BIOS 477/877
Bioinformatics and Molecular Evolution

Instructor: Etsuko Moriyama
(School of Biological Sciences)

Spring 2026 | Lecture 11

Today's topics

- Statistical Significance of Alignment Scores
- Similarity Search
  - FASTA and BLAST

Significance of alignment scores

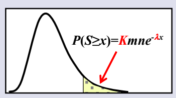
- $P(S \geq x|H_0)$: probability of getting the alignment score $S \geq x$ by chance

  Karlin-Altschul equation (Karlin and Altschul 1990)

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

  - K and λ: parameters based on a given scoring matrix and the amino acid composition of the sequences
  - m and n: lengths of sequences aligned

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



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

- E-value = $P(S \geq x|H_0) \times N$

  N: the number of sequences in the dataset

  Expected number of randomly selected sequences in the dataset that have alignment scores $\geq x$

E-value ≠ P-value

Estimation of K, λ, and P-value

- Estimation of K and λ from empirical distribution of random alignment scores (used in LALIGN and PRSS)
  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: $P(S \geq x) \approx Kmne^{-\lambda x}$

Simulation of random score distribution
(example)

[Input sequences]
- 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-QKALAAALGQIEKQFGKGSIMRLGEDRSMDVETISTGSLSLDIA     51
                    |.|.| .|.|..|...:...|...:...:   |..::..::|||.:||..
RAD51_YEAST     124 ISEAKADKLLNEAARLVPMGFVTAADFHM---RRSELICLTTGSKNLDTL    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 MSESRFSLIVVDSVMALY-RTDFSGRGELSARQMHLAKFM--RALQRLA-    314

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

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

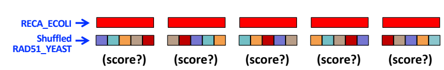
Using EMBOSS WATER @ EBI

Simulation of random score distribution
(example)

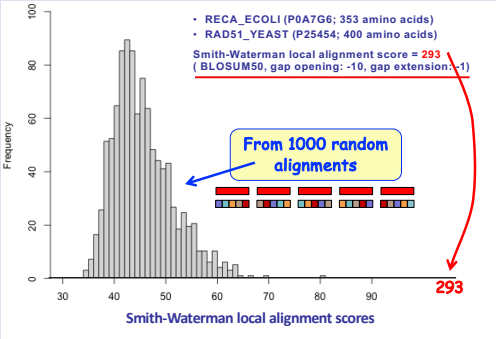
[Input sequences]
- RECA_ECOLI (P0A7G6; 353 amino acids)
- RAD51_YEAST (P25454; 400 amino acids)

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

1. Shuffle RAD51_YEAST 1000 times
   (generate 1000 random sequences)
2. Align with RECA_ECOLI
   (generate 1000 random local alignments)

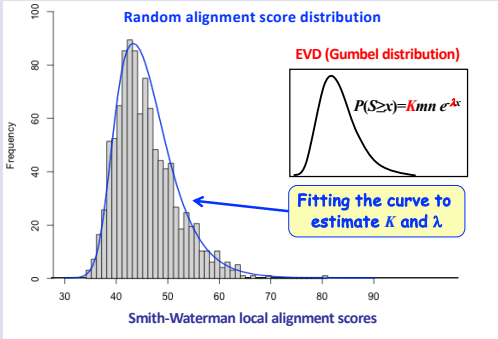


Simulation of random score distribution
(example)



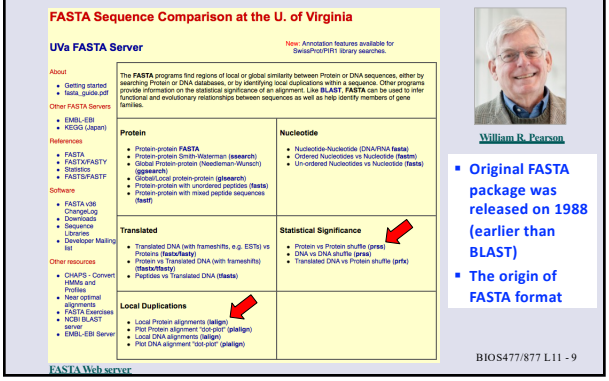
7

Simulation of random score distribution
(example)



