

BI694

Bioinformatics & Phylogenetics

Winter Semester 2017

WEEK 4

Overview

- BLAST: **Basic Local Alignment Tool**
 - Local alignment!
 - Used to compare sequences to database
 - Described by Altschul et al. 1990

Overview

- BLAST: **Basic Local Alignment Tool**
 - Identify highest scoring pair (HSP) between sequences
 - Produces ungapped alignments
 - Newer versions allow for gaps in alignment

BLAST algorithm

Database (e.g., genome)

C C A A G G T C A G T

BLAST algorithm

How do we find the best subset of sequences that match between query and database?

We want to find the best ungapped local alignment...

Database (e.g., genome)

C C A A G G T C A G T

A

C

C

Q G

u G

e

r d

 y^T

T

T

T

BLAST algorithm

1) Scoring matrix
with:

$$S_{ii} = +1$$
$$S_{ij} = -1$$
$$i \neq j$$

How do we find the best matching sub-sequences?

[illegible]

BLAST algorithm

1) Scoring matrix
with:

$$S_{ii} = +1$$
$$S_{ij} = -1$$
$$i \neq j$$

2) Determine highest segment score → Highest Scoring Pair (HSP)

[illegible]

BLAST algorithm

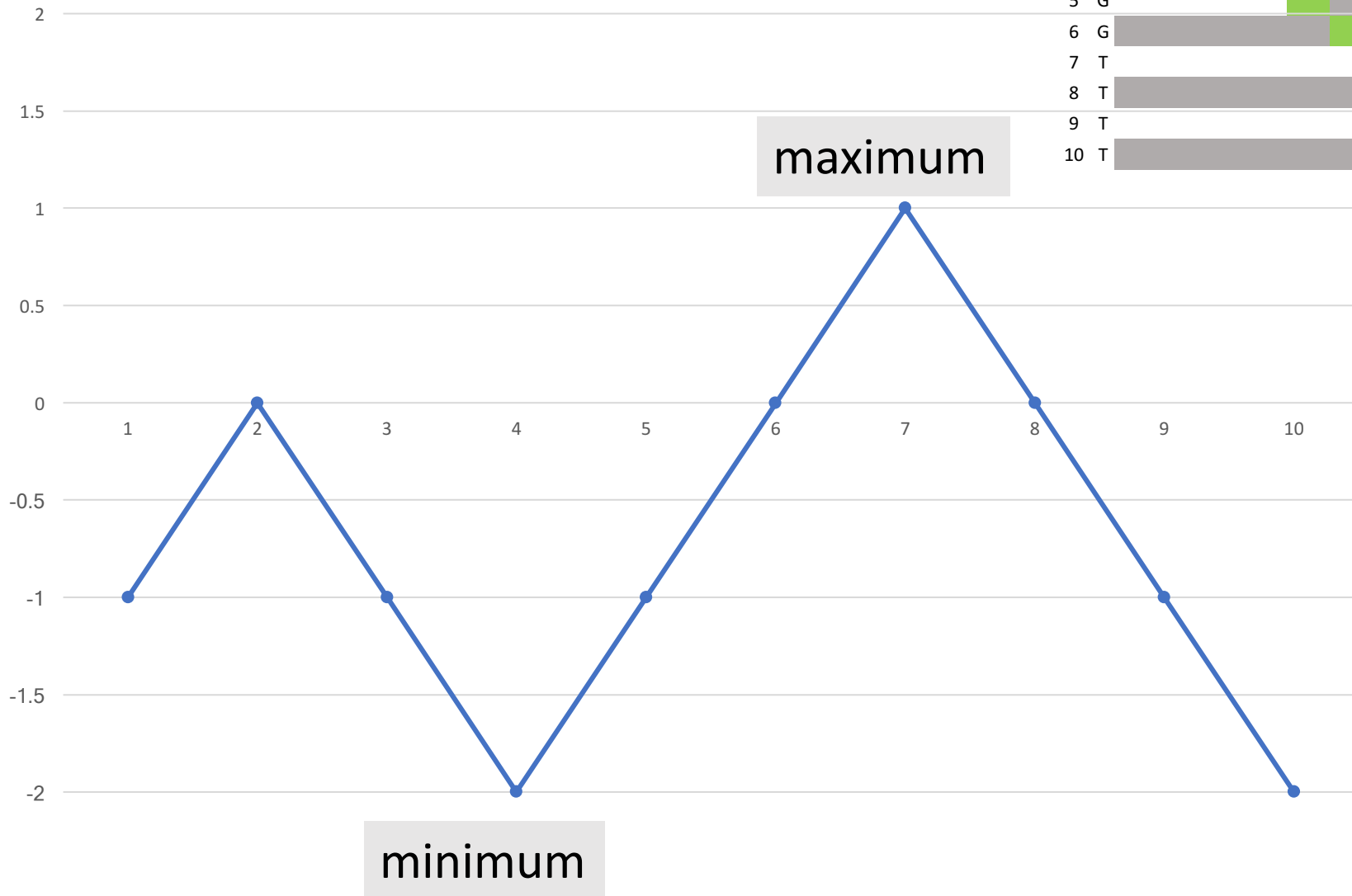
- 1) Scoring matrix
with:

