

Gilles Bernot
Professeur à l'Université Nice Sophia Antipolis
Université Côte d'Azur, CNRS
Laboratoire I3S, UMR 7271
Algorithmes-Euclide-B
2000, route des Lucioles
CS 40121
06903 Sophia Antipolis CEDEX
France

Tél : +33 4 92 94 27 76
Mail : bernot@unice.fr
<http://www.i3s.unice.fr/~bernot>

Sophia Antipolis, le 24 avril 2019.

RAPPORT

sur la thèse de doctorat en informatique de Monsieur Xinwei Chai

ayant pour sujet :

Reachability Analysis and Revision of Dynamics of Biological Regulatory Networks.

Dans les grandes lignes l'objectif de cette thèse est la mise au point (aussi) automatique (que possible) de modèles de réseaux de régulation génétiques pour que leur dynamique réponde à des propriétés d'atteignabilité ou de non atteignabilité d'ensembles d'états donnés du système. Pour ce faire, elle s'appuie sur, et améliore, bon nombre d'outils existants, allant de l'analyse de conditions nécessaires (resp. suffisantes) pour atteindre un état donné d'une variable, à des techniques d'apprentissage automatique.

L'introduction (**Chapitre 1**, 5 pages) donne le contexte de la thèse : modélisation des systèmes dynamiques que sont les réseaux de régulation biologiques, en insistant sur la modélisation discrète et asynchrone classique. Elle survole également les limitations des approches classiques (Model Checking et apprentissage automatique de transitions) et présente le focus de la thèse.

Le **Chapitre 2** (17 pages) présente l'état de l'art et introduit les bases nécessaires à la suite de la thèse. Il commence par un survol très rapide de la modélisation par ODE, et présente également très rapidement la modélisation discrète, les stratégies de mise à jour, les réseaux de Thomas, les réseaux Bayésiens : cette partie manque nettement de clarté. Le cas particulier des réseaux booléens est présenté de manière plus détaillée et plus claire, ainsi que la programmation logique, le process hitting et leur extension aux réseaux asynchrones. Les aspects plus fins à propos de la dynamique sont factorisés entre ces diverses approches, avec les questions classiques de synchronicité et les différentes variantes de stratégies de mise à jour : cette factorisation est judicieuse. Le Model Checking est ensuite survolé, puis une description intuitive du mode de fonctionnement de Pint et un éclairage rapide sur les questions d'atteignabilité sont fournis.

Enfin l'apprentissage et la révision de modèle sont abordés. Le côté « apprentissage » couvre dans ce panorama toute technique assistée par ordinateur d'identification des paramètres. Il y manque éventuellement certains aspects du doctorat d'Aurélien Rizk qui avait défini il y a une dizaine d'année, dans le cadre de BIOCHAM, un « degré de satisfaction » de formules temporelles qui pourrait avoir des liens avec les coefficients de corrélation du chapitre 4. Enfin le côté « révision de modèle » est résolument orienté vers les questions d'atteignabilité.

Le **Chapitre 3** (25 pages) décrit l'analyse d'atteignabilité par heuristiques. En fait les concepts présentés dans ce chapitre présenteront également de l'intérêt pour les chapitres suivants. Les définitions se suivent dans un ordre logique, même si les commentaires manquent parfois de clarté. Le concept de base de ABAN (réseau d'automates booléens avec mise à jour asynchrone) est sur le fond une définition « en extension » des ressources nécessaires à chaque transition de l'automate booléen correspondant. L'atteignabilité d'un état dit « partiel » (i.e. restreint à l'état d'un sous-ensemble

des automates) est définie de manière naturelle. Une première proposition simple (mais fortement utile par la suite) permet de réduire cette atteignabilité à celle d'*un seul* automate (il suffit que ses ressources soient justement l'ensemble des états partiels à atteindre).

La notion centrale de SLCG (Simplified Local Causality Graph) est ensuite définie et encode de manière judicieuse les conditions qui sont nécessaires pour atteindre l'état cible. Cet encodage ne définit qu'une « pseudo-atteignabilité » car la concurrence et l'asynchronicité peuvent rendre inexistant l'ordre des transitions à effectuer dans l'ABAN pour réaliser, le long d'un chemin, les conditions nécessaires encodées dans le SLCG. Quelques résultats permettent de simplifier les SLCG (suppression des cycles contenant l'état à atteindre, ou ne contenant aucune sortie). Je me suis convaincu de la correction de ces énoncés bien que les preuves, telles qu'elles sont rédigées, manquent de clarté.

Pour aller plus loin dans la simplification de SLCG, une heuristique « violente » est adoptée : elle consiste à effectuer un choix aléatoire chaque fois que le SLCG offre plusieurs solutions pour atteindre l'état cible (ce qui supprime les portes OU du SLCG). Les SLCG purement conjonctifs obtenus sont bien sûr pour la plupart irréalisables (au sens précédent) et l'idée est de jouer sur les permutations pour rétablir partiellement l'atteignabilité. L'algorithme *PermReach* qui implémente cette idée est évidemment d'une complexité énorme en raison du nombre de permutations mais réduit *notablement* l'ensemble des cas inconclusifs par rapport à l'analyse statique des SLCG de l'état de l'art courant.

