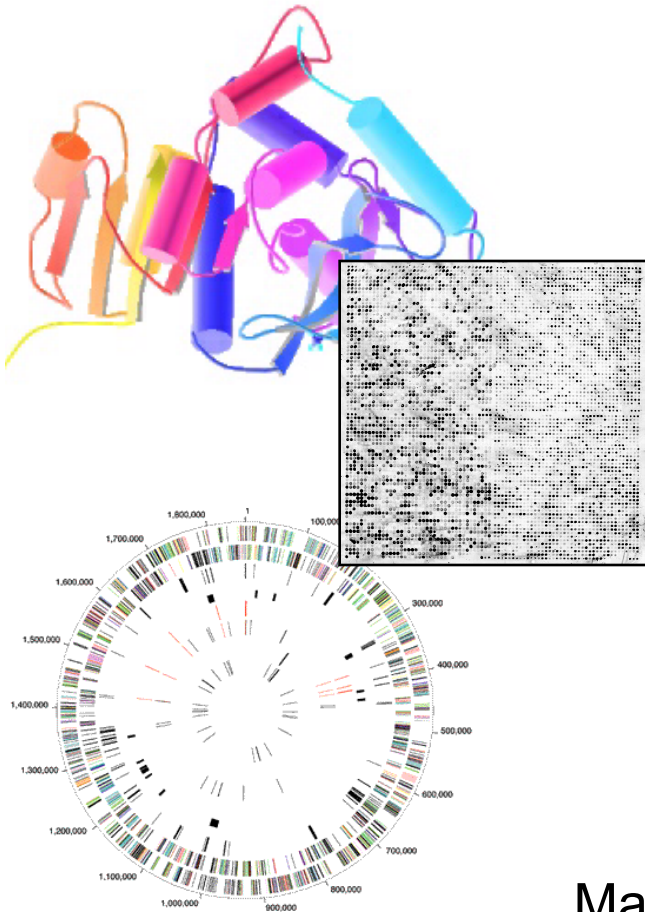


BIOINFORMATICS

Sequences I



Mark Gerstein, Yale University
gersteinlab.org/courses/452
(last edit in spring '11, final edit)

Sequence Topics (Contents)

- Basic Alignment via Dynamic Programming
- Suboptimal Alignment
- Gap Penalties
- Similarity (PAM) Matrices
- Multiple Alignment
- Profiles, Motifs, HMMs
- Local Alignment
- Probabilistic Scoring Schemes
- Rapid Similarity Search: Fasta
- Rapid Similarity Search: Blast
- Practical Suggestions on Sequence Searching
- Transmembrane helix predictions
- Secondary Structure Prediction: Basic GOR
- Secondary Structure Prediction: Other Methods
- Assessing Secondary Structure Prediction
- Features of Genomic DNA sequences

Aligning Text Strings

Raw Data ???

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

2 matches, 0 gaps

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

3 matches (2 end gaps)

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

4 matches, 1 insertion

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

4 matches, 1 insertion

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

Dynamic Programming

- What to do for Bigger String?

```
SSDSEREEHVKRFRQALDDTGMKVPMATNLFTHPVFKDGGFTANDRDVRRYALRKTIRNIDLAVELGAE'YVAVGGREGAESGGAKDVRDALDRMKEAFDLLGEYVTSQGYDIRFAI  
EP  
KPNEPRGDILLPTVGHALAFIERLERPELYGVNPEVGHEQMAGLNFPHGIAQALWAGKLFHIDLNGQNGIKYDQDLRFQAGDLRAAFWLVDLLESAGYSGPRHFDFKPPRTEDFDGVW  
AS
```

- Needleman-Wunsch (1970) provided first automatic method

- ◇ Dynamic Programming to Find Global Alignment

- Their Test Data (J→Y)

