

Model construction and selection for biological networks: use of domain knowledge and application to networks disturbed in diseases

Supervised by Malika SMAIL-TABBONE and Taha BOUKHOBZA
(Malika.Smail@loria.fr ; Taha.Boukhobza@univ-lorraine.fr)

This PhD thesis project was awarded with half-funding from the Charles Hermite Research Federation (FCH). The rest will come from the Region Grand Est.

Hosting Laboratories: LORIA and CRAN

Beginning of the thesis: September 2018 (Possibility of teaching activity)

Documents provided to apply (prior to an interview):

- Detailed CV
- Motivation letter
- Copy of diplomas and statements obtained
- Master thesis (or equivalent) or a description of the work in progress
- Recommendation letters can be sent directly by the signatory to the supervisors

Scientific Framework :

Biological systems are very complex compared to human-made systems. Developing a dynamic model of a cell in its entirety is still utopian today. However, the understanding of the pathways or networks¹ operating in a cell for the regulation of gene expression or for signalling events makes it possible to better understand the phenomena, especially those that lead to a disease. The control theory vision seems relevant to analyse the structure of biological systems because it consists of breaking down a complex system into a set of subsystems with good local properties, and then studying *a posteriori* with global properties due to the connection of these subsystems. Much work has been done in recent years on the construction and simulation of biological pathways from experimental data [10, 15]. Various formalisms [8,19] have been proposed to model these complex biological systems: Boolean networks, Bayesian networks, Petri nets, ordinary differential equations that can give nonlinear or linear time-varying models [11,17] or systems of stochastic equations [4]. With each formalism one is more or less able to express the specific characteristics of a particular type of pathway (signalling, regulation, or metabolism). Once the formalism has been chosen, a modelling approach can be used to build a model (or models) based on experimental data. However each model must be validated before it can be used for simulation or prediction.

In addition to the need to convert one formalism into another, some studies focus on the integration of different types of biological pathways (as in a cell for example) where each network is modelled in a specific formalism [14, 17]. It should be noted that differential equations constitute a generic formalism that makes it possible to build from experimental data, signalling networks, regulation networks as well as metabolic networks. These equations can be represented in the form of graphs where one can make a structural or topological analysis, allowing for example the ability to estimate the degree or the force of coupling / decoupling of sub-networks, to determine the number of points of stability, the subdivision and the hierarchy of the networks etc. When applied to gene expression, this type of analysis should lead to the characterization of existing regulations between genes, allowing us to answer in a generic way the problems of direct and reverse control: If one acts on a set of genes, what will be the consequences? If we want to modify the expression of a set of genes, which actions will make it possible? Various control strategies, whose

¹We consider here that the words *pathway* and *network* as synonyms.

objective is to intervene on the control of networks in order to avoid undesirable states of cells or to force the network to converge towards a desired state, have been proposed by taking inspiration from the optimal control theory [1].

The complexity of modelling biological networks should not make us ignore the existence of large amounts of data and annotations in many public biological databases. Indeed, it is now possible not only to exploit a wide variety of results of past biological experiments, but also to access and use both annotations and already established models [4]. Once resources for a given problem have been identified, a KDD (Knowledge Discovery from Databases) [7] process can be implemented to derive from these resources the knowledge needed to solve a problem. Recent years have seen the development of open and linked data (LOD) especially in the field of life sciences. Such data are represented in semantic web languages (RDF, RDFS) and described with minimal semantics, facilitating their integration into Ontology Web Language (OWL) knowledge bases. It is then possible to organize the data along with formalization of domain knowledge and to apply inference mechanisms in problem solving or decision support scenarios

Relevance, originality and objectives

This thesis project is motivated by two obstacles which hinders the current modelling approaches. Firstly, it is difficult to construct a complete descriptive model of a biological network when data is incomplete or uncertain [5]. We propose to introduce the notion of *oriented model*, meaning that we seek to build a model oriented by specific modelling objectives corresponding to carefully identified phenomena, possibly in the form of a set of protagonists and known parameters (genes, proteins, molecules, situations, disease, environment, treatment ...) for which experimental observation data are available.

Secondly, building many candidate models from a set of experimental data requires manual analysis by biologists for selecting the most promising. Examples of interesting work include model checking techniques for the validation of properties of interest from biologists' point of view in complex networks [16] or appropriate evaluation methods [13].

