Analyse du contexte génétique

chez les procaryotes

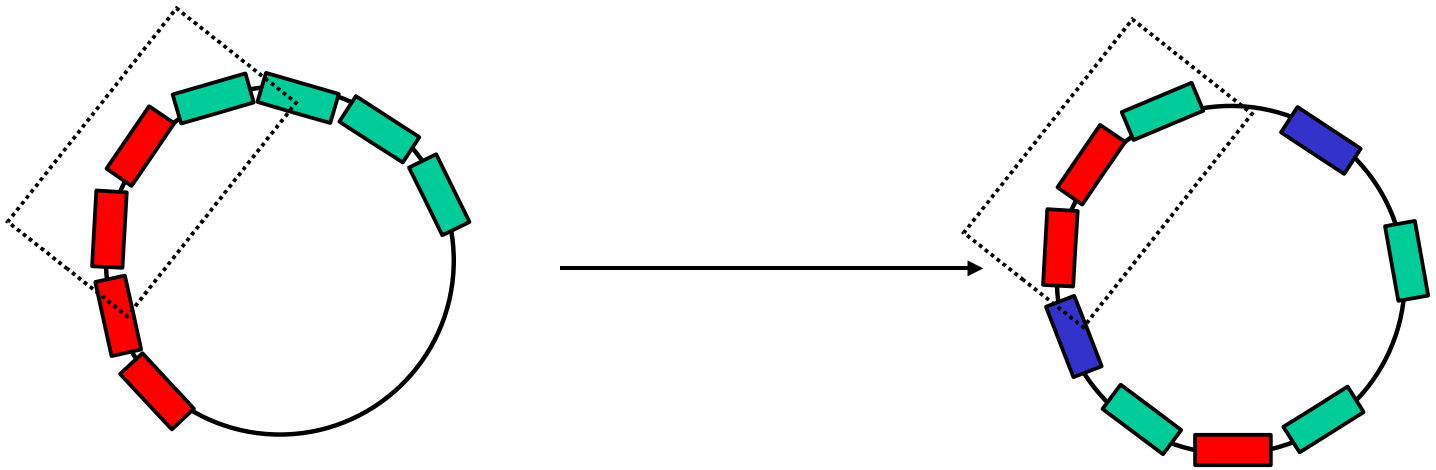
Frédéric Lemoine
Groupe Evolution Moléclaire et Bioinformatique des Génomes
Institut de Génétique et Microbiologie
UMR8621

Introduction

- Grande **plasticité/fluidité** des génomes procaryotes

  Pertes, gains, translocations, fusion et fission

- Mais **conservation** d'îlots dont l'ordre est préservé



Objectifs

Etude de la conservation de l'ordre des gènes

- Détection des “rares” régions de **synténie** dans les génomes procaryotes
- Étudier leur conservation à travers les espèces
    - Visualisation (observation des blocs de synténie)
    - Statistique (vitesse d’évolution,…)

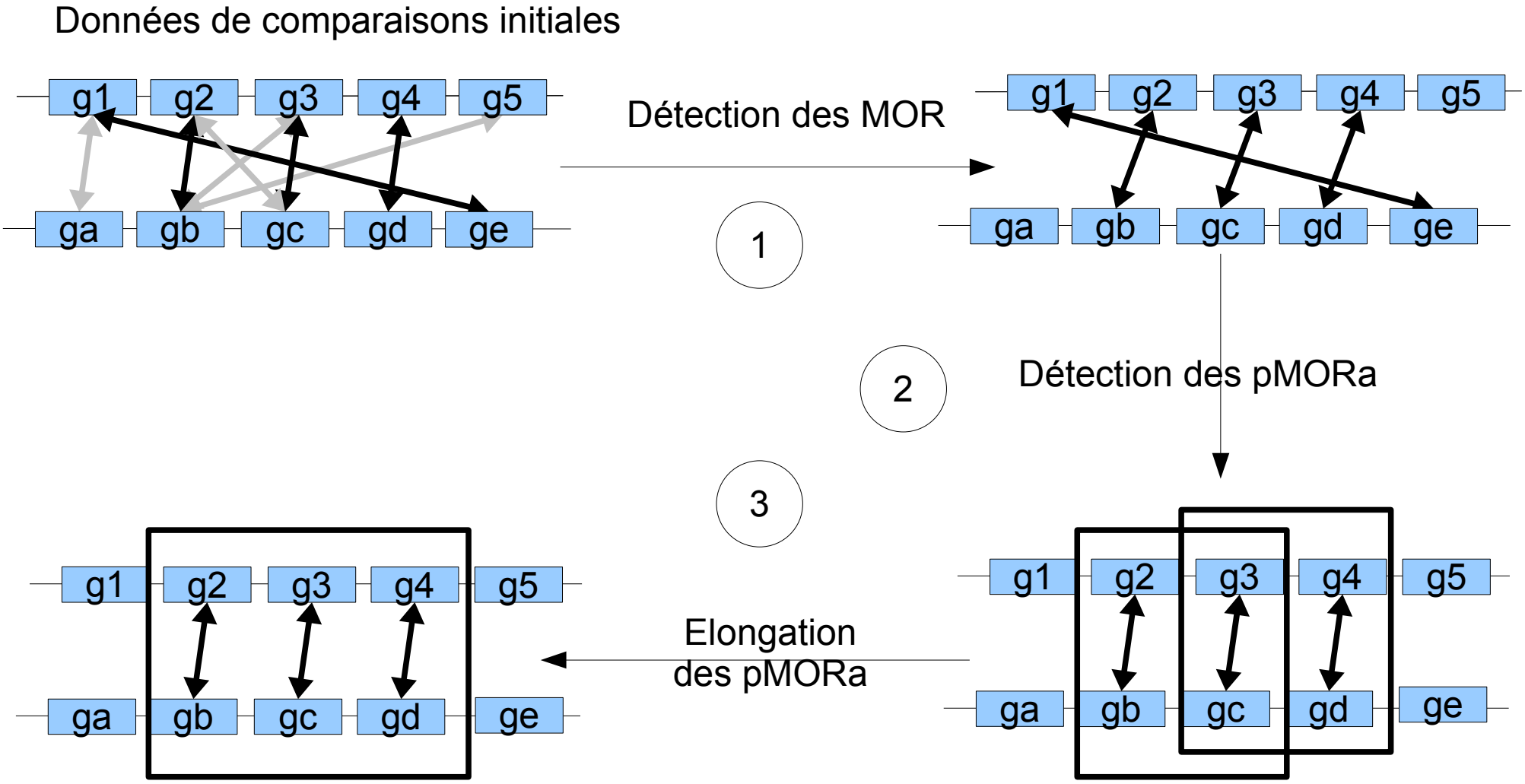
Approche

- **Originalité**:
  - Nombre d'espèces :107
  - Nombre de protéines: ~300 000 (Comparaisons 2 à 2)
  - Analyses statistiques de la synténie
- **Données** utilisées
  - Données primaires (génomes complets; Genbank)
  - Données de comparaison des génomes 2 à 2 (DARWIN; AllAll)