8

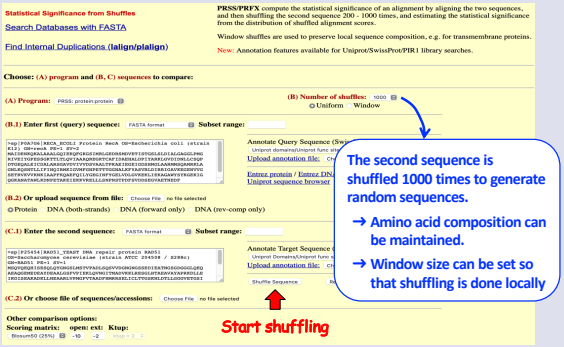
FASTA Web server by William Pearson



- Original FASTA package was released on 1988 (earlier than BLAST)
- The origin of FASTA format

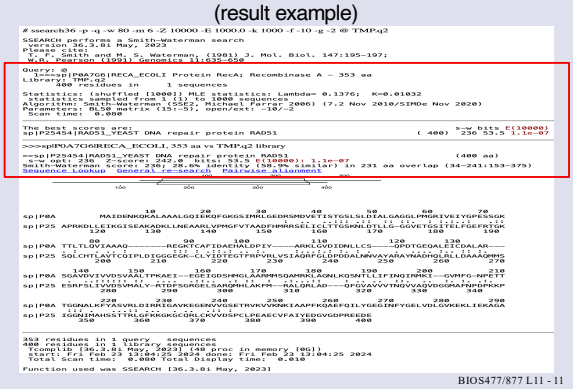
9

Statistical analysis of alignments by PRSS



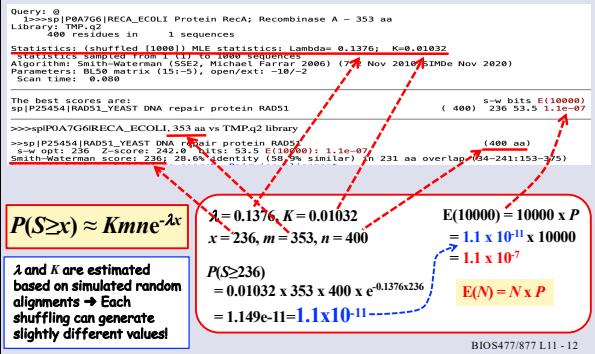
10

Statistical analysis of alignments by PRSS
(result example)



11

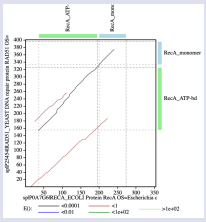
Statistical analysis of alignments by PRSS
(result example)



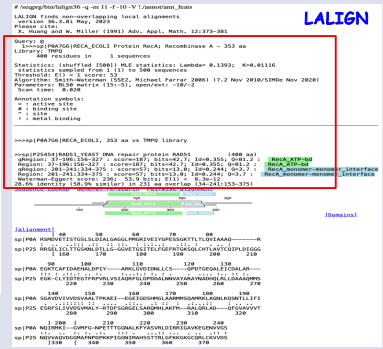
12

Pairwise alignment by LALIGN

PLALIGN (dot plot)



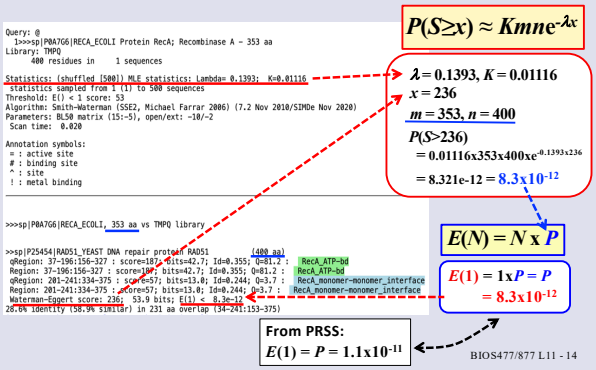
LALIGN



13

Pairwise alignment by LALIGN

$P(S \geq x) \approx Kmne^{-\lambda x}$



14

Estimation of K and λ in BLAST

➢ K and λ ➔ from Altschul & Gish (1996) table.

