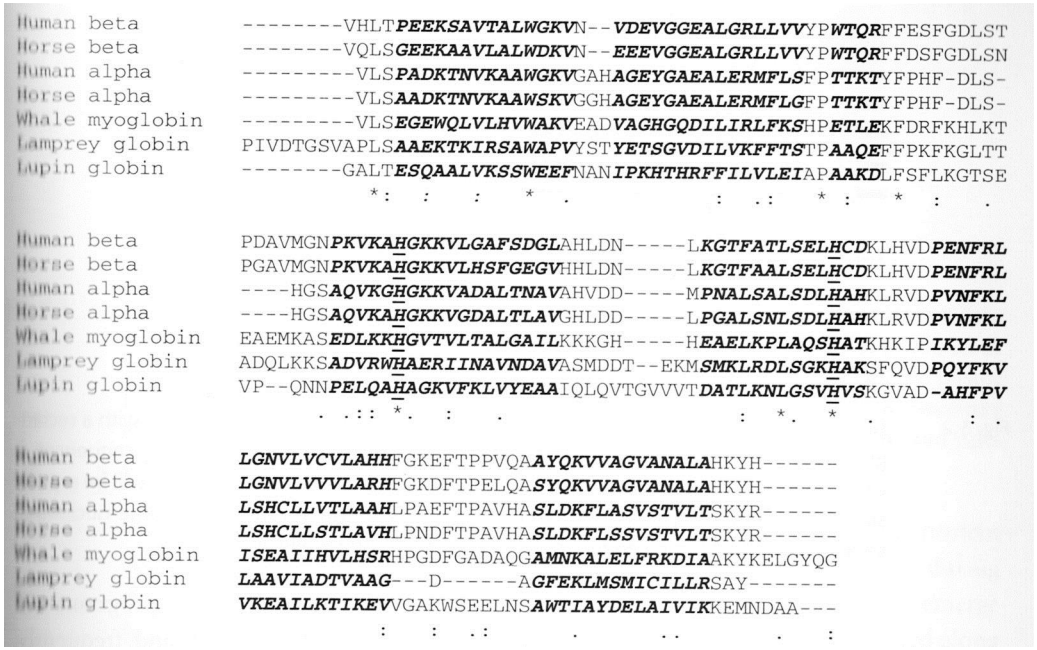
Introduction to Phylogenetics
Week 3

Alignment

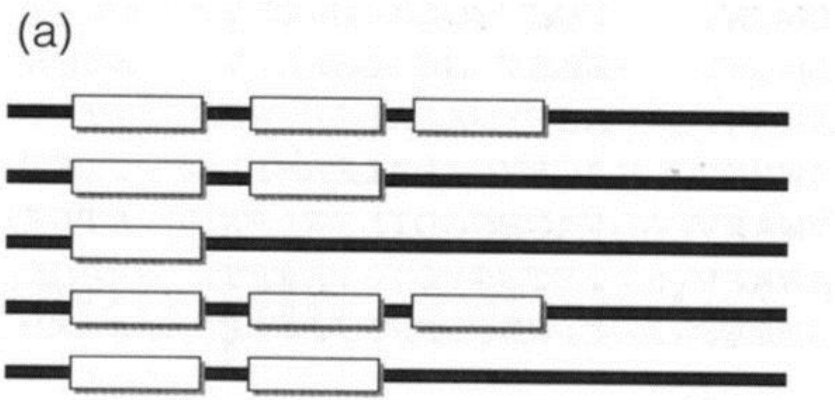
II. Alignment

- Protein allows for homology in alignment
  - More information for accurate alignment
  - * = identical positions
  - : = biochemically conserved
  - . = less conserved
  - = no homology



II. Alignment

- DNA only has 4 bases
    - Can minimize issues with structural sequences
    - - = gaps
    - . = missing data
- Repeats can cause
  big problems



(b)

```
accgtacc--gtaccgt
accgtaccccgtaccgt
accgtac---gtaccgt
accgtacac-gtaccgt
******  *******
```

II. Alignment

- Substitutions
  - The fewer, the easier to register the alignment
  - Polarity can weight substitutions
  - When identity drops <25%, difficult to solve
  - Functionally constrained regions help
  - Can drive regions of ambiguity
  - Likelihood substitutions – weight alignment

