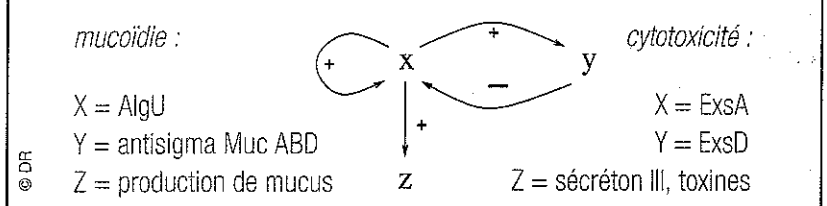
2 – Un exemple simple d'application : la découverte de *switch* épigénétiques chez *Pseudomonas aeruginosa* infectant des malades atteints de mucoviscidose



© 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 surévaluée.

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 $B) \Rightarrow C$ » 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*, Ovidiu Radulescu*, Michel Le Borgne***

* Irisa (CNRS, Inria, université de Rennes 1), Campus de Beaulieu, 35042 Rennes Cedex
** asiegel@irisa.fr

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 faire 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

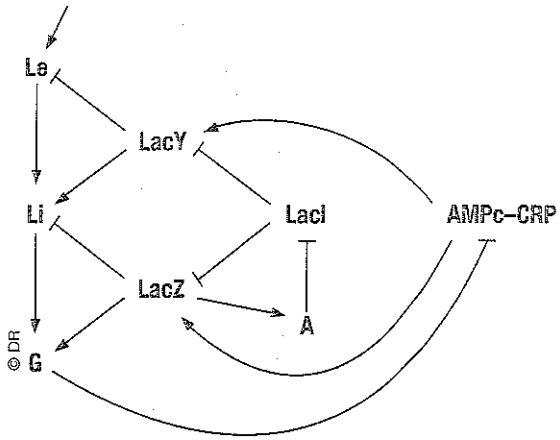


Figure 1 **Graphe d'interactions et contraintes qualitatives d'un modèle de l'opéron lactose**

Sur les flèches se terminant par ▶, la production du produit d'arrivée augmente. Sur celles se terminant par |, le produit de départ diminue. Les boucles de rétrorégulation négative ont été omises. En présence de lactose à l'extérieur de la cellule (**Le**), la perméase **LacY** transporte ce dernier à l'intérieur du milieu cellulaire (**Li**). Le lactose y est alors métabolisé, sous l'action de **LacZ**, pour produire du glucose (**G**) et de l'allolactose (**A**). Les protéines **LacY** et **LacZ** sont sous contrôle direct d'un inhibiteur appelé **LacI**. Ce dernier est lui-même régulé par la concentration de glucose, *via* l'action de l'AMP cyclique (**AMPc**). Cette régulation est utilisée pour inhiber la production des enzymes **LacY** et **LacZ** lorsque le glucose peut être acquis directement dans le milieu. Le lactose **Le** peut recevoir des influences extérieures non indiquées dans le réseau (nœud externe).

(5) Tyson JJ *et al.* (2001) *Nat Rev Mol Cell Biol* 2, 908-13

concentration d'une entrée d'un système initialement stable, et attend qu'il se stabilise à nouveau. On observe un déplacement d'équilibre sous l'effet d'une perturbation. Les techniques de production de données en masse renseignent sur ces déplacements d'équilibre, mais des observations se révèlent plus utiles que d'autres.

À partir d'une modélisation par graphe d'interactions, nous discutons et évaluons l'intérêt que présente l'observation d'un composant par rapport à un autre. On espère ainsi réduire considérablement le coût et augmenter l'efficacité des expérimentations futures.

Modélisation d'une expérimentation par des équations qualitatives

La représentation qualitative des connaissances est une caractéristique naturelle du fonctionnement du cerveau. Ainsi, les raisonnements qualitatifs relèvent tout simplement de ce qu'on appelle le bon sens et les équations qualitatives sont le résultat de la formalisation mathématique du bon sens. Cette formalisation se fera en plusieurs étapes décrites par la suite.

