Marta Lovino

Algorithms for complex systems in the life sciences

AI for gene fusion prioritization and multi-omics data integration

Due to the continuous increase in the number and complexity of the genomics and biological data, new computer science techniques are needed to analyse these data and provide valuable insights into the main features. The thesis research topic consists in designing and developing bioinformatics methods for complex systems in life sciences to provide informative models about biological processes.
The thesis is divided into two main sub-topics. The first sub-topic concerns machine and deep learning techniques applied to the analysis of aberrant genetic sequences like, for instance, gene fusions. The second one is the development of statistics and deep learning techniques for heterogeneous biological and clinical data integration.

Referring to the first sub-topic, a gene fusion is a biological event in which two distinct regions in the DNA create a new fused gene. Gene fusions are a relevant issue in medicine because many gene fusions are involved in cancer, and some of them can even be used as cancer predictors. However, not all of them are necessarily oncogenic. The first part of this thesis is devoted to the automated recognition of oncogenic gene fusions, a very open and challenging problem in cancer development analysis.

- In this context, an automated model for the recognition of oncogenic gene fusions relying exclusively on the amino acid sequence of the resulting proteins has been developed. The main contributions consist of: 1. creation of a proper database used to train and test the model; 2. development of the methodology through the design and the implementation of a predictive model based on a Convolutional Neural Network (CNN) followed by a bidirectional Long Short Term Memory (LSTM) network; 3. extensive comparative analysis with other reference tools in the literature; 4. engineering of the developed method through the implementation and release of an automated tool for gene fusions prioritization downstream of gene fusion detection tools.

- Since the previous approach does not consider post-transcriptional regulation effects, new biological features have been considered (e.g., micro RNA data, gene ontologies, and transcription factors) to improve the overall performance, and a new integrated approach based on MLP has explicitly been designed. In the end, extensive comparisons with other methods present in the literature have been made. These contributions led to an improved model that outperforms the previous ones, and it competes with state-of-the-art tools.

The rationale behind the second sub-topic of this thesis is the following: due to the widespread of Next Generation Sequencing (NGS) technologies, a large amount of heterogeneous complex data related to several diseases and healthy individuals is now available (e.g., RNA-seq, gene expression data, miRNAs expression data, methylation sequencing data, and many others). Each one of these data is also called omic, and their integrative study is called multi-omics. In this context, the aim is to integrate multi-omics data involving thousands of features (genes, micro-RNA) and identifying which of them are relevant for a specific biological process. From a computational point of view, finding the best strategies for multi-omics analysis and relevant features identification is a very open challenge.

- The first chapter dedicated to this second sub-topic focuses on the integrative analysis of gene expression and connectivity data of mouse brains exploiting machine learning techniques. The rational behind this study is the exploration of the capability to evaluate the grade of physical connection between brain regions starting from their gene expression data. Many studies have been performed considering the functional connection of two or more brain areas (which areas are activated in response to a specific stimulus). While, analyzing physical connections (i.e., axon bundles) starting from gene expression data is still an open problem. Despite this study is scientifically very relevant to deepen human brain functioning, ethical reasons strongly limit the availability of samples. For this reason, several studies have been carried out on the mouse brain, anatomically similar to the human one. The neuronal connection data (obtained by viral tracers) of mouse brains were processed to identify brain regions physically connected and then evaluated with these areas' gene expression data. A multi-layer perceptron was applied to perform the classification task between connected and unconnected regions providing gene expression data as input. Furthermore, a second model was created to infer the degree of connection between distinct brain regions. The implemented models successfully executed the binary classification task (connected regions against unconnected regions) and distinguished the intensity of the connection in low, medium, and high.

- A second chapter describes a statistical method to reveal pathology-determining microRNA targets in multi-omic datasets. In this work, two multi-omics datasets are used: breast cancer and medulloblastoma datasets. Both the datasets are composed of miRNA, mRNA, and proteomics data related to the same patients. The main computational contribution to the field consists of designing and implementing an algorithm based on the statistical conditional probability to infer the impact of miRNA post-transcriptional regulation on target genes exploiting the protein expression values. The developed methodology allowed a more in-depth understanding and identification of target genes. Also, it proved to be significantly enriched in three well-known databases (miRDB, TargetScan, and miRTarBase), leading to relevant biological insights.

- Another chapter deals with the classification of multi-omics samples. The literature's main approaches integrate all the features available for each sample upstream of the classifier (early integration approach) or create separate classifiers for each omic and subsequently define a consensus set rules (late integration approach). In this context, the main contribution consists of introducing the probability concept by creating a model based on Bayesian and MLP networks to achieve a consensus guided by the class label and its probability. This approach has shown how a probabilistic late integration classification is more specific than an early integration approach and can identify samples out of the training domain.

- To provide new molecular profiles and patients' categorization, class labels could be helpful. However, they are not always available. Therefore, the need to cluster samples based on their intrinsic characteristics is revealed and dealt with in a specific chapter. Multi-omic clustering in literature is mainly addressed by creating graphs or methods based on multidimensional data reduction. This field's main contribution is creating a model based on deep learning techniques by implementing an MLP with a specifically designed loss function. The loss represents the input samples in a reduced dimensional space by calculating the intra-cluster and inter-cluster distance at each epoch. This approach reported performances comparable to those of most referred methods in the literature, avoiding pre-processing steps for either feature selection or dimensionality reduction. Moreover, it has no limitations on the number of omics to integrate.