Sequence Homology Searches with BLAST

Julin Maloof

Some Slides courtesy of Venkatsean Sundaresan

The Scenario

- Let’s role back the clock to December, 2019.

The Scenario

- Let’s role back the clock to December, 2019.
- A strange new respiratory illness is rapidly spreading.
- Disease etiology suggests that it is caused by a virus.

The Scenario

- Let’s role back the clock to December, 2019.
- A strange new respiratory illness is rapidly spreading.
- Disease etiology suggests that it is caused by a virus.
- What kind of virus?
- Your colleagues purify viruses from an infected patient and assemble 8 viral genome sequences.

The Scenario

- Let’s role back the clock to December, 2019.
- A strange new respiratory illness is rapidly spreading.
- Disease etiology suggests that it is caused by a virus.
- What kind of virus?
- Your colleagues purify viruses from an infected patient and assemble 8 viral genome sequences.
- Your tasks:
  - Determine which of these 8 are the likely cause
  - Determine the evolutionary origin of the new virus

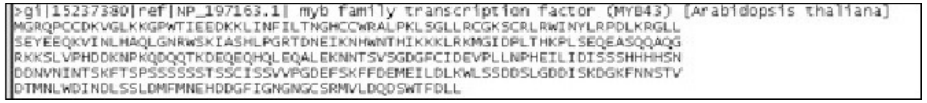
Methods

- Your tasks:
  - Determine which of these 8 are the likely cause
  - Determine the evolutionary origin of the new virus

- How?
  - Search for homologous sequences in a database of sequenced viral genomes
  - Build a phylogenetic tree of related sequences

BLAST (Basic Local Alignment Search Tool)

QUERY sequence(s)

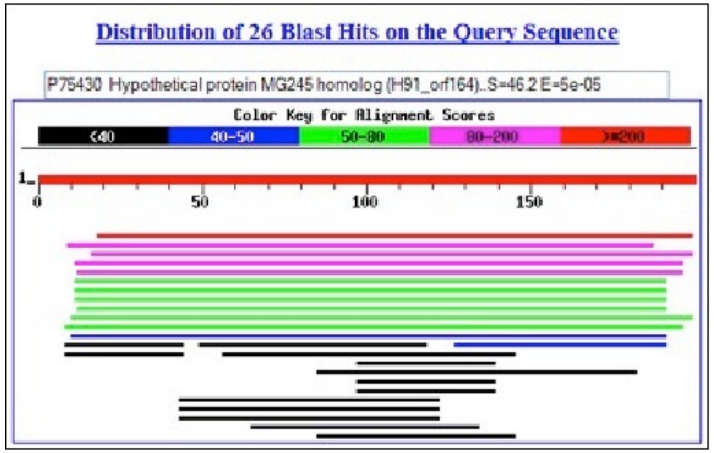


```
>gi|15237380|ref|NP_197163.1| myb family transcription factor (MYB43) [Arabidopsis thaliana]
MGRQPCCDKVGLKKGPWTIEEDKKLINFILTNGHCCWRALPKLSGLLRCGKSCRLRWINYLRPDLKRGLL
SEYEEQKVINLHAQLGNRWSKIASHLPGRTDNEIKNHWNTHIKKKLRKMGIDPLTHKPLSEQEASQQAQG
RKKSLVPHDDKNPKQDQQTKDEQEQHQLEQALEKNNTSVSGDGFCIDEVPLLNPHEILIDISSSHHHHSN
DDNVNINTSKFTSPSSSSSSTSSCISSVVPGDEFSKFFDEMEILDLKWLSSDDSLGDDISKDGKFNNSTV
DTMNLWDINDLSSLDMFMNEHDDGFIGNGNGCSRMVLDQDSWTFDLL
```

BLAST program

BLAST database

BLAST results



Distribution of 26 Blast Hits on the Query Sequence

P75430 Hypothetical protein MG245 homolog (H91_orf164)..S=46.2 E=5e-05

Color Key for Alignment Scores

<40	40-50	50-80	80-200	>=200

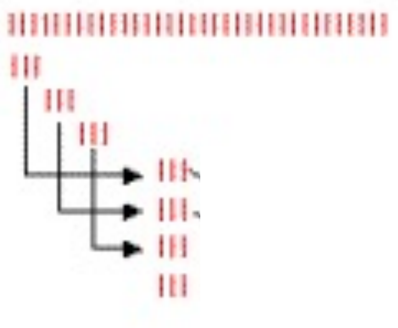
- Search for similarity to infer “homology”

BLAST

- BLAST is optimized to search large databases quickly.
- How does it do this?

