Intervalles communs de gènes

du modèle mathématique à l'arbre de la vie...

Annie Chateau

LIRMM (Montpellier)

29 avril 2008

Introduction

Le Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier :

- 95 enseignants-chercheurs et 29 chercheurs CNRS
- 28 ingénieurs, techniciens ou administratifs
- 134 doctorants
- 3 chercheurs contractuels (hors doctorants).

Deux formations doctorales

Équipe "Méthodes et Algorithmes pour la Bioinformatique" :

8 permanents, 3 techniciens, 7 membres associés, 1 post-doctorant, 11 doctorants, 2 stagiaires de master

Une plate-forme de calcul reconnue au niveau international (+1000 citations pour l'outil PhyML)

Introduction

C’est quoi la bioinformatique ?

Introduction

C’est quoi la bioinformatique ?

- C’est ce que font les bioinformaticiens ?

Introduction

C’est quoi la bioinformatique ?

- C’est ce que font les bioinformaticiens ?

Et que font les bioinformaticiens ? De la biologie et de l’informatique ?

Introduction

C’est quoi la bioinformatique ?

- C’est ce que font les bioinformaticiens ?

Et que font les bioinformaticiens ? De la biologie et de l’informatique ?

- Oui

Introduction

C’est quoi la bioinformatique ?

- C’est ce que font les bioinformaticiens ?

Et que font les bioinformaticiens ? De la biologie et de l’informatique ?

- Oui

- Et non. . .

Mais encore. . .

A On part d'un problème biologique, fourni par un biologiste.

Est-ce que deux morceaux d'ADN donnent la même protéine ?

Mais encore. . .

A On part d'un problème biologique, fourni par un biologiste.

Est-ce que deux morceaux d'ADN donnent la même protéine ?

B Le bioinformaticien créé un modèle mathématique de ce problème = travail à la fois mathématique et biologique

Un morceau d'ADN est un mot sur un alphabet de 4 lettres A, C, G, T

Mais encore. . .

- Le bioinformaticien met au point un algorithme de traitement des données modélisées = travail d'analyse informatique

  Algorithme qui donne un score de similitude entre deux séquences

Mais encore. . .

C Le bioinformaticien met au point un algorithme de traitement des données modélisées = travail d'analyse informatique

Algorithme qui donne un score de similitude entre deux séquences

D L'informaticien programme et produit les résultats du programme.

Les deux séquences obtiennent un score de similitude de 60%

Mais encore. . .

E Les résultats sont envoyés au biologiste qui les interprète **avec** le bioinformaticien.

Le bioinformaticien : est-ce que 60% ça veut dire que les protéines sont identiques ?
Le biologiste : 60% de quoi ?
. . .

En résumé

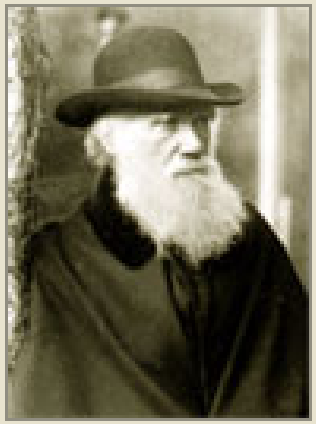
« Faire de la bioinformatique » c'est :

- Comprendre les notions élémentaires en biologie et savoir discuter avec un biologiste sans dictionnaire
- Avoir de solides notions mathématiques pour construire, analyser et valider des modèles mathématiques
- Avoir une connaissance de base en informatique pour pouvoir produire et si possible implémenter des logiciels, ou comprendre comment se servir des logiciels et bases de données proposés

Phylogénie

Phylogénie

Darwin (1859) a initié l'arbre comme support formel de la représentation des relations inter-espèces



"The affinities of all the beings of the same class have sometimes been represented by a great tree... As buds give rise by growth to fresh buds, and these if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications."

Charles Darwin, 1859

Phylogénie

Au début les modes de classifications des espèces étaient :

- Les comparaisons morphologiques
- Les comparaisons comportementales
- Les répartitions géographiques

Aujourd'hui les phylogénies sont obtenues à partir :

- des séquences moléculaires (phylogénie moléculaire)
- des caractères discrets
- des fréquences des gènes
- des traits quantitatifs

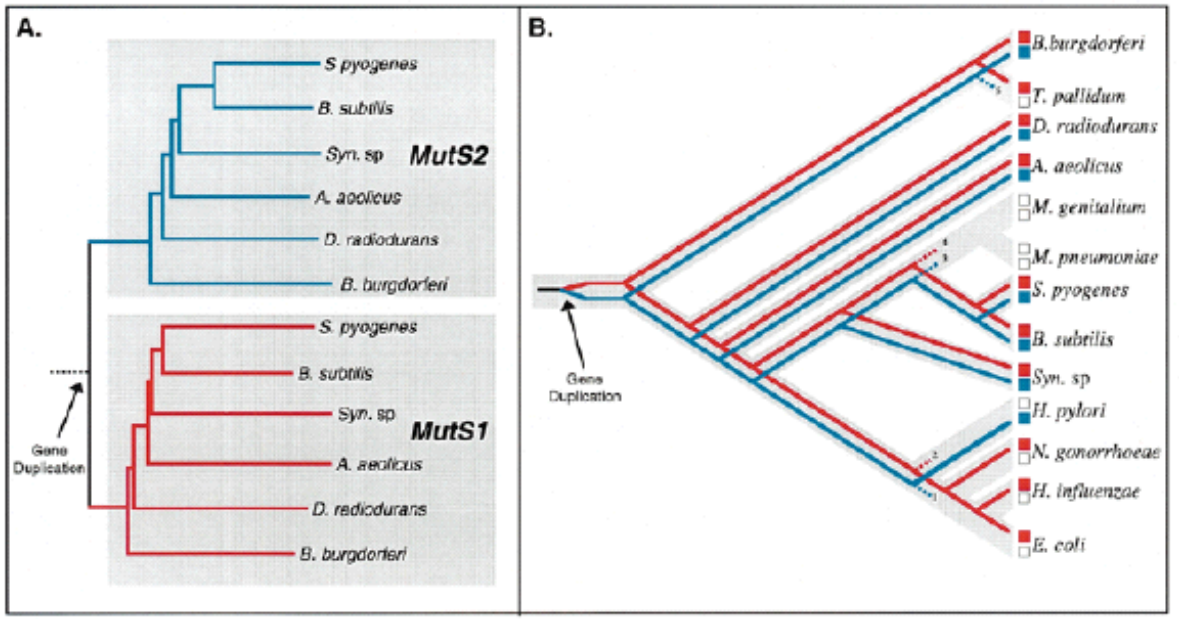
Applications

Bio-écologie

- Déplacement d'espèces
- Relation hôtes-parasites

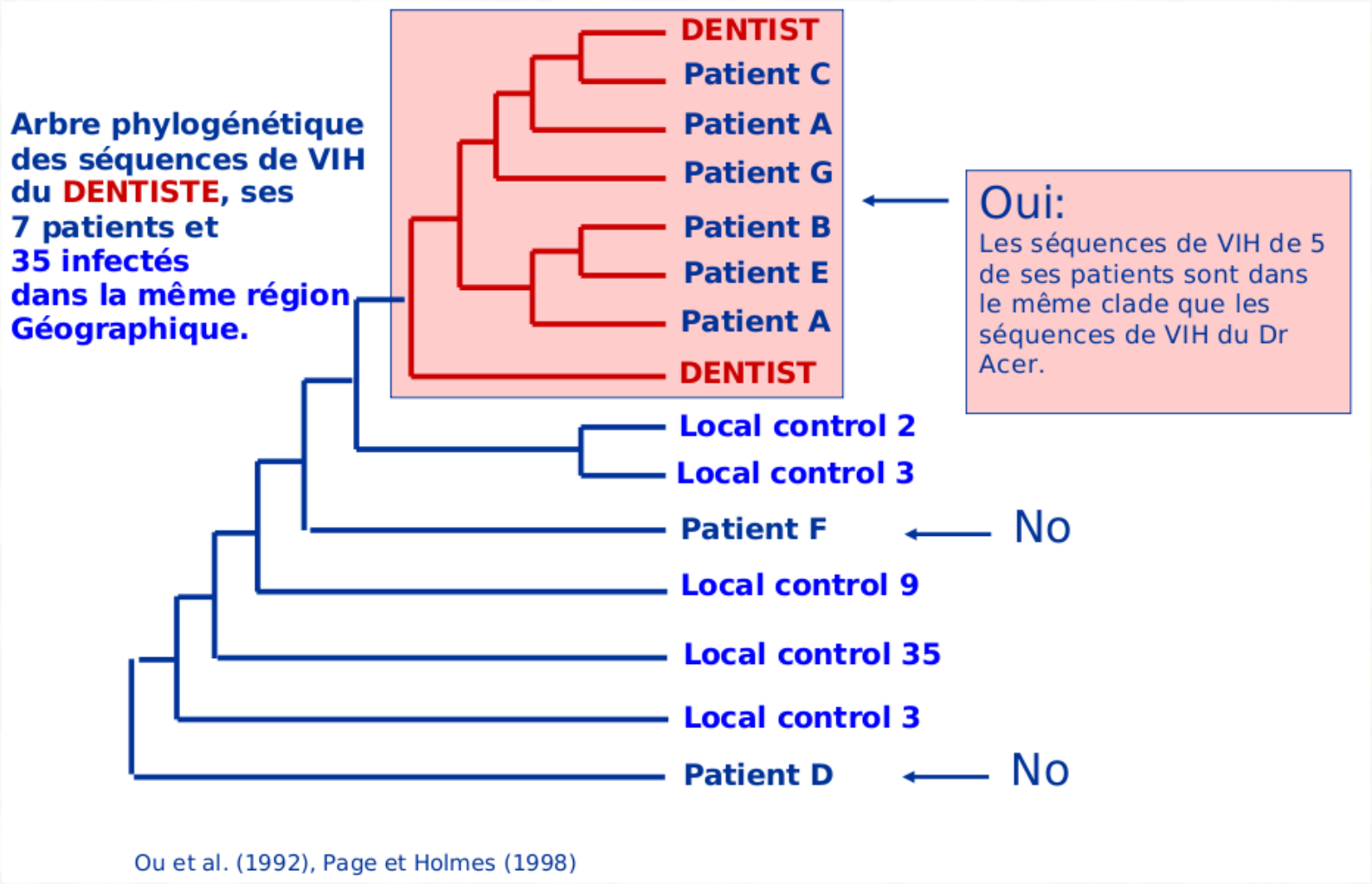
Épidémiologie : Tracer l'évolution d'un virus à travers ses différentes souches

Comprendre les phénomènes de duplications et pertes de gènes



Applications

Police scientifique : est ce que le Dr David Acer a contaminé ses patients ?



