Architectures souples pour le deep learning avec application en génétique des populations

. Thématique: apprentissage statistique (machine learning), génétique

. 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: TAO (Apprentissage et Optimisation) / BioInfo

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

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

. Présentation générale du domaine:

Ces dernières années, les réseaux de neurones ont connu un succès foudroyant, remportant de nombreux challenges en apprentissage statistique, et résolvant de nombreuses tâches d'apprentissage que l'on pensait jusqu'alors difficiles. Ce succès est dû en partie à l'augmentation des moyens de calculs (ce qui permet une meilleure optimisation des paramètres des neurones), mais aussi au choix de meilleures architectures, plus faciles à entraîner, ou plus adaptées au problème à résoudre. Par exemple, les réseaux convolutionnels sont adaptés au traitement des images et du texte, exploitant des invariances désirables (ex: invariance par translation de l'image) pour réduire significativement le nombre de paramètres à estimer et simplifiant ainsi le problème.

Pour chaque tâche à résoudre, on choisit donc d'abord une architecture de réseau de neurones adaptée (c'est-à-dire que l'on fixe le nombre de couches de neurones, le nombre de neurones par couche, le type des neurones, etc.) et ensuite on cherche à estimer les meilleurs paramètres possibles des neurones (les poids de connexion entre deux neurones) afin de mener la tâche à bien. Seulement, ceci suppose que toutes les données traitées ont **exactement la même taille** (chaque échantillon devant contenir autant d'éléments qu'il y a d'entrées au réseau de neurones). La question que l'on se pose ici est donc de savoir comment faire lorsque le format des données peut varier: comment rendre l'**architecture souple pour s'adapter aux données**? Comment apprendre à traiter à la fois des images 100x100 et des images 99x101, comment être capable de **généraliser la fonction** apprise à des **données de tailles différentes**?

. Objectifs du stage

Pour ce faire, plusieurs approches sont envisageables :

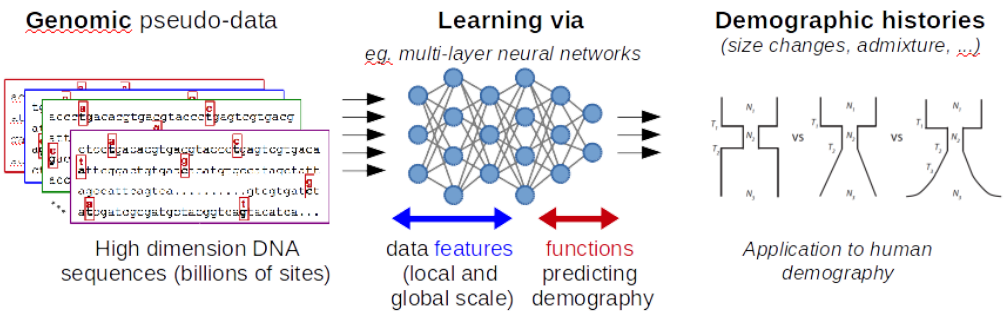
- **Méta-apprentissage:** Entraîner un méta-réseau de neurones, qui prend en entrée la taille du prochain échantillon et génère le réseau de neurones de la bonne taille qui lui sera appliqué.

Cela revient donc à apprendre conjointement une famille de réseaux de neurones (un réseau par taille), de façon générative.

- **Étude des invariants:** Décrire mathématiquement les propriétés d'invariance souhaitées de la fonction à apprendre (dans l'exemple ci-dessous: invariance du résultat sous permutation des brins d'ADN, invariance du traitement des bases d'ADN par translation le long des brins…) et en déduire la forme que doit prendre la fonction à apprendre pour ne pas dépendre de la taille des données en entrée.
- **Réseau de fonctions:** Choisir tout d'abord des familles de fonctions faciles à étendre à des tailles d'entrées variées (par exemple: la somme, la moyenne, le max, l'écart-type…), puis construire un réseau dont certains neurones sont de ce type et savent ainsi s'adapter à la taille de la couche précédente.

Le stage consistera à développer une ou plusieurs de ces approches.

L'application visée est en génétique des populations : comment, à partir de brins d'ADN de quelques individus actuels, **reconstruire l'évolution passée de la population**. En effet, la comparaison de brins d'ADN permet de détecter des mutations, et des statistiques sur celles-ci permettent d'inférer des facteurs comme l'ancienneté de ces mutations, la taille de la population à l'époque, son taux de métissage, etc. On souhaite donc apprendre, à l'aide d'un réseau de neurones et d'un ensemble d'apprentissage, la fonction qui associe à des brins d'ADN l'historique de la population. L'une des difficultés est que les brins sont en nombre variable et de longueur variable.



. Références bibliographiques:

- 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.
- 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.
- Sheehan S, Song YS. Deep learning for population genetic inference. *PLoS Comput Biol*. 2016 12(3):e1004845.
- Stanley, Kenneth O., David B. D'Ambrosio, and Jason Gauci. A hypercube-based encoding for evolving large-scale neural networks. *Artificial life* 15.2 (2009): 185-212.

. Compétences espérées:

- Mathématiques variées (statistiques, algèbre linéaire, permutations, analyse fonctionnelle…)
- Connaissances en apprentissage statistique
- Maîtrise d'un langage de programmation (ex: Python)

Non requises mais avantageuses:

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