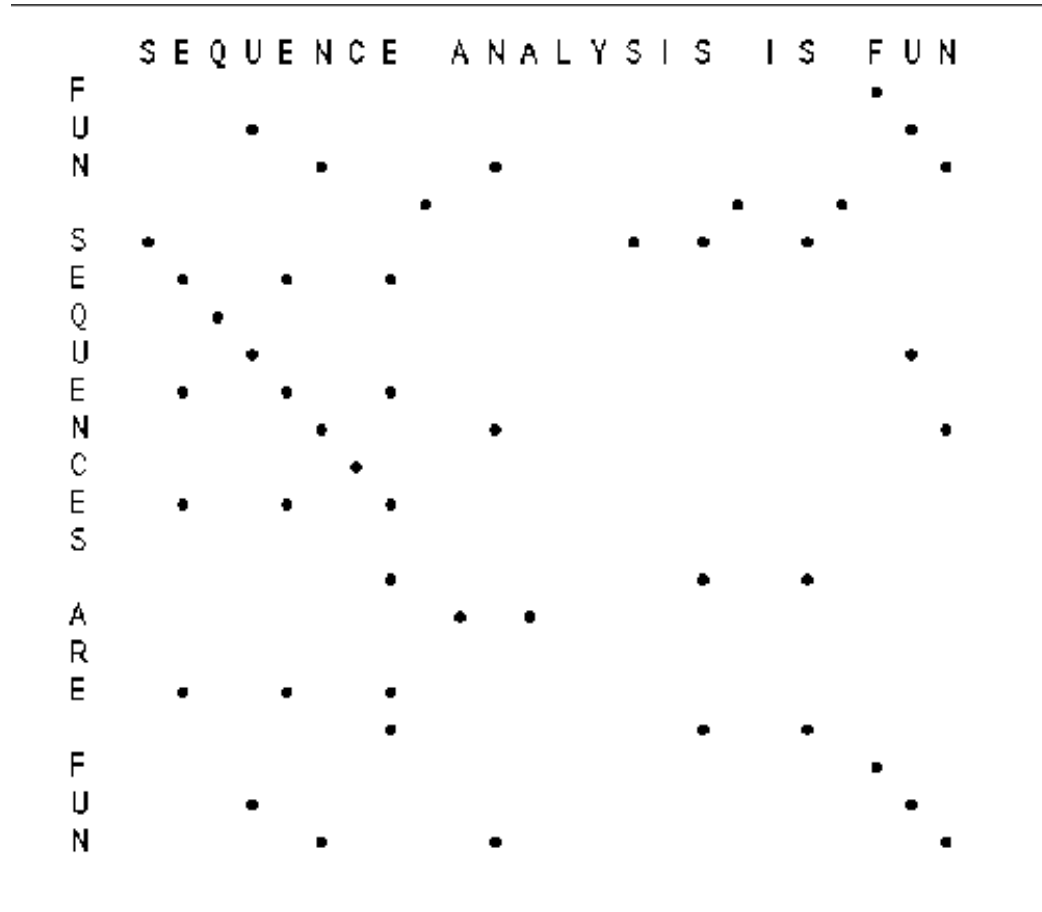
- ◇ ABCNYRQCLCRPM
AYCYNRCKCRBP

Step 1 -- Make a Dot Plot (Similarity Matrix)

Put 1's where characters are identical.

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

A More Interesting Dot Matrix



(adapted from R Altman)

Step 2 -- Start Computing the Sum Matrix

```

new_value_cell(R,C) <=
  cell(R,C)           { Old value, either 1 or 0   }
  + Max[
    cell (R+1, C+1),   { Diagonally Down, no gaps   }
    cells(R+1, C+2 to C_max), { Down a row, making col. gap }
    cells(R+2 to R_max, C+1) { Down a col., making row gap }
  ]

```

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0



Step 2 -- Start Computing the Sum Matrix

```

new_value_cell(R,C) <=
  cell(R,C)                { Old value, either 1 or 0 }
  + Max[
    cell (R+1, C+1),        { Diagonally Down, no gaps }
    cells(R+1, C+2 to C_max), { Down a row, making col. gap }
    cells(R+2 to R_max, C+1) { Down a col., making row gap }
  ]

