

Multiple Sequence Alignments



Introduction

The Questions

- What is a multiple sequence Alignment?
- What can it do for me?
- How Can I produce one of these?
- How Can I Use It?

What is A Multiple Sequence Alignment?

```

chite  ---ADKPKRPLSA YMLWLN SARESIKREN PDK- VTEVAKKGGELWRGLKD
wheat  --DPNKKPRAPSA F FVEMGE F REEFKQKNPKNK SVAAVGKAAGERWKSLS E
trybr  K KDSNAPKRAMT S F M F F S D F R S ---K H S D L S - I V E M S K A A G A A W K E I G P
mouse  ----KPKRPRSAYNI YVSE S F Q ---EAKDDS -A Q G K L K L V N E A W K N L I S P
          ***. . . . . : . . . . . * . . . . . *
chite  AATAKQNYIRALQ EYERNGG-
wheat  ANKLIKGEYNKAI IAA YNKGE S A
trybr  AEKDKER YKREM-----
mouse  AKDDRIRYDNEMK SWE E Q M A E
          * . . . . . : . . . . .

```

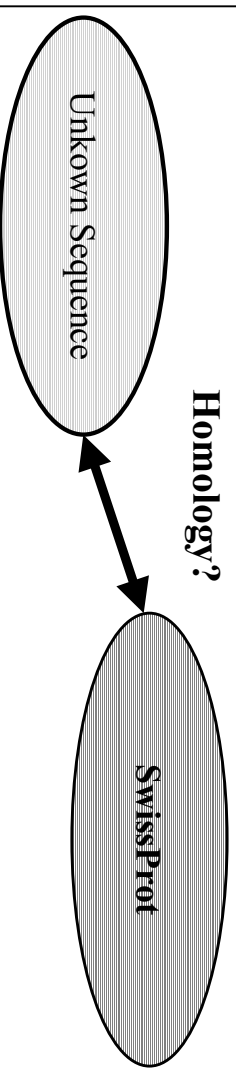
How Can I Use A Multiple Sequence Alignment?

```

chite  ---ADKPKRPLSA YMLWLN SARESIKREN PDK- VTEVAKKGGELWRGLKD
wheat  --DPNKKPRAPSA F FVEMGE F REEFKQKNPKNK SVAAVGKAAGERWKSLS E
trybr  K KDSNAPKRAMT S F M F F S D F R S ---K H S D L S - I V E M S K A A G A A W K E I G P
unknown ----KPKRPRSAYNI YVSE S F Q ---EAKDDS -A Q G K L K L V N E A W K N L I S P
          ***. . . . . : . . . . . * . . . . . *
chite  AATAKQNYIRALQ EYERNGG-
wheat  ANKLIKGEYNKAI IAA YNKGE S A
trybr  AEKDKER YKREM-----
unknown AKDDRIRYDNEMK SWE E Q M A E
          * . . . . . : . . . . .

```

Extrapolation



NiceProt View of SWISS-PROT: P40623

[General] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequences] [Tools]

General information about the entry

Entry name	HMGB_CHITTE
Primary accession number	P40623
Secondary accession number(s)	None
Entered in SWISS-PROT in	Release 31, February 1995
Sequence was last modified in	Release 31, February 1995
Annotations were last modified in	Release 32, November 1995

Name and origin of the protein

Protein name	MOBILITY GROUP PROTEIN 1B
Synonym(s)	None
Gene name(s)	HMGB1B
From	Chironomus tentans (Midge)
Taxonomy	Bekaryota, Metazoa, Arthropoda, Tracheata, Hexapoda, Insecta, Pterygota, Neoptera, Endopterygota, Diptera, Nematotermata, Chironomidae, Chironomidae, Chironomus

References

[1] SEQUENCE FROM N.A. TISSUE=EMBRIONIC EPITHELIUM; MEDLINE: 92381031 [NCBI] EXPASY: [Israel, Japan] Watanabe J.R., Schulte E.; "Insect proteins homologous to mammalian high mobility group protein 1. Characterization and DNA-binding properties."; J. Biol. Chem. 267:17170-17177(1992).

Comments

- **FUNCTION:** FOUND IN CONDENSED CHROMOMERES, BINDS PREFERENTIALLY TO AT-RICH DNA.
- **SUBCELLULAR LOCALIZATION:** NUCLEAR.
- **SIMILARITY:** BELONGS TO THE HMGB/HMG2 PROTEIN FAMILY.
- **SIMILARITY:** CONTAINS 1 HMGB BOX.

