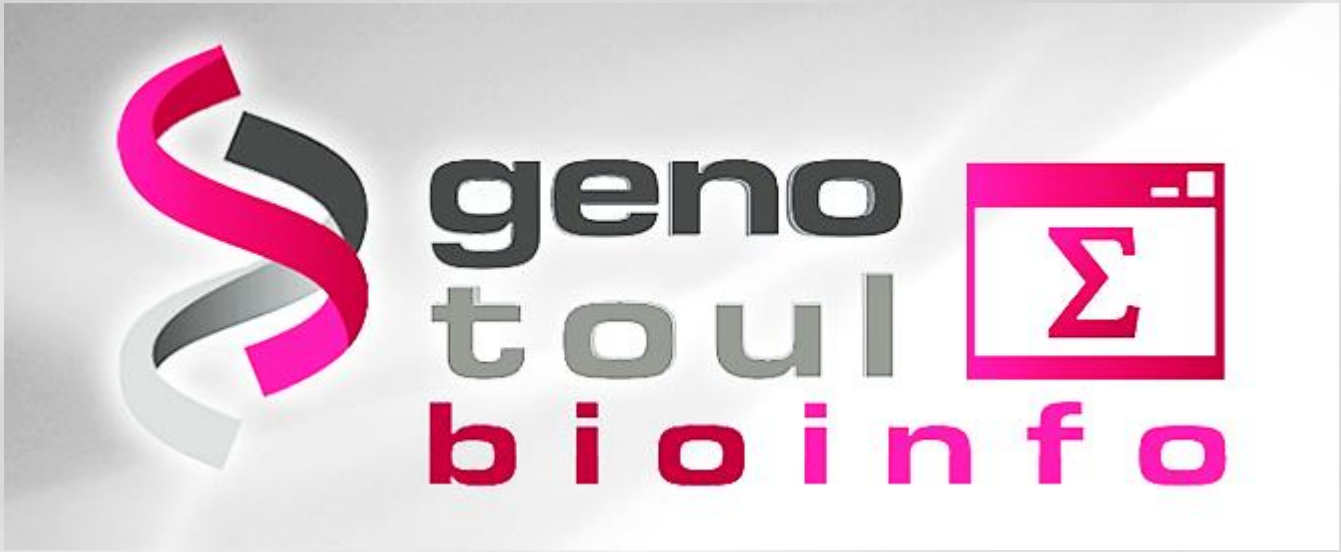
Plateforme Bioinformatique Midi-Pyrénées



Module 1: Initiation à l'alignement de séquence pour la phylogénie

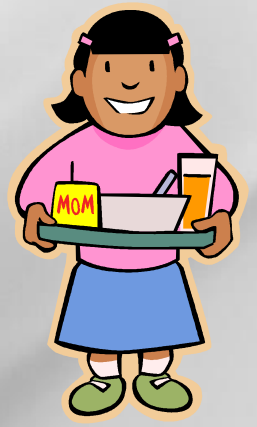
Géraldine PASCAL

FeedBack: Pourquoi vous intéressez-vous où aves-vous besoin de la reconstruction phylogénétique?

Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

2 pauses courtes
matin et après-midi

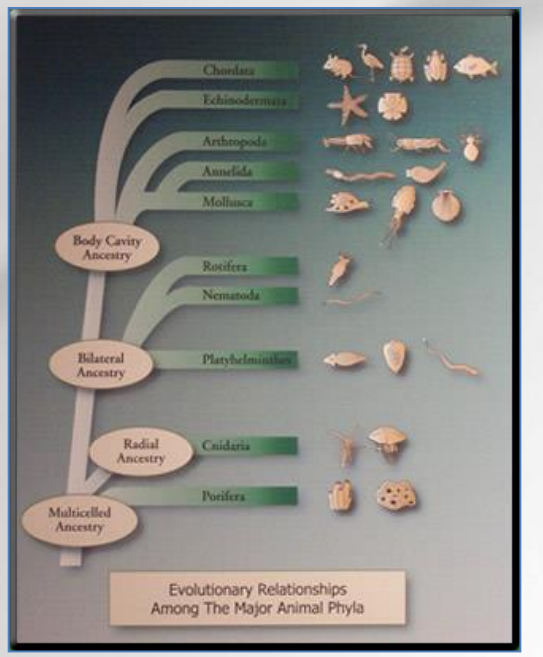


Déjeuner
12h30 à 13h30

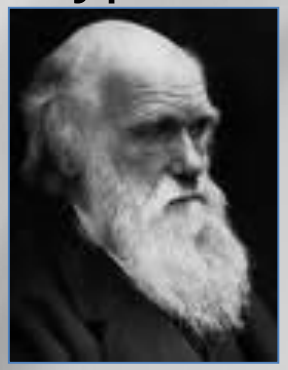
La théorie de l'évolution

Les espèces se modifient au cours du temps et donnent naissance à de nouvelles espèces.

Selon Jean-Baptiste Lamarck (1744-1829), les espèces évoluent en adoptant des caractères acquis par les individus au cours de leur vie.



Charles Darwin (1809-1882) émet l'hypothèse de la sélection du plus apte (ou sélection naturelle) parmi des individus naturellement variant.



La déduction par homologie, ou le «dogme central» de la bioinformatique

Si la bioinformatique «marche», c'est parce que l'évolution des gènes laisse une trace parfaitement visible lorsque l'on compare leur séquence

Évolution des gènes = mutations, insertions, délétion.

Les gènes des organismes modernes sont issus de remaniement de gènes ancestraux

On peut donc déduire la fonction de la plupart des gènes par comparaison avec les gènes «homologues» d'autres espèces.
(homologue = qui a un ancêtre commun)

La déduction par homologie, ou le «dogme central» de la bioinformatique

Les régions fonctionnelles des gènes (sites catalytique, de fixation, etc.) sont soumises à sélection.

Elles sont relativement préservées par l'évolution car des mutations trop radicales sont désavantageuses.

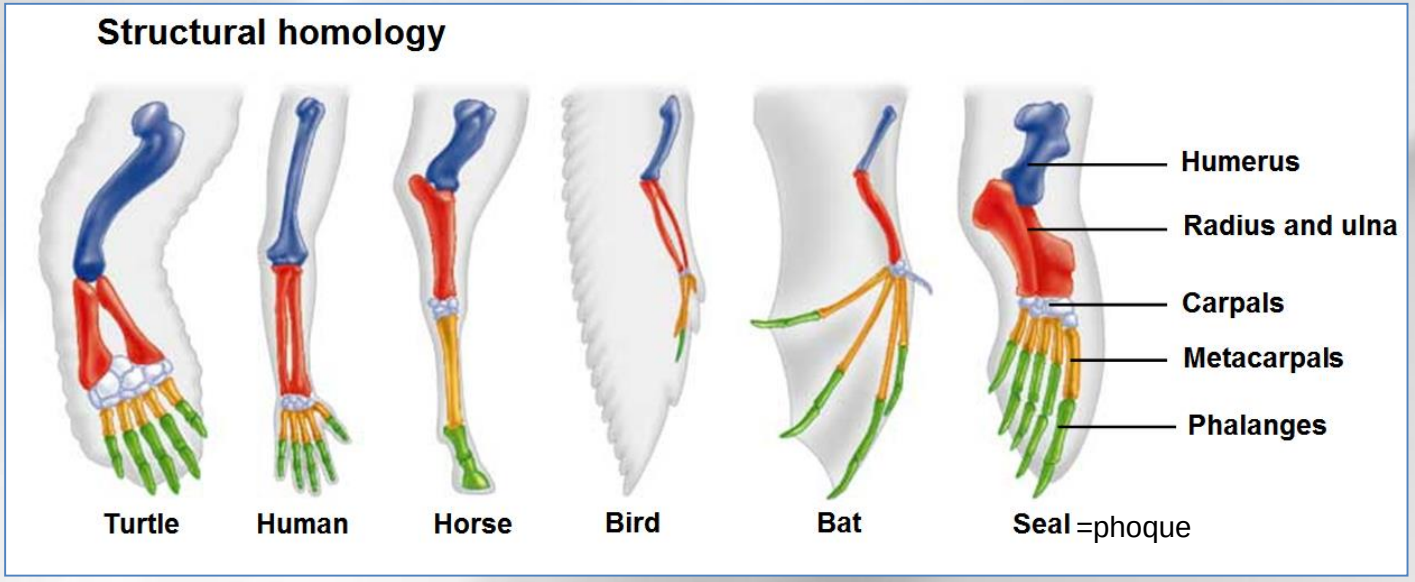
Les régions non fonctionnelles subissent peu de pressions de sélection et divergent rapidement au fur et à mesure que s'accumulent les mutations.

L'homologie de séquence

En bioinformatique aussi Homologie = parenté = ancêtre commun

L'aile de l'aigle est homologue à l'aile du perroquet (oiseaux) et à la patte de la tortue (reptile)

L'aile de la chauve souris est homologue à la patte du cheval et au bras de l'homme (chiroptère, ongulé et primate - mammifères).



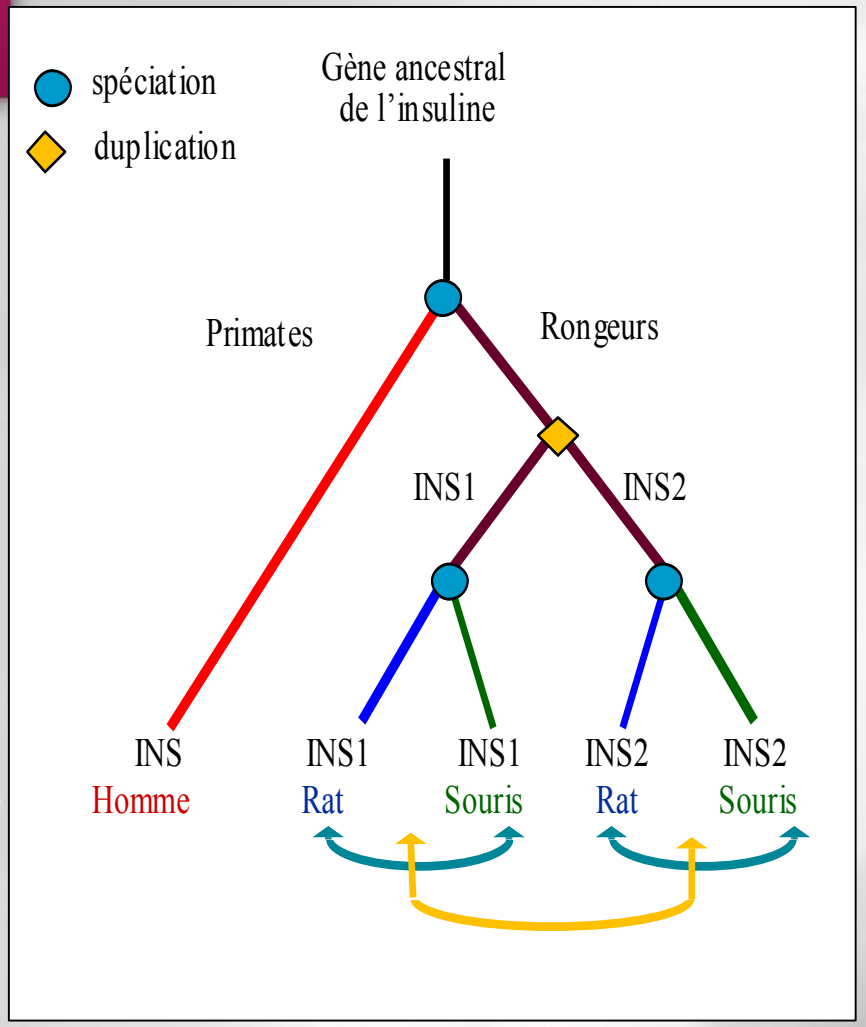
L'homologie de séquence

On est homologue ou on ne l'est pas.

Donc on ne dit pas: "~~très homologue~~", "~~faible homologie~~", ~~«22% d'homologie~~», etc.

Pour une notion quantitative, on parle de similitude ("très similaire", etc.) ou d'identité (28% d'identité)

Orthologie et paralogie



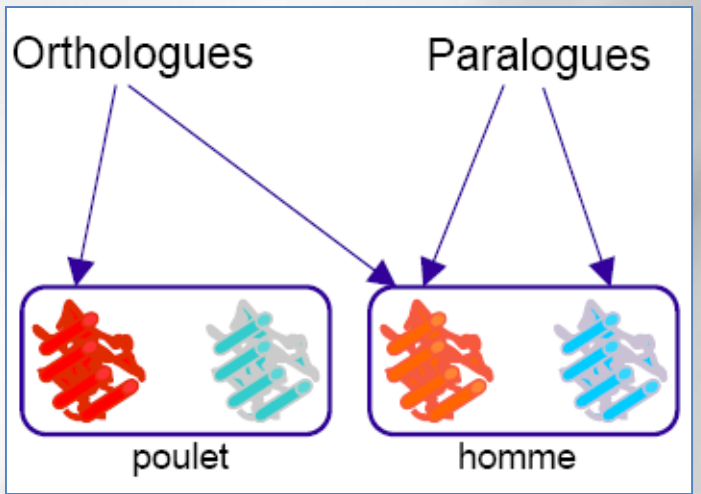
Homologie : deux gènes sont homologues si ils ont un ancêtre commun

Orthologie : deux gènes sont orthologues si ils ont divergé à la suite d'un évènement de spéciation

Paralogie : deux gènes sont paralogues si ils ont divergé à la suite d'un évènement de duplication

Fonction et homologie

- Homologie n'implique pas même fonction: par exemple l'aile de l'oiseau et le bras humain n'ont pas la même fonction
- Des orthologues rapprochés (p. ex. homme/souris) ont le plus souvent la même fonction dans l'organisme.
- Des orthologues distants (p. ex. homme/mouche) ont plus rarement le même rôle phénotypique, mais peuvent exercer le même rôle dans une voie donnée.
- Les paralogues acquièrent rapidement des fonctions différentes



La phylogénie, a quoi ça sert ?

La phylogénie, a quoi ça sert ?

- Histoire évolutive de familles de gènes
  - Analyse des duplications et des pertes de gènes.
  - Détection de transferts horizontaux.
- Histoire évolutive des organismes
- Ecologie
  - Phylogéographie.
  - Co-évolution hôte-parasite.
- Epidémiologie.
- Assignation taxonomique ou fonctionnelle.
- Identification de chimères.

La phylogénie, a quoi ça sert ?

- Histoire évolutive de familles de gènes
  - Analyse des duplications et des pertes de gènes.
  - Détection de transferts horizontaux.
- Histoire évolutive des organismes
- Ecologie
  - Phylogéographie.
  - Co-évolution hôte-parasite.
- Epidémiologie.
- Assignation taxonomique ou fonctionnelle.
- Identification de chimères.

Retracer l'histoire évolutive des espèces

- Une observation des êtres vivants révèle l'existence de nombreux points communs entre systèmes vitaux des organismes (respiration, circulation, excrétion, reproduction).

- Depuis l'Antiquité, l'homme a tenté de classifier la nature selon les ressemblances et les différences qu'il observait chez les animaux et les végétaux qui l'entouraient.

Retracer l'histoire évolutive des espèces

- Si certaines espèces se ressemblent beaucoup et d'autres moins, une unité du vivant n'a guère été remise en cause; elle a même été renforcée au cours du XXe siècle par la découverte d'un ADN unique supportant les codes génétiques de tout le vivant.

- Le phénomène d'évolution permet d'expliquer ces ressemblances par l'existence de liens généalogiques entre toutes les formes de vie: les organismes se ressemblent parce qu'ils partagent des caractères hérités d'un ancêtre commun.

Retracer l'histoire évolutive des espèces

Des faits observés dans plusieurs disciplines scientifiques corroborent cette théorie:

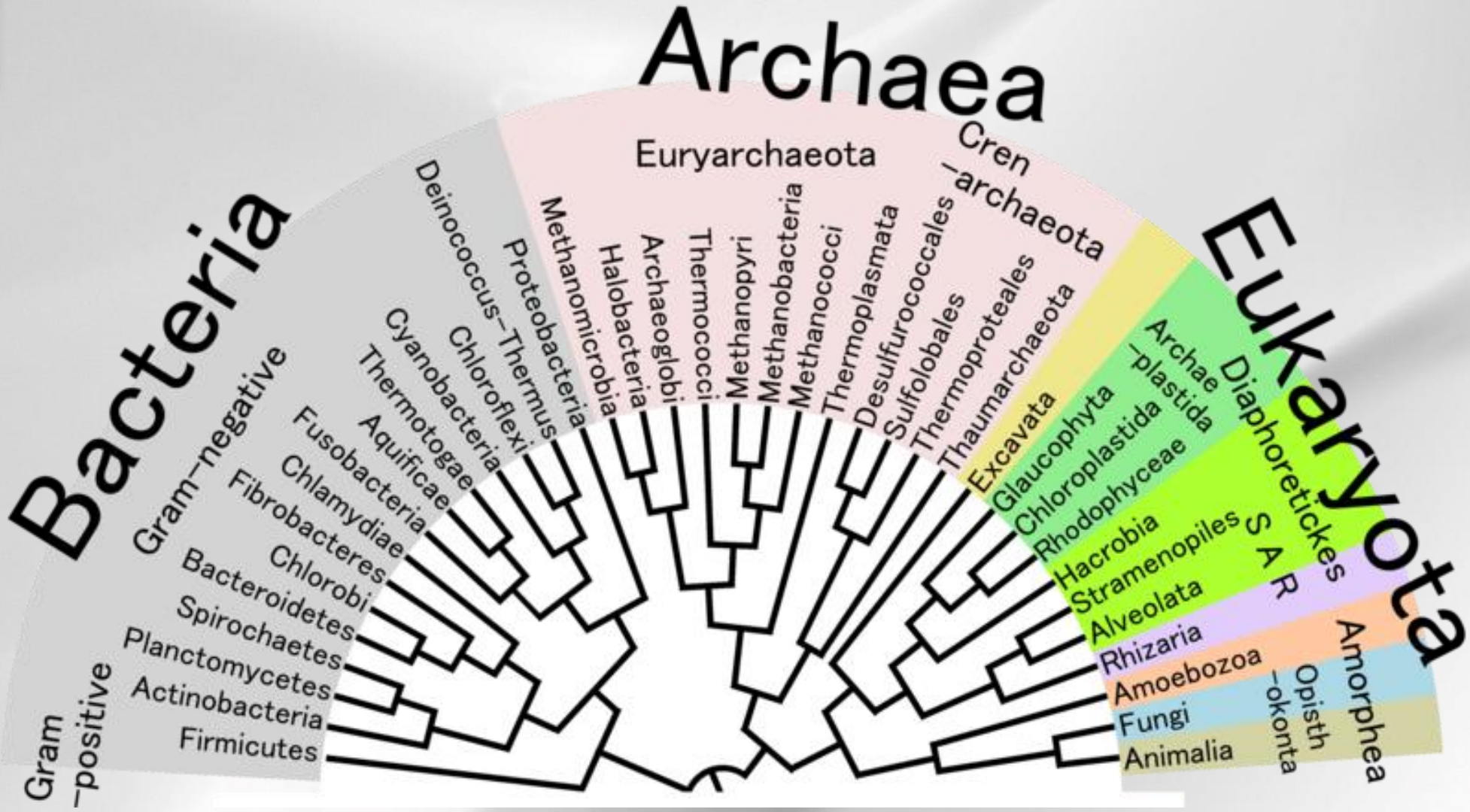
Faits anatomiques

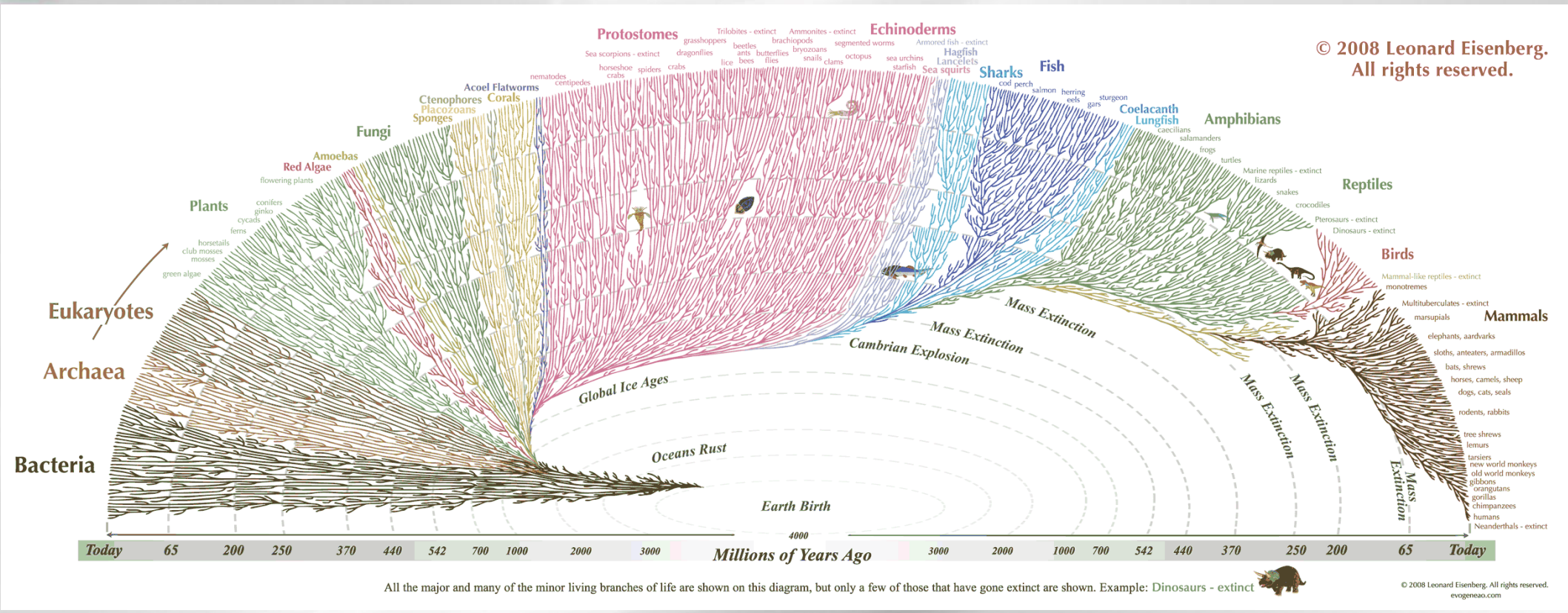
Faits biologiques

Faits paléontologiques

Faits génétiques

Arbre phylogénétique





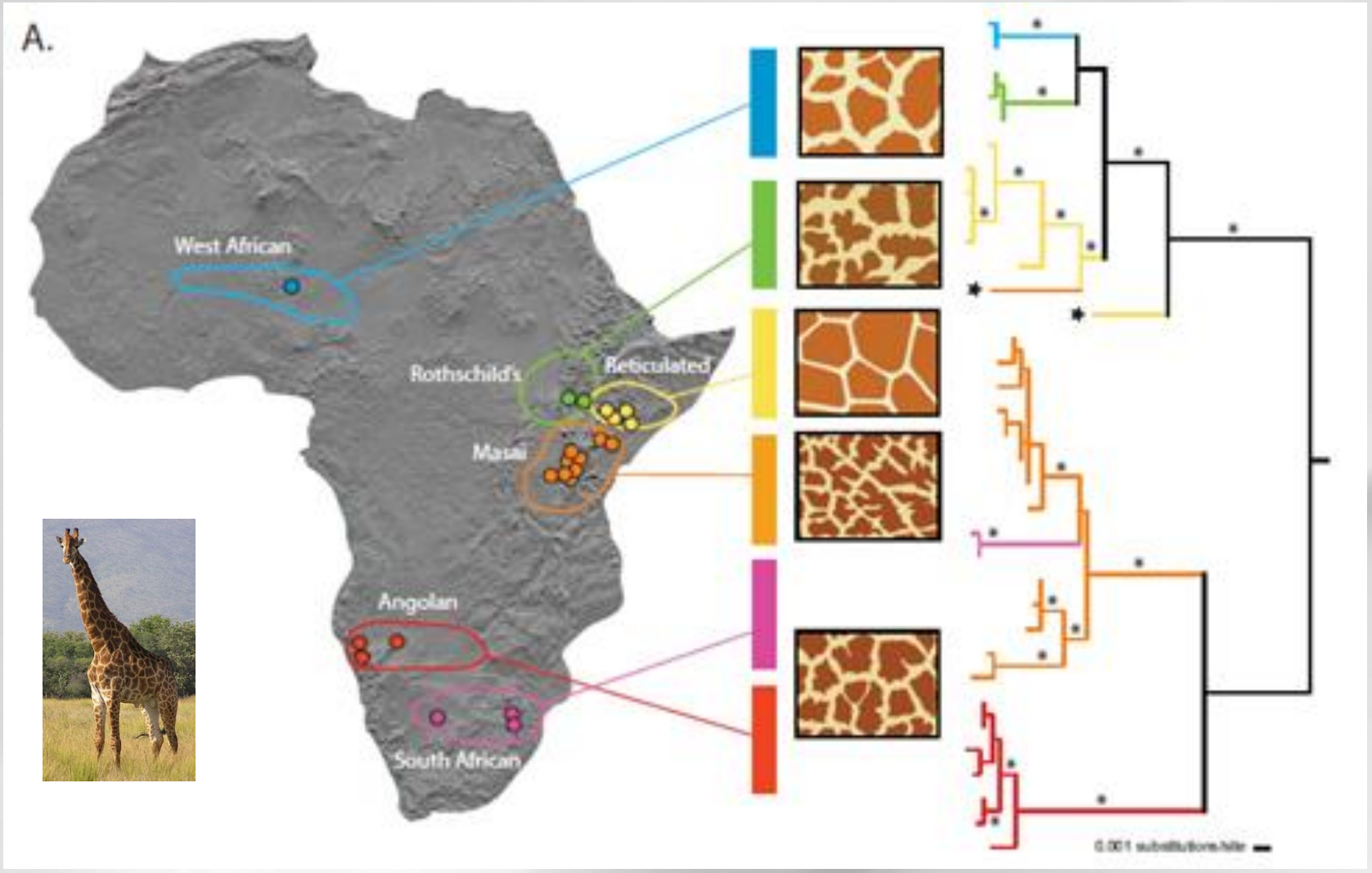
All the major and many of the minor living branches of life are shown on this diagram, but only a few of those that have gone extinct are shown. Example: Dinosaurs - extinct

Aujourd'hui
phylogénie = phylogénie moléculaire

Alliée à la puissance de calcul et à l'accessibilité toujours plus grandes des données moléculaires, la phylogénie moléculaire est majoritairement utilisée car puissante et précise.

La phylogénie, a quoi ça sert ?

- Histoire évolutive de familles de gènes :
  - Analyse des duplications et des pertes de gènes.
  - Détection de transferts horizontaux.
- Histoire évolutive des organismes les portant.
- Ecologie :
  - Phylogéographie.
  - Co-évolution hôte-parasite.
- Epidémiologie.
- Assignation taxonomique ou fonctionnelle.
- Identification de chimères.

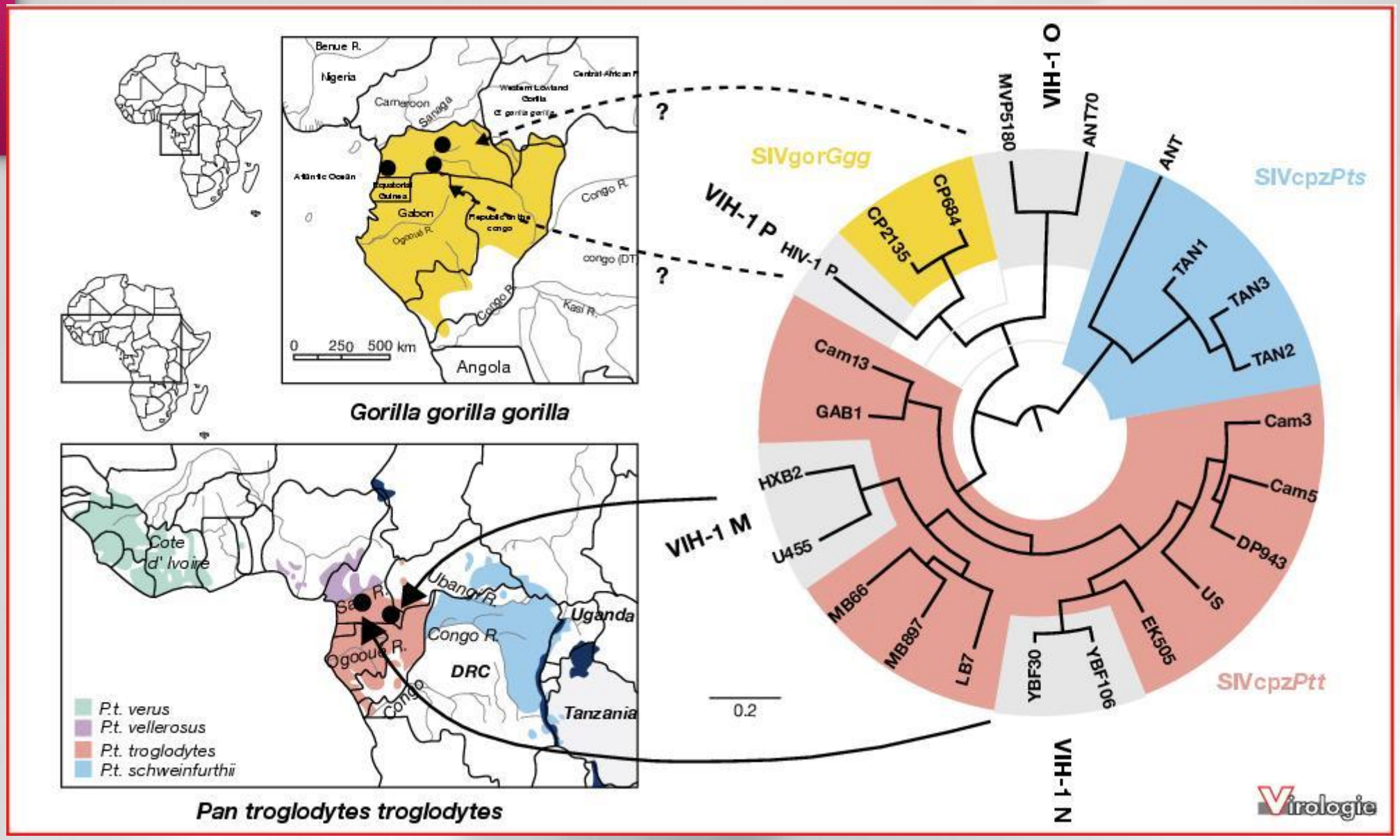


from Brown *et al*. (2007) - shows the approximate ranges and pelage patterns of the extant giraffe taxa

La phylogénie, a quoi ça sert ?

- Histoire évolutive de familles de gènes :
  - Analyse des duplications et des pertes de gènes.
  - Détection de transferts horizontaux.
- Histoire évolutive des organismes les portant.
- Ecologie :
  - Phylogéographie.
  - Co-évolution hôte-parasite.
- Epidémiologie.
- Assignation taxonomique ou fonctionnelle.
- Identification de chimères.

Origine du HIV – groupe 1



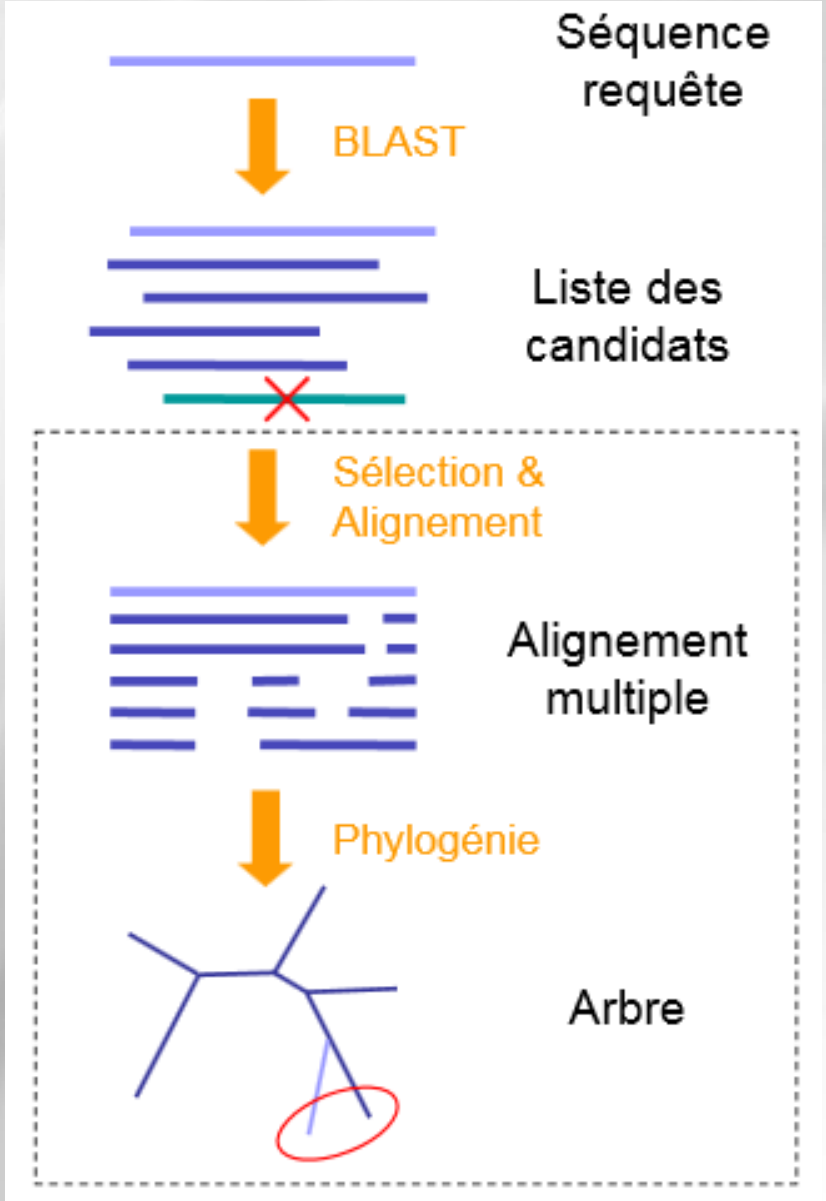
La phylogénie, a quoi ça sert ?

- Histoire évolutive de familles de gènes :
  - Analyse des duplications et des pertes de gènes.
  - Détection de transferts horizontaux.
- Histoire évolutive des organismes les portant.
- Ecologie :
  - Phylogéographie.
  - Co-évolution hôte-parasite.
- Epidémiologie.
- Assignation taxonomique ou fonctionnelle.
- Identification de chimères.

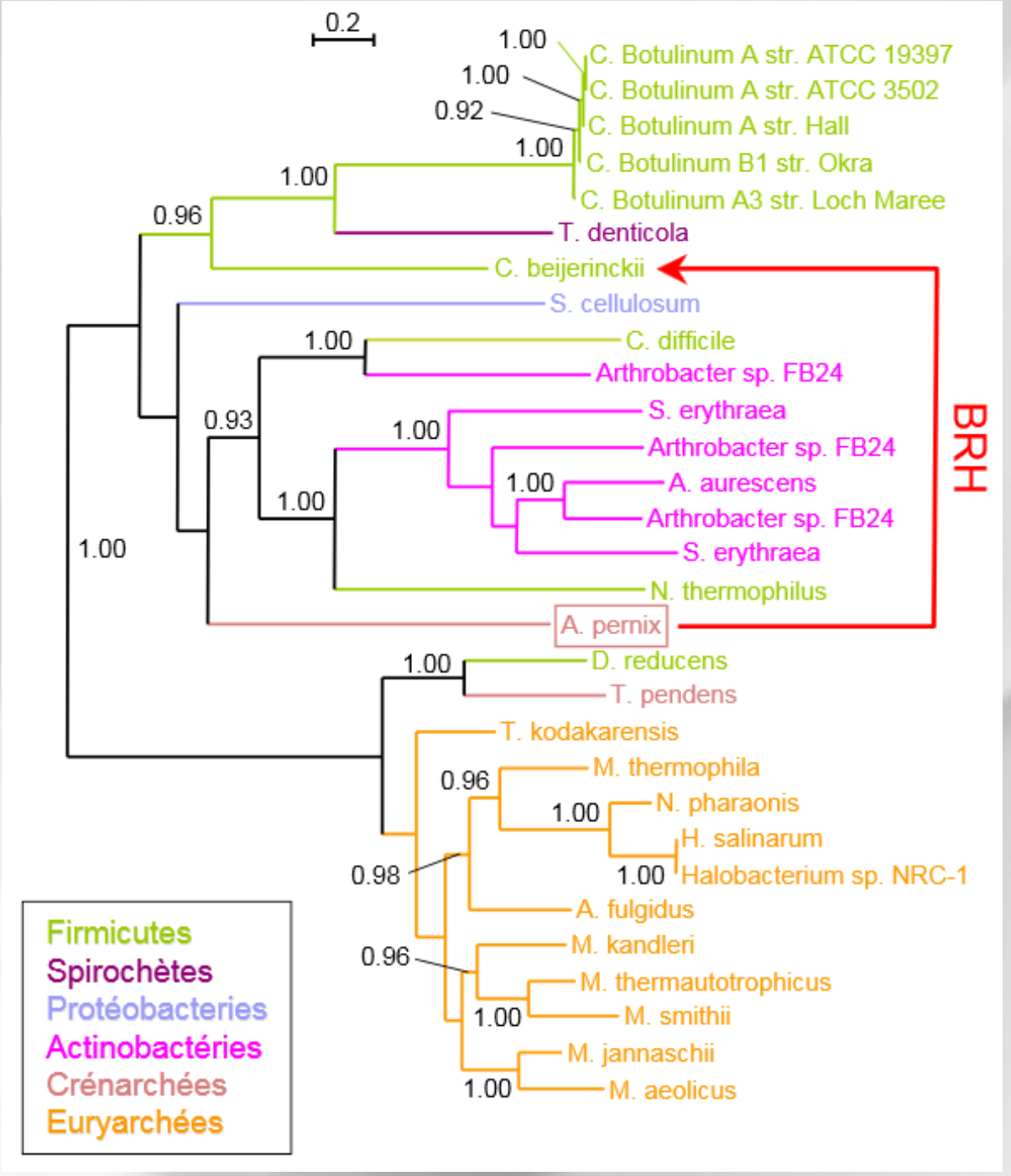
Identification taxonomique

La démarche:

1. Recherche par similarité (avec Blast par ex) séquences 16S/18S ou gènes de ménage (*rpoB, tuf, gyrB, sodA,recA, groES*…)
2. Sélection de candidats
3. Alignement multiple
4. Arbre phylogénétique



Identification taxonomique



Plateforme Bioinformatique Midi-Pyrénées

Les différentes étapes

de la reconstruction phylogénétique

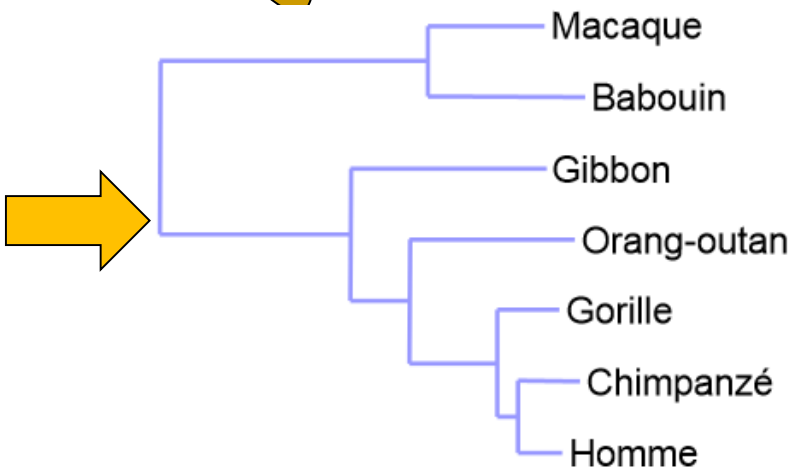
Les différentes étapes

Fichier fasta de séquences

```
>yneN
TTAATGCCTCTTCTCATTTCTTCTGCTGTCATCCGCACAGCAGAAGAATTCCTCATTGAC
TATTATTTCGCAATTTGCTCACATGGATTAAATTAAACTACATACTATAAGATATAAACT
TCTGCCTACAGCTGTAAGAAACTCCGCTCAGTACTGAAGCACCAGTCCTATTTCCTCTTT
TCTCCAGCCTGTTATATTAAGCATACTGATTAACGATTTTTAACGTTATCCGCTAAATAA
ACATATTTGAAATGCATGCGACCACAGTGAAAAACAAAATCACGCAAAGAGACAACTATA
A
>yegR
ACTAACGGCTGCCACCGATAAATTTCAAAAAAGAGCATATACCTAATATTCAACTAAACA
GTGGCATCTTCAATATAATATATTAAAGCCCCCATGGAGTTACCCTGAAGGGCCTCAATG
TCCGTAATTCCTACTTATGTAGGAAATGTTGTACAGAACATTTATTATAATCCTATTCAA
TTATAATAATCATGCCATTATTATATTTAAACACTAGAGAGTGTCGTTGGTATTTAATGG
GGGAAGGTGAGATGAAAAAGATAGCTGCTATATCATTAATTAGTATTTTTATTATGTCTG
G
>emrK
AAATCAGGGATTGTACCGATGATTTATAGTTTCAAGTTGGCACTATAAGTCTTCTTACTA
ATCCTACAGGCGTAAGAATTGTATTGCAAAAGCCACGGTTTAGTCCTCTGTTGTTTTTTT
TGCACCTCATTTAAATTAGGCCTCCAACGTTCCTGGGATAATGTGCAACACATGCACTGT
GTTTGATATGAAGAATGAATGCTCTTTTCATTCAATTCATAAATTTCATCTATGAGAAAT
GAGAGATAATAGTGGAACAGATTAATTCAAATAAAAAACATTCTAACAGAAGAAAATACT
T
>evgA
AATACAATTCTTACGCCTGTAGGATTAGTAAGAAGACTTATAGTGCCAACTTGAAACTAT
AAATCATCGGTACAATCCCTGATTTTATTGTTGACATTTCATTTATGCCGACTATTTATA
TGGTATACTTGTCGAATTATCTTAAAGGAAGCTCAGATTTTCTTATTTTTATTGAGAAAA
TGAGATGACGCCTTATGTCTGTATTACTACAGGGAGAAGGGAGATGCTTCATTGCAAAGG
GAATAATCTATGAACGCAATAATTATTGATGACCATCCTCTTGCTATCGCAGCAATTCGT
>yfdX
TGGCTGTATTTACATTTAATTAATCAGTATTTACATCGATATAATAAATGACATCTCTTT
GTGGTATATAAGAATAGTTCTCTGCGACAGGAAGCATATTCCTACAATTGTAAGACTAAA
ATACTTCTTGCGATAATAACTACAACTGTAAGATAACCCTTTCAAAATGACCGTTGCTCT
CTGATTTCTCATTTCATGCTCACCCAATATGATGGCGGCGTTTTCTAAAACTGTTAAAGA
ATGAGGTAAGTATGAAACGTTTAATTATGGC[illegible]ATGGTCACAGCAATTCTGGCATCTT
C
```

```
Gibbon      AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTT
Orang       AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCT
Gorille     AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCAT
Homme       AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGACTTACATCCT
Chimpanzé   AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCT
Macaque     AAGCTTTTCCGGCGCAACCATCCTTATGATCGCTCACGGACTCACCTCTT
Babouin     AAGCTTCTCCGGTGCAACCATCCTTATGATTGCCCACGGACTCACCTCTT
```

