Gene regulatory networks: prioritizing techniques to find gene candidates

PhD Student: **Leonardo Lenzini** Supervisor: **Roberto Livi**

A growing body of knowledge is now establishing a considerable role for non-coding DNA in gene expression and its regulation, thus implying its potential role in human diseases. The researchers are now in accordance when noting that predicting the consequences of genetic variation in these non-coding regions, especially those involved in gene regulation, will be challenging.

A promoter is the region of DNA which lays upstream of a gene and contains the information required for transcriptional initiation and regulation. At its most fundamental level, the function of a promoter is to recognize information about the status of the cell and, accordingly, regulate gene's expression. It is nowadays well established that promoters are crucial for gene regulation [1].

There have been proposed many methods trying to answer fundamental questions about promoters. In particular, base composition analysis (BCA) and specific entropic indicators have been employed for identifying structural similarities among different classes of promoters [2]. Moreover, the region around the transcription start sites (TSS) has been shown to exhibit a very distinctive structural profile, which seems to be actively maintained by non-neutral selective constraints [3]. In a recent paper [4] these features have been further investigated through a genome-wide analysis of *H. sapiens* promoters by exploiting a fully general mathematical procedure based on the combination of two spectral methods. This analysis showed that *H. sapiens* promoters can be pooled into four main groups. Indeed, because of the huge amount of data produced in last decades, many efforts are nowadays spent in order to identify the most promising genes among a larger pool of candidates. Gene prioritization aims to this purpose [5]. Its goal is to maximize the yield and biological relevance of further downstream screens, validation experiments or functional studies by focusing on the most promising candidates. Computational methods for gene prioritization have progressed quickly. They now demonstrably contribute to biological discovery. Their ability to gather and to integrate data from multiple sources provides a more thorough and less biased global assessment of candidate genes than can be manually achieved. Although prioritization methods have greatly improved in the past few years, some methodological improvements are still necessary. The challenging issue of the present project is to take advantage of promoters features in order to develop a new an innovative prioritization method.

Gene regulation play a central role also in the microbial world. Bacterial pathogens, for example, usually have precise control mechanisms for toxin production so that expression only takes place when required e.g. when the density of the bacterial population overcomes a certain threshold, or when the bacterium reaches a certain cell-type/organ. Similarly, environmental microbes sense and respond to changes in external conditions by means of gene and gene clusters regulation. In all these cases, promoter regions are crucial elements in the complex circuits determined by gene expression patterns. Accordingly, studying and modeling their structure together with their interactions with other cellular elements (e.g. regulatory proteins) may have important implications in different research areas, such as clinical, industrial and environmental microbiology.

We aim to broadening the field of analysis of our past research on gene's promoters going beyond the mere nucleotide sequence and taking into account biological information regarding interactions among promoters. The ambitious goal of this project is to take advantage of clustering applied to promoter sequences in order to develop a new tool for prioritizing genes. The main idea of this project is to define an interaction between promoters and then analyze the network obtained. The analyses we will carry out take inspiration from previous works [6,7], where notions of Random Matrix Theory (RMT) and percolation theory are applied to the study of gene expression networks.

We will apply the basic theoretical tools used to describe and analyze networks, most of which come from graph theory, the branch of mathematics that deals with networks. This step is devoted to measuring and quantifying the fundamental quantitative features of the network structure such as the degree, betweenness and significance. We will apply the methods of the RMT [8-11] to the study of the properties of our genetic network. The tools of RMT will allow us to build a procedure that identifies important nodes in the network, in order to extract biologically relevant information in the frame of gene prioritization problems. We will adapt the theory of percolation [12-15] to our multilayer network with the aim of investigating its robustness.

We will tackle promoter characterization in a different way, treating them as strings of a text. A procedure for characterizing promoter sequences can be worked out by suitable entropic indicators [16, 17]. We aim to investigate the correlations between structure and function in gene promoter sequences and to obtain an insight on putative selective constraints to randomness at promoter level that may be present due to the optimization of the interactions with the transcription factors, or in general because of functional constraints acting on the structure and composition of promoters. Regions subject to such constraints can be identified as those that exhibit a biased variability across promoters with respect to the nearby regions, as well as those showing a selection on the sequences content. We plan to treat promoters as strings of text to identify relevant information via text mining techniques. This analysis will result in a set of short DNA sequences that are good candidates to have an important biological role for promoter functions. This set of sequences will undergo an in-depth survey on their biological role searching for their correlation with biologically relevant functions.

References

[1] Lenhard, B., Sandelin, A., & Carninci, P. (2012) *Nature Reviews Genetics*, *13*(4), 233-245.
[2] Calistri E, Buiatti M, Livi R (2014) Journal of Theoretical Biology 363: 357 – 366.
[3] Calistri E, Livi R, Buiatti M (2011) Molecular Phylogenetics and Evolution 60: 228 – 235.
[4] Pettinato L, Calistri E, Di Patti F, Livi R, Luccioli S (2014). PLoS ONE 9: e85260.
[5] Moreau Y, Tranchevent LC (2012) Nature Reviews Genetics 13: 523–536.
[6] Jalan S, Solymosi N, Vattay G, Li B (2010) Phys Rev E 81: 046118.
[7] Luo F, Yang Y, Zhong J, Gao H, Khan L, et al. (2007) BMC Bioinformatics 8: 299.
[8] Mehta ML (2004) Random matrices, volume 142. Academic press.
[9] Guhr T, Muller-Groeling A, Weidenmuller HA (1998) Physics Reports 299: 189–425.
[10] Brody TA, Flores J, French JB, Mello P, Pandey A, et al. (1981) Reviews of Modern Physics 53: 385.
[11] Dyson FJ (1962) Journal of Mathematical Physics 3: 140–156.
[12] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Physics Reports 424: 175 - 308.
[13] Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E., & Havlin, S. (2010). *Nature*, 464(7291), 1025-1028.
[14] Gomez S, Diaz-Guilera A, et Al. (2013). *Physical review letters*, *110*(2), 028701.
[15] Bianconi, G. (2013). *Physical Review E*, *87*(6), 062806.
[16] Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E. *Gene* 389(2007) 52 - 65.
[17] Marsili M, Mastromatteo I, Roudi Y. *Journal of Statistical Mechanics: Theory and Experiment* (2013) P09003.