

Spring 2024
BIOS 477/877
Bioinformatics and Molecular Evolution
Lecture 11

BIOS477/877 L11 - 1

1

TODAY'S TOPICS

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

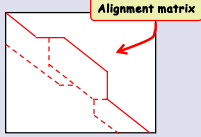
BIOS477/877 L11 - 2

2

Pairwise alignment summary

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

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



BIOS477/877 L11 - 3

3

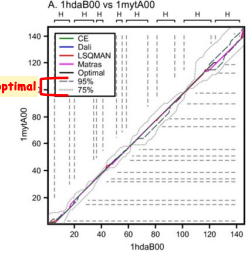
BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

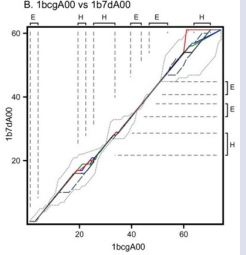
Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments

Michael L. Siek¹, Michael E. Smoot², Ellen J. Bass³, William R. Pearson^{1*}

A. 1hdaB00 vs 1myA00



B. 1bcgA00 vs 1b7dA00



suboptimal

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

BIOS477/877 L11 - 4

4

Pairwise alignment summary (continued)

- Optimal alignments and biologically meaningful alignments may not be the same
- Depending on the scoring system, unreasonable alignments can become optimal
- We need to choose a better (biologically reasonable) scoring system: level of divergence (scoring matrices), gap penalty (affine, etc.), algorithm (local, global, or semi-global)
- Manual adjustment may be necessary
- Test statistical significance of the alignment (is the alignment possible just by chance?)

BIOS477/877 L11 - 5

5

Significance of Alignment Scores

- Hypothesis testing (General)
 - Two hypotheses
 - Null-hypothesis
 - H_0 : The previous (original) belief is true
 - Alternative hypothesis
 - H_1 : The previous (original) belief is false; the new theory is true
 - S : Test statistic
 - Significance level is chosen a priori (e.g., 0.05)
 - P-value: $P(S|H_0 \text{ is true})$ Probability of getting S if H_0 is true
 - If $P < \text{Significance level}$, reject H_0

BIOS477/877 L11 - 6

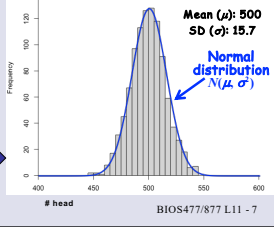
6

Significance of Alignment Scores

- **P-value:** $P(S|H_0 \text{ is true})$
 - Need to be calculated from the test statistic S
 - Need to know the probability distribution of the test statistic S under H_0

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

(Example)
Experiment: 1000 coin tossing
- Count the number of heads
- Repeat 1000 experiments
(Expect to see 500 heads/experiment)



Mean (μ): 500
SD (σ): 15.7
Normal distribution $N(\mu, \sigma)$

BIOS477/877 L11 - 7

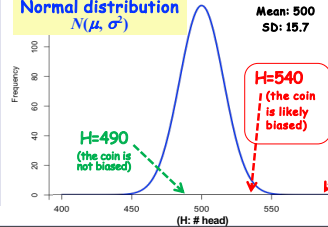
7

Significance of Alignment Scores

- **P-value:** $P(S|H_0 \text{ is true})$
 - Need to be calculated from the test statistic S
 - Need to know the probability distribution of the test statistic S under H_0

Normal distribution
 $N(\mu, \sigma^2)$
Mean: 500
SD: 15.7

S: # of head
 $P(S \geq 490|H_0) \gg 0.05$
 $P(S \geq 540|H_0) < 0.05$
 $P(S \geq 600|H_0) \ll 0.05$



BIOS477/877 L11 - 8

8

Significance of Alignment Scores

- Hypothesis testing for sequence alignment
 - Two hypotheses
 - Null-hypothesis
 H_0 : Two sequences are not related (random)
 - Alternative hypothesis
 H_1 : Two sequences are related
 - Test statistic: **alignment score (S)**
 - **Significance level** is chosen a priori (e.g., 0.05)
 - **P-value:** $P(S|H_0 \text{ is true})$
Probability of getting the alignment score S , even if the two sequences are not related but randomly matched
 - If $P < \text{Significance level}$, reject H_0
(The score should not be obtained just by aligning unrelated sequences)

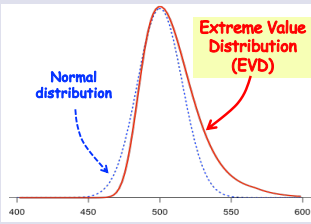
BIOS477/877 L11 - 9

9

Significance of Alignment Scores

- **P-value:** $P(S|H_0 \text{ is true})$
 - Need to be calculated from the test statistic S
 - Need to know the probability distribution of the test statistic S under H_0

Extreme Value Distribution (EVD)
Distribution of alignment scores follow
Extreme Value Distribution (Gumbel distribution)
The probability distribution of highest values in an experiment (e.g., optimal alignment scores)

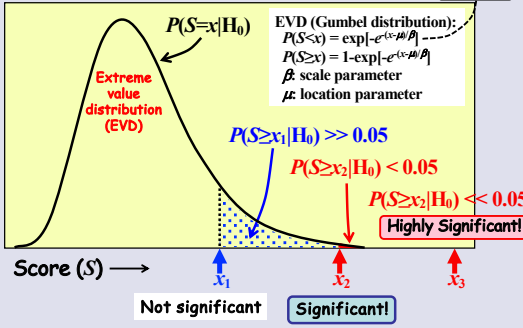


BIOS477/877 L11 - 10

10

Significance of Alignment Scores

EVD (Gumbel distribution):
 $P(S < x) = \exp[-e^{-(x-\mu)/\beta}]$
 $P(S \geq x) = 1 - \exp[-e^{-(x-\mu)/\beta}]$
 β : scale parameter
 μ : location parameter