a	b	λ	K	H (nats)	a	b	λ	K	H (nats)
∞	0–∞	0.232	0.11	0.34	11	8–11	0.197	0.05	0.21
					11	6, 7	0.190	0.04	0.19
16	4–16	0.222	0.08	0.31	11	5	0.184	0.04	0.17
16	3	0.213	0.06	0.27	11	4	0.177	0.031	0.15
16	2	0.207	0.05	0.24	11	3	0.167	0.028	0.11
16	1	0.180	0.024	0.15	11	2	0.130	0.009	0.06
					11	1		Linear	
15	8–15	0.222	0.09	0.31					
15	6, 7	0.219	0.08	0.29	10	8–10	0.183	0.04	0.17
15	4, 5	0.216	0.07	0.28	10	6, 7	0.178	0.035	0.16
15	3	0.210	0.06	0.25	10	5	0.168	0.026	0.13
15	2	0.202	0.05	0.22	10	4	0.156	0.020	0.10
15	1	0.166	0.018	0.11	10	3	0.139	0.013	0.07
					10	2	0.099	0.007	0.03
14	8–14	0.218	0.08	0.29	10	1		Linear	
14	5–7	0.214	0.07	0.27					
14	4	0.205	0.05	0.24	9	7–9	0.164	0.029	0.13
14	3	0.201	0.05	0.22	9	5, 6	0.152	0.021	0.10
14	2	0.188	0.034	0.17	9	4	0.134	0.014	0.07
14	1	0.140	0.009	0.07	9	3	0.107	0.008	0.04
					9	1, 2		Linear	
13	8–13	0.211	0.06	0.27					
13	5–7	0.205	0.05	0.24	8	8	0.139	0.017	0.08
13	4	0.202	0.05	0.22	8	7	0.134	0.015	0.07
13	3	0.188	0.034	0.18	8	6	0.127	0.013	0.06
13	2	0.174	0.025	0.13	8	5	0.117	0.011	0.05
13	1	0.114	0.006	0.04	8	4	0.101	0.009	0.03
					8	1–3		Borderline or linear	
12	7–12	0.205	0.06	0.24					
12	5, 6	0.197	0.05	0.21	7	7	0.100	0.010	0.04
12	4	0.192	0.04	0.18	7	6	0.094	0.010	0.03
12	3	0.178	0.028	0.15	7	1–5		Borderline or linear	
12	2	0.158	0.019	0.10					
12	1		Borderline		1–6	1–6		Linear	

← For BLOSUM50;
a: gap opening penalty
b: gap extension penalty
(affine gap penalty)

Based on 10,000 random sequence pairs,
➔ calculate optimal local alignment scores, S.
➔ estimate K and λ by fitting the distribution of S with $P(S \geq x) \approx Kmne^{-\lambda x}$

This is what BLAST does

15

Significance of alignment scores

➢ Alignment scores cannot be compared directly because they depend on: scoring matrix, gap penalty, algorithms used

➢ Statistical significance of alignments can be tested
- Null hypothesis (H_0): the given alignment score is possible by aligning two randomly selected (unrelated) sequences
- If H_0 is rejected ($P <$ threshold),
  ➔ The given score is not possible simply by chance
  ➔ The alignment is meaningful (the two sequences must be related).

➢ P-value or E-value (E=NP) can be compared regardless of the scoring systems used for alignments.
➔ $0 <$ P-value < 1 vs. $0 <$ E-value $< N$

NOTE: E-values change depending on the number of data used (if N is small, E-value becomes small)

16

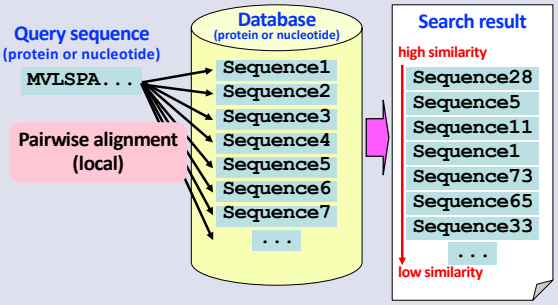
Similarity search

➢ What can you learn from similarity search?
- Related genes in another organisms
  ➔ Homologue candidates
- Possible functions of genes/protein

[From genomic sequence]
- Gene locations
- Against cDNA sequences
  ➔ exon-intron structure