BLAST: Heuristic algorithm

Query sequence of length L (this is the sequence with which you do a search)



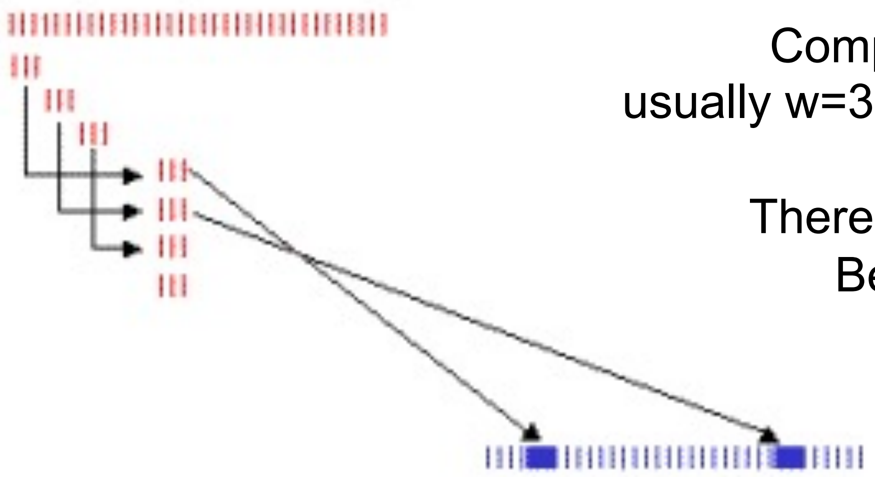
Compile list of words (w) from query
usually w=3 for proteins and 11-28 for nucleotides

There are L-w+1 words in sequence L
Begin with high scoring words

Galisson *EMBER* (2000)

BLAST: Heuristic algorithm

Query sequence of length L (this is the sequence with which you do a search)



Compile list of words (w) from query
usually w=3 for proteins and 11-28 for nucleotides

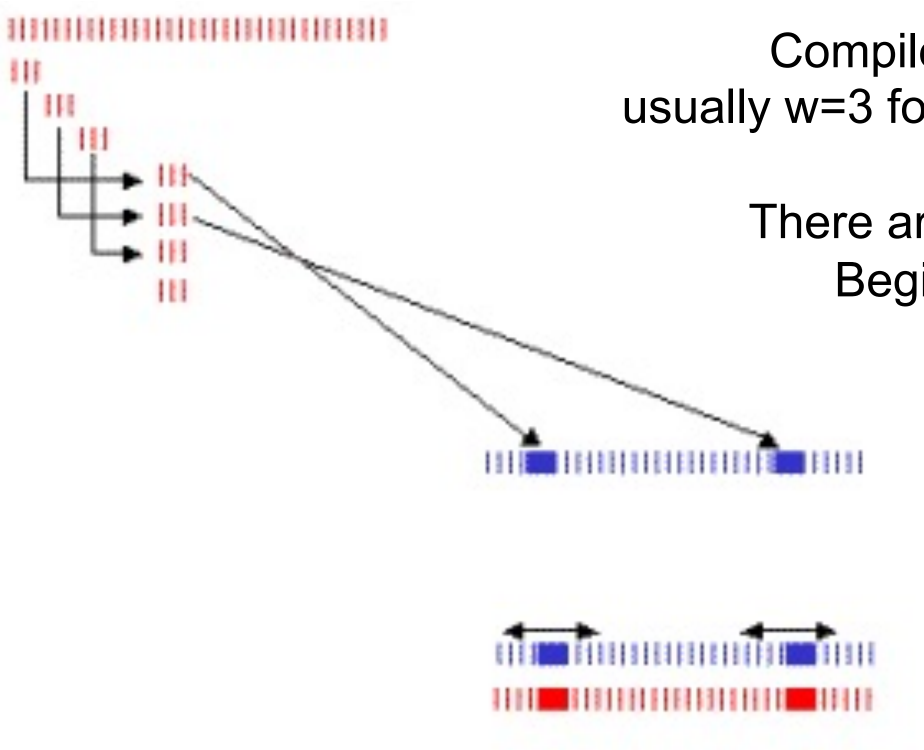
There are L-w+1 words in sequence L
Begin with high scoring words