Alignement



Arbre

Les différentes étapes

Point de départ :

- Un ensemble de séquences *homologues* alignées.
  - Avoir sa séquence et rechercher des séquences *similaires* dans les banques de données par blast par exemple

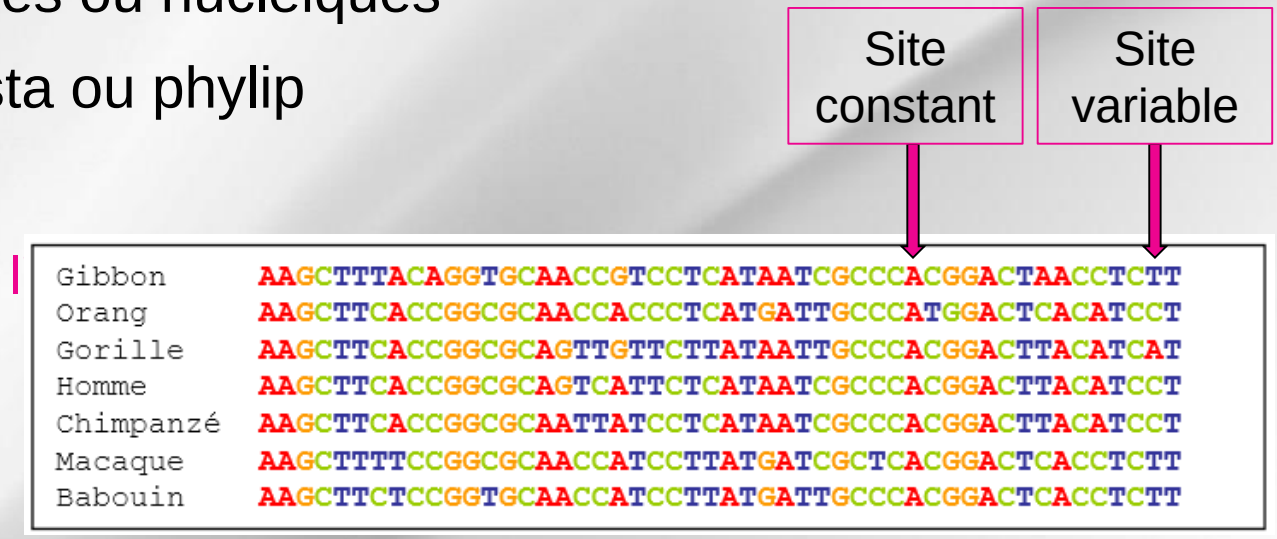
  OU

  - Avoir *sa propre liste* de séquences déduite d'expériences biologiques

- Séquences protéiques ou nucléiques
- Format standard fasta ou phylip

Alignement multiple

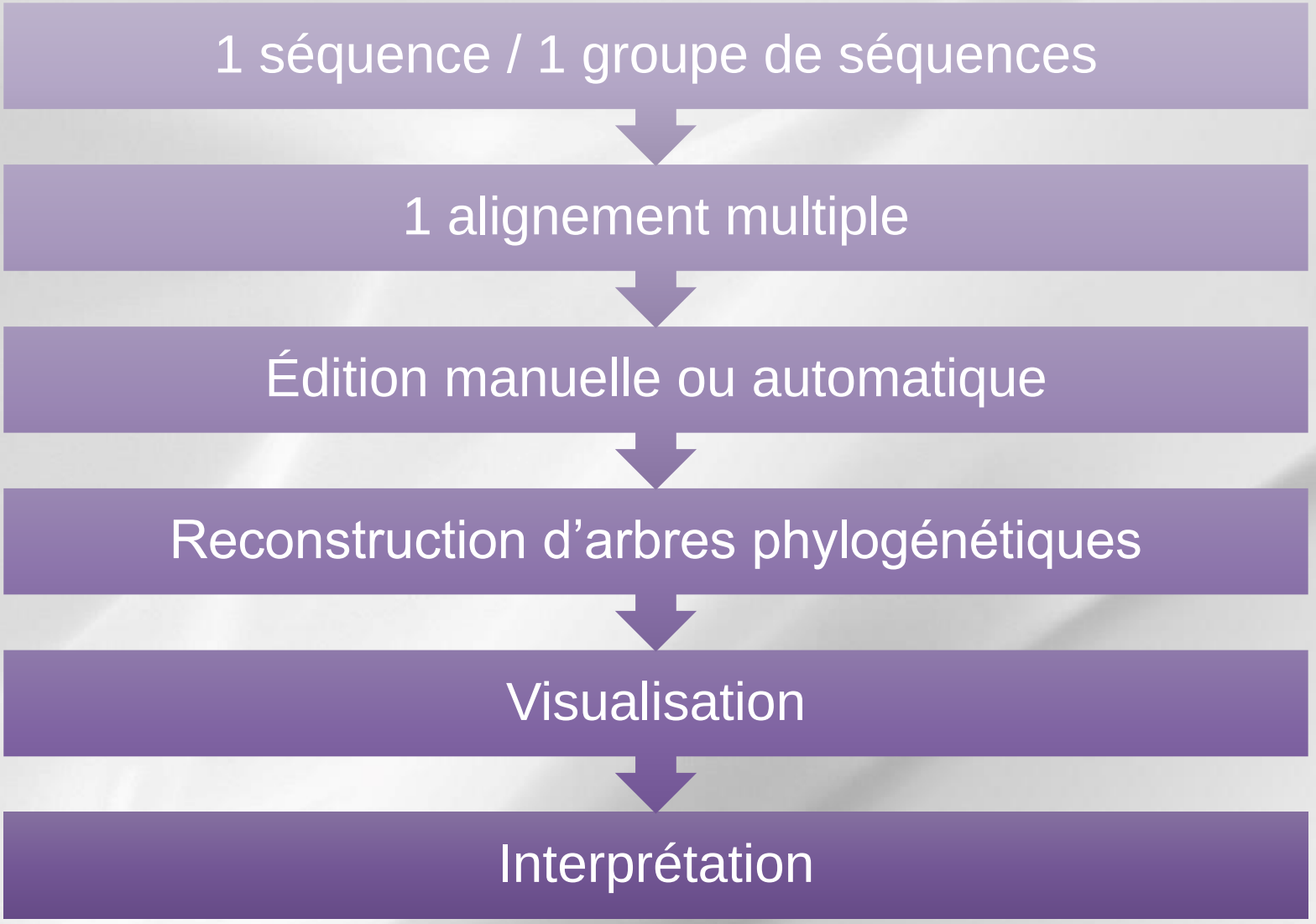
- Chaque position dans l



Résultat obtenu :

- Un arbre décrivant les relations évolutives entre les séquences
  i.e. un arbre phylogénétique.

Les différentes étapes



Homologues = similaires + variable

Séquences homologues = régions constante + régions variables

La phylogénie moléculaire observe la variabilité de conservation des caractères du groupe observé.

Définition d'un caractère héritable

- Attribut d'un membre d'une population ou d'un taxon par lequel il diffère d'un autre membre d'un groupe ou d'un autre taxon.

- Plus généralement, un caractère peut être tout attribut utilisé pour reconnaître, décrire, définir ou différencier les taxons.

- Dans le cadre de la phylogénie moléculaire un caractère est une base nucléique ou le codon.

La phylogénie moléculaire est fondée sur l'utilisation de séquences homologues.

Homologues, similaires ?

- Deux séquences sont dites homologues si et seulement si elles possèdent un ancêtre commun.
- Seuil variable suivant les circonstances :
  - Similarité sans homologie (homoplasie, répétitions).
  - Homologie avec faible similarité (limitation à quelques positions clés dans les séquences). Des séquences sans ressemblance apparente peuvent parfaitement être homologues (on le retrouve par ex. au niveau 3D)
- Etant donné la dimension de l'espace des séquences possibles, une ressemblance importante est généralement interprétée comme une homologie, et non pas comme une évolution convergente.

L'homoplasie: l'un des pièges de la similarité

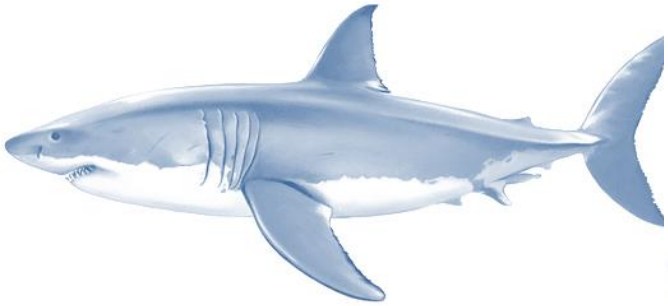
Ce sont les similarités non homologues

Résultat d'une évolution indépendante

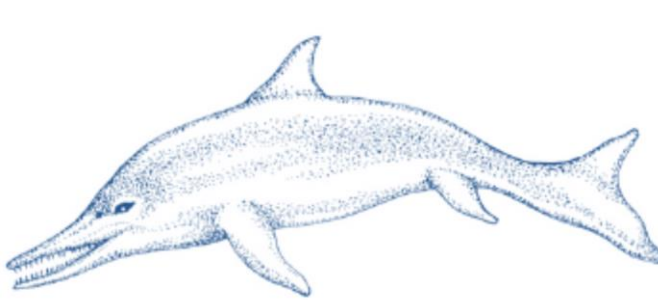
- Convergence évolutive
- Parallélisme (évolution parallèle)
- Réversion évolutive

Brouillent le signal phylogénétique : peuvent conduire à l'établissement de fausses relations de parenté

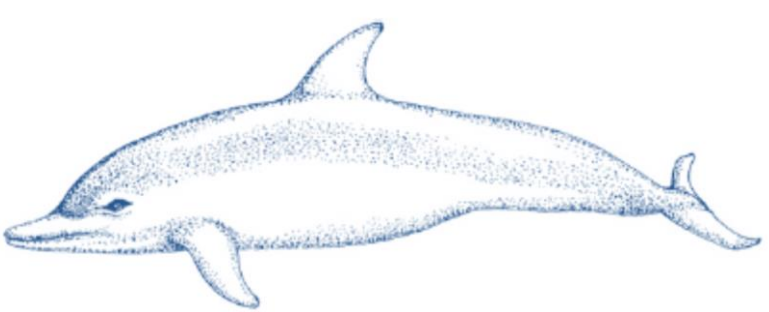
L'homoplasie: l'un des pièges de la similarité



REQUIN



ICHTHYOSAURE



DAUPHIN

ressemblances trompeuses entre les organismes adaptés à des modes de vie similaires

L'homoplasie: l'un des pièges de la similarité

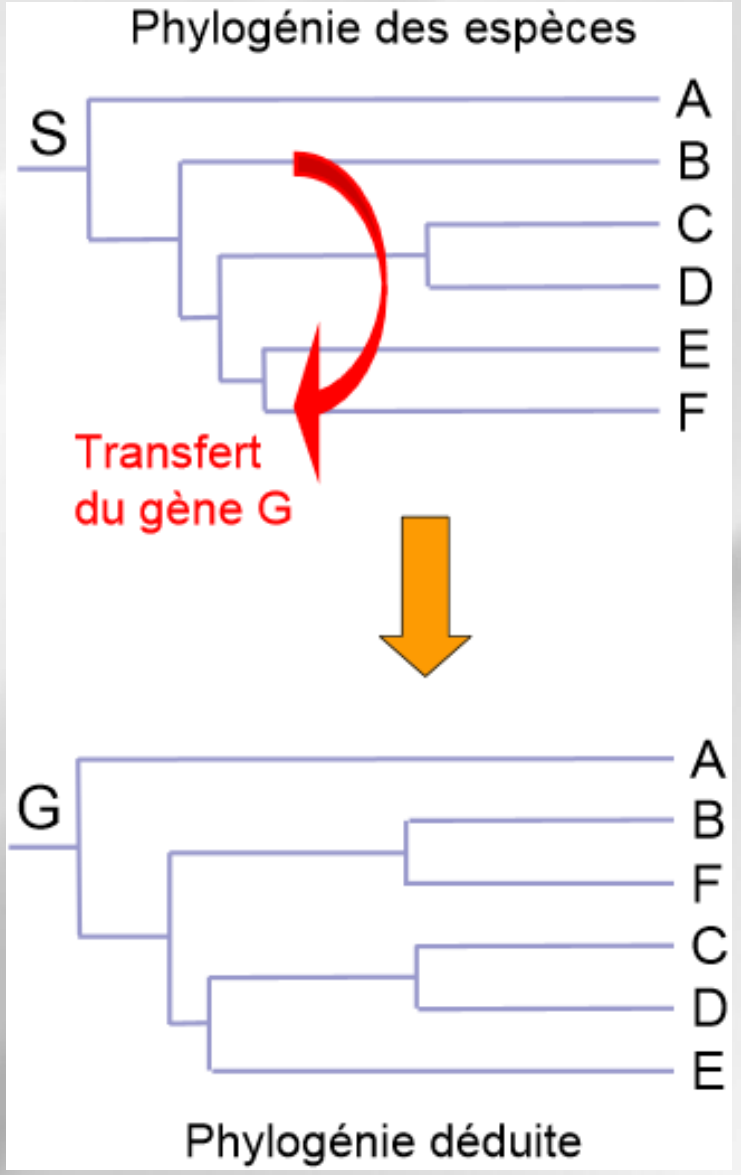
Au niveau moléculaire:

Homoplasie → difficile à détecter par alignement de séquences

Ex: cas d'homoplasie → transfert horizontal

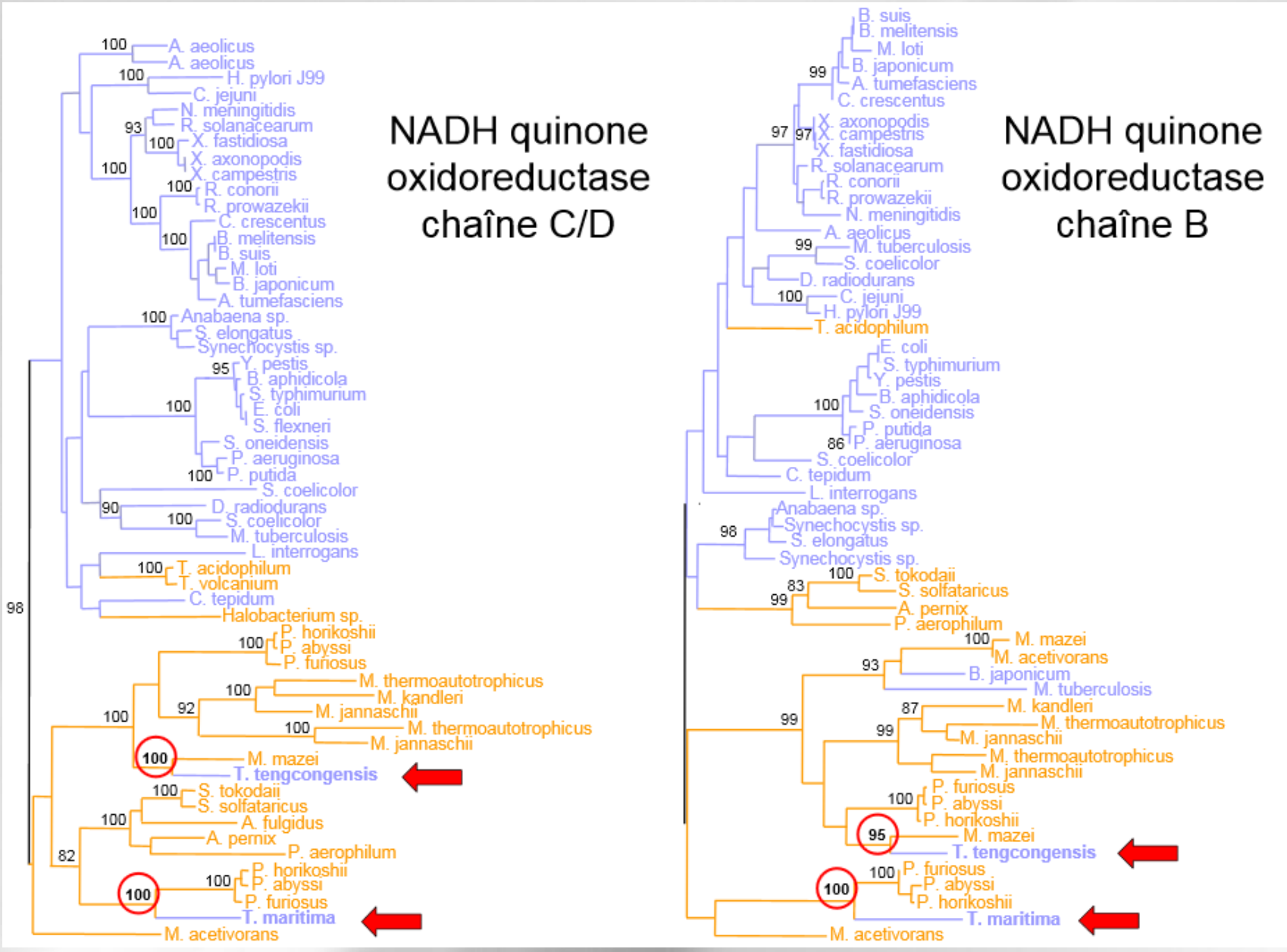
Attention au transfert horizontal

- Transmission de gènes entre taxons différents.
- Phénomènes supposés très fréquents chez les procaryotes :
  - Implication de différents mécanismes :
    - Transformation.
    - Conjugaison.
    - Transduction.
  - 17.6% des gènes d'*E. coli* auraient été obtenus par transfert.

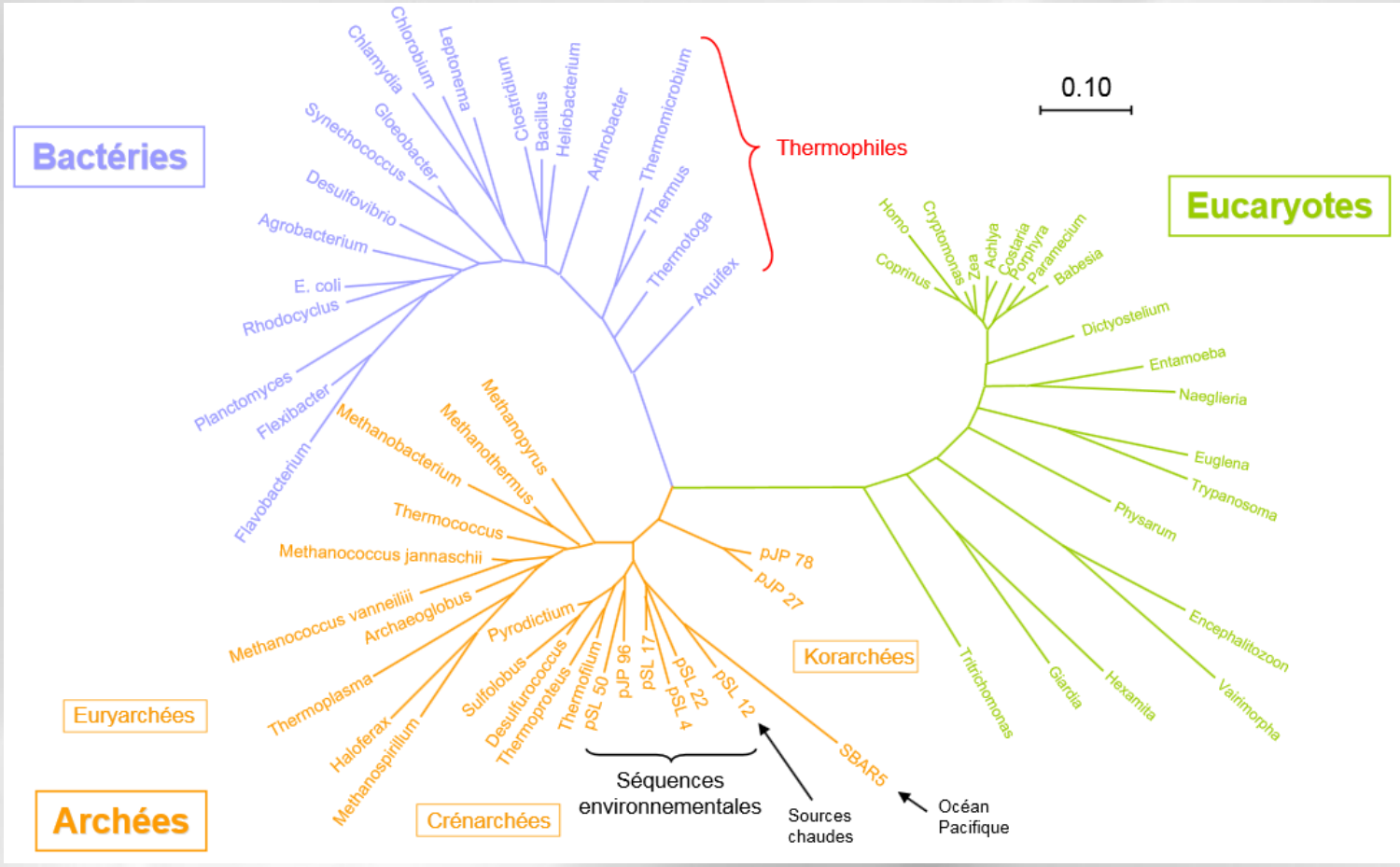


Attention au transfert horizontal

Origin *and* ***evolution*** *of the* ***sodium -pumping NADH****:* ***ubiquinone oxidoreductase****.*
Reyes-Prieto A, Barquera B, Juárez O.
PLoS One. 2014 May 8;9(5):e96696. doi: 10.1371/



Attention au transfert horizontal



L'homoplasie: l'un des pièges de la similarité

Au niveau moléculaire:

Homoplasie → difficile à détecter par alignement de séquences

Ex: cas d'homoplasie → transfert horizontal

Solution → reconstruire des arbres phylogénétique à partir de plusieurs gènes de l'organisme

Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

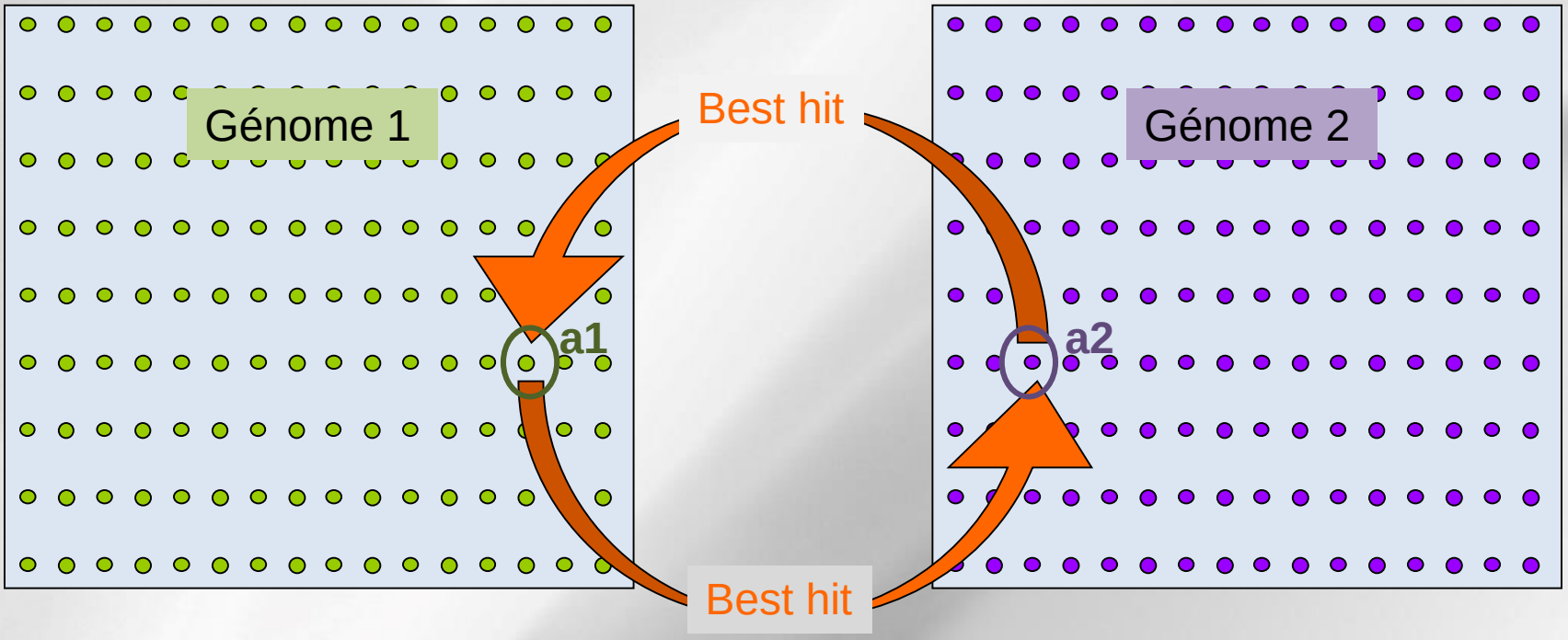
Constitution de son jeu de données

Comment choisir un orthologue potentiel ?

Le score, l'e-value de blast

La longueur de l'alignement

Le RBH (ou BBH) :Reciprocal Best Hit (Bidirectionnal Best Hit)

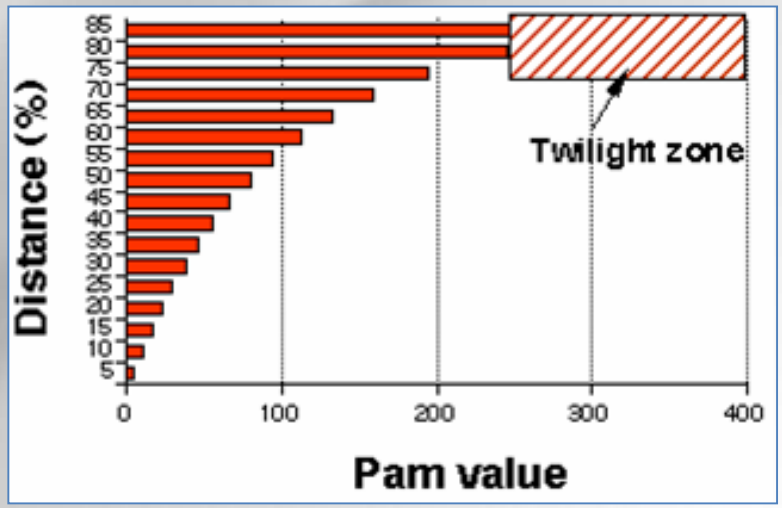


A quel point des séquences homologues se ressemblent-elles?

De 100% à quelques nucléotides/acides aminés en commun.

Il n'y a pas vraiment de limite, mais en dessous d'un certain niveau d'identité (twilight zone = zone nébuleuse), il devient difficile de distinguer une homologie d'une ressemblance fortuite.

Deux séquences d'ADN prises au hasard ont 25% de nucléotides communs.



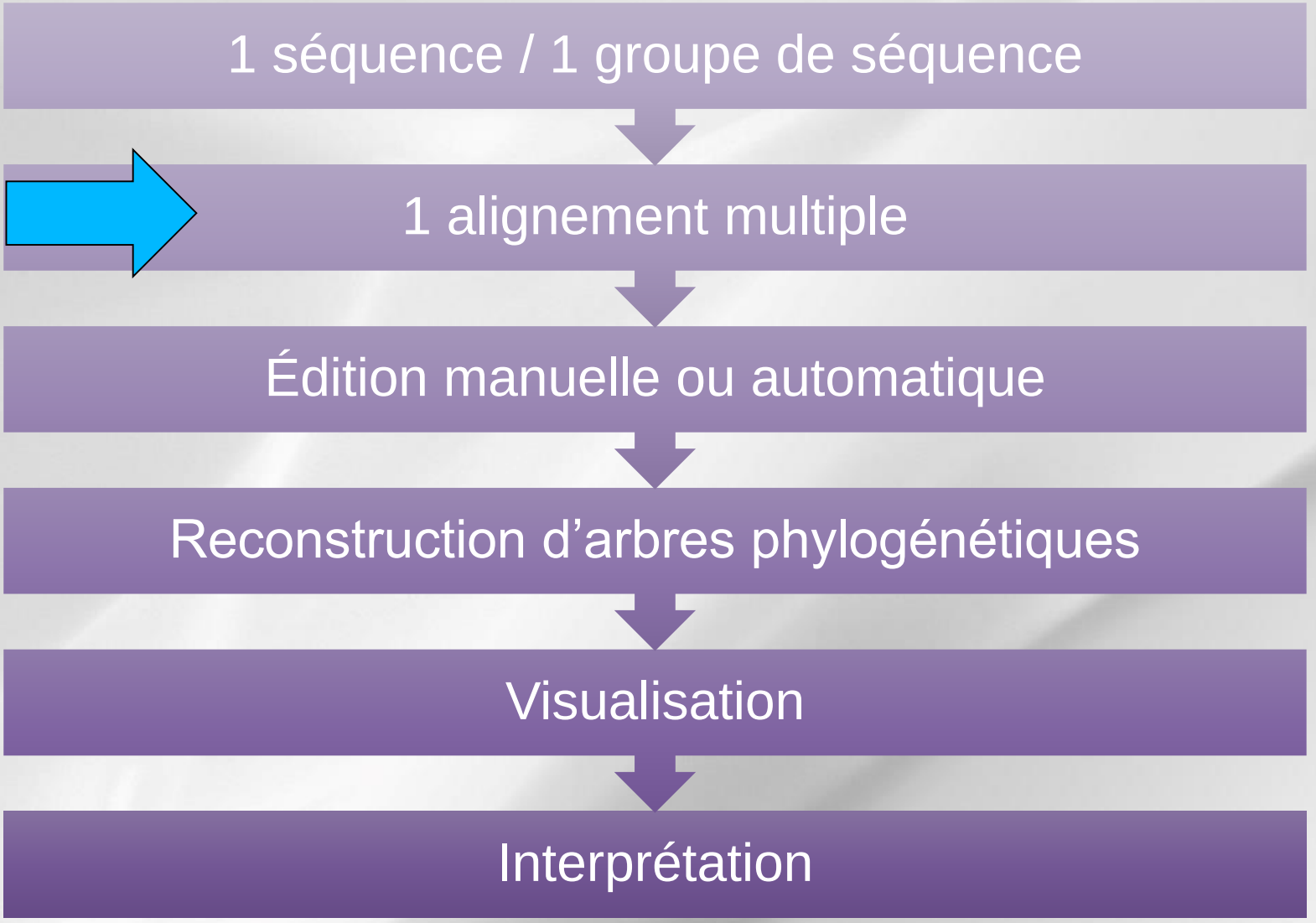
Constitution de son jeu de données

Suggestion:
30 % identité, 60% longueur (Subject et Query) and e-value < 1.e-3

Plus les espèces avec lesquelles le jeu de données sera constitué sont proches, plus il faut être sévère/strict sur les paramètres.
ex: 80% d'identité et 90% de la longueur.

Exercice 1

Les différentes étapes



Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

L’alignement par paires

Principe: comparaison de séquences

L’alignement des séquences est la principale méthode de comparaison.

Elle permet d’identifier des régions conservées.

On en déduit l'homologie.

D’autres méthodes existent:

Analyse statistique des «mots» contenus dans la séquence

Recherche de domaines ou motifs communs

L’alignement par paires

Comparer des séquences serait relativement simple si elles avaient toutes la même longueur.

Comme ce n'est pas le cas, il faut les aligner, c’est à dire trouver où se trouvent les insertions et délétions, représentées par des «indels» («gaps»)

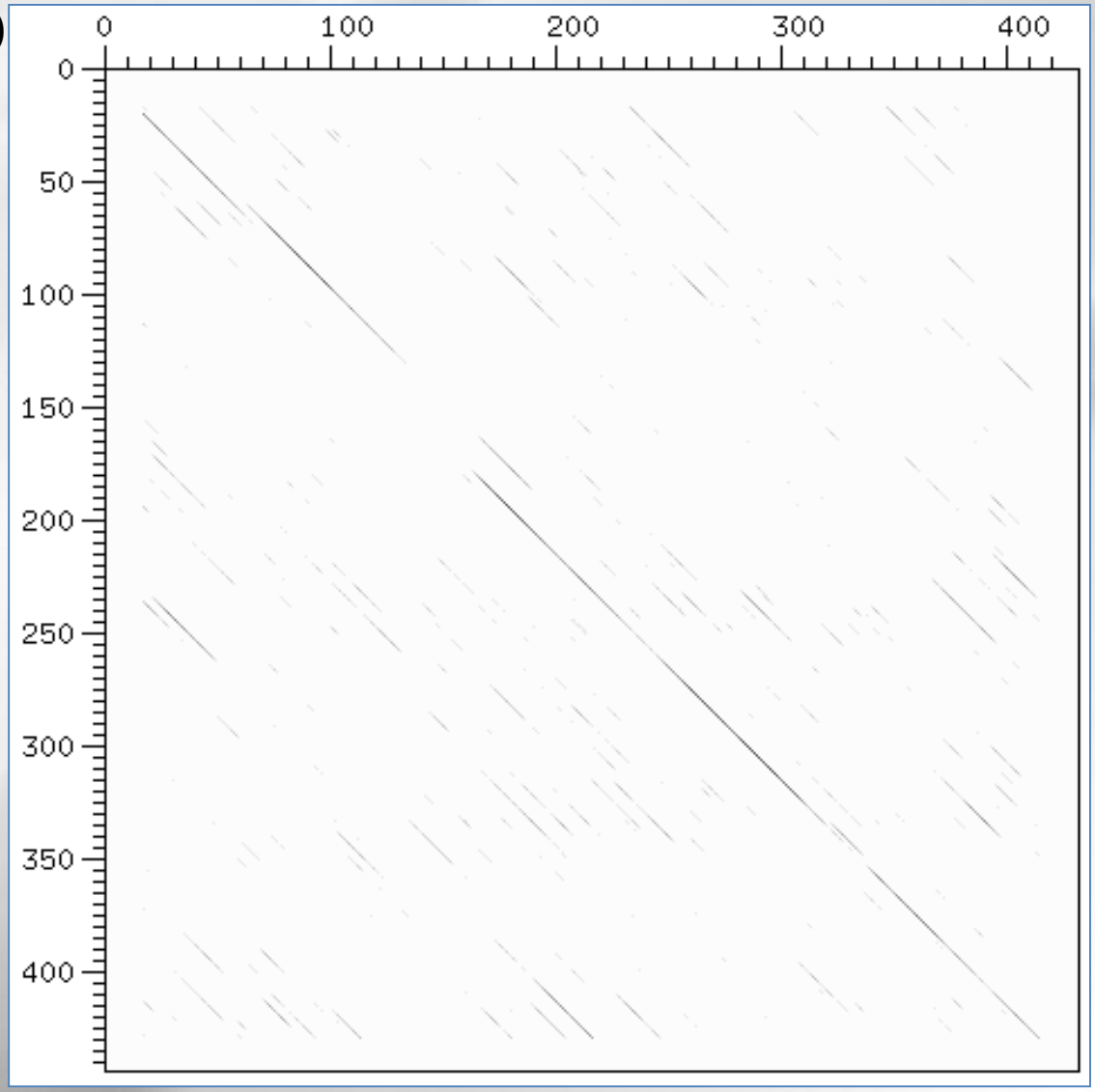
Les «dot plots»

Deux séquences (ici 2 gènes de globine)

une horizontalement,
l'autre verticalement

un point dans la matrice =
les deux positions sont identiques
ou n bases identiques dans une fenêtre
de N bases.

Lorsque des régions se ressemblent
apparition de diagonales

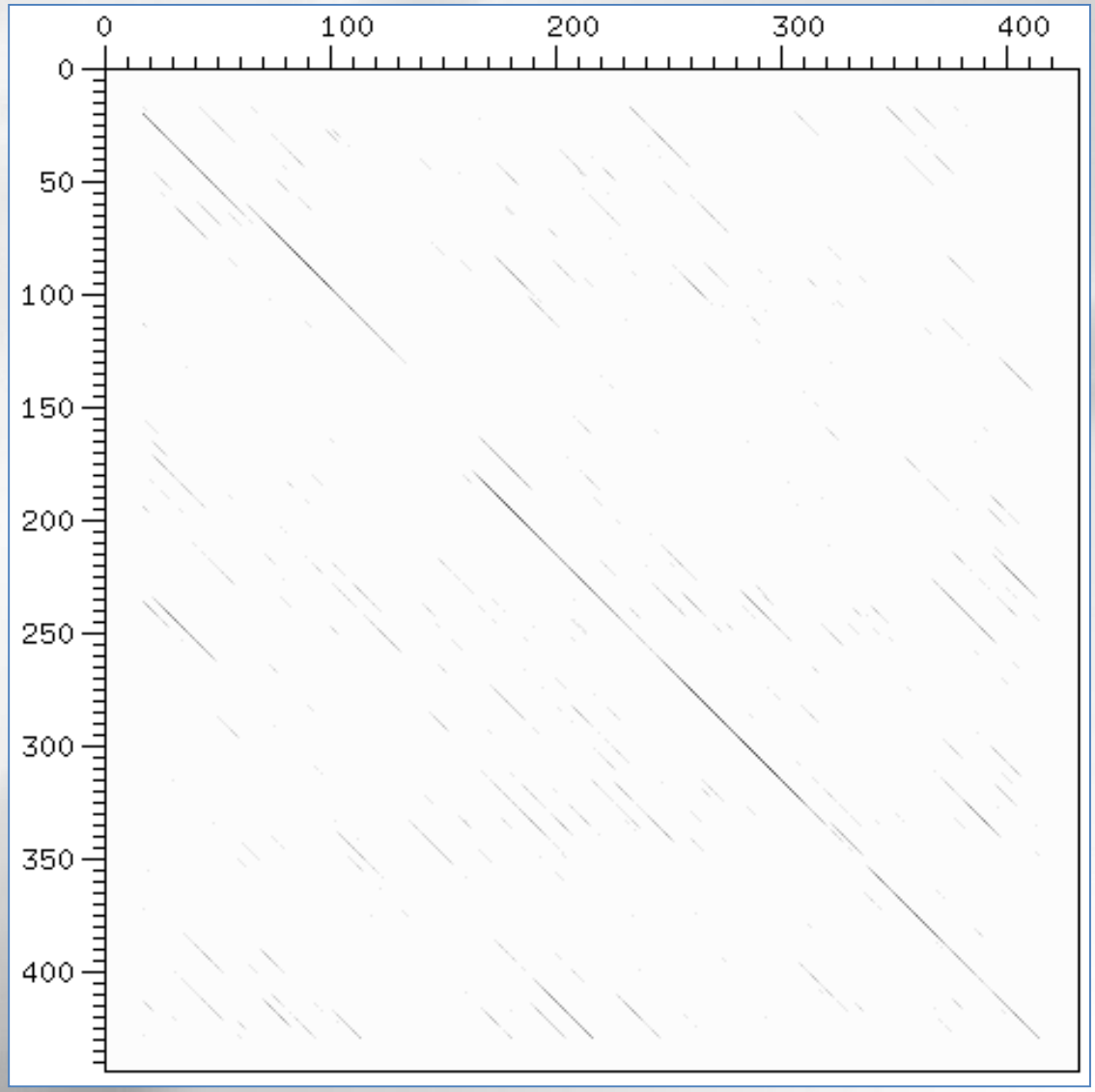


Les «dot plots»

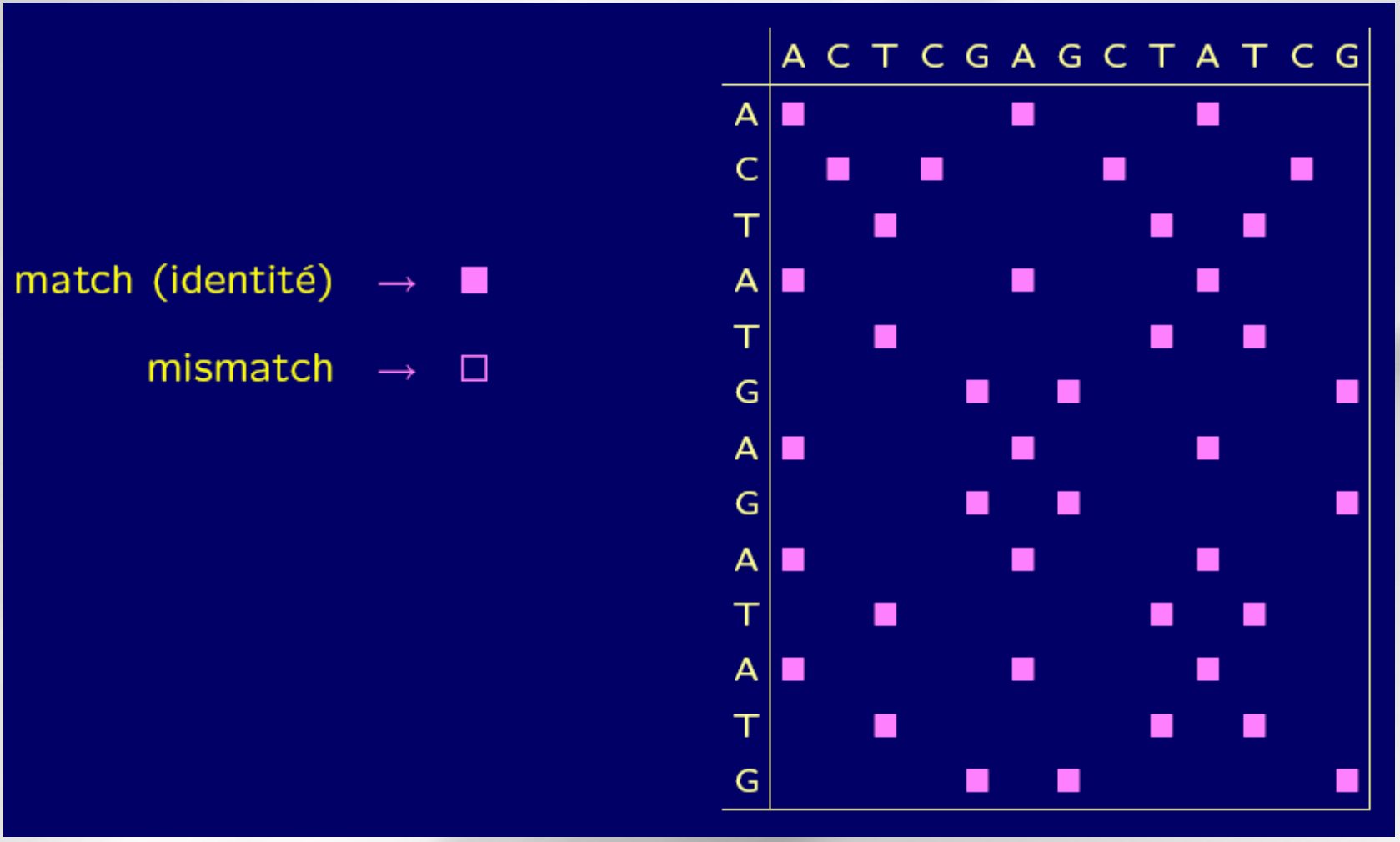
décalages entre les diagonales = insertions ou délétions

Plusieurs diagonales parallèles = une répétition

Les dot plots sur des génomes complets permettent de visualiser les évènements à grande échelle, la synténie, etc.