Copyright

This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the European Bioinformatics Institute. There are no restrictions on the use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See <http://www.ebi.ac.uk/infocentre/> or send an email to license@ebi.ac.uk).

Cross-references

EMBL	M93294, AAAA21713.1, EMBL / GenBank / DDBJ [CodonSequence]
HSSP	G05783, IHMA [HSSP ENTRY / SWISS-3DIMAGE / PDB]
PFAM	PF00505, HMGB_box_1
PRODOM	[Domain structure / List of seq sharing at least 1 domain]
BLOCKS	P40623
DOMO	P40623
PROTOMAP	P40623
PRESAGE	P40623
DIP	P40623
SWISS-2DPAGE	GET REGION ON 2D PAGE

Keywords

Nuclear protein, Chromosomal protein, DNA binding.

Features

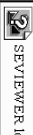
DNA BIND DOMAIN	5 .. 71	HMG BOX, ASP/GLU-RICH (ACIDIC).
	104 .. 110	

Sequence information

Length: 110 AA Molecular weight: 12150 Da [C]RC64: E3401735713333C4 [This is a checksum on the sequence]

10	20	30	40	50	60
MADPRERPL3	AYHLINSGAR	ESTIRENDF	RTEVYAKKGG	ELWRGLKDS	EWEAKAKTK
70	80	90	100	110	
QNYTRALQET	EMNGGGDDK	GKRKGAAPK	KGAKKAKNG	AHSDDDGDF	

P40623 in FASTA format



FT table viewer

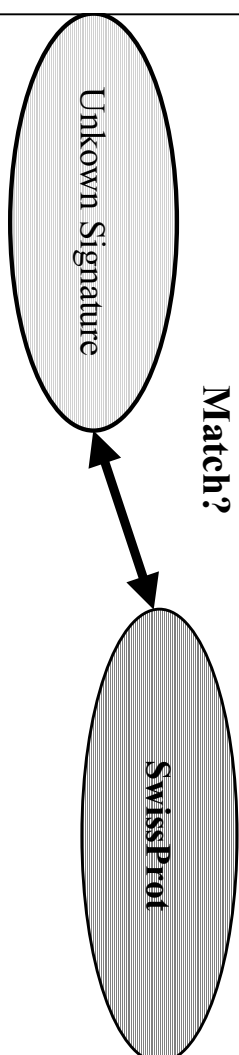
How Can I Use A Multiple Sequence Alignment?

```

chite  ---ADKPKRPLSAVMLWINSARESIRKRENPDK-VTEVAKKGGELWRGLKD
wheat  --DNPKPKRAPSFAFVFMGEFREFFKQKNPKKNSVAAVGKAAGERWKSLSLSE
trybr  KKD SNAPKRAMTSFMFFSSDFRS---KHSDIS-IVEMSKAAGAAWKEIGP
mouse  ----KPKRPSAYNIYVSESFQ---EAKDDS-AQGKTKLVNEFAWKNLSP
          *** . . . . . * . . . *
chite  AATAKQNYTRALQEYERNNG-
wheat  ANKLKGEYNKAIAAYNKGE SA
trybr  AEKDKERYKREM-----
mouse  AKDDRIRYDNEKMSWEEQMAE
          * . . . .
    
```

Extrapolation

Prosite Patterns



How Can I Use A Multiple Sequence Alignment?

```

chite  ---ADKPKRPLSA YMLINLSARESIKRENPPDFK-VTEVAKKGGELWRGLKD
wheat  --DPNKPKRAPSAFFVEMGEFREEFKQKNPKNKSVAAVGKAAGERWKSISE
trybr  KKD SNA PKRAMTSEMFSSDFRS---KHSDLS-IVEMSKAAGA AWKELGP
mouse  ----KPKRPRSAYNIYVSESFQ---EAKDDS-AQGKIKLVNEAWKNLSP
          ***. . . . . : . . . * . * : *
chite  AATAKQNYIRALQEYERNNG-
wheat  ANKIKGEYNKAI AAYNKGESA
trybr  AEKDKERYKREM-----
mouse  AKDDRIRYDNEKMSWEEQMAE
          * : . * . . :
  
```

Extrapolation

Motifs/Patterns

Profiles

Phylogeny

Struc. Prediction

PhD For secondary Structure Prediction: 75% Accurate.

Threading: is improving but is not yet as good.

