER the search]

Filtering the results does not affect the search space

BIOS477/877 L14 - 22

22

Database size and E-value

Search is NOT limited; result is filtered

Database	Score	E-value	% identity
WP_011279107.1	295	4e-88	48.15%

Search is limited to *Sulfolobus acidocaldarius*

Database	Score	E-value	% identity
WP_011279107.1	295	3e-93	48.15%

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

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

Search Databases with FASTA

Statistical Significance from Shuffles

Hydrophobic/Secondary-Structurelog

Retrieve result RID:

Choose: (A) Program, (B) Query (sequence/accession), (C) Database and (D) Start Search:

(A) Program: FASTA program:

(B) Query sequence: Subst range: Use Subst range:

(C) Database: Protein, DNA

(D) Start Search:

FASTA statistics

Query: 28678399 residues in 571282 sequences

Statistics: Expectation: 7.8838e-08, mu: 5.657, sigma: 0.888

mean_var=4.6386e+12, lambda=0.132 in 2/62

Parameters: BL50 matrix (151-5)x15, open/exp: -18/-2

Annotations:

- Show domain annotations for the query
- Show domain annotations for the database seqs
- To see the null distribution of alignment scores, check here

BIOS477/877 L14 - 25

FASTA statistics

Query: 28678399 residues in 571282 sequences

Statistics: Expectation: 7.8838e-08, mu: 5.657, sigma: 0.888

mean_var=4.6386e+12, lambda=0.132 in 2/62

Parameters: BL50 matrix (151-5)x15, open/exp: -18/-2

Instead of shuffling a sequence, pairwise alignments between the query vs. database entries are used to generate the null score distribution (sampled from 571,282 entries)

BIOS477/877 L14 - 26

25

26

FASTA statistics

Observed distribution (black squares)

Expected distribution (blue line)

Distribution of Smith-Waterman score

Query: human glucose transporter

Against: SwissProt (~84,000 seq)

