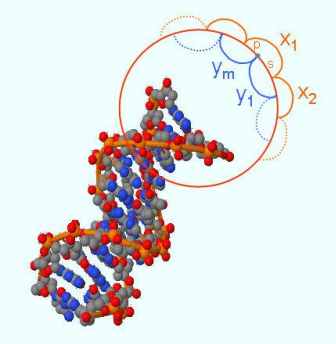

UFR Mathématique Informatique Strasbourg

Professeur Christian MICHEL
c.michel@unistra.fr

Projet de recherche en M1ILC

Big Data en bioinformatique

2016 - 2017

1. Contexte scientifique

Le projet scientifique Big Data en bioinformatique comporte deux objectifs.

Le premier objectif est l'acquisition de gènes et génomes à partir de la base de données génomiques Genbank (http://www.ncbi.nlm.nih.gov/genome/browse/). Il s'agit d'une base de données massive (en juin 2015: 193 milliards de lettres, 185 millions de séquences !) comportant trois sous-bases de données associées aux trois familles d'espèces (eucaryotes, procaryotes, virus).

Le deuxième objectif est le calcul de fréquences de quelques motifs génétiques, par exemple les 64 trinucléotides (motifs de longueur 3 sur un alphabet à 4 lettres), dans les gènes des trois familles d'espèces.

2. Enjeu

La difficulté informatique de ce projet réside dans la programmation web (Java) et la gestion de fichiers volumineux de données.

3. Méthodes informatiques

L'analyse informatique de votre projet portera sur divers axes:

- étude de la base de données génomiques Genbank: structuration des fichiers de génomes, balises web, adresses des génomes (Bioproject), etc.
- traitement de données massives sur PC (avec une mémoire centrale limitée);
- choix des structures de données, en particulier l'arborescence locale des fichiers;
- traitement des données non conformes (documentation: http://www.ncbi.nlm.nih.gov/books/NBK25501/);
- gestion des délais et/ou des coupures de connexion au site web (temporisateur);
- calcul des fréquences des motifs à la volée (lors de l'acquisition des gènes);
- écriture des résultats statistiques sur des fichiers Excel;
- possibilité de sauvegarder les génomes ou gènes dans des fichiers de type texte.

Un document complet donnant des informations détaillées est disponible (merci de me contacter par email).