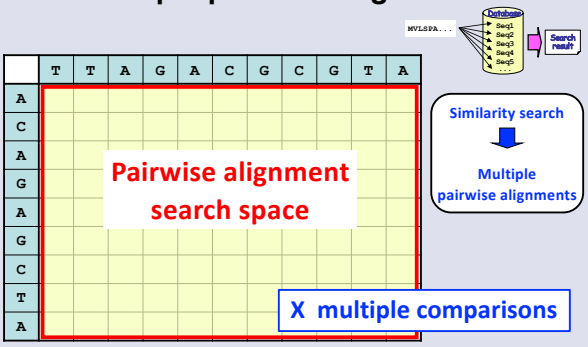
17

Similarity search



18

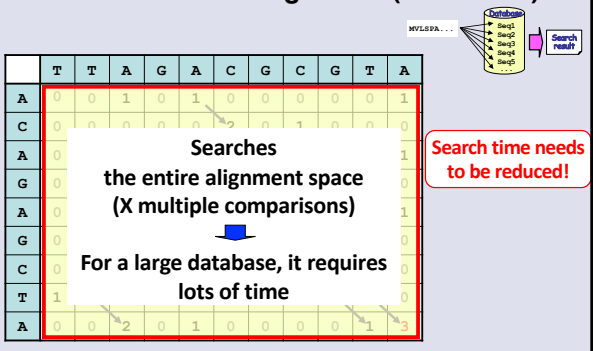
Similarity search as multiple pairwise alignments



BIOS477/877 L11 - 19

19

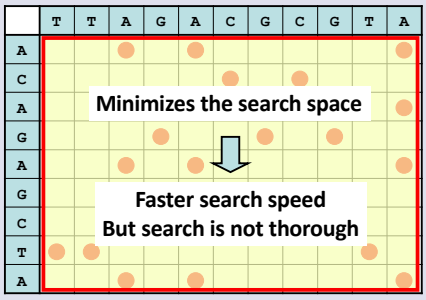
Similarity search using Smith-Waterman algorithm (SSEARCH)



BIOS477/877 L11 - 20

20

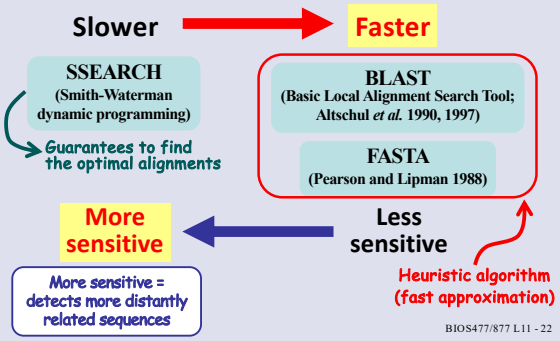
Heuristic algorithm (FASTA/BLAST)



