Sequence Analysis (part II)

BBSI 2006: Lecture #(χ+2)

Takis Benos (2006)

Outline

- Sequence variation
- Distance measures
- Scoring matrices
- Pairwise alignments (global, local)
- Database searches (BLAST, FastA)
- Multiple sequence alignments

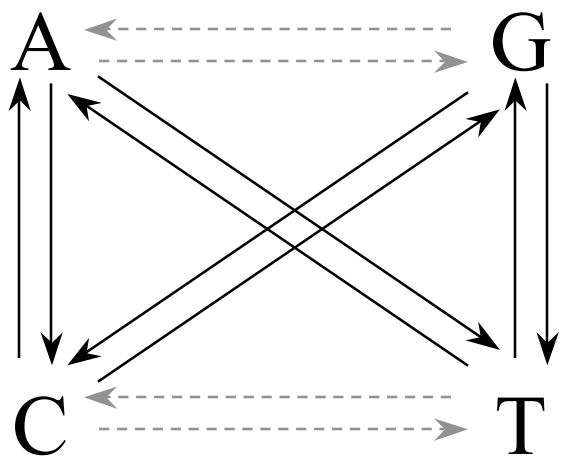
Sequence Variations

Sequence variation

- Base mutations: the source of sequence variation

Purines

Pyrimidines

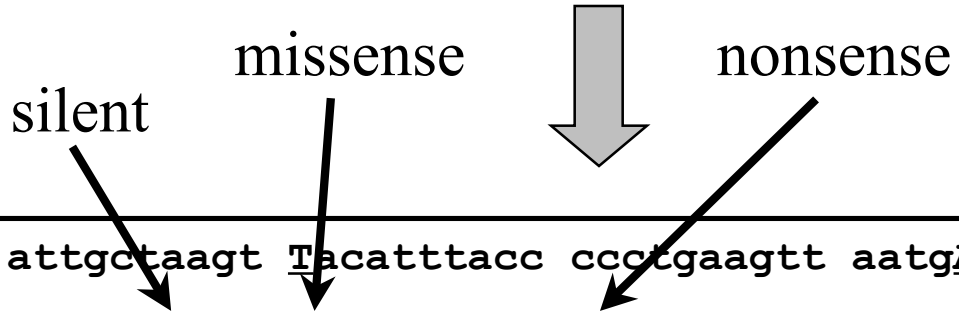


Transitions

Transversions

Sequence variation (cntd)

```
tggagctAtt attgctaagt Aacatttacc ccctgaagtt aatgGatcaa tcaagagaga 120
tgtgggctgt aatgaaTcgt Cttattgaat Taacaggttg gatcgttctt gtcgtttcag 180
                M   N   R   L   I   E   L
tcattcttct tggcgtggcg agtcacattg acaactatca gccacctgaa cagagtgctt 240
cggtacaaca caagtaagct ctgcacttgt ggagcgacat gctgcccgtc cgggtgcatg 300
```




```
tggagctGtt attgctaagt Tacatttacc ccctgaagtt aatgAatcaa tcaagagaga 120
tgtgggctgt aatgaaCcgt Gttattgaat Aaacaggttg gatcgttctt gtcgtttcag 180
                M   N   R   V   I   E
tcattcttct tggcgtggcg agtcacattg acaactatca gccacctgaa cagagtgctt 240
cggtacaaca caagtaagct ctgcacttgt ggagcgacat gctgcccgtc cgggtgcatg 300
```

Sequence variation (cntd)

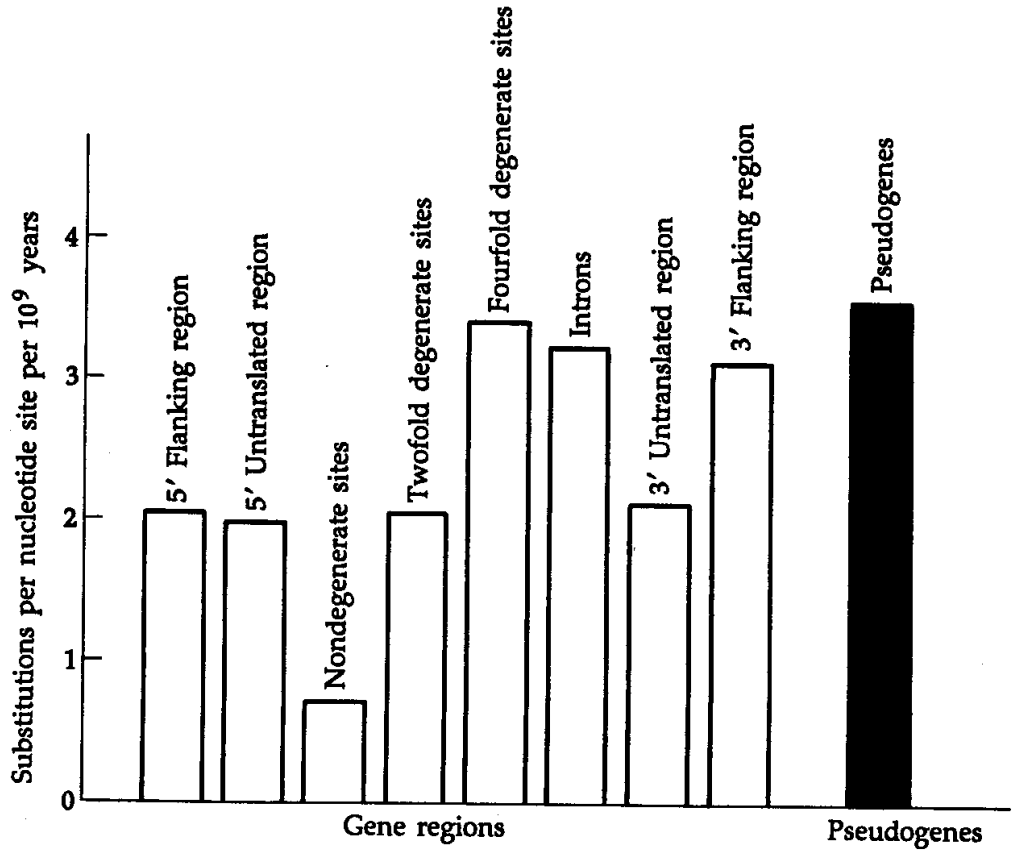


Figure 2. Average rates of substitution in different parts of genes and in pseudogenes.

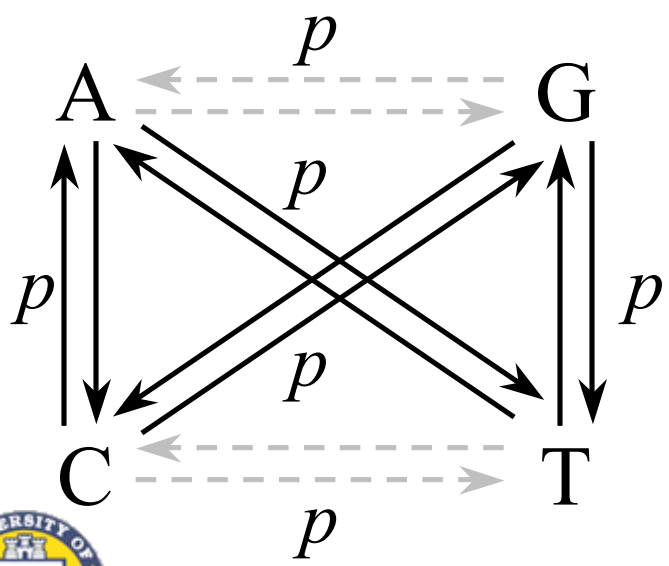
Source: Li & Graur "Fundamentals of Molecular Evolution", 1991, Sinauer Assoc.