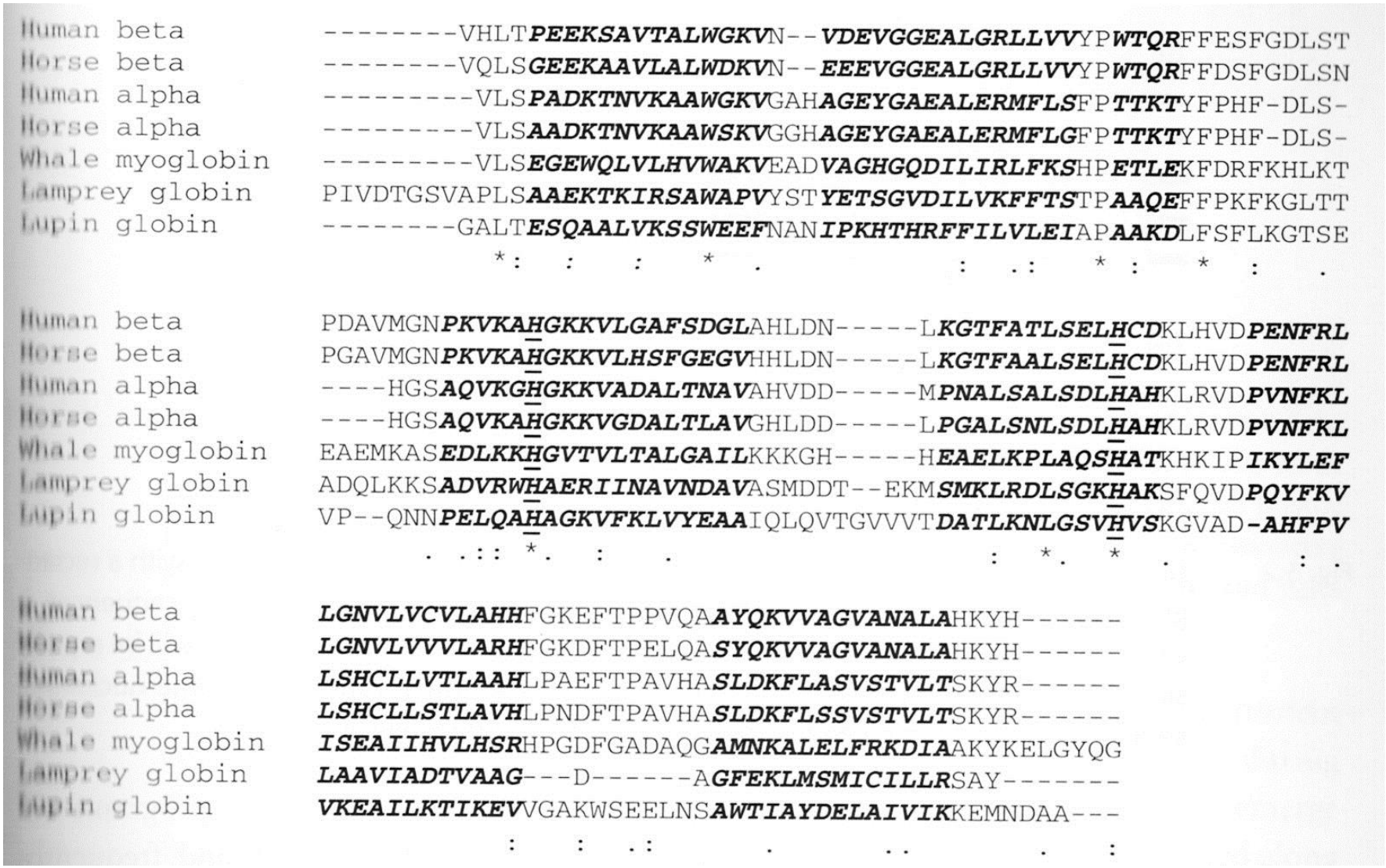
BLOSUM62

(Henikoff and Henikoff, 1992)

- Weights amino acid alignments
- Change regularly during evolution <n
- Uncommon change >n

