Élucider le fonctionnement d'un réseau de régulation biologique par l'informatique

En modélisation mathématique, l'ordinateur effectue souvent de gros calculs de simulations. En biologie des systèmes, lorsque les interactions non linéaires rendent impossible la détermination du comportement général d'un système à partir de quelques simulations, même lorsque ce système ne contient qu'un petit nombre d'éléments, la logique formelle informatique peut assister la découverte de modèles et suggérer des expériences « à la paillasse ».

Gilles Bernot*, Jean-Paul Comet*,, Janine Guespin*****

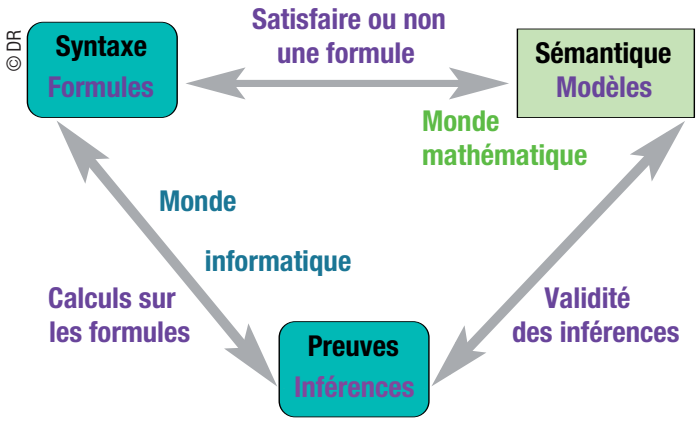
* Programme d'épigénomique, Génopole® et université d'Évry, 4 Bd François Mitterrand, 91025 Évry cedex
** IBISC, université d'Évry, 4 Bd F. Mitterrand, 91025 Évry cedex
*** Laboratoire de microbiologie du froid, université de Rouen, 76821 Mont Saint Aignan cedex

Le développement de la biologie moléculaire a produit de grandes quantités de données qu'il faut savoir exploiter. Cela a ouvert l'ère de la biologie des systèmes, dans laquelle la modélisation est une activité majeure. L'intrication des interactions entre molécules produit en effet un comportement global de la cellule imprévisible au vu de chaque interaction élémentaire. Augmenter la présence d'une molécule M qui agit *a priori* en faveur de l'expression d'un gène G peut parfois, paradoxalement, réduire l'activité de G en raison d'effets de rétroaction. Ainsi, prédire l'impact d'un médicament, l'intérêt de telle ou telle molécule, ou les conditions optimales d'obtention d'un phénotype ne relève pas de raisonnements de bon sens à partir d'un petit nombre d'expériences bien choisies. La modélisation mathématique est un passage obligé.

Effectuer exhaustivement *in vivo* ou *in vitro* les expériences permettant de s'assurer d'un rapport de cause à effet global est devenu totalement hors d'atteinte. Le succès de la biologie moléculaire se situe ailleurs : au niveau d'une énumération des interactions moléculaires élémentaires qui tend vers l'exhaustivité. Les modèles mathématiques servent alors à extraire des lois émergeant de la combinatoire des interactions, avec la difficulté des interactions non linéaires qui induisent une haute sensibilité des phénotypes à des variations minimes de paramètres. Mais la plupart de ces paramètres sont impossibles à mesurer avec une précision adéquate, à supposer qu'ils soient mesurables. Beaucoup d'entre eux sont même des paramètres purement théoriques ne correspondant à aucune réalité mesurable.

En regard de celui des expérimentations *in vivo* ou *in vitro*, le coût des expériences informatiques est quasi négligeable. Cela a conduit à une reconstruction du vivant « à tâtons », où les valeurs possibles des paramètres inconnus sont testées les unes après les autres, avec des études de robustesse et de stabilité des résultats pour certaines plages de valeurs de ces paramètres. De telles approches ont beaucoup apporté à la biologie des systèmes, en mettant l'accent sur les variables les plus pertinentes agissant sur le phénomène étudié, et en remettant en cause des idées reçues sur l'importance de tel ou tel objet biologique dans le système étudié.

