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?

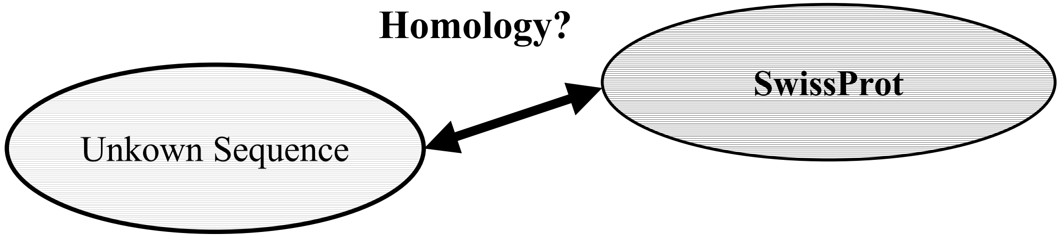
What is A Multiple Sequence Alignment?

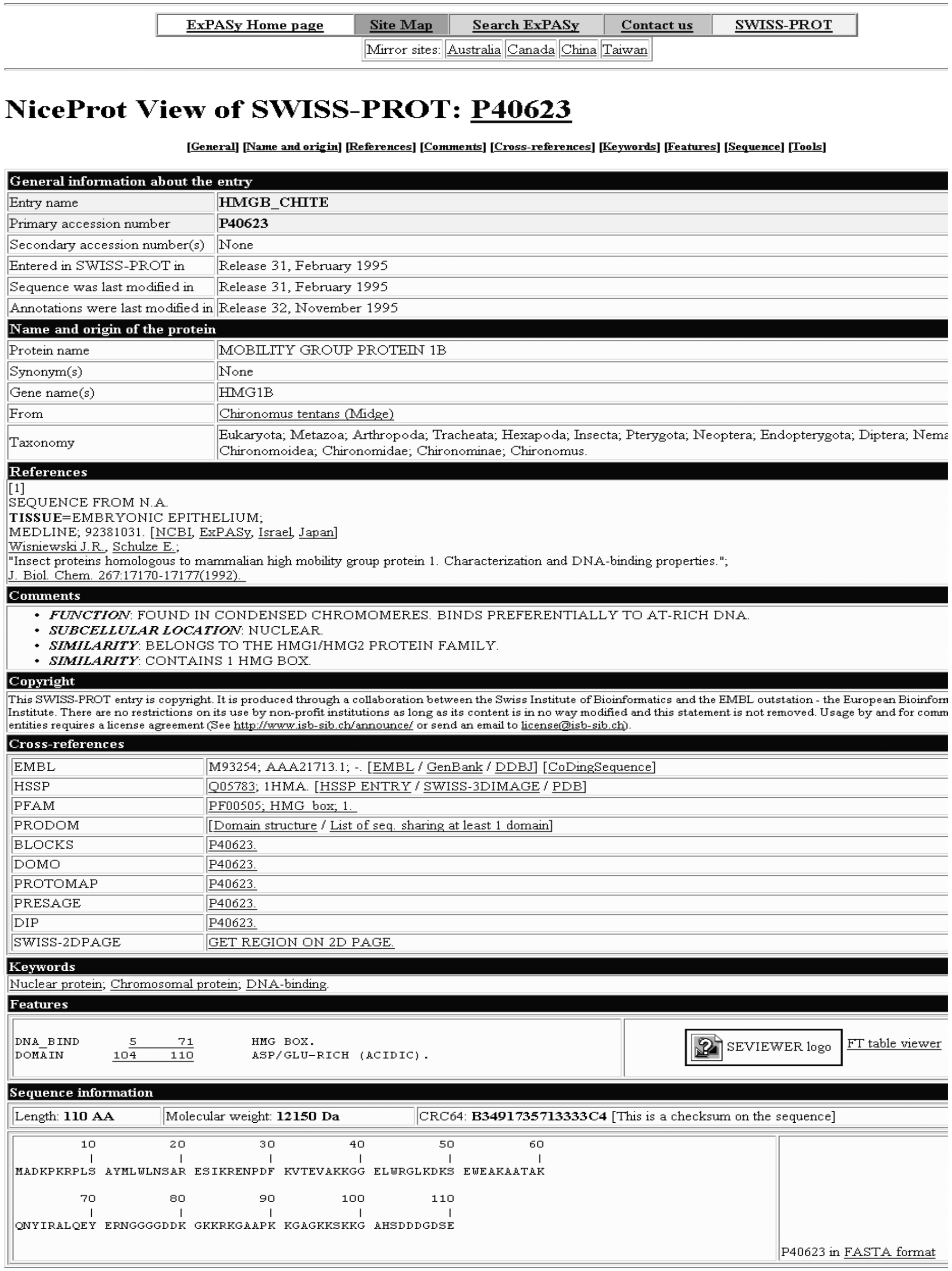
```
chite  ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat  --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr  KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse  -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
             ***. ::: .: ..  .    :  . .      *  .  *: *

chite  AATAKQNYIRALQEYERNGG-
wheat  ANKLKGEYNKAIAAYNKGESA
trybr  AEKDKERYKREM---------
mouse  AKDDRIRYDNEMKSWEEQMAE
       *   : .* . :
```

How Can I Use A Multiple Sequence Alignment?

```
chite    ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat    --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr    KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
unknown  -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
               ***. ::: .: ..  .    :  . .      *  .  *: *

chite    AATAKQNYIRALQEYERNGG-
wheat    ANKLKGEYNKAIAAYNKGESA
trybr    AEKDKERYKREM---------
unknown  AKDDRIRYDNEMKSWEEQMAE
         *   : .* . :
```

Extrapolation





ExPASy Home page | Site Map | Search ExPASy | Contact us | SWISS-PROT

Mirror sites: Australia Canada China Taiwan

NiceProt View of SWISS-PROT: P40623

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

General information about the entry	
Entry name	**HMGB_CHITE**
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)	HMG1B
From	Chironomus tentans (Midge)
Taxonomy	Eukaryota; Metazoa; Arthropoda; Tracheata; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Nema Chironomoidea; Chironomidae; Chironominae; Chironomus.

