Analyse phylogénétique

Cette séance traite de la prise en main de l'outil de calcul Galaxy, et de son utilisation pour différentes analyses de phylogénie.

Au préalable

Connectez vous sur:
http://gohelle.cirad.fr/galaxy/root/
en utilisant le login et le mot de passe fourni en début de séance.

Les deux fichiers de séquence de départ sont disponibles dans la librairie "Supagro2015", dans le répertoire phylogénie.

D'autres instructions d'installation et connexions à des outils seront fournis au file du sujet de TP.

Recherche d'homologues par BLAST

De la théorie...

Voici trois définitions à bien distinguer, les deux premières s'appliquent aux séquences, et la dernière s'applique aux objets biologiques:

1. Similarité: mesure mathématique de proximité entre deux séquences.
2. Couverture: longueur relative ou absolue le long de laquelle deux séquences sont comparables.
3. Homologie: deux molécules sont homologues si elles ont dérivé d'une molécule ancestrale commune.

Ces définitions impliquent que, pour trouver des molécules dont l'histoire évolutive commune a un sens, on modélise ces molécules par des séquences, et on tente d'attester de l'homologie des molécules sur la base de la similarité des séquences.

Il y a principalement deux grandes approches qui permettent d'aborder un problème de phylogénie:

1. La recherche d'homologues par BLAST (ou avec d'autres outils comme par exemple HMMer) dans les banques: partant d'une séquence d'intérêt, on cherche dans des banques généralistes un ensemble de séquences similaires. Ces séquences proviennent la plupart du temps d'autres espèces, plus ou moins proches de notre espèce d'intérêt.
2. Le clustering de séquences: partant d'un grand nombre de séquences d'intérêt, classer les séquences par groupes de similarité, dans le but de former des familles d'homologues.

Nous ne traiterons pas cette approche ici.

Notions abordées: "portée" de la reconstruction phylogénétique, similarité de séquences, homologie de molécules.

Technologies abordées: Bases de données de séquences, BLAST.

5 types de BLAST existent :

Query	Method	Subject
Nucléique	BLASTN	Nucléique
Nucléique traduit	BLASTX	Protéique
Protéique	BLASTP	Protéique
Nucléique traduit	TBLASTX	Nucléique traduit
Protéique	TBLASTN	Nucléique traduit

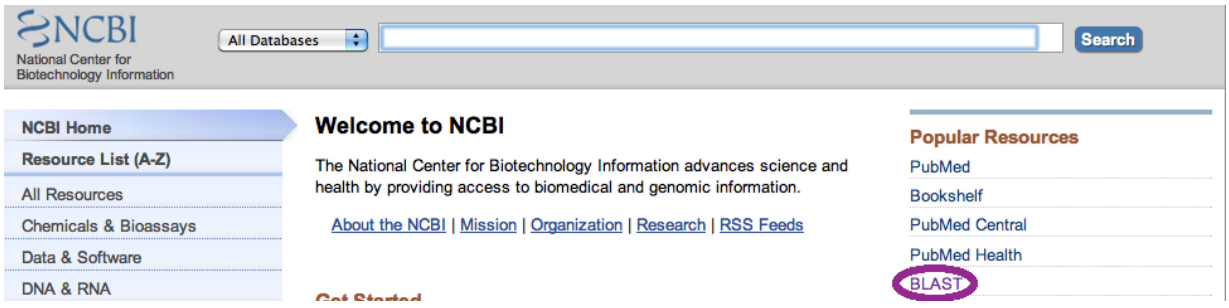
Résultats de Blast donnent, entre autres, des valeurs de :

- Score = valeur de 'qualité' de l'alignement, dépend de la longueur et de la similarité entre les séquences
- Query coverage = pourcentage de la séquence initiale retenue par l'alignement local
- E value = nombre de séquences de la base qui auraient donné la même valeur de score par hasard (pas une probabilité). Donc la e value dépend de la taille de la base (pour un même alignement, la e value est plus grande si la base est plus grande) et donc une e value faible est d'autant plus 'significative' que la base est grande.

