Modélisation et génome

ou

Bioinformatique

ou

Biologie "Computationnelle"

IDE'E D'UNE ECHELLE

DES ETRES NATURELS.

L'HOMME.

Orang-Outang.

Singe.

QUADRUPEDES.

Ecureuil volant.

Chauvesouris.

Autruche.

OISEAUX.

Oiseaux aquatiques.

Oiseaux amphibies.

Poissons volans.

POISSONS.

Poissons rampans.

Anguilles.

Serpens d'eau.

SERPENS.

Limaces.

Limaçons.

COQUILLAGES.

Vers à tuyau.

Teignes.

INSECTES.

Gallinsectes.

Tenia, ou Solitaire.

Polypes.

Orties de Mer.

Sensitive.

PLANTES.

Lychens.

Moisissures.

Champignons, Agarics.

Truffes.

Coraux & Coralloïdes.

Lithophytes.

Amianthe.

Talcs, Gyps, Sélénites.

Ardoises.

PIERRES.

Pierres figurées.

Crystallisations.

SELS.

Vitriols.

METAUX.

DEMI-METAUX.

SOUFRES.

Bitumes.

TERRES.

Terre pure.

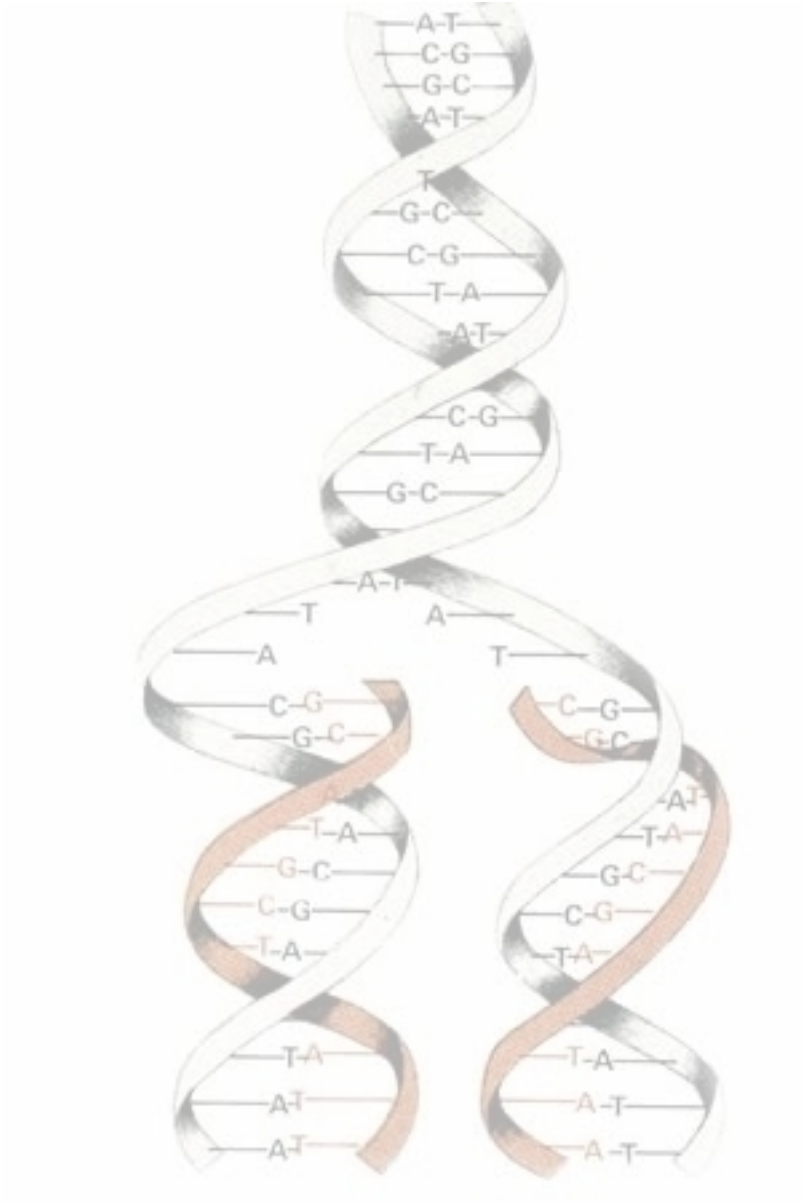
EAU.

AIR.

FEU.

Matieres plus subtiles.

Charles Bonnet, 1745



ADN

ARN

protéine

assemblée supramoléculaire

réseaux et voies

cellule

tissu

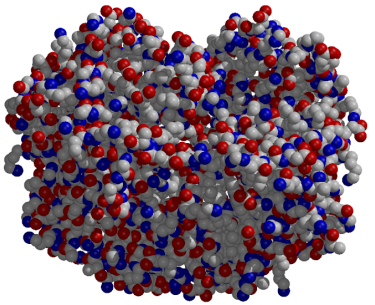
organisme

population

Comprendre la structure et fonction des biomolécules

AUGCGCUUAUAGCCAAGG:

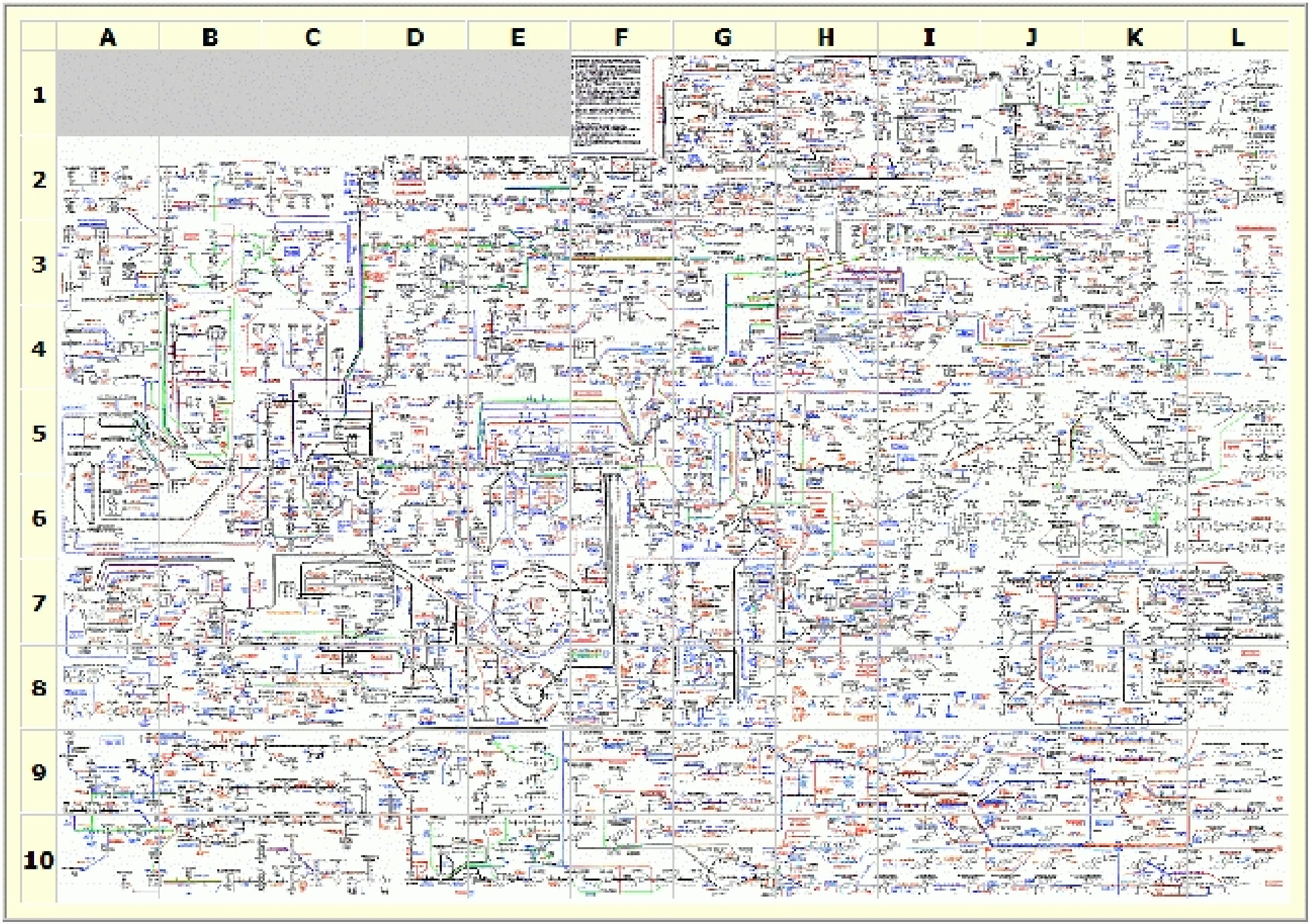
Séquence → **Structure**



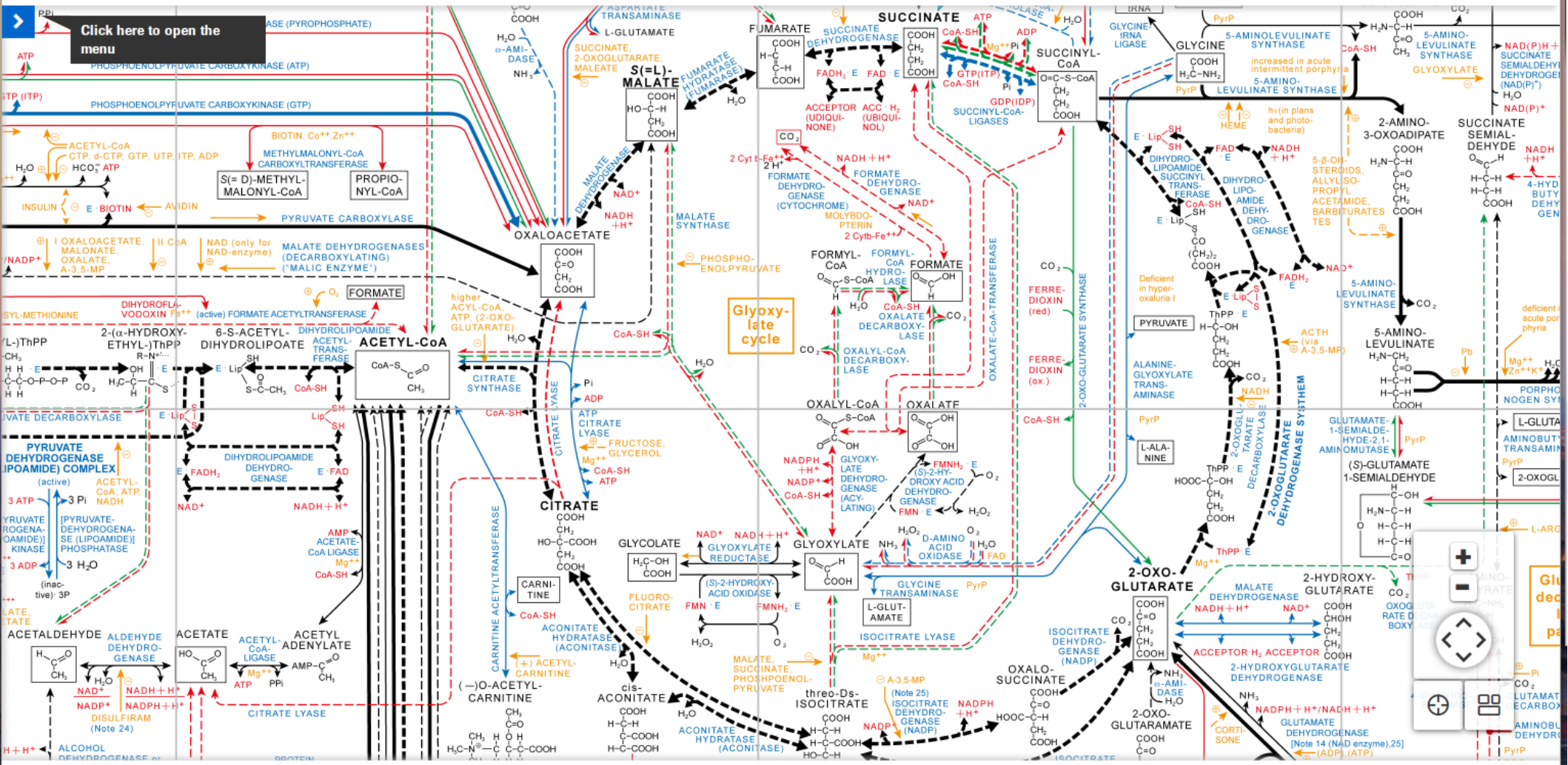
Fonction

Comprendre leur intégration dans la cellule (~10^9 macromolécules)

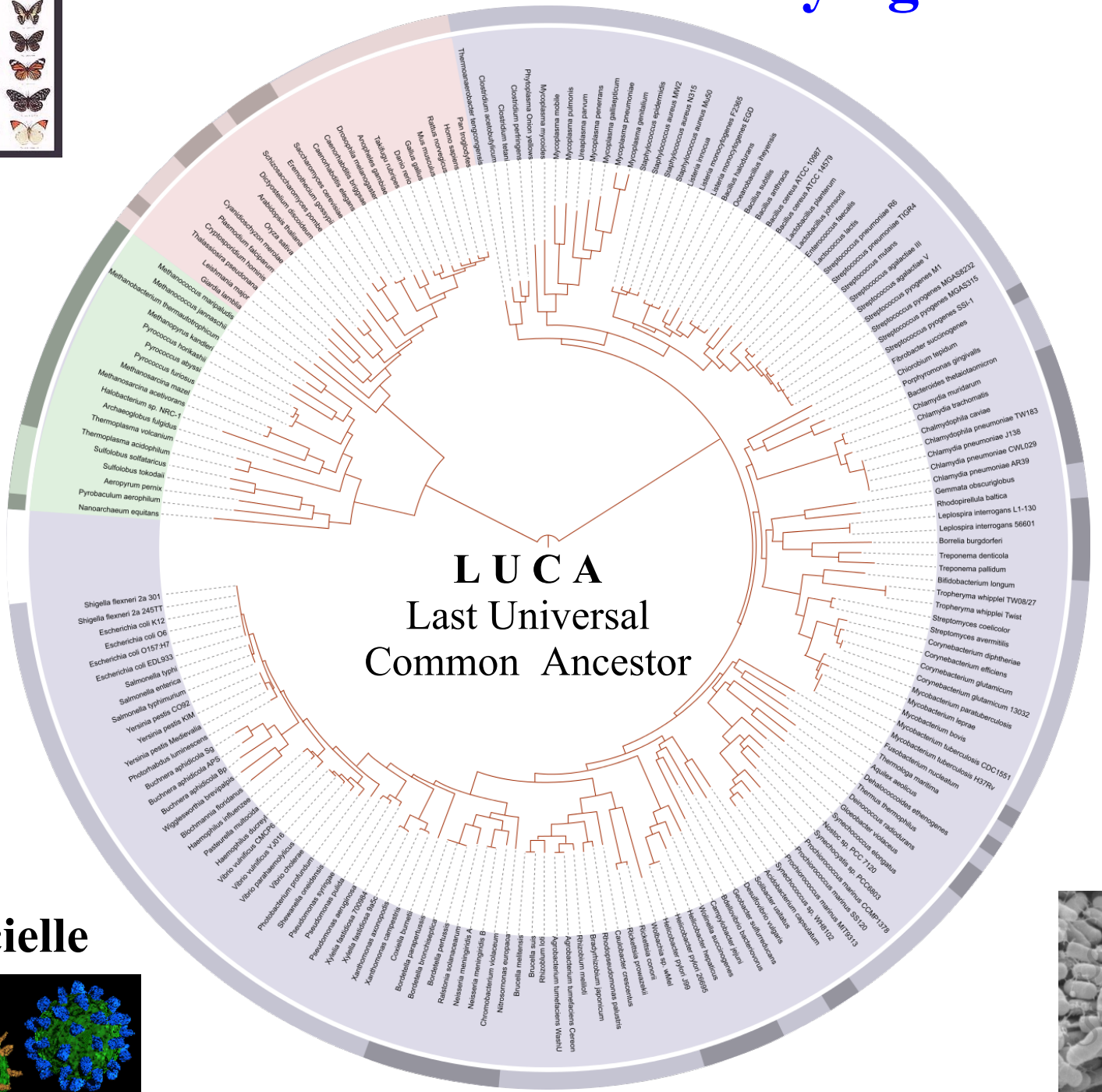
Biologie "intégrative" ou "systémique" (web.expasy.org/pathways)



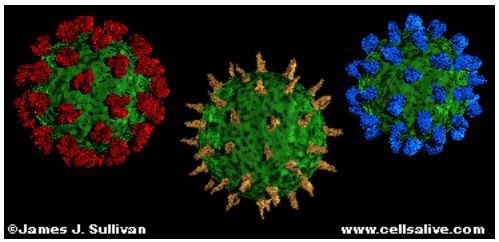
(web.expasy.org/pathways)