How Can I Use A Multiple Sequence Alignment?

```

chite  ---ADKPKRPLSA YMLINLSARESIKRENPPDFK-VTEVAKKGGELWRGLKD
wheat  --DPNKPKRAPSAFFVEMGEFREEFKQKNPKNKSVAAVGKAAGERWKSISE
trybr  KKD SNA PKRAMTSEMFSSDFRS---KHSDLS-IVEMSKAAGA AWKELGP
mouse  ----KPKRPRSAYNIYVSESFQ---EAKDDS-AQGKIKLVNEAWKNLSP
          ***. . . . . : . . . * . * : *
chite  AATAKQNYIRALQEYERNNG-
wheat  ANKIKGEYNKAI AAYNKGESA
trybr  AEKDKERYKREM-----
mouse  AKDDRIRYDNEKMSWEEQMAE
          * : . * . . :
  
```

Extrapolation

Motifs/Pateterns

Profiles

Phylogeny

Struc. Prediction

Automatic Multiple Sequence Alignment methods are not always perfect...

Caution!

Why Is It Difficult To Compute A multiple Sequence Alignment?

BIOLOGY

What is A GOOD Alignment?

```

chite  ---ADKPKRPLSAYMLWINSARESIRENDPK-VTEVAKKGGELWRGDKD
wheat  --DPNKPKRAPSAFFVFMGEFREEFKQKNPKKSVAAVVGKAAGERWKSISE
trybr  KDSNAPKRAMTSEMFSSDFRS-----KHSDL-S-IVEMSKAAGAAMKEILGP
mouse  -----KPKRPRSAAYNIYVSESFQ-----EAKDDS-AQGKLLVNEAMKNTLSP
      *** . : : : . : : . : : . : : . : : . : : . : : * * : *
  
```

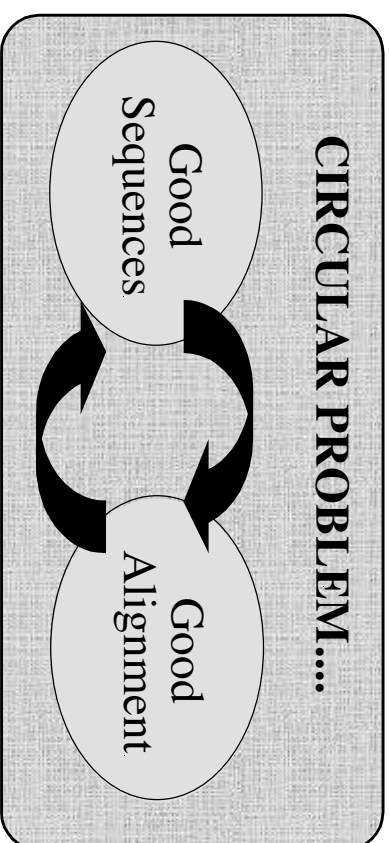
COMPUTATION

What is **THE** good Alignment?