1 – Relations d'addition et de multiplication pour les signes *[+]*, *[-]*, *[?]*

[+] + [-] = [?]	[+] + [+] = [+]	[-] + [-] = [-]
[?] + [-] = [?]	[?] + [+] = [?]	[?] + [?] = [?]
[+] × [-] = [-]	[+] × [+] = [+]	[-] × [-] = [+]
[?] × [-] = [?]	[?] × [+] = [?]	[?] × [?] = [?]

D'un modèle différentiel vers un graphe d'interactions

Considérons l'exemple classique de la modélisation de la production du glucose à partir du lactose chez *E. coli* (connu sous le nom d'opéron-lactose), détaillé sur la **figure 1.** On désigne les constituants du modèle par des indices *1, ..., i, ..., n*. La production d'une molécule i dépend des concentrations des autres molécules $X_1, X_2, ..., X_n$. Dans un modèle quantitatif différentiel, pour traduire cette information, on exprime les taux de production dX_i/dt sous la forme d'équations différentielles $dX_i/dt = F_i(X_1,..., X_n)$, où F_i est une fonction de $X_1, ..., X_n$.

On exploite d'autant mieux ce modèle qu'on connaît le plus précisément possible les fonctions F_i. La simulation et l'étude des réseaux métaboliques ont été portées par les connaissances cinétiques et biochimiques accumulées chez différents organismes. De même, à partir de nombreux jeux expérimentaux chez des mutants, un modèle différentiel très précis et prédictif a pu être construit pour le réseau contrôlant la division cellulaire chez les eucaryotes **(5).** En général, on dispose cependant de peu d'informations sur les F_i.

Même si la forme des fonctions F_i est inconnue, certaines informations qualitatives les concernant sont fournies par des expérimentations. En particulier, F_i dépend effectivement de X_j lorsque le taux de production du constituant *i* dépend de *j*. Par exemple, le taux de production de la protéine LacY dépend de la concentration de son inhibiteur LacI et de son inducteur AMPc. Cette dépendance est représentée par une relation graphique dénotant l'interaction entre *j* et *i*. On est ainsi amené à exploiter les connaissances les plus élémentaires (régulation d'expression de gènes, catalyse enzymatique...) dans une description qualitative du modèle sous la forme d'un graphe d'interactions : chaque nœud y représente un constituant ; un arc de *j* vers *i* signifie que *j* influence la production de i. On affecte un signe à l'arc $j \rightarrow i$ en fonction de l'action de *j* sur *i* : *[+]* si une augmentation de la concentration de *j* induit une augmentation de celle de *i*, *[-]* s'il y a diminution.

Formalisation d'une expérimentation comme déplacement d'état stationnaire

Les données recueillies lors d'une expérimentation représentent généralement les rapports des concentrations des constituants ou leurs variations entre deux états stationnaires. Or, à la stationnarité, les variables n'évoluent plus dans le temps : dans le modèle quantitatif, il s'agit de solutions du système d'équations (non linéaires) $F(X) = 0$. Le début et la fin d'une expérimentation se représentent ainsi par deux états $X_{éq}^1$ et $X_{éq}^2$ solutions de $F(X) = 0$. En particulier, on exclut le cas où le système présente des oscillations au lieu de se stabiliser.

Variations qualitatives

Dans un modèle qualitatif, les quantités sont remplacées par des relations. Ainsi, au lieu de s'intéresser aux valeurs numériques des variations $\Delta(i) = X_{éq}^2(i) - X_{éq}^1(i)$, on s'intéresse uniquement à leur signe. Par exemple, une variation positive *[+]* de niveau d'expression signifie qu'un gène non exprimé ou peu

exprimé dans l'état initial s'exprime dans l'état final. Certaines variations sont inconnues, il est donc utile d'utiliser également le signe indéterminé *[?]*.

Conformément à l'intuition, on peut effectuer des sommes et des multiplications sur ces signes **(tableau 1)**. En raison du signe *[?]*, on doit définir la notion d'égalité entre deux signes avec précaution : on considère que les relations $[?] \approx [+]$ et $[?] \approx [-]$ sont vraies, puisqu'il existe une possibilité pour que les variables qui sont ainsi représentées aient le même signe. On obtient ce qu'on appelle une algèbre de signes.

