ASAB – Week 5
Evolutionary sequence alignments (Dynamic Programming)

Algorithms for Sequence Analysis in Bioinformatics

Arnau Cordomí
arnau.cordomi@esci.upf.edu

Substitution matrices

SCORING SYSTEM

match score = 1

mismatch score = 0

THEFASTCAT

THELASTCAT 9

THELASTRAT 8

THEFASTCAT 10

	A	C	E	F	H	L	R	S	T
A	1	0	0	0	0	0	0	0	0
C	0	1	0	0	0	0	0	0	0
E	0	0	1	0	0	0	0	0	0
F	0	0	0	1	0	0	0	0	0
H	0	0	0	0	1	0	0	0	0
L	0	0	0	0	0	1	0	0	0
R	0	0	0	0	0	0	1	0	0
S	0	0	0	0	0	0	0	1	0
T	0	0	0	0	0	0	0	0	1

Identity matrix. Only appropriate for very similar sequences

	A	C	E	F	H	L	R	S	T
A	1	?	?	?	?	?	?	?	?
C	?	1	?	?	?	?	?	?	?
E	?	?	1	?	?	?	?	?	?
F	?	?	?	1	?	?	?	?	?
H	?	?	?	?	1	?	?	?	?
L	?	?	?	?	?	1	?	?	?
R	?	?	?	?	?	?	1	?	?
S	?	?	?	?	?	?	?	1	?
T	?	?	?	?	?	?	?	?	1

A much better matrix

How can we compare sequences ?

To compare sequences, we need to compare residues

We need to know how much it costs to substitute

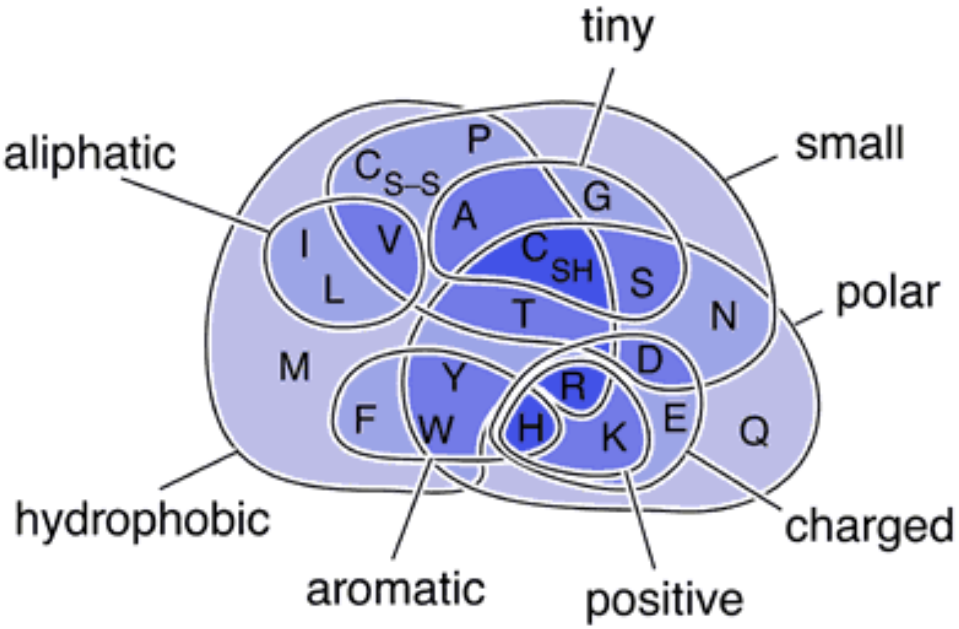
an **Alanine** into an **Isoleucine**
a **Tryptophan** into a **Glycine**
…

The table that contains the costs for all the possible substitutions is called the **SUBSTITUTION MATRIX**

How to derive that matrix?

https://commons.wikimedia.org/wiki/File:Amino_Acids_Venn_Diagram.png

Using **knowledge**?



… But we do not know enough about structure and evolution.

Much better if we use **data!!**

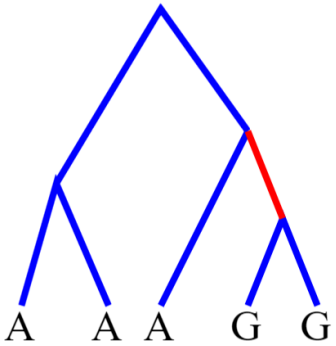
Margaret Dayhoff (1970')
PAM substitution matrices

- Took 71 pairs of Protein Sequences, easy to align (85% identical).
- Aligned them
- Counted all mutations in the alignments
    - 25 Tryptophan into Phenylalanine
    - 30 Isoleucine into Leucine
    - ...
- Computed all the scores

Which should be the MSA score of this column?

A
A
A
G
G

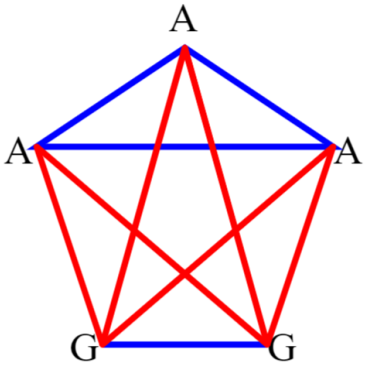
Model 1



How did this happen?
In one mutation?

cost (A-G) = 1

Model 2



cost (A→ G) = 6

Easy to compute

Over-estimation of the substitutions

S1: BABA
S2: AAAC
S3: AACC
S4: AABA
S5: AACC
S6: AABC

X X	Counts	Observed frequencies
A A	26 x 2 = 52	52/120
A B + B A	8 x 2 = 16	16/120
A C + C A	10 x 2 = 20	20/120
B B	3 X 2 = 6	6/120
B C + C B	6 X 2 = 12	12/120
C C	7 X 2 =14	14/120
Total	**60 x 2 = 120**	

Is this a lot?

BB: 1-4, 1-6, 4-6 4-1, 6-1, 6-4

BC: 1-3, 1-5, 3-4, 3-6, 4-5, 5-6 3-1, 5-1, 4-3, 6-3, 5-4, 6-5

Log odds ratio

For each mutation, set the substitution score to the log odds ratio:

$$\text{Log}\left(\frac{\text{Observed}}{\text{Expected by chance}}\right)$$

$$\log\left(\frac{p_{ij}}{q_i * q_j}\right)$$

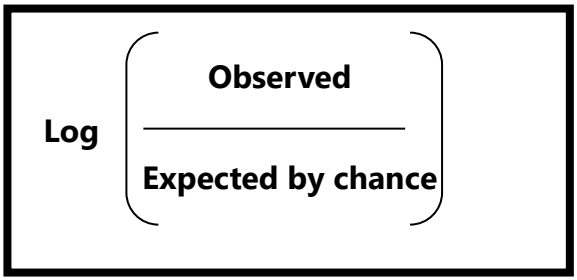
p_{ij} probability of aa_i -> aa_j transition
q_i probability of aa_i
q_j probability of aa_j

Hypothesis we wish to test; two amino acids are correlated because they are homologous.

$$\text{score} = \log \text{ odds ratio} = s_{AB} \propto \log\left(\frac{\text{observed}}{\text{expected}}\right)$$

Null hypothesis; two amino acids occur independently (and are uncorrelated and unrelated).

log odds ratio of the likelihood of a point accepted mutation from one amino acid to another to the likelihood that these amino acids were aligned by chance.



