

# Introduction to Phylogenetics

## Week 3

# Alignment

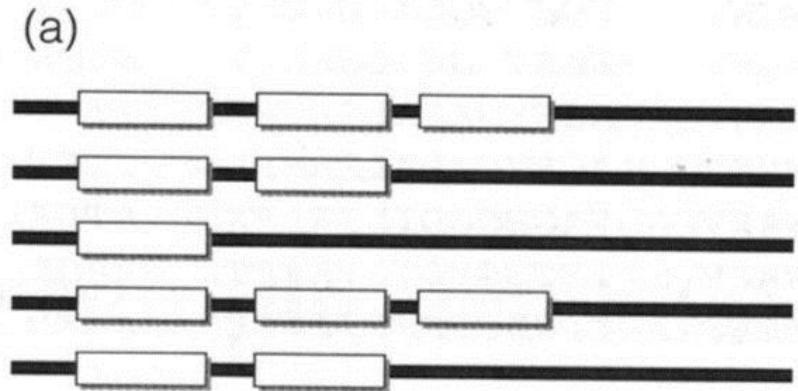
# II. Alignment

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

Human beta	-----VHLTPEEKSAVTALWGKVNV--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta	-----VQLSGEKAQVLAALWDKVN--EEVVGGEALGRLLVVYPWTQRFFDSFGDLST
Human alpha	-----VLSPADKTNVKAAWGKVGAGHAGEYGAELERMFLSFPTTKTYFPHF-DLS-
Horse alpha	-----VLSAADKTNVKAWSKVGGHAGEYGAELERMFLGFPPTTKTYFPHF-DLS-
Whale myoglobin	-----VLSGEGWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDREKHLKT
Lamprey globin	PIVDTGTSVAPLSAAEKTIRSAWAPVYSTYETSGVDILVKFFTSPPAAQEFPPKFKGLTT
Lupin globin	-----GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
	* : : * . : : * : * : .
Human beta	PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta	PGAVMGNPKVKAHGKKVLHSGEGVHHLDN-----LKGTFALSELHCDKLHVDPENFRL
Human alpha	----HGSAQVKGHGKKVADALTNVAHVDD----MPNALSALSDLHAHKLKRVDPVNFKL
Horse alpha	----HGSAQVKAHGKKVGDALTLAVGHLD----LPGALSNDLSDLHAHKLKRVDPVNFKL
Whale myoglobin	EAEMKASEDLKKHGVTVLTALGAILKKKGH----HEAELKPLAQSHATKKKIP IKYLEF
Lamprey globin	ADQLKKSADVVRWHAERIINAVNDVAVSMDDT--EKMSMKLRDLGKHAQSFQVDPQYFKV
Lupin globin	VP--QNNPELQAAGAGKVFVKLVYEAIIQLQVTGVVVTDATLKNLGSVHVSQGVAD-AHFPV
	. : : * . : . : * . * . : .
Human beta	LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH-----
Horse beta	LGNVLVVVLARHFGKDFTPPELQASQYQKVVAGVANALAHKYH-----
Human alpha	LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR-----
Horse alpha	LSHCLLVTLAVHLPNDFTPAVHASLDKFLSSVSTVLTISKYR-----
Whale myoglobin	ISEAIIHVLHSRHPGDFGADAQGMNKALELFRKDIAAKYKELGYQG
Lamprey globin	LAAVIADTVAAG--D-----AGFEKLMSMICILLRSAY-----
Lupin globin	VKEAIIKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
	: : : . . . : :

# II. Alignment

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



(b)

accgtacc	--	gtaccgt
accgtaccc	g	taccgt
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)

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
Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr Val																				

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

Aligned sequences need to MAXIMIZE this score

Human beta	-----VHLT <b>PEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST</b>
Horse beta	-----VQLS <b>GEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLNS</b>
Human alpha	-----VLS <b>PADKTNVKAAWGKVGHAAGEYGAEALERMFLSFPTTKTYFPHF-DLS-</b>
Horse alpha	-----VLS <b>AADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-</b>
Whale myoglobin	-----VLS <b>EGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFD</b> DRFKHLKT
Lamprey globin	PIVDTGSVAPLS <b>AAEKT</b> KIRSAWAPVYST <b>YETSGVDILVKFFTSTPAAQ</b> EFFPKFKGLTT
Lupin globin	-----GALT <b>ESQAALVKSSWEEFNANIPKH</b> THRFFILVLEIAP <b>AAKDL</b> FSFLKGTSE
	* :       :       * .                               :   .:       * :       * :       .