Méthodes existantes

Les méthodes de distance :

- on calcule une distance entre les espèces, soit en regardant la similarité entre les séquences de nucléotides (distance entre mots), soit en regardant d'autres critères de similarité $\Rightarrow$ matrice de distance
- à partir de la matrice de distance, on regroupe (méthode de clustering) les espèces les plus proches et on reconstruit l'arbre progressivement

Les méthodes d'optimisations :

- on se donne un critère de vraisemblance ou un critère probabiliste
- on parcourt l'espace des arbres (méthode exacte mais exponentielle) ou une sous-partie de cet espace (méthode heuristique mais plus rapide) en optimisant ce critère

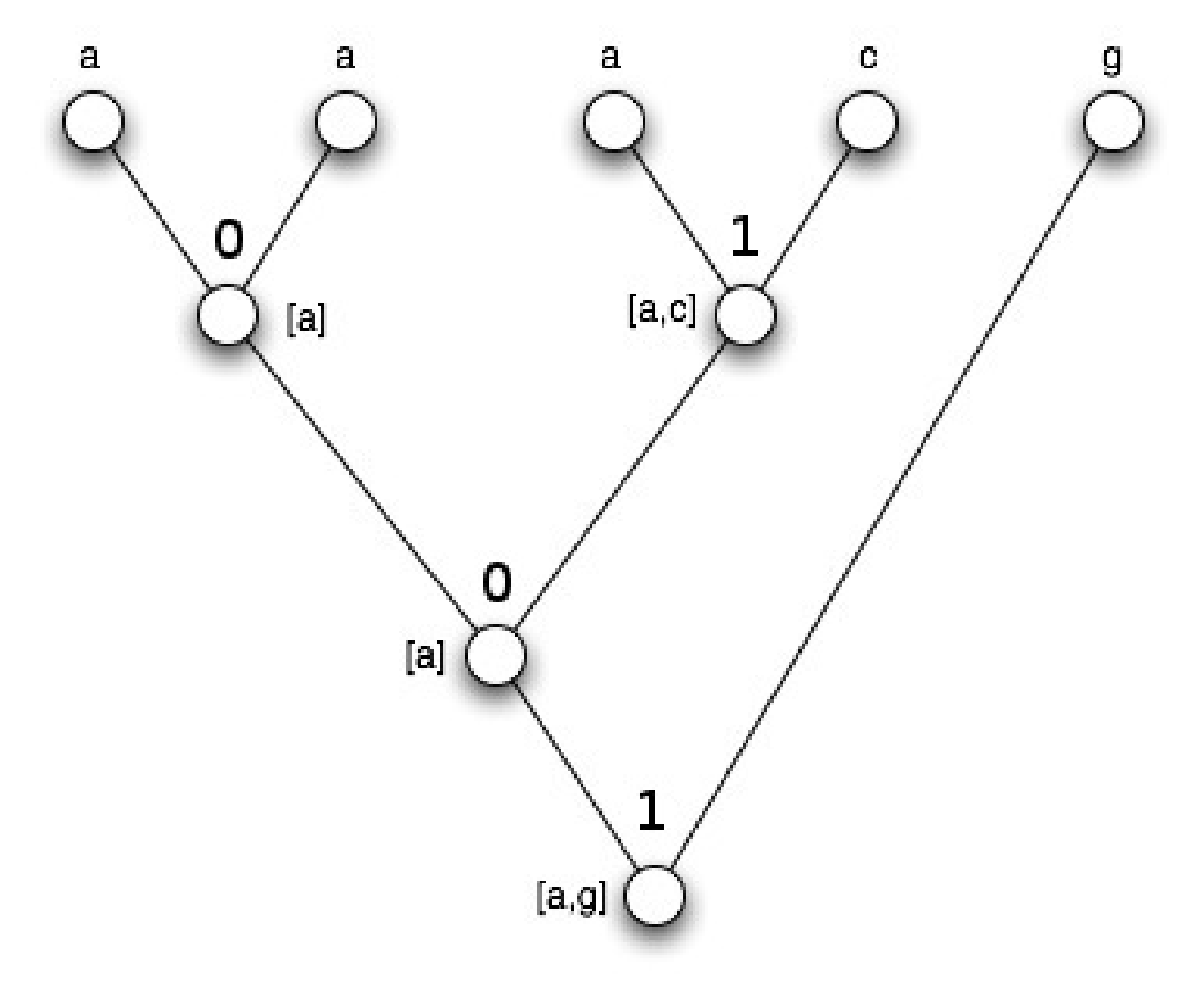
Reconstruction de génomes ancestraux

Problème : On se donne un arbre phylogénétique, et des génomes à chacune des feuilles. On veut connaître les génomes ancestraux à chaque nœud interne

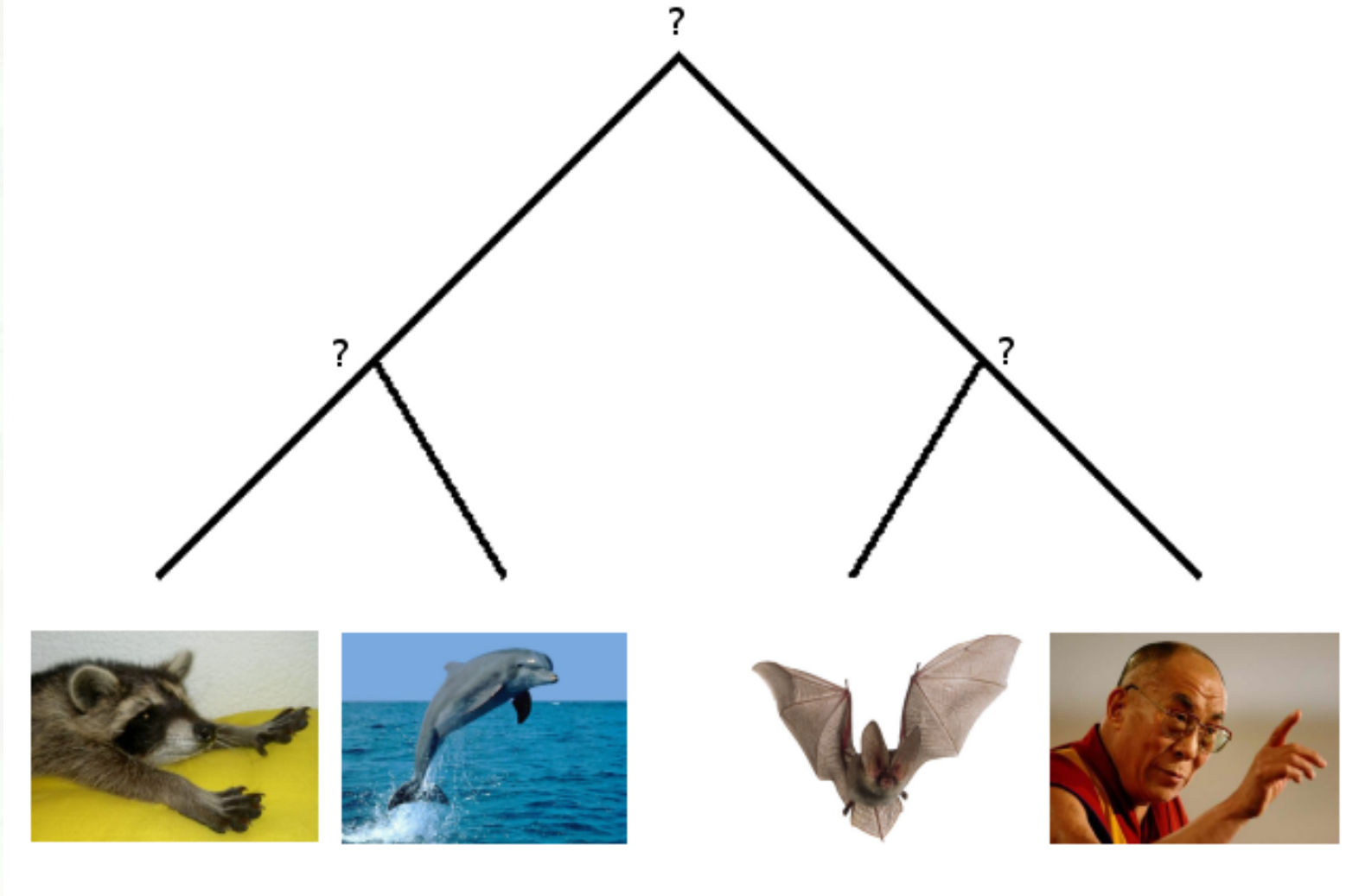
Un exemple de méthode : l'algorithme de Fitch