BIOS477/877 L11 - 11

11

Significance of Alignment Scores

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

Karlin-Altschul equation (Karlin and Altschul 1990)
 $P(S \geq x) = 1 - \exp[-Kmn e^{-\lambda x}] \approx Kmn e^{-\lambda x}$

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

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

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

$P(S \geq x) = Kmn e^{-\lambda x}$

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

BIOS477/877 L11 - 12

12

Significance of Alignment Scores

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

- 1) The second sequence is shuffled many times.
 (simulates random sequences)
- 2) Smith-Waterman local alignment score is calculated from each alignment: $P(S \geq x | H_0)$
- 3) The distribution is fitted to an extreme value distribution to obtain estimates of K and λ
- 4) P-value is estimated based on the K and λ , and the original alignment score x : $P(S \geq x) \approx Kmne^{-\lambda x}$

BIOS477/877 L11 - 13

13

Simulation of Alignment Scores

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

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

```

RECA_ECOLI  3  IDENK-QRALAALQETKDFPGGIMLGEDRMHVVYITGSELEDTA  51
RAD51_YEAST 124 YERARDELKIMAEKALVWQVYADKPKM---MRELKELTQSGRSDVPL  170
RECA_ECOLI  52  LGAGLWMCRIEIVGPRSSGRTT-----TLQV-IAAQRGKTCAPT  94
RAD51_YEAST 171 LG-QDYERGSITVELPEFGRFDRSGLCHLAVTQVPLDGGSDGSR-CELT  218
RECA_ECOLI  95  DAEKSLQDF---VAREKSVDFI---MELCGDFYVQVQLAEIEMILAM-  135
RAD51_YEAST 219 DFRTPTFPVPLVSIAGDFGLDDGMNLNVARATRADGRGRLGLDAAMQ  268
RECA_ECOLI  136 --SGNVVIVDQVALTPEAIT--EEEGDSBHLLAARMGQARMRLG  181
RAD51_YEAST 269 MBEERFSLVYDQVMALT-RTDFSGRGLSARMGKLAQFM--KALQLA-  314
RECA_ECOLI  182  NLAGQSTLLIFITQIMKRI--GVWPG-KPETPTDGNALKFAVAVLDIRK  228
RAD51_YEAST 315 --DQGVAVVYTRGVAQVGGMARNFDFPKKFGIMNMBEETRL----  358
RECA_ECOLI  229  IGAVRGEVYVGGESTVFFVVKHKLAAFPQAEFLYCGEII  269
RAD51_YEAST 359 --GPKRGE---GQRIKGVV-DKCLPEACEVFAI-YEDVQ  392
  
```

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

BIOS477/877 L11 - 14

14

Simulation of Alignment Scores

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

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

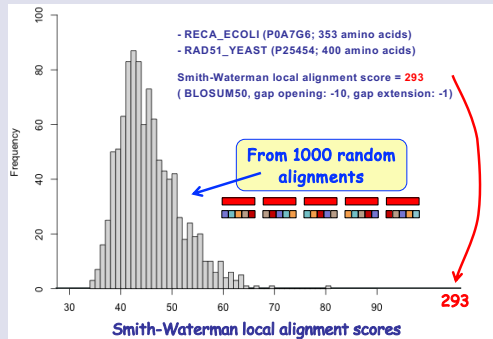
Shuffle RAD51_YEAST 1000 times
 (generate 1000 random sequences)
 ↓
 Align with RECA_ECOLI
 (generate 1000 random local alignments)

RECA_ECOLI → [red bars] (score?) (score?) (score?) (score?) (score?)
 shuffled
 RAD51_YEAST → [multicolored bars] (score?) (score?) (score?) (score?) (score?)

BIOS477/877 L11 - 15

15

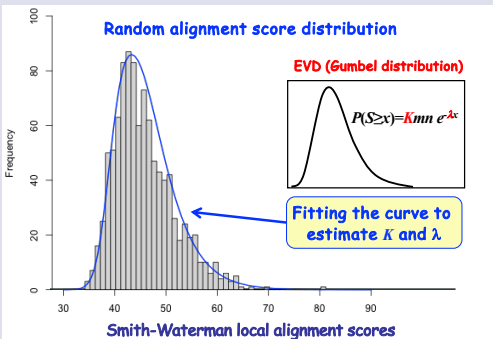
Simulation of Alignment Scores



BIOS477/877 L11 - 16

16

Simulation of Alignment Scores



BIOS477/877 L11 - 17

17

FASTA Web server by William Pearson

FASTA Sequence Comparison at the U. of Virginia

Uva FASTA Server

New Annotation features available for SwatchProt91 library searches.

About

- Getting started
- fasta_guide.pdf

Other FASTA Servers

- EMBL-EBI
- NCBI-Genbank

References

- Protein-protein FASTA
- Protein-protein Smith-Waterman (research)
- Global Protein-protein (Needleman-Neuwirth)
- (ggsearch)
- Global/local protein-protein (lgsearch)
- Protein-protein with unordered peptides (tsearch)
- Protein-protein with repeat peptide sequences (tsearch)

Software

- FASTA v06
- Changelog
- Downloads
- Sequence Library
- Developer/Mailing list

Other resources

- CHAPPE - Convert FASTA and Profiles
- Near optimal alignments
- FASTA Express
- NCBI-BLAST
- Swat
- EMBL-EBI Server

Protein

- Nucleotide Nucleotide (DNA/RNA Blast)
- Ordered Nucleotides vs Nucleotide (blast)
- Unordered Nucleotides vs Nucleotide (blast)

Translated

- Translated DNA (with frameshifts, e.g. GSThs vs Protein (blastx/txsl))
- Protein vs Translated DNA (with frameshifts) (blastx/txsl)
- Peptides vs Translated DNA (blastp)

Statistical Significance

- Protein vs Protein (local) (glocal)
- DNA vs DNA (local) (glocal)
- Translated DNA vs Protein (local) (glocal)

Local Duplications

- Local Protein alignments (align)
- Plot Protein alignment "dot plot" (align)
- Local DNA alignments (align)
- Plot DNA alignment "dot plot" (align)



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

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

The origin of FASTA format

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

BIOS477/877 L11 - 18

18

Statistical Analysis of Alignments by PRSS

Statistical Significance from Shuffles
Search Databases with FASTA
Find Internal Duplications (alignalign)

PRSSPREFIX compute the statistical significance of an alignment by aligning the two sequences, and then shuffling the second sequence 200 - 1000 times, and estimating the statistical significance from the distribution of shuffled alignment scores.
 Window shuffles are used to preserve local sequence composition, e.g. for transmembrane proteins.
 New: Annotation features available for UniProt/SwissProt/TrEMBL library searches.

