Construction et Sélection de modèles pour les réseaux biologiques :
Utilisation des connaissances du domaine et application aux réseaux perturbés dans les pathologies

Thèse proposée par Malika SMAIL-TABBONE et Taha BOUKHOBZA
(Malika.Smail@loria.fr ; Taha.Boukhobza@univ-lorraine.fr)

Cette offre de thèse a obtenu un demi-financement par la Fédération de recherche Charles Hermite(FCH), l'autre moitié du financement viendra de la région Grand-Est.
Laboratoires d'accueil : LORIA et CRAN
Début de la thèse : Septembre 2018 (Possibilité d'effectuer des heures d'enseignement)
Documents à fournir pour candidater (préalablement à un entretien) :

- CV détaillé
- Lettre de motivation
- Copie des diplômes et relevés des notes obtenues
- Mémoire de Master (ou équivalent) ou une description des travaux en cours
- Au moins une lettre de recommandation à adresser directement par la/le signataire aux encadrants de la thèse.

Contexte scientifique

Les systèmes biologiques sont très complexes comparés aux systèmes conçus par l'Homme. Développer un modèle dynamique de la cellule dans sa totalité reste utopique à ce jour. Toutefois, la compréhension des réseaux opérant dans la cellule pour la *régulation* de l'expression des gènes ou pour la *signalisation* permet de mieux cerner les phénomènes qui conduisent à une maladie.

La vision automatique semble pertinente pour analyser la structure des systèmes biologiques car elle consiste d'une part à décomposer un système complexe en un ensemble de sous-systèmes possédant de bonnes propriétés locales et d'autre part à étudier *a posteriori* les propriétés globales résultant de la connexion de ces sous-systèmes. De nombreux travaux ont été consacrés ces dernières années à la construction et à la simulation de réseaux biologiques à partir de données expérimentales [10, 15]. Divers formalismes [8,19] ont été proposés pour modéliser ces systèmes biologiques complexes : réseaux booléens, réseaux Bayésiens, réseaux de Pétri, équations différentielles ordinaires pouvant donner des modèles non linéaires ou linéaires variant dans le temps [11,17] ou encore des systèmes d'équations stochastiques [4]. Chaque formalisme est plus ou moins apte à exprimer les caractéristiques spécifiques d'un type particulier de réseaux (de signalisation, de régulation, ou métabolique). Une fois le formalisme choisi, une approche de modélisation permet de construire un (ou des) modèle(s) à partir de données expérimentales. Le modèle retenu doit être validé avant d'être utilisé en simulation ou en prédiction.

En plus de la nécessité de convertir un formalisme dans un autre, des études portent sur l'intégration de différents types de réseaux biologiques (tels que cela se présente dans une cellule par exemple) où chaque réseau est modélisé dans un formalisme propre [14, 17]. Notons que les équations différentielles constituent un formalisme générique qui permet de construire, à partir de données expérimentales, aussi bien des réseaux de signalisation, de régulation que des réseaux métaboliques. Ces équations sont représentables sous forme de graphes sur lesquels on peut faire une analyse structurelle ou topologique qui permet par exemple d'estimer le degré ou la force de couplage/découplage de sous-réseaux, de déterminer le nombre de points de stabilité, la subdivision

et la hiérarchisation des réseaux… Appliqué à la régulation de l'expression génétique, ce type d'analyse devrait conduire à la caractérisation des régulations existant entre gènes et permettre de répondre de façon générique aux problèmes de contrôle direct et inverse : Si l'on agit sur cet ensemble de gènes, quelles en seront les conséquences ? Si l'on souhaite modifier l'expression d'un ensemble de gènes, quelles sont les actions qui permettent de le faire ? Diverses stratégies de contrôle, dont l'objectif est d'intervenir sur les réseaux de régulation afin d'éviter des états indésirables de la cellule ou de forcer le réseau à converger vers un état désiré, ont été proposées en s'inspirant de la théorie de la commande optimale [1].