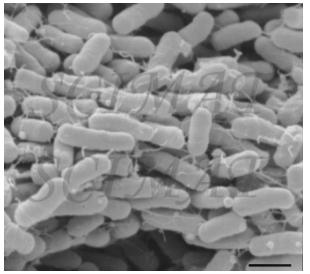


Phylogénie et évolution

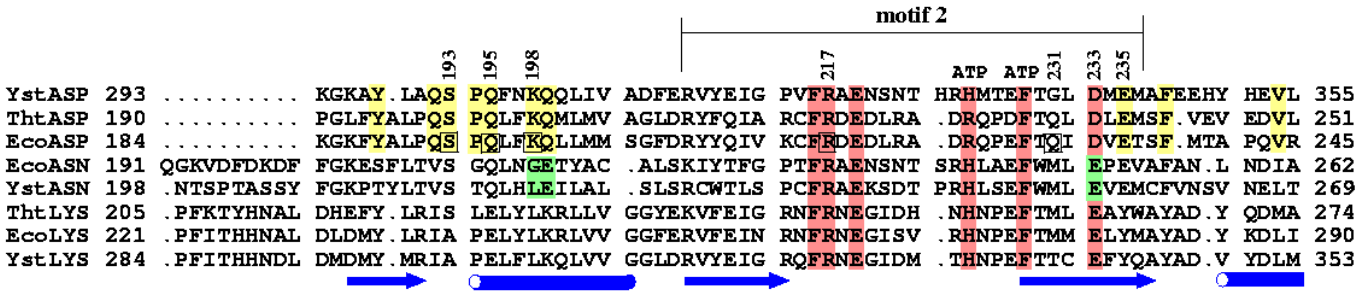


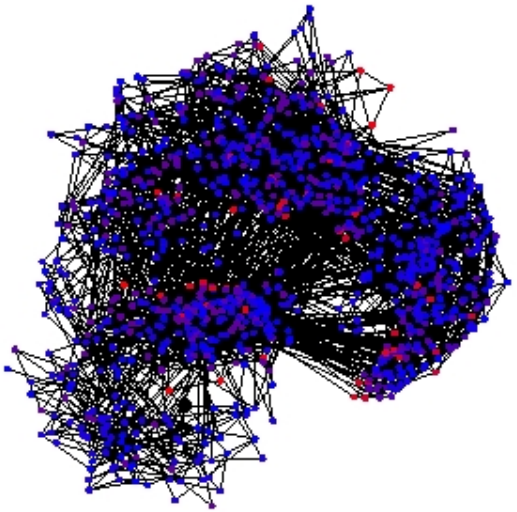
Vie artificielle

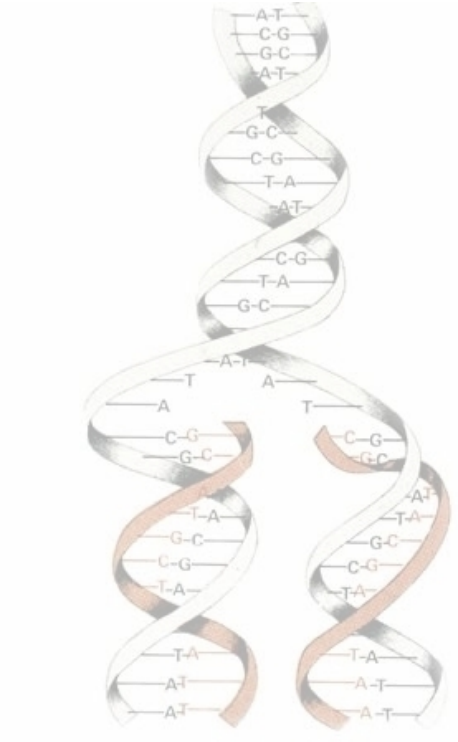


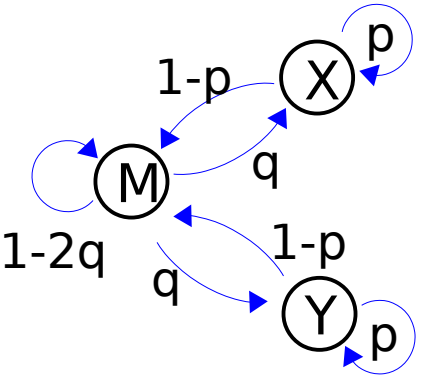


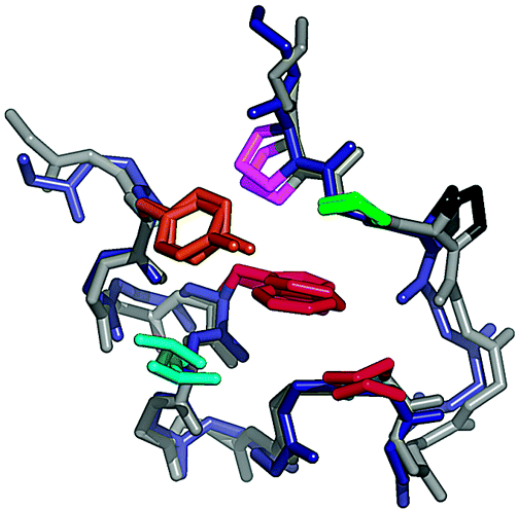
Les outils de l'informatique, des statistiques, de la chimie, de la physique











Organisation du cours

Thomas Simonson (Dépt Bio)
thomas.simonson@polytechnique.fr

2/12	Comparaisons de séquences, I
9/12	Comparaisons de séquences, II
16/12	Protéines
6/1	Prédiction de structures d'ARN
13/1	Phylogénie (T Gaillard)
20/1	Analyse statistique du génome
27/1	Organisation et séquençage du génome
3/2	Réseaux cellulaires
10/2	Evolution moléculaire

Petites classes

Amphi Curie

Apportez vos ordinateurs pour les PC 2, 3

thomas.simonson@polytechnique.edu

thomas.gaillard@polytechnique.edu

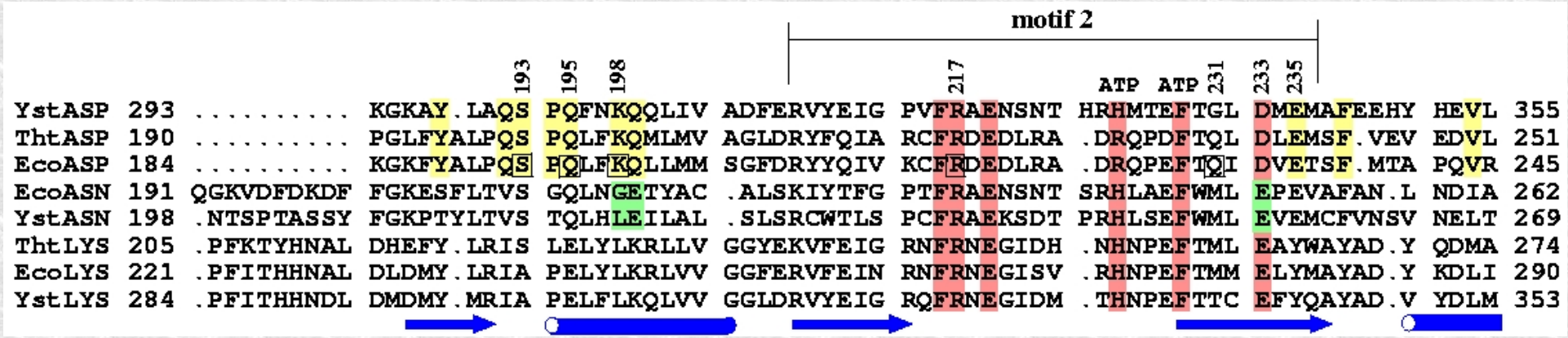
Document écrit (chaps 1-7)

Cf aussi biology.polytechnique.fr/biocomputing
cf. biblio en fin de cours

Contrôle écrit

Langue = français + anglais

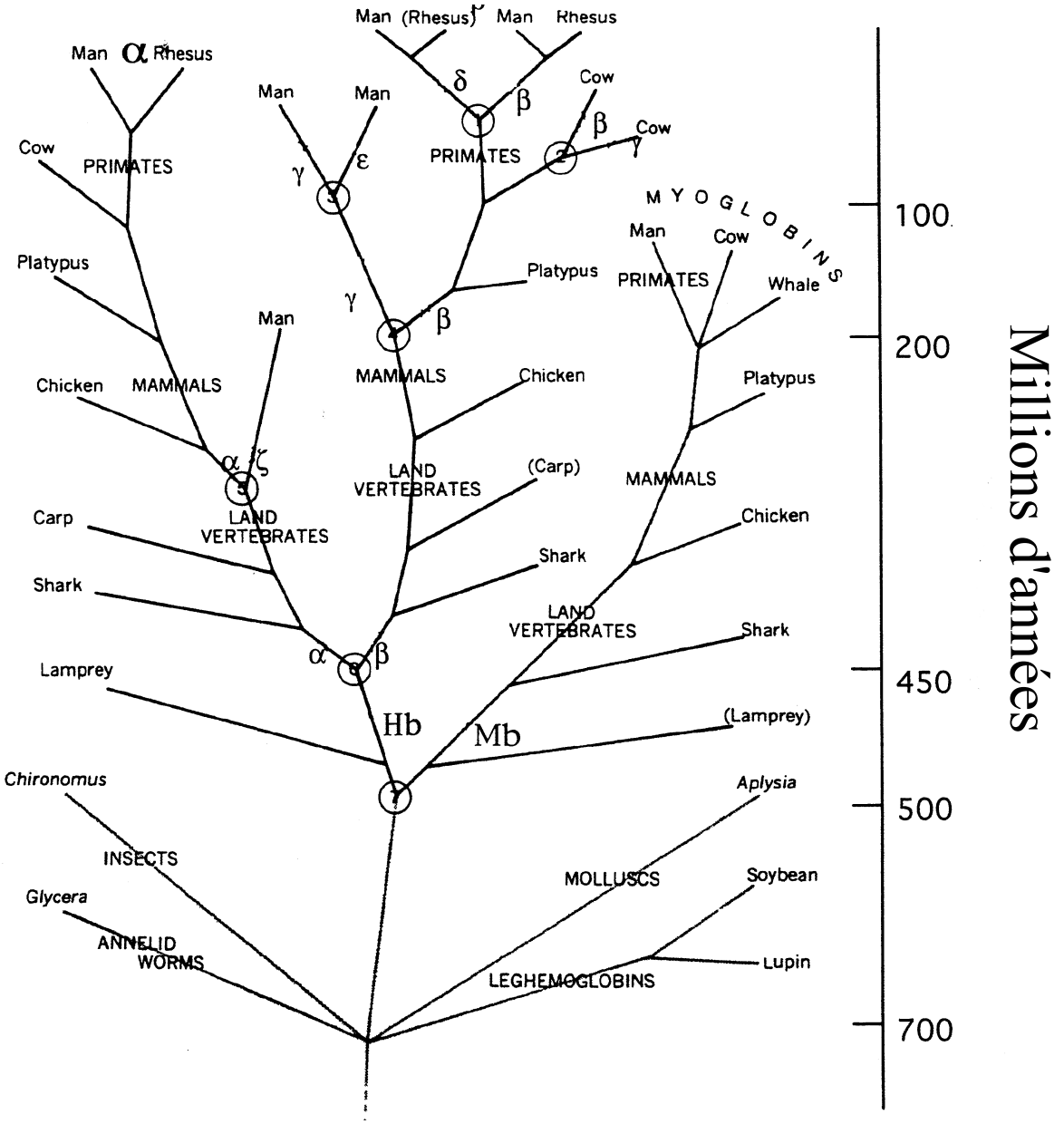
Comparaison et alignement de séquences. I.



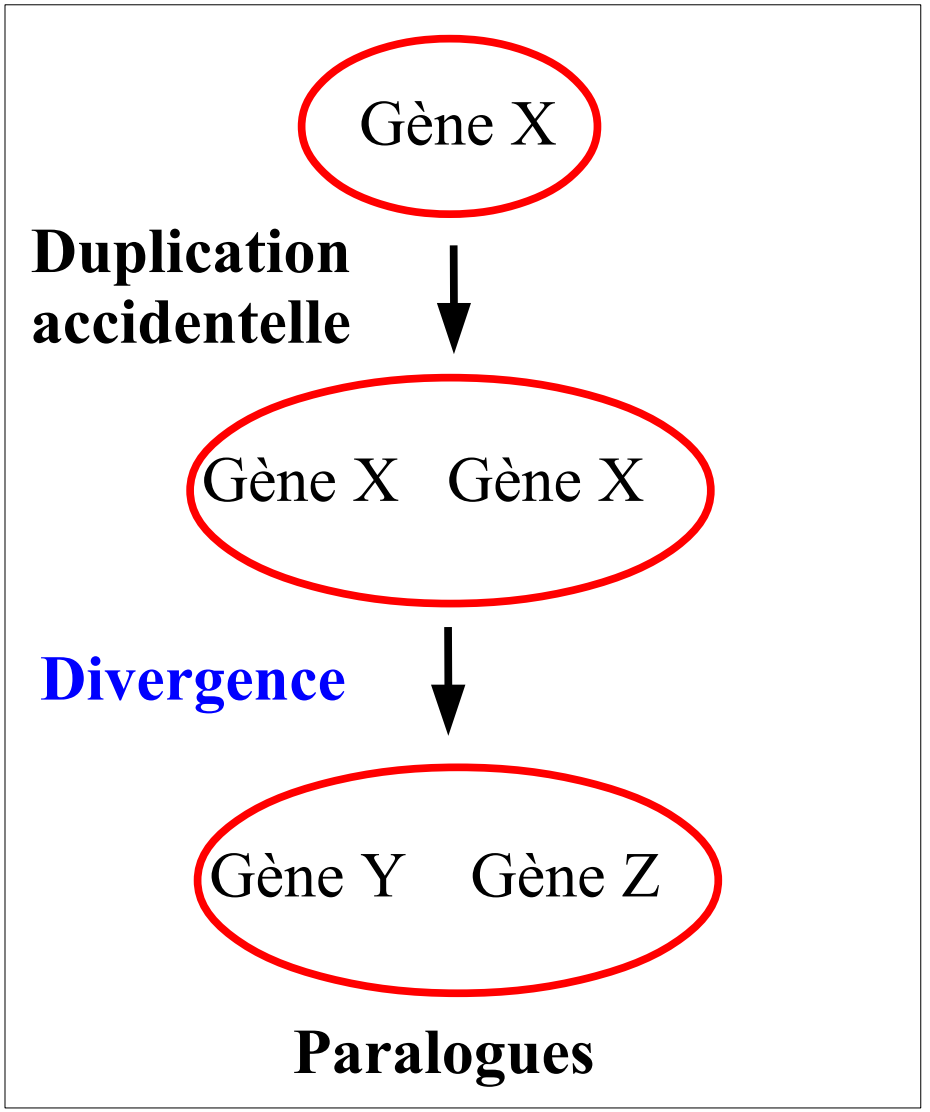
- **Comparaisons de séquences: pourquoi faire?**
- **L'alignement de séquences comme modèle d'un processus évolutif**
- **Alignement de séquences: algorithmes**

1) Les protéines similaires ont généralement un ancêtre commun: Exemple des globines

Arbre phylogénétique retraçant l'évolution des globines

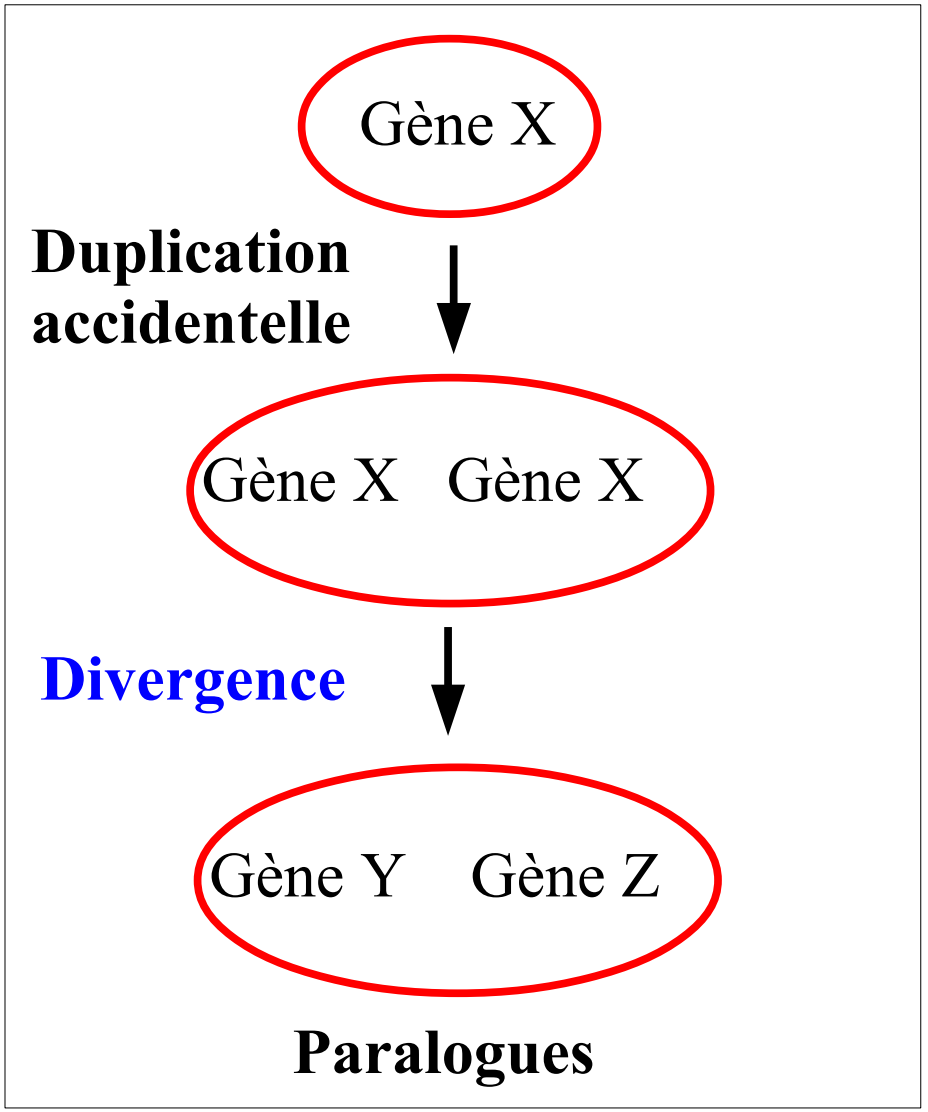


La duplication conduit à des gènes paralogues



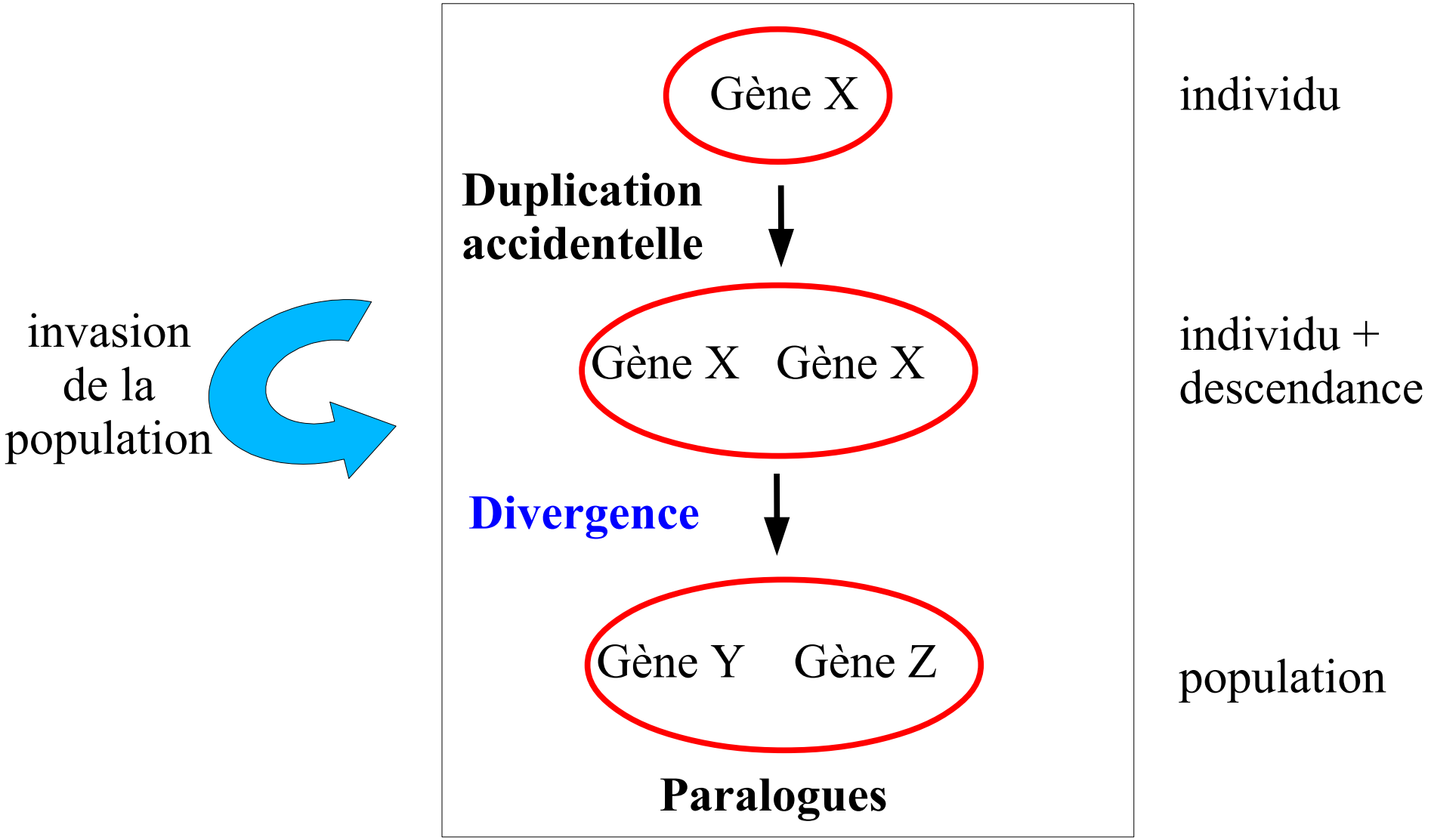
Les paralogues changent généralement de fonction