Choose: (A) program and (B,C) sequences to compare:

(A) Program: prss_prefix
 (B) Number of shuffles: 1000
 (C) Uniform Window

(B.1) Enter first (query) sequence: FASTA format | Subst range: Annotate Query Sequence (SwissProt / TrEMBL / UniProt)

(B.2) Or upload sequence from file: Upload sequence (FASTA)

(C.1) Enter the second sequence: FASTA format | Subst range: Annotate Target Sequence (SwissProt / TrEMBL / UniProt)

(C.2) Or choose file of sequences/accessions: Upload sequence (FASTA)

Other comparison options:
 Scoring matrix: BLOSUM62
 Gap opening: -10
 Gap extension: -2

The second sequence is shuffled 1000 times to generate random sequences.
→ Amino acid composition can be maintained.
→ Window size can be set so that shuffling is done locally

BIOS477/877 L11 - 19

Statistical Analysis of Alignments by PRSS

```
# search36 q -t -w 80 -m 6 -Z 10000 -E 10000 -f 10 -g 2 -E TMP.q2
SEARCH program a Smith-Waterman search
Version: 36.3-81 Rev. 2823
URL: http://www.smb.jku.at/~waterman/1199937/, Mol. Biol. 147:195-197;
W.R. Pearson 1993; J. Comput. Biol. 6:1-11
OUTDIR: /tmp/PRA7G6/RECA_ECOLI Protein RecA Recombinase A - 353 aa
Library: TMP.q2
The best scores are:
sp|P24454|RAD51_YEAST DNA repair protein RAD51 (400) s.w. bits E(10000)
sp|P0A7G6|RECA_ECOLI Protein RecA Recombinase A - 353 aa s.w. bits E(10000)
>>>sp|P0A7G6|RECA_ECOLI Protein RecA Recombinase A - 353 aa vs TMP.q2 library
Statistics: (shuffled 1000) NLE statistics: Lambda= 0.1376; K=0.01032
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.2 Nov 2016/SMDc Nov 2020)
Parameters: BLO matrix (15-5), open/ext: -10/-2
Scan time: 0.080
```

BIOS477/877 L11 - 20

19

20

Statistical Analysis of Alignments by PRSS

```
Query: 0
1->sp|P0A7G6|RECA_ECOLI Protein RecA; Recombinase A - 353 aa
Library: TMP.q2
400 residues in 1 sequences
Statistics: (shuffled 1000) NLE statistics: Lambda= 0.1376; K=0.01032
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.2 Nov 2016/SMDc Nov 2020)
Parameters: BLO matrix (15-5), open/ext: -10/-2
Scan time: 0.080
```

The best scores are:
 sp|P24454|RAD51_YEAST DNA repair protein RAD51 (400) s.w. bits E(10000)
 sp|P0A7G6|RECA_ECOLI Protein RecA Recombinase A - 353 aa s.w. bits E(10000)

$P(S > x) \approx Kmne^{-\lambda x}$
 $\lambda = 0.1376, K = 0.01032$
 $x = 236, m = 353, n = 400$
 $E(10000) = 10000 \times P = 1.1 \times 10^{-7}$
 $P(S > 236) = 0.0116 \times 353 \times 400 \times e^{-0.1376 \times 236} = 1.149 \times 10^{-11} = 1.1 \times 10^{-11}$

BIOS477/877 L11 - 21

Pairwise alignment by LALIGN

PLALIGN (dot plot)

```
# program: align36-q-d-11-1-10-V-Union.com from
ALIGN From: user-specified local alignments
Version: 36.3-81 Rev. 2823
URL: http://www.smb.jku.at/~waterman/1199937/, Mol. Biol. 147:195-197;
W.R. Pearson 1993; J. Comput. Biol. 6:1-11
OUTDIR: /tmp/PRA7G6/RECA_ECOLI Protein RecA Recombinase A - 353 aa
Library: /tmp/PRA7G6/RECA_ECOLI Protein RecA Recombinase A - 353 aa
Statistics: (shuffled 1000) NLE statistics: Lambda= 0.1393; K=0.01116
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.2 Nov 2016/SMDc Nov 2020)
Parameters: BLO matrix (15-5), open/ext: -10/-2
Scan time: 0.028
```

BIOS477/877 L11 - 22

21

22

Pairwise alignment by LALIGN

```
Query: 0
1->sp|P0A7G6|RECA_ECOLI Protein RecA; Recombinase A - 353 aa
Library: TMP.q2
400 residues in 1 sequences
Statistics: (shuffled 1000) NLE statistics: Lambda= 0.1393; K=0.01116
Threshold: E| < score: 53
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.2 Nov 2016/SMDc Nov 2020)
Parameters: BLO matrix (15-5), open/ext: -10/-2
Scan time: 0.028
```

$P(S > x) \approx Kmne^{-\lambda x}$
 $\lambda = 0.1393, K = 0.01116$
 $x = 236$
 $m = 353, n = 400$
 $P(S > 236) = 0.01116 \times 353 \times 400 \times e^{-0.1393 \times 236} = 8.321 \times 10^{-12} = 8.3 \times 10^{-12}$

