

Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 14

BIOS477/877 L14 - 1

1

TODAY'S TOPICS

- **BLAST & FASTA statistics**
- **Multiple Alignment**
- **Progressive alignment (Clustal W)**
- **Assignment 7**

BIOS477/877 L14 - 2

2

BLAST Search Set vs. Format Option

Can we find similar proteins from other *Sulfolobus* species/genomes?

BIOS477/877 L14 - 3

3

BLAST Search Set vs. Format Option

[Limit the search space **BEFORE** starting the search]

BIOS477/877 L14 - 4

4

BLAST Search Set vs. Format Option

[Limit the search space **BEFORE** starting the search]

BIOS477/877 L14 - 5

5

BLAST Search Set vs. Format Option

[Limit the search result **AFTER** the search]

BIOS477/877 L14 - 6

6

BLAST Search Set vs. Format Option

After the search, results are filtered for "Sulfolobus" sequences

Search Parameters

Program: blastp
Word size: 5
Expect value: 0.05
Hitlist size: 5000
Gapcosts: 11,1
Matrix: BLOSUM62
Filter string: F
Genetic Code: 1
Window Size: 40
Threshold: 0
Composition-based stats: 2

Database

Posted date: Mar 3, 2024 2:30 AM
Number of letters: 278,407,168,794
Number of sequences: 722,992,963
Entrez query: None

Karlin-Altschul statistics

Lambda: 0.319428
K: 0.186251
H: 0.398828
Alpha: 0.7916
Alpha_v: 4.96466
Sigma: 43.6362

Search space is NOT affected

BIOS477/877 L14 - 7

7

BLAST Search Set vs. Format Option

Search is NOT limited;
results are filtered

Search is limited

Database

Posted date: Mar 3, 2024 2:30 AM
Number of letters: 278,407,168,794
Number of sequences: 722,992,963
Entrez query: None

Database

Posted date: Mar 3, 2024 2:30 AM
Number of letters: 14,101,522
Number of sequences: 151,369
Entrez query: Includes Sulfolobus (taxid:2284)

(Database size is ~20,000 times larger)

	Score	Query cov	E-value	% ident	>	Score	Query cov	E-value	% ident
WP_011278902.1	432	99%	2e-143	44.44%	>	432	99%	9e-148	44.44%
WP_198968373.1	60.1	26%	7e-06	28.29%	>	60.1	26%	4e-10	28.29%

(E-values are ~10⁴ times larger)

$E = Km'n'e^{-\lambda S}$
 E-value is affected by the database size!

BIOS477/877 L14 - 8

8

BLASTP Search Summary

Query: Q58746.1
Query length: 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: Mar 3, 2024 2:30 AM
Number of letters: 278,407,168,794
Number of sequences: 722,992,963
Entrez query: None

Karlin-Altschul statistics

Lambda: 0.319897
K: 0.137272
H: 0.416015
Alpha: 0.7916
Alpha_v: 4.96466
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)

$\lambda, K, \text{ and } H$ are pre-estimated for a combination of the scoring matrix and gap penalties

for gapped alignment

BIOS477/877 L14 - 9

9

FASTA

https://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml
(includes also SSEARCH)

<https://www.ebi.ac.uk/jdispatcher/sss/fasta>
(includes also SSEARCH)
With graphic output
Results can be obtained through email

<https://www.genome.jp/tools/fasta/>
(search can be done against KEGG databases)

BIOS477/877 L14 - 10

10

FASTA Similarity Search

Search Databases with FASTA

Search Proteomes/Genomes

Statistical Significance from Shuffles

Find Internal Duplications (align/plalign)

Hydropathy/Secondary-Structure

Retrieve result IDs:

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

(A) Program: FASTA_protein_protein

(B) Query: Q58746.1

(C) Database: DNAS (SwissProt)

(D) Start Search: Search Database

Default DB: PIR (small, limited, only for demo use)
Choose SwissProt for actual search

