Deep Learning for the reconstruction of protein-protein interactions between paralogs and mutated proteins

Advisor: Alessandra Carbone - alessandra.carbone@lip6.fr

Context and motivations:

Genomics widely uses machine learning to capture dependencies in data and derive new biological hypotheses. By effectively leveraging large data sets, in the last couple of years, **deep learning (DL) has started to transform genomics** the same way it worked on computer vision and natural language processing. Numerous genomics modelling tasks have been proposed, including predicting the impact of genetic variation on gene regulatory mechanisms such as DNA accessibility and splicing. New sequence-based models have been developed to predict protein structures, protein interactions and the effect of mutations in proteins. **These models provide a significantly higher accuracy than state-of-the-art methodologies and offer a truly revolutionary opportunity to accelerate and create innovative applications in genomics.** At the moment, they have very complex architectures and leave the results biologically uninterpretable, but the dream of building machines that reach the level of accuracy for use in the clinical setting becomes a realistic and exciting challenge.

Protein-protein interaction (PPI) networks play a key role in biology and medicine in the interpretation of protein functions in cellular processes. In the past two decades, working with networks has significantly advanced our understanding of the relationships between molecules. This was possible thanks to many computational attempts and high throughput experimental methods such as yeast two-hybrid or tandem purification that have been extensively developed. A particular concern is their level of noise and incompleteness. Indeed, physical interaction networks obtained by high-throughput techniques are found to include numerous non-functional PPI and at the same time many missing true interactions.

Besides, studies giving insights into the precise spatial organization and dynamic temporal remodelling of local protein interaction networks within the cell, highlight that PPI networks are only “projections" of particular spatio-temporal PPI realisations. For instance, within each individual, **genomic alterations** contribute to different PPI realisations. Accordingly, understanding any biological process, necessitates defining three parameters: the composition of the “underlying protein network", its organization in space, and its evolution over time. Future technological developments will bring an overwhelming amount of precise information on these genomics-spatio-temporal dimensions and sophisticated computational tools for extracting information from them are mandatory. Ultimately, PPI networks should be understood within biological frameworks including transcriptomic and epigenetic data. **The construction of the “underlying protein network" which will serve as the basis of more sophisticated and realistic reconstructions makes the first step of this development.**

Because of the experimental difficulties explained above, which are intrinsic to experimental data, *ab initio* computational reconstruction of PPIs are expected to bring invaluable information leading to the discovery of potential molecular interactions.

Advances in the field.

Two deep neural network architectures, DPPI (Hashemifar et al., 2018) and PIPR (Chen et al., 2019), outperformed the state of the art in PPI prediction. In both cases, siamese networks combining layers of convolutional neural networks (CNN) were designed. Additionally, PIPR used recurrent gated network units to account for the sequential nature of the data and alleviate the constraint of fixed length input windows, by showing an increase in performance. Both methods consider a protein sequence as a whole, without taking into account that protein interactions take place through specific binding sites on a protein surface where interaction motifs, that is short amino acid (aa) patterns in sequences, locates.

Recently, we developed IMPRINT (L.David, H.Richard, A.Carbone, IMPRINT: motIfs and augMented sequence space for PRotein partner IdeNTification, to be submitted shortly, 2021), an *ab initio* computational approach to protein partners identification, **coupling Deep Neural Networks (DNN)** and **protein evolution** to achieve an optimal identification of protein partners spanning a large spectrum of biological functions and leading to the reconstruction of PPIs. IMPRINT is based on a new data augmentation schema and explicitly integrates protein interaction motifs in a first layer of its architecture which consists of an array of convolutional filters learned in the training step and partly initialised from a pool of known linear interaction patterns. A biologically guided data augmentation and the integration of prior knowledge in the DNN initialisation includes IMPRINT into the family of DNN architectures that inject biological knowledge into their design and that have already shown great success for functional predictions of genomic regulation signals (Kelley et al., 2016) while providing interpretable DNA motifs.

IMPRINT is the first DNN architecture whose initialisation step is designed by integrating prior knowledge on motifs interaction sites, making it close to the biological mechanisms.

The design of IMPRINT architecture makes it possible to find **protein partners among tens of thousands of proteins in a very reasonable time**. IMPRINT scoring of the 28224 protein pairs of the Mintseris dataset (Mintseris et al., 2005; this dataset is made of protein structures covering a broad spectrum of protein interactions associated to a large variety of functional classes) provided a AUC of 0.97. Moreover, IMPRINT can learn biologically significant interaction motifs and identify the binding site of the interaction for the two proteins. Its data augmentation scheme significantly improves the performance of the IMPRINT architecture compared to other methods.

This thesis: The interest of this thesis is twofold. On the one hand, it tackles a fundamental biological problem and on, the other hand, it develops Deep Learning approaches in genomics, which will become fundamental in years to come. Among other architectural aspects, we want to explore the use of DL to understand patterns produced by a DL architecture.

