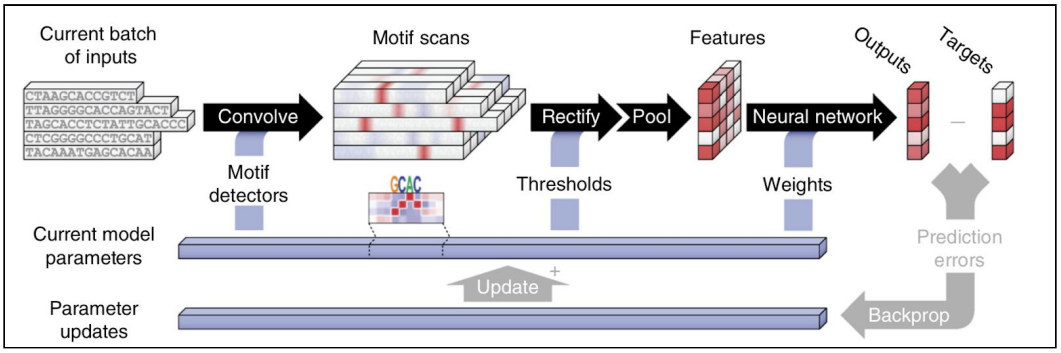
Jérémie KALFON, jkobject.com, jkobject@gmail.com,

Academics

I am a Biomedical Engineer in Computer Science & Electronics *[1](résumé)* from France, thanks to my sundry training, I have gained consistent knowledge about many different topics *[2](courses_list)*. However, French engineering schools do not focus on in-depth training in any precise topic and it is -as I have explained in my *[3](coverletter)*- why I have chosen a second master at the University of Kent in Computational Intelligence (aka Machine Learning).
My Master Project has recently been defined as a study, in collaboration with Tobias Von Der Haar, Dominique Chu and Yun Deng about the data mining of codon usage statistics (entropy values) in fungus using unsupervised learning techniques to try to decipher the problem of "codon usage bias" which shows that, sets encoding the same amino acid are not present in equal proportion in the genome (closely related to the GC bias problem). This task which should result into a joint publication seems to represent a great introduction to my research objectives.

I want to pursue my research career in this direction, focusing on data science and statistical learning techniques. More specifically, what has been capturing my attention, this last couple of years, is advanced Machine Learning, artificial neural networks and their constantly increasing diversification, from the latest RNNs architectures *[4](Graves et al.)* to recently created GANs *[5](Goodfellow et al.)* and VAEs, etc. Their applications of which the only limit currently seems to be our imagination, keep increasing in the domains of science where the amount of data requires such powerful tools. Their recent utilisation on the transcriptome *[6](Alipanahi et al.)*, to predict DNA- and RNA-binding protein, to quantify the function of DNA sequences *[7](Quang et al.)* ,or even to generate DNA sequences with given properties *[8](Killoran et al.)* shows an emerging landscape of research with direct translational values.



Architecture of a neural network to extract motifs/information from DNA samples.
This allows to predict potential protein binding locations. Work by Babak Alipanahi
Nature biotechnology All rights reserved.

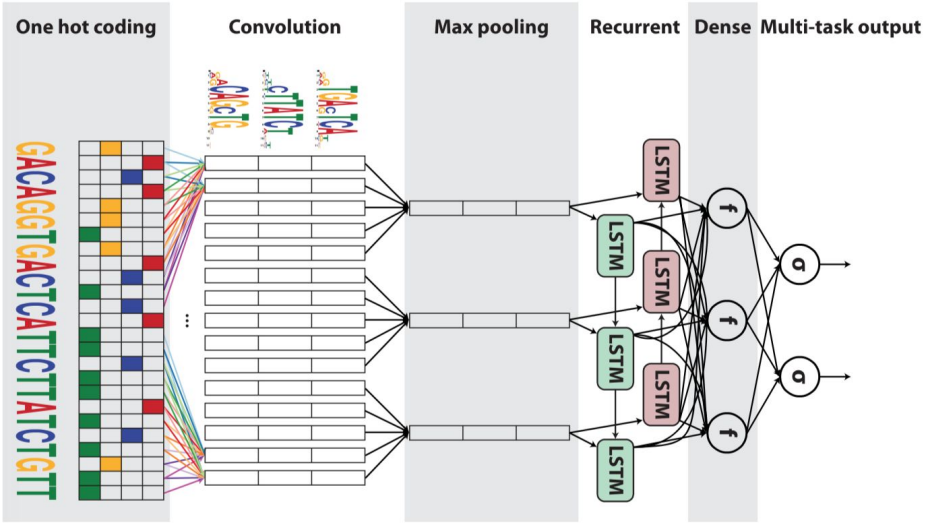
However in addition to raw performance, a wide range of challenges needs to be overcome. For example, as black box algorithms, they lack interpretability. An ability to give meaningful answers that researchers and doctors can use is a requirement before any widespread usage. Bayesian Networks, Neural Coders and many other approaches *[9](Distill, Interpretability)* ought to be tried in the biomedical context.

