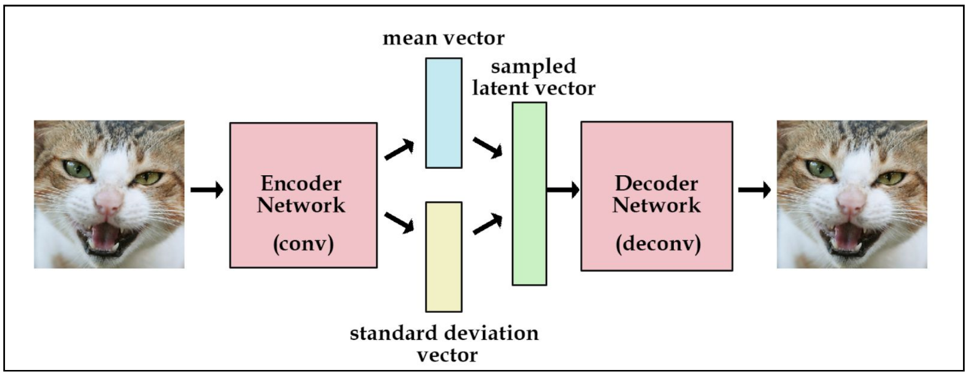
Jérémie KALFON, Deplancke Lab

Academics

I am a Biomedical Engineer in Computer Science & Electronics *(resume)*, thanks to my sundry training, I have gained a knowledge about many different topics *(courses_list)*. However, French engineering schools do not focus on academic tasks and it is -as I have explained it in my *(motivation letter)*- why I have chosen a second master at the University of Kent in Computational Intelligence.
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 artificial neural networks and their forever more diversification, from new RNNs architectures *(Graves et al.)* to GANs *(Goodfellow et al.)*, VAEs, etc. Their applications of which the only limit is our imagination, are increasing in domains of science where the amount of data requires such powerful tool. Their recent utilisation on the transcriptome *(Alipanahi et al.)* demonstrates their unparalleled inferential potential.



Simple architecture of a VAE property of Kevin Frans

However through the black box problem, they lack the ability to give answers that researchers can use to understand and explain the causality behind such inferences. Bayesian Networks, Neural Coders and other similar approaches together with a turbo-charged research and innovation is alleviating this problem.

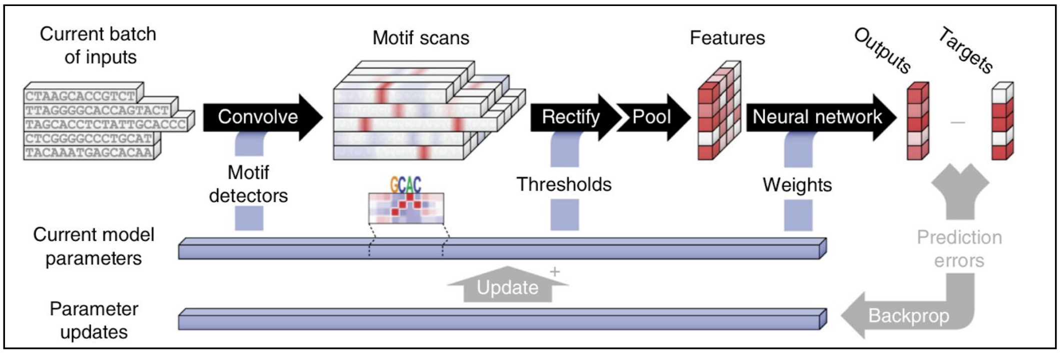
I wish to bring and transform this ever expanding toolkit to the life-improving world of biology, more precisely, genomics and the transcriptome. This encompass the entire regulation network of the cell which -through the eyes of a computer scientist- can be seen as the program that it is running. Understanding it, means to understand the cellular language and how to fully interact with it.
The translational power of this research to the biomedical industry is big and game changing.

This research objective is highly interdisciplinary, the potency of which lies in uncountable examples throughout history. The incessant crossings between neural networks and neuroscience *(Hassabis et al.)* for example, is to me a key example of how disruptive interdisciplinary research can be. I have witnessed it many times from my involvement in CaImAn[1] *(Giovannucci et al. -in review-)* and my participation to HBP workshops.
One of the first working proof of applying neural nets to do inference in the transcriptome has been given by renown computer scientists recently *(Alipanahi et al.)*.

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.

Objectives

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).



Architecture of a neural network to extract information from codons.
It is however done in a supervised fashion. published by Babak Alipanahi
Nature biotechnology All rights reserved.

This task which should result into a joint publication seems to represent a great introduction to my research objectives.

Today a great deal of work in computational biology is focused on inferring structure folding and function from simple information on the linear chain of amino acids that is given by the translation of RNA sequences. However, using known proteins and their interaction in the network of cellular pathways, we can already evaluate how, cis and trans promoters and inhibitors, together with translational regulators can affect the processes and activities of the cell. Thus helping bridging the gap between genotype and phenotype.

According to me, the integration of data science in life sciences and systems biology is slowly happening and will boom; and the open sourcing of data *(NAR’s list...)*, the rapid increase of its production through Next Generation Sequencing and Nationwide health databases makes it a great direction to take for future researches and applications.

Plan

- A first part of this project is to continue my review of the literature
  - This is attained by creating a comprehensive map of the regulation processes with its grey areas and questions with their respective theories, assumptions and the type of data required and available. In an Engineering fashion.
  - 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 (BioPython) to work with.

- 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.
    - There are many parallels made between gene feature extraction models and language processing methods *(DanQ, gene2vec)*. It is very likely that they will appear in many pre-processings.
    - 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 open source projects such as Scikit-learn and Pytorch. Which are firstly the result of one or two researchers and have expanded to encompass many uses and users because of their versatility, modularity and simple, carefully crafted documentation.

Addition :

1. This plan requires that I am able to surround myself with different researchers with various knowledges 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 (UI, API, documentation) and versatile (simplicity, scalability, modularity) as to be verified, tested, modified, improved and reused by other groups across disciplines and objectives.

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