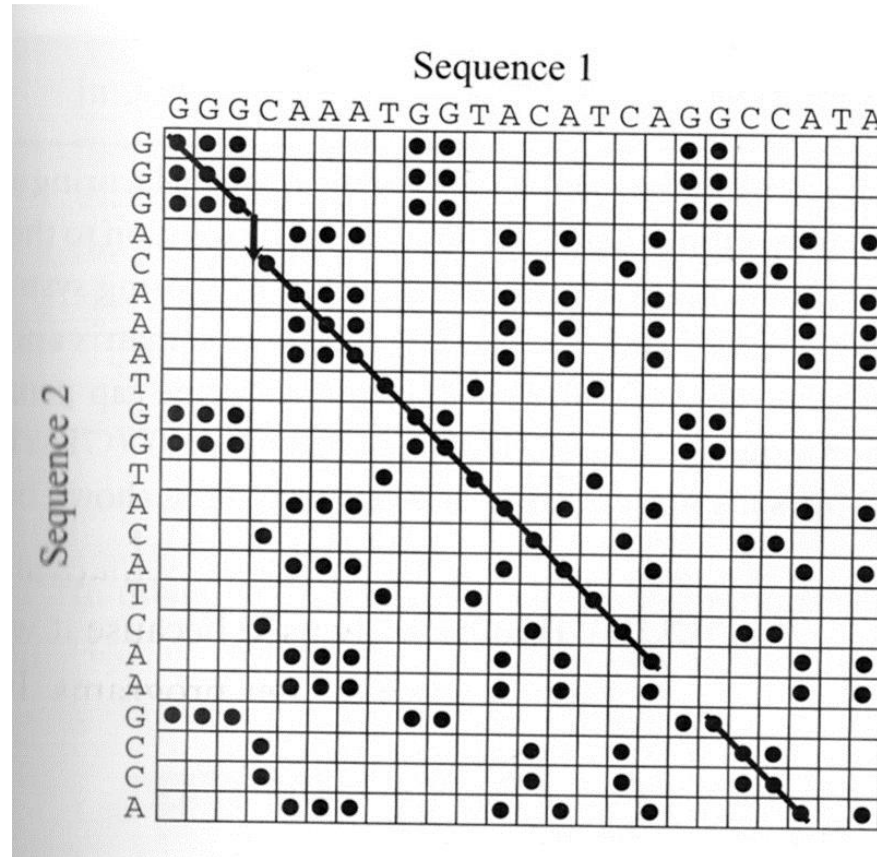
Human beta	PDAVMGN <b>PKVKAHGKKVLGAFSDGLA</b> HLDN----- <b>LKGT</b> FATLSEL <b>HCD</b> KLHVD <b>PENFRL</b>
Horse beta	PGAVMGN <b>PKVKAHGKKVLHSFG</b> EGVHHLDN----- <b>LKGT</b> FAALSEL <b>HCD</b> KLHVD <b>PENFRL</b>
Human alpha	----HGSA <b>QVKGHGKKVADAL</b> TNAVAVHDD----- <b>MPNAL</b> SALS <b>SDL</b> <b>H</b> AHKL RVD <b>PVNFKL</b>
Horse alpha	----HGSA <b>QVKAHGKKVGDAL</b> TLAVGHLDD----- <b>LP</b> GALSNLS <b>SDL</b> <b>H</b> AHKL RVD <b>PVNFKL</b>
Whale myoglobin	EAEMKAS <b>EDLKKHGVT</b> VL <b>TALGAIL</b> KKKGH----- <b>HEAEL</b> KPLA <b>QSH</b> ATKHKIP <b>IKYLEF</b>
Lamprey globin	ADQLKKS <b>ADV</b> R <b>WHAERI</b> INAVND <b>AVAS</b> MDDT--EKM <b>SMKLRDL</b> SG <b>KHAKS</b> FQVD <b>PQYFKV</b>
Lupin globin	VP--QNN <b>PELQA</b> HAGKVFKLVY <b>EAAI</b> QLQVTGVVVT <b>DATL</b> KNLGS <b>VHV</b> SKGVAD- <b>AHFPV</b>
	. .:: * .       :       .                               :   * .       *       .       : .

