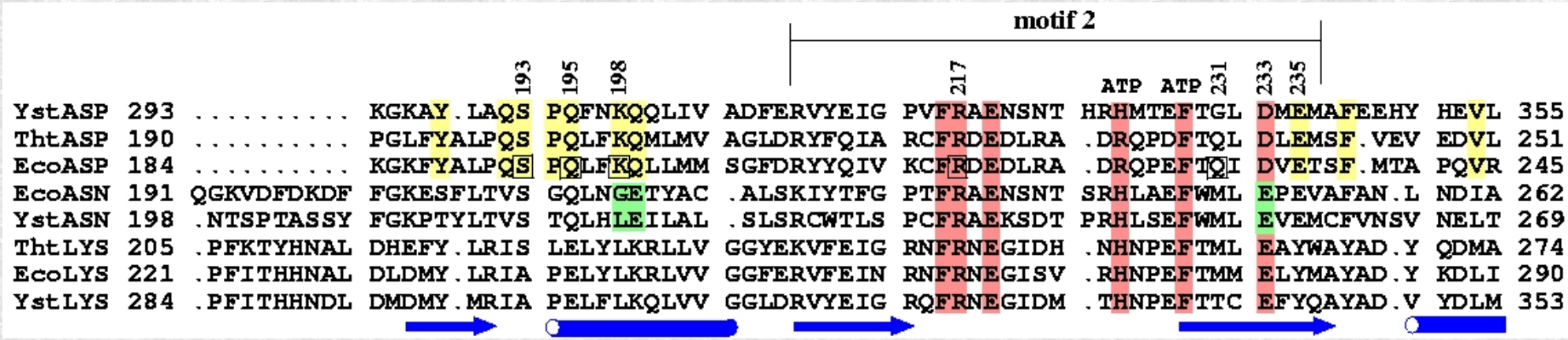
Sequence comparison and alignment



Mastères Ingénierie et Chimie des BioMolécules et Biologie et Santé

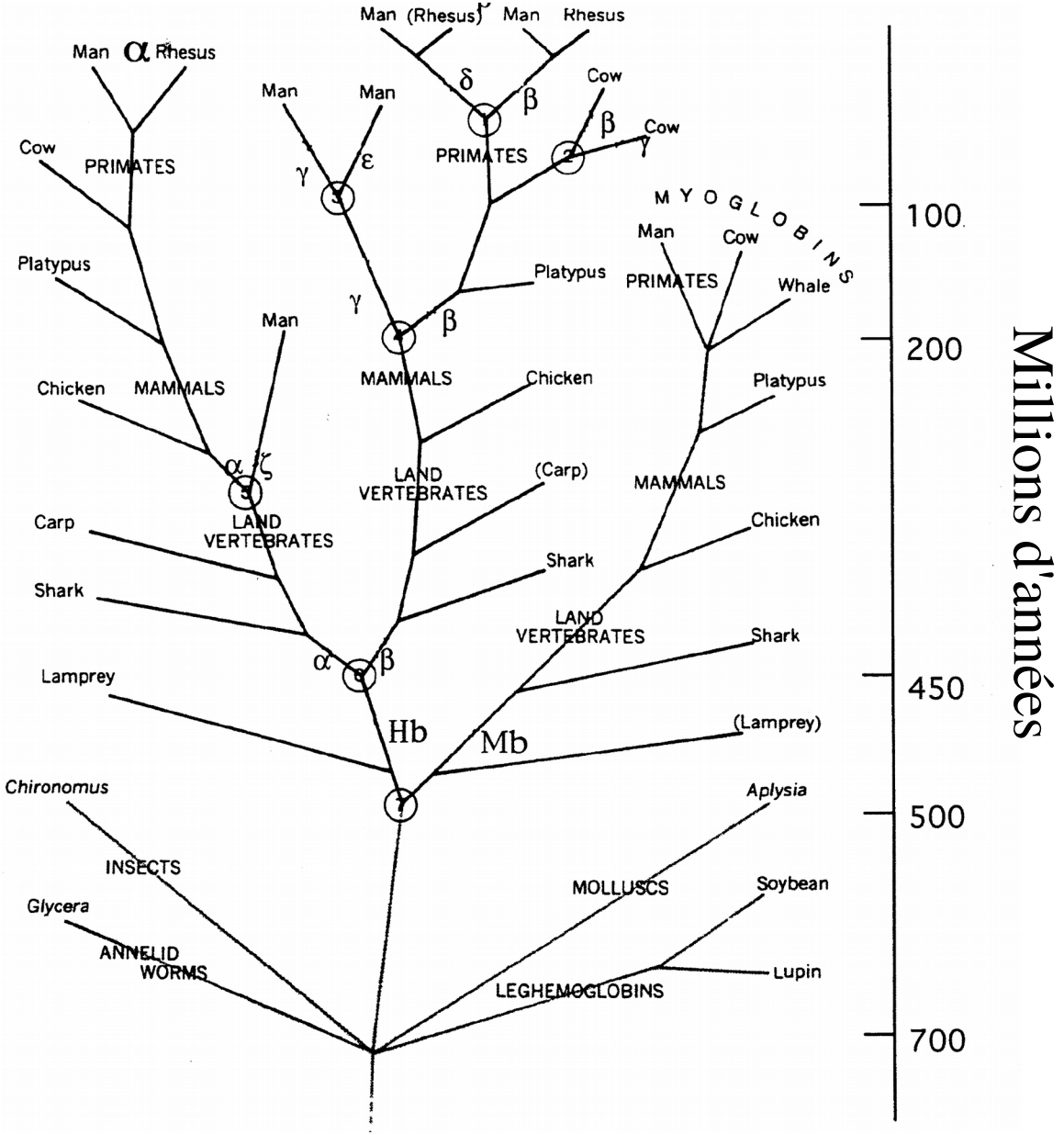
Module Bioinformatique structurale. Cours I

Novembre 2022

T. Simonson, Ecole Polytechnique

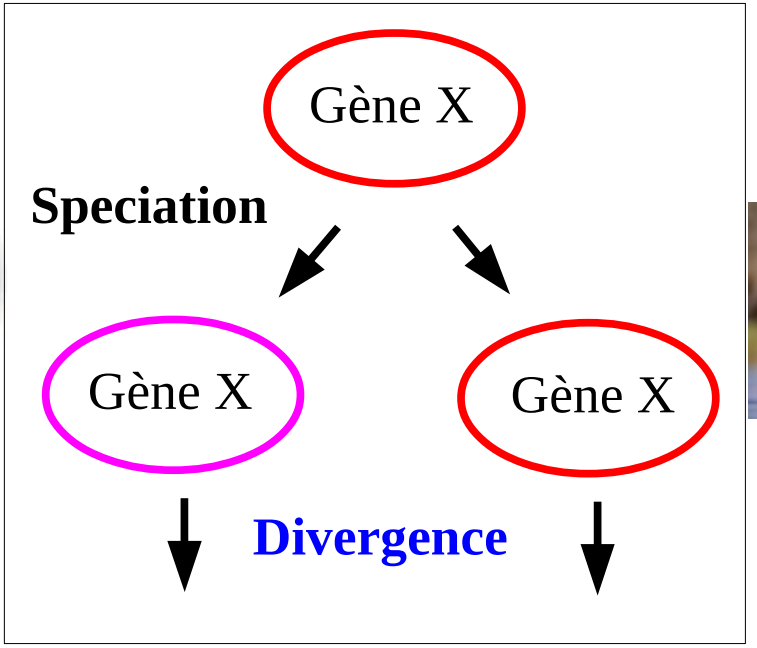
- **Sequence comparisons: what for?**
- **Sequence alignment as a model of evolution**
- **Sequence alignment: algorithms**

