

# AtlanSTIC

Réponse Appel à projets 2006

## VÉRIFICATION FORMELLE DE PROPRIÉTÉS POUR UN SYSTÈME DYNAMIQUE STOCHASTIQUE

### 1 Intervenants :

#### 1.1 Coordonnateurs :

Olivier Roux†, IRCCyN, équipe Modélisation et Vérification des Systèmes embarqués (MoVeS)  
Christine Sinoquet‡, Lina, équipe Combinatoire et Bio-Informatique (ComBi)

†Olivier.Roux@irccyn.ec-nantes.fr

‡Christine.Sinoquet@univ-nantes.fr

#### 1.2 Participants :

Jérémie Bourdon, ComBi Jeremie.Bourdon@univ-nantes.fr

Didier Lime, MoVeS Didier.Lime@irccyn.ec-nantes.fr

Jamil Ahmad, MoVeS Jamil.Ahmad@irccyn.ec-nantes.fr

### 2 Objectifs :

#### 2.1 Motivation biologique :

L'expression d'un gène – ou synthèse protéique – comporte les deux étapes de la transcription et de la traduction : grâce à l'intervention d'une molécule d'ARN polymérase, le segment d'ADN correspondant au gène est transcrit en une molécule d'ARN messenger ; cette molécule d'ARN messenger est ensuite traduite en la protéine codée par le gène. L'expression génique est un processus complexe susceptible de régulations à divers instants de cette synthèse. Une régulation peut consister en une activation ou une inhibition. Comme les protéines remplissant cette fonction régulatrice sont produites par des gènes, il y a lieu de parler de réseau de régulation génique, structuré par un réseau d'interactions entre molécules d'ADN, d'ARN, de protéines et d'autres molécules de plus petite taille, les métabolites.

La production à grande échelle de données issues de l'analyse du transcriptome, du protéome et du métabolome a fortement accéléré l'essor de travaux destinés à modéliser la dynamique du vivant. Une modélisation appropriée rend possibles des simulations, dont sont attendues la vérification de propriétés, l'identification des étapes sensibles ou au contraire robustes de la dynamique des systèmes analysés. Lorsque le système concerné est un réseau de régulation de gènes, tout travail prospectif peut s'appuyer sur un socle à trois composantes, avec retours d'expériences et affinages successifs du modèle : (i) apprentissage du modèle par observations massivement parallèles, spatio-temporelles, du niveau d'expression des gènes (puces à ADN, à oligonucléotides...), (ii) modélisation des réactions ou interférences entre molécules, (iii) simulation. Pour vérifier certaines propriétés du réseau, une alternative à la simulation peut être exploitée : la vérification formelle de propriétés.

Parmi les questions auxquelles biologistes et biochimistes peuvent souhaiter voir apporter une réponse figurent les suivantes : Quels sont les états par lesquels peut transiter le système étudié ? Quels sont les états stables, instables ? En combien de temps en moyenne les atteint-on à partir d'un état initial donné ? Quels sont les divers chemins possibles, à partir d'un état initial donné ? A partir d'un état, peut-on atteindre un autre état ? Si oui, en combien de temps en moyenne ? Peut-on identifier des classes de comportements du système, reliées à des caractéristiques données du système ? Par exemple, lorsque l'état initial appartient à telle classe, il est certain que le système passe par tel état, ou encore lorsque l'état initial appartient à telle classe, il est certain que le système aboutit à tel état.

Une modélisation au moyen d'un graphe permet déjà d'inférer certaines informations sur le réseau de régulation modélisé. Par exemple, l'absence de chemin reliant deux gènes qu'on sait par ailleurs être en interférence met en évidence un défaut du modèle (interactions manquantes). L'identification de chemins multiples reliant deux gènes informe sur les redondances du système biologique étudié. La présence de cycles peut révéler des boucles de rétro-action (feedback).

Les questions précédentes correspondent à des modèles simples. Dès lors que l'on introduit des paramètres destinés à une description plus fine du système biologique, par exemple la notion de niveau d'expression d'un gène (passage du qualitatif au quantitatif), celle de délai de transition d'un état à un autre, ou encore la notion d'alternative, associée à une probabilité de choix d'alternative, d'autres questions apparaissent. Par exemple, on peut se poser la question d'identifier, si possible, les domaines de valeurs pour les paramètres du système qui induisent des comportements similaires.

