Magistère Biotechnologies

Analyse in silico de génomes, protéomes et transcriptomes

« Génomique comparative » V.2012.1

Protocole TD

Notes :

Scripts et données sur : http://rna.igmors.u-psud.fr/gautheret/cours/AnalInSilico

---

1. Introduction

Prérequis :
- Prise en main d'Unix/Linux. Commandes Unix de base

Présentation de la symbiose Buchnera aphidicola / puceron: voir le support pdf

2. Récupération des protéomes

- Récupération des données sur le site `http://www.ebi.ac.uk/genomes/bacteria.html`
  - Prendre les protéomes (format fasta) de: Buchnera aphid. Subsp Acyrthosiphon pisum (souche 5A) et E. coli K12 (souche MG1655)
- Vérifier le contenu des fichiers avec un éditeur de texte (protéines au format Fasta)
- Compter le nombre de protéines de chaque espèce à l'aide de la commande egrep (cf TD Unix).

3. Utilisation programme Perl: compter les séquences, taille moyenne.

- Utilisation du programme Perl `check-datafile` :

check_datafile.pl permet de verifier la cohérence des fichiers fasta (fichier.fasta) téléchargés
usage : perl check_datafile.pl [fichier.fasta]

- Nombre de protéines coli et Buch ? taille moyenne ?
- Y a-t-il une différence entre les tailles des protéines de ces deux génomes ?

4. Comparaison de 2 protéomes dans le but de trouver les orthologues

- Recherche d'orthologues :
  - duplications / paralogues /orthologues et les conséquences sur Blast
- Principe du « best reciprocal hit »
- Rappels Blast (si nécessaire)
  - (k-words), formattage de la base,
  - Run en ligne de commande. Options de base.
  - Interprétation du fichier de sortie. E-value.

- Comparer « manuellement » les protéomes de *coli* et *Buchnera* à l'aide de Blast. Lancer Blast avec l'option `-m 9`, dans les deux directions : *coli* vs. *buchnera* et *buchnera* vs. *coli*.
  Attention : c'est long ! Redirigez les résultats dans un fichier de sortie avec « > »
  Regardez à quoi ressemble le fichier de sortie et comprenez bien le format.
- A l'aide des scripts Perl ci-dessous, obtenir la liste des meilleurs hits Blast dans chaque direction, puis la liste des couples présents dans les deux listes (meilleurs hits blast réciproques).

first_line.pl — permet de filtrer les resultat de blast (genome1_vs_genome2) en ne gardant que le meilleur hit si sa E-value est < 1e-5.
usage : first_line.pl [fichier-blast] > [fichier]
Attention : requiert des résultats de Blast obtenu avec l'option –m 9

bbh.pl — permet de comparer des fichiers BLAST filtrés. Seuls les BLAST reciproques sont retenus.
usage : bbh.pl [fichier firstline 1] [fichier firstline 2] > [fichier]

- Combien de couples d'orthologues Buch/coli ?

5. Récupération des gènes spécifiques de *Buchnera*

- A l'aide des résultats précédents, réaliser à la main un diagramme de Venn des gènes communs aux deux espèces et spécifiques de chaque espèce.
- Pour trouver les gènes du génome de *Buchnera* absents dans le génome de *coli*, comparer buch.fasta avec les résultats du bbh, à l'aide du programme suivant :

missing_names.pl — recherche les noms de gènes d'un fichier fasta absents dans un fichier bbh
usage: missing_names [fichier fasta] [fichier bbh]

- A l'aide du script `get_seq.pl` et récupérer les descriptions (titre long) des protéines spécifiques de *Buchnera*

get_seq.pl — permet d'extraire d'un fichier fasta les séquences dont la ligne de commentaires contient la chaîne de caracteres recherchée
usage : get_seq.pl [fichier fasta] [fichier d'identifiant protéine]
(le fichier d'identifiants peut être celui généré par missing_names.pl)

6. Analyse des gènes spécifiques à *Buchnera*

- Parmi les gènes spécifiques à *Buchnera*, deux sont des gènes liés à la formation du flagelle. Recherchez parmi les protéines de *Buchnera* combien possèdent le mot flagelle dans leur nom,

puis utilisez un test de Fisher (par ex : http://www.langsrud.com/fisher.htm) pour savoir si les gènes de flagelle sont plus représentés parmi les gènes spécifiques à *Buchnera* que dans l'ensemble des gènes de *Buchnera*.

7. Identifier les duplications ou pertes récentes

Si le meilleur hit de X est Y et que le meilleur hit de Y n'est pas X, alors il y a eu perte ou duplication récente dans l'un des génomes (paralogue unique chez une espèce). Nous recherchons les gènes dans ce cas.

- Rechercher les duplications ou pertes récentes à l'aide du script `diff.pl`

diff.pl permet de rechercher et d'afficher les lignes différentes entre deux fichiers
usage : diff.pl [fichier firstline] [fichier bbh]
(affiche les lignes trouvées uniquement dans le fichier 1, puis celles trouvées uniquement dans le fichier 2)

- Attention : utiliser diff avec des fichiers dans lesquels les noms de gènes sont inscrits dans le même ordre.
- Combien de paralogues uniques à coli ? à Buchnera ?

8. Protéome ancestral

Comment vérifier si *Buchnera* a évolué par perte de gènes par rapport à l'ancêtre des gammaprotéobactéries, ou si c'est *coli* qui a évolué par gain de gènes ? Pour répondre à cette question il faut récupérer un protéome d'une espèce (espèce X) située dans un groupe extérieur au groupe de *Buchnera* et *coli* et comparer le protéome X à celui de *coli* et de *Buchnera*.