1) Similar proteins usually share a common ancestor: the globin example

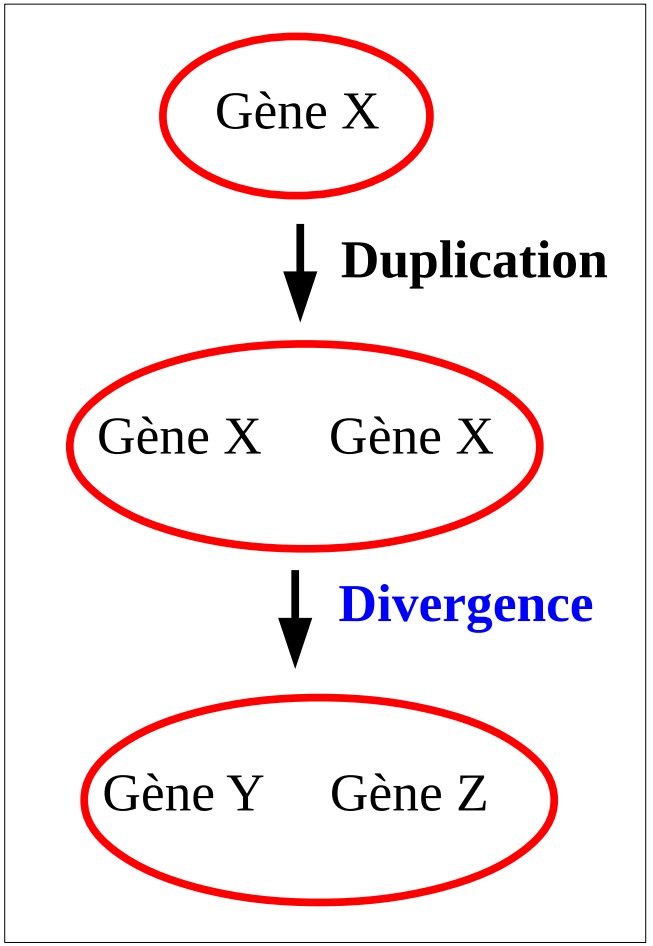


Phylogenetic tree for globin evolution

Similar proteins usually share a common ancestor

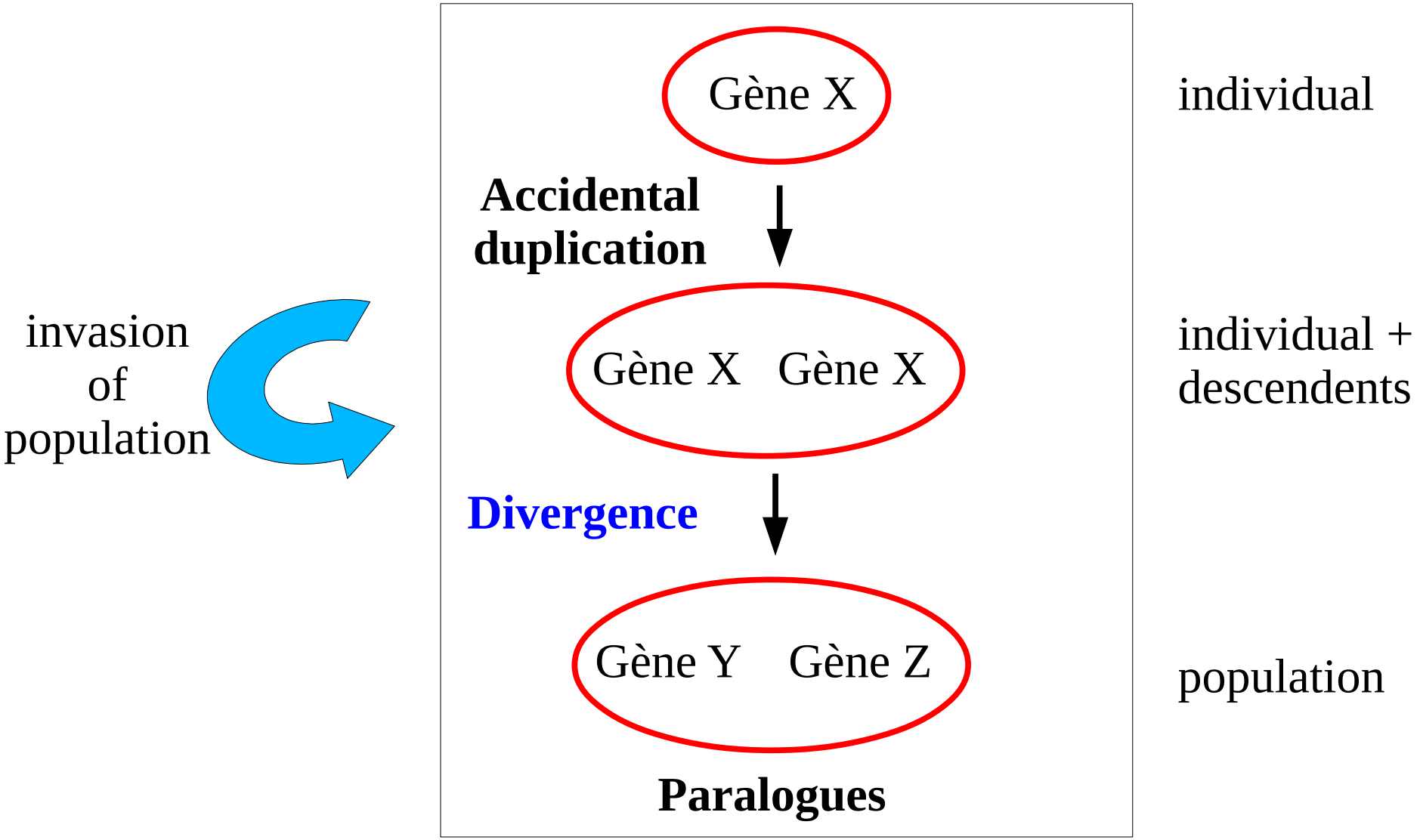


Orthologues



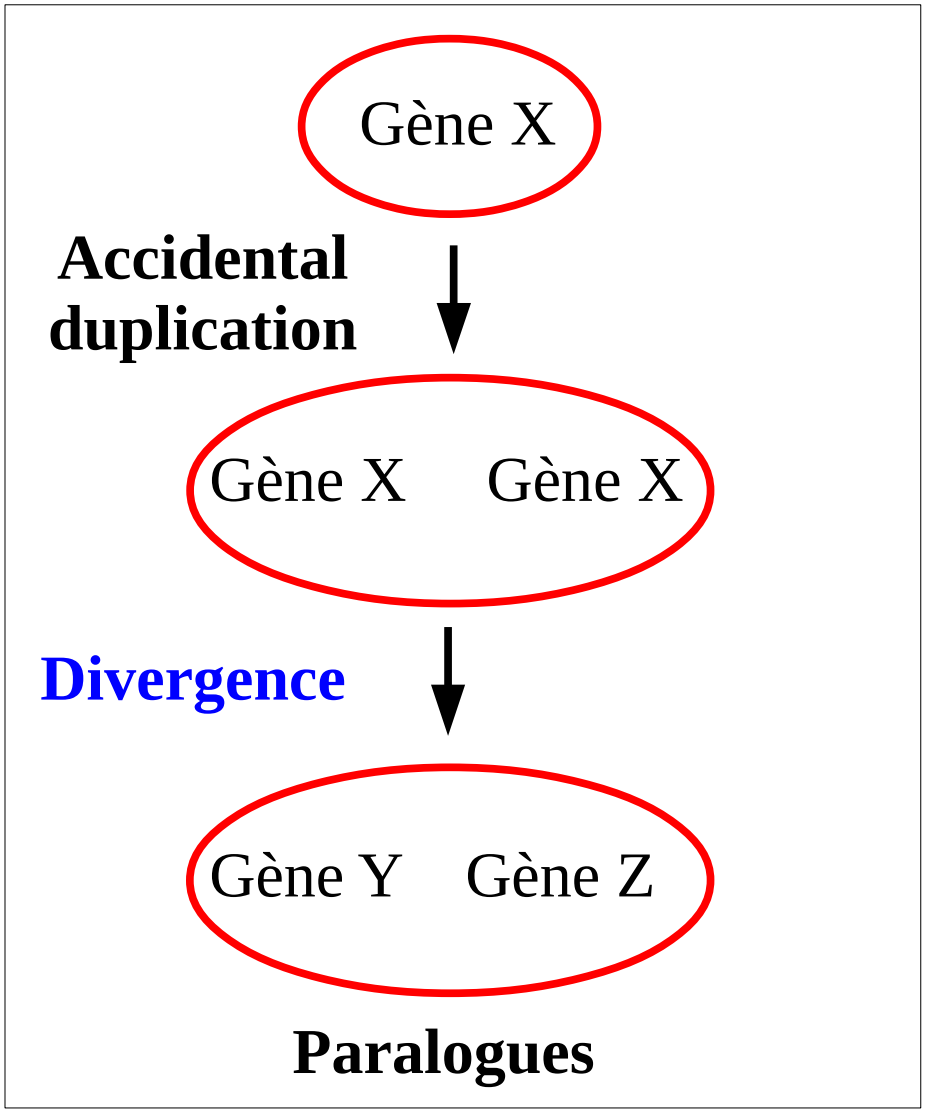
Paralogues

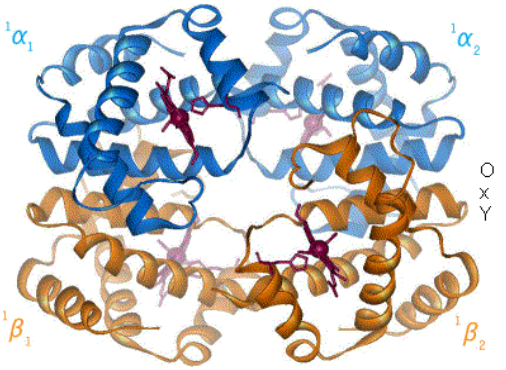
Duplication leads to paralogues



Paralogues usually change their function

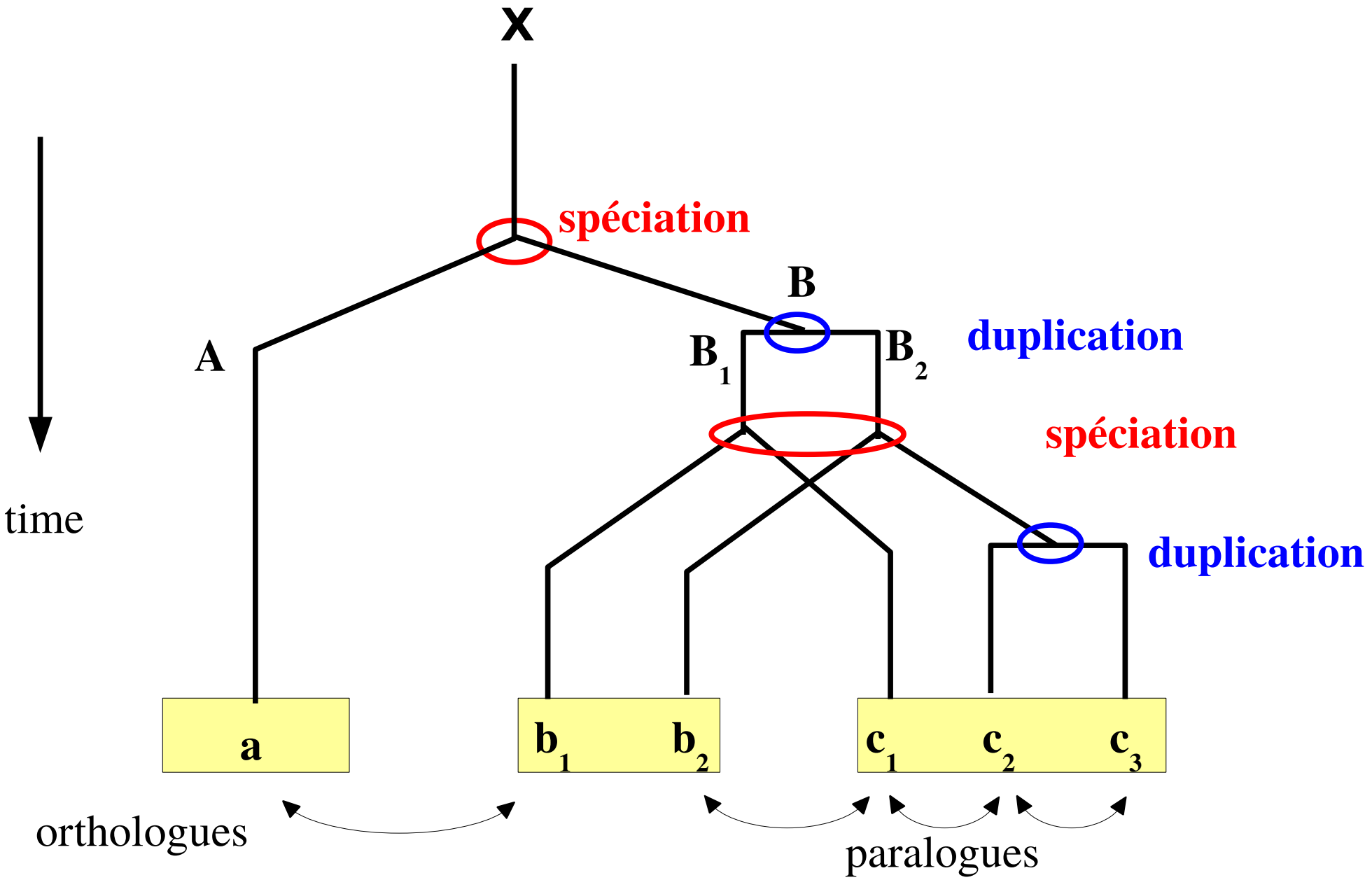
Duplication leads to paralogues





Paralogues usually change their function

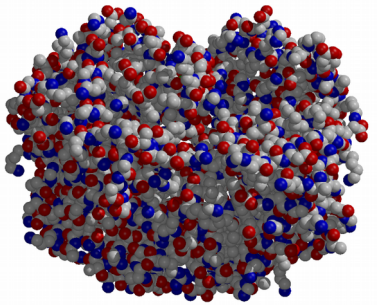
Evolution of a gene by speciation, duplication and divergence



2) Similar proteins usually have similar structures and often have similar functions

K
L
H
G
G
P
M
L
D
S
D
Q
K
F
W
R
T
P
A
A
L

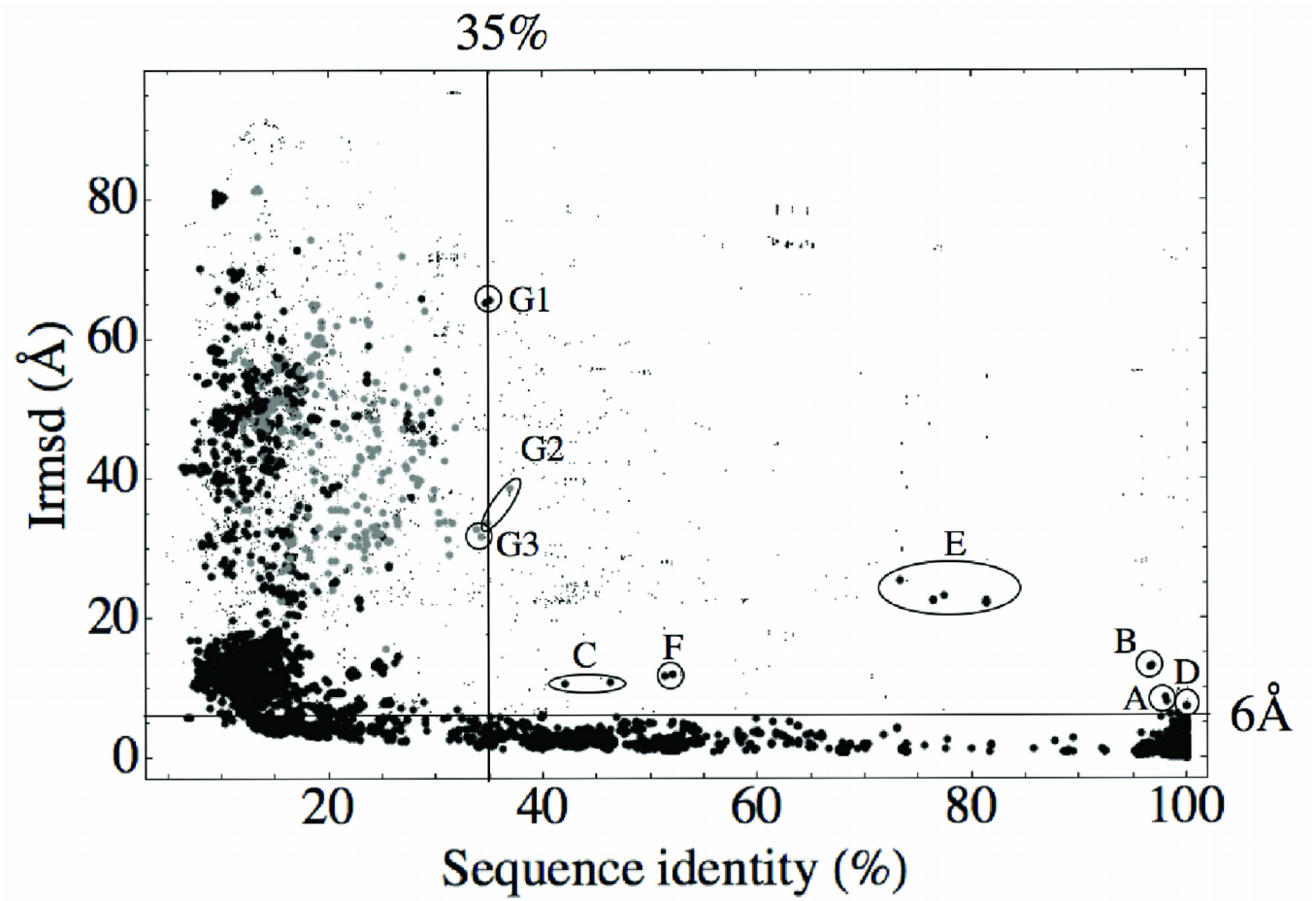
Sequence → **Structure**



Function

Above 60% sequence identity, two homologs normally have the same function.

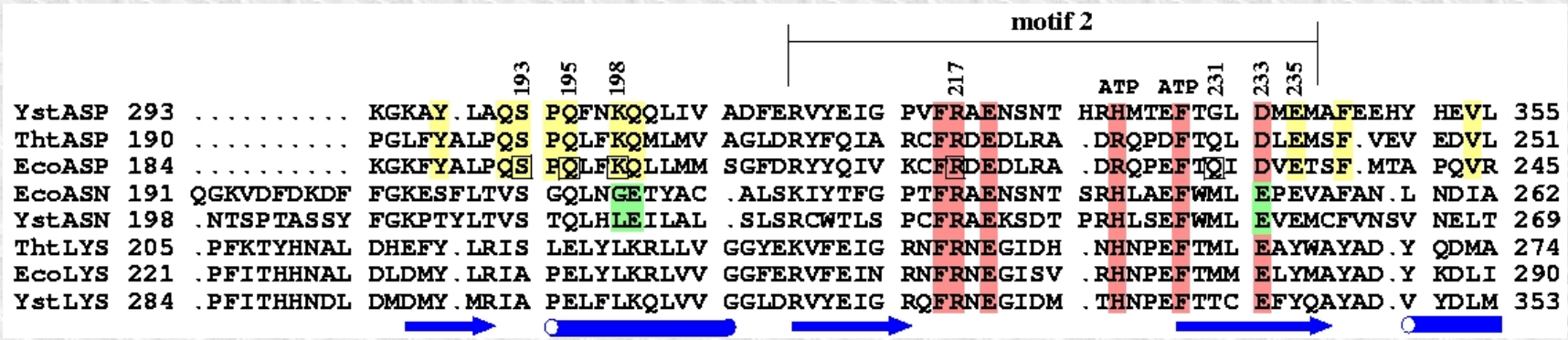
Above 35% sequence identity, homologous complexes usually interact in the same manner.



Launay & Simonson
BMC Bioinformatics
2008

Figure 1
The relationship between sequence and structural similarity. 743 complexes from 40 interacting superfamily groups (ISGs) were analyzed. All pairs within each ISG were compared, for a total of 9630 pairwise comparisons. Small points correspond to comparisons involving at least one complex with either a small interface (buried area < 600 Å^2) or a weak association energy (E_{int} > -10 kcal/mol; see text). Points labelled A-G are discussed in the text. The horizontal line corresponds to Irmsd = 6 Å; the vertical line corresponds to a 35% sequence identity. Gray points correspond to comparisons where the MATRAS structural alignment provided fewer than 80% of the equivalent residues used for the Irmsd calculation. All the gray points lie below the 35% similarity threshold.

Sequence alignment as an evolutionary model



Alignment as a model of evolution

Hypothesis: two similar proteins always share a common ancestor

Alignment as a hypothesis of a "parsimonious" common ancestor

T C L I C G D E A S G C H Y	Androsterone receptor
T C L V C G D E A T G Y H Y	**Hypothetical common ancestor**
L C V V C G D K A T G Y H Y	Thyroid hormone receptor

Hypothetical mutations in red

Is the hypothesis likely?

A probabilistic model of evolution

$P(x_i, y_j)$ = probability to observe x_i aligned with y_j

$P(x_i \neq y_j)$ = P("x_i mutated to y_j" or "y_j mutated to x_i")

$P(x_i = y_j)$ = P("x_i conserved")

sequence x: T C L I C G D E A S G C H Y

sequence y: L C V V C G D K A T G Y H Y

- Probabilities estimated from test alignements
- Assume positions are equivalent and independent

The likelihood or probability of an alignment

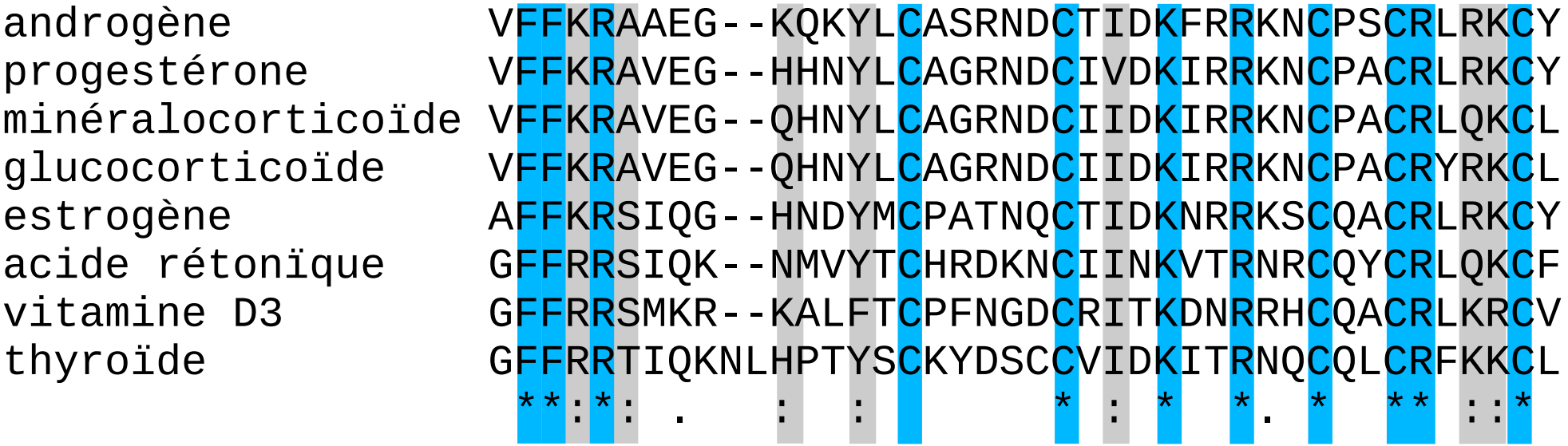
T C L I C G D E A S G C H Y Récepteur de l'androstérone

L C V V - G D K A T G Y H Y Récepteur de l'hormone thyroïdienne

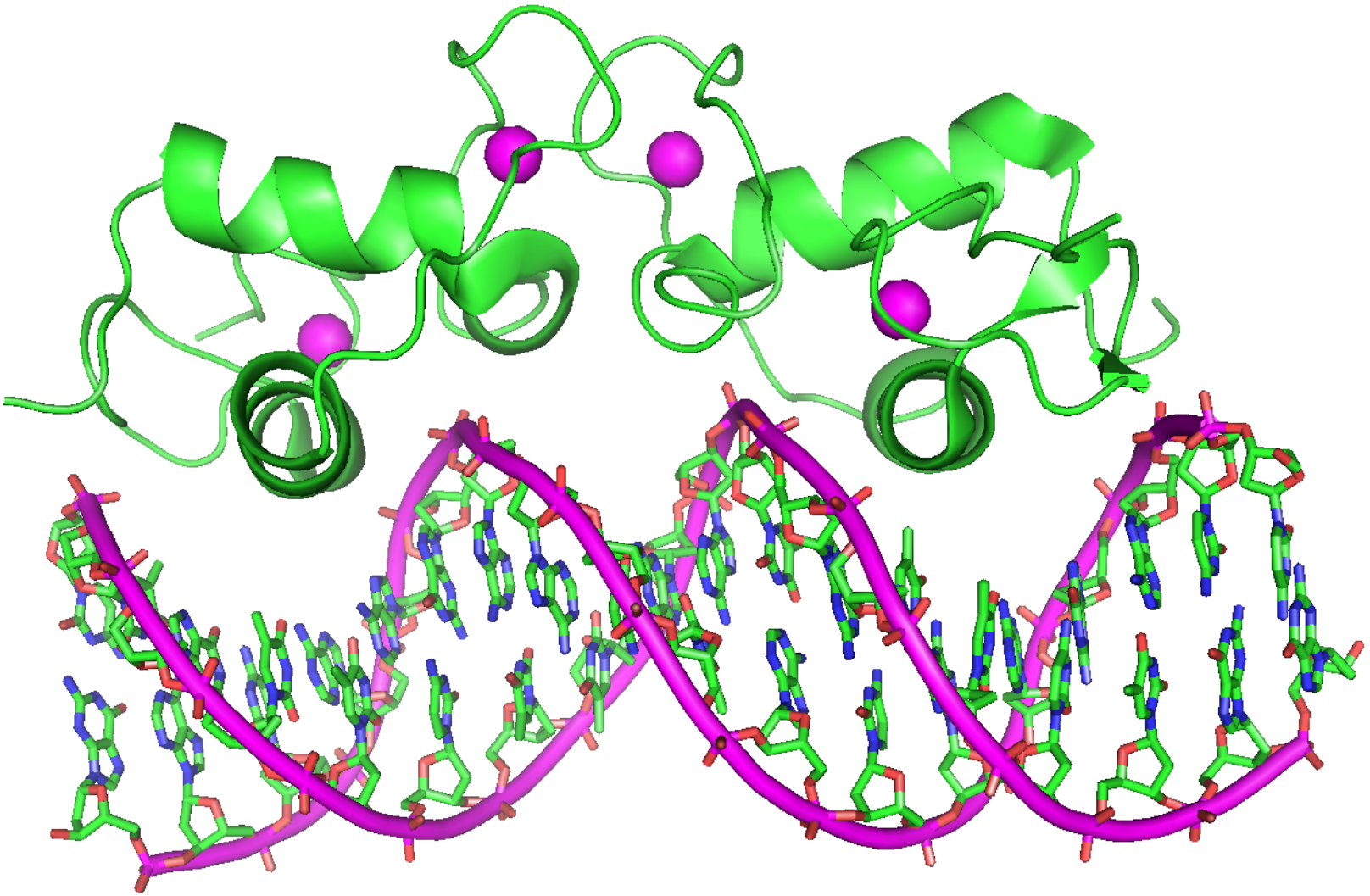
Probable mutations → probable alignment

Improbable mutations → improbable alignment

Comparing *homologous* proteins, we see that mutations have different probabilities