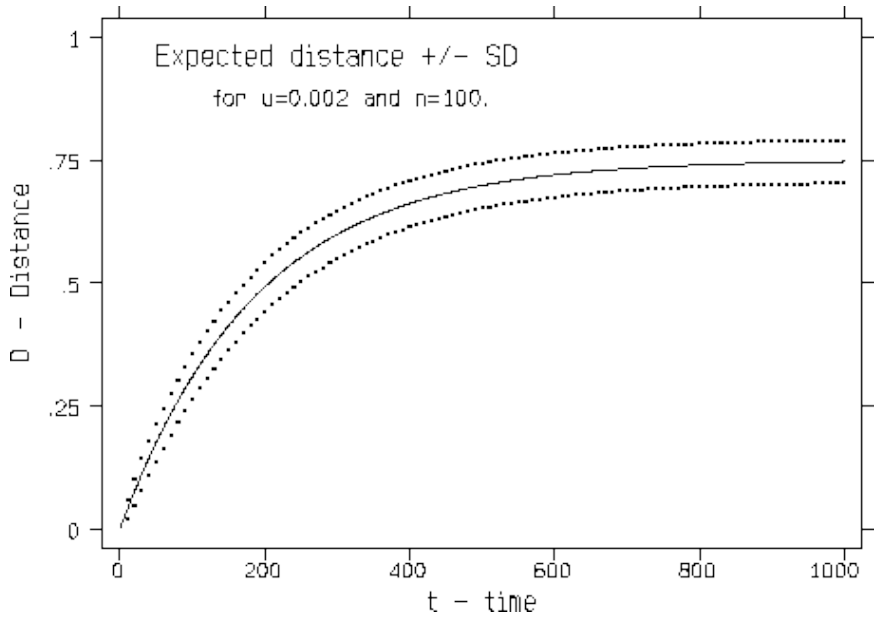
Distance measures

Nucleic acid distances

- No selection - no correction:

$$D = k / N$$



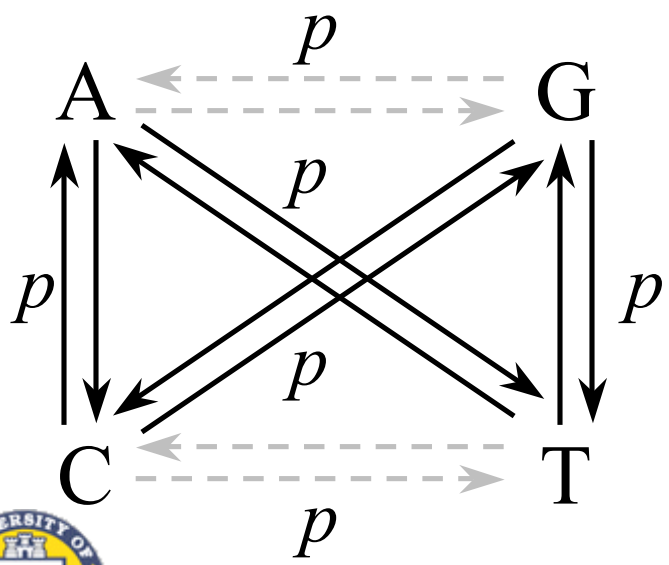


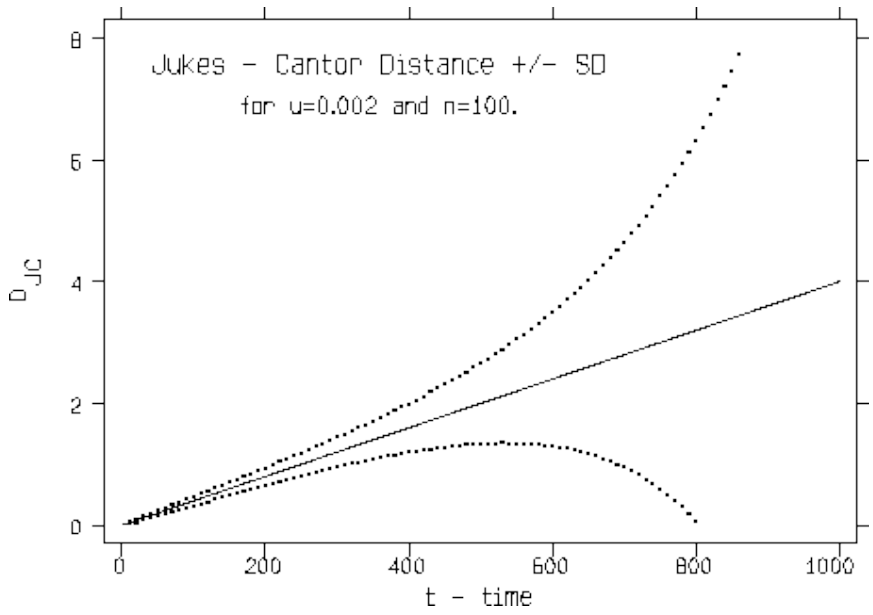
Source: http://helix.biology.mcmaster.ca/721/distance/distance.html.

Nucleic acid distances (cntd)

- Jukes-Cantor correction:

$$D_{JC} = -0.75 \ln (1 - D/0.75)$$



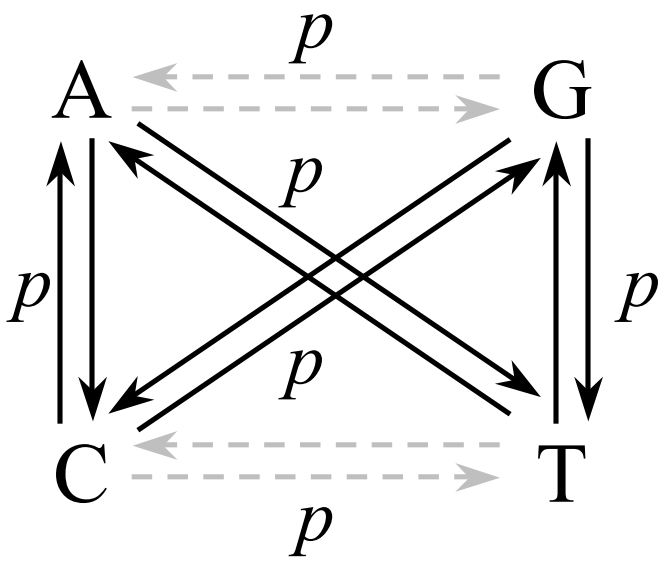


Source: http://helix.biology.mcmaster.ca/721/distance/distance.html.

Nucleic acid distances (cntd)

- Kimura's 2-parameter model:

$$D_{K2P} = -0.5 \ln (1 - 2P - 2Q) - 0.25 \ln(1 - 2Q)$$



Source: http://helix.biology.mcmaster.ca/721/distance/distance.html.

Scoring matrices

Nucleic acid distances (cntd)

- Nucleotide substitution matrices.

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

Identity

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

BLAST

	A	T	C	G
A	0	5	5	1
T	5	0	1	5
C	5	1	0	5
G	1	5	5	0

Transition/ Transversion

Amino acid distances: PAM

- **P**ercent **A**ccepted **M**utations (PAM) matrices:
  - Frequency substitution matrix from aligned sequences (Dayhoff, 1978).
  - M(*i*,*j*): no. of a.a. *i* to *j* mutations
  - 71 groups of closely related proteins *(why?)*; 1,572 changes.
  - PAM*n*: the aligned sequences have *n* a.a. substitutions per 100 residues.

Amino acid distances: PAM (cntd)

- Assumptions of the PAM model:
  - Replacement at any site depends only on the a.a. on that site, given the mutability table.
  - Sequences in the training set (and those compared) have average a.a. composition.

Amino acid distances: PAM (cntd)

Table 1 - The log odds matrix for 250 PAMs (multiplied by 10)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	2	-2	0	0	-4	1	-1	-1	-1	-2	-1	0	1	0	-2	1	1	0	-6	-3
C		12	-5	-5	-4	-3	-3	-2	-5	-6	-5	-4	-3	-5	-4	0	-2	-2	-8	0
D			4	3	-6	1	1	-2	0	-4	-3	2	-1	2	-1	0	0	-2	-7	-4
E				4	-5	0	1	-2	0	-3	-2	1	-1	2	-1	0	0	-2	-7	-4
F					9	-5	-2	1	-5	2	0	-4	-5	-5	-4	-3	-3	-1	0	7
G						5	-2	-3	-2	-4	-3	0	-1	-1	-3	1	0	-1	-7	-5
H							6	-2	0	-2	-2	2	0	3	2	-1	-1	-2	-3	0
I								5	-2	2	2	-2	-2	-2	-2	-1	0	4	-5	-1
K									5	-3	0	1	-1	1	3	0	0	-2	-3	-4
L										6	4	-3	-3	-2	-3	-3	-2	2	-2	-1
M											6	-2	-2	-1	0	-2	-1	2	-4	-2
N												2	-1	1	0	1	0	-2	-4	-2
P													6	0	0	1	0	-1	-6	-5
Q														4	1	-1	-1	-2	-5	-4
R															6	0	-1	-2	2	-4
S																2	1	-1	-2	-3
T																	3	0	-5	-3
V																		4	-6	-2
W																			17	0
Y																				10

$$Score(i,j) = \log_{10} M(i,j)/f(i)$$

Source: http://helix.biology.mcmaster.ca/721/distance/node9.html

Amino acid distances: PAM (cntd)

- Sources of error in the PAM model:
  - Many proteins depart from the average a.a. composition.
  - The a.a. composition can vary even within a protein (e.g., transmembrane proteins).
  - A.a. positions are not "mutated" equally probably; especially in long evolutionary distances.