$E(N) = N \times P$
 $E(1) = 1 \times P = P$
 $E(1) = 8.3 \times 10^{-12}$

From PRSS:
 $P = E(1) = 1.1 \times 10^{-11}$

BIOS477/877 L11 - 23

Significance of Alignment Scores

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

← For BLOSUM50;
 a : gap opening penalty
 b : gap extension penalty (affine gap penalty)
 Based on 10,000 random sequence pairs,
 → calculated optimal local alignment scores, S .
 → K and λ are estimated by fitting the distribution of S with $P(S > x) \approx Kmne^{-\lambda x}$
 This is what BLAST does!

a	b	K (bits)	λ (bits)	H (bits)
0-10	0.252	0.11	0.54	73
11	0.197	0.05	0.21	11
11-6	0.222	0.08	0.31	11
16	0.215	0.06	0.27	11
16-2	0.207	0.05	0.24	11
16-1	0.180	0.024	0.15	11
15	0.222	0.09	0.31	10
15-6	0.219	0.08	0.29	10
15-4	0.216	0.07	0.26	10
15	0.210	0.06	0.25	10
15-2	0.202	0.05	0.22	10
15-1	0.166	0.018	0.11	10
14	0.218	0.08	0.29	10
14-7	0.214	0.07	0.27	10
14	0.205	0.05	0.24	9
14-3	0.201	0.05	0.23	9
14-2	0.188	0.034	0.17	9
14-1	0.140	0.009	0.07	9
13	0.211	0.06	0.27	8
13-7	0.205	0.05	0.24	8
13-4	0.203	0.05	0.23	8
13-3	0.188	0.034	0.18	8
13-2	0.187	0.03	0.17	8
13-1	0.114	0.006	0.04	8
12	0.205	0.06	0.24	7
12-6	0.197	0.05	0.21	7
12-4	0.192	0.04	0.18	7
12-3	0.178	0.03	0.15	7
12-2	0.158	0.019	0.10	7
12-1	-	-	-	7
1-6	1-6	-	-	Linear

BIOS477/877 L11 - 24

23

24

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
 - Is the alignment possible just by chance?
 - If the alignment score is statistically significant (not possible simply by chance), the alignment is meaningful.
 - P-value or E-value ($E=NP$) can be compared regardless of the scoring systems used for alignments.

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

$$0 < P\text{-value} < 1 \text{ vs. } 0 < E\text{-value} < N$$

BIOS477/877 L11 - 25

25

Similarity Search

- Why do you want to perform similarity search?
 - To find related genes in another organisms
 - Homologue candidates
 - To identify a possible function of a gene/protein

[From genomic sequence]

- To predict gene structure: against cDNA sequences → exon-intron structure
- To predict gene locations

BIOS477/877 L11 - 26

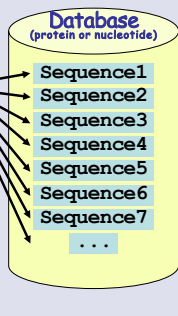
26

Similarity Search

Query sequence
(protein or nucleotide)

MVLSPA...

Pairwise alignment
(local)



Search result

high similarity

Sequence28
Sequence5
Sequence11
Sequence73
Sequence65
Sequence33

low similarity

BIOS477/877 L11 - 27

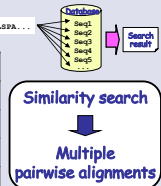
27

Similarity Search = Multiple Pairwise Alignments

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

Pairwise alignment
search space

X multiple comparisons



BIOS477/877 L11 - 28

28

Smith-Waterman Local Alignment (used in SSEARCH)

	T	T	A	G	A	C	G	C	G	T	A
A	0	0	1	0	1	0	0	0	0	0	1
C	0	0	0	0	0	0	2	0	1	0	0
A											
G											
A											
G											
C	0	0	0	0	1	2	1	1	0	0	0
T	1	1	0	0	0	0	1	1	2	2	0
A	0	0	2	0	1	0	0	0	0	1	3

Searches the entire alignment space
(X multiple comparisons)

For a large database, it requires lots of time

Search time needs to be reduced!

BIOS477/877 L11 - 29

29

Heuristic algorithm (FASTA/BLAST)

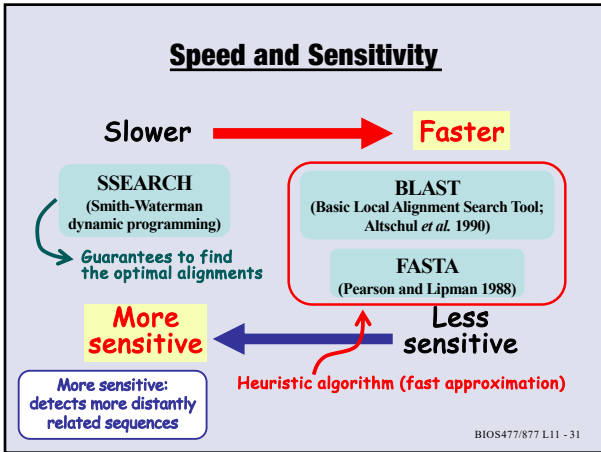
	T	T	A	G	A	C	G	C	G	T	A
A											
C											
A											
G											
A											
G											
C											
T											
A											

Minimizes the search space

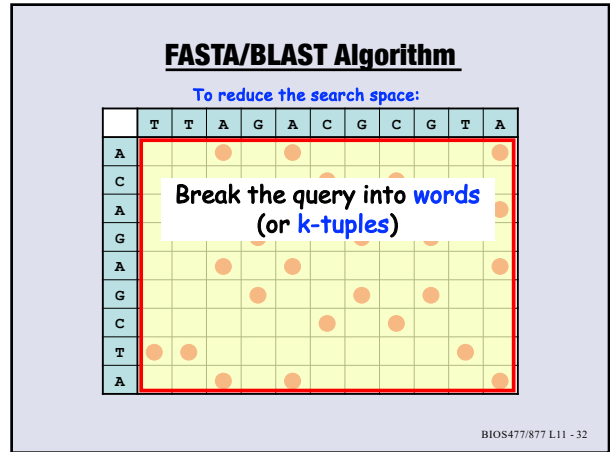
Faster search speed
But search is not thorough

