POLITECNICO DI TORINO

Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering

XXIX cycle

Advisor

Prof. Alfredo Benso

Computational Methods For the Modeling of Complexity Of Biological Systems

PhD Candidate: *Alessandro Vasciaveo*

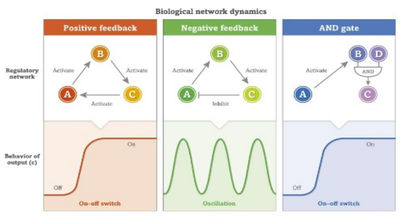
Introduction

Nowadays, the availability and the amount of data, generated from the most diverse ways, spread throughout all research fields. From the traffic generated by communication networks to sensors in the Internet of Things, as well as from Social Networks records to the data stored on the various Cloud platforms, also the data generated by biological High-Throughtput devices is growing in size and in availability acquiring day by day more importance for **Systems Biology and Precision Medicine**, just to cite few examples in Life Science research fields. The data generated by these biological assays (e.g. cDNA Microarrays or the newest RNA-Seq technology) allows scientists to get a profile either of the genetic code that is programming the life of a living organism or is providing to researchers a snapshot of *what is going on* within a cell of a organism body (e.g. humans or mice).
The recent advances, in those technologies that are able to capture the biological state of cell populations in **Terabyte of data**, did not help to automatically uncover the molecular mechanisms in a systematic manner. Moreover, this huge amount of information is also often heterogeneous and cannot be handled directly by hand from a pool of biologists, but it is necessary to use ad hoc computational methods in order to **extract the knowledge** which is hidden by the **complexity of the data**.

Gene Regulatory Networks Simulator

Gene Regulatory Networks (GRN) are models of the molecular interactions that represent the behavior of regulatory mechanisms in the cells of living organisms. These networks are composed mainly by connections between genes and the understanding of the dynamics of the change in time of the states of this complex system allow researchers to better understand the arise and the progression of diseases.
Here, at Politecnico di Torino, we modeled the network dynamics of GRNs using the Boolean Network paradigm. The result of this work has been the design of algorithms and their implementation in C++ in a software tool called Enhanced Boolean Network Toolkit (EBNT).
Recently, researchers discovered new components as small transcript of RNA called MicroRNAs (or in general non-coding RNA) that are not containing the message for the synthesis of any protein but are still actors that take part in regulatory mechanism.
Thus, one of the main feature that we implement in our tool is the ability of modeling post-transcriptional mechanisms (i.e. using MicroRNA as nodes in the network topology).
Further details about this approach have been published here:

- Benso, A., Di Carlo, S., Politano, G., Savino, A., & **Vasciaveo, A**. (2014). An extended gene protein/products boolean network model including post-transcriptional regulation. *Theoretical Biology and Medical Modelling*, *11*(1)
- Politano, G., Savino, A., Benso, A., Di Carlo, S., Rehman, H. U., & **Vasciaveo, A**. (2014). Using Boolean networks to model post-transcriptional regulation in gene regulatory networks. *Journal of Computational Science*, *5*(3), 332-344.



Motivation

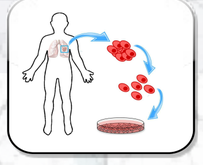
The huge amount of data generated by biological assays in the last years raised the need for sofisticate software tools for several reasons. For example, after the acquisition and the collection of the assays, the signal has to be detected and separated from the noise which is always present. Mathematical and statistical models are needed to account for this issue. Another reason is found in *how to make useful* out of the data for the research purposes: statistical frameworks and information theoretic approaches are valid tools to model properly these very intricate biological phenomena.
Thus, thanks to my background in computer engineering, one of my research aims has been the study of algorithms and models to manage such a complexity, giving new tools to the Computational Biology community and new insights to the Systems Biology community.

On Cloud Platform for the Simulation of Biological Networks

With the collaboration of the Instituto Superiore Mario Boella, we created an ad-hoc platform using Apache Hadoop and the **MapReduce algorithm** to simulate the dynamics of complex biological networks with thousands of components and high degree of node connectivity.
For further references see:

Vasciaveo, A., Benso, A., Di Carlo, S., Politano, G., Savino, A., Bertone, F., ... & Terzo, O. (2015). A cloud-based approach for Gene Regulatory Networks dynamics simulations. In *4th Mediterranean Conference on Embedded Computing (MECO)* (pp. 72-76). IEEE.