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

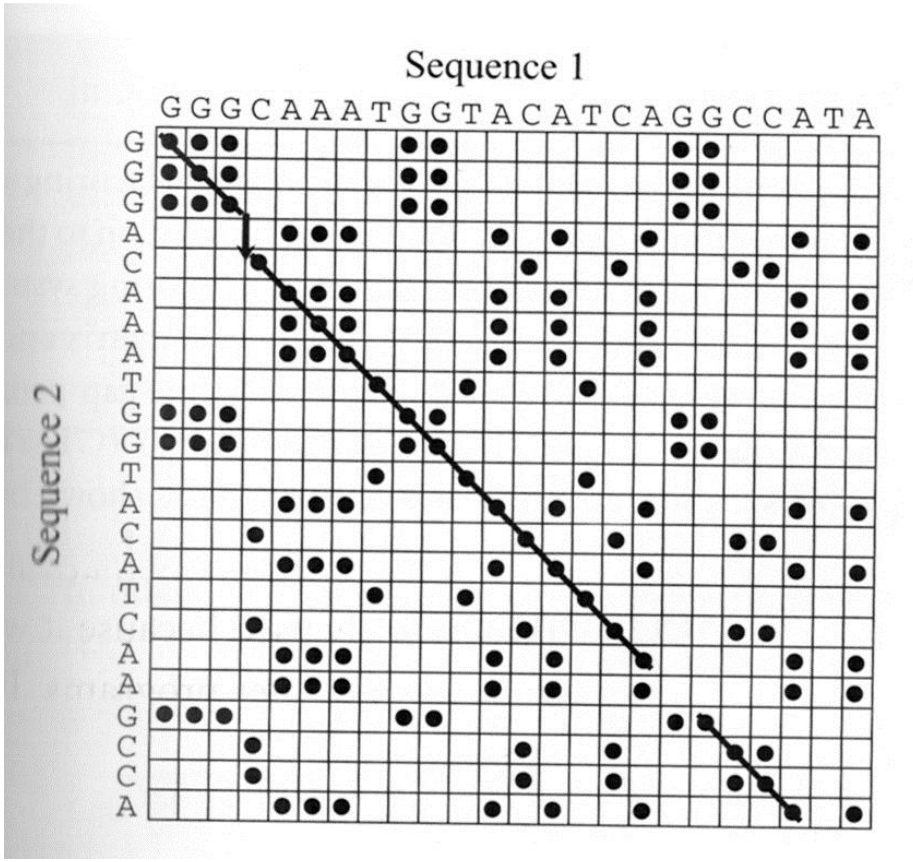
Aligned sequences need to MAXIMIZE this score


Human beta
Horse beta
Human alpha
Horse alpha
Whale myoglobin
Lamprey globin
Lupin globin
Human beta
Horse beta
Human alpha
Horse alpha
Whale myoglobin
Lamprey globin
Lupin globin
Human beta
Horse beta
Human alpha
Horse alpha
Whale myoglobin
Lamprey globin
Lupin globin