- Récupérer le protéome de l'espèce X

- Que suggère la taille du protéomes X et des protéomes de *Buchnera* et de *coli* ?

- Rechercher les orthologues entre protéome X et chacun des deux autres protéomes. Pour faire les Blast plus rapidement, nous utiliserons le script autoblast.pl :

autoblast.pl permet de blaster toutes les protéines de [genome1] contre la banque [genome 2]. Réalise automatiquement le formatdb et lance Blast avec l'option –m 9 (sortie tabulée)
usage : autoblast.pl [genome1] [genome2]
(pas la peine de rediriger les sortie, le script crée fichier genome1_vs_genome2.bl)

- En vous basant sur les nombres d'orthologues entre les espèces prises deux à deux, concluez sur le protéome ancestral et le mode d'évolution de *Buchnera*.

9. Adaptation à l'hote : Comparaison de 3 Buchnera

Quels gènes conférent à *Buchnera* la capacité de symbiose avec son hôte spécifique ?
Il y a maintenant plusieurs génomes de *Buchnera* entièrement séquencés. Si ces souches coexistent avec des hôtes différents, chacune doit avoir un jeu de gènes spécifique de son hôte. On va donc faire une comparaison 3 à 3.

- Récupération des génomes de *Buchnera* de puceron de l'orge (souche sg ou *Schizaphis graminum*) et de puceron du cèdre (souche *Cinara cedri)* sur le site EBI.

- A l'aide du script autoblast, des scripts précédents et du script Venn.pl, réaliser un diagramme de Venn 3x3 pour faire apparaitre les orthologues des 3 protéomes et des protéomes 2 à 2.

venn.pl permet de trouver la liste des protéines communes à trois fichiers de type bbh
usage : venn.pl [fichier bbh 1] [fichier bbh 2] [fichier bbh 3]

- Gènes en commun : Y a-t-il plus de gènes en commun entre *Buchnera aphidicola* et *Escherichia coli* qu'entre les trois souches de *Buchnera* ? Discutez la variabilité intra-souche et la variabilité inter-espèces ?
- Quelles est la souche la plus divergente ? Quelles sont les souches les plus ressemblantes ?

10. Gènes spécifiques à une souche

- A l'aide des scripts missing-names.pl et get_seq.pl, établissez la liste des gènes spécifiques à chacune des 3 souches de Buchnera. Il faudra pour cela concaténer les fichier bbh pertinents à l'aide de la commande cat, afin de créer des listes de gènes à exclure. Attention à l'ordre des gènes dans les fichiers pour la commande missing_names.

- Pouvez-vous trouver des fonctions sur-représentées dans ces gènes spécifiques ?

11. Adaptation à l'hôte : Comparaison de 3 *E. coli*

Nous allons maintenant comparer 3 souches de *E. coli*. La souche de laboratoire K12 et les deux souches pathogènes O157-H7 et coli CFT073.

- Question bibliographique : quelles sont les spécificités d'hôte des 3 souches? (donner références)
- Récupérer à l'EBI les protéomes de : coli O157-H7 (st sakai) et coli CFT073.
- Réaliser un diagramme de Venn pour faire apparaitre les orthologues des 3 protéomes et des protéomes 2 à 2.

- Quelles sont les souches les plus divergentes ? Explication possible ?
- Comparer les diagrammes de Venn obtenus entre les 3 souches de *Buchnera* et entre les 3 souches de *E. coli*. Que peut-on en dire ?

- En combinant judicieusement les scripts diff.pl, get_seq.pl et/ou missing_names.pl recherchez la liste des gènes communs aux deux pathogènes et absents chez coli K12 et celle des gènes spécifiques de l'un des pathogènes.

- Trouvez-vous une classe de gènes sur-représentée dans les pathogènes ou dans l'un des pathogènes ?

12. Comparaisons au niveau génomique

On se pose ici la question : comment ont émergé les gènes responsables de la pathogénicité des souches de coli? par des insertions ponctuelles ? en bloc ? Pour répondre, il faut étudier la synténie. Nous allons comparer K12 et O157 :H7 au niveau de la séquence génomique, avec un programme de dot-plot : Gepard.

- Télécharger les génomes (.fna) de K12 et O157 :H7 sur le site de l'EBI.
- Lancer gepard en ligne de commande « sh <chemin>/gepard.sh » (s'il est installé sur vos machines) ou en mode Java Webstart à l'adresse : http://mips.helmholtz-muenchen.de/services/analysis/gepard (choisir version normale, 512 MB).
- Charger les deux génomes et afficher le dotplot

- Question : comment sont arrivés les gènes responsables de la pathogénicité : par bloc ou par insertions ponctuelles?

- Effectuer dans la fenêtre inférieure du programme l'alignement des deux régions homologues.

13. Arbre des 6 espèces

Question : le nombre de gènes communs est-il un indicateur de proximité évolutive ? Non. On peut avoir une perte massive dans un temps court, ou au contraire une divergence de longue durée sans perte de gènes. Voir par ex. le nombre de gènes communs entre les *Buchnera* et entre *Buchnera* et *coli*. Pour confirmer ce fait, nous allons faire un arbre phylogénétique des 6 espèces étudiées.

- Classe séparée entre 2 groupes.

- Groupe 1 : Arbre 16S. Prendre 16S de *E. coli* (fichier 16s.seq). Blaster contre chaque génome. Récupérer les six 16S. Réaliser l'arbre général sur phylogeny.fr

- Groupe 2 : Arbre metaprotéines. Trouver des gènes communs aux 6 espèces (à partir des Venn). Concaténer les séquences protéiques en 5 « metaprotéines ». Réaliser l'arbre général sur phylogeny.fr. Attention aux tailles des protéines (ne pas prendre des protéines trop grandes).