Amino acid distances: PAM (cntd)

- Sources of error in the PAM model *(cntd)*:
  - Rare replacements are observed too infrequently and…
  - …errors in PAM1 are magnified in PAM250.

A.a. distances: BLOSUM

- **Blo**cks **Su**bstitution **M**atrices (BLOSUM):
  - Log-likelihood matrix (Henikoff & Henikoff, 1992)
  - BLOCKS database of aligned sequences used as primary source set.

A.a. distances: BLOSUM (cntd)

```
AKAGDA---GGCDA
DRALDAFG-GSSDA
GKLGDAI--GSSAF
AKAGGA---GGTAG
CRIGFRC-DGTTDH
AKAKDA--DHSSCI
```

$$Score(i,j) = 2 \log_2 q_{i,j} / e_{i,j}$$

$$e_{i,j} = p_i^2 \qquad for\ i=j$$

$$e_{i,j} = 2\, p_i\, p_j \qquad for\ i \neq j$$

$$p_i = 0.5\ (q_{ii} + \sum q_{ij})$$

A.a. distances: BLOSUM (cntd)

- Weighted contribution of similar(*) sequences in order to reduce redundancy.
- BLOSUM62 is more closely related to PAM120.

(*) *n*% similar; the *n* in BLOSUM*n*

A.a. distances: BLOSUM (cntd)

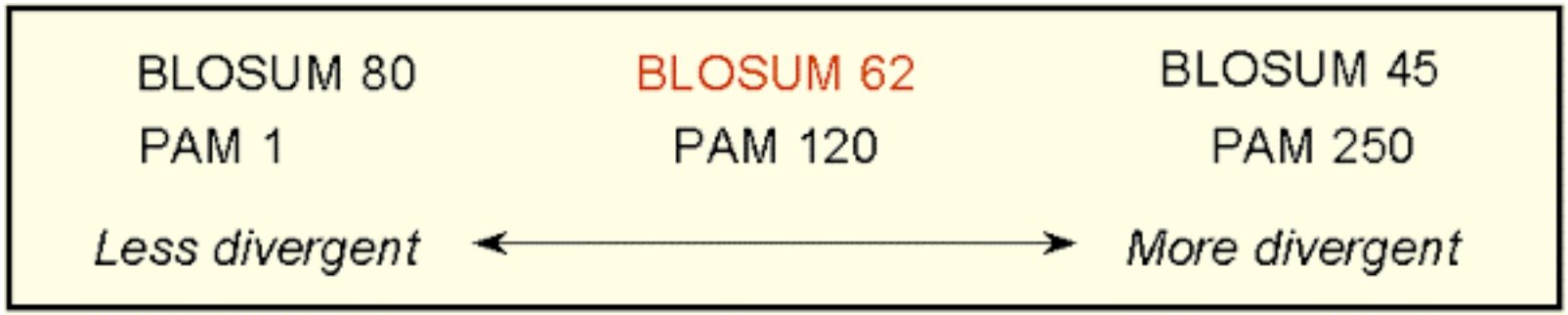
Table 2 - The log odds matrix for BLOSUM 62

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

Source: http://helix.biology.mcmaster.ca/721/distance/node10.html

Substitution matrices: comparison

- PAM *vs* BLOSUM



Source: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Scoring2.html

- Matrices of choice:
  - BLOSUM62: the all-weather matrix
  - PAM250: for distant relatives

Substit. matrices: comparison (cntd)

- PAM *vs* BLOSUM *(cntd)*
  - Lower PAM/higher BLOSUM matrices identify shorter local alignments of highly similar sequences
  - Higher PAM/lower BLOSUM matrices identify longer local alignments of more distant sequences

Substit. matrices: comparison (cntd)

PAM10

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7																			
R	-10	9																		
N	-7	-9	9																	
D	-6	-17	-1	8																
C	-10	-11	-17	-21	10															
Q	-7	-4	-7	-6	-20	9														
E	-5	-15	-5	0	-20	-1	8													
G	-4	-13	-6	-6	-13	-10	-7	7												
H	-11	-4	-2	-7	-10	-2	-9	-13	10											
I	-8	-8	-8	-11	-9	-11	-8	-17	-13	9										
L	-9	-12	-10	-19	-21	-8	-13	-14	-9	-4	7									
K	-10	-2	-4	-8	-20	-6	-7	-10	-10	-9	-11	7								
M	-8	-7	-15	-17	-20	-7	-10	-12	-17	-3	-2	-4	12							
F	-12	-12	-12	-21	-19	-19	-20	-12	-9	-5	-5	-20	-7	9						
P	-4	-7	-9	-12	-11	-6	-9	-10	-7	-12	-10	-10	-11	-13	8					
S	-3	-6	-2	-7	-6	-8	-7	-4	-9	-10	-12	-7	-8	-9	-4	7				
T	-3	-10	-5	-8	-11	-9	-9	-10	-11	-5	-10	-6	-7	-12	-7	-2	8			
W	-20	-5	-11	-21	-22	-19	-23	-21	-10	-20	-9	-18	-19	-7	-20	-8	-19	13		
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-9	-8	10	
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8

Substit. matrices: comparison (cntd)

PAM250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

Substit. matrices: comparison (cntd)

BLOSUM62

	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

Pairwise alignments

Alignment: the problem

Given two sequences, *S* and *T*, and a *scoring matrix* find their relative arrangement with the highest "score".

```
Seq. #1: G A A T T C A G T T A
Seq. #2: G G A T C G A
```

Alignment: the problem (cntd)

```
  G A A T T C A G T T A
  | |
G G A T C G A
```

```
G A A T T C A G T T A
|   | |     |
G G A T C G A
```

```
G A A T T C - A G T T A
|   |   | |   |
G G A - T C G A
```

Alignment: the problem (cntd)

- Scoring schemes: three possible situations…
  - Match
  - Mismatch
  - Gap
    - Gap initiation
    - Gap extension

Penalise???

Penalise

How much??

