Pairwise sequence alignments & BLAST

Goals

Goals

- Basics of alignments
- DotPlots
- Scoring alignments
- Global vs local alignments
- Approximate alignment searches
- E values

The point of sequence alignment

- If you have two or more sequences, you may want to know
  - How similar are they? (A quantitative measure)
  - Which residues correspond to each other?
  - Is there a pattern to the conservation/variability of the sequences?
  - What are the evolutionary relationships of these sequences?

```
Human: ccatcctcagatccgtcttcagaaccaccttcccctcgatccacggctccattttcatcc
       |||||||||||| |||||||||||| || |||||   ||| ||| |||| ||| | || |||
Mouse: ccatcctcagacccgtcttcagagcccccttc---tcggtccccggccccactgtcttcc
```

```
String A =   a b c d e
String B =   a c d e f

A (good) alignment would be:

String A =   a b c d e —
             |   | | |
String B =   a - c d e f
```

Many alignments are possible, we want to find the best

```
g c t g a a c g
g t a t a a t c
```

Bad:

```
g c t g a a c g - - - - - - -
- - - - - - - c t a t a a t c
```

Many alignments are possible, we want to find the best

```
g c t g a a c g
c t a t a a t c

Better?

g c t g a a - - - c g
  | |   |         |
- c t - a t a a t c
```

Many alignments are possible, we want to find the best

```
g c t g a a c g
c t a t a a t c

Better?

g c t g - a a - c g
  | |     | |   |
- c t a t a a t c
```

To decide which alignment is best we need

- A way to examine all possible alignments
- A way to compute a score that gives the quality of the alignment

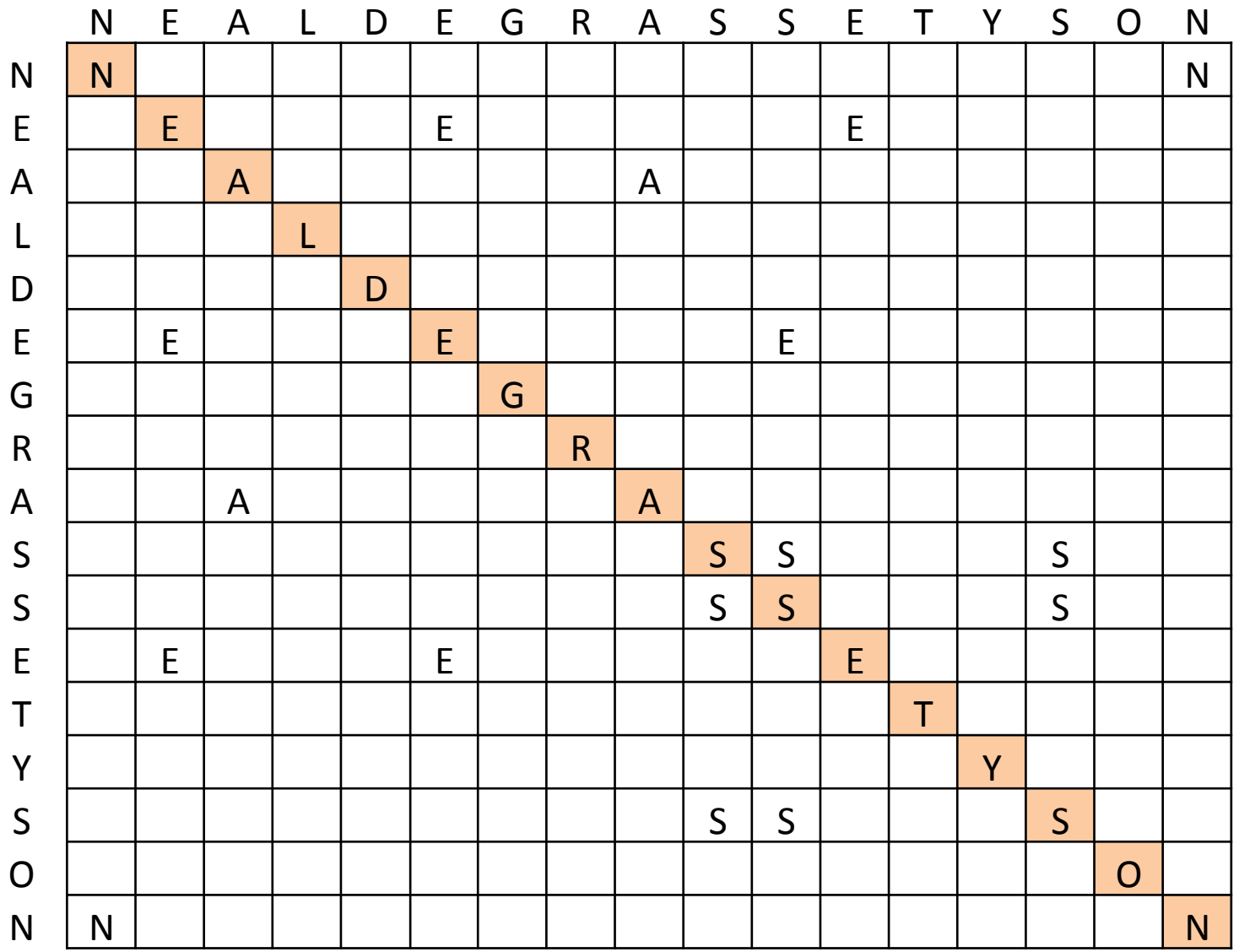
DotPlot - An easy to build representation of the relationship between two sequences.

Option 1 – use the same sequences.