Human beta	<b>LGNVLVCVLAHHF</b> GKEFTPPVQ <b>AA</b> YQKV <b>VAGVANALAH</b> KYH-----
Horse beta	<b>LGNVLVVVLARHF</b> GKDFTPELQ <b>AS</b> YQKV <b>VAGVANALAH</b> KYH-----
Human alpha	<b>LSHCLLVTLAAH</b> LPAEFTPAVHA <b>SLDKFLASVSTVL</b> TSKYR-----
Horse alpha	<b>LSHCLLSTLAVH</b> LPNDFTPAVHA <b>SLDKFLSSVSTVL</b> TSKYR-----
Whale myoglobin	<b>ISEAIIHVLHSR</b> H <b>PGDFGADAQ</b> G <b>AMNKALELFRKDIA</b> AKYKELGYQ <b>G</b>
Lamprey globin	<b>LA</b> AVIAD <b>TVAAG</b> ---D----- <b>AGFEKL</b> MS <b>MICILL</b> RSAY-----
Lupin globin	<b>VKEAILKTIKEV</b> VGAKWSEELNS <b>AW</b> TIAY <b>DELAIVIK</b> KEMNDAA---
	:       :       .:       .       .       .       :

# III. Gaps

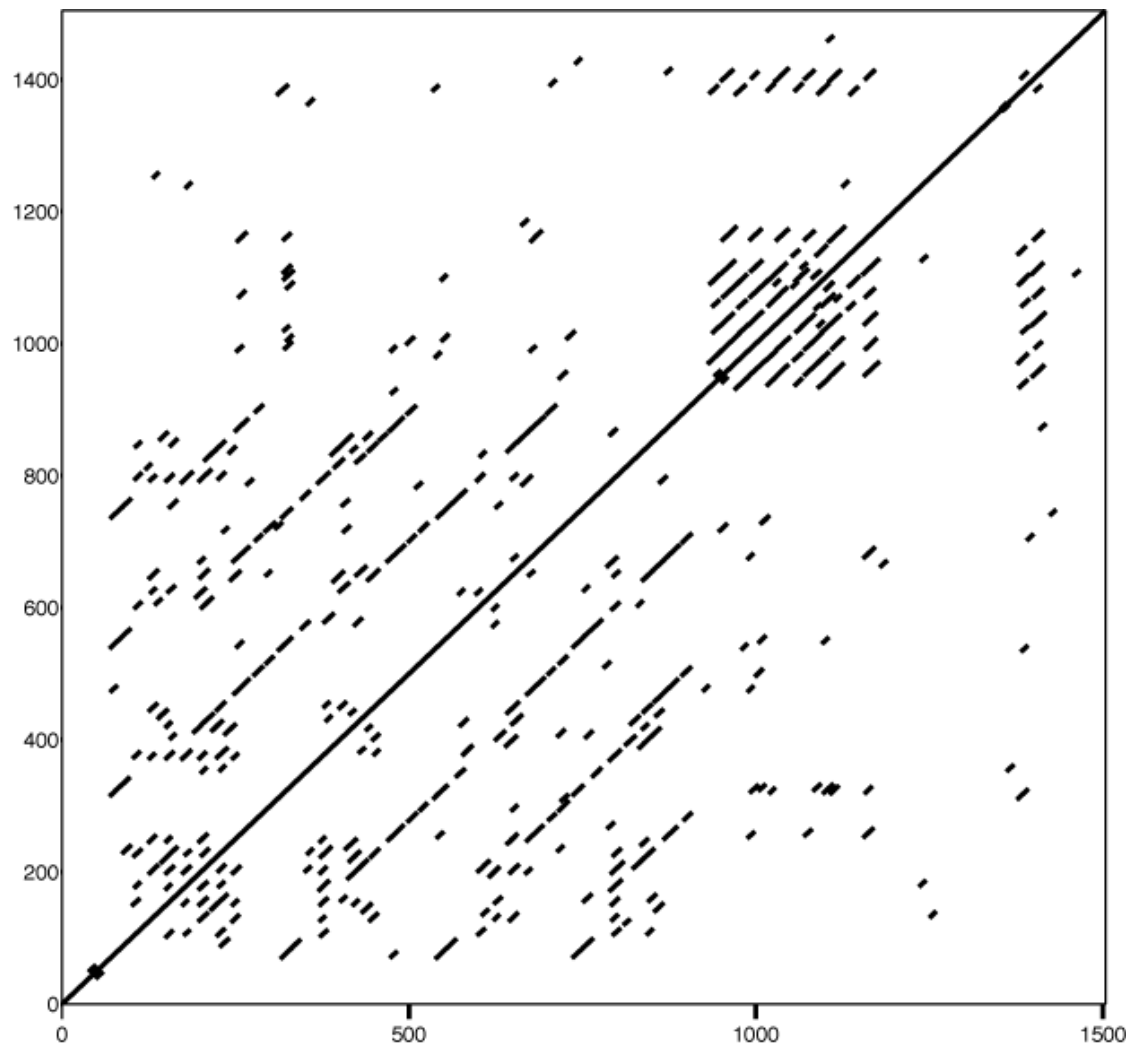
- Tend to occur in regions of less evolutionary conservation
  - Usually the result of insertions/deletions (indels)
  - Inclusion of gaps makes it difficult to align
  - Requires pairwise sequence alignment
  - Possible to generate more gaps than sequence

