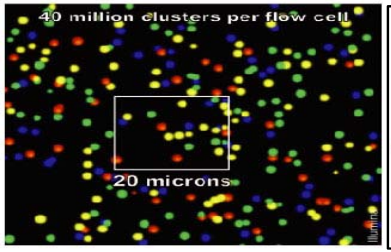
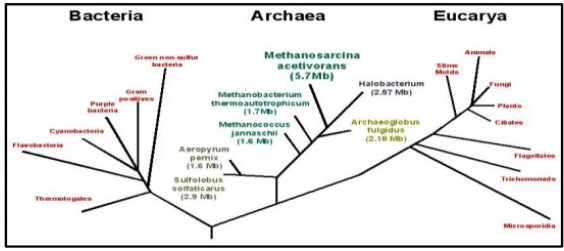
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



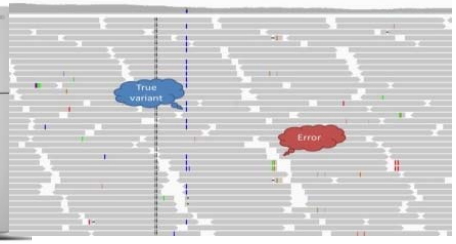


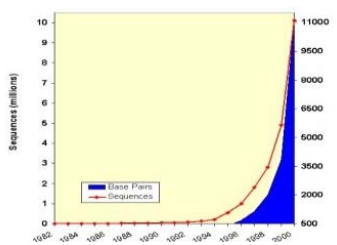
Sequence Alignment

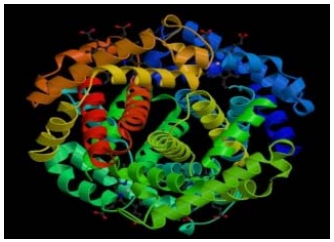
北京大学生物信息学中心 高歌

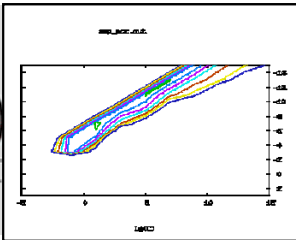
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University

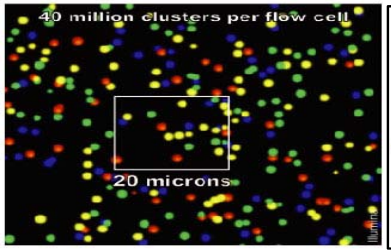


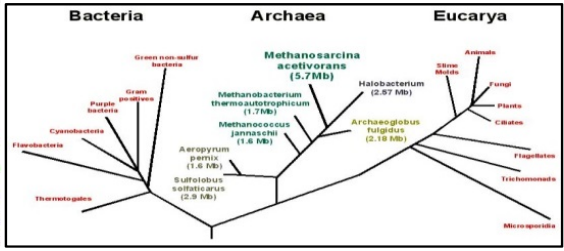






TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



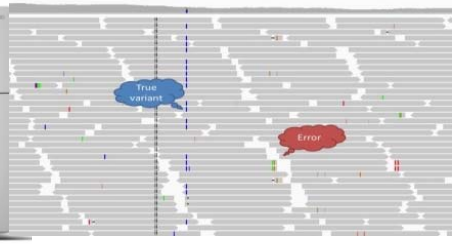


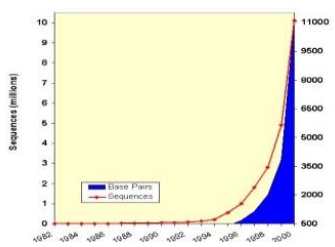
Unit 2: Global Alignment by Dynamic Programming

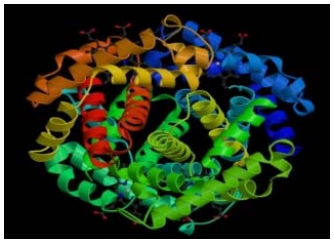
北京大学生物信息学中心 高歌

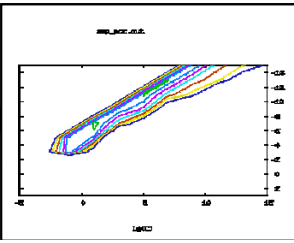
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University









Pairwise Sequence Alignment: in Maths

- Input data:
  - Two sequences S1 and S2
- Parameter(s)
  - A scoring function f for
    - Substitutions
    - Gaps

- Output:
  - The optimal alignment of S1 and S2, which has the maximal score.