The aim of the PhD thesis is to formalize and evaluate the notion of oriented models with known methods for building biological networks and to design and test mechanisms for model reduction or selection in an automated way guided by formalized knowledge, using semantic web languages with formal semantics such as RDF (S), OWL 2 EL, OWL2 QL, and OWL2 RL.

This doctoral project is both interdisciplinary and ambitious. It should lead the student to acquire dual expertise in data & knowledge management construction and the construction and analysis of quantitative models from experimental data.

Application framework:

This project can be applied to the study of various types of regulatory and signalling networks of interest for cancers. In the continuity of work already undertaken, it will be possible to begin by modelling the known receptor regulatory networks to better understand those that specifically involve an oestrogen receptor variant described as a factor of bad prognosis but whose regulation and activity remain not well described. The methodology developed to construct and validate oriented models of the genetic regulation network of oestrogen receptors can also be tested with data on the mineralocorticoid receptor (MR), in the case of heart failure in the context of the Hospital Research Project "Fight Heart Failure" in which two LORIA teams are involved.

References

- [1] Bouaynaya N, Sheterenberg R and Schonfeld D (2012) Methods for optimal intervention in gene regulatory networks. *IEEE Signal Processing Magazine*, 29(1): 158-163.
- [2] Boukhobza T (2008) « Analyse structurelle des propriétés d'observabilité et de diagnosticabilité des systèmes linéaires et bilinéaires – Approche graphique ». Habilitation à diriger des recherches, Université Henri Poincaré.
- [3] Datta A, Pal R, Choudhary A, Dougherty E (2007) What approaches have been developed for addressing the issue of intervention? *IEEE Signal Processing Magazine*, 24(1): 54-63.
- [4] Devignes M-D and Smaïl-Tabbone M (2009) « Maîtriser les ressources numériques, biologie *in silico* ». In *Biologie l'Ere Numérique*, Editions du CNRS, Magali Roux ed., pp 189-222.
- [5] De Jong H, Ropers D (2006) Strategies for dealing with incomplete information in the modeling of molecular interaction networks. *Brief Bioinform.*, 7(4):354-63.
- [6] De Jong H, Gouzé J, Hernandez C, Page M, Sari T, Geiselmann J (2004) Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bull Math Biol* 66(2):301-340.
- [7] Fayyad U, Piatetsky-Shapiro G, Smyth P (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.
- [8] Fey D, Kuehn A, Kholodenko B.N (2016) On the personalised modelling of cancer signalling, *IFAC-PapersOnLine*, 49 (26),312.
- [10] González-Vargas A.M, Cinquemani E, Ferrari-Trecate G (2016). Validation methods for population models of gene expression dynamics, *IFAC-PapersOnLine*, 49(26),114.
- [11] Halter W, Tuza Z.A, Allgöwer F (2017) Signal differentiation with genetic networks. *IFAC-PapersOnLine*, 50 (1), 10938.
- [13] Kiani NA, Zenil H, Olczak J, Tegnér J (2016) Evaluating network inference methods in terms of their ability to preserve the topology and complexity of genetic networks. *Semin Cell Dev Biol*, 51:44-52.
- [14] Machado D, Costa R, Rocha M, Ferreira E, Tidor B and Roch I (2011) Modeling formalisms in Systems Biology. *AMB Express*: 1-45.
- [15] Medley J.K, Goldberg A.P, Karr J.R (2016). Guidelines for reproducibly building and simulating systems biology models. *IEEE Transactions on Biomedical Engineering* 63 (10).
- [16] Monteiro P.T, Abou-Jaoudé W, Thieffry D, Chaouiya C (2014). Model Checking Logical Regulatory Networks, *IFAC Proceedings*, Volume 47 (2), 170.
- [17] Otero-Muras I, Banga J.R (2016) Exploring Design Principles of Gene Regulatory Networks via Pareto Optimality. *IFAC-PapersOnLine*, 49 (7).809.
- [18] Smaïl-Tabbone M (2014). Contributions à l'extraction de connaissances à partir de données biologiques. Habilitation à diriger des recherches, Université de Lorraine.
- [19] Sugavanewaran L (2017). Mathematical Modeling of Gene Networks. *Reference Module in Biomedical Sciences*.