Show domain annotations for the query
For the database seqs

To see the null distribution of alignment scores, check here

https://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml

BIOS477/877 L14 - 11

11

FASTA Statistics

Query: TM0...
Library: SwissProt (Uniprot)
565254 residues in 565254 sequences

one = represents 128 library sequences

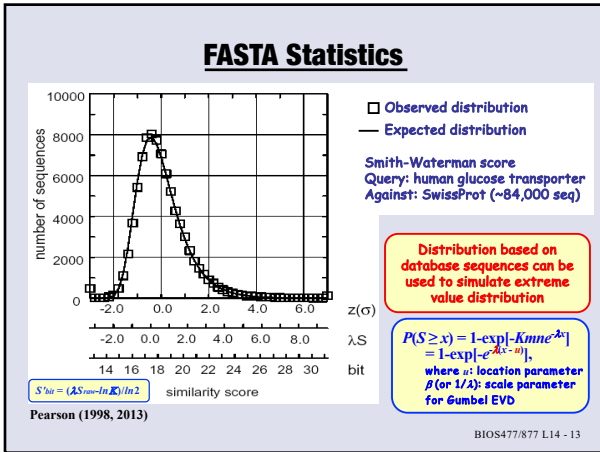
Extreme value distribution

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

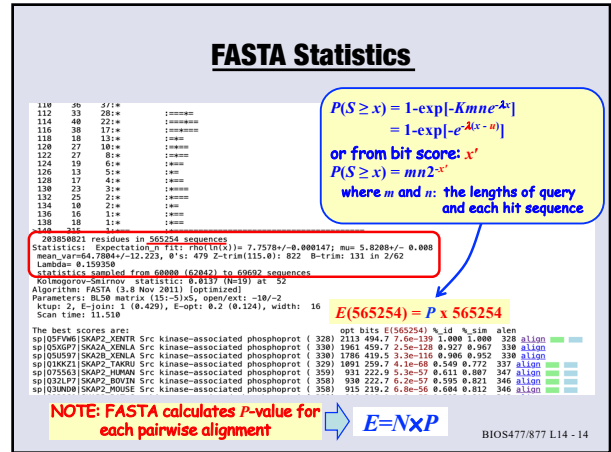
Instead of shuffling a sequence, pairwise alignments between the query vs. all database entries are used to generate the null score distribution (565,254 entries from SwissProt)