Alignment: a naïve approach

Seq #1

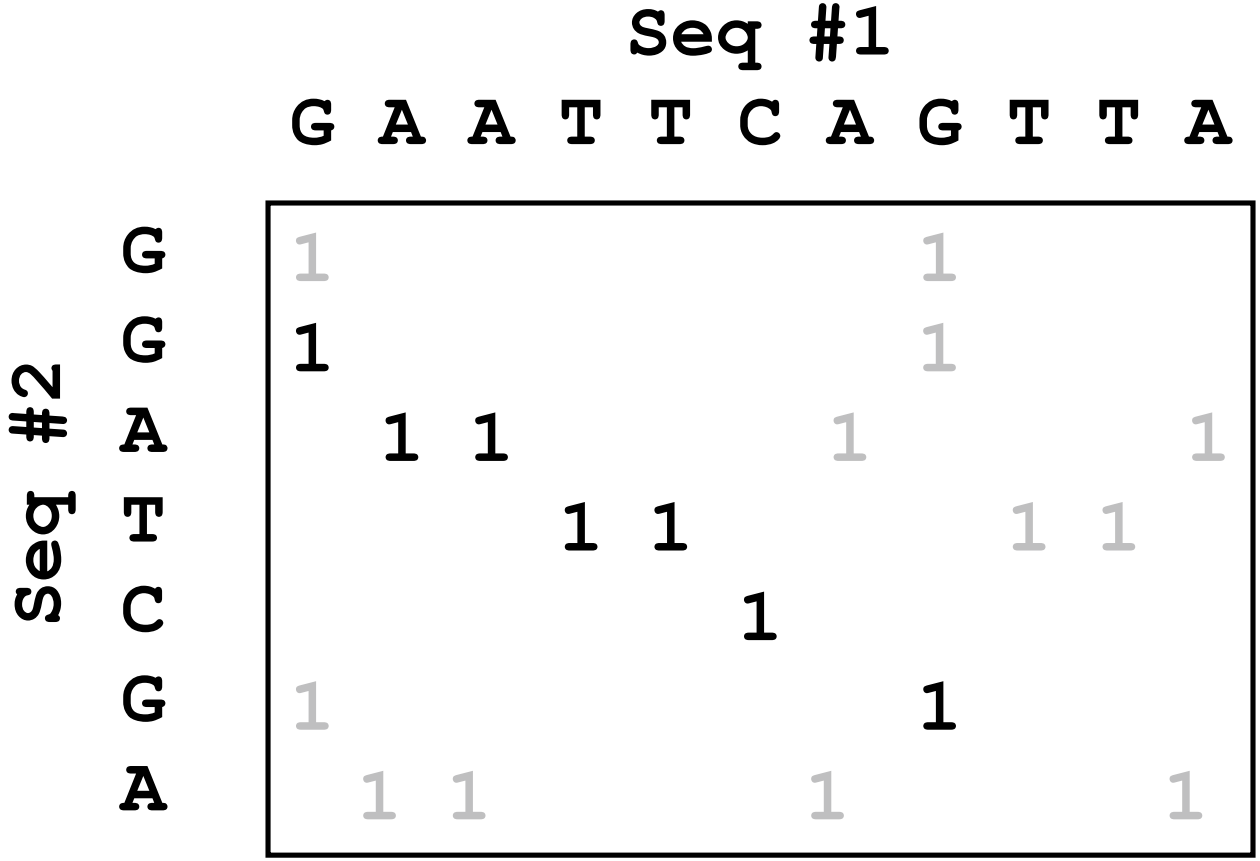
Seq #2

	G	A	A	T	T	C	A	G	T	T	A
G	1							1			
G	1							1			
A		1	1				1				1
T				1	1				1	1	
C						1					
G	1							1			
A		1	1				1				1

insertion seq. #2

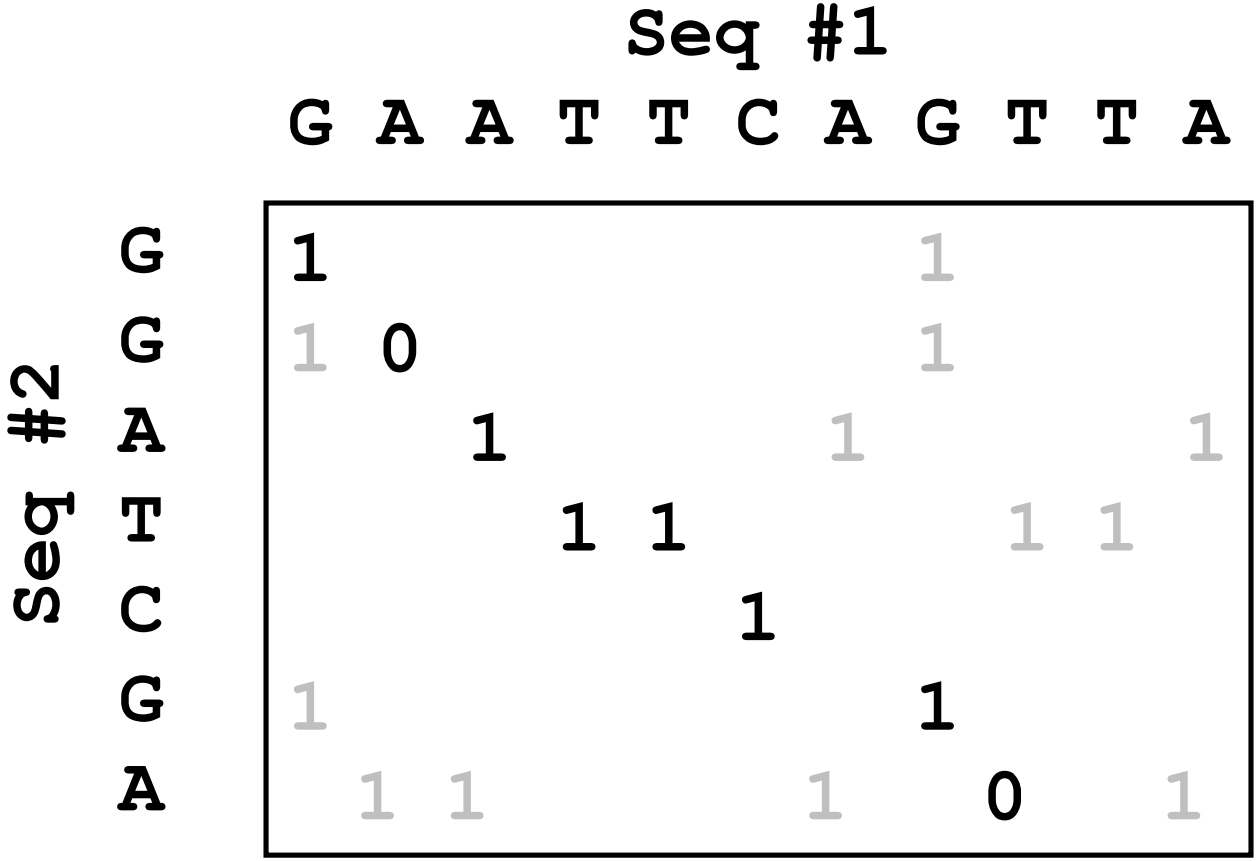
insertion seq. #1

Alignment: a naïve approach



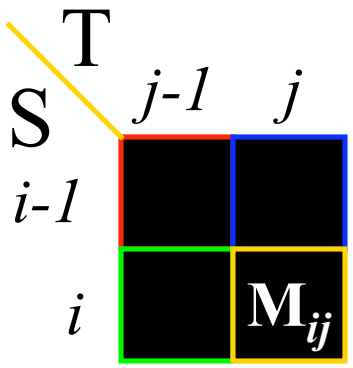

```
  G A A T T C - A G T T A
  | |   |   |   |
G G A - T - C G A
```

Alignment: a naïve approach (cntd)




```
G A A T T C A G T T A
|   | |   |   |
G G A T - C - G A
```

Global alignment



PAM

DNA matrix BLOSUM

$$M_{i,j} = \text{MAX} \begin{cases} M_{i-1,j-1} + \text{Score}(S_i, T_j) \\ M_{i,j-1} + w \\ M_{i-1,j} + w \end{cases}$$

Gap penalty

Needleman & Wunsch, 1970

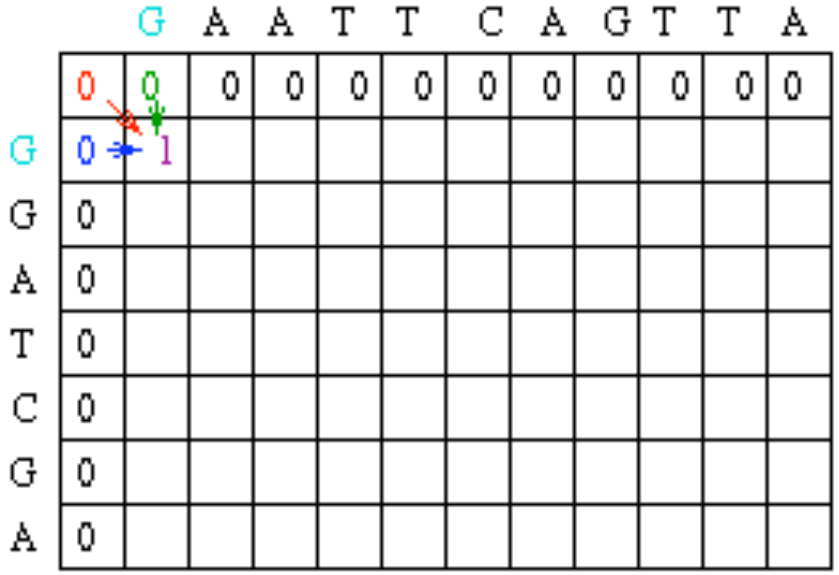
Alignment: adding scores (cntd)