La duplication conduit à des gènes paralogues



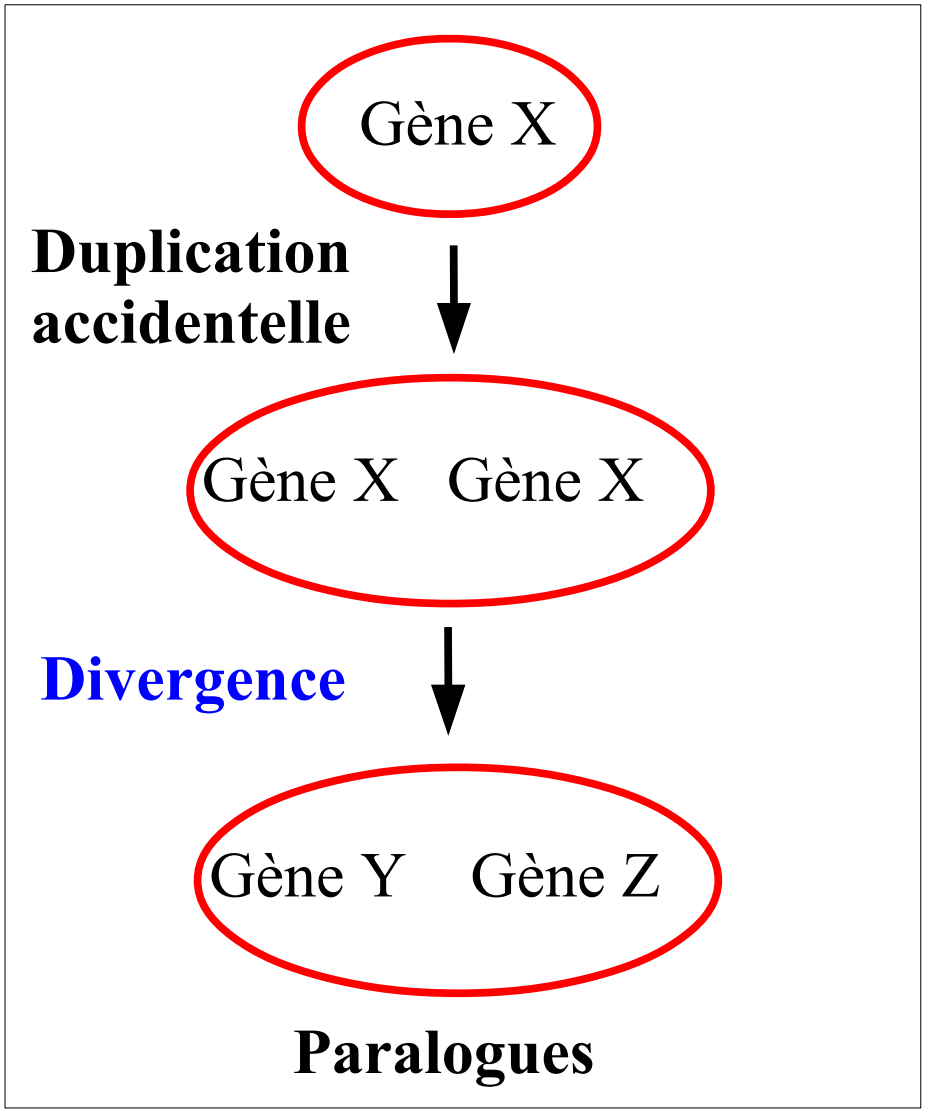
Les paralogues changent généralement de fonction

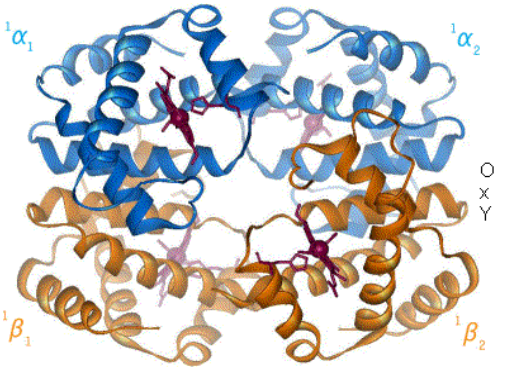
La duplication conduit à des gènes paralogues



Les paralogues changent généralement de fonction

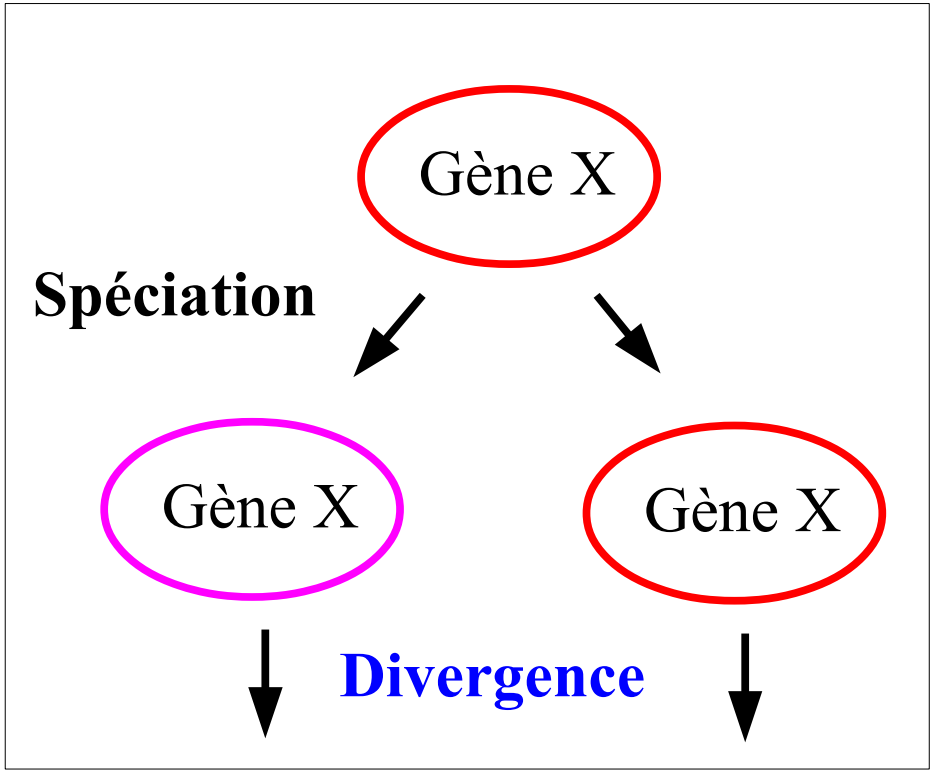
La duplication conduit à des gènes paralogues





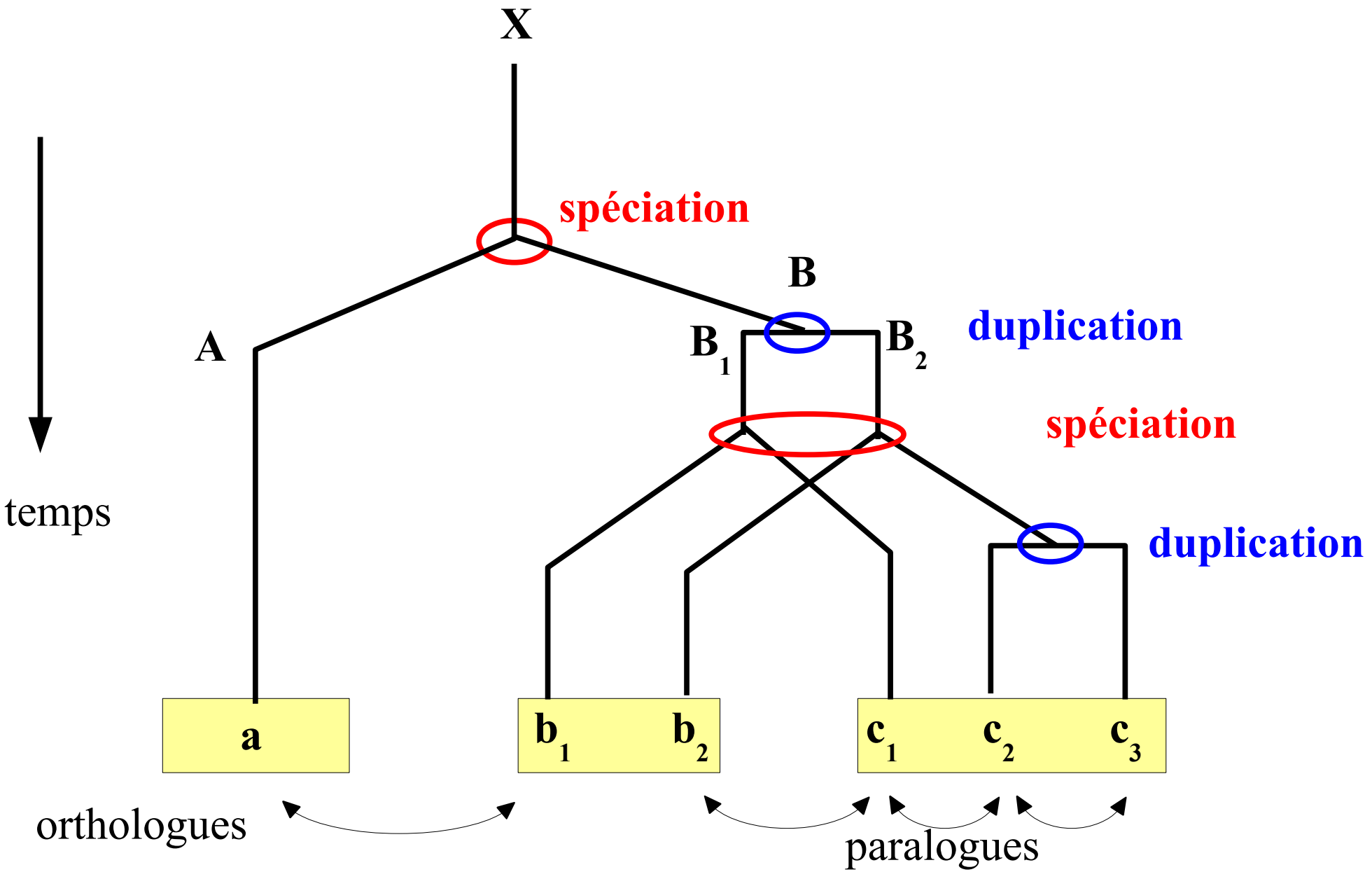
Les paralogues changent généralement de fonction

La spéciation produit des gènes orthologues

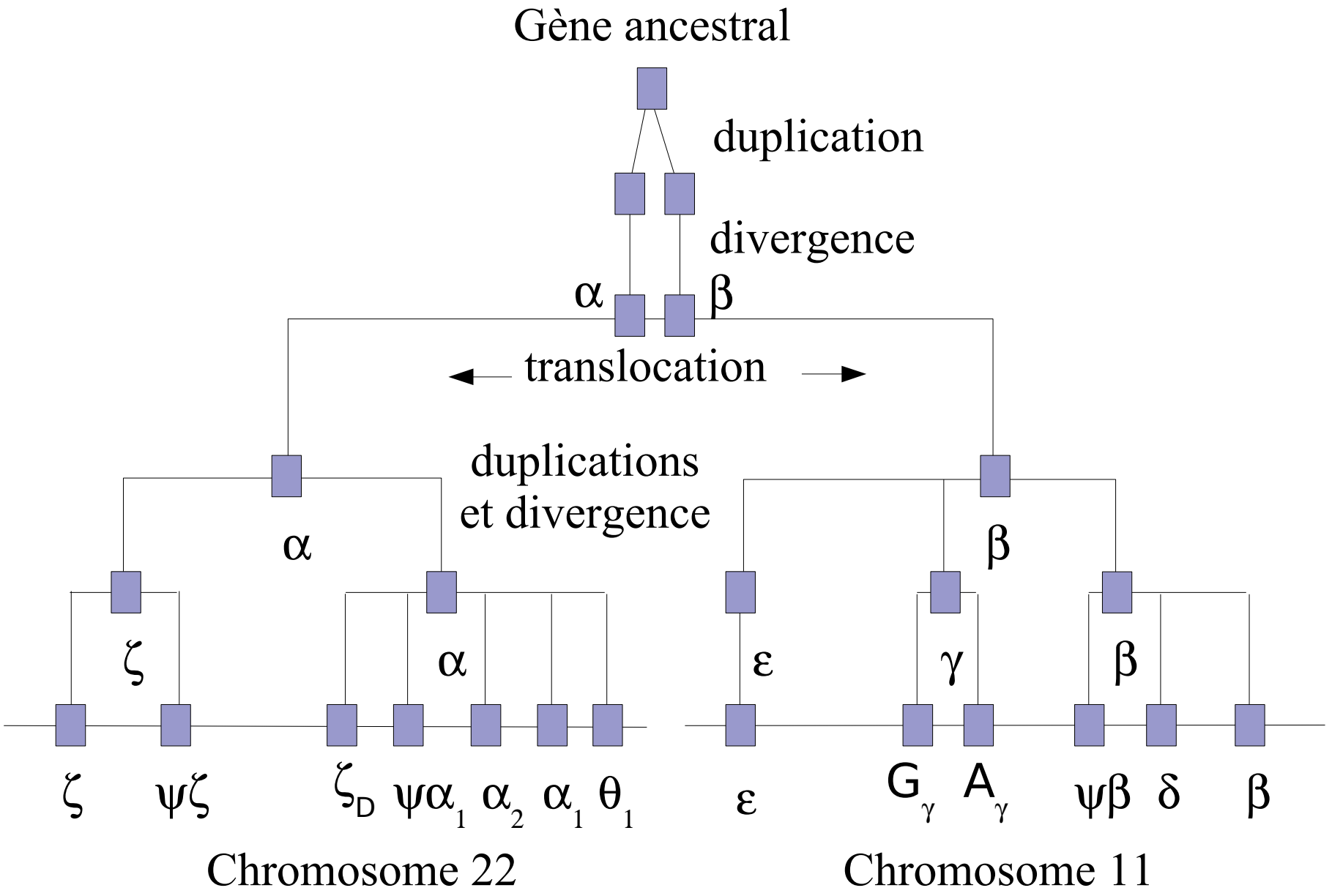


Les gènes orthologues conservent la fonction

Evolution d'un gène par spéciation, duplication et divergence

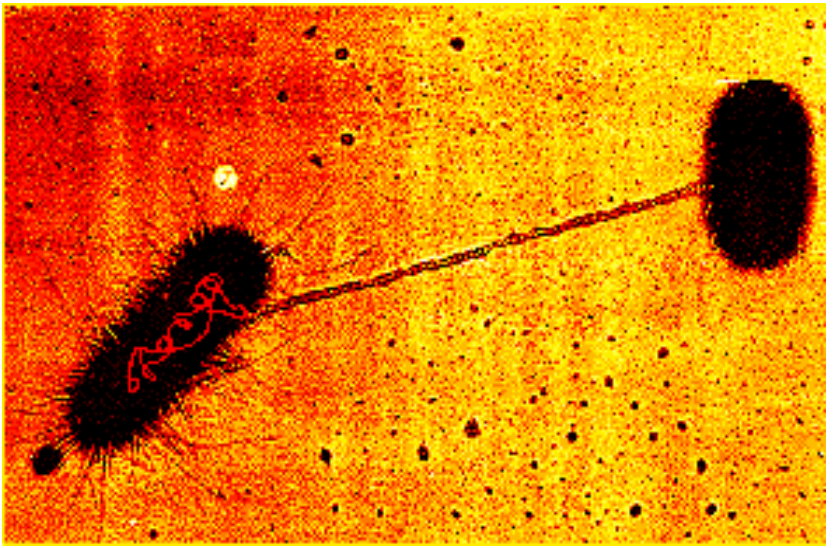


Gènes et pseudogènes d'hémoglobine



Transfert horizontal chez les bactéries

Un plasmide en cours de transfert d'une bactérie à une autre



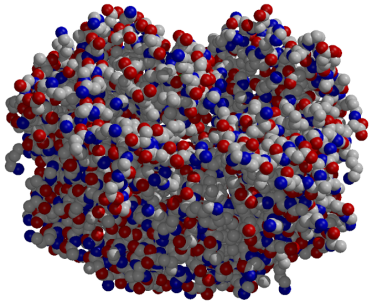
~25% des gènes d'*Escherichia coli* acquis par transfert horizontal

Ce processus ne respecte pas les arbres phylogénétiques...

