Gene Expression vs. Network Attractors

Gene Expression vs Network Attractors

Gianfranco Politano, Alessandro Savino, Alessandro Vasciaveo

Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
email: <firstname>.<lastname>@polito.it
web: www.sysbio.polito.it

Extended Abstract

Abstract: Microarrays, RNA-Seq, and Gene Regulatory Networks (GRNs) are common tools used to study the regulatory mechanisms mediating the expression of the genes involved in the biological processes of a cell. Whereas microarrays and RNA-Seq provide a snapshot of the average expression of a set of genes of a population of cells, GRNs are used to model the dynamics of the regulatory dependencies among a subset of genes believed to be the main actors in a biological process. In this paper we discuss the possibility of correlating a GRN dynamics with a gene expression profile extracted from one or more wet-lab expression experiments. This is more a position paper to promote discussion than a research paper with final results.

1. Objectives

Microarrays, RNA-Seq, and Gene Regulatory Networks (GRNs) are widely used to study and model gene expression status and mechanisms in a cell. Whereas microarrays and RNA-Seq techniques are the output of wet-lab experiments, GRNs are used for in-silico analysis and simulation of biological pathways [8-9]. GRN models, commonly known as "pathways", are available in several web repositories as Kegg, WikiPathway, Reactome [10-12]. Nevertheless, despite focusing on the same biological process, gene expression technologies and GRN are intrinsically different. The first big difference is in the type of information they can provide. Gene expression experiments provide an indication (more qualitative or quantitative depending on the technology) of the expression level of a large amount of genes in a particular physiological condition. Nevertheless they cannot provide any reliable information about causal relationships between genes (which gene activates which). On the contrary, GRNs are used to model regulatory relationships between genes, rather than focusing on the precise expression status of a single gene in a particular physiological condition.

A second important difference is the source of information. A GRN ideally models the regulatory network dynamics of a single cell, whereas gene expression technologies show the "average" expression status of all the cells present in the sample used in

the experiment. This can be misleading and for this reason in certain wet-lab experiments cells are "synchronized" in order to increase the probability of having most of them in the same state.

Despite these differences, GRNs and gene expression experiments are very often used together to try to elucidate several regulatory mechanisms involved in the life of a cell. In this case the obvious assumption is that the Microarray or the RNA-Seq show a "live" view of the regulatory mechanisms modeled by the GRN. The problem is that there is no measure to quantify how much a GRN model is "compliant" or "compatible" with the gene expression profiles result of a lab experiment. If they provide data referring to the same biological process, then there should be a way to correlate them. To our knowledge, in literature there is no formal way designed to understand if a gene expression experiment is actually showing expression data compatible with a GRN model. "Compatible" in this case means: "that it has a high probability of having been generated by a set of genes regulated as described by the GRN model".

The goal of this paper is therefore to discuss the following problem: "Is it possible to find a relation between a GRN model and the expression profile of its genes extracted from one or more lab experiments?" In this paper we are not presenting final and conclusive data (yet); we try instead to formalize the problem and explaining the methodology we are applying to find an answer.

2. Methods

The problem stated before requires subdividing the investigation into several steps. The first and most important one is to understand which data needs to be used, and how to make the data extracted from a GRN and from an expression experiment statistically comparable. The second step is to choose the correct statistical method to compare the data, and the third and final one is to understand if and which part of the network dynamics extracted from the GRN are correlated to the ones showing in the expression experiments.

Gathering the correct data

When running a gene expression experiment, especially on a tissue sample, the resulting data are not a clear image of the expression profile of the target genes. They are instead the quantification of the average expression of certain genes in the cells of the sample. In general, the activation of a particular gene in the sample is either static or dynamic. It is static if the expression of that gene does not change in time, for example in housekeeping genes. It is dynamic if the gene is involved in one or more biological processes that are active at the time of the experiment. In this case, the expression of the same gene in different cells is likely to be different, depending on the particular status of the cell. As an analogy, think about taking several different pictures of the same street always from the same position but at different times. In the street only two cars are present. One is blue and it is always parked (static expression gene), and one is red and always moving (dynamic expression gene). Imagine now overlapping all the pictures and trying to pinpoint the position of the blue and red car. The blue parked car will always be in focus and still, while the red moving car will

appear in several different positions. It will not be possible to pinpoint its exact position, but only to have an "average" idea of its position.