If observed

= chance → 0

> chance → positive

< chance → negative

log (10) = 1 log (0.1) = -1

log (100) = 2 log (0.01) = -2

log (1000) = 3 log (0.001) = -3

Log space makes values symmetrical

1000x more likely vs. 1000x less likely

Compute a substitution matrix: expected frequencies

S1: BABA
S2: AAAC
S3: AACC
S4: AABA
S5: AACC
S6: AABC

X	Counts	Freqs
A	14	14/24
B	4	4/24
C	6	6/24
Total	**24**	**1**

Amino acid frequencies

X X	Expected frequencies
A A	(14/24)*(14/24) = 0.34
A B = B A	(14/24)*(4/24) = 0.10
A C = C A	(14/24)*(6/24) = 0.15
B B	(4/24)*(4/24) = 0.03
B C = C B	(4/24)*(6/24) = 0.04
C C	(6/24)*(6/24) = 0.06

Expected frequencies

Compute a substitution matrix

S1: BABA
S2: AAAC
S3: AACC
S4: AABA
S5: AACC
S6: AABC

XX	Observed Frequency (O)	Expected Frequency (E)	log (O / E) x 10
A A	52/120	(14/24)*(14/24)	1.04
A B + B A	16/120	(14/24)*(4/24)	-1.64
A C + C A	20/120	(14/24)*(6/24)	-2.43
B B	6/120	(4/24)*(4/24)	2.55
B C + C B	12/120	(4/24)*(6/24)	0.79
C C	14/120	(6/24)*(6/24)	2.71

S1: BABA
S2: AAAC
S3: AACC
S4: AABA
S5: AACC
S6: AABC

The substitution Matrix

	A	B	C
A	1.04	-1.64	-2.43
B	-1.64	2.55	0.79
C	-2	0.79	2.71

	A	B	C
A	1	-2	-2
B	-2	3	1
C	-2	1	3

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	O	T	Y	V
A	2																			
R	−2	6																		
N	0	0	2																	
D	0	−1	2	4																
C	−2	−4	−4	−5	12															
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	−3	−4	−3	−6	−4	−5	−5	−5	−2	1	2	−5	0	9						
P	1	0	0	−1	−3	0	−1	0	0	−2	−3	−1	−2	−5	6					
O	1	0	1	0	0	−1	0	1	−1	−1	−3	0	−2	−3	1	2				
T	1	−1	0	0	−2	−1	0	0	−1	0	−2	0	−1	−3	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

Scores for each possible substitution

The diagonal indicates preservation

"Some residues are more/less likely to align with other than expected"

Why half matrix?
Is A → W the same as W → A?

PAM250 substitution matrix

Cysteins form disulphide bridges

small

acidic/ polar

basic

aliphatic

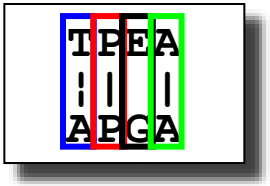
aromatic

	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	

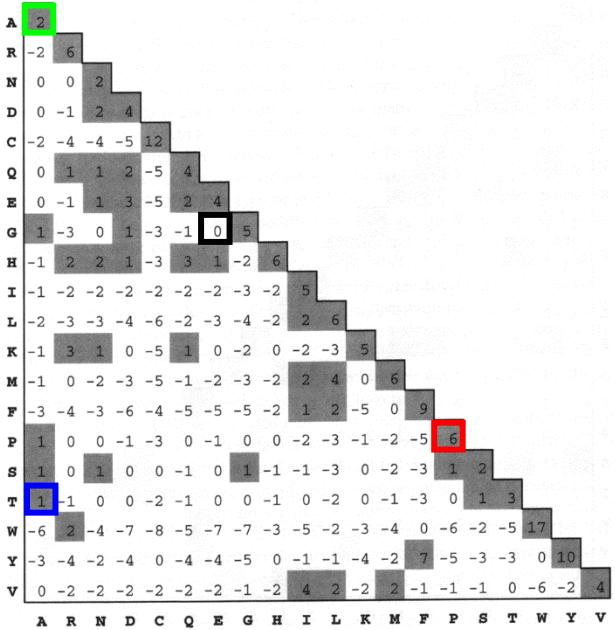
Tryptophan is large

BLOSUM62 substitution matrix

Scoring an Alignment



Score = 1 + 6 + 0 + 2 = 9



Is it a lot?

Is it possible to get such a good alignment by chance only?

PAM is the unit of evolutionary distance between two sequences.

1PAM: 1 point accepted mutation in 100 aa.

$$PAMn = (PAM\ 1)^n$$

30 PAM = 30 point mutations/100 aa

160 PAM = 160 point mutations/100 aa

250 PAM = 250 point mutations/100 aa

} Mutations could occur multiple times at any given position

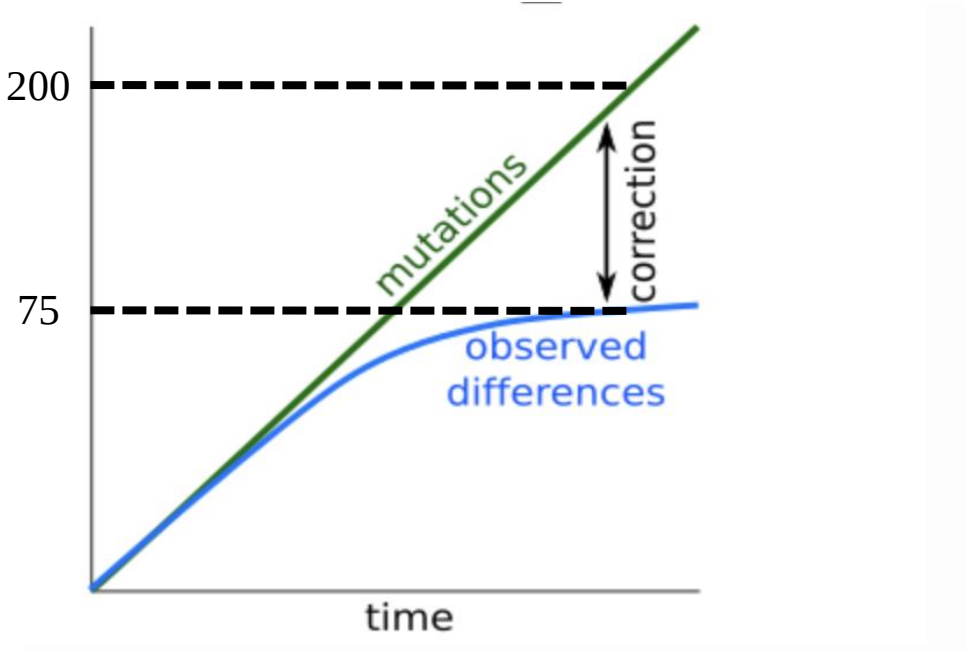
PAM family of matrices:

PAM1 - PAM30 - PAM70 - PAM160 - PAM250

Evolutive time – Sequence divergence

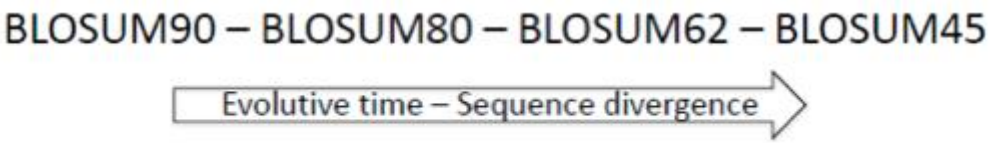
Margaret Dayhoff 1978

PAM200: 200 point mutations in 100 amino acids



- Developed by using **multiple sequence alignments of conserved regions** (BLOCKs database) in evolutionary divergent proteins

- Obtained from alignments created by clustering sequences that were more similar than a given % (indicated in the matrix)



- For closely related sequences: low PAM or high BLOSUM
- For distantly related sequences: high PAM or low BLOSUM
- BLOSUM matrices usually perform better than PAM matrices

Steven and Jorja Henikoff 1992

Each base, amino acid, portion of a genome …. is unique

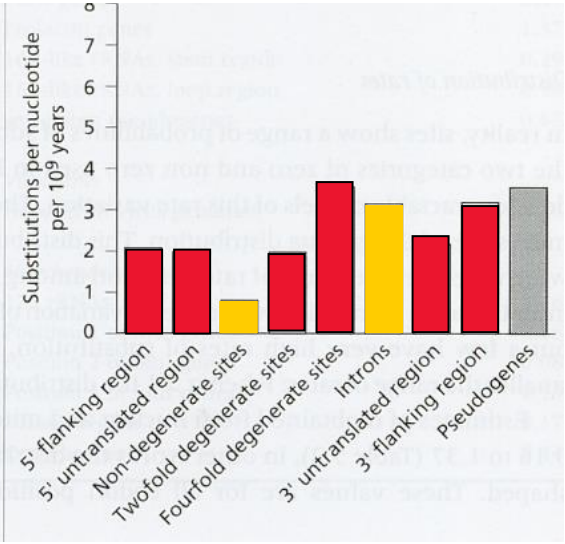
Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences.