BIOS477/877 L11 - 21

21

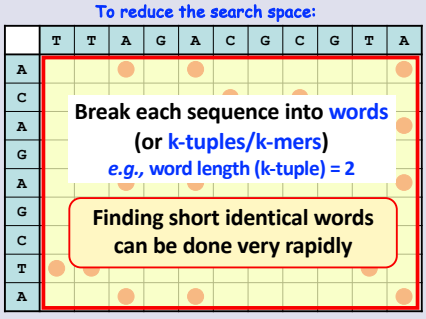
Search speed and sensitivity



BIOS477/877 L11 - 22

22

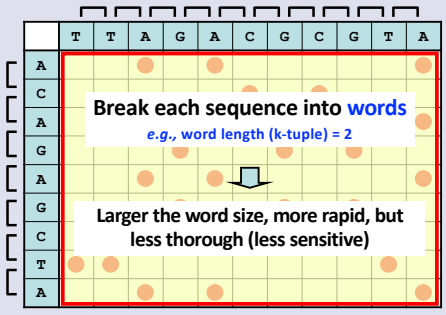
Word/k-tuple matching



BIOS477/877 L11 - 23

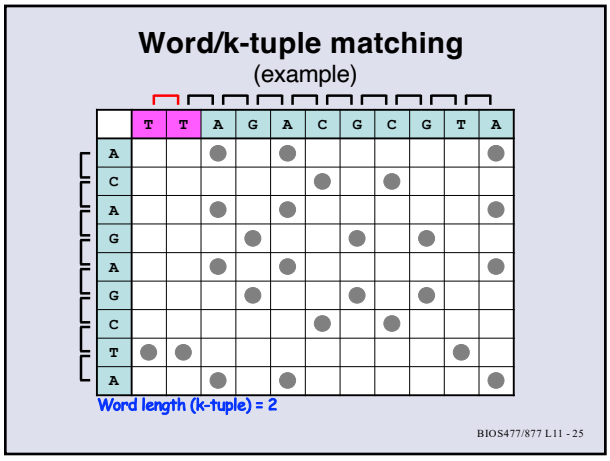
23

Word/k-tuple matching

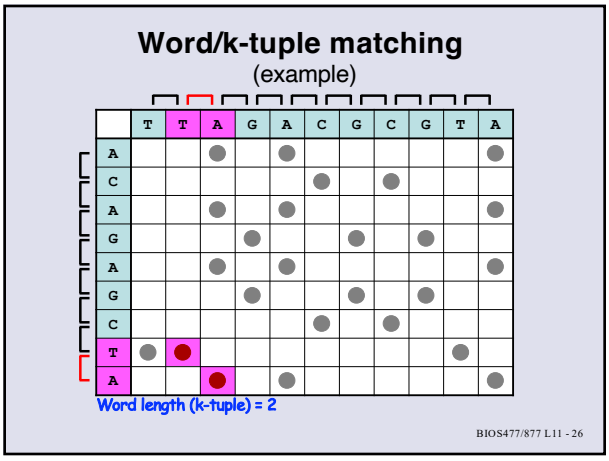


BIOS477/877 L11 - 24

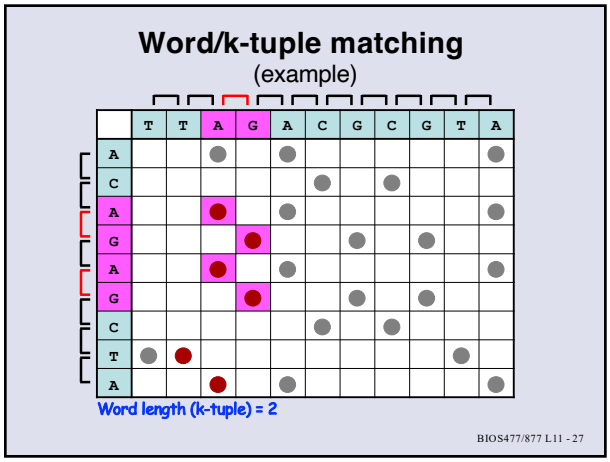
24



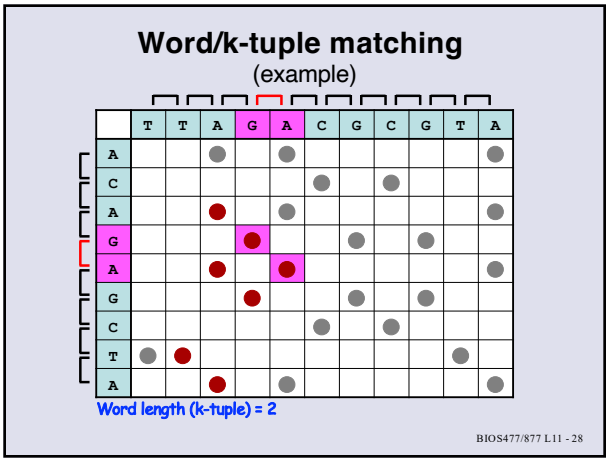
25



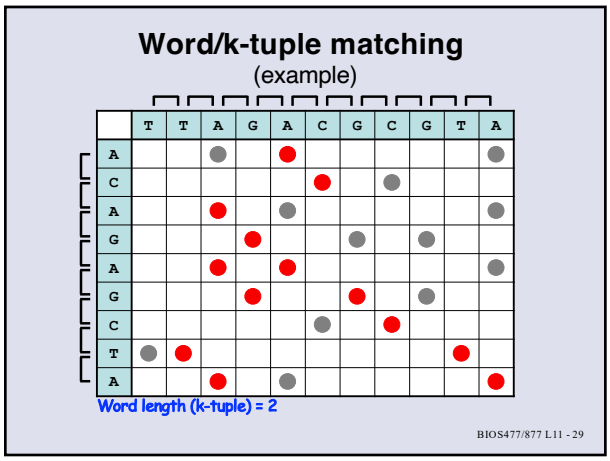
26



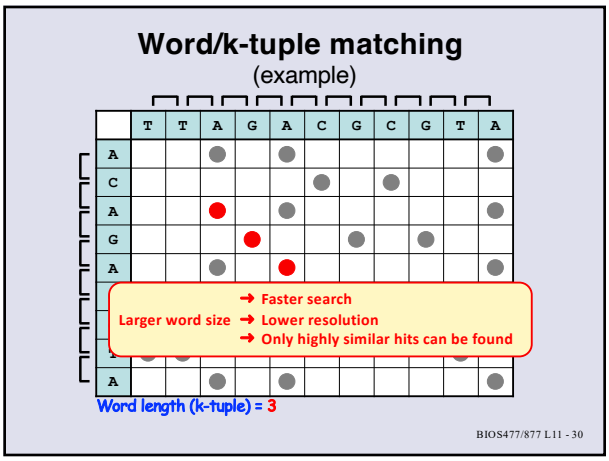
27



28



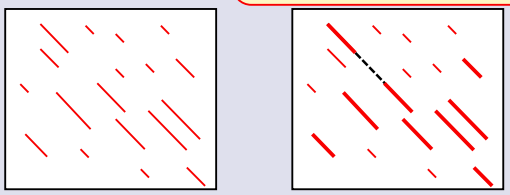
29



30

FASTA algorithm

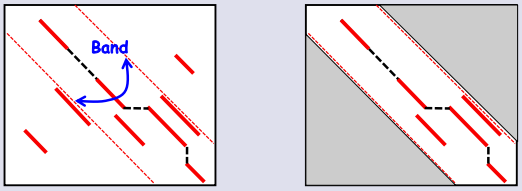
1. Find identities using k-tuples
   (k-tuples = words)
2. Join diagonals without gaps
3. Choose top 10 diagonals using a scoring matrix (e.g., BLOSUM62)
   (init1: the top diagonal score)



31

FASTA algorithm

4. Join again with gaps
   (initn: score of the longer diagonal)
5. A diagonal band is defined
   (width: 32 if k=1, 16 if k=2 for protein)
6. Find optimal local alignment using dynamic programming algorithm within the band
   (opt: the final score)



32

FASTA similarity search

- Ranking