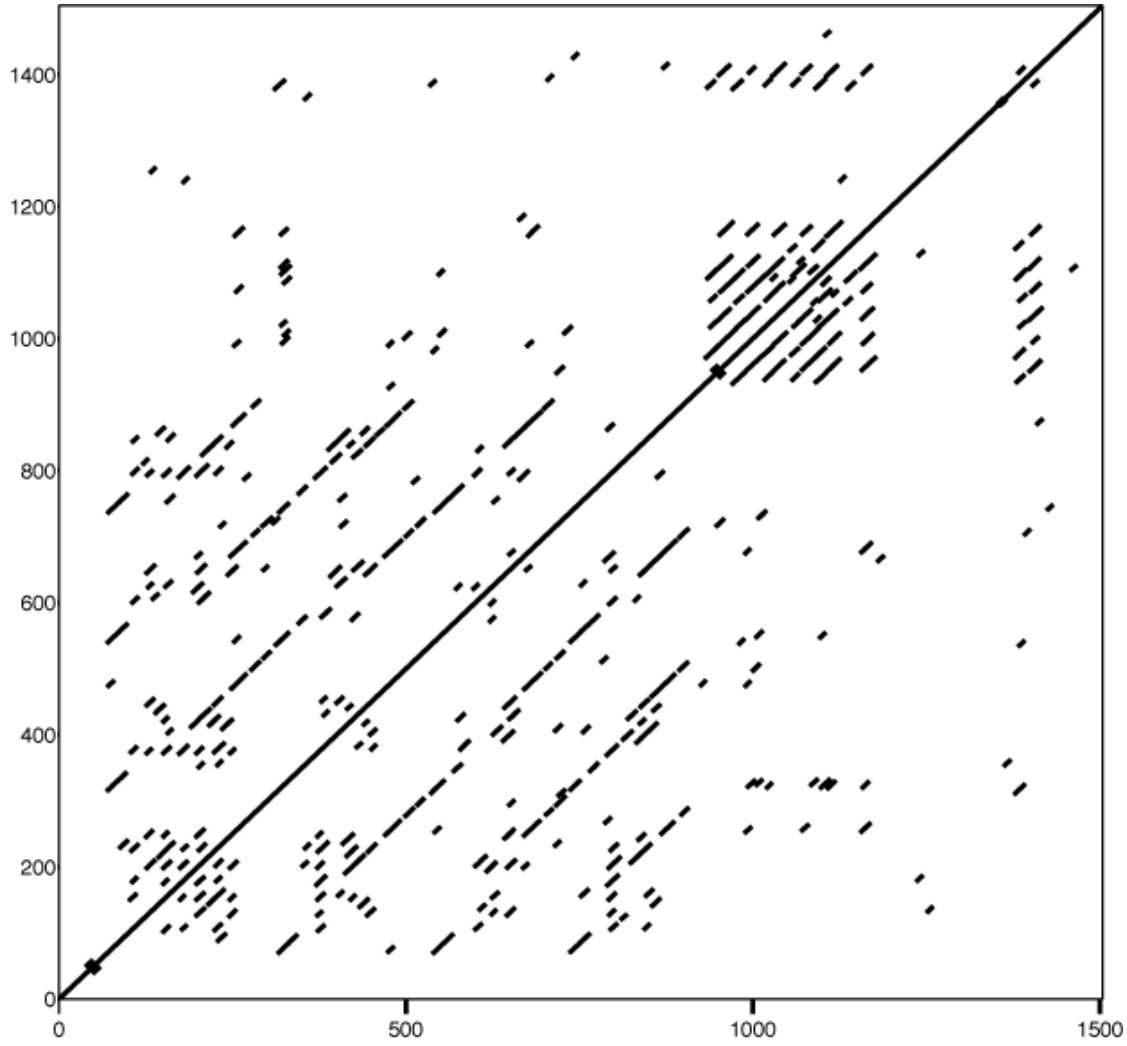
III. Gaps

- Tend to occur in regions of less evolutionary conservation
  - Usually the result of insertions/deletions (indels)
  - Inclusion of gaps makes it difficult to align
  - Requires pairwise sequence alignment
  - Possible to generate more gaps than sequence

III. Pairwise Sequence Alignment




```
GGG-CAAATGGGACAGCAGGCCATA
||| ||||||| |||  | ||||
GGGACAAATGGTACATGAAGCCA--
```


1400
1200
1000
800
600
400
200
0
0
500
1000
1500

A

FFESFGDLSTPDAVMGN
YFPHDLSHGS

FFDSFGDLSNPGAVMGN
YFPHF-DLS-----HGS

B

FFESFGDLSTPDAVMGN
LPNDFTPAVHA

FFESFGDLSTPDAVMGN---
-L--PND-FTP-AV---HA

III. Gaps

- To prevent excessive gaps – use gap penalties (GPs).

$$GP = g + e(l - 1)$$

- l = length of gap
- g = gap opening penalty
- e = gap-extension penalty
- end-gaps are generally free

(you get to decide values)

III. Dynamic programming

- Scores all possible pair alignments
- Penalizes gaps
- Bellman's principle of optimality – any sub-solution of optimal solution is a solution

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(X_i, Y_i), \\ F(i-1,j) - g, \\ F(i,j-1) - g \end{cases}$$