Récepteur de l'androstérone

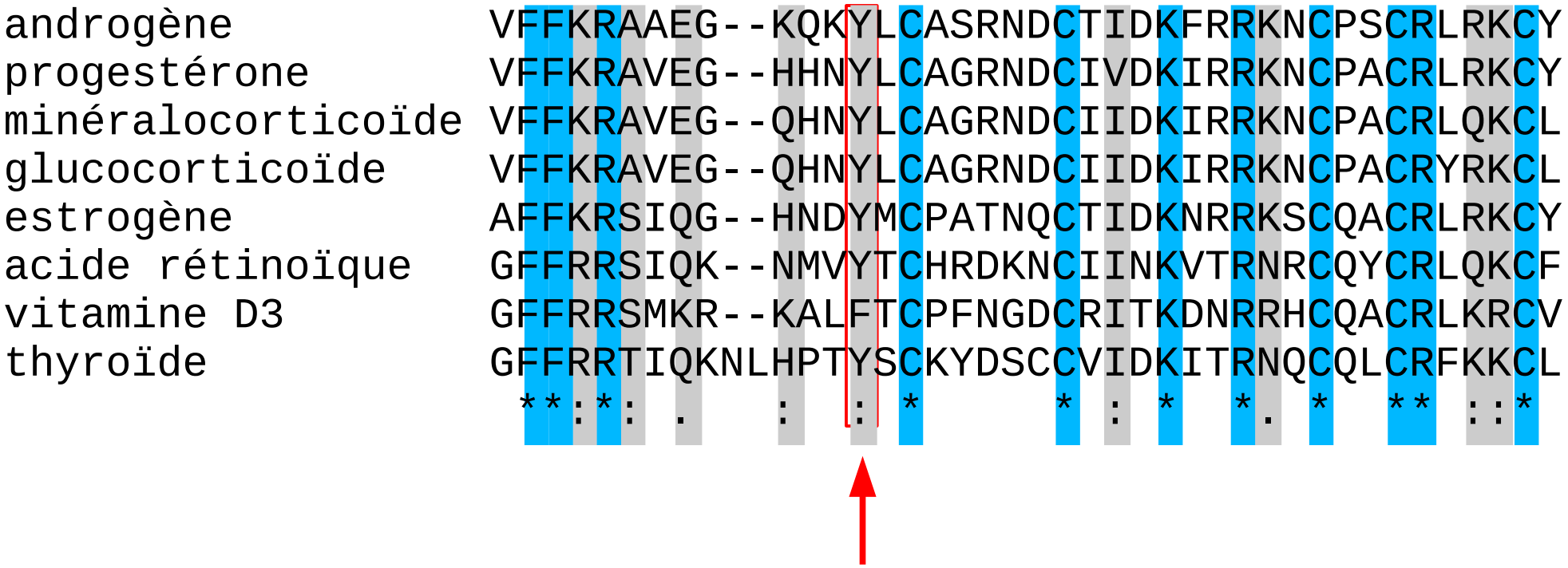


Nuclear receptors

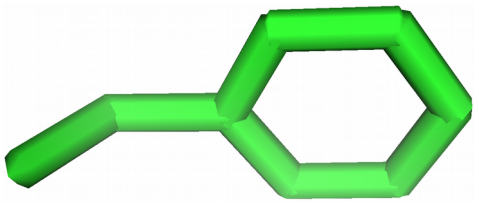
Bind DNA and regulate gene expression, under the control of small ligands: steroids, vitamins, ...

```
androgen        CLICGDEASGAHYGALTCGSCKVFFKRAAEGKQKYL-CASRNDCTIDKFRRKNCPSCRLRKCYEAGMTLGA
Rev Erb         CKVCGDVASGFHYGVLACEGCKGFFRRSIQQNIQYKRCLKNENCSIVRINRNRCQQCRFKKCLSVGMSRD-
glucocorticoid  CLVCSDEASGCHYGVLTCEGCKAFFKRAVEGQHNYL-CKYEGKCIIDKIRRKNCPACRYRKCLQAGMNLEA
retinoic acid   CAICGDRSSGKHYGVYSCEGCKGFFKRTVRKDLTYT-CRDNKDCLIDKRQRNRCQYCRYQKCLAMGM---
```

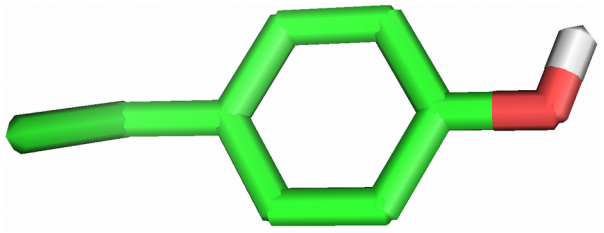
Mutations have different probabilities



Different amino acid structures yield different probabilities



PHE
F



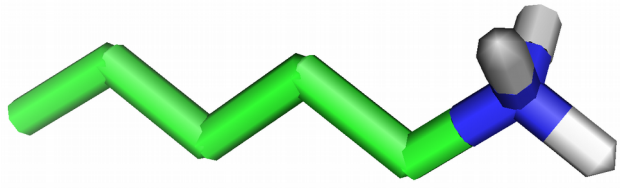
TYR
Y

KYL
NYL
NYL
NYL
DYM
VYT
LFT
TYS
:

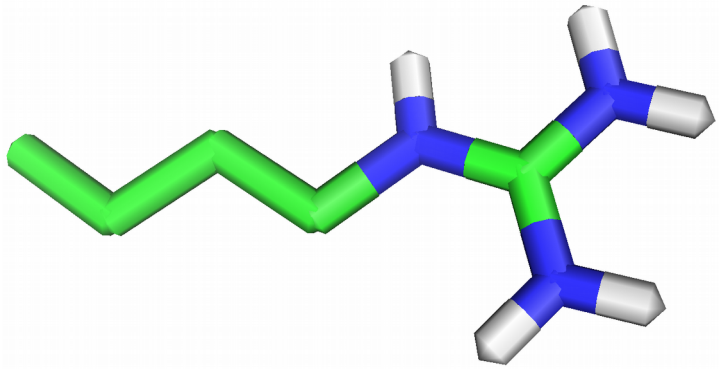
homologues

Different amino acid structures yield different probabilities

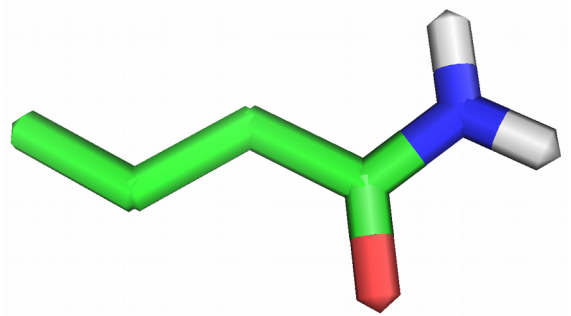
LYS
K



ARG
R



GLN
Q



homologues

```
RKCY
RKCY
QKCL
RKCL
RKCY
QKCF
KRCV
KKCL
  : :*
```

A probabilistic model of evolution

$P(x_i, y_j)$ = probability to observe x_i aligned with y_j

$P(x_i \neq y_j)$ = P("x_i mutated to y_j" or "y_j mutated to x_i")

$P(x_i = y_j)$ = P("x_i conserved")

sequence x: T C L I C G D E A S G C H Y

sequence y: L C V V C G D K A T G Y H Y

- Probabilities estimated from test alignements
- Assume positions are equivalent and independent

Problem: what is a large probability?

Alignement avec le récepteur humain de la progestérone

```
PQKTCLICGDEASGAHYGALTCGSCKVFFKRAAEGKQKYLCASRNDCTIDKFRRKNCPSC
PQRVCVICGDEASGCHYGVLTCGSCKVFFKRAVEGHHQYLCAGRNDCIVDKIRRKNCPAC
**' *'********* *** ************** **  '**** **** '** *******'*
```

Alignement avec le récepteur humain de l'hormone thyroïdienne

```
PQKTCLICGDEASGAHYGALTCGSCKVFFKRAAEG--KQKYLCASRNDCTIDKFRRKNCPSC
KDEQCVVCGDKATGYHYRCITCEGCKGFFRRTIQKNLHPTYSCKYDSCCVIDKITRNQCQLC
    *''****'*'* **  '**  ** **'*  '       * *     * ***  *  *  *
```

Alignement avec la ferrédoxine de la bactérie *Proteus vulgaris*

```
PQKTCLICGDEASGAHYGTLTCGSCKVFFKRAAEGKQKYLCASRNDCTIDKFRRKNCPSC
DQDKCIGCKTCVLACPYGTMEVVSRPVMRKLTALNTIEAFKAEANKCDLCHHRAEG-PAC
 *   *' *          ***'   *   *   *   *          *   * * '   * '  *'*
```

“Random” modem as a reference

Evolutionary hypothesis: two sequences have a (parsimonious) common ancestor

Null hypothesis: they have no biological relation: compute probability as if the sequences were random

$Q(x_i,y_j)$ = probability to observe x_i and y_j in two independent proteins

= $q_{xi}\, q_{yj}$ Natural frequencies of x_i, y_j

x	q_x	
W	1.3 %	of amino acids
L	9.0 %	

$$\begin{matrix} W & & L \\ W & \neq & L \end{matrix}$$

Likelihood of an alignment column

We compute the ratio between the probabilities P (evolutionary case) and Q (random case):

$$M(x_i,y_j) \overset{\text{def.}}{=} \log P(x_i,y_j)/Q(x_i,y_j)$$

$$= \log [\, P(x_i,y_j)/q_{xi}\, q_{yj} \,]$$

x_i

T C L I C G D E A S G C H Y

L C V V C G D K A T G Y H Y

y_i

M = scoring matrix

Matrix example: BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2	C
S		4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3	S
T			5	-1	0	-2	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	0	-2	-2	-2	T
P				7	-1	-2	-2	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4	P
A					4	0	-2	-2	-1	-1	-2	-1	-1	-1	-1	-1	0	-2	-2	-3	A
G						6	0	-1	-2	-2	-2	-2	-2	-3	-4	-4	-3	-3	-3	-2	G
N							6	1	0	0	1	0	0	-2	-3	-3	-3	-3	-2	-4	N
D								6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4	D
E									5	2	0	0	1	-2	-3	-3	-2	-3	-2	-3	E
Q										5	0	1	1	0	-3	-2	-2	-3	-1	-2	Q
H											8	0	-1	-2	-3	-3	-3	-1	2	-2	H
R												5	2	-1	-3	-2	-3	-3	-2	-3	R
K													5	-1	-3	-2	-2	-3	-2	-3	K
M														5	1	2	1	0	-1	-1	M
I															4	2	3	0	-1	-3	I
L																4	1	0	-1	-2	L
V																	4	-1	-1	-3	V
F																		6	3	1	F
Y																			7	2	Y
W																				11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

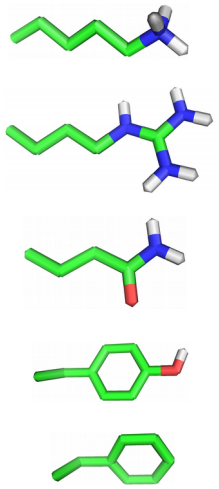
Matrix example: BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2	C
S		4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3	S
T			5	-1	0	-2	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	0	-2	-2	-2	T
P				7	-1	-2	-2	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4	P
A					4	0	-2	-2	-1	-1	-2	-1	-1	-1	-1	-1	0	-2	-2	-3	A
G						6	0	-1	-2	-2	-2	-2	-2	-3	-4	-4	-3	-3	-3	-2	G
N							6	1	0	0	1	0	0	-2	-3	-3	-3	-3	-2	-4	N
D								6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4	D
E									5	2	0	0	1	-2	-3	-3	-2	-3	-2	-3	E
Q										5	0	1	1	0	-3	-2	-2	-3	-1	-2	Q
H											8	0	-1	-2	-3	-3	-3	-1	2	-2	H
R												5	2	-1	-3	-2	-3	-3	-2	-3	R
K													5	-1	-3	-2	-2	-3	-2	-3	K
M														5	1	2	1	0	-1	-1	M
I															4	2	3	0	-1	-3	I
L																4	1	0	-1	-2	L
V																	4	-1	-1	-3	V
F																		6	3	1	F
Y																			7	2	Y
W																				11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Penalty for gaps: see below

Exc: compute P(W,W) and P(L,L) and comment.

Mutations have different probabilities



	K	R	Q	Y	F
K	6	3	2	-2	-4
R		7	1	-1	-3
Q			7	-1	-4
Y				8	4
F					8

BLOSUM50 matrix

Likelihood or probability of an alignment

$P(x_i,y_j)$ = probability to observe x_i aligned with y_j

= P("x_i conserved" or "x_i mutated to y_j" ou "y_j mutated to x_i")

```
T C L I C G D E A S G C H Y
L C V V - G D K A T G Y H Y
```

- Probabilities from test'alignments
- Assume equivalent positions
- Assume independent positions: $P(\text{alignement}) = \Pi_{(i,j)} P(x_i,y_j)$

Likelihood or probability of an alignment

$$s(x_i,y_j) = \log P(x_i,y_j)/Q(x_i,y_j)$$

$$= \log [\, P(x_i,y_j)/q_{xi}\, q_{yj} \,]$$

$$s(x,y) = \Sigma_{(i,j)}\, s(x_i,y_j)$$

T C L I C G D E A S G C H Y

L C V V - G D K A T G Y H Y

Mutation probability depends on time

P(x,y) = P(x,y;T) = probability of a mutation during time T

Cytochromes c from chimpanzee and human
are closer than those from
human and Escherichia coli...

A similarity matrix is valid for a certain time scale,
or similarity scale

BLOSUM50 ≠ BLOSUM62

Better gap treatment

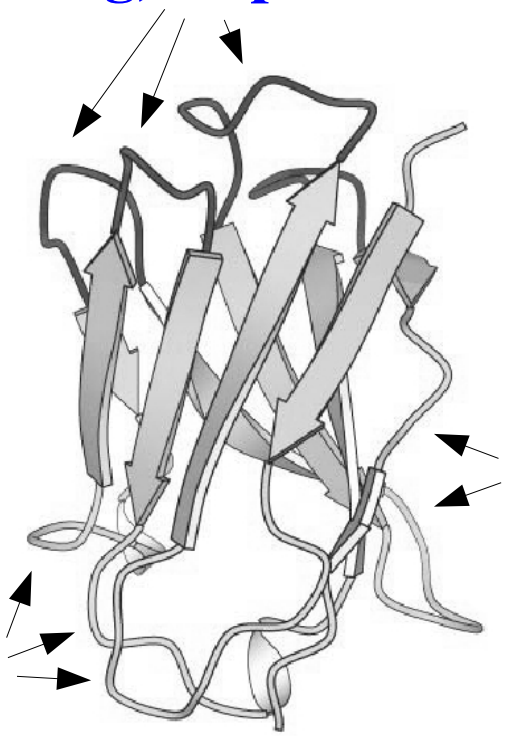
It is easier to *extend* a gap than open a new one

- Open a gap: -d
- Extend by one residue: -e (e<d)

- Cost of a gap of length m:

$$g(m) = -d - (m-1)\,e$$

Eg, loops



Positions are no longer equivalent nor independent

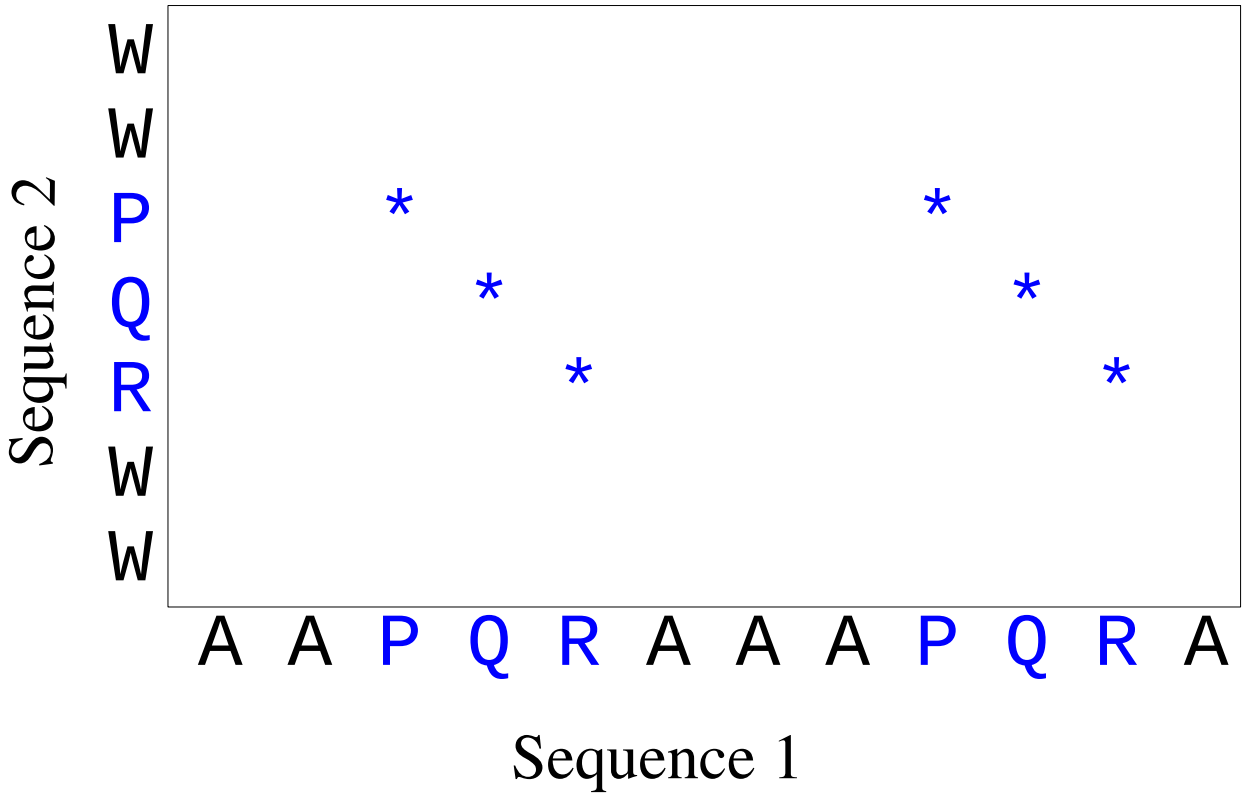
An alignment represents a model of evolution, with:

