Nantes Université

HABILITATION À DIRIGER DES RECHERCHES
HDR

NANTES UNIVERSITE

Spécialité : INFO

Par
Carito GUZIOLOWSKI

Modeling Biological Networks as Logic Programs

Travaux présentés et soutenus à Nantes, le 25/01/2024
Unité de recherche : Laboratoire des Sciences du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Elisabeth REMY	Directrice de Recherche, CNRS, Université d'Aix Marseille
Pedro MONTEIRO	Associate Professor, University of Lisboa, Portugal
Mohamed ELATE	Professeur, Université de Lille

Composition du Jury :

Président :	Prénom Nom	Fonction et établissement d'exercice *(à préciser après la soutenance)*
Examinateurs :	Jérémie BOURDON	Professeur, Nantes Université
	Damien EVEILLARD	Professeur, Nantes Université
	Marie-France SAGOT	Directrice de Recherche, INRIA, Université Claude Bernard, Lyon 1
	Anne SIEGEL	Directrice de Recherche, IRISA, Université de Rennes 1

Nantes Université

Titre : Modélisation de réseaux biologiques à l'aide des programmes logiques

Mots clés : Programmation logique, réseaux de régulation, modélisation de systèmes biologiques

Résumé : Dans ce manuscrit nous explorons deux représentations d'un système biologique en utilisant des modélisations informatiques. Les résultats de ces deux représentations ont été diffusés et valorisés au travers de publications scientifiques méthodologiques et applicatives; et pour certains d'entre eux, au travers de projets de recherche en étroite collaboration avec des biologistes.

Notre première contribution a été dans la modélisation par la *consistance des signes*. Ici, un réseau de régulation (graphe orienté et signé) est confronté à des données expérimentales contenant un ensemble d'observations des gènes du système dans deux conditions différentes. Cette confrontation a été implémentée dans un programme logique écrit dans le paradigme de la Programmation par Ensemble Réponses. Ce programme détecte une mesure de consistance entre le graphe et le jeu de données expérimentales, et propose des corrections automatiques minimales dans le cas d'inconsistance. Une fois que la consistance est établie pour le système, une liste de déductions peut être énumérée, ces déductions logiques seront appelées les prédictions du système.
Notre outil se nomme Iggy et utilise le solveur *clasp*. Nous avons appliqué cette modélisation à plusieurs systèmes biologiques, notamment chez l'humain. Nous détaillons, dans ce manuscrit, son application dans la modélisation du Myélome Multiple, qui nous a permis de mettre en lumière des espèces clés dans le réseau de régulation. Les prédictions de ces espèces clés nous ont permit de discriminer des patients ayant une meilleur survie. Dans une autre étude, nous avons étendu Iggy pour développer un nouveau système, MajS, dont la finalité est de faire le lien entre modélisation d'un réseau des gènes et métabolisme.

Une deuxième contribution a été dans l'apprentissage des familles des réseau booléens (RBn). Ce formalisme consiste à apprendre de familles de RBs à partir d'une connaissance préliminaire de régulation (ou *prior knowledge network*, PKN) en la confrontant avec des données d'expression (de gènes ou protéines) issues de multiples perturbations expérimentales. Les familles de RBs seront optimales car elles auront une taille minimale et qu'elles expliqueront de façon optimale les observations obtenues à travers les multiples perturbations. Le premier système conçu a été caspo, également implémenté avec des programmes logiques.Une extension de caspo a été conduite pour proposer des expériences de perturbations pour réduire la taille de la famille des RBs apprise. Plus tard, nous avons proposé caspo-ts qui peut assimiler des séries temporelles des données et qui propose des RBs dynamiques. caspo-ts a été appliqué sur les données d'une défi internationale, nommé le *HPN-DREAM challenge*, appliqué à des lignées cellulaires du cancer du sein. L'objectif était de déterminer des mécanismes de régulations ou des fonctions booléennes différentes qui s'expriment dans des lignées cellulaires. En parallèle de ces travaux, nous avons proposé une méthode (basé sur la programmation logique), pour générer de données de *pseudo-perturbations* à partir de données expérimentales pour des systèmes où il n'est pas possible de réaliser des perturbations pour des raisons éthiques. Cette méthode a été utilisée pour analyser des données issues de patients ayant développé une Leucémie Myeloïde Aiguë, et pour discriminer les patients ayant une meilleure réponse au traitement employé. Nous sommes actuellement en train d'adapter cette méthode pour l'appliquer à des données de *single cell*, dans une étude du développement embryonnaire chez l'humain.

Title : Modeling Biological Networks as Logic Programs

Keywords : Logic programming, regulatory networks, biological systems modeling

In this manuscript it is proposed to explore two representations of a biological system using computational modeling. These representations both gave birth to several methodological publications, and in some cases research projects in close collaboration with biologists.

One is done through the *sign-consistency* modeling. In this approach a regulatory network (signed directed graph) is combined with a dataset of gene expression observations, using a logic program. This logic program, written in Answer Set Programming, expresses a rule that has to be valid for each species in the network, which relates the *sign* of a network species with its direct predecessors *influences and signs*. This rule is tested in a global way, through all the network species by using an efficient solver, clasp. The sign-consistency modeling framework we proposed is named Iggy. Iggy performs as well automatic and optimal correction of sign inconsistencies. The sign-consistency modeling framework has been applied to different biological case studies. For example, the signaling pathway of Hepatocyte Growth Factor, where some of the computational predictions of our model were validated experimentally. A case-study well described in this manuscript is the modeling of Multiple Myeloma patients gene expression data. Our main results on this system was to propose Multiple Myeloma markers, that is, species in the network, coupled with our computational predictions, that allow to identify patients having a better survival. Iggy has inspired MajS, our last sign-consistency modeling framework contribution. In this on-going research project we plan to integrate gene regulatory and metabolic network modeling.

A second modeling approach is *learning Boolean network families*. In this framework, given a regulatory network (also called Prior Knowledge Network, PKN) and a set of network species observations upon multiple perturbations over the system, our framework learns a family of Boolean Networks (BNs), compatible with the PKN topology, that fits the perturbation data with minimal error. The first system we conceived is named caspo. It is also implemented using Answer Set Programming. An extension of caspo was implemented, so that new experimental designs (*i.e.* new experimental perturbations) can be proposed to decrease the number of learned BNs. Later, we proposed a system named caspo-ts, which deals with perturbation time-series data, and the output of this system is a family of dynamic BNs. caspo-ts has been applied to the data of HPN-DREAM challenge, concerning Breast cancer cell lines. Our objective was to identify the different BNs underlying the four Breast Cancer cell lines considered. In parallel, since multiple perturbation data, essential for caspo or caspo-ts, is sometimes hard to obtain in Human systems because of ethical reasons; we have begun a research subject towards the extraction of multiple pseudo-perturbations from non perturbed datasets, such as proteomics or RNA-Seq datasets. This method has been applied to discriminate Acute Myeloid Leukemia patients having different treatment prognosis. Currently, we are exploring to extract multiple pseudo-perturbations from single cell data, in the study of Human embryo development.