- Données : un arbre, avec à chaque feuille un caractère (exemple, un nucléotide A, C, G, T)
- Critère : un coût est associé à chaque transformation du caractère (A $\leftrightarrow$ T coûte 1 par exemple). Pour un arbre dont tous les nœuds sont étiquetés, on peut calculer le coût total de transformation le long des branches. C'est la *parcimonie*. C'est le critère que l'on veut optimiser

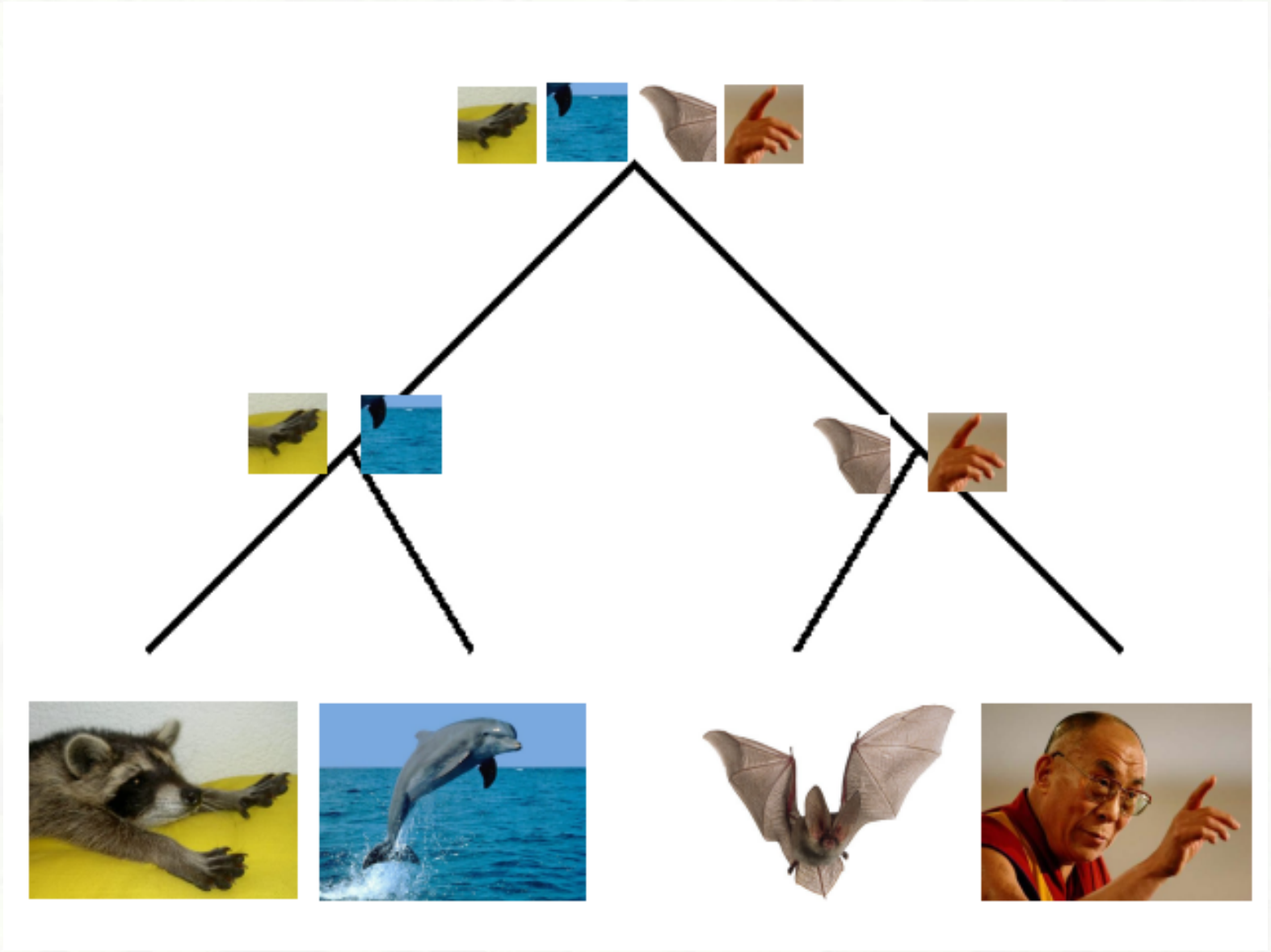
Reconstruction de génomes ancestraux



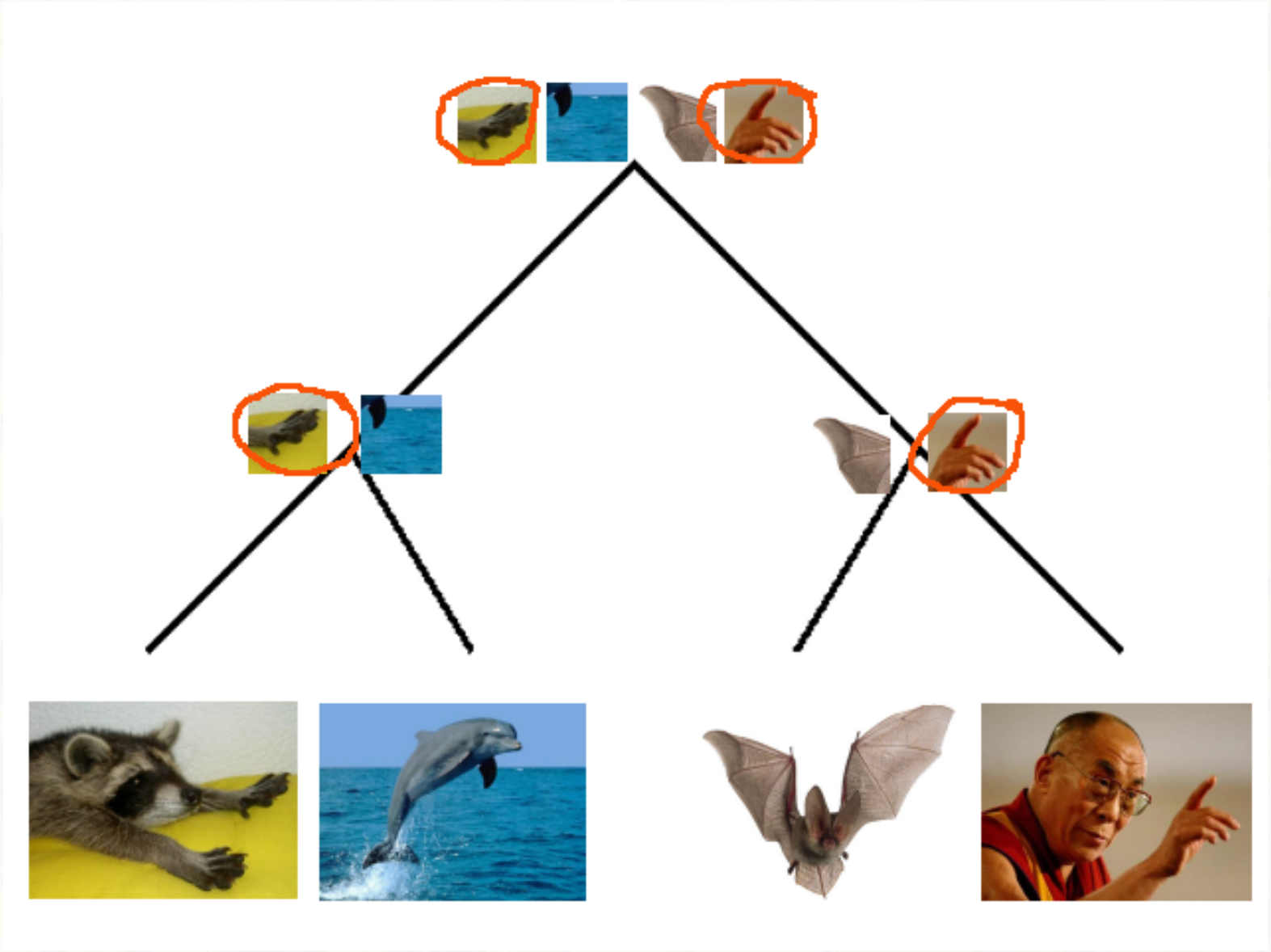
Reconstruction de génomes ancestraux



Reconstruction de génomes ancestraux



Reconstruction de génomes ancestraux



Intervalles communs

On considère des génomes à contenus identiques
Exemple : ADN de mitochondries et de chloroplastes