- Minimal divergence from a common ancestor
- Empirical mutation probabilities
- Assumption of equivalent and independent positions
- A simple Null hypothesis

Sequence alignment algorithms

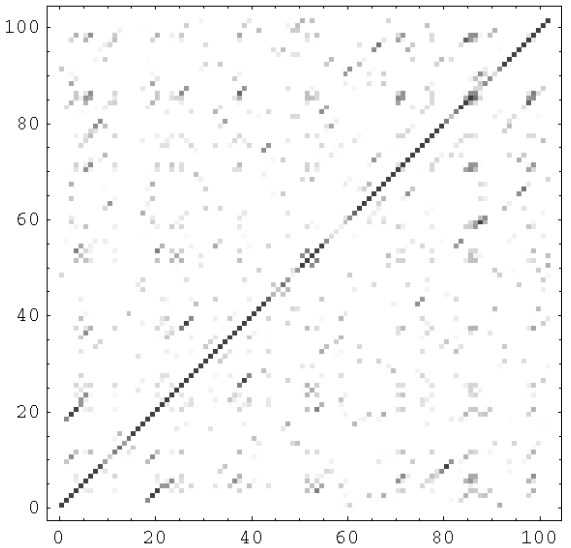
Graphical sequence comparison

Each sequence is written on the edge of a table or matrix.
For similar residue pairs, one inserts a similarity symbol:



Graphical sequence comparison

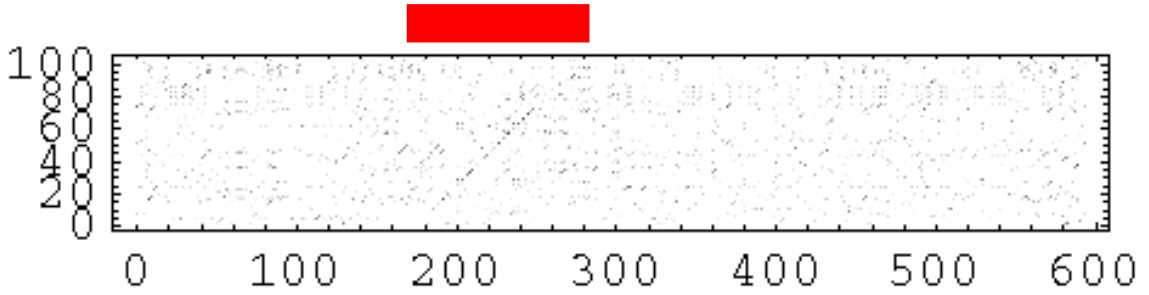
horse



Cytochromes c

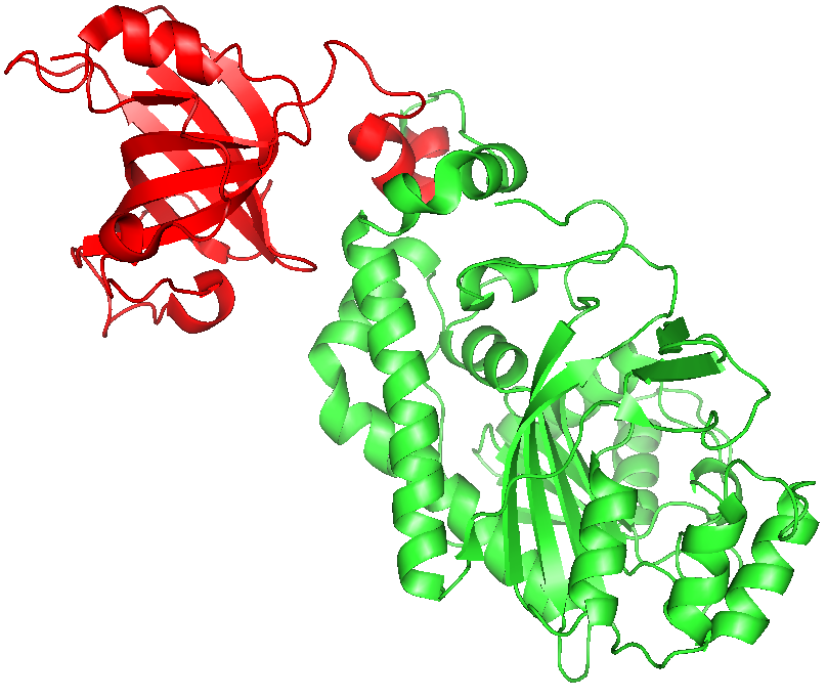
human

Récepteur androgène



Récepteur estrogène

Often only part of a protein is conserved: Need for a "local" alignment

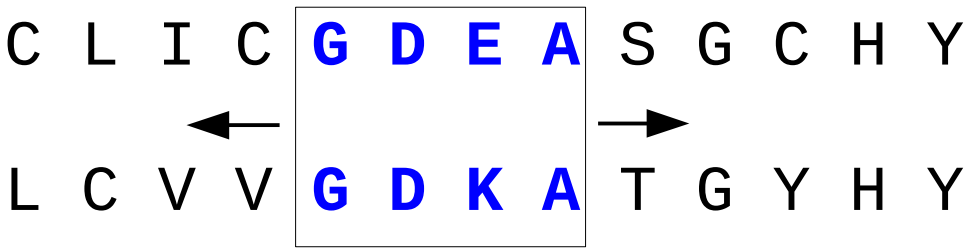


Aspartyl-tRNA synthetase from yeast

Active site domain, Aspartyl-tRNA synthetase from E coli

BLAST: *Basic Local Alignment Search Tool*

- Find homologous tetrapeptides
- Extend each peptide as long as similarity > threshold



Récepteur de l'androstérone

Récepteur de l'hormone thyroïdienne

Tetrapeptides homologous to a query tetrapeptide

G D E A	score BLOSUM62	
G D E A	21	= 6+6+5+4
G D D A	18	
G D Q A	18	
G E E A	17	
G D E G	17	
G D K A	17	
G D E V	17	

BLOSUM62

	V	A	G	D	E	Q	K
A	0	4					
G			6				
D				6	2		
E					5	2	1

7 homologues (using BLOSUM62 and a threshold of 17)

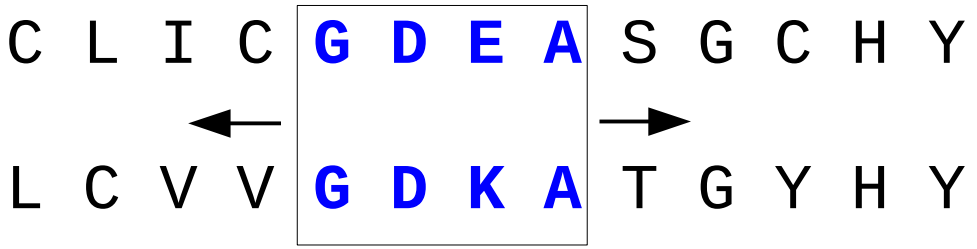
Find tetrapeptides that are homologous to query

G D E A
G D D A
G D Q A
G E E A
G D E G
G D K A
G D E V

Tetrapeptides homologous to **GDEA**

```
>sp|P10827|THA_HUMAN Thyroid hormone receptor alpha Homo sapiens.
MEQKPSKVECGSDPEENSARSPDGKRKRKNGQCSLKTSMSGYIPSYLDKDEQCVVCGDKA
TGYHYRCITCEGCKGFFRRTIQKNLHPTYSCKYDSCCVIDKITRNQCQLCRFKKCIAVGM
AMDLVLDDSKRVAKRKLIEQNRERRRKEEMIRSLQQRPEPTPEEWDLIHIATEAHRSTNA
QGSHWKQRRKFLPDDIGQSPIVSMPDGDKVDLEAFSEFTKIITPAITRVVDFAKKLPMFS
ELPCEDQIILLKGCCMEIMSLRAAVRYDPESDTLTLSGEMAVKREQLKNGGLGVVSDAIF
ELGKSLSAFNLDDTEVALLQAVLLMSTDRSGLLCVDKIEKSQEAYLLAFEHYVNHRKHNI
PHFWPKLLMKEREVQSSILYKGAAAEGRPGGSLGVHPEGQQLLGMHVVQGPQVRQLEQQL
GEAGSLQGPVLQHQSPKSPQQRLLELLHRSGILHARAVCGEDDSSEADSPSSSEEEPEVC
EDLAGNAASP
```

Extend “micro-alignment” in each direction, as long as score > threshold



Récepteur de l'androstérone

Récepteur de l'hormone thyroïdienne

- Each tetrapeptide leads to a local alignment without gaps
- Keep the best alignments

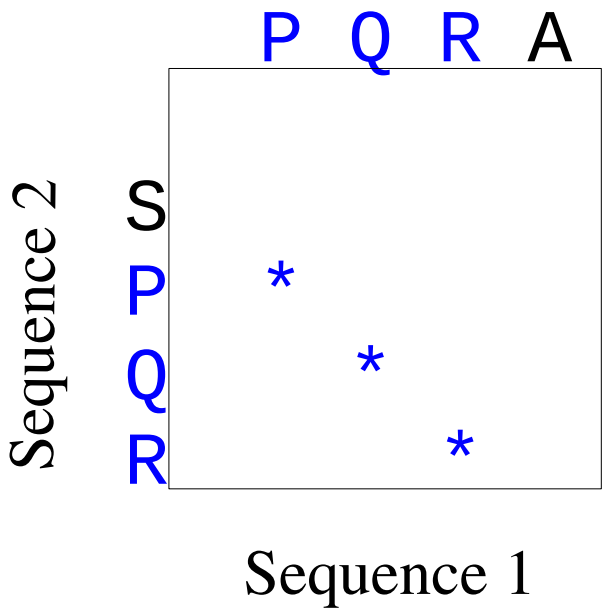
Homologues of the androsterone receptor identified using BLAST

#	ID Swissprot	Hit	Description	Score (bits)	E *	% Identity	Match Length
1	P15207	ANDR_RAT	Androgen receptor.	162	1e-40	100	73
6	P19091	ANDR_MOUSE	Androgen receptor.	162	1e-40	100	73
14	Q63449	PRGR_RAT	Progesterone receptor (PR)	136	1e-32	80	72
17	P06401	PRGR_HUMAN	Progesterone receptor (PR)	136	1e-32	80	72
21	P08235	MCR_HUMAN	Mineralocorticoid receptor (MR)	136	1e-32	79	72
33	P04150	GCR_HUMAN	Glucocorticoid receptor (GR)	131	3e-31	77	72
41	Q9YH32	ESR2_ORENI	Estrogen receptor beta (ER-beta)	99	3e-21	58	72
42	Q9YH33	ESR1_ORENI	Estrogen receptor (ER-alpha)	98	4e-21	55	72
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
343	Q9N4Q7	NH13_CAEEL	Nuclear hormone receptor nhr-13	54	8e-08	39	66
344	Q23294	NH11_CAEEL	Nuclear hormone receptor nhr-11	54	8e-08	42	66
345	O45460	NH54_CAEEL	Nuclear hormone receptor nhr-54	54	1e-07	37	67
346	Q09565	NH20_CAEEL	Nuclear hormone receptor nhr-20	51	7e-07	34	66
347	Q09587	NH22_CAEEL	Nuclear hormone receptor nhr-22	45	5e-05	32	66
349	P17672	E75B_DROME	Ecdysone-induced protein 75B	40	0.001	37	47
351	P20659	TRX_DROME	Trithorax protein.	31	0.74	26	49
355	P98164	LRP2_HUMAN	Lipoprotein receptor.	30	1.7	27	65

*E = expectation of number of random alignments with a higher score

A rigorous method, inspired by the previous “matrix”

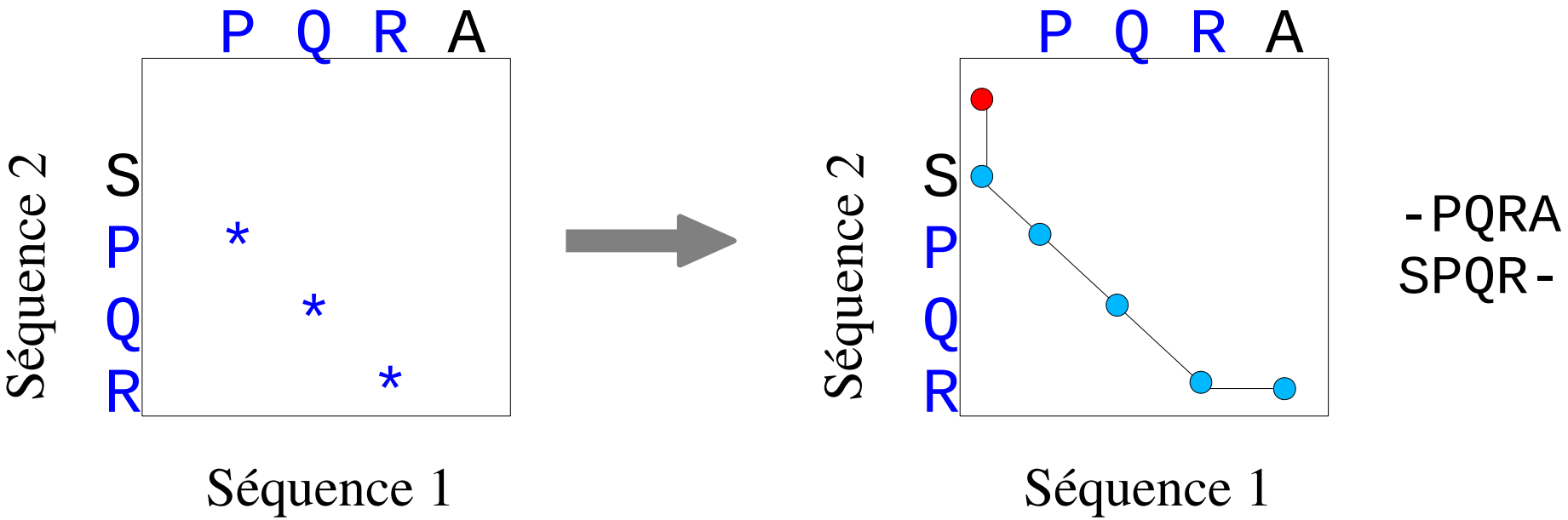
The sequences are written along the edges of the matrix:



A rigorous method; partial description

The sequences are written along the edges of the matrix.

An alignment is a path through the matrix:



Exc: represent the alignments:

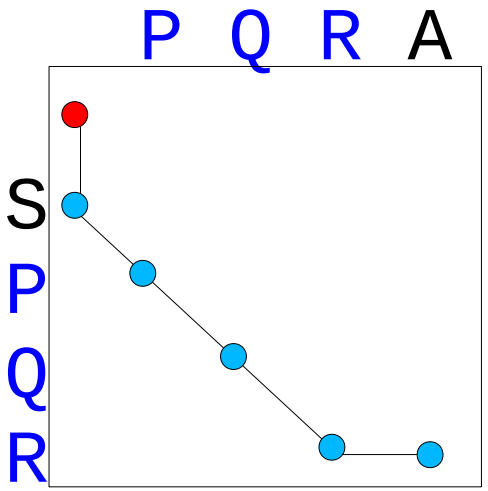
```
P-QRA
SPQR-
```

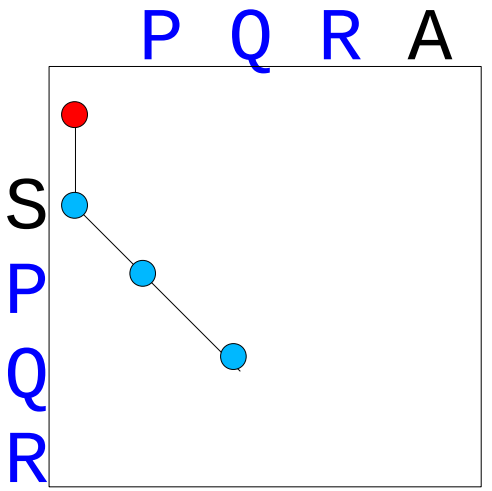
```
P--QRA
SPQ-R-
```

A rigorous method; partial description

The sequences are written along the edges of the matrix.