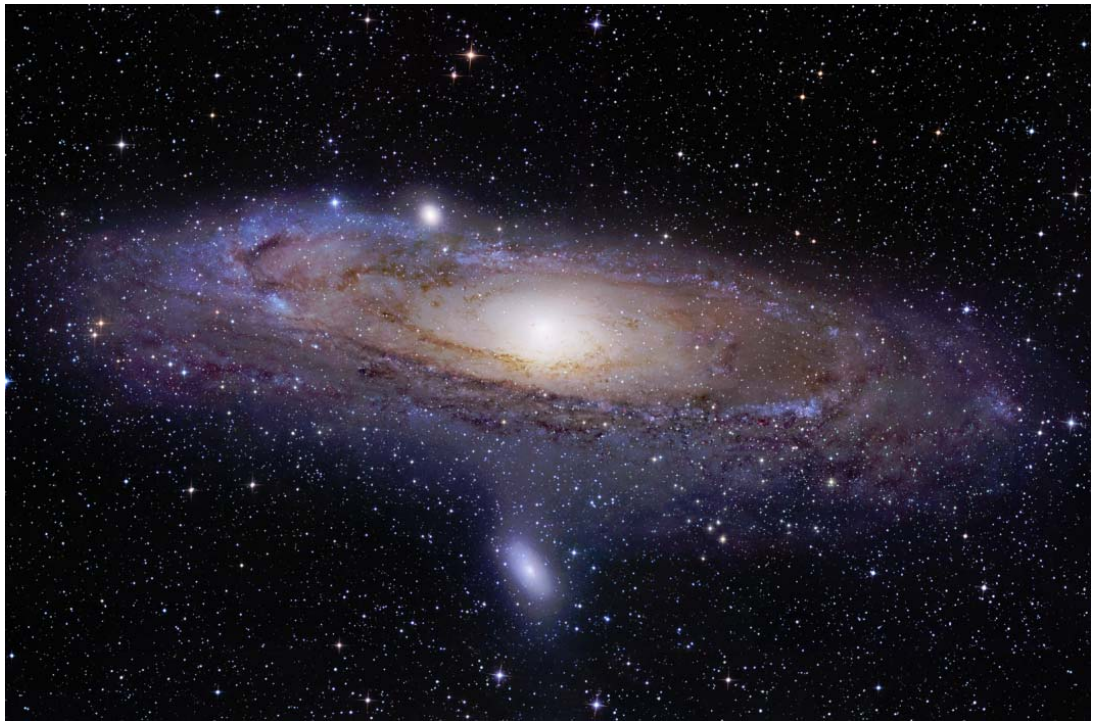
$$\arg\max_{ali}(f(ali(S1, S2))$$

Sequence Alignment: Enumerate?

```
LSPADK      L-SPADK     L-SPADK
LTPEEK      LTPEEK-     LT-PEEK
------LSPADK  L-S-P-A-D-K-
LTPEEK------  -L-T-P-E-E-K
```

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2}$$

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} = \frac{(2*300)!}{(300!)^2} \approx 7\times10^{88}$$



(Source: http://discoverystudentadventures.blogspot.com/2011/11/10-mind-bending-facts-about-universe.html)

The visible universe is estimated to contain 10^{78} ~ 10^{80} atoms (Source: wikianswers) only !

Sequence Alignment:
What is the computational **A**lgorithm?

MV-LSP
MVHLTP

```
HBA_HUMAN       1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D     48
                  || |:|.:|:.|.|.||||  :..|.|.|||.|:.:.:|.|:.:|..| |
HBB_HUMAN       1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD     48

HBA_HUMAN      49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR     93
                  ||     .|:.:||.|||||..|.::.:||:|::....:.||:||..||.
HBB_HUMAN      49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH     98

HBA_HUMAN      94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                  |||.||:||.:.|:..||.|...|||.|.|:..|.:|.|:..|..||.
HBB_HUMAN      99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
```

A residue can either

- Align to other residue, or
- Align to a gap

S	S	-
T	-	T

```
#=======================================
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBB_HUMAN
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:           9/149 ( 6.0%)
# Score: 292.5
#
#
#=======================================

HBA_HUMAN         1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D     48
                    || |:|.:|:.|.|.||||  :..|.|.|||.|:.:.:|.|:.:|..| |
HBB_HUMAN         1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD     48

HBA_HUMAN        49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR     93
                    ||     .|:.:||.|||||..|.::.:||:|::....:.||:||..||.
HBB_HUMAN        49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH     98

HBA_HUMAN        94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                    |||.||:||.:.|:..||.|...||||.|.|:..|.:|.|:..|..||.
HBB_HUMAN        99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
```

