DEEP LEARNING FOR GENOMICS: FUNCTIONAL CLASSIFICATION OF PROTEIN-CODING GENES IN OCEAN MICROBIOMES – SUPERVISORS: ALESSANDRA CARBONE (LCQB, SORBONNE UNIVERSITÉ) AND CHRIS BOWLER (IBENS, ENS PARIS)

Now that the issue of predicting protein structures has been largely "resolved" by the unbelievable advances of Deep Learning approaches that lead to AlphaFold [1], determining what proteins do when they interact is the next frontier. This thesis takes on this new challenge to decipher the complexity of the interaction between proteins and other molecules from the perspective of function.

Proteins are key molecules in living cells. They are responsible for nearly every task of cellular life and are essential for the maintenance of the structure, function, and regulation of the unicellular organisms in any ecosystem, from tissues and organs in the human body to the ocean. Cells can produce thousands of different types of proteins (the so-called proteome), which perform a plethora of diverse functions, all crucial for cell viability in their environment. Assigning functions to the vast array of proteins present in cells remains a challenging task in cell biology. This question applies to the multitude of organisms interacting in the ocean and constituting the **ocean microbiome**, that is a highly dilute microbial system that covers the majority of Earth's surface and extends an average of 3600 m down to the seafloor. Broadly, the myriad of protein coding sequences accumulating in our databases coming from ecosystems have no identified function and their **functional classification constitutes the critical bottleneck in their understanding and in our control on their health**.

**In this project, we want to design and train a novel Deep Learning (DL) architecture, which is able to classify sets of sequences by function, discover possibly new functions and functional subclasses.** We shall take advantage of the huge amounts of sequences present in our databases, protein Language Models [2] and multi-view DL approaches [3], and the recent inhouse approach ProfileView [4] devoted **to domain functional classification. The method should allow 1. to infer a function on sequences sharing similar sequence patterns by transferring functional labels from those few sequences where the function is already characterized, 2. to discover the existence of new functions by exploiting new sequence patterns, and 3. to identify functional determinants, that is the ensemble of residues that allows a protein to realize the function.**

The **intrinsic difficulty of the functional annotation problem** relies on the fact that very similar sequences (say, 90% identical) might have different functions. On the opposite extreme, very divergent sequences, presenting low sequence identity, might share the same function even though their divergence might have reasonably led to functional differentiation. The determination of the crucial differences and similarities in protein sequences supporting *functional determinants*, that is the key inter-dependent residues encoding functional diversity and which are conserved within a protein subfamily, becomes a challenge. Other important difficulties of the problem rely on several facts: 1. the number of functions for a protein family are not known, 2. the architecture of a protein, possibly constituted by distinct building blocks called "domains", might influence its functional activity through functional sequence patterns that are shared by multiple domains, 3. some subgroups of homologous sequences are expected to be characterized by specific patterns in their sequences while other subgroups might not, due to an under-representation of protein sequences in the subclass, a possible consequence of "missing" sequences. These different aspects of the problem make it a highly challenging multi-class imbalanced classification problem in genomics and AI.

**Preliminary work on which the thesis will rely upon.** My very recent approach, offered through ProfileView [3], provides a proof-of-concept for this thesis. ProfileView has achieved unprecedented results in accurately classifying seven widespread protein families involved in the interaction with nucleic acids, amino acids and small molecules, and in a large variety of functions and enzymatic reactions. Agreement was found with existing functional data in the literature regarding organization into functional subgroups and residues characterizing functions. ProfileView is based on a novel mathematical concept that defines a formal functional sequence space based on similarities and differences of sequences from profile models of known domains, allowing it to classify sequences and identify functional determinants with no prior knowledge. **This project will reimagine ProfileView with cutting-edge AI techniques and protein-related data** that **did not exist even a year ago**. The availability of vast amounts of protein sequences, the AI advancements in representation learning, protein Language Models (pLMs) [2] and Multiview learning (MV) [3] lead, within this project, to the **new AI learning paradigm of "multiview relational learning"**, making unsupervised functional classification and functional proteomes reconstruction possible, while **scaling-up classification to millions of sequences**. In genomics, multiview learning has been applied to integrate omics data of different origins [3], but has never before been used to classify protein families. Here, we will re-orient the capacities of MV learning into a ***relational learning*** **paradigm**, avoiding supervised *training*. Instead, we exploit differences and similarities within sequences holding a common history, by guiding, based on them, the learning process through a multiplicity of views encoding these relations.