Compare word list with sequences
in database and identify matches

Galisson *EMBER* (2000)

BLAST: Heuristic algorithm

Query sequence of length L (this is the sequence with which you do a search)



Compile list of words (w) from query
usually w=3 for proteins and 11-28 for nucleotides

There are L-w+1 words in sequence L
Begin with high scoring words

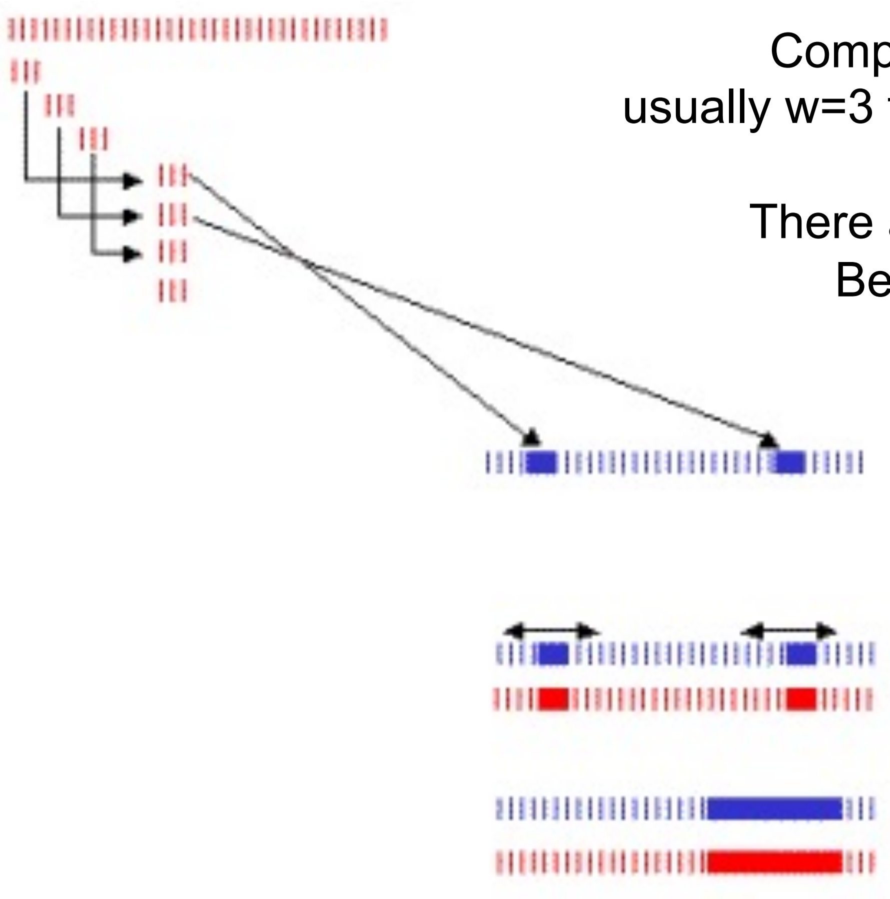
Compare word list with sequences
in database and identify matches

Extend matches in both directions
until further extension causes the
score to drop by a certain amount

Galisson *EMBER* (2000)

BLAST: Heuristic algorithm

Query sequence of length L (this is the sequence with which you do a search)



Compile list of words (w) from query
usually w=3 for proteins and 11-28 for nucleotides