	N	E	A	L	D	E	G	R	A	S	S	E	T	Y	S	O	N
N	N																N
E		E				E						E					
A			A						A								
L				L													
D					D												
E		E				E					E						
G							G										
R								R									
A			A						A								
S										S	S				S		
S										S	S				S		
E		E				E						E					
T													T				
Y														Y			
S										S	S				S		
O																O	
N	N																N

DotPlot - An easy to build representation of the relationship between two sequences.

Option 1 – use the same sequences.



Comparing the same sequence is good for finding repeats

N E A L N E A L N E A L

vs.

N E A L N E A L N E A L

Comparing the same sequence is good for finding repeats

	N	E	A	L	N	E	A	L	N	E	A	L
N	N				N				N			
E		E				E				E		
A			A				A				A	
L				L				L				L
N	N				N				N			
E		E				E				E		
A			A				A				A	
L				L				L				L
N	N				N				N			
E		E				E				E		
A			A				A				A	
L				L				L				L

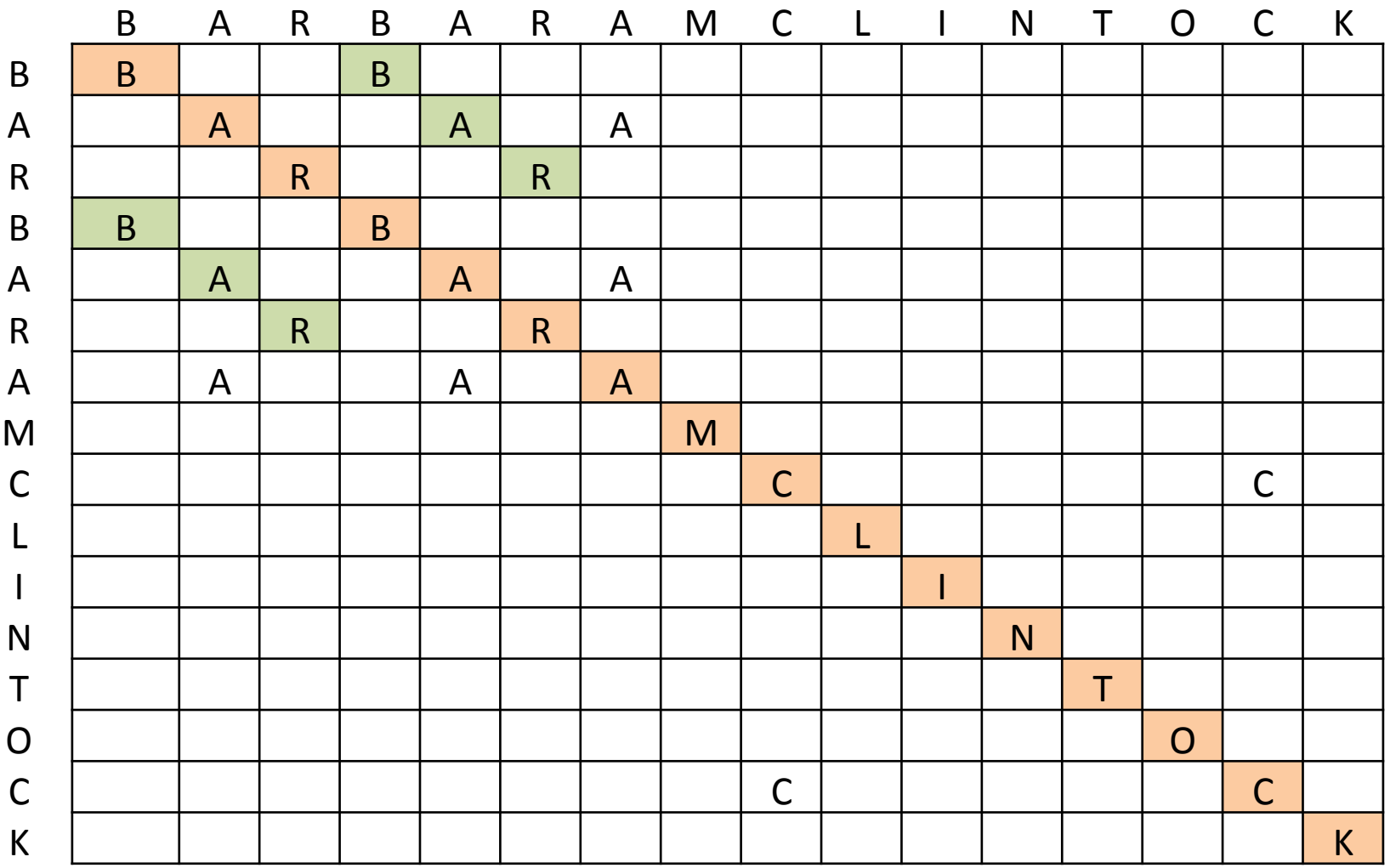
A smaller repeat

B A R B A R A M C L I N T O C K

vs.

B A R B A R A M C L I N T O C K

A smaller repeat



Dotplots are also useful when the sequences aren't the same

Deletion

N E A L D E G R A S S E T Y S O N

vs.