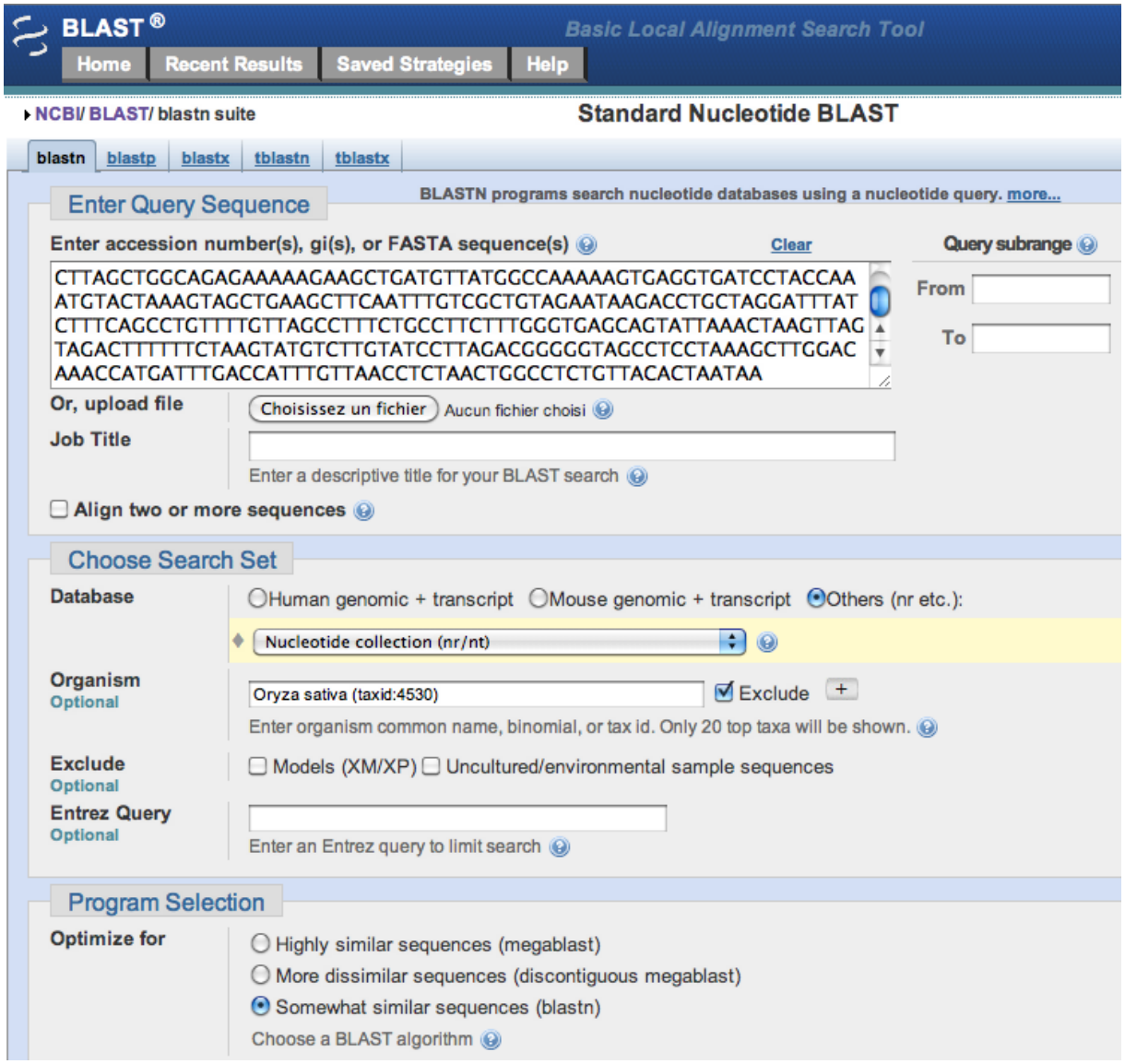
A la pratique...

Cette première mise en pratique traite de la recherche de séquences similaires par BLAST dans différentes espèces et création d'un fichier multifasta pour alignement nucléique.

1. **charger la page du NCBI (http://www.ncbi.nlm.nih.gov/)**
2. **aller sur le lien BLAST, puis nucléotide blast**

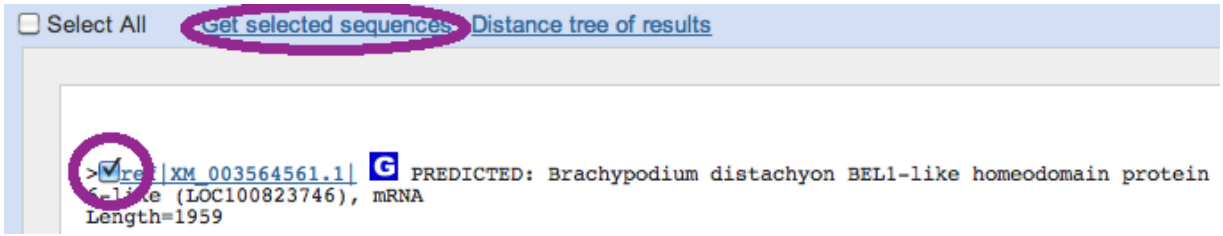


3. **copier la séquence d'intérêt dans l'encart prévu**
4. **régler les paramètres de query (others/nr)**
5. **régler les paramètres de query sur "Exclude Oryza sativa".**
6. **régler le programme de recherche sur BlastN**