BIOS477/877 L14 - 12

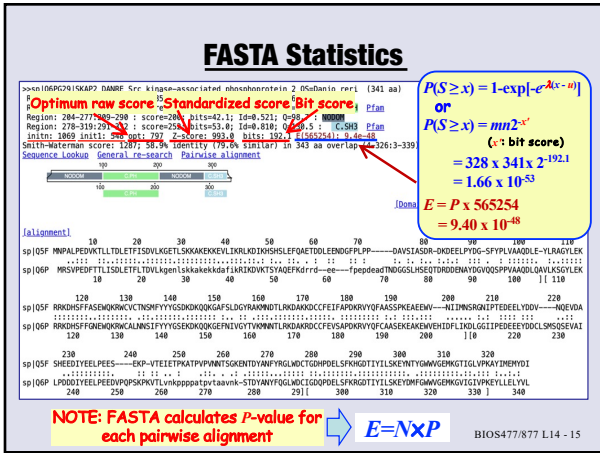
12



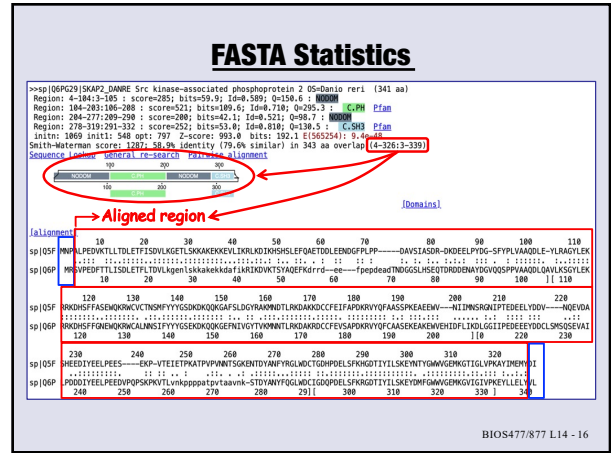
13



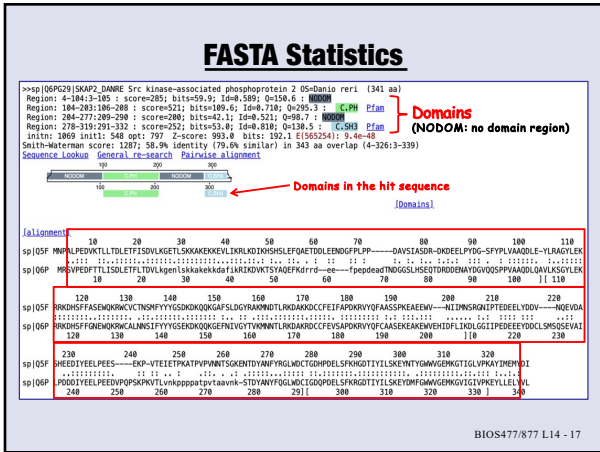
14



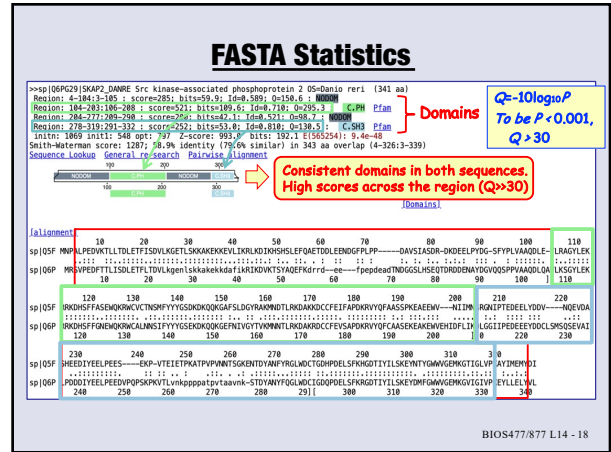
15



16



17



18

FASTA Statistics

```

>>sp|Q9Y2H5|PROM4_HUMAN Pleckstrin homology domain-containing family A member (1848 aa)
Region: 78-183:26-59 : score=21; bits=5.4; Id=0.118; Q=0.0 : NODOM
Region: 184-282:68-158 : score=36; bits=8.4; Id=0.276; Q=0.6 : C:PH Pfam Domains
Region: 286-262:159-218 : score=44; bits=11.2; Id=0.348; Q=0.9 : NODOM
Initn: 58 init1: 58 opt1: 186 Z-score: 223.1; bits: 51.9 E(65224): 7.3e-85
Smith-Waterman score: 281; 24.5% identity (55.1% similar) in 196 aa overlap (78-262:26-218)
Sequence Lookup General_re-search Pairwise alignment
  
```

Only SH3 domain region has high scores (Q>30); Possibly correctly aligned

Other regions do not contribute to the score (Q=0). Alignment overextended.

```

Alignment!
  30 40 50 60 70 80 90 100 110 120 130 140
sp|Q9Y2H5|PROM4_HUMAN SSKAKKEKVELIKLIKDKIRHSISLEFQMETDLEENGF...
sp|Q9Y2H5|PROM4_HUMAN MSNKTGGKRPATINSIDIPHMVSY...
  
```

FASTA uses domain information to indicate possible incorrect alignment
See Mills and Pearson (2013)

BIOS477/877 L14 - 19

19

FASTA Statistics

```

>>sp|Q9Y2H5|PROM4_HUMAN Pleckstrin homology domain-containing family A member (1848 aa)
Region: 78-183:26-59 : score=21; bits=5.4; Id=0.118; Q=0.0 : NODOM
Region: 184-282:68-158 : score=36; bits=8.4; Id=0.276; Q=0.6 : C:PH Pfam Domains
Region: 286-262:159-218 : score=44; bits=11.2; Id=0.348; Q=0.9 : NODOM
Initn: 58 init1: 58 opt1: 186 Z-score: 223.1; bits: 51.9 E(65224): 7.3e-85
Smith-Waterman score: 281; 24.5% identity (55.1% similar) in 196 aa overlap (78-262:26-218)
Sequence Lookup General_re-search Pairwise alignment
  
