PROGRAMME INTITUTS ET INITIATIVES

Appel à projet – campagne 2021

Proposition de projet de recherche doctoral (PRD)

SCAI - Sorbonne Center of Artificial Intelligence

Intitulé du projet de recherche doctoral (PRD):
Generation of genomic DNA sequences with GANs

Directeur.rice de thèse porteur.euse du projet (titulaire d'une HDR) :

NOM : **MOZZICONACCI** Prénom : **Julien**
Titre : Professeur des Universités ou
e-mail : julien.mozziconacci@mnhn.fr
Adresse professionnelle : CP26, Rue Cuvier, MNHN, 75005 Paris
(site, adresse, bât., bureau)

Unité de Recherche :
Intitulé : Structure et instabilité des génomes
Code *(ex. UMR xxxx)* : UMR 7196

École Doctorale de rattachement de l'équipe (future école doctorale du.de la doctorant.e) : **ED227-Sciences vie homme : évolution écolog**

Doctorant.e.s actuellement encadré.e.s par la.e directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) : 1 doctorant depuis 2018 (3e année)

Co-encadrant.e :

NOM : **BOULE** Prénom : **Jean Baptiste**
Titre : Chargé de Recherche ou HDR ☒
e-mail : jbboule@mnhn.fr

Unité de Recherche :
Intitulé : Structure et instabilité des génomes
Code *(ex. UMR xxxx)* : UMR 7196

École Doctorale de rattachement : **ED227-Sciences vie homme : évolution écologie**
Ou si ED non Alliance SU :

Doctorant.e.s actuellement encadré.e.s par la.e co-directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) :

Co-encadrant.e :

NOM : Prénom :
Titre : Choisissez un élément : ou HDR ☐
e-mail :

Unité de Recherche :
Intitulé :
Code *(ex. UMR xxxx)* :

École Doctorale de rattachement : **Choisissez un élément :**
Ou si ED non Alliance SU :

Doctorant.e.s actuellement encadré.e.s par la.e co-directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) :

Cotutelle internationale : ☐ Non ☐ Oui, précisez Pays et Université :

Selon vous, ce projet est-il susceptible d'intéresser une autre Initiative ou un autre Institut ?
☐ Non ☐ Oui, précisez Choisissez l'institut ou l'initiative :

SORBONNE UNIVERSITÉ

Description du projet de recherche doctoral *(en français ou en anglais) :*

Ce texte sera diffusé en ligne : il ne doit pas excéder 3 pages et est écrit en interligne simple.

Détailler le contexte, l'objectif scientifique, la justification de l'approche scientifique ainsi que l'adéquation à l'initiative/l'Institut.

Le cas échéant, préciser le rôle de chaque encadrant ainsi que les compétences scientifiques apportées. Indiquer les publications/productions des encadrants en lien avec le projet.
Préciser le profil d'étudiant(e) recherché.

Generation of genomic DNA sequences with GANs

1) Scientific Context

The improvement of DNA sequencing techniques lead to an explosion in the number and completeness of fully sequenced genomes (i.e. stretched of billions of A/C/T/G letters). However the mere genome sequence of an organism is far too little to understand how this sequence is interpreted by the molecular machinery within the cell. In eukaryotes, organisms in which the genome is sequestered inside a micron size nucleus, the long DNA molecule (sometimes up to meters) that constitutes the chromosomes is wrapped around protein complexes, called nucleosomes, to form an array like beads on a string. Nucleosomes have two roles. The first role is structural: 147 base pairs of DNA are wrapped within a nucleosome and this wrapping softens the DNA molecule, reducing its persistence length by a factor of ten and allowing these long DNA molecules to be compacted in small volumes. The second role is functional: nucleosomes occupy specific positions on the genome sequence and they can locally modify the binding affinity of DNA for transcription factors, the proteins that read the DNA motifs and regulate the expression of the genes. In cells expressing the same genes, nucleosomes occupy the same positions [1] and an important question in the field is to understand the mapping that relates the DNA sequence and the position of nucleosomes along this sequence. This function cannot be explicitly described using a mathematical formula but we have recently shown that a deep neural network can be used to model it [2].

As Early as 2015, pioneering studies [3, 4] demonstrated the efficiency of deep neural networks to learn protein binding site, or more generally any type of annotation, directly from DNA sequencing experiments. The application of deep neural networks to genomics has been since growing at a high pace and it can now be considered as a state of the art computational approach to learn and predict genomic annotations at the genome wide level [5, 6, 7, 8, 9]. Deep Convolution Neural Networks (CNN) are in particular very efficient at detecting sequence features since they rely on the optimisation of convolution filters that can be directly matched to DNA motifs [10]. Stacking several of these convolutional layers together can lead to the detection of nested motifs at larger scales. A game changing advantage of these tools is their ability to predict a learned annotation on a variation of the genome, i.e. to predict the effect of mutations. As a first proof of concept of the ability to predict mutations on large scales, we developed the mutasome approach, in which all bp of a single genome are mutated individually to see the effect on a given genome annotation. In our case we successfully predicted the effect of all the possible mutations on the nucleosome positions on the yeast Saccharomyces cerevisiae genome [2]. We then used used this information to identify the

sequence motifs that were responsible for the precise positioning of nucleosomes.