Why Is It Difficult To Compute A multiple Sequence Alignment

BIOLOGY

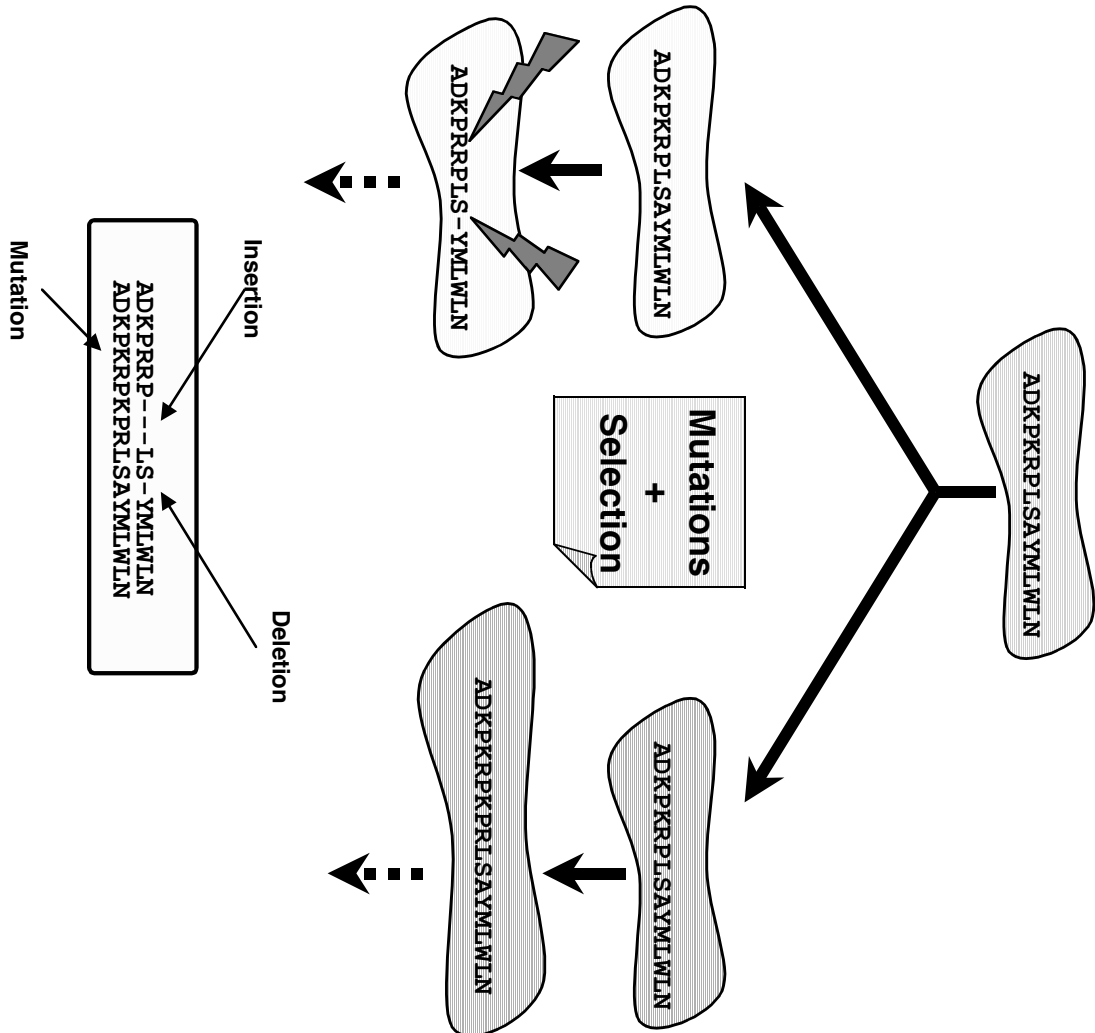
COMPUTATION



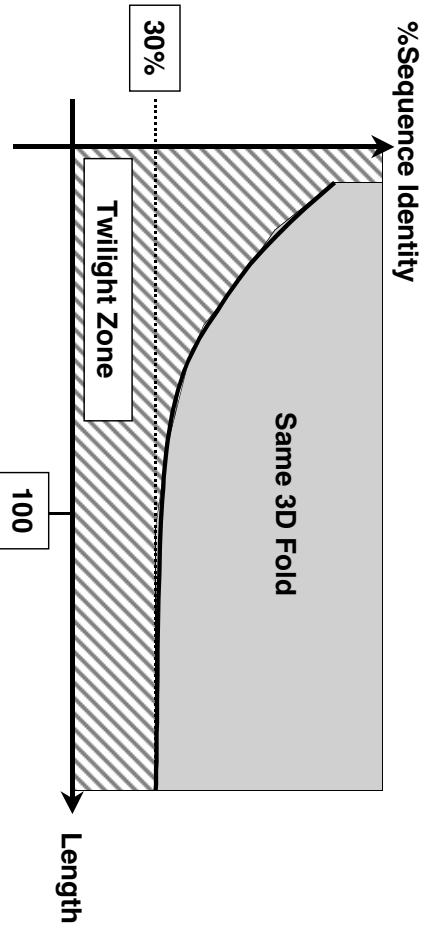
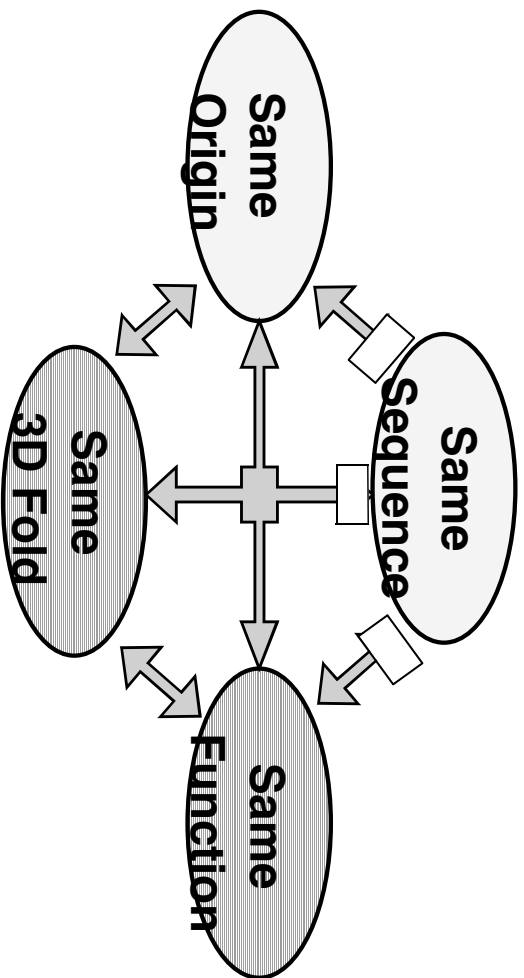
What Do I Need To Know To Make A good Multiple Sequence Alignment?