Makalowski W, Boguski MS.

Proc Natl Acad Sci U S A. **1998** Aug 4;95(16):9407-12.

PMID: 9689093 **Free PMC Article** Paperpile

Similar articles



One particular **ADENINE**

- Coding vs. noncoding region
- Binding to the transcriptional machinery, regulatory elements.
- Binding to histones
- Hypersensitive sites
- DNA methylation, histone modifications
- Responsible for the 3D organization of the genome
- …

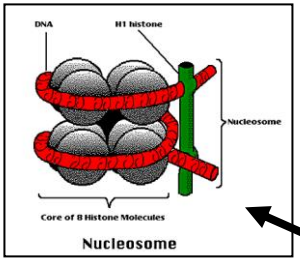
Each protein is unique

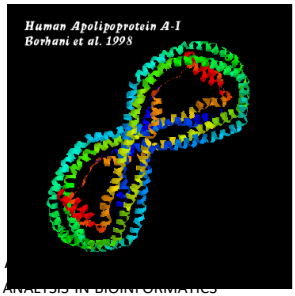
Constrained genome positions evolve slowly

Every protein family has its own level of constraint

Rates in Substitutions/site/Billion Years as measured on Mouse Vs Human (80 Million years)

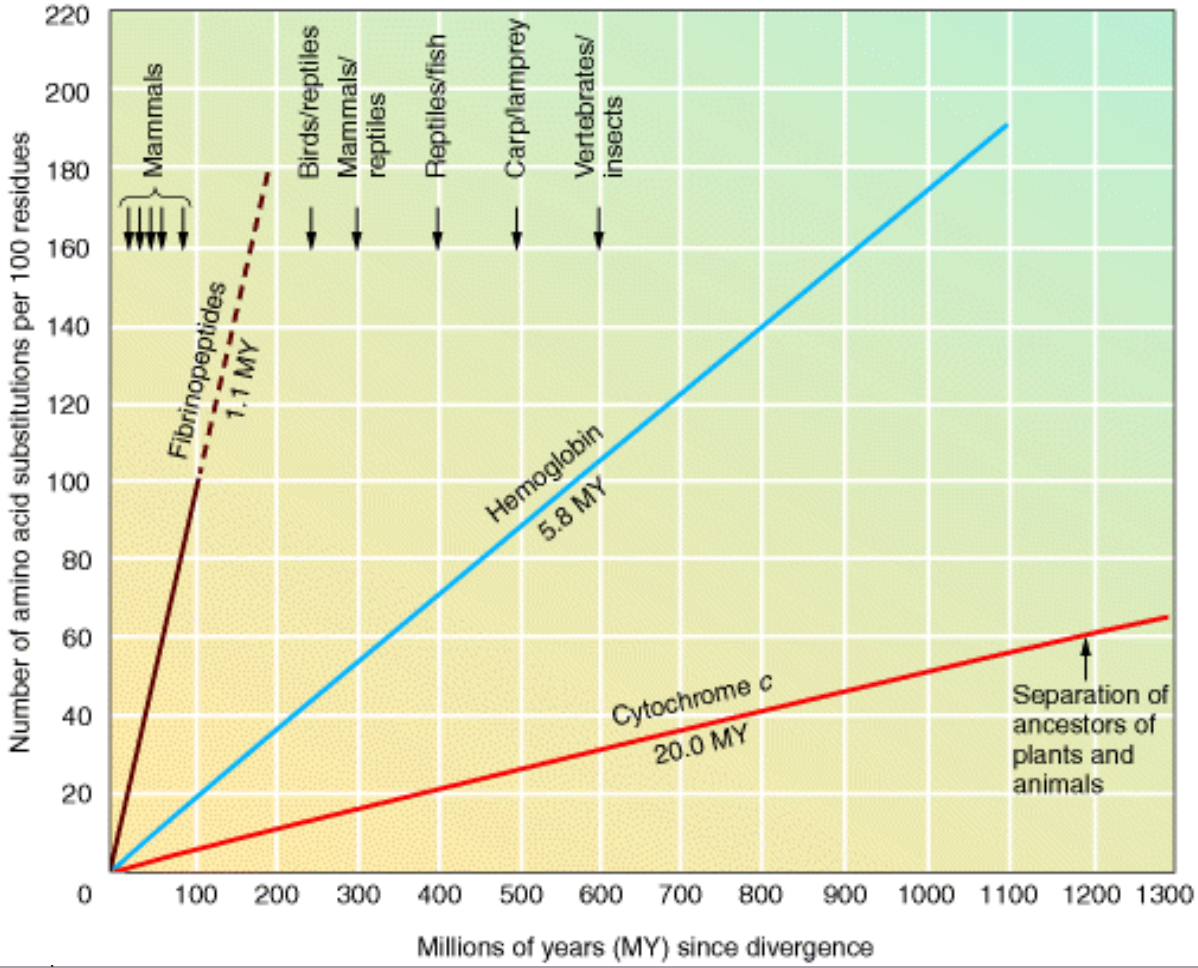
Ks Synonymous Mutations **Ka** Non-Neutral.