Generate substitution matrix (F)

j \ *i*		1	2	3	4	5	6	7	8
		*	G	R	Q	T	A	G	L
1	*								
2	G								
3	T								
4	A								
5	Y								
6	D								
7	L								

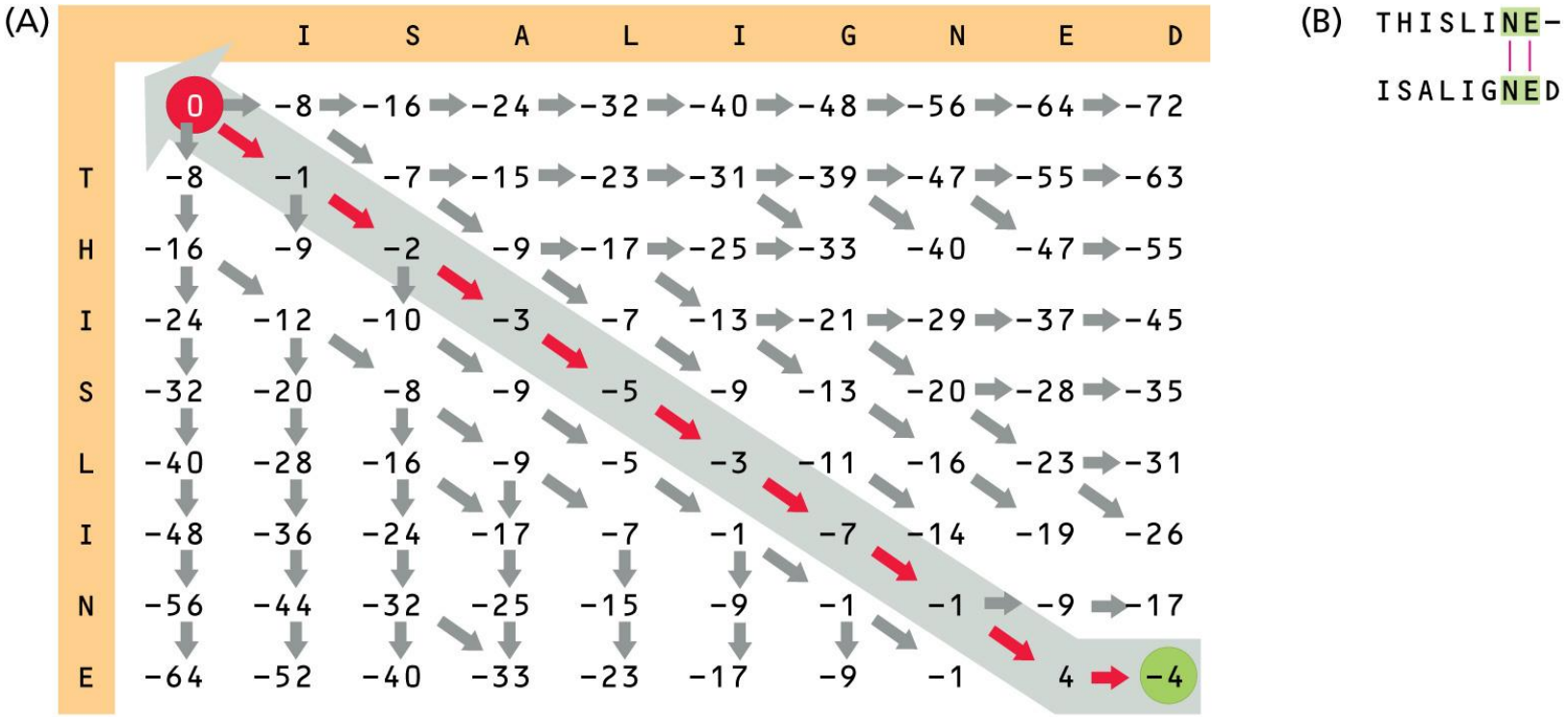
gap penalty (g) = -8

BLOSUM62 substitution values

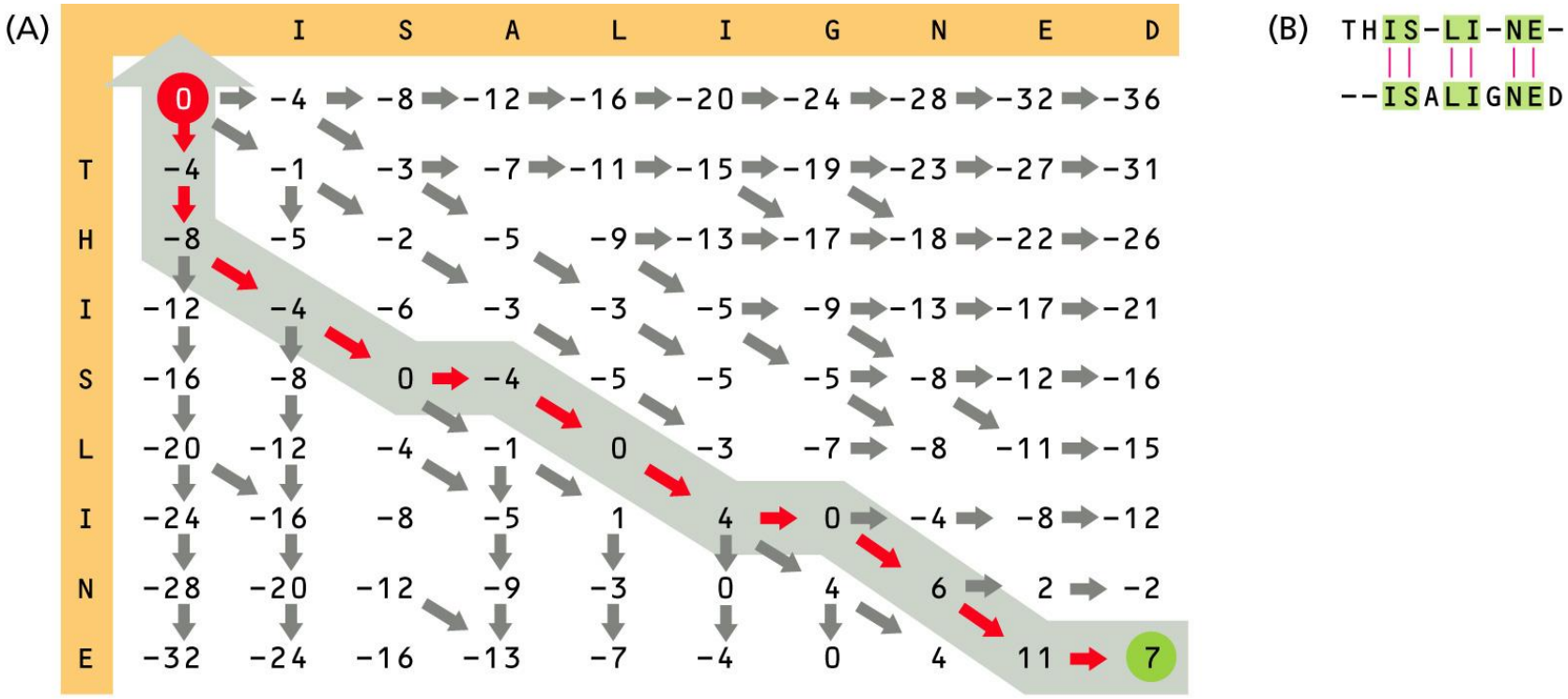
$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(X_i, Y_i), \\ F(i-1, j) - g, \\ F(i, j-1) - g \end{cases}$$

j \ i		1	2	3	4	5	6	7	8
		*	G	R	Q	T	A	G	L
1	*								
2	G								
3	T								
4	A								
5	Y								
6	D								
7	L								

g = -8


(A)
I S A L I G N E D
0 -8 -16 -24 -32 -40 -48 -56 -64 -72
T -8 -1 -7 -15 -23 -31 -39 -47 -55 -63
H -16 -9 -2 -9 -17 -25 -33 -40 -47 -55
I -24 -12 -10 -3 -7 -13 -21 -29 -37 -45
S -32 -20 -8 -9 -5 -9 -13 -20 -28 -35
L -40 -28 -16 -9 -5 -3 -11 -16 -23 -31
I -48 -36 -24 -17 -7 -1 -7 -14 -19 -26
N -56 -44 -32 -25 -15 -9 -1 -1 -9 -17
E -64 -52 -40 -33 -23 -17 -9 -1 4 -4
(B)
THISLINE-
ISALIGNED