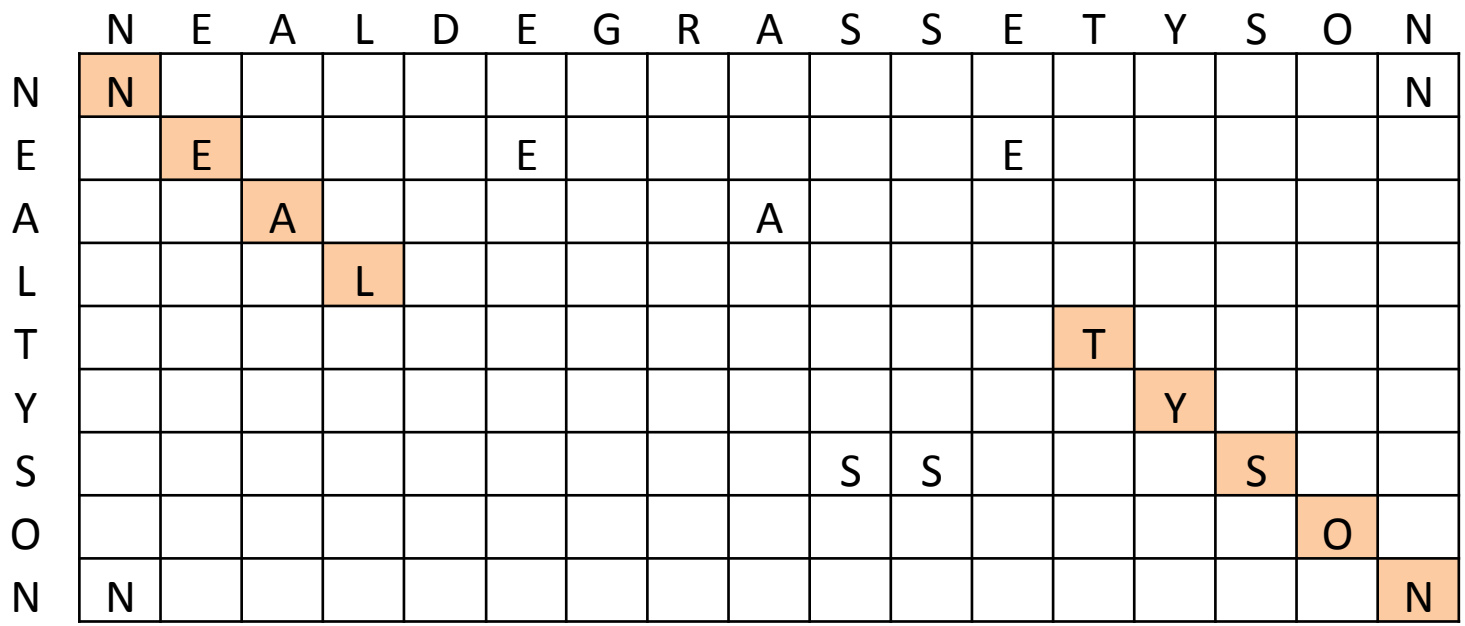
N E A L T Y S O N

Dotplots are also useful when the sequences aren't the same

	N	E	A	L	D	E	G	R	A	S	S	E	T	Y	S	O	N
N	N																N
E		E				E						E					
A			A						A								
L				L													
T													T				
Y														Y			
S										S	S				S		
O																O	
N	N																N