```

This is the only aligned region

These are outside of the aligned region. Should not be considered to be similar regions!

```

Alignment!
  30 40 50 60 70 80 90 100 110 120 130 140
sp|Q9Y2H5|PROM4_HUMAN SSKAKKEKVELIKLIKDKIRHSISLEFQMETDLEENGF...
sp|Q9Y2H5|PROM4_HUMAN MSNKTGGKRPATINSIDIPHMVSY...
  
```

BIOS477/877 L14 - 20

20

Multiple Sequence Alignment (MSA)

BIOS477/877 L14 - 21

21

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 - 22

22

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 - 23

23

Multiple alignment: complexity

- **Dynamic programming algorithm**

Pairwise alignment

	T	T	G	A	C	G	T	G
T								
C								
G								
A								
C								
T								
G								

Search Space

$O(mn)$ or $O(n^2)$

Multiple alignment (3x)

$O(n^3)$

Complexity can be expressed with big-O notation

BIOS477/877 L14 - 24

24

Multiple alignment: complexity

➤ **Dynamic programming algorithm**

Pairwise alignment
3 ways to align

Multiple alignment for 3
7 ways to align

$O(m \times n)$ or $O(n^2)$ ← $O(n^3)$

Complexity can be expressed with big-O notation

BIOS477/877 L14 - 25

25

Multiple alignment: complexity

➤ **Dynamic programming algorithm**

Pairwise alignment

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

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

Impossible for more than 5 - 6 sequences!

BIOS477/877 L14 - 26

26

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

A

MEGP-GE	MEGP-GE	A_{12}
MDGPPPQ	MDGPPPQ	
MEG-GE	MEGP-GE	A_{13}
MEA--AD	MEA--AD	
	MDGPPPQ	A_{23}
	MEA--AD	

$S(A) = S(A_{12}) + S(A_{13}) + S(A_{23})$

→ $S(A)$ has no statistical justification

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

BIOS477/877 L14 - 27

27

Multiple alignment: Divide and Conquer (DCA)

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

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.

Stoye (1998)

BIOS477/877 L14 - 28

28

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 - 29

29

Multiple alignment: Divide and Conquer (DCA)

DCA:

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

↓

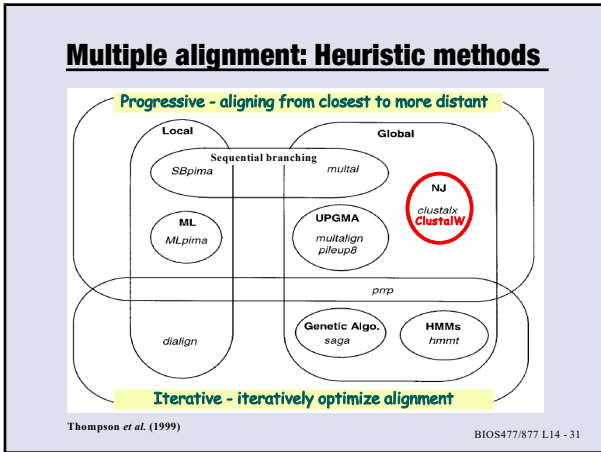
DIALIGN (local fragment alignment) can be used to find cutting positions (based on highly conserved fragments)

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

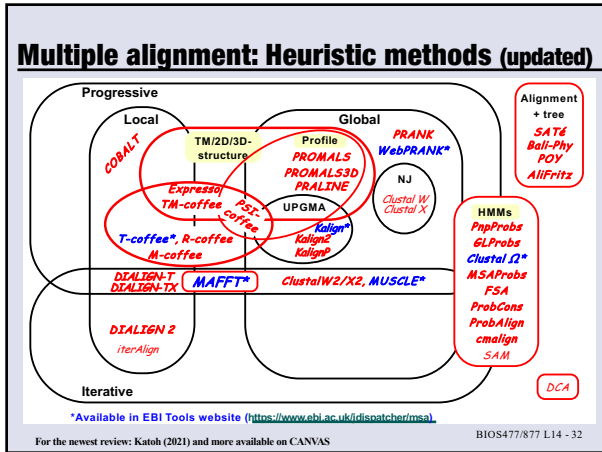
<https://bibiserv.cebitec.uni-bielefeld.de/dialign/>