  - Database sequences are ranked based on z-values or OPT scores
    → z-value = the number of standard deviations from the mean (standardized score)
- Alignment
  - The high scored sequence pairs are aligned using the full Smith-Waterman dynamic programming algorithm
    → Not just from the diagonal band → better alignment
- FASTA/SSEARCH website (more on CANVAS)
  Original FASTA server @ U Virginia (By W. Pearson)
  FASTA website @ EBI

33

BLAST similarity search

https://blast.ncbi.nlm.nih.gov/Blast.cgi

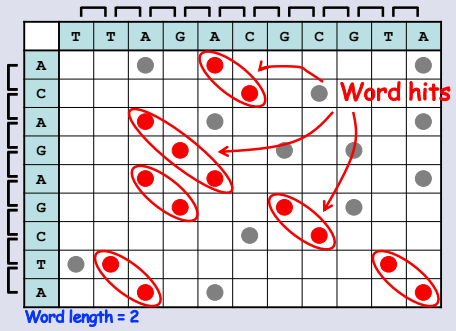
34

BLAST resources

- BLAST Help
  https://blast.ncbi.nlm.nih.gov/doc/blast-help/
  - Guide to BLAST home and search pages
  - BLAST Command Line User Manual
  - BLAST Tutorials on YouTube
- BLAST Statistics
  https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html
- "BLAST" (online version)
  (see more on CANVAS)

35

BLAST algorithm: seeding



36

BLAST algorithm: seeding
(including neighborhood)

- Using words reduces the search space
- Neighborhood increases the sensitivity

Match = 2
Mismatch (Ts) = -1
Mismatch (Tv) = -5

TC / TC = 2 + 2 = 4

TC / TT = 2 - 1 = 1

TC / CC = -1 + 2 = 1

...