2) Les protéines aux séquences similaires ont souvent des structures et des fonctions similaires

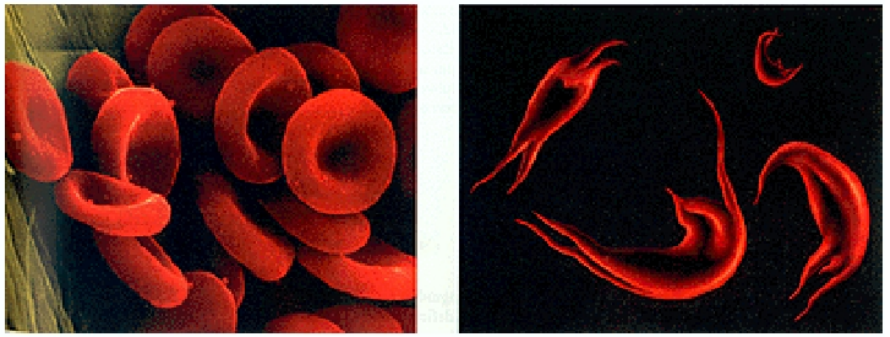
K
L
H
G
G
P
M
L
D
S
D
Q
K
F
W
R
T
P
A
A
L

Séquence → **Structure**

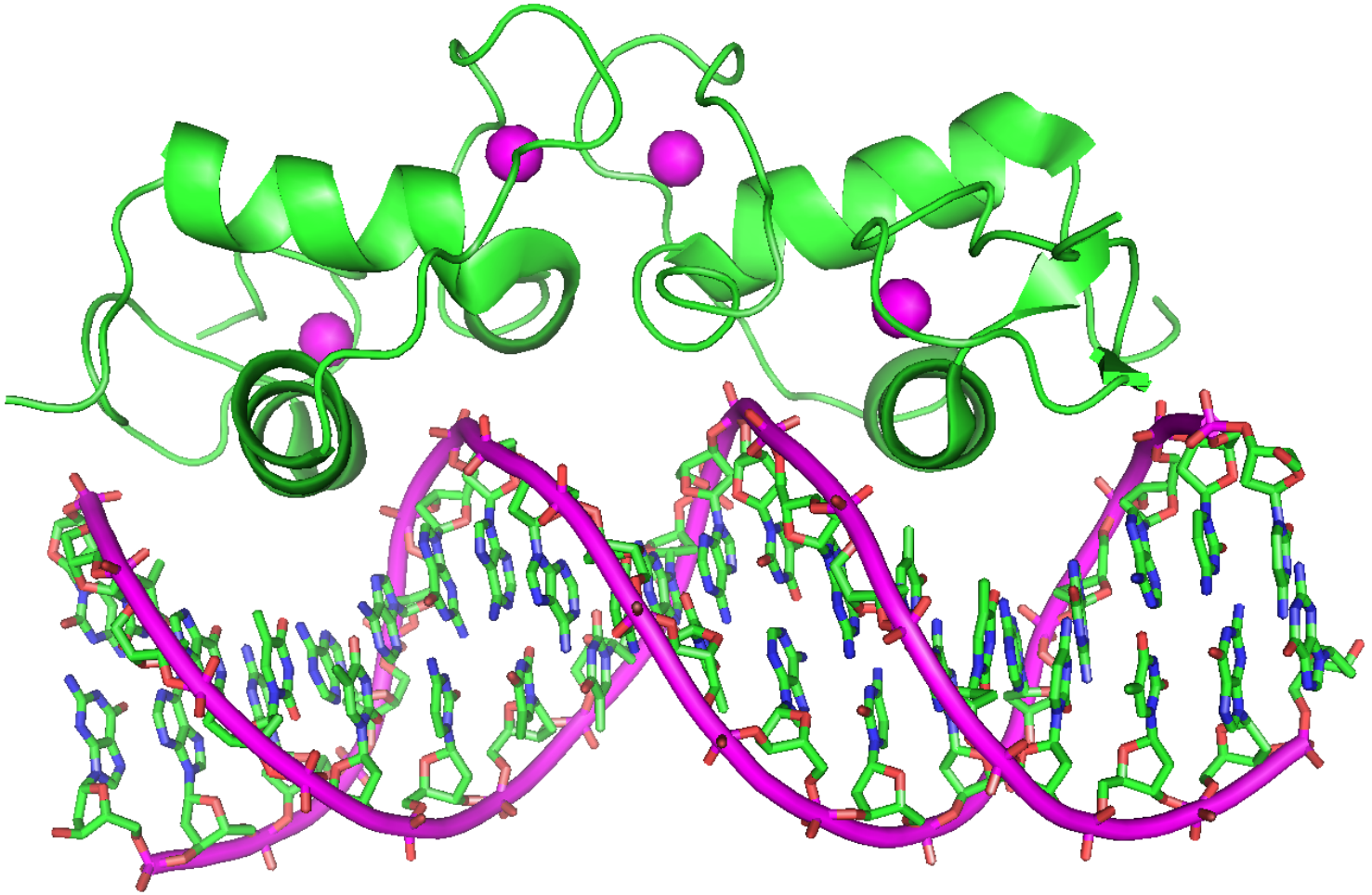


↓

Fonction

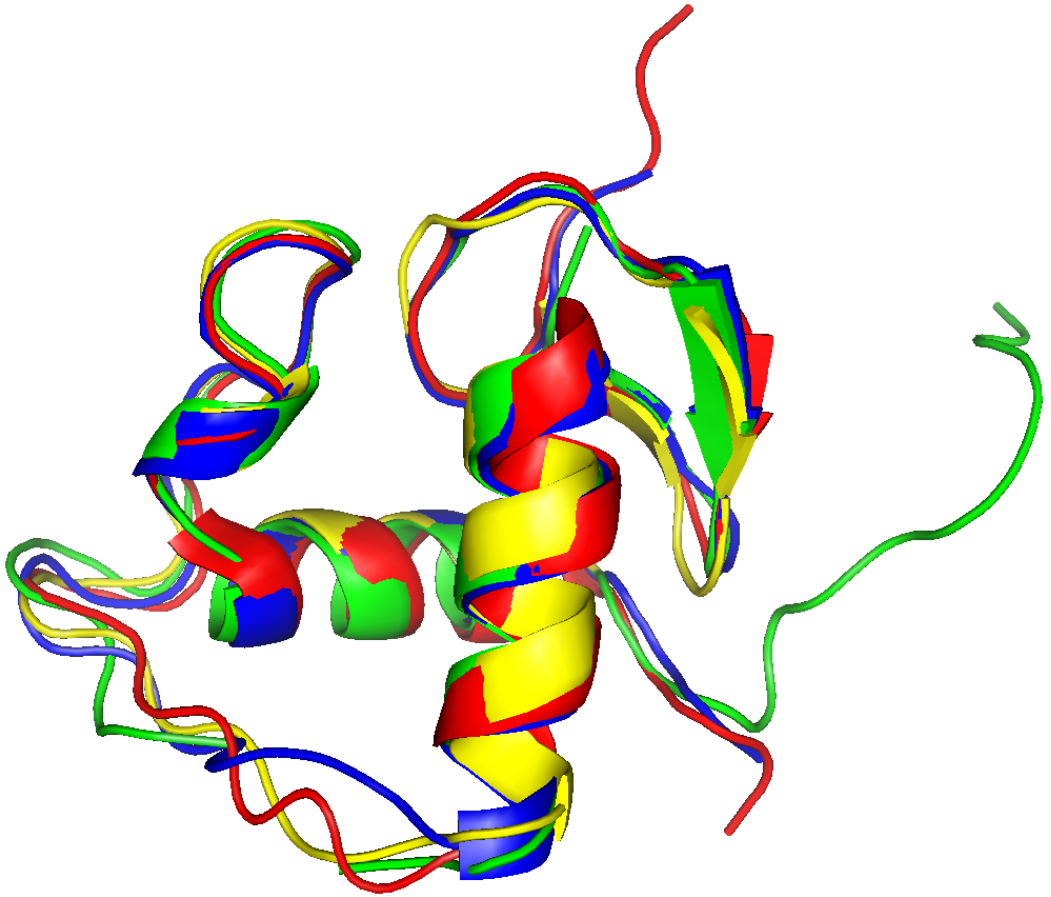


Récepteur de l'androstérone



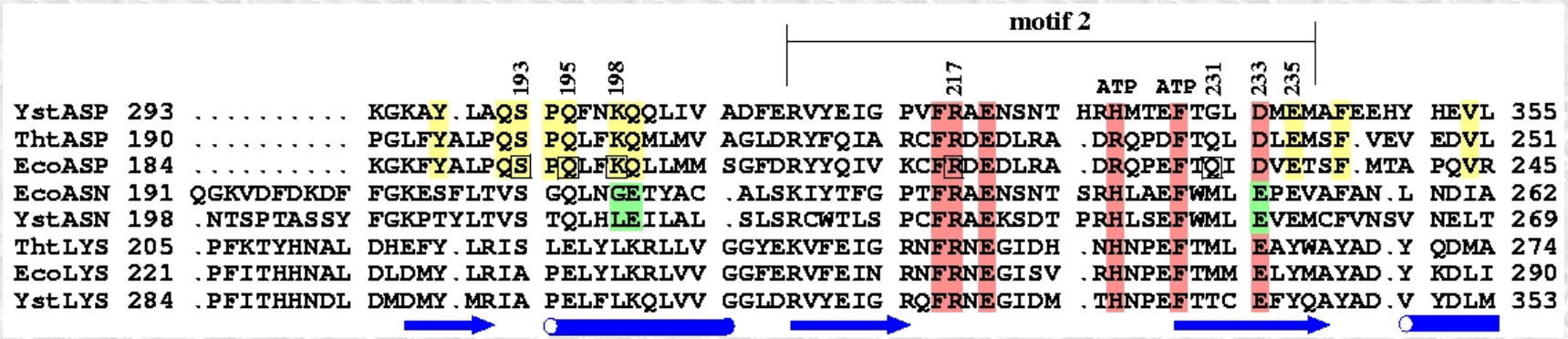
Récepteurs nucléaires

Se fixent sur l'ADN et régulent l'expression génétique sous le contrôle de petits ligands: stéroïdes, vitamines, hormones, ...



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

L'alignement de séquences comme modèle d'un processus évolutif



Récepteur de l'androstérone

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
 *  *' *         ***'   *  *  *  *         *  * * '   * '  *'*
```

L'alignement comme modèle d'un processus évolutif

Hypothèse: deux protéines similaires ont toujours un ancêtre commun

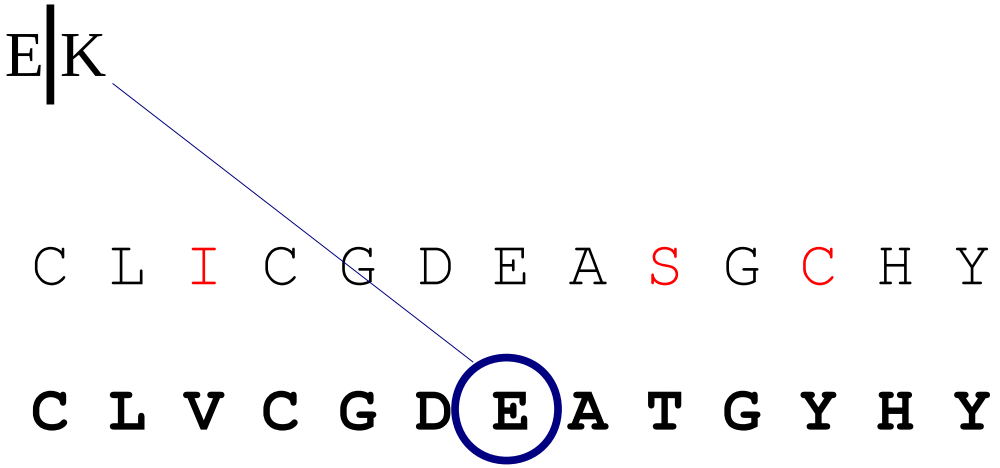
L'alignement comme hypothèse d'un ancêtre commun “parsimonieux”

T C L I C G D E A S G C H Y	Récepteur de l'androstérone
T C L V C G D E A T G Y H Y	**Ancêtre commun hypothétique**
L C V V C G D K A T G Y H Y	Récepteur de l'hormone thyroïdienne

Mutations hypothétiques en rouge

L'hypothèse ci-dessus est-elle vraisemblable?

L'alignement comme hypothèse d'un ancêtre commun “parsimonieux”



Récepteur de l'androstérone

Ancêtre commun hypothétique

Récepteur de l'hormone thyroïdienne

Mutations hypothétiques en rouge

L'hypothèse ci-dessus est-elle vraisemblable?

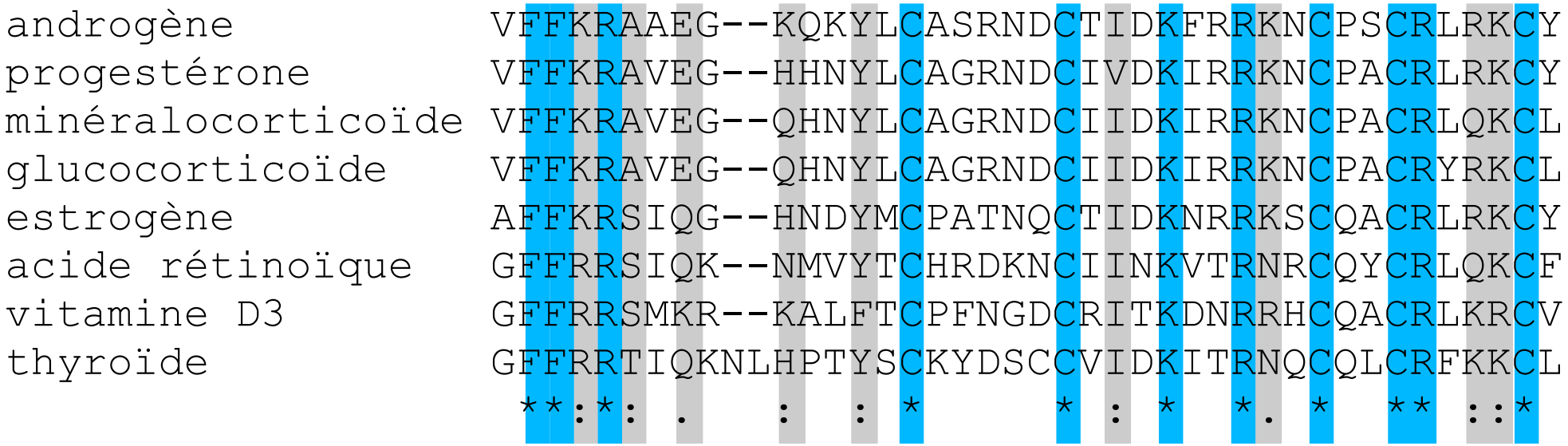
L'alignement comme hypothèse d'un ancêtre commun “parsimonieux”

T C L I C G D E A S G C H Y	Récepteur de l'androstérone
T C L V C G D E A T G Y H Y	**Ancêtre commun hypothétique**
L C V V C G D K A T G Y H Y	Récepteur de l'hormone thyroïdienne

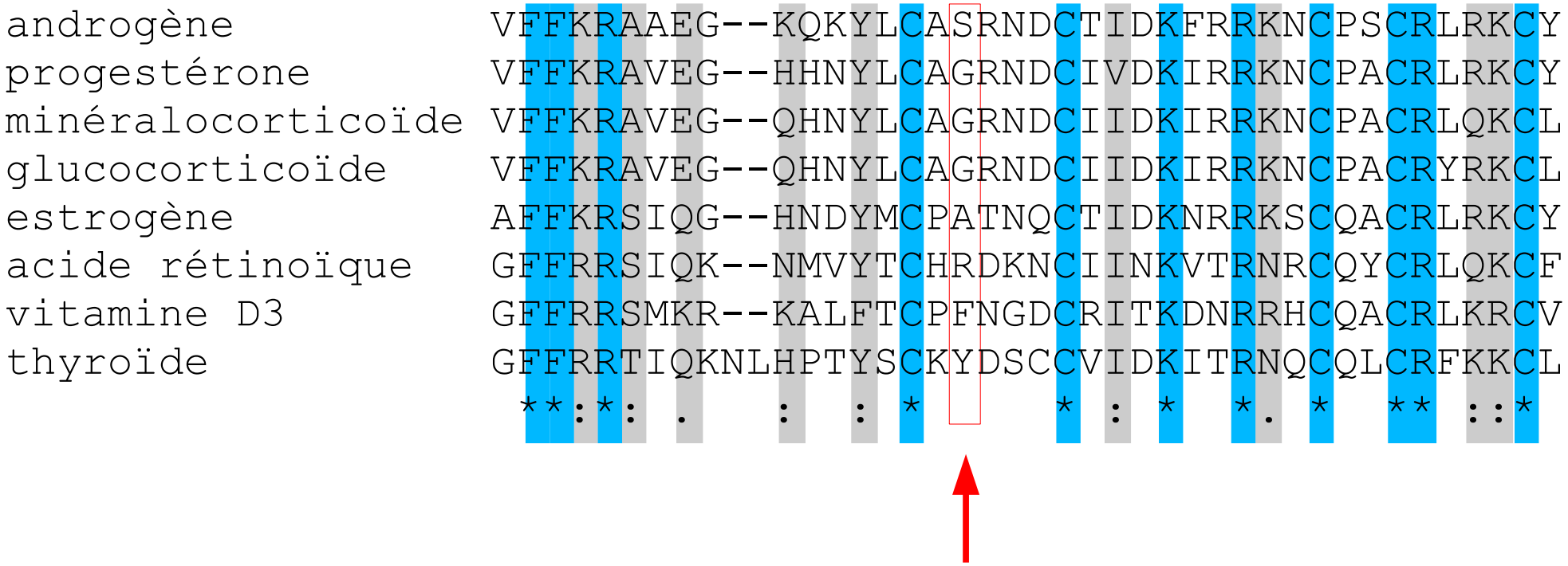
Mutations hypothétiques en rouge

L'hypothèse ci-dessus est-elle vraisemblable?

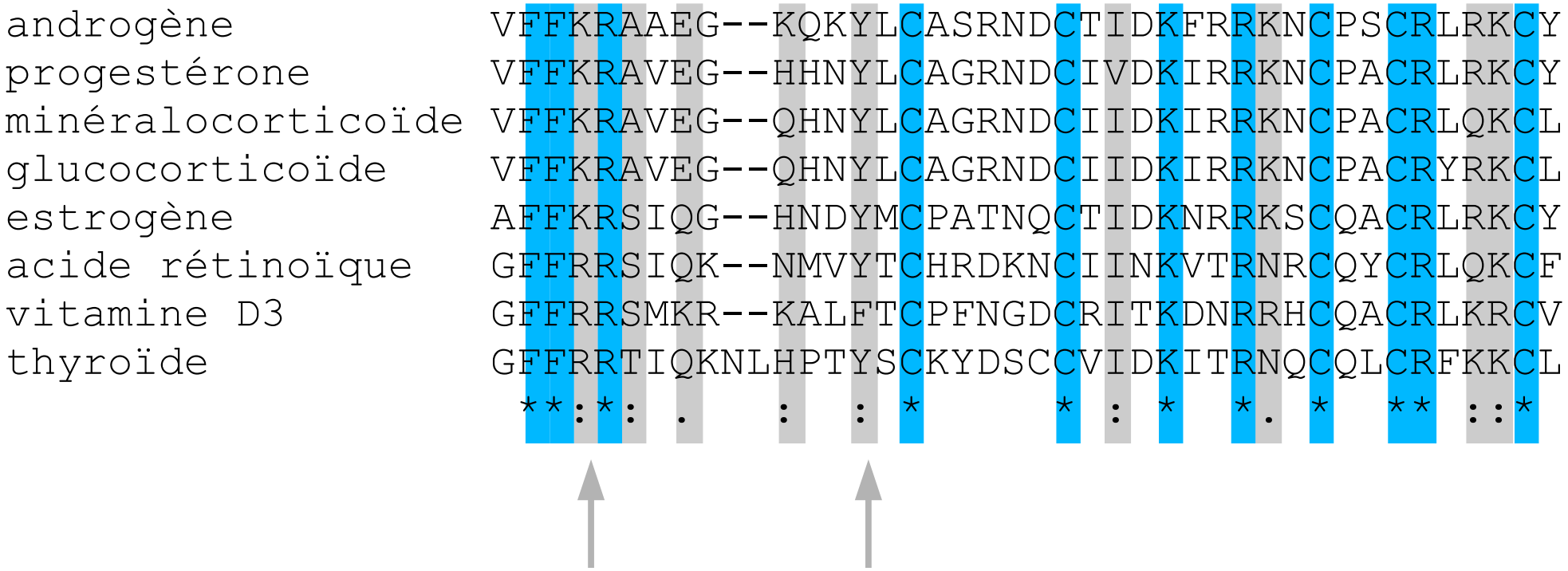
Les mutations ont des probabilités différentes



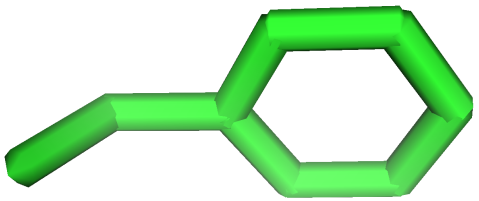
Les mutations ont des probabilités différentes



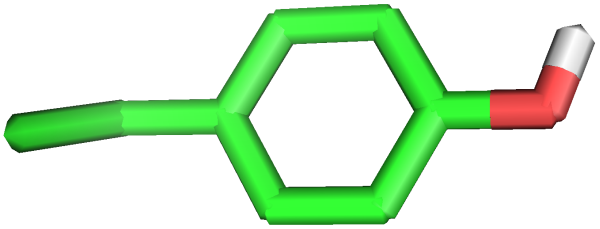
Les mutations ont des probabilités différentes



Les mutations ont des probabilités différentes



PHE
F

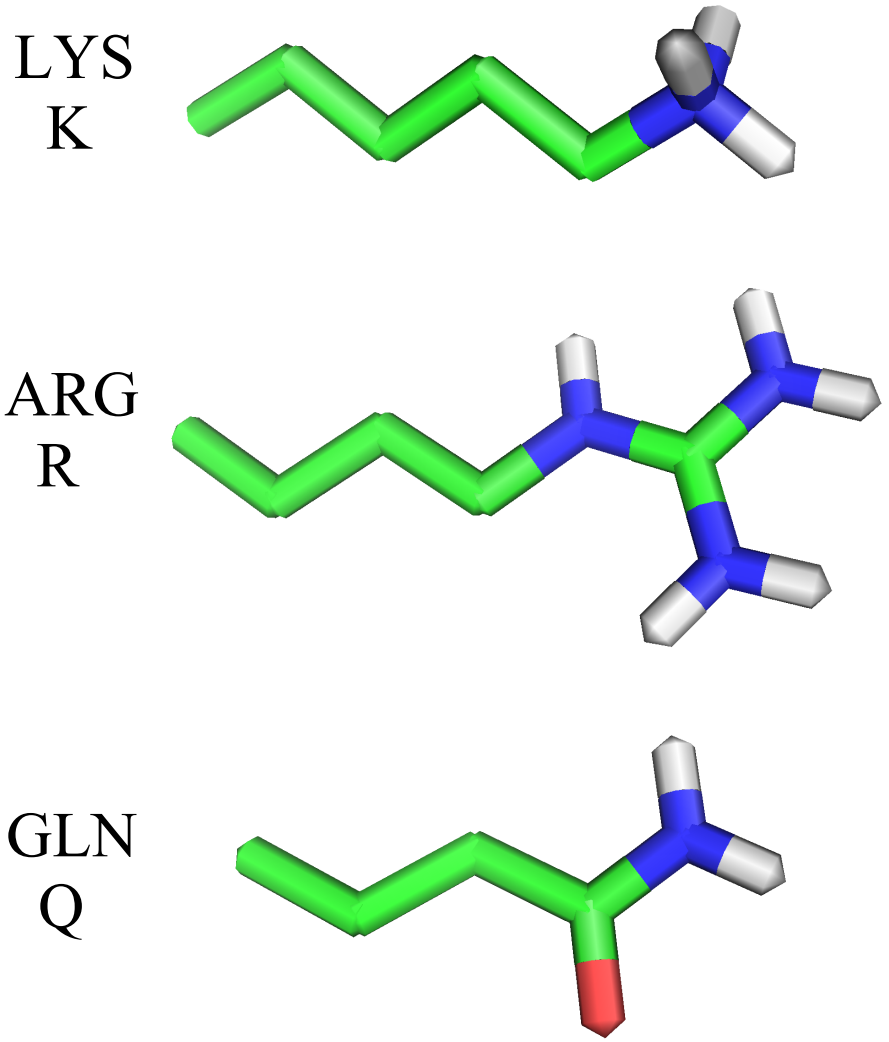


TYR
Y

Similaires
ou
“homologues”

KYL
NYL
NYL
NYL
DYM
VYT
LFT
TYS
:

Les mutations ont des probabilités différentes



homologues

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

Un modèle probabiliste d'évolution

$P(x_i,y_j)$ = probabilité d'observer x_i aligné avec y_j

= P("x_i muté en y_j" ou "y_j muté en x_i")

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

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

- Probabilités estimées à partir d'alignements tests
- Hypothèse de positions équivalentes et indépendantes

Récepteur de l'androstérone

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
 *  *' *         ***'   *  *  *  *         *  * * '   * '  *'*
```