BIOS477/877 L14 - 30

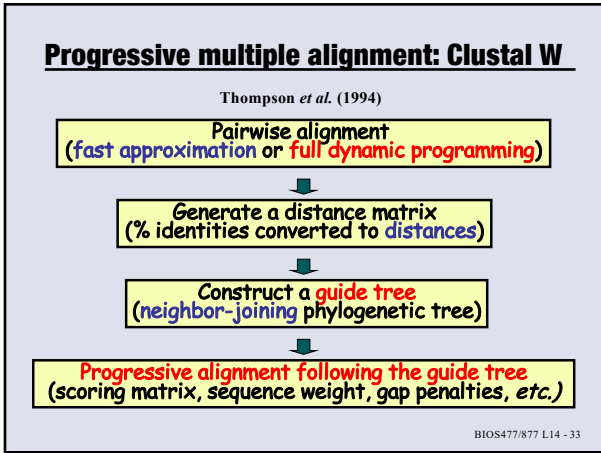
30



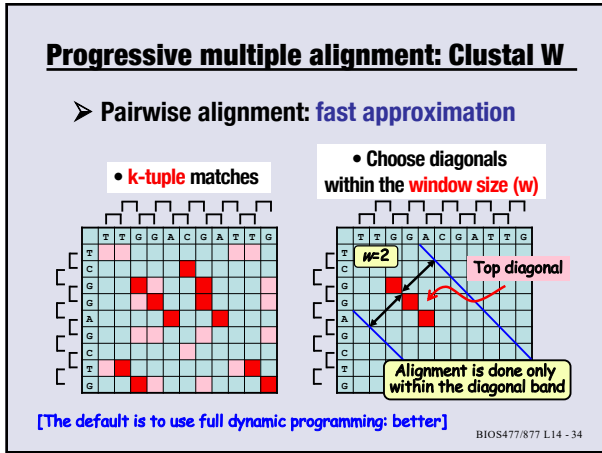
31



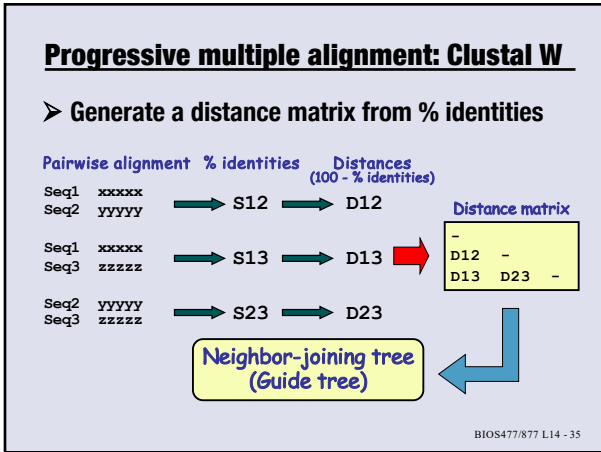
32



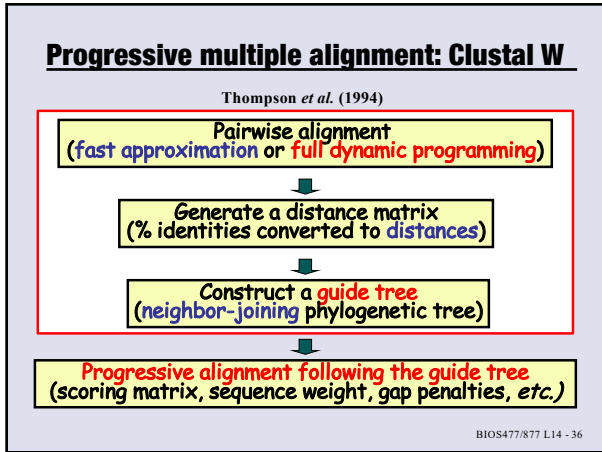
33



34



35



36

Progressive multiple alignment: Clustal W

➤ **Progressive alignment** following the guide tree

Closest sequences are aligned first

Pairwise alignment can be done by using simply the dynamic programming algorithm

Next closest sequence is aligned against the existing alignment

→ Profile alignment (can be an alignment between alignments)