1 – La logique formelle en informatique



Un ordinateur est plus qu'un « calculateur » (computer). C'est un manipulateur de symboles. On peut calculer en manipulant des symboles (chiffres) bien sûr, mais on peut de plus raisonner dans l'abstrait. La logique formelle en informatique distingue ainsi trois aspects de son champ disciplinaire :

- les suites de symboles qui signifient quelque chose pour le domaine considéré : c'est l'ensemble des « phrases » autorisées et que l'ordinateur acceptera et saura manipuler. La définition de cet ensemble, et les moyens d'en reconnaître les éléments appelés les formules, constituent la **syntaxe** ;
- les règles de manipulation des phrases autorisées (transformations et productions de formules) : ce sont les étapes élémentaires de raisonnement logique qui permettent de produire les conséquences d'un ensemble d'affirmations écrites par des formules. Il s'agit là d'une vision élargie de la notion de **calcul** que l'on nomme aussi l'inférence ;
- les deux précédentes étapes sont purement conventionnelles : on pourrait les définir selon des manipulations parfaitement aléatoires (de même que l'on pourrait inventer des tables de multiplication remplies aléatoirement). Il faut donc s'assurer que les règles d'inférence utilisées aient un sens cohérent. On le fait en donnant une **sémantique** mathématique aux formules, avec des modèles mathématiques qui peuvent satisfaire, ou non, les formules logiques.

Cependant, la recherche fondamentale reste sur sa faim à l'issue de ce processus : le modèle retenu dans ce processus d'essais-erreurs est-il le seul valide ? Ce modèle, qui a résisté à la comparaison avec les comportements connus lors des simulations précédentes, résistera-t-il à la prochaine simulation ?

Pour un réseau de régulation génétique, les expériences *in vivo*, qui permettent de calibrer le modèle mathématique, sont souvent des *knock-out* de certains gènes. Il n'est pas rare que le *knock-out* d'un nouveau gène conduise à reconsidérer les plages de paramètres précédemment supposées valides. On aimerait donc, à chaque instant, connaître exhaustivement les plages de paramètres qui restent en accord avec toutes les expériences connues.

Des environnements informatiques d'aide à la modélisation

Réduire le nombre de paramètres nécessaires à la modélisation est la première étape vers une validation exhaustive des modèles mathématiques. Depuis de nombreuses années, en informatique, l'abstraction est la clef pour manipuler des modèles validables (ou réfutables) de gros systèmes. Il s'agit de regrouper en un seul objet formel, aux propriétés bien établies, un ensemble compliqué de processus élémentaires. Ainsi, on peut raisonner globalement sur un nombre modeste d'objets abstraits, sans gérer les processus internes de chaque objet.

La théorie logique des réseaux de régulation de René Thomas (1) est un exemple du genre. La notion de gène régulateur est abstraite, jusqu'à cacher entièrement les mécanismes moléculaires de régulation : un niveau d'expression abstrait est affecté à chaque gène à un instant donné et les lois d'évolution des niveaux d'expression ne dépendent que de quelques paramètres qui peuvent prendre un nombre très restreint de valeurs entières. Seuls les comportements qualitatifs sont modélisés, de sorte que l'on peut faire un lien assez direct entre la nature des réseaux de régulation considérés et les phénotypes observés (états d'équilibre stables, comportements périodiques, différentiations, etc.). On peut alors considérer l'ensemble de tous les modèles reflétant des régulations connues ou supposées, en s'appuyant sur la théorie des systèmes de transitions assez bien étudiée en informatique.

Dominer le nombre de paramètres dont dépendent les modèles théoriques ouvre la porte à une méthode raisonnée de construction de modèles mathématiques. Il devient possible de caractériser exhaustivement les modèles potentiels du processus biologique étudié. On manipule à chaque instant non plus un modèle unique utilisé pour les simulations et adapté au fil de leurs résultats, mais l'ensemble des modèles parmi lesquels on choisira le modèle adéquat *in fine*. La question n'est plus de trouver à tâtons un modèle acceptable mais de restreindre petit à petit l'ensemble des modèles possibles.

Une définition suffisamment rigoureuse de l'espace des possibles permet de gérer par ordinateur l'ensemble des modèles potentiels tout au long du processus de modélisation. On peut ainsi mettre en place un environnement informatique dédié à la modélisation des réseaux de régulation biologique. C'est le cas du logiciel SMBioNet (2), qui réalise cette approche dans le cadre des réseaux de régulation selon la théorie de René Thomas.

