Homology ⟶ Similarity

mismatch

# of exact matches

score

	A	C	G	T
A	1			0.9
C		1		
G			1	
T				1

$$S = \sum_{i=1}^{n} s(x_i, y_i)$$

Sum of scores model

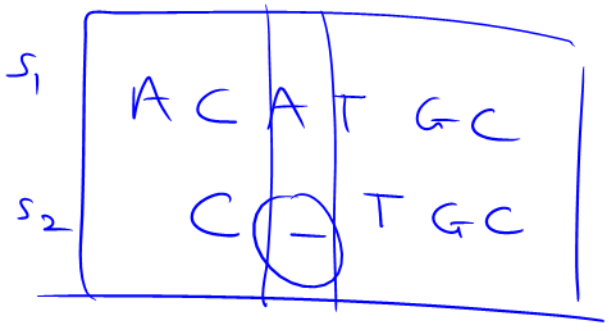
ACATGC
CTGC

ACATGC
CTGC

ACATGC
CTGC

ACATGC
| | |
CTGC

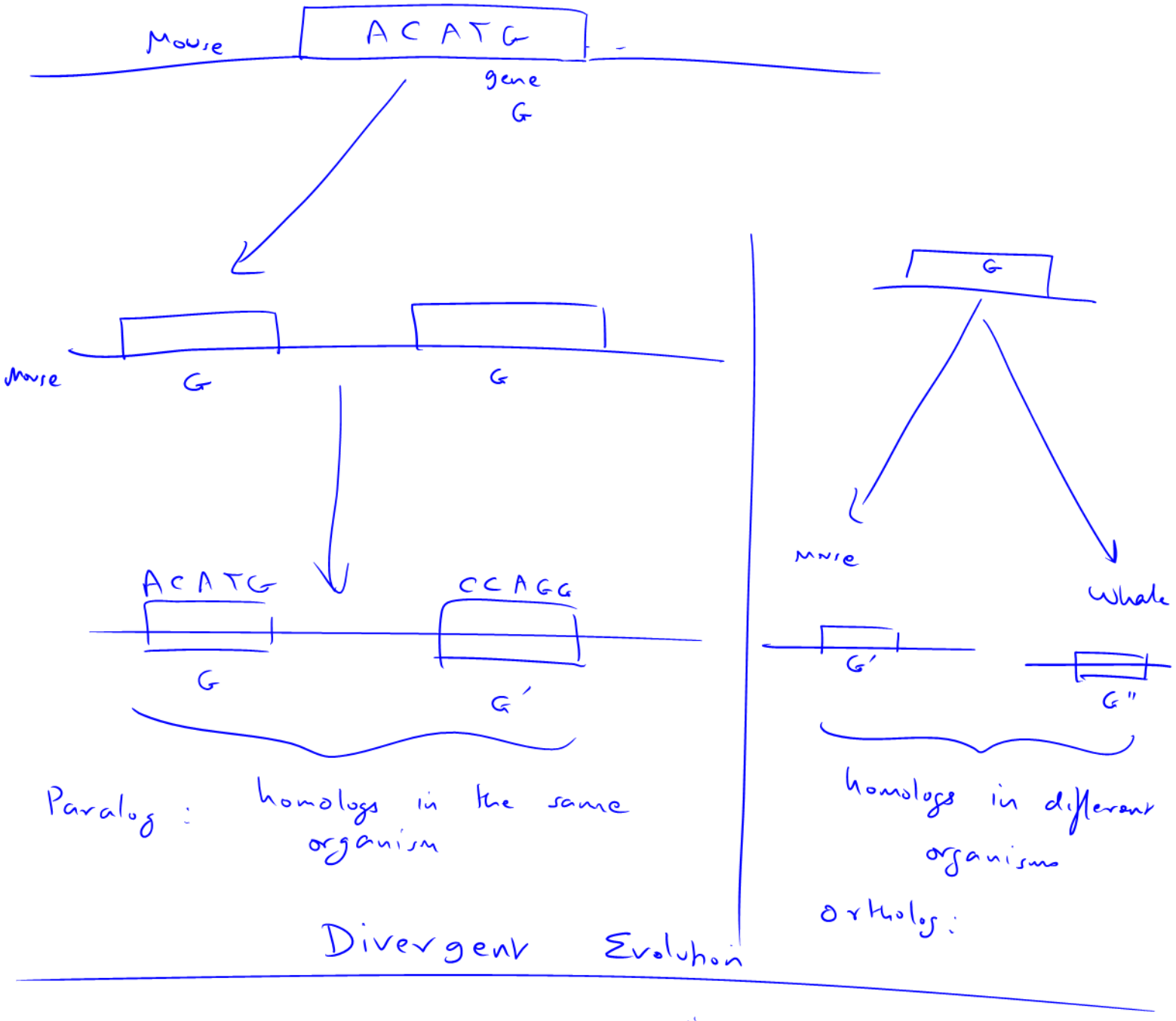
gapless



gapped alignment

A is an insertion in S_1

A is a deletion in S_2

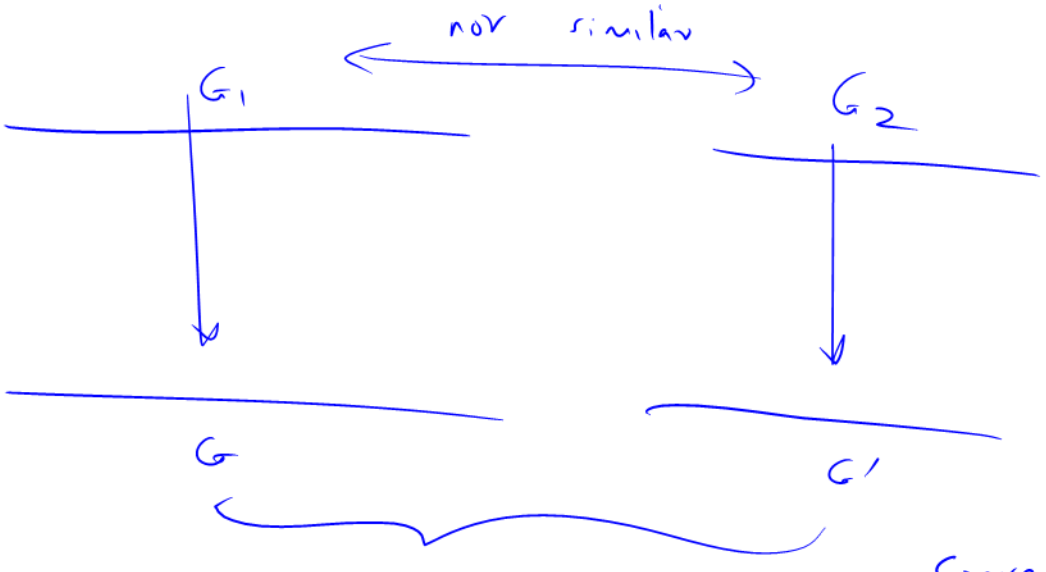


Paralog : homologs in the same organism

homologs in different organisms

ortholog :