Les «dot plots»

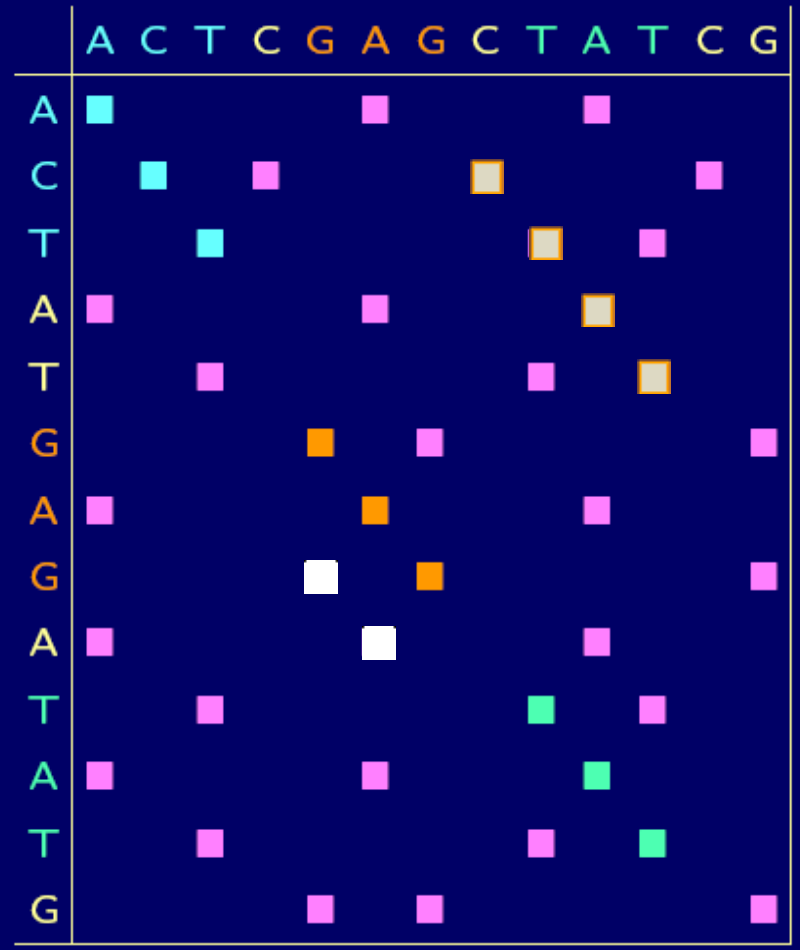


Les «dot plots»

match (identité) → ■

mismatch → □

diagonale = région similaire

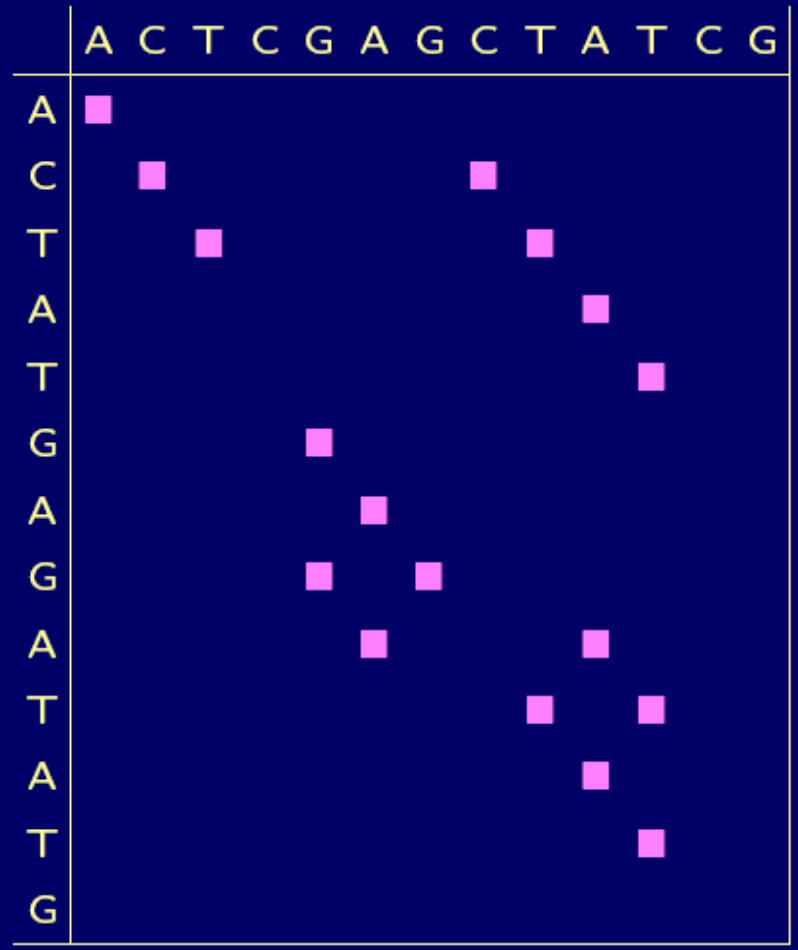


Les «dot plots»

match (identité) → ■

mismatch → □

Fenêtre de taille 2

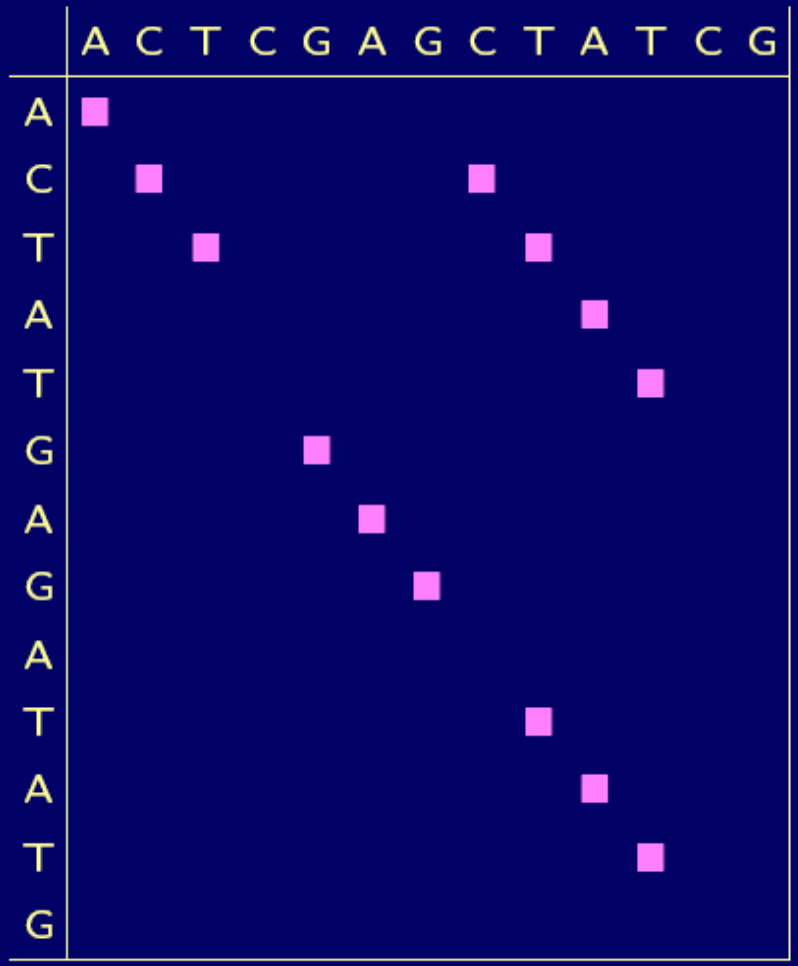


Les «dot plots»

match (identité) → ■

mismatch → □

Fenêtre de taille 3

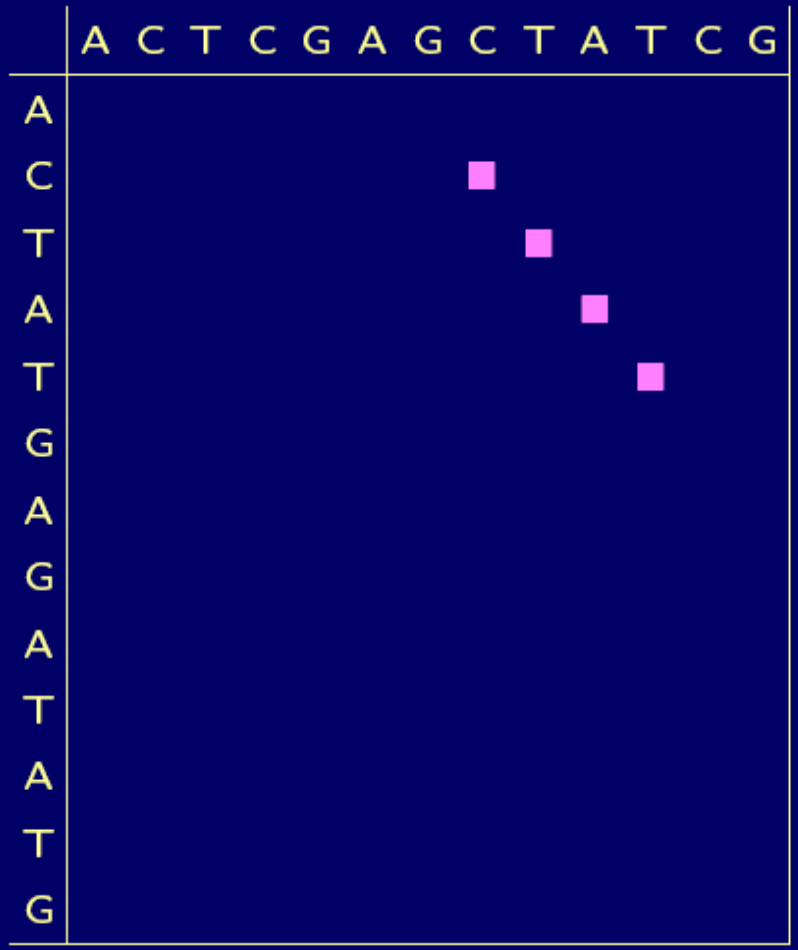


Les «dot plots»

match (identité) → ■

mismatch → □

Fenêtre de taille 4

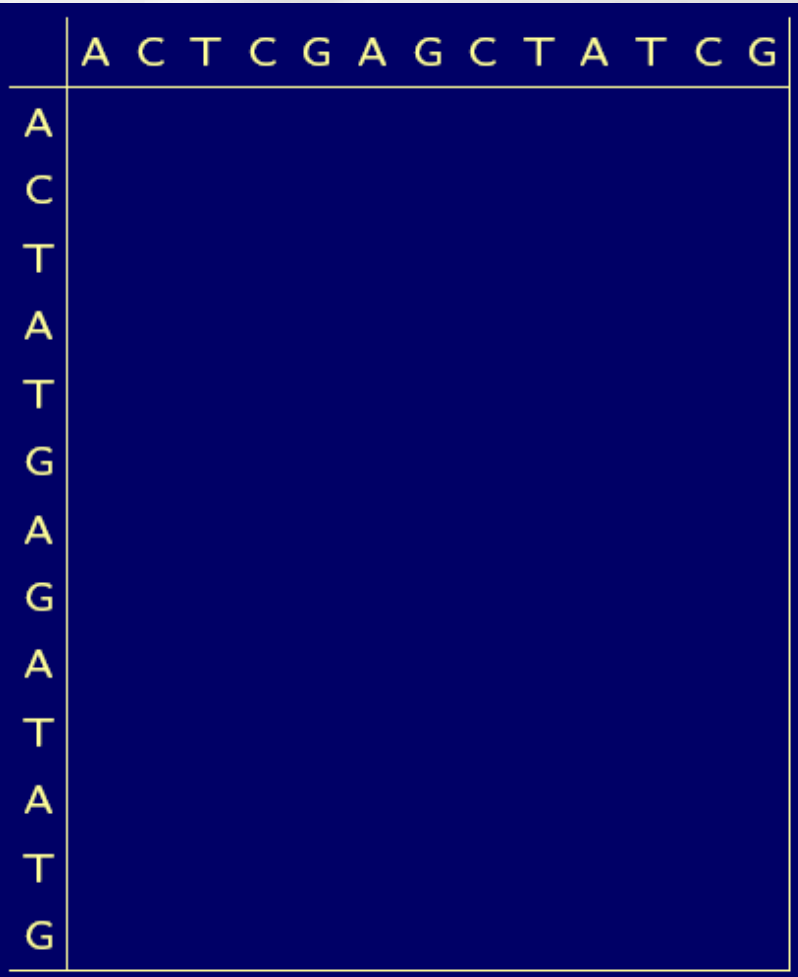


Les «dot plots»

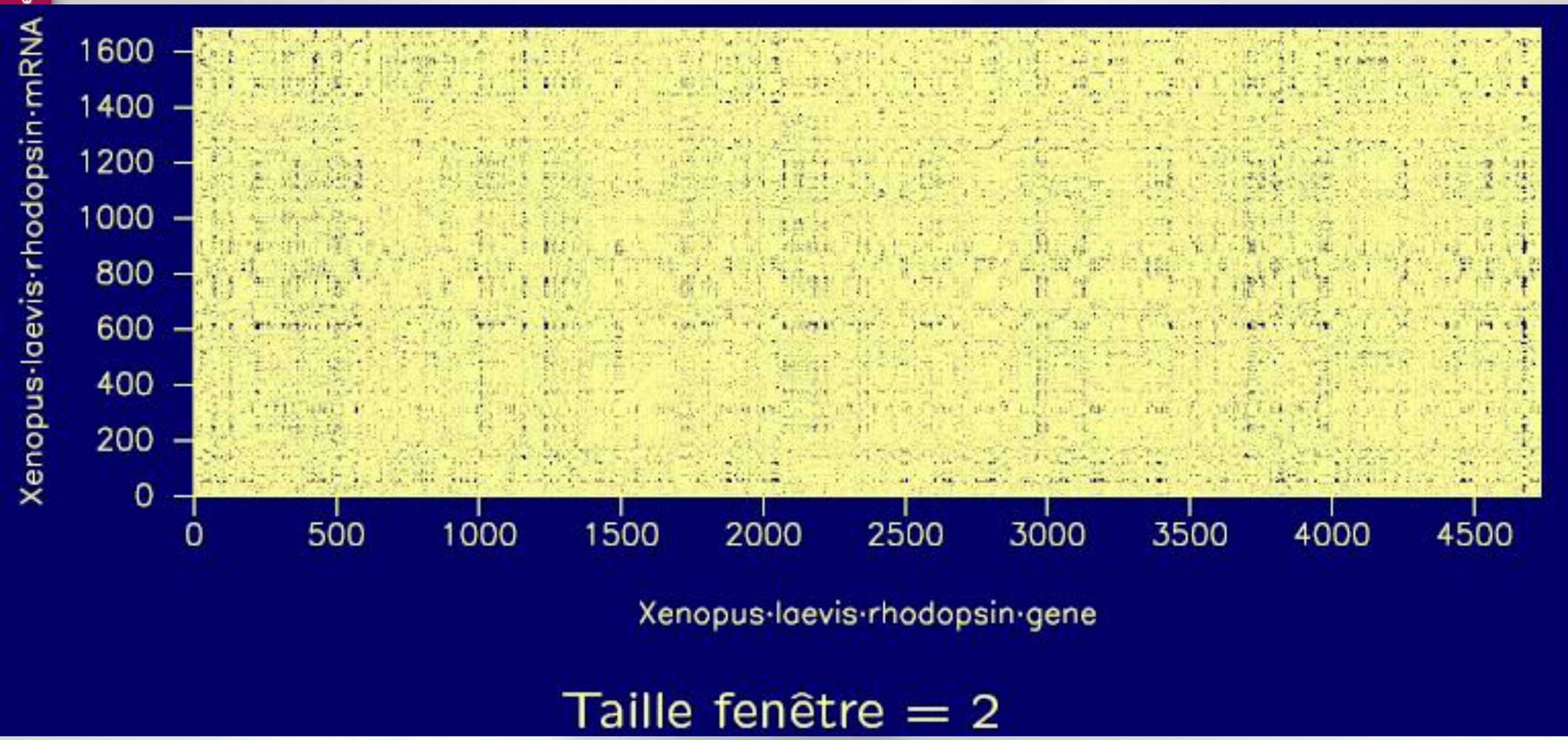
match (identité) → ■

mismatch → □

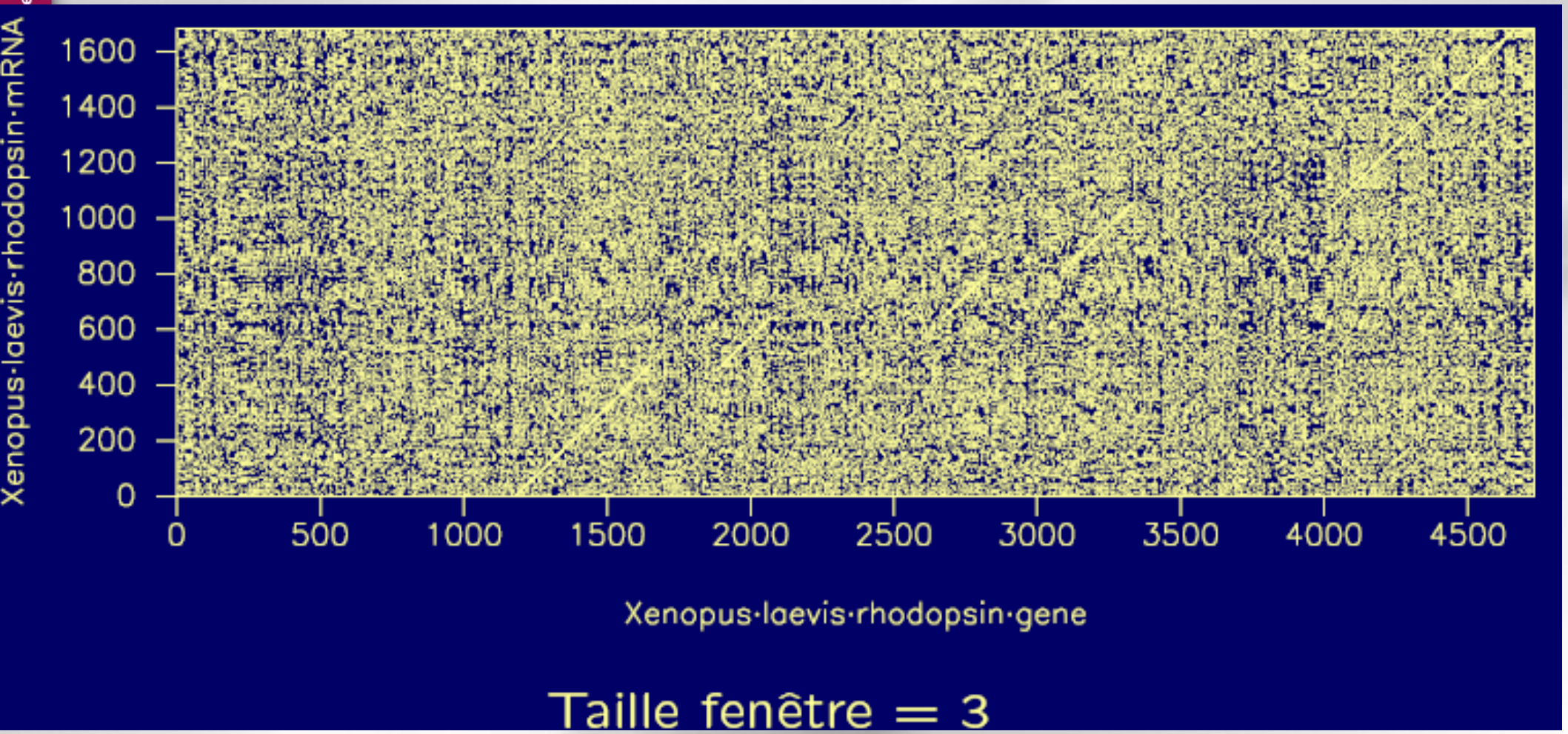
Fenêtre de taille 5



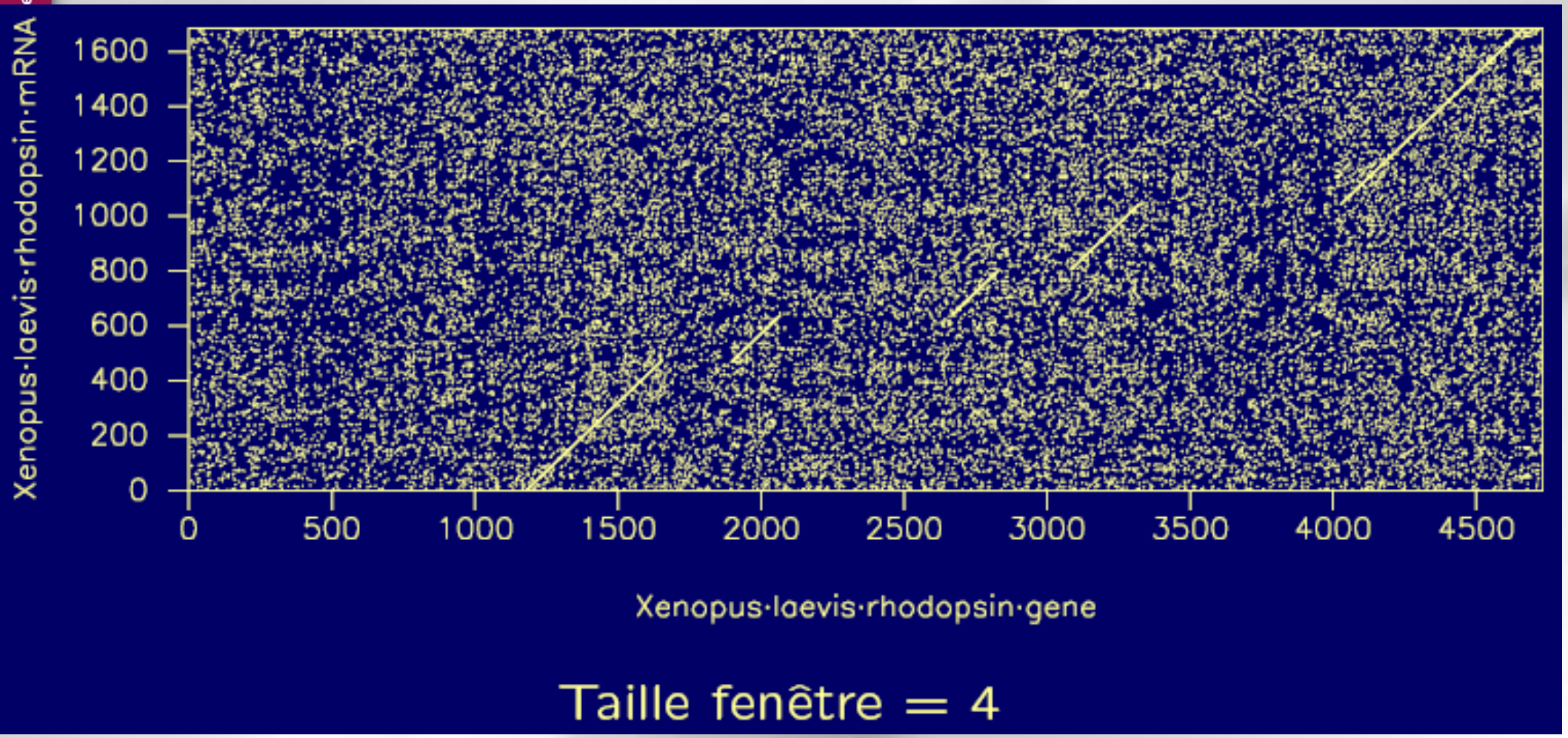
Les «dot plots»



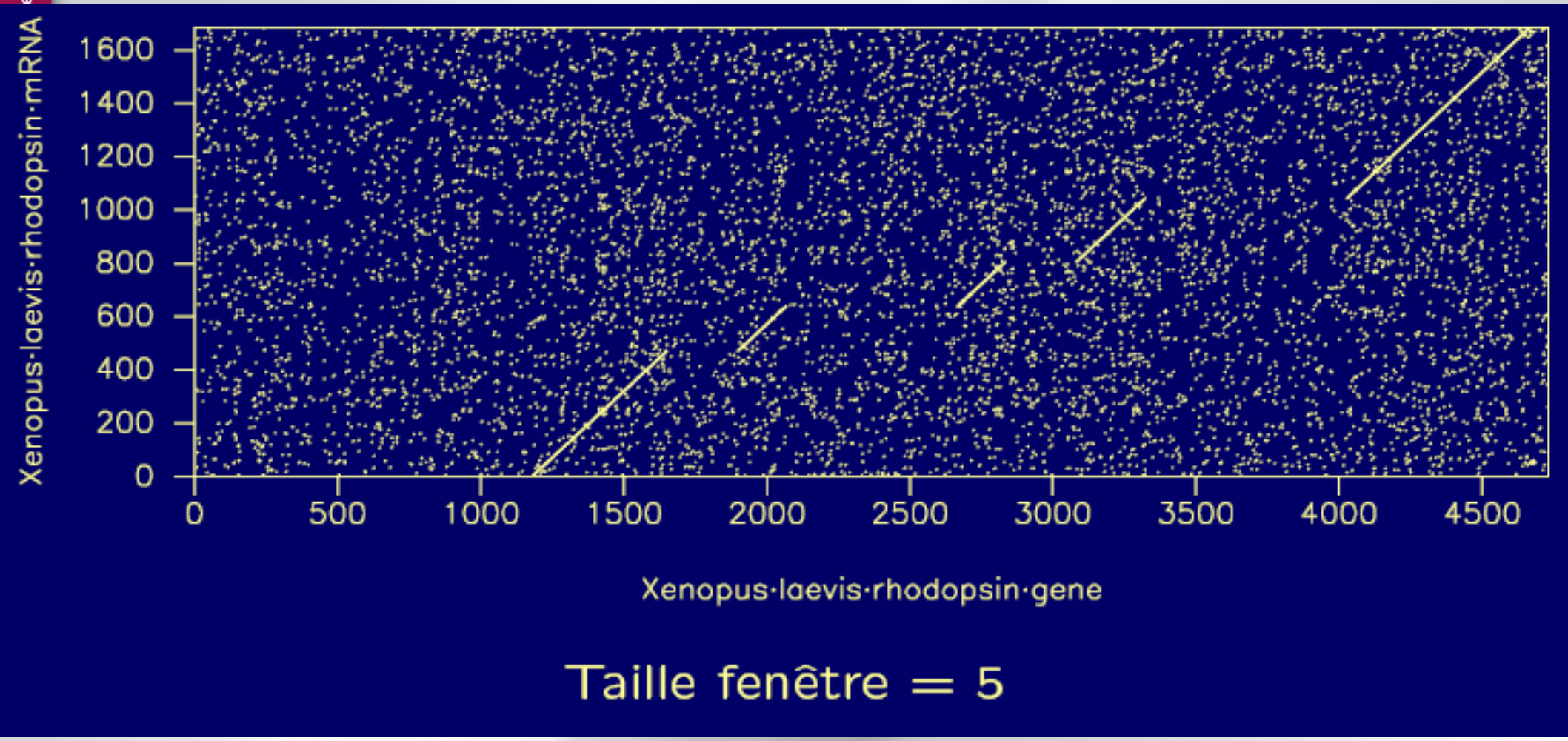
Les «dot plots»



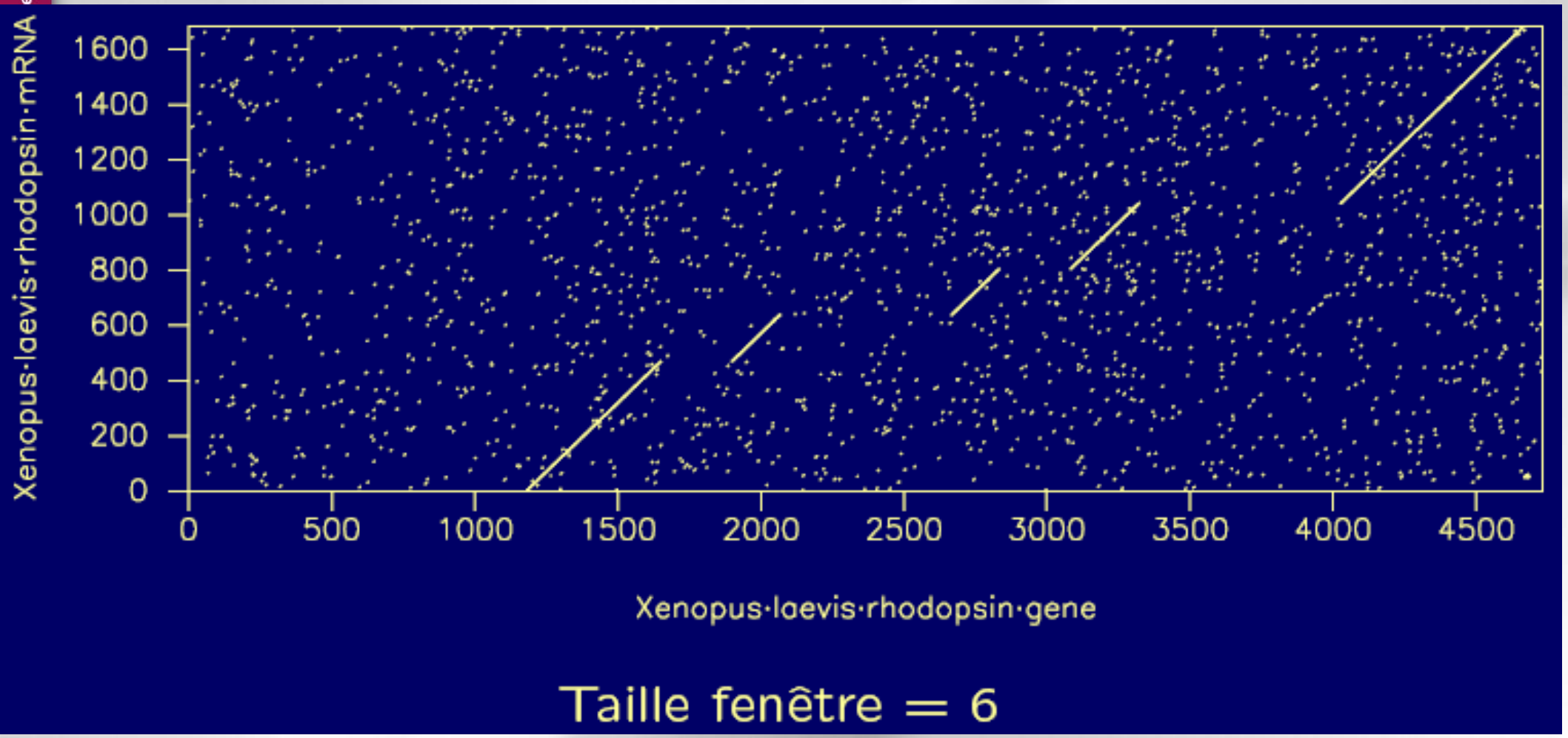
Les «dot plots»



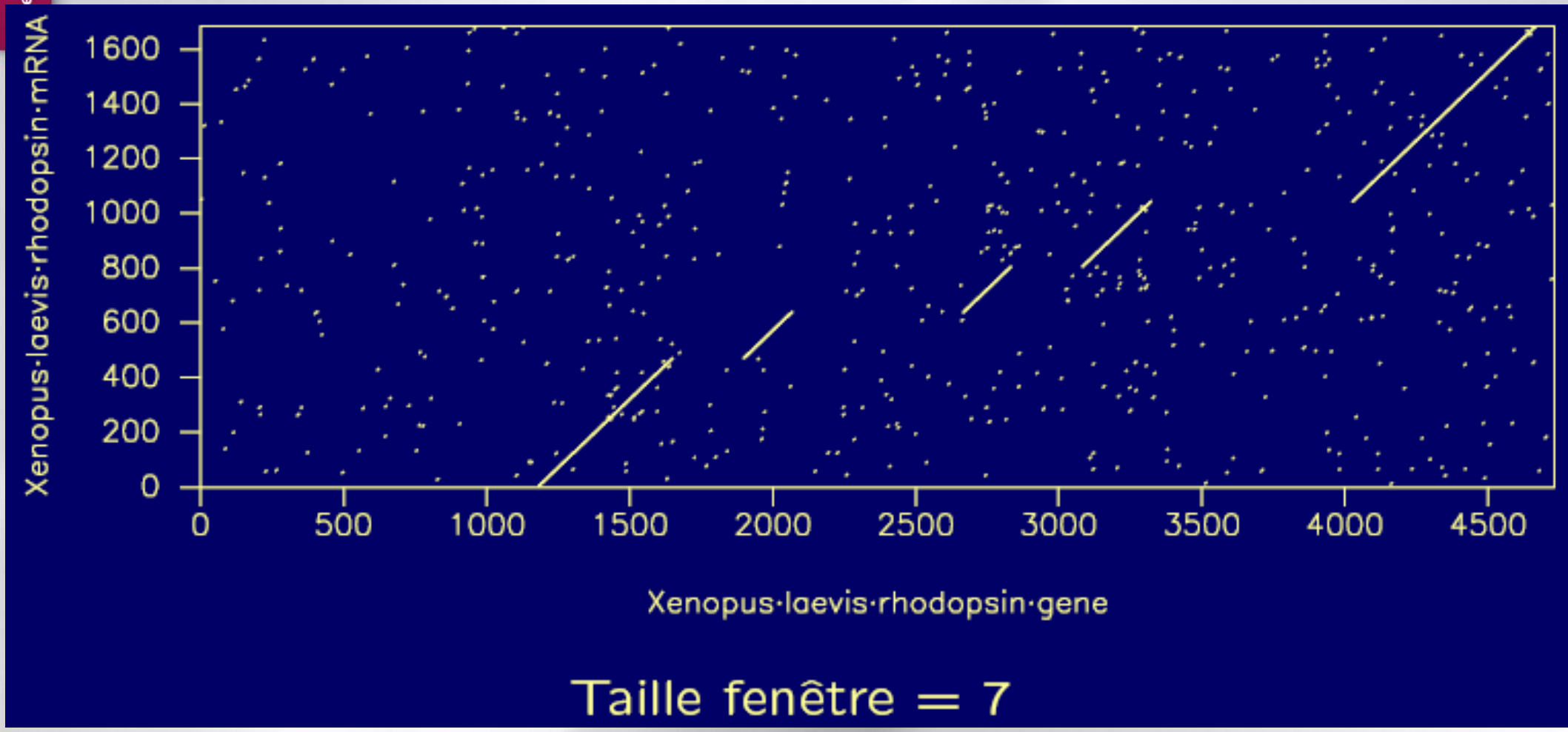
Les «dot plots»



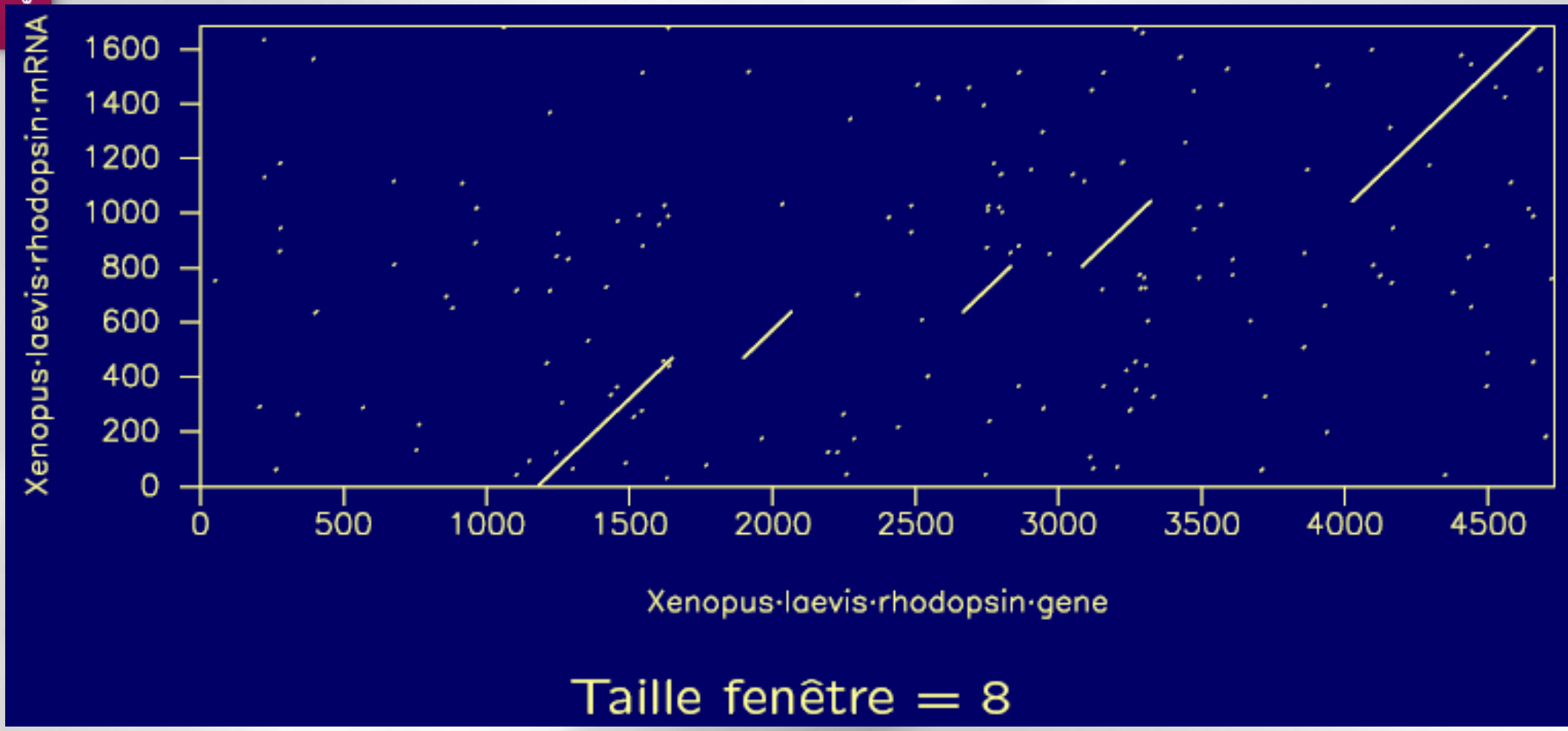
Les «dot plots»



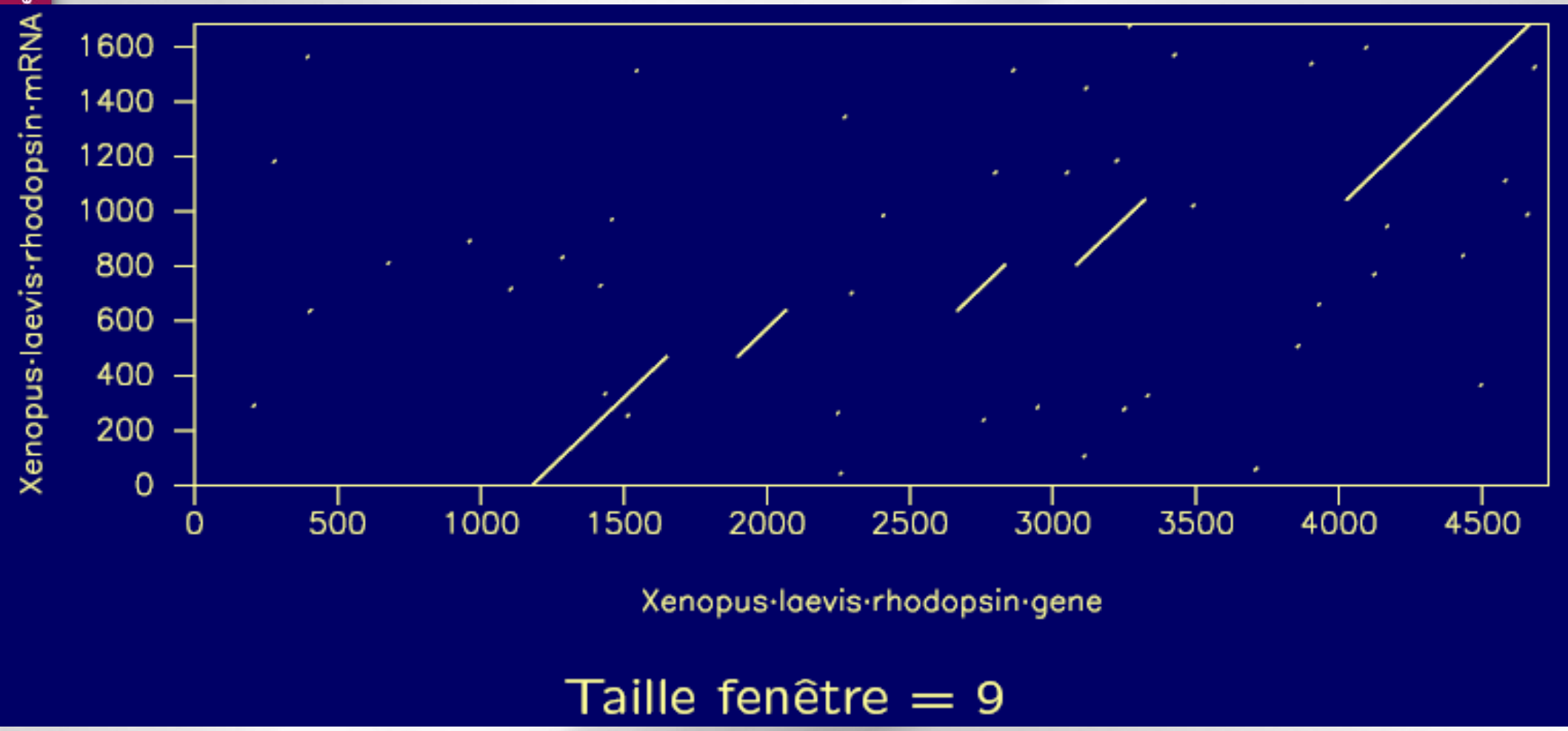
Les «dot plots»