BIOS477/877 L11 - 30

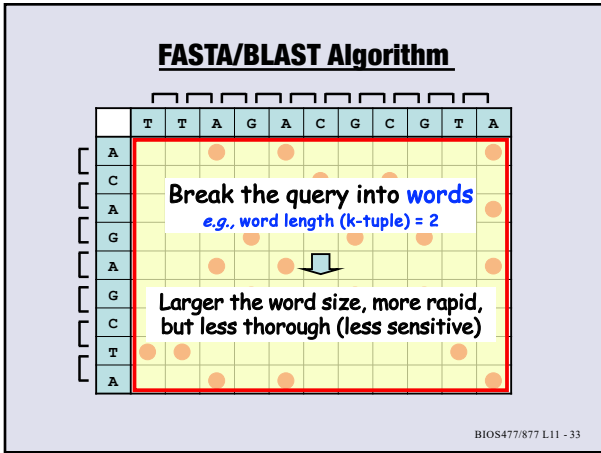
30



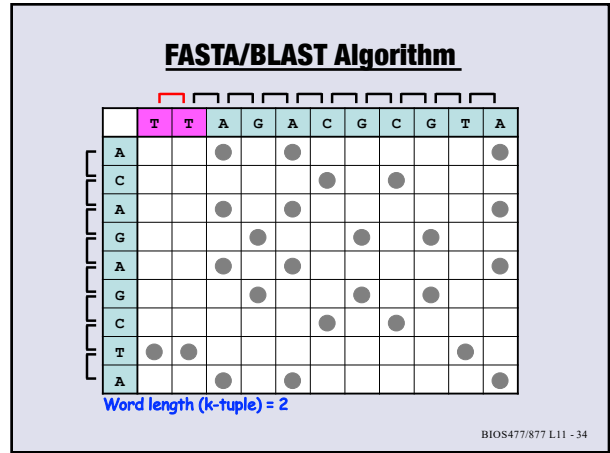
31



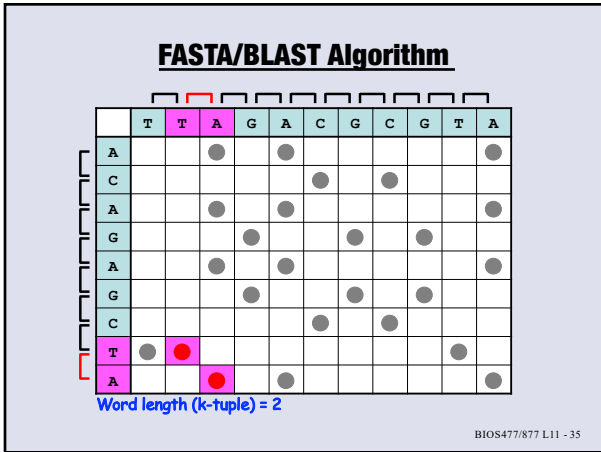
32



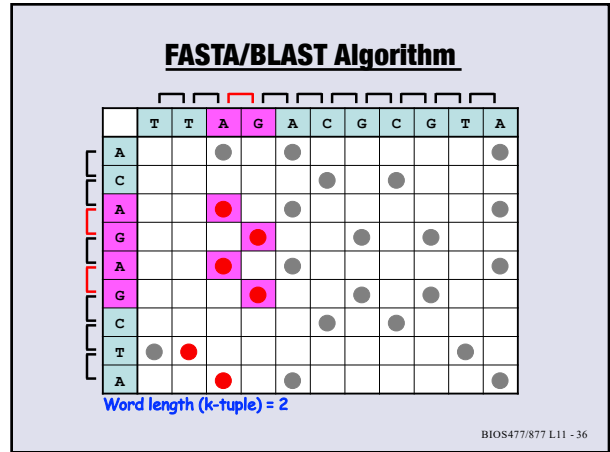
33



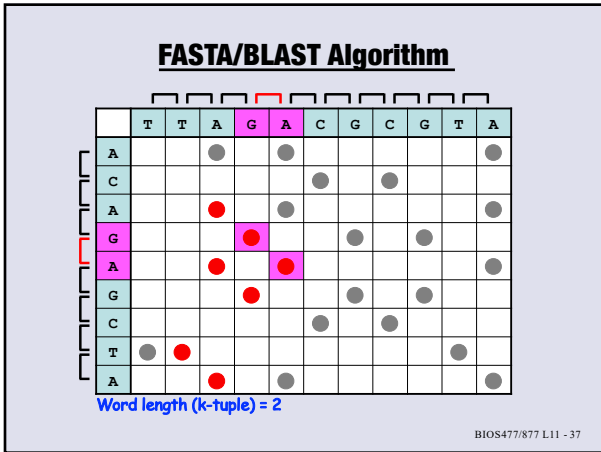
34



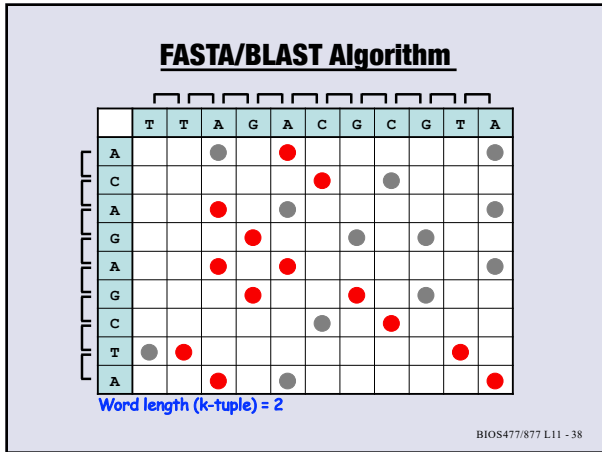
35



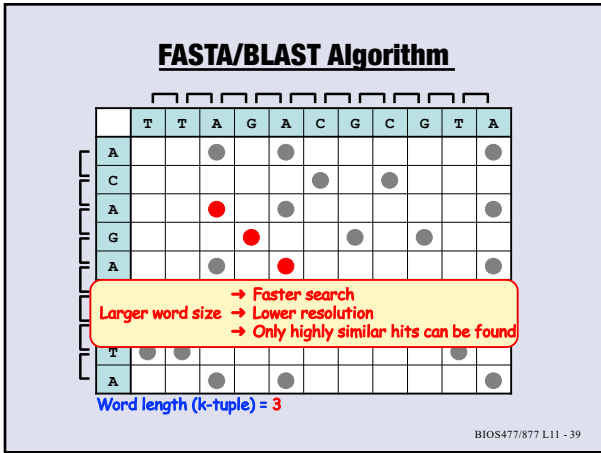
36



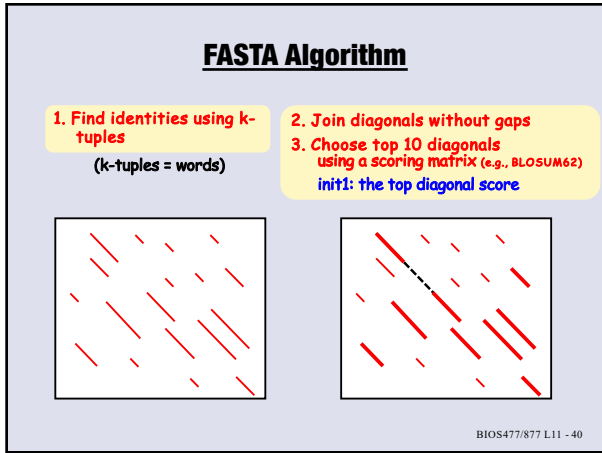
37



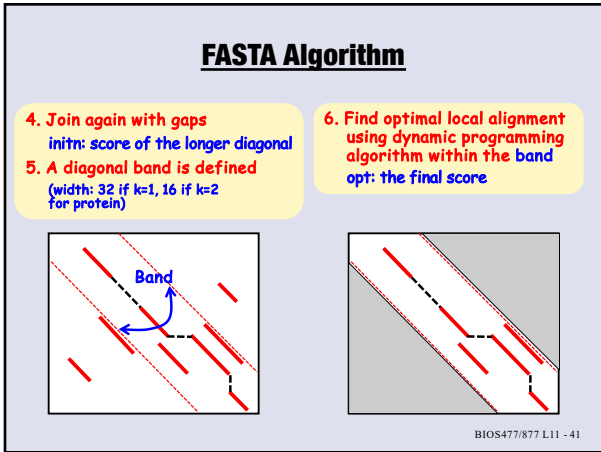
38



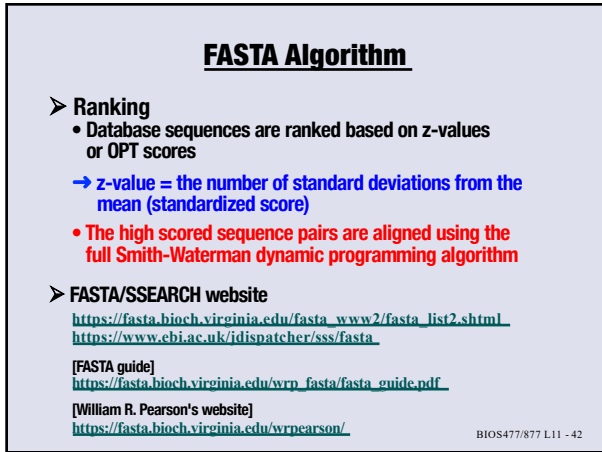
39



40



41



42

BLAST Similarity Search

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

43

BLAST resources

➤ **BLAST**
<https://blast.ncbi.nlm.nih.gov/Blast.cgi>
 [Guide to BLAST home and search pages]
ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf
 (Also available on Canvas)

[BLAST Report Description]
https://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf
 (Also available on Canvas)

[BLAST Statistics]
<https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-L.html>

[BLAST Command Line User Manual]
<https://www.ncbi.nlm.nih.gov/books/NBK279690/>

[BLAST YouTube Tutorials]
 (Link is available from NCBI Help page or from Canvas)

44

BLAST Algorithm: seeding

Word length = 2

45

BLAST Algorithm: seeding

Word length = 2

46

BLAST Algorithm: seeding

➤ Using **words** reduce the search space
 ➤ **Neighborhood** increases the sensitivity

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

TC = 2 + 2 = 4
 TC = 2 + 2 = 4

TC = 2 - 1 = 1
 TC = -1 + 2 = 1
 ...

TC = 2 - 5 = -3
 TA = 2 - 5 = -3
 TC = 2 - 5 = -3
 TC = -5 + 2 = -3
 AC = -5 + 2 = -3
 TC = -5 - 1 = -6
 AT = -5 - 1 = -6

Neighborhood

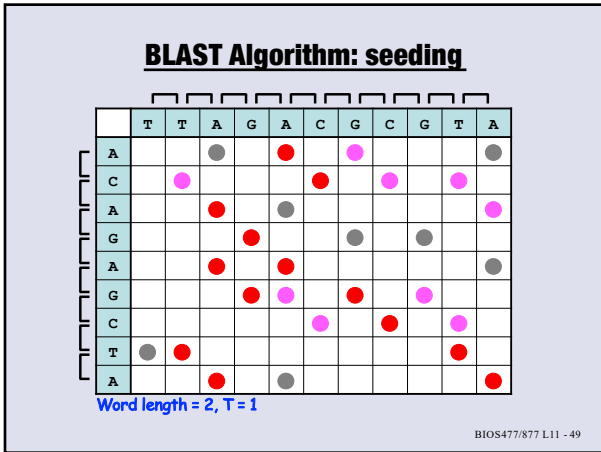
Neighborhood Threshold (T) = 1
 [minimum score allowed to be the neighborhood]

