PhD projects on Deep Learning for Population Genetics

Supervisors: Flora Jay (CR CNRS), Guillaume Charpiat (CR INRIA),
Direct collaborators: Burak Yelmen (postdoc U Paris-Saclay), Cyril Furtlehner (CR INRIA), Aurélien Decelle (Complutense University of Madrid)
Contacts: flora.jay@lri.fr **; guillaume.charpiat@inria.fr**
web: https://flora-jay.blogspot.com/ ; https://www.lri.fr/~gcharpia/

Location: LISN (Paris-Saclay University)
machine learning and bioinformatics groups
1, rue Raimond Castaing
91190 Gif-sur-Yvette, France (Paris suburban area)

When/Duration: September/Oct 2022 (or even earlier), 3 years.

We are looking for **one highly motivated candidate to do a PhD** in our lab and suggest two potential projects. They are follow-up research of two of the lab papers:
- "Creating artificial human genomes using generative neural networks", Yelmen et al 2021
- “Deep learning for population size history inference: design, comparison and combination with approximate Bayesian computation”, Sanchez et al 2020
Please note that we would happily talk with candidates that have an alternative project in mind, as long as it falls in the scope of deep learning for population genetics.

In their motivation letter, the candidates should explain which of these topics they are interested in (it could be both).

We can provide further details about the projects upon request.

Keywords
deep learning, population genetics, generative models, interpretability, inference, evolution
apprentissage profond, génétique des populations, modèles génératifs, interprétabilité, inférence, évolution

Requirements
The ideal candidate should be good at python scripting and machine learning/statistics concepts. Experience with deep learning is a plus. Familiarity with some of the following topics: genomics, population genetics, generative models, bash scripting and high-performance computing is not mandatory but a clear plus. Being curious and autonomous is highly recommended for any PhD. Being able to communicate, read and write in **English** (French not required).

Salary: regular PhD stipend in academia. The PhD fellowship is fully funded (ANR grant) for 3 years.

Application: Ideally the candidate **should provide** a CV, motivation letter, past scores/ranks, names and emails of previous mentors that can be contacted. Send your application to: flora.jay@lri.fr ; guillaume.charpiat@inria.fr Deadline: a.s.a.p., preferably before May 26th, however we will keep reading applications received after this date until the position is filled.

Lab environment

The PhD candidate will interact with two teams with the LISN, a machine learning team and a bioinformatic team. They host many permanent researchers, postdocs and PhD students with whom to interact. There are weekly meetings on population genetics (informal working groups), bioinformatics and machine learning topics.

Projects

* **Subject 1: Creating artificial human genomes using generative neural networks**

Context : Using machine learning, we could generate synthetic genomes that successfully mimic the real ones but are not identical to any of them. We relied on two type of neural network architectures that we trained on human genetic databases: (1) Generative Adversarial Networks, that were a breakthrough in the domain of computer vision, allowing the generation of extremely realistic images; (2) Restricted Boltzmann Machines, another family of generative models capable of learning complex data distributions. We measured the quality of the generated genomes in terms of data hidden structure, population structure, linkage disequilibrium, haplotype diversity, etc. and demonstrated that they provided an accurate representation of the real ones. Without duplicating any of the individuals, most key characteristics of the data were conserved. Additionally, we showed that releasing these genomes would lead to a privacy gain. A direct implication is the increase in richness of public datasets, e.g. with populations still under-represented in genetic studies.

PhD tasks

The candidate will work in close collaboration with a postdoctoral researcher. The multidisciplinary project combines genomics, population genetics and machine learning with several focal points:

1) **Improve the proposed generative models; design, implement and train new neural networks for large scale genomic data.** In particular, the original architectures need to be scaled up to process very large genomic data (~million base pairs). The candidate will propose and implement enhancements of the currently used GAN, VAE and/or RBM networks.

2) **Extend the applicability of generative models/generated genomes for various other genomic tasks** such as imputation and other tasks applied daily by the population genetics communities.

3) **Control privacy leakage** (information leaked about real individuals that contributed DNA to the original training dataset).

* **Subject 2: Inferring the evolutionary past of populations from genomic data and interpreting neural networks.**

Context. Our lab and others have recently developed inference methods based on deep learning to infer evolutionary models directly from genetic data (e.g. for reconstructing demographic history, Sanchez et al 2020). We have also developed a generic software, dnadna, for implementing and applying neural networks in population genetics (Sanchez, Bray et al). Despite their good performances, neural networks are often (understandably) criticized for their lack of interpretability. Interpretability is not only crucial for explaining a prediction, but also for avoiding artifact and biases. As for neural networks, theoretical and applied research in this direction is moving fast. Many methods address the question of what information the network uses globally to construct a model, or what information contributed to the output for a particular example. Yet very few have been applied to DNN based on genetic data for population genetic inference.
On the other hand, interpretability has been investigated in a more traditional setup, where SNP data are reduced into handcrafted expert statistics. In some ML settings one can easily evaluate the relative contribution of each statistic to the prediction which gives a clue on the information used by the method.

PhD tasks

1) **Adapt and test existing interpretability methods used in computer vision or genetics** on DNNs solving evolutionary inference tasks (such as inferring the adaptation or the past demography of a population from a genomic sample).

2) **Develop a novel interpretability approach for DNNs in population genetics** based on dimensionality reduction techniques, knowledge-based features and neural activations. Exploit the results to provide meaningful feedback to population geneticists.

3) **Compare the interpretability of architectures** previously proposed by the population genetics community and **design novel architectures** that are more easily interpretable (for solving evolutionary inference tasks)

4) **Incorporate interpretability into our dnadna python package** (a framework facilitating the use/sharing/reproducibility of DL approaches in population genetics).

* **Subject 3: Your project**
Feel free to contact us if you have a strong opinion on the PhD project that you would like to pursue, as long as it remains in the scope of machine learning and population genetics!

*** References**

B Yelmen, A Decelle, L Ongaro, D Marnetto, C Tallec, F Montinaro, C Furtlehner, L Pagani, F Jay.

Creating Artificial Human Genomes Using Generative Models. PLoS genetics, 17(2), e1009303. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009303

T Sanchez, J Cury, G Charpiat, F Jay (2020). Deep learning for population size history inference: design, comparison and combination with approximate Bayesian computation. Molecular Ecology Ressources DOI:10.1111/1755-0998.13224
Link:
https://www.researchgate.net/profile/Flora-Jay/publication/342822922_Deep_learning_for_population_size_history_inference_Design_comparison_and_combination_with_Approximate_Bayesian_Computation/links/5f353ad192851cd302f1829c/Deep-learning-for-population-size-history-inference-Design-comparison-and-combination-with-Approximate-Bayesian-Computation.pdf

T Sanchez, EM Bray*, P Jobic, J Guez, G Charpiat, J Cury°, F Jay° (2021) Dnadna: Deep Neural Architecture for DNA - A deep learning framework for population genetic inference*
https://hal.archives-ouvertes.fr/hal-03352910

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences." International Conference on Machine Learning. PMLR, 2017.