Équations qualitatives

L'intuition biologique et le bon sens nous disent qu'il doit y avoir des contraintes à satisfaire par les signes des variations. Ainsi, dans le modèle de l'opéron lactose, lorsqu'on sait que *LacI* diminue, *LacZ* doit augmenter à moins qu'il reste indéterminé. On peut écrire ce qu'on appelle une équation qualitative $LacZ \approx -LacI$. Ainsi, la connaissance du graphe d'interactions permet d'écrire un système d'équations qualitatives. Plus précisément, on cherche à comprendre l'effet d'une perturbation sur la variation $\Delta(i)$. On note $j_1, j_2, \ldots, j_k$ les produits qui ont une influence directe sur le constituant i, ce qui revient à dire qu'il existe une flèche de $j_1, j_2, \ldots, j_k$ vers i dans le graphe d'interaction. En utilisant le modèle différentiel, on montre que les signes des variations vérifient la relation suivante **(6, 7)** :

$$\text{signe}(\Delta(i)) \approx \text{signe}(j_1 \rightarrow i) \times \text{signe}(\Delta(j_1)) + \ldots + \text{signe}(j_k \rightarrow i) \times \text{signe}(\Delta(j_k))$$

On peut considérer cette relation comme une équation vérifiée par les variables signe$(\Delta(i))$, signe$(\Delta(j_1))$, signe $(\Delta(j_k))$. Il s'agit de l'expression mathématique rigoureuse d'une intuition : toutes les influences sur un nœud donné arrivent à travers ses premiers voisins **(6, 7)**.

Les équations associées au modèle de l'opéron lactose sont données dans le **tableau 2**. Comme pour un système d'équations numériques, on recherche des solutions à ce système en donnant la valeur *[+]* ou *[-]* à chaque variable et en vérifiant que toutes les équations sont satisfaites.

Ainsi, pour l'opéron lactose, il y a 2^8 (soit 256) valeurs possibles pour le jeu de variations (*Le, LacI, A, LacZ, Li, G, AMPc, LacY*). Parmi tous ces jeux de valeurs, seuls 18 sont effectivement solutions du système qualitatif. Nous explicitons ces 18 solutions dans le **tableau 2**. Par exemple, si on observe que les concentrations de LacI et A augmentent, on écrit $LacI = [+]$ et $A = [+]$, et on constate que l'équation (E1) du tableau, $LacI \approx -A$, n'est pas vérifiée. Le modèle prédit ainsi que LacI et A ne peuvent pas augmenter simultanément pendant l'expérimentation.

Inversement, la première ligne du **tableau 2** se lit $Le = [-]$, $LacI = [-]$, $A = [+]$, $LacZ = [+]$, $Li = [+]$, $G = [+]$, $AMPc = [-]$, $LacY = [+]$ et on s'assure que toutes les équations sont vérifiées dans ce cas.

La résolution de ces systèmes d'équations est un problème difficile : on parle de problème NP-complet, ce qui signifie qu'il existe des cas pour lesquels les calculs mettront un temps déraisonnable. Néanmoins, la structure des équations issues des graphes biologiques semble être suffisamment simple pour éviter de tels cas. Nos algorithmes **(8)** de résolution utilisent les redondances des contraintes et simplifient le problème. Ainsi, on arrive à traiter en quelques minutes des réseaux comportant initialement plusieurs milliers de produits, comme par exemple le réseau modélisant les régulations transcriptionnelles et les régulations associées aux facteurs sigma qui influencent la bactérie *E. coli* (3 883 interactions entre 1 529 molécules fournies par la base *RegulonDB* en mars 2006 **(9)**).

2 – Systèmes de contraintes relatives au réseau de l'opéron lactose présenté figure 1 et liste des 18 jeux expérimentaux (parmi les 256 possibles) solutions de ce système

Le	LacI	A	LacZ	Li	G	AMPc	LacY	
-	-	+	+	+	+	-	+	(S1)
-	-	+	+	-	+	-	+	(S2)
-	-	+	+	-	+	-	-	(S3)
-	-	+	+	-	-	+	+	(S4)
-	+	-	-	+	-	+	-	(S5)
-	+	-	-	+	-	+	+	(S6)
-	+	-	-	-	-	+	-	(S7)
-	+	-	-	-	-	+	+	(S8)
-	+	-	-	+	+	-	-	(S9)
+	-	+	+	-	-	+	+	(S10)
+	-	+	+	+	+	-	+	(S11)
+	-	+	+	-	+	-	-	(S11)
+	-	+	+	+	+	-	-	(S13)
+	-	+	+	-	+	-	+	(S14)
+	+	-	-	+	-	+	-	(S15)
+	+	-	-	+	-	+	+	(S16)
+	+	-	-	-	-	+	-	(S17)
+	+	-	-	+	+	-	-	(S18)