47

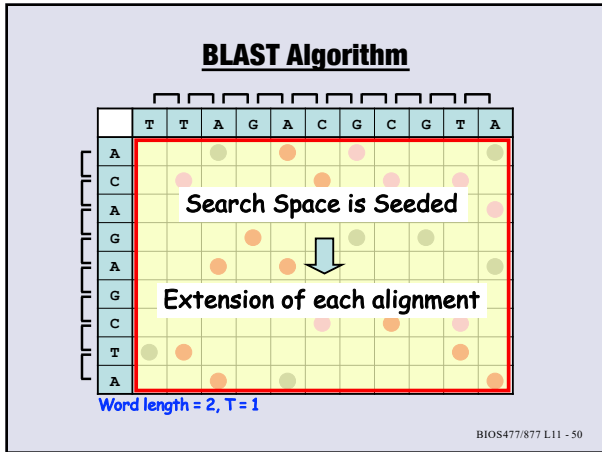
BLAST Algorithm: seeding

Word length = 2, T = 1

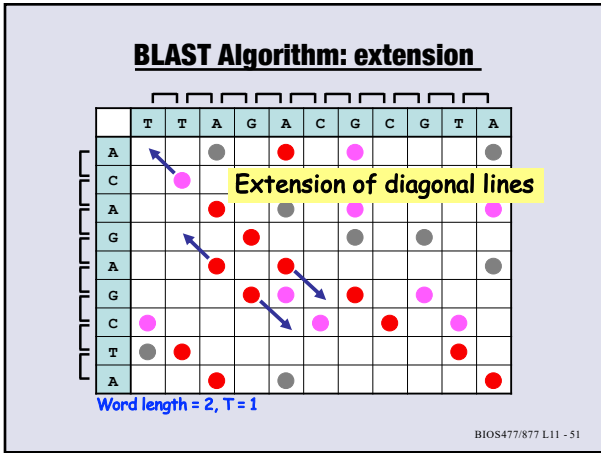
48



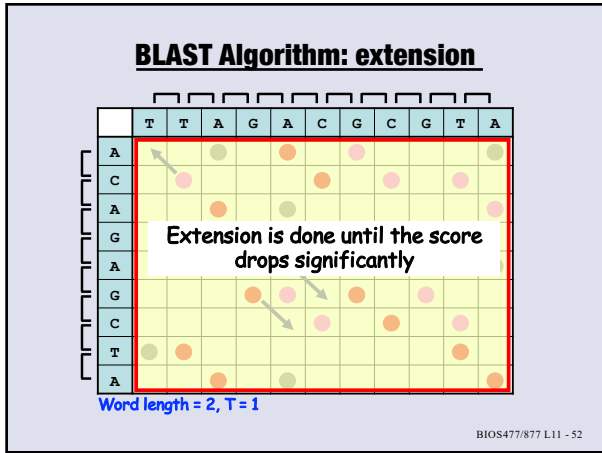
49



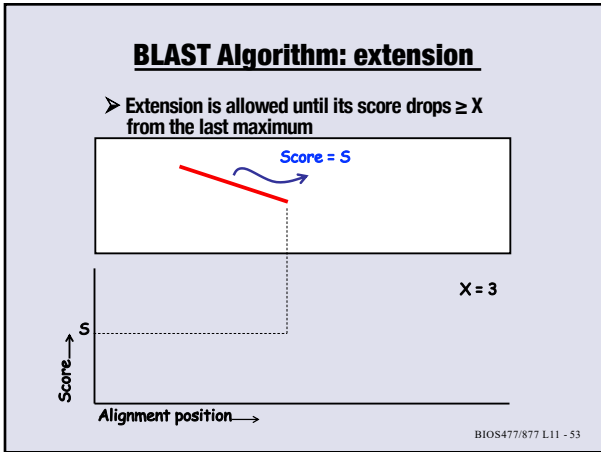
50



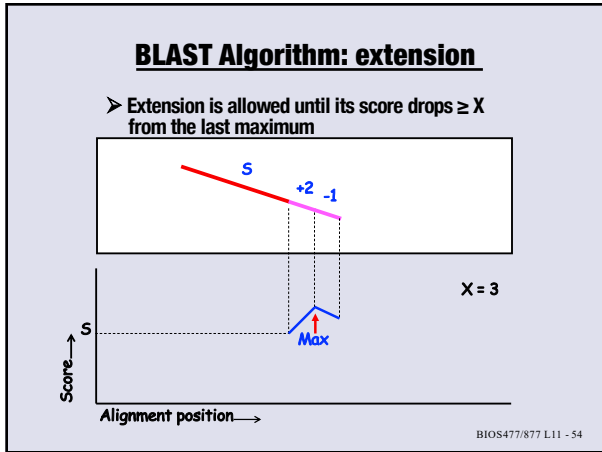
51



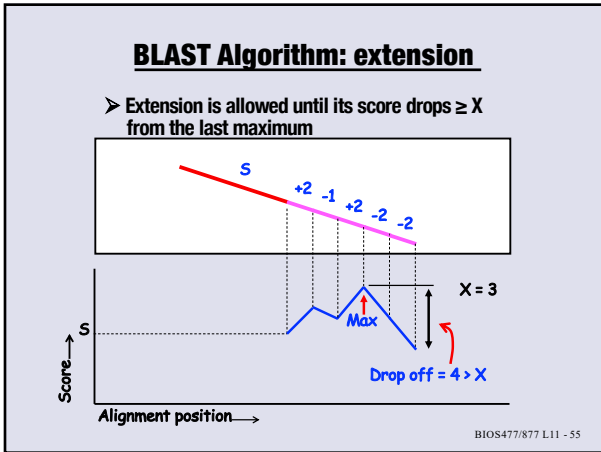
52



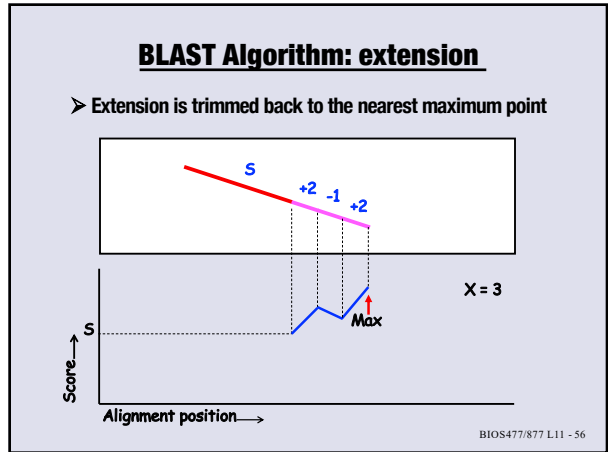
53



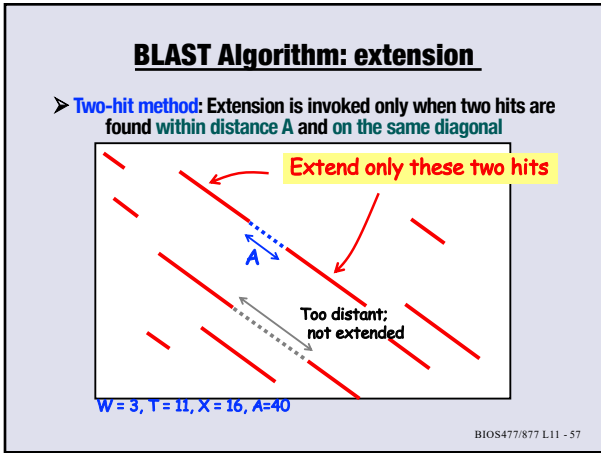
54



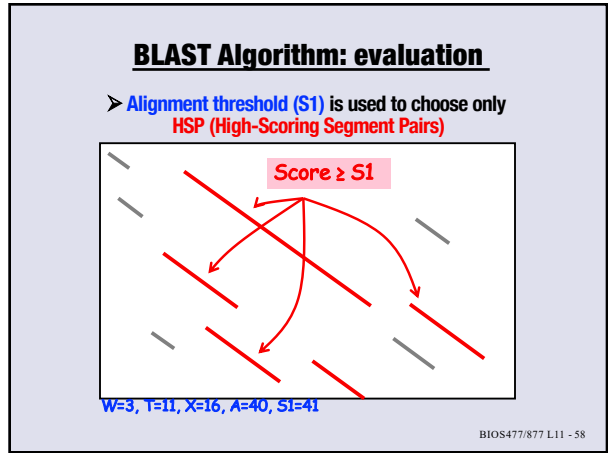
55



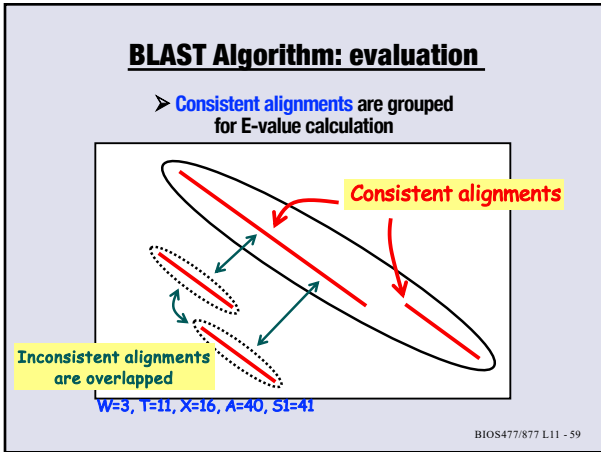
56



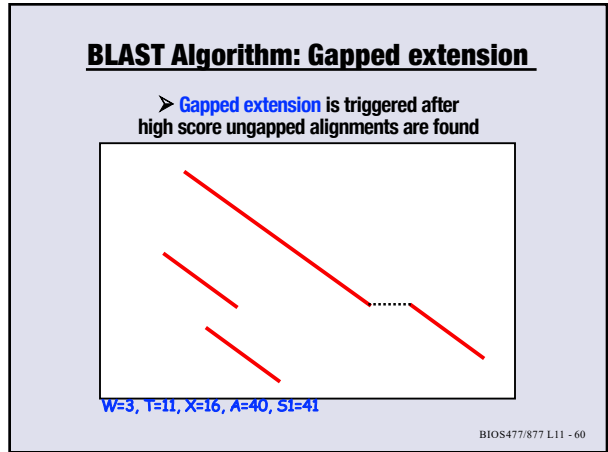
57



58



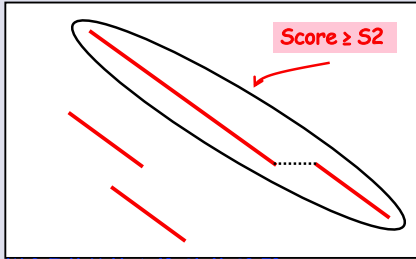
59



60

BLAST Algorithm: Gapped extension

- Another threshold for gapped alignment (S_2) is used to choose the final set of HSPs



$W=3, T=11, X=16, A=40, S_1=41, S_2=70$

BIOS477/877 L11 - 61

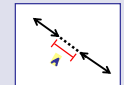
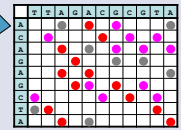
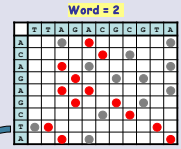
61

BLAST Algorithm

- 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 (X)
 - Two-hit method
 - (A : distance b/w 2 hits)
- HSP selection
 - (ungapped alignment threshold: S_1)
- Gapped HSP extension
 - (gapped alignment threshold: S_2)

T=11	
TC	vs.
TC	= 4
TT	= 1
CC	= 1
CT	= -2
TA	= -3
TG	= -3

Neighborhood



BIOS477/877 L11 - 62

62