An incomplete path represents an incomplete alignment:





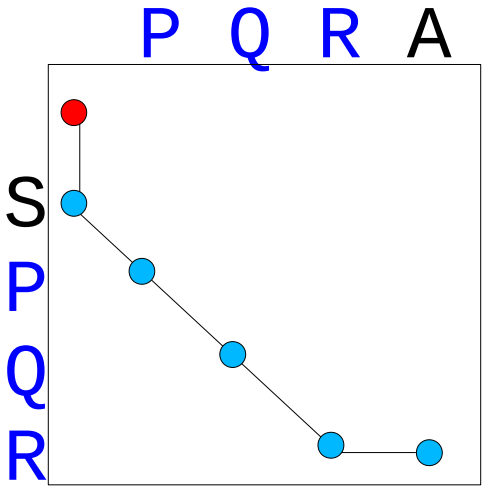
-PQRA
SPQR-

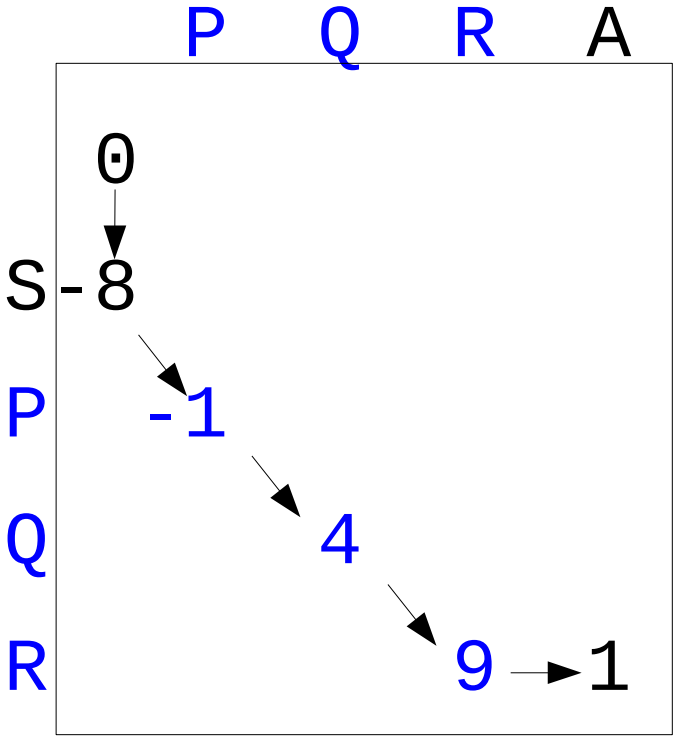
-PQ
SPQ

A rigorous method; partial description

The sequences are written along the edges of the matrix.

We annotate the matrix with the scores:



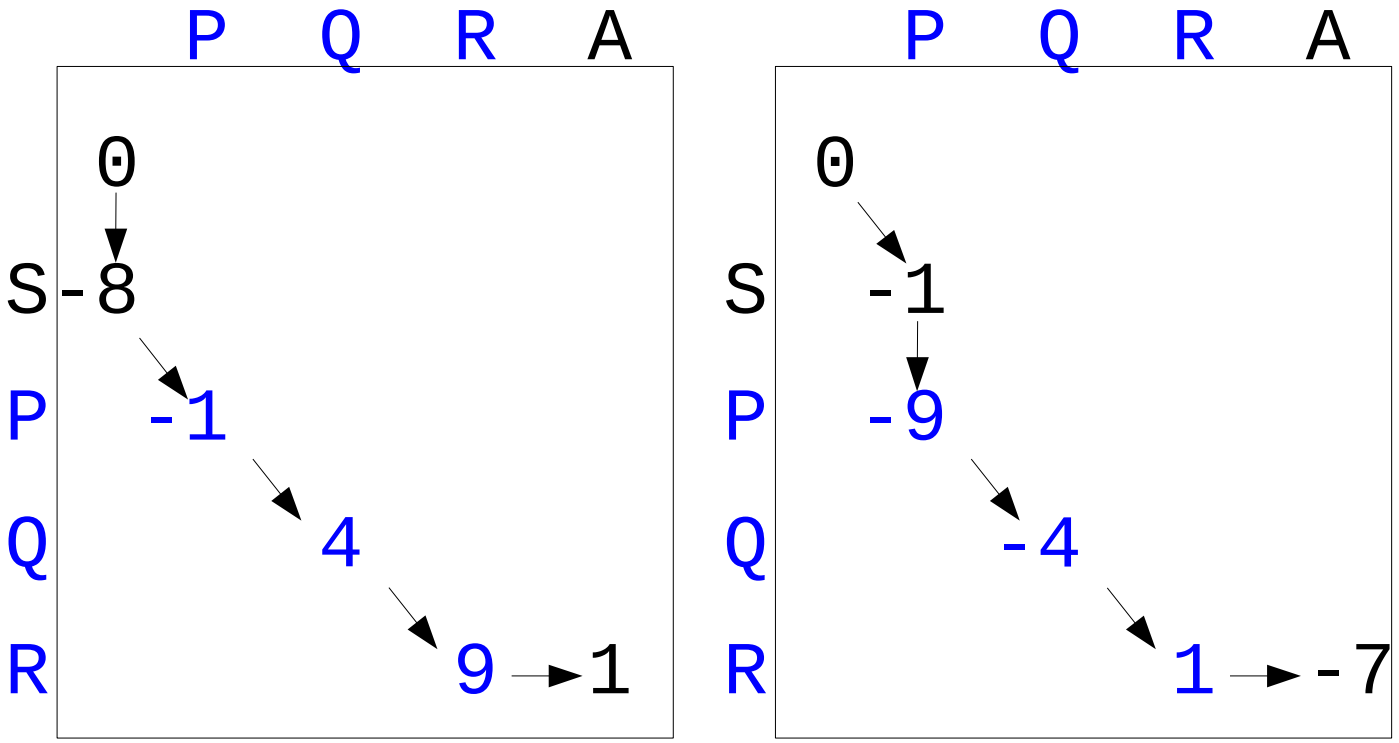


	S	P	Q	R
P	-1	7	-1	-2
Q	0	-1	5	1
R	-1	-2	1	5
A	1	-1	-1	-1

```
-PQRA
SPQR-
```

A rigorous method; partial description

The best alignment competes with many others:

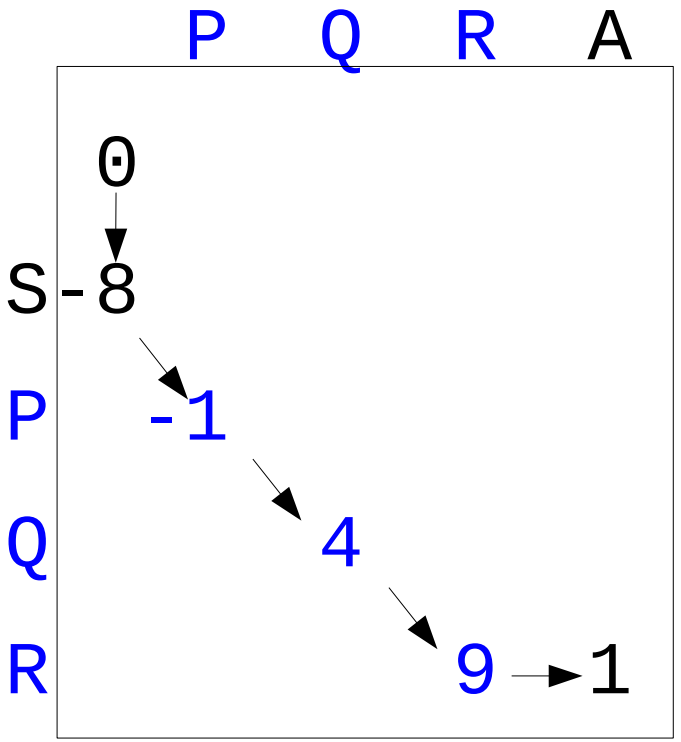


-PQRA
SPQR-

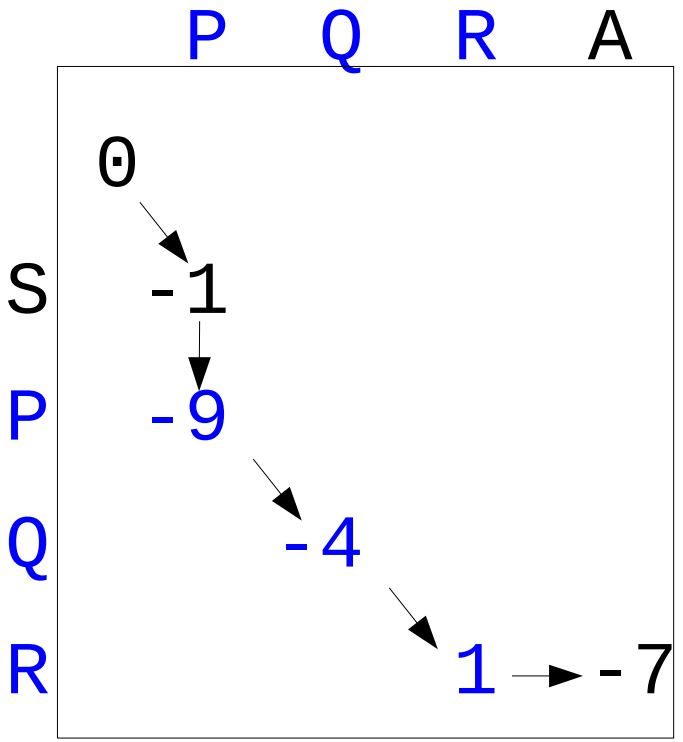
P-QRA
SPQR-

A rigorous method; partial description

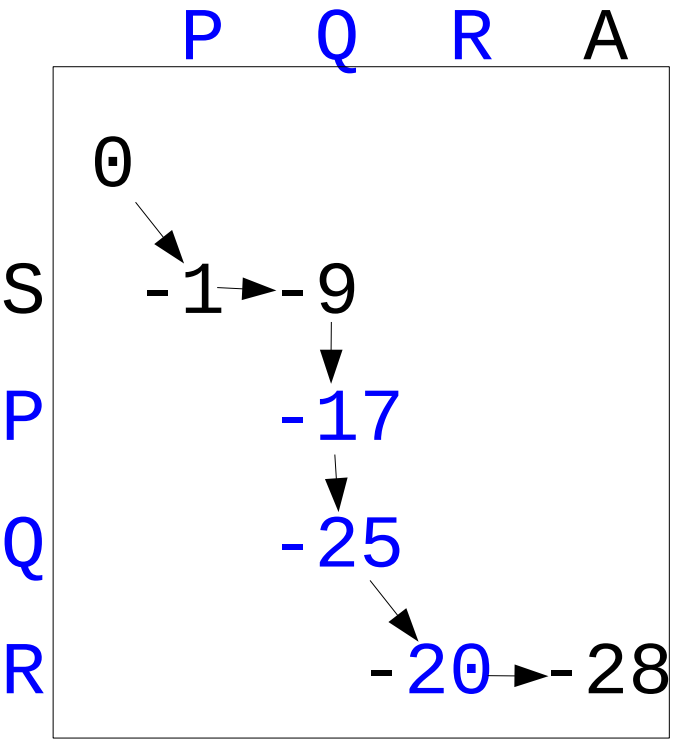
The best alignment competes with many others:



-PQRA
SPQR-



P-QRA
SPQR-



PQ--RA
S-PQR-

How to find the best path through the table?

Suppose we know the scores of 3 particular alignments:

	P	Q	R	A
S	0			
P				
Q			-1	-9
R			9	?

Each blue score corresponds to an incomplete alignment, unknown so far.

From these scores, we deduce the missing score.

How?

How to find the best path through the table?

Suppose we know the scores of 3 particular alignments:

		P	Q	R	A
	0				
S					
P					
Q				-1	-9
R				9	?

From the blue scores, we deduce the missing score.

3 possibilités:

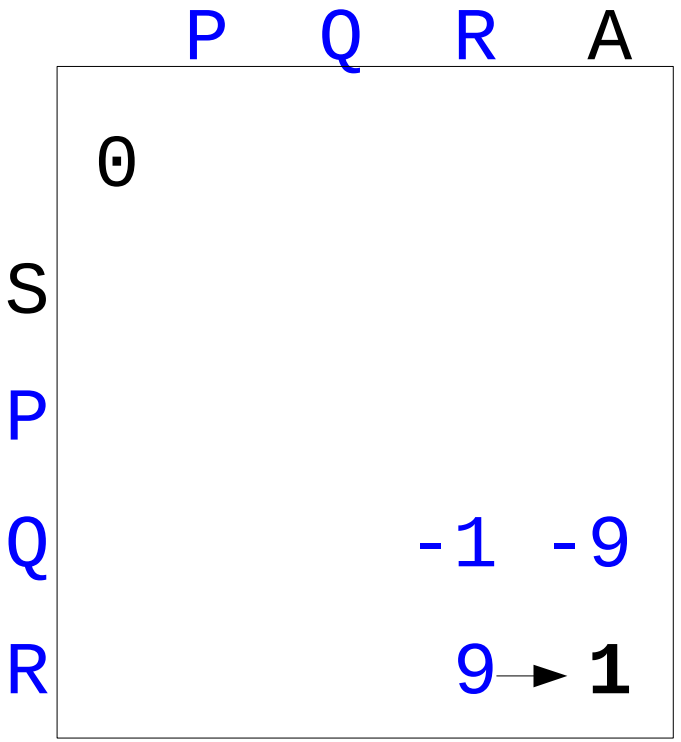
↘ score -1 -1 = -2

→ score 9 - 8 = 1

↓ score -9 - 8 = -17

How to find the best path through the table?

Suppose we know the scores of 3 particular alignments:



From the blue scores, we deduce the missing score.

3 possibilités:

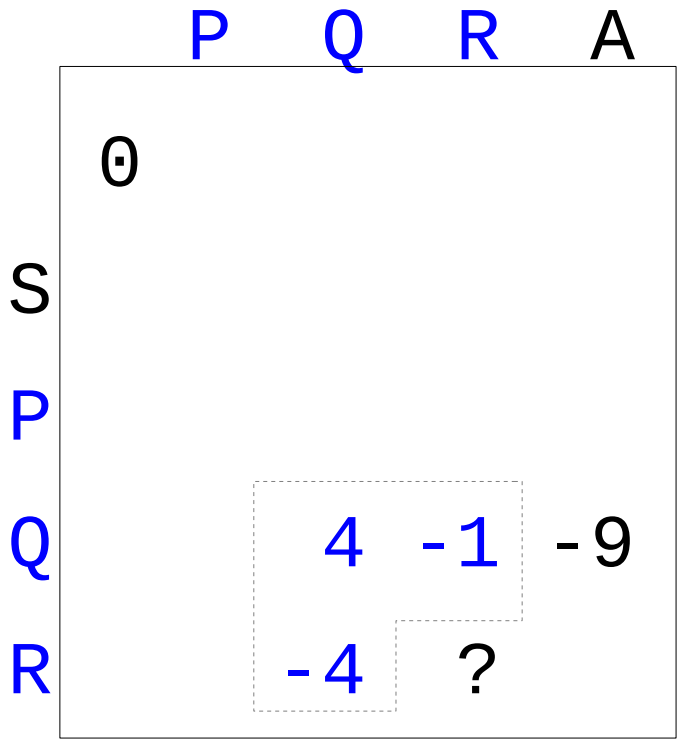
↘ score -1 -1 = -2

→ score 9 - 8 = 1

↓ score -9 - 8 = -17

How to find the best path through the table?

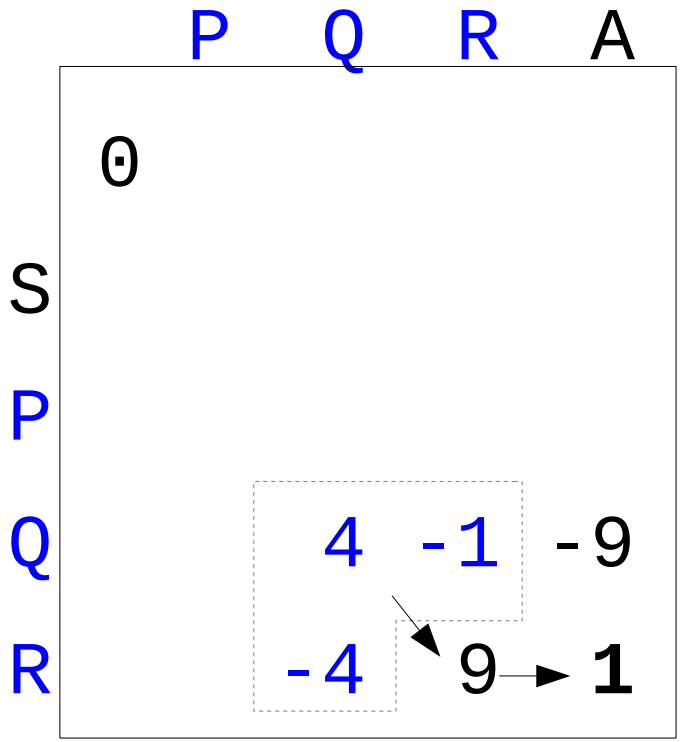
Suppose we know the scores of 3 particular alignments:



Now back up and repeat.