Dotplots are also useful when the sequences aren't the same

INSERTION/DELETION



Dotplots are also useful when the sequences aren't the same

Inversion

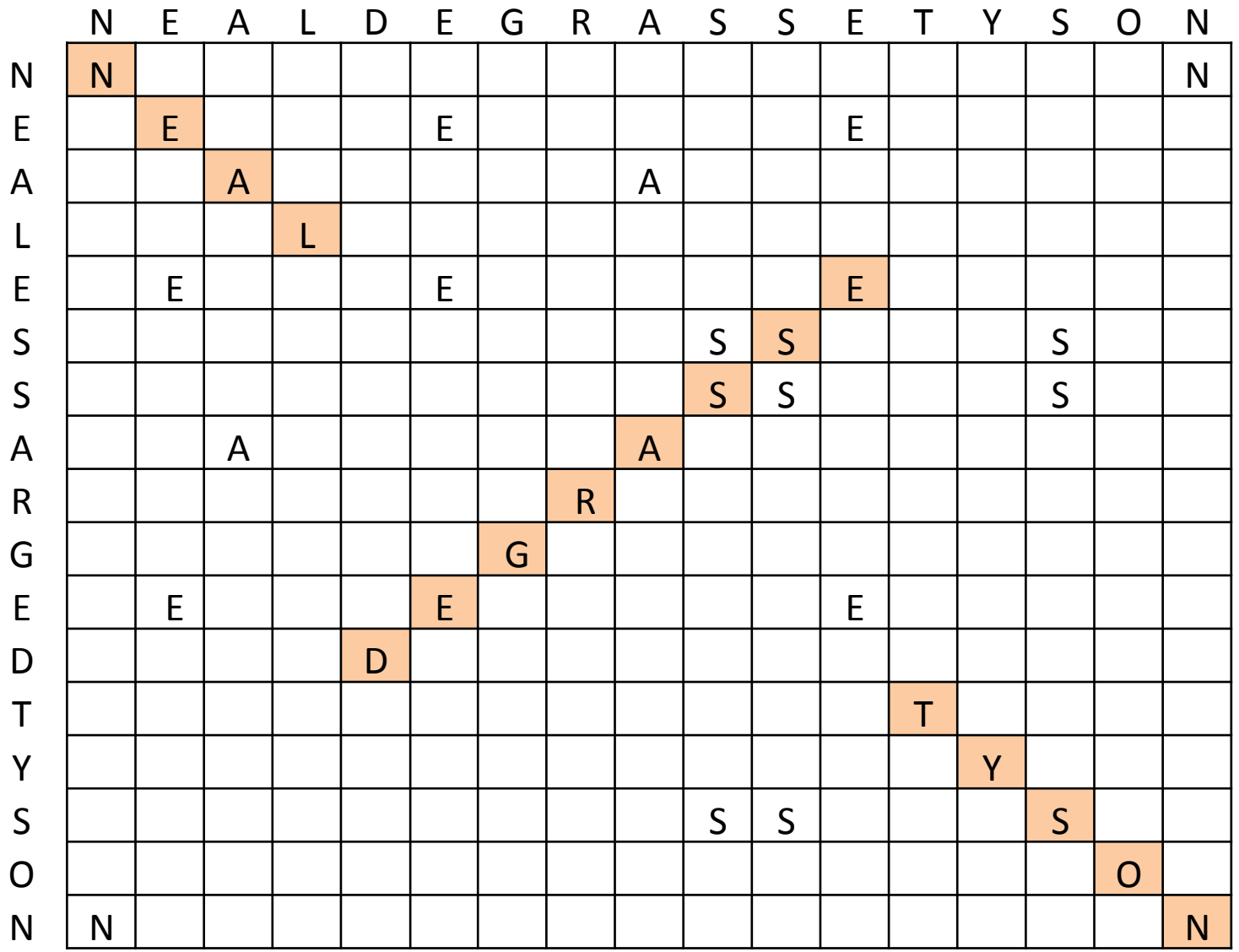
N E A L E S S A R G E D T Y S O N

vs.

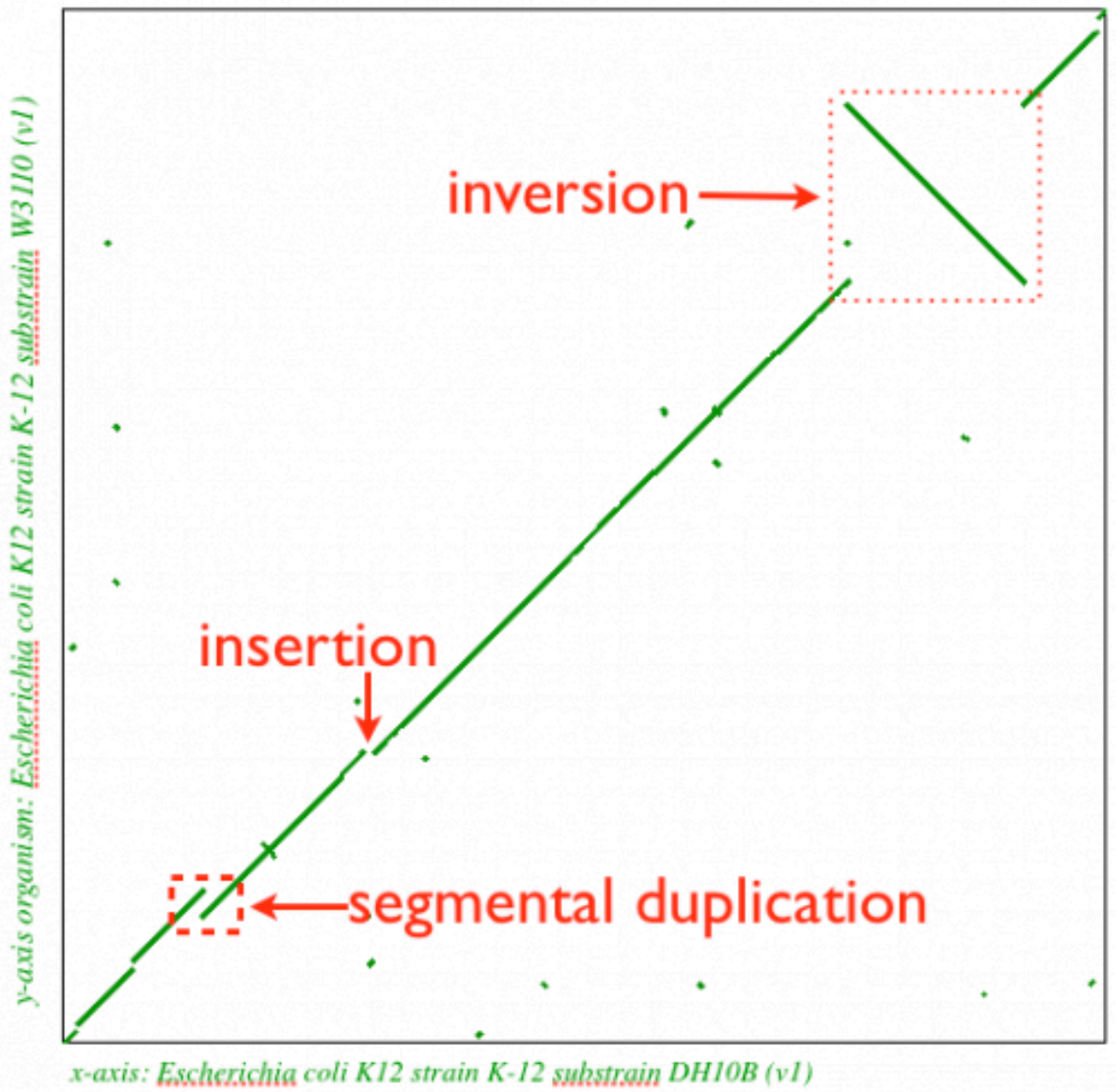
N E A L D E G R A S S E T Y S O N

Dotplots are also useful when the sequences aren't the same

INVERSION



Practical Example



Two substrains of E. coli showing various patterns of genome evolution.

Can be used for whole genomes or genes

(Provided by SynMap software: https://genomevolution.org/CoGe/SynMap.pl)

How is a dotplot relevant to sequence alignment?