Plan: Démarche

1. **Détection** des régions génomiques dont l'ordre est conservé
2. **Visualisation** des résultats
3. Étude de la **taille** des blocs de synténie
4. Étude des modes **d'évolution** des protéines

Détection des régions dont l'ordre est conservé



Détection des régions dont l'ordre est conservé

1 - Détection des Meilleurs Orthologues Réciproques

Homologie: Si les protéines ont une **distance** PAM < 250

et alignement de **longueur** > 80 aa

Distance PAM: Représente le nombre de mutations ponctuelles par 100 aa

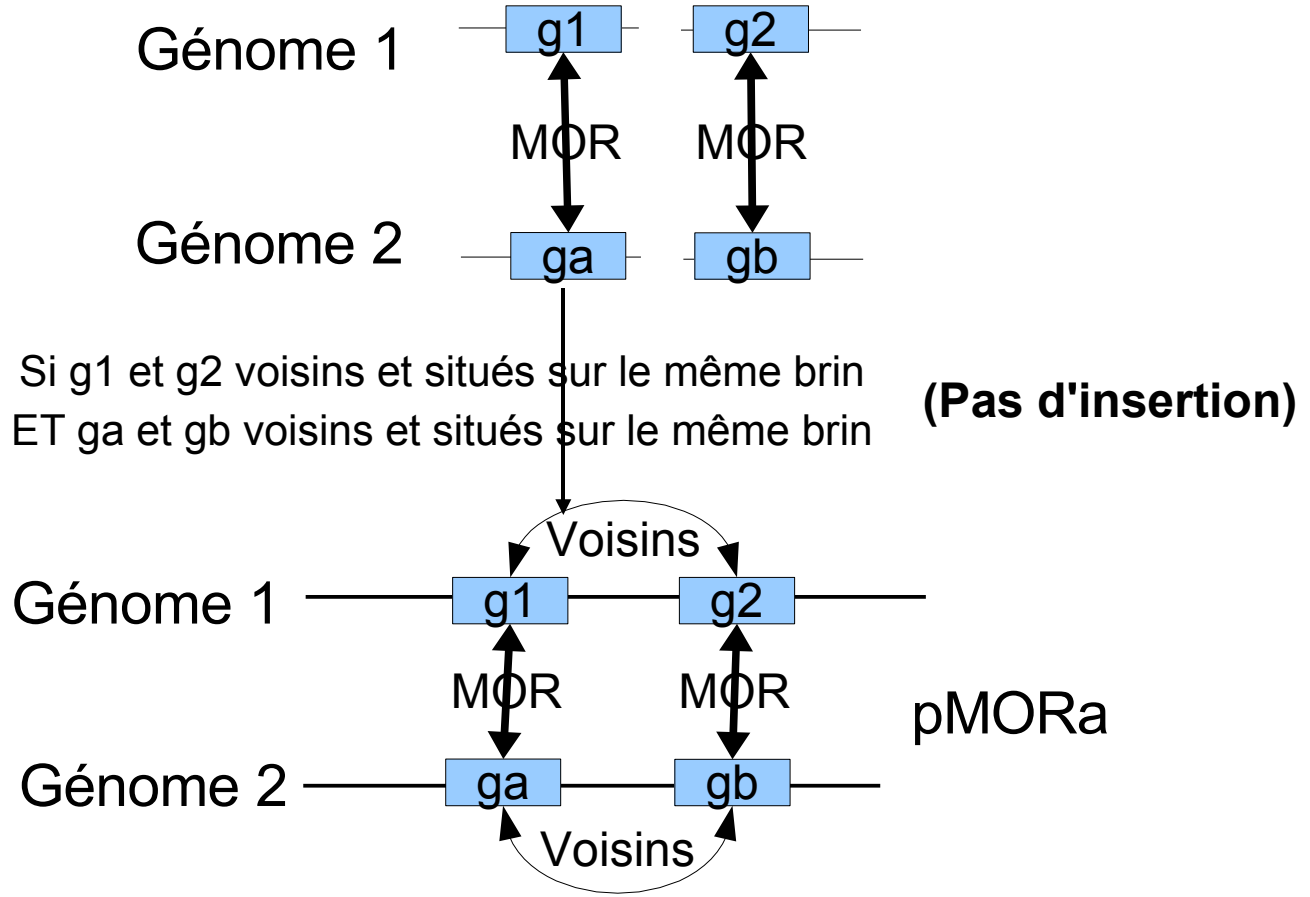
Données de comparaisons

Meilleurs **O**rthologues **R**éciproques

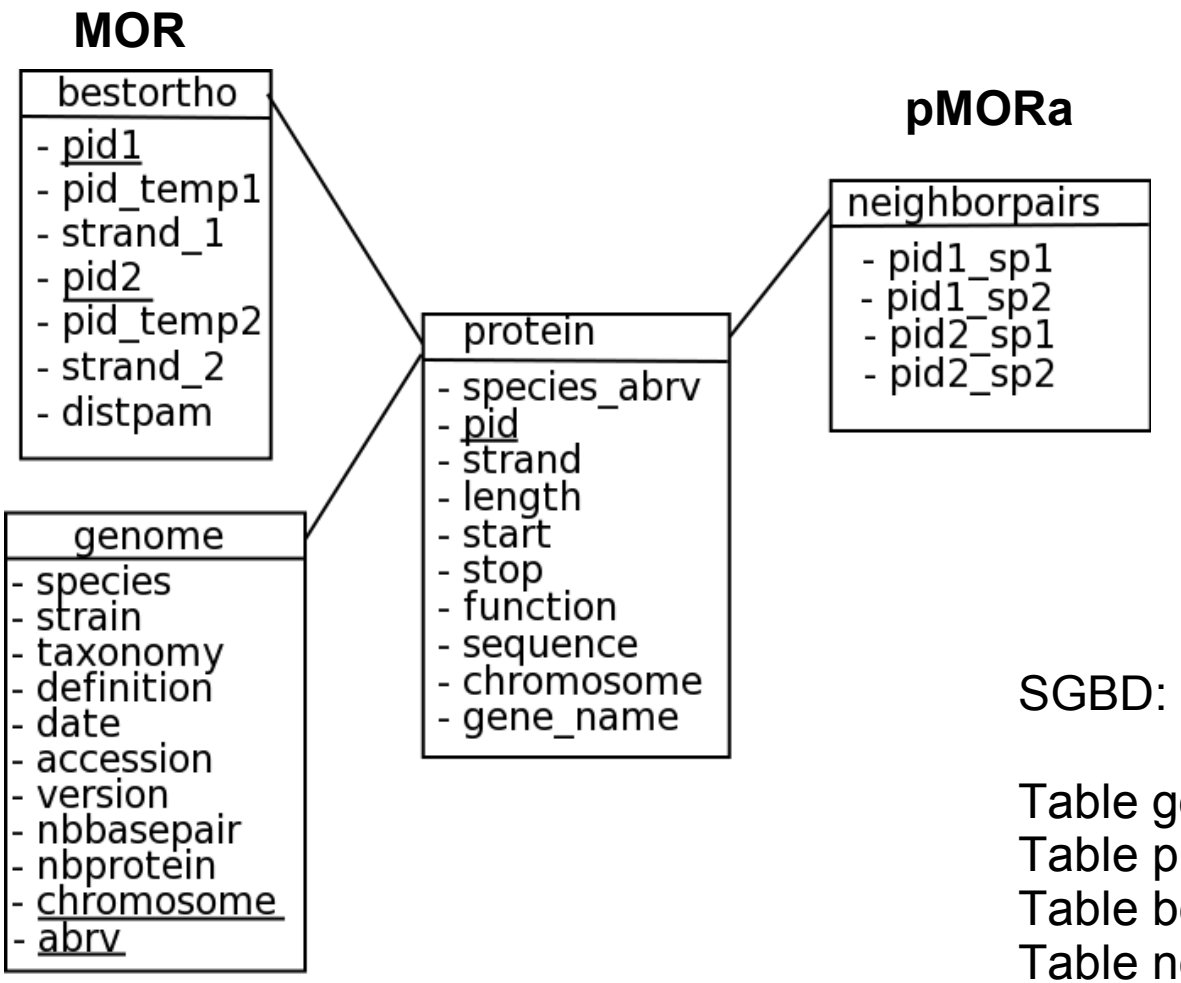
Programme *findbro*

Détection des régions dont l'ordre est conservé

2 - Détection des **paires** de Meilleurs Orthologues Réciproques **adjacentes**



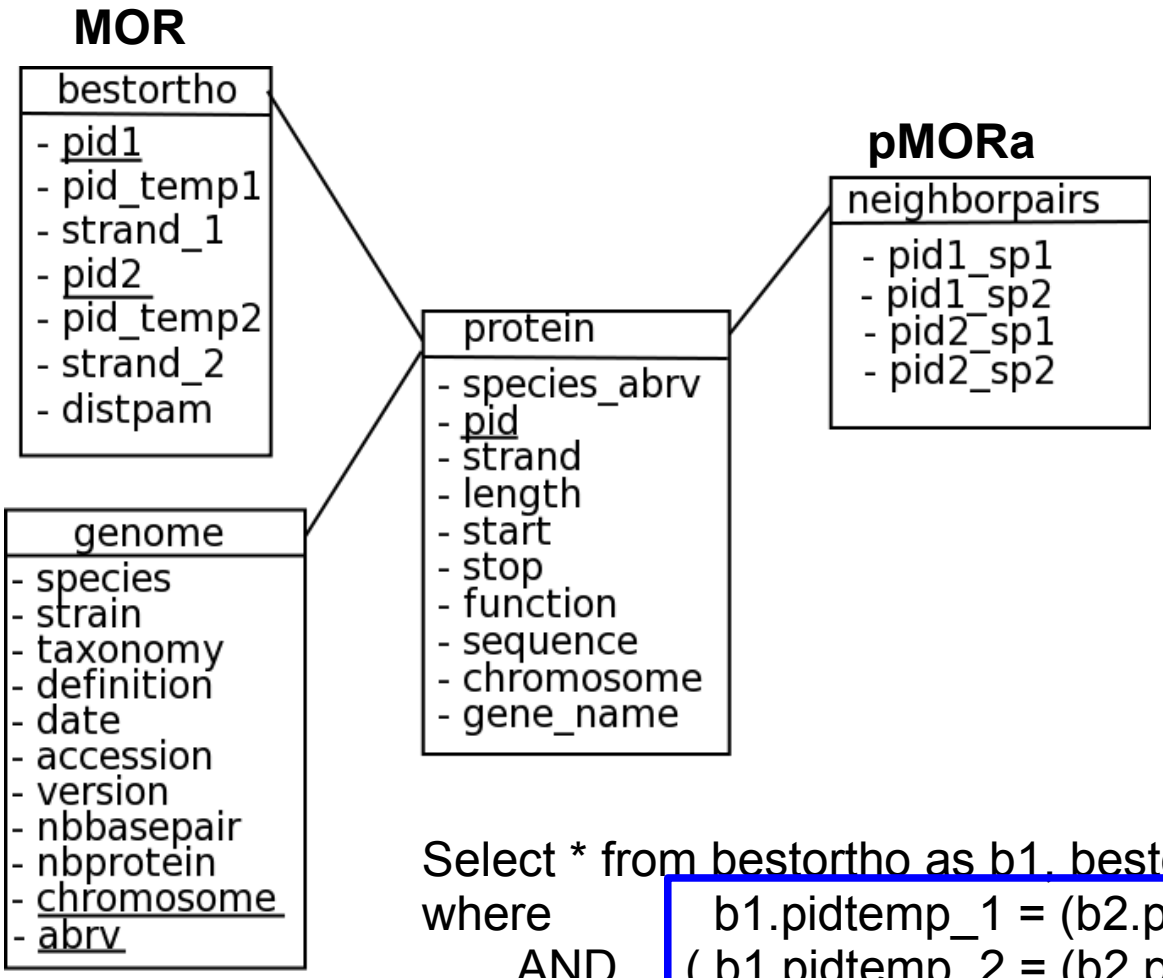
Construction de *synteBase*



SGBD: postgreSQL

Table genome: 107 entrées
Table protein: 295 608 entrées
Table bestortho: 3 908 954 entrées
Table neighborpairs: 392 222 entrées