# III. Pairwise Sequence Alignment



GGG-CAAATGGGACAGGCCATA  
||| ||||| ||| | |||  
GGGACAAATGGTACATGAAGCCA--





**A**

FFESFGDLSTPDAVMGN  
YFPHDLSHGS

FFDSFGDLSNPGAVMGN  
YFPHF-DLS-----HGS

**B**

FFESFGDLSTPDAVMGN  
LPNDFTP AVHA

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)

<div><div><div><div></div><div><i>j</i></div></div><div><div><i>i</i></div><div></div></div></div></div>		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								

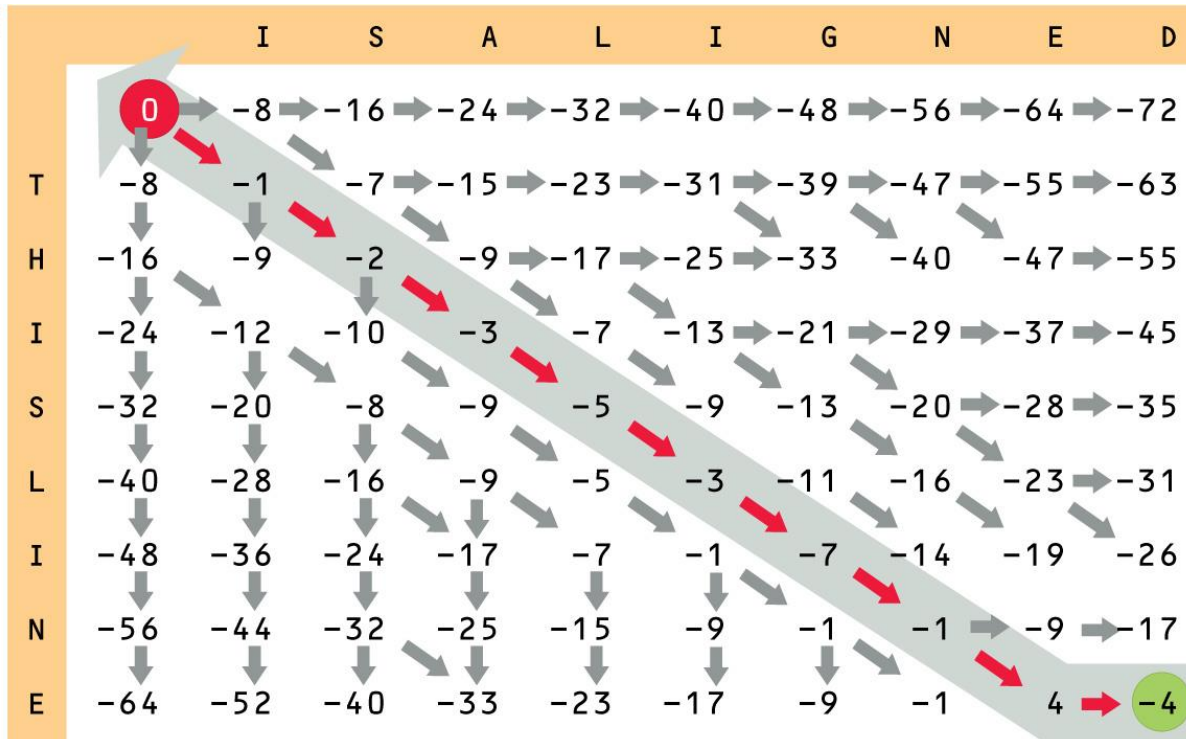
gap penalty (g) = -8  
BLOSUM62 substitution values