- A dotplot explains the main algorithm for comparing sequences
- By constructing this plot, a score can be built for every path
- The path with the best score is the best alignment

Scoring sequence similarity

- A simple scheme
  - +1 for a match
  - -1 for a mismatch

```
String A =      a b c d e
                |   | | |
String B =      a c c d e
```

```
+ 4
- 1

Total Score: 3
```

Scoring based on Biology

- Nucleotides are not mutated randomly
- Transition mutations are more common
  - Purine (A/G) to purine (A/G)
  - Pyrimidine (C/T) to pyrimidine (C/T)
- Transversion mutations are less common
- Can build a scoring scheme to reflect this:
  - Residue is the same = +1
  - Residue undergoes transition = 0
  - Residue undergoes transversion = -1

Scoring Based on Biology

- Amino Acids are not mutated at random either
- Those of similar physicochemical types are more likely to replace each other
- Instead of guessing what these rates might be, can measure empirically

Scoring Based on Biology

- Margaret Dayhoff (1978)
  - Collected statistics on protein substitution frequencies
  - Built the first set of protein substitution matrices
  - Point accepted mutation (PAM) matrices
  - PAM1
  - 1 for 1% substitutions

PAM1

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Scoring Based on Biology

- PAM1 only works well for 1% substitutions
- The substitutions compound over time and become more likely…
- PAM30, PAM70
- PAMn
    - N = the number of mutations per 100 amino acids

PAM70

```
# Entries for the PAM70 matrix at a scale of ln(2)/2.0.
   A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V   B   J   Z   X   *
A  5  -4  -2  -1  -4  -2  -1   0  -4  -2  -4  -4  -3  -6   0   1   1  -9  -5  -1  -1  -3  -1  -1 -11
R -4   8  -3  -6  -5   0  -5  -6   0  -3  -6   2  -2  -7  -2  -1  -4   0  -7  -5  -4  -5  -2  -1 -11
N -2  -3   6   3  -7  -1   0  -1   1  -3  -5   0  -5  -6  -3   1   0  -6  -3  -5   5  -4  -1  -1 -11
D -1  -6   3   6  -9   0   3  -1  -1  -5  -8  -2  -7 -10  -4  -1  -2 -10  -7  -5   5  -7   2  -1 -11
C -4  -5  -7  -9   9  -9  -9  -6  -5  -4 -10  -9  -9  -8  -5  -1  -5 -11  -2  -4  -8  -7  -9  -1 -11
Q -2   0  -1   0  -9   7   2  -4   2  -5  -3  -1  -2  -9  -1  -3  -3  -8  -8  -4  -1  -3   5  -1 -11
E -1  -5   0   3  -9   2   6  -2  -2  -4  -6  -2  -4  -9  -3  -2  -3 -11  -6  -4   2  -5   5  -1 -11
G  0  -6  -1  -1  -6  -4  -2   6  -6  -6  -7  -5  -6  -7  -3   0  -3 -10  -9  -3  -1  -7  -3  -1 -11
H -4   0   1  -1  -5   2  -2  -6   8  -6  -4  -3  -6  -4  -2  -3  -4  -5  -1  -4   0  -4   1  -1 -11
I -2  -3  -3  -5  -4  -5  -4  -6  -6   7   1  -4   1   0  -5  -4  -1  -9  -4   3  -4   4  -4  -1 -11
L -4  -6  -5  -8 -10  -3  -6  -7  -4   1   6  -5   2  -1  -5  -6  -4  -4  -4   0  -6   5  -4  -1 -11
K -4   2   0  -2  -9  -1  -2  -5  -3  -4  -5   6   0  -9  -4  -2  -1  -7  -7  -6  -1  -5  -2  -1 -11
M -3  -2  -5  -7  -9  -2  -4  -6  -6   1   2   0  10  -2  -5  -3  -2  -8  -7   0  -6   2  -3  -1 -11
F -6  -7  -6 -10  -8  -9  -9  -7  -4   0  -1  -9  -2   8  -7  -4  -6  -2   4  -5  -7  -1  -9  -1 -11
P  0  -2  -3  -4  -5  -1  -3  -3  -2  -5  -5  -4  -5  -7   7   0  -2  -9  -9  -3  -4  -5  -2  -1 -11
S  1  -1   1  -1  -1  -3  -2   0  -3  -4  -6  -2  -3  -4   0   5   2  -3  -5  -3   0  -5  -2  -1 -11
T  1  -4   0  -2  -5  -3  -3  -3  -4  -1  -4  -1  -2  -6  -2   2   6  -8  -4  -1  -1  -3  -3  -1 -11
W -9   0  -6 -10 -11  -8 -11 -10  -5  -9  -4  -7  -8  -2  -9  -3  -8  13  -3 -10  -7  -5 -10  -1 -11
Y -5  -7  -3  -7  -2  -8  -6  -9  -1  -4  -4  -7  -7   4  -9  -5  -4  -3   9  -5  -4  -4  -7  -1 -11
V -1  -5  -5  -5  -4  -4  -4  -3  -4   3   0  -6   0  -5  -3  -3  -1 -10  -5   6  -5   1  -4  -1 -11
B -1  -4   5   5  -8  -1   2  -1   0  -4  -6  -1  -6  -7  -4   0  -1  -7  -4  -5   5  -5   1  -1 -11
J -3  -5  -4  -7  -7  -3  -5  -7  -4   4   5  -5   2  -1  -5  -5  -3  -5  -4   1  -5   5  -4  -1 -11
Z -1  -2  -1   2  -9   5   5  -3   1  -4  -4  -2  -3  -9  -2  -2  -3 -10  -7  -4   1  -4   5  -1 -11
X -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1  -1 -11
* -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11 -11   1
```