Dans le cas où le SLCG à traiter est purement conjonctif et sans cycles (mais rappelons que la suppression de cycles précédente n'est que partielle), il est alors proposé d'utiliser ASP pour encoder l'atteignabilité le long du SLCG. L'algorithme correspondant, *ASPR* *Reach*, est assez fin et exploite proprement les idées précédentes. Les algorithmes *PermReach* et *ASPR* *Reach* sont fournis en annexes ainsi que la preuve de correction des réponses (*False*, *True* ou *Inconclusive*) de ces algorithmes.

Enfin une extension au cas multivalué est décrite mais elle impose une restriction forte de monotonie des variations à partir de l'état initial, ce qui de mon point de vue est rédhibitoire pour la biologie puisqu'elle interdit des comportements oscillatoires. Mais cela n'enlève rien, naturellement, à la pertinence du cas booléen.

Le **Chapitre 4** (33 pages) aborde l'inférence et la révision de modèles. Il s'agit d'un domaine encore presque vierge et l'approche décrite est originale et motivante. Elle étend judicieusement les méthodes partiellement conclusives du chapitre précédent. Le problème de la complétion de modèles ABAN est défini sans suppression de transitions puis avec une extension des « cut sets » de Paulevé & al. Des stratégies de sur- et sous-approximation reposant sur un ensemble de R de « transitions candidates » sont décrites et des exemples pédagogiques permettent de comprendre les définitions quelque peu touffues.

Il est alors proposé d'utiliser des méthodes statistiques pour inférer l'ensemble R précité à partir de données expérimentales quantitatives. La méthode commence par remplacer les ODE, utilisées classiquement pour modéliser les réseaux de régulation, par des équations différentielles *avec délai*. On connaît toutes les difficultés de résolution de telles équations différentielles mais en revanche leur expression est considérablement plus simple. Néanmoins les justifications données pour atteindre des équations différentielles avec délai de forme encore plus simple sont erronées : les termes ignorés dans la forme avec délais par rapport aux ODE classiques ne sont pas liés à des autorégulations comme affirmé, mais au niveau basal des gènes et à la dégradation des protéines. C'est heureux d'ailleurs car s'ils s'agissait d'autorégulations, il est bien connu que de telles boucles ne seraient pas négligeables, puisqu'elles peuvent engendrer des changements de configuration notables de l'espace des états. Bref, la simplification proposée est sans doute raisonnable (d'autant plus qu'elle préserve justement les autorégulations puisqu'une variable peut encore appartenir à l'ensemble de ses prédécesseurs) mais pas pour les raisons données dans le manuscrit. L'énorme avantage de cette nouvelle forme avec délais est de faciliter la définition de coefficients de corrélation : ces derniers sont soigneusement définis et la méthode appliquée, qui regroupe de nombreuses étapes successives, est bien détaillée et donne lieu à l'algorithme appelé *CRAC*.

Enfin la révision de modèle, à partir de données cette fois discrétisées, est abordée. Il s'agit essentiel-

lement d'élargir la méthode dite « LFIT » à une sémantique asynchrone, bien adaptée aux réseaux de régulation. L'approche se formalise à partir d'une logique propositionnelle et un algorithme de révision *M2RIT* est développé pas à pas, en passant par un encodage sous forme de programme du SLCG. L'avantage de cette approche par révision est de ne plus nécessiter l'ensemble *R* de transitions candidates.


Le **Chapitre 5** (9 pages) évalue la mise en œuvre des outils précédents sur des exemples. *PermReach* et *ASPR* sont comparés aux outils existants avec un mode opératoire bien décrit, en laissant les détails d'implémentation en annexe. On constate sans surprise que PermReach est plus conclusif que Pint, et l'est moins que ASPReach. Tous sont évidemment moins conclusifs qu'un Model Checker classique. . . lorsque ce dernier peut conclure sans tomber en memory-out. Ces évaluations sont donc tout à fait concluantes (modulo les quelques hypothèses faites sur le SLCG) et montrent bien le gain sur la taille des problèmes abordables.

En l'absence d'outils concurrents *CRAC* est testé sur des exemples aléatoirement engendrés auxquels on a retiré 20% des transitions. Là encore le mode opératoire est sérieux et bien décrit et les résultats montrent également une capacité de « rasoir d'Ockham » de CRAC qui ne redécouvre que des transitions indispensables pour l'atteignabilité. *M2RIT* est testé de manière similaire. Cette fois les trajectoires discrètes sont engendrées par des parcours aléatoires des chemins du modèle d'origine. M2RIT fournit des modèles éventuellement différents du modèle initial caché, mais équivalents du point de vue de l'atteignabilité. La grande taille des exemples traités par CRAC et M2RIT est de mon point de vue une particularité remarquable.

Les conclusion et perspectives (**Chapitre 6**, 5 pages) se contentent de restituer le contexte « des données aux modèles », de récapituler les contributions, et fournissent ensuite quelques pistes d'améliorations techniques suggérées par les limitations déjà mentionnées dans la thèse.

En bref, ce travail est sans conteste original et l'apport est clairement du niveau d'une thèse. La rédaction (en anglais) manque parfois de clarté et de recul sur les aspects formels, mais il y a un réel effort pédagogique au travers d'exemples bien choisis.

En conséquence, je déclare que la thèse de Monsieur Xinwei Chai mérite d'être soutenue.



Gilles Bernot, Professeur d'informatique à Polytech'Nice-Sophia,
Directeur du Conseil Scientifique & Pédagogique de l'Ecole Universitaire de Recherche DS4H
Chercheur au laboratoire I3S, équipe SPARKS.