```

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1	1				
Y					1								
N				1									
R						1					1		
C			1					1	1				
K													
C			1					1	1				
R						1					1		
B		1											
P												1	

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1	1				
Y					1								
N				1									
R						1					1		
C			1					1	1				
K													
C			1					1	1				
R						1					2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Step 3 -- Keep Going

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1	1				
Y					1								
N				1									
R						1					1		
C			1					1	1				
K													
C			1					1	1				
R						1					2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1	1				
Y					1								
N				1									
R						5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Step 4 -- Sum Matrix All Done

Alignment Score is 8 matches.

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0



Step 5 -- Traceback

Find Best Score (8) and Trace Back

A B C N Y - R Q C L C R - P M
 A Y C - Y N R - C K C R B P

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Hansel & Gretel

Step 6 -- Alternate Tracebacks

A B C - N Y R Q C L C R - P M
 A Y C Y N - R - C K C R B P

Also,
 Suboptimal
 Alignments

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

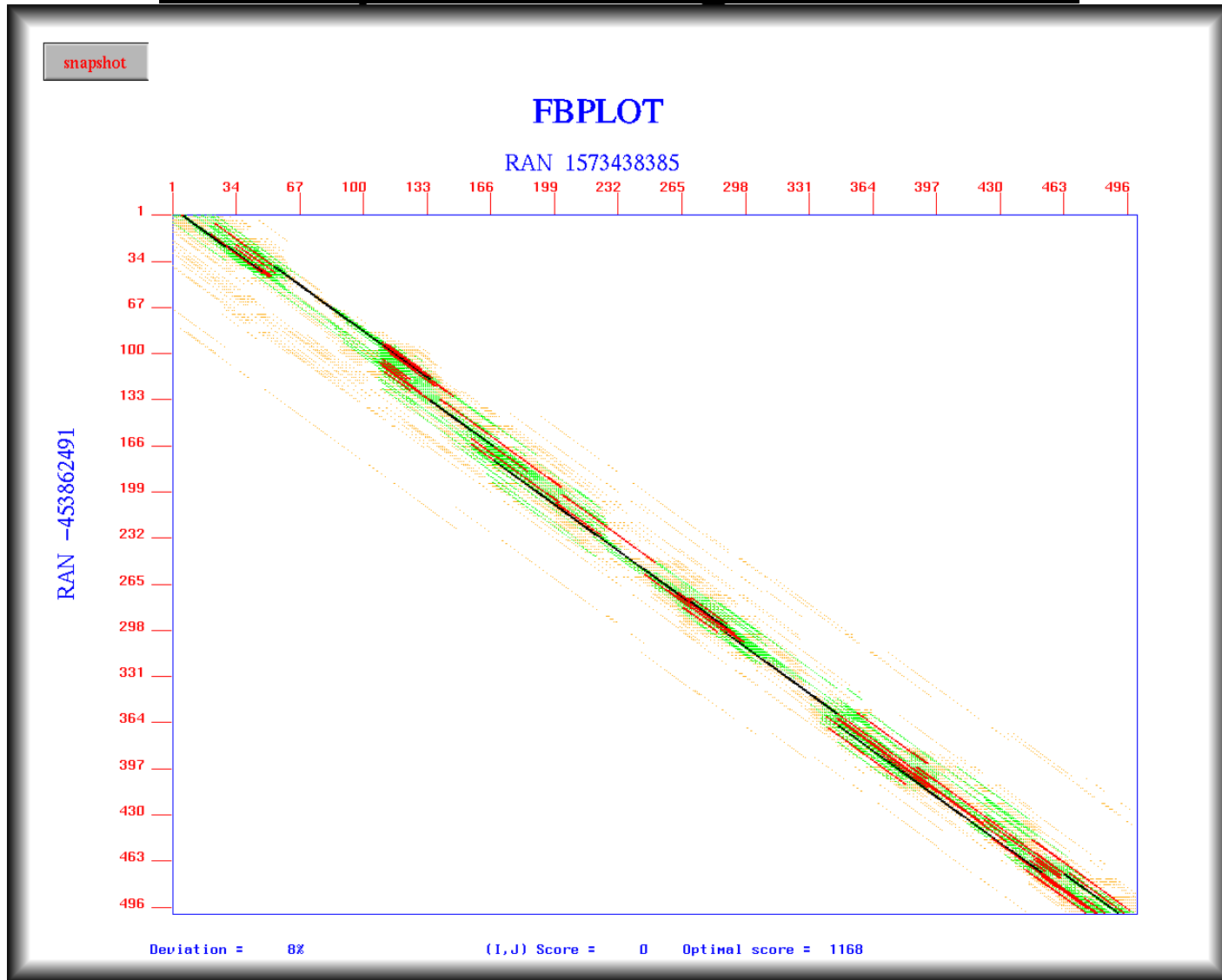
Suboptimal Alignments

```
;
; Random DNA sequence generated using the seed :   -453862491
;
;   500 nucleotides
;
; A:C:G:T =   1 :   1 :   1 :   1
;
RAN  -453862491
AAATGCCAAA TCATACGAAC AGCCGACGAC GGGAGCAACC CAAGTCGCAG TTCGCTTGAG CTAGCGCGCT