PAM70

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	J	Z	X	*
A	5	-4	-2	-1	-4	-2	-1	0	-4	-2	-4	-4	-3	-6	0	1	1	-9	-5	-1	-1	-3	-1	-1	-11
R	-4	8	-3	-6	-5	0	-5	-6	0	-3	-6	2	-2	-7	-2	-1	-4	0	-7	-5	-4	-5	-2	-1	-11
N	-2	-3	6	3	-7	-1	0	-1	1	-3	-5	0	-5	-6	-3	1	0	-6	-3	-5	5	-4	-1	-1	-11
D	-1	-6	3	6	-9	0	3	-1	-1	-5	-8	-2	-7	-10	-4	-1	-2	-10	-7	-5	5	-7	2	-1	-11
C	-4	-5	-7	-9	9	-9	-9	-6	-5	-4	-10	-9	-9	-8	-5	-1	-5	-11	-2	-4	-8	-7	-9	-1	-11
Q	-2	0	-1	0	-9	7	2	-4	2	-5	-3	-1	-2	-9	-1	-3	-3	-8	-8	-4	-1	-3	5	-1	-11
E	-1	-5	0	3	-9	2	6	-2	-2	-4	-6	-2	-4	-9	-3	-2	-3	-11	-6	-4	2	-5	5	-1	-11
G	0	-6	-1	-1	-6	-4	-2	6	-6	-6	-7	-5	-6	-7	-3	0	-3	-10	-9	-3	-1	-7	-3	-1	-11
H	-4	0	1	-1	-5	2	-2	-6	8	-6	-4	-3	-6	-4	-2	-3	-4	-5	-1	-4	0	-4	1	-1	-11
I	-2	-3	-3	-5	-4	-5	-4	-6	-6	7	1	-4	1	0	-5	-4	-1	-9	-4	3	-4	4	-4	-1	-11
L	-4	-6	-5	-8	-10	-3	-6	-7	-4	1	6	-5	2	-1	-5	-6	-4	-4	-4	0	-6	5	-4	-1	-11
K	-4	2	0	-2	-9	-1	-2	-5	-3	-4	-5	6	0	-9	-4	-2	-1	-7	-7	-6	-1	-5	-2	-1	-11
M	-3	-2	-5	-7	-9	-2	-4	-6	-6	1	2	0	10	-2	-5	-3	-2	-8	-7	0	-6	2	-3	-1	-11
F	-6	-7	-6	-10	-8	-9	-9	-7	-4	0	-1	-9	-2	8	-7	-4	-6	-2	4	-5	-7	-1	-9	-1	-11
P	0	-2	-3	-4	-5	-1	-3	-3	-2	-5	-5	-4	-5	-7	7	0	-2	-9	-9	-3	-4	-5	-2	-1	-11
S	1	-1	1	-1	-1	-3	-2	0	-3	-4	-6	-2	-3	-4	0	5	2	-3	-5	-3	0	-5	-2	-1	-11
T	1	-4	0	-2	-5	-3	-3	-3	-4	-1	-4	-1	-2	-6	-2	2	6	-8	-4	-1	-1	-3	-3	-1	-11
W	-9	0	-6	-10	-11	-8	-11	-10	-5	-9	-4	-7	-8	-2	-9	-3	-8	13	-3	-10	-7	-5	-10	-1	-11
Y	-5	-7	-3	-7	-2	-8	-6	-9	-1	-4	-4	-7	-7	4	-9	-5	-4	-3	9	-5	-4	-4	-7	-1	-11
V	-1	-5	-5	-5	-4	-4	-4	-3	-4	3	0	-6	0	-5	-3	-3	-1	-10	-5	6	-5	1	-4	-1	-11
B	-1	-4	5	5	-8	-1	2	-1	0	-4	-6	-1	-6	-7	-4	0	-1	-7	-4	-5	5	-5	1	-1	-11
J	-3	-5	-4	-7	-7	-3	-5	-7	-4	4	5	-5	2	-1	-5	-5	-3	-5	-4	1	-5	5	-4	-1	-11
Z	-1	-2	-1	2	-9	5	5	-3	1	-4	-4	-2	-3	-9	-2	-2	-3	-10	-7	-4	1	-4	5	-1	-11
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-11
*	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	1

BLOSUM

- BLOSUM (BLOck SUbstitution Matrix) - Henikoff and Henikoff
- A new substitution matrix, preferred today
- Much better for more divergent species (constructed using divergent species alignments)
- BLOSUM62 is the matrix used by default in most recent alignment applications such as BLAST.

BLOSUM62

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

Scoring Gaps

- What about gaps?
- Usually, a gap opening is more of a penalty than a gap extension
- Why? A single mutational even may insert more than one base.