BIOS477/877 L14 - 37

37

Progressive multiple alignment: Clustal W

➤ **Progressive alignment** following the guide tree

A sequence (S3) is aligned against the alignment (S1, S2)

Profile alignment: alignment vs. sequence, alignment vs. alignment

BIOS477/877 L14 - 38

38

Progressive multiple alignment: Clustal W

➤ **Progressive alignment** following the guide tree

From a scoring matrix

S1 vs. S3 = S(P,E)
S2 vs. S3 = S(G,E)

S1/S2 vs. S3 = (S(P,E)+S(G,E))/2

Simple average from all pairwise scores

Profile alignment: alignment vs. sequence, alignment vs. alignment

BIOS477/877 L14 - 39

39

Progressive multiple alignment: Clustal W

➤ **Progressive alignment** following the guide tree

Sequence weights

[Simple average (without weighting)]
Score(P/G,E) = (S(P,E) + S(G,E)) / 2

Or

[Weighted average]
Score(P/G,E) = (W1×W3×S(P,E) + W2×W3×S(G,E)) / 2

If W1=W2=W3=1, the same as simple average

Thompson et al. (1994)

Closely related sequences share information
→ The sequences with duplicated information should get smaller weights

BIOS477/877 L14 - 40

40

Progressive multiple alignment: Clustal W

➤ **Sequence weights**: based on branch lengths

Shorter branch lengths → Fewer changes

Shorter branch lengths → Fewer changes (e.g., S1 and S2 are more similar than S1 and S3)

Distance from the root

Rooted at the midpoint

BIOS477/877 L14 - 41

41

Progressive multiple alignment: Clustal W

➤ **Sequence weights**: based on branch lengths

Shorter branch lengths → Fewer changes

$W1 = \frac{0.02 + 0.15/2 + 0.09/3}{2} = 0.125$

Divided by the number of sequences sharing the branch

Closely related sequences share information
→ The sequences with duplicated information should get smaller weights

BIOS477/877 L14 - 42

42

Progressive multiple alignment: Clustal W

➤ **Sequence weights:** based on branch lengths

Shorter branch lengths → Fewer changes

$$W1 = 0.02 + 0.15/2 + 0.09/3 = 0.125$$

$$W2 = 0.08 + 0.15/2 + 0.09/3 = 0.185$$

Divided by the number of sequences sharing the branch

Closely related sequences share information
→ The sequences with duplicated information should get smaller weights

BIOS477/877 L14 - 43

43

Progressive multiple alignment: Clustal W

➤ **Sequence weights:** based on branch lengths

Shorter branch lengths → Fewer changes

$$W1 = 0.02 + 0.15/2 + 0.09/3 = 0.125$$

$$W2 = 0.08 + 0.15/2 + 0.09/3 = 0.185$$

$$W3 = 0.38 + 0.09/3 = 0.410$$

$$W4 = 0.46$$

Divided by the number of sequences sharing the branch

*Weights are normalized so that the max weight (0.46) becomes 1.0

Closely related sequences share information
→ The sequences with duplicated information should get smaller weights

BIOS477/877 L14 - 44

44

Progressive multiple alignment: Clustal W

➤ How sequence weighting works: Example 1

[Simple average] $Score(L/V,P) = \{S(L,P) + S(V,P)\} / 2 = 0.5 S(L,P) + 0.5 S(V,P)$

[Weighted average] $Score(L/V,P) = \{11 \times 12 \times S(L,P) + 5 \times 12 \times S(V,P)\} / 2 = 66 \times S(L,P) + 30 \times S(V,P)$

$S(L,P) \gg S(V,P)$

[Aligning the 3rd sequence (P) to the first 2 sequences (L, V) previously aligned]

Closely related sequences share information
→ The sequences with duplicated information should get smaller weights

BIOS477/877 L14 - 45

45

Progressive multiple alignment: Clustal W

➤ How sequence weighting works: Example 2

How should 'VG' in S4 be aligned against the S1/S2/S3 alignment?