References

[1]
SEQUENCE FROM N.A.
TISSUE=EMBRYONIC EPITHELIUM;
MEDLINE; 92381031. [NCBI, ExPASy, Israel, Japan]
Wisniewski J.R., Schulze 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 LOCATION***: NUCLEAR.
- ***SIMILARITY***: BELONGS TO THE HMG1/HMG2 PROTEIN FAMILY.
- ***SIMILARITY***: CONTAINS 1 HMG BOX.

Copyright

This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinform Institute. There are no restrictions on its 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 comm entities requires a license agreement (See http://www.isb-sib.ch/announce/ or send an email to license@isb-sib.ch).

Cross-references	
EMBL	M93254; AAA21713.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
HSSP	Q05783; 1HMA. [HSSP ENTRY / SWISS-3DIMAGE / PDB]
PFAM	PF00505; HMG_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      5     71       HMG BOX.
DOMAIN      104    110       ASP/GLU-RICH (ACIDIC).
```

SEVIEWER logo FT table viewer

Sequence information

Length: **110 AA**	Molecular weight: **12150 Da**	CRC64: **B3491735713333C4** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
         |          |          |          |          |          |
MADKPKRPLS AYMLWLNSAR ESIKRENPDF KVTEVAKKGG ELWRGLKDKS EWEAKAATAK

        70         80         90        100        110
         |          |          |          |          |
QNYIRALQEY ERNGGGGDDK GKKRKGAAPK KGAGKKSKKG AHSDDDGDSE
```

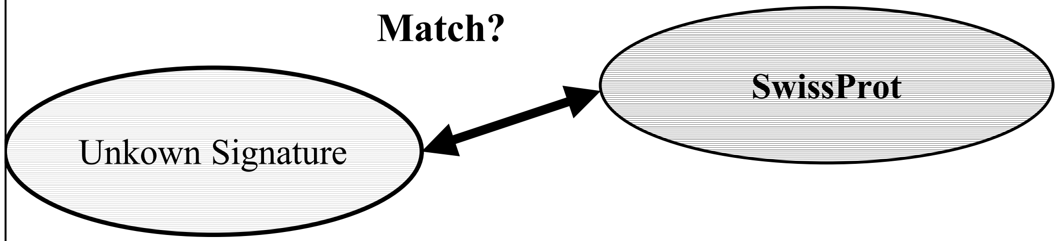
P40623 in FASTA format

How Can I Use A Multiple Sequence Alignment?

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

Extrapolation

Prosite Patterns



How Can I Use A Multiple Sequence Alignment?

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-IQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

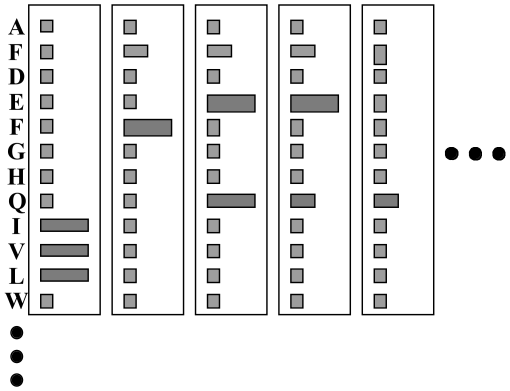
L?

K>R

Extrapolation

Prosite Patterns

Prosite Profiles



-More Sensitive
-More Specific

How Can I Use A Multiple Sequence Alignment?

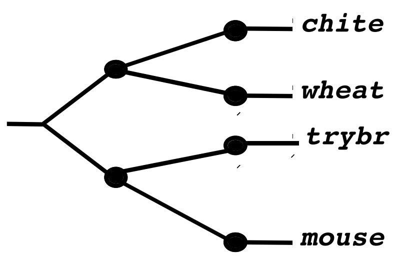
```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

Extrapolation

Motifs/Patterns

Profiles

Phylogeny



-Evolution
-Paralogy/Orthology

How Can I Use A Multiple Sequence Alignment?

```
chite  ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat  --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr  KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse  -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
             ***. ::: .: ..  .    : . .      *  .  *: *

chite  AATAKQNYIRALQEYERNGG-
wheat  ANKLKGEYNKAIAAYNKGESA
trybr  AEKDKERYKREM---------
mouse  AKDDRIRYDNEMKSWEEQMAE
       *   : .* . :
```

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  ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat  --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr  KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse  -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
             ***. ::: .: ..  .    : . .      *  .  *: *

chite  AATAKQNYIRALQEYERNGG-
wheat  ANKLKGEYNKAIAAYNKGESA
trybr  AEKDKERYKREM---------
mouse  AKDDRIRYDNEMKSWEEQMAE
       *   : .* . :
```