- Commonly used is the affine gap penalty:

  Gap opening penalty of 11

  Gap extension penalty of 1 for each additional residue

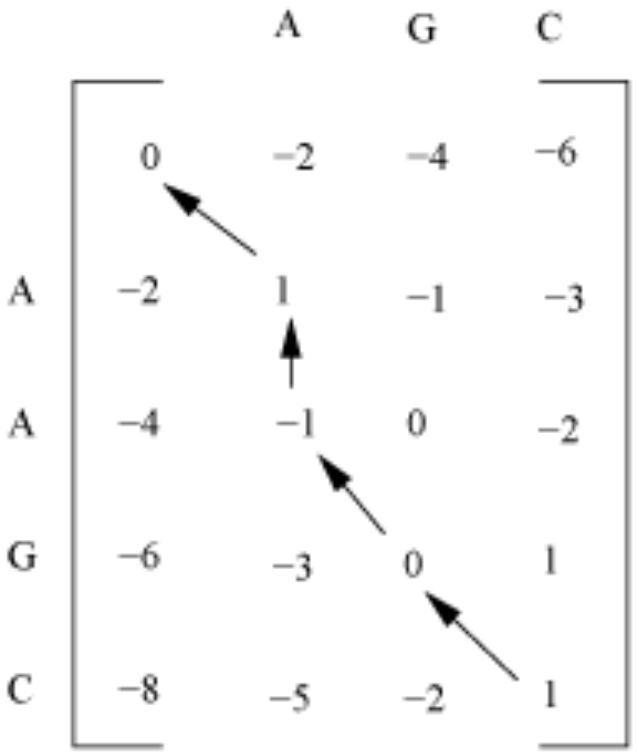
Scoring Wrap Up

- Now we have good a way to score a particular alignment
  1. Score substitutions appropriately reflecting biology
  2. Score gaps appropriately reflecting biology
- But how to generate all the possible alignments?

1970

Needleman and Wunsch

- Dynamic programming algorithm - method by which a larger problem may be solved by first solving smaller, partial versions of the problem
- Basically – uses the dotplot concept, traces all paths, and looks for best scoring path
- Needleman and Wunsch
  - Guaranteed to find the optimal global alignment
  - Many alignments may give the same score
  - Extremely computationally intensive and slow




```
Global alignment:

A A G C
A - G C
```

Global vs Local

- Needleman Wunsch is a global alignment algorithm
- Requires the entirety of both sequences to be examined and scored
- Local alignment became an obvious next needed step
  - Some proteins only share regions of homology
  - Comparing a short sequence (gene) to a very large sequence (genome)

```
Global  FTFTALILLAVAV
        F--TAL-LLA-AV

Local   FTFTALILL-AVAV
        --FTAL-LLAAV--
```

Smith and Waterman

1981

- Proposed a variation of Needleman and Wunsch to create local alignments
- Arbitrary-length segments of each sequence can be aligned
- No penalty for the unaligned portions of the sequences at the ends
- Still fairly time consuming

Rosalind.org

Approximate Methods

- Need more speed!
- Approximate methods have been developed that are
  - Great at detecting close relationships
  - Inferior to exact methods for picking up distant relationships
  - Approximate! (IE no guarantee that the optimal match is found)
- Start with "words"
  - Called k-tuples or k-mers
  - Use these words to quickly find perfect matches
  - Then use the more slow dotplot methods to grow the matches
- BLAST works this way

Heuristic – any that employs a practical methodology not guaranteed to be optimal or perfect, but sufficient for the immediate goals

BLAST

- Basic Local Alignment Search Tool
- Altschul, et al 1990
- Has been cited over 61,000 times
- The most highly cited scientific paper in the entire decade of the 1990s

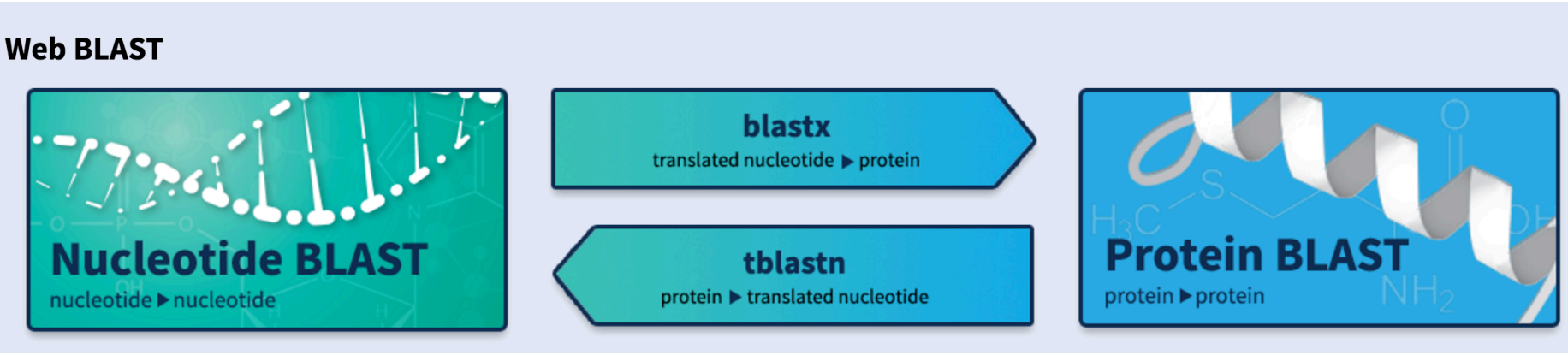
J. Mol. Biol. (1990) 215, 403–410

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

BLAST

- Compares a QUERY sequence to a DATABASE of sequences (also called SUBJECT sequences)
- nucleotide or protein sequences
- Calculates statistical significance
- Available as an online web server , for example, at NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi)