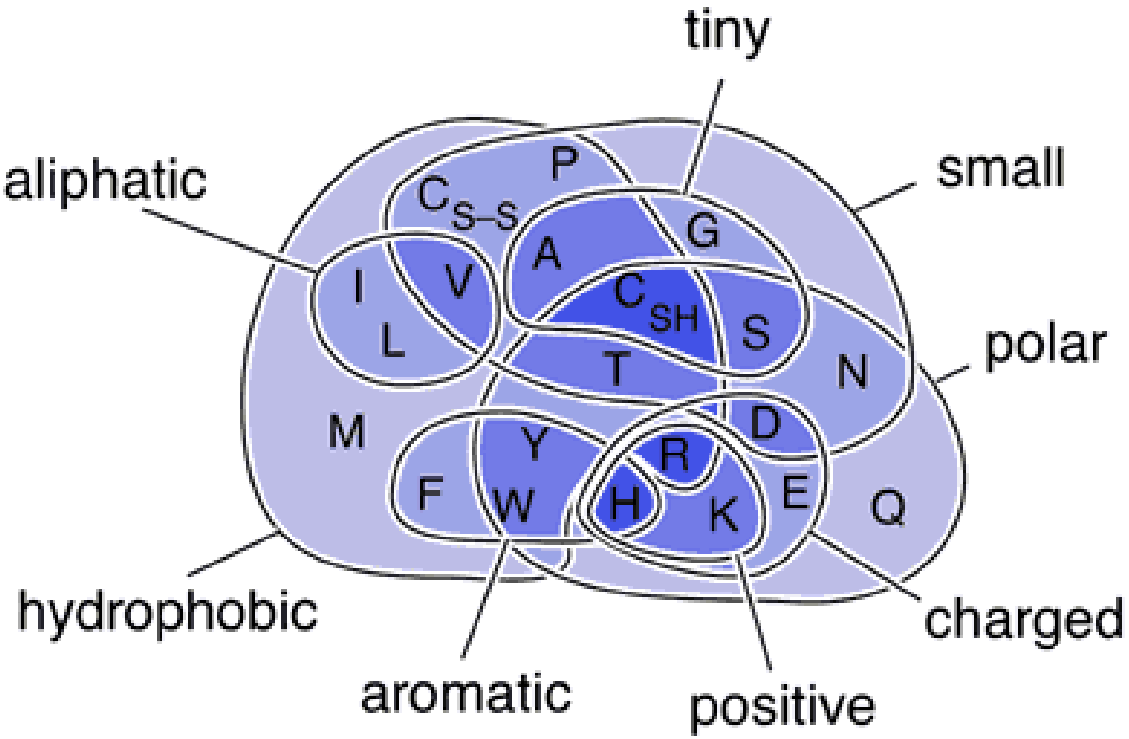


Family	K_S	K_A
Histone3	6.4	0
Insulin	4.0	0.1
Interleukin I	4.6	1.4
α-Globin	5.1	0.6
Apolipoprot. AI	4.5	1.6
Interferon G	8.6	2.8

Different molecular clocks for different proteins

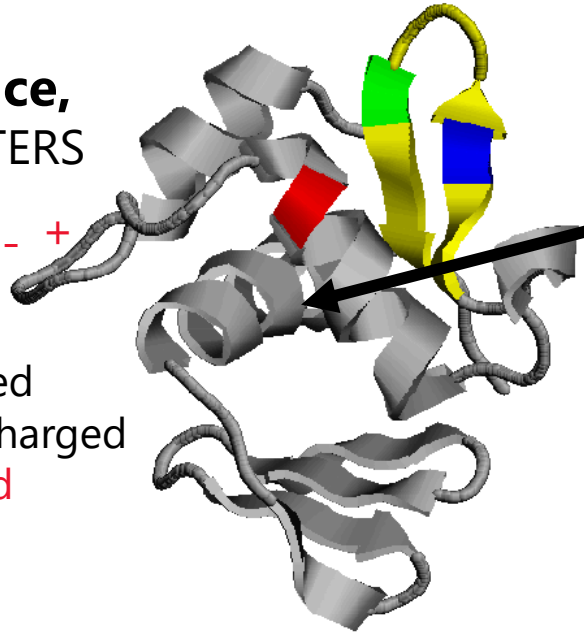


Every residue has its own personality

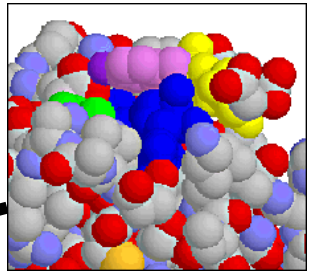


Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation.
Livingstone CD, **Barton** GJ.
Comput Appl Biosci. 1993 Dec;9(6):745-56.
PMID: 8143162 Paperpile

On the surface,
CHARGE MATTERS



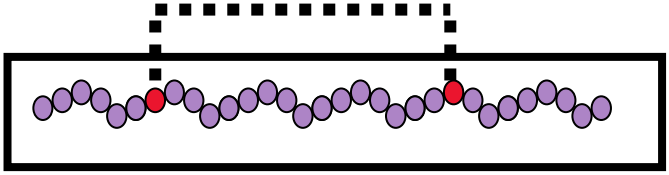
Charged → Charged
Uncharged → Uncharged
Indels are tolerated



In the core,
SIZE MATTERS

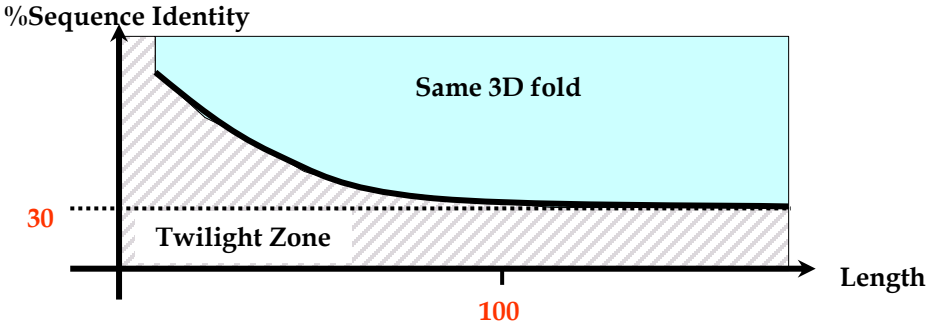
Big -> Big
Small ->Small
Indels are less tolerated

- Ignore non-local interactions



- Assume the evolution rate to be constant
  - *Not the case in different proteins*
  - *Not the case in different regions of the same protein*

- Substitution matrices only work well with similar sequences (> 30% id).



Global alignment
Needleman and Wunsch

Alignments up to $|s_1| + |s_2|$ characters long

```
THEFASTCAT---------
----------THEFATCAT
```

$$\binom{L_1 + L_2}{L_1} = \frac{(L_1 + L_2)!}{L_1!\, L_2!}$$

$L_1 = 10$

$L_2 = 9$

alignments = 92378

1o letters

```
THEFASTCAT
 THEFATCAT
  THEFATCAT
   THEFATCAT
    THEFATCAT
     THEFATCAT
      THEFATCAT
       THEFATCAT
        THEFATCAT
         THEFATCAT
```

```
THEFATCAT
```

9 letters

- DP invented in the 50s by Bellman

  Programming ⇔ Tabulation

- Has found applications in aerospace, engineering, economics ….
- Re-invented in 1970 by Needleman and Wunsch

  It took 10 year to find out…

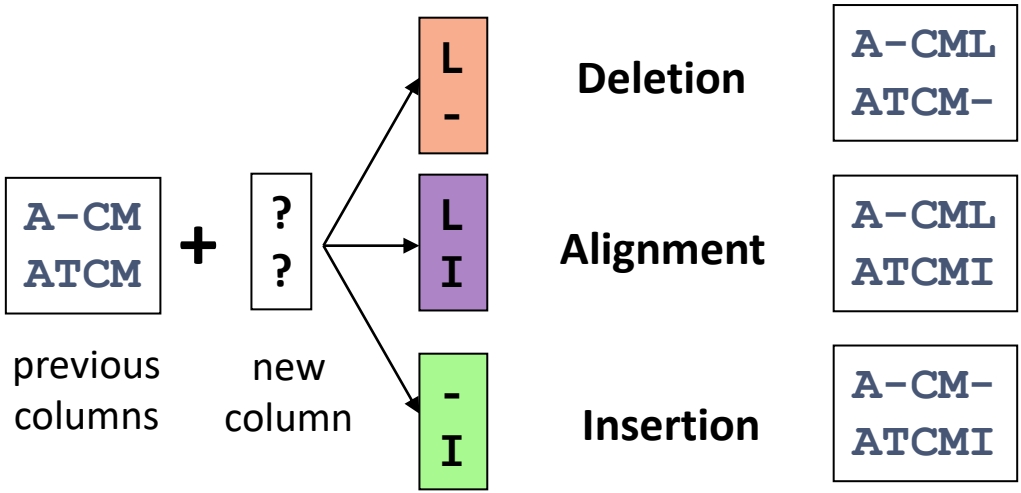
Dynamic programming usually refers to simplifying a decision by breaking it down into a sequence of decision steps over time

Makes foolish assumptions:

- The score of each column of the alignment is independent of the rest of the alignment