Extrapolation

Motifs/Patterns

Profiles

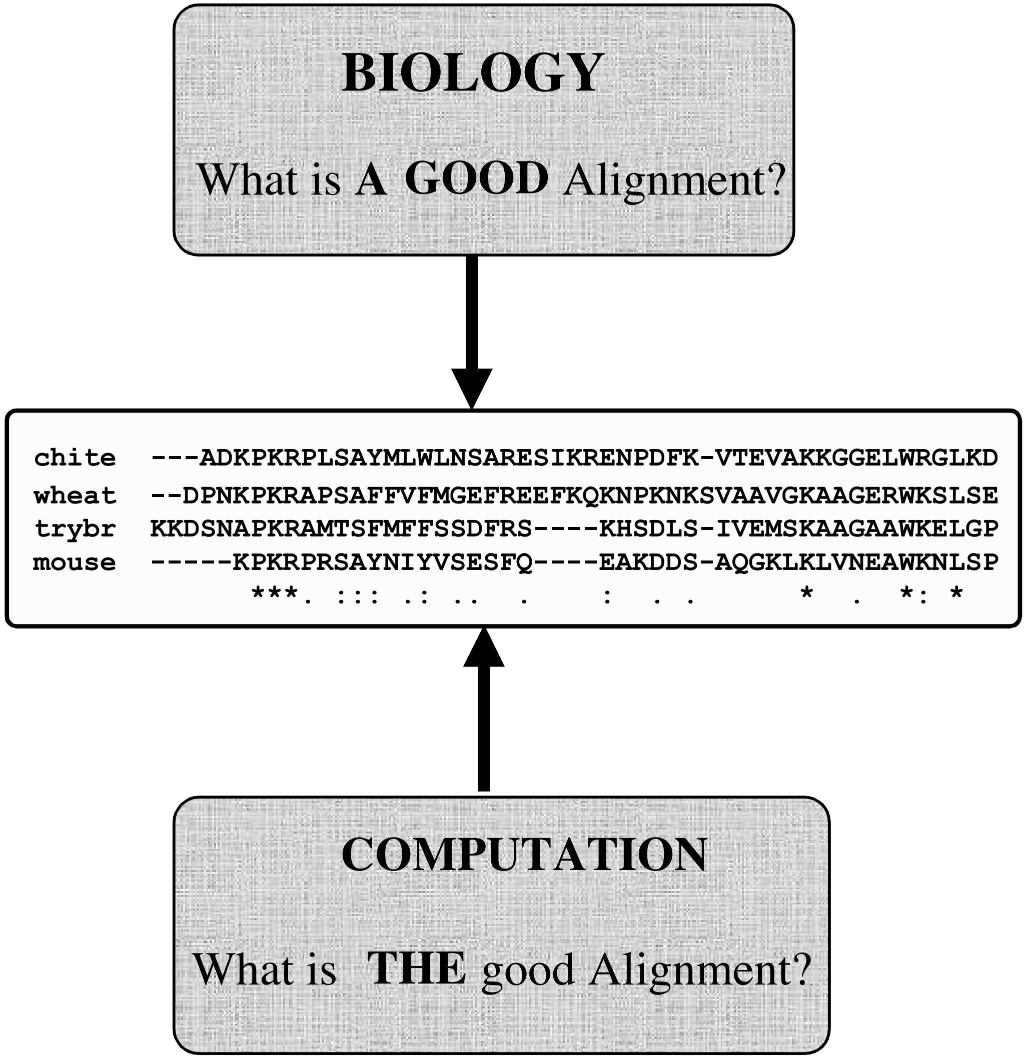
Phylogeny

Struc. Prediction

Caution!

Automatic Multiple Sequence Alignment methods are not always perfect…

Why Is It Difficult To Compute A multiple Sequence Alignment?



BIOLOGY

What is **A GOOD** Alignment?

```
chite  ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat  --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr  KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse  -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
             ***. ::: .: ..  .    :  . .      *  .  *: *
```

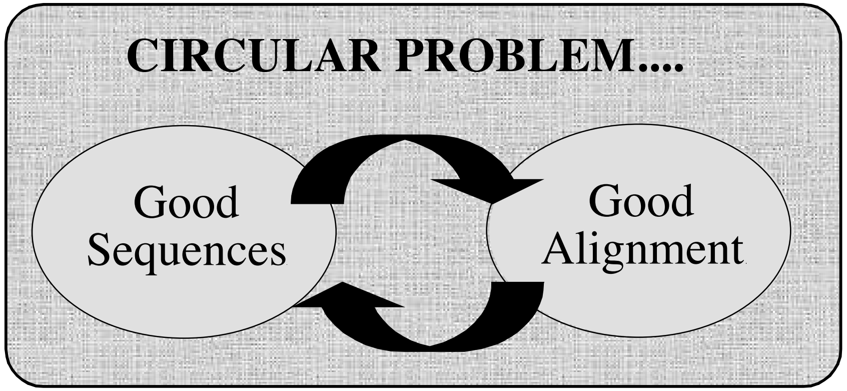
COMPUTATION

What is **THE** good Alignment?