Microarrays provide a less quantitative data than RNA-Seq [6]. In Microarrays the expression of a gene is converted into a real number corresponding to the intensity of the corresponding spot. Nevertheless, because of technological biases, the expression of a gene is not easily comparable with the expression of a different gene. For this reason Microarrays are usually the choice for differential expression experiments, where the focus is on the difference of expression of a set of genes between two different phenotypes [1-2]. RNA-Seq technology instead is able to provide a much more reliable quantitative value of the expression of a gene. Databases like Gene Atlas [3] store large sets of experiments that characterize the average expression of most known genes in different tissues and in different conditions (baseline and pathological). For these reasons, RNA-Seq data appears to be better suited for the study discussed in this paper.

Data concerning the network dynamics collectable from a GRN requires making a number of assumptions. To make the simulation of a realistic network (tens of nodes) computationally feasible in a reasonable time, it is possible to use a Boolean model to limit the possible states of each node/gene to only two possible values. Despite this simplification, this allows studying the network dynamics in terms of "steady-states" or "equilibrium-states" or, more, formally, "attractors". An attractor is a set of states (one or more) towards which the network tends to converge. Once an attractor is reached (and the inputs of the network remain steady) it is not possible to transition out of it, unless external perturbations are applied. In case of a point attractor, the system's state freezes whenever the network enters the attractor. Differently, cyclic attractors (most common in GRNs) show a cyclic behavior of the system: once a cell falls into one of the states belonging the attractor, the system keeps cyclically moving among the attractor's states. Attractors are also present in multi-valued or even continuous models (where the state of each gene is a real number), but their computation is computationally unfeasible for networks of more than a few nodes.

From the definition of "attractor", it is reasonable to assume that they represent high probable states for a cell [7], and therefore can be considered the states that contribute more to the expression of the genes. "Attractors" are therefore the network dynamics data that we consider more suitable to be compared with a gene expression profile.

Making data statistically comparable

As we discussed before, if the pathway is correct (i.e. it reliably represents the real regulatory dynamics of the genes included in the model), then we can make the assumption that, without strong external perturbations, at any given moment a cell has a high probability of being into an attractor state. If this assumption holds, then the expression profile of a gene expression experiment should be correlated to the attractors in which these cells were during the experiment.

To prove this, we need to generate, from both the GRN model and the expression experiments, two statistically comparable datasets.

From the GRN model it is possible to compute the set of attractor states by running the Enhanced Boolean Network Simulator presented in [4][8]. The result is a set of `n` Boolean arrays. Each element of the arrays represent a gene in the network, and its value ('0' or '1') its expression value. These steps are those depicted in violet in Figure 1.

Given a subset of attractors '`a`', for each gene (array position) it is possible to compute the Attractor Expression Frequency (AEF_a) as the number of '1' that the gene shows in the attractors set '`a`'. Now, if (most of) the cells participating in the expression experiment are in one of the attractor states of '`a`', then the average expression of a particular gene should be correlated with its AEF_a. Obviously, the set '`a`' of attractors to be used to compute AEF has to be carefully selected. This step will be explained later.

From the gene expression experiment point of view, shown in orange in Figure 1, and for the particular purpose of this work, the normalized value of FPKM (Fragments Per Kilobase of transcript per Million mapped reads) available from the Expression Atlas [5] of the EMBL European Bioinformatics Institute currently appears to be the best choice, since it represents an "expression frequency" that can be statistically compared with the AEF measure extracted from the GRN model.

Statistically comparing data

Once these two datasets are obtained, we need to statistically compare them. To perform this comparison it is worth to consider that the normalized FPKM of each gene can be considered as a frequency of the appearance of the gene expressed in the specific experiment. The gene profile composed of FPKM values therefore represents a profile of **expected frequencies** for all genes. On the other hand the AEF represents a set of **observed frequencies** from the simulation of the GRN. In order to compare the two set of frequencies a chi-square test can be used to determine whether there is a significant difference between the expected frequencies and the observed frequencies. The acceptance of the test's null hypothesis would be a confirmation that the two set of frequencies are not statistically different, giving an indication that the AEFs are somehow linked to the gene expression profiles obtained in the lab.

Looking for a correlation

If we compared the expression profile with the complete set of attractors of the target GRN, we would probably not obtain any significant result. From a theoretical point of view, this is intuitive: if the attractor set represents a set of the network possible states, it is very likely that the number of '1' and '0' in each position (and therefore the expression of the corresponding gene) will be very similar and therefore the AEF would be probably close to 0,5. But biologically, considering all attractors of the network does not make a lot of sense: of all the possible attractors of a network, probably only a subset is biologically valid and/or significant. Moreover, within this subset, not all attractors are equally probable. It is instead very likely that there are very probable attractor states, as well as very rare ones. This bias could affect the result of

the statistical tests. It is therefore necessary to devise an algorithm able to find the best set of attractors that better correlates with the gene expression profile. We plan to solve the problem with an evolutionary algorithm (in green in Figure 1), since they are very efficient in analyzing very large solution spaces.