$$S_{ii} = +1$$

$$S_{ij} = -1$$

$$i \neq j$$

- 2) Determine highest segment score → Highest Scoring Pair (HSP)

[illegible]

BLAST algorithm

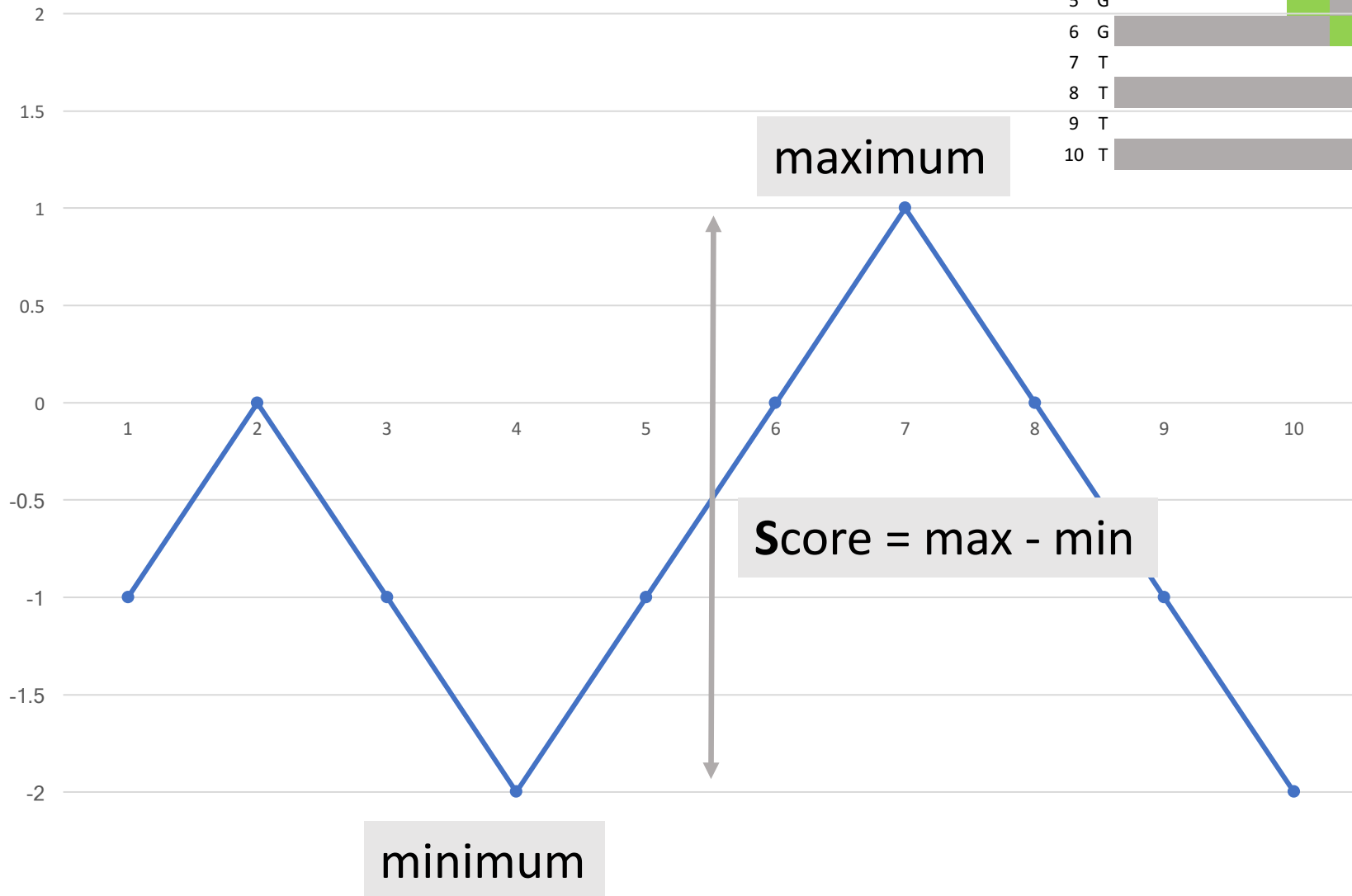
- 1) Scoring matrix
with:

$$S_{ii} = +1$$

$$S_{ij} = -1$$

$$i \neq j$$

- 2) Determine highest segment score → Highest Scoring Pair (HSP)

[illegible]

BLAST algorithm

- 1) Scoring matrix
with:

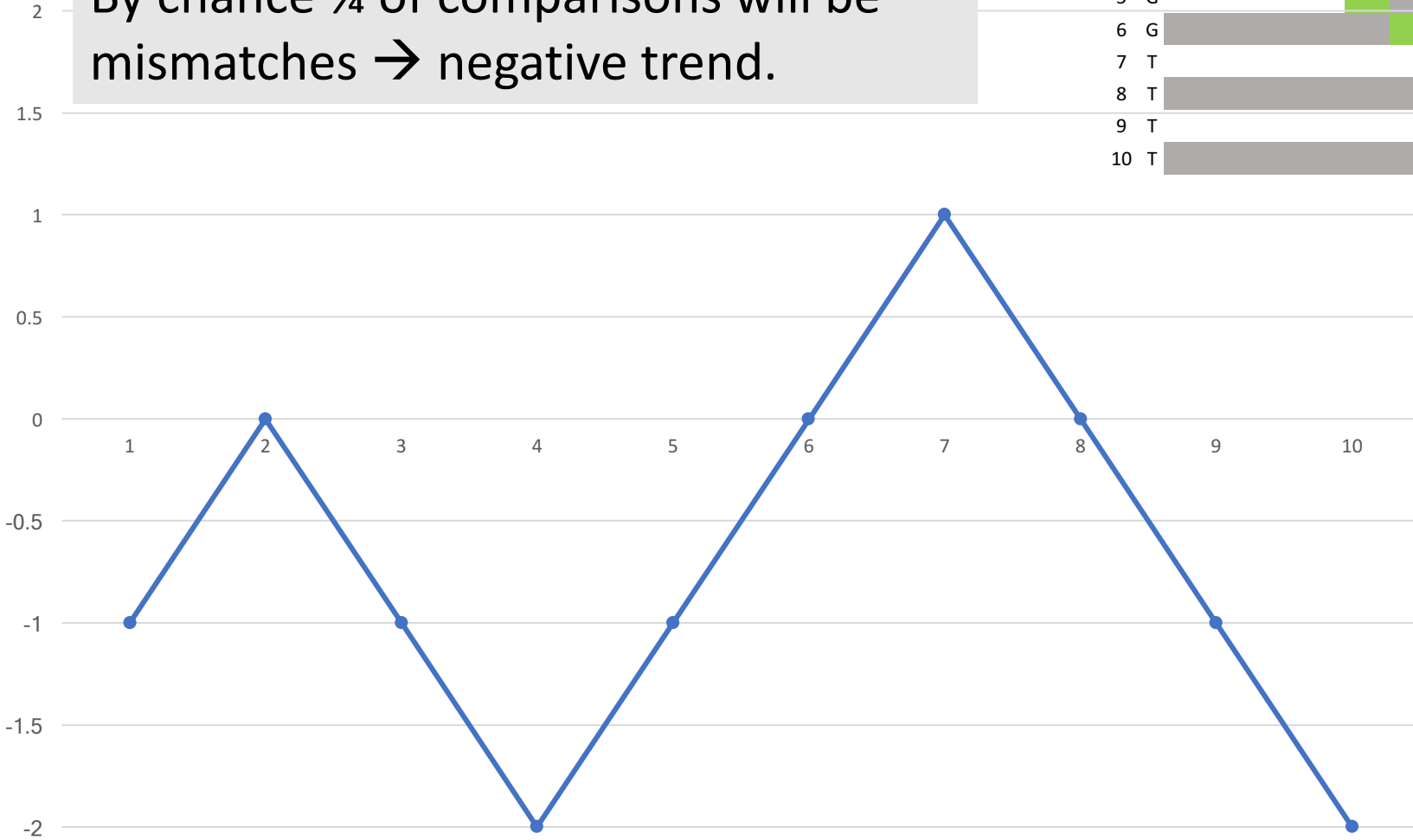
$$S_{ii} = +1$$

$$S_{ij} = -1$$

$$i \neq j$$

- 2) Determine highest segment score → Highest Scoring Pair (HSP)

By chance $\frac{3}{4}$ of comparisons will be mismatches \rightarrow negative trend.



		1	2	3	4	5	6	7	8	9	10	11
		C	C	A	A	G	G	T	C	A	G	T
1	A											
2	C											
3	C											
4	G											
5	G											
6	G											
7	T											
8	T											
9	T											
10	T											

BLAST algorithm

- 1) Scoring matrix with:

$$S_{ii} = +1$$

$$S_{ij} = -1$$

$$i \neq j$$

- 2) Determine highest segment score → Highest Scoring Pair (HSP)

Without going into much detail: we can calculate our confidence in the HSP by comparing the best **Score** to the background.



		1	2	3	4	5	6	7	8	9	10	11
		C	C	A	A	G	G	T	C	A	G	T
1	A											
2	C											
3	C											
4	G											
5	G											
6	G											
7	T											
8	T											
9	T											
10	T											

BLAST algorithm

- 1) Scoring matrix
with:

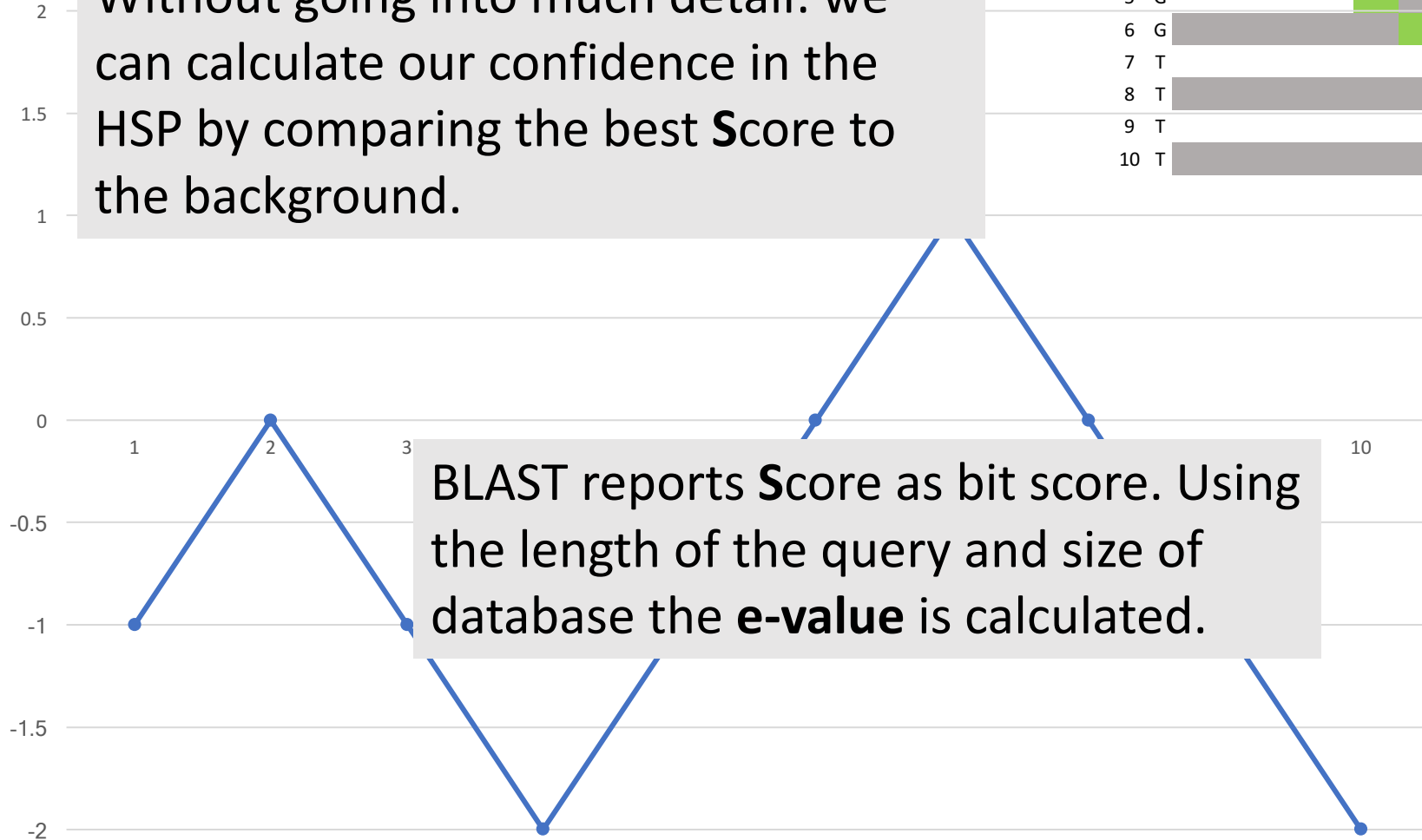
$$S_{ii} = +1$$

$$S_{ij} = -1$$

$$i \neq j$$

- 2) Determine highest segment score → Highest Scoring Pair (HSP)

Without going into much detail: we can calculate our confidence in the HSP by comparing the best **Score** to the background.



BLAST reports **Score** as bit score. Using the length of the query and size of database the **e-value** is calculated.