Construction de *synteBase*



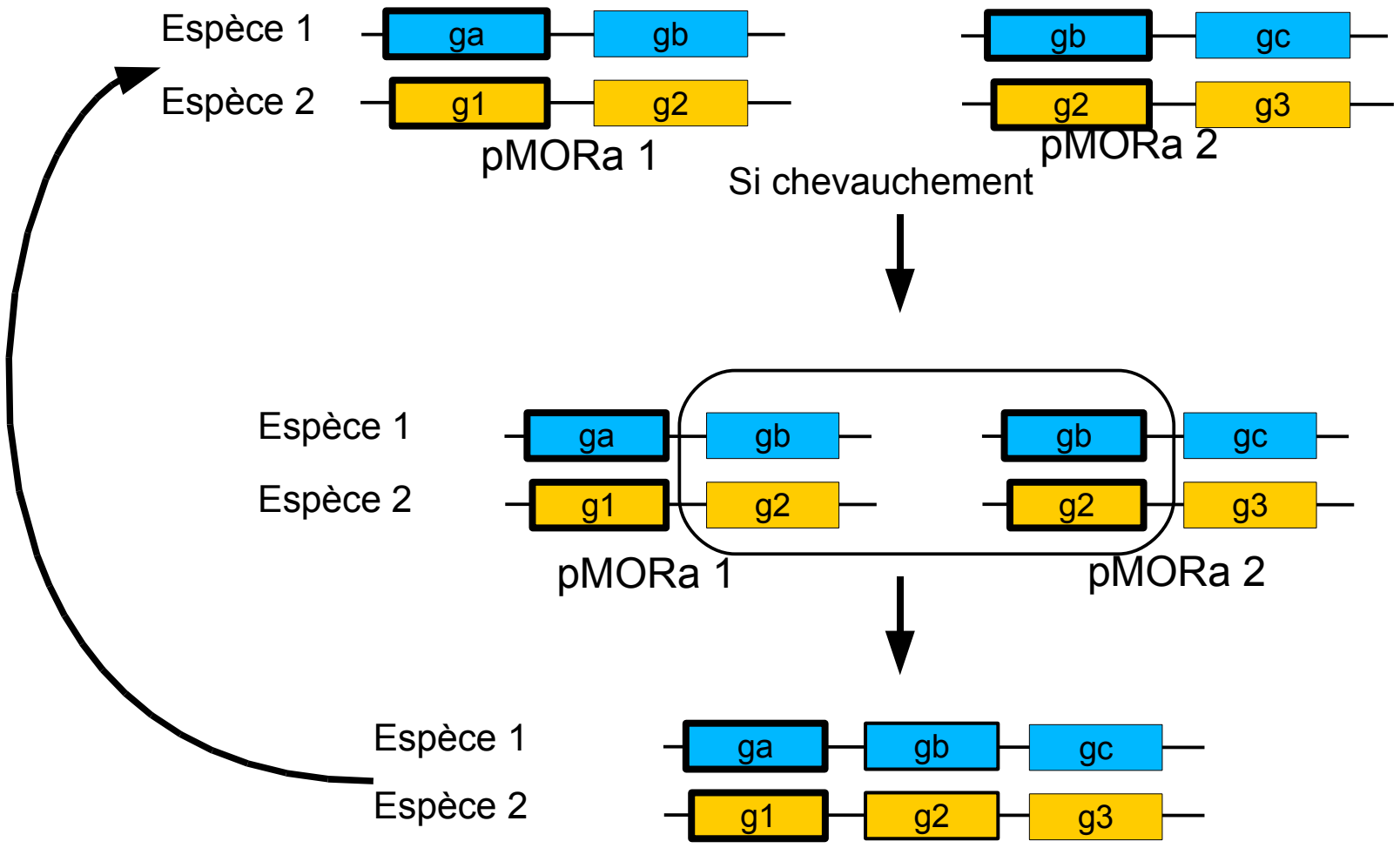

```
Select * from bestortho as b1, bestortho
where      b1.pidtemp_1 = (b2.pidtemp_1+1)
     AND ( b1.pidtemp_2 = (b2.pidtemp_2+1)
           OR
           b1.pidtemp_2 = (b2.pidtemp_2-1))
     AND   b1.strand_1 = b2.strand_1
     AND   b1.strand_2 = b2.strand_2;
```

Consécutifs

Même brin

Détection des régions dont l'ordre est conservé

3 - Élongation des pMORa



Programme synteblock

Détection des régions dont l'ordre est conservé

- Algorithmes implémentés en langage **Perl**
- Temps de calcul sur les 107 génomes:
  - Comparaisons 2 à 2 : 3 ans
  - Détection des MOR : 44 min
  - Calcul des pMORa : 8 min
  - Elongation des régions de synténie: 26 min