Il est important de pouvoir vérifier des propriétés sur un système biologique sans recourir à la simulation d'un nombre de cas d'autant plus élevé que le modèle est complexe. Ainsi apparaît-il essentiel de disposer d'une part d'un modèle décrivant de manière suffisamment fine la réalité biologique, et d'être capable d'autre part de traduire ce modèle à l'aide d'un formalisme, basé sur des automates ou des processus de type markovien en l'occurrence. Le formalisme est destiné à permettre la vérification de propriétés, au moyen d'un raisonnement basé sur une théorie. Parmi ce dernier type de raisonnements, figure le raisonnement symbolique, sur lequel s'appuient des outils de vérification de propriétés dans les systèmes.

#### 2.2 Quelques repères en modélisation de système biologique :

De nombreuses compilations présentent les formalismes adoptés pour modéliser les réseaux de régulation géniques [6, 7, 12, 20, 17].

La modélisation par réseau booléen [14, 21, 15] associe à chaque gène une variable désignant son état (actif/inactif). A l’instant  $t + 1$ , l’état d’un gène est modélisé comme une fonction booléenne des états à l’instant  $t$ , des gènes activateurs ou inhibiteurs dont il dépend. Dans de nombreux cas, le type de modèle précédent est trop restrictif puisqu’il est approximatif (il ne rend pas compte des niveaux d’expression intermédiaires d’un gène) et est basé sur l’hypothèse de temps synchrone.

Palliant les deux problèmes précédents, les équations différentielles représentent le formalisme le plus couramment utilisé pour décrire *quantitativement* la cinétique d’un système dynamique. La cinétique de la régulation d’un réseau de gènes est alors modélisée par un ensemble d’équations exprimant le taux de production de chaque composant du système (ARN, protéine, autre molécule) comme une fonction des concentrations des autres composants. Mais la non-linéarité des fonctions utilisées rend ce type de formalisme réfractaire à l’analyse mathématique. Le seul recours qui reste pour identifier des propriétés du système étudié est alors la simulation numérique. De telles simulations, opérées en variant divers paramètres, permettent de mettre en relation des comportements divers du système étudié et des états physiologiques connus de la cellule, ainsi que des états non encore observés expérimentalement. Cependant, *in vivo* comme *in vitro*, la difficulté à obtenir des mesures permettant d’étalonner ce type de modèle limite leur utilisation à quelques cas très bien étudiés.

Pour tenter d’identifier des classes de comportement (ou trajectoire) du système modélisé, correspondant à des domaines de valeurs prises par les paramètres du modèle, il est intéressant d’adopter une approche intermédiaire entre la modélisation par variables à valeurs continues et la modélisation booléenne. Ce type de modélisation permet d’échapper à la nécessité de mesures fines et de s’en tenir à des niveaux d’expression de gènes. Notamment, l’approche de Thomas [22] modélise les niveaux d’expression d’un gène  $g$  au moyen d’une variable logique  $x_i$ , et repose sur le concept d’attracteur. La variable discrète  $x_i$  associée à un gène  $g$  prend ses valeurs dans un ensemble fini d’entiers  $V_i = \{0, 1, 2, \dots\}$ . Ces valeurs correspondent à autant de seuils d’expression du gène  $g$ . La variation de  $x_i$  est déterminée par la donnée (i) d’un ensemble de gènes activateurs et inhibiteurs du gène  $g$ , (ii) des seuils à partir desquels ces gènes sont respectivement activateurs et inhibiteurs pour le gène  $g$ , (iii) du niveau d’expression  $K$  vers lequel tend (est “attiré”) le gène  $g$  sous l’influence de ces seuls gènes. L’ensemble de ces données définit un “attracteur”. Dans cette logique, en l’absence de tout autre changement dans le système biologique, la valeur  $x_i$  est décrétementée ou incrémentée, au cours du temps, jusqu’à atteindre  $K$ . Sur le graphe de transitions du système, il est alors possible de détecter des états, en particulier des états stables et aussi des comportements (ou trajectoires) qui seraient passés inaperçus avec un formalisme à variables booléennes.