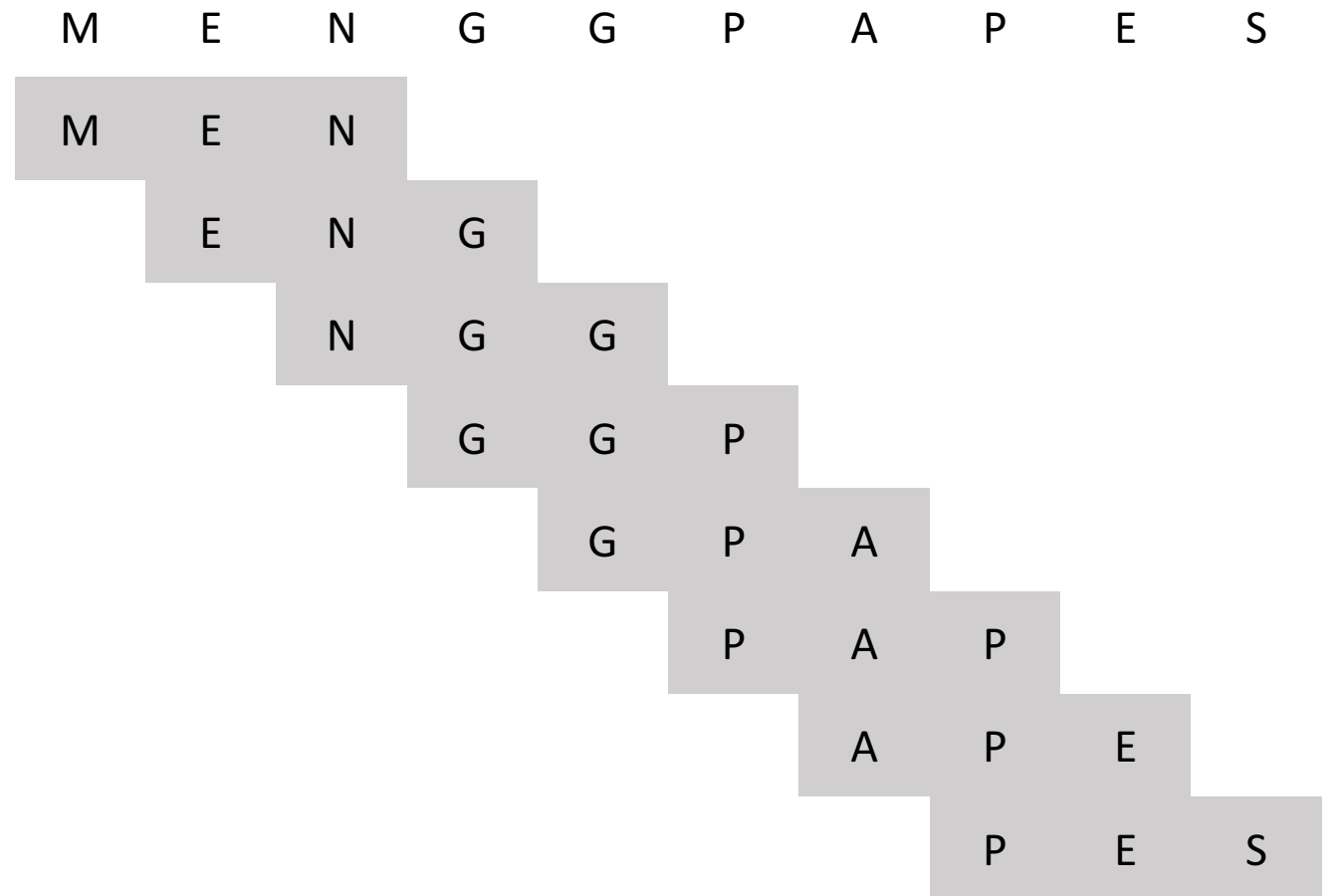
[illegible]

BLAST algorithm

M E N G G P A P E S

BLAST algorithm

- 1) Create all **Words** of length 3 (kmers)



BLAST algorithm

- 1) Create all **W**ords of length 3 (kmers)
- 2) Find exact matches for all **W**

I P A G G P A P E S

M E N G G P A P E S

M E N

ENIG

N G G

G G P

G P A

P A P

A P E

P E S

BLAST algorithm

- 1) Create all **W**ords of length 3 (kmers)
- 2) Find exact matches for all **W**
- 3) Score all hits

[illegible]

BLAST algorithm

- 1) Create all **Words** of length 3 (kmers)
- 2) Find exact matches for all **W**
- 3) Score all hits
- 4) Keep hits above **Threshold (19)**

[illegible]

BLAST algorithm

- 1) Create all **W**ords of length 3 (kmers)
- 2) Find exact matches for all **W**
- 3) Score all hits
- 4) Keep hits above **Threshold (19)**
- 5) Extend alignment until **T** drops