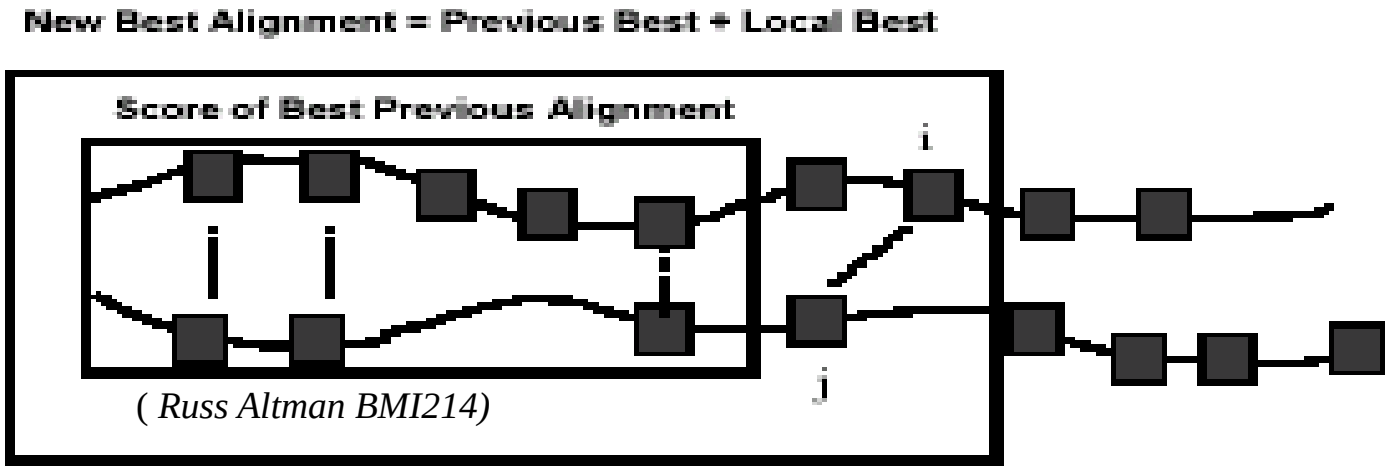
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1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

Affine gap penalty: opening a gap receives a penalty of d; extending a gap receives a penalty of e. So the total Penalty for a gap with length n would be:

Penalty = d + (n-1)* e

Final Score = (sum of substitution scores) + (-1) * (sum of Gap Penalty)

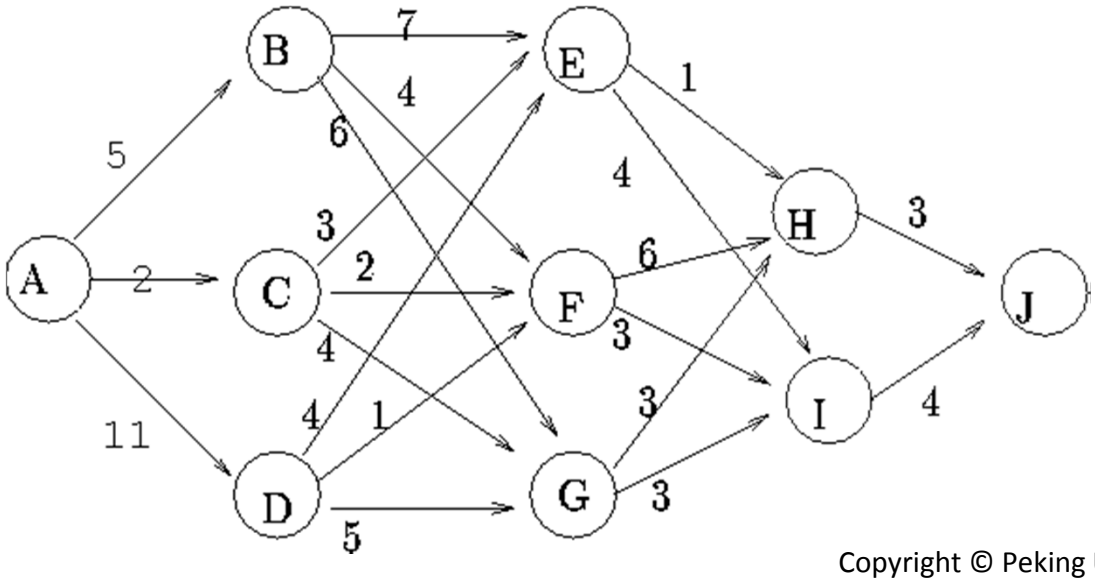
The best alignment that ends at a given pair of symbols is the best alignment of the sequences up to that point, plus the best alignment for the two additional symbols.