Why Is It Difficult To Compute A multiple Sequence Alignment

BIOLOGY

COMPUTATION

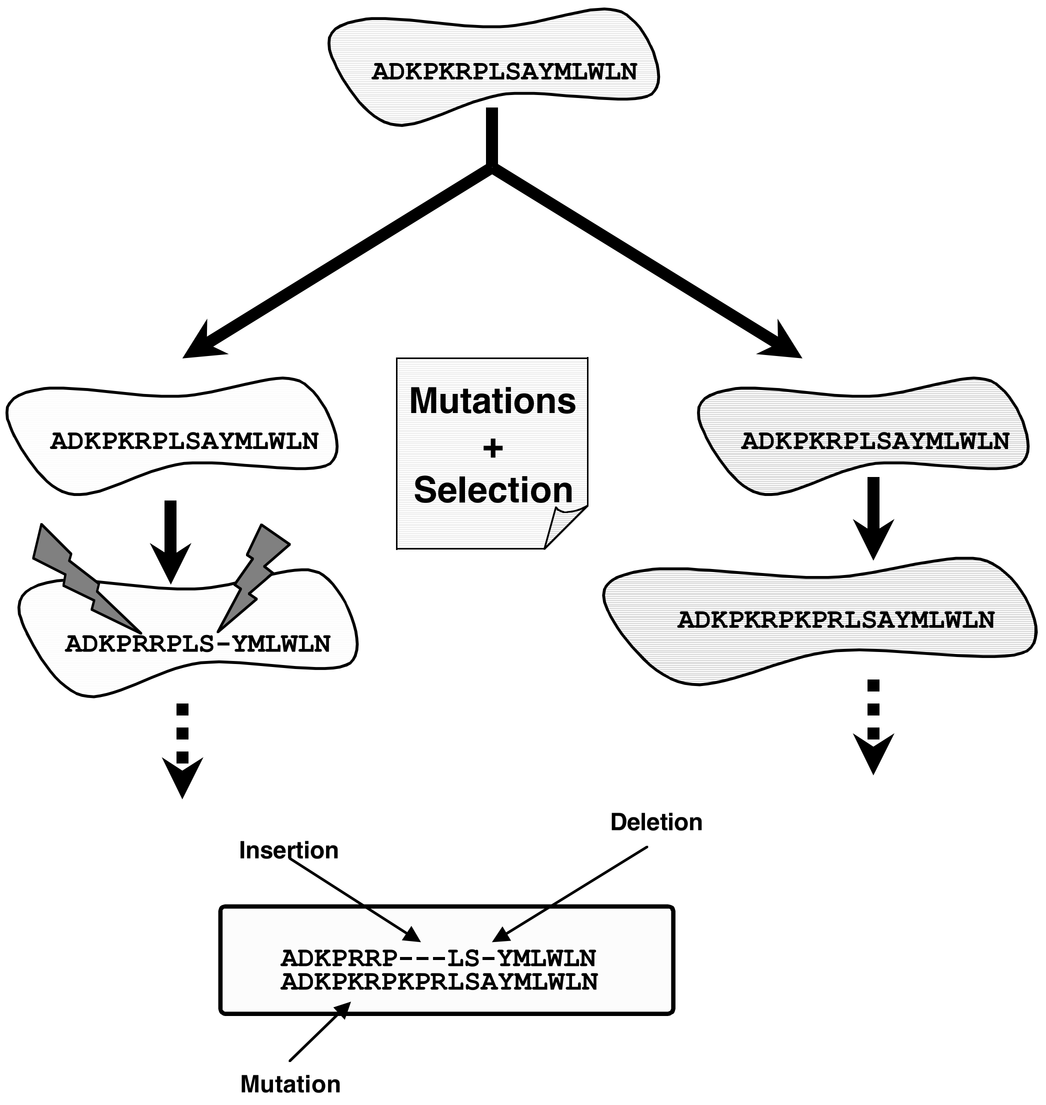


CIRCULAR PROBLEM....

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

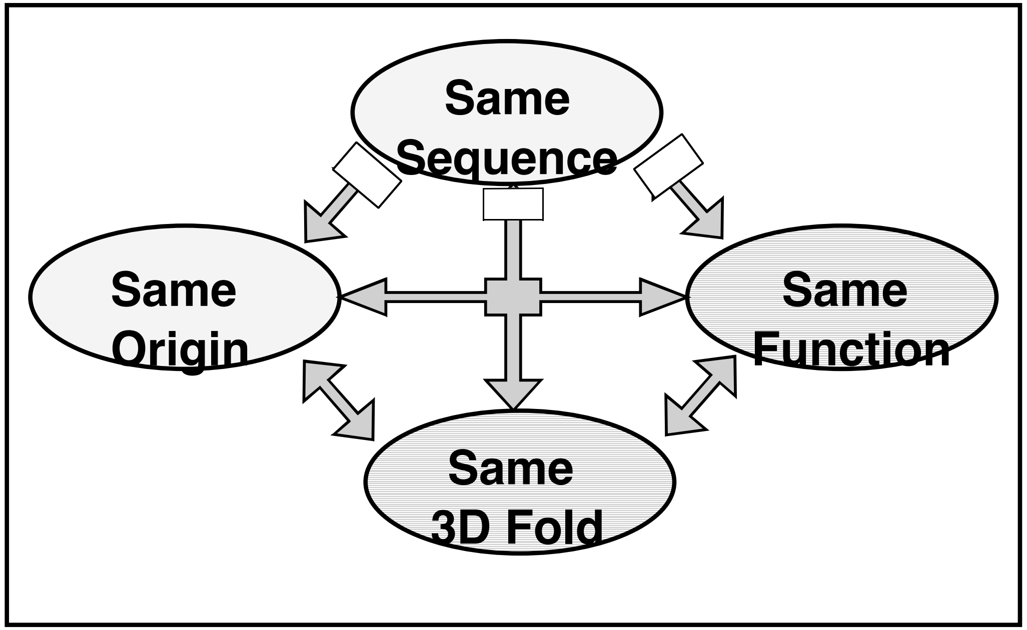
- How Do Sequences Evolve?
- How Does The Computer Align The Sequences?
- How Can I Choose My Sequences?
- What is The Best Program?
- How Can I Use My Alignment?

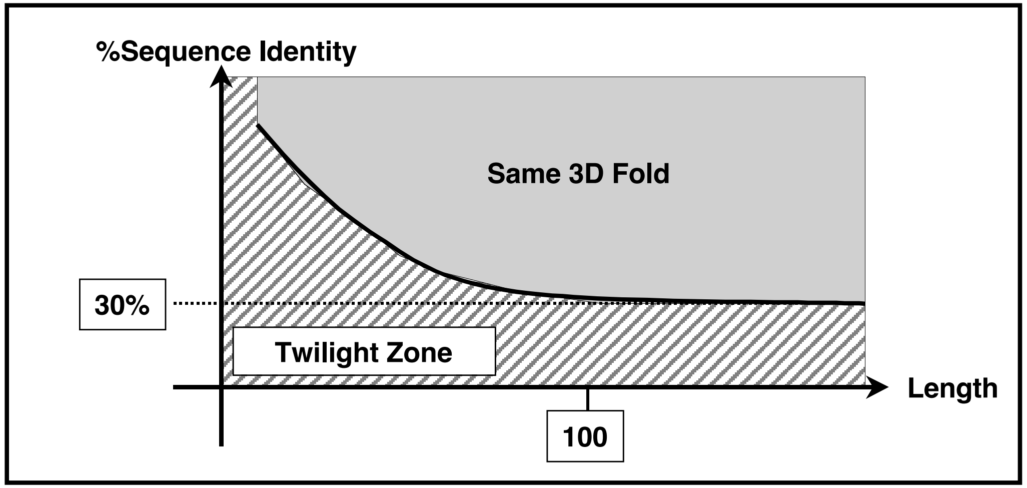
An Alignment is a **STORY**



```
ADKPRRP---LS-YMLWLN
ADKPKRPKPRLSAYMLWLN
```

