. Titre : **Reconstruire notre passé : apprentissage statistique (deep learning) pour la génétique des populations**

. Thématique : Génétique et apprentissage statistique (machine learning)

. Laboratoire, institution et université :
Laboratoire de Recherche en Informatique (LRI), CNRS/INRIA/Université Paris Saclay

. Ville et pays : Gif-sur-Yvette, France

. Équipe ou projet dans le labo : BioInfo / TAO (Apprentissage et Optimisation)

. Nom et adresse électronique du directeur de stage
Flora Jay <flora.jay@lri.fr>
Guillaume Charpiat <guillaume.charpiat@inria.fr>

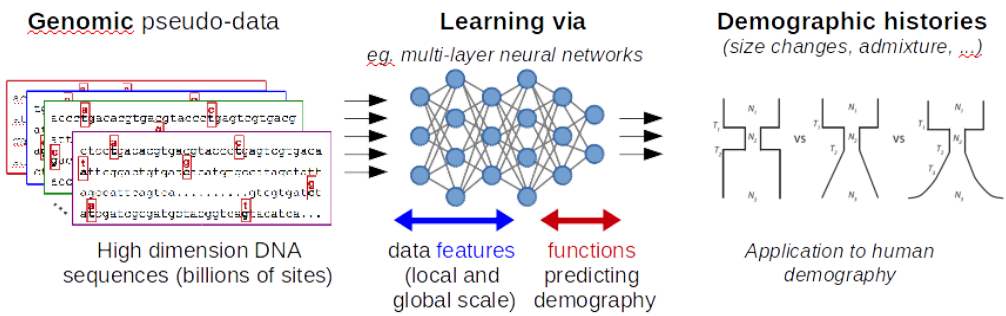
. Nom et adresse électronique du directeur du laboratoire
Yannis Manoussakis <yannis.manoussakis@lri.fr>

. Présentation générale du domaine

Avec la récente explosion des technologies de séquençage, de plus en plus de données génétiques sont disponibles, ce qui ouvre la porte à une connaissance approfondie de l'histoire évolutive et démographique des populations. Cette histoire peut en effet être reconstruite sur une échelle de plusieurs milliers d'années : la comparaison de l'ADN d'individus vivant actuellement permet d'identifier les mutations génétiques qui les différencient, d'estimer les fréquences de ces mutations, leur corrélation à différentes échelles du génome, et en conséquence de comprendre les relations liant ces individus et les caractéristiques des populations auxquelles ils appartiennent, comme le temps de séparation de deux populations ou leurs tailles (en terme de nombre d'individus) à différentes époques. Toutefois extraire l'information intéressante des données génomiques de manière efficace reste un problème ouvert.

. Objectifs du stage

Il s'agit de développer une nouvelle méthode d'inférence démographique à partir de données génomiques, c'est-à-dire une méthode permettant l'estimation de multiples paramètres d'importance biologique, comme les tailles de population, le taux de croissance et la date d'une expansion, le taux de métissage entre populations, etc.
Pour se faire plusieurs étapes seront réalisées :



- Proposer et implémenter une représentation des données génétiques satisfaisante en terme de complexité de calcul, de stockage et de perte d'information.
- Étudier l'applicabilité de différentes techniques d'apprentissage profond actuellement utilisées dans les domaines du langage et de l'image (ex: réseaux de neurones récurrents, réseaux convolutifs, ...) pour construire une méthode d'inférence démographique.
- Appliquer la méthode à des données génomiques de populations humaines et comparer aux connaissances actuelles en anthropologie génétique.

. Références bibliographiques

Sheehan S, Song YS. Deep learning for population genetic inference. PLoS Comput Biol. 2016 12(3):e1004845.

Boitard S, Rodriguez W, Jay F, et al. Inferring population size history from large samples of genome-wide molecular data-an approximate Bayesian computation approach. PLoS Genet. 2016 12(3):e1005877.

Durbin R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). Bioinformatics. 2014 30(9):1266-72.

Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence. 2013 35(8):1798-828.

. Compétences espérées

- Maîtrise d'un langage de programmation (ex: Python)
- Connaissances en apprentissage statistique

Non requises mais avantageuses :

- Expérience/connaissance des réseaux de neurones et l'apprentissage profond
- Connaissances en génétique, intérêt pour la génétique des populations