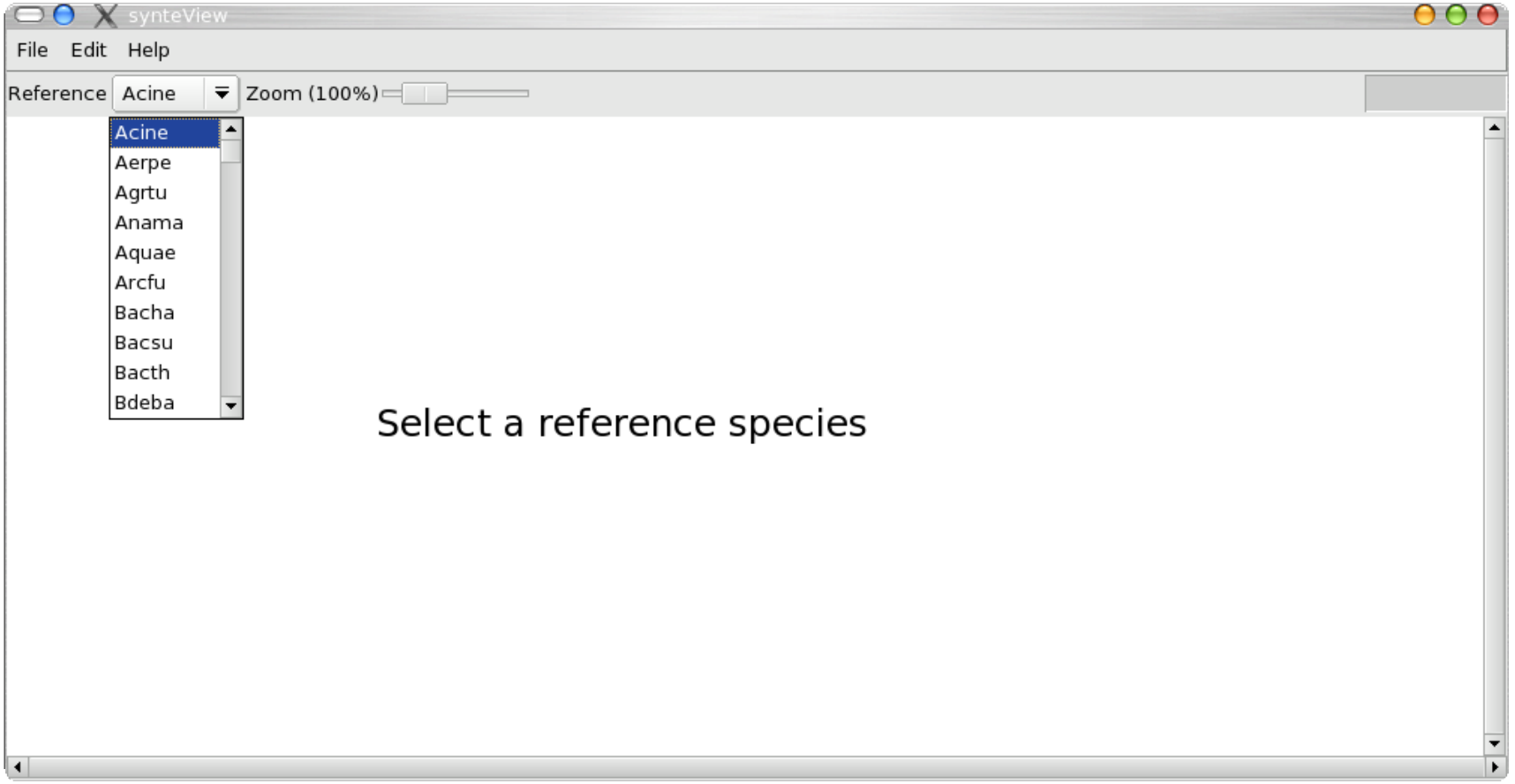
~ 1h20

Plan: Démarche

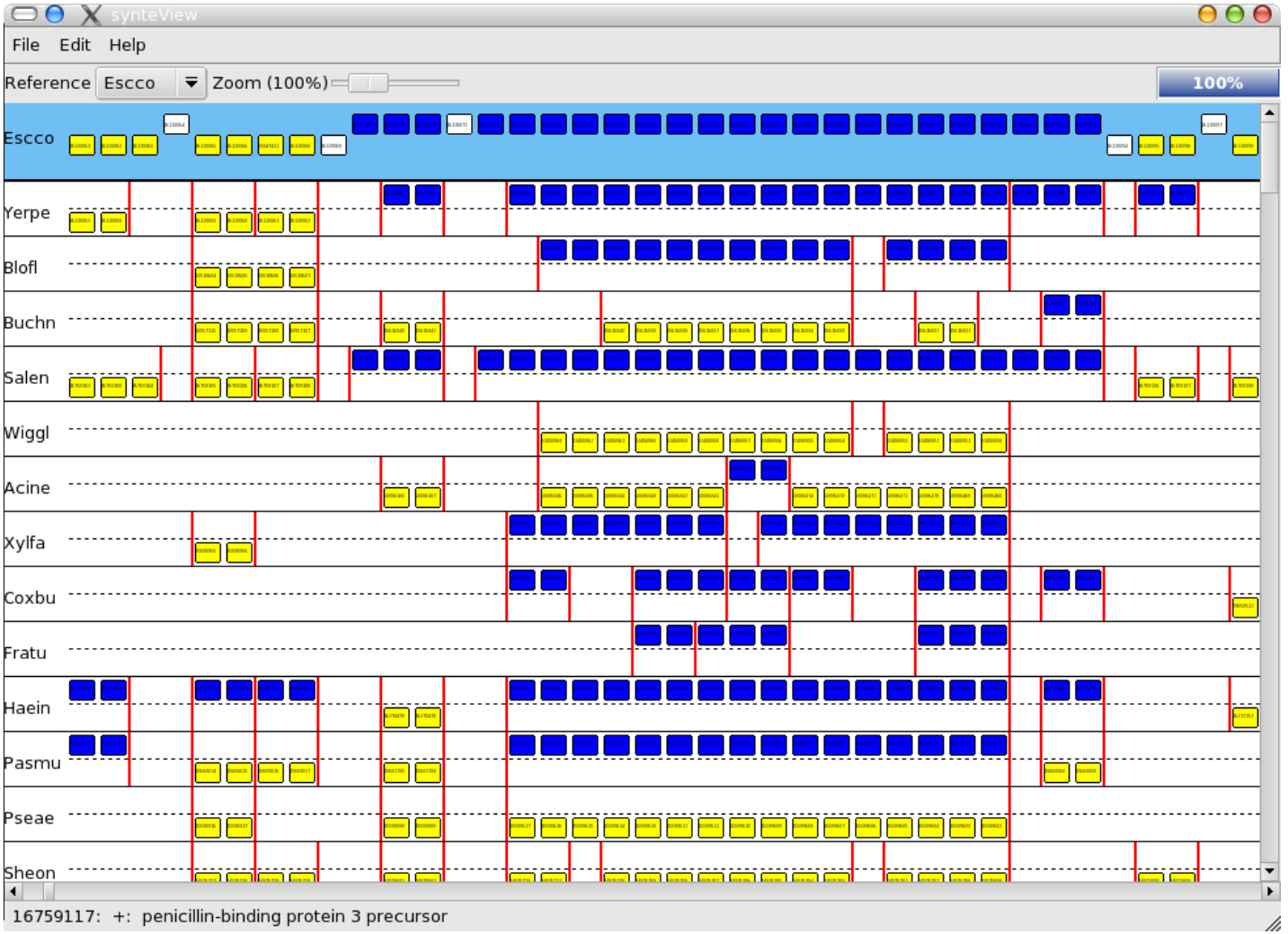
1. **Détection** des régions génomiques dont l'ordre est conservé
2. **Visualisation** des résultats
3. Étude de la **taille** des blocs de synténie
4. Étude des modes **d'évolution** des protéines

Visualisation des résultats

Sélection d'une espèce de référence



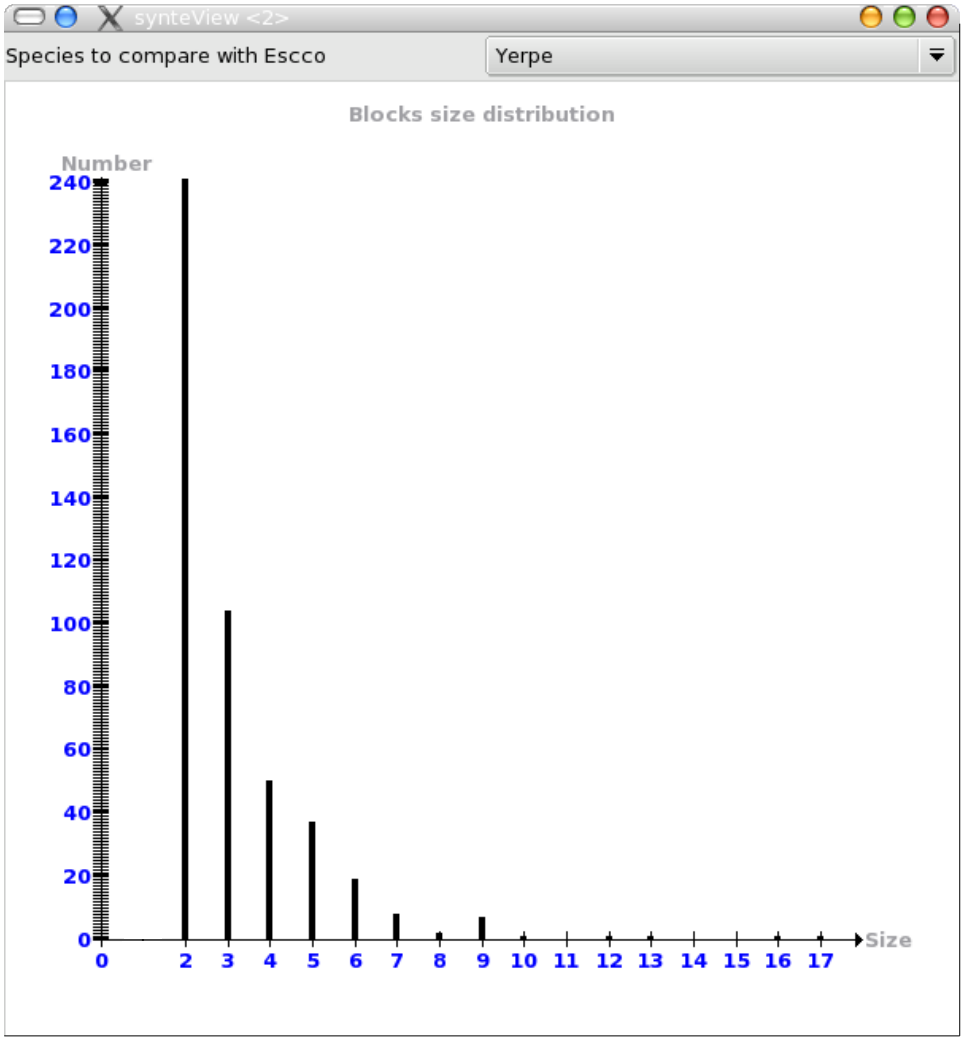
Visualisation des résultats



- Affichage des blocs conservés dans chaque espèce comparé à l'espèce de référence