## 2.3 Vérification formelle de propriétés sur un réseau de régulation :

Sur la base du modèle de Thomas, l’équipe MoVeS a d’abord proposé une sémantique formelle pour décrire la régulation par attracteurs, dans un modèle asynchrone [3]. Le vérificateur de propriétés HyTech [13] a permis ensuite l’identification des états possibles du système, des états stables et instables (circuits), pour quelques cas, à partir de la seule description des attracteurs et d’un état initial du système. HyTech est un outil de vérification de propriétés pour automates hybrides, ou automates à comportement contrôlé par des modifications discrètes et continues. HyTech utilise donc une logique temporelle : une caractéristique de HyTech est sa capacité à conduire une analyse paramétrique, c’est-à-dire à déterminer les valeurs des paramètres du modèle pour lesquelles un automate satisfait à une spécification exprimée en logique temporelle. Comme HyTech considère des automates hybrides linéaires, la vérification symbolique est conduite automatiquement par calcul sur les polyèdres associés aux diverses contraintes mises en oeuvre. Plus récemment, un raffinement du modèle a consisté à associer un délai à chaque transition [1]. Ces délais sont liés aux temps de transcription, traduction et diffusion dans la cellule. Les systèmes biologiques ont alors été décrits à l’aide d’automates temporisés, classe particulière des automates hybrides linéaires. Un automate temporisé est un automate étendu avec des horloges (variables dont la valeur réelle positive augmente uniformément avec le temps) et dont les noeuds et arcs sont étiquetés avec des contraintes d’horloge, par exemple “( $x \geq 5$ ) et ( $y < 1$ )”, appelées invariants et gardes. Les invariants spécifient les conditions sous lesquelles un automate peut rester dans un état, laissant le temps s’écouler. Les gardes indiquent l’instant où une transition peut être empruntée. Les transitions sont instantanées et induisent parfois une remise à l’heure des horloges du type  $x := 0$ . Ainsi, l’utilisation de HyTech a permis de générer automatiquement, pour un chemin donné du réseau, la condition temporelle (sur les délais) qui permet d’emprunter et de suivre ce chemin.

## 2.4 Objectifs du projet de recherche

Pour répondre à un besoin de modélisation de système biologique non-déterministe, nous souhaitons étendre l’étude de la faisabilité de la vérification formelle de propriétés à des systèmes temporisés stochastiques. En effet, les mécanismes de régulation de gènes mis en oeuvre chez les bactéries sont caractérisés par de faibles concentrations intra-cellulaires ainsi que des vitesses de réaction relativement lentes. Les concentrations sont faibles car la majorité des molécules régulatrices sont produites en faibles quantités dans chaque cellule [11] et la plupart des gènes (et donc leur région régulatrice, site de fixation des molécules régulatrices) sont présents en un ou deux exemplaires dans chaque cellule. Dans ces conditions, la cinétique usuellement décrite, basée sur une évolution continue et déterministe des concentrations n’est plus applicable. Les modifications apportées dans un tel système interviennent donc en nombre entiers de molécules, et sont la conséquence d’événements aléatoires (réaction, fixation). Il a été montré que l’évolution d’un tel système est un processus stochastique de type markovien [9, 23]. Ainsi, selon le trajet qu’un métabolite donné empruntera dans la cellule, il pourra provoquer une réaction plutôt qu’une autre (bifurcation des voies métaboliques) [2, 18]. Or, si des algorithmes de simulation stochastique permettent de connaître les divers comportements possibles du système [10], nous souhaitons au contraire étudier une piste basée sur la vérification formelle, qu’elle soit symbolique (“model checking”) ou théorique (processus de type markovien), de réseaux de régulation biologiques stochastiques. Par ailleurs, sauf pour les systèmes à équations différentielles, il est impossible de modéliser l’apport de métabolites par des sources extérieures. Nous souhaitons examiner s’il est possible de tenir compte d’apports de métabolites, dont la variation de la concentration au cours du temps suit une loi probabiliste donnée.

# 3 Verrous scientifiques et technologiques

Une gageure consiste à identifier si un vérificateur formel de propriétés actuellement disponible permettrait de décrire et analyser des réseaux d’interaction de taille relativement élevée, en tenant compte de délais et de choix de nature stochastique.

La prise en compte d’apports extérieurs de métabolites au réseau de régulation génique peut être décrite par un modèle à équations différentielles. Or, ce dernier type de modèle est réservé à l’analyse de systèmes biologiques de cinétique très précisément connue. Doit donc être examinée la possibilité de vérifier des propriétés, de façon formelle, lorsqu’est pris en compte un apport extérieur stochastique.

## 4 Projet scientifique détaillé :

Le model checking probabiliste intègre l'analyse probabiliste et le model checking classique en un seul outil [16, 19]. Il constitue une alternative à la simulation et à l'approche analytique, la première approche étant souvent de complexité temporelle élevée et la seconde représentant le système à un niveau de description insuffisamment détaillé. Les model checkers usuels prennent en entrée (i) une description du modèle, représentant un système de transitions entre états, et (ii) une spécification, typiquement une formule exprimée dans une logique temporelle. Ces model checkers renvoient le résultat "oui" ou "non", selon que le modèle vérifie ou non la propriété. Dans le cas du model checking probabiliste, les modèles sont probabilistes (variantes des chaînes de Markov), en ce sens qu'ils codent la probabilité de réalisation d'une transition entre deux états donnés, plutôt que la seule existence d'une telle transition. Un espace de probabilités, induit sur les comportements du système, permet alors le calcul de la vraisemblance de l'occurrence de certains événements pendant l'activité du système. Ceci permet ensuite d'établir des propriétés quantitatives sur le système, en complément des propriétés qualitatives habituellement établies par les model checkers classiques.

