Athénaïs VAGINAY
(01/10/18 - 30/09/21)

Directeurs de thèse : Taha Boukhobza (CRAN) Malika Smaïl-Tabbone (LORIA)

Titre : Sélection et analyse de modèles pour les réseaux biologiques : utilisation des connaissances du domaine et application aux réseaux perturbés dans les pathologies

Mots-clés : sélection de modèles, base de connaissances, réseaux bayésiens, bases de données biologiques, analyse toplogique/structurelle de graphes, SBML/cellML/mathML, OWL, SWRL

I- Etat de l'art

Les systèmes biologiques sont très complexes comparés aux systèmes conçus par l'Homme. Plus de 10 000 variables d'état sont nécessaires pour décrire des organismes simples comme des bactéries alors que quelques dizaines suffisent à modéliser un avion ou une transmission automatique intégrant un moteur thermique ou électrique. Développer un modèle dynamique de la cellule dans sa totalité reste utopique à ce jour. 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 (de Jong 2002). De nombreux formalismes (Machado et al. 2011, Pal et al. 2012) ont été utilisés 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 (Kabir 2010) ou encore des systèmes d'équations stochastiques (Chasagnole et al. 2002, De Jong et al. 2004). Une approche récente préconise de construire une base de règles expertes permettant de produire (par déduction) un modèle de réseau à partir de données expérimentales (Aslaoui-Errafi 2012). Le pouvoir d'expression important de cette approche est contrebalancé par l'effort de construction de la base de règles.

Ainsi, il existe plusieurs formalismes, chacun étant plus ou moins apte à exprimer les caractéristiques d'un type particulier de réseaux (de signalisation, de régulation, ou métabolique). Une fois le formalisme choisi, une approche de modélisation permet d'inférer 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 actuellement 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 (Machado et al. 2011). Notons à ce sujet 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 (Datta et al. 2007, Qian et al. 2009, Vahedi et al. 2008, Bouaynaya et al. 2012, Kabir et al. 2010).

La complexité de ces approches pour modéliser les réseaux biologiques ne doit pas faire oublier l'existence de quantités très importantes de données et d'annotations dans les bases de données biologiques. En effet, depuis l'entrée de la biologie dans l'ère numérique, 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 (parties de) modèles déjà décrits (Devignes et Smail-Tabbone, 2009). L'essor des bases de données en biologie est tel que nous avons jugé utile de proposer des outils pour leur organisation et leur recherche (Devignes et al., 2010). Une fois identifiées les ressources nécessaires pour un problème donné, un processus de KDD (« Knowledge Discovery from Databases », Fayyad et al., 1996) peut être mis en œuvre pour tirer de ces ressources les connaissances utiles pour la résolution du problème (Smaïl-Tabbone, 2014). Ces dernières années ont vu l'essor du mouvement 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 exposées 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 d'intérêt et d'appliquer des mécanismes d'inférence à des fins de résolution de problème ou d'aide à la décision.

II- Problématique de la thèse

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. Nous proposons donc 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, lequel peut se présenter sous forme d'un ensemble des protagonistes et de paramètres connus (gènes, protéines, molécules, situations…) 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 et fastidieuse par des biologistes est 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.

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.

III- Eléments méthodologiques et ressources utiles

Diverses méthodes d'inférence de modèles dynamiques de réseaux de régulation ou de signalisation à partir de différents types de données expérimentales dynamiques ont été proposées depuis quelques années (Wang et al. 2006, Shamaiah et al. 2012). Une fois les données expérimentales prétraitées et la structure de modèle choisie, des algorithmes d'optimisation classiques (méthodes des gradients, algorithmes génétiques, programmation linéaire ou quadratique, sont en général utilisées) permettent d'instancier le modèle. Les difficultés majeures restent le nombre de données expérimentales et la non unicité des instanciations de modèles obtenus (problème d'optimisation non convexe). Aussi, la construction de modèles à partir de données expérimentales n'est plus un challenge en soi ou ne le sera bientôt plus. En revanche, la question est le plus souvent la sélection du meilleur modèle puis son utilisation en vue de la prédiction, du contrôle ou du diagnostic.

Par ailleurs, les données disponibles dans les sources publiques couvrent un large spectre qui va des données expérimentales (ex : le niveau de transcription des gènes dans des multitudes de situations dans la base GEO, les interactions physiques entres protéines dans les bases Intact , BIND, DIP…) à des modèles complets de réseaux de gènes (KEGG, Reactome, …) en passant par diverses données d'annotation (ex : fonction des gènes dans la base UniProt , associations gène-maladie dans la base OMIM …). Des initiatives internationales concertées ont conduit à la construction de ressources à grande valeur ajoutée pour décrire les modèles biologiques publiés dans la littérature. La base de données BioModels[1] en est le prototype et fait l'objet de processus d'annotation et de vérification très soignés qui font appel à des vocabulaires contrôlés mais aussi à des formalismes structurés standard de description de réseaux biologiques tels que SBML[2], CellML[3] (Li et al., 2010).