There are L-w+1 words in sequence L
Begin with high scoring words

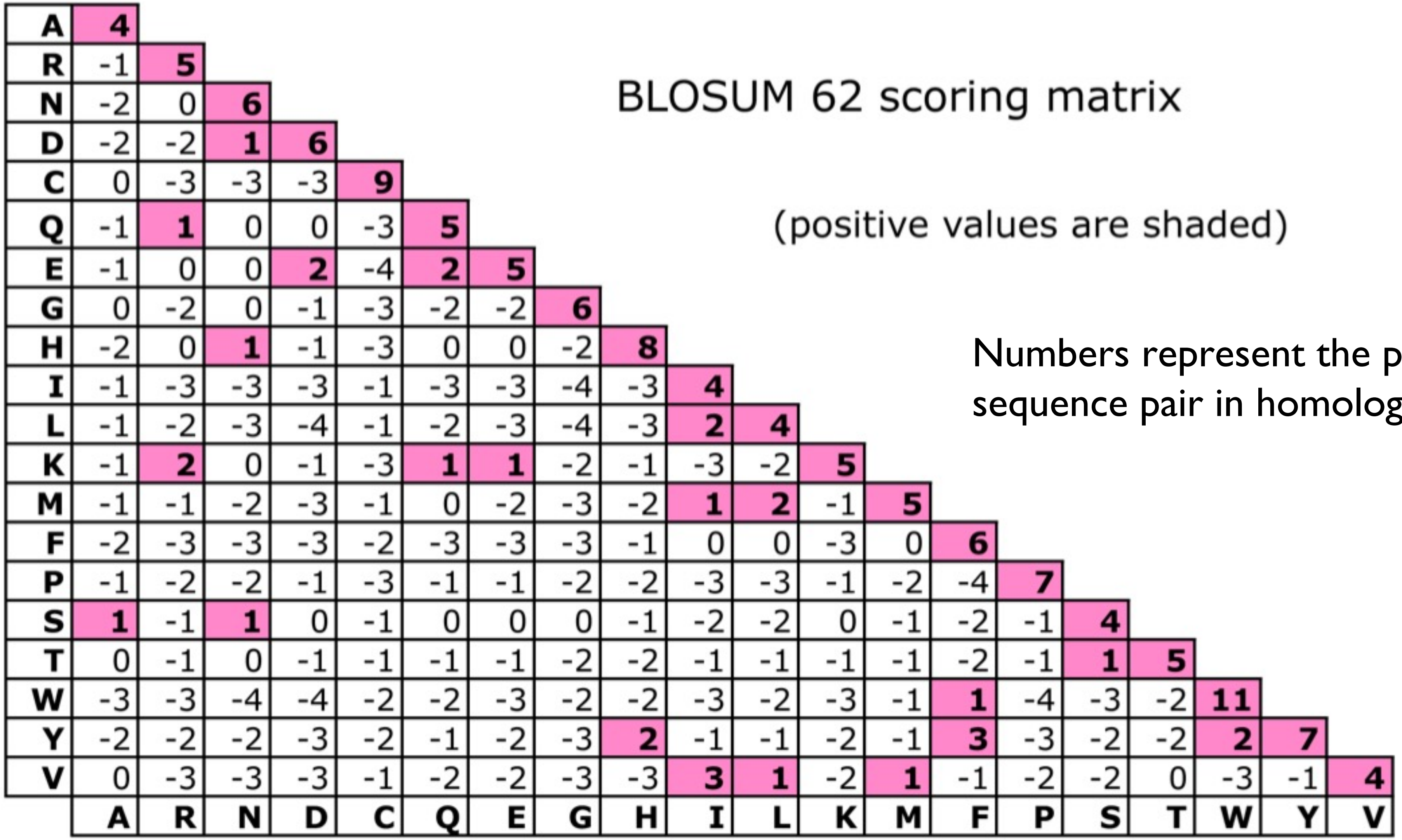
Compare word list with sequences
in database and identify matches

Extend matches in both directions
until further extension causes the
score to drop by a certain amount

High scoring segment pair HSP

Galisson *EMBER* (2000)

