

## **Sujet de stage de recherche pour Master Recherche en informatique :**

### **reconstruction de super-arbres enracinés, application à la reconstruction phylogénétique multi-gènes.**

Mots clés : Bio-informatique, Super arbres, Phylogénies, Complexité.

Les méthodes de reconstruction phylogénétique ont pour but de reconstruire l'histoire évolutive d'un ensemble d'espèces à partir de données moléculaires (séquences d'ADN ou d'acides aminés).

L'approche classique consiste à 1) sélectionner un gène particulier dont on connaît la séquence pour chacune des espèces étudiées, 2) utiliser un modèle d'évolution qui décrit le processus d'évolution de ce gène et 3) chercher l'histoire évolutive qui, pour ce modèle d'évolution, a le plus de chances d'avoir produit les séquences observées (principe du maximum de vraisemblance).

De plus en plus d'études phylogénétiques s'appuient non plus sur un seul gène mais sur un ensemble de gènes. L'objectif d'une telle approche est d'essayer de reconstruire au mieux la phylogénie des espèces étudiées et non celle d'un gène particulier. De plus, certains événements évolutifs peuvent avoir eu lieu sans laisser de trace dans les séquences d'un gène donné. En utilisant plusieurs gènes, on procède un peu comme pour une enquête policière où plusieurs témoins sont nécessaires pour avoir une vue complète de la situation et pour pouvoir valider les faits en croisant leurs différents témoignages.

Une phylogénie est traditionnellement représentée par un arbre dont les feuilles sont étiquetées avec les noms des espèces étudiées. Le cas d'une analyse multigénique peut alors être vu comme un problème où l'on dispose en entrée d'un ensemble d'arbres (phylogénies obtenues à partir de chacun des gènes) et où l'on cherche l'arbre qui résume le mieux cette collection (phylogénie des espèces).

Lorsque chaque arbre de la collection contient le même ensemble de feuilles on parle d'arbre consensus. En revanche, si les espèces présentes dans chacun des arbres ne sont que partiellement chevauchantes on parle de super-arbre. Dans le cadre d'une analyse multi-gènes, on est généralement face à un problème de super-arbre. En effet, il est rare que tous les gènes aient été séquencés pour toutes les espèces. Dans le cas d'une approche super-arbre, on peut distinguer deux sous-cas : soit les arbres de départ sont enracinés, soit ils ne le sont pas. La plupart des méthodes existantes supposent que les arbres ne sont pas enracinés, bien qu'ils le soient presque toujours dans les analyses phylogénétiques.

L'objectif de ce stage est d'étudier et d'améliorer une méthode de super-arbres enracinée existante, fondée sur la décomposition des topologies sources en triplets d'espèces, puis sur leur réassociation selon un critère optimisé. Il faudra notamment étudier ses propriétés théoriques (consistance de la méthode), proposer un algorithme ayant une complexité optimale en temps,  $O(n^3)$  avec  $n$  le nombre d'espèce étudiées, et raisonnable en espace.

Le stage sera effectué à l'ISEM (Bâtiment 22 de la faculté des sciences – UMII). Il sera encadré par : Vincent Ranwez ([ranwez@isem.univ-montp2.fr](mailto:ranwez@isem.univ-montp2.fr) tel : 04 67 14 36 97) et co-encadré par Vincent Berry ([Vincent.Berry@lirmm.fr](mailto:Vincent.Berry@lirmm.fr)) et Emmanuel Douzery ([douzery@isem.univ-montp2.fr](mailto:douzery@isem.univ-montp2.fr)).

Une présentation plus détaillée des méthodes de super-arbres est disponible sur le site d'Alexis Criscuolo : <http://www.lirmm.fr/~criscuolo/superarbre.html>