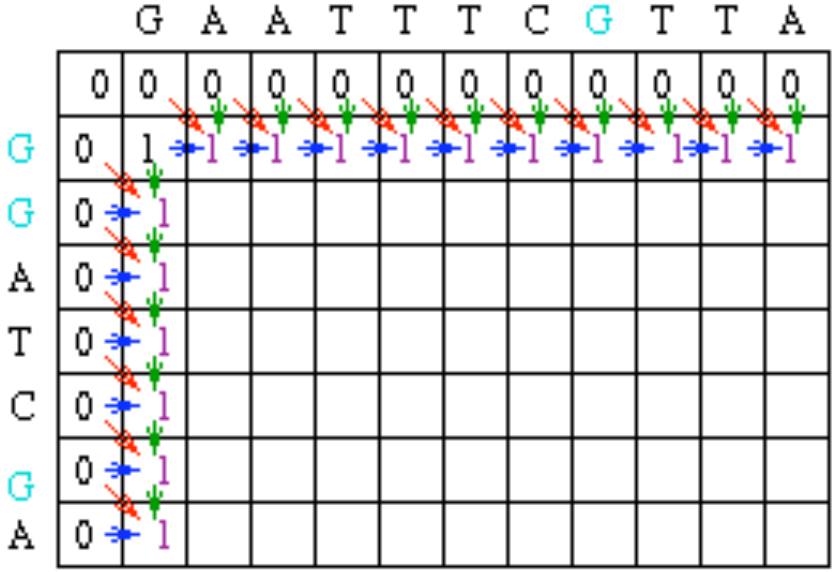
- In the following example: match=1, mismatch=gap=0.

- In each step we need to keep track only the scores of the (*i*,*j*) position and its immediate neighbours: (*i-1*,*j-1*), (*i-1*,*j*) and (*i*,*j-1*).

- We backtrack from the right-down corner to find the actual alignment.

Alignment: adding scores (cntd)





S(1,1) = MAX{
  S(0,0)+1=1,
  S(0,1)+w=0,
  S(1,0)+w=0}=1

Source:
http://www.sbc.su.se/~per/molbioinfo2001/dynprog/dynamic.html

Alignment: adding scores (cntd)

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

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	3	4	4	4	4	4
G	0	1	2	2	3	3	3	4	4	5	5	5
A	0	1	2	3	3	3	3	4	5	5	5	**6**

Source:
http://www.sbc.su.se/~per/molbioinfo2001/dynprog/dynamic.html

Alignment: adding scores (cntd)

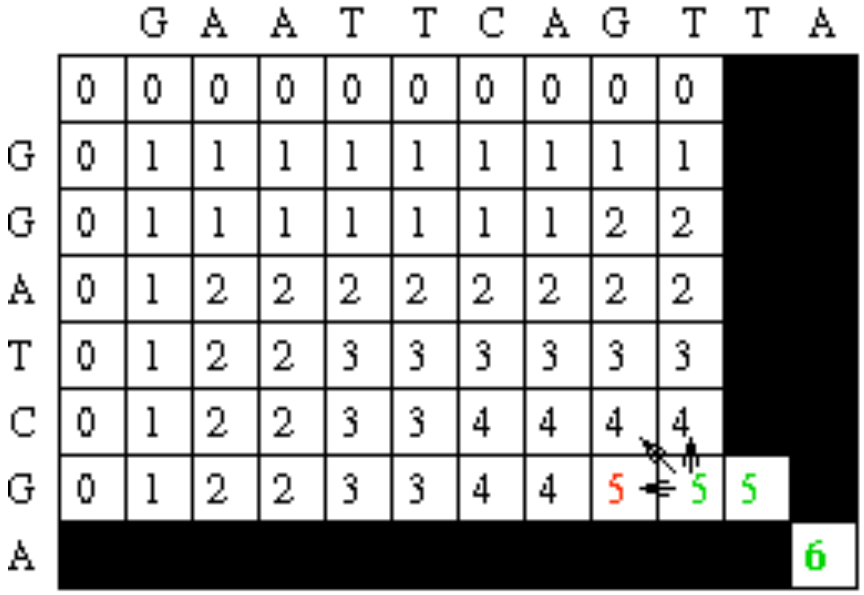
		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	1	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5	5
A	0	1	2	3	3	3	4	5	5	5	5	6

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	
G	0	1	1	1	1	1	1	1	1	1	1	
G	0	1	1	1	1	1	1	1	2	2	2	
A	0	1	2	2	2	2	2	2	2	2	2	
T	0	1	2	2	3	3	3	3	3	3	3	
C	0	1	2	2	3	3	4	4	4	4	4	
G	0	1	2	2	3	3	4	4	5	5	5	
A												6

```
                (Seq #1)                A
Alignment:                              |
                (Seq #2)                A
```

Source:
http://www.sbc.su.se/~per/molbioinfo2001/dynprog/dynamic.html

Alignment: adding scores (cntd)

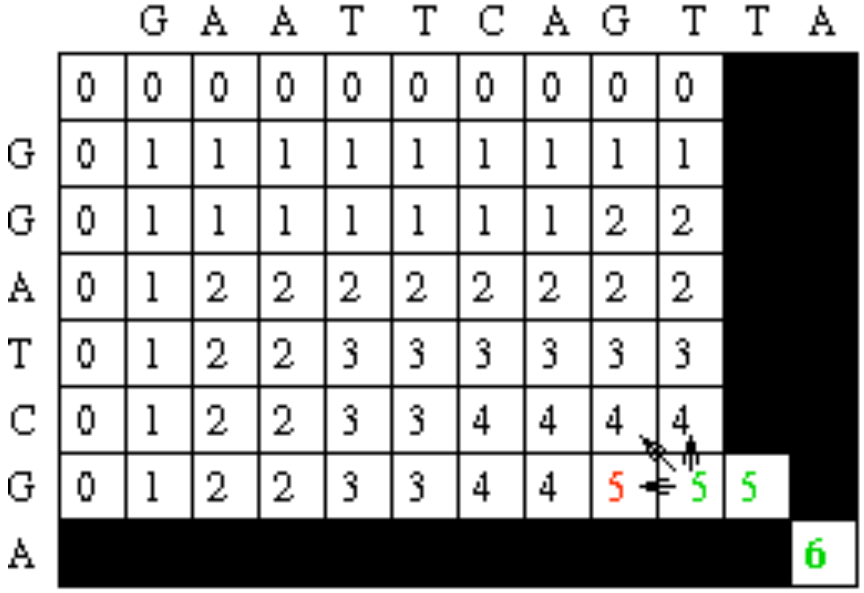


		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0		
G	0	1	1	1	1	1	1	1	1	1		
G	0	1	1	1	1	1	1	1	2	2		
A	0	1	2	2	2	2	2	2	2	2		
T	0	1	2	2	3	3	3	3	3	3		
C	0	1	2	2	3	3	4	4	4	4		
G	0	1	2	2	3	3	4	4	5	5	5	
A												6

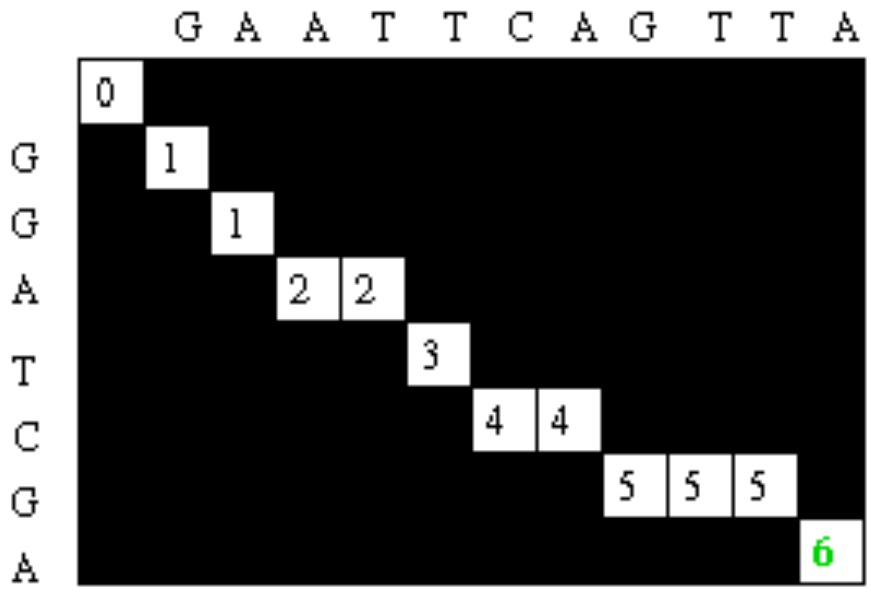
```
                (Seq #1)                T A
Alignment:                                |
                (Seq #2)                - A
```