Which is the best alignment, (a) or (b) or (c)?

(a) 123
S1 VAG
S2 VGA
S3 IVG
S4 VG-

(b) 123
S1 VAG
S2 VGA
S3 IVG
S4 V-G

(c) 123
S1 VAG
S2 VGA
S3 IVG
S4 -VG

[Aligning S4 to the first 3 sequences (S1, S2, and S3) previously aligned]

BIOS477/877 L14 - 46

46

Progressive multiple alignment: Clustal W

➤ How sequence weighting works: Example 2

[Simple average] (match=2, mismatch=-1)

Alignment (a) [Score = 1 + gap penalty]
1: Score = $\{S(V,V) + S(V,V) + S(L,V)\} / 3 = (2 + 2 - 1) / 3 = 1$
2: Score = $\{S(A,G) + S(G,G) + S(V,G)\} / 3 = (-1 + 2 - 1) / 3 = 0$
3: Score = (gap penalty $\times 3$) / 3

Alignment (b) [Score = 2 + gap penalty]
1: Score = $\{S(V,V) + S(V,V) + S(L,V)\} / 3 = (2 + 2 - 1) / 3 = 1$
2: Score = (gap penalty $\times 3$) / 3
3: Score = $\{S(G,G) + S(A,G) + S(V,G)\} / 3 = (2 - 1 + 2) / 3 = 1$

Alignment (c) [Score = 1 + gap penalty]
1: Score = (gap penalty $\times 3$) / 3
2: Score = $\{S(A,V) + S(G,V) + S(V,V)\} / 3 = (-1 - 1 + 2) / 3 = 0$
3: Score = $\{S(G,G) + S(A,G) + S(V,G)\} / 3 = (2 - 1 + 2) / 3 = 1$

[Aligning S4 to the first 3 sequences (S1, S2, and S3) previously aligned]

BIOS477/877 L14 - 47

47

Progressive multiple alignment: Clustal W

➤ How sequence weighting works: Example 2

[Weighted average] (match=2, mismatch=-1)

Alignment (a)
1: Score = $\{1.1 \times 7 \times S(V,V) + 1.1 \times 7 \times S(V,V) + 6.1 \times 7 \times S(L,V)\} / 3 = \{7.7 \times 2 + 7.7 \times 2 + 42.7 \times (-1)\} / 3 = -3.97$
2: Score = $\{1.1 \times 7 \times S(A,G) + 1.1 \times 7 \times S(G,G) + 6.1 \times 7 \times S(V,G)\} / 3 = \{7.7 \times (-1) + 7.7 \times 2 + 42.7 \times (-1)\} / 3 = -11.67$
3: Score = (gap penalty $\times 3$) / 3

Alignment Score = -15.64 + gap penalty

[Aligning S4 to the first 3 sequences (S1, S2, and S3) previously aligned]

BIOS477/877 L14 - 48

48

Progressive multiple alignment: Clustal W

➤ How sequence weighting works: Example 2

[Weighted average] (match=2, mismatch=-1)

Alignment (b)

1: Score = $\{1.1 \times 7 \times 5(V,V) + 1.1 \times 7 \times 5(V,V) + 6.1 \times 7 \times 5(I,V)\} / 3$
 $= \{7.7 \times 2 + 7.7 \times 2 + 42.7 \times (-1)\} / 3 = -3.97$

2: Score = (gap penalty $\times 3$) / 3

3: Score = $\{1.1 \times 7 \times 5(G,G) + 1.1 \times 7 \times 5(A,G) + 6.1 \times 7 \times 5(G,G)\} / 3$
 $= \{7.7 \times 2 + 7.7 \times (-1) + 42.7 \times (2)\} / 3 = 31.03$