g = -4


(A)
I S A L I G N E D
0 -4 -8 -12 -16 -20 -24 -28 -32 -36
T -4 -1 -3 -7 -11 -15 -19 -23 -27 -31
H -8 -5 -2 -5 -9 -13 -17 -18 -22 -26
I -12 -4 -6 -3 -3 -5 -9 -13 -17 -21
S -16 -8 0 -4 -5 -5 -5 -8 -12 -16
L -20 -12 -4 -1 0 -3 -7 -8 -11 -15
I -24 -16 -8 -5 1 4 0 -4 -8 -12
N -28 -20 -12 -9 -3 0 4 6 2 -2
E -32 -24 -16 -13 -7 -4 0 4 11 7
(B)
THIS-LI-NE-
--ISALIGNED

Gap penalty (g) = -8

Gap extention (e) = -2

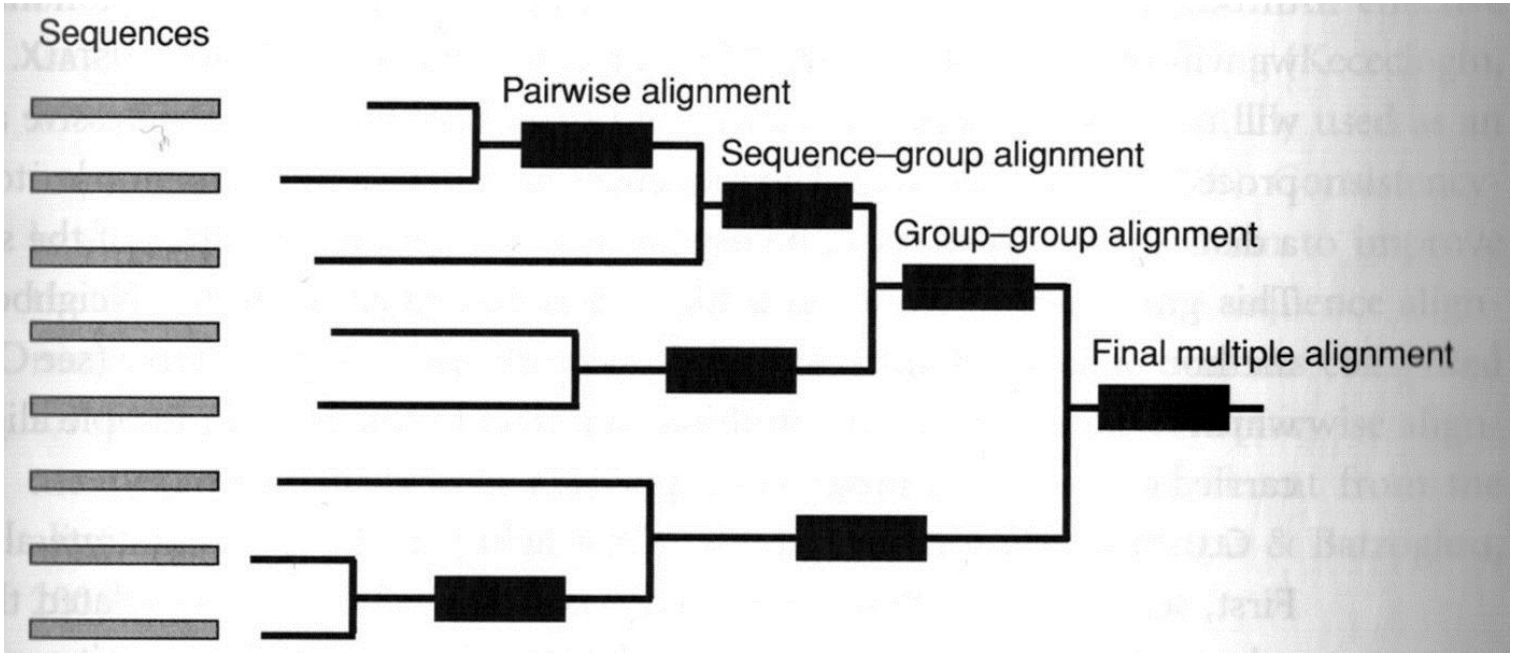
$$GP = g + e(l - 1)$$

j \ i		1	2	3	4	5	6	7	8
		*	G	R	Q	T	A	G	L
1	*	0	-8	-10	-12	-14	-16	-18	-20
2	G	-8	6	-2	-4	-6	-8	-10	-12
3	T	-10	-2	5	-3	1	-6	-10	-11
4	A	-12	-4	-3	4	-3	5	-3	-11
5	Y	-14	-6	-5	-4	2	-3	2	-4
6	D	-16	-8	-7	-5	-3	0	-4	-2
7	L	-18	-10	-9	-9	-6	-4	-4	0