FASTA statistics: E-value

$P(S \geq x) = 1 - \exp[-Kmn^\lambda e^{-\beta x^\alpha}]$

$P(S \geq x) = mn2^{-x^\alpha}$

$E(571,282) = P \times 571,282$

BIOS477/877 L14 - 27

FASTA statistics: E-value

Query: 28678399 residues in 571282 sequences

Statistics: Expectation: 7.8838e-08, mu: 5.657, sigma: 0.888

mean_var=4.6386e+12, lambda=0.132 in 2/62

Parameters: BL50 matrix (151-5)x15, open/exp: -18/-2

FASTA statistics: E-value

$P(S \geq x) = 1 - \exp[-Kmn^\lambda e^{-\beta x^\alpha}]$

$P(S \geq x) = mn2^{-x^\alpha}$

$E(571,282) = P \times 571,282$

BIOS477/877 L14 - 28

27

28

FASTA statistics: E-value

Optimum raw score: 333

Standardized score: 1.81

Bit score: 10.0

Region: 184-203:186-208 | scores=521; b=1189.6; Id=8.451; Q=95.7; P=0.0002

Region: 278-318:291-331 | scores=516; b=1188.5; Id=8.646; Q=91.7; P=0.0002

Region: 266-326:279-339 | scores=387; b=77.2; Id=8.622; Q=288.4; P=0.0002

Smith-Waterman score: 1287; 58.9% identity (79.8% overlap) in 343 aa overlap

FASTA statistics: E-value

$P(S \geq x) = 1 - \exp[-Kmn^\lambda e^{-\beta x^\alpha}]$

$P(S \geq x) = mn2^{-x^\alpha}$

$E = P \times 571282 = 10.0 \times 10^{-48}$

BIOS477/877 L14 - 29

FASTA alignment

Optimum raw score: 333

Standardized score: 1.81

Bit score: 10.0

Region: 184-203:186-208 | scores=521; b=1189.6; Id=8.451; Q=95.7; P=0.0002

Region: 278-318:291-331 | scores=516; b=1188.5; Id=8.646; Q=91.7; P=0.0002

Region: 266-326:279-339 | scores=387; b=77.2; Id=8.622; Q=288.4; P=0.0002

Smith-Waterman score: 1287; 58.9% identity (79.8% overlap) in 343 aa overlap

FASTA alignment

BIOS477/877 L14 - 30

29

30

FASTA alignment

$Q = -10 \log_{10} P$
 To be $P < 0.001$,
 $Q > 30$

Domains in hit & query

Domains in the query sequence

Domains in the hit sequence

Domains

Alignment

```

>>>sp|06G29|SKAP2_DANRE Src kinase-associated phosphoprotein 2 OS=Danio rerio (341 aa)
Region: 184-283:184-283 : score=521; bits=189.6; Id=0.651; Q=295.2 ; C:SP1 Pfam
qRegion: 183-286:184-211 : score=516; bits=188.5; Id=0.646; Q=293.7 ; IM InterPro
Region: 278-319:291-332 : score=252; bits=83.8; Id=0.626; Q=138.5 ; C:SH3 Pfam
qRegion: 266-320:279-339 : score=367; bits=117.2; Id=0.822; Q=208.4 ; SH3 InterPro
init: 1869 init: 548 opt: 797 Z-score: 992.8 bits: 192.0 E(1571282): 9.8e-48
Smith-Waterman score: 1287.58.9% Identity (19.6% similar) in 343 aa overlap (4-326:3-339)
Sequence Lookup Re-search w/subject Pairwise Alignment
  
```

BIOS477/877 L14 - 31

31

FASTA alignment

$Q = -10 \log_{10} P$
 To be $P < 0.001$,
 $Q > 30$

Domains

Consistent domains in both sequences. High scores across the region ($Q > 30$)

Alignment

```

>>>sp|06G29|SKAP2_DANRE Src kinase-associated phosphoprotein 2 OS=Danio rerio (341 aa)
Region: 184-283:184-283 : score=521; bits=189.6; Id=0.651; Q=295.2 ; C:SP1 Pfam
qRegion: 183-286:184-211 : score=516; bits=188.5; Id=0.646; Q=293.7 ; IM InterPro
Region: 278-319:291-332 : score=252; bits=83.8; Id=0.626; Q=138.5 ; C:SH3 Pfam
qRegion: 266-320:279-339 : score=367; bits=117.2; Id=0.822; Q=208.4 ; SH3 InterPro
init: 1869 init: 548 opt: 797 Z-score: 992.8 bits: 192.0 E(1571282): 9.8e-48
Smith-Waterman score: 1287.58.9% Identity (19.6% similar) in 343 aa overlap (4-326:3-339)
Sequence Lookup Re-search w/subject Pairwise Alignment
  
```

BIOS477/877 L14 - 32

32

FASTA alignment

$Q = -10 \log_{10} P$
 To be $P < 0.001$,
 $Q > 30$

Domains

Only SH3 region has high scores ($Q > 30$)

→ Possibly correctly aligned

Other regions do not contribute to the score ($Q < 0$)

→ Alignment overextended.

Alignment

```

>>>sp|P27878|IVAV_MOUSE Proto-oncogene vav OS=Mus musculus 00-18998 QMvavi PE (845 aa)
Region: 95-311:95-321 : score=115; bits=37.1; Id=0.591; Q=68.0 ; C:SH3 Pfam
qRegion: 183-286:623-725 : score=35; bits=9.3; Id=0.245; Q=4.1 ; C:SH3 Pfam
Region: 158-226:671-745 : score=11; bits=3.9; Id=0.183; Q=0.8 ; C:SH3 Pfam
Region: 272-319:788-834 : score=125; bits=43.1; Id=0.333; Q=53.7 ; C:SH3 Pfam
qRegion: 266-320:783-841 : score=11; bits=3.9; Id=0.183; Q=0.8 ; C:SH3 Pfam
init: 176 init: 95 opt: 184 Z-score: 222.6 bits: 58.9 E(1571282): 7.8e-49
Smith-Waterman score: 332.24.3% Identity (40.6% similar) in 253 aa overlap (88-326:688-841)
Sequence Lookup Re-search w/subject Pairwise Alignment
  
```

FASTA uses domain information to indicate possible incorrect alignment

See Mills and Pearson (2013)

BIOS477/877 L14 - 33

33

FASTA alignment

$Q = -10 \log_{10} P$
 To be $P < 0.001$,
 $Q > 30$

Domains

This is the only aligned region

This is outside of the aligned region. Should not be considered to be similar regions!