HOMOLOGY

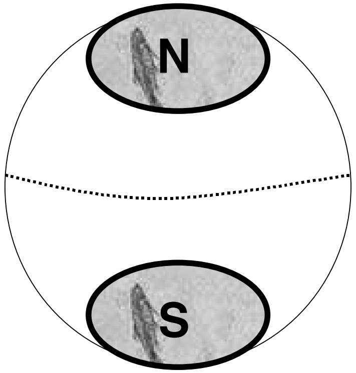




Convergent Evolution

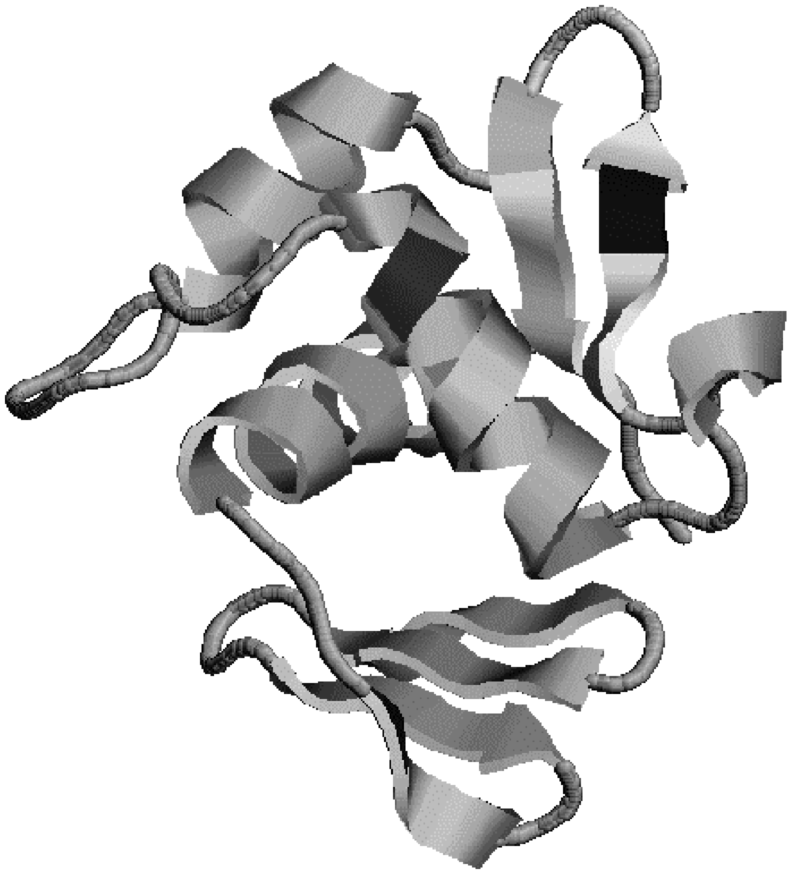
Chen et al, 97, PNAS, 94, 3811-16

AFGP with (ThrAlaAla)n
Similar To Trypsynogen



AFGP with (ThrAlaAla)n
NOT
Similar to Trypsinogen

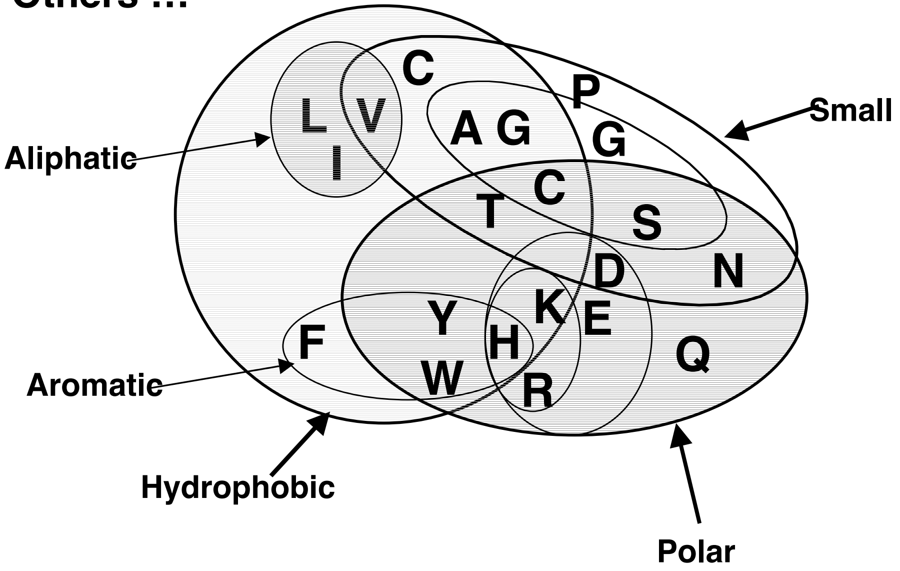
Structures and Mutations…