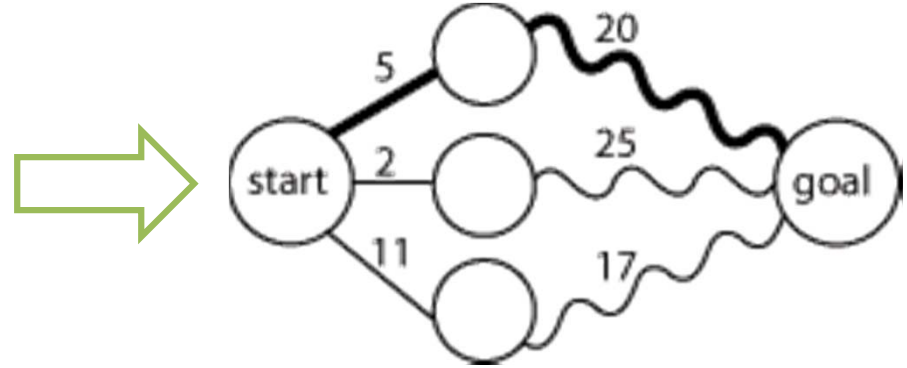


S	S	-
T	-	T

Dynamic Programming

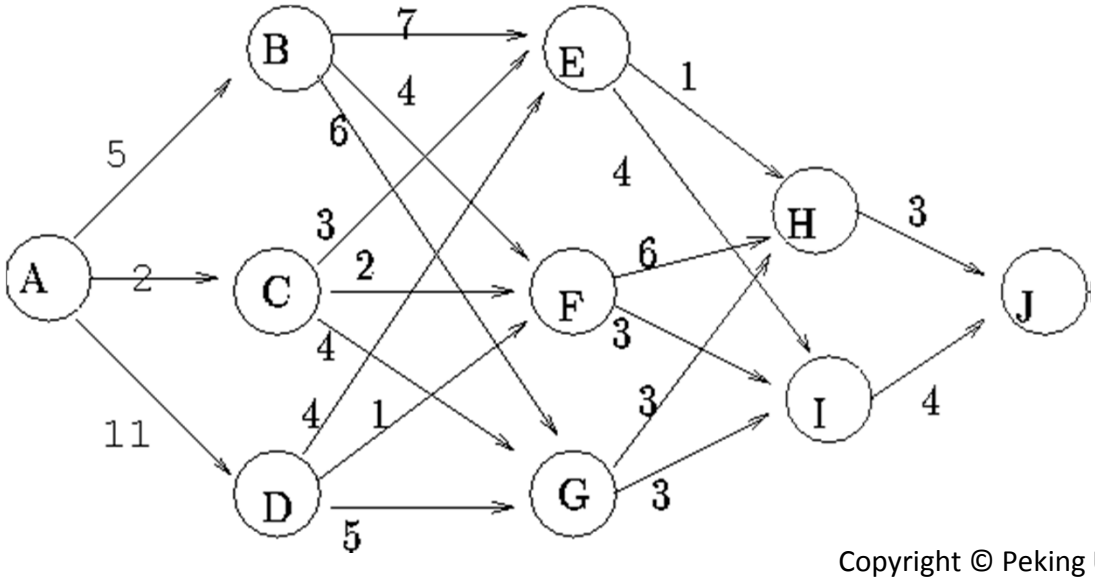
Dynamic Programming solves problems by combining the solutions to sub-problems.

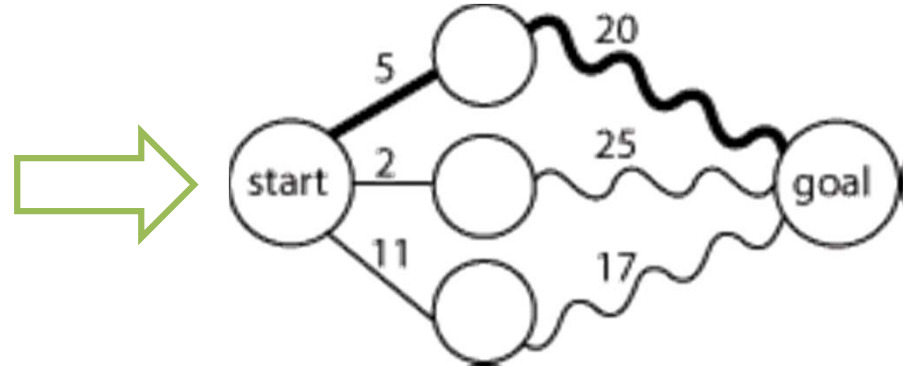




(Modified from Wikipedia)

1. Break the problem into smaller sub-problems.
2. Solve these sub-problems optimally recursively.
3. Use these optimal solutions to construct an optimal solution for the original problem.