Source:
http://www.sbc.su.se/~per/molbioinfo2001/dynprog/dynamic.html

Alignment: adding scores (cntd)



		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0		
G	0	1	1	1	1	1	1	1	1	1		
G	0	1	1	1	1	1	1	1	2	2		
A	0	1	2	2	2	2	2	2	2	2		
T	0	1	2	2	3	3	3	3	3	3		
C	0	1	2	2	3	3	4	4	4	4		
G	0	1	2	2	3	3	4	4	5	5	5	
A											6	

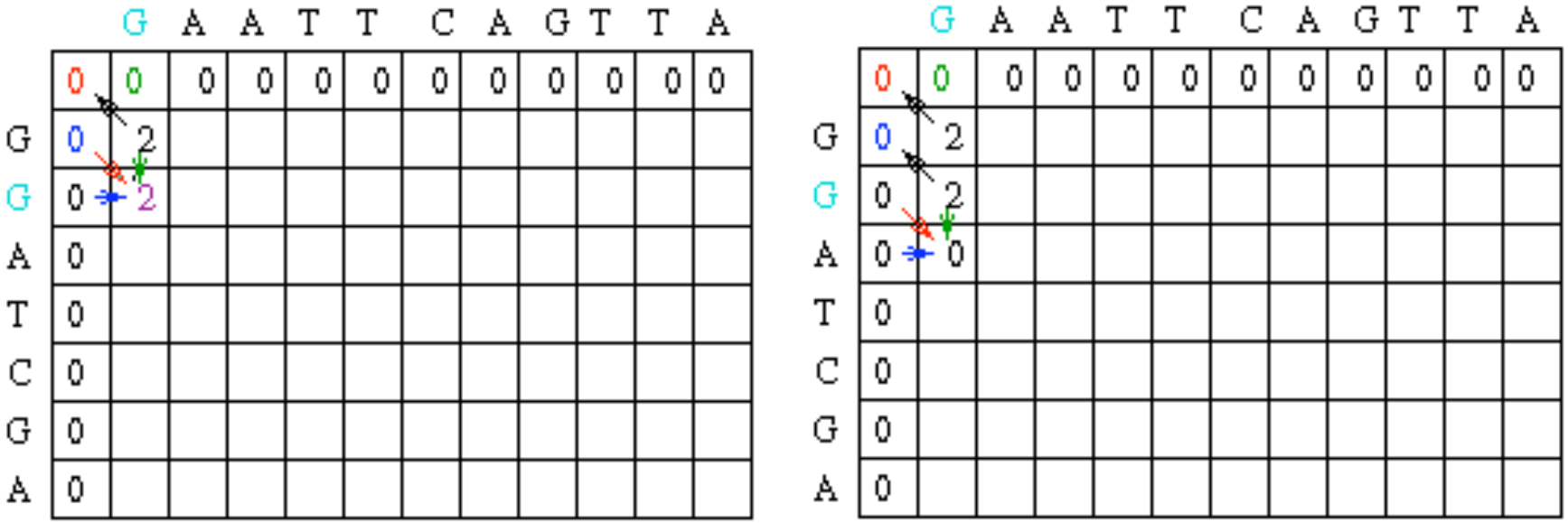


		G	A	A	T	T	C	A	G	T	T	A
	0											
G		1										
G			1									
A				2	2							
T						3						
C							4	4				
G									5	5	5	
A												6

```
                (Seq #1) G A A T T C A G T T A
                         |   |   | | |   |     |
Alignment:               |   |   | | |   |     |
                (Seq #2) G G A - T C - G - - A
```

Source:
http://www.sbc.su.se/~per/molbioinfo2001/dynprog/dynamic.html

Alignment: another example



Now: match=2, mismatch=-1, gap=-2

Source:
http://www.sbc.su.se/~per/molbioinfo2001/dynprog/adv_dynamic.html

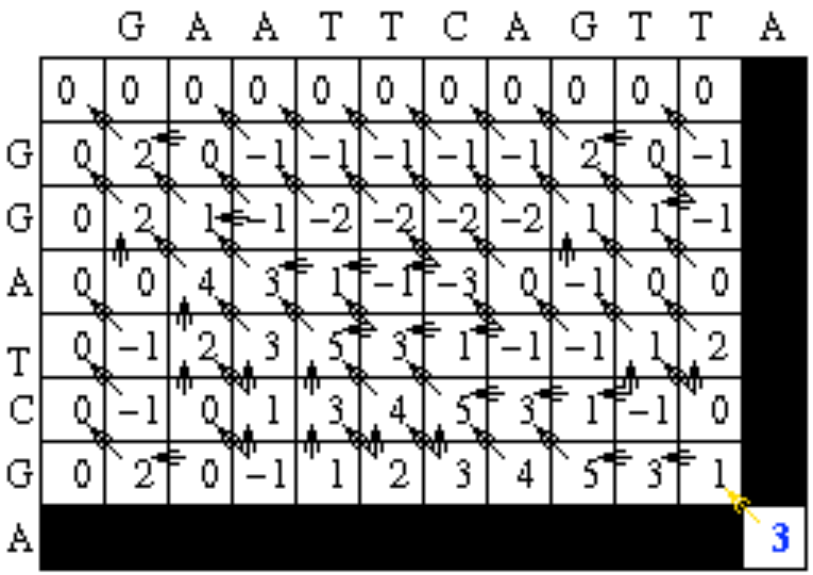
Alignment: another example (cntd)

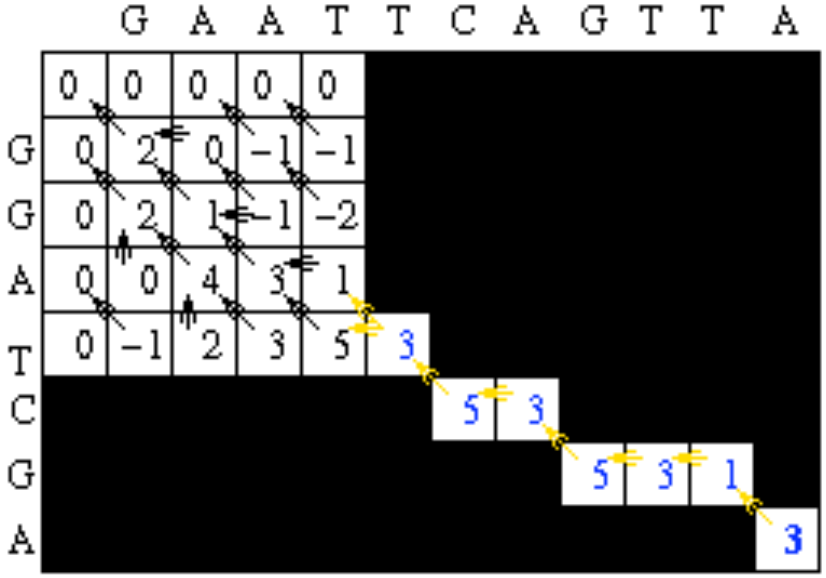
		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	2	0	−1								
G	0	2	1	−1								
A	0	0	4									
T	0	−1	2									
C	0	−1	0									
G	0	2	0									
A	0	0	4									

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	2	0	−1	−1	−1	−1	−1	2	0	−1	−1
G	0	2	1	−1	−2	−2	−2	−2	1	1	−1	−2
A	0	0	4	3	1	−1	−3	0	−1	0	0	1
T	0	−1	2	3	5	3	1	−1	−1	1	2	0
C	0	−1	0	1	3	4	5	3	1	−1	0	1
G	0	2	0	−1	1	2	3	4	5	3	1	−1
A	0	0	4	2	0	0	1	5	3	4	2	3

Source:
http://www.sbc.su.se/~per/molbioinfo2001/dynprog/adv_dynamic.html

Alignment: another example (cntd)