A scoring matrix is used to evaluate matches



BLOSUM 62 scoring matrix

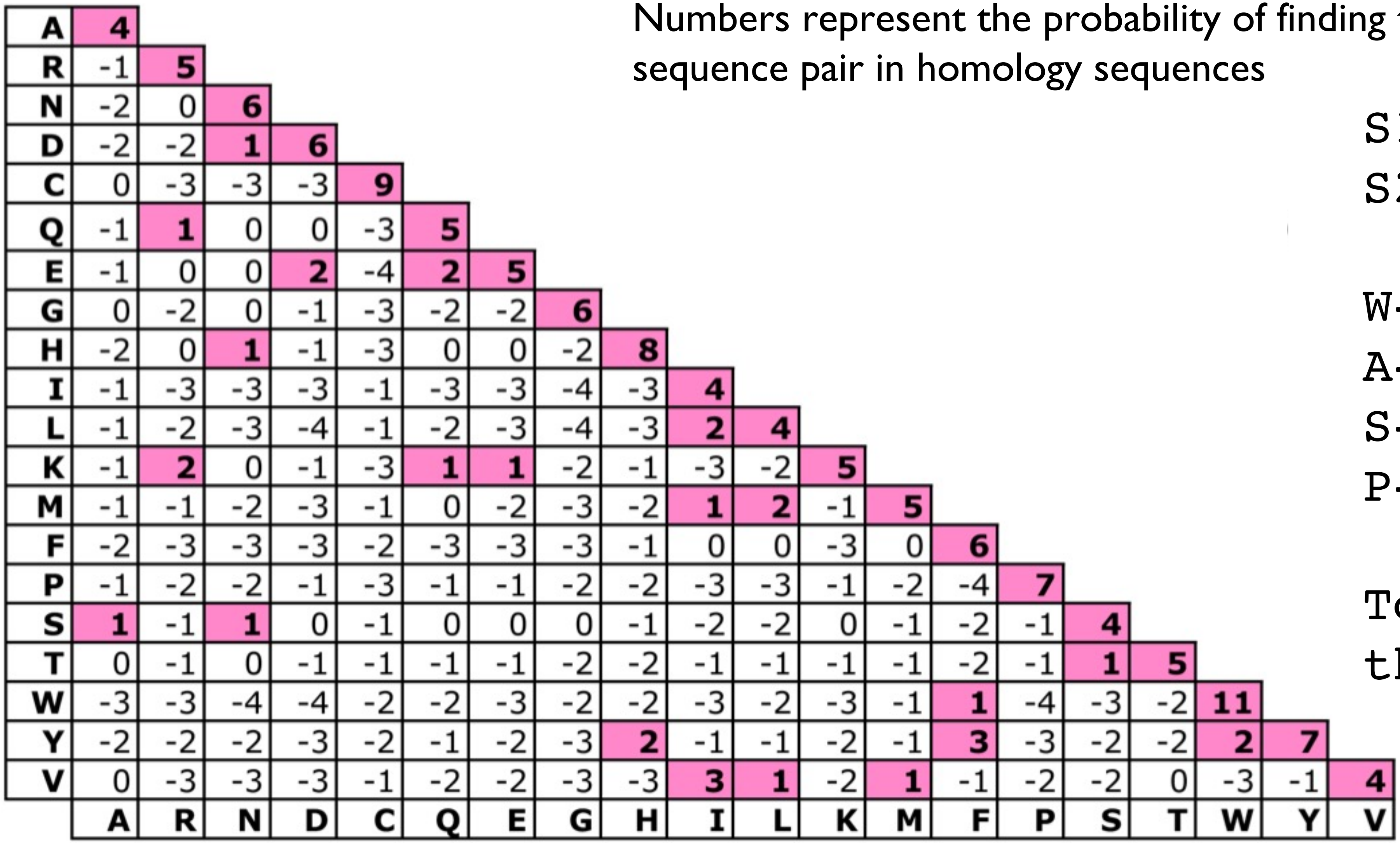
(positive values are shaded)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
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	-3	**1**	**1**	-2	-1	-3	-2	**5**								
M	-1	-1	-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**

Numbers represent the probability of finding that sequence pair in homology sequences

The values for amino acid substitutions were obtained from Henikoff S & Henikoff JG (1992) Amino acid substitutions matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**: 10915-10919.

A scoring matrix is used to evaluate matches



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	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-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**

The values for amino acid substitutions were obtained from Henikoff S & Henikoff JG (1992) Amino acid substitutions matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**: 10915-10919.