(Modified from Wikipedia)

MV-LSP
MVHLTP

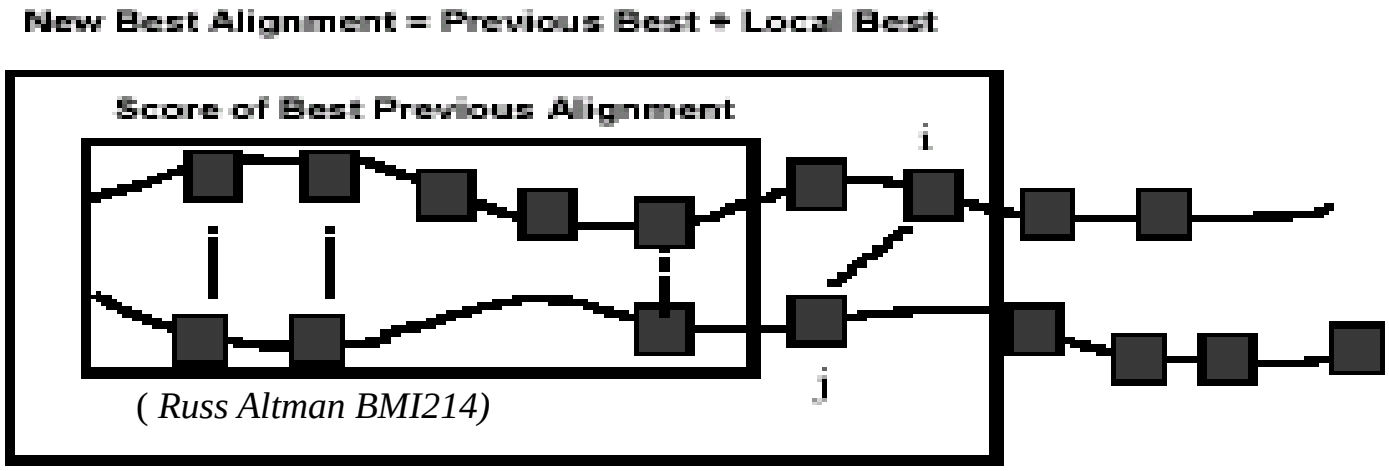
```
HBA_HUMAN       1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D     48
                  || |:|.:|:.|.|.||||  :..|.|.|||.|:.:.:|.|:.:|..| |
HBB_HUMAN       1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD     48

HBA_HUMAN      49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR     93
                  ||     .|:.:||.|||||..|.::.:||:|::....:.||:||..||.
HBB_HUMAN      49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH     98

HBA_HUMAN      94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                  |||.||:||.:.|:..||.|...|||.|.|:..|.:|.|:..|..||.
HBB_HUMAN      99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
```

A residue can either

- Align to other residue, or
- Align to a gap

S	S	-
T	-	T

The best alignment that ends at a given pair of symbols is the best alignment of the sequences up to that point, plus the best alignment for the two additional symbols.



S	S	-
T	-	T

Sequence alignment with Dynamic Programming: the Formula

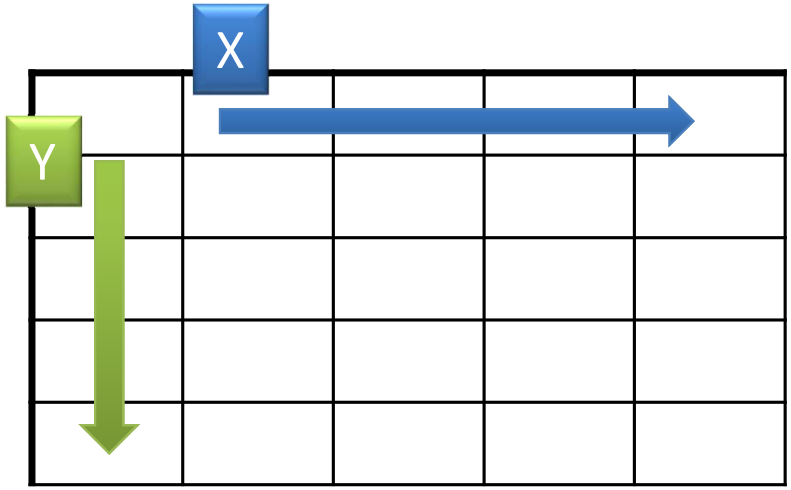
- Align two sequences: x and y
  - F (i,j) is the score of the best alignment between $x_{1\ldots i}$ and $y_{1\ldots j}$
  - s(A,B) is the score for substituting A with B; d is the (linear) gap penalty