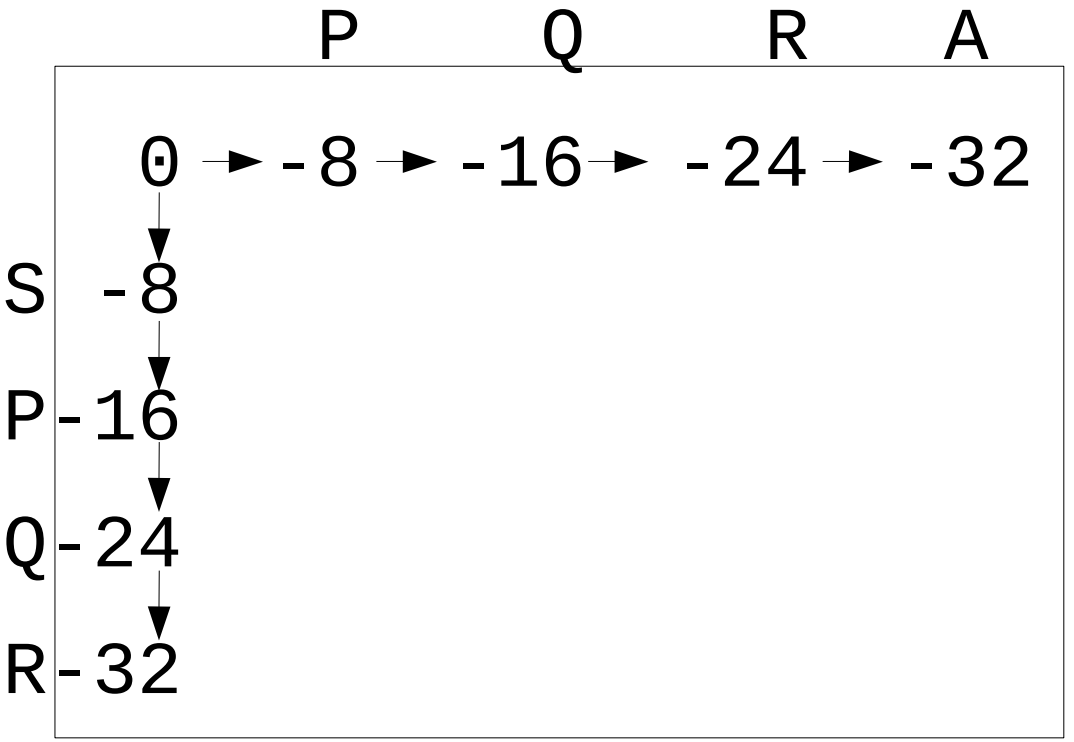
How to find the best path through the table?

Suppose we know the scores of 3 particular alignments:

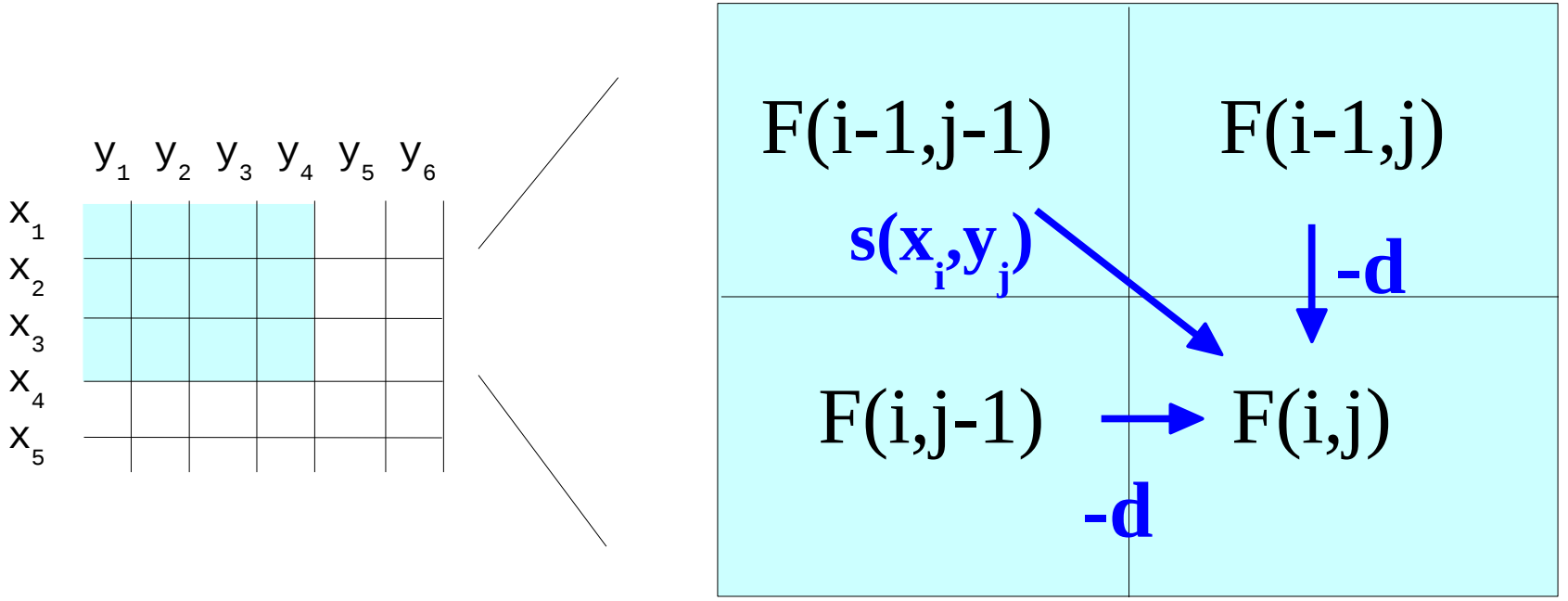


Now back up and repeat.

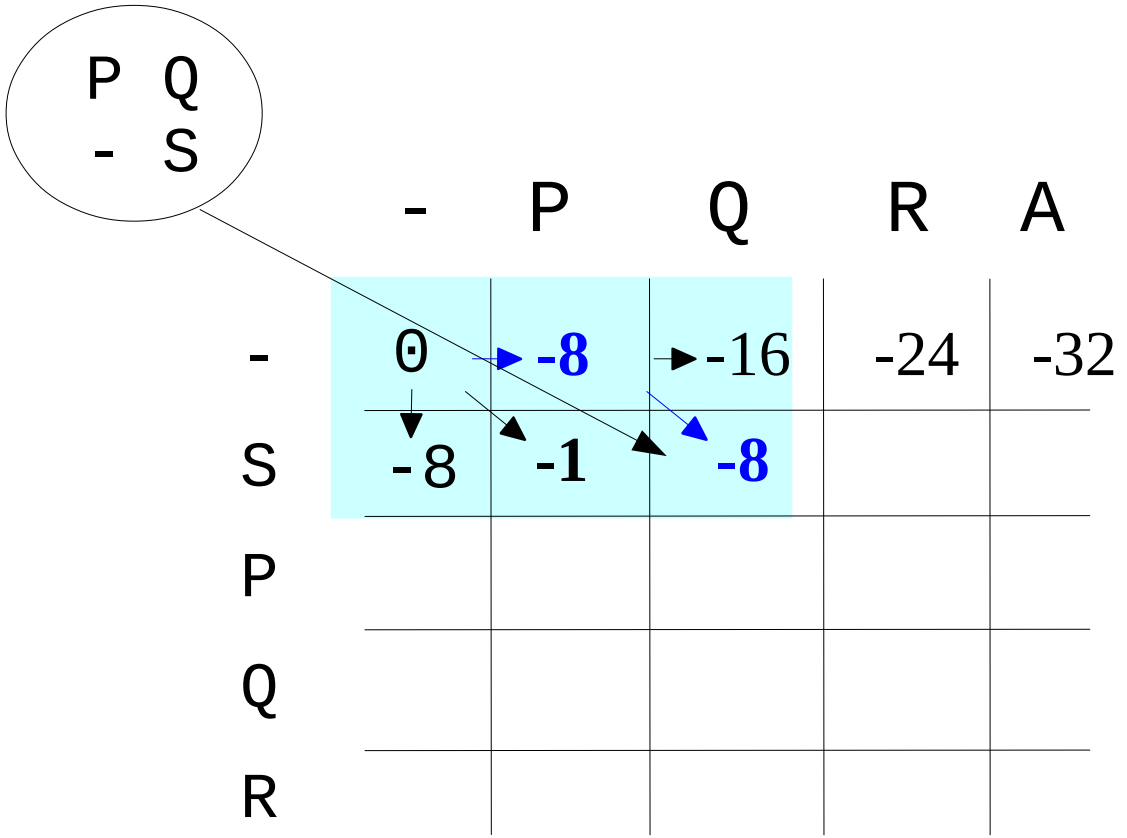
Needleman-Wunsch method: initialization



Recursive calculation of the score F(i,j)



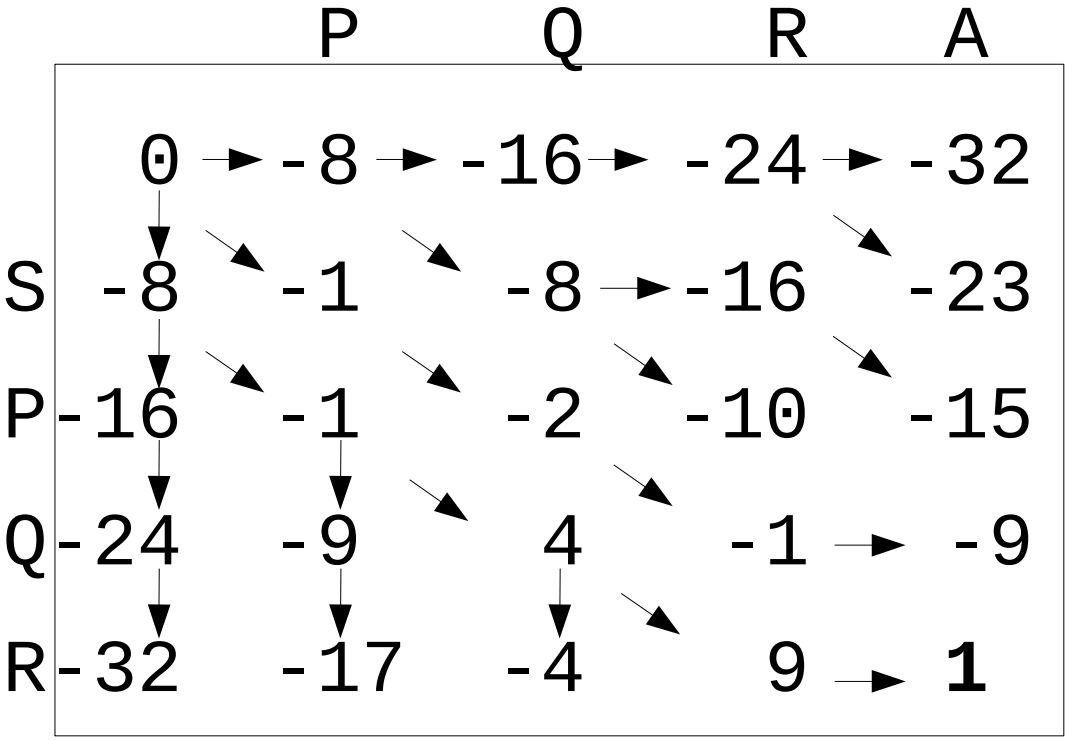
Extending F



Similarity matrix

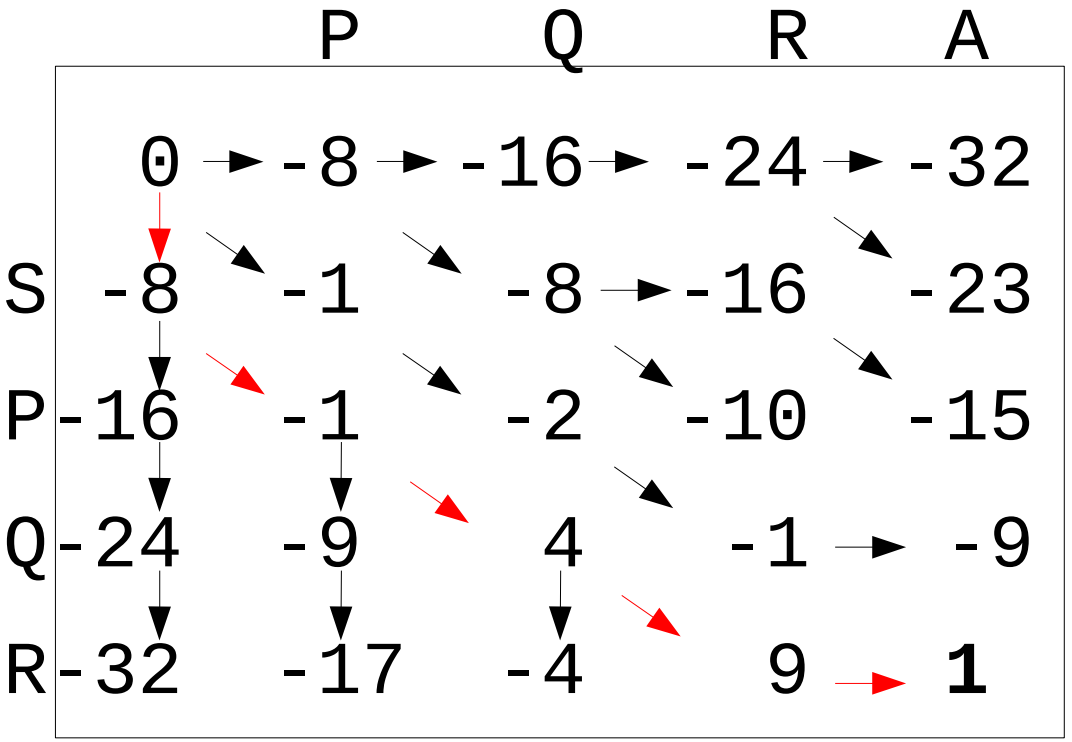
	P	Q
S	-1	0

Needleman-Wunsch method

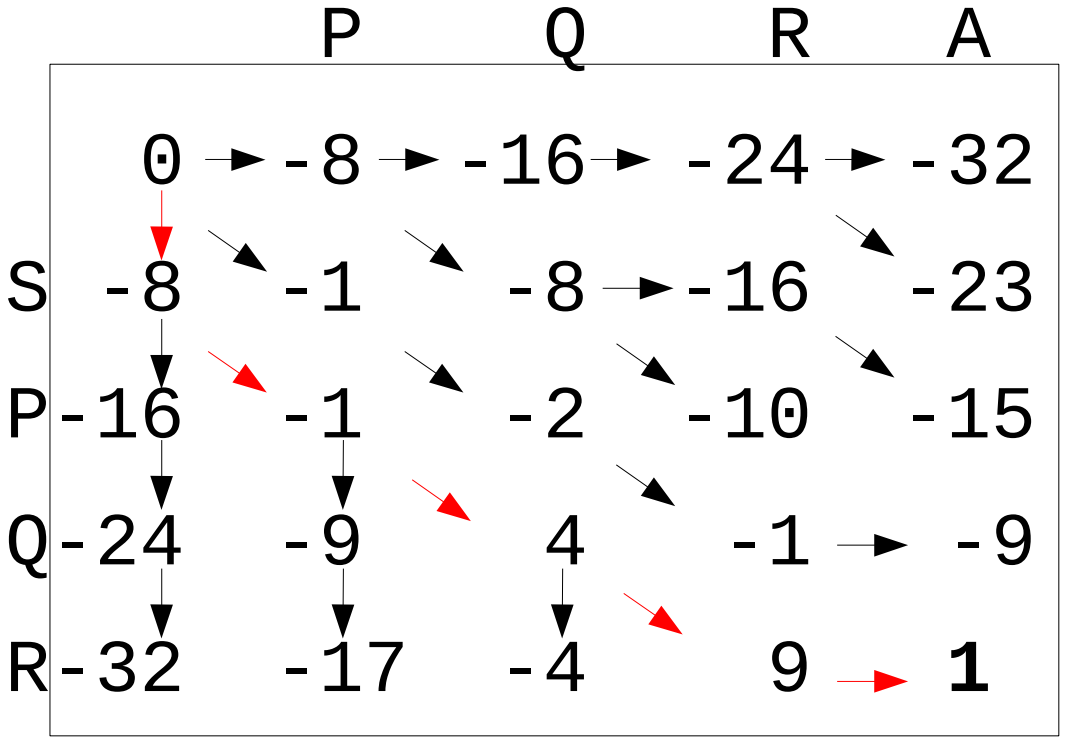


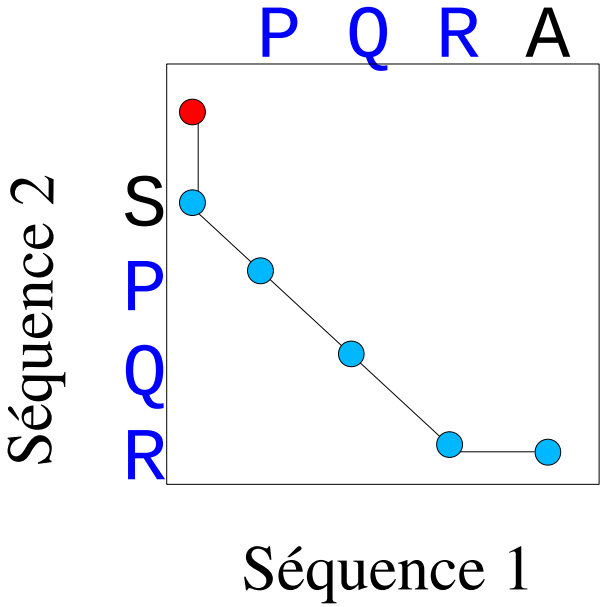
For more details, see Dardel & Képès book

