18
NOV.
2024

14h
Salle de conférences de la Bibliothèque Universitaire de la Doua

THÈSE

Soutenance de thèse de Luca Nesterenko

Apprentissage automatique pour l'evolution moléculaire

Composition du jury:

Céline Scornavacca (CNRS Montpellier) - Rapportrice
Tal Pupko (Université Tel Aviv Israël) - Rapporteur
Stéphane Robin (Sorbonne Université Paris) - Rapporteur
Simona Cocco (CNRS Paris) - Examinatrice
Flora Jay (CNRS Gif-sur-Yvette) - Examinatrice
Laurent Guegen (Université Lyon 1) - Examinateur
Bastien Boussau (CNRS Lyon) - Directeur de thèse
Laurent Jacob (CNRS Paris - Co-directeur de thèse

La présentation aura lieu en anglais et durera 45 minutes, suivie par les questions/discussions avec les membres du jury.

La soutenance sera suivie d'un pot dans la salle de pause au 1er étage du bâtiment Mendel.

Mots-clés: Phylogenie,Apprentissage automatique,Bioinformatique

Résumé:

Comprendre l'histoire évolutive d'un groupe d'organismes est une tâche centrale en biologie. En particulier, étant donné un ensemble de séquences codant pour la même protéine chez plusieurs espèces, un objectif important est de reconstruire l'arbre décrivant leur évolution à partir d'un ancêtre commun. Bien qu'étant une étape cruciale dans plusieurs pipelines bioinformatiques, cela représente un problème difficile en soi, car le nombre d'arbres possibles augmente de manière superexponentielle avec le nombre d'espèces.

Les méthodes de pointe reposent sur des modèles probabilistes de l'évolution des séquences et cherchent à maximiser la vraisemblance de l'arbre correspondant. Cette stratégie n'est réalisable qu'avec des modèles très simplifiés. De plus, elle est extrêmement coûteuse en termes de calculs et conduit parfois à des estimations imprécises, notamment dans le cas de mauvaise specification du modèle.

D'un autre côté, les simulations permettent de générer facilement de grandes quantités de données à partir de ces modèles. L'objectif principal de cette thèse a été d'explorer une approche d'apprentissage supervisé pour résoudre ce problème dans un cadre d'inférence basé sur la simulation, sans recours à la vraisemblance. Au lieu de maximiser la vraisemblance d'un modèle d'évolution des séquences, nous avons généré des arbres phylogénétiques ainsi que des séquences ayant évolué selon ces modèles, et les avons utilisés pour apprendre une fonction, paramétrée par un réseau de neurones profond, qui transforme un ensemble de séquences homologues en un ensemble de distances évolutives.

L'arbre lui-même peut ensuite être reconstruit à partir de ces distances via les méthodes dites basées sur la distance. Bien que des progrès notables aient été réalisés ces dernières décennies pour améliorer ces méthodes, l'estimation des distances est encore généralement effectuée dans le cadre du maximum de vraisemblance par de simples comparaisons par paires, ce qui ne permet pas d'exploiter pleinement l'information contenue dans l'alignement multiple des séquences en entrée et qui conduit finalement à des précisions de reconstruction inférieures par rapport à une approche de maximum de vraisemblance complète. Le présent travail vise donc à combler cette lacune en proposant une prédiction conjointe de toutes les distances évolutives, en tirant parti des développements récents et des succès de l'apprentissage profond dans le traitement de données à haute dimension et de séquences.

Nous montrons que ce nouveau paradigme peut améliorer les méthodes de reconstruction phylogénétique existantes ou aboutir à des précisions similaires pour de grands ensembles d'espèces pour lesquelles les méthodes actuelles seraient trop coûteuses en ressources. Cette approche ouvre également la voie à l'adoption de modèles d'évolution plus complexes et réalistes, pour lesquels l'inférence, avec les méthodes basées sur la vraisemblance, serait intractable.

Nous discutons des avantages et de la flexibilité offerts par l'architecture de réseau de neurones développée, qui peut facilement être adaptée pour traiter différentes tâches d'inférence biologique connexes, démontrant ainsi son efficacité dans l'analyse des données de séquences moléculaires.