Alignment Score = 27.06 + gap penalty

[Aligning S4 to the first 3 sequences (S1, S2, and S3) previously aligned]

BIOS477/877 L14 - 49

49

Progressive multiple alignment: Clustal W

➤ How sequence weighting works: Example 2

[Weighted average] (match=2, mismatch=-1)

Alignment (c)

1: Score = (gap penalty $\times 3$) / 3

2: Score = $\{1.1 \times 7 \times 5(A,V) + 1.1 \times 7 \times 5(G,V) + 6.1 \times 7 \times 5(V,V)\} / 3$
 $= \{7.7 \times (-1) + 7.7 \times (-1) + 42.7 \times (2)\} / 3 = 23.33$

3: Score = $\{1.1 \times 7 \times 5(G,G) + 1.1 \times 7 \times 5(A,G) + 6.1 \times 7 \times 5(G,G)\} / 3$
 $= \{7.7 \times 2 + 7.7 \times (-1) + 42.7 \times (2)\} / 3 = 31.03$

Alignment Score = 54.36 + gap penalty

[Aligning S4 to the first 3 sequences (S1, S2, and S3) previously aligned]

BIOS477/877 L14 - 50

50

Progressive multiple alignment: Clustal W

➤ How sequence weighting works: Example 2

[Simple average]

(a) Alignment score = 1 + gap penalty
 (b) Alignment score = 2 + gap penalty
 (c) Alignment score = 1 + gap penalty

[Weighted average]

(a) Alignment score = -15.64 + gap penalty
 (b) Alignment score = 27.06 + gap penalty
 (c) Alignment score = 54.36 + gap penalty

[Aligning S4 to the first 3 sequences (S1, S2, and S3) previously aligned]

BIOS477/877 L14 - 51

51

Progressive multiple alignment: Clustal W

➤ How to choose **scoring matrix**:

- Choose only the scoring matrix series (BLOSUM, PAM, etc.)
- Specific matrix is determined based on distance between sequences

80 - 100% identity → Blosum80
 60 - 80% identity → Blosum62
 30 - 60% identity → Blosum45
 0 - 30% identity → Blosum30

Thompson *et al.* (1994)

BIOS477/877 L14 - 52

52

Progressive multiple alignment: Clustal W

➤ How **gap penalties** are determined:

- Initial gap penalties: **GOP** (gap opening) and **GEP** (gap extension) → set by the user
- Weight (scoring) matrix dependent gap penalties
- Similarity level dependent gap penalties
- Sequence length dependent gap penalties
- Position specific gap penalties
 - if gaps already exist
 - residue specific (e.g., hydrophilic stretches)

Thompson *et al.* (1994)

BIOS477/877 L14 - 53

53

Progressive multiple alignment: Clustal W

Pairwise alignment

S1 → D12
 S2 → D13
 S1 → D14
 S4 → D14

Guide tree

Profile alignment

- Progressive alignment
 - Greedy (finds local optima, but no guarantee for global optima)
 - Errors (incorrect gap positions) in the early alignments cannot be rectified later
- Global alignment only (local similarity may be missed)



BIOS477/877 L14 - 54



54

Clustal Web servers

<http://www.clustal.org/> (Clustal original website)

Clustal: Multiple Sequence Alignment
Multiple alignment of nucleic acid and protein sequences





Clustal Omega

- Latest version of Clustal - fast and scalable (can align hundreds of thousands of sequences in hours), greater accuracy due to new HMM alignment engine
- Command line/ web server only (GUI public beta available soon)

ClustalW/ClustalX

- "Classic Clustal"
- GUI (ClustalX), command line (ClustalW), web server versions available

Use Clustal Ω instead of Clustal W

[Legacy version]
<https://galaxy.pasteur.fr/> (part of Galaxy@Pasteur)
<https://www.ezname.in/tools-bin/clustalw>

BIOS477/877 L14 - 55