CCCACCGGGA TATACACTAA TCATTACAGC AGGTCTCCTG GGCGTACAGA CTAGCTGAAC GCGCTGCGCC
AATTCCAACT TCGGTATGAA GGATCGCCTG CGGTATCGC TGACTTGAGT AACCAGATCG CTAAGGTTAC
GCTGGGGCAA TGATGGATGT TAACCCCTTA CAGTCTCGGG AGGGACCTTA AGTCGTAATA GATGGCAGCA
TTAATACCTT CGCCGTTAAT ATACCTTTAA TCCGTCTTG TCAATGCCGT AGCTGCAGTG AGCCTTCTGT
CACGGGCATA CCGCGGGGTA GCTGCAGCAA CCGTAGGCTG AGCATCAAGA AGACAAACAC TCCTCGCCTA
CCCCGGACAT CATATGACCA GGCAGTCTAG GCGCCGTTAG AGTAAGGAGA CCGGGGGGCC GTGATGATAG
ATGGCGTGTT 1
;
; Random DNA sequence generated using the seed :   1573438385
;
;   500 nucleotides
;
; A:C:G:T =   1 :   1 :   1 :   1
;
RAN  1573438385
CCCTCCATCG CCAGTTCCTG AAGACATCTC CGTGACGTGA ACTCTCTCCA GGCATATTAA TCGAAGATCC
CCTGTGCTGA CGCGGATTAC GAGGGGATGG TGCTAATCAC ATTGCGAACA TGTTTCGGTC CAGACTCCAC
CTATGGCATC TTCCGCTATA GGGCACGTAA CTTTCTTCGT GTGGCGGCGC GGCAACTAAA GACGAAAGGA
CCACAACGTG AATAGCCCGT GTCGTGAGGT AAGGGTCCCG GTGCAAGAGT AGAGGAAGTA CGGGAGTACG
TACGGGGCAT GACGCGGGCT GGAATTCAC ATCGCAGAAC TTATAGGCAG CCGTGTGCCT GAGGCCGCTA
GAACCTCAA CGCTAACTAG TGATAACTAC CGTGTGAAAG ACCTGGCCCG TTTGTCCCT GAGACTAATC
GCTAGTTAGG CCCCATTGT AGCACTCTGG CGCAGACCTC GCAGAGGGAC CGGCCTGACT TTTCCGGCT
TCCTCTGAGG 1

Parameters: match weight = 10, transition weight = 1, transversion weight = -3
Gap opening penalty = 50   Gap continuation penalty = 1
Run as a local alignment (Smith-Waterman)
```

(courtesy of Michael Zucker)

Suboptimal Alignments II



(courtesy of Michael Zucker)

Gap Penalties

The score at a position can also factor in a penalty for introducing gaps (i. e., not going from i, j to $i-1, j-1$).

