I Introduction

Gene expression consists in the synthesis of proteins from messenger RNAs (mRNAs), which are transcribed from the DNA. This process is regulated by transcription factors (TFs), proteins that bind to the DNA, regulating the synthesis of the mRNAs. One TF can control the expression of multiple genes, and the expression of one gene can be regulated by several TFs. This gives rise to ‘Gene Regulatory Networks’ (GRNs), that control cell differentiation and development, response of the organisms to the environment and in general all cellular functions.

Gene expression is a noisy process: at the single-cell level, the main noise sources are the stochasticity of individual molecular events and the diffusive nature of the arrival of molecules to their targets (such as the recognition of the promoter on the DNA by the TF); these processes are identified as intrinsic noise.

In addition, there are extrinsic noise sources, such as the presence of other molecular species involved in the expression of a given gene, the stage in the cell cycle and environmental factors. Both classes of noise cause concur in the imposition of physical limits to the reliability with which cells can regulate the expression of the genes. Moreover, they cause the expression of a given gene to be different across genetically identical cells and this cell-to-cell variability is today a central object of study in Molecular Biology, thanks to the development of single-cell genomics and transcriptomics.

The emergence of these techniques, together with the identification of network motifs, is giving the possibility to quantify cellular noise and to build models of the regulation of gene expression. Viewing cells as information-processing devices, the presence of noise sets a limit to the maximal amount of information that can be transmitted from regulator molecules to targets (e.g., the copy number of a protein and the TF regulating the expression of the gene producing it). Recently, the optimization of information flow has been hypothesized as the theoretic principle that governs networks in regulating gene expression. This principle is supported by several experimental evidences that real networks operates close to the theoretically predicted limits.

In my Ph.D. project I will focus on different aspects of the regulation of gene expression, using Statistical Mechanics and Information Theory approaches. Firstly, I will study information flow in GRNs in the presence of extrinsic noise on the network parameters. This aspect has not yet been considered from a theoretical point of view, so it could be interesting to quantify its effect on the ability of networks to transmit information. Regarding cell-to-cell variability of gene expression, I would develop a method to infer the set of ‘critical’ variables from the noisy background in single-cell transcriptome data. Finally, I would address a biologically relevant problem: modeling a recently discovered RNA network that controls muscle cell differentiation.

II Information flow in GRNs with extrinsic noise

Information transmission between input and output layers in GRNs is limited by the presence of noise and by the distribution of the input concentrations. In a recent work on small GRN [1], Tkačik et al. assumed the maximization of mutual information as a starting point to find the optimal parameters of the network. Modeling the shape of the input/output relations and of the intrinsic noise sources in the circuit, they were able to find an expression of the mutual information, in the case of one TF regulating multiple outputs, in the limit in which noise level on the outputs is small. The maximization of this expression with respect to the probability distribution of the input concentration provides an...
expression of the optimal mutual information, $I_{opt} = \log_2 Z$, where $Z$ depends on the parameters of the input/output relations and the maximal input concentration.

Starting from these results, I would investigate the role of extrinsic noise on the parameters of the network, assuming that the thresholds of the input/output relations have a certain probability distribution. Initially, I will consider a network composed by one TF activating the expression of $M$ targets. The first goal of the model is the analytical computation of the annealed average of the optimal mutual information and the comparison with a numerical computation of the quenched average. Then, the analytical result will be extended to the presence of repressors, giving the possibility to study the optimal proportion between activators and repressors in the network.

Moreover, it could be interesting to study which probability distributions of the thresholds maximize the mutual information, if the introduction of correlations between the thresholds changes the way in which the targets use the input dynamic range and what analytical progress is possible if couplings between the targets are introduced in the model.

**III Inferring critical variables from transcriptome data**

Until a few years ago, experimental techniques for the study of the transcriptome allowed only the observation of its averaged behaviour. The recent advent of single-cell techniques gives insights into cell-to-cell variability of gene expression, leading to the discovery of new cell subtypes and to an improved characterization of diseases, such as cancer.

I am interested in the analysis of single-cell transcriptome data to infer ‘critical’ variables, intended as the set of most representative from the point of view of their variability across samples, from the noisy background. In particular, I will study the role of cell-to-cell variability of gene expression in Multiple Sclerosis (MS), a multifactorial autoimmune disease of the central nervous system, whose etiology is still largely unknown. I will work on a dataset obtained from single-cell RNA sequencing of a pool of cells belonging to healthy and ill patients.

In order to extract the set of most critical transcripts, I would implement a stochastic optimization algorithm, using as objective function an entropy defined on a subset of the data, which quantifies the heterogeneity in the expression of the transcripts belonging to that subset across cells. This function has been defined firstly in [2], [3] in an information theoretical framework, with the aim of finding the most informative clustering of a dataset. Recently, it has been applied successfully to a biological problem, namely to the identification of critical residues from Multiple Sequence Alignment of protein families [4].

This method will probably enrich the set of standard Bioinformatics tools for transcriptome data analysis, since it will give the possibility to discover novel biomarkers in problems in which variability of gene expression plays a central role.

**IV Minimal model of a real regulatory network**

I will focus on a recently discovered RNA network constituted by both coding (mRNAs) and non coding (miRNAs and long non coding RNAs (lncRNAs)) RNAs, controlling the timing of skeletal muscle cell differentiation. It has been studied experimentally in [5], [6], where it has been found that the lncRNA linc-MD1 competes with MAML1, MEF2C (two TFs needed for the expression of muscle-specific genes) and the human antigen R (HuR) protein for the binding of two miRNAs, miR-133 and miR-135. This competition gives rise to an effective crosstalk between the miRNA targets, called ‘ceRNA (competing endogenous RNA) effect’. Moreover, this network is characterized by a feedforward loop controlled by HuR, that affects the levels of the lncRNA and the miRNA miR-133, and by the presence of two channels of miR-133 production, both from a precursor common to linc-MD1 and from independent genomic loci.

Relying on two recent works, in which the steady state [7] and the dynamics [8] of a generic ceRNA network have been studied, I would model the system through mass-action kinetic equations. Then, the main goals are the numerical study of the ‘ceRNA effect’ at steady state and the characterization of the two main regulatory mechanisms of the system, namely the feedforward loop controlled by HuR and the presence of the alternative channels of miRNA transcription.

In order to validate the model, I will compare the numerical solutions of the dynamic equations with time-resolved experimental data presented in [5], [6].

A further improvement of the model could consist in the introduction of molecular noise. Indeed, though a deterministic model is suitable to quantify the ‘ceRNA effect’, the presence of intrinsic noise could be useful to study the effectiveness of control more precisely.
References