2) Scientific objectives and approach

The first part of this PhD project is to extend this study using a larger variety of species in order to understand the evolution of nucleosome positioning rules. While the yeast has been under detailed scrutinising in the past, few nucleosome positioning data-sets exist for other species. In order to increase this number we will rely on the Museum's RBCell collections and will focus our analysis on unicellular eukaryotes and fungal strains. The collections of the RBCell ensemble represent a strong heritage interest for the conservation of the diversity of living things. These collections come from samples taken from the environment and stored in a functional form (e.g. living or cryopreserved cells). They are the support for research activities on the cellular mechanisms of living organisms. Experiments that will reveal the nucleosome positions on the genome of these organisms will be carried in the group of our collaborator Jean Baptiste Boule (JBB) at MNHN. This unique opportunity to bring together Julien Mozziconacci's (JM) expertise on deep genomics, JBB'expertise on yeast genetics and the use of the MNHN collection resources will empower us to decipher the rules of nucleosome positioning in many diverse unicellular organisms.

The second part of the PhD project will go beyond the mere prediction of the effect of mutations on nucleosome positions. The rational behind this extension of the project is that if one can successfully predict the effect of mutations, it becomes also possible to design new sequences with controlled properties, a field now known as genome writing [11]. Genome writing is in its early years and our group is involved in the international GP-write consortium which aims at designing synthetic pieces of genomes that will be synthesised and tested experimentally. In the present project, we wish to leverage the use of Generative Adversarial Networks (GANs) for genome writing. Our goal is to design short (around 200 bp) sequences that will position nucleosomes precisely in their centre in vivo. A typical GAN model is composed of the input random vector, the generator, and the discriminator. The generator and discriminator are implicit function expressions, implemented by deep neural networks. In short, the generator used will be a network that generates a DNA sequence of the given length from random variables. The generator will use two cost functions. The first one will be computed the predicted nucleosome density on the sequence and will be minimum when this density will display at peak at the sequence centre. The second cost function will be given by the discriminator network. The discriminator will try to discriminate these generated sequences from real yeast DNA sequences of the same length. The parameters of both networks will be optimised alternatively while keeping the other network fixed. We will thus train the generator to generate yeast-like sequences that will position one nucleosome in their centre. This strategy has been recently proposed on theoretical grounds [12] and has been applied to generate artificial variations of human splice sites [13]. It requires the use of the Wasserstein algorithm [14] in order to efficiently sample the sequence space. We plan to leverage this strategy in order to generate sequences (of varying length from 167 to 237 bp) that will position one nucleosomes using a network trained on yeast Saccharomyces cerevisiae nucleosome position data. These sequences will be then assembled as tandem arrays of hundreds of repeats to form regular nucleosomal array with different spacing between nucleosomes. We thus expect to form regular arrays of nucleosome with various spacing between them in vivo. The efficiency of the design, will be experimentally tested in collaboration with JBB at MNHN who will also co-supervise this thesis.

3) Condidate profile

The recruited PhD student will be involved both in training the neural networks on multi species nucleosome data (first part) and training the GANs to generate synthetic DNA sequences (second part) as well as analysing the experimental results obtained in the JBB team. The project lies at the frontiers between synthetic biology and computer science. It will require strong computational and theoretical skills. We will therefore seek for a candidate with background either in computer science or bioinformatics that will acquire the relevant knowledge in deep learning and biology in our two

groups. This project will be one of the very first world-wild that uses deep learning techniques to design in silico an entirely new DNA sequence with desired properties and to test these properties in living organisms.

References (**** publications from the PhD supervisor)

[1] Alexander M Tsankov, Dawn Anne Thompson, Amanda Socha, Aviv Regev, and Oliver J Rando. The role of nucleosome positioning in the evolution of gene regulation. PLoS biology, 8(7):e1000414, 2010.

[2] **** Etienne Routhier, Edgard Pierre, Ghazaleh Khodabandelou, and Julien Mozziconacci. Genome-wide prediction of dna mutation effect on nucleosome positions for yeast synthetic genomics. Genome Research, pages gr–264416, 2020.

[3] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. Nature biotechnology, 33(8):831, 2015.

[4] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. Nature methods, 12(10):931, 2015.

[5] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome research, 26(7):990– 999, 2016.

[6] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. Nature genetics, 51(1):12–18, 2019.

[7] William Jones, Kaur Alasoo, Dmytro Fishman, and Leopold Parts. Computational biology: deep learning. Emerging Topics in Life Sciences, 1(3):257–274, 2017.

[8] **** Ghazaleh Khodabandelou, Etienne Routhier, and Julien Mozziconacci. Genome annotation across species using deep convolutional neural networks. PeerJ Computer Science, 6:e278, 2020.

[9] **** Etienne Routhier, Ayman Bin Kamruddin, and Julien Mozziconacci. keras dna: a wrapper for fast implementation of deep learning models in genomics. Bioinformatics, 2020.

[10] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface, 15(141):20170387, 2018.

[11] Nili Ostrov, Jacob Beal, Tom Ellis, D Benjamin Gordon, Bogumil J Karas, Henry H Lee, Scott C Lenaghan, Jeffery A Schloss, Giovanni Stracquadanio, Axel Trefzer, et al. Technological challenges and milestones for writing genomes. Science, 366(6463):310–312, 2019.

[12] Nathan Killoran, Leo J Lee, Andrew Delong, David Duvenaud, and Brendan J Frey. Generating and designing dna with deep generative models. arXiv preprint arXiv:1712.06148, 2017.

[13] Nicholas Bogard, Johannes Linder, Alexander B Rosenberg, and Georg Seelig. A deep neural network for predicting and engineering alternative polyadenylation. Cell, 2019.

[14] Martin Arjovsky, Soumith Chintala, and L´eon Bottou. Wasserstein GAN. arXiv preprint arXiv:1701.07875, 2017.

Merci d'enregistrer votre fichier au format PDF et de le nommer :
«ACRONYME de l'initiative/institut – AAP 2021 – NOM Porteur.euse Projet »

Fichier envoyer simultanément par e-mail à l'ED de rattachement et au programme :
cd_instituts_et_initiatives@listes.upmc.fr avant le 20 février.