JOBIM mini-symposium proposal:

“Deep learning for omics”

List of abstracts

Laurent Jacob, LBBE, Lyon: *Learning with pangenomes.*

25mn

The growing number of annotated biological sequences available makes it possible to learn models predicting phenotypes from sequences with increasingly high accuracy. The success of predictive models crucially depends on data representation, i.e., on our ability to describe sequences by a set of descriptors which are predictive of the phenotype.

A well-established line of work coined kernels methods has resorted to using a very large number of descriptors, typically relying on the presence/absence profile of short words in each sequence. Storing these large representations or using them in learning algorithms is not manageable, but many learning algorithms can exploit them implicitly because they only use data through pairwise dot products, which can often be computed efficiently. However by forfeiting their explicit access to sequence descriptors, kernel methods also make it more difficult to select the most relevant descriptors from the full representation.

Another, more recent line of work to predict phenotypes from biological sequences is convolutional neural networks (CNNs). Originally designed for computer vision tasks, their distinctive feature is their ability to build a data representation that is optimized for a learning task at hand. They were successfully adapted to sequences, where they reach state of the art accuracies by while producing trained models that allow some level of interpretability.

Here I will present an attempt to cast CNNs for biolgical sequences in the kernel framework. The resulting method inherits the ability of CNNs to learn a data representation. Because it is cast in the well-defined framework of kernels, it also allows for alternative algorithms that are better suited to medium- and small-scale datasets.

Raphael Mourad, CBI, Toulouse: *Deep learning approaches to predict regulatory sequences: a review.*

20mn

Gene expression but also many other complex and crucial cell processes such as DNA replication, 3D structure or recombination are governed by regulatory sequences. These sequences mostly lie into non-coding regions that were previously thought to be « junk DNA ». Understanding how these sequences regulate such cell processes is of great importance in order to decipher mechanisms involving transcription factors. Moreover, reguatory sequence analysis can eventually help to elucidate the roles of specific genome-wide association study (GWAS) SNPs that are difficult to decipher otherwise, since over 95%

of associated SNPs are located outside coding sequences. In this review, I will present current machine/deep learning approaches to predict regulatory sequences, especially k-mer based and convolutional neural network-based approaches. I will also show how such methods are the key to assess the impact of GWAS SNPs in the context of common complex diseases.

Vera Pancaldi, CRCT, Toulouse: *Deep learning to integrate multi-modal phenomics and transcriptomics.*

20mn

A very important application of machine learning approaches in computational biology in the era of personalized medicine and single-cell approaches is represented by multi-omics data integration. The feeling that collecting more data would provide more understanding is crushed by the complexity of biological processes and the difficulty in obtaining different data from exactly the same samples. It is therefore interesting to explore methods to connect the different levels of regulation and detection of biological processes, for example by jointly analysing protein and mRNA levels at bulk or single-cell level.

We are involved in projects describing and modelling interactions between different cells that are found in tumour samples, including immune cells that can either promote or hinder tumour growth. More specifically, we are working on a detailed characterization of tumour infiltrating lymphocytes in blood and bone marrow of multiple myeloma patients and healthy controls.

Measurement of protein abundance via flow cytometry or RNA quantities thanks to single cell RNA sequencing techniques has helped to discover new cell types and has provided new insights in cellular heterogeneity and biological processes in healthy tissues and diseases. We will present recent results on CITE-seq, a technique enabling the joint measurement of both surface proteins and RNA abundances in single cells. The paired transcriptomic and proteomic data produced by this method is a promise for new discoveries, yet, several challenges remain, like the technical biases specific to each omics datasets, and the integration of these data in a unified space.

We will present ongoing results on how recent Deep Learning models based on Variational Auto Encoders can be used to learn a representation of both transcriptomics and proteomics and denoise the data for downstream analyses, such as dimensionality reduction and cell clustering.