OmpR, Cter Domain

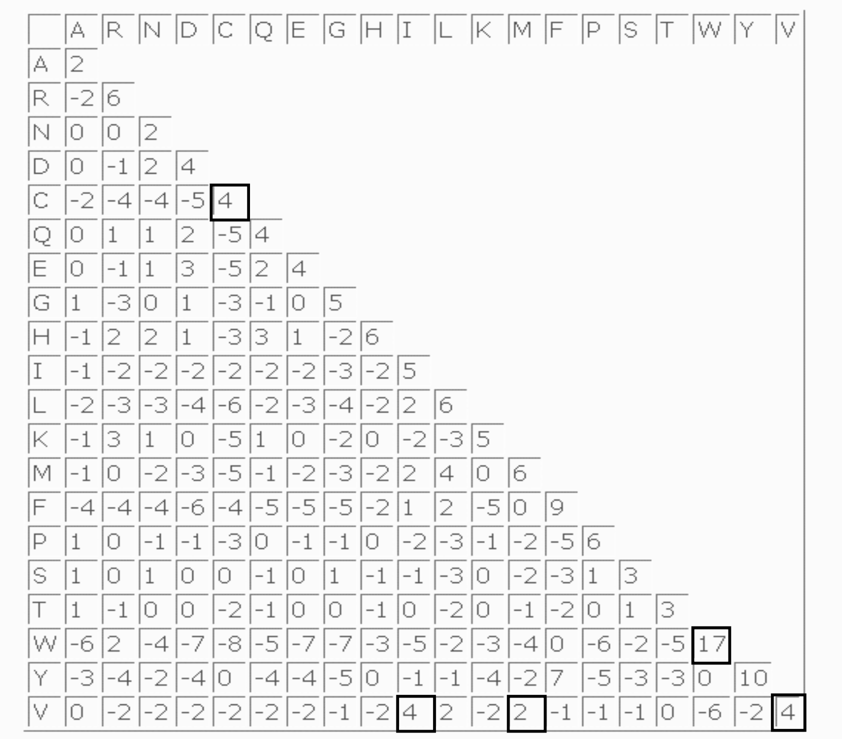
Residues et Mutations…

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

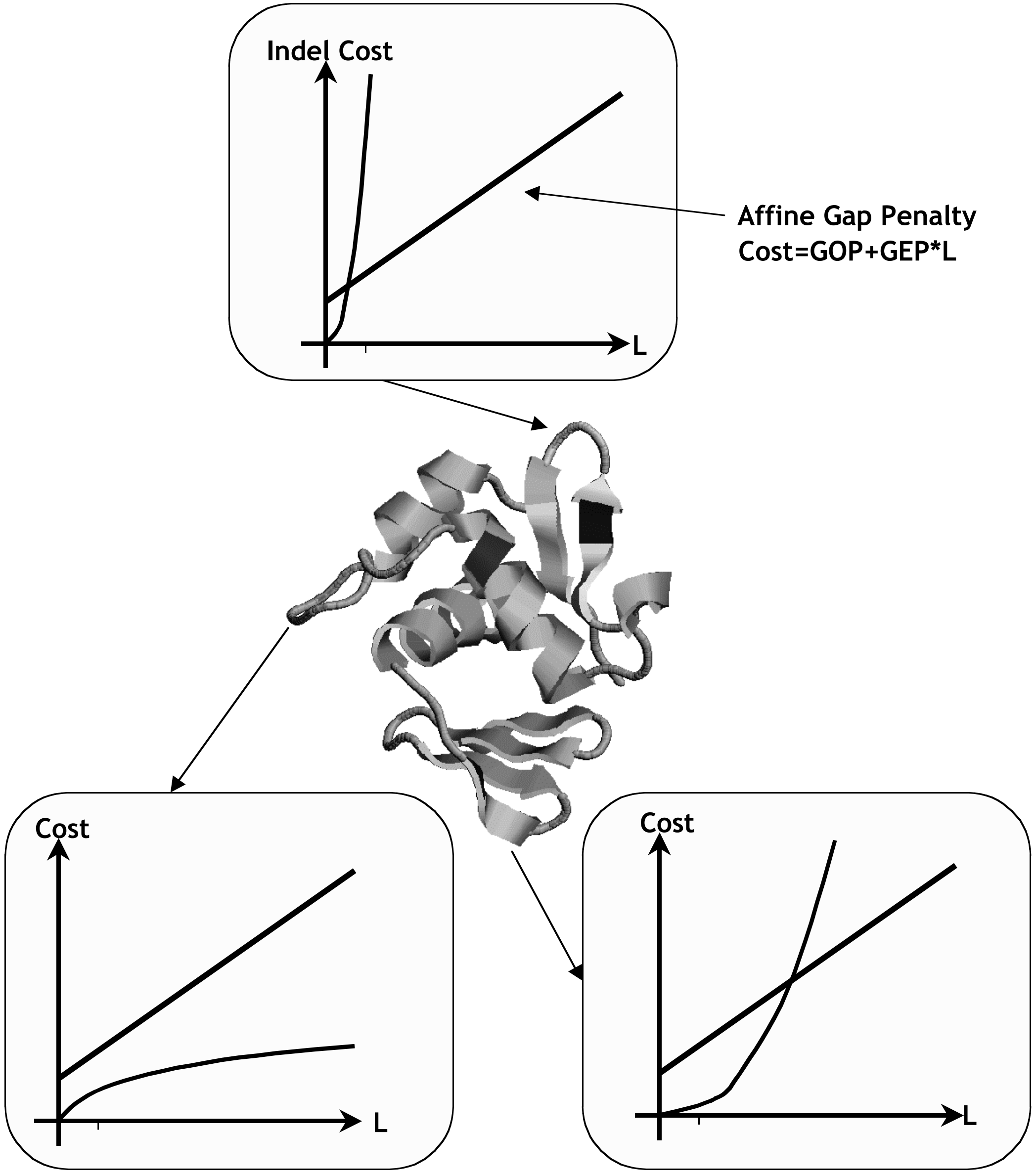
Family	S	NS
Histone3	6.4	0
Insulin	4.0	0.1
Interleukin I	4.6	1.4
α–Globin	5.1	0.6
Apolipoprot. AI	4.5	1.6
Interferon G	8.6	2.8