Notre explication (hypothèse) est plus ou moins probable: oui, mais...

Ce n'est pas parce qu'un modèle reproduit les données expérimentales qu'il est juste

Ce n'est pas parce qu'un modèle reproduit les données expérimentales qu'il est juste

Il faut aussi exclure les autres modèles ou explications a priori plausibles.

Modèle “aléatoire” comme hypothèse concurrente

$Q(x_i,y_j)$ = probabilité d'observer x_i et y_j
dans deux protéines indépendantes

$\approx$ $q_{xi}\, q_{yj}$ fréquences naturelles de x_i, y_j

	x	q_x	
tryptophane	W	1.3 %	des acides aminés
leucine	L	9.0 %	

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

Vraisemblance d'une colonne dans un alignement

déf.

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

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

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

s = matrice de “score”

Une matrice empirique s très utilisée

	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	

BLOSUM62

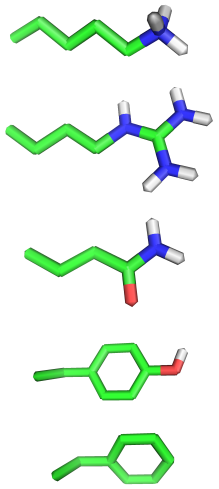
Une matrice empirique s très utilisée

	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	

BLOSUM62

Pénalité pour gaps: voir plus loin

Les mutations ont des vraisemblances différentes



	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**

Extrait de la matrice empirique BLOSUM50

La vraisemblance ou probabilité d'un alignement

$P(x_i,y_j)$ = probabilité d'observer x_i aligné avec y_j

= P("x_i muté en y_j" ou "y_j muté en 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

- Probabilités estimées à partir d'alignements tests
- Hypothèse de positions équivalentes
- Hypothèse de positions indépendantes: $P(\text{alignement}) = \Pi_{(i,j)} P(x_i,y_j)$

Vraisemblance ou “score” d'un alignement

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

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

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

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

La probabilité de mutation dépend du temps

P(x,y) = P(x,y|T) = probabilité d'une mutation pendant le temps T

Les cytochromes c de chimpanzee et d'humain sont plus similaires que les cytochromes c d'humain et d'Escherichia coli...

Si on compare des protéines de mammifère, par exemple, les vrais homologues seront très similaires: scores élevés. Si on compare mammifère et bactérie, de vrais homologues seront nettement moins similaires: scores plus faibles si on utilise la même matrice

La probabilité de mutation dépend du temps

P(x,y) = P(x,y|T) = probabilité d'une mutation pendant le temps T

Les cytochromes c de chimpanzee et d'humain sont plus similaires que les cytochromes c d'humain et d'Escherichia coli...

Une matrice de similarité n'est optimale que pour une échelle de temps, ou de similarité donnée.

BLOSUM40 ≠ BLOSUM50 ≠ BLOSUM62 ≠ BLOSUM80

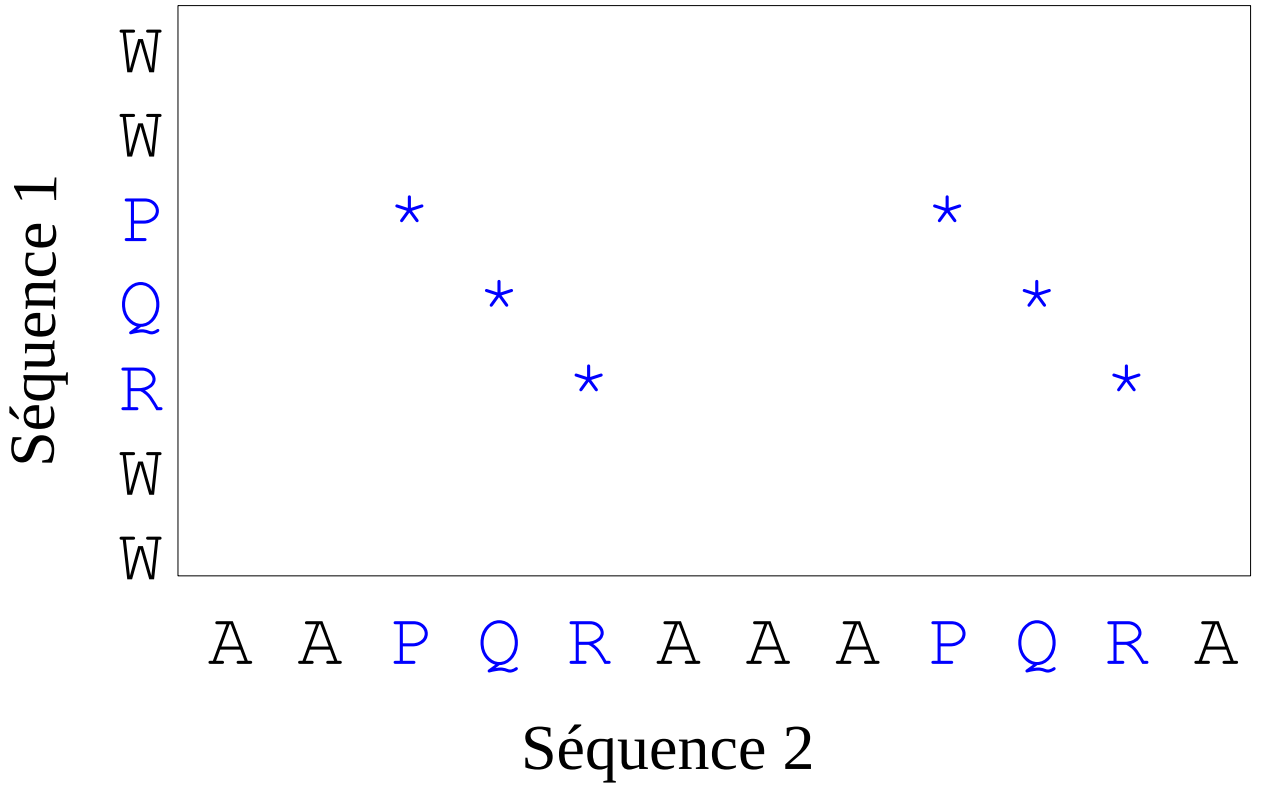
En résumé, un alignement correspond à un modèle d'évolution, qui comprend:

- Hypothèse de divergence “minimale” depuis un ancêtre commun
- Détermination empirique des probabilités de mutations
- Hypothèse de positions équivalentes et indépendantes
- Un modèle de référence (hypothèse “nulle”)

L'alignement de séquences: algorithmes

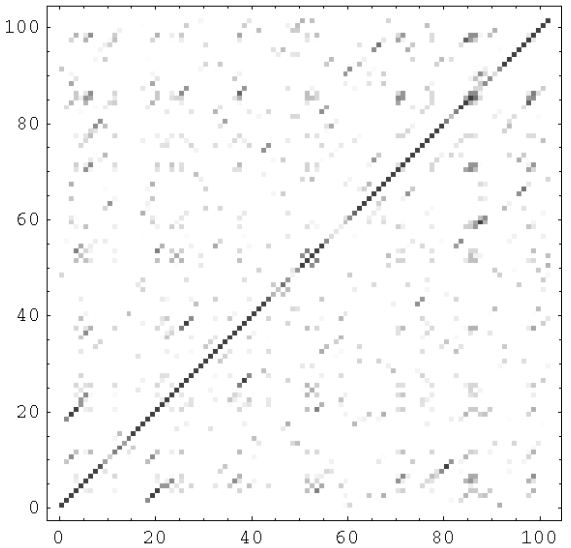
Comparaison graphique de deux séquences par une matrice densité

On dispose les deux séquences sur les bords d'une matrice. Pour chaque paire de résidus similaires ou identiques, on dessine un point dans la matrice:



Comparaison graphique de deux séquences par une matrice densité

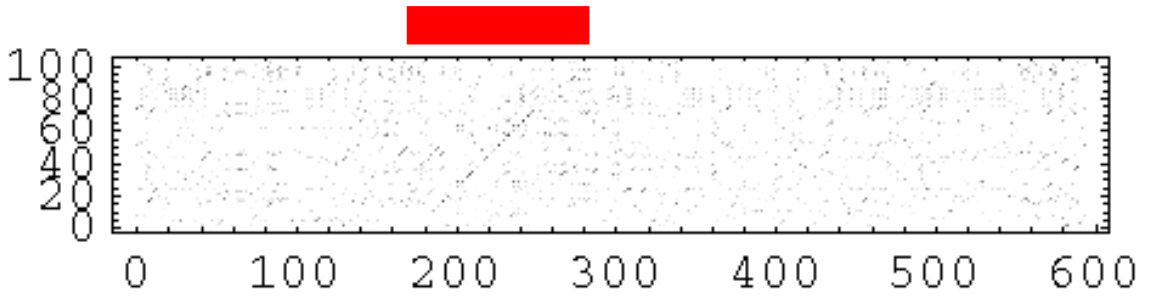
cheval



Cytochromes c

humain

Récepteur androgène



Récepteur estrogène

Comment trouver l'alignement le plus vraisemblable?

Pour deux séquences de longueur n:

$$N \approx \binom{2n}{n} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$ alignements possibles!

N	1	2	3	4	5	6	7	8	9	....	100
n	2	6	20	70	252	924	3.432	17.160	48.620	.....	10^{59}

Recherche exhaustive impossible!

Existence d'un algorithme récursif

Un alignement donné, A, comprend de nombreux alignements plus petits, A_1, A_2, A_3, ...