Needleman-Wunsch method



Needleman-Wunsch method





-PQRA
SPQR-

Multiple alignements: principles

Homologues of the androsterone receptor identified using BLAST

#	ID Swissprot	Hit	Description	Score (bits)	E *	% Identity	Match Length
1	P15207	ANDR_RAT	Androgen receptor.	162	1e-40	100	73
6	P19091	ANDR_MOUSE	Androgen receptor.	162	1e-40	100	73
14	Q63449	PRGR_RAT	Progesterone receptor (PR)	136	1e-32	80	72
17	P06401	PRGR_HUMAN	Progesterone receptor (PR)	136	1e-32	80	72
21	P08235	MCR_HUMAN	Mineralocorticoid receptor (MR)	136	1e-32	79	72
33	P04150	GCR_HUMAN	Glucocorticoid receptor (GR)	131	3e-31	77	72
41	Q9YH32	ESR2_ORENI	Estrogen receptor beta (ER-beta)	99	3e-21	58	72
42	Q9YH33	ESR1_ORENI	Estrogen receptor (ER-alpha)	98	4e-21	55	72
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
343	Q9N4Q7	NH13_CAEEL	Nuclear hormone receptor nhr-13	54	8e-08	39	66
344	Q23294	NH11_CAEEL	Nuclear hormone receptor nhr-11	54	8e-08	42	66
345	O45460	NH54_CAEEL	Nuclear hormone receptor nhr-54	54	1e-07	37	67
346	Q09565	NH20_CAEEL	Nuclear hormone receptor nhr-20	51	7e-07	34	66
347	Q09587	NH22_CAEEL	Nuclear hormone receptor nhr-22	45	5e-05	32	66
349	P17672	E75B_DROME	Ecdysone-induced protein 75B	40	0.001	37	47
351	P20659	TRX_DROME	Trithorax protein.	31	0.74	26	49
355	P98164	LRP2_HUMAN	Lipoprotein receptor.	30	1.7	27	65

*E = expectation of number of random alignments with a higher score

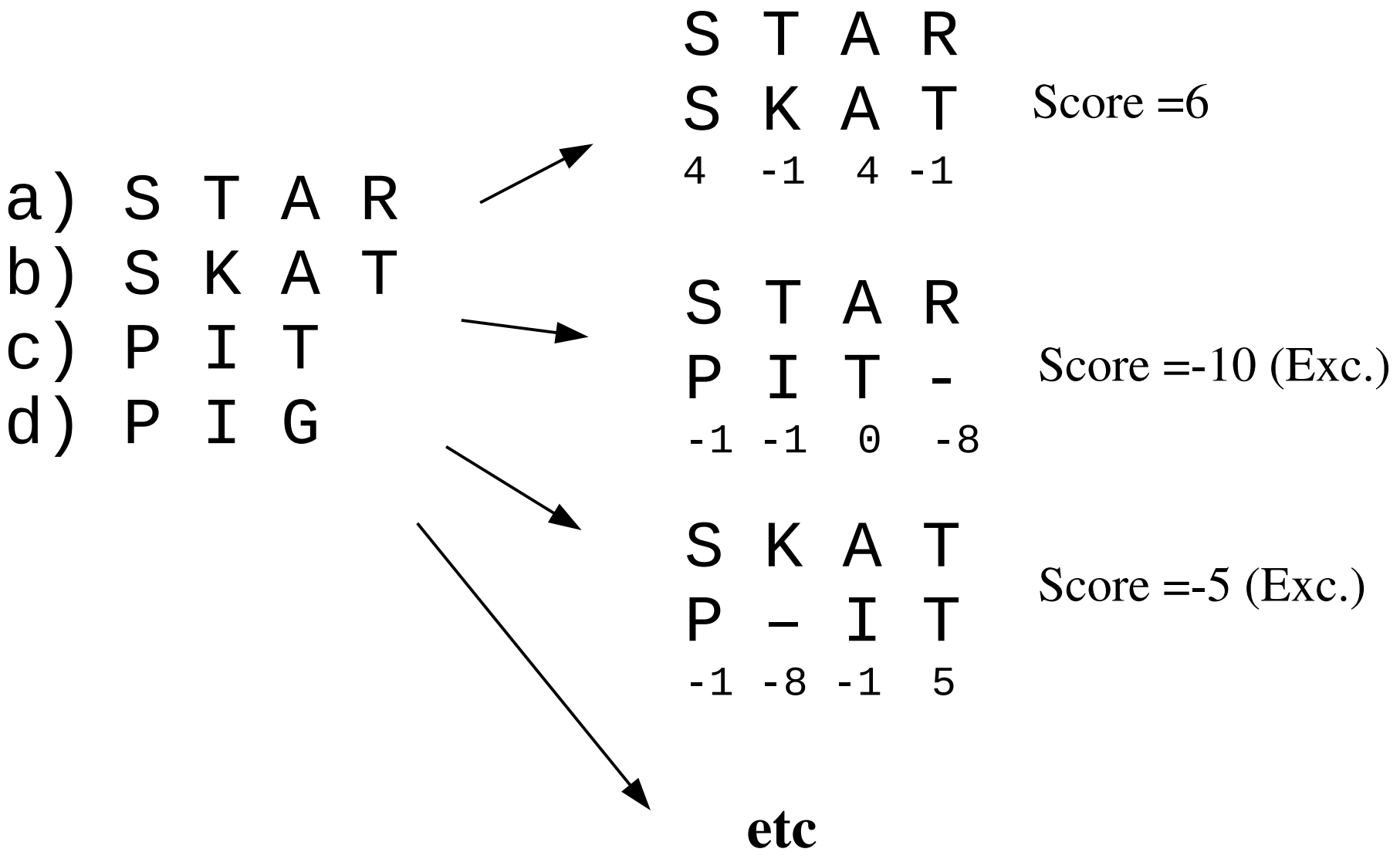
A heuristic method for multiple sequence alignment in three stages

1) Align each pair of sequences

2) ***Classify*** the sequences by similarity: “guide” tree

3) Incorporate the sequences progressively into an alignment, in a sensible order (determined by the tree)

Stage 1: align all sequence pairs

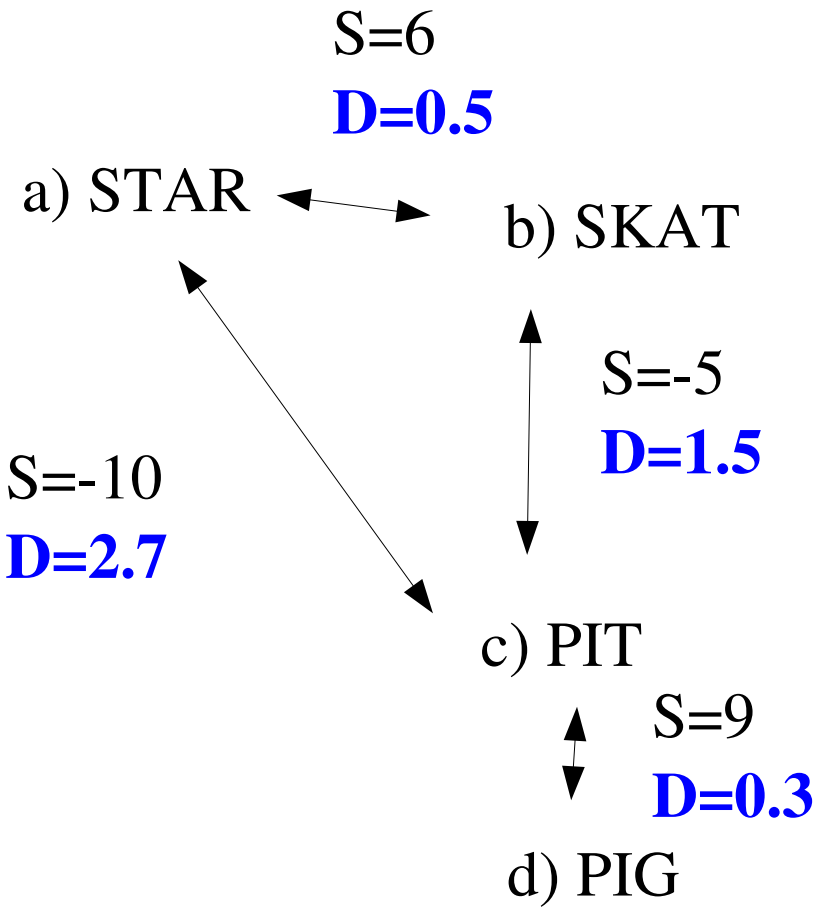


Stage 2: Classify the sequences by similarity: “guide” tree

Stage 3: Incorporate the sequences progressively into the alignment, in a sensible order

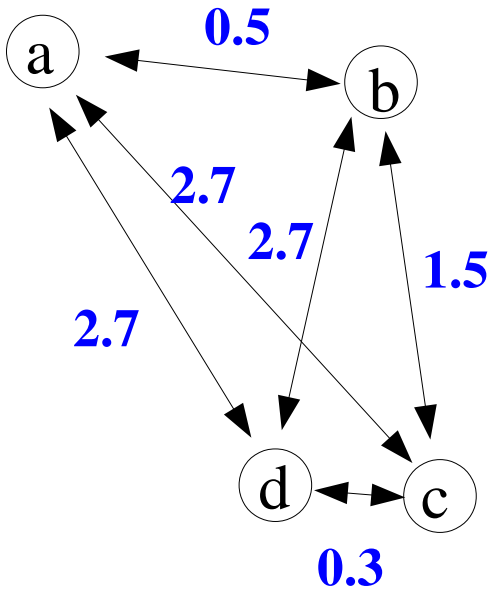
Stage 2: compute "distances" between sequences

Classification methods use distances, rather than similarity

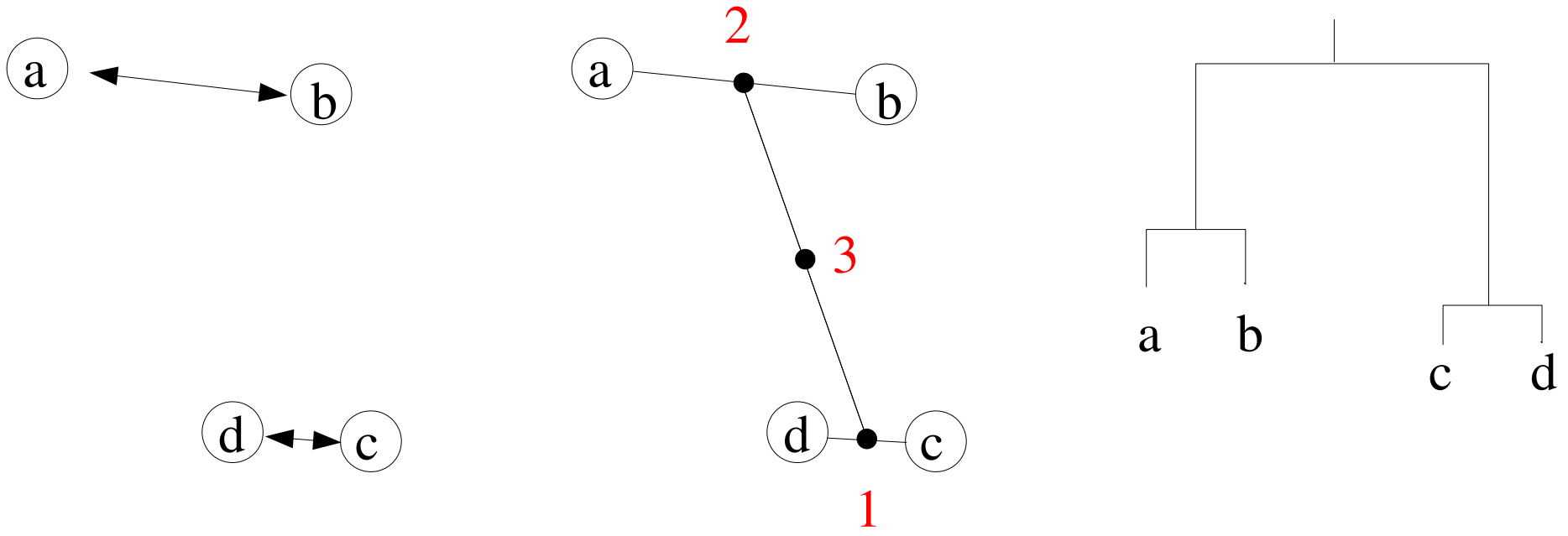


What would be your choice of distance?

Stage 2: compute “distances” between sequences

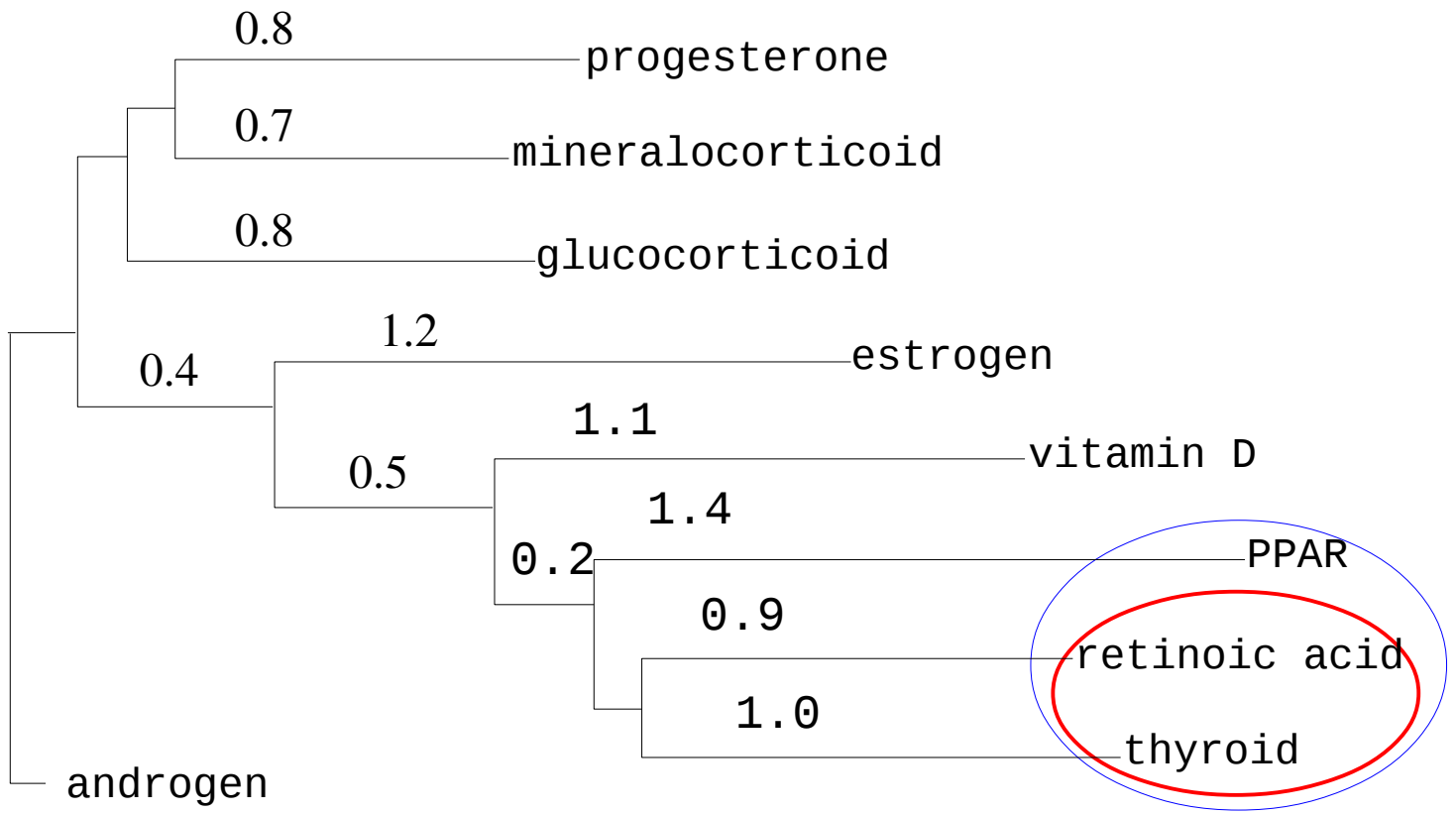


Stage 2: hierarchical classification or “guide” tree



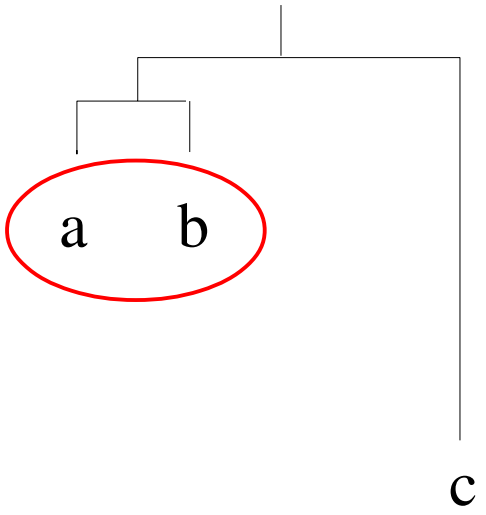
“Unweighted Pair Group Joining with Arithmetic Mean”