$LacI \approx -A$	(E1)	$G \approx Li + LacZ$	(E5)
$A \approx LacZ$	(E2)	$AMPc \approx -G$	(E6)
$LacZ \approx AMPc - LacI$	(E3)	$LacY \approx AMPc - LacI$	(E7)
$Li \approx Le + LacY - LacZ$	(E4)		

Par exemple, l'équation (E1) signifie que la variation de **LacI** lors d'un déplacement d'équilibre est de signe opposé à celle de **A**. Chaque ligne du tableau de signes est une solution du système d'équations.

Application à l'étude d'un modèle à partir de données expérimentales

Nous sommes ainsi capables de calculer la liste des solutions d'un système d'équations qualitatives. Pour exploiter concrètement ces solutions, nous proposons une démarche en plusieurs temps : d'abord, validation et correction d'un modèle, suivies de l'étude de ses prédictions, et enfin identification des meilleures expériences à faire pour valider les prédictions et améliorer le modèle.

Validation et correction d'un modèle

Nos méthodes permettent de tester la validité d'un modèle à partir d'un jeu de données expérimentales. Ainsi, nous avons déjà vu que LacI et A ne peuvent pas augmenter tous les deux pendant une expérience. Autrement dit, il n'y a aucun jeu dans le **tableau 2** pour lequel $LacI = A = [+]$.

De manière un peu moins évidente, si G et LacI augmentent tandis que Li diminue, alors les équations (E3), (E6) et (E7) ne peuvent pas être vérifiées simultanément. Dans ce cas, soit le modèle est faux, soit les observations sont erronées.

(6) Siegel A *et al.* (2006) *BioSystems* 84, 153-74
(7) Radulescu O *et al.* (2006) *J Roy Soc Interface* 3(6), 185-96
(8) Veber P *et al.* (2004/5) *Complexus* 2, 140-51
(9) Salgado *et al.* (2006) *Nucleic Acids Res* 34, 394-7

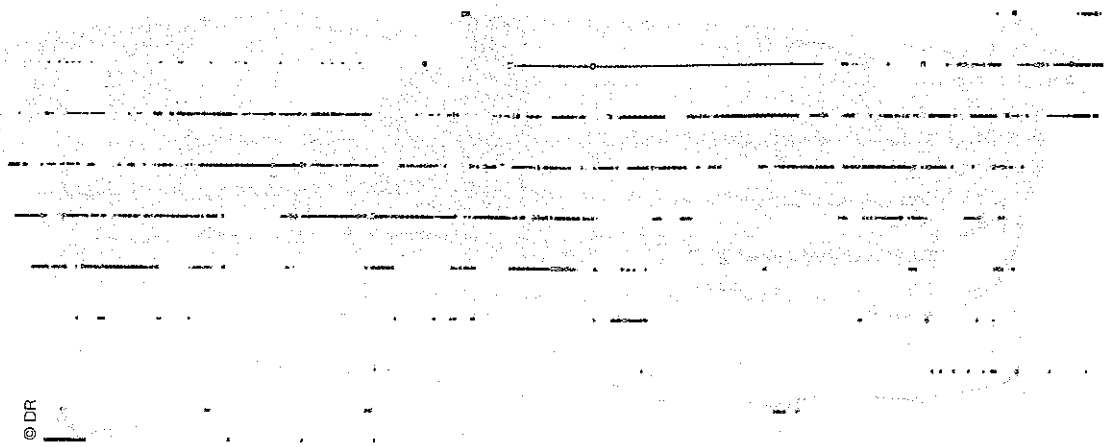


Figure 2 Prédictions sur le comportement du réseau décrivant les régulations transcriptionnelles et celles des facteurs sigma chez *E. coli* (1 529 variables, 3 883 interactions)
Les variations de 40 molécules sous l'effet d'un stress nutritionnel sont validées par la littérature (carrés bleus/baisse et verts/augmentation). Elles permettent de prédire le comportement de 381 molécules supplémentaires (carrés rouges), c'est-à-dire l'augmentation ou la baisse de leur concentration sous l'effet d'un stress nutritionnel.