- How Do Sequences Evolve?
- How Does The Computer Align The Sequences?
- How Can I Choose My Sequences?
- What is The Best Program?
- How Can I Use My Alignment?

An Alignment is a STORY



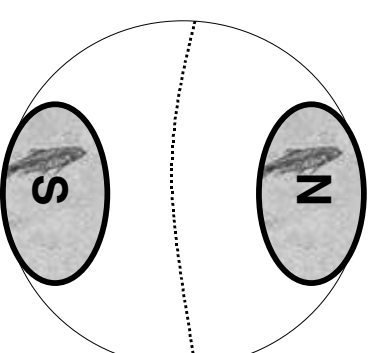
HOMOLOGY



Convergent Evolution

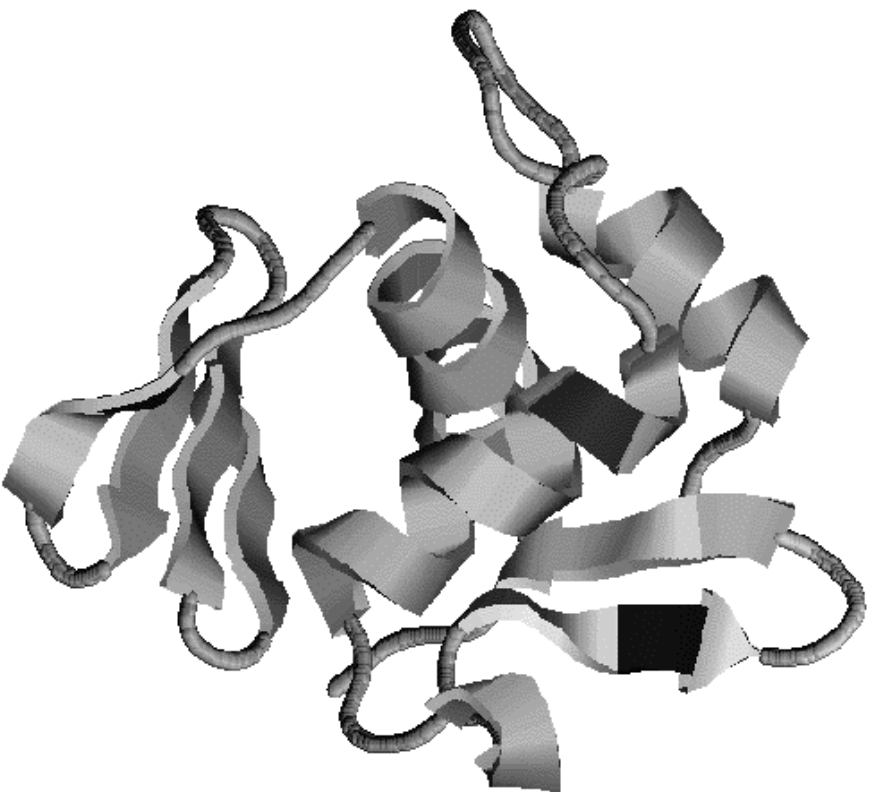
Chen et al, 97, PNAS, 94, 3811-16

**AFGP with (ThrAlaAla)_n
Similar To Trypsinogen**



**AFGP with (ThrAlaAla)_n
NOT
Similar to Trypsinogen**

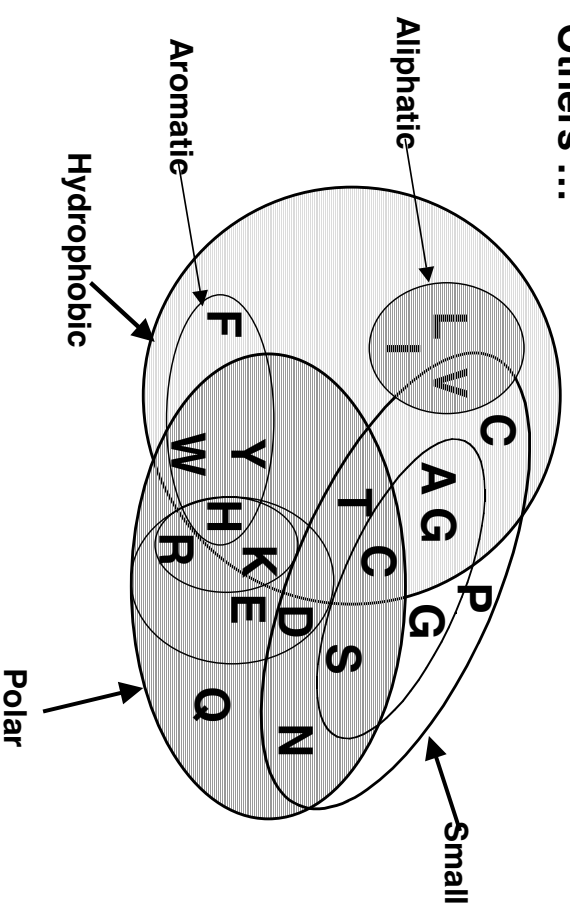
Structures and Mutations...



OmpR, Cter Domain

Residues et Mutations...

All Residues are Equal, But some More Than Others ...



Accurate Matrices are Data Driven Rather Than Knowledge Driven.