- Affichage de la fonction, si connue, des protéines

Visualisation des résultats



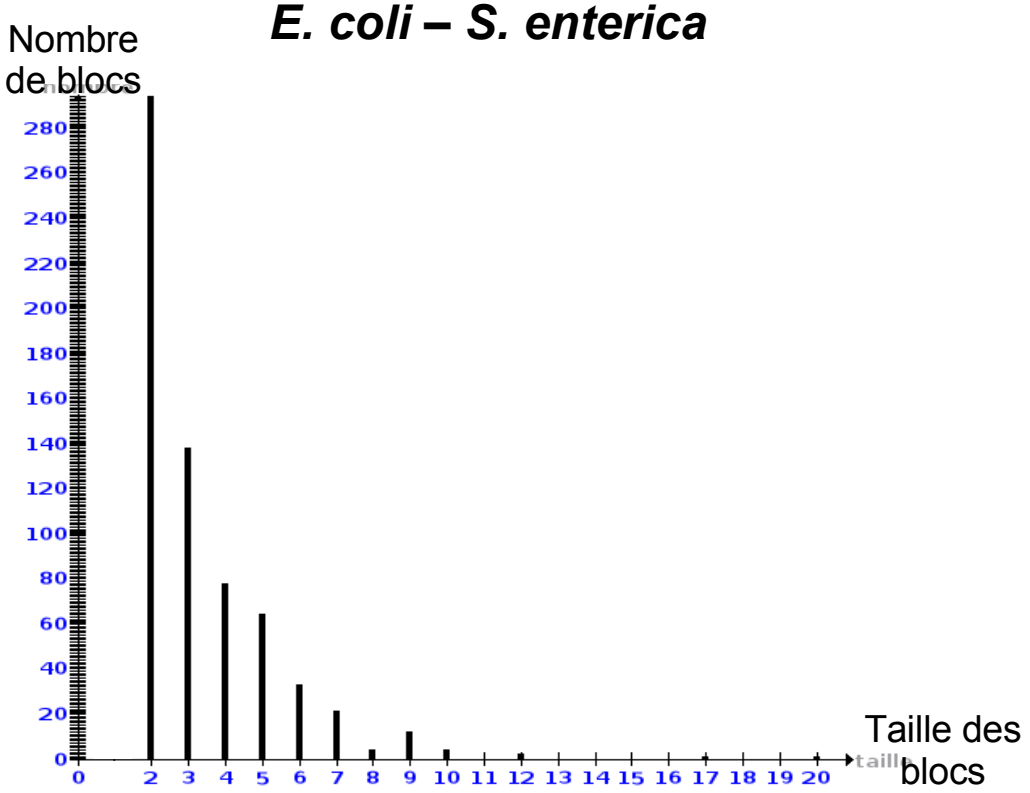
Possibilité d'afficher l'histogramme représentant la répartition des tailles des blocs (en nombre de protéines)

Plan: Démarche

1. **Détection** des régions génomiques dont l'ordre est conservé
2. **Visualisation** des résultats
3. Étude de la **taille** des blocs de synténie
4. Étude des modes **d'évolution** des protéines

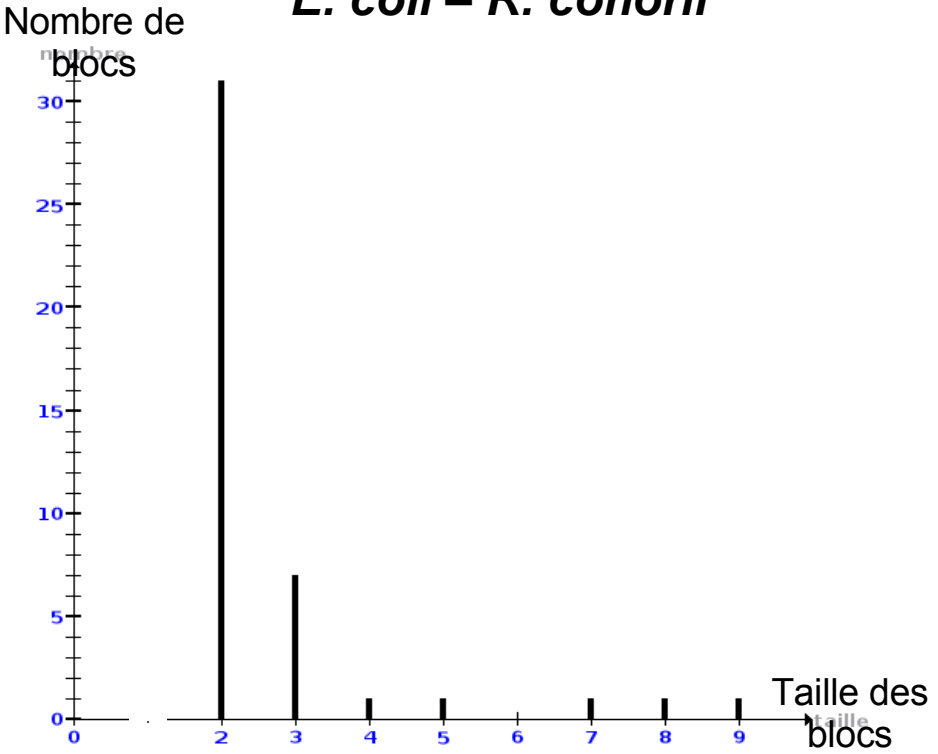
Distribution de la taille des blocs de synténie

Répartition de la taille des blocs
E. coli – S. enterica



- Nombre total de blocs élevé (>630 blocs)
- Beaucoup de petits blocs, et très peu de grands