I	P	A	G	G	P	A	P	E	S
1	-2	-2	6	6	7	4	7	5	4
M	E	N	G	G	P	A	P	E	S



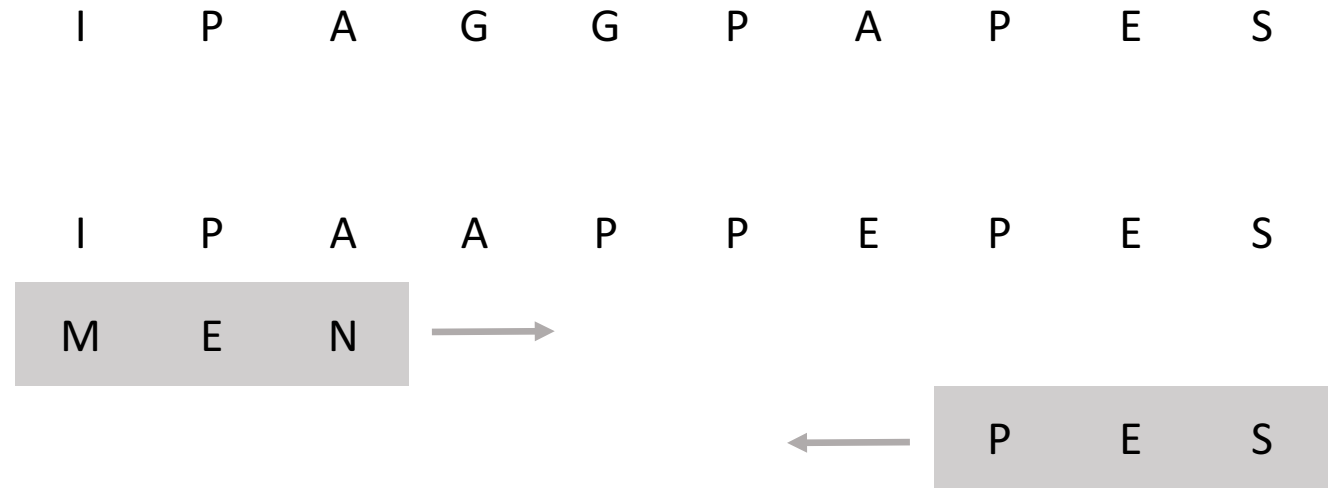
BLAST algorithm

Current versions produce
gapped alignments

Multiple seeds of short
kmer matches are
kept during initial
search

These matches are
extended

Final alignment reported
is a Smith-Waterman
alignment



BLAST

BLOSUM62 Scoring Matrix

- **BLO**ck **SU**bstitution **M**atrix
- By Henikoff and Henikoff (1992)
- Default scoring matrix for pairwise alignment of sequences using BLAST
- Based on empirical observations of distantly-related proteins

A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-1	1	1	-2	-1	-3	-2	5								
M	-1	-2	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

BLAST

Query	NLYENFVQATFNALTAEKV
	NY ENF+Q+ + L +
Subject	NYAENTIQSIISTVEPAQR

Centerline provides the following information:

- Letter designates identity of high similarity
- Plus indicates similarity but not very similar
- No symbol equals low similarity

BLAST

Nucleotide scoring matrices for ungapped (left) and gapped alignments (right)

	A	T	C	G
A	5			
T	-4	5		
C	-4	-4	5	
G	-4	-4	-4	5

	A	T	C	G
A	1			
T	-3	1		
C	-3	-3	1	
G	-3	-3	-3	1

Gap opening score: -11
Gap extension: -1

BLAST statistics

- Score (Bits)
 - a conversion of of summed substitution scores
- Expect(e) value
 - Function of the score and database size
 - 1 alignment using a query of this size will by chance produce a score of this value in a database of this size
 - e-value is specific to a database of a certain size

BLAST statistics

- Rules of thumb

- e value of $\leq e^{-5}$: often used for annotating genes
- E value of $\leq e^{-30}$: strong evidence of homology
- Length of hit is important when evaluating e values

BLAST programs

Search

blastn

blastx

tblastx

blastp

Query

nucleotide

translated nucleotide in all 6 frames

translated nucleotide in all 6 frames

protein

Database

nucleotide

protein

translated nucleotide in all 6 frames

protein

BLAST programs

Search

blastn
blastx
tblastx
blastp

Query

nucleotide
translated nucleotide in all 6 frames
translated nucleotide in all 6 frames
protein

Database

nucleotide
protein
translated nucleotide in all 6 frames
protein

		Second letter				
		U	C	A	G	
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G
	A	AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G