In this thesis, we wish to apply statistical inference and Deep Learning to protein sequences to scale up in the problem of PPI reconstruction and bring it closer to biological reality. Namely, we want to focus on similar sequences that are **either** homologous proteins, that is sequences with a common ancestor, **or** paralogous sequences, that is sequences with a common ancestor that have been obtained by a duplication of a gene within the same genome, **or** simply protein sequences defined by one or a few mutations:

- Homologs and paralogs are characterised by sequences which evolved with time by influencing the associated PPI network topology due to a change of function. In consequence, the **discrimination of partners** between proteins from a pair of homolog/paralog families or within a family of homologs/paralogs, known to share similar partners over the same interaction site, becomes a challenging problem in the **reconstruction of correct topologies**.
- Mutated proteins are characterized by sequences where even a single mutation might induce a disrupted protein-protein interaction and, as a consequence, a change of the PPI topology of an individual. Identifying deleterious mutations or mutations inducing changes in binding affinity between two proteins is crucial to **identify changes in the topology** of the PPI network which are due to an absence of an expected PPI or to the weakening of the PPI affinity within a network. This is a fundamental question for a personalized medicine.

The two types of sequences are related by evolutionary processes characterised by long and short time-scales respectively, while they both maintain conserved interaction patterns by disrupting the geometry of the interaction sites of the proteins. The phenotypic effect might be deleterious or induce a complete change of function.

We want to develop end-to-end DL architectures that take (two or more) proteins as input and that allow us to distinguish interaction patterns between proteins of similar sequences and identify differences in the interaction **at the scale of a single mutation**. For this, we shall work on data augmentation schema driven by biological information and on the integration of biological information in the architecture such as:

1. charge-charge interactions, generally a good discriminant for paralogue families that are less divergent and a major determinant of specificity in such systems.
2. compensatory mutations (eg coevolution) allowing to identify contact points between amino acids in protein sequences. This contribution will be achieved at a domain level, thus providing information on protein domain interaction.
3. the notion of “functional representation space” of sequences introduced in (Vicedomini et al 2021 under revision) to revisit the notion of evolutionary conservation and taking into consideration functional patterns.

We shall explore the possibility to use transfer learning, introduced in Natural Language Processing (NLP) tasks, utilizing embeddings, and/or representation extraction from multiple pre-trained supervised models, to enrich embeddings with domain specific knowledge.

Moreover, building upon our recent development with IMPRINT, we shall extract and compare “profiles of interaction” produced by IMPRINT, that is 2D curves describing scores of interaction computed by the IMPRINT architecture based on local changes along the protein sequence. These profiles of interaction are partner dependent and describe the interaction site with the partner. Different partners, known to interact over different regions of the protein sequence are detected to interact differently by IMPRINT, that display different profiles. In this thesis, we shall explore how to construct a DL architecture that takes, as input, profiles of interaction and decides whether they are underlying a real interaction or not and whether the binding affinity is weakened in a mutant compared to the wild-type/reference protein (i.e. the $\Delta\Delta G$). This will be a first realization, at our knowledge, of a DL architecture extracting patterns from complex objects produced by a DL architecture.

Consequences of this research. The questions are of interest for the international bioinformatics and genomics community. Indeed, we will couple Deep Neural Networks (DNN) and evolution to achieve what we could not have done a few years ago. We will make it possible to analyze *ab initio* the interactions between **tens of thousands of proteins** and, possibly, bring individual PPI networks into clinics. Not least, we will also contribute to the creation of an environment of DL models for genomics.

Data available. Publicly available datasets will be used to validate our computational approach. Among them: 1. A dataset comprising trio (parents and child) exome sequence data from 4,293 probands from the DDD Study (Deciphering Developmental Disorder study; www.ddduk.org) with severe developmental disorders for pathogenic postzygotic mosaicism in the child or a clinically-unaffected parent. Variant pathogenicity was assessed for each variant and a full list of validated mosaic variants was compiled (Study DDD et al 2017, Wright et al 2019). All diagnostic variants linked to phenotypes are available at https://decipher.sanger.ac.uk/. DDD Study data is available under managed access via the European Genome-phenome Archive (https://ega-archive.org/studies/EGAS00001000775). These exomes are particularly relevant to our main question of explaining mutational effects from the genetic background. (More such anonymized datasets are expected to be made available in the coming years at https://www.ebi.ac.uk/ega/ and https://www.ncbi.nlm.nih.gov/gap/). 2. A comprehensive analysis of oncogenic driver genes and mutations in >9000 tumors across 33 cancer types (Bayley et al 2018). Identification (by PanSoftware) of 299 cancer driver genes, and *in silico* identification of 3400 cancer mutations coupled with experimental validation. As reference genome, we shall use the latest version of the human reference genome available (GRCh38 or more recent; ENCODE project, 1000 Genomes Project). 3. The reconstruction of the PPI networks for the ~2500 individuals of the 1000 Genome project will provide a starting dataset for sophisticated analysis of the structural effects of mapped mutations over protein structures within networks.