Numbers represent the probability of finding that sequence pair in homology sequences

```
S1: W-A-S-P
S2: W-E-S-T

W-W = 11
A-E = -1
S-S =  4
P-T = -1

Total score for
this alignment: 13
```

Q :ROBJOEZACANNLIZ

Break this up into 3 letter words

ROB,OBJ,BJO,...,ZAC,…ANN,…NLI,LIZ

Q :ROBJOEZACANNLIZ

Break this up into 3 letter words

ROB,OBJ,BJO,...,ZAC,…ANN,…NLI,LIZ

Search sequences S1, S2, etc. in database

Find a match with the word ZAC then extend on both sides until no or weak matches

Q :ROBJOEZACANNLIZ

Break this up into 3 letter words

ROB,OBJ,BJO,...,ZAC,…ANN,…NLI,LIZ

Search sequences S1, S2, etc. in database
Find a match with the word ZAC then extend on both sides until no or weak matches

Q :ROBJOEZACANNLIZ
S1:TOMZOEZACANNLIA

Q :ROBJOEZACANNLIZ
S2:TOMZOEZACAMYLEA

```
Q :ROBJOEZACANNLIZ
```

Break this up into 3 letter words

```
ROB,OBJ,BJO,...,ZAC,…ANN,…NLI,LIZ
```

Search sequences S1, S2, etc. in database
Find a match with the word ZAC then extend on both sides until no or weak matches

```
Q :ROBJOEZACANNLIZ
S1:TOMZOEZACANNLIA
```

```
Q :ROBJOEZACANNLIZ
S2:TOMZOEZACAMYLEA
```

Search with high scoring words first for better chance of high scoring alignments

```
Q:LVAAVGVCWDILRAAA
```

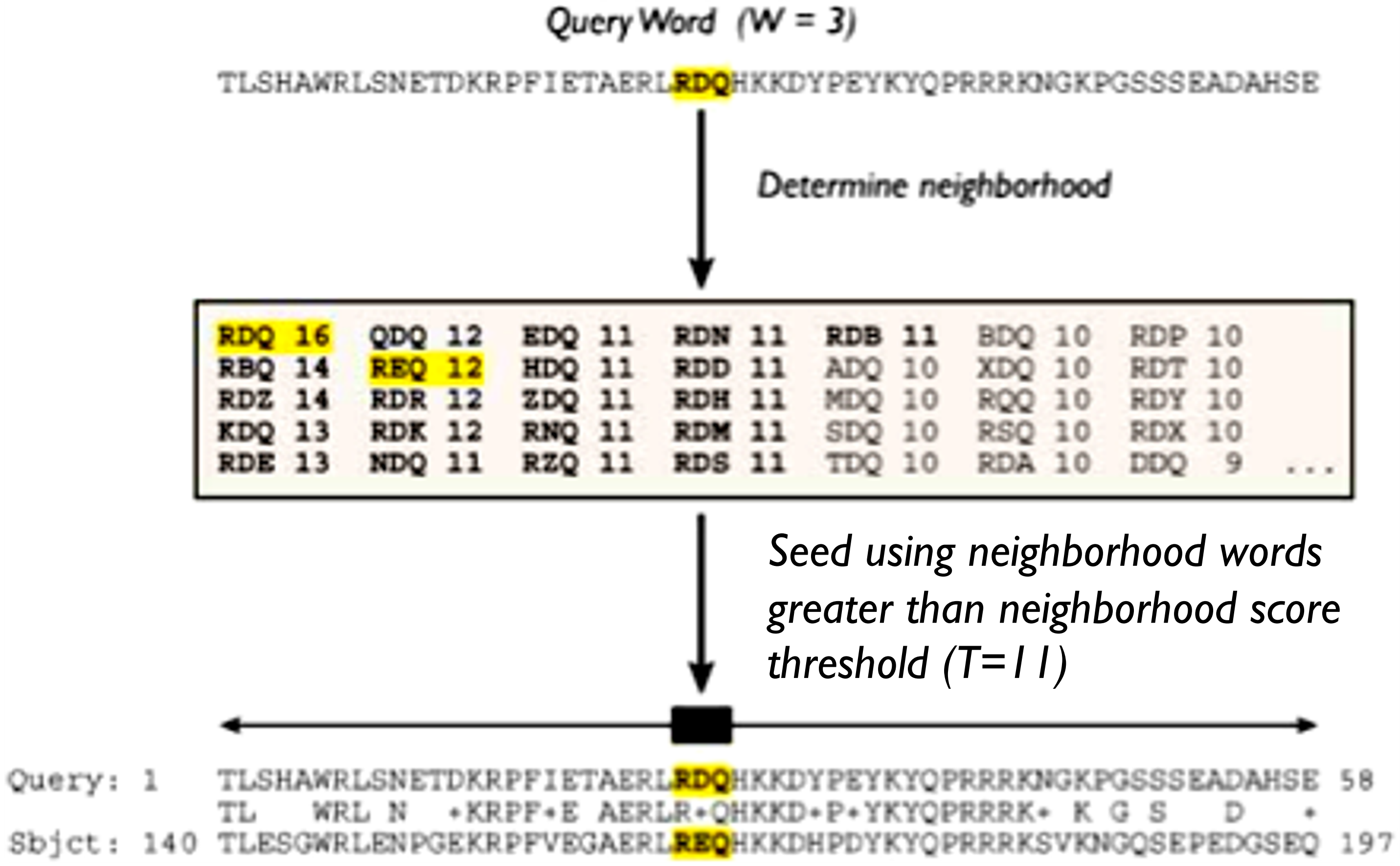
In the above example, BLOSUM62 scores for matches to LVA and CWD are 12 and 26 respectively, so search with CWD

```
Q:LVAAVGVCWDILRAAA
     || |||||| |
S:AGGAVVVCWDILKAGG
```