Gap penalties are often of linear form:

$$\text{GAP} = a + bN$$

GAP is the gap penalty

a = cost of opening a gap

b = cost of extending the gap by one (affine)

N = length of the gap

(Here assume $b=0$, $a=1/2$, so $\text{GAP} = 1/2$ regardless of length.)

ATGCAAAT

ATG-AAAAT .5

ATG--AAAT .5 + (1) b [$b=.1$]

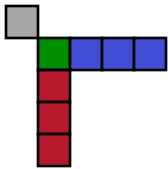
ATG---AAT .5 + (2) (.1) = .7

Step 2 -- Computing the Sum

Matrix with Gaps

```

new_value_cell(R,C) <=
  cell(R,C)           { Old value, either 1 or 0 }
  + Max[
    cell(R+1,C+1),    { Diagonally Down, no gaps }
    cells(R+1,C+2 to C_max) - GAP, { Down a row, making col. gap }
    cells(R+2 to R_max,C+1) - GAP { Down a col., making row gap }
  ]
  
```



	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1.5	0	0
B	1	1.5	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

GAP
= 1/2

All Steps in Aligning a 4-mer

	C	R	P	M
C	1			
R		1		
B				
P			1	

	C	R	P	M
C	1			
R		2	0	0
B	1	1	0	0
P	0	0	1	0

	C	R	P	M
C	3	1	0	0
R	1	2	0	0
B	1	1	0	0
P	0	0	1	0

C R B P

C R P M

- C R P M

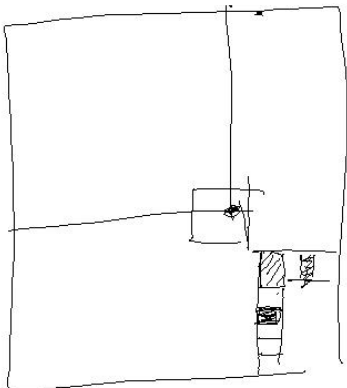
C R - P M

	C	R	P	M
C	3	1	0	0
R	1	2	0	0
B	1	1	0	0
P	0	0	1	0

Bottom right hand corner of previous matrices

Key Idea in Dynamic Programming

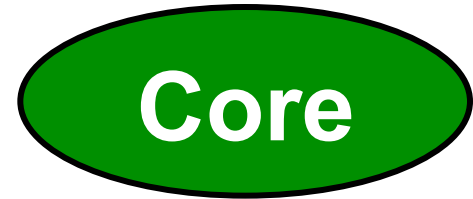
- ◇ The best alignment that ends at a given pair of positions (i and j) in the 2 sequences is the score of the best alignment previous to this position PLUS the score for aligning those two positions.
- ◇ An Example Below
 - Aligning R to K does not affect alignment of previous N-terminal residues. Once this is done it is **fixed**. Then go on to align D to E.
 - How could this be violated?
Aligning R to K changes best alignment in box.



ACSQRP--LRV-SH	R SENCV
A-SNKPQLVKLMTH	V K DFCV

ACSQRP--LRV-SH	-R	S ENCV
A-SNKPQLVKLMTH	VK	D FCV

Similarity (Substitution)



Matrix

- Identity Matrix

- ◊ Match L with L => 1
- Match L with D => 0
- Match L with V => 0??

- S(aa-1,aa-2)

- ◊ Match L with L => 1
- Match L with D => 0
- Match L with V => .5

- Number of Common Ones

- ◊ PAM
- ◊ Blossum
- ◊ Gonnet

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