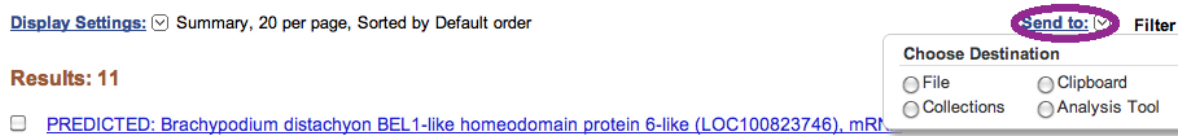


7. **lancer le blast**

Une fois les résultats obtenus, sélectionnez une dizaine de séquences d'espèces variées:



et les télécharger au format FASTA:



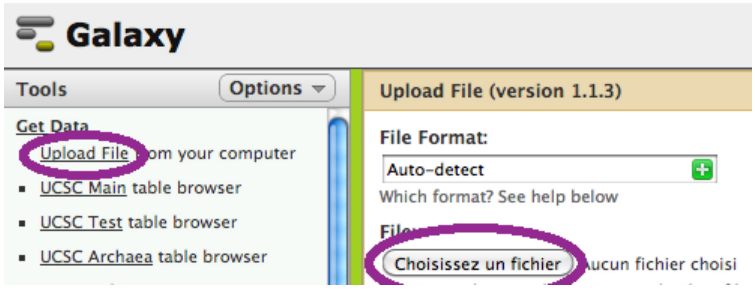
Notes technique:
Le fichier obtenu contient en général des noms de séquences longs et verbeux, qui nécessitent un nettoyage à la main. Vous pouvez récupérer un fichier propre et prêt à l'emploi en ajouter un "s" à "sequence" dans l'adresse de la séquence d'intérêt:
http://gohelle.cirad.fr/phylogeny/formation/sequences.fasta

Alignement de séquences, et nettoyage de l'alignement

De la théorie...
Aligner des séquences a pour objectif de mettre en correspondance les portions homologues des molécules, afin de retrouver de la façon la plus cohérente possible le signal phylogénétique. La plupart des méthodes d'alignement multiple sont basées sur l'algorithme de l'alignement progressif: son principe est d'aligner les séquences par paires de blocs, en suivant un arbre guide. Le nettoyage de l'alignement (*masking*, curation) consiste à sélectionner dans un alignement multiple imparfait l'ensemble des sites effectivement représentatifs de l'homologie.
Notions abordées: artefacts d'alignements, biais de curation, qualité d'alignement, signal phylogénétique.
Technologies abordées: alignement avec MAFFT, sélection de blocs avec GBlocks, édition avec Seaview.

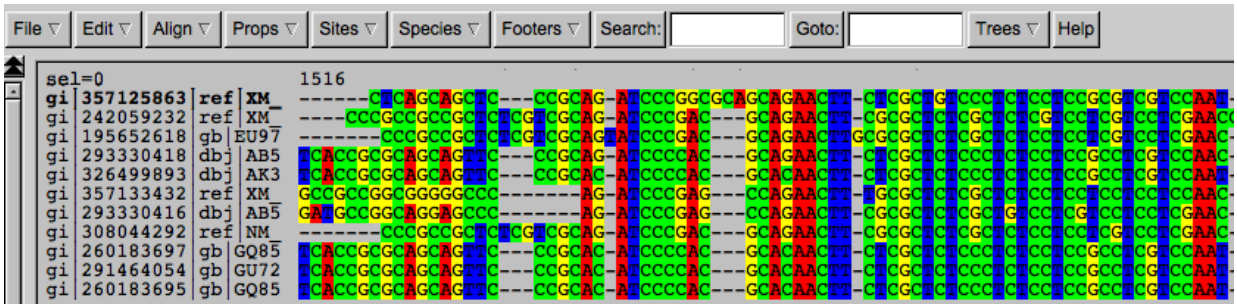
… A la pratique
Cette étape nécessite la prise en main de l'interface Galaxy, ainsi que l'utilisation d'un logiciel tiers à installer sur votre ordinateur:
http://pbil.univ-lyon1.fr/software/seaview.html
Les étapes de pratique sont décomposables comme suit:

1. **Placer dans votre historique Galaxy les fichier séquence:**



2. **Utiliser la brique "MAFFT", dans la catégorie "Sequence comparisons", sur votre jeu de données. Faire varier les pénalités d'ouverture et de prolongation de *gap* (*gap opening penalty* et *gap extension penalty*). Comparez les résultats obtenus en utilisant "Seaview".**
3. **Utiliser la brique "GBlocks", dans la catégorie "Sequence comparisons", et testez**

différents paramétrages (taille des blocs détectés, similarité des blocs, et quantité de *gaps* tolérés). Comparez les résultats obtenus avec Seaview ou un navigateur web.



Notes techniques:

MAFFT contient diverses options ayant une grande influence sur l'alignement résultat:

1. La matrice de dissimilarité, qui n'est utile que pour les données protéiques.
2. Le nombre d'itérations, qui augmente le temps calcul et la qualité des résultats.
3. La pénalité d'ouverture et de prolongation de *gaps* permettent de moduler l'inférence des délétions et insertions (0 de pénalité de prolongation permet par exemple un coup fixe à chaque *gap* quel que soit leurs longueurs).
4. La méthode de distances utilisée, qui influe sur les temps calculs et sur le résultat, selon l'adéquation de la méthode de distance aux données.

GBlocks est également un logiciel très paramétrable, qui va nettoyer de façon plus ou moins stricte, les alignements de séquences. Les principaux paramètres sont:

1. Le nombre minimum de séquences conservées, pour un bloc et pour une position à côté, ils peuvent être réglés à 51% et 52% pour être le moins strict possible. Les paramètres par défaut sont davantage sélectifs.
2. La taille minimum d'un bloc d'homologie est le paramètre le plus porteur de sens. Il permet de formuler une hypothèse d'homologie: un bloc conservé est considéré homologue à condition qu'il comprenne un nombre de site minimum.
3. La quantité de *gaps* autorisée, soit aucun, soit tous, soit la moitié des séquences. Cela permet de considérer, ou pas, valides les séquences partiellement couvrantes.

Recherche d'homologues en utilisant HMMer

De la théorie...

Une fois l'étape d'alignement maîtrisée, il est possible d'appliquer une recherche d'homologues en utilisant les modèles de Markov cachés. L'avantage est qu'on ne cherche pas de ressemblance à une séquence unique dans une base, mais la ressemblance à un profile de plusieurs séquences. La conséquence est que l'homologie est alors plus représentative de la diversité de la famille.

2 versions de HMMer existent: le HMMer standard qui travaille sur des séquences protéiques, et nHMMer qui travaille sur des séquences nucléiques.

Dans l'exemple que nous allons traiter, nous avons à notre disposition quelques séquences chez le bananier identifiées comme étant de la famille des expansines, et nous souhaitons inférer

l'histoire évolutive de cette famille chez les monocotes, en échantillonnant le riz asiatique, le bananier, et divers palmiers.

… A la pratique

Importer les séquences protéiques dans votre historique Galaxy. Une fois cela fait, vous pouvez constituer votre jeu de données famille comme suit:

1. **Alignez vos séquences représentatives avec MAFFT.**
2. **Utilisez le programme HmmBuild sur l'alignement pour en construire un profile**
3. **Utilisez le programme "HmmSearch multispecies" sur le profile, en échantillonnant le bananier (MUSAC), le palmier à huile (ELAGV), le palmier datier (PHODC), et le riz asiatique (ORYSJ).**
4. **Extraire un fasta cohérent, en utilisant "HmmSearch to fasta multispecies", en filtrant par exemple les séquences avec un seuil de E-valeur de 1.0e-10**
5. **Nettoyez le fasta obtenu avec le programme "format fasta header"**
6. **Reproduire l'étape d'alignement/nettoyage sur ce nouveau fasta bien propre.**

Reconstruction phylogénétique

De la théorie...

Un arbre phylogénétique retrace l'histoire évolutive d'une famille de molécules homologues. Il existe pour une phylogénie à n feuilles un grand nombre d'arbres possibles et complètement résolus:

$$\prod_{i=3}^{n}(2i-5) = 1 \times 3 \times 5 \times \ldots \times (2n-5)$$

Soit 3 arbres pour 4 feuilles, 15 pour 5 feuilles, et déjà plus de 2 millions pour 10 feuilles.