useful parameters

- Word size: the size of the chunks that the query sequence is chopped into
- Threshold: minimum score for a word match to be considered to seed an extension

How BLAST works



Query Word (W = 3)

```
TLSHAWRLSNETDKRPFIETAERLRDQHKKDYPEYKYQPRRRKNGKPGSSSEADAHSE
```

Determine neighborhood

RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10	
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10	
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10	
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10	
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9	...

Seed using neighborhood words greater than neighborhood score threshold (T=11)

```
Query: 1   TLSHAWRLSNETDKRPFIETAERLRDQHKKDYPEYKYQPRRRKNGKPGSSSEADAHSE 58
           TL   WRL N  +KRPF+E AERLR+QHKKD+P+YKYQPRRRK+ K G S   D   +
Sbjct: 140 TLESGWRLENPGEKRPFVEGAERLREQHKKDHPDYKYQPRRRKSVKNGQSEPEDGSEQ 197
```

HSP = High-scoring Segment Pair – a segment pair whose score will not increase by further extension or by trimming

Score (S) = measures alignment quality (scoring matrix - gaps)

E value (E) = number of different alignments with score S that are expected to occur by chance in a search of that database

Nucleotide vs Protein BLAST

- blastn: nucleotide blast. Comes in different flavors
  - megablast: optimized for nearly identical sequences
  - dc-megablast: discontinuous megablast…more distant sequences

Nucleotide vs Protein BLAST

- blastn: nucleotide blast. Comes in different flavors
  - megablast: optimized for nearly identical sequences
  - dc-megablast: discontinuous megablast…more distant sequences
- Seeding:
  - Default word size is 28 (megablast) or 11 (dc-megablast)
  - No threshold for seeding, requires exact match

Nucleotide vs Protein BLAST

- blastn: nucleotide blast. Comes in different flavors
  - megablast: optimized for nearly identical sequences
  - dc-megablast: discontinuous megablast…more distant sequences
- Seeding:
  - Default word size is 28 (megablast) or 11 (dc-megablast)
  - No threshold for seeding, requires exact match
- Scoring matrix
  - Exact match: +1 (megablast); +2 (dc-megablast)
  - Any mismatch: -2 (megablast); -3 (dc-megablast)

BLAST Summary

- Computes regions of high “similarity” in local alignments of 2 sequences
- Break search into “chunks” by finding all subsequences (stretches of similarity, or “words”) of length k that occur in both seqs
- Build score on matches (scoring matrix, gap cost)
- Extend subsequences to see if score increases
- Compute total score (when no more extensions are possible)
- Then compare BLAST score against precomputed expected scores for all sequences in database
- Then rank score

Command Line BLAST

- You are probably familiar with the web interface for BLAST
- We will use a command-line version of the program
- Why would one want to do this?
  - Overcome web version limitations on query size
    - E.g. BLAST one genome against another
  - Can use custom database
  - Easier to test the effect of changing parameters
  - Torture