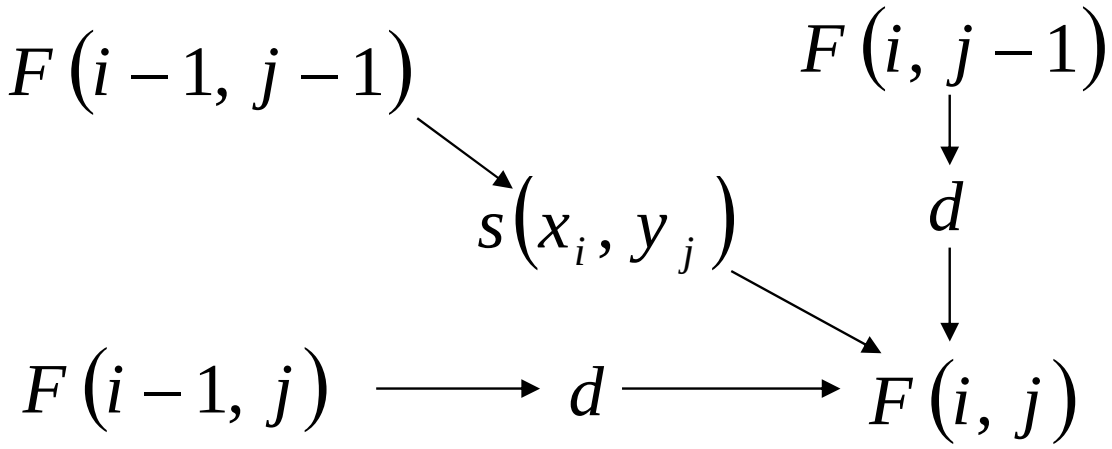
$$F(0,0) = 0$$

$$F(i,\ j) = \max \begin{cases} F(i-1,\ j-1) + s(x_i, y_j) & \textbf{x_i aligned to y_j} \\ F(i-1,\ j) + d & \textbf{x_i aligned to \textit{a gap}} \\ F(i,\ j-1) + d & \textbf{y_j aligned to \textit{a gap}} \end{cases}$$

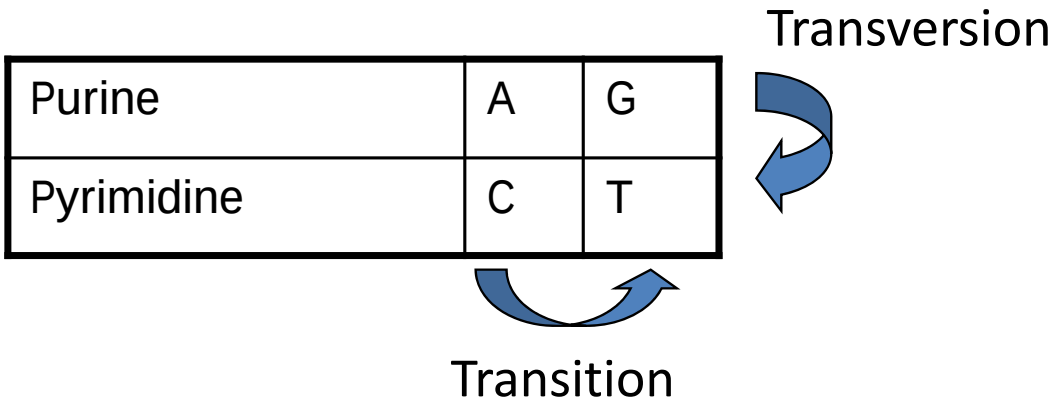
$$F(0,0) = 0$$

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) & \textbf{x_i aligned to y_j} \\ F(i-1, j) + d & \textbf{x_i aligned to \textit{a gap}} \\ F(i, j-1) + d & \textbf{y_j aligned to \textit{a gap}} \end{cases}$$





Scoring Nucleotide



A nucleotide substitution matrix:

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Input Sequence 1: AAG

Input Sequence 2: AGC

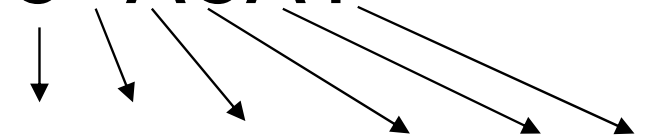
	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