Trouver le meilleur arbre parmi tout les arbres possibles est donc une tâche qui ne peut être qu'approximée, en optimisant des critères mathématiques. Il existe 3 grandes classes de méthodes de reconstruction phylogénétiques:

1. Les méthodes "de distance" optimisant la longueur totale de l'arbre (UPGMA, NJ, BioNJ, FastME). Ces méthodes sont très rapides.
2. Les méthodes "de parcimonie" optimisant le nombre de mutations nécessaires pour expliquer l'histoire évolutive.
3. Les méthodes "de maximum de vraisemblance" optimisant la vraisemblance de l'arbre selon les données (l'alignement de séquences) et un modèle d'évolution définit. Ces méthodes sont beaucoup plus lentes. En maximum de vraisemblance sur séquences nucléiques, le modèle GTR (optimisant à partir des données les coûts de substitutions entre bases) avec loi gamma (considérant 4 catégories de vitesse de mutation sur l'ensemble des sites) est le plus standard.

Notions abordées: méthodes de reconstruction phylogénétiques, supports de branche, topologies, longueurs de branches.

Technologies abordées: PhyML, visualiseurs d'arbres.

… A la pratique

Cette étape nécessite l'installation d'un logiciel de visualisation d'arbres, que l'on peut télécharger ici:

http://southgreen.cirad.fr/sites/all/files/Dendroscope_windows_3_0_14beta.exe

Ensuite, retour sur Galaxy:

1. **Convertissez les fichiers FASTA obtenus en sortie de GBlocks en Phylip grâce à la brique "Fasta2Phylip" dans la catégorie "Convert formats".**
2. **Produire une phylogénie par maximum de vraisemblance, en utilisant le modèle GTR + Gamma pour le jeu nucléique, et LG+Gamma pour le jeu protéique. Essayer de les produire d'abord avec les supports SH-like.**
3. **Afficher ces arbres avec l'aide de Dendroscope. Choisir une racine, et essayer de comprendre les hypothèses portées par l'arbre.**

Notes techniques:

PhyML est un logiciel complet d'optimisation d'arbres phylogénétiques par maximum de vraisemblance. Pour un jeu nucléique, le modèle GTR avec loi gamma à 4 catégories est le plus standard. Pour un jeu protéique, on utilise en général LG+gamma ou WAG+gamma. On peut également citer quelques paramètres d'importance:

1. La méthode d'optimisation de la topologie (par NNI, par SPR, ou par NNI+SPR). Un NNI (Nearest Neighbor Interchange) est un mouvement topologie de faible portée, consistant à perturber la topologie autour d'une branche. Un SPR (Subtree Pruning and Regraphing) consiste à débrancher un sous-arbre et le rebrancher ailleurs. Le second est plus coûteux en temps calcul, mais est plus puissant pour éviter de tomber dans un minimum local lors de l'optimisation de la topologie. Employer les deux stratégies et prendre la meilleure des deux en vraisemblance est la méthode produisant la meilleure qualité de résultat.
2. La méthode de calcul de support de branche. La méthode SH rajoute 10% au temps calcul de l'arbre, tandis que le bootstrap multiplie par 100 ou 1000 le temps calcul. La seconde méthode, plus ancienne, est bien établie comme référence dans la communauté des bioanalystes.

Réconciliation d'arbres, racinement et détection de paralogies/pertes

De la théorie...

Un arbre phylogénétique produit par un logiciel comme PhyML n'est pas annoté, et n'a pas de racine définie. La comparaison de cet arbre avec l'arbre des espèces, connu par ailleurs, permet de détecter les évènements de duplications et de pertes, et permet également de proposer une racine minimisant la quantité inférée de ces évènements.

Il existe assez peu de logiciels permettant de faire de la réconciliation automatique, et l'essentiel de la recherche dans le domaine concerne la détection d'évènements plutôt rares chez les plantes: les transferts horizontaux.

Nous allons donc utiliser un logiciel se consacrant à la détection des seuls duplications et pertes: RAP-Green.

… A la pratique

Commencer par récupérer dans votre historique un arbre des espèces de référence, en important "viridiplantae.phyloxml" qui se trouve dans les données Galaxy (shared data / data libraries / taxonomies).

Utiliser ensuite le logiciel RAP-Green (dans la catégorie "evolution"), en utilisant les paramètres par défaut.

Visualiser les arbres résultats "Newick" et "Reconciled newick" à l'aide de Dendroscope.

Notes techniques:

RAP-Green permet de visualiser l'arbre des gènes raciné et annoté, mais aussi beaucoup d'autres types de résultats, comme en particulier l'arbre réconcilié (figurant les compromis topologiques entre arbre des gènes et des espèces, et les pertes).