Répartition de la taille des blocs
E. coli – R. conorii



- Nombre faible blocs (<45 blocs)
- Plus de petits blocs que de grands

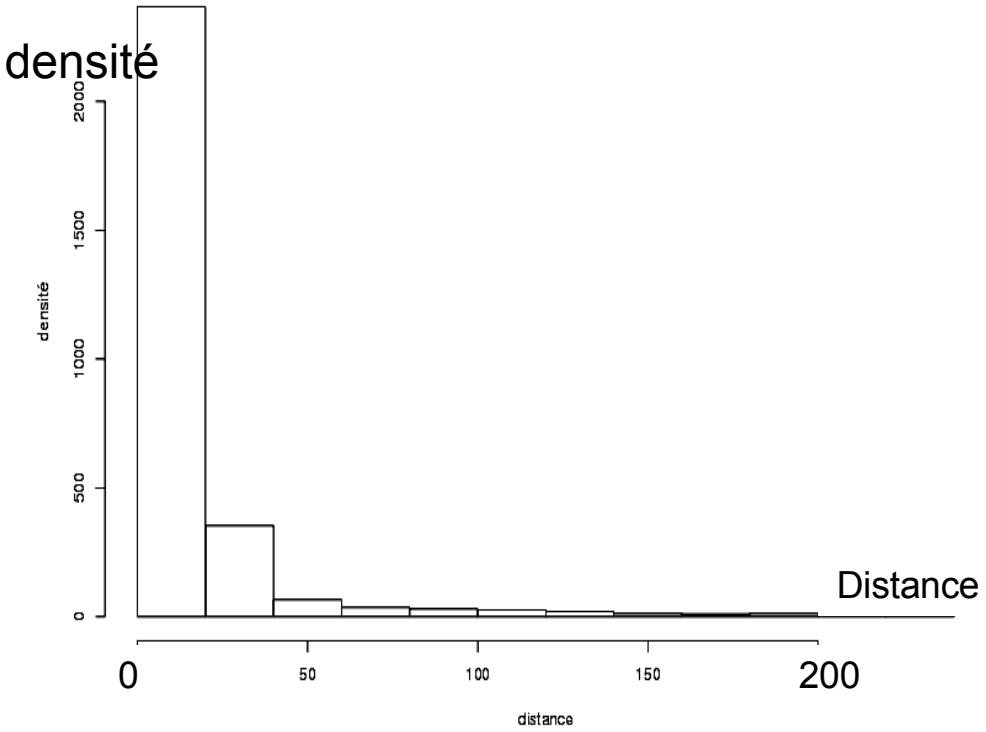
Plan: Démarche

1. **Détection** des régions génomiques dont l'ordre est conservé
2. **Visualisation** des résultats
3. Étude de la **taille** des blocs de synténie
4. Étude des modes **d'évolution** des protéines

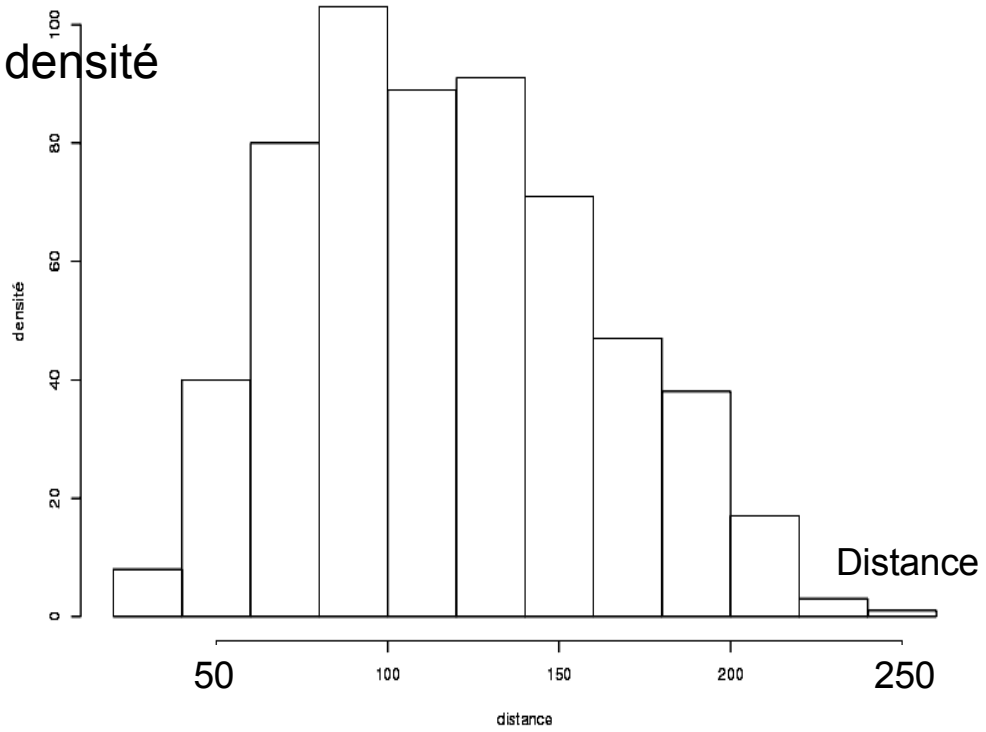
Modes d'évolution des protéines

Sur tous les MOR

Répartition des distances PAM entre *E. coli* et *S. enterica*



Répartition des distances PAM entre *E. coli* et *R.conorii*

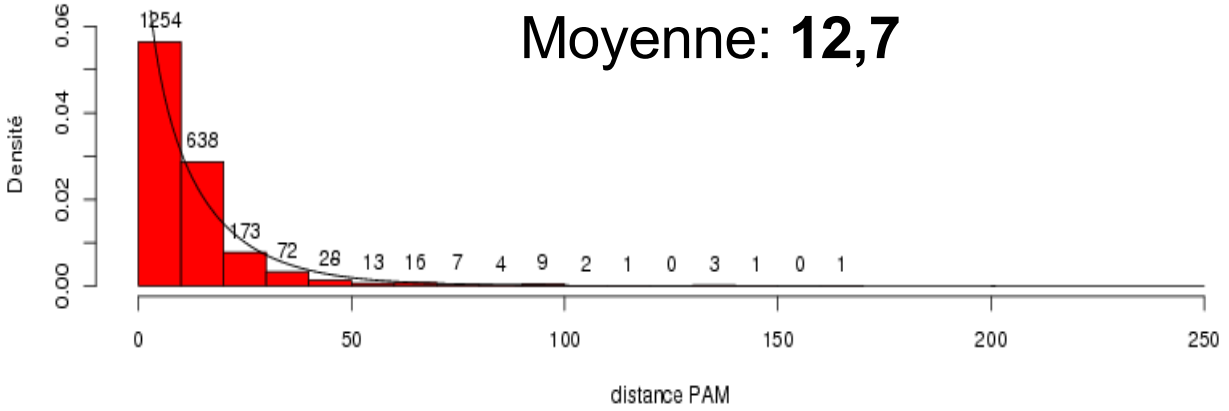


Confirmation: La distribution des distances PAM entre les protéines reflète la distance taxonomique