Substitution Matrices...

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

Different Flavors:

- Pam: 250, 350
- Blosum: 45, 62
- ...

What is the Best Substitution Matrix?

Mutations Rates Depend on Families...

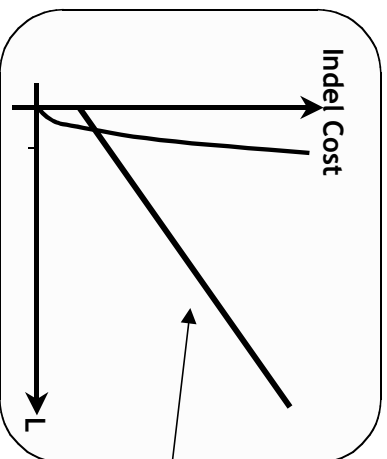
Family	S	NS
Histone3	6.4	0
Insulin	4.0	0.1
Interleukin I	4.6	1.4
α -Globin	5.1	0.6
Apolipoprot. AI	4.5	1.6
Interferon G	8.6	2.8

Rates in Substitutions/site/Billion Years as measured on Mouse Vs Human (0.08 Billion years)

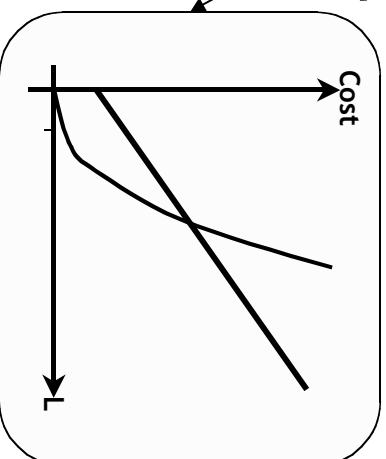
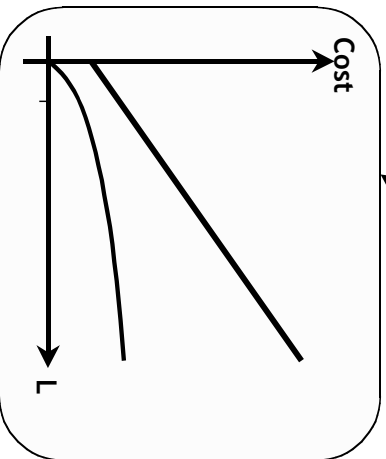
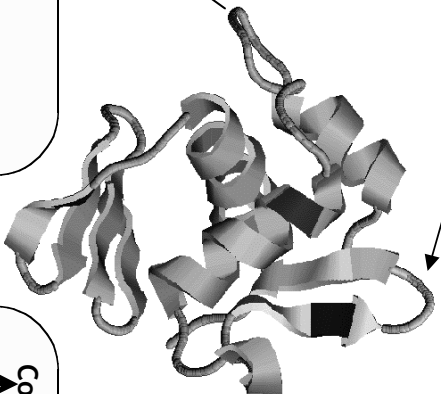
Choosing The Right Matrix may be Tricky...