To the eyes of an engineer, the regulatory network of cells presents itself as the biological program that it is running. Decoding and understanding it, means to speak the cellular language, and enables us to fully interact with it. Research on the topic of network inference and embeddings[15,16](*Bo Wang et. al.*, *Niall M. Mangan et. al.*), and on the gathering of multiple type of NGS analysis[17](*Ricar Arguelaguet et. al.*) are big steps in this direction. The translational power of this research to the biomedical industry is important and game changing. The data is typically noisy, patterned, interpreted as sequences or networks and often, time varying with a lot of correlations across multiple modalities. This makes it, to any data scientist, one of the best and most challenging "playing" field. It is demonstrated by the many research papers relating ML and biomedicine.

Objectives

As my interest, my research objective is interdisciplinary, the potency of which lies in uncountable examples throughout history. The incessant crossings between neural networks and neuroscience *[10](Hassabis et al.)* for example, is to my mind a key example of how disruptive interdisciplinary research can be. I have repetitively witnessed it, from my involvement in CaImAn[1] *[11](Giovannucci et al. -preprint-)*, my participation to HBP workshops *[12](about it)* and my Master thesis *[13](github PyCUB)*.

I aim at understanding the cellular language by teaching algorithms to do so, using data from multiple sources. It may be seen as trying to bridge the gap between genotype and phenotype. To drive this research and objectives, I am considering the current work of the teams of Pr Frey, Serafim, Bonneau, Zhang, Berejano, Vert, Azencott, Segal, Rubin, Carpenter, Hammer, Engelhardt, Kundaje, Gerstung, Petsalaki, Deplancke, Saria… and the teams from uncountable companies and startup across the world (Deep Genomics, Ayasdi, Sancare, Verge Bioscience, IBM, SenS, Calico, AtomWise, Verily, Cofactor, Freenome, Clear Data, Clarify, etc…) which I found very inspiring.



Architecture of a Convolutional + Recurrent neural network to extract Regulatory features from ncDNA sequences. This allows them to predict non-coding function de novo. Work by Daniel Quang
Nucleic Acids Research. All rights reserved.

[1] It uses gaussian kernels and a constrained matrix factorization (using Autoregressive processes, Constrained deconvolution with sparsity control, noise estimation and overlaps management) to divide the input image into a spatial location and a temporal activity of the neurons from a deconvolution of their calcium dynamics.

Today a great deal of work in machine learning and computational biology is focused on inferring protein structure/folding and function from simple information on the linear chain of amino acids that is given by the translation of RNA sequences. Another is on drug discovery to also infer molecules functions and effects. Understanding more about core features of DNA sequences will drive the research these fields as well.

Systems biology has always been at the forefront of data sciences. According to me a boom is still waiting to happen after which we will see the pervasiveness of the field to computer scientists and an increase in the open sourcing of data *[14](NAR's list...)*. All in all, It shows a great direction to take for future researches.