A

```
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
```

A_1

```
T C L I C G D
L C V V - G D
```

A_2

```
E A S G C H Y
K A T G Y H Y
```

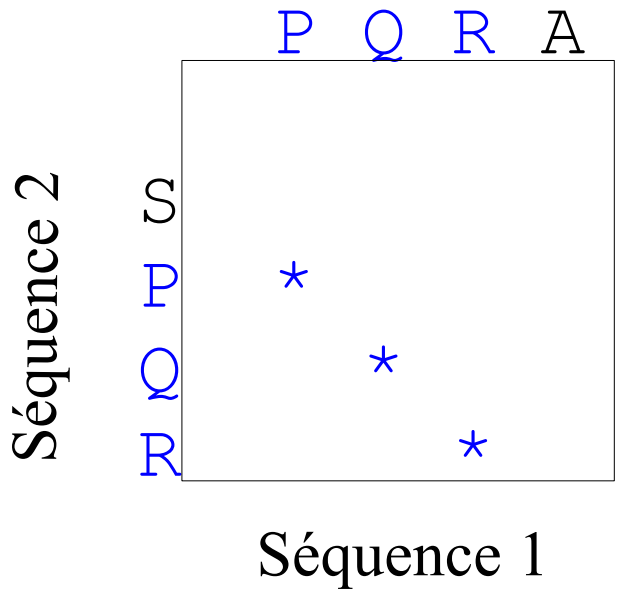
A_3

```
I C G D E A
V - G D K A
```

➡ **Pour évaluer un alignement A, on peut utiliser les résultats déja acquis pour des sous-alignements (scores additifs): algorithme récursif**

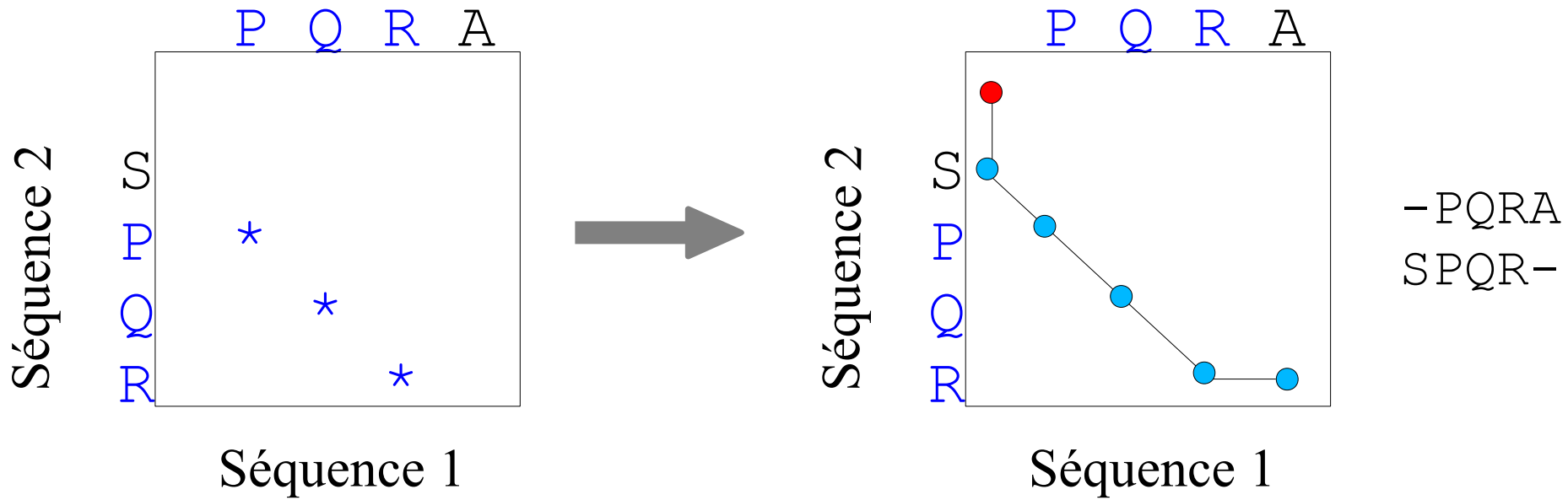
Une méthode qui s'inspire de la “matrice” déja vue

On dispose les deux séquences sur les bords d'une matrice:



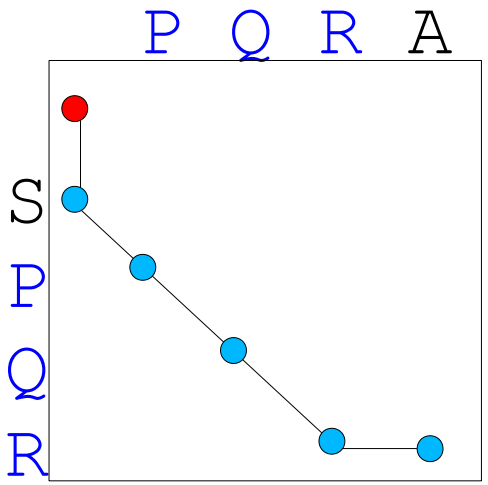
Une méthode qui s'inspire de la “matrice” déja vue

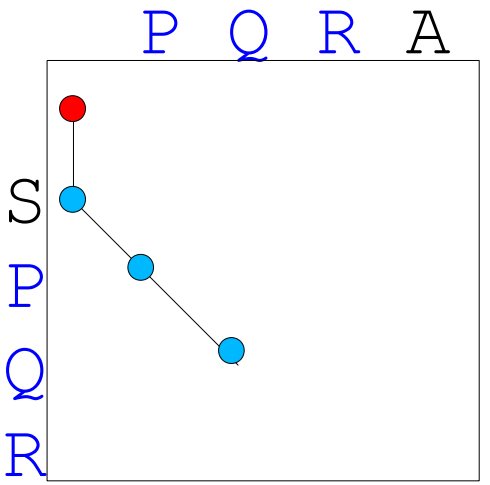
On dispose les deux séquences sur les bords d'une matrice.
Un alignement correspond à une ligne à travers la matrice:



Une méthode qui s'inspire de la “matrice” déja vue

On dispose les deux séquences sur les bords d'une matrice.
Un alignement correspond à une ligne à travers la matrice:



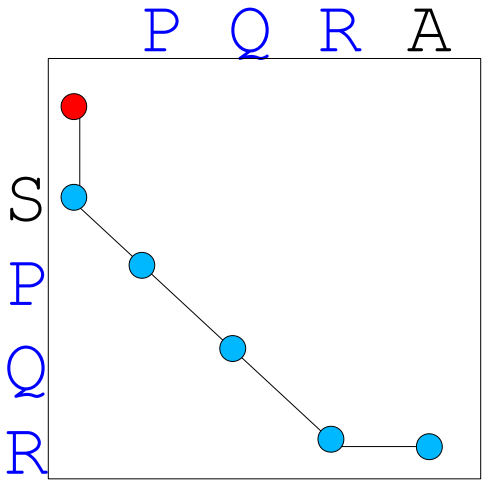


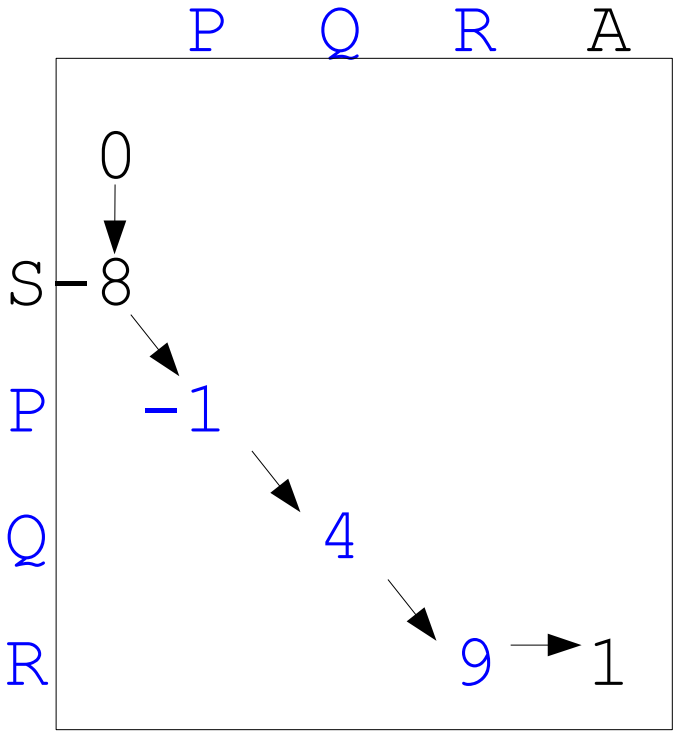
-PQRA
SPQR-

-PQ
SPQ

Une méthode qui s'inspire de la “matrice” déja vue

Un alignement correspond à une ligne à travers la matrice.
On annote la table avec les scores:



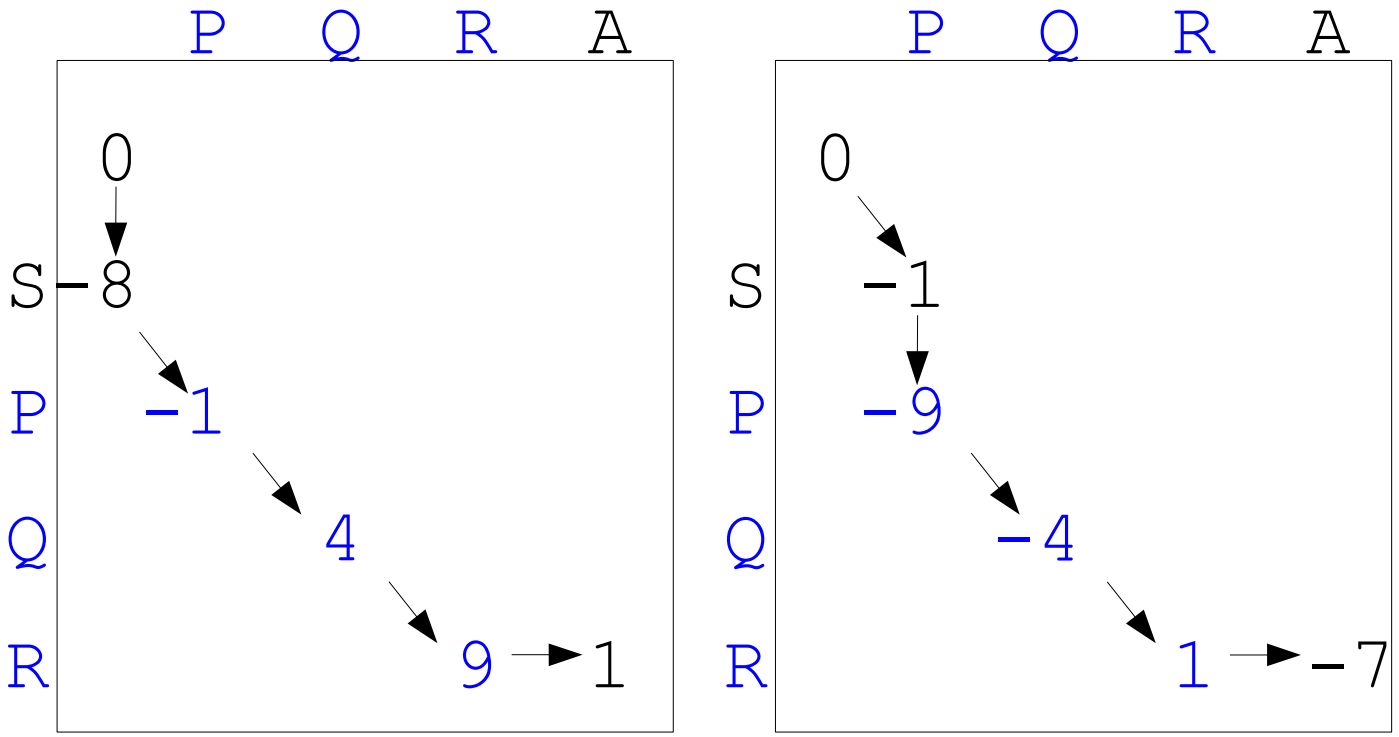


	S	P	Q	R
P	-1	7	-1	-2
Q	0	-1	5	1
R	-1	-2	1	5
A	1	-1	-1	-1

```
-PQRA
SPQR-
```

Une méthode qui s'inspire de la “matrice” déja vue

L'alignement optimal est en compétition avec beaucoup d'autres alignements possibles:

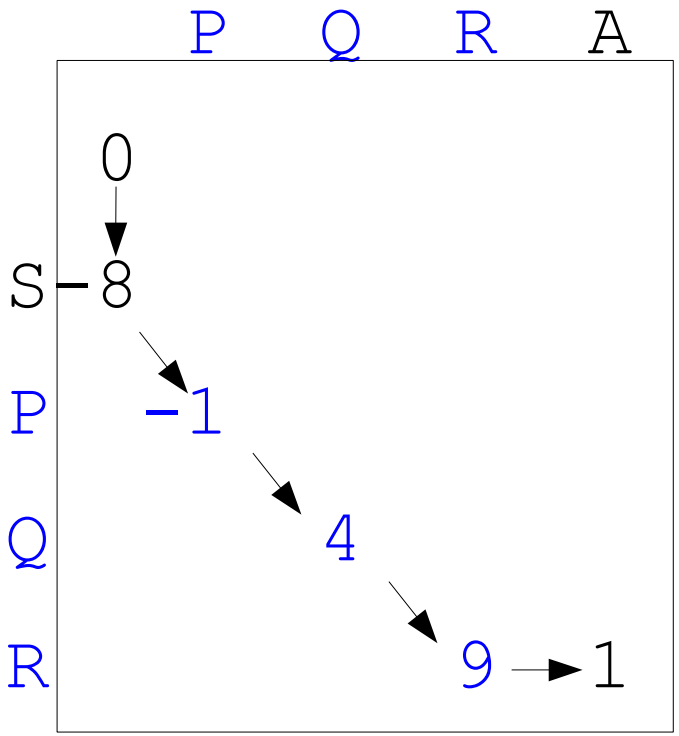


-PQRA
SPQR-

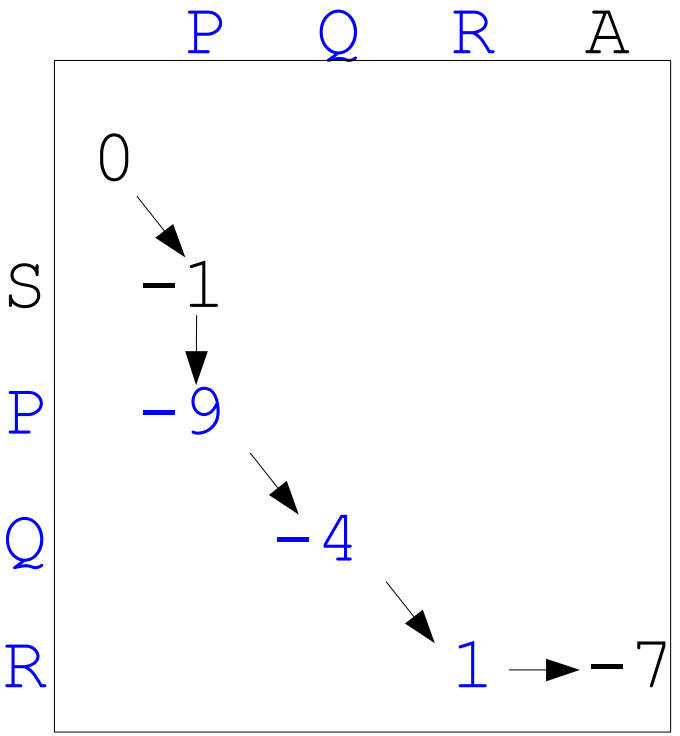
P-QRA
SPQR-

Une méthode qui s'inspire de la “matrice” déja vue

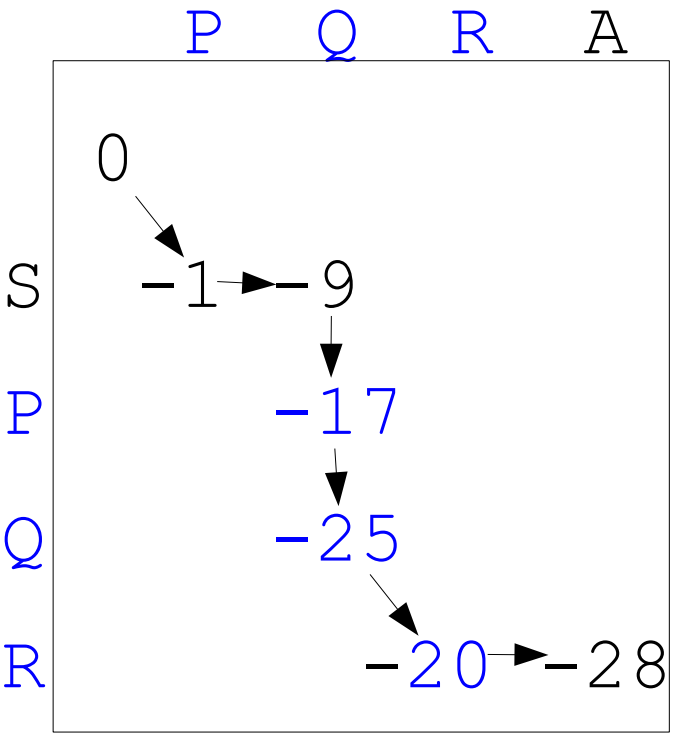
L'alignement optimal est en compétition avec beaucoup d'autres alignements possibles:



-PQRA
SPQR-



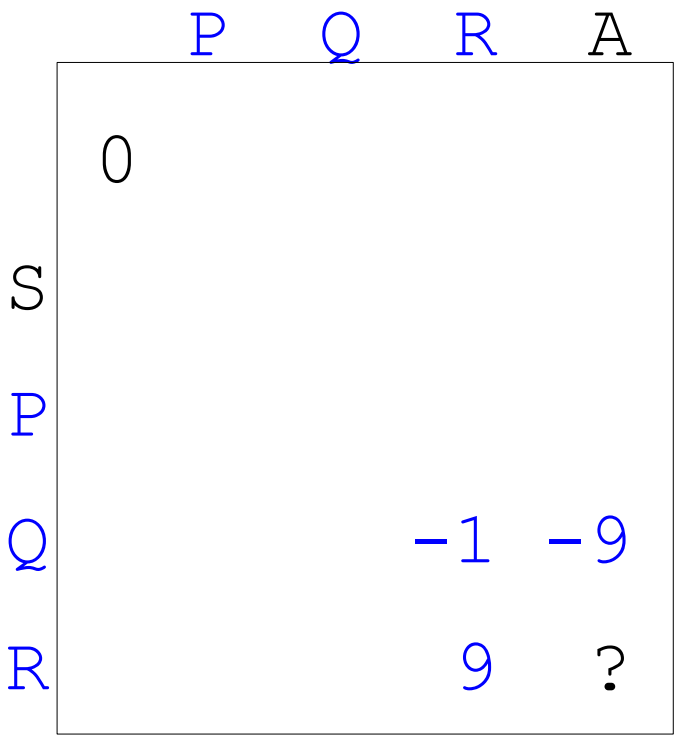
P-QRA
SPQR-



PQ--RA
S-PQR-

Comment trouver le meilleur chemin à travers la table?

Supposons que nous connaissons les scores de trois alignements particuliers:



Chaque score bleu correspond à un alignement incomplet, inconnu à ce stade.

A partir de ces 3 scores, je peux trouver le score manquant

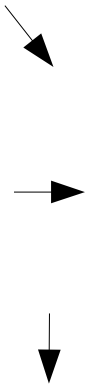
Comment trouver le meilleur chemin à travers la table?

Supposons que nous connaissons les scores de trois alignements particuliers:

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

A partir des 3 scores bleus, je peux trouver le score manquant

3 possibilités:



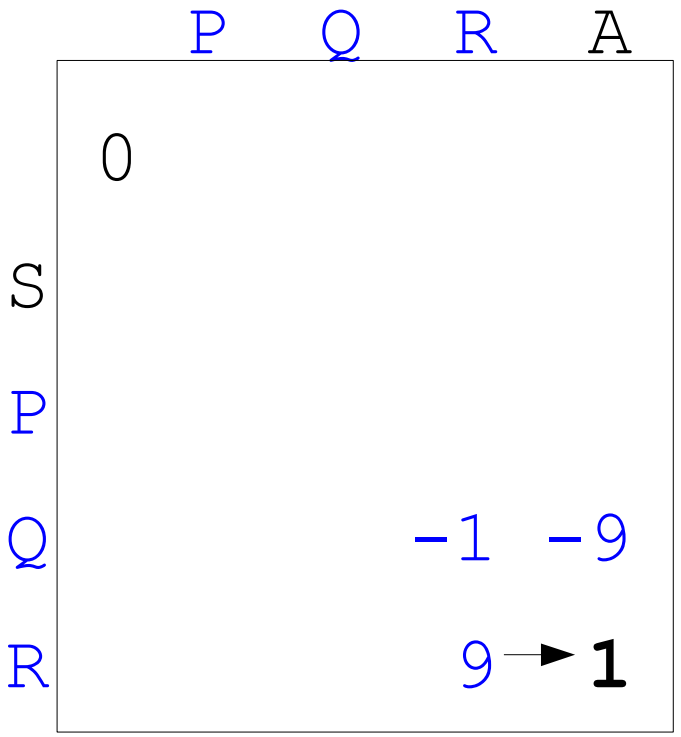
score -1 -1 = -2

score 9 - 8 = 1

score -9 - 8 = -17

Comment trouver le meilleur chemin à travers la table?