Une analyse structurelle du modèle obtenu permettra ensuite de caractériser les réseaux étudiés : importance de certains groupes de gènes, stabilité, résolution de certains problèmes liés à la régulation génique. Les outils d'analyse structurelle utilisés se fondent essentiellement sur la théorie des graphes permettent de dégager des propriétés structurelles importantes (Boukhobza, 2008).

Nous postulons qu'il est désormais incontournable de s'appuyer sur l'ensemble des données et des connaissances disponibles dans les ressources publiques afin de construire une base de connaissances (BC) qui serait capable de fournir des éléments pour aider à la construction de modèles correspondant à un réseau biologique mais aussi à la sélection de modèles candidats ou à leur classement. Nous proposons donc d'exploiter la BC, pour (i) en amont, établir un ensemble de contraintes sur la base desquelles peut se faire la construction des modèles orientés candidats et (ii) en aval, apporter des éléments en faveur (respectivement en défaveur) de tel ou tel modèle orienté.

[1] http://www.ebi.ac.uk/biomodels
[2] Systems Biology Markup Language (http://sbml.org)
[3] Cell Markup Language (http://cellml.org)

L'approche que nous proposons d'explorer est pertinente lorsque les biologistes cherchent à modéliser des réseaux biologiques dans des situations pathologiques telles que la survenue d'un cancer ou d'une insuffisance cardiaque. Dans ce cas, il est peu vraisemblable que le modèle visé soit présent dans les bases de modèles. L'approche pourra être validée sur quelques réseaux liés à certains types de cancer connus (publiés dans la littérature et pour lesquels des données expérimentales sont disponibles) que nous tenterons de reconstruire par composition et agrégation d'éléments trouvés dans les réseaux présents dans les bases de modèles précitées.

Ce sujet de thèse est ambitieux et interdisciplinaire et nécessite clairement une expertise pour la gestion des données symboliques et des connaissances (côté 27ème section du CNU, LORIA) et une expertise (côté 61ème section du CNU, CRAN) pour la construction de modèles quantitatifs et orientés à partir de données expérimentales mais aussi pour l'analyse des propriétés structurelles de ces modèles avec une vision système qui permettrait d'aborder des questions essentielles de thérapie ou de diagnostic/pronostic en les traduisant en problèmes de contrôle direct ou inverse. De plus, chaque partenaire apporte au projet un atout supplémentaire, d'une part à travers l'expérience de l'équipe Orpailleur dans la formalisation des connaissances et l'application du processus de KDD à des données biologiques, d'autre part à travers la présence au CRAN de biologistes spécialistes de certains cancers et concernés par les applications du projet.

IV- Etat de collaboration entre les deux équipes et positionnement du projet

Les équipes des deux encadrants ont collaboré entre 2012 et 2015 dans le cadre d'un projet financé par l'Agence Nationale de Sécurité Sanitaire (ANSES) intitulé *Validation de ERα36 comme marqueur prédictif de susceptibilité aux nonylphénols in vivo et in vitro* (NONYLER36) porté par Hélène Dumond. Ce projet avait pour objectif de valider *in vitro* et *in vivo* la fonction du gène ERα36 comme marqueur de susceptibilité aux nonylphénols et de modéliser les mécanismes de perturbation endocrinienne engendrée par les nonylphénols qui dépendent de ERα36 (Chamard-Jovenin et al., 2017).

La demande de thèse fait suite cette collaboration et se propose d'étendre l'étude à divers types de réseaux de régulation et de signalisation pouvant s'avérer intéressants pour le cancer. Ainsi, il sera intéressant de modéliser les réseaux de régulation des récepteurs connus tels que ERα66 et ERβ pour mieux comprendre celui qui implique le récepteur ERα36.

Un second cadre applicatif des travaux de cette thèse est celui du projet RHU[4] *Fight Heart Failure* dans lequel deux équipes du LORIA sont impliquées. La méthodologie développée pour construire et valider des modèles orientés du réseau de régulation génétique des récepteurs aux estrogènes pourra être testée sur les données relatives au récepteur des mineralocorticoïdes (MR) dans le cas de l'insuffisance cardiaque.

Au-delà de ces scénarios applicatifs, l'objectif du projet est d'arriver à une approche de modélisation alliant les avantages des méthodes numériques et symboliques pour améliorer la pertinence des modèles et qui soit généralisable à d'autres réseaux biologiques voire à d'autres phénomènes biologiques complexes.