Divergent Evolution



high similarity

analogs

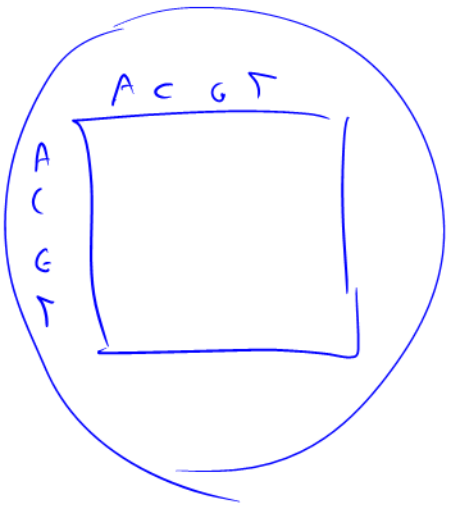
Convergent Evolution

x: ACTAG

y: A-TGG

$S = s(A,A) + s(C,-) + s(T,T) + s(A,G) + s(G,G)$

$s(C,-)$ ← SCP model



1) position independence

$$P(AB) = P(A) \times P(B)$$

if A & B are independent

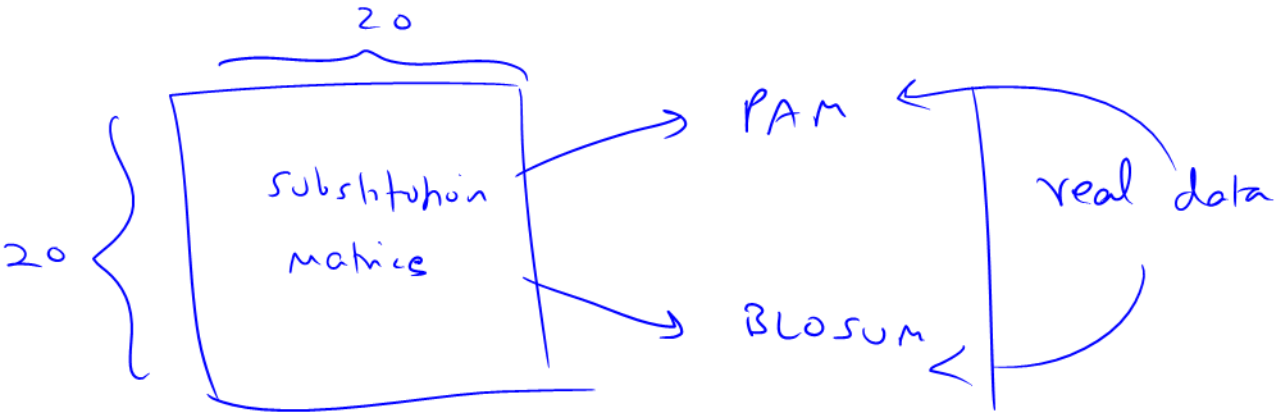
$$\log P(AB) = \log P(A) + \log P(B)$$

$$\log P(\text{alignment}) = \log\left(\prod_{i=1}^{n} \text{OR}(x_i, y_i)\right)$$

odds ratio

$$S = \sum_{i=1}^{n} s(x_i, y_i)$$

Compare / align protein sequence



Score of the alignment
Observe vs. random chance → statistically significant alignments

Ungapped alignment of a given length n.

	1	2	3	4	5
X :	A	C	A	T	G
Y :	C	G	A	G	G

$n = 5$

$$P(x, y \mid R) = \prod_{i=1}^{n} q_{x_i} \times \prod_{i=1}^{n} q_{y_i}$$

null model { random model

q_A
q_C
q_G
q_T

background or random probabilities

$$P(x,y \mid M) = \prod_{i=1}^{n} p_{x_i y_i}$$

how? PAM Blosum

$P(x,y \mid M)$: alternative hypothesis / model

Odds ratio:

$$\frac{P(x,y \mid M)}{P(x,y \mid R)} = \frac{\prod_{i=1}^{n} p_{x_i y_i}}{\prod_{i=1}^{n} q_{x_i} q_{y_i}}$$

Observed / expected

$$= \prod_{i=1}^{n} \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}$$

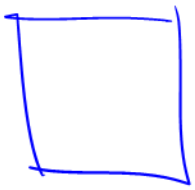
$$S = \log_2 \left(\frac{P(x,y \mid M)}{P(x,y \mid R)} \right) = \sum_{i=1}^{n} \log \left(\frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \right)$$

$$S = \sum_{i=1}^{n} s(x_i, y_i)$$

Alignment Algorithms

x : length n

y : length m



substitution matrix / scoring matrix

+ Gap model

- linear model: gd
- affine model: d (opening penalty), e (extension penalty)

$$d + (g-1)e$$

A C A G T A − C G

− C − − G A − G −

$g = 4$

$d = -10$

$d + e$

-40

of possible alignments?

$$\binom{n+m}{n} \text{ or } \binom{n+m}{m}$$

e.g. if $m = n$

then $\binom{2n}{n} \approx \frac{2^{2n}}{\sqrt{\pi n}}$

exponential # of possible alignments

Dynamic Programming Algorithm

Optimal sub problem property

x : A C A G

y : G G

	A	C	G	T
A	1		0	
C		1		
G	0		1	
T				1

gap = −1

F(i, j) = optimal score of aligning x[1 ... i] with y[1 ... j]

then F(n, m) gives the final score

x : A C A G

y : G G

$\begin{pmatrix} ACAG \\ GG \end{pmatrix}$ — —

Case 1

$\begin{pmatrix} ACA \\ G \end{pmatrix}$ $\begin{matrix} G \\ | \\ G \end{matrix}$ (i, j)

