Initiation à Galaxy - mise en autonomie sur un sujet simple

L'objectif de ce TP est de vous proposer un exercice avec des consignes minimum, afin que vous puissiez gagner en autonomie sur l'utilisation de Galaxy. Le sujet en est la phylogénie, car il fallait un sujet biologique, mais cette notion n'est qu'illustrative (on tentera néanmoins de ne pas dire de bêtises sur ce thème).

La théorie

La phylogénie moléculaire est l'étude de l'histoire évolutive d'une famille de gènes homologues. C'est-à-dire qu'on considère un ensemble de gènes qui se ressemblent suffisamment pour qu'on leur suppose un ancêtre commun, et qui proviennent d'espèces différentes. Partant des séquences de ces gènes, on reconstruit un arbre retraçant l'histoire évolutive, qu'on appelle l'arbre phylogénétique.

La pratique

Les séquences d'ADN d'une famille de gènes homologues ont été déposées dans une librairie publique appelée "Formation Galaxy". Votre mission, que vous avez déjà accepté malgré vous, est de reconstruire l'arbre phylogénétique correspondant.

Partant d'un fichier de séquences, on cherche à obtenir un arbre. Plusieurs étapes sont nécessaires pour y parvenir:

- ➔ Il faut d'abord aligner les séquences, c'est-à-dire insérer des caractères de *gaps* afin de faire correspondre les portions similaires. Un des logiciels les plus performants d'alignement de séquences s'appelle MAFFT.
- ➔ Si vous voulez afficher dans de bonnes conditions l'alignement obtenu, vous devrez installer un logiciel sur votre ordinateur en local, et l'utiliser pour ouvrir le résultat. Vous pouvez par exemple utiliser le logiciel Seaview (développé au PBIL).
- ➔ Un alignement multiple contient la plupart du temps des portions artefactuelles, c'est-à-dire des colonnes ne correspondant pas à des portions homologues entre les séquences, mais simplement à du bruit. Dans ce cas on utilise en général un outil spécialisé, GBlocks ou Trimal étant les existants les plus connus. Ces outils vont enlever de l'alignement donné en entrée les sites (ou les colonnes) les moins fiables.
- ➔ L'alignement multiple peut alors servir à inférer un arbre phylogénétique. La taille modeste de nos données nous permet d'employer un logiciel probabiliste, en l'occurrence PhyML.

Note: La configuration par défaut d'un programme est souvent celle considérée comme la plus générique par les auteurs.

Subsidiaire 1

Il est possible d'afficher un arbre dans Galaxy. Vous pouvez par exemple utiliser le module "Southgreen visualisation".

Subsidiaire 2

Pourquoi ne pas créer un *workflow* afin de retracer cette démarche, et la partager avec un collaborateur (ou au pire un voisin de tablée) ?

Subsidiaire 3

Vous pourriez créer un clone du premier workflow, le configurer pour qu'il fonctionne avec des données protéiques, et pourquoi pas l'essayer sur la famille "famille_Proteine.fasta" de la librairie "Formation Galaxy' ?