VI- Plan de réalisation de la thèse

[4] Recherche Hospitalo-Universitaire

Voici une esquisse de plan de la thèse :

- Octobre 2018 – Février 2019 : Etude de l'état de l'art sur les différents types de modèles/données et leur exploitation
- Mars 2019 – Septembre 2019 : Analyse des différentes méthodes d'agrégation de modèles et de données
- Octobre 2019-Février 2020 : Proposition et mise en œuvre d'une approche de modélisation alliant numérique et symbolique. Test sur les données relatives aux récepteurs ERα66 et ERβ.
- Mars 2020 – Février 2021 : Application de l'approche aux réseaux liés aux tumeurs mammaires et/ou au récepteur des mineralocorticoïdes. Scénarios d'utilisation des modèles et validation.
- Mars 2021 – Septembre 2021 : Rédaction de la thèse et soutenance

Des publications pourront être réalisées sur les aspects méthodologiques et sur les aspects applicatifs.

Références biobliographiques

Aslaoui-Errafi E, Cohen-Boulakia S, Froidevaux C, Gloaguen P, Poupon A, Rougny A and Yahiaoui M (2012) "Towards a logic-based method to infer provenance-aware molecular networks". In Proc. of the 1st ECML/PKDD International workshop on Learning and Discovery in Symbolic Systems Biology:103-110.

Bouaynaya N, Sheterenberg R and Schonfeld D (2012) "Methods for optimal intervention in gene regulatory networks", IEEE Signal Processing Magazine 29(1): 158-163.

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.

Büchel F, Wrzodek C, Mittag F, Dräger A, Eichner J, Rodriguez N, Le Novère N and Zell A. (2012) "Qualitative translation of relations from BioPAX to SBML qual". Bioinformatics. 28(20): 2648-53.

Chamard-Jovenin C, Thiebaut C, Chesnel A, Bresso E, Morel C, Smaïl-Tabbone M, Devignes MD, Boukhobza T, Helene Dumond H (2017) "Low-dose alkylphenol exposure promotes mammary epithelium alterations and transgenerational developmental defects, but does not enhance tumorigenic behaviour of breast cancer cells". , Frontiers in Endocrinology, 23 October 2017.

Chassagnole C, Noisommit-Rizzi N, Schmid J, Mauch K and Reuss M. (2002) "Dynamic modeling of the central carbon metabolism of Escherichia coli". Biotechnol. Bioeng. 79(1): 53-73.

Datta A, Pal R, Choudhary A and Dougherty E. (2007) "What approaches have been developed for addressing the issue of intervention? ". IEEE Signal Processin Magazine. 24(1): 54-63.

Devignes MD, Franiatte P, Messai N, Bresso E, Napoli A and Smaïl-Tabbone M (2010) "BioRegistry: Automatic extraction of metadata for biological database retrieval and discovery". Int. Journal on Metadata Semantics and Ontologies 5 : 184-193.

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.

De Jong H, Gouzé J, Hernandez C, Page M, Sari T and Geiselmann J (2004) "Qualitative simulation of genetic regulatory networks using piecewise-linear models". Bull Math Biol 66(2):301-340.

De Jong H (2002) "Modeling and simulation of genetic regulatory systems: a literature review". Journal of Computational Biology; 9(1):67-103.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Habibi M, Eslahchi C and Wong L (2010) "Protein complex prediction based on k-connected subgraphs in protein interaction network". *BMC Systems Biology* 2010, **4**:129.

Kabir M, Noman N and Iba H(2010) "Reverse engineering gene regulatory network from microarray data using linear time-variant model". *BMC Bioinformatics*, 201; 11(Suppl 1):S56.

Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, Li L, He E,Henry A, Stefan MI, Snoep JL, Hucka M, Le Novère N, Laibe C. (2010) BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst Biol. 4:92.

Machado D, Costa R, Rocha M, Ferreira E, Tidor B and Roch I (2011) “Modeling formalisms in Systems Biology”, AMB Express: 1-45.

Pal R, Bhattacharya S and Caglar M U (2012) “Robust approaches for genetic regulatory network modeling and intervention”, IEEE Signal Processing Magazine 29(1): 66-76.

Qian X, Ivanov I, Ghaffari N and Dougherty E (2009) “Intervention in gene regulatory networks via greedy control policies based on long-run behavior”. BMC Syst Biol 3(1): 61-77.

Shamaiah M, Lee S H and Vikalo H (2012) “Graphical models and inference on graph genomics”, IEEE Signal Processing Magazine 29(1): 51-66.

Schulz M, Krause F, Le Novère N, Klipp E, Liebermeister W. (2011) Retrieval,alignment and clustering of computational models based on semantic annotations. Mol Syst Biol. 7:512.

Smaïl-Tabbone M (2014). Contributions à l'extraction de connaissances à partir de données biologiques. Apprentissage. Thèse HDR, Université de Lorraine, 2014.

Vahedi G, Faryabi B, Chamberland J, Datta A and Dougherty E (2008) “Intervention in gene regulatory networks via stationary mean-first-passage-time control policy”. IEEE Trans Biomed Eng 55(10) : 2319-2331.

Wang Y, Joshi T, Zhang X-S, Xu D and Chen L (2006) “Inferring gene regulatory networks from multiple microarray datasets”. Bioinformatics 22(19) : 2413-2420.