La complexité de ces approches pour modéliser les réseaux biologiques ne doit pas occulter l'existence de quantités très importantes de données et d'annotations dans les bases de données biologiques. En effet, il est aujourd'hui possible non seulement d'exploiter une grande variété de résultats d'expériences biologiques passées, mais aussi d'accéder et d'utiliser des annotations et des modèles déjà décrits [4]. Une fois que l'on a identifié les ressources nécessaires pour un problème donné, un processus de KDD (« *Knowledge Discovery from Databases* ») [7] peut être mis en œuvre pour tirer de ces ressources les connaissances utiles pour la résolution du problème. Ces dernières années ont vu l'essor des données ouvertes et liées (LOD, *Linked Open Data*) en particulier dans le champ des sciences de la vie. Ces données sont représentées dans les langages du web sémantique (RDF, RDFS) et sont décrites avec une sémantique minimale, ce qui facilite leur intégration dans des bases de connaissances OWL (*Web Ontology Langage*). Il est alors possible d'organiser ces données selon une formalisation plus expressive des connaissances du domaine et d'appliquer des mécanismes d'inférence au service de la résolution de problèmes ou d'aide à la décision.

Pertinence, originalité et objectifs

Cette proposition de thèse est motivée par deux obstacles sur lesquelles butent les approches actuelles de modélisation des réseaux biologiques. Le premier est qu'il est difficile de construire un modèle descriptif complet d'un réseau biologique lorsque les données sont incomplètes ou incertaines [5]. Nous proposons d'introduire la notion de *modèle orienté* qui correspond au fait que nous cherchons à construire un modèle orienté par l'objectif spécifique de la modélisation d'un nombre de phénomènes identifiés, et qui peut se présenter sous forme d'un ensemble des protagonistes et de paramètres connus (gènes, protéines, molécules, situations, pathologie, environnement, traitement…) pour lesquels on dispose de données d'observation expérimentales.

Le second obstacle réside dans le fait qu'il est possible de construire de nombreux modèles candidats à partir d'un ensemble de données expérimentales. Une analyse manuelle par des biologistes semble alors nécessaire afin de choisir le modèle qui semble le plus prometteur par rapport à leur expertise souvent fondée sur une excellente connaissance de la littérature dans un domaine assez circonscrit. Des exemples de travaux intéressants font appel aux techniques de *model checking* pour la validation de propriétés intéressantes du point de vue des biologistes dans des réseaux complexes [16] ou encore à des méthodes d'évaluation idoines [13].

L'objectif de la thèse est donc de formaliser et évaluer la notion de modèle orienté –avec certaines méthodes de construction de réseaux biologiques- et de concevoir et tester des mécanismes de réduction ou de sélection de modèles de façon automatisée et guidée par les connaissances formalisées à l'aide des langages du web sémantique dotés d'une sémantique formelle que sont RDF(S), OWL 2 EL, OWL2 QL, OWL2 RL.

Ce projet doctoral est à la fois interdisciplinaire et ambitieux. Il permettra d'acquérir une double expertise dans la gestion des données et des connaissances et dans la construction de modèles quantitatifs et orientés à partir de données expérimentales ainsi que l'analyse des propriétés structurelles de ces modèles.

Contextes applicatifs

Le présent projet peut s'appliquer à l'étude des divers types de réseaux de régulation et de signalisation intéressants pour les cancers. Dans la continuité de travaux déjà engagés, il sera possible de commencer par modéliser les réseaux de régulation des récepteurs connus pour mieux comprendre ceux qui impliquent spécifiquement un variant du récepteur aux estrogènes décrit comme facteur de mauvais pronostic mais dont la régulation et l'activité restent peu décrites.

La méthodologie développée pour construire et valider des modèles orientés du réseau de régulation génétique des récepteurs aux estrogènes pourra également être testée sur les données relatives au récepteur des mineralocorticoïdes (MR) dans le cas de l'insuffisance cardiaque et du projet de Recherche Hospitalo-Universitaire *Fight Heart Failure* dans lequel deux équipes du LORIA sont impliquées.

Références bibliographiques

[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] Sugavaneswaran L (2017). Mathematical Modeling of Gene Networks. *Reference Module in Biomedical Sciences*.