Figure 1 - Flow chart of the overall method

A simple algorithm consists in the evaluation of the statistical correlation between AEF and FKPM measures performed by a fitness function. If the correlation could be considered statistically strong (i.e. the Chi-square value is greater than a threshold, or a rho value between 0.5 and 1 if the correlation test is the Spearman's rank-order), then a statistical hypothesis test should be performed in order to confirm the statistical significance of the result. Otherwise, if no strong relationship is found, then a new attractors set is collected within the selection phase and new AEF measures are computed. The termination conditions and genetic operators (not shown in Figure 1 for brevity) are the same of classical evolutionary algorithms [13].

3. Conclusions

The main reason for writing this paper is to stimulate other researchers to look into this idea. The use of biological network already proved to be successful in several Life Sciences areas, but their full potential has not surfaced yet mainly because there is no way of understanding how well a GRN is modeling the actual regulatory mechanisms in the cell. This study goes into the direction of formalizing a methodology to investigate if making this correlation is possible. The possible outcomes are significant. First of all having a realistic measure of a pathway model "biological compatibility" would allow to dramatically increase the quality of the pathway network models, and, concurrently, making them actually usable for understanding the complex systems that regulate cell life.

4. Acknowledgments

This work has been partially supported by grants from Regione Valle d'Aosta (for the project: "Open Health Care Network Analysis" - CUP B15G13000010006), from the Italian Ministry of Education, University & Research (MIUR) (for the project MIND - PRIN 2010, and FIRB Giovani RBFR08F2FS-002 FO), from the Compagnia di San Paolo, Torino (DT), and from AIRC 2010 (IG 10104 DT).

5. References

[1] Seita, J., Sahoo, D., Rossi, D. J., Bhattacharya, D., Serwold, T., Inlay, M. a, … Weissman, I. L. (2012). Gene Expression Commons: an open platform for absolute gene expression profiling. PloS One, 7(7), e40321. doi:10.1371/journal.pone.0040321

[2] McCall, M. N., Uppal, K., Jaffee, H. a, Zilliox, M. J., & Irizarry, R. a. (2011). The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. Nucleic Acids Research, 39 (Database issue), D1011–5. doi:10.1093/nar/gkq1259

[3] Petryszak, Robert, et al. "Expression Atlas update—a database of gene and transcript expression from microarray-and sequencing-based functional genomics experiments." Nucleic acids research 42.D1 (2014): D926-D932.

[4] Benso A., Di Carlo S., Politano G., Savino A., Vasciaveo A., An Extended Gene Protein/Products Boolean Network Model Including Post-Transcriptional Regulation, THEORETICAL BIOLOGY AND MEDICAL MODELLING, Vol.11 (Suppl 1), pp.1-17, ISSN: 1742-4682

[5] http://www.ebi.ac.uk/gxa/home (last visit December 2014)

[6] Nagalakshmi, Ugrappa, Karl Waern, and Michael Snyder. "RNA-Seq: A Method for Comprehensive Transcriptome Analysis." Current Protocols in Molecular Biology (2010): 4-11.

[7] Huang, S., Eichler, G., Bar-Yam, Y., & Ingber, D. E. (2005). Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network. Physical Review Letters, 94(12), 128701. doi:10.1103/PhysRevLett.94.128701

[8] Benso A., Di Carlo S., Rehman H.U. ,Politano G.,Savino A., Squillero G.,Vasciaveo A., Benedettini S., Accounting for Post-Transcriptional Regulation in Boolean Networks Based Regulatory Models, in International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO 2013, pp.397-404

[9] Politano G., Benso A., Di Carlo S., Savino A., Ur Rehman H., Vasciaveo A. (2014) *Using Boolean Networks to Model Post-transcriptional Regulation in Gene Regula tory Networks.* In: JOURNAL OF COMPUTATIONAL SCIENCE, vol. 5 n. 3, pp. 332-344. - ISSN 1877-7503

[10] Kanehisa, Minoru, and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic acids research* 28.1 (2000): 27-30.

[11] Kelder, Thomas, et al. "WikiPathways: building research communities on biologi cal pathways." *Nucleic acids research* 40.D1 (2012): D1301-D1307.

[12] Croft, David, et al. "Reactome: a database of reactions, pathways and biological processes." *Nucleic acids research* (2010): gkq1018.

[13] Bäck, Thomas, David B. Fogel, and Zbigniew Michalewicz, eds. “Evolutionary computation 1: Basic algorithms and operators.” Vol. 1. CRC Press, 2000.