- It is possible to model the relationship between two sequences with a substitution matrix + a gap penalty

*If you **optimally** extend an **optimal** alignment of two sub-sequences, the result remains an **optimal** alignment*



Deletion/insertion in the second sequence relative to the first

Formalizing the algorithm (I)

Sequence 1: [1, 2 , 3, … i]
Sequence 2: [1, 2 , 3, … j]

Deletion (gap in seq 2)

1…i-1 1…j	+	i -

Aligned

1…i-1 1…j-1	+	i j

Insertion (gap in seq 1)

1…i 1…j-1	+	- j

Three ways to finish the alignment

1…i 1…j

-Sequence 1: **[1, 2 , 3, … i]**

-Sequence 2: **[1, 2 , 3, … j]**

F(i, j) = best

- F(i-1, j) + **GEP** → **1…i-1 / 1…j** + **X / -** i
- F(i-1, j-1) + **Mat(i, j)** → **1…i-1 / 1…j-1** + **X / Y** i j
- F(i, j-1) + **GEP** → **1…i / 1…j-1** + **- / Y** j

GEP = Gap (Extension) Penalty

Arranging everything in a table

seq2 [1, … j]

seq1 [1, … i]

	–	F	A	T
–				
F		1…I-1 1…J-1	1…I-1 1…J	
A		1…I 1…J-1	1…I 1…J	
S				
T				

score of aligning substrings (FA with FA)