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

<div style="display: inline-block; transform: rotate(-45deg);"> <i>i</i>  <i>j</i> </div>		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								

$$g = -8$$

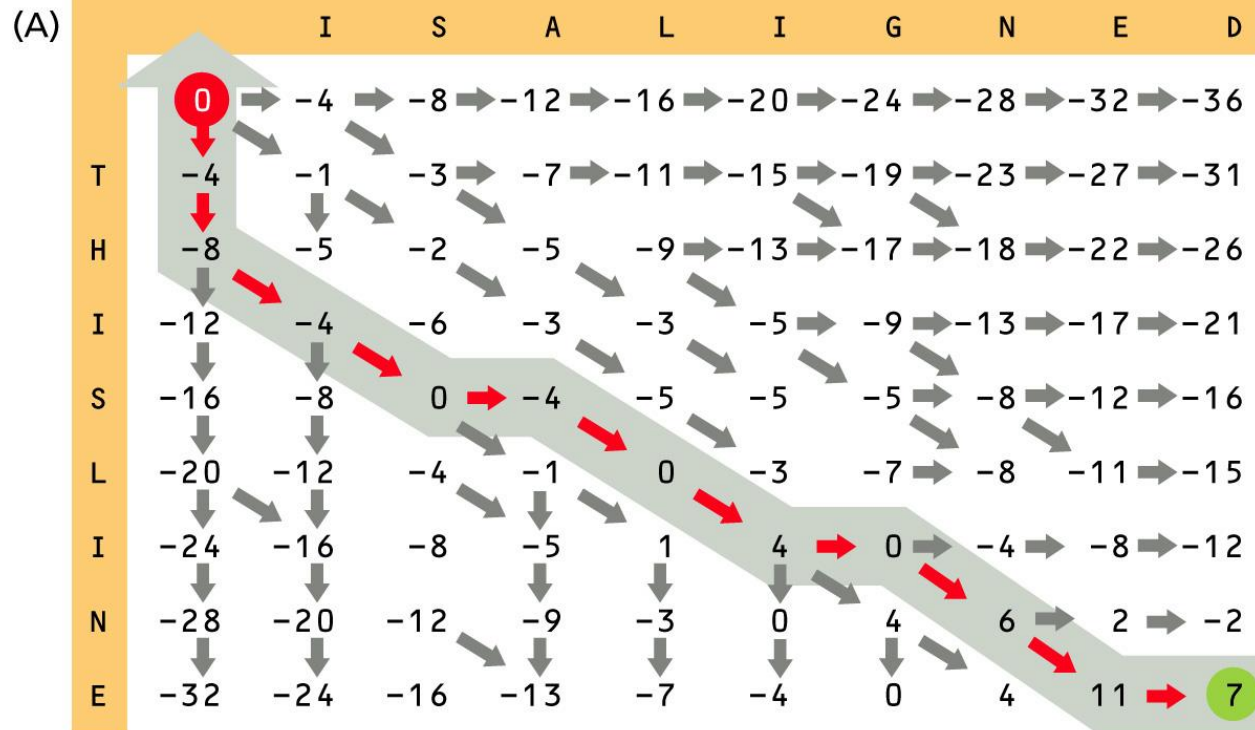
(A)