Supposons que nous connaissons les scores de trois alignements particuliers:



A partir des 3 scores bleus, je peux trouver le score manquant

3 possibilités:

score -1 -1 = -2

score 9 - 8 = 1

score -9 - 8 = -17

Si les chemins aboutissant aux cases bleues sont optimaux, le chemin final aussi

Algorithme de "programmation dynamique" de Needleman-Wunsch

Pour aligner deux séquences:

$x_1\ x_2\ x_3\ x_4\ \ldots\ x_p$

$y_1\ y_2\ y_3\ y_4\ y_5\ \ldots\ y_q$

On construira les **alignements optimaux** de toutes les tailles:

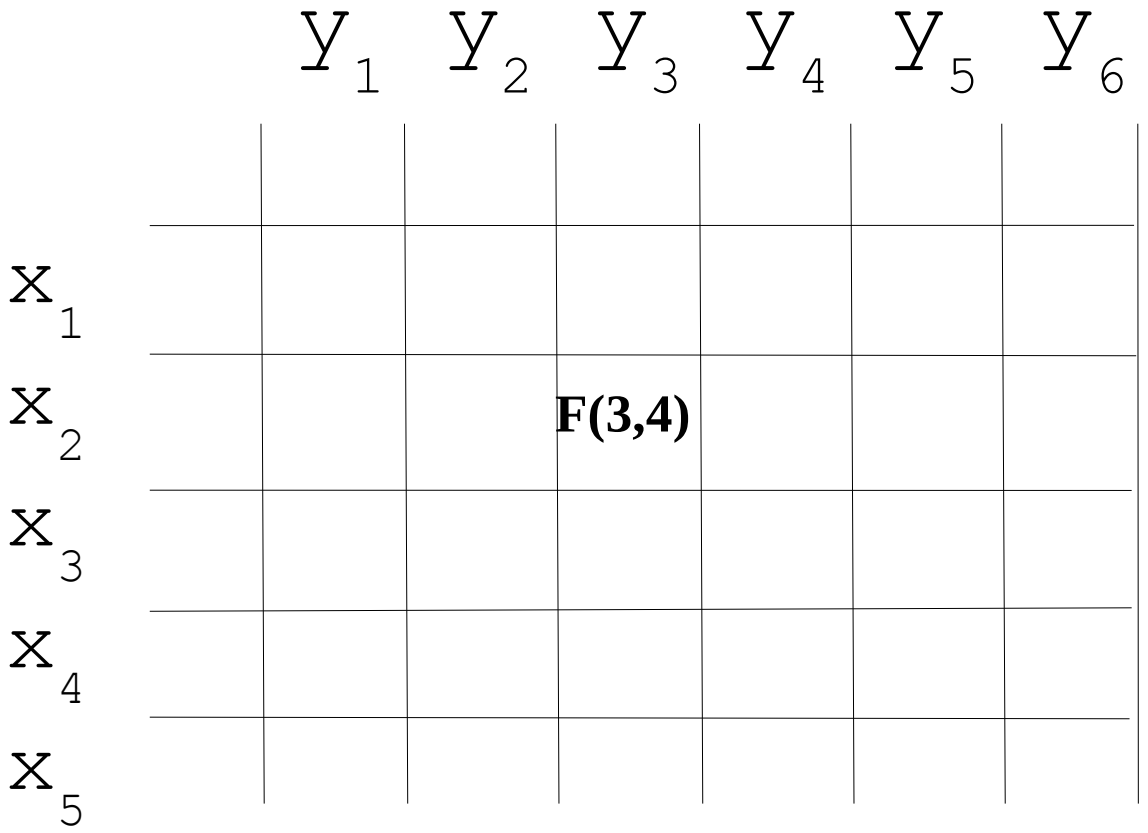
alignement optimal de $x_1, \ldots, x_i$ avec $y_1, \ldots, y_j$

pour chaque $i \leq p,\ \ j \leq q$

Matrice F de scores d'alignements partiels

Soit: $F(i,j) \overset{\text{déf.}}{=}$ score du meilleur alignement de $x_1, \ldots, x_i$ avec $y_1, \ldots, y_j$

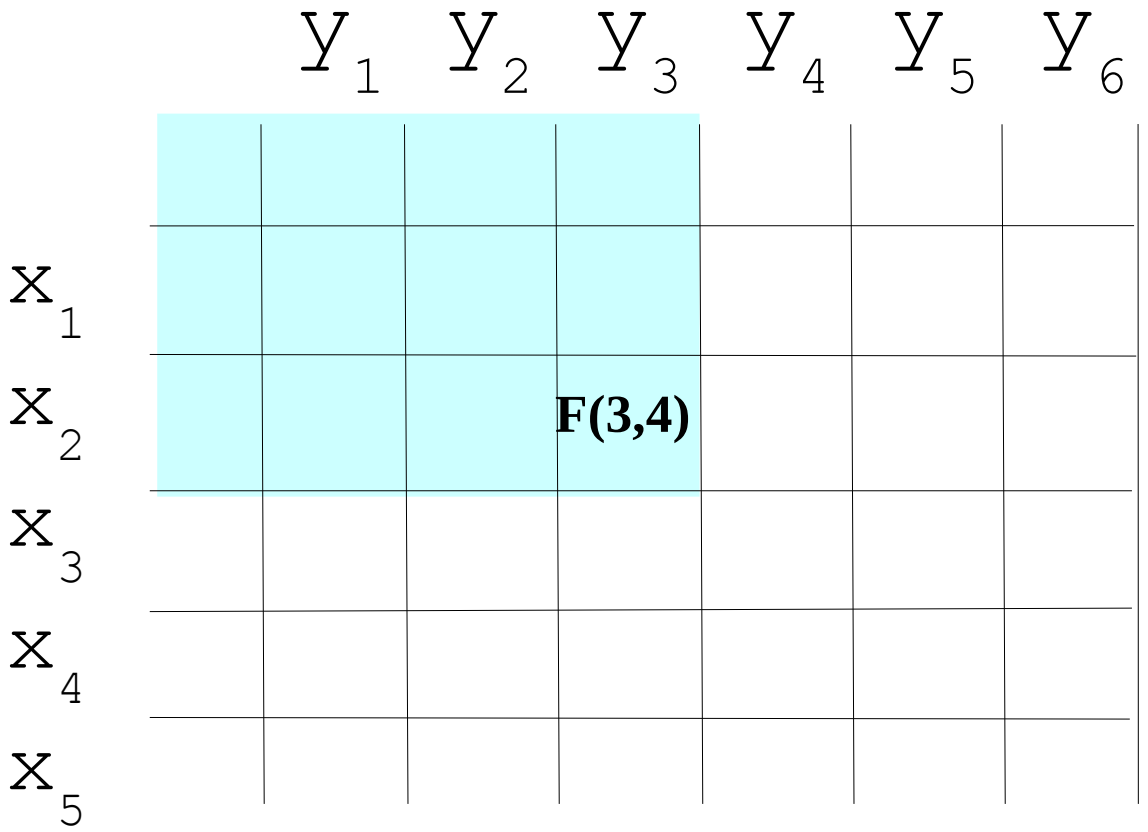
$A(i,j) \overset{\text{déf.}}{=}$ l'alignement correspondant



Matrice F(i,j) de scores optimaux

Algorithme de programmation dynamique de Needleman-Wunsch

On va déduire F(i,j) des scores d'alignements (optimaux) plus petits



Calcul récursif de F(i,j)

Un alignement de

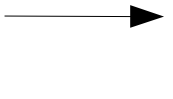
$x_1\ x_2\ \ldots\ x_i$

avec

$y_1\ y_2\ \ldots\ y_j$

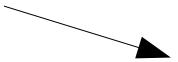
peut être obtenu en ajoutant x_i ou y_j ou les deux au bout d'un alignement plus petit.

I G A x_i
L G V y_j



A I G x_i
G V y_j –

création ou extension d'un “gap”



G A x_i – –
S L G V y_j

création ou extension d'un “gap”

Calcul récursif de F(i,j)

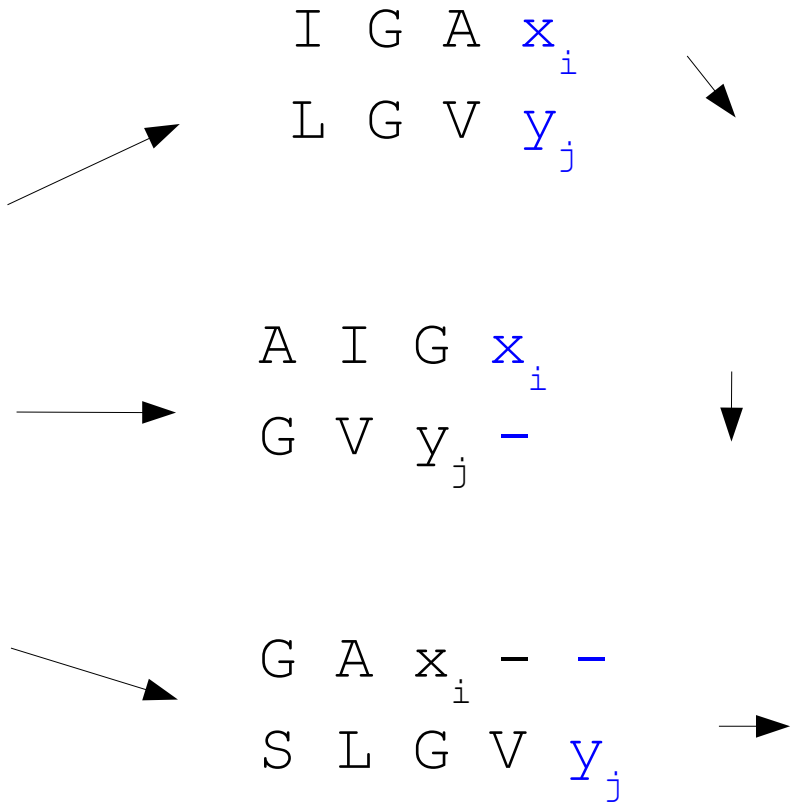
L'alignement **optimal** de

$x_1\ x_2\ \ldots\ x_i$

avec

$y_1\ y_2\ \ldots\ y_j$

peut être obtenu en ajoutant x_i ou y_j ou les deux au bout d'un alignement **optimal** plus petit.



Calcul récursif de F(i,j)

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i,y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

I G A x_i
L G V y_j

A I G x_i
G V y_j –

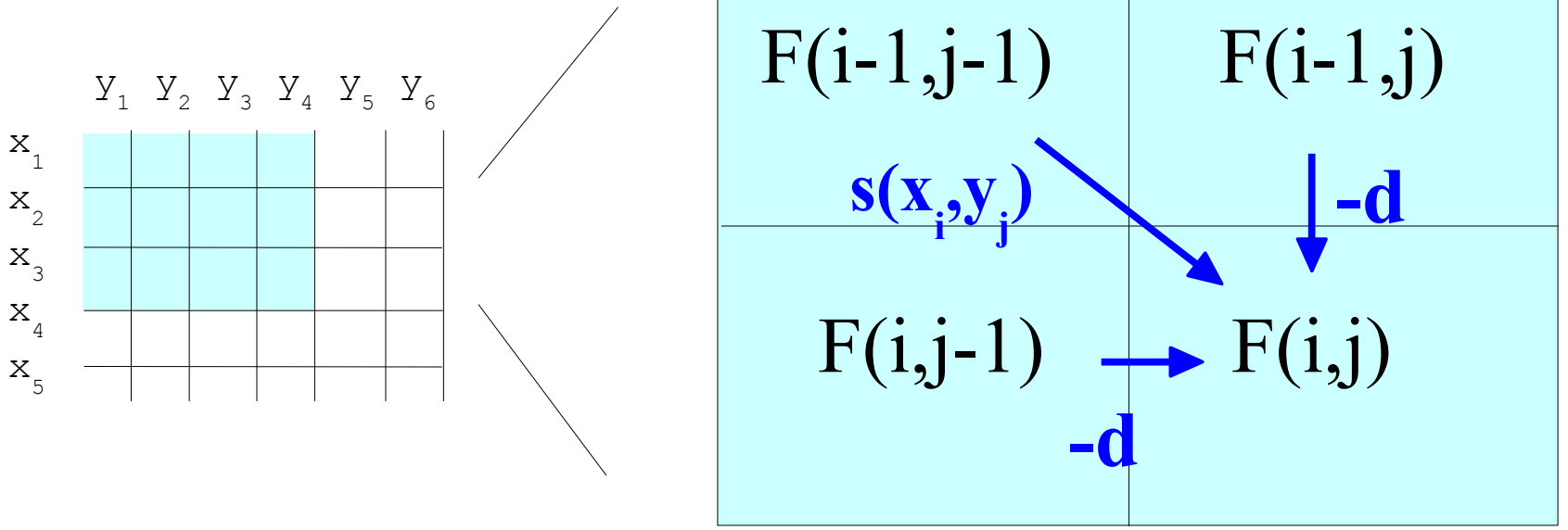
création ou extension d'un “gap”

G A x_i – –
S L G V y_j

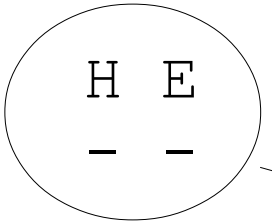
création ou extension d'un “gap”

s est la matrice de score employée (eg, BLOSUM62)

Calcul récursif de F(i,j)



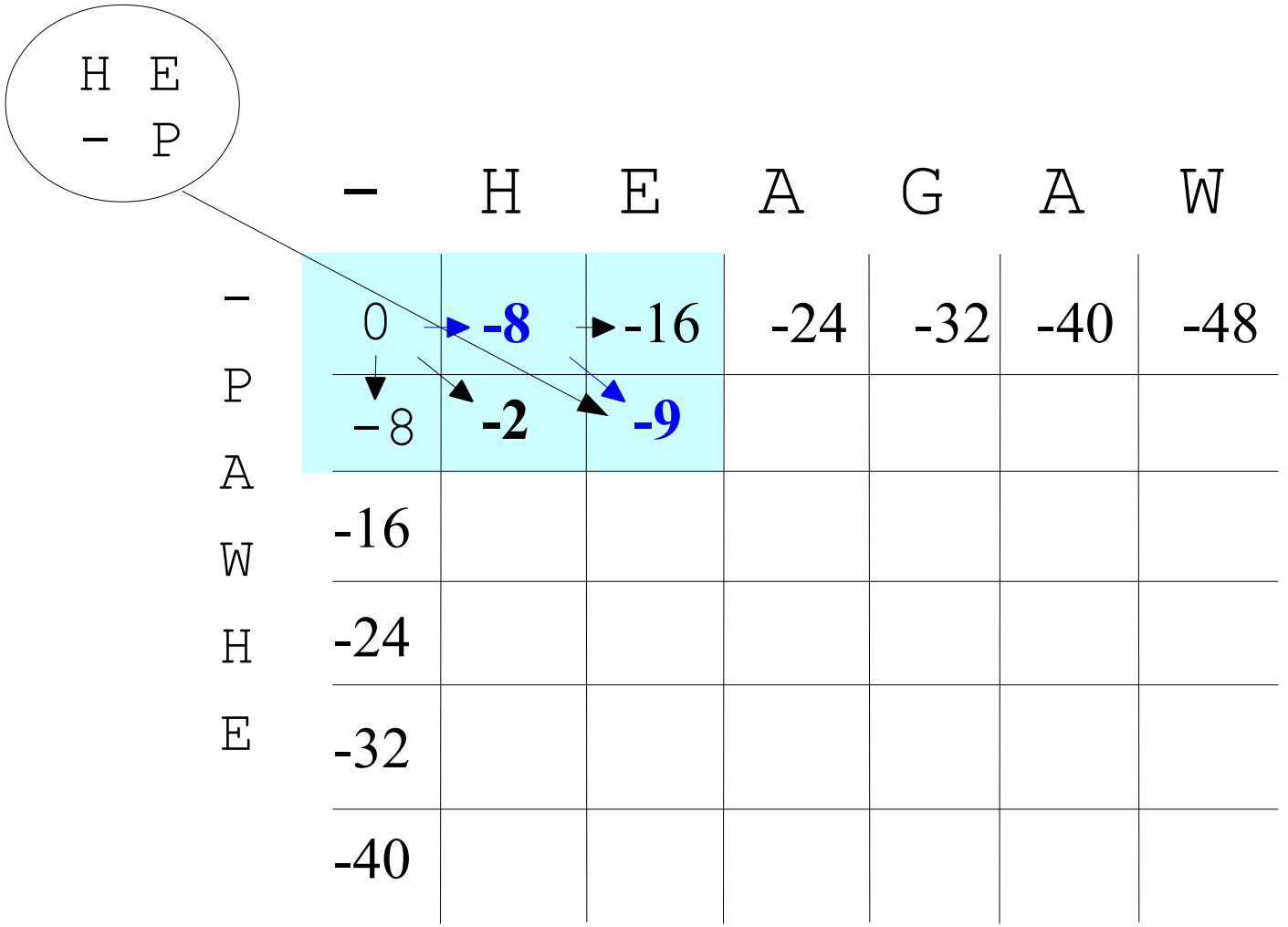
Initialisation de F



	–	H	E	A	G	A	W
–	0	-d	-2d	-3d	-4d	-5d	-6d
P	-d						
A	-2d						
W	-3d						
H	-4d						
E	-5d						