TC / TA = 2 - 5 = -3

TC / TG = 2 - 5 = -3

TC / AC = -5 + 2 = -3

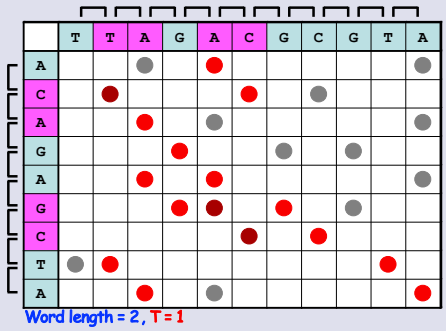
TC / AT = -5 - 1 = -6

Neighborhood

Neighborhood Threshold (T) = 1
[minimum score allowed to be the neighborhood]

37

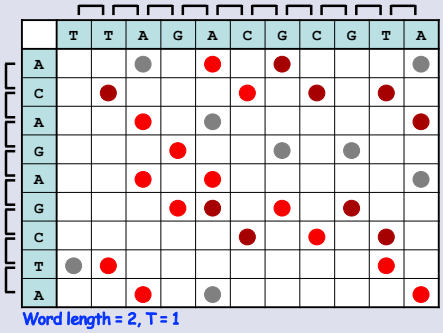
BLAST algorithm: seeding
(including neighborhood)



Word length = 2, T = 1

38

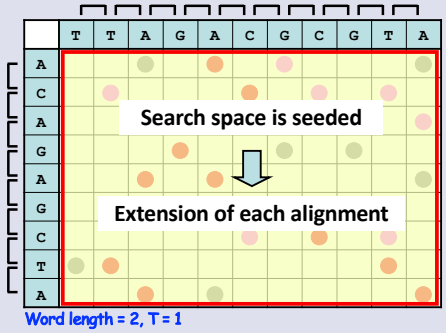
BLAST algorithm: seeding
(including neighborhood)