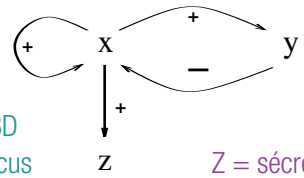
Pour outiller valablement la démarche de construction de modèles pour la biologie, on doit prendre en compte non seulement les modèles, mais aussi les connaissances biologiques issues des expériences et des observations, ainsi que les hypothèses scientifiques qui motivent l'étude biologique. Il s'agit de propriétés comportementales de l'objet ou de la fonction biologique étudié, propriétés connues ou hypothétiques. En science informatique, la manipulation de propriétés passe par des logiques formelles.

Les logiques temporelles

Chacun connaît la logique élémentaire qui repose essentiellement sur des tables de vérité. On y considère des affirmations élémentaires *A*, *B*, *C*, etc., qui peuvent être chacune soit vraie, soit fausse, et on construit des affirmations composées (formules) en combinant les affirmations élémentaires avec des connecteurs logiques

(1) Thomas R, d'Ari R (1990) *Biological feedback*, CRC press, Boston (http://hal.ccsd.cnrs.fr/ccsd-00087681)
(2) Bernot G *et al.* (2004) *JTB* 229 (3), 339-47

2 – Un exemple simple d'application : la découverte de *switch* épigénétiques chez *Pseudomonas aeruginosa* infectant des malades atteints de mucoviscidose



© DR

La cause principale de mortalité due à la mucoviscidose est l'infection des malades par la bactérie *P. aeruginosa*, qui y acquiert deux phénotypes nouveaux, la cytotoxicité (contre les défenses de l'hôte), et la mucoïdie (production d'un épais mucus et résistance aux antibiotiques).

Les réseaux de régulation contrôlant ces phénotypes présentent chacun un régulateur clef, impliqué dans deux circuits de rétroaction, l'un positif, l'autre négatif. De plus, les souches qui ont acquis ces phénotypes chez les malades les conservent *ex vivo*. Nous avons donc fait l'hypothèse que l'acquisition de chacun de ces phénotypes est due, non à une mutation, mais à un switch épigénétique entre deux états stationnaires possibles résultant du fonctionnement des réseaux de régulation **(4)**.

Cette hypothèse peut être formulée à la fois sous la forme d'un graphe simplifié du réseau de régulation et des formules CTL qui formalisent l'hypothèse des deux états stables. Dans les deux cas, l'étude du réseau de régulation des gènes impliqués montre une boucle de rétroaction positive et un circuit de rétroaction négatif. On peut tracer, en abstrayant au maximum, le même graphe formel de régulation (les variables X, Y et Z ayant une interprétation différente dans chaque cas). La boucle de régulation positive permet, en s'appuyant sur le théorème de multistationnarité de René Thomas, de poser l'hypothèse de la nature épigénétique de ces modifications.

Modèles et formules ont été testés grâce au logiciel SMBioNet **(2)**, qui a d'abord prouvé la cohérence de l'hypothèse (existence de modèles biologiquement cohérents présentant ces états stables) **(5)**. Cette modélisation a aussi permis d'établir qu'une seule expérience, consistant à fournir un pulse de la protéine clef, suffit pour prouver ou falsifier l'hypothèse. Cette expérience a été réalisée et l'hypothèse prouvée dans le cas de la cytotoxicité **(6)**.

Classiquement, pour se débarrasser de cette bactérie, on doit utiliser des antibiotiques qui, dans ce cas, sont très peu efficaces. Notre découverte devrait permettre d'envisager d'autres types de traitements.

comme le « *et* », le « *ou* », la négation (« *non* ») ou l'implication (« $\Rightarrow$ »). Par exemple, « (A ou B) $\Rightarrow A$ » est une formule, et en consultant les tables de vérité du « *ou* » et de l'implication, on peut montrer qu'elle est vraie sauf lorsque A est fausse et B est vraie **(encadré 1)**.

On peut ainsi conduire des raisonnements par ordinateur, par exemple tenter de savoir si B est vraie ou fausse en connaissant la véracité de la formule précédente et celle de A. En pratique, on gère des cas de figure où l'on sait qu'une formule est fausse (ou vraie) à la suite d'une expérience biologique qui la contredit (ou la valide), et où l'on connaît, par les conditions d'expérience, la véracité de certaines affirmations élémentaires ; on exploite ensuite les capacités de raisonnement pour en déduire d'autres affirmations inconnues jusqu'alors.