- GONNET 250 > BLOSUM62 > PAM 250.
- But This will depend on:
 - The Family.
 - The Program Used and Its Tunning.
- Insertions, Deletions?

Insertions and Deletions?

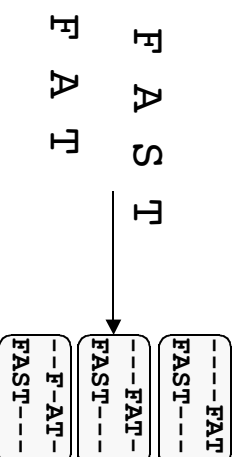


Affine Gap Penalty
Cost = GOP + GEP * L



HOW CAN I ALIGN TWO SEQUENCES

Brut Force Enumeration



$$\binom{(L1+L2)!}{(L1)! * (L2)!}^2$$

Dynamic Programming (Needleman and Wu)

Match=1 Mismatch=-1 Gap=-1

	F	A	S	T	
F	0	-1	-2	-3	-4
A	-1	0	-2	-3	-4
S	-2	-1	0	-2	-3
T	-3	-2	-1	0	-2

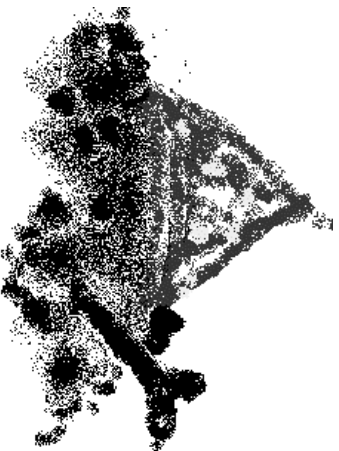
	F	A	S	T	
F	0	-1	-2	-3	-4
A	-1	0	-2	-3	-4
S	-2	-1	0	-1	-2
T	-3	-2	-1	0	-1

	F	A	S	T	
F	0	-1	-2	-3	-4
A	-1	0	-2	-3	-4
S	-2	-1	0	-1	-2
T	-3	-2	-1	0	-1

F A S T
F A - T

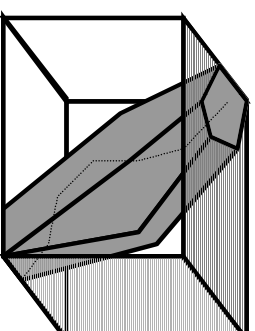
HOW CAN I ALIGN *MANY* SEQUENCES

7 Globins => 1000 years



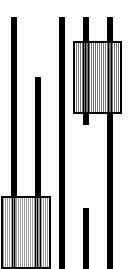
Existing Methods

1-Carillo and Lipman:



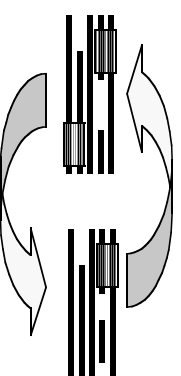
- MSA, DCA.
- Few Small Closely Related Sequence.
- Do Well When They Can Run.

2-Segment Based:



- DIALIGN, MACAW.
- May Align Too Few Residues

3-Iterative:



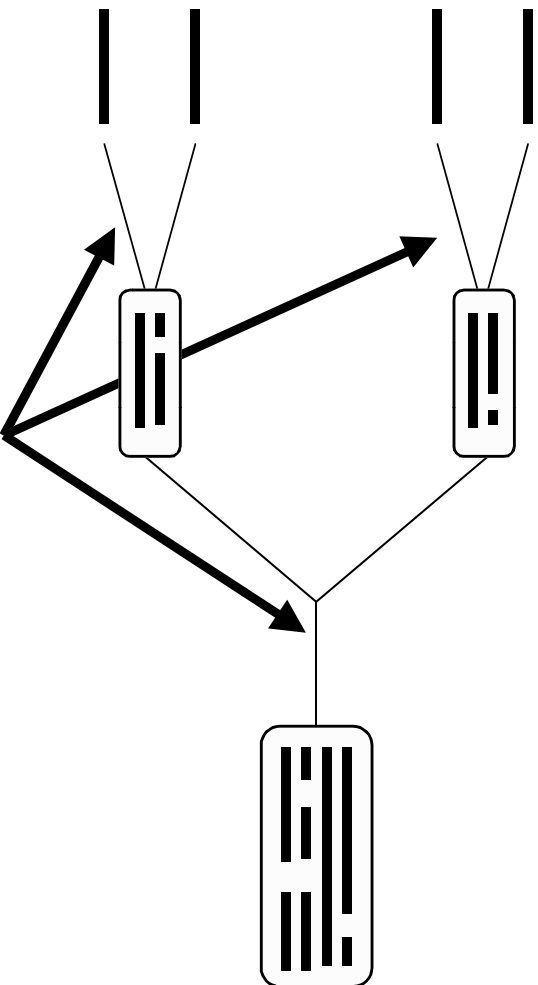
- HMMS, HMMER, SAM.
- Slow, Sometimes Inaccurate
- Good Profile Generators

