Ph.D. subject Inria-INSERM

Title: Dynamic gene network inference for cancer genetic program discovery

Location: INRIA Nancy Grand Est research center --- IECL Laboratory, Vandoeuvre-lès-Nancy, France

Supervision

Nicolas Champagnat (nicolas.champagnat@inria.fr), Pierre Vallois (pierre.vallois@univ-lorraine.fr). IECL, Université de Lorraine and Inria Nancy – Grand Est

Scientific context

In most cancers, the accumulation of genetic and epigenetic aberrations progressively alters the behavior of the cells. Various genetic programs sustain these aberrant behaviors. Statistical methods have been proposed to reverse-engineer the gene regulatory networks underlying these genetic programs. An important therapeutic motivation for this statistical inference is the search for the best genetic targets that need to be influenced in order to modulate the genetic program of proliferation of cancer cells.

In the specific case of chronic lymphocytic leukemia, the most frequent leukemia in adults and currently incurable, it is possible to experimentally induce the expression of the cancerous genetic program ex vivo. It is then possible to collect *temporal* gene expression and proteomic data, which can be used to infer the *dynamical* behavior of the gene regulatory network. This was done in Vallat et al. (2013) based on micro-array gene expression data.

Missions

The method used in Vallat et al. (2013) consists first in clustering the set of genes in order to classify temporally their activation, and second to infer a linear model of gene network taking into account the different classes and the temporal dependencies between genes. The validation of the inferred network and its predictions—typically regarding the effects of silencing a given genetic target on the gene network—are important questions that remain to be answered. This is the global objective of the Ph.D.

The difficulty of this problem lies in the form of data: although they are high dimensional because they give the expression of more than 20,000 genes, they usually only contain few patients and few time measurements. It is then unrealistic to expect a reliable estimation of the full gene network. Therefore, the validation of the inferred network and its predictions can be validated by experimental silencing of a given gene or a given set of genes. However, these are costly experiments, which need to be carefully planned.

The Ph.D. thesis will focus on these two issues: first, improve the reliability of the inferred gene network by jointly performing the clustering with the inference step and by inferring the interaction matrix of aggregates expressions of groups of genes instead of the full interaction matrix; second, build a theoretical framework for the validation of the inferred gene network to take into account experimental data on cells where one or several genes have been turned off .

Methods

Statistical methods of gene inference include information theoretic methods, Bayesian methods, learning methods and penalized regression methods (Allouche et al., 2013).

Among those, the last set of methods are the more suited to our problem. They usually take advantage of a parsimony assumption, for example using Lasso techniques, Dantzig selector, or group Lasso. These methods need to be combined with appropriate methods of computation of confidence scores, for example using bootstrapping. A starting point will be the CASCADE-R library (Jung et al., 2013).

Environment

The Ph.D. will take place in IECL under the supervision of Nicolas Champagnat and Pierre Vallois. The Institut Élie Cartan de Lorraine (IECL) is the laboratory of Mathematics of Université de Lorraine. The Probability and Statistics group, composed of more than 30 permanent members, is the largest one in east part of the France. Two Inria projects belong to this group: the first one called BIGS (Biology, Genetics and Statistics) works on statistics and stochastic modeling for Biology and Medicine; TOSCA (TO Simulate and CAlibrate stochastic models) is the second one, with field of research stochastic modeling, control and stochastic numerical methods. The Ph.D. will be co-supervised by Laurent Vallat (Inserm Strasbourg). The statistical work will be performed on the dynamic microarray dataset of gene expression of chronic lymphocytic leukemia cells used in Vallat et al. (2013) and on a recent dataset.

Bibliography

- T. Bastogne, R. Keinj, P. Vallois. Tumor growth modeling based on cell and tumor lifespans. Journal of Theoretical Biology 312, 76-86 (2012).
- T. Bastogne, J.-L. Marchand, S. Pinel, P. Vallois. A branching process model of heterogeneous DNA damages caused by radiotherapy in in vitro cell cultures. Mathematical Biosciences (2017).
- L. Vallat, C.A. Kemper, N. Jung, M. Maumy-Bertrand, F. Bertrand, N. Meyer, A. Porcheville, J.W. Fisher III, J.G. Gribben, S. Barham. Reverse-engineering the genetic circuity of a cancer cell with predicted intervention in chronic lymphocytic leukemia. PNAS 110(2), 459–464, 2013.
- D. Allouche, C. Cierco-Ayrolles, S. de Givry, G. Guillermin, B. Mangin, T. Schiex, J. Vandel, M. Vignes. A panel of learning methods for the reconstruction of gene regulatory networks in a systems genetics context. In: A. de la Fuente (ed.), Gene Network Inference: Verification of Methods for Systems Genetics Data, Springer-Verlag Berlin, 2013.
- N. Jung, F. Bertrand, S. Bahram, L.Vallat, M. Maumy-Bertrand. Cascade: a R-package to study, predict and simulate the diffusion of a signal through a temporal gene network. Bioinformatics, 30(4), 571-3, 2014.

Skills and profile

Required qualification: Master in applied mathematics. Specific knowledge on stochastic modeling and statistics are required. A strong interest in biological applications is also important.

Funding

Application for Ph.D. funding from the Inria-INSERM Ph.D. program. Please contact the supervisors to prepare the application file. Deadline: Tuesday **27th June.**