Nous référant aux travaux déjà menés sur le model checking probabiliste, le model checking et les formalismes adaptés aux réseaux de régulation de gènes [5, 8], nous essaierons de proposer un formalisme étendu permettant de modéliser les réseaux d'interaction de gènes, en intégrant délais et choix probabilistes pour les transitions. Il ne semble pas que les modèles à réseaux bayésiens soient particulièrement bien adaptés à la vérification formelle de propriétés, en matière de réseaux de régulation de gènes, mais une étude plus poussée devra confirmer ou infirmer cette assertion.

En parallèle, un outil de vérification formelle, basé sur une étude probabiliste, sera élaboré, sur la base de travaux antérieurs [4] sur des processus apparentés aux chaînes de Markov. Nous étudierons dans quelle mesure ces précédents travaux peuvent intégrer simultanément délais et probabilités de choix. L'objectif est de pouvoir répondre à des questions du type : quelle est la durée moyenne pour atteindre un état donné ? La probabilité d'atteindre un état donné est-elle inférieure à 0.05 ? Les réponses à ces questions sont connues pour les chaînes de Markov. Pour le processus de type markovien en lequel nous envisageons de traduire un modèle biologique donné, nous apporterons des réponses à ce type de questions en calculant moyennes, variances et éventuellement distributions de probabilités limites des variables (aléatoires) d'intérêt.

## 5 Résultats attendus :

Après avoir identifié un model checker probabiliste, nous étudierons la faisabilité de la traduction automatique d'une description d'un réseau d'interactions avec délais et probabilités de choix vers le type d'automate analysé par ce model checker.

Parallèlement, nous basant sur les possibilités de vérification formelle reposant sur une étude probabiliste, nous développerons un programme informatique destiné à générer automatiquement les réponses aux questions suivantes, pour un réseau de régulation donné :

Après quelle durée, en moyenne, atteint-on pour la première fois un état donné ? Est-il vrai que la probabilité d'atteindre un état donné est au plus égale à 0.05 ? En une durée donnée, combien de fois en moyenne passe-t-on dans un état donné ? Quelle est la probabilité de passer  $k$  fois dans un état donné, en une durée donnée ? L'assertion suivante est-elle vraie ou fausse ? La probabilité pour que le niveau d'expression d'un gène donné atteigne un seuil donné, dans n'importe quel état final du système, est au moins égale à 0.85. Connaissant les lois régissant les apports extérieurs au système, quel est le niveau moyen d'expression d'un gène donné ? Peut-on identifier des configurations de valeurs de paramètres (probabilités de choix) qui conduisent aux mêmes comportements du réseau de régulation ? Quelle est la probabilité pour que deux alternatives possibles, dans un même réseau, conduisent en moyenne au même niveau d'expression d'un gène donné, pour le même état final ?

Cette approche probabiliste nécessitera le calcul de caractéristiques (valeurs propres, vecteurs propres) des matrices de transitions associées aux automates. Lors de l'implémentation informatique correspondante, les algorithmes appropriés seront mis œuvre afin de prendre en compte le fait que ces matrices peuvent être de taille importante.

## 6 Historique de la collaboration :

Dans le cadre de l'annonce d'un appel à projets programmé pour le printemps 2006, les équipes MoVeS et ComBi se sont rapprochées dès décembre 2005 afin d'examiner s'il était possible de fédérer les compétences de ces deux équipes. Spécialisée dans la spécification et la vérification des systèmes réactifs et temporisés, l'équipe MoVeS avait déjà étendu l'application de ses thématiques habituelles au cas de systèmes de régulation biologiques. Il est apparu une complémentarité évidente avec ComBi, spécialisée en combinatoire pour la bio-informatique, et amenée à adopter des approches probabilistes ou stochastiques pour échapper à une complexité trop élevée. Les deux équipes apportent notamment à la collaboration leurs compétences en matière d'étude de systèmes dynamiques, dans des registres différents (approche par logique temporelle, approche probabiliste). La seconde équipe a introduit l'aspect stochastique. Dès décembre 2005, comme la répartition des charges des uns et des autres était connue, le calendrier du montage scientifique du projet avait été établi, pour pallier l'impossibilité de tenir des réunions fréquentes durant le premier trimestre 2006. Un consensus sur les objectifs, le projet scientifique et son déroulement, les résultats attendus a été atteint rapidement. Pour mémoire, deux des intervenants ont déjà été amenés à travailler ensemble (Olivier Roux, Christine Sinoquet, Etude des liens entre les langages réactifs asynchrone et synchrone : application à Electre et SIGNAL).