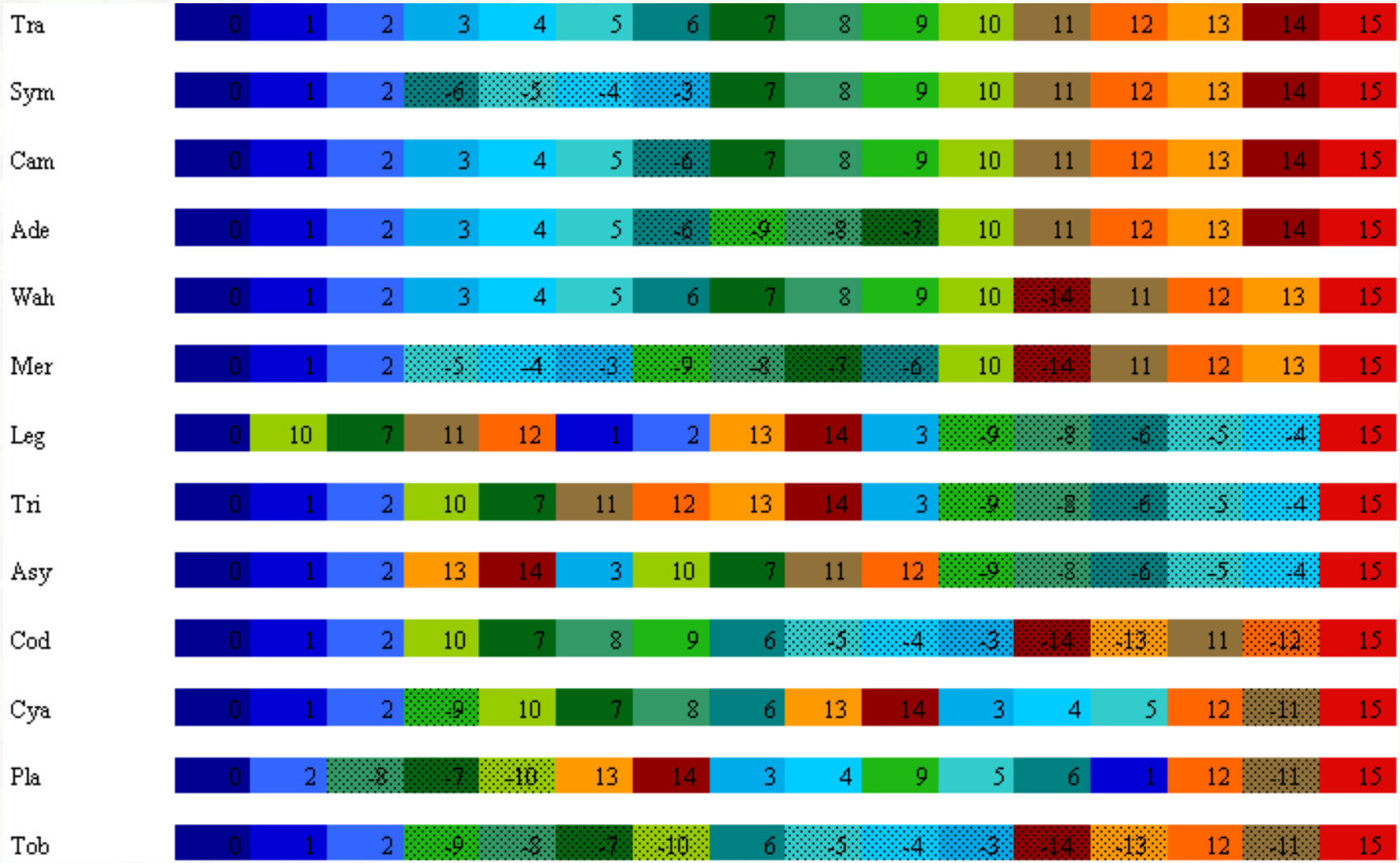
On regarde l'ordre d'apparition des gènes $\Rightarrow$ permutations

On tient compte du brin sur lequel ils apparaissent (orientation)
$\Rightarrow$ permutations signées

L'ordre et l'orientation des gènes servent à :

- Mesurer une distance d'évolution entre les espèces
- Déterminer la phylogénie d'un ensemble d'espèces
- Déterminer les ordres de gènes ancestraux

Définition


Tra 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Sym 0 1 2 -6 -5 -4 -3 7 8 9 10 11 12 13 14 15
Cam 0 1 2 3 4 5 -6 7 8 9 10 11 12 13 14 15
Ade 0 1 2 3 4 5 -6 -9 -8 -7 10 11 12 13 14 15
Wah 0 1 2 3 4 5 6 7 8 9 10 -14 11 12 13 15
Mer 0 1 2 -5 -4 -3 -9 -8 -7 -6 10 -14 11 12 13 15
Leg 0 10 7 11 12 1 2 13 14 3 -9 -8 -6 -5 -4 15
Tri 0 1 2 10 7 11 12 13 14 3 -9 -8 -6 -5 -4 15
Asy 0 1 2 13 14 3 10 7 11 12 -9 -8 -6 -5 -4 15
Cod 0 1 2 10 7 8 9 6 -5 -4 -3 -14 -13 11 -12 15
Cya 0 1 2 -9 10 7 8 6 13 14 3 4 5 12 -11 15
Pla 0 2 -8 -7 -10 13 14 3 4 9 5 6 1 12 -11 15
Tob 0 1 2 -9 -8 -7 -10 6 -5 -4 -3 -14 -13 12 -11 15

Définition

Définition

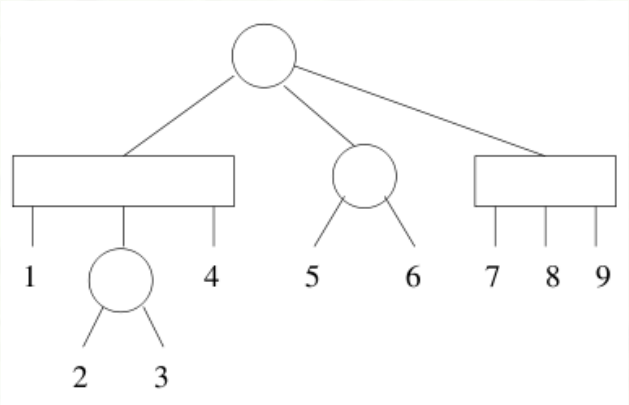
Un intervalle commun à un ensemble de permutations $\mathcal{P}$ sur $\{1, \ldots, n\}$ est une partie I de $\{1, \ldots, n\}$ telle que I est un intervalle dans chacune des permutations.

Proposition

L'ensemble $I(\mathcal{P})$ des intervalles communs d'un ensemble de permutations $\mathcal{P}$ sur $\{1, \ldots, n\}$ est une famille faiblement partitive, c'est-à-dire :

- $\{1, \ldots, n\} \in I(\mathcal{P}), \emptyset \notin I(\mathcal{P}), \forall i \in \{1, \ldots, n\}, \{i\} \in \{1, \ldots, n\}$
- *Pour tous $A, B \in I(\mathcal{P})$ qui se chevauchent :*
  1. $A \cup B \in I(\mathcal{P})$
  2. $A \cap B \in I(\mathcal{P})$
  3. $A \backslash B \in I(\mathcal{P})$
  4. $B \backslash A \in I(\mathcal{P})$

Arbres PQ



Les intervalles communs qui ne chevauchent aucun autre sont des *intervalles forts*

Les intervalles forts s'organisent en un arbre d'inclusion avec des nœuds P (ronds) et Q (rectangulaires) $\Rightarrow$ Arbre PQ

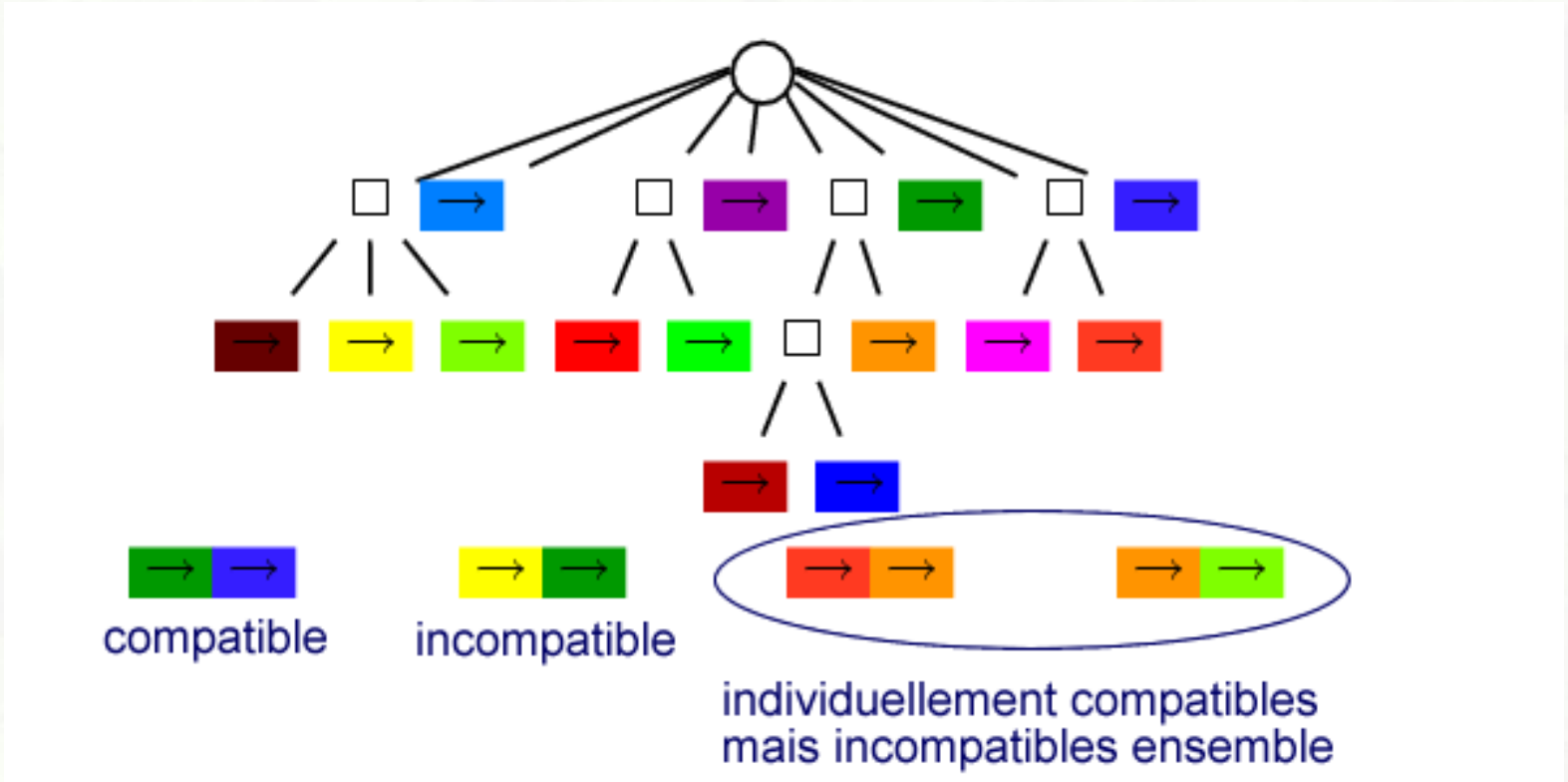
Les intervalles communs sont soit des nœuds P, soit des nœuds Q, soit des unions de fils consécutifs de nœuds Q