Word length = 2, T = 1

39

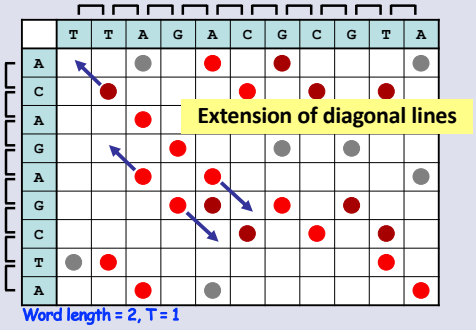
BLAST algorithm



Word length = 2, T = 1

40

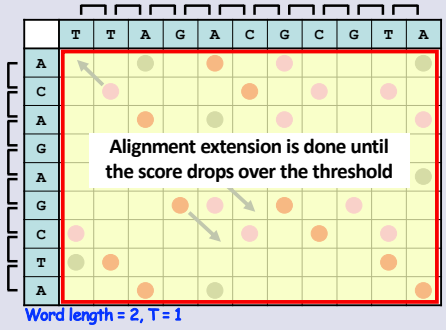
BLAST algorithm: extension



Word length = 2, T = 1

41

BLAST algorithm: extension

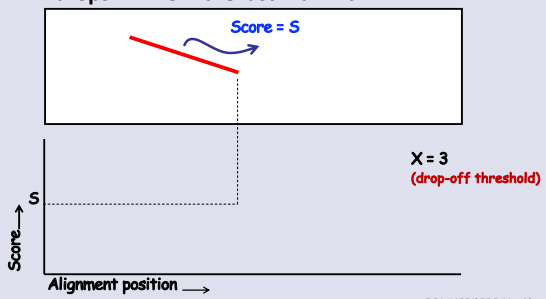


Word length = 2, T = 1

42

BLAST algorithm: extension

- Alignment extension is done until its score drops ≥ X from the last maximum

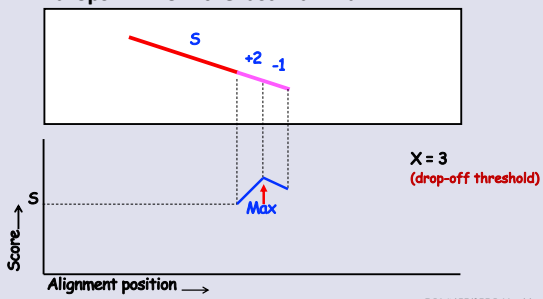


BIOS477/877 L11 - 43

43

BLAST algorithm: extension

- Alignment extension is done until its score drops ≥ X from the last maximum

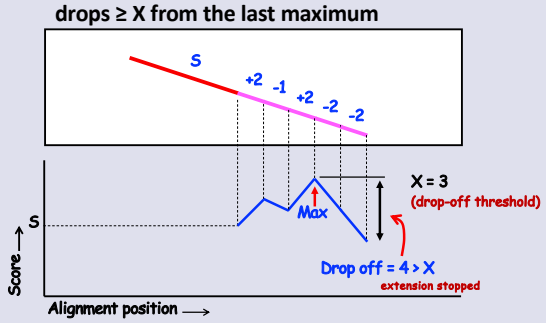


BIOS477/877 L11 - 44

44

BLAST algorithm: extension

- Alignment extension is done until its score drops ≥ X from the last maximum

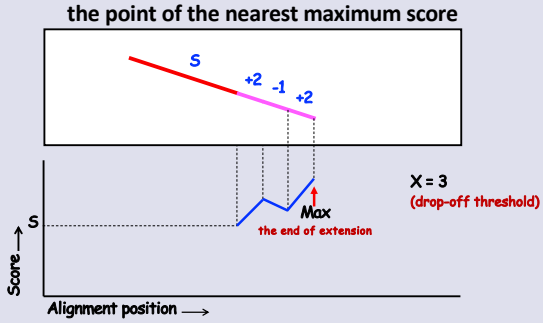


BIOS477/877 L11 - 45

45

BLAST algorithm: extension

- Alignment extension is trimmed back to the point of the nearest maximum score

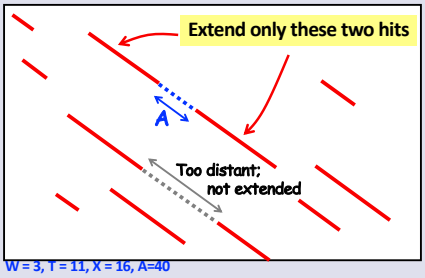


BIOS477/877 L11 - 46

46

BLAST algorithm: extension
(two-hit method)

- Two-hit method: Extension is invoked only when two hits are found within distance A and on the same diagonal



W = 3, T = 11, X = 16, A=40

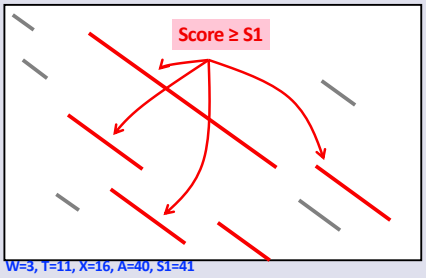
Many thresholds are set!

BIOS477/877 L11 - 47

47

BLAST algorithm: HSP selection

- HSPs (High-scoring Segment Pairs) scoring higher than the threshold (S1) are chosen



W=3, T=11, X=16, A=40, S1=41

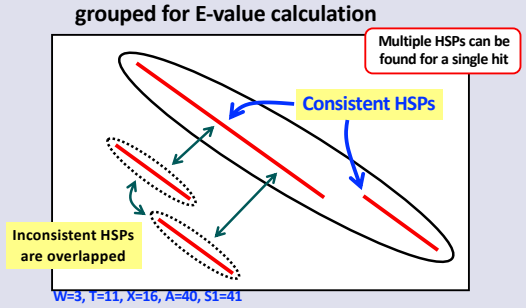
Even more thresholds!

BIOS477/877 L11 - 48

48

BLAST algorithm: HSP selection

- Consistent (non-overlapping) HSPs are grouped for E-value calculation



W=3, T=11, X=16, A=40, S1=41

49

BLAST algorithm: gapped extension

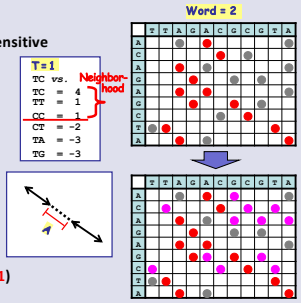
- Gapped extension is done only for selected HSPs
- Gapped alignments scoring higher than the threshold (S2) is chosen as the final set

W=3, T=11, X=16, A=40, S1=41, S2=70

50

BLAST algorithm: summary

- Word-matching size (W)
  - → Longer words: faster but less sensitive
- Neighborhood threshold (T)
  - → Lower T:
    - ○ detects weaker similarities
    - ○ slower but more sensitive
- Extension
  - Drop-off score threshold (X)
  - Two-hit method (A: distance b/w 2 hits)
- HSP selection (ungapped alignment threshold: S1)
- Gapped HSP extension (gapped alignment threshold: S2)



51