Christine Sinoquet participe à l'école thématique "Modélisation de systèmes biologiques complexes dans le contexte de la génomique", organisé à Bordeaux, du 3 au 7 avril 2006 (<http://epigenomique.free.fr/programme.php>).

## Références

- [1] J. Ahmad, A. Richard, G. Bernot, J.-P. Comet, and O. Roux. Delays in biological regulatory networks (BRN). In *Proceedings of IWBRA 2006, LNCS*, page to appear, 2006.
- [2] A. Arkin, J. Ross, and H.H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *escherichia coli* cells. *Genetics*, 149(4) :1633–48, 1998.
- [3] G. Bernot, F. Cassez, J.\*P. Comet, F. Delaplace, C. Müller, O. Roux, and O. Roux. Semantics of biological networks. *Bio-CONCUR 2003*, 8888(8888) :8888, 2003.
- [4] J. Bourdon and B. Vallée. Pattern matching statistics on correlated sources. In *Proceedings of LATIN 2006*, volume 8888, pages 224–237, 2006.
- [5] F. Chabrier and F. Fages. Symbolic model checking of biochemical networks, 2003.

- [6] H. De Jong. Modeling and simulation of genetic regulatory systems : a literature review. *Journal of Computational Biology*, 9(1) :67–103, 2000.
- [7] D. Endy and R. Brent. Modelling cellular behavior. *Nature*, 409 :391–395, 2001.
- [8] F. Fages. Constraint-based model checking of non-deterministic hybrid systems : A first experiment in systems biology, 2004.
- [9] D.T. Gillespie. *Markov Processes : An introduction for Physical Scientists*. Academic Press, San Diego, 1992.
- [10] D.T. Gillespie. A rigorous derivation of the chemical master equation. *Phys. A*, 188 :404–425, 1992.
- [11] P. Guptasarma. Does replication-induced transcription regulate synthesis of the myriad low copy number proteins of escherichia coli? *BioEssays*, 17 :987–997, 1995.
- [12] J. Hastay, D. McMillen, F. Isaacs, and J.J. Collins. Computational studies of gene regulatory networks. *Nat. Rev. Genet.*, 2 :268–279, 2001.
- [13] T. Henzinger, P.-H. Ho, and H. Wong-Toi. HyTECH : A model checker for hybrid systems. *Software Tools for Technology Transfer*, 1 :110–122, 1997.
- [14] S. Huang. Gene expression profiling, genetic networks, and cellular states : An integrating concept for tumorigenesis and drug discovery. *J. Mol. Med.*, 77 :469–480, 1999.
- [15] S.A. Kaufmann. *The origins of Order : Self-Organization and Selection in Evolution*. Oxford University Press, New York, 1993.
- [16] M. Kwiatkowska. Model checking for probability and time : from theory to practice. In *Proceedings of the 18th Annual IEEE Symposium on Logic in Computer Science (LICS'03)*, lics, page 351, 2003.
- [17] H.H. McAdams and A. Arkin. Simulation of prokaryotic genetic circuits. *Ann. Rev. Biophys. Biomol. Struct.*, 27 :199–224, 1998.
- [18] C.J. Morton-Firth, T.S. Shimizu, and Bray D. A free-energy-based stochastic simulation of the tar receptor complex. *J. Mol. Biol.*, 286 :1059–1074, 1999.
- [19] D. Parker. Implementation of symbolic model checking for probabilistic system, 2002.
- [20] P. Smolen, D.A. Baxter, and Byrne J.H. Modeling transcriptional control in gene networks : Methods, recent results and future directions. *Bull. Math. Biol.*, 62 :247–292, 2000.
- [21] R. Somogyi and C.A. Sniegoski. Modeling the complexity of genetic networks : Understanding multigenic and pleiotropic regulation. *Complexity*, 1(6) :45–63, 1996.
- [22] R. Thomas, D. Thieffry, and M. Kaufman. Dynamical behaviour of biological regulatory networks : I. biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull. Math. Biol.*, 57(2) :247–276, 1995.
- [23] N.G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam, 1992.