Un arbre PQ $\Leftrightarrow$ un ensemble de permutations

Reconstruction de génomes ancestraux

Idée : reprendre l'idée de l'algorithme de Fitch en propageant les intervalles communs du bas vers le haut.

Premier problème : comment conserver la notion d'arbre PQ à chaque étape ? $\Rightarrow$ *intervalles conflictuels*



Reconstruction de génomes ancestraux

Deuxième problème : comment ne pas éliminer ”trop” de candidats ?

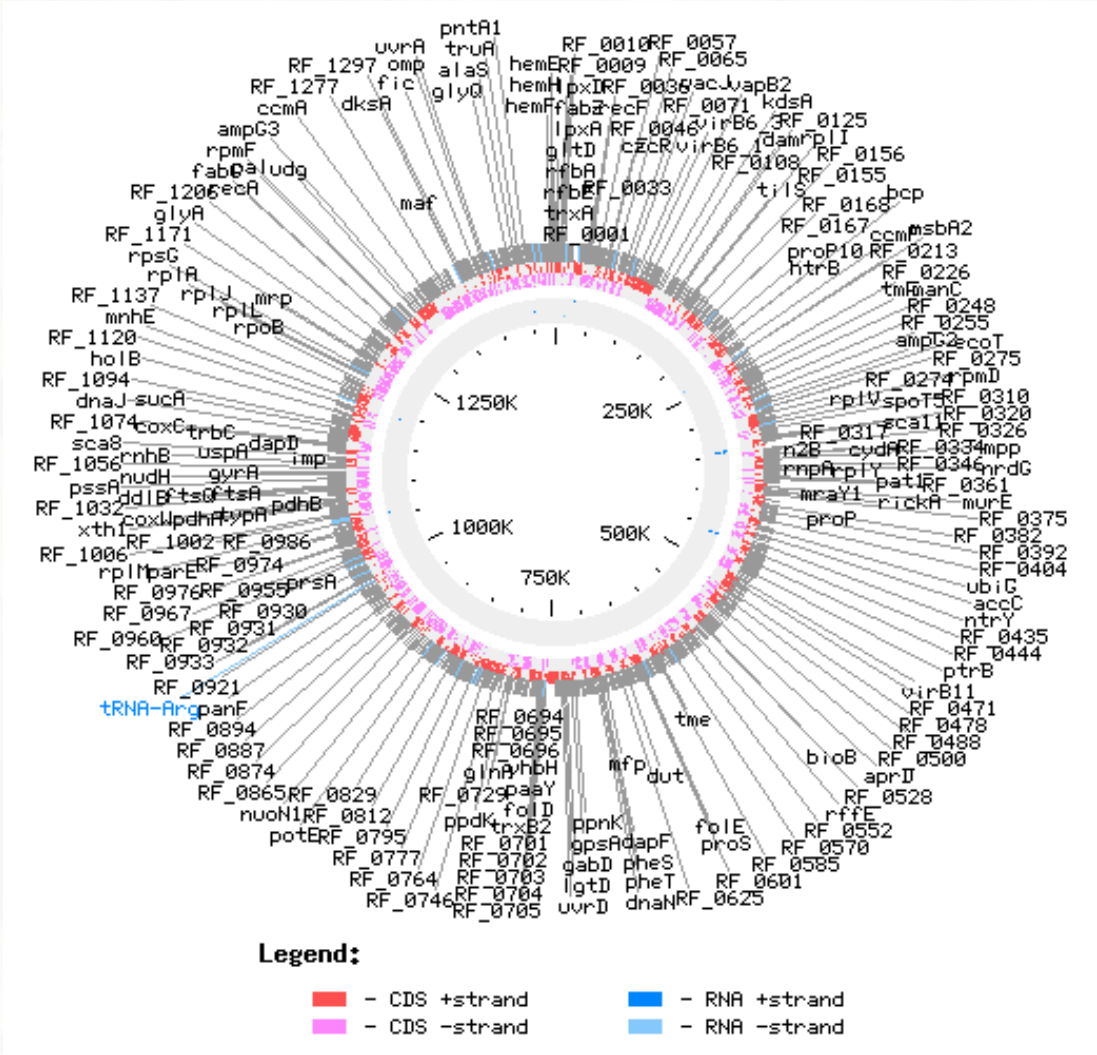
Conserver le maximum d'intervalles, et déterminer de manière *quantitative* (critère d'optimisation) ou *qualitative* (comprendre la structure sous-jacente) les vrais intervalles ancestraux

Théorème bioinformatique

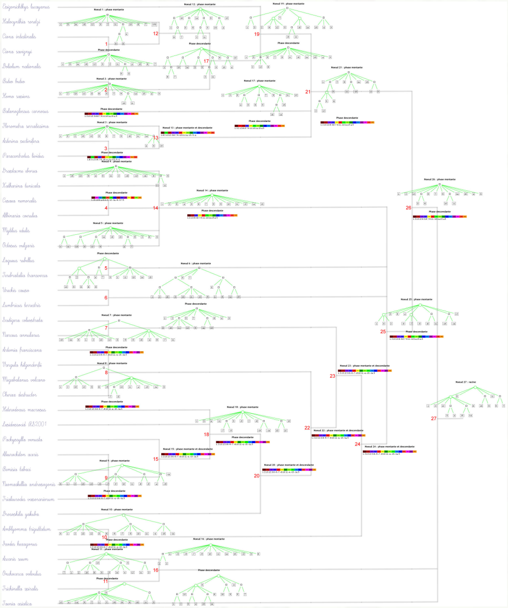
Si les intervalles associés à un nœud interne d'une phylogénie sont tous ancestraux, alors ils correspondent à un arbre PQ.

Reconstruction de génomes ancestraux

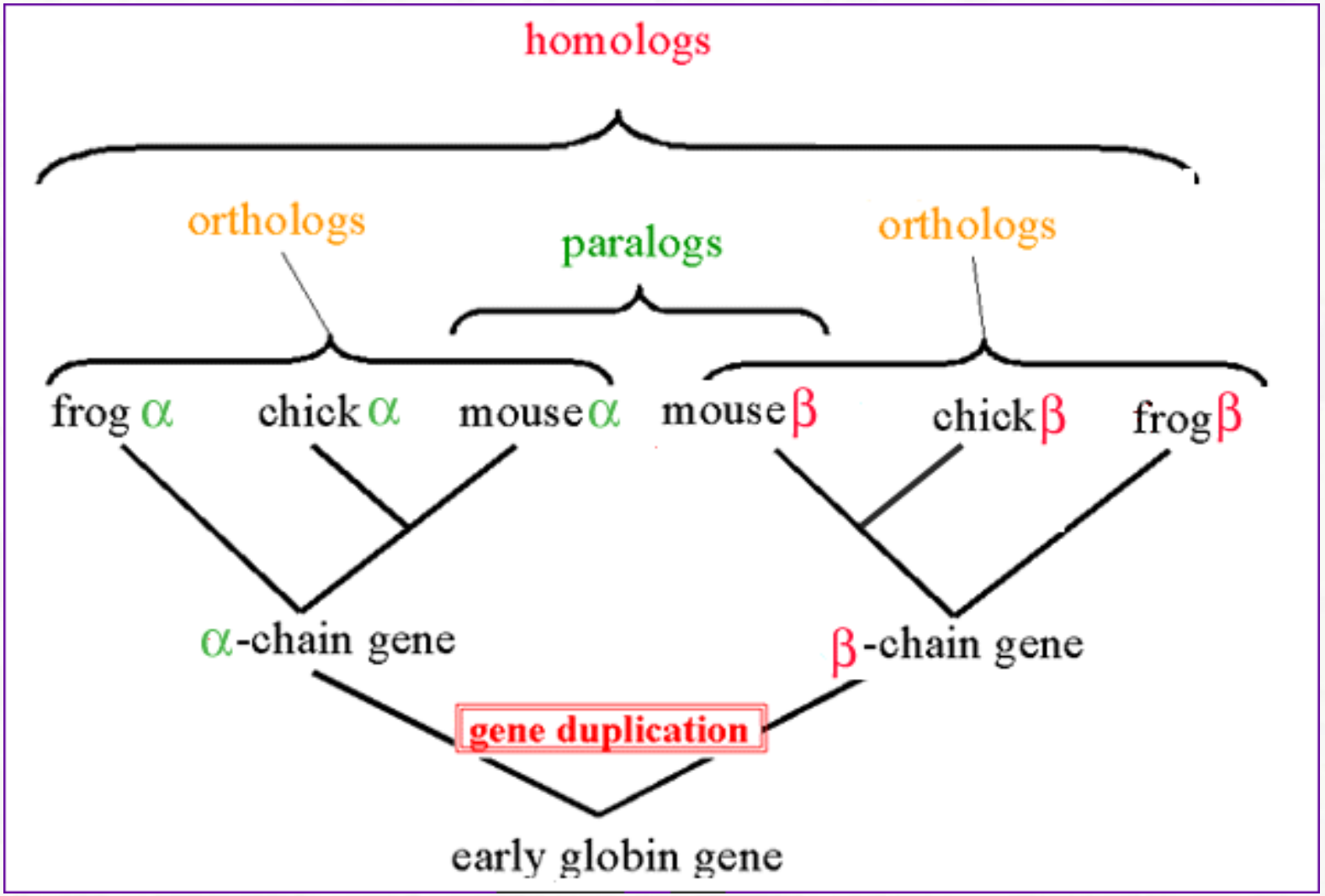
Application : Prospection autour de génomes de mitochondries dans les espèces animales, à la recherche de la bactérie ancestrale.