On peut aussi montrer que le réseau transcriptionnel incluant les facteurs sigma chez la bactérie *E. coli* n'est pas compatible avec les observations expérimentales sur le stress nutritionnel induisant un passage en phase stationnaire, détaillées dans la base *RegulonDB* **(9)** et portant sur 40 composants. En analysant les solutions du système, on montre aussi que cette incompatibilité provient d'une erreur de retranscription des observations expérimentales sur deux produits dans la base *RegulonDB*. Après correction du jeu de données, le modèle est validé par les observations **(10)**.

9) Guziolowski C *et al.* 006) à paraître

Pouvoir de validation d'un jeu de données

Il faut noter que la validation d'un modèle à l'aide de certains jeux de données n'a parfois aucune signification. Par exemple, une observation du système de l'opéron lactose limitée aux nœuds *(Le,G,A)* ne peut pas mettre en défaut le modèle proposé. En effet, pour chaque signe affecté au triplet *(Le,G,A)*, on trouve dans le **tableau 2** un ensemble de signes pour *(Li,LacY,LacZ,LacI,AMPc)* qui est solution du système. Et il ne s'agit pas d'un cas isolé : parmi les 56 possibles, 22 triplets de composants du modèle de l'opéron lactose ne permettent aucunement de valider le modèle. Inversement, parmi les huit valeurs possibles du triplet *(LacI,A,LacZ)*, seuls les jeux *([+],[-],[-])* et *([-],[+],[+])* peuvent être complétés en une solution du système : ce jeu de données s'avère très contraignant pour la validité du modèle. Parmi les triplets de constituants, il s'agit en fait du jeu qui est le plus astreignant. Dans ce contexte, un expérimentateur ne pouvant faire que trois mesures pour valider son modèle aura tout intérêt à tester en priorité les composants LacI, A, et LacZ.
Plus généralement, on attribue à un jeu de p constituants un pouvoir de validation qui est d'autant plus proche de 1 que les composants sont à même de valider le modèle (voir un exemple **tableau 3** pour l'opéron lactose). Si on vient de construire un modèle qui doit être validé par des expérimentations, on peut ainsi rechercher, pour de petites valeurs de p (comprises entre 10 et 20), quels sont les p constituants les plus pertinents pour cette validation. L'explosion combinatoire des calculs empêche de choisir des jeux expérimentaux de plus grande taille.

Prédictions qualitatives

Supposons maintenant qu'une expérimentation sur l'opéron lactose ait montré que Le décroît et LacI augmente. On se retrouve ainsi dans une des solutions *(S5), (S6), (S7), (S8), (S9)* du **tableau 2.** En examinant ces cinq solutions, on constate qu'on a toujours $A = LacZ = [-]$. Autrement dit, le modèle prédit que A et LacZ diminuent pendant l'expérimentation.
Plus généralement, on appelle prédiction du modèle en rapport avec une expérimentation l'ensemble des constituants dont la variation est identique dans toutes les solutions du système qualitatif qui étendent le jeu

3 – Pouvoir de validation en fonction du groupe de sommets observés sur le réseau de l'opéron lactose

Observation	Pouvoir de validation
(LacZ,LacI,G,A)	0,75
(AMPc,LacI,G,A)	0,75
(AMPc,LacZ,G,A)	0,75
(LacZ,LacY,LacI,A)	0,75
(Le,LacZ,LacI,A)	0,75
(AMPc,LacZ,LacI,G)	0,75
(Li,Le,LacY,G)	0
(AMPc,Li,Le,LacY)	0
(Li,Le,LacY,A)	0,125
(Li,Le,LacY,LacI)	0,125
(Li,Le,LacZ,LacY)	0,125
(AMPc,Le,LacY,A)	0,25