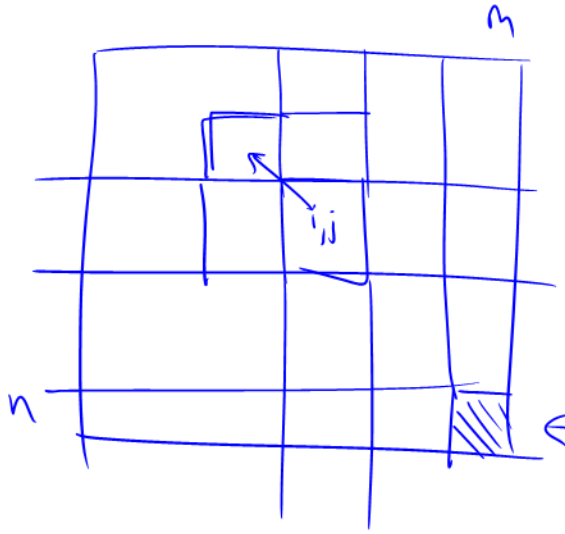
$\begin{pmatrix} ACA \\ GG \end{pmatrix}$ $\begin{matrix} G \\ - \end{matrix}$ (i, j)

$\begin{pmatrix} ACAG \\ G \end{pmatrix}$ $\begin{matrix} - \\ G \end{matrix}$ (i, j)

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + S(x_i, y_j) & \text{case 1} \\ F(i-1, j) + d & \leftarrow \text{case 2} \\ F(i, j-1) + d & \leftarrow \text{case 3} \end{cases}$$

$O(nm)$ time algorithm

$n \cdot m$ cells



trace back marker

← final score

match = 1

mismatch = 0

gap = −1

		-	A	C	A	G
		0	-1	-2	-3	-4
G		-1	0	-1	-2	-2
		-2	-1	0	-1	-1

X

Y

F

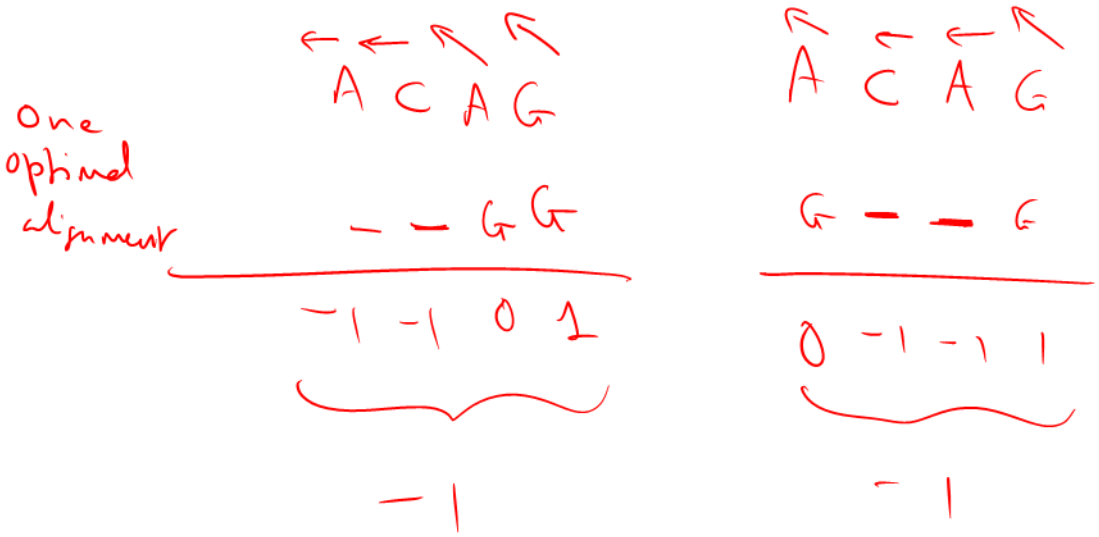
$(n+1) \times (m+1)$

size

$S(A, G) = 0$

A C A G
- - - -
- - - - G G

best score = −1


One optimal alignment
A C A G
- - G G
-1 -1 0 1
-1
A C A G
G - - G
0 -1 -1 1
-1