$s(x_i,-) = s(-,y_j) = -d$ = coût d'un gap

Extension de F

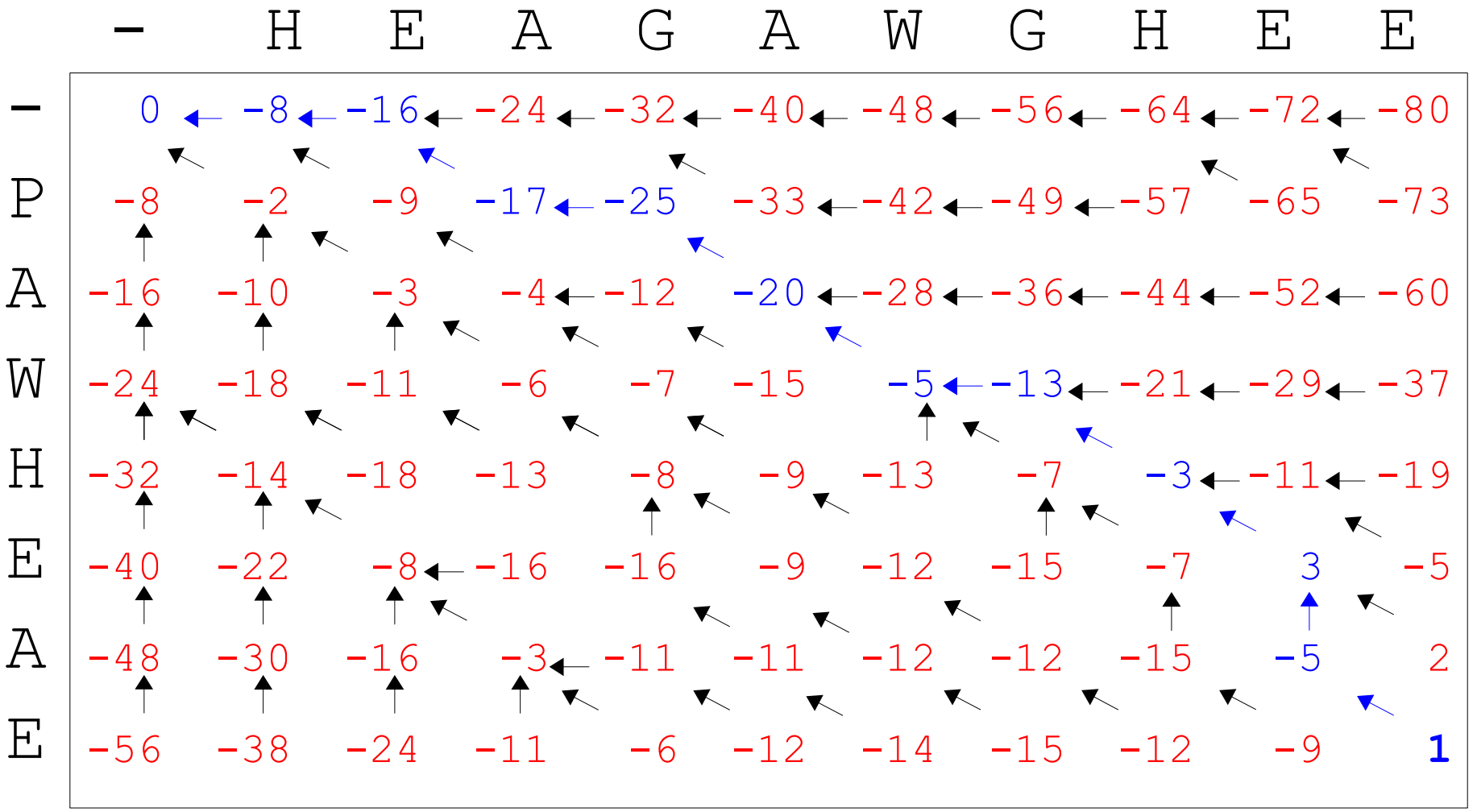


Matrice de similarité

	H	E	A
P	-2	-1	-1
A	-2	-1	5

Coût μ pq en mémoire et calcul

Identification de l'alignement optimal par “tracé inverse”



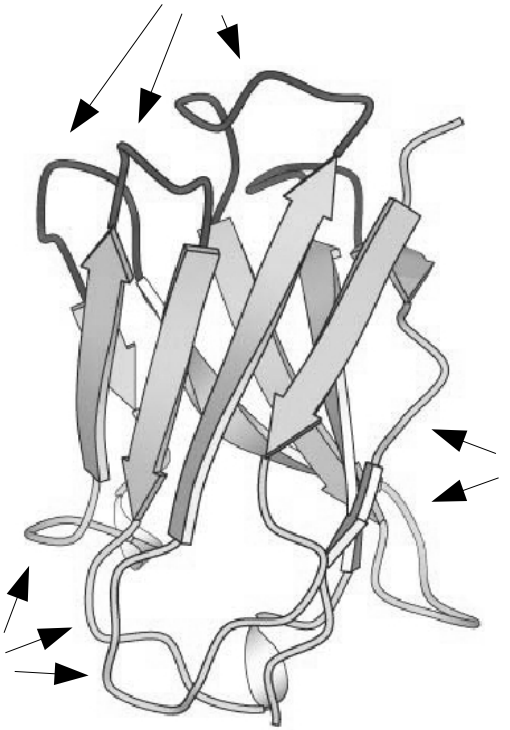
Alignement optimal:

HEAGAWGHE-E

--P-AW-HEAE

Il est plus facile d'allonger un gap existant que d'ouvrir un nouveau gap

Eg, boucles



- Ouvrir un gap: -d
- L'allonger d'un résidu: -e (e<d)

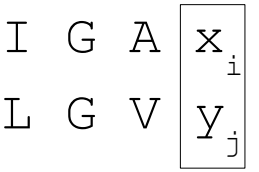
- Coût d'un gap de longueur m:

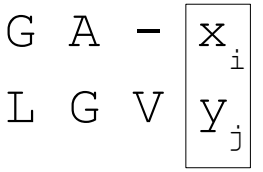
$$g(m) = -d - (m-1)\,e$$

≠ positions équivalentes et indépendantes

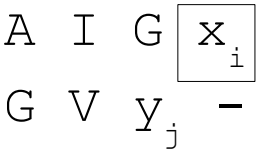
Programmation dynamique avec 3 états

$A(i,j)$ = score du meilleur alignement de $x_1, \ldots, x_i$ avec $y_1, \ldots, y_j$ qui se termine sur un **appariement**

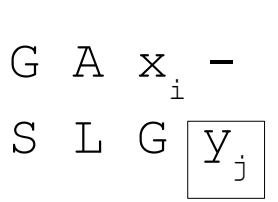




$I_x(i,j)$ = score du meilleur alignement de $x_1, \ldots, x_i$ avec $y_1, \ldots, y_j$ qui se termine sur une **insertion**



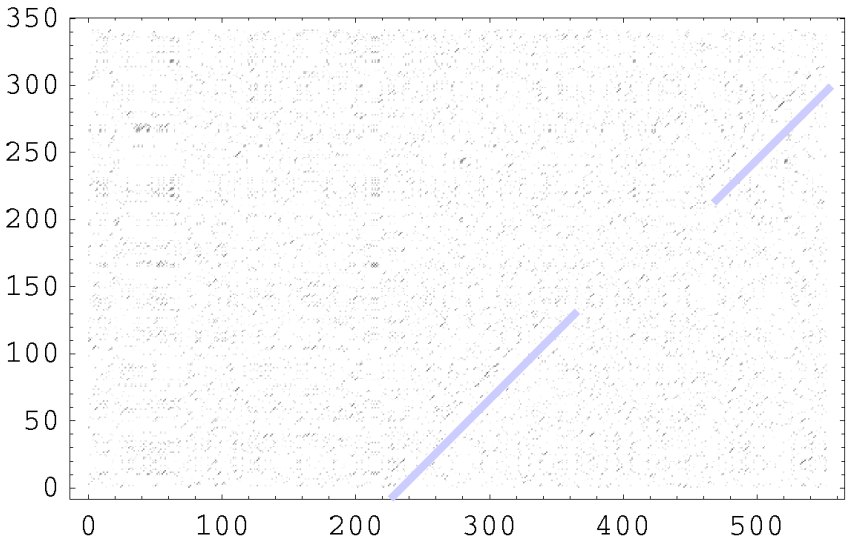
$I_y(i,j)$ = score du meilleur alignement de $x_1, \ldots, x_i$ avec $y_1, \ldots, y_j$ qui se termine sur une **insertion**

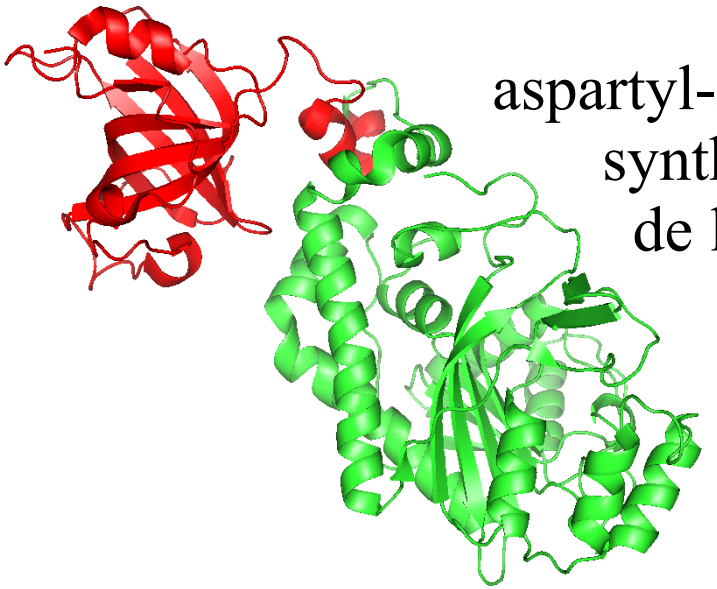


Exercice: trouver l'algorithme correspondant

C'est souvent une partie d'une protéine qui est conservée: il faut chercher un alignement “local”

Domaine du site actif, aspartyl-ARNt synthétase d'E coli





aspartyl-ARNt synthétase de levure

Dans cet exemple, chaque protéine possède un domaine absent chez l'autre (un seul est visible dans les vues 3D)

Alignement local: méthode de Smith-Waterman

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i,y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \\ 0 \end{cases}$$

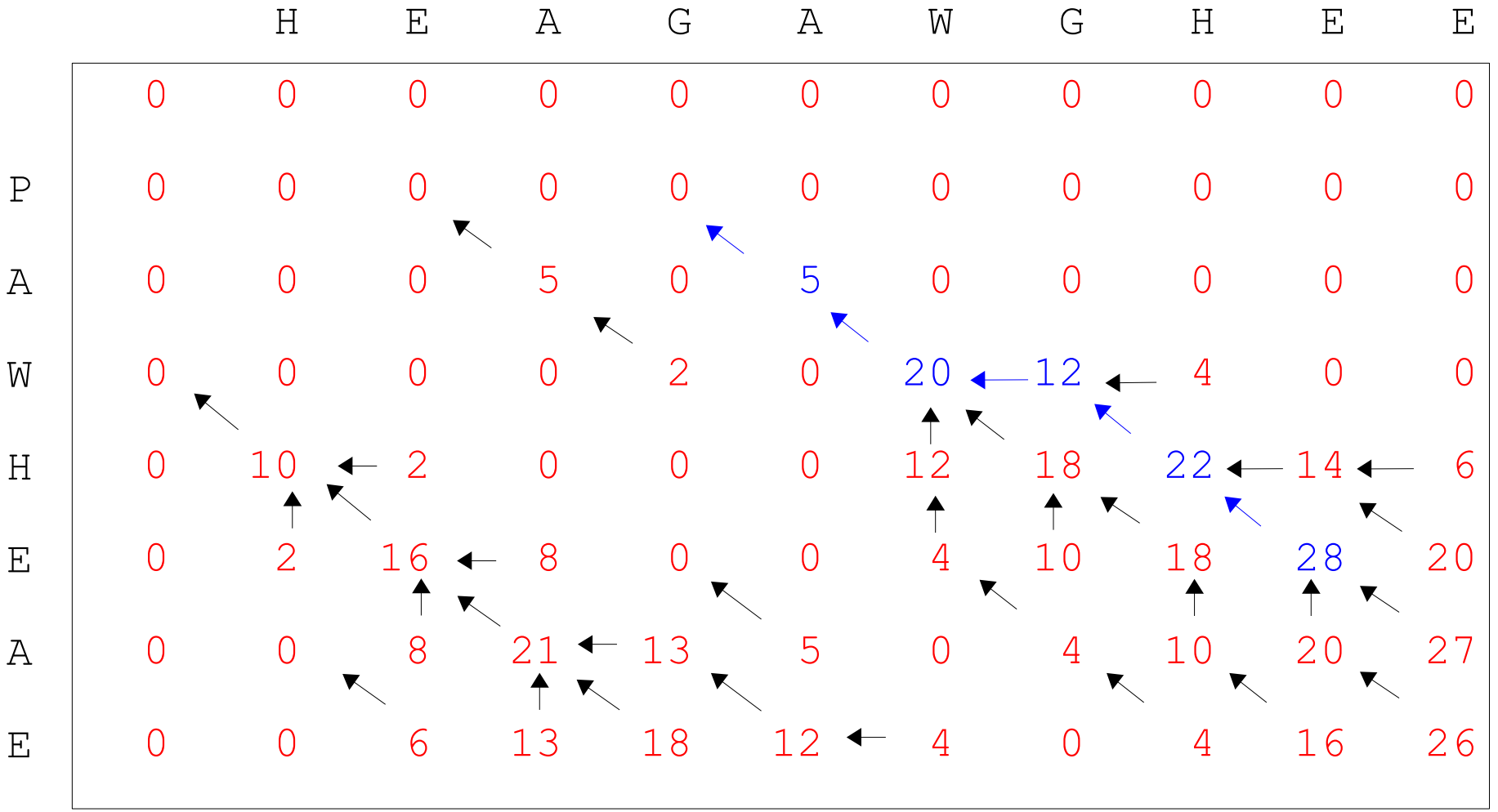
I G A x_i
L G V y_j

A I G x_i
G V y_j –

G A x_i – –
S L G V y_j

Nouvel alignement

On ne pénalisera pas les morceaux non-alignés en début et fin de séquence.



Alignement local optimal:

AWGHE

AW−HE

La méthode Smith-Waterman est utilisée en stratigraphie, pour aligner les prélèvements rocheux.

En résumé:

- **Comparaisons de séquences pour prédire la structure, la fonction, et l'origine évolutive des protéines**

- **L'alignement de séquences comme modèle d'un processus évolutif**

- **Alignement de séquences: algorithmes récursifs exacts et algorithmes heuristiques**

L'alignement de séquences: algorithmes heuristiques

Needleman-Wunsch: “exacte” mais coûteuse

- Atteint le maximum absolu de F(p,q)

Mais:

- Coût μ pq en mémoire et calcul
- Hypothèse de positions équivalentes
- Hypothèse de positions indépendantes:

$$P(\text{alignement}) = \Pi_{i,j}\ P(x_i, y_j)$$

Méthodes heuristiques rapides

BLAST: *Basic Local Alignment Search Tool*

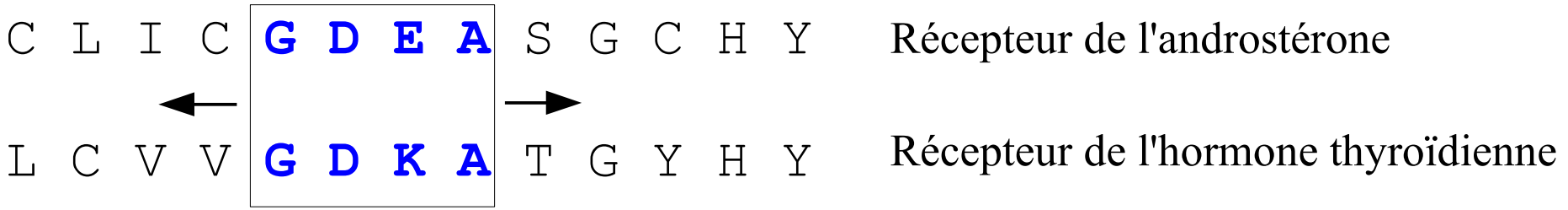
- recherche de tétrapeptides homologues
- extension de chaque tétrapeptide tant que la similarité > seuil

FASTA

- recherche de dipeptides homologues
- extension tant que la similarité > seuil
- alignement exact avec gaps, limité à une bande autour de l'extension trouvée

BLAST: ***Basic Local Alignment Search Tool***

- recherche de tétrapeptides homologues
- extension de chaque tétrapeptide tant que la similarité > seuil



Tétrapeptides homologues à un tétrapeptide de référence

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	

Extrait de BLOSUM62

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

7 homologues (avec BLOSUM62 et un seuil de 17).

On identifie dans la séquence y un tétrapeptide homologue à un tétrapeptide de la séquence x

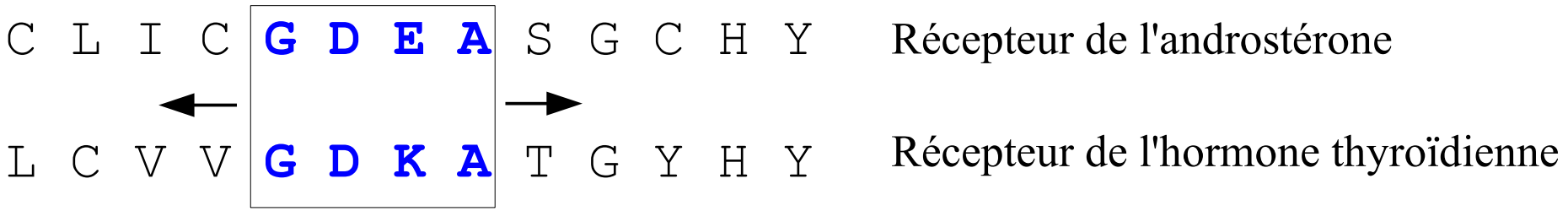
G D E A
G D D A
G D Q A
G E E A
G D E G
G D K A
G D E V

Eg: Tétrapeptides homologues à la référence **GDEA**

Séquence y

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

On étend le “micro-alignement” dans chaque direction, tant que le score > seuil



- Chaque tétrapeptide conduit à un alignement local sans gaps
- On retient les meilleurs alignements

Homologues du récepteur de l'androstérone identifiés par recherche BLAST dans la banque SwissProt

#	ID Swissprot	Hit	Description	Score	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

FASTA: intermédiaire entre BLAST et Needleman-Wunsch

- recherche de dipeptides homologues
- extension tant que la similarité > seuil
- alignement exact avec gaps, limité à une bande autour de l'extension trouvée

Cf. document écrit

En résumé:

- **Comparaisons de séquences pour prédire la structure, la fonction, et l'origine évolutive des protéines**

- **L'alignement de séquences comme modèle d'un processus évolutif**

- **Alignement de séquences: algorithmes récursifs exacts et algorithmes heuristiques**