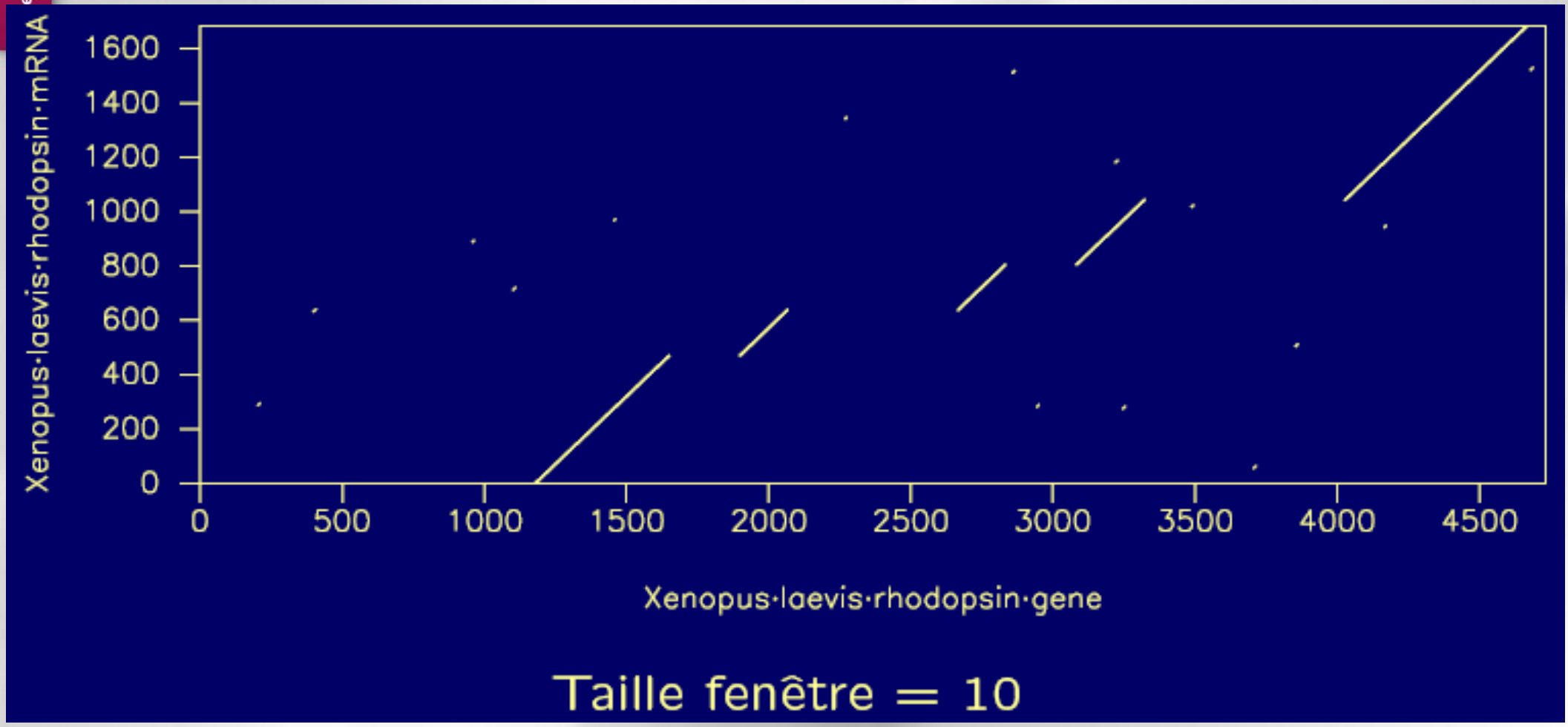
Les «dot plots»



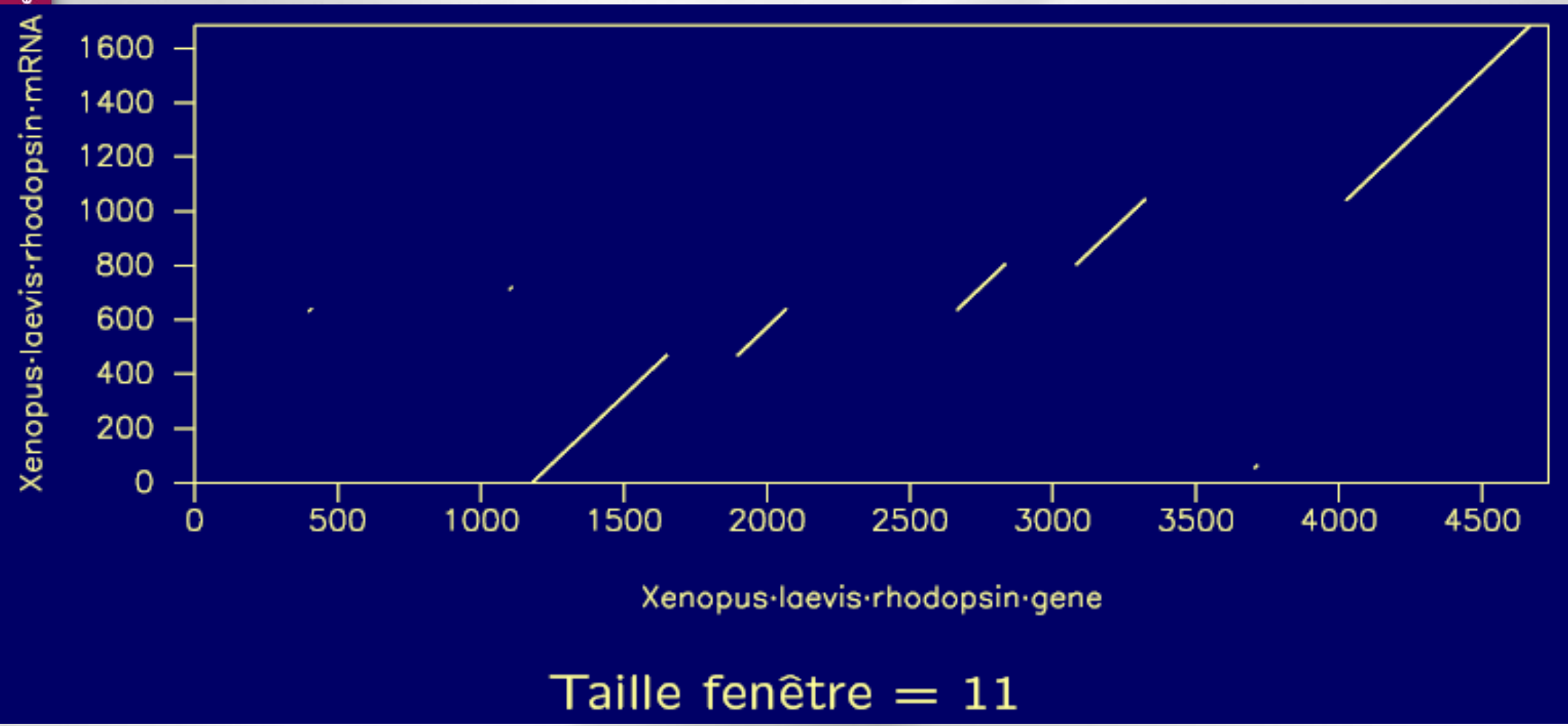
Les «dot plots»



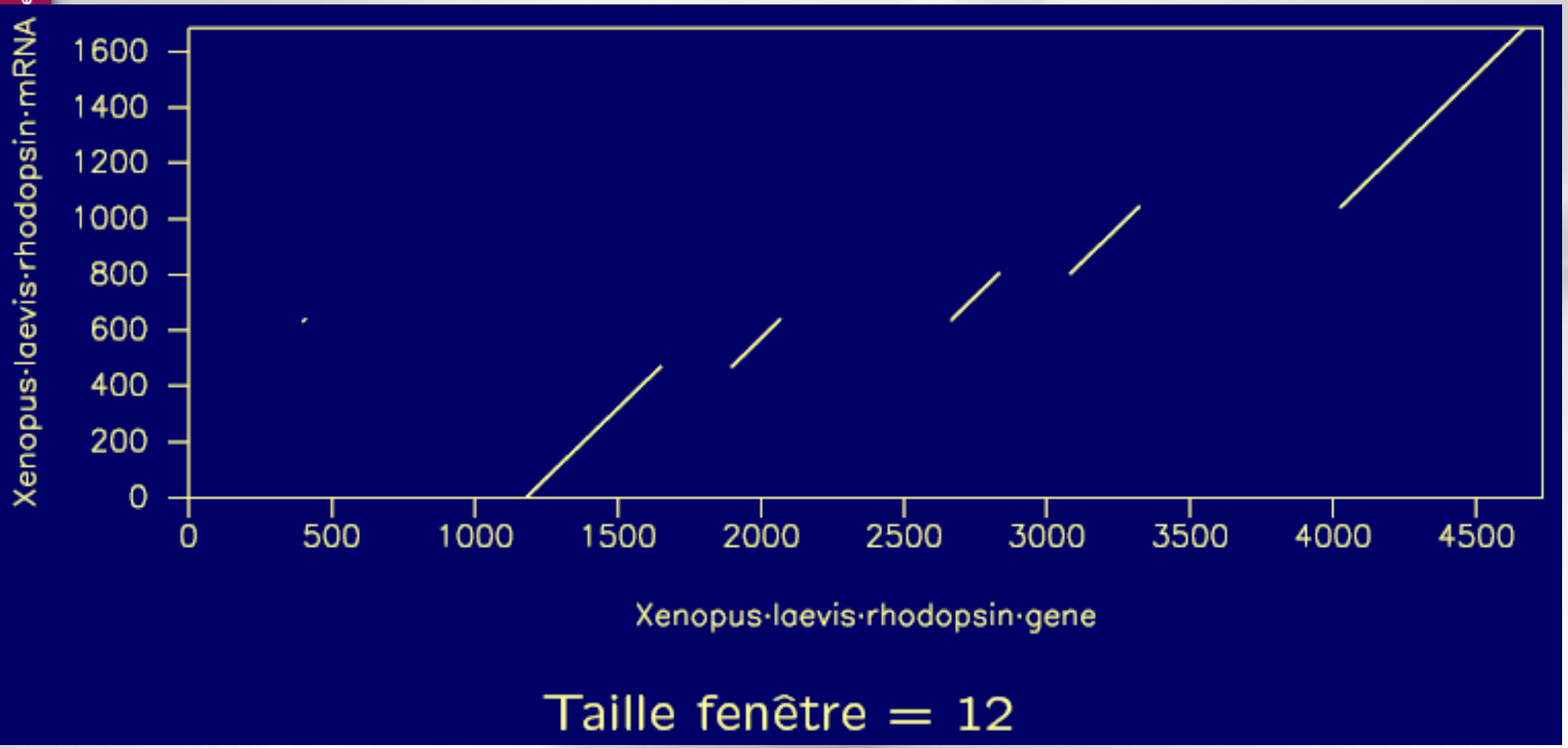
Les «dot plots»



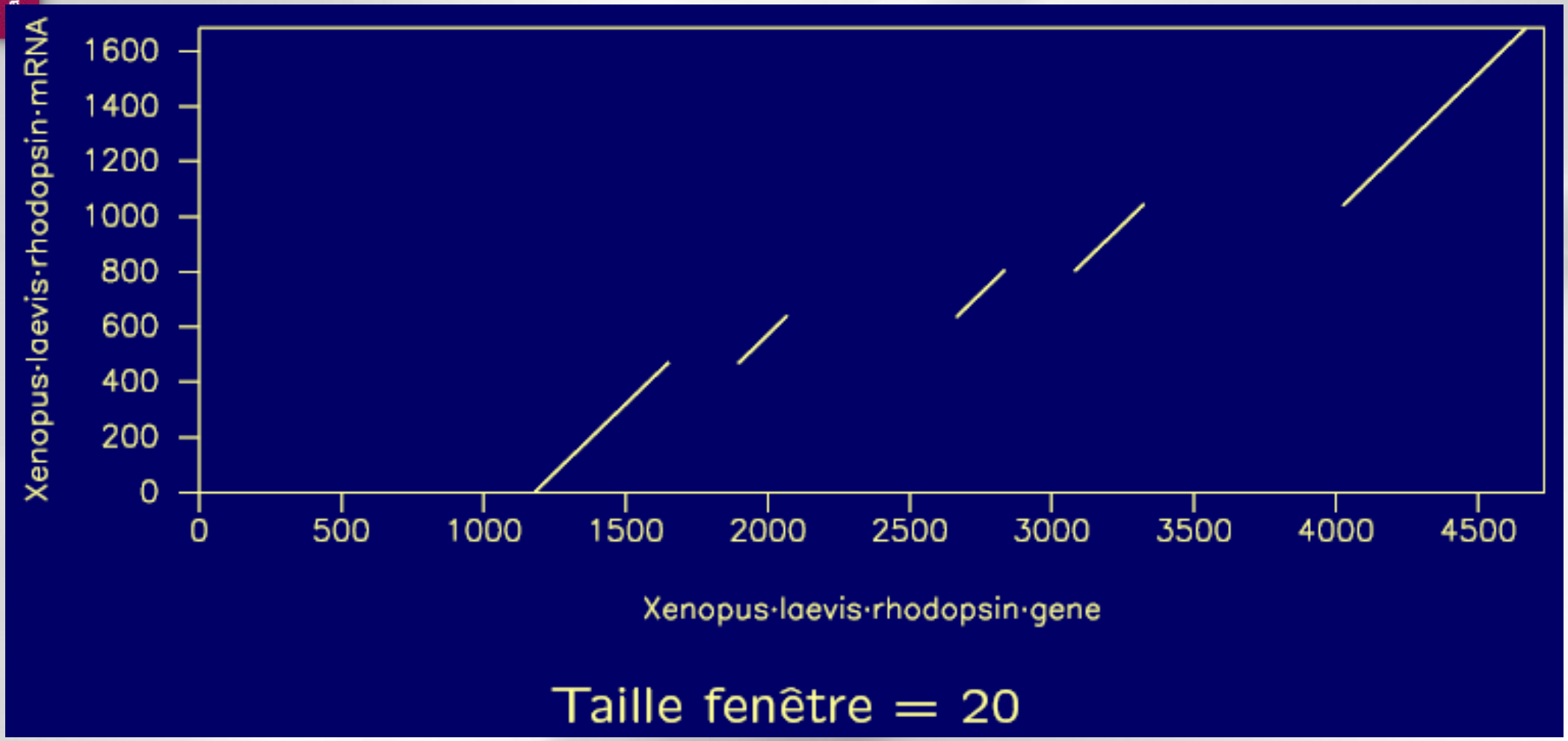
Les «dot plots»



Les «dot plots»

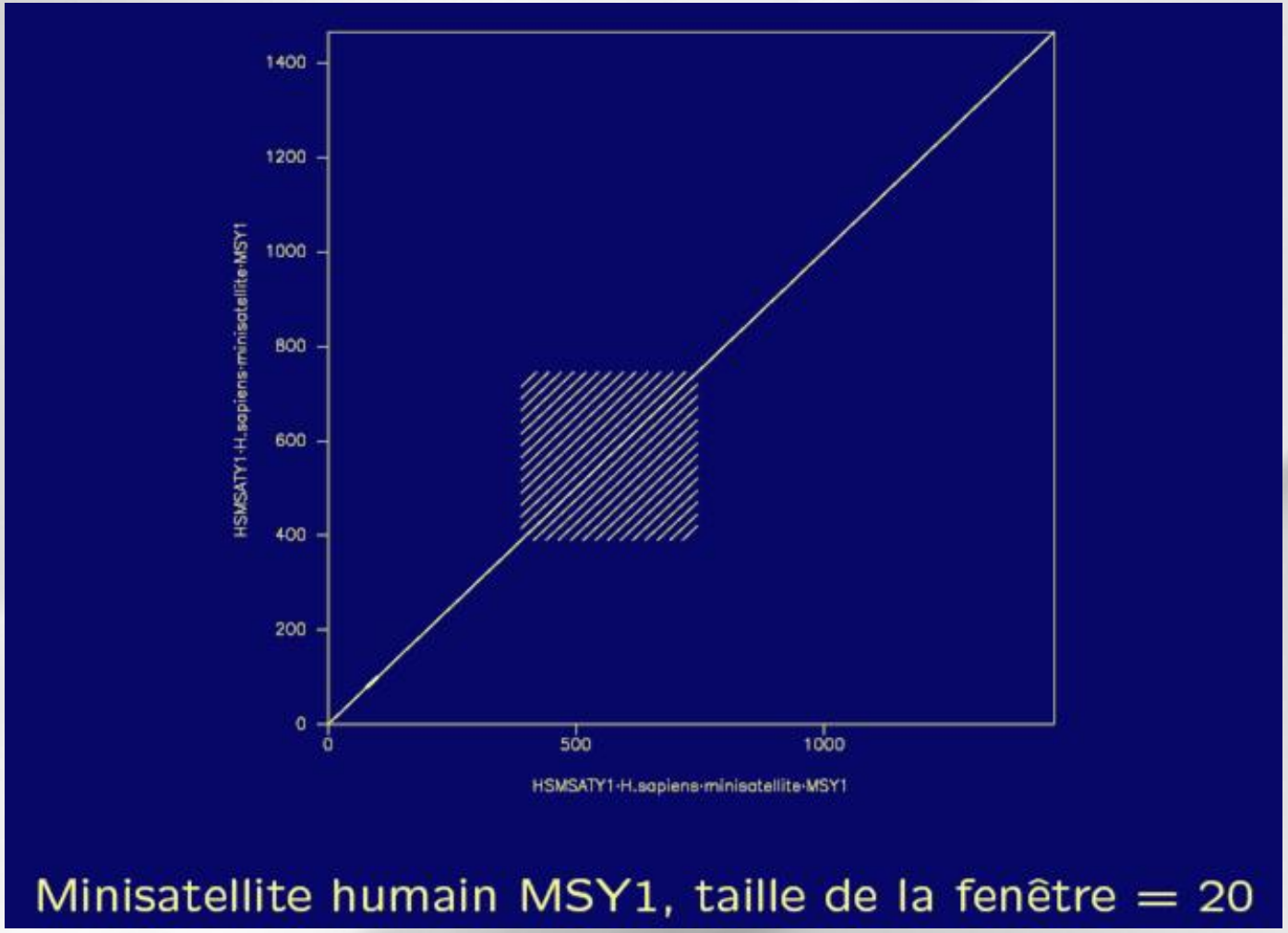


Les «dot plots»



Alignement = trouver le meilleur chemin dans ce graphe

Les «dot plots» exemple de visualisation de répétitions



Minisatellite humain MSY1, taille de la fenêtre = 20

Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

Alignement de séquences: le Score

Le bon alignement est celui qui minimise les opérations à réaliser pour passer d'une séquence à l'autre.

Alignement de séquences: le Score

Opérations:

conservation, remplacement/mutation, délétion, insertion.

Une pénalité peut être affectée à chaque opération,

par exemple c=0, m=1, d=2, i=2. La distance finale entre les deux séquences (distance d'édition) est la somme de ces pénalités.

	Seq 1	CAGTGGT-GC
	Seq 2	CA-TCGTAGC
		c=0, m=1, d=2, i=2.
Ou,	distance	ccicmccdcc = 0+0+2+0+1+0+0+2+0+0 = 5
variante:	ressemblance	ccicmccdcc = 2+2-1+2-1+2+2-1+2+2 = 11
		c=2, m=-1, d=-1, i=-1.

Une délétion à l'intérieur d'une séquence est considérée comme une insertion dans la séquence lui faisant face.

Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

Matrices de Substitution - Nucléique

Matrice 4X4 (nucléotides) ou 20x20 (acides aminés) décrivant la distance ou la similitude entre résidus.

Estiment le coût ou le taux de remplacement d'un résidu par un autre (distance).

Le choix d'une matrice affecte fortement le résultat de l'analyse.

Chaque matrice de score représente implicitement une théorie évolutive donnée

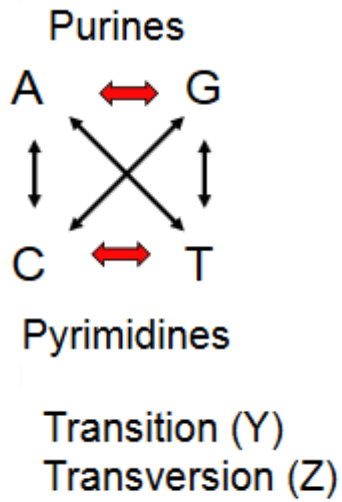
Matrices DNA

	A	C	G	T
A	2	-1	-1	-1
C	-1	2	-1	-1
G	-1	-1	2	-1
T	-1	-1	-1	2

Matrice identité

	A	C	G	T
A	3	-1	1	-1
C	-1	3	-1	1
G	1	-1	3	-1
T	-1	1	-1	3

Matrice transition/transversion



Matrices de Substitution - Protéique

400 changements possibles (20x20) pour les acides aminés mais non équivalents

1- Matrices fondées sur le code génétique

Les scores sont déterminés en fonction du nombre commun de nucléotides présents dans les codons des acides aminés, ce qui revient à considérer le minimum de changements nécessaires en bases pour convertir un acide aminé en un autre.

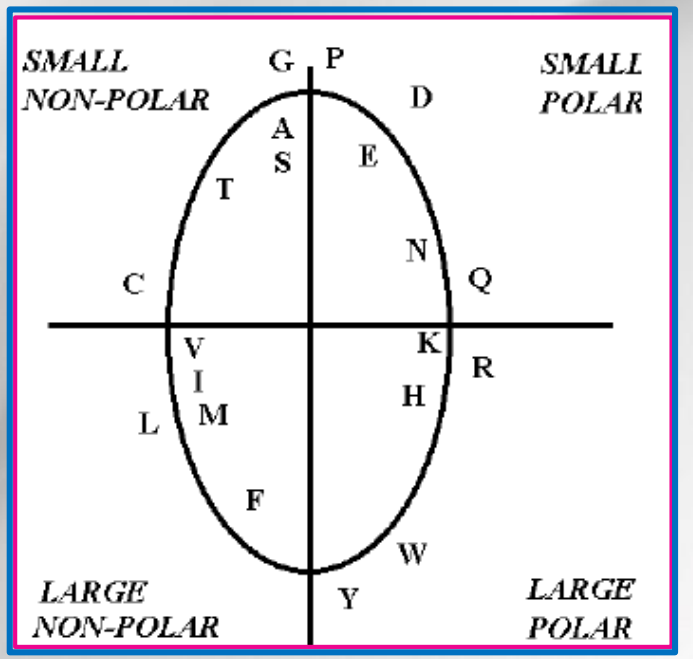
Matrices de Substitution - Protéique

2- Matrices fondées sur les propriétés physicochimiques

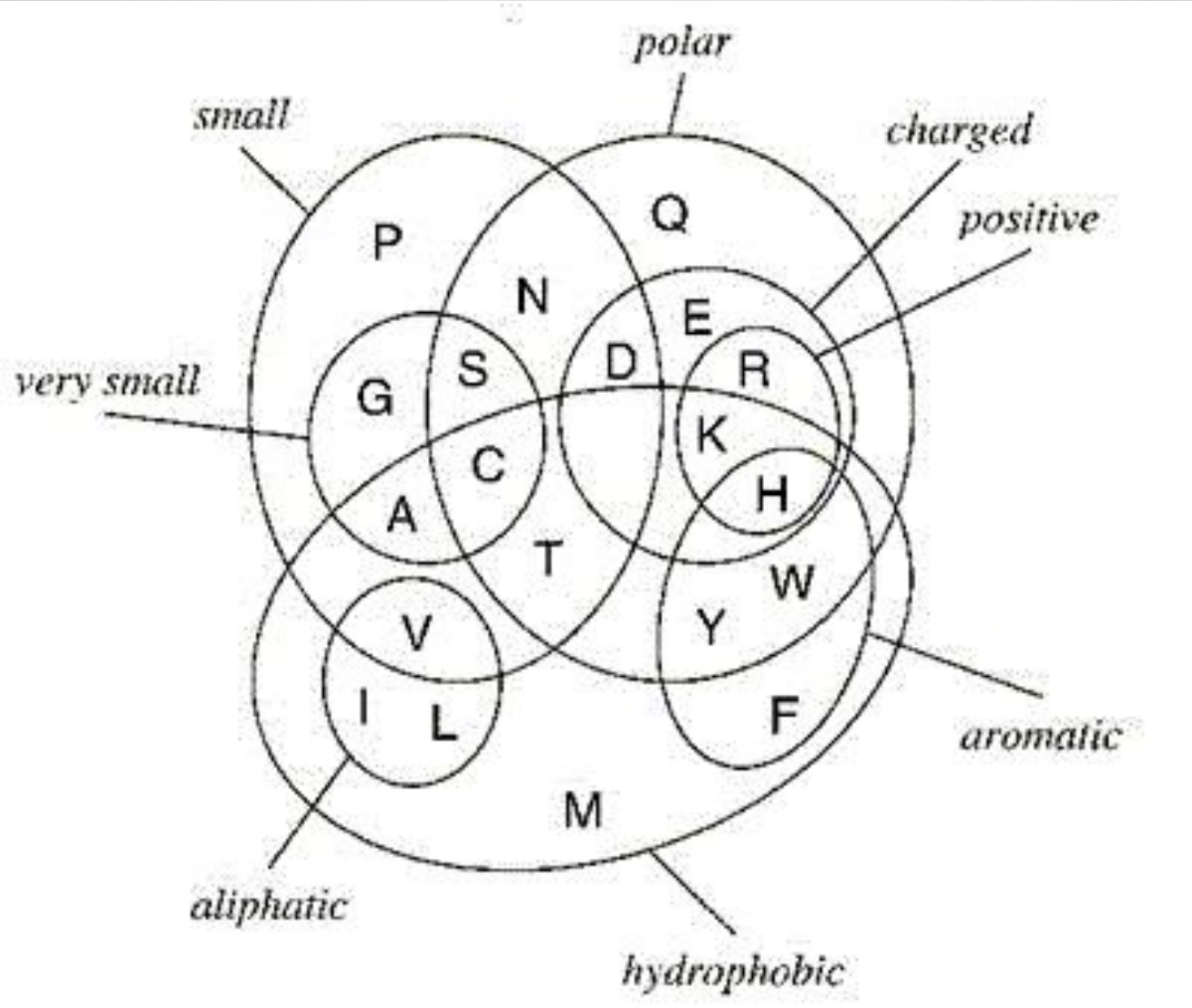
Les plus courantes sont celles basées sur le caractère hydrophile ou hydrophobe des protéines.

Matrices peu utilisées.

3- Matrices fondées sur l'évolution (ex: Dayhoff)



Une représentation bidimensionnelle des propriétés des aa calculée d'après la matrice de Dayhoff par G. Vriend, Centre for Molecular and Biomolecular Informatics, University of Nijmegen



Taylor (1986) The Classification of Amino Acid Conservation, *J. Theor. Biol. 119, 205-218*

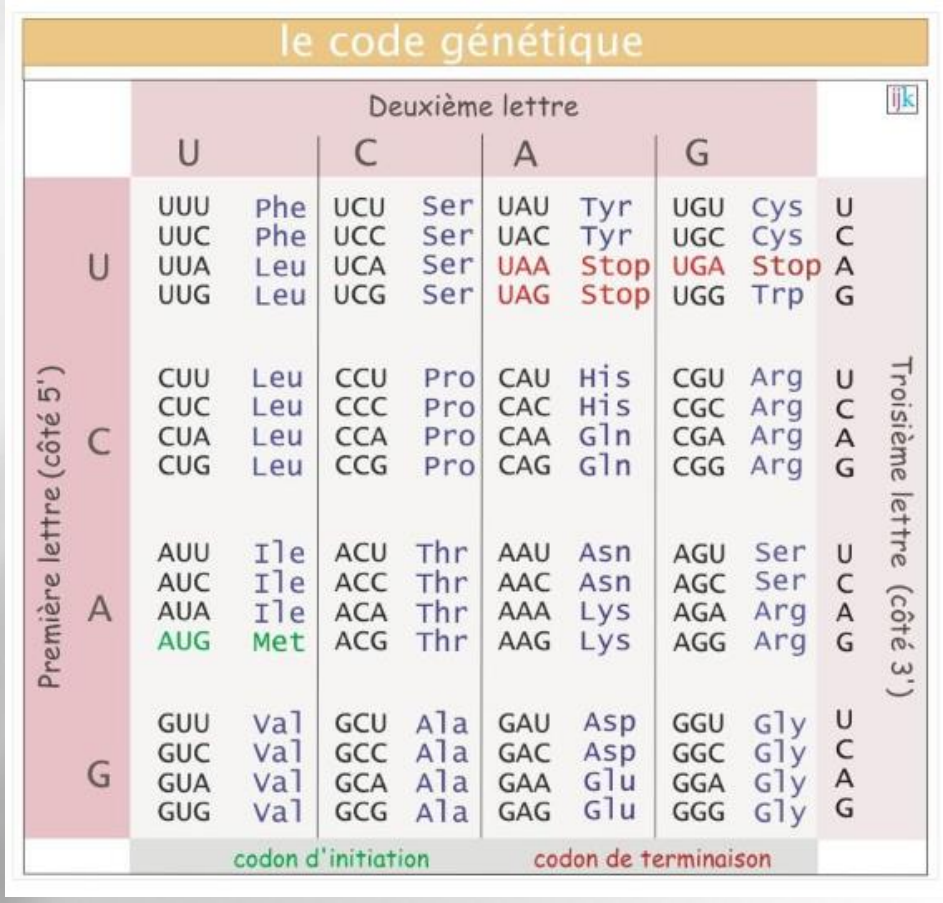
LES MATRICES PROTÉIQUES LIÉES AU CODE GÉNÉTIQUE

Basée sur le code génétique:

une substitution d'un acide aminé en un autre se produit d'autant plus rarement que cela nécessite un plus grand nombre de mutations au niveau de l'ADN.

Matrice génétique (Fitch, 1966)

identité: +3
1 mutation ADN = 2 nt identique: +2
2 mutation ADN = 1 nt identique: +1
3 mutation ADN = 0 nt identique: +0



LES MATRICES PROTÉIQUES LIÉES AU CODE GÉNÉTIQUE

Nombre de mutations nécessaires pour passer du codon d'un acide aminé au codon d'un autre acide aminé

Mutation GLU ⟶ LYS

d 'où

GAA ⟶ **AAA**
GAG ⟶ **AAG**

↓

1 mutation sur la première base du codon = +1

geno toul bioinfo

Asymmetric step matrix for non-synonymous nucleotide substitutions based on vertebrate mitochondrial genetic code.

	Amino acid	PHE	LEU	LEU		ILE	MET	VAL		SER		PRO		THR		ALA		TYR	HIS	GLN	ASN	LYS	ASP	GLU	CYS	TRP	ARG		SER	GLY	
	Codon	UUY	UUR	CUY	CUR	AUY	AUR	GUY	GUR	UCY	UCR	CCY	CCR	ACY	ACR	GCY	GCR	UAY	CAY	CAR	AAY	AAR	GAY	GAR	UGY	UGR	CGY	CGR	AGY	GGY	GGR
	Symbol	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	1	2	3	4
Amino acid	Codon																														
PHE	UUY	0	1	1	1	1	2	1	1	1	1	2	2	2	2	2	2	1	2	3	2	3	2	3	1	2	2	2	2	2	2
LEU	UUR	1	0	1	1	2	1	1	1	1	1	2	2	2	2	2	2	2	3	2	3	2	3	2	2	1	2	2	2	2	2
LEU	CUY	1	2	0	0	1	2	1	1	2	2	1	1	2	2	2	2	2	1	2	2	3	2	3	2	3	1	1	2	2	2
	CUR	2	1	0	0	2	1	1	1	2	2	1	1	2	2	2	2	3	2	1	3	2	3	2	3	2	1	1	3	2	2
ILE	AUY	1	2	1	1	0	1	1	1	2	2	2	2	1	1	2	2	2	2	3	1	2	2	3	2	3	2	2	1	2	2
MET	AUR	2	1	1	1	2	0	2	2	2	2	2	2	1	1	2	2	3	3	2	2	1	3	2	3	2	2	2	2	2	2
VLA	GUY	1	2	1	1	1	2	0	0	2	2	2	2	2	2	1	1	2	2	3	2	3	1	2	2	3	2	2	2	1	1
	GUR	2	1	1	1	2	1	0	0	2	2	2	2	2	2	1	1	3	3	2	3	2	2	2	3	2	2	2	3	1	1
SER	UCY	1	2	2	2	2	3	2	2	0	0	1	1	1	1	1	1	1	2	3	2	3	2	3	1	2	2	2	2	2	2
	UCR	2	1	2	2	3	2	2	2	0	0	1	1	1	1	1	1	2	3	2	3	2	3	2	2	1	2	2	3	2	2
PRO	CCY	2	3	1	1	2	3	2	2	1	1	0	0	1	1	1	1	2	1	2	2	3	2	3	2	3	1	1	2	2	2
	CCR	3	2	1	1	3	2	2	2	1	1	0	0	1	1	1	1	3	2	1	3	2	3	2	3	2	1	1	3	2	2
THR	ACY	2	3	2	2	1	2	2	2	1	1	1	1	0	0	1	1	2	2	3	1	2	2	3	2	3	2	2	1	2	2
	ACR	3	2	2	2	2	1	2	2	1	1	1	1	0	0	1	1	3	3	2	2	1	3	2	3	2	2	2	2	2	2
ALA	GCY	2	3	2	2	2	3	2	2	1	1	1	1	1	1	0	0	2	2	3	2	3	1	2	2	3	2	2	2	1	1
	GCR	3	2	2	2	3	2	1	1	1	1	1	1	1	1	0	0	3	3	2	3	2	2	1	3	2	2	2	3	1	1
TYR	UAY	1	2	2	2	2	3	2	2	1	1	2	2	2	2	2	2	0	1	2	1	2	1	2	1	2	2	2	2	2	2
HIS	CAY	2	3	1	1	2	3	2	2	2	2	1	1	2	2	2	2	1	0	1	1	2	1	2	2	3	1	1	2	2	2
GLN	CAR	3	2	1	1	3	2	2	2	2	2	1	1	2	2	2	2	2	1	0	2	1	2	1	3	2	1	1	3	2	2
ASN	AAY	2	3	2	2	1	2	2	2	2	2	2	2	1	1	2	2	1	1	2	0	1	1	2	2	3	2	2	1	2	2
LYS	AAR	3	2	2	2	2	1	2	2	2	2	2	2	1	1	2	2	2	2	1	1	0	2	1	3	2	2	2	2	2	2
ASP	GAY	2	3	2	2	2	3	1	1	2	2	2	2	2	2	1	1	1	1	2	1	2	0	1	2	3	2	3	2	1	1
GLU	GAR	3	2	2	2	3	2	1	1	2	2	2	2	2	2	1	1	2	2	1	2	1	1	0	3	2	2	2	3	2	2
CYS	UGY	1	2	2	2	2	3	2	2	1	1	2	2	2	2	2	2	1	2	3	2	3	2	3	0	1	1	1	1	1	1
TRP	UGR	2	1	2	2	3	2	2	2	1	1	2	2	2	2	2	2	2	3	2	3	2	3	2	1	0	1	1	2	1	1
ARG	CGY	2	3	1	1	2	3	2	2	2	2	1	1	2	2	2	2	2	1	2	2	3	2	3	1	2	0	0	1	1	1
	CGR	3	2	1	1	3	2	2	2	2	2	1	1	2	2	2	2	3	2	1	3	2	3	2	2	1	0	0	2	1	1
SER	AGY	2	3	2	2	1	2	2	2	2	2	2	2	1	1	2	2	2	2	3	1	2	2	3	1	2	1	1	0	1	1
GLY	GGY	2	3	2	2	2	3	1	1	2	2	2	2	2	2	1	1	2	2	3	2	3	1	3	1	2	1	1	1	0	0
	GGR	3	2	2	2	3	2	1	1	2	2	2	2	2	2	1	1	3	3	2	3	2	2	2	2	1	1	1	2	0	0

Les matrices protéiques liées aux caractéristiques physico-chimiques au code génétique

Distance basée sur les propriétés des acides aminés :

- composition, polarité, volume moléculaire (Grantam, 1974)
- matrice d'hydrophobicité de Levitt (1976)
- matrice de structure secondaire (Levin, 1986)
- polarité, hydrophobicité, structure secondaire (Rao, 1987)

Les matrices protéiques fondées sur les fréquences de substitution des acides aminés au cours de l'évolution

Principe:

Les séquences homologues ont conservées des fonctions similaires

Deux acides aminés se ressembleront d'autant plus que la fréquence de substitution observée est grande puisque ces substitutions n'auront pas modifié la fonction de la protéine.

Il est possible d'estimer la fréquence avec laquelle un acide aminé est remplacé par un autre au cours de l'évolution à partir de séquences alignées.

Les matrices protéiques fondées sur les fréquences de substitution des acides aminés au cours de l'évolution

Principales approches:

Comparaison directe de séquences (alignement global): matrice PAM (Dayhoff, 1978)

Comparaison de domaines protéiques (régions les plus conservées au cours de l'évolution): matrice BLOSUM (Henikoff et Henikoff, 1992)

Alignement de séquences en comparant leurs structures secondaire et tertiaire

Matrices de Dayhoff ou PAM

Margaret Dayhoff, 1978

PAM = Percentage/Point of Accepted Mutation

Elle rend compte de deux processus

1. L'apparition de substitutions
2. Leur passage au travers le crible de la sélection

Matrices de Dayhoff ou PAM

Margaret Dayhoff, 1978

- Si deux séquences appartiennent au même processus évolutif, et qu'un acide aminé de l'une a été muté pour donner l'autre, alors on peut supposer que les deux acides aminés sont similaires :
  - les mutations sont dites acceptées (Point Accepted Mutation)
  - elles ont été conservées au cours de l'évolution de part leur caractère à ne pas altérer la fonction de la protéine.

Matrices de Dayhoff ou PAM

Margaret Dayhoff, 1978

Les protéines évoluent via des successions de mutations ponctuelles indépendantes les unes des autres et acceptées dans la population.

Matrices de Dayhoff ou PAM

Margaret Dayhoff, 1978

La matrice PAM contient la probabilité d'observer la mutation i → j après un temps évolutif donné.

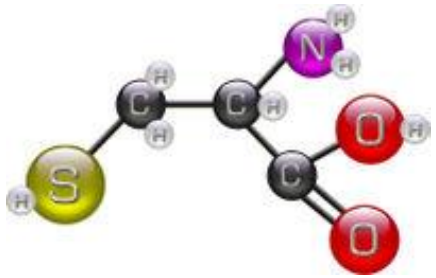
Basée sur alignement global de ~1300 protéines conservées à plus de 85% appartenant à 71 familles de protéines.

Aujourd'hui actualisées : 16 130 séquences appartenant à 2 621 familles de protéines

Matrices de Dayhoff ou PAM Margaret Dayhoff, 1978

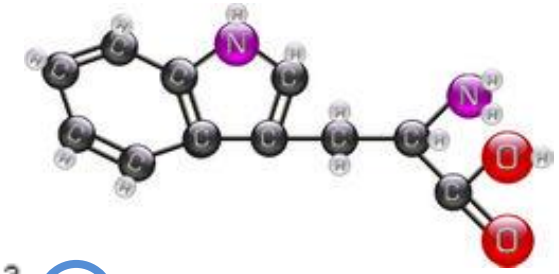
Dans cette matrice de similitude, plus la valeur est négative, plus la probabilité est faible, plus le remplacement est rare.

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



Cystéine

Acides aminés peu permutables



Tryptophane

matrices de substitution BLOSUM

BLOSUM (***BLO****cks of Amino Acid* ***SU****bstitution* ***M****atrix*)

Le but: détecter des relations entre protéines plus éloignées.

Avec les matrices PAM, les valeurs pour des protéines éloignées sont extrapolées.

Avec BLOSUM, ces valeurs sont obtenues en comparant des blocs facilement « alignables » (alignement multiple local sans brèche/trous/gaps) dans des familles de protéines très éloignées.

Ces matrices sont reconnues pour mettre en valeur les similitudes biologiquement importantes (celles qui sont présentes dans les régions alignées sans gaps).

matrices de substitution BLOSUM

Avantages par rapport aux matrices PAM :

contrairement aux matrices PAM, les matrices BLOSUM pour différentes distances évolutives sont obtenues directement avec des séquences plus ou moins divergentes

l'utilisation de blocs plutôt que de séquences complètes : modélise les contraintes uniquement sur les régions conservées obtenues à partir d'un plus grand jeu de données (>2000 blocks, > 500 familles)

Quelle matrice doit-on utiliser?

Les matrices BLOSUM = matrices par défaut
car fréquences de substitution directement calculées à partir de l'alignement.

La BLOSUM62 (ou PAM120) = bon compromis
(ou distances évolutives pas connues)

La BLOSUM80 (ou PAM40) = séquences évolutivement proches
Trouve alignements courts fortement similaires.

La BLOSUM30 (ou PAM350) = séquences éloignées
Trouve longs alignements locaux de faible conservation.

Autre matrice de substitution

Matrices d'après alignement 3D

Basées sur les structures secondaire ou tertiaire.

Évaluent la propension d'un acide aminé à adopter une certaine conformation. Fiables car fondées sur le meilleur alignement possible.

Encore incomplètes en raison de la taille des banques de données 3D.

Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local

4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

Le score d'un alignement

Score = Σ score élémentaire - Σ pénalité d'insertions/délétions

Score est fonction de la longueur de la séquence: + séquence longue + score élevé.

Dépendant de la matrice de substitution utilisée.

Dépendant des poids des pénalités

Dépendant des longueurs des insertion ou des délétion

Traitement des insertions et des délétions

1. **Pénalité simple pour chaque insertion quelle que soit sa longueur**

Traitement des insertions et des délétions

2. Pénalité fixe pour toute insertion + pénalité pour étendre l'insertion

Pénalité moins lourde mais permet de prendre en compte la longueur de gap

$$P = x + yL$$

P: pénalité pour une insertion de longueur L

x: pénalité fixe d'insertion indépendante de la longueur

y: pénalité extension pour l'élément (~10 fois moins que x)

Traitement des insertions et des délétions

Concordance avec les évènements biologiques observés

Poids des pénalités peut être établi selon les endroits où elles se trouvent afin d'améliorer la sensibilité de la recherche (feuillets béta, hydrophobicité…)

Alignement de séquence optimal = chronophage

Temps de comparaison de deux séquences de longueur équivalente N est proportionnel à N^2

L'exploration de chaque position de chaque séquence pour la détermination éventuelle d'une insertion augmente d'un facteur 2N le temps de calcul.

La programmation dynamique est un moyen de limiter cette augmentation pour conserver un temps de calcul de l'ordre de N^2.

Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

Méthode de programmation dynamique

Elle est basée sur le fait que tous les évènements sont possibles et calculables mais que la plupart sont rejetés en considérant certains critères.

Needleman et Wunsch (1970) ont introduit les premiers ce type d'approche pour un problème biologique et leur algorithme reste une référence dans le domaine.