Stage 3: sequence alignment, moving up the tree



Stage 3: sequence alignment, moving up the tree

Example with 3 sequences

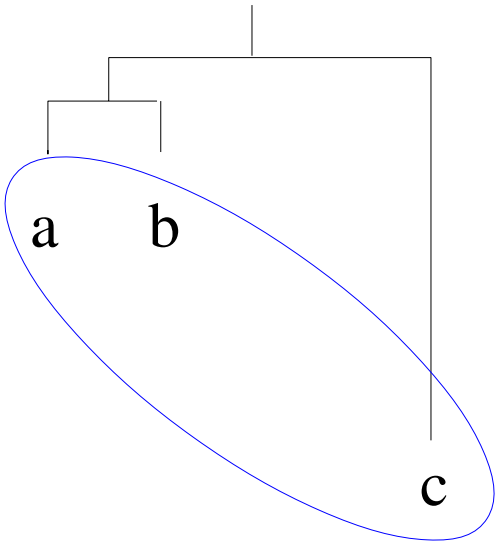


a with b

a) S T A R
b) S K A T

Stage 3: sequence alignment, moving up the tree

Example with 3 sequences



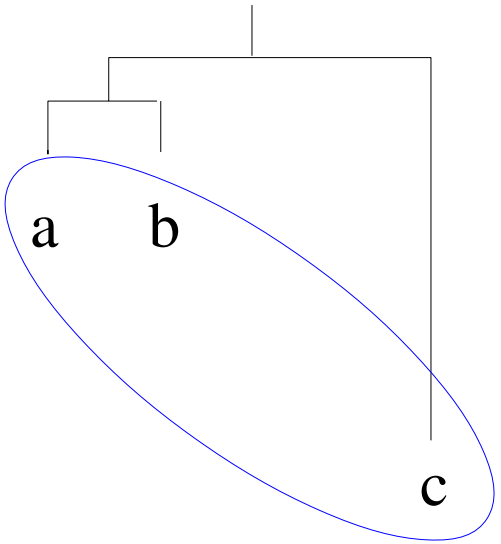
a with b

a) S T A R
b) S K A T

c with {a, b}

Stage 3: sequence alignment, moving up the tree

Example with 3 sequences



a with b

```
a) S T A R
b) S K A T
```

c with {a, b}

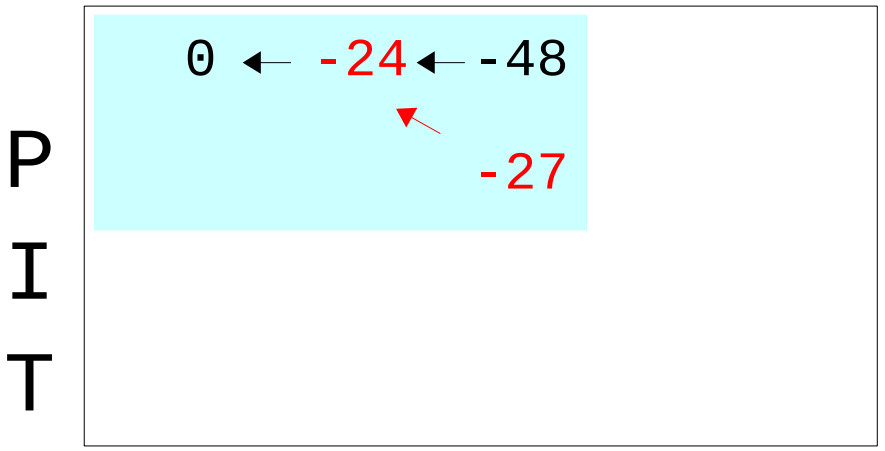
How to align a sequence with an alignment?

“sequence-profile” alignment

We need to generalize slightly our pairwise alignment method....

Sequence-“profile” alignment : dynamic programming, as before

S T A R
S T I R
S K A T



	P
T	-1
T	-1
K	-1
total	-3

gap penalities are tripled

Aligning a sequence with an alignment: computing a "mean" score

An alignment of 3 sequences:

```
S  [T]  A  R
S  [T]  I  R
S  [K]  A  T
```

→

	P
T	-1
T	-1
K	-1
total	-3

4th sequence y, to be aligned:

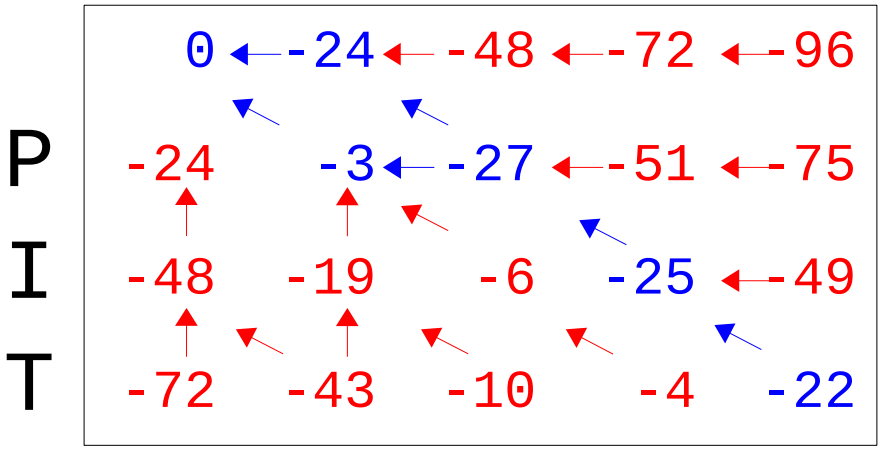
```
[P] I  T
```

$$s(a, \begin{bmatrix} b \\ b' \\ b'' \end{bmatrix}) = s(a,b) + s(a,b') + s(a,b'')$$

Sum over pairs

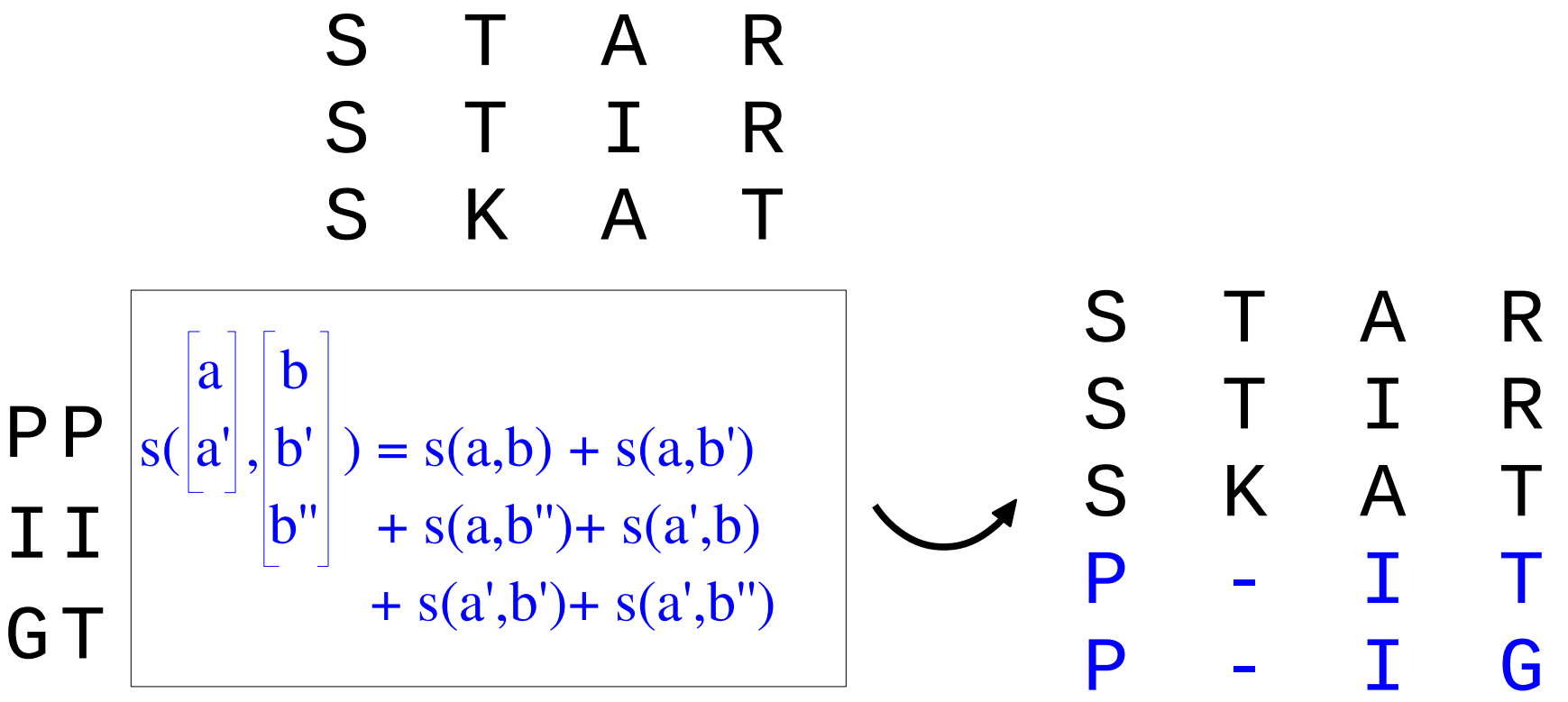
Sequence-“profile” alignment : dynamic programming

S T A R
S T I R
S K A T



S T A R
S T I R
S K I T
P - I T

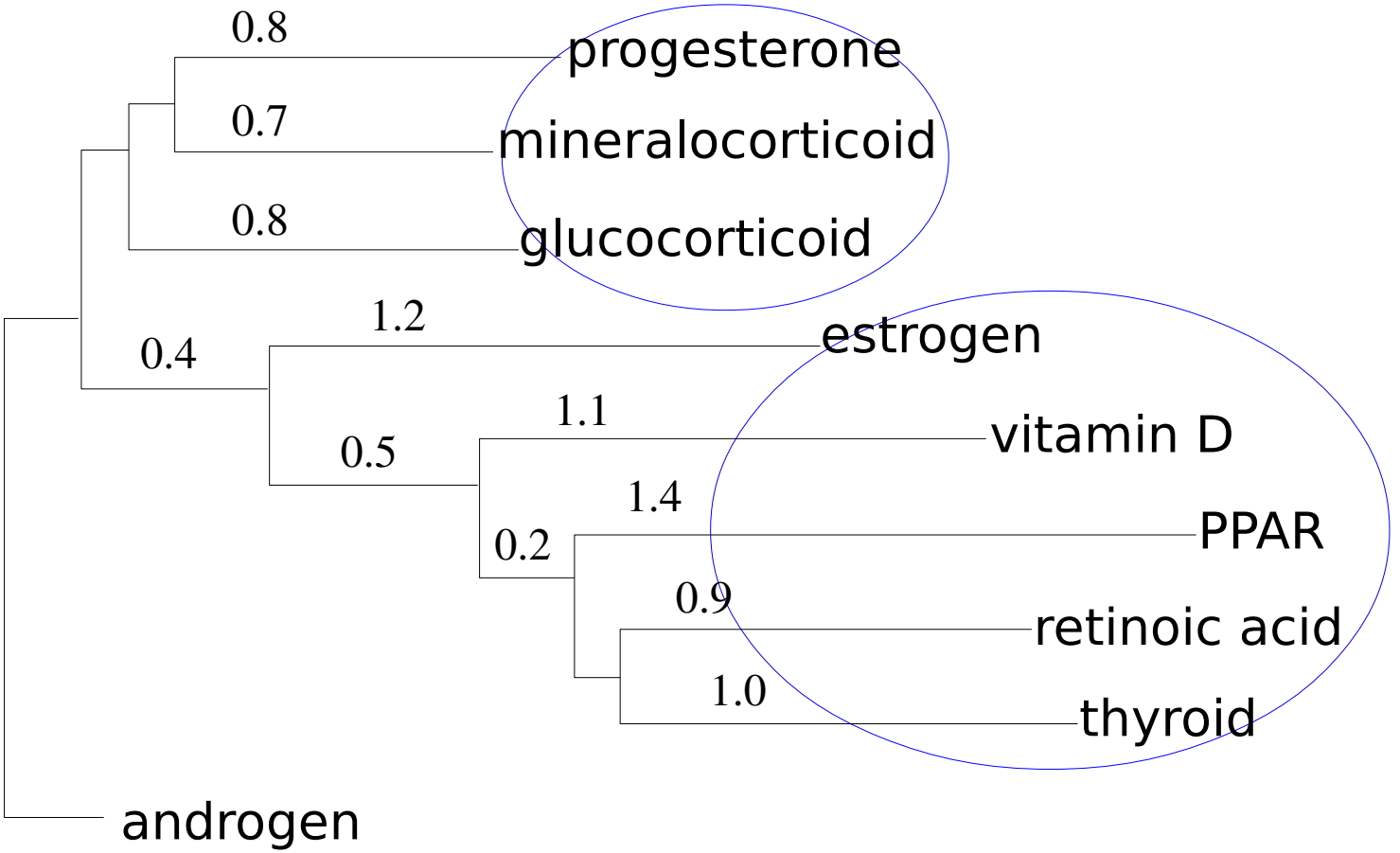
Aligning two alignments: “profile-profile” alignment



Sum over pairs

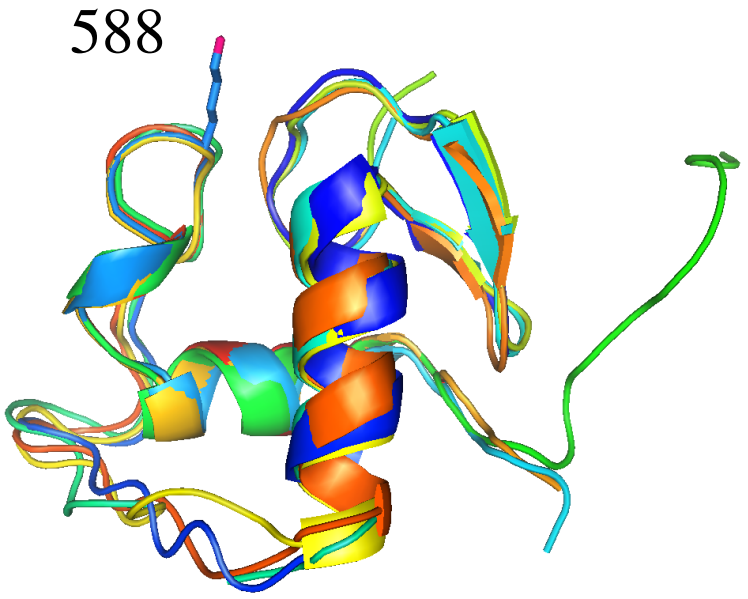
slight generalization of sequence alignment

Stage 3: align the sequences in the order of the guide tree



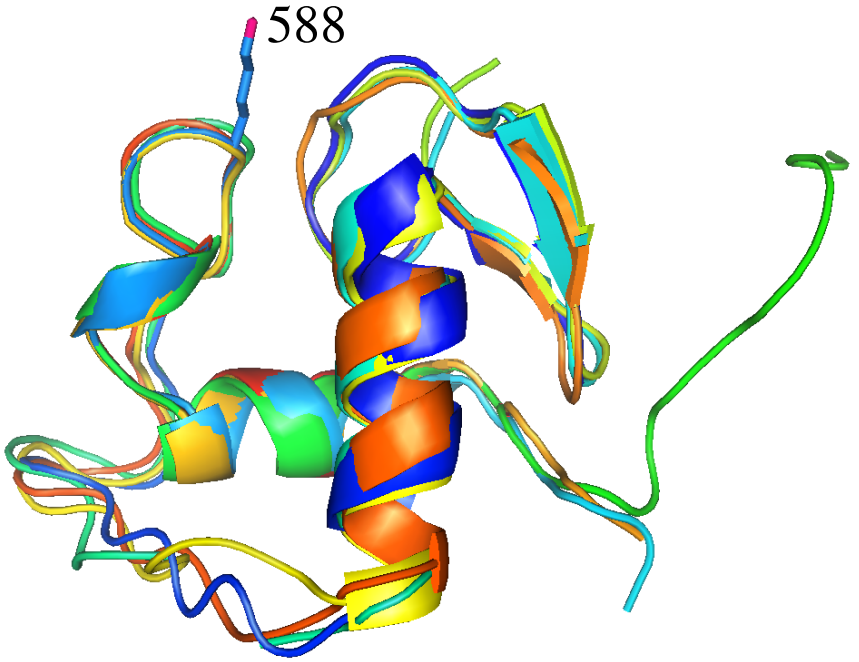
The final result

```
                                                584   588
androgen            VFFKRAAEG--KQKYLCASRNDCTIDKFRRKNCPSCRLRKCY
progesterone        VFFKRAVEG--HHNYLCAGRNDCIVDKIRRKNCPACRLRKCY
Mineralocorticoïd   VFFKRAVEG--QHNYLCAGRNDCIIDKIRRKNCPACRLQKCL
Glucocorticoïd      VFFKRAVEG--QHNYLCAGRNDCIIDKIRRKNCPACRYRKCL
Estrogen            AFFKRSIQG--HNDYMCPATNQCTIDKNRRKSCQACRLRKCY
Retinoic acid       GFFRRSIQK--NMVYTCHRDKNCIINKVTRNRCQYCRLQKCF
Vitamin D3          GFFRRSMKR--KALFTCPFNGDCRITKDNRRHCQACRLKRCV
Thyroid             GFFRRTIQKNLHPTYSCKYDSCCVIDKITRNQCQLCRFKKCL
                     **:*: .   :  :       * : *  *. *  ** ::*
```



Experimental testing is needed

- Site-directed mutagenesis of conserved residues
- Detecting an interaction with a substrate or inhibitor
- Determination of the 3D-structure!



Matinées 9h30 – 12h30 Après midis 14h – 17h30

dates	chapitres de la matinée (salles)	après-midi
17/11	Alignements de séquence (SI31)	TP (SI31)
24/11	Modélisation (SI31)	TP (SI31)
28/11	Reconnaissance moléculaire 1 + TD	TP (SI35)
29/11	Modélisation par homologie + TD	TP (SI35)
3/1	Reconnaissance moléculaire 2 + TD	TP (SI31)
5/1	TP	TP (SI31)
12/1	TD	TP (SI31)

19/1 **Evaluation:** rapport écrit sur un des TPs + contrôle oral

thomas.simonson@polytechnique.edu thomas.gaillard@polytechnique.edu
Laboratoire de Biologie Structurale de la Cellule, Ecole Polytechnique

http://biology.polytechnique.fr/biocomputing/teach.html