**Plan of the thesis.** First step: we shall produce **ground truth information** by using ProfileView on all protein domain families in the international database InterPro collecting all known domain families. As a result of this first main step, we shall generate a database of functional classifications which is unique. It will contain sequence classification and functional determinants that are characteristic of specific functional classes, that is the ensemble of residues in a protein sequence that are susceptible to play a role in the function.

Second step: we shall work on the design of a Deep Learning architecture for genomic sequences that will produce a functional classification of protein families. In this second step, we shall first construct protein language models (pLMs) encoding **structural** and **functional** information from sets of sequences and from the database of functional ground truth generated in the first step. This will make **a new generation of pLMs** never generated before.

Third step: given a set of sequences to classify, we take a sample from it and construct a representative ensemble of pLMs for the set. We shall define a DL architecture, based on a new ***relational learning*** **paradigm,** that classifies the complete set of sequences based on the pLMs. Intuitively, the pLMs represent the structural and functional characteristics of a protein family and will be used as "views" for the family. Indeed, the ensemble of constructed "views" will be used to determine the similarities and differences of the sequences in the set from the views. The "collaborative" outcome of the views will be sufficient for unsupervised classification, as already demonstrated in ProfileView. This architecture will allow to determine functional classes and their functional determinants. We shall test several supervision strategies for our architecture, namely self-supervision and near-neighbor supervision.

Time permitting, we shall develop other important aspects of our architecture concerning:

1. **the prediction of double functions**. Some protein family might be described by subgroups of sequences that realize a double function. In this case, the functional classification problem becomes a multi-label classification problem, possibly imbalanced. The determination of double functions is a very challenging task that seems to be out of reach today. A relatively small number of experimentally tested examples are known.

2. **proteins with complex domain architecture**. As shown in [4], one can operate under the assumption that the analysis of a single domain is sufficient to functionally classify very different protein families, possibly having complex domain architecture. On the other hand, we cannot exclude that domain multiplicity and/or domain order might play a role for functional classification of some protein families. We expect our DL architecture to treat protein sequences with arbitrary domain complexity and motifs "lying across" multiple domains.

**Experimental validations "in the environment" of the computational biology approach**. Unraveling the functional determinants for protein families will bring insights on the flux of materials, both within microbes and between microbes and their surroundings, constituting microbial function and determining its response to external changes [5]. To show this important consequence of our computational method, Chris Bowler lab will perform expression analyses to see how genes encoding specific protein families, whose function has been inferred by the new approach, **behave in the environment**. This can help determine whether a particular gene may be involved in responding to nutrients, light, temperature etc. The logic behind is that if a gene of interest is expressed with the same pattern as a gene encoding an iron uptake protein, for instance, then we can hypothesize that our gene may also encode a protein involved in iron metabolism confirming, in this way, the computational hypothesis. The experiment will be conceived to target multiple protein coding genes identified computationally. Note that experiments realized "in the environment" are of crucial importance in genomics. On the third year, the student will participate to the design of the experiments.

**Role of each supervisor / skills provided:** AC (https://www.ihes.fr/~carbone/) will advise the student on the computation part for the development of a deep learning architecture and the statistical analysis of (meta)genomic data. CB (https://www.ibens.ens.fr/spip.php?article215&lang=fr) will advise the student on the biological interpretation of his/her results, and guide him/her with the biogeographical and evolutionary studies. We intend to have regular weekly meetings between the three of us to follow the progress of the thesis. The complementarity of the two advisors, one coming from computer science and the other from biology, will allow for an optimal realization of this project that requires both a competence in the development of Deep Learning architectures for dealing with large quantities of (meta)genomic data and a competence in the biological analysis of these data to infer meaningful biological results.

## References


1. Jumper, John, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596.7873 (2021): 583-589.
2. Unsal S., Atas H., Albayrak M., Turhan K., Acar A. C., Doğan T. (2022). Learning functional properties of proteins with language models. *Nature Machine Intelligence*, *4*(3), 227-245.
3. Yan X., Hu S., Mao Y., Ye Y., Yu H. (2021). Deep multi-view learning methods: a review. *Neurocomputing*, *448*,106-129.
4. Vicedomini R., Bouly J.P., Laine E., Falciatore A., **Carbone A.** (2022). Multiple profile models extract features from protein sequence data and resolve functional diversity of very different protein families. *Molecular Biology and Evolution*. *39*(4):msac070.
5. Sunagawa S., Coelho L. P., Chaffron S., Kultima J. R., Labadie K., Salazar G., ... **Bowler C.**, … Bork P. (2015). Structure and function of the global ocean microbiome. *Science*, *348*(6237), 1261359.