(B) THISLINE-

ISALIGNED

$$g = -4$$



(B)

```

  T H I S - L I - N E -
      | | | | |
  - - I S A L I G N E D
  
```



Gap penalty ( $g$ ) = -8  
 Gap extension ( $e$ ) = -2

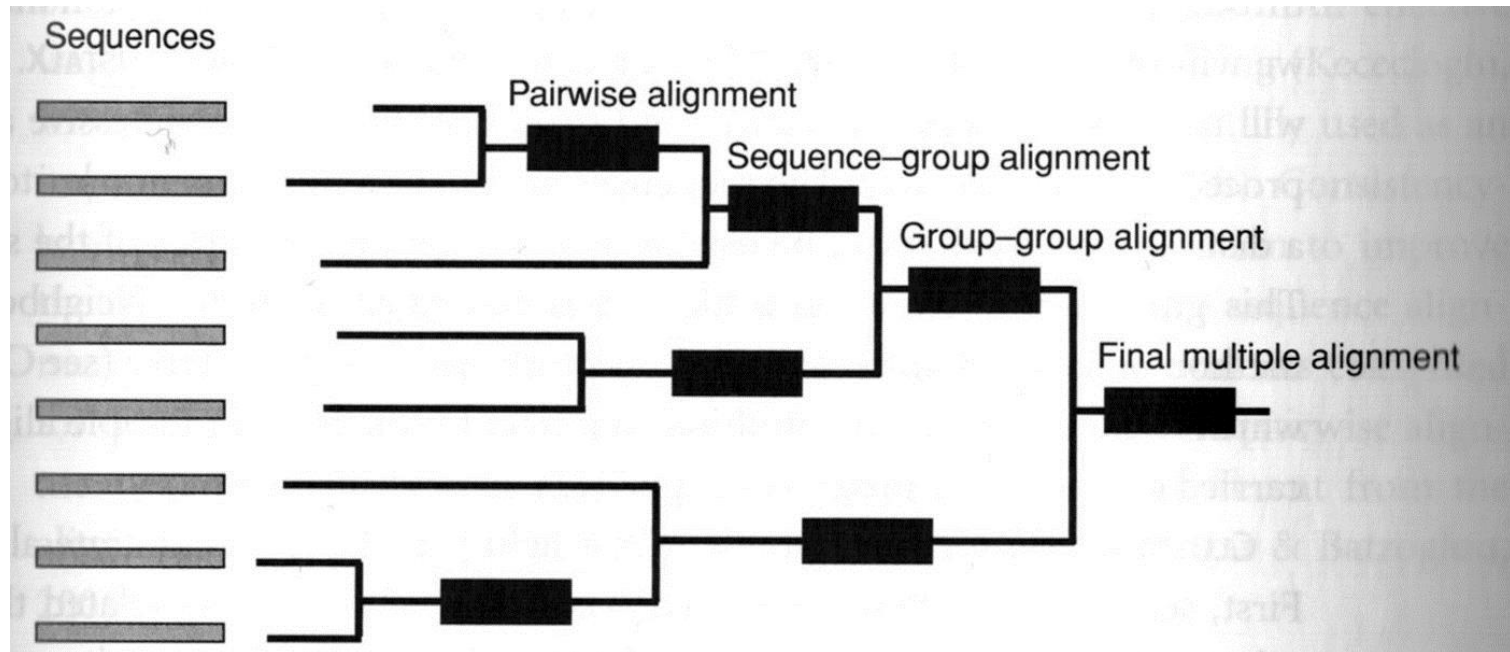
$$GP = g + e(l - 1)$$

<div> <div><math>j \searrow i</math></div> </div>		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)

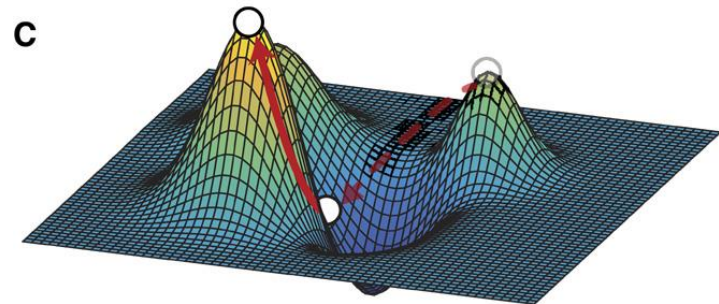
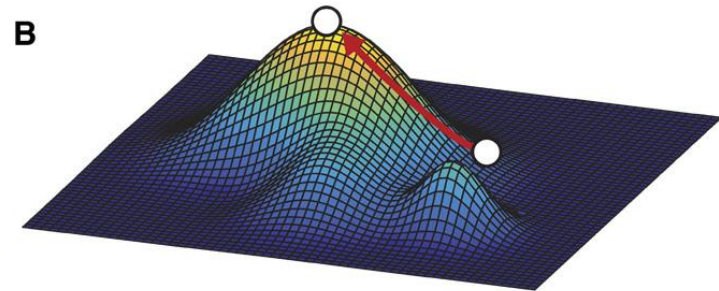
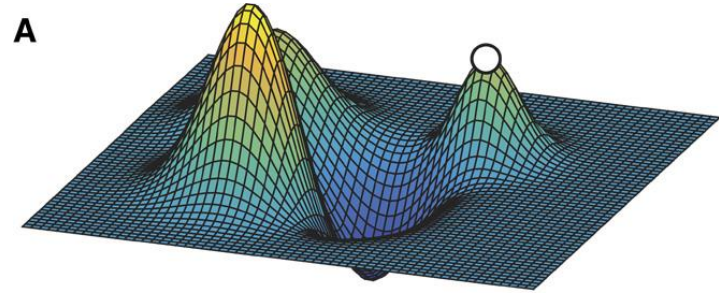
# III. Progressive Alignments



## IV. Programs: CLUSTALX

- Difference between X/W interface
- Progressive alignment based on unrooted 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  $K_{\text{mer}}$  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)