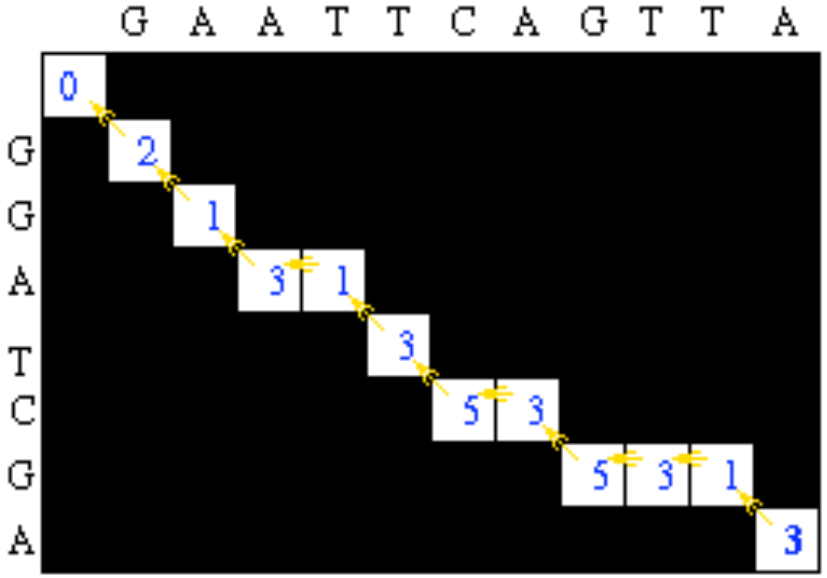


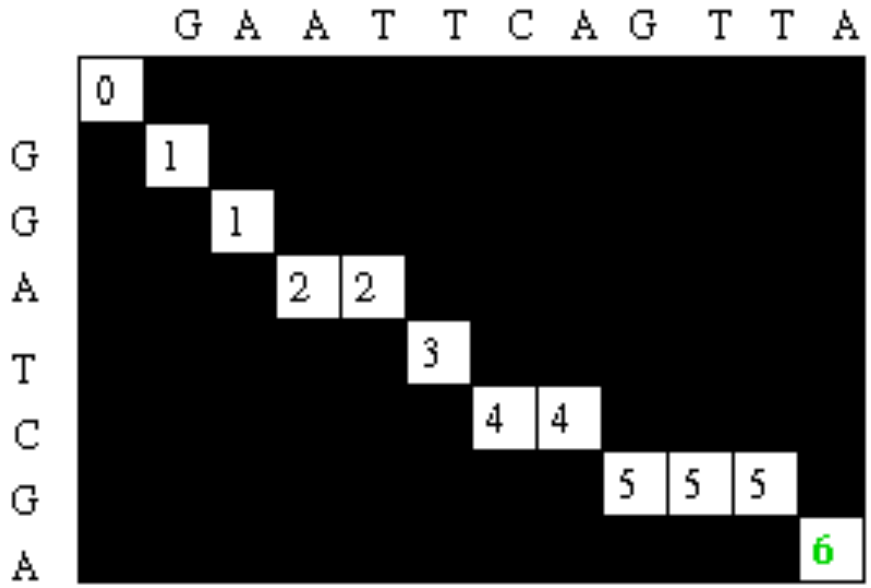



```
                (Seq #1)        T C A G T T A
                                | |   |     |
Alignment:
                (Seq #2)        T C - G - - A
```

Source: http://www.sbc.su.se/~per/molbioinfo2001/dynprog/adv_dynamic.html

Alignment: another example (cntd)

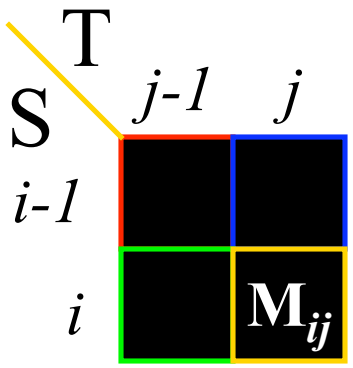





```
                (Seq #1) G A A T T C A G T T A
                         |   |   | | |   |     |
Alignment:      (Seq #2) G G A - T C - G - - A
```

Source: http://www.sbc.su.se/~per/molbioinfo2001/dynprog/adv_dynamic.html

Local alignment



PAM

DNA matrix BLOSUM

$$M_{i,j} = \text{MAX} \begin{cases} 0 \\ M_{i-1,\,j-1} + \text{Score}(S_i, T_j) \\ M_{i,j-1} + w \\ M_{i-1,j} + w \end{cases}$$

Gap penalty

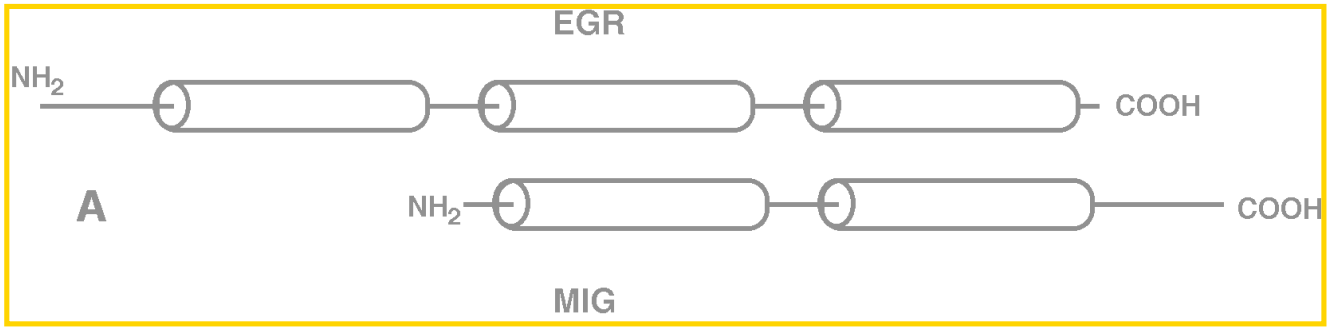
Smith & Waterman, 1981

Local alignment

Given two sequences, *S* and *T*, find two subsequences, *s* and *t*, whose alignment has the highest "score" amongst all subsequence pairs.

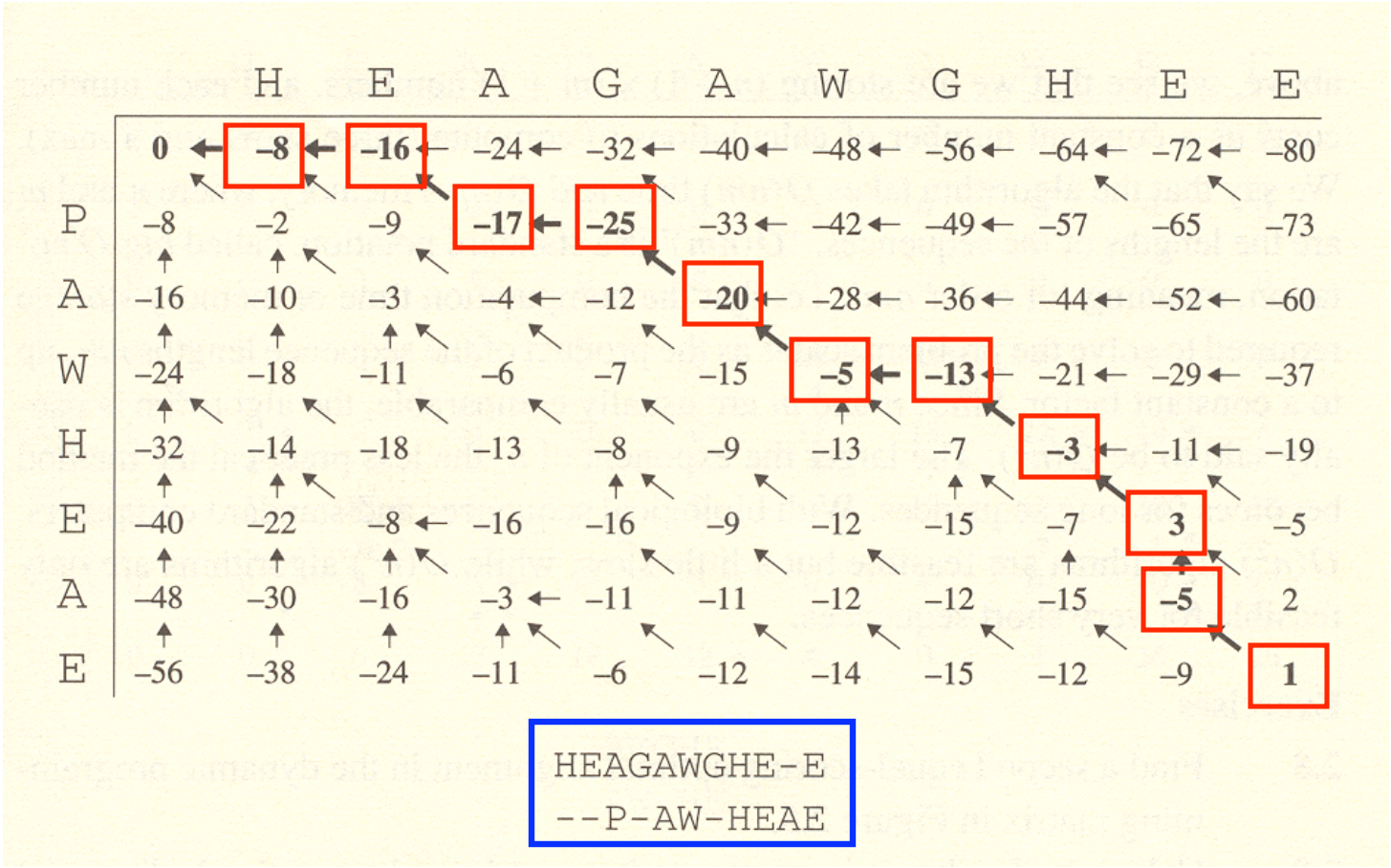
Why do we need local alignment, if we have the global one?