ISMB
Istituto Superiore Mario Boella



Precision Medicine

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

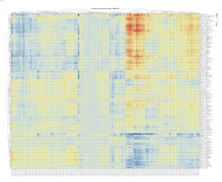
By using Mutual Information to find relationships between genes and their products (i.e. proteins) it is possible to reconstruct Gene Regulatory Networks from the biological data samples of gene expression (i.e. RNA-Seq assays).
This process is called **reverse engineering** of cellular networks.
These networks are used to build *interactomes*, context-specific networks that characterize the protein activity of the biological phenomenon described by gene expression values from the samples.
In my research, I am building and querying these interactomes using data from each patient (e.g. from TCGA online database and other sources) to get the profile of her protein activity in order to find the right drug or drug combinations that play a synergistic effect in the reverse of the aberrant protein activity that characterize the single patient disease (tumor most of the time).
This field is a very new research trend and it is very promising because its aim is to provide a optimized treatment for a single patient based on its genomic profiles, reducing the collateral drug effects of a non personalized treatment.

Conclusions

Here are shown few highlights of the research activity I have been able to conduct at Politecnico di Torino in the last years. The numerous collaborations abroad that I performed in renowned research institutions, such as the German Cancer Center in Heidelberg and the Columbia University of New York, are a confirmation of the increasing importance that researchers in Life Science are giving to computer science and engineering approaches. These methods and algorithms to solve large-scale problems are the only tools able to account for the huge complexity that the biological phenomena are showing to us.
Thus, it is undoubted that a research career in Systems Biology or Computational Biology is enjoyable and fruitful for a enthusiastic scientist with a background in **engineering.**
We have the tools and the methods to observe, model, understand and enhance the world in which we are living.

Methods

Since **networks** are useful to model **complex systems**, they are heavily adopted in many research areas, and they find applications in Systems Biology too. Furthermore, scientists discovered that genes never act alone in a biological system, but they partecipate in a cascade of networks. More specifically, interactions among biological compounds are mapped into a particular kind of networks referred to us as pathways. These networks are constantly growing by manual annotations due to scientific validated discoveries and with predictions made by machine learning algorithms that try to exploit the huge amount of information contained in the High-Throughtput biological screenings mentioned above. The so called Genes Regulatory Networks (GRNs) are the most investigated ones becoming one of the principal subjects of my research.



Cancer Systems Biology

The understanding of cell regulatory mechanisms is extremely useful to model the disease progression. In my research, in a collaboration at Columbia University of New York, we are looking at cancer tissue samples for the characterization of proteins that drive tumor subtypes. In particular, we are following an hypothesis that assumes the phenotypic change to be driven by a small set of proteins named Master Regulators. These actors are modeled by regulatory networks.
Control theory and Information Theoretic approaches are used to unravel these mechanisms.
In Particular, feedback loops in the network topology of a GRN are responsible of the resistant behavior of tumor cells when their are targets of drugs in chemotherapy. By modeling these loops (e.g. using Dynamic Bayesian Networks) it is virtually possible to break them using the right combinations of drugs on those targets that are implementing the feedback control of the system.

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

Gene Therapy

In Gene Therapy, the main goal is to cure a disease by the elimination of the wrong genetic instructions with the placement of new working ones. The purpose is very close to bug corrections in software development. For example, when a organism has a disease (e.g. cancer in humans or a inherited genetic disease like hemophilia) it is because the instructions encoded in the DNA are not able to execute the right mechanisms that allow a flawless work at the molecular level within the cells of the body. An example is the hemophilia diseases in which *"patients born with Hemophilia are not able to induce blood clots and suffer from external and internal bleeding that can be life threatening"*. This is a very hot research field in which the understanding of how to correct the ***bug*** is of crucial relevance.
The wrong placement of the new working instructions are often causes of tumor, thus a better understanding of how the delivery of the corrected genes can degenerate in **cancer** is of paramount importance.
The work has been conducted in collaboration with the German Cancer Center (DKFZ) and the National Center For Tumor Diseases (NCT) of Heidelberg in Germany.

dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

A Graph-Based Framework for the Analysis of Viral Integration Sites

NCT NATIONAL CENTER FOR TUMOR DISEASES HEIDELBERG
supported by German Cancer Research Center (DKFZ) Heidelberg University Medical Center Hospital for Thoracic Diseases German Cancer Aid

In my research, the focus has been on the design and development of a new framework based still on a **network approach** for the analysis of the clusters formed by viral integration sites in the genome of mammalian organisms such as humans and mice. The framework has been implemented in Java and the R language has been used for the building of prototypes of the **statistical models**. Once validated, a Java application was released ready to be used by biologists who may query the results of the pipeline and generate consistent reports.
This software is in testing by laboratories whose work in translational medicine is necessary to asses the saefty of viral vectors. This is extremely important because those vectors are carrying the genomic features whose missing is causing the disease.

Further details about this research have been published on international journals this year:

- Fronza, R., **Vasciaveo, A.**, Benso, A., & Schmidt, M. (2016). A graph based framework to model virus integration sites. *Computational and structural biotechnology journal*, *14*, 69-77.
- **Vasciaveo, A.**, Velevska, I., Politano, G., Savino, A., Schmidt, M., & Fronza, R. (2016). Common integration sites of published datasets identified using a graph-based framework. *Computational and structural biotechnology journal*, *14*, 87-90.