L'algorithme de Needleman et Wunsch

Transformation de la matrice de comparaison initiale en matrice de comparaison transformée selon l'algorithme de Needleman et Wunsch

	V	T	E	E	R	D	A	F
L	2	-2	-3	-3	-3	-4	-2	2
T	0	3	0	0	-1	0	1	-3
S	-1	1	0	0	0	0	1	-3
H	-2	-1	1	1	2	1	-1	-2
E	-2	0	4	4	-1	3	0	-5
A	0	1	0	0	-2	0	2	-4
L	2	-2	-3	-3	-3	-4	-2	2

a) Matrice initiale obtenue à partir de la matrice de substitution utilisée pour l'alignement (ici la matrice PAM250 de Dayhoff)

	V	T	E	E	R	D	A	F
L	14	7	6	6	4	4	0	2
T	10	12	9	9	6	4	3	-3
S	8	10	9	9	7	4	3	-3
H	6	7	9	8	9	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2

b) Matrice transformée construite à partir de la matrice initiale

L'algorithme de Needleman et Wunsch

$$S(i,j) = se(i,j) + \max S(x,y) \qquad \text{avec } i<x\leq m \text{ et } y=j+1 \text{ ou } x=i+1 \text{ et } j<y\leq n \qquad (3)$$

		V	T	E	E	R	D	A	F
		2	-2	-3	-3	-3	-4	-2	2
L		2	-2	-3	-3	-3	-4	-2	2
T		0	3	0	0	-1	0	1	-3
S		-1	1	0	0	(0)	4	3	-3
H		6	7	9	8	9	5	1	-2
E		2	4	8	8	3	7	2	-5
A		2	3	2	2	0	2	4	-4
L		2	-2	-3	-3	-3	-4	-2	2

		V	T	E	E	R	D	A	F
		2	-2	-3	-3	-3	-4	-2	2
L		2	-2	-3	-3	-3	-4	-2	2
T		0	3	0	0	-1	0	1	-3
S		-1	1	0	0	(7)	4	3	-3
H		6	7	9	8	9	5	1	-2
E		2	4	8	8	3	7	2	-5
A		2	3	2	2	0	2	4	-4
L		2	-2	-3	-3	-3	-4	-2	2

c) On montre ici la matrice transformmée en cours de construction. La séquence horizontale est indicée en i et la séquence verticale en j. A gauche, la matrice avant le calcul du score somme à la position i=5 et j=3 et à droite, après ce calcul. Le score somme est obtenue à partir de l'expression 3 décrite dans le texte, c'est-à-dire ici en additionnant le score de substitution de R par S (0) et le score maximum de la zone grisée (7).

Chaque score de la matrice transformée est obtenu par la somme du score actuel et du score maximum déjà obtenu

L'algorithme de Needleman et Wunsch

		V	T	E	E	R	D	A	F
	L	14	7	6	6	4	4	0	2
	T	10	12	9	9	6	4	3	-3
	S	8	10	9	9	7	4	3	-3
	H	6	7	9	8	9	5	1	-2
	E	2	4	8	8	3	7	2	-5
→	A	2	3	2	2	0	2	4	-4
	L	2	-2	-3	-3	-3	-4	-2	2

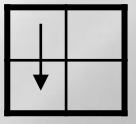
NB: Pas de backtracking dans cet exemple

```
VT-EERDAF
LTSHE--AL
```

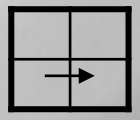
Résultat de l'alignement

Établissement du chemin des scores maximum dans la matrice transformée. Le chemin est établi en partant du score somme le plus élevé, ici 14. Les flèches indiquent les endroits où il est nécessaire de faire des insertions/délétions pour un alignement global optimum.

Match ou mismatch

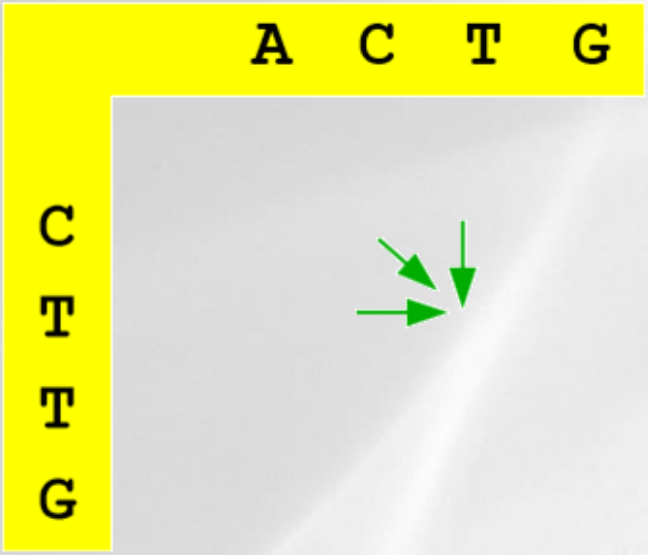


insertion



délétion

Programmation dynamique: un exemple étapes par étapes

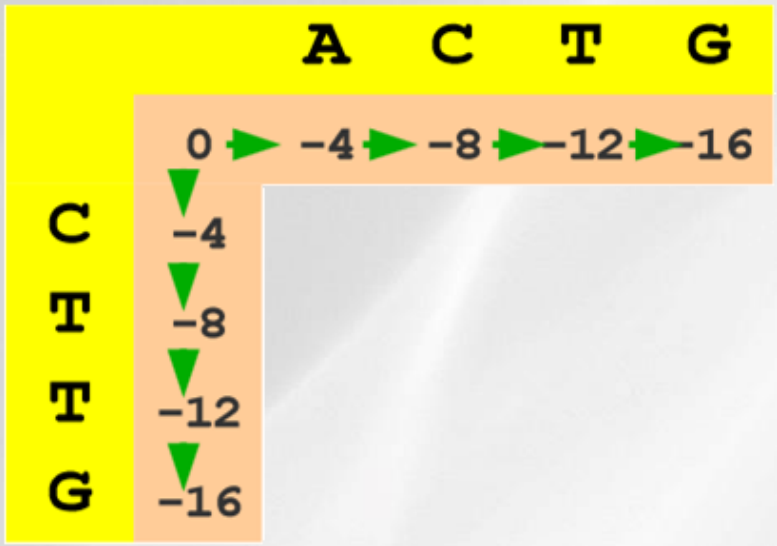


- **Règle 1:**
  chaque case va contenir un score; le score de l'alignement sera celui de la case en bas à droite
- **Règle 2:**
  le score d'une case se déduit à partir de celui des cases au-dessus, à gauche ou en diagonale
- **Règle 3:**
  un pas horizontal/vertical coûte 1 gap
  un pas diagonal coûte 1 position alignée (match ou mismatch)

Programmation dynamique: un exemple étapes par étapes

Etape 1 :

on remplit la première ligne et la première colonne



Score:

gap: -4

Etape 2:

on remplit toutes les cases en gardant en mémoire le mouvement qui donne le meilleur score

		A	C	T	G
	0	-4	-8	-12	-16
C	-4				
T	-8				
T	-12				
G	-16				

Score:

gap: -4 mismatch: -4

alignement AC → score = 0-4 = -4

insertion de gap → score = -4-4 = -8

insertion de gap → score = -4-4 = -8

Programmation dynamique: un exemple étapes par étapes

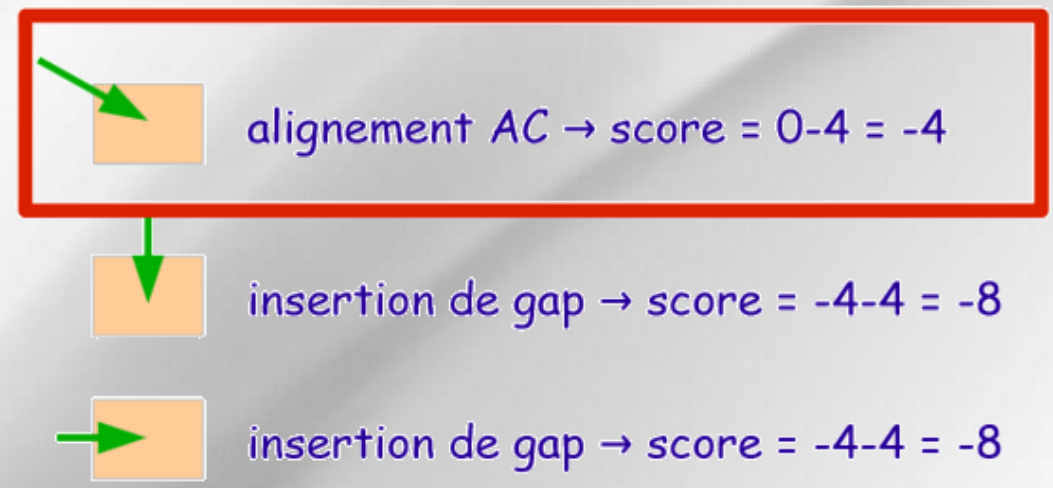
		A	C	T	G
	0	-4	-8	-12	-16
C	-4	-4			
T	-8				
T	-12				
G	-16				

Etape 2:

on remplit toutes les cases en gardant en mémoire le mouvement qui donne le meilleur score

Score:

gap: -4 mismatch: -4



Programmation dynamique: un exemple étapes par étapes

		A	C	T	G
	0	-4	-8	-12	-16
C	-4	-4			
T	-8				
T	-12				
G	-16				

Score:

gap: -4 mismatch: -4

match: +4

alignement CC → score = -4+4 = 0

insertion de gap → score = -8-4 = -12

insertion de gap → score = -4-4 = -8

Programmation dynamique: un exemple étapes par étapes

		A	C	T	G
	0	-4	-8	-12	-16
C	-4	-4	0		
T	-8				
T	-12				
G	-16				

Score:

gap: -4 mismatch: -4

match: +4

alignement CC → score = -4+4 = 0

insertion de gap → score = -8-4 = -12

insertion de gap → score = -4-4 = -8

Programmation dynamique: un exemple étapes par étapes

		A	C	T	G
	0	-4	-8	-12	-16
C	-4	-4	0		
T	-8				
T	-12				
G	-16				

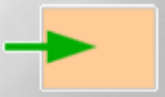
Score:

gap: -4 mismatch: -4

match: +4

alignement AT → score = -4-4 = -8

insertion de gap → score = -4-4 = -8



insertion de gap → score = -8-4 = -12

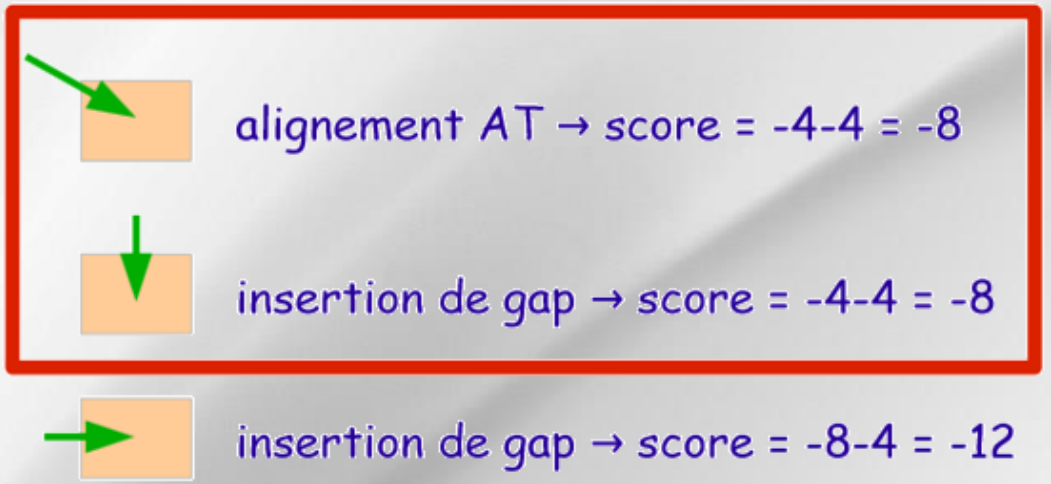
Programmation dynamique: un exemple étapes par étapes

		A	C	T	G
	0	-4	-8	-12	-16
C	-4	-4	0		
T	-8	-8			
T	-12				
G	-16				

Score:

gap: -4 mismatch: -4

match: +4



Programmation dynamique: un exemple étapes par étapes

		A	C	T	G
	0	-4	-8	-12	-16
C	-4	-4	0	-4	-8
T	-8	-8	-4	4	0
T	-12	-12	-8	0	0
G	-16	-16	-12	-4	4

meilleur score

Score:

gap: -4

mismatch: -4

match: +4

Programmation dynamique: un exemple étapes par étapes

Etape 3:

On part du score en bas à droite, et on remonte le cours des flèches pour trouver l'alignement (« **backtracking** »)

		A	C	T	G
	0	-4	-8	-12	-16
C	-4	-4	0	-4	-8
T	-8	-8	-4	4	0
T	-12	-12	-8	0	0
G	-16	-16	-12	-4	4

2 chemins =
2 alignements **optimaux**:

```
AC-TG      ACT-G
-CTTG      -CTTG
```

score: +4

Bilan:
- 24 scores calculés
- 3^{4+4} = 6561 chemins possibles

Alignement **global**:
on aligne les 2 séquences
du début à la fin

Alignement global

		A	T	G	C	A	T	C	C	C	A	T	G	A	C
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14
T	-1	-2	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
C	-2	-3	0	-1	2	1	0	-1	-2	-3	-4	-5	-6	-7	-8
T	-3	-4	-1	-2	1	0	3	2	1	0	-1	-2	-3	-4	-5
A	-4	-1	-2	-3	0	3	2	1	0	-1	2	1	0	-1	-2
T	-5	-2	1	0	-1	2	5	4	3	2	1	4	3	2	1
A	-6	-3	0	-1	-2	1	4	3	2	1	4	3	2	5	4
T	-7	-4	-1	-2	-3	0	3	2	1	0	3	6	5	4	3
C	-8	-5	-2	-3	0	-1	2	5	4	3	2	5	4	3	6
C	-9	-6	-3	-4	-1	-2	1	4	7	6	5	4	3	2	5
G	-10	-7	-4	-1	-2	-3	0	3	6	5	4	3	6	5	4
T	-11	-8	-5	-2	-3	-4	-1	2	5	4	3	6	5	4	3

Scores:

gap: -1

mismatch: -3

match: +2

```
ATGC-ATC-CCATGAC
-T-CTATATCCGT---
```

ces gaps coûtent cher, alors que les 2 séquences sont de longueurs différentes...

Alignement semi-global

les gaps au début de chaque séquence ont un coût nul

		A	T	G	C	A	T	C	C	C	A	T	G	A	C
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	-1	2	1	0	-1	2	1	0	-1	-1	2	1	0	-1
C	0	-1	1	0	3	2	1	4	3	2	1	1	0	-1	2
T	0	-1	1	0	2	1	4	3	2	1	0	3	2	1	1
A	0	2	1	0	1	4	3	2	1	0	3	2	1	4	3
T	0	1	4	3	2	3	6	5	4	3	2	5	4	3	2
A	0	2	3	2	1	4	5	4	3	2	5	4	3	6	5
T	0	1	4	3	2	3	6	5	4	3	4	7	6	5	4
C	0	0	3	2	5	4	5	8	7	6	5	6	5	4	7
C	0	-1	2	1	4	3	4	7	10	9	8	7	6	5	6
G	0	-1	1	4	3	2	3	6	9	8	7	6	9	8	7
T	0	-1	1	3	2	1	4	5	8	7	6	9	8	7	6

le score de l'alignement correspond au meilleur score de la dernière colonne ou la dernière ligne

		A	T	G	C	A	T	C	C	C	A	T	G	A	C
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	-1	2	1	0	-1	2	1	0	-1	-1	2	1	0	-1
C	0	-1	1	0	3	2	1	4	3	2	1	1	0	-1	2
T	0	-1	1	0	2	1	4	3	2	1	0	3	2	1	1
A	0	2	1	0	1	4	3	2	1	0	3	2	1	4	3
T	0	1	4	3	2	3	6	5	4	3	2	5	4	3	2
A	0	2	3	2	1	4	5	4	3	2	5	4	3	6	5
T						3	6	5	4	3	4	7	6	5	4
C						4	5	8	7	6	5	6	5	4	7
C						3	4	7	10	9	8	7	6	5	6
G						2	3	6	9	8	7	6	9	8	7
T						1	4	5	8	7	6	9	8	7	6

il y a un score meilleur ici:

alignement local

```
ATCC
ATCC
```

alignement (semi) global

```
ATGC-ATC-CCATGAC
-T-CTATATCCGT---
```

Où en sommes nous ?

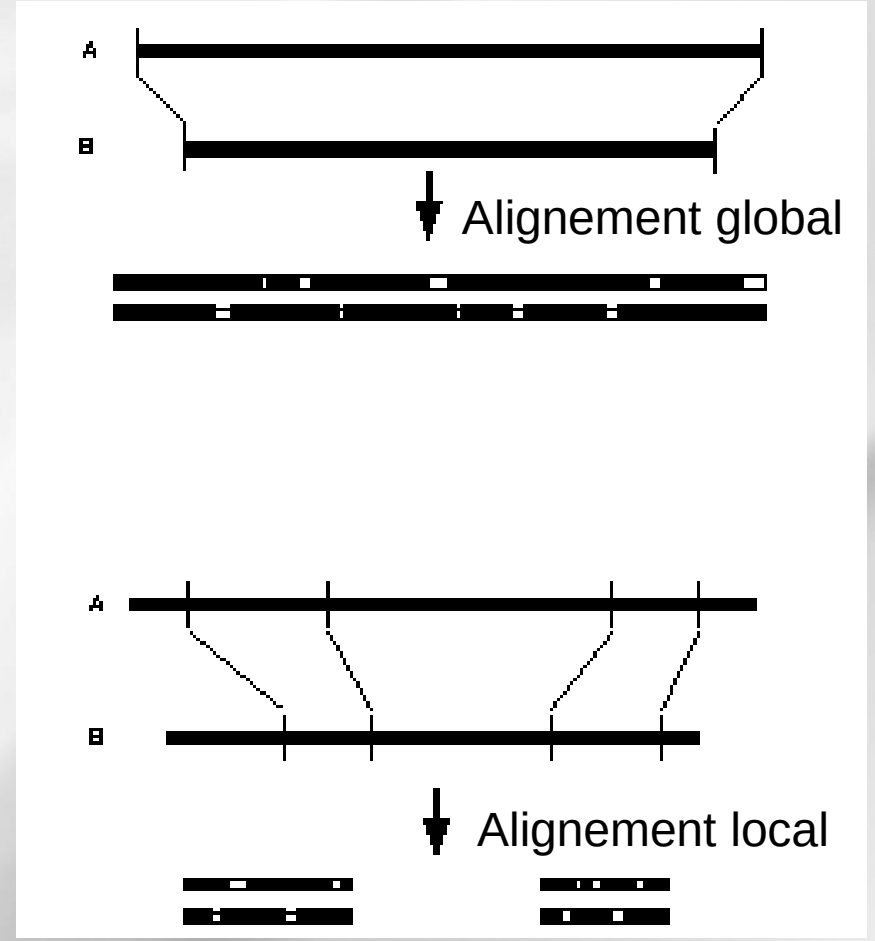
1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

Alignement local ou global

Des finalités très différentes:

L'alignement global est conçu pour comparer des séquences homologues sur toute leur longueur.

L'alignement local est conçu pour rechercher des régions semblables entre A et B.



Les programmes d'alignement global

Méthode employée pour aligner des séquences dont on soupçonne l'homologie.

L'alignement est optimisé sur toute la longueur des séquences.

L'algorithme de référence est celui de Needleman & Wunsh (1970).

Utilisé principalement aujourd’hui dans le cadre de l’alignement multiple

Les programmes d'alignement local

Aligne seulement les régions dont le score est supérieur à un seuil donné.

Utilisé lorsque l'on veut aligner deux séquences de taille très différente. (par ex. dans une recherche de sous séquence).

Beaucoup plus rapide que l'alignement global.

Ex: Smith et Warteman, Fasta, Blast

Programmation dynamique avec arrêt de la procédure quand le score devient trop faible.

Sélection du meilleur alignement local.

L'algorithme de Smith et Waterman

Résultat de l'alignement

EERDAF
TSHEAL

		V	T	E	E	R	D	A	F
	L	2	-2	-3	-3	-3	-4	-2	2
	T	0	3	0	0	-1	0	1	-3
	S	-1	1	0	0	0	0	1	-3
	H	-2	-1	1	1	2	1	-1	-2
	E	-2	0	4	4	-1	3	0	-5
	A	0	1	0	0	-2	0	2	-4
	L	2	-2	-3	-3	-3	-4	-2	2

a) Matrice initiale obtenue à partir de la matrice de substitution utilisée pour l'alignement (ici la matrice PAM250 de Dayhoff)

		V	T	E	E	R	D	A	F
	L	2							2
	T	0	3	0	0		0	1	
	S		1	0	0	0	0	1	
	H			1	1	2	1		
	E		0	4	4		3	0	
	A	0	1	0	0		0	2	
	L	2							2

b) Matrice initiale où sont représentés uniquement les scores positifs ou nuls. C'est à dire toutes les positions susceptibles d'être un premier point de départ pour la transformation de la matrice initiale.

$$S(i,j) = \max \begin{vmatrix} se(i,j) + S(i+1,j+1) \\ se(i,j) + \max S(x,j+1) - P \\ se(i,j) + \max S(i+1,y) - P \\ 0 \end{vmatrix} \quad \begin{array}{ll} \text{avec} & i+2<x\leq m \\ \text{et} & j+2<y\leq n \end{array}$$

NB: Pas de backtracking dans cet exemple

		V	T	E	E	R	D	A	F
	L	8	7	0	0	0	0	0	2
	T	6	6	9	3	0	0	1	0
	S	2	6	3	9	1	0	1	0
	H	0	3	5	2	9	1	0	0
	E	0	0	4	4	0	7	0	0
	A	0	1	0	0	0	0	4	0
	L	2	0	0	0	0	0	0	2

c) matrice transformée construite à paritr de la matrice initiale.L'expression 4 décrite dans le texte est utilisée pour le calcul de chaque score somme avec une pénalité fixe de 6.

		V	T	E	E	R	D	A	F
	L	8	7	0	0	0	0	0	2
	T	6	6	(9)	3	0	0	1	0
	S	2	6	3	(9)	1	0	1	0
	H	0	3	5	2	(9)	1	0	0
	E	0	0	4	4	0	(7)	0	0
	A	0	1	0	0	0	0	(4)	0
	L	2	0	0	0	0	0	0	(2)

d) construction du chemin qui correspond à l'alignement local optimal. Cet alignementl local débute à la position où se trouve le score maximum de la matrice transformée, c'est-à-dire le score 9.

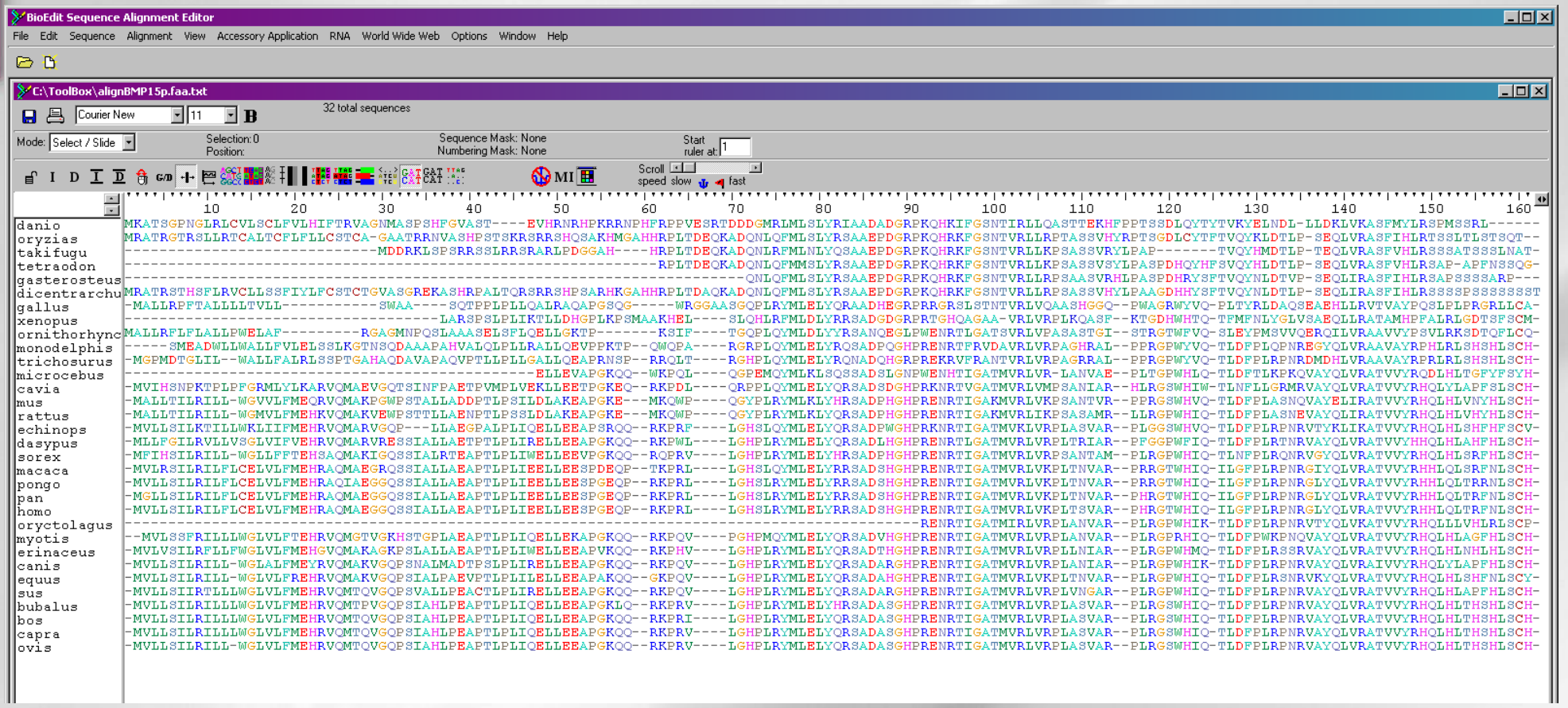
Exercice 2

Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

L’alignement multiple

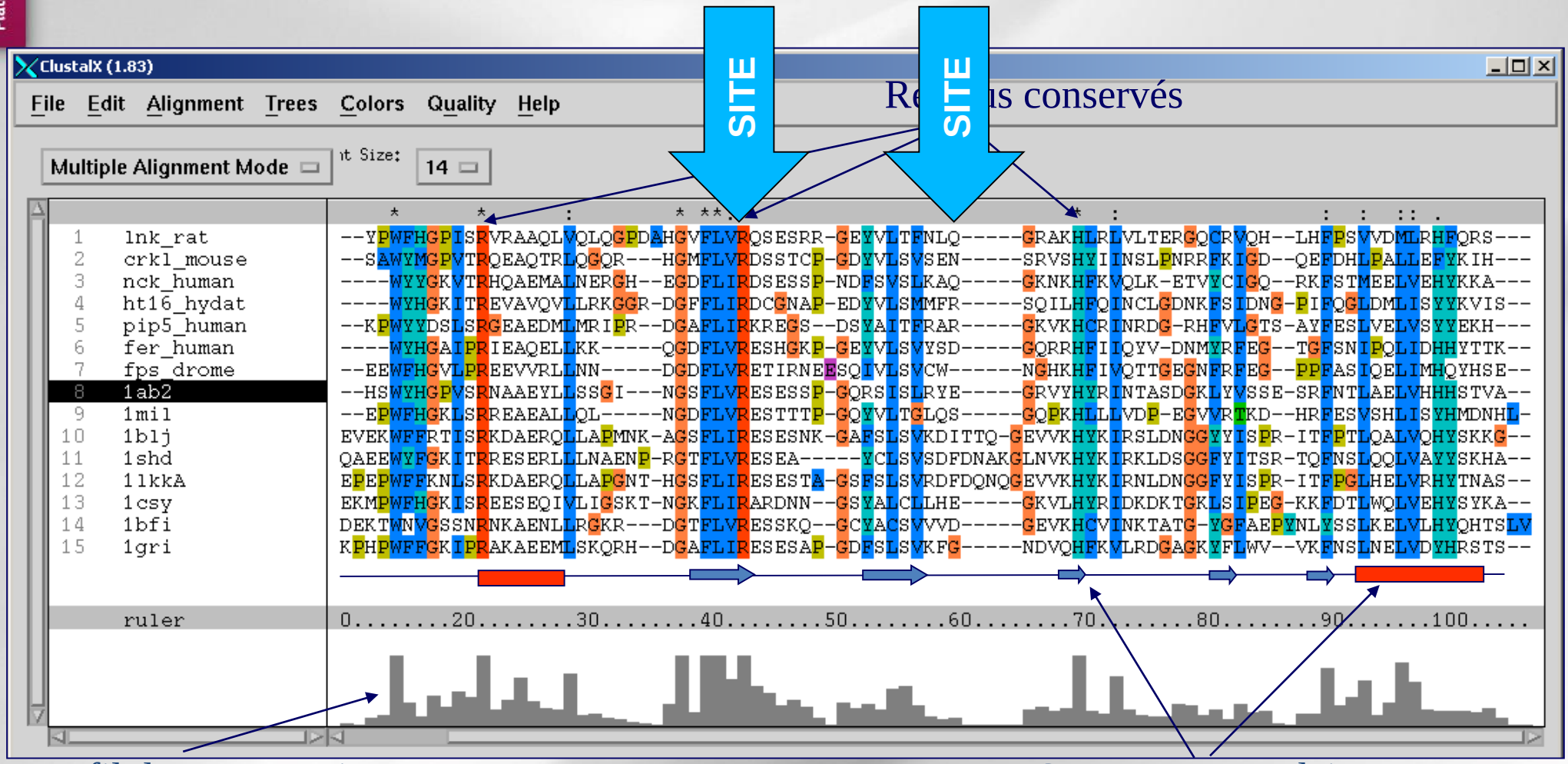
L'alignement multiple dans un éditeur



http://www.mbio.ncsu.edu/BioEdit/bioedit.html

Qu'est-ce qu'un alignement multiple ?

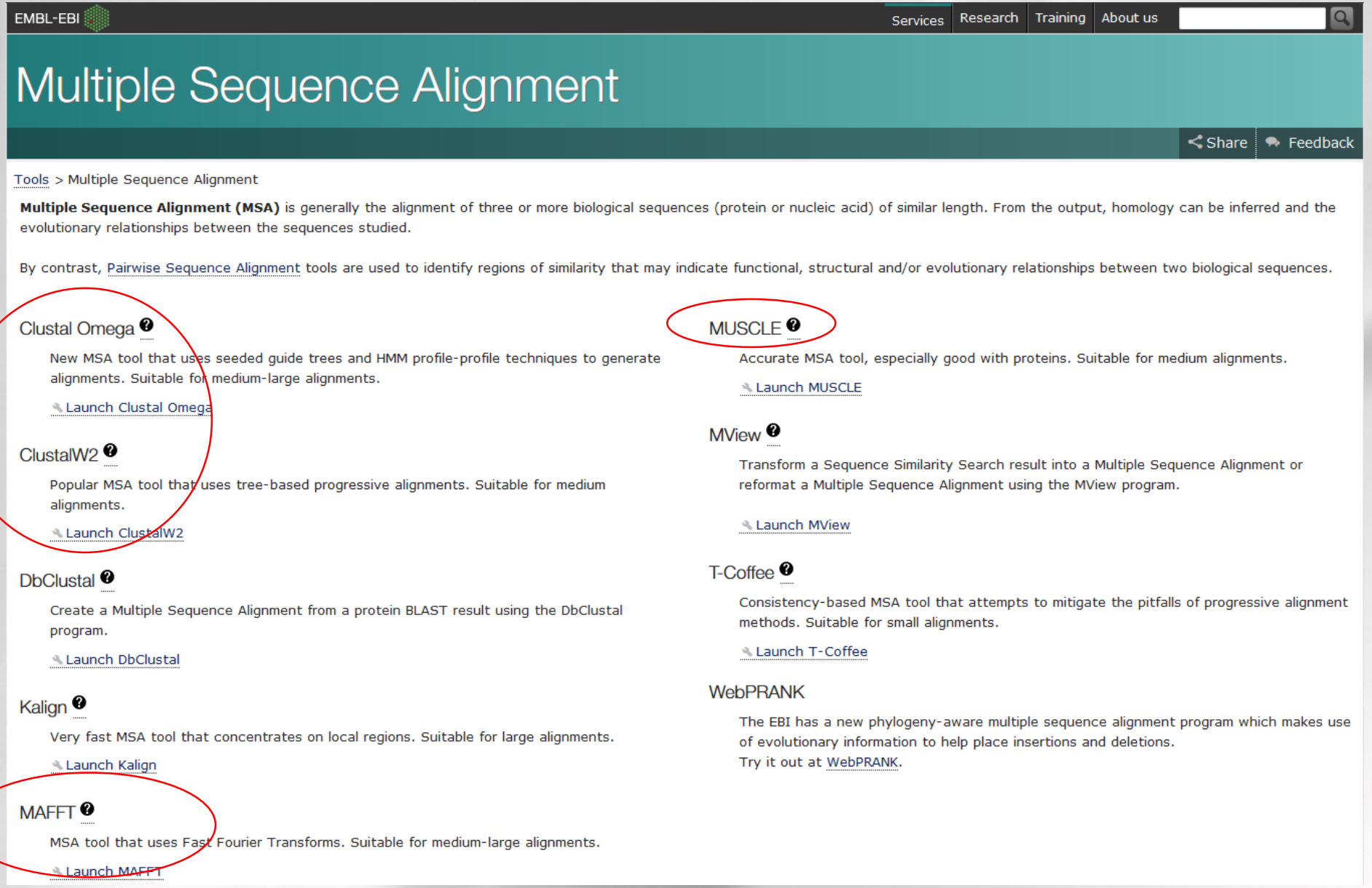
Une représentation d'un ensemble de séquences, dans lesquelles les résidus équivalents (d'un point de vue fonctionnel ou structural) sont alignés en colonnes (un site).



L'alignement multiple au format fasta

```
>danio
MKATSGPNGLRLCVLSCLFVLHIFTRVAGNMASPSHFGVAST----EVHRNRHPKRRNPH
FRPPVESRTDDDGMRLMLSLYRIAADADGRPKQHKIFGSNTIRLLQASTTEKHFPPTSSD
LQYTYTVKYELNDL-LLDKLVKASFMYLRSPMSSRL---------------PYICEASV
TSLQNPLEGDRITMGPRSRWTETDVTDHV--------SESKDGHVSFFARYWCTKPEHKR
SV----AHRKRF---------PPQHHLRAPLLLLFLEENKHPVEWG--------------
-KSFPPLSRPRTRR
>oryzias
MRATRGTRSLLRTCALTCFLFLLCSTCA-GAATRRNVASHPSTSKRSRRSHQSAKHMGAH
HRPLTDEQKADQNLQFMLSLYRSAAEPDGRPKQHRKFGSNTVRLLRPTASSVHYRPTSGD
LCYTFTVQYKLDTLP-SEQLVRASFIHLRTSSLTLSTSQT----------VEPPQCRAQI
ASF---GGKSLAVLEPHEQWTETDITAHVSAHILQGRDINEEGHLTLTAQYWCTEPGKPD
-----VDERKRW---------NSEPHLEAPSLLLYLEEERENST--LELKDSFLDALN--
-SPTSSS-------
>takifugu
----------------------------MDDRKLSPSRRSSLRRSRARLPDGGAH----
HRPLTDEQKADQNLRFMLNLYQSAAEPDGRPKQHRKFGSNTVRLLKPSASSVRYLPAP--
-----TVQYHMDTLP-TEQLVRASFVHLRSSSATSSSLNAT-------QGAKPPRCEARI
TSL---GQESLVTLEPHEQWMETDITAHVR----QDKNQSPGKFLTLTGQYWCAAEDALV
HSEEDVGLKWWWTRLQERGRRSGEHHLEVPSLLLYLEEQREEQR---------------
---------RHRR-
>tetraodon
------------------------------------------------------------
-RPLTDEQKADQNLQFMMSLYRSAAEPDGRPKQHRKFGSNTVRLLKPSASSVSYLPASPD
HQYHFSVQYHLDTLP-SEQLVRASFVHLRSAP-APFNSSQG---------APPPRCRARV
APL---GRESLVVLEPHQRWTETDITAHVR----QRQDQSPGGALTLTAQYRCTAPMAAQ
GGG---GLPRPW------AQRRGGQHLEVPSLLLYLEEERDGNVWMGDLLG---------
-----PEQRRRRRR
>gasterosteus
------------------------------------------------------------
-----------QNLQFMLSLYRSAAEPDGRPKQHRKFGSNTVRLLRPSAASVRHLPASPD
HRYSFTVQYNLDTVP-SEQLIRASFIHLRSAPSSSSARP-----------LRPPRCRAQI
---PSLGKASLVTLEPHERWTETDITAHVR----RGRSRGPGGHLTLTAQYWCTAWGGGF
--------------------PSTGGSPTSRSRLFFYTWRRS-----------------
-GRTPPQRTPRTRR
```

Les outils les plus populaires



EMBL-EBI Services Research Training About us

Multiple Sequence Alignment

Share Feedback

Tools > Multiple Sequence Alignment

Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied.

By contrast, Pairwise Sequence Alignment tools are used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences.

Clustal Omega

New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments.

Launch Clustal Omega

ClustalW2

Popular MSA tool that uses tree-based progressive alignments. Suitable for medium alignments.

Launch ClustalW2

DbClustal

Create a Multiple Sequence Alignment from a protein BLAST result using the DbClustal program.

Launch DbClustal

Kalign

Very fast MSA tool that concentrates on local regions. Suitable for large alignments.

Launch Kalign

MAFFT

MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments.

Launch MAFFT

MUSCLE

Accurate MSA tool, especially good with proteins. Suitable for medium alignments.

Launch MUSCLE

MView

Transform a Sequence Similarity Search result into a Multiple Sequence Alignment or reformat a Multiple Sequence Alignment using the MView program.

Launch MView

T-Coffee

Consistency-based MSA tool that attempts to mitigate the pitfalls of progressive alignment methods. Suitable for small alignments.

Launch T-Coffee

WebPRANK

The EBI has a new phylogeny-aware multiple sequence alignment program which makes use of evolutionary information to help place insertions and deletions.
Try it out at WebPRANK.

Construire un alignement multiple

Approches traditionnelles

A. alignement multiple optimal
B. alignement multiple progressif
C. alignement multiple itératif

Où en sommes nous ?

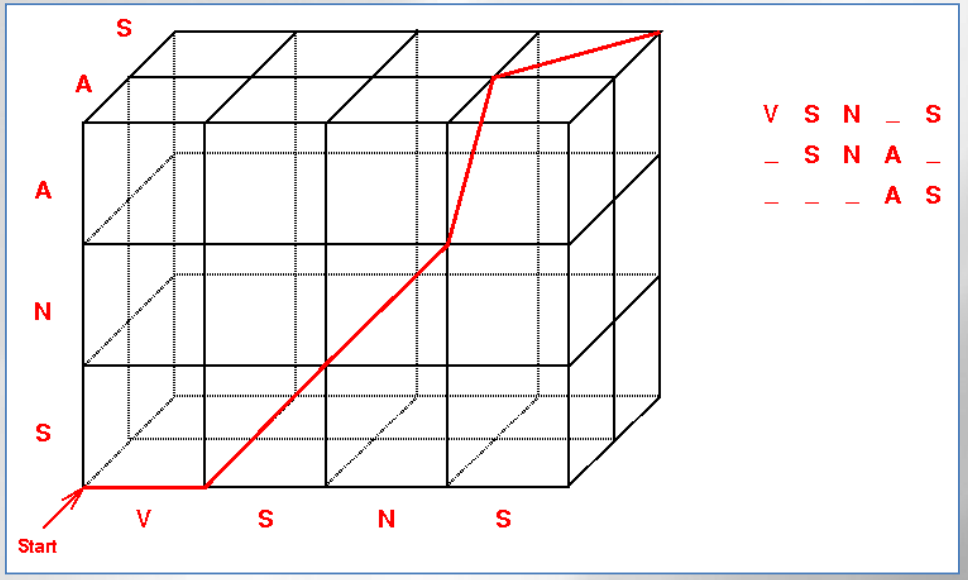
1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

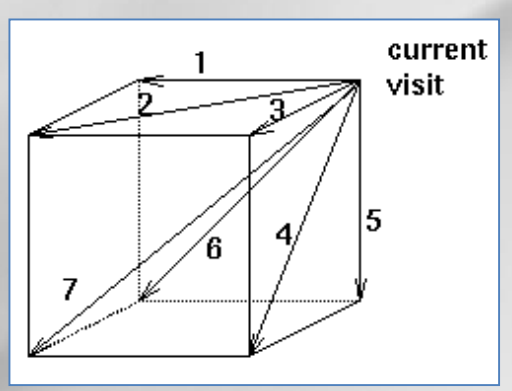
A. Alignement multiple optimal

Extension directe des programmes dynamiques des alignements de séquences par paires à N dimensions (Sankoff, 1975).

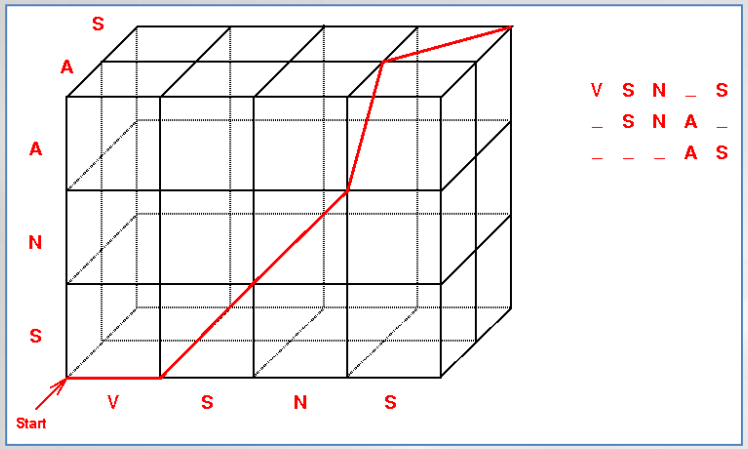
Examine l'ensemble des alignements possibles afin de trouver l'alignement optimal

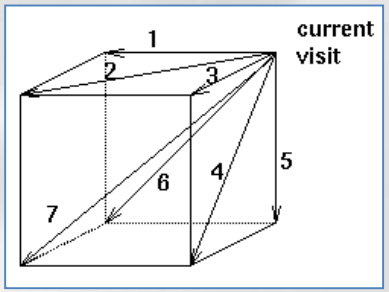
Exemple: alignement de 3 séquences





A. Alignement multiple optimal





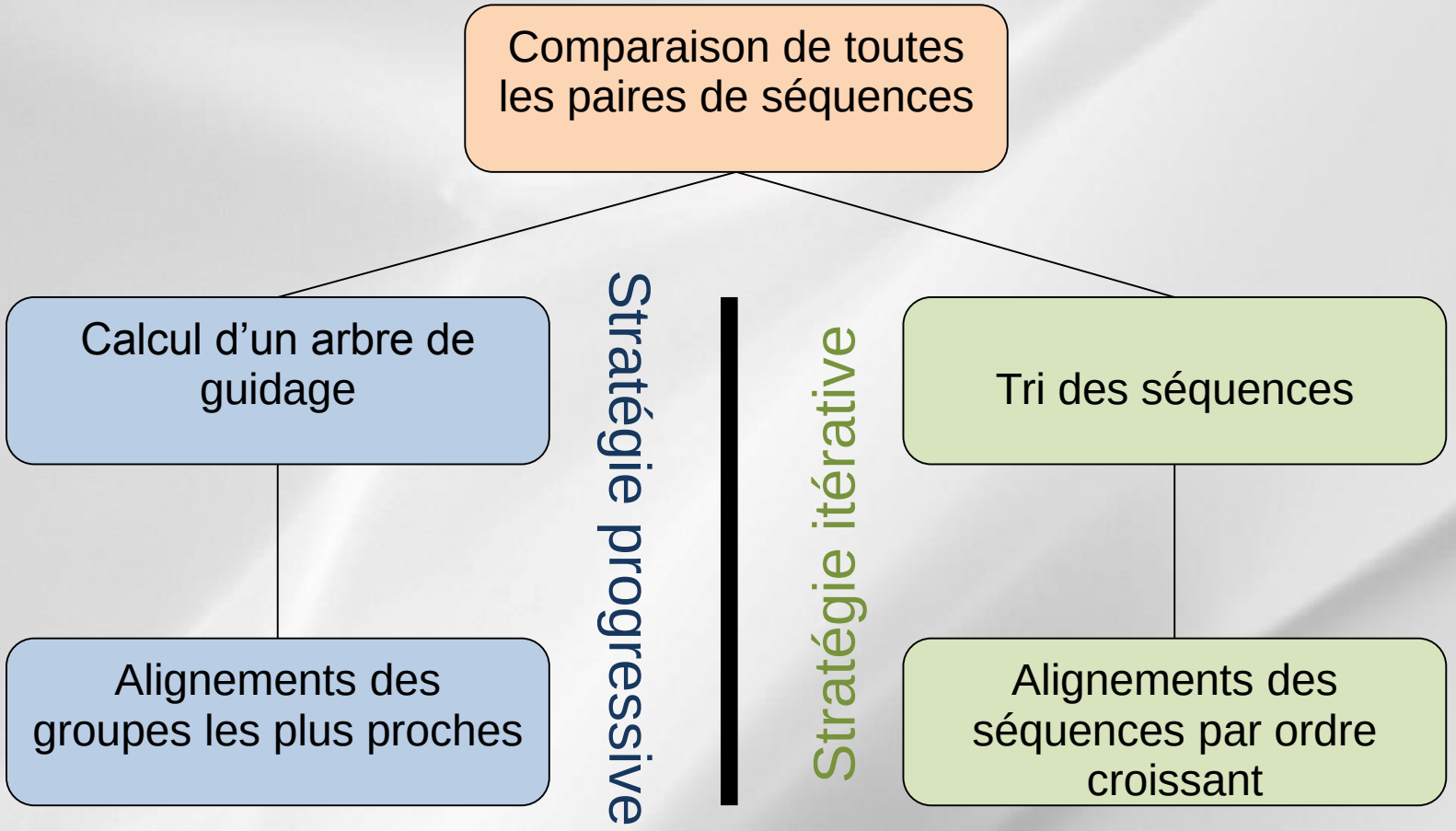
Problème

L'alignement mathématique optimisé n'est pas nécessairement l'alignement biologique optimal.