Les déductions utiles à la biologie sont cependant plus compliquées que la simple exploitation de tables de vérité, parce que les systèmes étudiés sont dynamiques et donc les propriétés intéressantes sont souvent temporelles. Cela conduit à faire appel à des logiques dites temporelles offrant, en plus des connecteurs logiques habituels, des symboles parlant du temps : « A sera vraie *la prochaine fois que...* », « *il est possible que plus tard A soit vraie* », « *il est nécessaire qu'un jour ou l'autre A soit vraie* », « *A sera vraie jusqu'à ce que B le devienne* », etc.

Ainsi les environnements informatiques d'aide à la modélisation manipulent deux ensembles de connaissances de natures distinctes :
- les modèles potentiels, qui sont souvent décrits *via* des graphes d'interactions ;
- les propriétés biologiques, connues ou hypothétiques, exprimées en logique temporelle.

Deux questions s'imposent alors :
- la cohérence à chaque étape de modélisation : existe-t-il au moins un des modèles potentiels qui satisfasse les propriétés biologiques ?
- la validation : ce n'est pas parce qu'il existe des modèles qui valident les connaissances et hypothèses biologiques que l'objet biologique *in vivo* correspond à l'un d'eux ; il pourrait correspondre au contraire à l'un des autres. C'est là qu'intervient le « retour à la paillasse » piloté par la modélisation.

La cohérence des modèles avec les connaissances et hypothèses biologiques

La difficulté à raisonner sur un système biologique complexe, principalement due aux interactions non linéaires et aux boucles de rétroaction, fait de l'ordinateur et de la logique formelle des outils indispensables pour vérifier la cohérence. Cette étape peut éviter des expériences coûteuses liées à de mauvaises questions.

L'informatique offre des techniques automatiques sophistiquées de vérification de cohérence des modèles avec un ensemble de formules temporelles (appelées *model checking* **(3)**, résolution de contraintes, produits d'automates, etc.). En pratique, il n'est pas rare que des incohérences soient soulevées au cours du processus de modélisation. Elles imposent alors une étroite collaboration entre chercheurs informaticiens et biologistes : connaissance ou hypothèse biologique maladroitement encodée en logique temporelle, graphe d'interactions incomplet, importance d'une interaction dans le graphe sous- ou sur-valuée.

Cette étape de mise au point de modèles par « simple » cohérence constitue en pratique un premier processus de découverte important en biologie des systèmes. Elle met le doigt sur les éléments clefs liés aux hypothèses biologiques étudiées.

La validation des hypothèses par des plans d'expérience calculés

Formaliser les hypothèses biologiques en formules de logique temporelle est un travail pluridisciplinaire exigeant, mais l'investissement est rentable car il confère aux propriétés affirmées une structure syntaxique qui peut être largement exploitée.

Supposons par exemple que la formule « (A ou B) $\Rightarrow C$ » soit une hypothèse sur un système biologique (A, B et C étant des propriétés biologiques, qui peuvent être vraies ou fausses selon l'état de la cellule). Cette formule est structurée en une prémisse « A ou B » et une conclusion « C », et l'on sait, par sa table de vérité, qu'une implication dont la prémisse est fausse est quant à elle toujours vraie. Comme l'a explicité Karl Popper, une expérience biologique n'aura d'intérêt pour cette question que si elle est apte à réfuter l'hypothèse, c'est-à-dire à rendre fausse la formule. Il faut donc que la prémisse soit vraie. Une simple exploitation des tables de vérité du « *ou* » permet à l'ordinateur d'indiquer que trois classes d'expériences sont à explorer : A vraie et B fausse, A fausse et B vraie, enfin A et B vraies. Dans

(3) Huth MR, Ryan MD (2000) *Logic in Computer Science: Modelling and Reasoning about Systems, Cambridge University Press* (www.cs.bham.ac.uk/research/projects/lics)
(4) Guespin-Michel J, Kaufman M (2001) *Acta Biotheoretica* 49 (4), 207-18
(5) Guespin-Michel J *et al.* (2004) *Acta Biotheoretica* 52, 379-90
(6) Filopon D *et al.* (2006) *BMC Bioinformatics* 7, 272-82

les trois cas, la table de vérité de l'implication nous dit qu'il faudra vérifier si C est vraie à l'issue de l'expérience.