Local alignment: an example



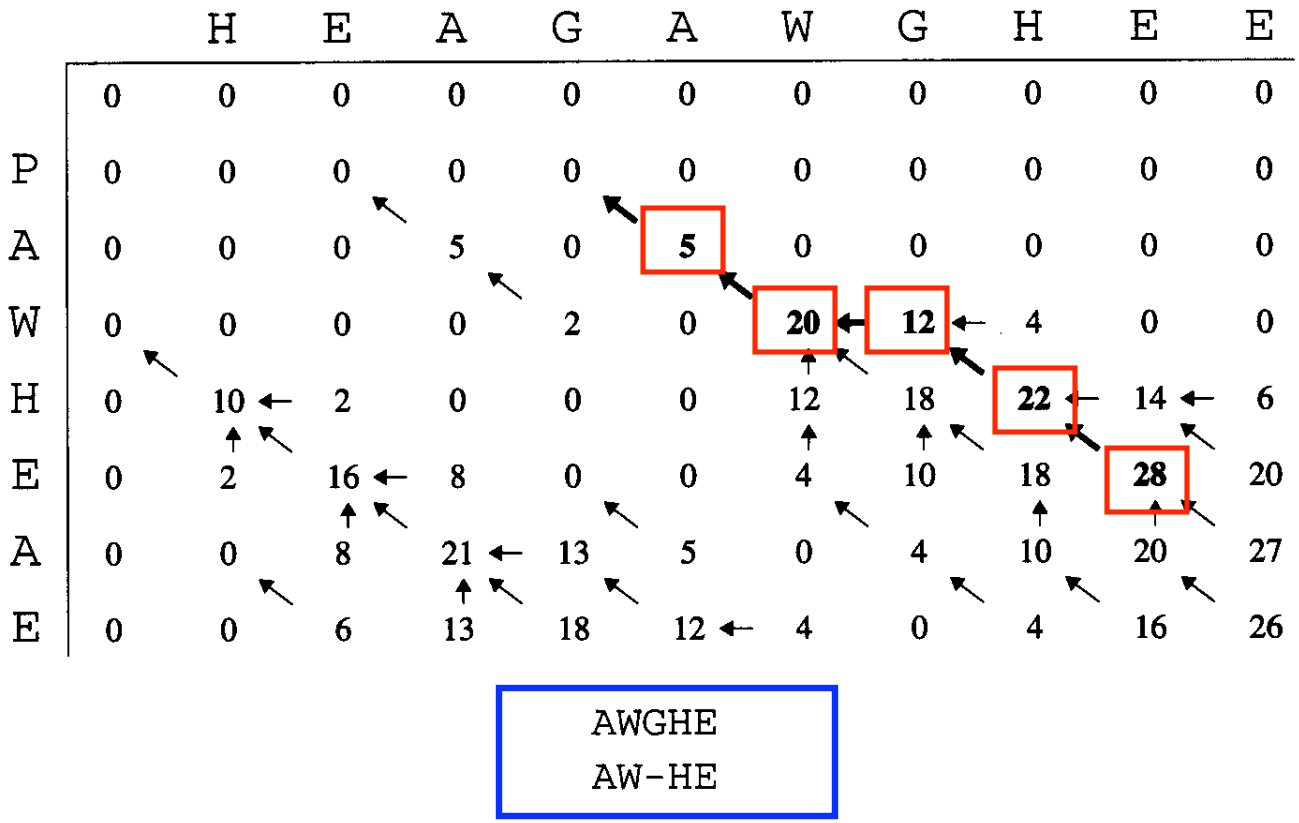

```
EGR4_HUMAN    KA [FACPVESCVRSFARSDELNRHLRIH] TGHKP [FQCRICLRNFSRSDHLTSHVRTH] TGEKP [FACDV--CGRRFARSDEKKRHSKVH]
EGR4_RAT      KA [FACPVESCVRTFARSDELNRHLRIH] TGHKP [FQCRICLRNFSRSDHLTTHVRTH] TGEKP [FACDV--CGRRFARSDEKKRHSKVH]
EGR3_HUMAN    RP [HACPAEGCDRRFSRSDELTRHLRIH] TGHKP [FQCRICMRSFSRSDHLTTHIRTH] TGEKP [FACEF--CGRKFARSDERKRHAKIH]
EGR3_RAT      RP [HACPAEGCDRRFSRSDELTRHLRIH] TGHKP [FQCRICMRSFSRSDHLTTHIRTH] TGEKP [FACEF--CGRKFARSDERKRHAKIH]
EGR1_HUMAN    RP [YACPVESCDRRFSRSDELTRHIRIH] TGQKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDI--CGRKFARSDERKRHTKIH]
EGR1_MOUSE    RP [YACPVESCDRRFSRSDELTRHIRIH] TGQKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDI--CGRKFARSDERKRHTKIH]
EGR1_RAT      RP [YACPVESCDRRFSRSDELTRHIRIH] TGQKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDI--CGRKFARSDERKRHTKIH]
EGR1_BRARE    RP [YACPVETCDRRFSRSDELTRHIRIH] TGQKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACEI--CGRKFARSDERKRHTKIH]
EGR2_RAT      RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDY--CGRKFARSDERKRHTKIH]
EGR2_XENLA    RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDY--CGRKFARSDERKRHTKIH]
EGR2_MOUSE    RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDY--CGRKFARSDERKRHTKIH]
EGR2_HUMAN    RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDY--CGRKFARSDERKRHTKIH]
EGR2_BRARE    RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNFSRSDHLTTHIRTH] TGEKP [FACDF--CGRKFARSDERKRHTKIH]
MIG1_KLULA    -- [-------------------------] ---RP [YVCPICQRGFHRLEHQTRHIRTH] TGERP [HACDFPGCSKRFSRSDELTRHRRIH]
MIG1_KLUMA    -- [-------------------------] ---RP [YMCPICHRGFHRLEHQTRHIRTH] TGERP [HACDFPGCAKRFSRSDELTRHRRIH]
MIG1_YEAST    -- [-------------------------] ---RP [HACPICHRAFHRLEHQTRHMRIH] TGEKP [HACDFPGCVKRFSRSDELTRHRRIH]
MIG2_YEAST    -- [-------------------------] ---RP [FRCDTCHRGFHRLEHKKRHLRTH] TGEKP [HHCAFPGCGKSFSRSDELKRHMRTH]
                 [                         ]    :* [. *  * * * * :* . *:* *] ***:* [. *    * : *:**** .** : *]
```

Local vs. global alignment



Source: Durbin et al "Biological Sequence Analysis", 1998, Cambridge University Press

Local vs. global alignment (cntd)



		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	**5**	0	0	0	0	0
W	0	0	0	0	2	0	**20**	**12**	4	0	0
H	0	10	2	0	0	0	12	18	**22**	14	6
E	0	2	16	8	0	0	4	10	18	**28**	20
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

```
AWGHE
AW-HE
```

Source: Durbin et al "Biological Sequence Analysis", 1998, Cambridge University Press

Local alignment (cntd)

- Characteristics of local alignments:
  - The alignment can start/end at any point in the matrix.
  - No negative scores.
  - The mean value of the scoring matrix (e.g. PAM, BLOSUM) should be negative.
  - There should be positive scores in the scoring matrix.