Le temps CPU (temps de calcul sur ordinateur) et la mémoire requise sont prohibitifs pour un usage classique (temps requis est proportionnel à N^k avec k séquences de longueur N).

En pratique, moins de 10 séquences peuvent être alignées.

Stratégies d'alignements non-optimales



Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

B. Alignement multiple progressif

Evite le calcul de l'ensemble des alignements possibles

Non garantie d'obtenir l'alignement optimal

Principe :

Les séquences (ou groupe de séquences) sont alignées progressivement par paires

B. Alignement multiple progressif

Problème :

Quelles sont les deux premières séquences à aligner?

Dans quel ordre aligner les séquences ?

On aligne en premier les deux séquences les plus proches

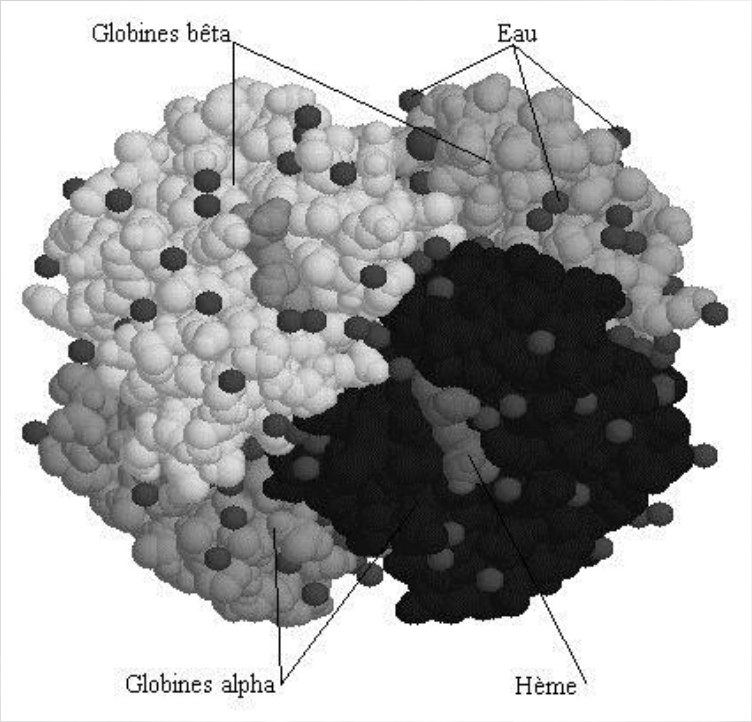
Comment estimer la distance entre deux séquences ?

- Aligner toutes les paires de séquences
- Calculer la matrice de distance à partir des alignements par paires
- Construire un arbre guide à partir de la matrice de distance
- Alignement multiple progressif selon l'ordre des branches de l'arbre

Alignement multiple progressif

Exemple :

Alignement de 7 globines (Hbb_human, Hbb_horse, Hba_human, Hba_horse, Myg_phyca, Glb5_petma et Lgb2_lupla)



Alignement multiple progressif

Étape 1: alignement par paire de toutes les séquences

```
Hbb_human    1 VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLST ...
               | |. |||.|| ||| ||| :||||||||||||||||||||:||||||
Hbb_horse    2 VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSN ...
```

```
Hbb_human    1 LTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFESFGDLST ...
               |.| :|. | | ||||  .  | | ||| |: . :| |. :|  | |||
Hba_human    3 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF.DLS. ...
```

```
Hba_human    3 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF.DLSH ...
               ||  :|  | | | ||     | | ||| |: . :| |. :|  | |||.
Hbb_horse    2 LSGEEKAAVLALWDKVNEE..EVGGEALGRLLVVYPWTQRFFDSFGDLSN ...
```

Les alignement peuvent être obtenus avec:

- méthodes globales ou locales
- Programmation dynamiques ou méthodes heuristiques (non optimales)

Alignement multiple progressif

Étape 2: construction de la matrice de distance

Dans Clustalw:

$$\text{Distance entre deux séquences} = 1 - \frac{\text{Nb de résidus identiques}}{\text{Nb de résidus comparés}}$$

Ex : Hbb_human vs Hbb_horse = 83% identité = distance de 17%

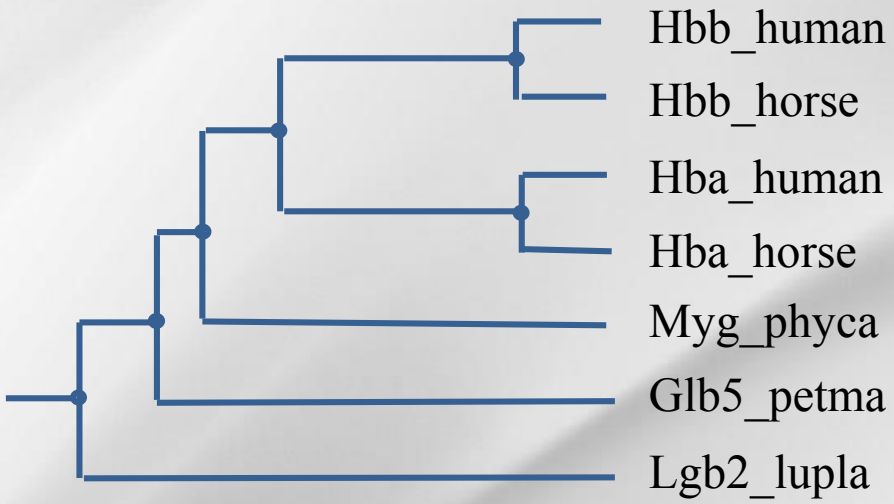
Hbb_human	1	-						
Hbb_horse	2	.17	-					
Hba_human	3	.59	.60	-				
Hba_horse	4	.59	.59	.13	-			
Myg_phyca	5	.77	.77	.75	.75	-		
Glb5_petma	6	.81	.82	.73	.74	.80	-	
Lgb2_lupla	7	.87	.86	.86	.88	.93	.90	-
		1	2	3	4	5	6	7

Alignement multiple progressif

Étape 3: construction de l'arbre guide

Hbb_human	1	-					
Hbb_horse	2	.17	-				
Hba_human	3	.59	.60	-			
Hba_horse	4	.59	.59	.13	-		
Myg_phyca	5	.77	.77	.75	.75	-	
Glb5_petma	6	.81	.82	.73	.74	.80	-
Lgb2_lupla	7	.87	.86	.86	.88	.93	.90	-
		1	2	3	4	5	6	7

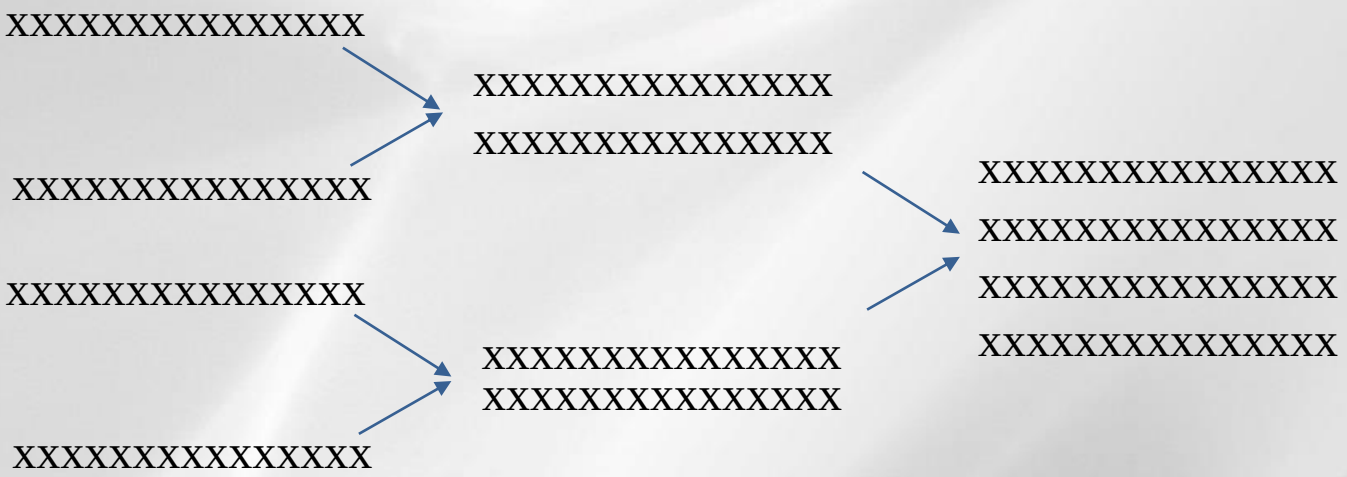
Arbre guide

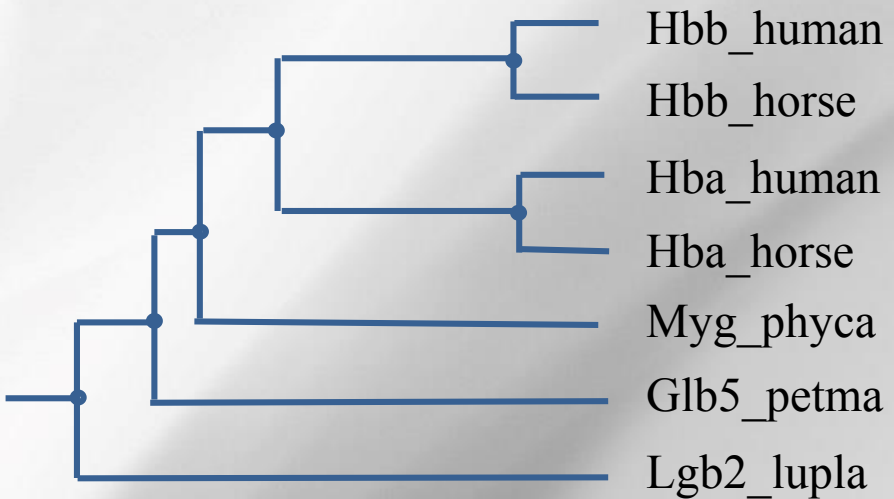


1. Joint les deux séquences les plus proches
2. Calcul à nouveau les distances et
   joint les deux séquences les plus proches ou les noue
3. Répétition de l'étape 3 jusqu'à ce que toutes les séquences soient jointes.

Alignement multiple progressif

Étape 4: Alignement progressif selon l'ordre des branches de l'arbre guide





Où en sommes nous ?

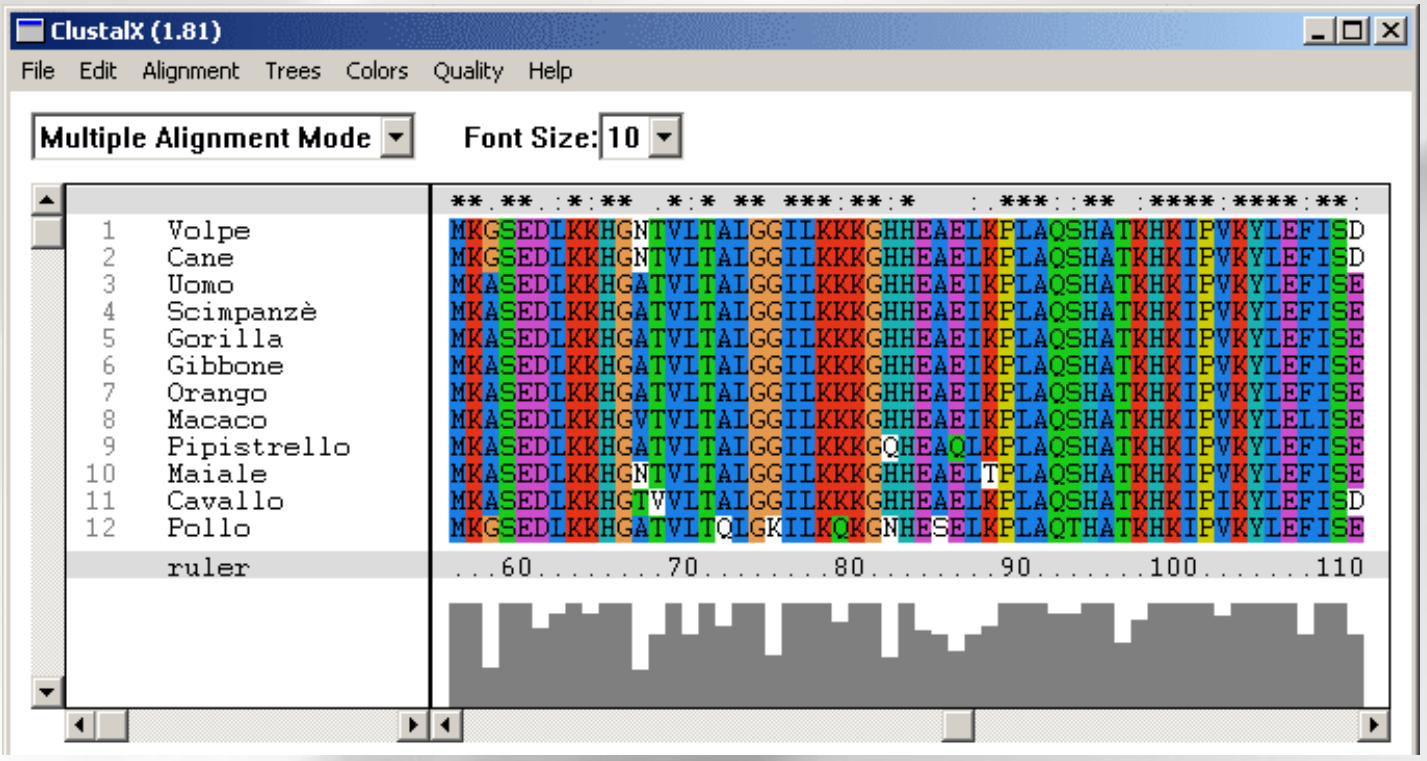
1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

ClustalW

ClustalW (de Des Higgins) est le programme d'alignement multiple le plus employé.

Version actuelle: CLUSTAL Omega - Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments.

ClustalX c'est ClustalW avec interface graphique



ClustalW

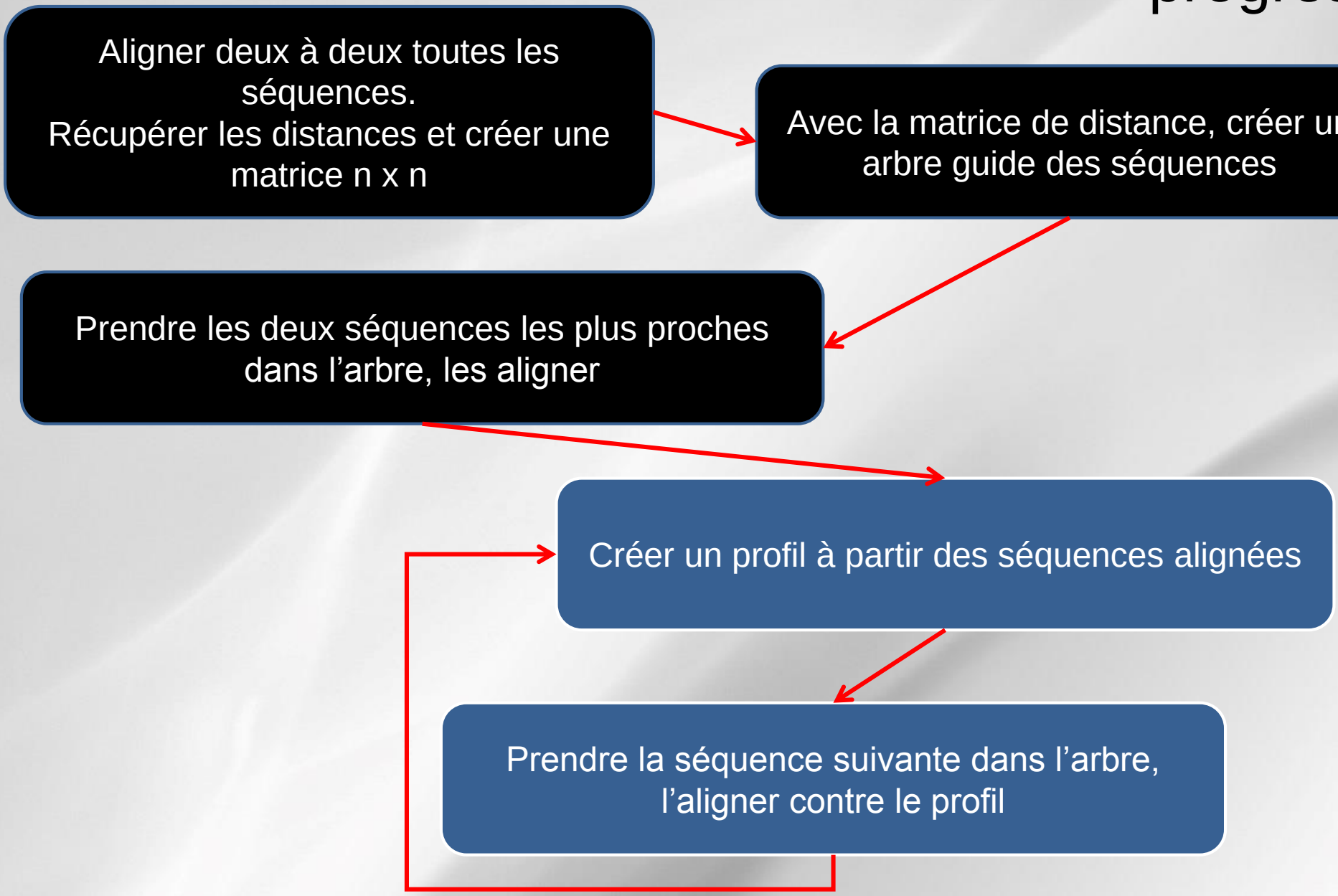
ClustalW utilise les profils.

Les séquences déjà alignées servent de profil pour diriger la suite de l'alignement.

Les profils sont représentés sous forme de tableau dans lequel sont données pour chaque position la fréquence observée de chaque lettre.

Chaque nouvelle séquence est alignée contre le profil des séquences déjà alignées.

Algorithme de ClustalW, un algorithme progressif



Les particularités de ClustalW

Pondération des séquences en fonction de leur sur/sous représentation (Sequence weighting).

Lorsque l'alignement contient plusieurs séquences très proches, celles-ci vont prendre plus d'importance dans le profil qu'une séquence isolée éloignée de celles-ci.

Pour éviter que de tels groupes de séquences proches biaisent l'alignement, l'importance de chaque séquence dans le profil est pondérée en fonction du nombre de séquences proches, telles que déterminés par un arbre.

Les particularités de ClustalW

Adaptation des matrices de similitudes au fil de l'algorithme en fonction de la divergence des séquences à aligner

Blosum 80 pour aligner des séquences proches

Blosum 50 pour aligner des séquences distantes

Les particularités de ClustalW

Pénalités de gaps spécifiques à chaque résidu.

Par exemple, les Glycines sont davantage susceptibles d'avoisiner un gap que les Valines.

Pénalités de gaps réduites dans les régions hydrophiles

Encourage la formation de gaps dans des boucles plutôt que dans des régions structurées.

Pénalités de gaps augmentées dans le voisinage d'autres gaps

Evite la formation de petits gaps voisins, au profit de longs gaps.

Attention

Attention l'arbre guide généré dans ClustalW n'est pas un arbre phylogénétique

Algorithme de ClustalW utilisation de 2 méthodes

Les scores d'alignement calculés par clustalW utilisent 2 méthodes pour les alignements par paires:

Algorithme de ClustalW utilisation de 2 méthodes pour l'alignement par paires

1 - la programmation dynamique qui a pour avantage d'être optimale mais lente.

La programmation dynamique sera plutôt utilisée pour des jeux de courtes séquences mais deviendra extrêmement lente pour des jeux de données supérieures à 100 séquences de 1000 acides aminés chacune.

N.B. : Bien qu'implémentant de la programmation dynamique (algorithme de Needleman et Wunsch) pour l'alignement par paires, clustalw n'utilise pas la programmation dynamique pour l'alignement multiple global.

Algorithme de ClustalW utilisation de 2 méthodes pour l'alignement par paires

2- l'algorithme de *Wilbur et Lipman*, qui est très rapide mais plus approximatif.

Le programme n'utilise que les meilleures diagonales, c'est à dire celles présentant le plus de fragments d'appariements exacts.

Attention

L'ordre des séquences dans le fichier d'entrée a un impact important sur les résultats finaux.

Conclusion sur ClustalW

Grand succès de ClustalW

D'autres programmes fondés sur d'autre algorithmes ou heuristiques donnent souvent de meilleurs résultats dans les cas difficiles.

Si vos séquences sont difficiles à aligner (peu de similarités, longueurs différentes), il est impératif d 'essayer d'autres programmes comme DIALIGN, MAFFT ou MUSCLE (méthode itérative).

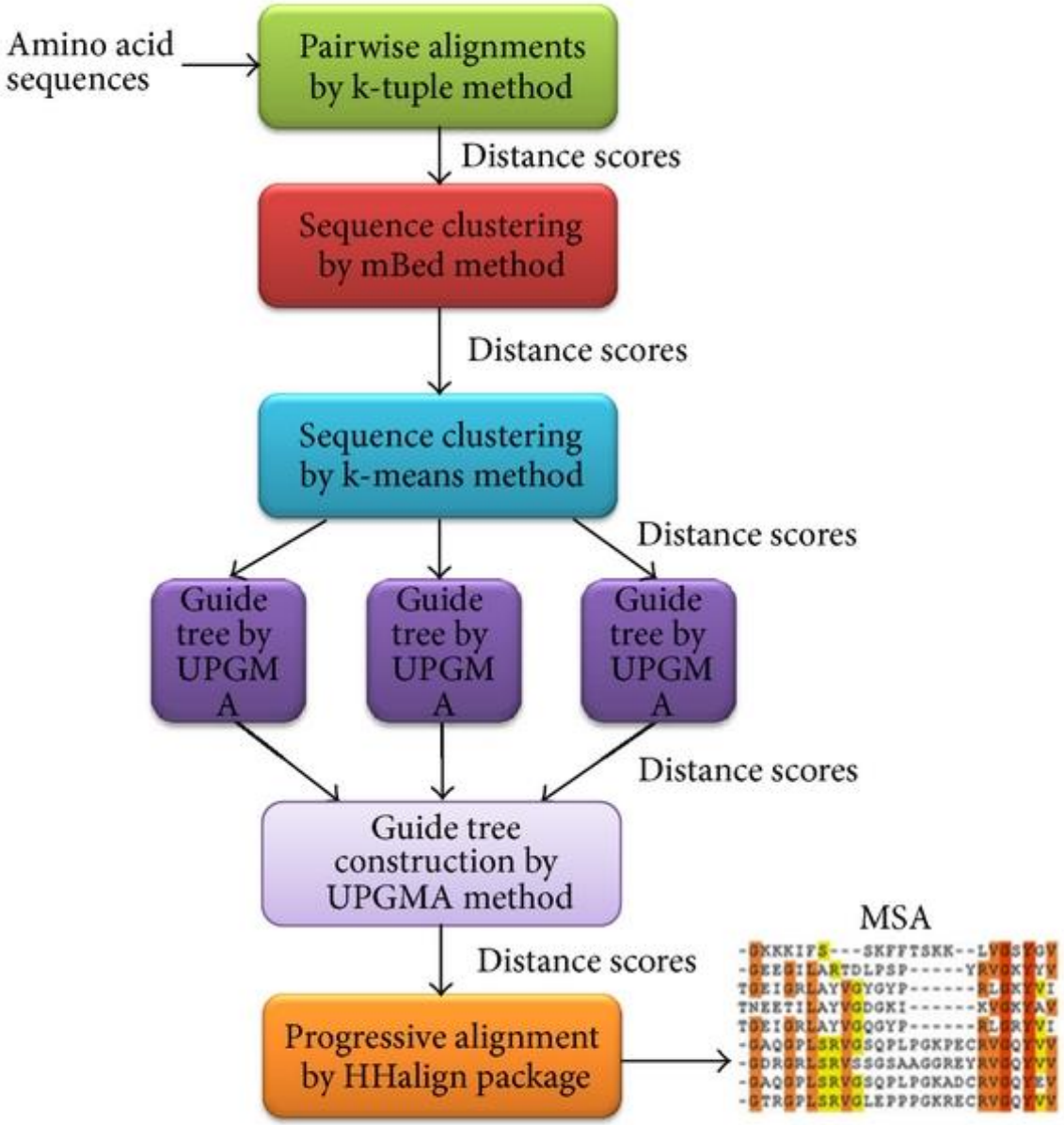
Conclusion sur ClustalW

N'oubliez pas que clustalW – et il n'est pas le seul – souffre d'un défaut congénital gravissime: certes il effectue un alignement multiple, mais il le fait à partir de l'alignement des paires de séquences.

Autrement dit, quand il aligne la *j*-ième séquence avec la *n*-ième séquence pour calculer leur score d'alignement et construire l'arbre de guidage, il ignore « royalement » toute l'information contenue dans les autres séquences.

Un court motif commun à deux séquences peut ne pas être repéré, même s'il est commun à toutes les séquences…

Clustal Omega



CLUSTAL Omega - Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments.

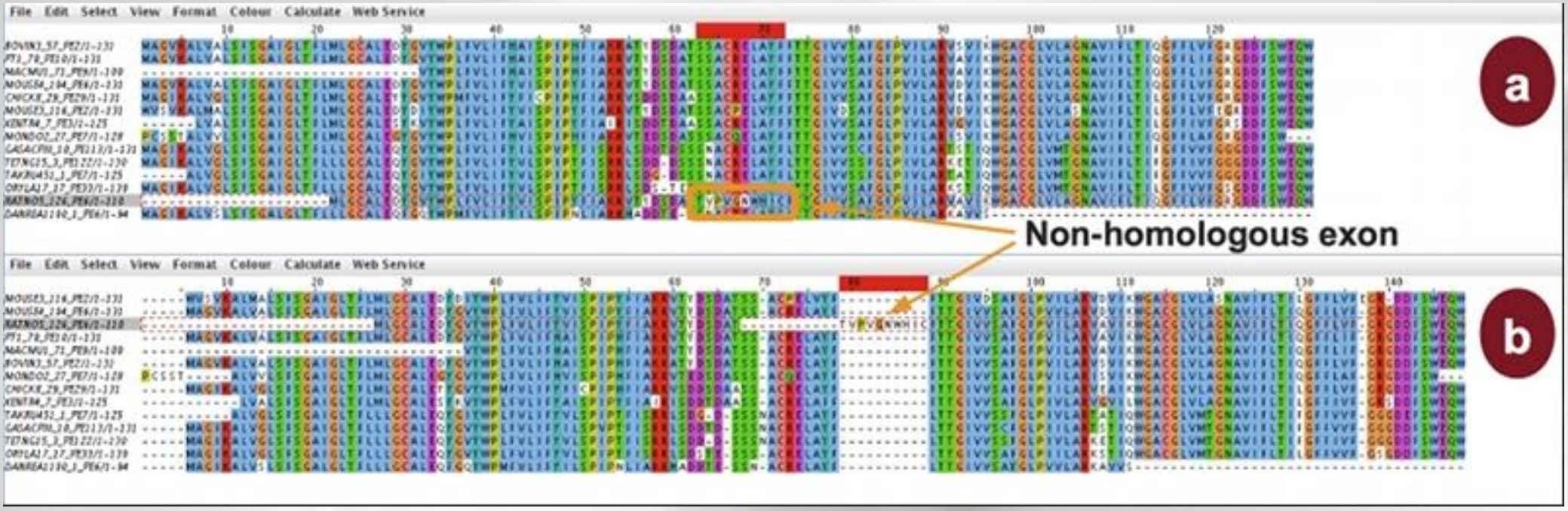
Clustal Omega algorithm, which works by taking an input of amino acid sequences (or DNA) completing a pairwise alignment using the k-tuple method, sequence clustering using mBed method, and k-means method, guide tree construction using the UPGMA method, followed by a progressive alignment using HHalign package to output a multiple sequence alignment.

Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

Prank

Non-homologous exon alignment by an iterative method (a), and by a phylogeny-aware method (b)



Source: Wikipédia

L'exon non homologue ne doit pas être aligné (car non homologue !) or seules les méthodes fondées sur la phylogénie ne l'alignent pas

Prank

Differences to other alignment methods

It uses evolutionary information for the placement of gaps and modelling of the substitution process.

When this information is correct, PRANK makes superior alignments compared to other progressive methods.

However, when its assumptions are violated, the program's performance may be significantly affected.

If your phylogenetic tree is wrong, your MSA will be more false !

Prank

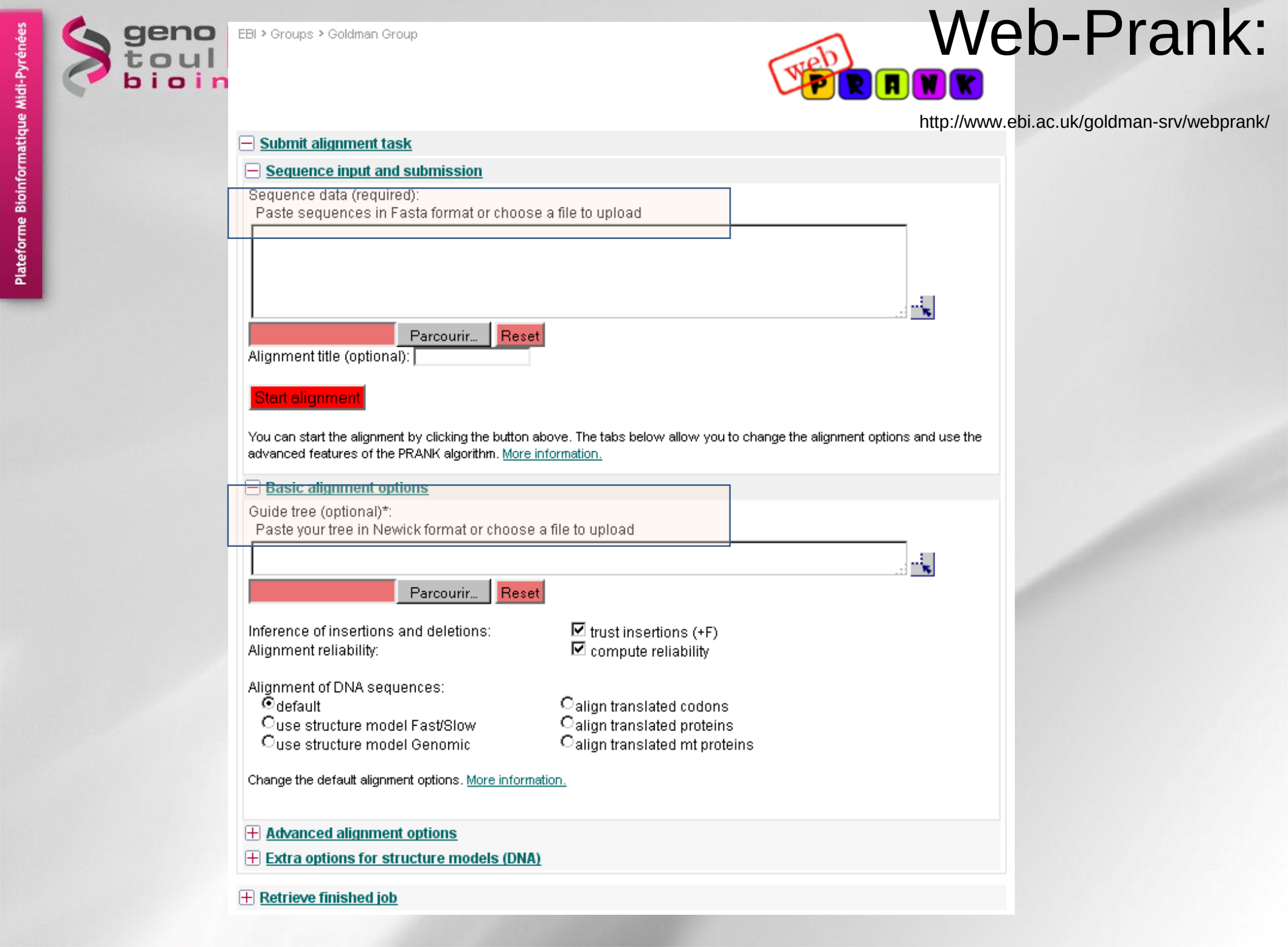
The reconstruction of evolutionary homology -- including the correct placement of insertion and deletion events -- is only feasible for rather closely-related sequences.

PRANK is not meant for the alignment of very diverged protein sequences.

If sequences are very different, the correct homology cannot be reconstructed with confidence and PRANK may simply refuse to match them.

Web-Prank:

http://www.ebi.ac.uk/goldman-srv/webprank/



Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

C. Principe de la méthode itérative

1. Dans une première phase,

   on calcule un score de similarité entre toutes les paires de séquences par comparaison des séquences deux à deux;

   on obtient un ensemble de scores d'alignement qui sont regroupés dans une matrice dites de similarités

C. Principe de la méthode itérative

2. Cette matrice est utilisée pour trier les séquences,

   généralement de plus proches ou similaires aux plus éloignées

C. Principe de la méthode itérative

3. Cette liste est parcourue itérativement pour construire l'alignement multiple final, (pas d'arbre guide)

   c'est-à-dire que les deux plus proches séquences sont alignées (itération 1).

   A partir de cet alignement, on calcule un « profil », qui est en quelque sorte une séquence consensus,

   puis on aligne la troisième séquence profil (itération 2).

   Un nouveau profil est calculé avec ces trois séquences, et la quatrième séquence est alignées.

Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

DIALIGN

extrait du livre « bioinformatique principes d'utilisation des outils »

DIALIGN est un programme d'alignement multiple qui repose sur une méthode très différente de celle employée par ClustalW.

Il s'agit ici d'un algorithme itératif utilisant une approche locale pour calculer les alignements.

DIALIGN

extrait du livre « bioinformatique principes d'utilisation des outils »

Dans le fonctionnement d'un algorithme itératif, la première étape consiste à comparer toutes les paires de séquences.

Dans le cas de DIALIGN, cette étape consiste à rechercher tous les fragments pour ne retenir que ceux qui sont compatibles.

Un fragment consiste en une suite (la plus grande possible) de résidus (bases, acides aminées) consécutifs, similaires entre deux séquences.

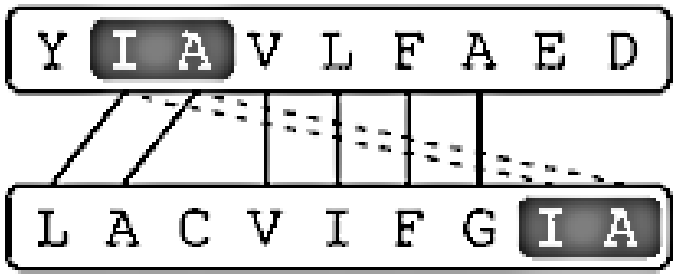
Selon cette définition, on constate qu' un fragment ne peut pas contenir d'indels.

DIALIGN

extrait du livre « bioinformatique principes d'utilisation des outils »

Ensuite, on ne retient que ceux qui sont compatibles, c'est-à-dire des fragments qui ne se croisent pas.

Notez que les fragments de DIALIGN s'appellent des mots dans d'autres programmes comme BLAST ou FASTA.



Exemple d'un fragment DIALIGN commun à deux séquences.

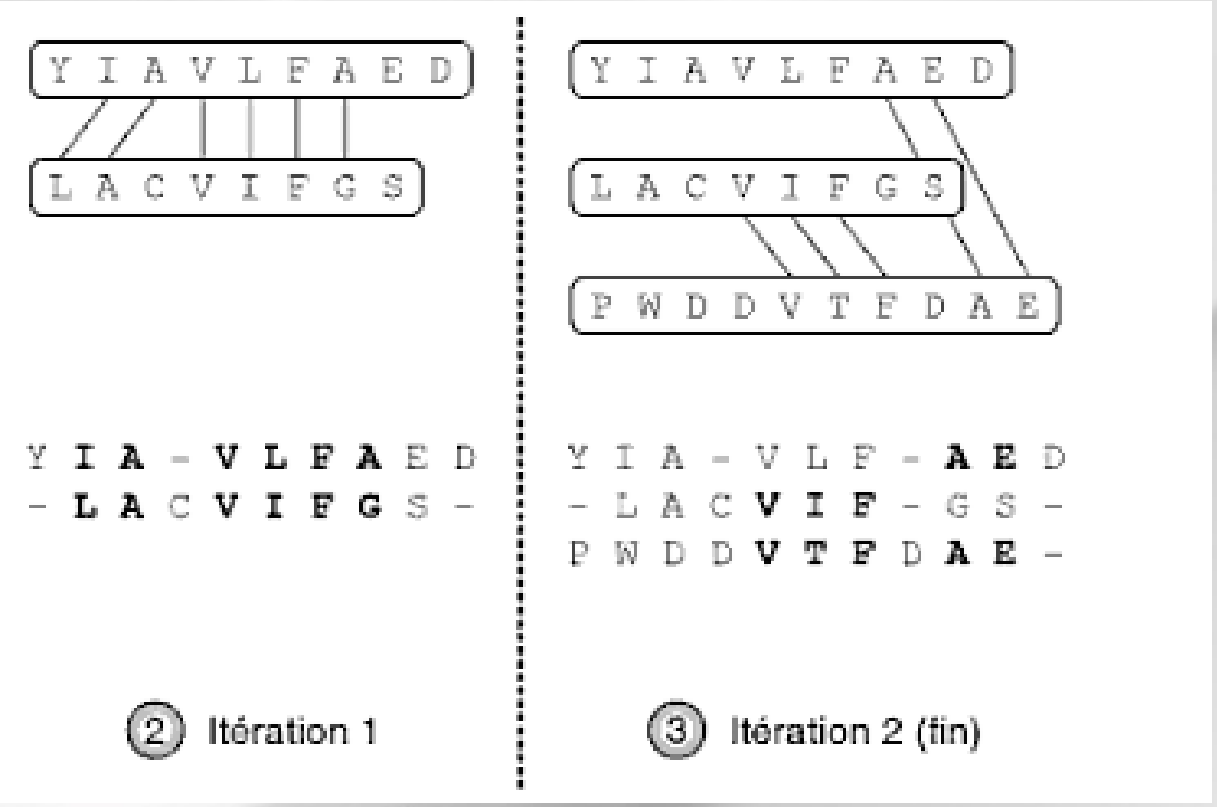
Par exemple, les deux séquences **IAVLFA** et **LACVIFG** sont similaires mais de longueurs différentes; ce ne sont donc pas des fragments au sens de la définition ci-avant.

En revanche, **IA/LA**, **IA/IA** et **VLFA/VIFG** sont des fragments,

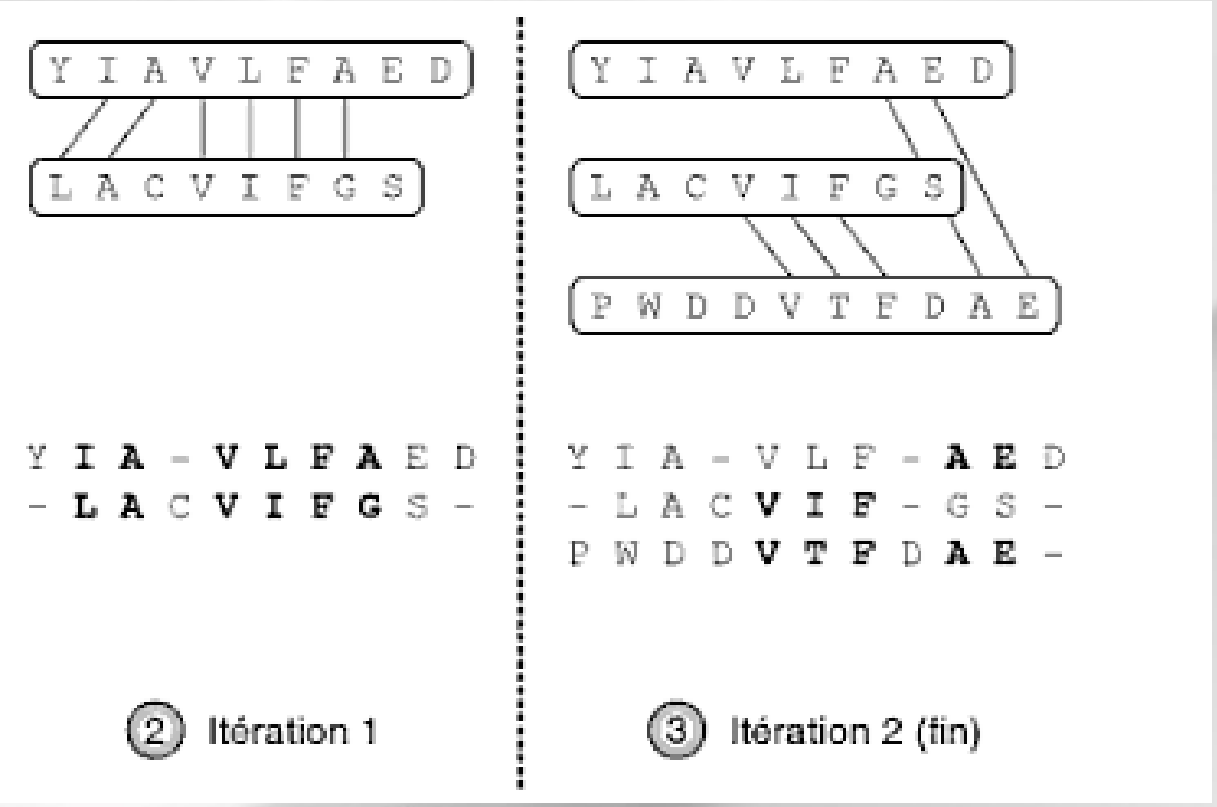
mais **IA** n'est pas compatible puisqu'à gauche dans un cas et à droite dans l'autre.