Reconstruction de génomes ancestraux

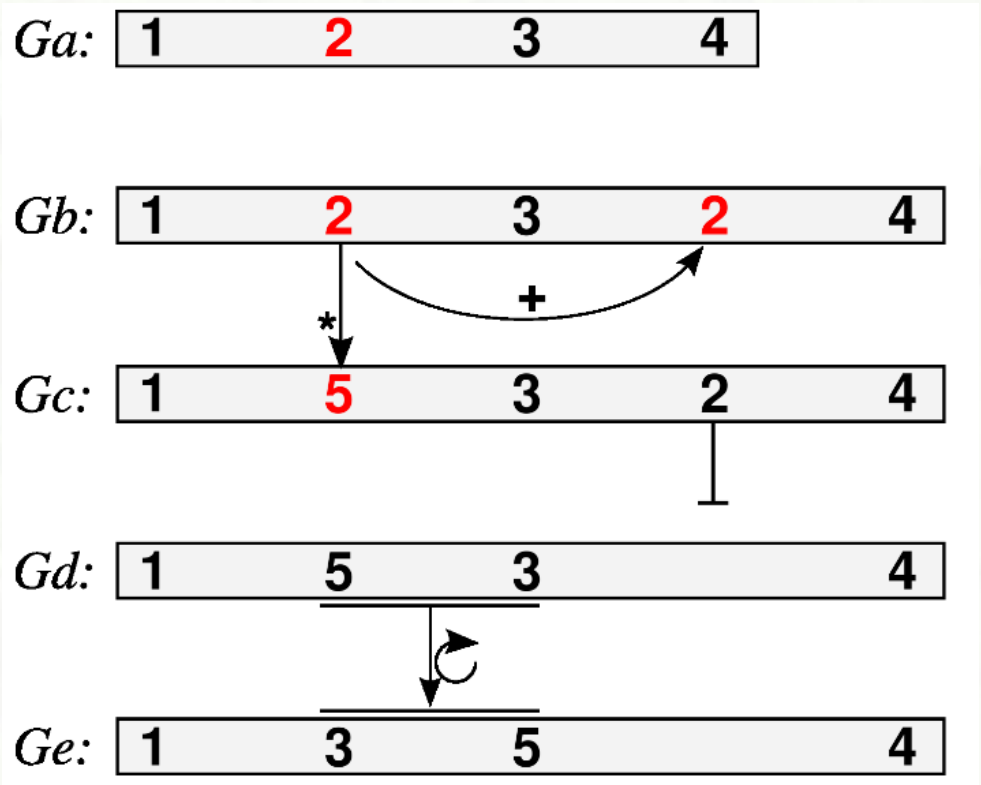


Détermination de marqueurs orthologues



http ://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html

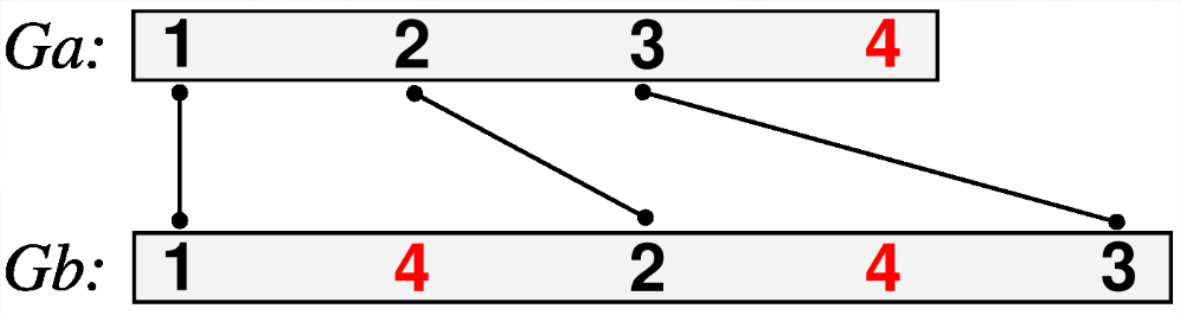
Détermination de marqueurs orthologues



Réarrangements génomiques $\Rightarrow$ contenus différents et avec duplication

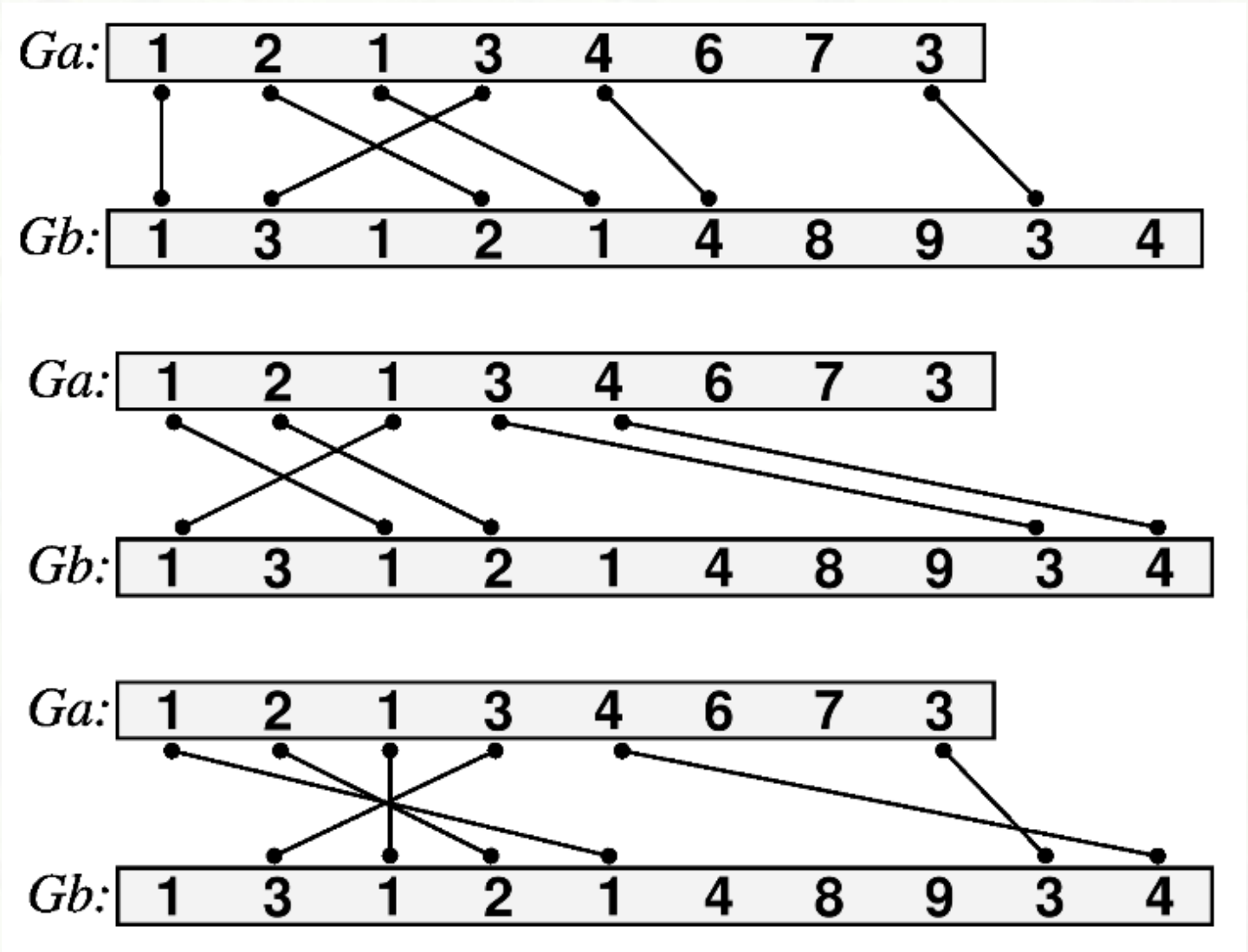
En phylogénie, il est important de comparer des *vrais orthologues*

Détermination de marqueurs orthologues



Une seule copie dans chaque gènome : le couplage est facile
Copies multiples : quelles copies choisir ?

