Projets d'Études et de Recherche 2013-2014

Master 2 Informatique

Classification supervisée de chaînes de caractères : Identification de séquences ADN inconnues

Contexte

Les technologies actuelles peuvent lire l'ADN avec un débit qui permet l'exploration « en vrac » du matériel génétique de tous les organismes vivants contenus dans un échantillon d'eau, de terre, etc. Parmi ces organismes se trouvent de nombreux inconnus, que les biologistes souhaitent attribuer à une grande famille connue.

Les organismes vivants proches les uns des autres ont beaucoup de gènes en commun, donc leurs séquences (modélisées sous forme de chaînes de caractères) seront très semblables. Ainsi, afin de comparer les séquences entre elles, plusieurs mesures de distances ont été développées.

La plupart sont basées sur le comptage de k-mers (toutes les sous-chaînes de taille k qui composent une chaîne), mais nous explorons d'autres pistes, comme la segmentation de chaîne de caractères, semblable à ce qui se fait en imagerie.

Travail demandé

L'objectif de ce projet est d'étudier l'algorithme de classification LMAT à travers l'article dans lequel il est décrit ainsi que son code source, et de modifier son comportement afin qu'il utilise les résultats produits par la segmentation de séquences d'ADN au lieu de la composition en k-mers de ces séquences.

Ressources

- Article décrivant LMAT : http://bioinformatics.oxfordjournals.org/content/early/2013/07/04/bioinformatics.btt389.abstract
- Code source : http://sourceforge.net/projects/lmat/
- Article décrivant la segmentation de séquences d'ADN : http://www.ncbi.nlm.nih.gov/pubmed/10890398
- Il existe une implémentation en C de l'algorithme de segmentation décrit, qui n'est pas disponible en ligne mais qui vous sera envoyée par mail.

Encadrants

Macha Nikolski (macha@labri.fr), Louise-Amélie Schmitt (lschmitt@labri.fr)