Après avoir repéré tous les fragments compatibles, les séquences sont triées en fonction du nombre total de fragments communs entre elles.

La dernière étape de l'algorithme consiste à aligner itérativement les séquences, c'est-à-dire de la première à la dernière séquence de la liste.



A chaque itération, des insertions sont ajoutées de manière à ce que les différents résidus soient correctement alignées.



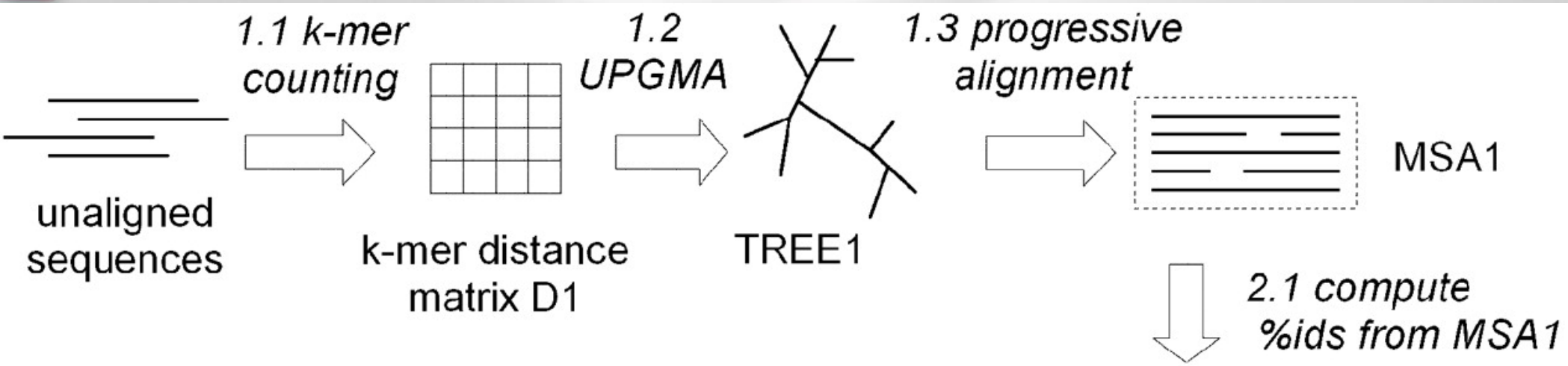
Exercice 3

Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

MUSCLE: méthode itérative (et progressive)

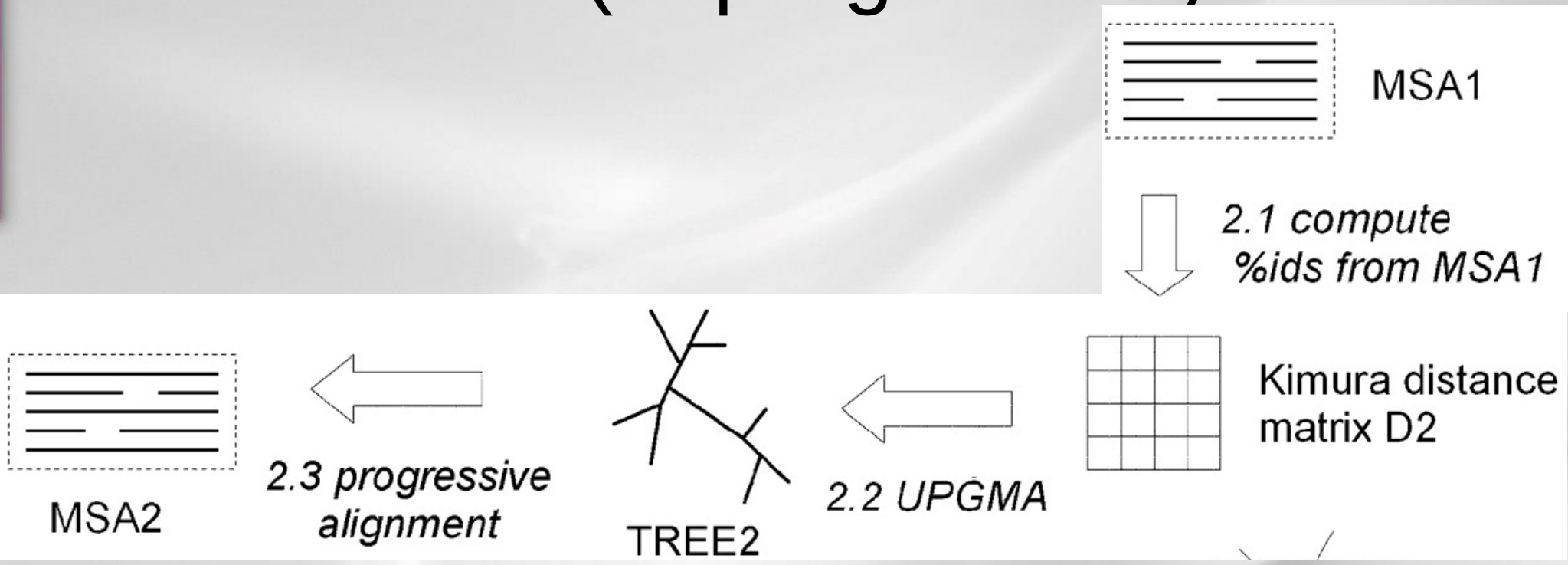
MUSCLE: multiple sequence alignment by log-expectation



Un arbre guide est construit à partir de la matrice de distance qui a été calculé par UPGMA ou NJ (Neighbor Joinning), et une racine est identifiée.

Un alignement progressif est construit en suivant l'ordre des branches de l'arbre guide, produisant un alignement multiple de l'ensemble des séquences.

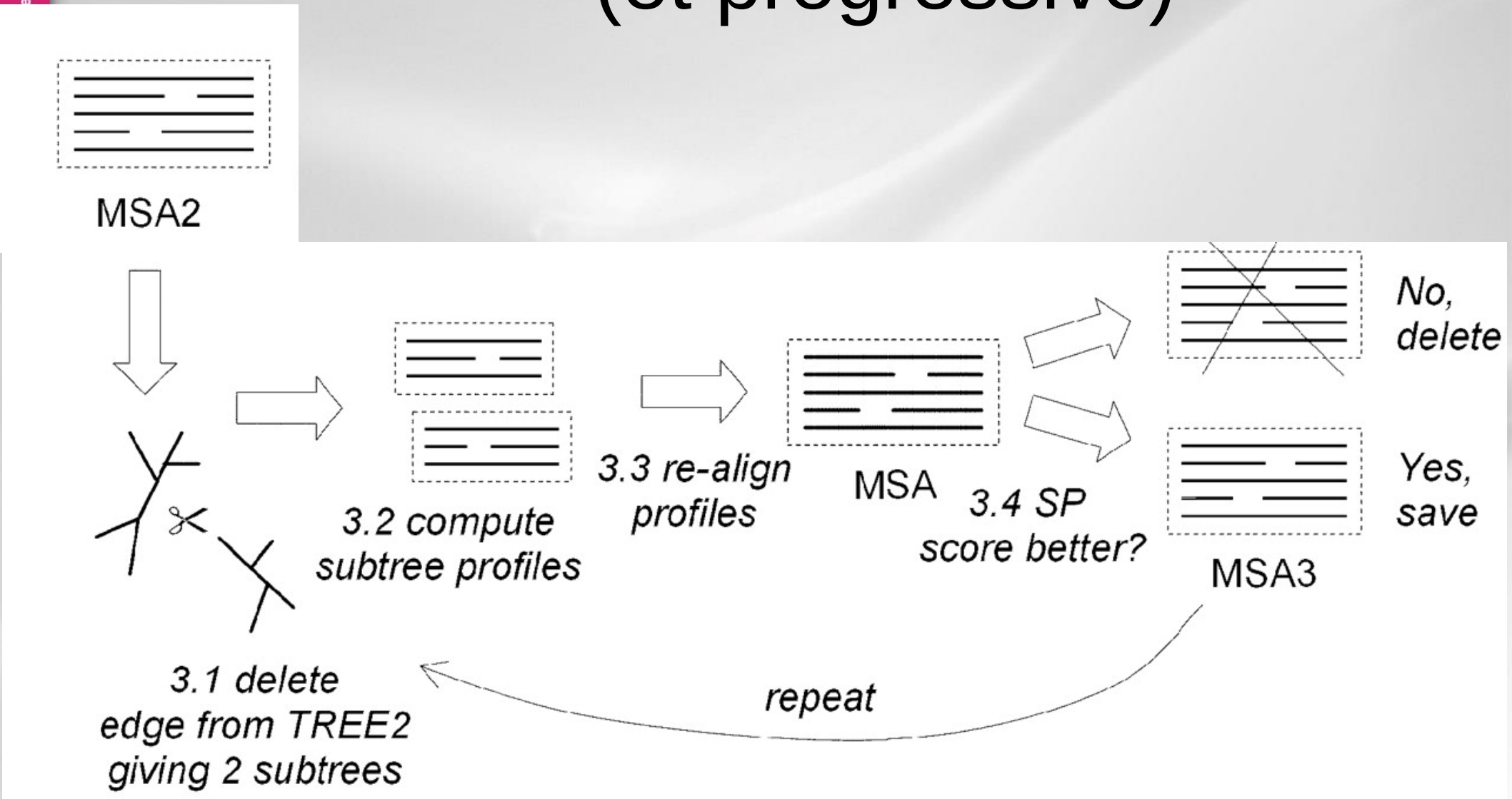
MUSCLE: méthode itérative (et progressive)



La deuxième étape tente d'améliorer l'arbre guide et construit un nouvel alignement progressif selon cet arbre.

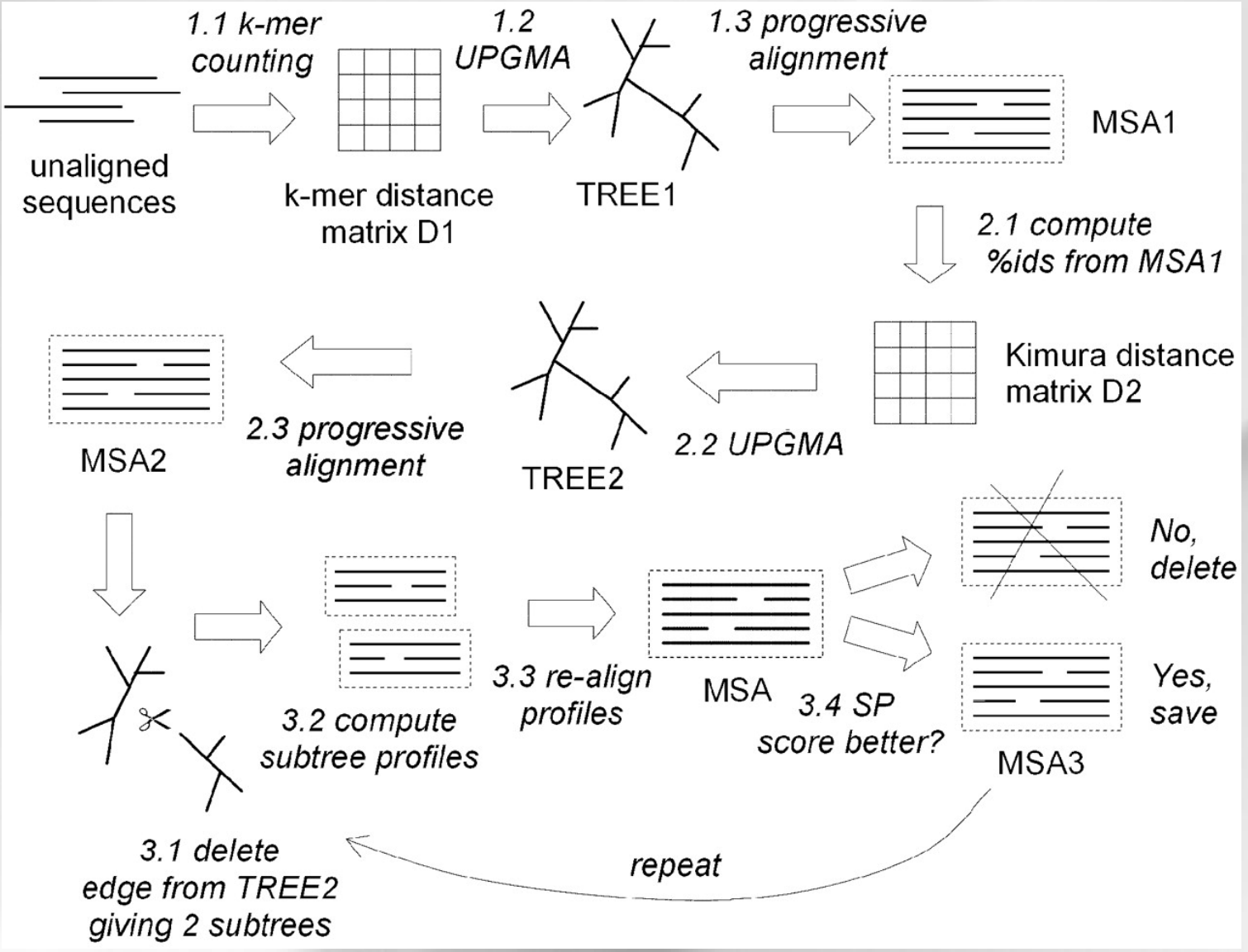
Cette étape peut être répétée (méthode itérative).

MUSCLE: méthode itérative (et progressive)



Troisième étape : raffinement

MUSCLE: méthode itérative (et progressive)



MUSCLE: méthode itérative (et progressive)

MUSCLE uses a much faster, but somewhat more approximate,

method to compute distances: it counts the number of short sub-sequences (known as *k*-mers, *k*-tuples or words) that two sequences have in common, without constructing an alignment.

This is typically around 3,000 times faster that CLUSTALW's method, but the trees will generally be less accurate.

Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

MAFFT: méthode itérative (Katoh et al, 2002) d'alignement multiple

MAFFT = programmes de nouvelle génération

MAFFT a été écrit dans le but explicite d'accélérer considérablement le processus d'alignement multiple,

permettant ainsi d'aligner un grand nombre de séquences sans pour autant sacrifier à la qualité de l'alignement.

MAFFT: méthode itérative (Katoh et al, 2002) d'alignement multiple

3 grandes étapes :

1ère ETAPE

chaque acide aminé est décrit par sa polarité et son volume, les séquences sont réécrites dans ce système.

suite de lettres (chaque séquence) → une suite de valeurs numériques.

Les nucléotides sont recodées en utilisant les fréquences locales des quatre bases.

Puis les segments de similarité entre chaque paire de séquences sont repérés au moyen d'un algorithme de calcul appelé transformée de Fourier rapide, ou Fast Fourier Transform (FFT) en anglais.

Les paires de séquences sont ensuite alignées sur la base de ces segments de similarité (cf. DIALIGN). Sauf pour des séquences très divergentes, ce procédé permet d'aligner toutes les paires de séquences environ 10 fois plus vite que ClustalW;

MAFFT: Alignement multiple

2ème ETAPE

Un arbre de guidage (cf. ClustalW et MUSCLE) est ensuite calculé à partir des alignements précédents.

Ici, le calcul des distances entre les séquences est simplifié et accéléré en recodant les séquences protéiques dans un alphabet réduit à 6 lettres: par exemple, les acides aminés hydrophobes I, L, M et V forment un seul groupe, de même que D, E,N et Q, etc.

La distance entre 2 séquences est estimée à partir du nombre de mots de 6 lettres (*k-mers*) que ces séquences partagent dans ce nouvel alphabet (cf. MUSCLE);

MAFFT: Alignement multiple

3ème ETAPE

Les séquences sont ensuite alignées progressivement en suivant l'ordre indiqué par l'arbre de guidage.

MAFFT: Alignement multiple

Plusieurs programmes sont proposés sur la page de garde du site Web de MAFFT.

Ce que nous venons de décrire correspond à l'option FFT-NS-1.

Ce nom un peu barbare signifie Fast Fourier Transform-New Scoring matrix-1 step.

MAFFT: Alignement multiple

Contrairement à ClustalW mais comme MUSCLE, MAFFT peut occasionnellement procéder à un deuxième passage.

Dans ce dernier, l'alignement réalisé précédemment sert à recalculer la distance entre chaque paire de séquences, un nouvel arbre de guidage et un nouvelle alignement multiple.

Cette option s'appelle FFT-NS-2.

MAFFT: Alignement multiple

De manière similaire à MUSCLE, MAFFT peut procéder à un raffinement de l'alignement.

Dans ce cas, l'arbre de guidage est scindé en deux, puis les deux moitiés sont réalignées.

On recommence ainsi tant que le score d'alignement s'améliore.

On procède alors à un nombre *i* d'itérations *i* étant inconnu *a priori*.

Cette option porte le nom de FFT-NS-i.

On peut, sur la page de garde de MAFFT, limiter à deux le nombre d'itérations (« two cycles only »).

MAFFT: Alignement multiple

Il faut par ailleurs noter que le site de MAFFT propose des programmes fondés non pas sur la transformées de Fourier rapide, mais sur l'algorithme de programmation dynamique.

Ainsi le programme nommé G-INS-i aligne les paires de séquences suivant l'algorithme global de Needlman-Wunsch, comme ClustalW, calcule un arbre de guidage, aligne toutes les séquences suivant cet arbre et procède enfin à un raffinement de l'alignement comme décrit ci-avant.

Les programmes L-INS-i et E-INS-i procèdent de la même façon, mais avec l'algorithme d'alignement local de Smith-Waterman.

Bien entendu, ces programmes, nettement plus lents, ne conviennent pas pour un grand nombre de séquences.

Enfin, le programme Q-INS-i est spécifiquement dédié à l'alignement de séquences d'ARN.

MAFFT

Nombreuses options de MAFFT

1. Mode basique, rapide — juste progressif
   a) FFTNS1 (fftns --retree 1)
   b) FFTNS2 (fftns) (same as mafft --retree 2)

OK jusqu'à 1 000 séquences facilement alignables

2. Mode intermédiaire — progressif + itérations
   a) FFTNSI (fftnsi) default two cycles, or e.g. fftnsi --maxiterate 1000
   b) NWNSI (nwnsi) same as FFTNSI, but no FFT, Needleman-Wunsch only.

OK entre 100 et 500 séquences

3. Mode avancé — progressif + itérations + consistance (cf. T-Coffee)
   a) EINSI (einsi) Smith-Waterman (plusieurs régions similaires même ordre)
   b) LINSI (linsi) Smith-Waterman stricte (1 région similaires)
   c) GINSI (linsi) global Needleman-Wunsch

Exercice 4

Types of multiple sequence alignment and corresponding algorithms.

Types of MSA alignment	MSA algorithms
Pairwise alignment	Needleman-Wunsch, k-mer, k-tuple, and Smith-Waterman algorithms.
Progressive alignment	Clustal Omega, ClustalW, MAFFT, Kalign, Probalign, MUSCLE, Dialign, ProbCons, and MSAProbs.
Iterative progressive alignment	PRRP, MUSCLE, DIALIGN, SAGA, and T-COFFEE.
Homology search tools	BLAST, PSI-BLAST, and FASTA.
Structure incorporating alignment	3D-COFFEE, EXPRESSO, and MICAlign.
Motif alignment	PHI-Blast, GLAM2.
Short-read alignment	Bowtie, Maq, and SOAP.

http://www.hindawi.com/journals/isrn/2013/615630/

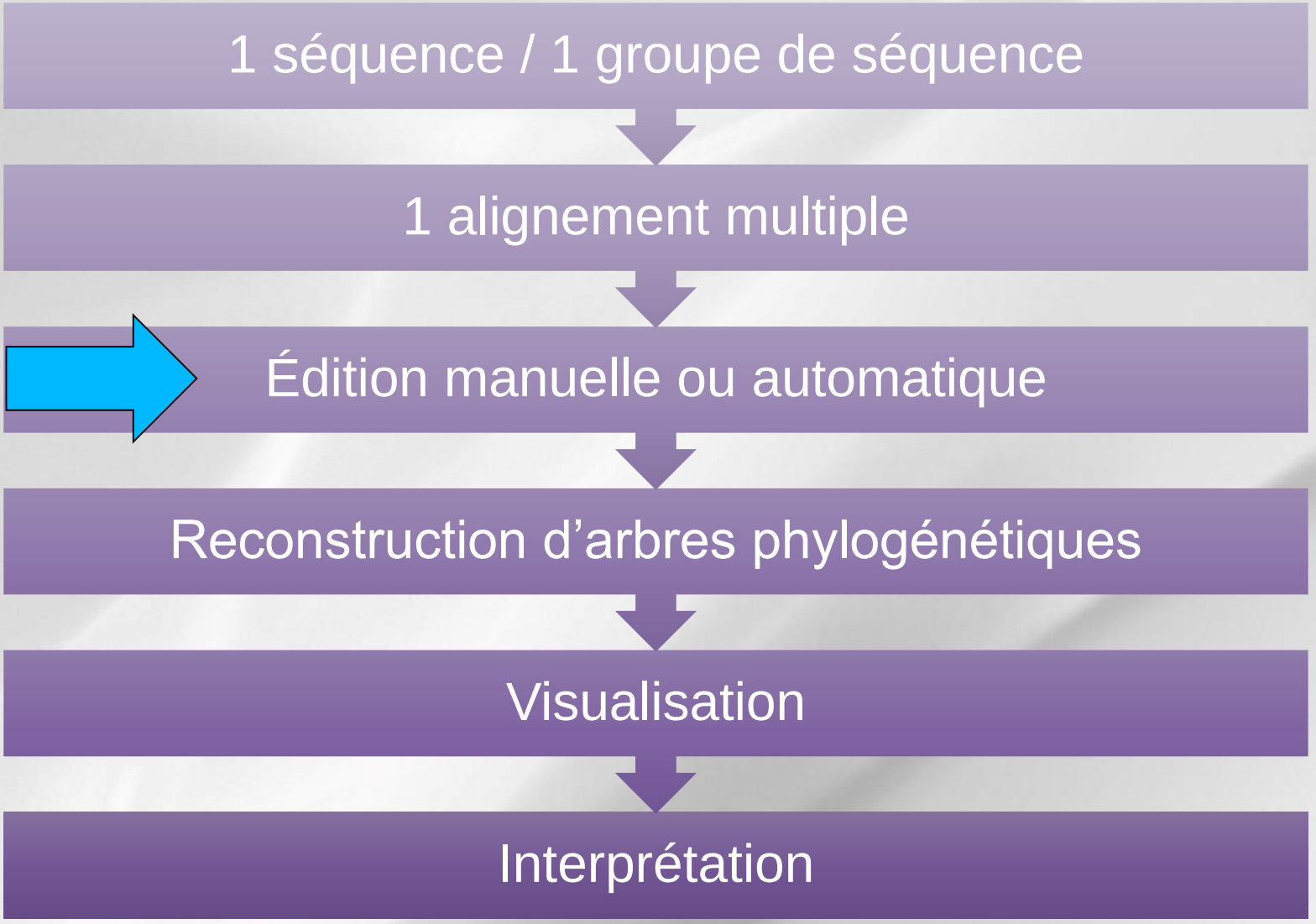
Où trouver les aligneurs?

- ClustalO, (http://www.clustal.org/omega/)
- ClustalW2, v2.1 (http://www.clustal.org)
- DIALIGN 2.2.1 (http://dialign.gobics.de/)
- FSA 1.15.5 (http://sourceforge.net/projects/fsa/)
- Kalign 2.04 (http://msa.sbc.su.se/cgi-bin/msa.cgi)
- MAFFT 6.857 (http://mafft.cbrc.jp/alignment/software/source.html)
- MSAProbs 0.9.4 (http://sourceforge.net/projects/msaprobs/files/)
- MUSCLE version 3.8.31 posted 1 May 2010 (http://www.drive5.com/muscle/downloads.htm)
- PRANK v.100802, 2 August 2010 (http://www.ebi.ac.uk/goldman-srv/prank/src/prank/)
- Probalign v1.4 (http://cs.njit.edu/usman/probalign/)
- PROBCONS version 1.12 (http://probcons.stanford.edu/download.html)
- T-Coffee Version 8.99 (http://www.tcoffee.org/Projects_home_page/t_coffee_home_page.html#DOWNLOAD)

Où en sommes nous ?

1. **Introduction générale à la phylogénie.**
2. **Acquisition du jeu de données.**
3. **L'alignement en détaillant les notions de:**
   - d'alignement par paires,
   - de score,
   - de matrice de substitution,
   - de programmation dynamique,
   - d'alignement global et local
4. **Les algorithmes d'alignements multiple**
   a) alignement multiple optimal
   b) alignement multiple progressif
      - ClustalW
      - Prank
   c) alignement multiple itératif
      - Dialign
      - Muscle
      - Mafft
5. **Edition des alignements multiples**

Les différentes étapes



Édition de l'alignement multiple

« Nothing in Biology makes sense except in the light of evolution »
Theodosius Dobzhansky (1900-1975)

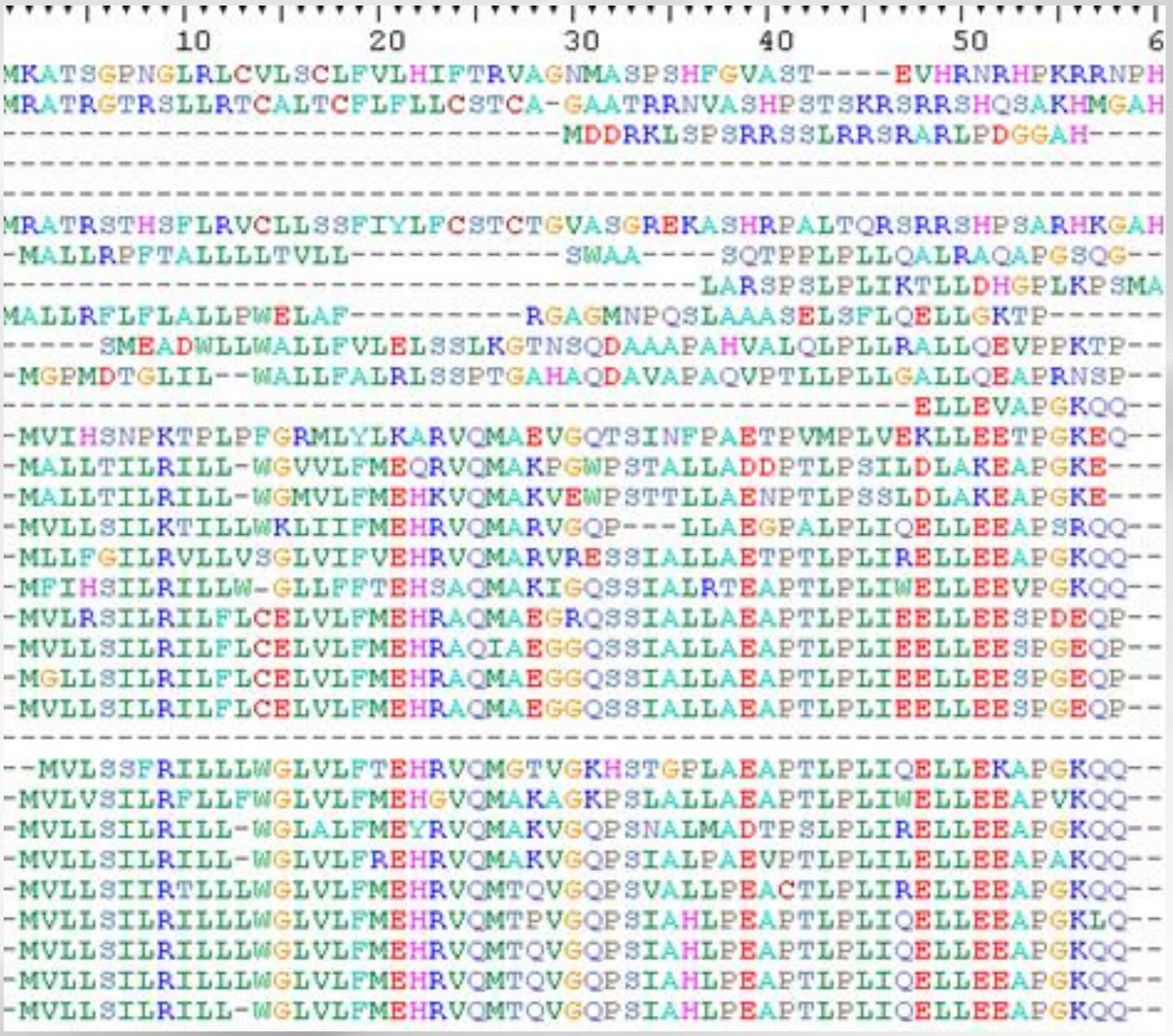
« Nothing in Evolution makes sense except in the light of a good phylogenetic tree » Vincent Laudet.

Édition de l'alignement multiple

- La qualité de l'alignement est essentielle
- Chaque colonne de l'alignement (site) = résidus homologues (nucléotides, aminoacides)
- Parties non « fiables » de l'alignement ⇨ supprimées de l'analyse phylogénétique
- La plupart des méthodes de reconstruction d'arbres ne tiennent compte que des substitutions, les brèches ou gaps (événements d'insertion/délétion) ne sont pas utilisées

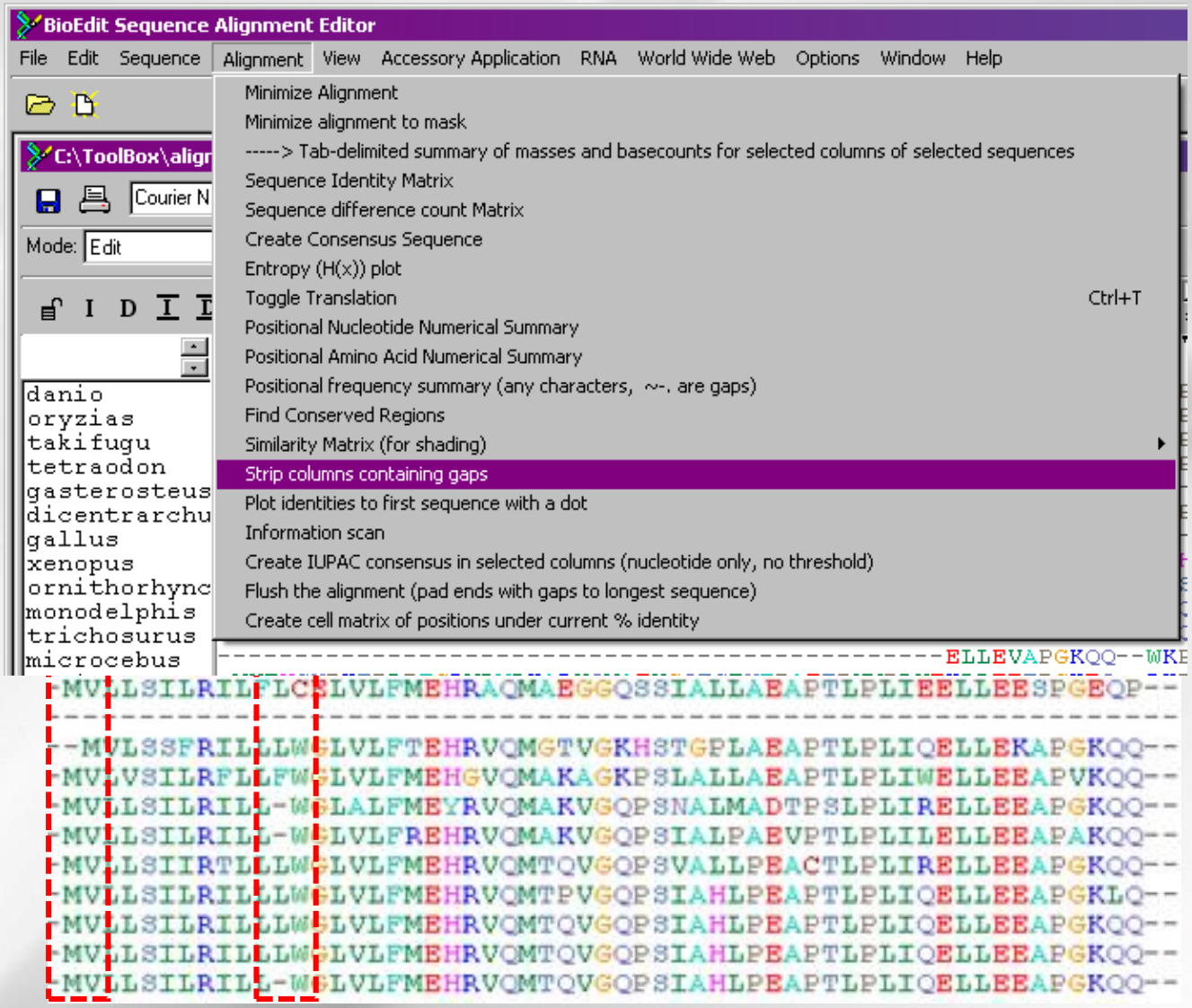
Édition manuelle

1. Vérification de l'alignement et comparaison de méthodes (MUSCLE et ClustalO par ex.)
2. Suppression des sites où il y a au moins un gap
3. Suppression des zones trop divergentes (étape difficile à la main)



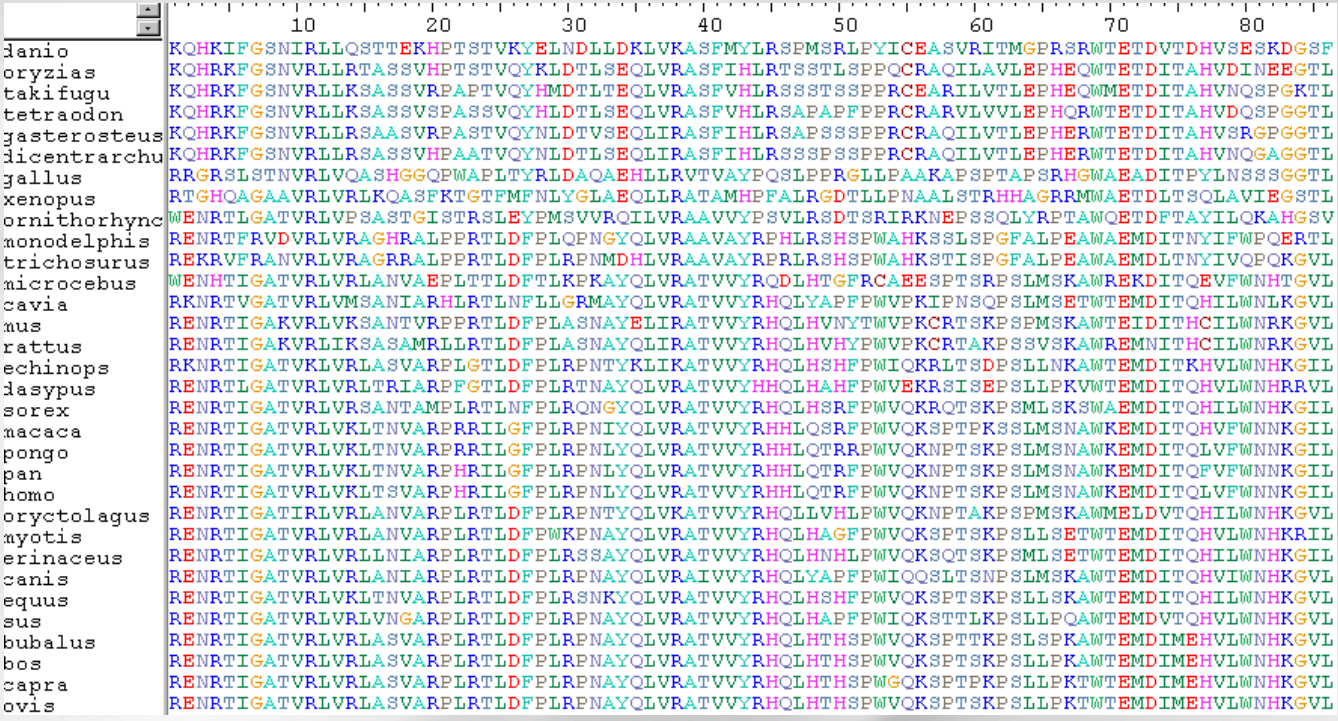
Édition manuelle

1. Vérification de l'alignement et comparaison de méthodes (MUSCLE et ClustalO par ex.)
2. Suppression des sites où il y a au moins un gap
3. Suppression des zones trop divergentes (étape difficile à la main)



Édition manuelle

1. Vérification de l'alignement et comparaison de méthodes (MUSCLE et ClustalO par ex.)
2. Suppression des sites où il y a au moins un gap
3. Suppression des zones trop divergentes (étape difficile à la main)



Édition automatique

GBLOCKS

http://molevol.cmima.csic.es/castresana/Gblocks.html

Le programme élimine les régions avec gaps et les régions trop divergentes

Car positions peuvent être non-homologues ou saturées par de multiple substituions

Conservation de blocs comme on peut le faire à la main mais ici fait de manière reproductible

Édition automatique

http://molevol.cmima.csic.es/castresana/Gblocks_server.html

Gblocks Server

Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis

About the Gblocks Server

Version 0.91b, January 2002

Copyright © Jose Castresana

Gblocks eliminates poorly aligned positions and divergent regions of a DNA or protein alignment so that it becomes more suitable for phylogenetic analysis. This server implements the most important features of the Gblocks program to make its use as simple as possible without loosing the functionality that it is necessary in most of the cases. Other options can be changed in the stand-alone program. You can see here an example output file showing the blocks selected from a protein alignment. Further information can be found in the online documentation. Please see the Gblocks page for citations.

Gblocks Server

Paste an alignment in NBRF/PIR or FASTA format:

Or upload an alignment file:

Parcourir...

Type of sequence:

DNA ○ || Protein ◉ || Codons ○

Options for a less stringent selection:

☐ Allow smaller final blocks
☐ Allow gap positions within the final blocks
☐ Allow less strict flanking positions

Options for a more stringent selection:

☐ Do not allow many contiguous nonconserved positions

Get Blocks Clear

Édition automatique

```
Gblocks nom_align -t=p -b1= -b2= -b3= -b4= -b5=n -b6=y -s=y -p=y
```

Stringent:

```
Gblocks nom_align -t=p -b1= -b2= -b3=8 -b4=10 -b5=n -b6=y -s=y -p=y -a=y
```

Relaxé:

```
Gblocks nom_align -t=p -b1= -b2= -b3=10 -b4=5 -b5=n -b6=y -s=y -p=y
```

Paramètres :

variables selon l'alignement — variables selon le nombre de séquences — variables selon stringent/relaxé

-b1: nombre minimum de séquences pour une position conservée
Par défaut: 50% du nombre de séquences +1
Valeurs autorisées: > à 50% du nombre de séquences ou ≤ au nombre total de séquences

-b2: nombre minimum de séquences pour une position « flanquante »
Par défaut: 85% du nombre de séquences
Valeurs autorisées: = ou > à b1

-b3: nombre maximum de positions contigües non conservées

-b4: longueur minimum d'un bloc

-b5: autorisation des gaps (n=none)

-b6: utilisation d'une matrice de similarité (yes par défaut)

-s: sélection des blocs (yes par défaut)

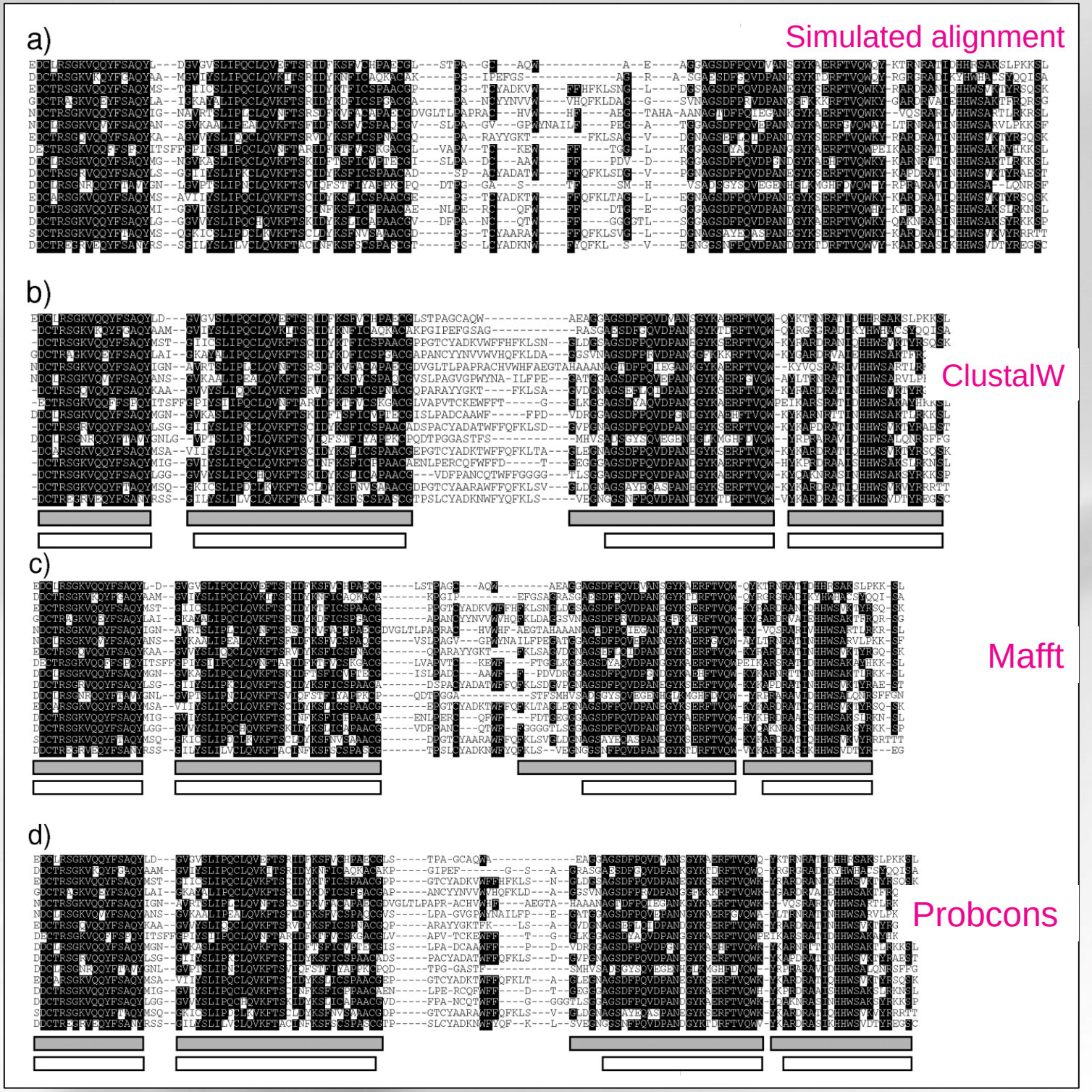
-p: résultats et fichiers de paramètres (yes par défaut)

Plateforme Bioinformatique Midi-Pyrénées

a) Fragment of a simulated alignment
b) the realignment of the same sequences (after gap removal) by ClustalW
c) Mafft
d) Probcons

blocks = fragments selected by **Gblocks** with **relaxed conditions (grey blocks)** and with **stringent conditions (white blocks).**

Positions of the alignments where more than 50% of the sequences are identical are shown with black boxes.