For simplicity, let's set (i.e. linear gap penalty)

gap OPEN (d) = gap EXTEND (e) = -5

GAC - AT

C - ACAT



(-7) + (-5) + (-7) + (-5) + 2 + 2= -20

Dynamic Programming Matrix

		A	A	G
A				
G				
C				

DP for sequence alignment: Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

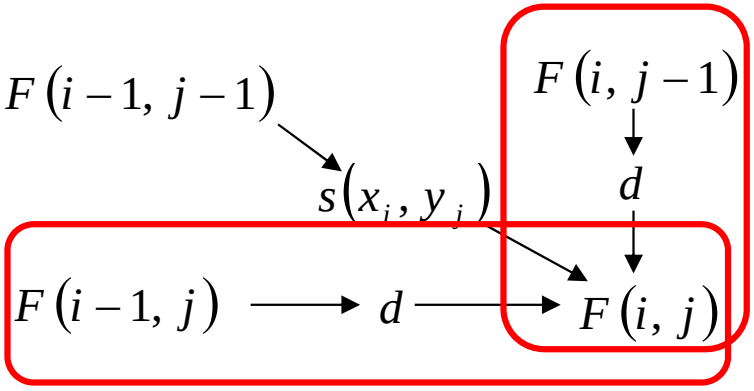
		A	A	G
	0			
A				
G				
C				

$$F(0,0) = 0$$

$$F(i,j) = \max\begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) + d \\ F(i,j-1) + d \end{cases}$$

DP for sequence alignment: Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

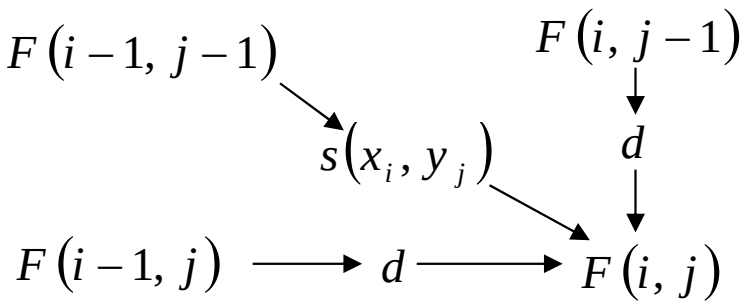


Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

		A	A	G
	0 →	-5 →	-10 →	-15
A	-5			
G	-10			
C	-15			

DP for sequence alignment: Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2



Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

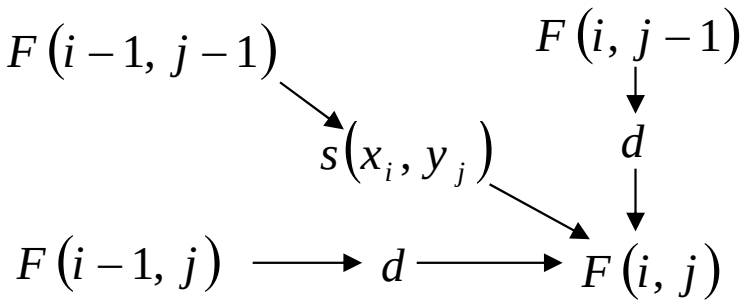
		A	A	G
	0	-5	-10	-15
A	-5	2	-3	-8
G	-10	-3	-3	-1
C	-15	-8	-8	**-6**