Détermination de marqueurs orthologues



Détermination de marqueurs orthologues

Plusieurs approches :

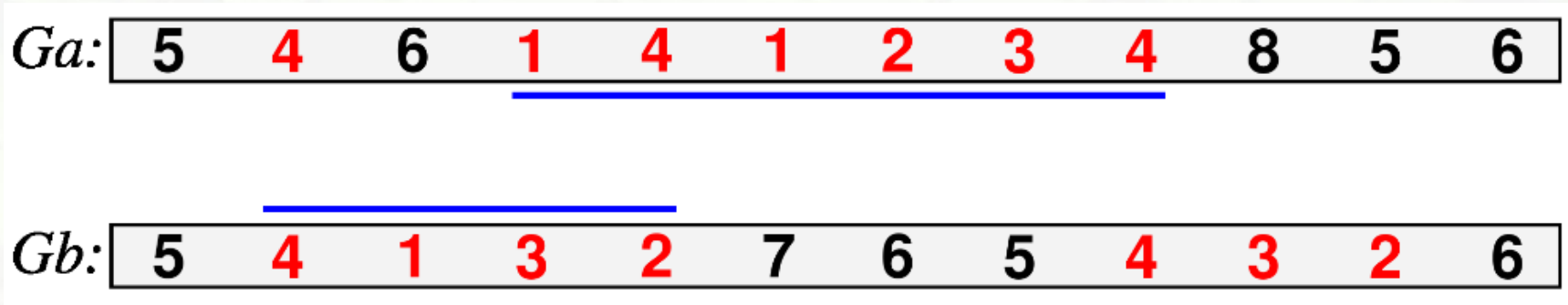
- Approche exemplaire : on ne garde qu'une seule copie de chaque gène
- Optimisation d'un critère, par exemple le nombre d'adjacences conservées

Problème : le problème général est NP-complet $\Rightarrow$ approches heuristiques

- Segments conservés (Longest Common Substring)
- Intervalles communs

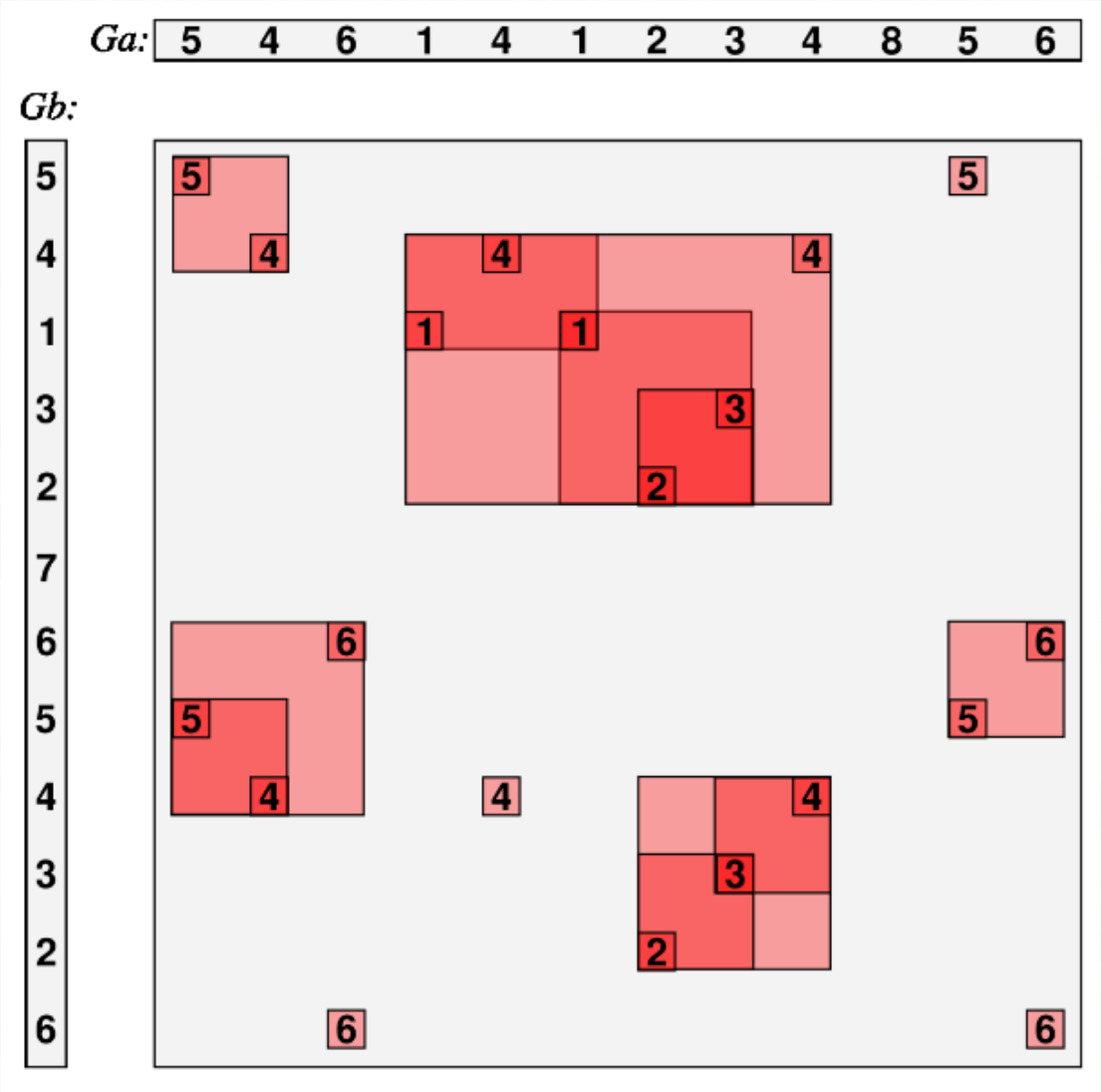
Détermination de marqueurs orthologues

Généralisation des intervalles communs aux *mots* (et non plus aux permutations)

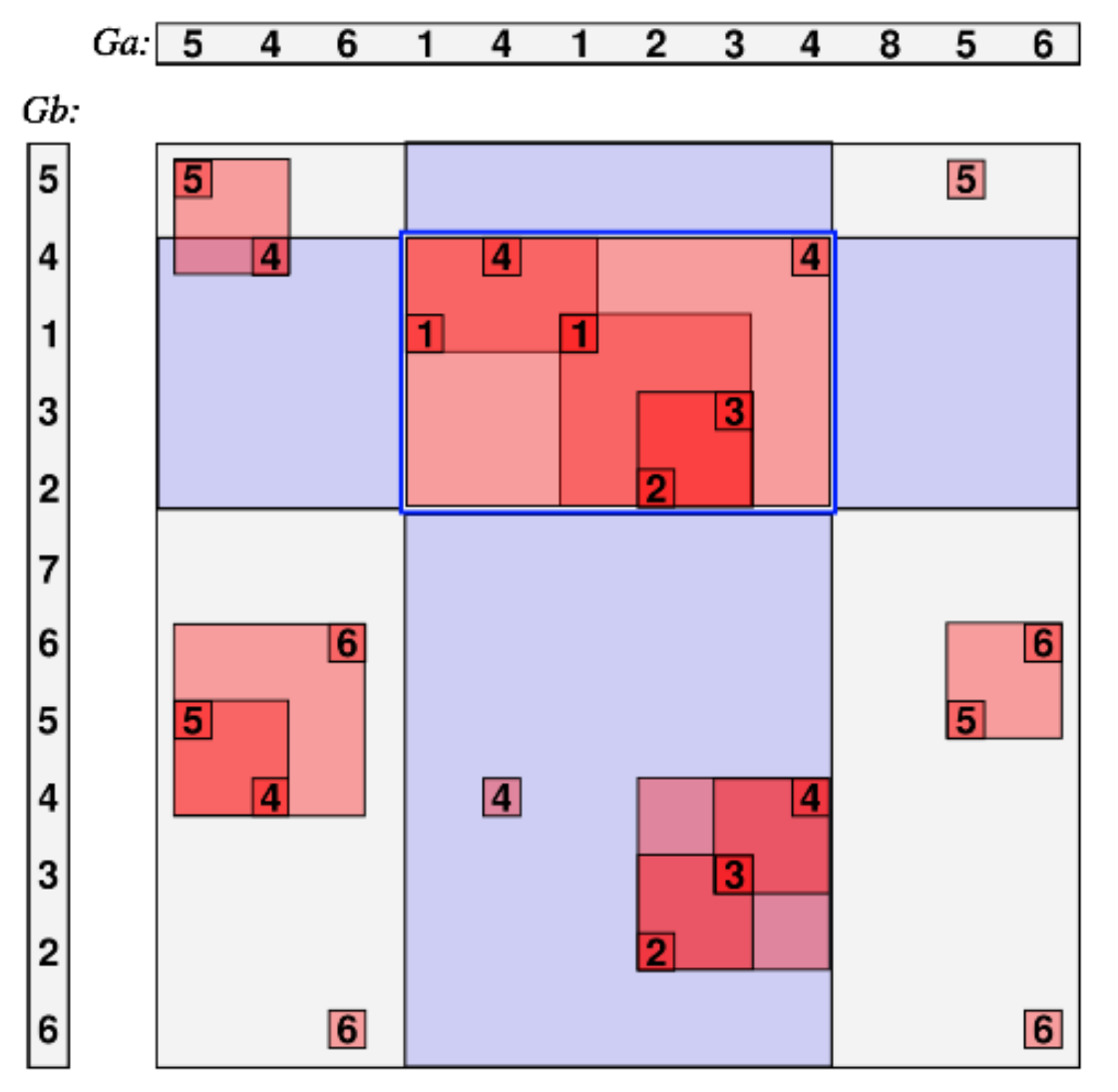


Calcul de tous les intervalles communs en temps quadratique (Schmidt & Stoye, 2004)