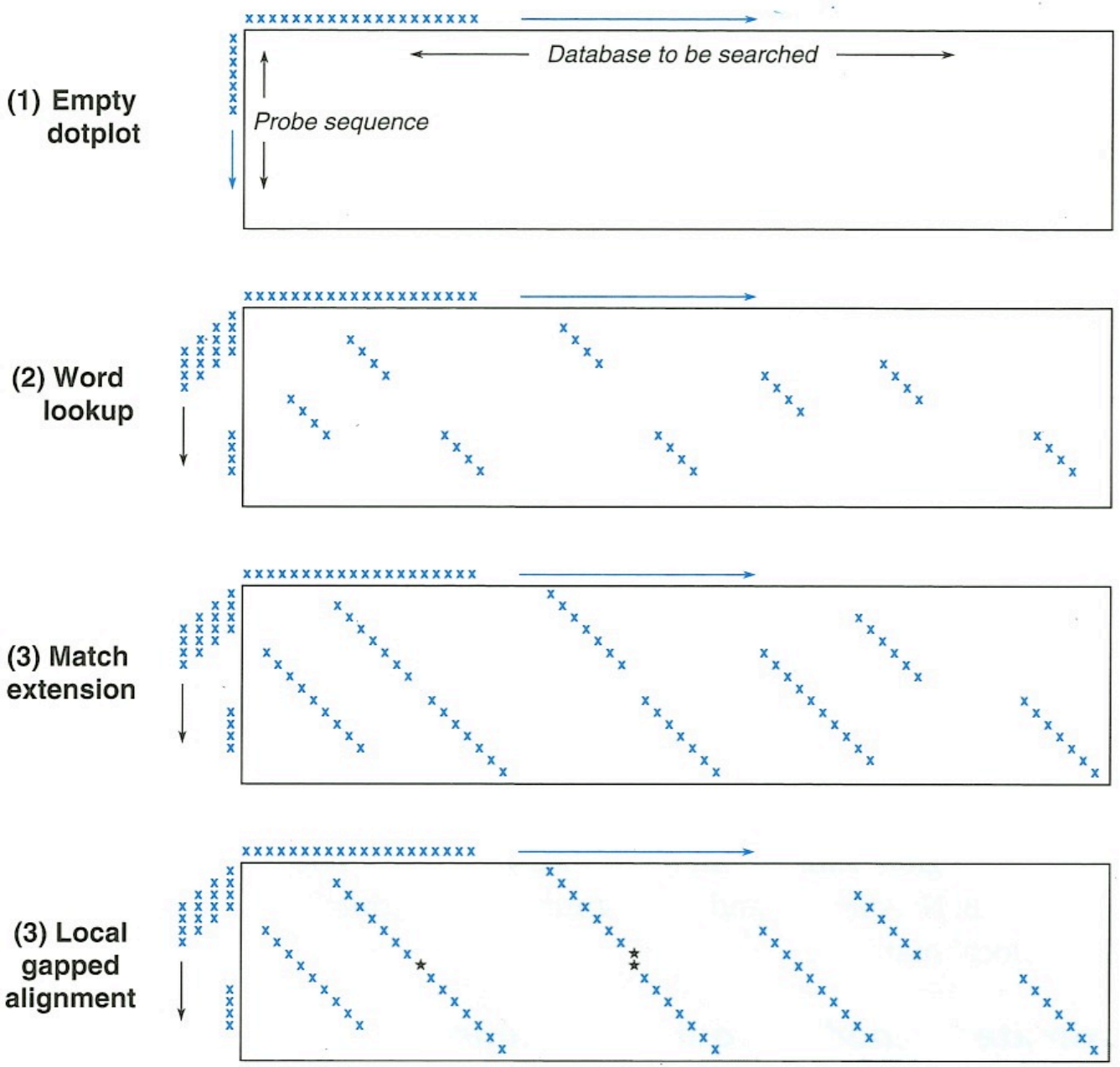
BLAST programs

Program	Query	Database
blastp	protein	protein
blastn	nucleotide	nucleotide
blastx	nucleotide translated to protein	protein
tblastn	protein	nucleotide translated to protein
tblastx	nucleotide translated to protein	nucleotide translated to protein

Why would we want to use translated nucleotides?

BLAST

- Also available as a command line tool (guess which one we'll be using???)

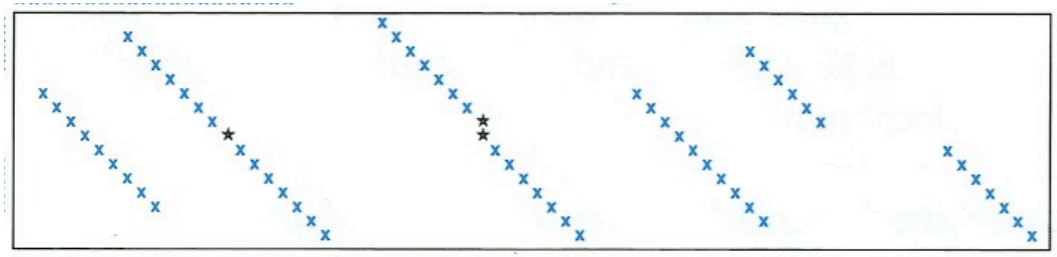


From
Lesk
"Introduction to
Bioinformatics"
3rd Edition

BLAST

High-scoring segment pairs (HSP)

- A query and a match sequence can have more than one HSP



Significance of Alignments

- Now we can find the best scoring alignment (or at least approximately if using BLAST)
- But is it significant in the statistical sense?
  - What is the likelihood that you are observing true biological similarity (evolution) vs random chance?
- E (expect) value = the number of hits one can "expect" to see by chance when searching a database of a particular size
- Lower = more biologically meaningful

E values

E Value	How many random alignments just as good?
1	1 in 1
.2	1 in 5
1e-5	1 in 100,000
1e-9	1 in 1,000,000,000
0	0%

Review

- Dotplots
- Scoring alignments
- Needleman and Wunsch
- Smith and Waterman
- BLAST
- E values