Vincent Rocher and Elissar Nasseredine, CBI, Toulouse: *DeepG4: prediction of in vivo g-quadruplexes with deep learning.*

Vincent Rocher, Mathieu Genais, Elissar Nasseredine and Raphael Mourad

20min

DNA is a complex molecule carrying the instructions an organism needs to develop, live and reproduce. In 1953, Watson and Crick discovered that DNA is composed of two chains forming a double-helix. Later on, other structures of DNA were discovered and shown to play important roles in the cell, in particular G-quadruplexes (G4s). Several bioinformatic algorithms were developed to map G4s in vitro based on a canonical sequence motif, G-richness and G-skewness or alternatively sequence features including k-mers. Here, we propose instead a convolutional neural network (DeepG4) to map G4s by assessing its ability to form both in vitro and in vivo (active G4s). State-of-the-art algorithms fail to predict

accurately G4 activity in vivo, while DeepG4 is very accurate. Moreover, DeepG4 identifies key DNA motifs that are predictive of G4 activity. We found that active G4s do not follow a very flexible sequence pattern as previous algorithms seek. Instead, active G4s are determined by numerous specific motifs. Moreover, we identified known transcription factor DNA motifs which could play important roles in G4 activity by either forming G4 structure themselves or by participating in G4 formation in the vicinity.

BREAK :
15 min

Mathys Grapotte, SANOFI, IGMM, LIRMM, Montpellier: Deep neural network to predict repeat transcription.
20min

Using the Cap Analysis of Gene Expression technology, the FANTOM5 consortium provided one of the most comprehensive maps of Transcription Start Sites (TSSs) in several species. Strikingly, ~ 72% of them could not be assigned to a specific gene and initiate at unconventional regions, outside promoters or enhancers. Here, we probe these unassigned TSSs and show that, in all species studied, a significant fraction of CAGE peaks initiate at short tandem repeats (STRs). In order to study these sequences, belonging to the « dark matter » of the genome, we decided to train a convolutional neural network on each of the STR classes. A specific architecture was able to reach Spearman correlations (observed versus predicted CAGE signal) of > 70% on several STR classes. In this talk, we will show how CNN models can be used to acquire novel biological knowledge. We will provide two examples aimed at measuring the impact of genetic variants and evaluating inter-species conservation. Our approach can be extended to other types of deep learning models.

Arnaud Abreu, ICUBE: *Automatic histopathology diagnosis with CNN: hype or reality?*
20min

Deep neural networks have been tackling, with very impressive results, most of the pattern recognition challenges of the past few decades. With the advent of efficient digitization techniques for microscopy slides, deep learning frameworks have been adapted to automate many complex analyses on histopathology images, including diagnosis. Together with a team of expert pathologists, we successfully developed our own framework for distinguishing Follicular Lymphoma from benign Follicular Hyperplasia and tried to identify and address the major limitations that hinder the implementation of these techniques in routine hospital practice.

Julien Mozziconacci: *Genome annotation across species using deep convolutional neural networks*
Ghazaleh Khodabandelou, Etienne Routhier, Julien Mozziconacci
10mn

Convolutional neural networks have recently been exploited for identifying the functional role of short genomic sequences. These approaches rely on gathering large sets of sequences with known functional role, extracting those sequences from whole-genome-annotations. These sets are then

split into learning, test and validation sets in order to train the networks. While the obtained networks perform well on validation sets, they often perform poorly when applied on whole genomes in which the ratio of positive over negative examples can be very different than in the training set. We here address this issue by assessing the genome-wide performance of networks trained with sets exhibiting different ratios of positive to negative examples. As a case study, we use sequences encompassing gene starts from the RefGene database as positive examples and random genomic sequences as negative examples. We than demonstrate that models trained using data from one organism can be used to predict gene-start sites in a related species, when using training sets providing good genome-wide performance. This cross-species application of convolutional neural networks provides a new way to annotate any genome from existing high-quality annotations in a related reference species. It also provides a way to determine whether the sequence motifs recognised by chromatin-associated proteins in different species are conserved or not.