Le pouvoir de validation a été calculé pour tous les groupes de quatre sommets ; dans le tableau ne figurent que les lignes dont le score est minimal ou maximal. Si le taux d'un groupe est proche de 1, une observation des molécules compatible avec les équations qualitatives du réseau valide fortement le réseau. Notons qu'il ne semble pas y avoir de règle simple pour deviner, à partir du graphe d'interactions, quels groupes de sommets sont les plus importants à observer. Nos méthodes permettent précisément d'explorer la combinatoire des interactions, et d'appréhender les relations complexes entre sommets du réseau.

expérimental. Ceci revient à propager dans le graphe d'interactions l'information fournie par les expérimentations. On prédit ainsi le comportement d'un certain nombre de constituants non observés.
En pratique, sur le réseau d'interactions d'*E. coli* incluant les facteurs sigma, les données concernant le stress nutritionnel portent sur 40 produits et prédisent la variation de 381 molécules supplémentaires, qui sont validées à 70 % par des données « transcriptome « **(figure 2) (10).** Les 30 % de prédictions non validées indiquent des défauts du modèle qui doit être précisé.

Mesure alternative

Partant de connaissances et de données incomplètes sur un système, cette démarche permet d'évaluer la validité d'un modèle, puis de guider les expérimentations qui permettront de le préciser, en quantifiant l'importance des produits pour la validation du modèle. Plus généralement, ces méthodes suggèrent une mesure alternative de l'importance fonctionnelle d'un groupe de composants d'un réseau, basée sur le pouvoir prédictif de leur observation et prenant en compte le pourcentage du réseau qui est contraint par l'observation d'un jeu de variables.
On pourra en particulier comparer cette approche à la théorie statistique des réseaux biologiques **(11),** où l'importance d'un sommet est en rapport avec le nombre total de connexions. Les réseaux transcriptionnels des procaryotes suggèrent ainsi que les nœuds de grande valence sont les plus conservés au cours de l'évolution. Il sera intéressant de voir si le pouvoir prédictif défini par les contraintes qualitatives confirme ces suggestions. ●

(11) Barabasi AL, Albert R (1999) *Science* 286, 509-12

Structure et modélisation des réseaux métaboliques

La connaissance du génome d'un organisme, associée au travail antérieur des biochimistes, permet une bonne connaissance de son métabolisme. La complexité des réseaux métaboliques oblige cependant à se concentrer, dans un premier temps, sur l'étude de leur structure, indépendamment des propriétés cinétiques des enzymes. Ces analyses ainsi limitées produisent cependant des résultats extrêmement informatifs et pertinents.

Jean-Pierre Mazat*,, Christine Nazaret***, Sabine Pérès***

* Inserm U688 et Programme d'épigénomique, Genopole®, Évry ; université Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux cedex
** jpm@u-bordeaux2.fr
*** CNRS 5466, université Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux cedex

En principe, la connaissance du génome entraîne la connaissance du métabolisme de l'organisme **(figure 1).** Les biologistes ont pendant longtemps suivi une approche « réductionniste » dans laquelle les composants individuels du système vivant ont été étudiés séparément. Avec l'arrivée des données de masse (séquençage, puce à ADN, etc.), la procédure s'inverse et permet désormais d'étudier comment ces composants interagissent pour former un système complexe en utilisant une approche « intégrée ». L'objectif de la biologie intégrative est ainsi de comprendre comment les différents éléments interagissent dans un système donné afin d'établir des relations qui conduisent d'un génotype à son expression phénotypique. C'est pourquoi l'étude du métabolome est une étape nécessaire.

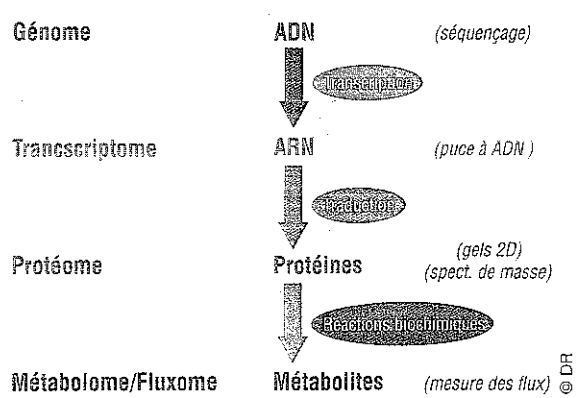


Figure 1 **Du génome au métabolome**