F A
F A

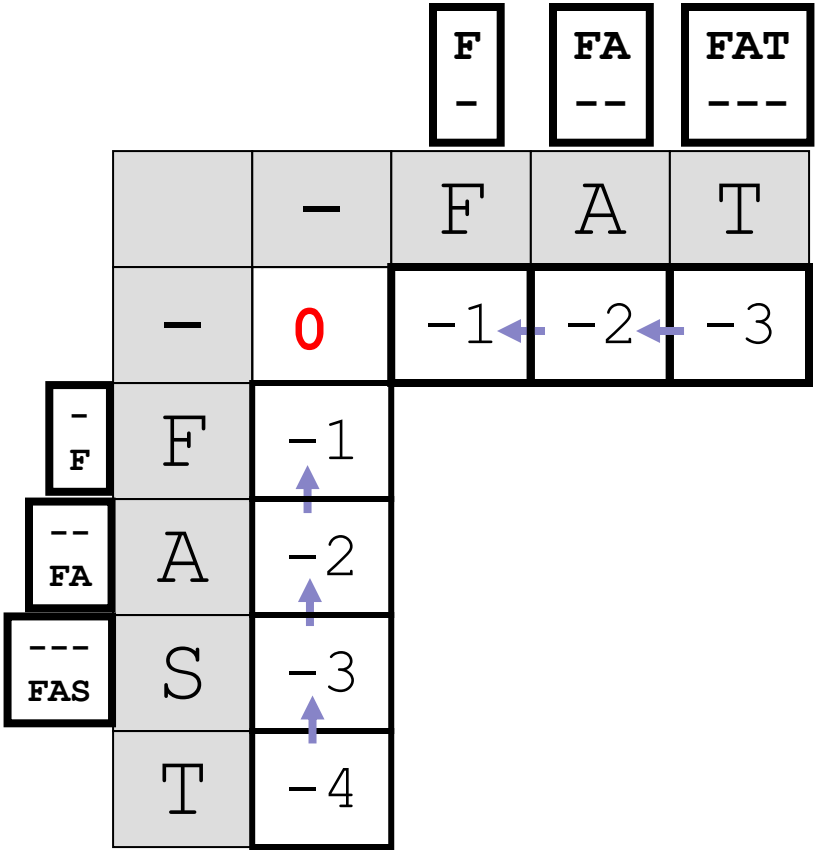
[FA]- Gap in seq2

F A Ins. in seq1

F A Ins. in seq2

[FA]- Gap in seq1

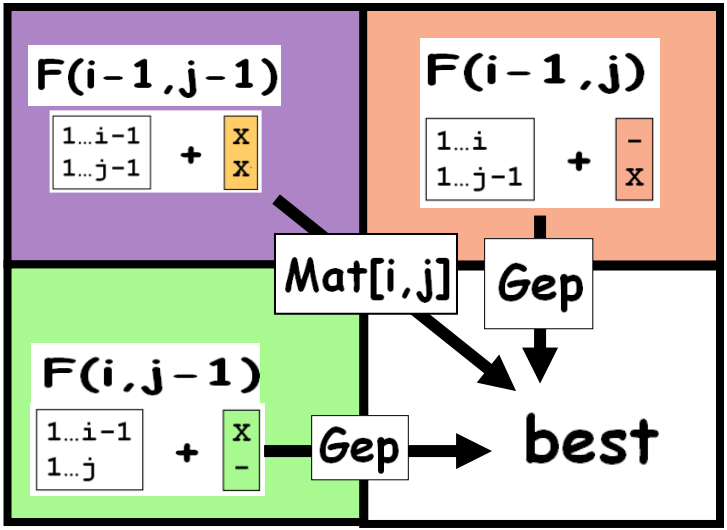
F(I, J): row, col



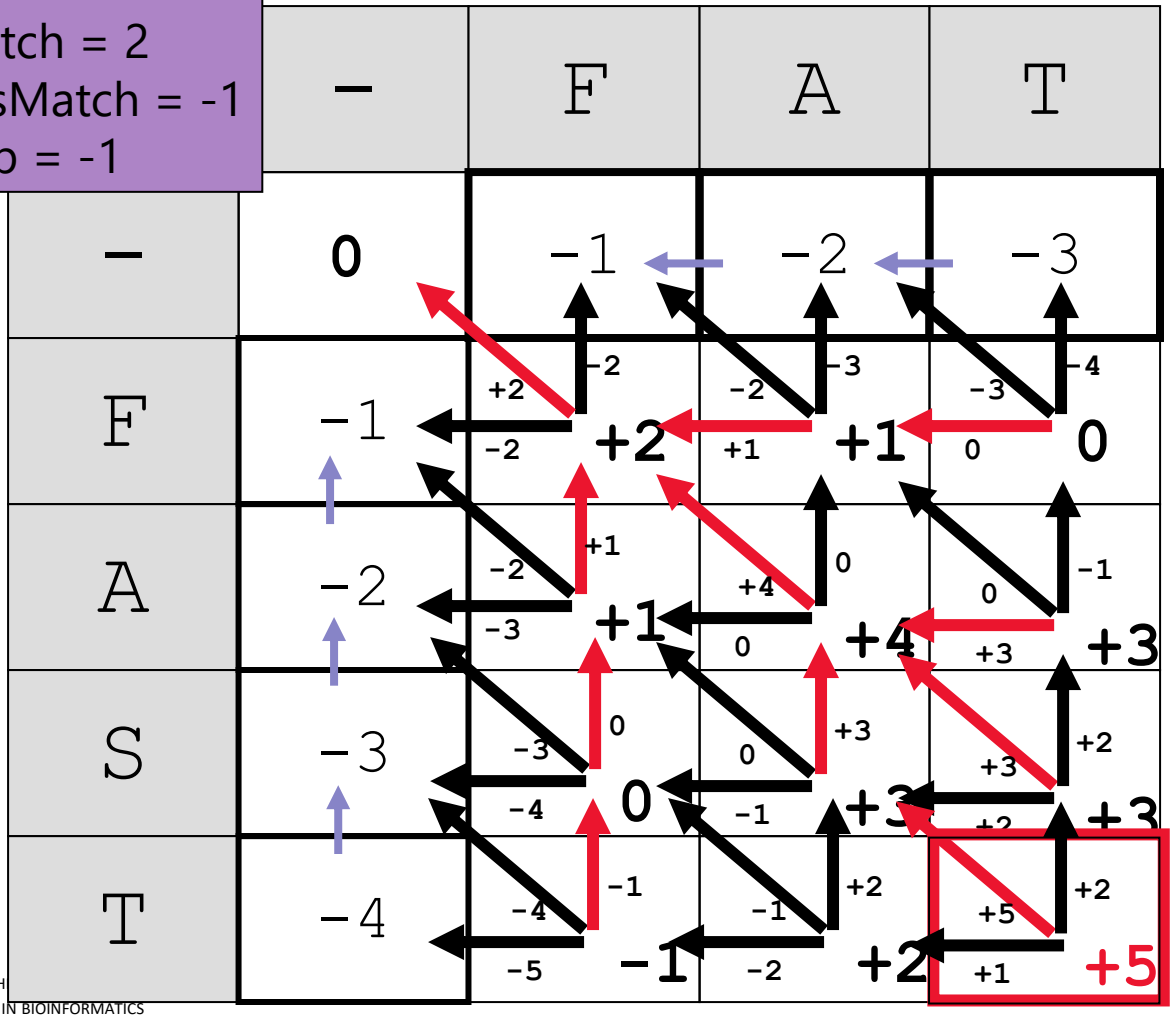
Match = 2
MisMatch = -1
Gap = -1

Filing Up The Matrix

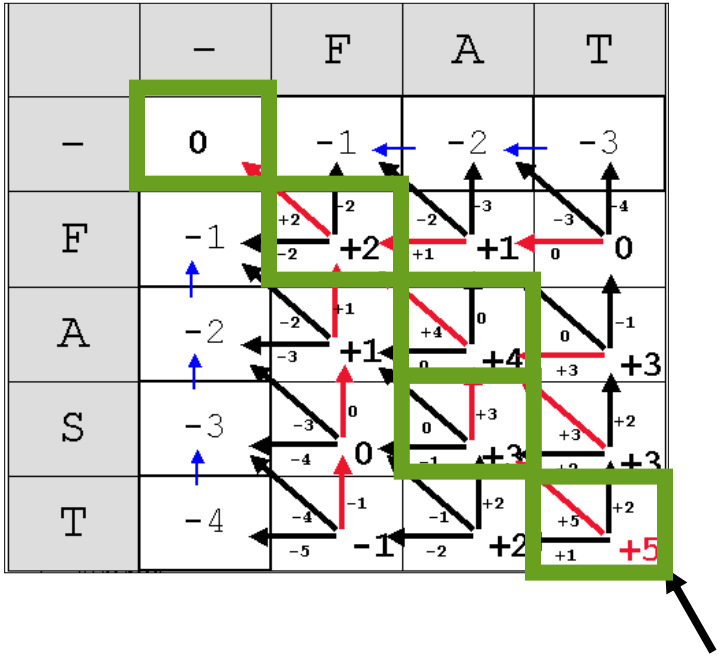
$$F(i,j) = \text{best} \begin{cases} F(i,j-1) + Gep & \begin{matrix}1\ldots i \\ 1\ldots j-1\end{matrix} + \begin{matrix}-\\ x\end{matrix} \\ F(i-1,j-1) + Mat[i,j] & \begin{matrix}1\ldots i-1 \\ 1\ldots j-1\end{matrix} + \begin{matrix}x\\ x\end{matrix} \\ F(i-1,j) + Gep & \begin{matrix}1\ldots i-1 \\ 1\ldots j\end{matrix} + \begin{matrix}x\\ -\end{matrix} \end{cases}$$



Match = 2
MisMatch = -1
Gap = -1



Delivering the alignment: Trace-back



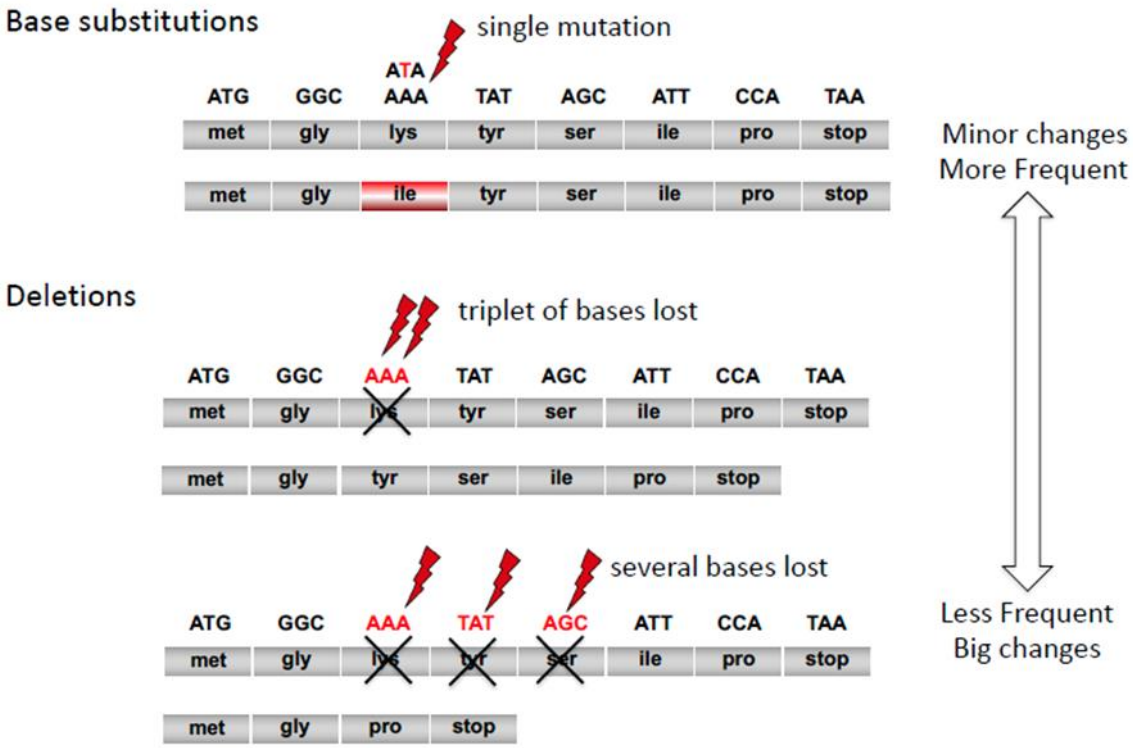
T	–	A	F
T	S	A	F

Optimal Aln Score: Score of FAT vs FAST

Needleman and Wunsch algorithm:
Global alignment without affine penalties

Adding affine penalties (Gotoh algorithm)

Substitutions vs. Indels



Indels (in proteins) involve more drastic changes than substitutions

GOP GOP

AVT--GFTGH
AVATAGFTGH

This requires one event

GOP GOP

AV-T-GFTGH
AVATAGFTGH

This requires two events

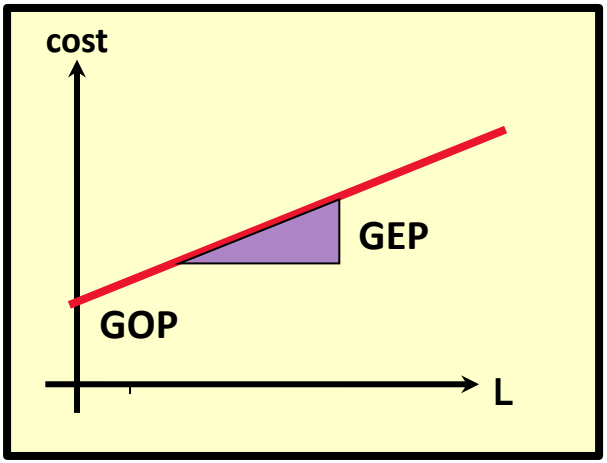
When a gap occurs, several adjacent residues may be involved

Hypothesis: Evolution follows maximum parsimony. The simplest path (fewer changes) is the most likely.