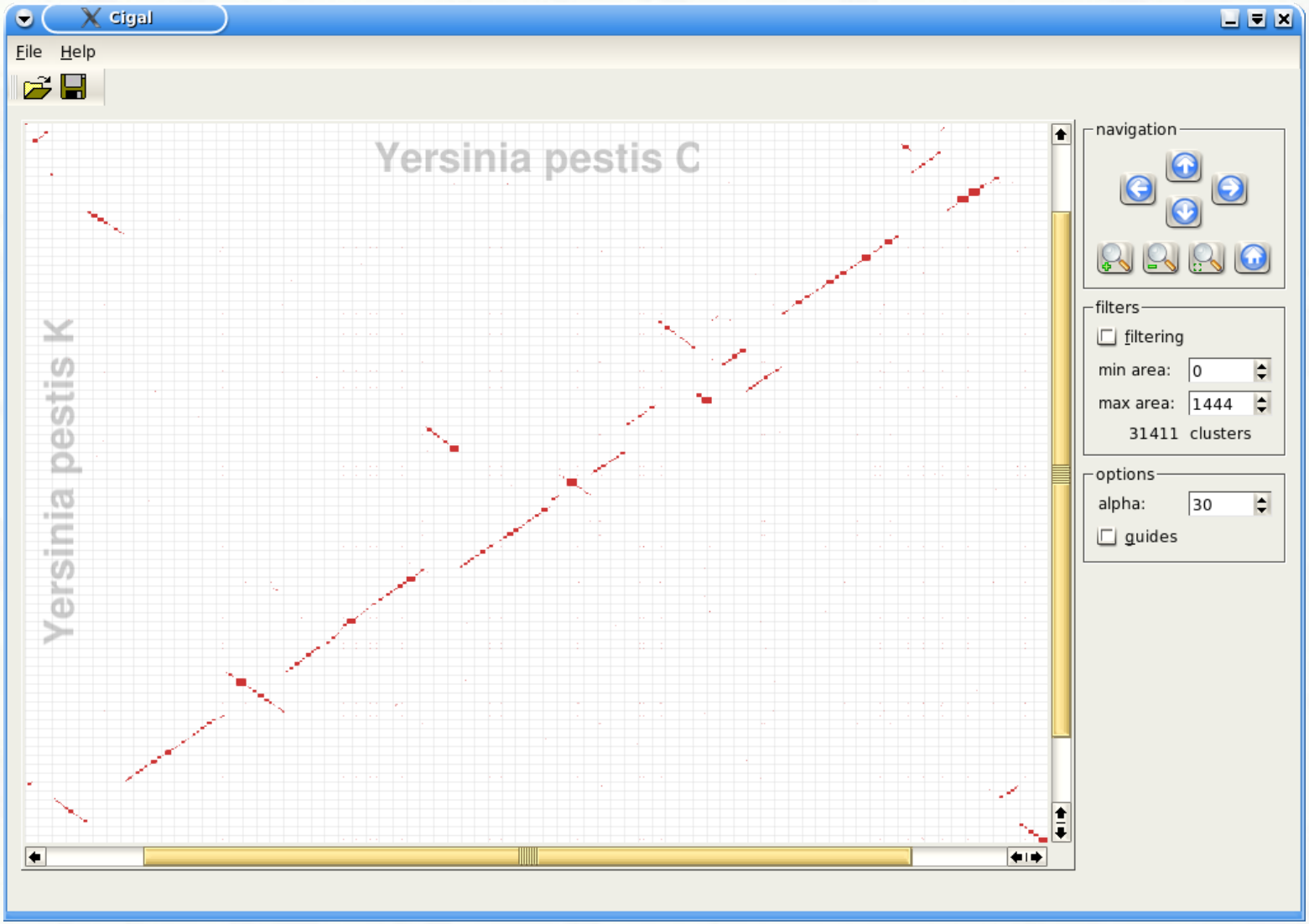
Détermination de marqueurs orthologues



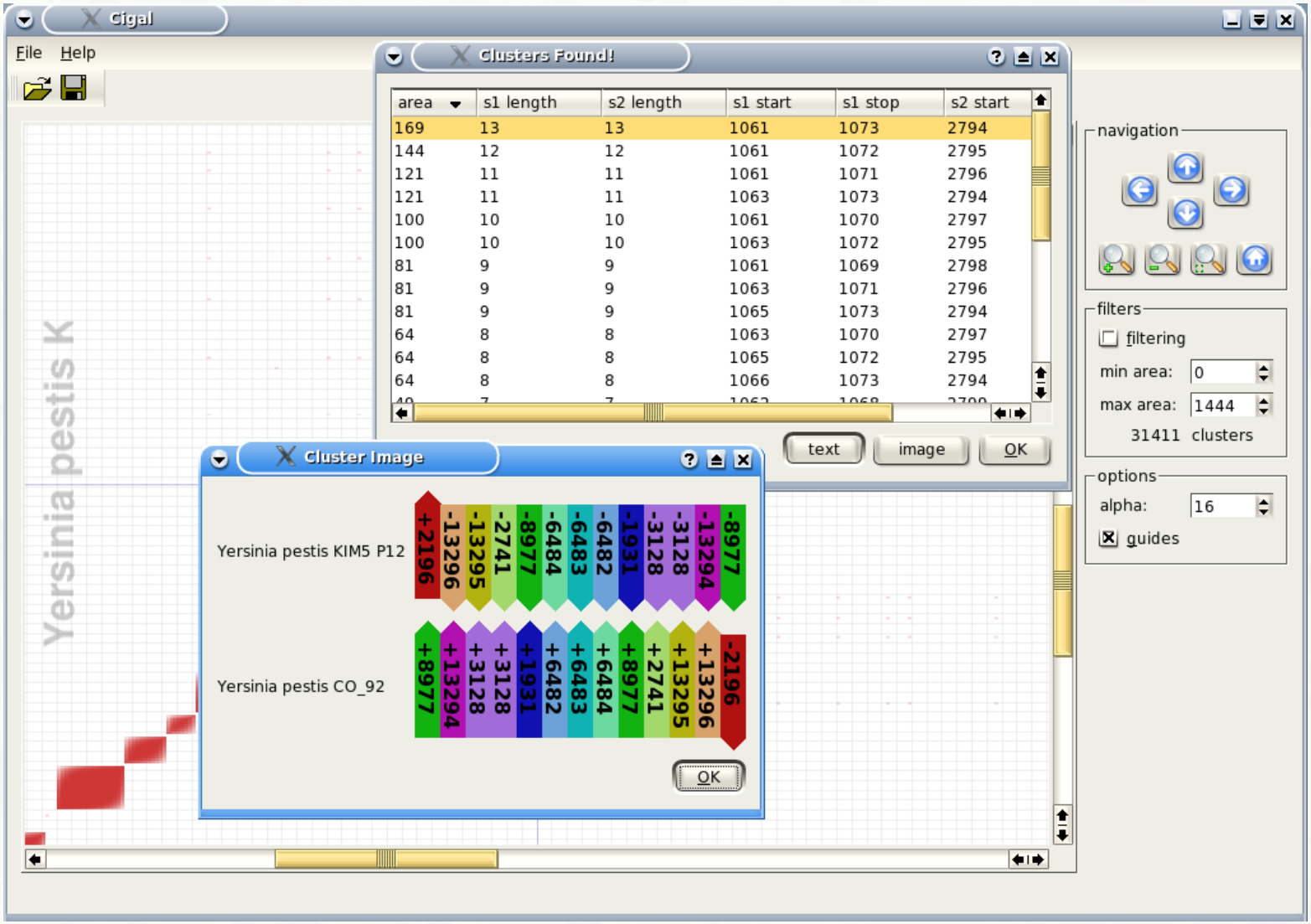
Détermination de marqueurs orthologues



Détermination de marqueurs orthologues



Détermination de marqueurs orthologues



Détermination de marqueurs orthologues

Application à une famille de $\gamma-$protéobactéries :

	LCS	**Intervalles Communs**
Vrais positifs	19142	18875
Faux positifs	3045	3324
Composantes	3439	3539
Consistance	2907	3117
% Consistance	85%	88%
VP dans CC	14954	10729
Comp. parfaite	1531	1628
% Comp. parfaite	53%	52%

- LCS : Longest Common Substrings (G. Blin, C. Chauve, G. Fertin, 2005)
- Vrais positifs : gènes ayant le même nom dans la base Uniprot

Détermination de marqueurs orthologues

- Résultats encourageants
- Heuristique à améliorer : tenir compte de la ”complexité” d’un intervalle commun plutôt que de son aire
- Détection des grandes duplications (génomes de plantes ayant jusqu’à six copies)

Conclusion

En bioinformatique, on se donne des problèmes concrêts

On modélise ces problèmes

On théorise sur les modèles

On conçoit des algorithmes

On répond au problème

On recommence car le problème est plus complexe que le modèle de départ. . .

On ne s'ennuie jamais !