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

F A S T
F A T

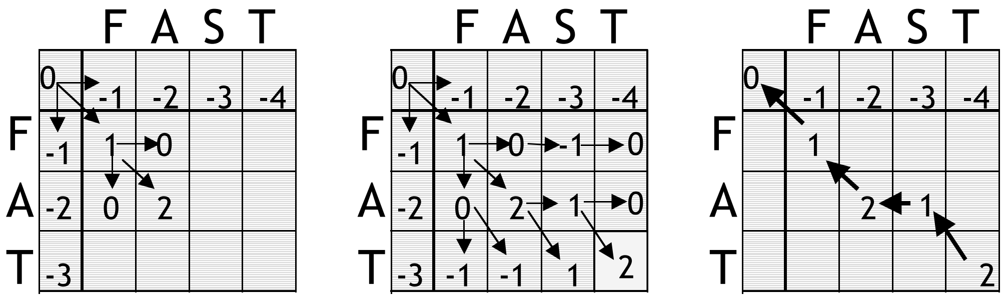
```
----FAT
FAST---

---FAT-
FAST---

--F-AT-
FAST---
```

⋮

$$\left(\frac{(L1+l2)!}{(L1)!*(L2)!}\right)^2$$

Dynamic Programming (Needlman and Wu

Match=1 MisMatch=-1 Gap=-1



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

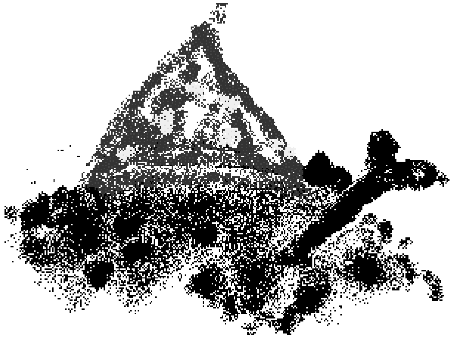
		F	A	S	T
	0	-1	-2	-3	-4
F	-1	1	0	-1	0
A	-2	0	2	1	0
T	-3	-1	-1	1	2

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

F A S T
F A - T

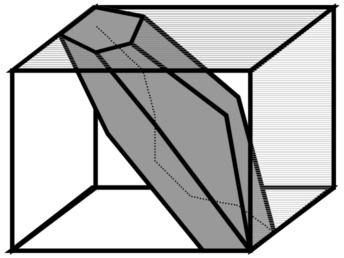
HOW CAN I ALIGN *MANY* SEQUENCES

7 Globins =>1000 years



Existing Methods

1-Carillo and Lipman:

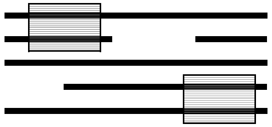


-MSA, DCA.

-Few Small Closely Related Sequence.

-Do Well When They Can Run.

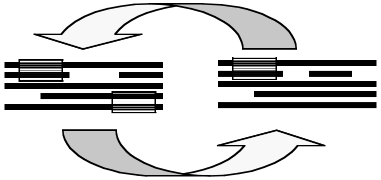
2-Segment Based:



-DIALIGN, MACAW.

-May Align Too Few Residues

3-Iterative:



-HMMs, HMMER, SAM.

-Slow, Sometimes Innacutrate

-Good Profile Generators

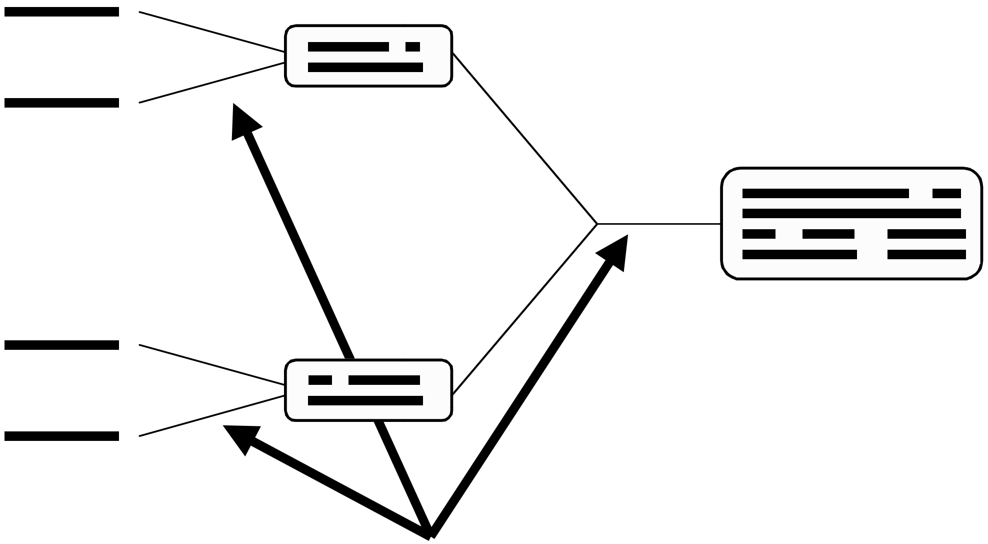
4-Progressive:

-ClustalW, Pileup, Multalign…

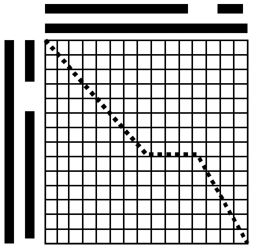
-Fast and Sensitive

Progressive Alignment

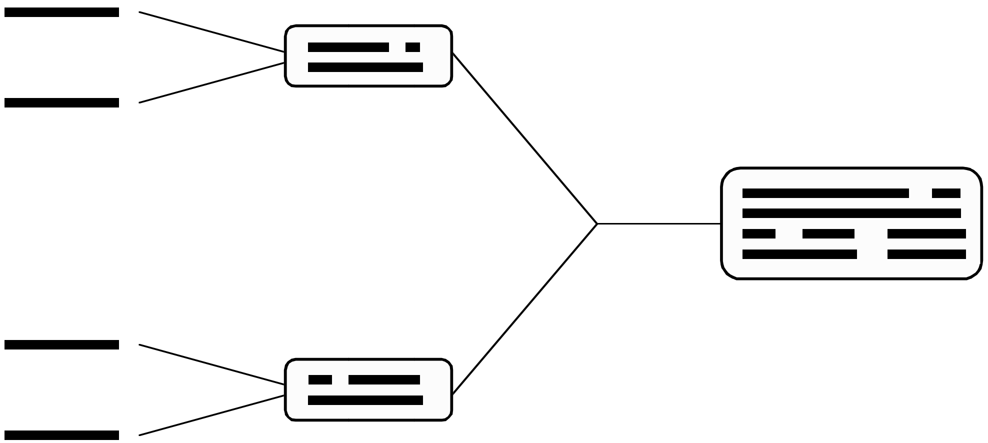
Feng and Dolittle, 1980; Taylor 1981



Dynamic Programming Using A Substitution Matrix



Progressive Alignment



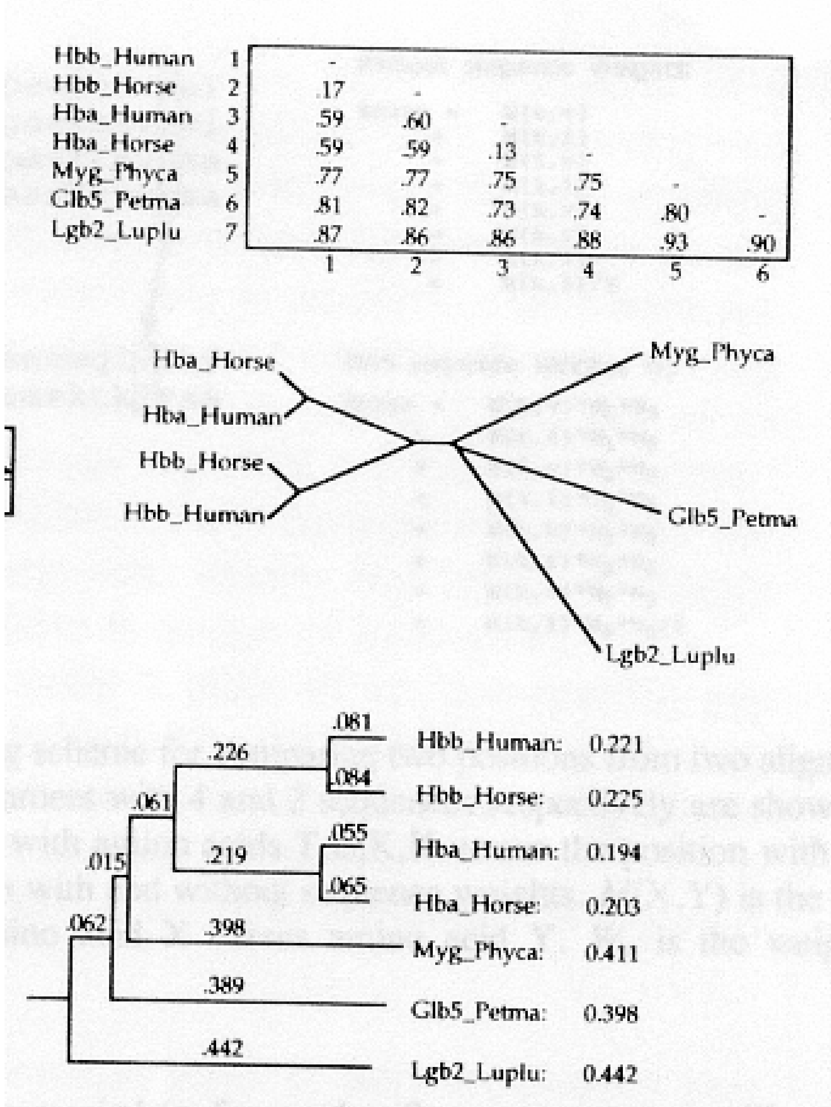
-Depends on the CHOICE of the sequences.

-Depends on the ORDER of the sequences (Tree).

-Depends on the PARAMETERS:

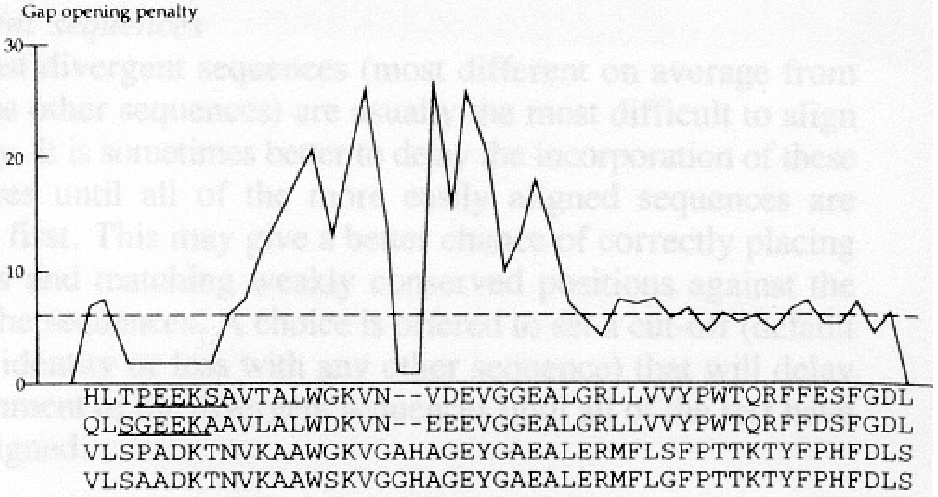
- Substitution Matrix.
- Penalties (Gop, Gep).
- Sequence Weight.
- Tree making Algorithm.

Weighting Within ClustalW



		1	2	3	4	5	6
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	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90

Position Specific GOP



```
HLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDL
QLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDL
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLS
```