Lorsque des symboles temporels sont à prendre en compte, la génération de classes d'expériences intéressantes est plus subtile, mais la technique utilisant la logique pour guider les expériences subsiste.

Il faut souligner enfin que la difficulté principale réside dans l'adéquation entre les capacités expérimentales et les propriétés élémentaires, comme *A* ou *B*, que l'ordinateur suggère de rendre alternativement vraies ou fausses. On doit définir préalablement les affirmations logiques qui peuvent être atteintes expérimentalement. Si par exemple *A* n'est pas contrôlable expérimentalement, alors il faut exploiter les capacités de preuve de l'ordinateur pour suggérer une ou plusieurs autres propriétés aptes à remplacer *A* et contrôlables expérimentalement. Le principe logique est de faire une preuve automatique indirecte de « (*A* ou *B*) ⇒ *C* » en s'appuyant sur des formules expérimentables *in vivo* **(encadré 2)**.

Des outils très actuels

Comprendre le vivant à partir des connaissances de la biologie moléculaire passe par une « reconstruction du vivant » *via* des modèles mathématiques. L'informatique joue alors un rôle qui dépasse la force brute de calcul des simulations : elle aide à raisonner sur les objets biologiques et aide à choisir les expériences in vivo optimales. Les réseaux de régulation biologiques sont ainsi fortement d'actualité en bioinformatique car ils bénéficient de lois de fonctionnement établies avec suffisamment de rigueur pour que l'ordinateur puisse y appliquer ces raisonnements complexes mais automatisés. ●

Optimiser un plan d'expérience à partir de modèles qualitatifs ?

La « biologie systémique » développe des méthodes pour interpréter et exploiter les données massives produites par l'observation d'une cellule. Elle vise plus précisément à comprendre le comportement qui résulte des réseaux d'interactions. Elle construit pour ce faire des modèles dynamiques dont les prédictions, obtenues par simulation, doivent être confrontées, via des expérimentations ciblées, aux données disponibles. Mais comment déterminer efficacement ces expérimentations ?

Anne Siegel*,, Carito Guziolowski*, Philippe Veber*, Olidiu Radulescu*, Michel Le Borgne***

* Programme d'épigénomique, Génopole® et université d'Évry, 4 Bd François Mitterrand, 91025 Évry cedex
** IBISC, université d'Évry, 4 Bd F. Mitterrand, 91025 Évry cedex
*** Laboratoire de microbiologie du froid, université de Rouen, 76821 Mont Saint Aignan cedex

Au sein d'un large spectre de formalismes, on distingue deux types d'approches pour construire des réseaux et étudier leur comportement. Il existe tout d'abord des méthodes quantitatives, dont les prédictions sont numériques et dépendent de la connaissance d'un grand nombre de paramètres. De manière complémentaire, on peut fait appel à des méthodes qualitatives, qui ne demandent pas de paramètres numériques, et dont les prédictions expriment des relations d'ordre ou de dépendance : une valeur est-elle plus grande qu'une autre ? Une valeur est-elle fonction d'une autre ? **(1, 2, 3)**.

Les méthodes quantitatives et qualitatives sont bien entendu reliées. Nous allons illustrer comment un problème quantitatif, tel que l'étude des variations des niveaux d'expression de gènes et les concentrations de protéines entre deux états d'une cellule, peut être traduit dans un modèle qualitatif qui prend en compte uniquement les signes de ces variations. Notre approche, inspirée par la « physique qualitative » de Kuipers **(4)** utilisée par exemple en cognition et en intelligence artificielle, s'adapte aux données souvent imprécises et relationnelles produites en masse par les techniques expérimentales en génomique. Il ne s'agit pas, comme suggéré dans une fameuse phrase de Rutherford, de faire du « pauvre quantitatif », mais de structurer et d'améliorer la fiabilité de nos connaissances sur des systèmes de complexité très grande.

Pour illustrer l'intérêt de cette démarche, nous nous plaçons dans la situation où un biologiste modifie la

(1) De Jong H *et al.* (2005) *Biofutur* 252, 36-40
(2) De Jong H (2002) *J Comput Biol* 9, 67-103
(3) Covert MW *et al.* (2004) *Nature* 429, 92-6
(4) Kuipers B (1994) *Qualitative reasoning*, MIT Press