LMNTG----NT
LMNTGGGGGNT

Gap open penalty **+ 3 x** **Gap extension penalty**

GOP > GEP



Affine Gap Penalty

In the context of BLOSUM62, typically:

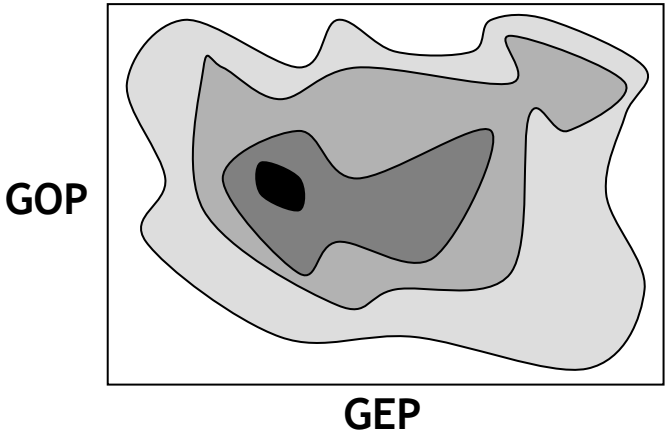
GOP = -11
GEP = -1

$$cost = GOP + L * GEP$$

or

$$cost = GOP + (L - 1) * GEP$$

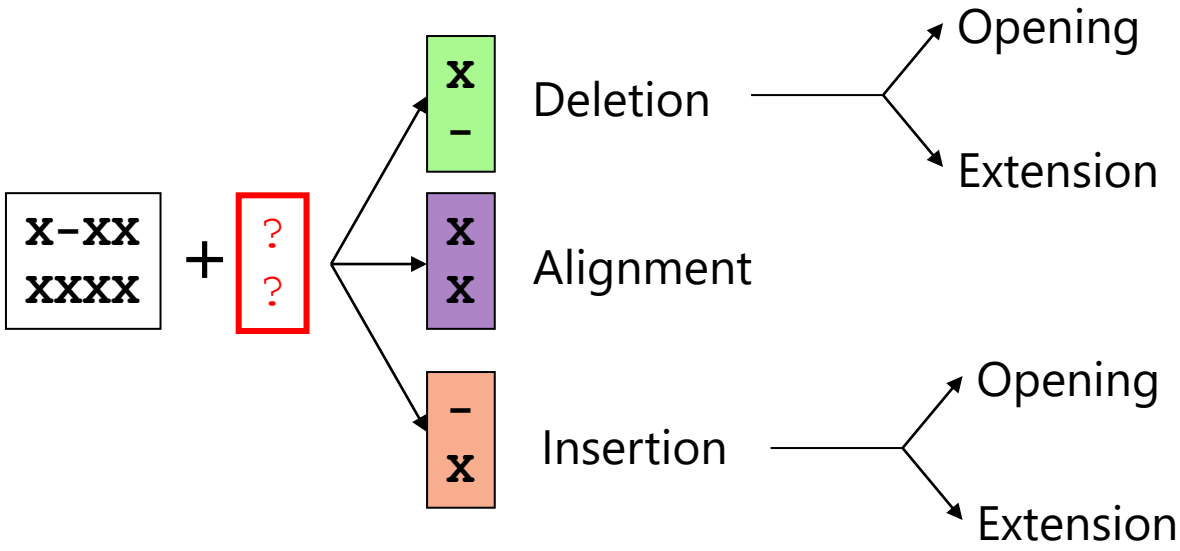
Which values of GEP and GOP should I use?



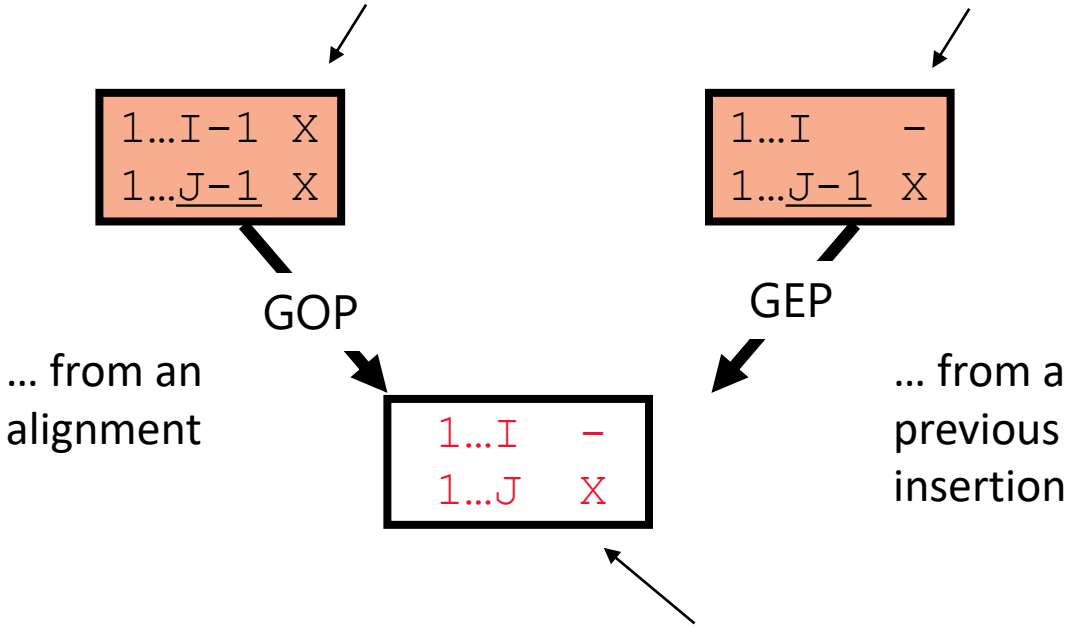
Global Alignments are very sensitive to gap Penalties

The same as Needlemann-Wunsch, but harder.

More than 3 ways to extend an alignment



What is the cost of an insertion ?



Solution: maintain 3 tables

M: Table that contains the score of every optimal alignment 1…I vs 1…J that finishes with an alignment between sequence X and Y

Ix: Table that contains the score of every optimal alignment 1…i vs 1…j that finishes with an Insertion in sequence X.

Iy: Table that contains the score of every optimal alignment 1…I vs 1…J that finishes with an Insertion in sequence Y.

$$M(i,j) = \text{best} \begin{cases} M(i-1,j-1) + Mat(i,j) \\ Ix(i-1,j-1) + Mat(i,j) \\ Iy(i-1,j-1) + Mat(i,j) \end{cases}$$

$$\begin{bmatrix} 1\ldots i-1 \\ 1\ldots j-1 \end{bmatrix} + \begin{bmatrix} X \\ X \end{bmatrix}$$

Three possible values!

$$Ix(i,j) = \text{best} \begin{cases} M(i-1,j) + gop \\ Ix(i-1,j) + gep \end{cases}$$

$$\begin{bmatrix} 1\ldots i-1 & X \\ 1\ldots j & X \end{bmatrix} + \begin{bmatrix} X \\ - \end{bmatrix} \qquad \begin{bmatrix} 1\ldots i-1 & X \\ 1\ldots j & - \end{bmatrix} + \begin{bmatrix} X \\ - \end{bmatrix}$$

$$Iy(i,j) = \text{best} \begin{cases} M(i,j-1) + gop \\ Iy(i,j-1) + gep \end{cases}$$

$$\begin{bmatrix} 1\ldots i & X \\ 1\ldots j-1 & X \end{bmatrix} + \begin{bmatrix} - \\ X \end{bmatrix} \qquad \begin{bmatrix} 1\ldots i & - \\ 1\ldots j-1 & X \end{bmatrix} + \begin{bmatrix} - \\ X \end{bmatrix}$$