Guidance: un aligneur éditeur

Nucleic Acids Res. 2010 Jul;38(Web Server issue):W23-8. doi: 10.1093/nar/gkq443. Epub 2010 May 23.

GUIDANCE: a web server for assessing alignment confidence scores.

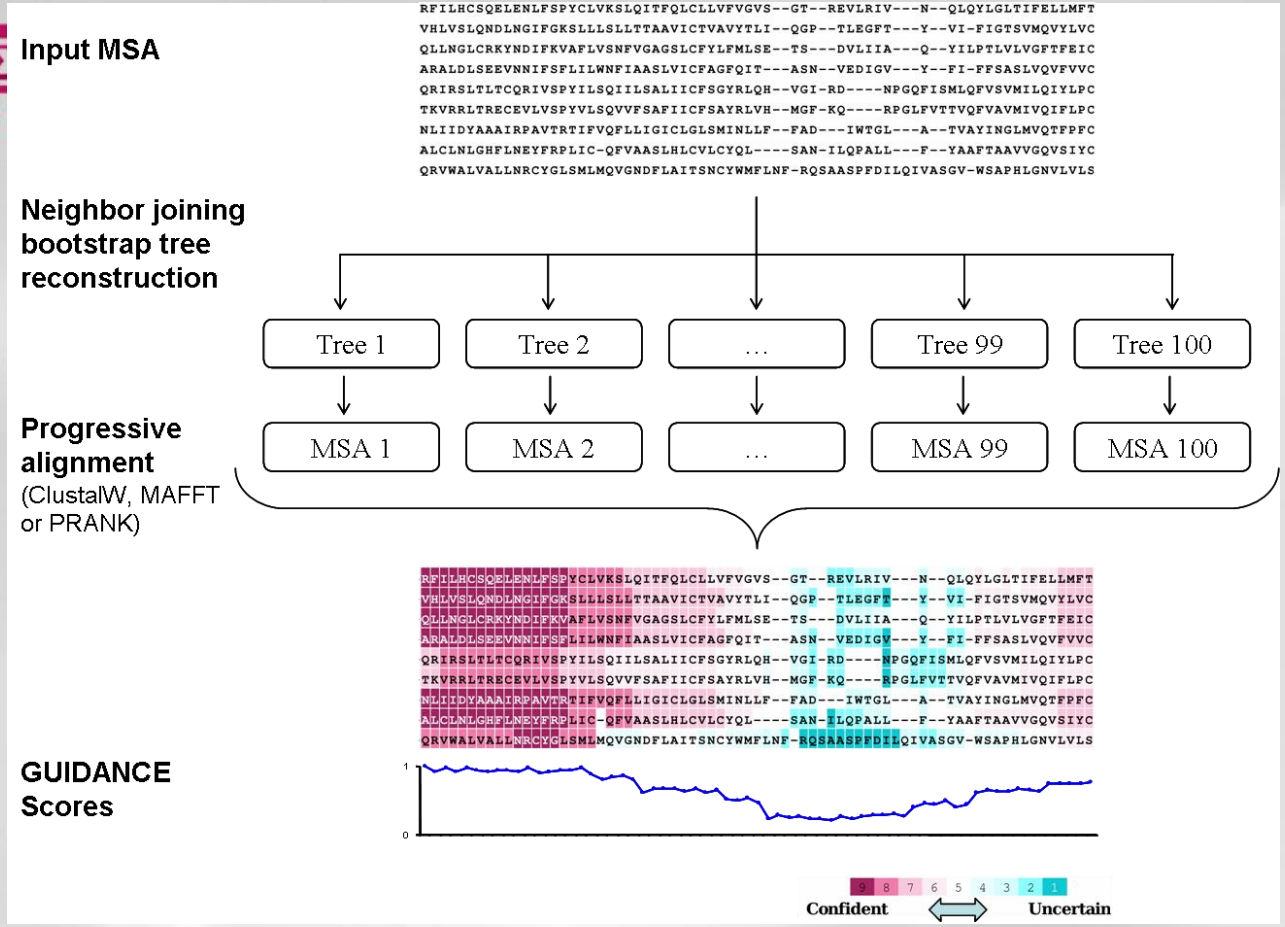
Penn O[1], Privman E, Ashkenazy H, Landan G, Graur D, Pupko T.

Guidance: un aligneur éditeur

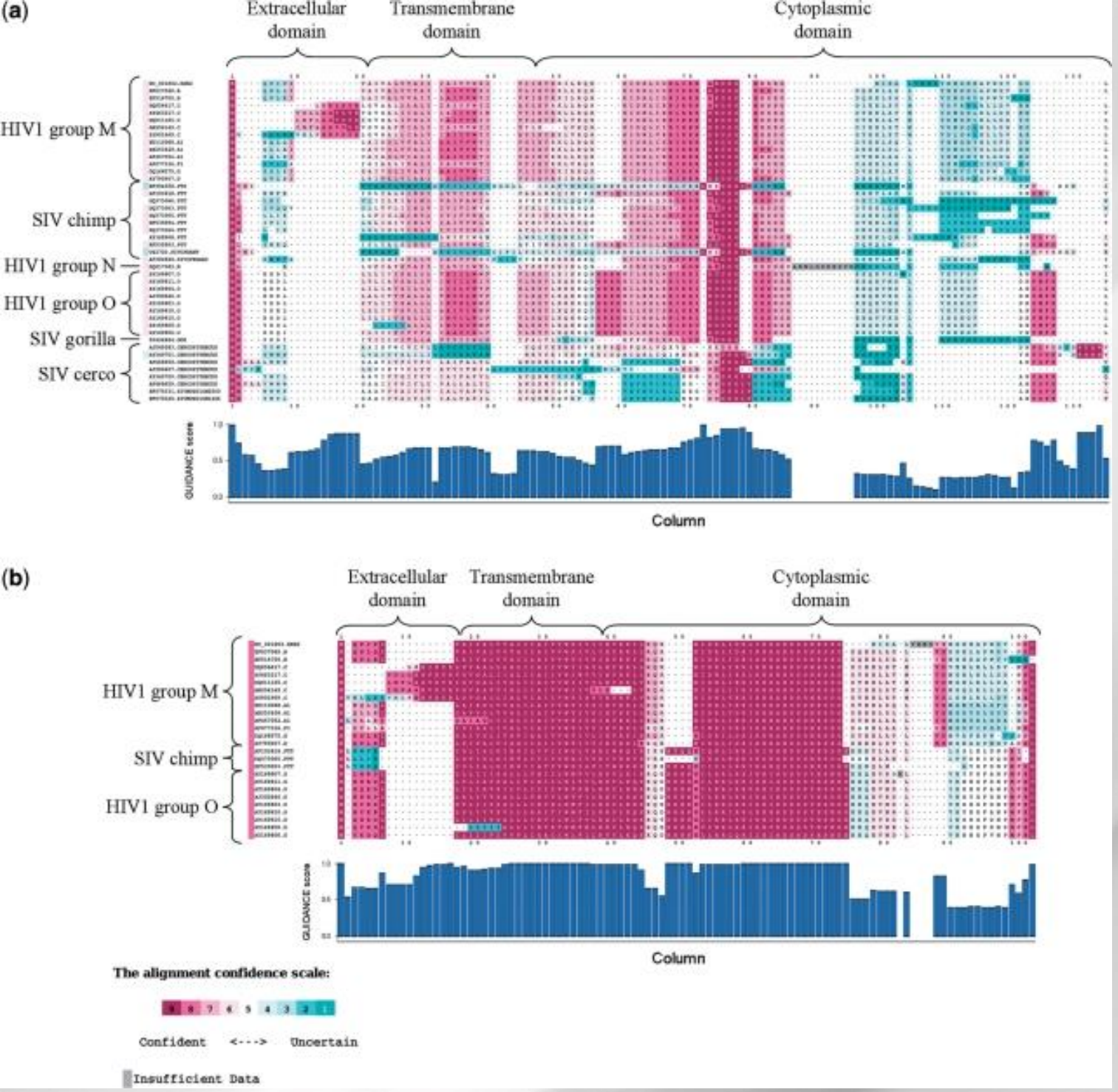
http://guidance.tau.ac.il/

Guidance prend en entrée des séquences non alignées et le serveur les alignent (ClustalW, MAFFT, MUSCLE, PAGAN et PRANK).

Il multiplie la production d'alignements multiples et score ensuite la « récurence » des sites alignés.

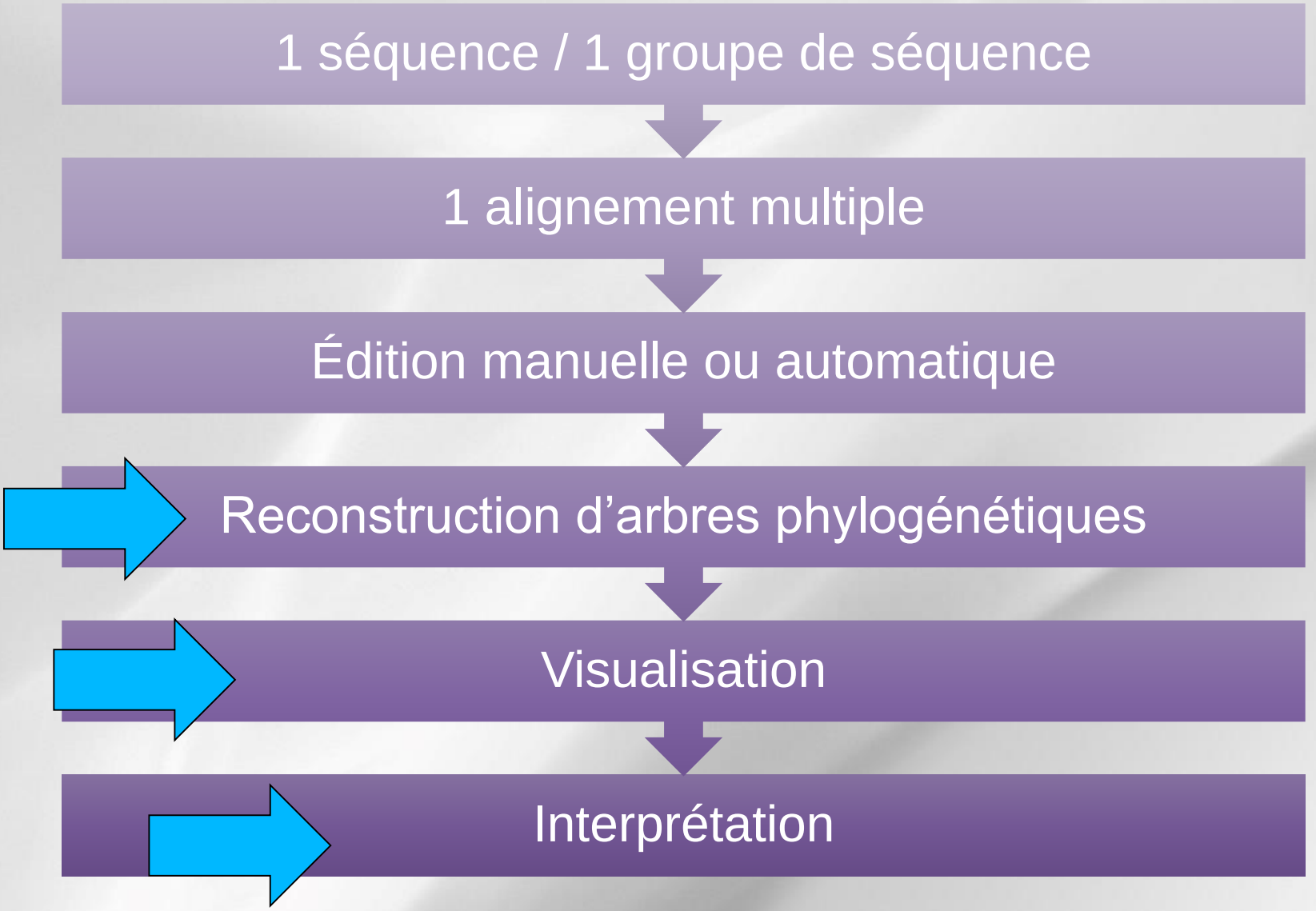


A schematic flowchart of the GUIDANCE algorithm. A base MSA is produced by any progressive alignment method. Bootstrap neighbor joining (NJ) trees are reconstructed and given as guide trees to the progressive alignment program, producing a set of MSAs. GUIDANCE scores are then calculated by comparing each MSA to the base MSA, and are color coded on each residue in the alignment.



An example of the GUIDANCE output. (**a**) Residue confidence scores are projected onto the MAFFT alignment of Vpu protein sequences from human and simian immunodeficiency viruses (HIV and SIV). Confidently aligned residues are colored in shades of magenta and pink, while uncertain residues are colored in shades of blue. Column scores are plotted below the alignment. (**b**) Dramatically improved alignment confidence after filtering low-scoring sequences and re-running GUIDANCE. Note the color-coding next to the sequence names before and after re-alignment.

Les différentes étapes



Reconstruction d'arbres phylogénétiques

4 familles principales de méthodes :

- Parcimonie
- Méthodes de distance
- Méthodes du maximum de vraisemblance
- Méthodes bayésiennes

Consensus d'arbre

Bootstrap

Enracinement

Plateforme Bioinformatique Midi-Pyrénées

Programme formation:Methods for phylogenetic trees construction

09:00 am to 10:30 am

Introduction
- Basic concepts
- Genetic distances and nucleotide substitution models

10:45 am to 12:45 pm

Phylogenetic inference methods (part 1)
- Distance methods
- Parcimony methods

14:00 pm to 15:45 pm

Phylogenetic inference methods (part 2)
- Maximum likehood methods
- Introduction to Bayesian methods

16:00 pm to 17:00 pm

Conclusion
- Testing tree topologies (bootstrap)
- How to choose a method ?

Exercice 1

Exercice 1 : Recherche d'homologues probables de BCRCA2

1. Sur le site http://blast.ncbi.nlm.nih.gov/Blast.cgi/ essayer de trouver des séquences similaires à la séquence **BCRCA2** dans différentes espèces plus ou moins proches de l'homme.
   Noter si vous utiliser la séquence nucléique ou protéique de BRCA2
   Quelles sont vos conclusions ?
   Dans quels organismes trouvez-vous des homologues possibles ?

2. Relancer la question précédente avec un blast différent
   Quelles sont les différences entre ces deux outils?

Exercice 1 : Recherche d'homologues probables de BCRCA2

3. Observez certains alignements
   Pourquoi n'a-t-on pas tous les alignements sur toute la longueur de la séquence de départ ?
   à quoi correspondent les + dans les alignements ?
   quels résultats vous paraissent significatifs ?
   Sur quoi vous basez-vous?
   Un alignement avec une E-value de l'ordre de e-50, vous parait-il un "bon" alignement ?

4. Relancer la question précédente avec le blast de votre choix en sélectionnant les séquences de poissons cartilagineux (taper "cartilaginous fishes (taxid:7777)" dans Organism)
   Qu'observez-vous ?
   Ces résultats vous paraissent-ils significatifs ?

5. Testons le RBH, relancer le blastp avec XP_007889510.1 de Callorhinchus milii contre les séquences « human »
   Est-ce que XP_007889510.1 est le RBH de votre séquence BCAR2 ?
   Vérifier si les ref sont identiques tel que le /db_xref="CCDS:CCDS9344.1 ou l'accession number NP_000050.2 etc…

Exercice 2

Exercice 2 : comparaison alignement local et global

Comparaison des séquences prot1 et prot2

```
>prot1
MVMEYLVLEKRLKRLREVLEKRQKDLIVFADNVKNEHNFSAIVRTCDAVGVLYLYYYHAEGKKAKINEGI
TQGSHKWVFIEKVDNPVQKLLEFKNRGFQIVATWLSKESVNFREVDYTKPTVLVVGNELQGVSPEIVEIA
DKKIVIPMYGMAQSLNVSVATGIILYEAQRQREEKGMYSRPSLSEEEIQKILKKWAYEDVIKERKRTLST
S
>prot2
MVMEYLVLEKRLKRLREVLEKRQKDLIVFADNVKNEHNFSAIVRTCDAVATWLSKESVNFREVDYTKPTV
LVVGNELQGVSPEIVEIAVGVLYLYYYHAEGKKAKINEGIS
```

Exercice 2 : comparaison alignement local et global

On utilisera la suite "EMBOSS" sur ce site http://emboss.toulouse.inra.fr/ , http://emboss.bioinformatics.nl/ ou sur le portail bioinfo de l'Institut Pasteur mobyle.pasteur.fr/ .

1.Faire un dotplot de ces 2 séquences (dotmatcher)
qu'observez-vous ? Modifiez les paramètres et regardez les résultats.

2. Faire un alignement semi-global avec needle ou global avec stretcher
Qu'observez-vous ?
Combien y a t-il de gaps ?
A quoi correspond le pourcentage de similarité ?
Quels sont les paramètres de calcul du score ?
Modifiez-les et regardez en quoi l'alignement change.

Exercice 2 : comparaison alignement local et global

On utilisera la suite "EMBOSS" sur ce site http://emboss.bioinformatics.nl/ ou sur le portail bioinfo de l'Institut Pasteur mobyle.pasteur.fr/ .

1. Faire un dotplot de ces 2 séquences (dotmatcher)
   qu'observez-vous ? Modifiez les paramètres et regardez les résultats.

La prot 2 est fait de 3 domaines ABC qui se retrouvent sur prot 1 (qui est plus longue) selon le schéma ACB. En baissant la taille du mot on aperçoit un séquence répétée dans le domaine A.

2. Faire un alignement semi-global avec needle ou global avec stretcher
Qu'observez-vous ? On ne voit pas la conservation du domaine B
Combien y a t-il de gaps ? 106
A quoi correspond le pourcentage de similarité ? Mismatch + identité
Quels sont les paramètres de calcul du score ? Gap penaliy +gap extension + matrice de substitution.
Modifiez-les et regardez en quoi l'alignement change. Variation du score, du nb gap, de similarité et d'identité

Exercice 2 : comparaison alignement local et global

3. Faire un alignement local avec matcher
Qu'observez-vous ?
Demandez à voir d'autres alignements.
Puis modifier les paramètres pour impacter sur le score.

Comparez et expliquez les différences obtenues entre une méthode d'alignement global (stretcher ou needle) et une méthode d'alignement local (matcher).

Exercice 2 : comparaison alignement local et global

3. Faire un alignement local avec matcher
Qu'observez-vous ? 1 seul alignement local
Demandez à voir d'autres alignements. On demande pour matcher plusieurs résultats (au moins 4) et ainsi les autres alignement locaux apparaissent.
Puis modifier les paramètres pour impacter sur le score. On doit changer les pénalité de gap (ouverture et extension)

Comparez et expliquez les différences obtenues entre une méthode d'alignement global (stretcher ou needle) et une méthode d'alignement local (matcher).
Finalité différentes.

Exercice 3

Exercice 3.1

Voici 3 protéines : celle de Escherichia coli possède 2 fonctions enzymatiques (EC 4.1.1.48 et EC 5.3.1.24) et 2 protéines de Xylella fastidiosa ayant chacune une de ces 2 fonctions :

```
>trpC, EC:4.1.1.48 et EC:5.3.1.2, E. coli
MMQTVLAKIVADKAIWVEARKQQQPLASFQNEVQPSTRHFYDALQGARTAFILECKKASP
SKGVIRDDFDPARIAAIYKHYASAISVLTDEKYFQGSFNFLPIVSQIAPQPILCKDFIID
PYQIYLARYYQADACLLMLSVLDDDQYRQLAAVAHSLEMGVLTEVSNEEEQERAIALGAK
VVGINNRDLRDLSIDLNRTRELAPKLGHNVTVISESGINTYAQVRELSHFANGFLIGSAL
MAHDDLHAAVRRVLLGENKVCGLTRGQDAKAAYDAGAIYGGLIFVATSPRCVNVEQAQEV
MAAAPLQYVGVFRNHDIADVVDKAKVLSLAAVQLHGNEEQLYIDTLREALPAHVAIWKAL
SVGETLPAREFQHVDKYVLDNGQGGSGQRFDWSLLNGQSLGNVLLAGGLGADNCVEAAQT
GCAGLDFNSAVESQPGIKDARLLASVFQTLRAY

>EC:5.3.1.24, Xfa
MALAYGSECMNISPYRTRIKFCGMTRVGDVRLASELGVDAVGLIFASGSSRLLTVSAACA
IRRTVAPMVNVVALFQNNSADEIHTVVRTVRPTLLQFHGEEEDAFCRTFNVPYLKAIPMA
GAEAKRICTRTLYLKYPNAAGFIFDSHLKGGTGQTFDWSRLPIDLQHPFLLAGGITPENV
FDAIAATVPWGVDVSSGIELQPGIKDGDKMRQFVEEVRRADGRRLFGVA

>EC:4.1.1.48, Xfa
MSNILTKIIAWKVEEIAERLLHVSQAELVARCADLPTPRGFAGALQATIAHGDPAVIAEI
KKASPSKGVLREDFRPAEIAISYELGGASCLSVLTDVHFFKGHDDYLSQARDACTLPVLR
KDFTIDPYQVYEARVLGADCILLIVAALDDAQLVDLSGLALQLGMDVLVEVHDIDELERA
IQISAPLIGINNRNLSTFNVSLETTLTMKGLVPRDRLLVSESGILTSADVQRLRAAGVNA
FLVGEAFMRATEPGESLREMFFIT
```

Exercice 3.1

Aligner ces 3 séquences avec ces 3 programmes d'alignement multiple : ClustalW, Dialign, ClustalO

Lequel retourne l'alignement attendu ?

Tester Dialign avec Threshold=4 (seuil sur les segments initiaux)

Exercice 3.2

Aligner ces 4 séquences de tRNA synthétase de E. coli, avec différents programmes d'alignement et comparez-les.

```
>gi|2507434|sp|P00956.5|SYI_ECOLI RecName: Full=Isoleucine--tRNA ligase; AltName: Full=Isoleucyl-tRNA synthetase;
Short=IleRS
MSDYKSTLNLPETGFPMRGDLAKREPGMLARWTDDDLYGIIRAAKKGKKTFILHDGPPYANGSIHIGHSV
NKILKDIIVKSKGLSGYDSPYVPGWDCHGLPIELKVEQEYGKPGEKFTAAEFRAKCREYAATQVDGQRKD
FIRLGVLGDWSHPYLTMDFKTEANIIRALGKIIGNGHLHKGAKPVHWCVDCRSALAEAEVEYYDKTSPSI
DVAFQAVDQDALKAKFAVSNVNGPISLVIWTTTPWTLPANRAISIAPDFDYALVQIDGQAVILAKDLVES
VMQRIGVTDYTILGTVKGAELELLRFTHPFMGFDVPAILGDHVTLDAGTGAVHTAPGHGPDDYVIGQKYG
LETANPVGPDGTYLPGTYPTLDGVNVFKANDIVVALLQEKGALLHVEKMQHSYPCCWRHKTPIIFRATPQ
WFVSMDQKGLRAQSLKEIKGVQWIPDWGQARIESMVANRPDWCISRQRTWGVPMSLFVHKDTEELHPRTL
ELMEEVAKRVEVDGIQAWWDLDAKEILGDEADQYVKVPDTLDVWFDSGSTHSSVVDVRPEFAGHAADMYL
EGSDQHRGWFMSSLMISTAMKGKAPYRQVLTHGFTVDGQGRKMSKSIGNTVSPQDVMNKLGADILRLWVA
STDYTGEMAVSDEILKRAADSYRRIRNTARFLLANLNGFDPAKDMVKPEEMVVLDRWAVGCAKAAQEDIL
KAYEAYDFHEVVQRLMRFCSVEMGSFYLDIIKDRQYTAKADSVARRSCQTALYHIAEALVRWMAPILSFT
ADEVWGYLPGEREKYVFTGEWYEGLFGLADSEAMNDAFWDELLKVRGEVNKVIEQARADKKVGGSLEAAV
TLYAEPELSAKLTALGDELRFVLLTSGATVADYNDAPADAQQSEVLKGLKVALSKAEGEKCPRCWHYTQD
VGKVAEHAEICGRCVSNVAGDGEKRKFA

>gi|2507435|sp|P07813.2|SYL_ECOLI RecName: Full=Leucine--tRNA ligase; AltName: Full=Leucyl-tRNA synthetase;
Short=LeuRS
MQEQYRPEEIESKVQLHWDEKRTFEVTEDESKEKYYCLSMLPYPSGRLHMGHVRNYTIGDVIARYQRMLG
KNVLQPIGWDAFGLPAEGAAVKNNTAPAPWTYDNIAYMKNQLKMLGFGYDWSRELATCTPEYYRWEQKFF
TELYKKGLVYKKTSAVNWCPNDQTVLANEQVIDGCCWRCDTKVERKEIPQWFIKITAYADELLNDLDKLD
HWPDTVKTMQRNWIGRSEGVEITFNVNDYDNTLTVYTTRPDTFMGCTYLAVAAGHPLAQKAAENNPELAA
FIDECRNTKVAEAEMATMEKKGVDTGFKAVHPLTGEEIPVWAANFVLMEYGTGAVMAVPGHDQRDYEFAS
KYGLNIKPVILAADGSEPDLSQQALTEKGVLFNSGEFNGLDHEAAFNAIADKLTAMGVGERKVNYRLRDW
GVSRQRYWGAPIPMVTLEDGTVMPTPDDQLPVILPEDVVMDGITSPIKADPEWAKTTVNGMPALRETDTF
DTFMESSWYYARYTCPQYKEGMLDSEAANYWLPVDIYIGGIEHAIMHLLYFRFFHKLMRDAGMVNSDEPA
KQLLCQGMVLADAFYYVGENGERNWVSPVDAIVERDEKGRIVKAKDAAGHELVYTGMSKMSKSKNNGIDP
QVMVERYGADTVRLFMMFASPADMTLEWQESGVEGANRFLKRVWKLVYEHTAKGDVAALNVDALTENQKA
LRRDVHKTIAKVTDDIGRRQTFNTAIAAIMELMNKLAKAPTDGEQDRALMQEALLAVVRMLNPFTPHICF
TLWQELKGEGDIDNAPWPVADEKAMVEDSTLVVVQVNGKVRAKITVPVDATEEQVRERAGQEHLVAKYLD
GVTVRKVIYVPGKLLNLVVG
```

Exercice 3.2

```
>gi|135149|sp|P00959.2|SYM_ECOLI RecName: Full=Methionine--tRNA ligase; AltName: Full=Methionyl-tRNA synthetase;
Short=MetRS
MTQVAKKILVTCALPYANGSIHLGHMLEHIQADVWVRYQRMRGHEVNFICADDAHGTPIMLKAQQLGITP
EQMIGEMSQEHQTDFAGFNISYDNYHSTHSEENRQLSELIYSRLKENGFIKNRTISQLYDPEKGMFLPDR
FVKGTCPKCKSPDQYGDNCEVCGATYSPTELIEPKSVVSGATPVMRDSEHFFFDLPSFSEMLQAWTRSGA
LQEQVANKMQEWFESGLQQWDISRDAPYFGFEIPNAPGKYFYVWLDAPIGYMGSFKNLCDKRGDSVSFDE
YWKKDSTAELYHFIGKDIVYFHSLFWPAMLEGSNFRKPSNLFVHGYVTVNGAKMSKSRGTFIKASTWLNH
FDADSLRYYYTAKLSSRIDDIDLNLEDFVQRVNADIVNKVVNLASRNAGFINKRFDGVLASELADPQLYK
TFTDAAEVIGEAWESREFGKAVREIMALADLANRYVDEQAPWVVAKQEGRDADLQAICSMGINLFRVLMT
YLKPVLPKLTERAEAFLNTELTWDGIQQPLLGHKVNPFKALYNRIDMRQVEALVEASKEEVKAAAAPVTG
PLADDPIQETITFDDFAKVDLRVALIENAEFVEGSDKLLRLTLDLGGEKRNVFSGIRSAYPDPQALIGRH
TIMVANLAPRKMRFGISEGMVMAAGPGGKDIFLLSPDAGAKPGHQVK

>gi|2507438|sp|P07118.2|SYV_ECOLI RecName: Full=Valine--tRNA ligase; AltName: Full=Valyl-tRNA synthetase; Short=ValRS
MEKTYNPQDIEQPLYEHWEKQGYFKPNGDESQESFCIMIPPPNVTGSLHMGHAFQQTIMDTMIRYQRMQG
KNTLWQVGTDHAGIATQMVVERKIAAEEGKTRHDYGREAFIDKIWEWKAESGGTITRQMRRLGNSVDWER
ERFTMDEGLSNAVKEVFVRLYKEDLIYRGKRLVNWDPKLRTAISDLEVENRESKGSMWHIRYPLADGAKT
ADGKDYLVVATTRPETLLGDTGVAVNPEDPRYKDLIGKYVILPLVNRRIPIVGDEHADMEKGTGCVKITP
AHDFNDYEVGKRHALPMINILTFDGDIRESAQVFDTKGNESDVYSSEIPAEFQKLERFAARKAVVAAVDA
LGLLEEIKPHDLTVPYGDRGGVVIEPMLTDQWYVRADVLAKPAVEAVENGDIQFVPKQYENMYFSWMRDI
QDWCISRQLWWGHRIPAWYDEAGNVYVGRNEDEVRKENNLGADVVLRQDEDVLDTWFSSALWTFSTLGWP
ENTDALRQFHPTSVMVSGFDIIFFWIARMIMMTMHFIKDENGKPQVPFHTVYMTGLIRDDEGQKMSKSKG
NVIDPLDMVDGISLPELLEKRTGNMMQPQLADKIRKRTEKQFPNGIEPHGTDALRFTLAALASTGRDINW
DMKRLEGYRNFCNKLWNASRFVLMNTEGQDCGFNGGEMTLSLADRWILAEFNQTIKAYREALDSFRFDIA
AGILYEFTWNQFCDWYLELTKPVMNGGTEAELRGTRHTLVTVLEGLLRLAHPIIPFITETIWQRVKVLCG
ITADTIMLQPFPQYDASQVDEAALADTEWLKQAIVAVRNIRAEMNIAPGKPLELLLRGCSADAERRVNEN
RGFLQTLARLESITVLPADDKGPVSVTKIIDGAELLIPMAGLINKEDELARLAKEVAKIEGEISRIENKL
ANEGFVARAPEAVIAKEREKLEGYAEAKAKLIEQQAVIAAL
```

Ces séquences présentent 2 motifs propres aux tRNA synthétases de type I : *HIGH* et *KMSKS*.
Les trouvez-vous ?

Comparaison DIALIGN CLUSTALW

```
SYI_ECOLI/1-938        51  filhdgPPYA NGSIHIGHSV NKILKDIIVK SKGLSGYDSP YVPGWDCHGL
SYL_ECOLI/1-860        42  -------PYP SGRLHMGHVR NYTIGDVIAR YQRMLGKNVL QPIGWDAFGL
SYM_ECOLI/1-677        15  -------PYA NGSIHLGHML EHIQADVWVR YQRMRGHEVN FICADDAHGT
SYV_ECOLI/1-951        41  ------PPNV TGSLHMGHAF QQTIMDTMIR YQRMQGKNTL WQVGTDHAGI
                               0000000555 7999999978 8888888888 8888886665 5555554444

   ..........


SYI_ECOLI/1-938       589  VLTHGFTVDG QGRKMSKSIG NTVSPQDVMN K--------- ----------
SYL_ECOLI/1-860       617  ---------- -MSKMSKSKN NGIDPQVMVE R--------- ----------
SYM_ECOLI/1-677       331  ---------- -GAKMSKSRG TFIKASTWLN H--------- ----------
SYV_ECOLI/1-951       541  VYMTGLIRDD EGQKMSKSKG NVIDPLDMVD gislpellek rtgnmmqpql
                               1111111111 1224444444 4443333333 2000000000 0000000000
```

Exemple de sortie d'alignement effectué sur le site Web de DIALIGN.

Toutes ces tRNA synthétases sont de type I et ont la particularité d'être globalement similaires, avec cependant la présence de longues insertions pour *SYL_ECOLI*. Par ailleurs, ces séquences sont connues pour posséder deux motifs caractéristiques propres aux tRNA synthétases de type I, **HIGH** et **KMSKS**.

Comparaison DIALIGN CLUSTALW

```
SYL_ECOLI/1-860     1 --------------MQEQYRPEEIESKVQLHWDEKRTFEVT--EDESKEKYYCLSMLPYPSGRLHMGHVRNYTIGDV  61
SYV_ECOLI/1-951     1 --------------MEKTYNPQDIEQPLYEHWEKQGYFKPN--GDESQESFCIMIPPPNVTGSLHMGHAFQQTIMDT  61
SYI_ECOLI/1-938     1 MSDYKSTLNLPETGFPMRGDLAKREPGMLARWTDDDLYGIIRAAKKGKKTFILHDGPPYANGSIHIGHSVNKILKDI  77
SYM_ECOLI/1-677     1 ------------------------------------------MTQVAKKILVTCALPYANGSIHLGHMLEHIQADV  34
                        .........

SYL_ECOLI/1-860   335 VMAVPGHDQRDYEFASKYGLNIKPVILAADGSEPDLSQQALTEKGVLFNSGEFNGLDHEAAFNAIADKLTAMGVGER 411
SYV_ECOLI/1-951   343 AVVAAVDALGLLEEIKPHDLTVPYGDRGGVVIEPMLTDQWYVRADVLAKPAVEAVENGDIQFVPKQYENMYFSWMR- 418
SYI_ECOLI/1-938   382 IVVALLQEKGALLHVEKMQHSYPCCWRHKTPIIFRATPQWFVSMDQKGLRAQSLKEIKGVQWIPDWGQARIESMVAN 458
SYM_ECOLI/1-677   286 STAELYHFIG-KDIVYFHSLFWPAMLEGSN---FRKPSNLFVHGYVTVNGAKMSKSRGTFIKASTWLNHFDADSLR- 357

SYL_ECOLI/1-860   412 KVNYRLRDWGVSRQRYWGAPIPMVTLEDGTVMPTPDD---QLPVILPEDVVMDGITSPIKADPEWAKTTVN---GMP 482
SYV_ECOLI/1-951   419 ----DIQDWCISRQLWWGHRIPAWYDEAGNVYVGRNEDEVRKENNLGADVVLRQDEDVLDTWFSSALWTFS---TLG 488
SYI_ECOLI/1-938   459 -----RPDWCISRQRTWGVPMSLFVHKDTEELH-------PRTLELMEEVAKRVEVDGIQAWWDLDAKEILGDEADQ 523
SYM_ECOLI/1-677   358 --------------YYYTAKLSSRIDDID--------------LNLEDFVQRVNADIVNKVVNLASRNAGFINKRF 405

SYL_ECOLI/1-860   483 ALRETDTFDTFMESSWYYARYTCPQYKEG--------MLDSEAANYWLPVD-IYIGGIEHAIMHLLYFRFFHKLMRD 550
SYV_ECOLI/1-951   489 NPENTDALRQFHPTSVMVSGFDIIFFWIARMIMMTMHFIKDENGKPQVPFHTVYNTGLIRDDEGQKMSKSKGNVIDP 565
SYI_ECOLI/1-938   524 YVKVPDTLDVWFDSGSTHSSVVDVRPEFAG--HAADMYLEGSDQHRGWFMSSLMISTAMKGKAPYRQVLTHGFTVDG 598
SYM_ECOLI/1-677   406 DGVLASELADPQLYKTFTDAAEVIG----------------------------------------------------- 430

SYL_ECOLI/1-860   551 AGMVNSDEPAKQLLCQG--MVLADAFYYVGENGERNWVS----------------------PVDAIVERDEKGRIVK 603
SYV_ECOLI/1-951   566 LDMVDGISLPELLEKRTGNMMQPQLADKIRKRTEKQFPNGIEPHGTDALRFTLAALASTGRDINWDMKRLEGYRNFC 642
SYI_ECOLI/1-938   599 QGRKMSKSICNTVSPQD-------VMNKLGADILRLMVASTDYTG----------------EMAVSDEILKRAADSY 652
SYM_ECOLI/1-677   431 ------------------------------------------------------------EAWESREFGKAVREI 445

SYL_ECOLI/1-860   604 AKDAAGHELVYTGHSKMSKSKNNGI-DPQVMVERYGADTVRLFMMFASPADMTLEWQESG----VEGANRFLKRVWK 675
SYV_ECOLI/1-951   643 NKLWNASRFVLMNTEGQDCGFNGGE-MTLSLADRWILAEFNQTIKAYREALDSFRFDIAAGILYEFTWNQFCDWYLE 718
SYI_ECOLI/1-938   653 RRIRNTARFLLANLNGFDPAKDMVKPEEMVVLDRWAVGCAKAAQEDILKAYEAYDFHEVVQRLMRFCSVEMGSFYLD 729
SYM_ECOLI/1-677   446 MALADLANRYVDEQAPWVVAKQEGRDADLQAICSMGINLFRVLMTYLKPVLPKLTERAEAFLNTELTWDGIQQPLLG 522
```

Alignement effectué par ClustalW montrant le non-alignement du motif KMSKS.

Exercice 4

Exercice 4

Compare the performance of 5 different multiple alignment methods (mafft, muscle, dialign, clustalw, clustalO) for aligning a set of proteins

Exercice 4

Comparison of multiple alignment programs

In this part of the exercise you will use a data set of 11 alternatively-spliced gene products from the human erythrocyte membrane protein band 4.1 (EPB).

In addition to the unaligned data set we also have a pre-aligned optimal alignment.

The goal of this short exercise is to compare how well five different popular multiple alignment programs perform when attempting to align a set of proteins that are identical except for having different deletions.

You can run the 5 MSA tools via http://mobyle.pasteur.fr/cgi-bin/portal.py#welcome or with following command lines:

Exercice 4

1. **Align the EPB sequences using mafft:**
   mafft EPB.fasta > EPB.mafft.fasta
2. **Align the EPB sequences using muscle:**
   muscle -in EPB.fasta -out EPB_muscle.fasta
3. **Align the EPB sequences using dialign:**
   on webserver http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::dialign
4. **Align the EPB sequences using clustalw:**
   clustalw2 -infile=EPB.fasta.txt -type=protein -outfile=EPB.clustalw.fasta -output=fasta
5. **Align the EPB sequences using clustalO:**
   clustalo -i EPB.fasta.txt -o EPB.clustalo.fasta

Exercice 4

6. **Have a look at the original, unaligned data:**

   with a your favorite text editor (seaview or clustalX if installed) look EPB.fasta

   This is the unaligned set of sequences.
   The sequences are displayed on the screen with names on the left hand side, and the sequences themselves on the right.
   With clustalX:
   Residues are colored according to amino acid characteristics.
   It is possible to resize the window by "pulling" at the edges, so you can fit all lines in one window. Make sure to pull the window so it's as wide as possible. A scroll-bar at the bottom of the window allows you to move along the alignment (the sequences are too long to fit in the horizontal direction).
   Beneath the sequences there is a ruler starting at 1 for the first residue position.
   Below this is a graphical indication of the degree of conservation in each column of the alignment. A high score indicates a well-conserved column; a low score indicates low conservation. Since these sequences are not aligned (they are all just lined up starting with their first residues), most values are quite low.
   NOTE: In this part of the exercise we are only using clustalx as a sequence/alignment viewer, not as an alignment program.

Exercice 4

7. **Have a look at the optimally aligned data:**

   on mobile Pasteur / display / alignment / mview_alignment

This is the optimally aligned alignment.

Drag your window and scroll along to see the entire alignment.

Keep this window hanging around somewhere so you can compare the other alignments to it.

Exercice 4

8. Have a look at the four different alignments:

with mobile Pasteur / display / alignment / mview_alignment

clustalw: clustalx EPB.clustalw.fasta &

clustalo: clustalx EPB.clustalo.fasta &

dialign: clustalx EPB.dialign.fasta &

muscle: clustalx EPB.muscle.fasta &

mafft: clustalx EPB.mafft.fasta &

Question: Which, if any, of the five alignment methods got the alignment entirely correct?

You should note that this was just one particular form of test. On a different problem the relative performance of the alignment methods could well be different. However, you should also note that this was a fairly simple problem, and one where you could easily see artefacts. That will not be the case for most real biological data sets.

Sources

Cours: http://pbil.univ-lyon1.fr/members/perriere/cours/EGOISt/tp.html

file:////147.99.97.157/home/grpascal/Documents/Redaction/Enseignements/Enseignements_Catherine/TP1.html

http://evomics.org/learning/phylogenetics/paml/

ftp://pbil.univ-lyon1.fr/pub/cours/lesecque/TP_PAML/TP_EvoMol_n4_ETU.pdf

ftp://pbil.univ-lyon1.fr/pub/cours/lesecque/TP_PAML/tp_evol_Mol.pptx.pdf

ftp://biomol.univ-lyon1.fr/pub/cours/marais/ms_HDR_aout09.pdf

ftp://pbil.univ-lyon1.fr/pub/cours/fablet/EvolutionMoleculaire/sujet_TP-ET.pdf

http://en.wikipedia.org/wiki/Multiple_sequence_alignment

http://www.unige.ch/sciences/biologie/biani/msg/files/public/CodonModels.pdf

http://biologie.univ-mrs.fr/upload/p202/Cours3_alignement.pdf

Cours de Manolo Gouy (PBIL)

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0096696

Sources

Livre: « bioinformatique principes d'utilisation des outils » Ouvrage coordonné par **Denis Tagu** et **Jean-Loup Risler,** publié en 2010 aux éditions **Quæ** dans la collection **Savoir faire** –

See more at: http://bioinfo-fr.net/jai-lu-bio-informatique-principe-dutilisation-des-outils#sthash.dRYts9Cx.dpuf

Pierre Darlu et Pascal Tassy version électronique de « *La Reconstruction phylogénétique. Concepts et Méthodes* »

Livre : « Concepts et méthodes en phylogénie moléculaire » Auteur(s) : Guy Perrière , Céline Brochier-Armanet Editeur : Springer

Livre: "Phylogenetic handbook", http://www.kuleuven.be/aidslab/phylogenybook/home.html - See more at: http://bioinfo-fr.net/jai-lu-concepts-et-methodes-en-phylogenie-moleculaire#sthash.dBHIfXty.dpuf

W. Mount. Bioinformatics: Sequence and Genome Analysis. (2004) pp. 692. http://www.bioinformaticsonline.org/ (Code BU: 572.86 MOU)

Perrière et Brochier-Armanet: Concepts et méthodes en phylogénie moléculaire, 2010, Springer (BU:570.11 PER)