DP for sequence alignment: Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

		A
	0 →	-5
A	-5	2

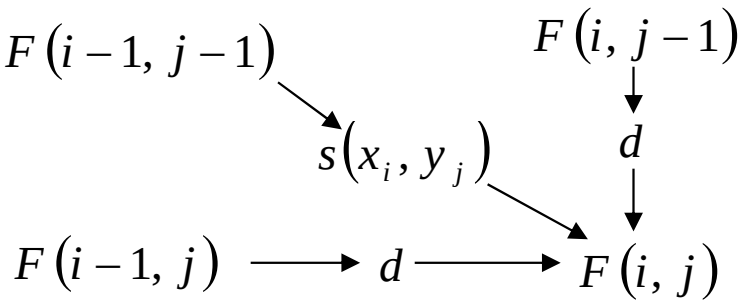


DP for sequence alignment: Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

		A
	0 →	-5
A	-5 →	2



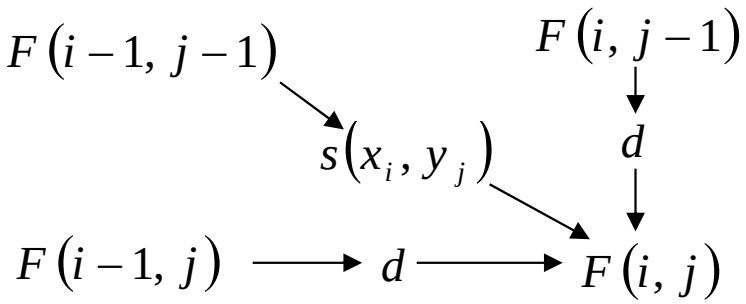
-5 + (-5) = -10

0 + 2 = 2

-5 + (-5) = -10

DP for sequence alignment: Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

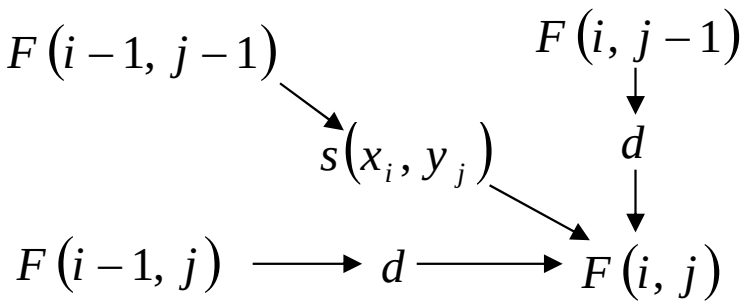


Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

		A	A	G
	0 →	-5 →	-10 →	-15
A	-5 ↓	2 →	-3 →	-8
G	-10 ↓	-3 ↓	-3 ↓	-1 ↓
C	-15	-8	-8	**-6**

DP for sequence alignment: Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2



Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

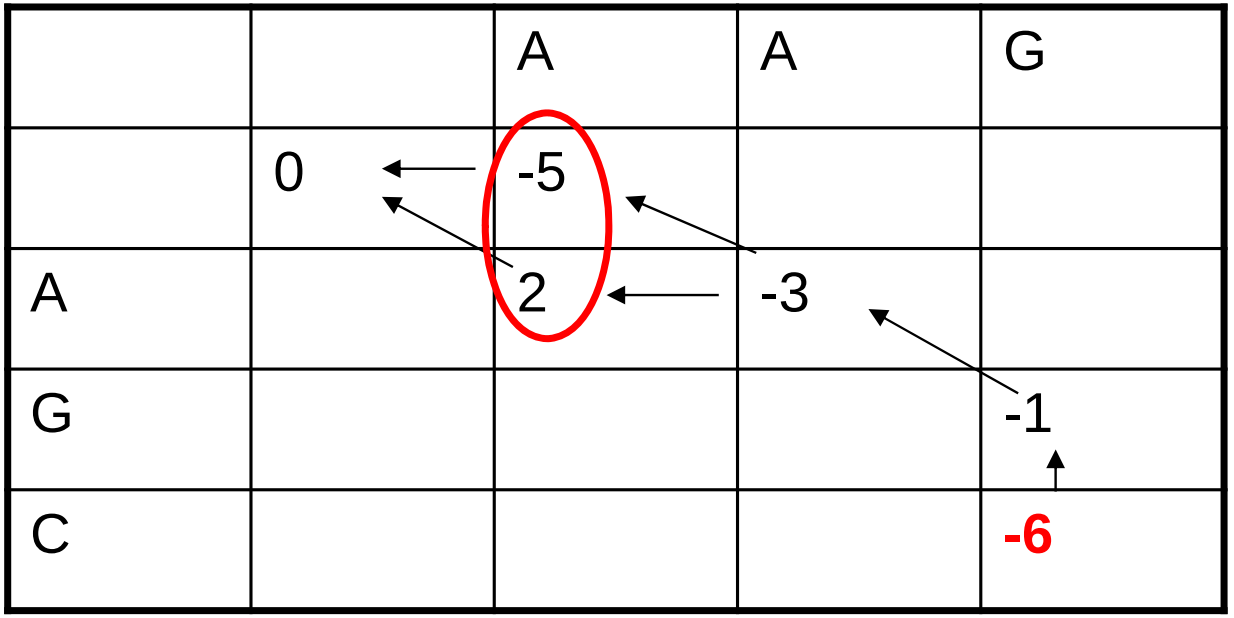
		A	A	G
	0 →	-5 →	-10 →	-15
A	-5 ↓	2 →	-3 →	-8
G	-10 ↓	-3 ↓	-3 ↓	-1 ↓
C	-15	-8	-8	**-6**

Traceback: Decode the “Real” Alignment

- Trace back to the upper left. Each arrow introduces one symbol at the end of each aligned sequence.

A A G -
- A G C

A A G -
A - G C



Summary Questions

- Why can we build pairwise alignment by dynamics programming algorithm?
  - Is there any assumption(s)?
- Do you think these assumptions reasonable?
  - Why?

生物信息学：导论与方法
Bioinformatics: Introduction and Methods