4-Progressive:

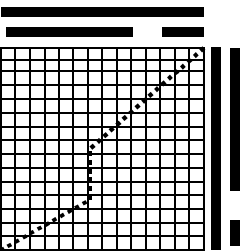
- ClustalW, Pileup, Multalign...
- Fast and Sensitive

Progressive Alignment

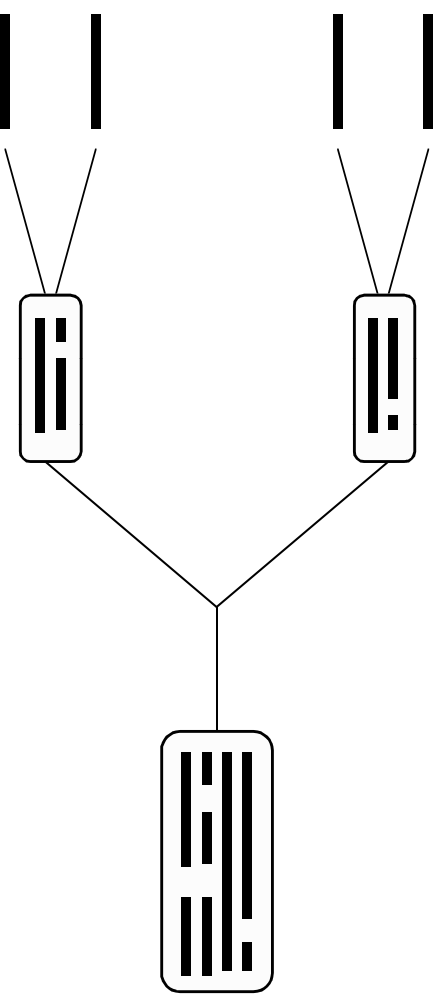
Feng and Dolittle, 1980; Taylor 1981



Dynamic Programming Using A Substitution Matrix



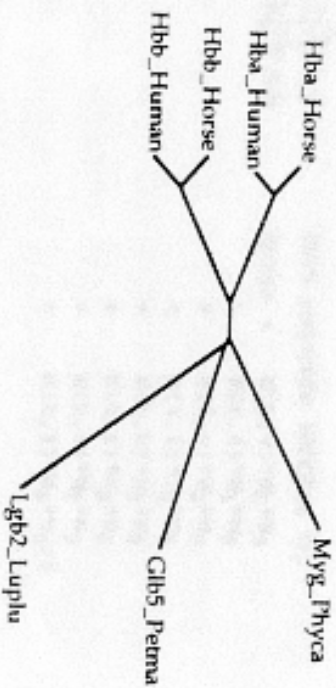
Progressive Alignment



- Depends on the CHOICE of the sequences.
- Depends on the ORDER of the sequences (Tree).
- Depends on the PARAMETERS:
 - Substitution Matrix.
 - Penalties (Gop, Gep).
 - Sequence Weight.
 - Tree making Algorithm.

Weighting Within ClustalW

Hbb_Human	1	-	-	-	-	-
Hbb_Horse	2	.17	-	-	-	-
Hba_Human	3	.59	.60	-	-	-
Hba_Horse	4	.59	.59	.13	-	-
Myg_Phyca	5	.77	.77	.75	.75	-
Cib5_Petma	6	.81	.82	.73	.74	.80
Lgb2_Luplu	7	.87	.86	.86	.88	.93
		1	2	3	4	5
						6



		.081	Hbb_Human:	0.221
		.084	Hbb_Horse:	0.225
	.226	.055	Hba_Human:	0.194
	.219	.065	Hba_Horse:	0.203
	.015		Myg_Phyca:	0.411
	.062		Cib5_Petma:	0.398
	.389		Lgb2_Luplu:	0.442
	.442			

Position Specific GOP