Filing up a SW matrix

x FAT

y FAAAT

Match = 1
MisMatch = -1
GOP = -3
GEP = -1

Iy		F_1	A_2	A_3	A_4	T_5
		-∞	-∞	-∞	-∞	-∞
F_1	_	-8	-9	-10	-11	-12
A_2	_	-3	-7	-8	-9	-10
T_3	_	-4	-2	-6	-7	-8

M		F_1	A_2	A_3	A_4	T_5
	0	-4	-5	-6	-7	-8
F_1	-4	1	-3	-4	-5	-6
A_2	-5	-3	2	-2	-3	-4
T_3	-6	-4	-2	1	-3	-2

Ix		F_1	A_2	A_3	A_4	T_5
		_	_	_	_	_
F_1	-∞	-8	-3	-4	-5	-6
A_2	-∞	-9	-7	-2	-3	-4
T_3	-∞	-10	-8	-6	-3	-4

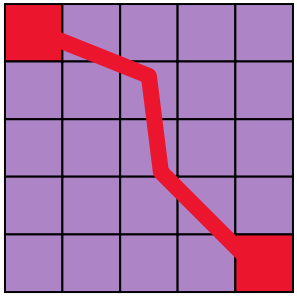
Start From BEST { M(i,j), Ix(i,j), Iy(i,j)

```
F__AT
*  **
FAAAT
```

```
FA__T
**  *
FAAAT
```

Local alignment
Smith and Waterman

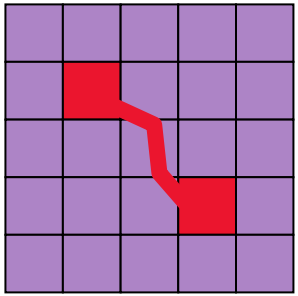
GLOBAL Alignment



- End-to-end alignment (contains all letters from both sequences)
- Suitable for closely related sequences (homologous genes; similar sequence of similar length)

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||||    |||||||  ||||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

LOCAL Alignment



- Aligns a substring to a substring
- Finds local regions with the highest similarity (ignoring the rest)
- Suitable for aligning distantly related sequences

```
5'          TACTTACGGATCAGGTACTTTAGAGGCT          3'
            |||| |||||| ||||||||||||||||
       5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

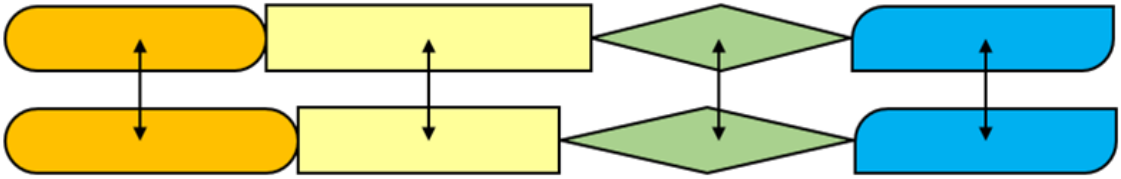
GLOBAL ALIGNMENT

Align all letters

Needleman & Wunsch

Seq1

Seq2



THEFASTFASTCAT
THEFATFASTRAT

THEFASTFASTCAT
THEFA-TFASTRAT

Smith & Waterman

LOCAL ALIGNMENT

Align some letters (one domain)

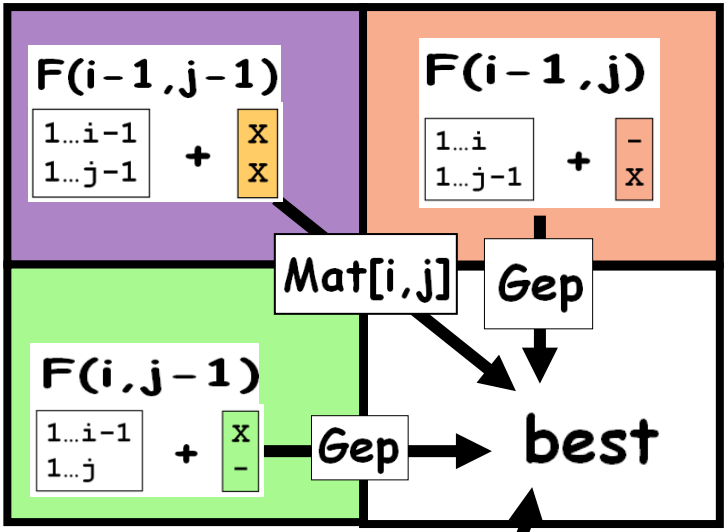
Seq1

Seq2

AVEYRYFASTCAT
AFATNICERAT

FAST
FA-T

Smith And Waterman (1981) = variation of Needleman and Wunsch to do LOCAL alignment



Ignore the rest of the previous cells and start a local alignment

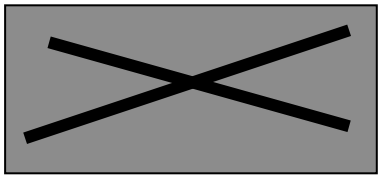
0

→ negative scoring cells are set to zero

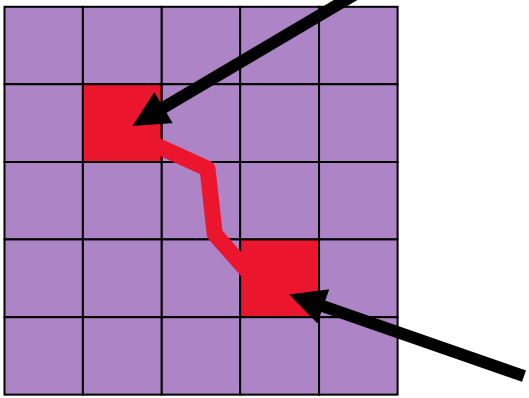
Formalizing the algorithm

$$F(i, j) = \text{best} \begin{cases} F(i-1, j) + \textbf{GEP} \\ F(i-1, j-1) + \textbf{Mat(i, j)} \\ F(i, j-1) + \textbf{GEP} \\ 0 \end{cases}$$

F(i-1, j) + **GEP**	1…i-1 1…j	+	X -	i
F(i-1, j-1) + **Mat(i, j)**	1…i-1 1…j-1	+	X Y	i j
F(i, j-1) + **GEP**	1…i 1…j-1	+	- Y	j



start fresh



Ignore the previous cells and start a local alignment

score = 0

local alignments NEVER start/end with a gap… (GEP is < 0)

The alignment terminates when the score starts to decrease

score = the highest in the matrix

Filing up the SW matrix and traceback

THECATISFAST

AFASTCAT

Match = 1
MisMatch = -1
Gap = -2

Beginning of the trace-back: at the best local score

The matrix of scores only contains positive numbers and 0

Limits: local alignments NEVER start/end with a gap…

S		A_1	F_2	A_3	S_4	T_5	C_6	A_7	T_8
	0	0	0	0	0	0	0	0	0
T_1	0	0	0	0	0	1	0	0	1
H_2	0	0	0	0	0	0	0	0	0
E_3	0	0	0	0	0	0	0	0	0
C_4	0	0	0	0	0	0	1	0	0
A_5	0	1	0	1	0	0	0	2	0
T_6	0	0	0	0	0	1	0	0	3
I_7	0	0	0	0	0	0	0	0	1
S_8	0	0	0	0	1	0	0	0	0
F_9	0	0	1	0	0	0	0	0	0
A_{10}	0	1	0	2	0	0	0	1	0
S_{11}	0	0	0	0	3	1	0	0	0
T_{12}	0	0	0	0	1	4	2	0	1
Score: 4									