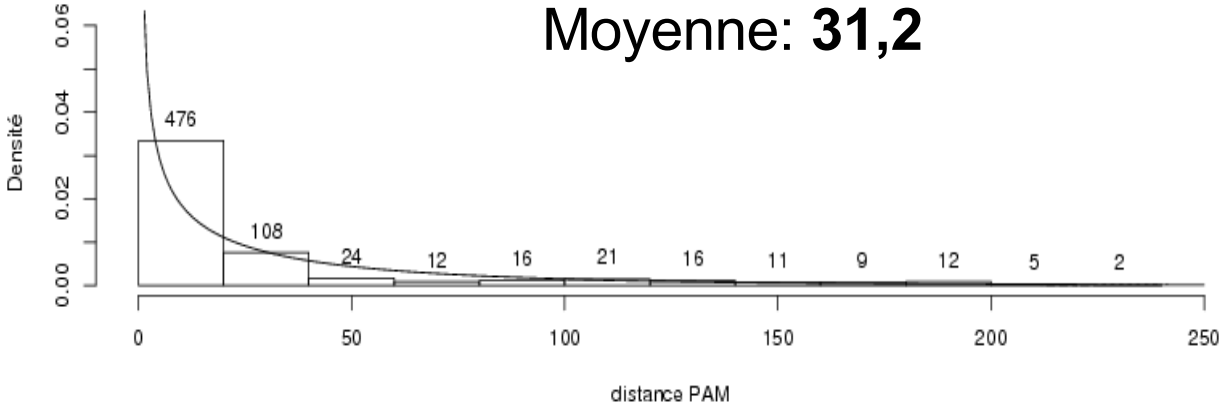
Modes d'évolution des protéines

Distinction entre les MOR appartenant à un bloc de synténie et les autres MOR

E.Coli – S.enterica. Blocs
Moyenne: **12,7**



E.Coli – S.enterica. Hors Blocs
Moyenne: **31,2**

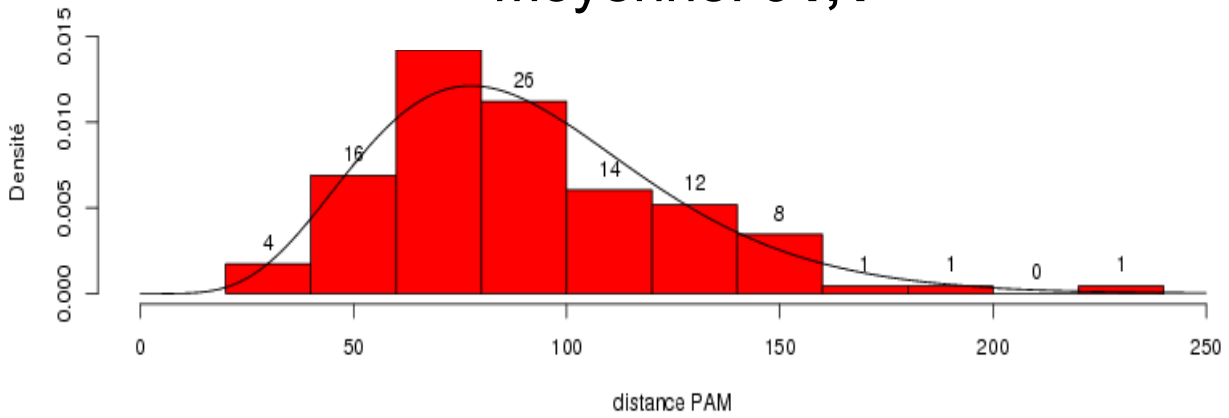


- Distances PAM **faibles**
- Moyennes **différentes**

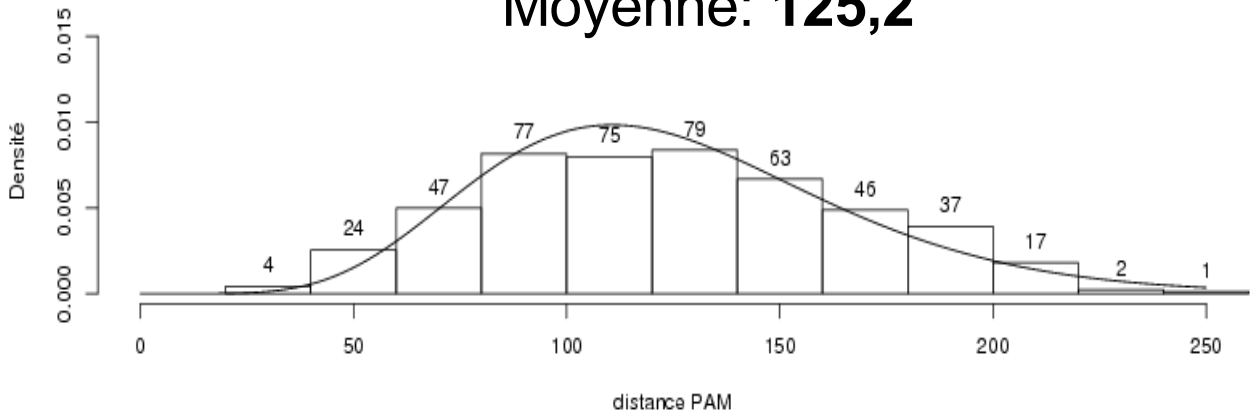
Modes d'évolution des protéines

Distinction entre les MOR appartenant à un bloc de synténie et les autres MOR

E.Coli – R.conorii. Blocs
Moyenne: **91,1**



E.Coli – R.conorii. Hors Blocs
Moyenne: **125,2**



- Distributions **différentes**
- Plus de distances PAM **élevées**
- Moyennes **différentes**

Modes d'évolution des protéines

Distinction entre les MOR appartenant à un bloc de synténie et les autres MOR

- Test statistique H0: E(X1)=E(X2) / H1: E(X1)<E(X2)
  - X1 : distances dans les blocs de synténie,
  - X2 : distances à l'extérieur
- On ne compare pas la statistique de test à une loi de Student, mais à une loi **Bootstrap**

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\hat{S}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T^* \qquad \text{Avec} \qquad T^{*(b)} = \frac{\bar{X}_1^{*(b)} - \bar{X}_2^{*(b)} - (\bar{X}_1 - \bar{X}_2)}{S^{*(b)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Modes d'évolution des protéines

Distinction entre les MOR appartenant à un bloc de synténie et les autres MOR

- Dans 5563 comparaisons inter espèces sur 5671 (98%), H0 est rejetée
- Dans les 108 autres cas (2%): Espèces très éloignées

- **Il existe des forces de sélection différentes qui s'exercent sur les protéines appartenant aux blocs de synténie par rapport aux autres**
- **Interaction des produits**
- **Synthèse simultanée**

→ **Contrainte**

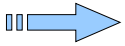
Conclusion

• L'outil de visualisation mis au point permet **d'observer** les différentes **régions** de manière aisée et d'observer les **fonctions** des protéines impliquées dans ces relations de voisinage

• Le **nombre de régions** conservées décroît très rapidement avec la distance taxonomique. La distribution des **distances** entre les protéines et la distribution des **tailles** des blocs de synténie dépendent en partie de la taxonomie

•Les gènes qui constituent les blocs sont plus contraints.

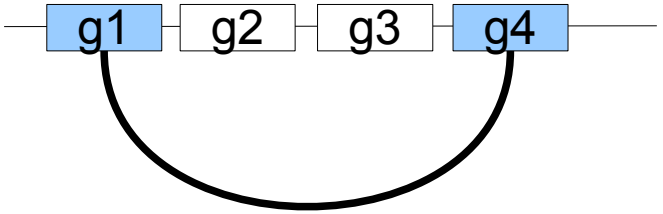
- Pression sur la séquence
- Pression sur l'ordre des gènes



Contexte génétique

Perspectives

- Définition moins stricte du voisinage



- Gènes paralogues
- Analyse fonctionnelle (répartition des fonctions, annotation)
- Utiliser Ka/Ks
- Annotation