Plan

- A part of any Ph.D. project is learning, not only about or from the literature, but about or from other researchers and their work (create a network). Being a good researcher also means to write and present research to the many and to organize and manage oneself to be the most efficient. This is why it is part of my objectives.
    - Creating an efficient work schedule and pipeline.
    - Reading research on the topic of my PhD project with a drive to select and focuses on what is important.
    - Getting in the first year, up to speed with any important topic that I would have overlooked during my studies.
    - Going to many conferences.
    - Visiting other labs.
    - Writing as much as possible (Papers, Articles..).
- The Project would belong, but not be restrained to, what I have displayed in my objectives. Whatever the precise title may be, it would have to involve advanced computational learning techniques and help the understanding of fundamental cellular mechanisms whilst retaining a strong translational potential. I can thus draw a first plan of my PhD project:
    - Following what my engineer background taught me, I will have a real "Systems" approach to the cellular mechanism I am studying, with its grey areas and questions together with their respective theories, assumptions and the type of data required and available.
    - Then find and assess focus points in this map according to the extensive data mining tool in both the data science community and the -more specialized ones of the- Bioinformatics community and get to know the datasets and best computational tools (e.g. BioPython, scikit learn, tensorflow) to work with.
    - The process of developing the project is an iterative one requiring feedbacks from my mentors, other PhD students, similar research (conferences, papers) and possibly other players (doctors, patients, experimentalist, companies)
    - There are many parallels made between the recent gene feature extraction models and language processing methods. It is very likely that other similar parallels and ideas can be drown from other data science fields.

- The next part is to present, according to the first one, new tools and packages for the research community and the biomedical industry to build upon.
    - An incremental approach will help having a worthy project rapidly in time and to adapt new decisions and opportunities to change.
    - I have been inspired by some great machine learning and bioinformatics open source projects such as Scikit-learn, CaImAn, GATK, deepVariant, which are firstly the result of one or two researchers and have expanded to encompass many uses and users (and thus contributors) because of their versatility, modularity and simple, carefully crafted documentation.
    - Finally I hope to display the ability of the softwares by always answering some biological questions with them and see if they are achieving their objectives.

Addition :

1. This plan requires for me to be able to surround myself with different researchers with various knowledge and backgrounds. An interdisciplinary project requires good collaboration (project management) to improve the exchange of information and the success of the endeavour.

2. I wish to produce results that are as consistent, reproducible, usable (open sourcing of the code, open sourcing of the data, API, documentations) and scalable to big datasets as to be verified, tested, used, modified and improved by other groups.

3. While trying to be as precise as possible, this is a first research plan from a pre-Doc. It mostly tries to demonstrate an ability to plan, to understand some valuable concepts and a knowledge of the discipline.

References:

[1] Résumé available on GDrive at https://drive.google.com/file/d/1Pv7sNHQ5is0PZ-DxzDDcbC9QNa6UmJlm

[2] List of my courses available on my website at http://jkobject.com/papers/courses.html

[3] Motivation letter for a Ph.D. available on GDrive at https://drive.google.com/open?id=1enkEXh-rqRfz2b5XggaAWnyk3f_-mLKZ

[4]"Neural Turing Machines: Convergence of Copy Tasks" , arXiv:1612.02336v1

[5]"Generative Adversarial Nets", arXiv:1406.2661v1

[6] "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning" Nature Biotechnology volume 33, pages 831–838 (2015)

[7] "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences." doi: 10.1093/nar/gkw226.

[8]"Generating and designing DNA with deep generative models", arXiv:1712.06148v1

[9] "The building blocks of interpretability", doi:10.23915/distill.00010

[10] "Neuroscience-Inspired Artificial Intelligence", DOI:10.1016/j.neuron.2017.06.01

[11] "CaImAn: An open source tool for scalable Calcium Imaging data Analysis", doi: 10.1101/339564

[12] More information available on my website at http://jkobject.com/project/hbp.html

[13] Project available here https://github.com/jkobject/PyCUB, paper to come in the next few months. Master thesis: "PyCUB: a Machine exploration of the Codon Usage Bias", doi:10.13140/RG.2.2.35259.26400

[14] "The 2018 Nucleic Acids Research database issue and the online molecular biology database collection", doi:10.1093/nar/gkx1235

[15] "Vicus: Exploiting local structures to improve network-based analysis of biological data", doi:0.1371/journal.pcbi.1005621

[16] "Inferring Biological Networks by Sparse Identification of Nonlinear Dynamics", doi:10.1109/TMBMC.2016.2633265

[17] "Multi-Omics factor analysis - a framework for unsupervised integration of multi-omic data sets" , doi: 10.1101/217554