III. Progressive Alignments

- Generate phylogenetic tree
- Use tree in progressive alignment (heuristic)
- Dramatically speeds up math
- Tree generated from pair-wise alignment (Neighbor Joining)

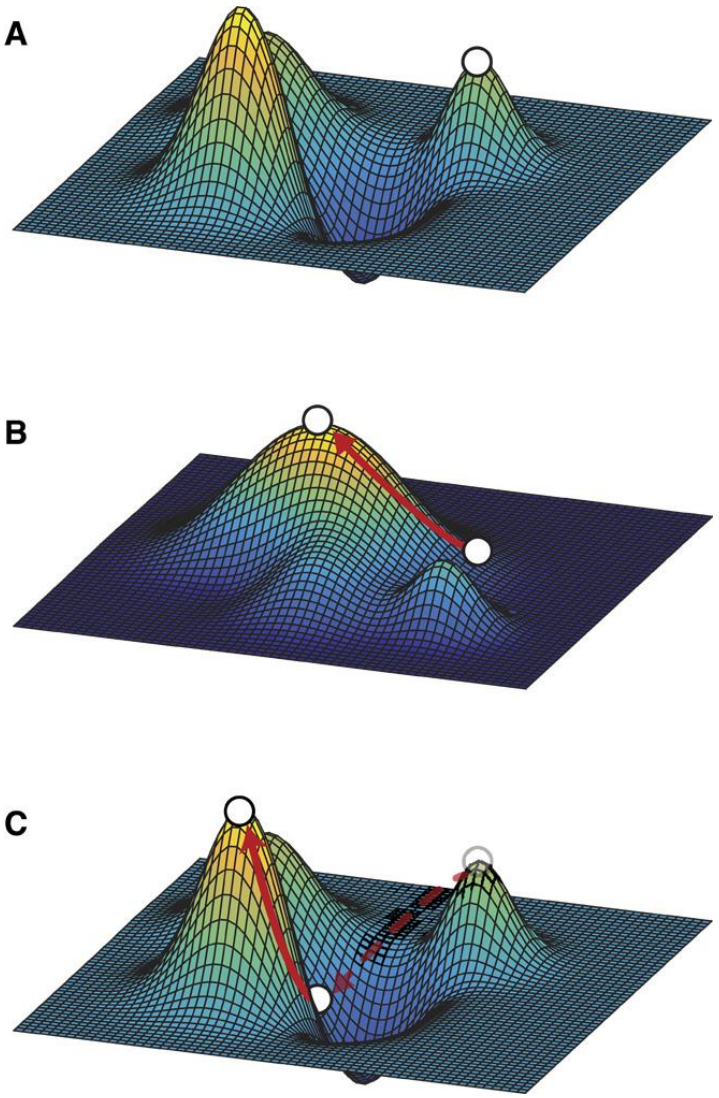
III. Progressive Alignments



IV. Programs: CLUSTALX

- Difference between X/W interface
- Progressive alignment based on un-rooted NJ tree
- Reduce weights on aligned sequences
- Uses BLOSUM or PAM
- Varies GP values, including position-specific GP values

The local minimum problem



IV. Programs: T-COFFEE

- Tree-based consistency objective function for alignment evaluation
- Consistency based scoring
- Relies more on initial weighted sum of pairs (WSP) values (A-B, B-C vs. A-C)
- Uses different algorithms for alignment
- Provides alignment that’s most consistent

IV. Programs: MUSCLE

- Uses fast distance estimations
- Relies on K_{mer} values
- Provides progressive alignment
- Uses tree for iterative alignments

IV. Comparing alignments

- Visual alignment
- Comparative programs such as ALTAVIST
- Edit alignments for tree building
  - Within CLUSTALX
  - BIOEDIT (DOS)
  - SE-AL (MacOS)