Alignment

```

>>>sp|P27878|IVAV_MOUSE Proto-oncogene vav OS=Mus musculus 00-18998 QMvavi PE (845 aa)
Region: 95-311:95-321 : score=115; bits=37.1; Id=0.591; Q=68.0 ; C:SH3 Pfam
qRegion: 183-286:623-725 : score=35; bits=9.3; Id=0.245; Q=4.1 ; C:SH3 Pfam
Region: 158-226:671-745 : score=11; bits=3.9; Id=0.183; Q=0.8 ; C:SH3 Pfam
Region: 272-319:788-834 : score=125; bits=43.1; Id=0.333; Q=53.7 ; C:SH3 Pfam
qRegion: 266-320:783-841 : score=11; bits=3.9; Id=0.183; Q=0.8 ; C:SH3 Pfam
init: 176 init: 95 opt: 184 Z-score: 222.6 bits: 58.9 E(1571282): 7.8e-49
Smith-Waterman score: 332.24.3% Identity (40.6% similar) in 253 aa overlap (88-326:688-841)
Sequence Lookup Re-search w/subject Pairwise 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

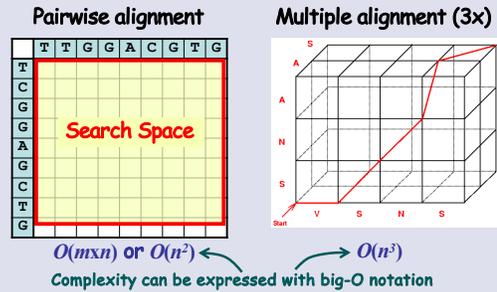
- **Dynamic programming algorithm**
 - Guarantees to find the optimal alignment based on the scoring system
- Optimal alignments are searched based on **alignment score**
 - Match/mismatch (S_{ij}) and gap penalties

BIOS477/877 L14 - 37

37

Multiple alignment: complexity

- **Dynamic programming algorithm**



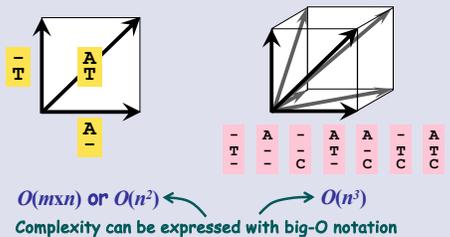
BIOS477/877 L14 - 38

38

Multiple alignment: complexity

- **Dynamic programming algorithm**

Pairwise alignment: 3 ways to align
Multiple alignment for 3: 7 ways to align

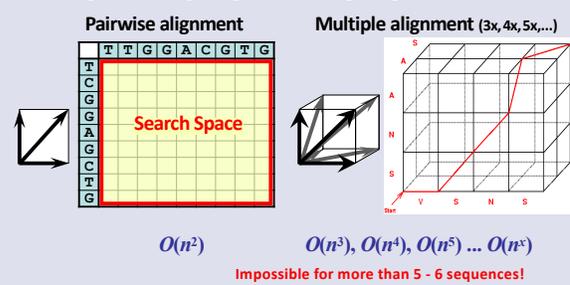


BIOS477/877 L14 - 39

39

Multiple alignment: complexity

- **Dynamic programming algorithm**



BIOS477/877 L14 - 40

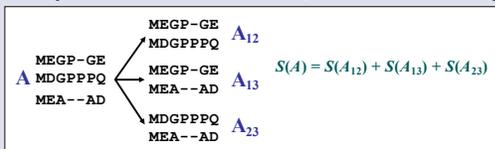
40

How to score multiple alignment

- **Sum of pairs score**

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

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



- $S(A)$ has no statistical justification

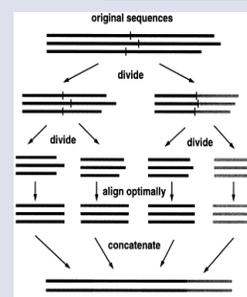
There is no single good method that can measure the overall quality of multiple alignments!

BIOS477/877 L14 - 41

41

Multiple alignment: Divide and Conquer (DCA)

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



- Divide and Conquer algorithm**
- Breaks down a problem into small sub-problems,
 - Solves each sub-problem independently,
 - Combines the solutions to the sub-problems to give a solution to the original large problem.

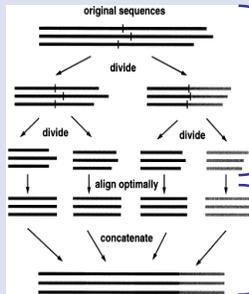
Stoye (1998)

BIOS477/877 L14 - 42

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.

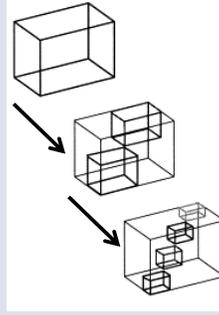
3. Each group of short subsequence is aligned optimally (dynamic programming algorithm)
4. Short alignments are concatenated, yielding a solution of the original multiple alignment problem.

Stoye (1998)

BIOS477/877 L14 - 43

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>

BIOS477/877 L14 - 44

44