

Une méthodologie pour l'analyse qualitative des réseaux biologiques : de la base de données à la vérification formelle

Y. BASTIDE * S. LAGARRIGUE † M. LE BORGNE ‡ A. SIEGEL ‡ P. VEBER ‡
O. RADULESCU § A. LE BECHEC **

* ENSA Rennes

† UMR de génétique animale, INRA, Rennes

‡ Projet Symbiose, IRISA, Rennes

§ IRMAR UMR-CNRS 6625, Rennes

** INSERM U620, Rennes

Courriel : {bastide, lagarrigue}@agrocampus-rennes.fr,

{leborgne, asiegel, pveber}@irisa.fr,

{ovidiu.radulescu, antony.lebechec}@univ-rennes1.fr

Résumé

Nous présentons ici un ensemble d'outils et méthodes permettant de tester la cohérence de modèles de réseaux d'interactions géniques et métaboliques. Nous montrons comment des données qualitatives portant sur le signe des interactions peuvent être collectées et exploitées dans des modèles de taille importante. Cette démarche est fondée sur une formalisation mathématique qui permet d'en délimiter le champ d'application.

Mots clés : Réseaux biologiques, modèle différentiel, base de données d'interactions, résolution de systèmes d'équations qualitatives

Abstract

We present tools and methods for checking the consistency of gene and metabolic networks models. Qualitative informations on interaction signs and on variations of concentration are collected and used in models of significant size. This approach is based on a mathematical foundation whose hypotheses delimit the applicability.

Keywords: Biological networks, interaction database, solving a system of qualitative equations

Introduction : modèles et données non quantitatives

L'utilisation de plus en plus courante des puces à ADN pour l'étude des régulations cellulaires est jusqu'à présent orientée vers la détection de gènes co-régulés et l'inférence de régulations intergènes. L'intégration des données de puces aux connaissances biologiques portant sur d'autres éléments du fonctionnement cellulaire, comme le métabolisme, passe par l'utilisation de nouveaux modèles formels et des logiciels correspondant. Ajoutons que l'arrivée de données issues d'autres techniques émergentes tel que la spectrométrie de masse rend encore plus nécessaire le développement de méthodes adaptées.

La nature des modèles à utiliser dépend fortement de la qualité des données expérimentales : des modèles sophistiqués sont inapplicables si leurs paramètres ne sont pas accessibles à partir des résultats de l'expérimentation. Les données de puces à ADN ont ainsi deux caractéristiques contraignantes : elles sont de nature différentielle (deux situations sont comparées), et qualitative (en terme de sur ou sous-expression de gènes).

Un grand nombre de connaissances biologiques sont également de nature qualitative. Elles ont été établies par des expériences où peu de produits ont été mesurés ; les interactions qu'elles ont mises en évidence sont donc dispersées dans la littérature. Quoiqu'il existe des efforts nombreux pour rassembler ces connaissances, nous n'avons pas trouvé dans les bases de données existantes (comme Bind [1], IntAct [4], Amaze [9]) les informations nécessaires à notre modélisation.

À partir de ces constatations, nous avons développé un environnement comportant une base de données, des logiciels d'extraction de modèles, d'exploration et de preuve de propriétés des systèmes. L'ensemble

s'appuie sur un formalisme mathématique [13, 15] dans lequel le problème de la cohérence entre les données expérimentales et les modèles est résolu sous des hypothèses clairement énoncées. Ce formalisme débouche sur des modèles qualitatifs, et nous proposons une méthode originale de résolution de ces systèmes.

Le formalisme mathématique est détaillé dans la section 1. La modélisation à priori s'appuie essentiellement sur l'extraction manuelle et la gestion dans une base de données relationnelle des informations qualitatives présentes dans la littérature scientifique ; la méthodologie retenue est décrite dans la section 2. Nous présentons ensuite, dans la section 3, une nouvelle approche de résolution de systèmes d'équations qualitatives. Cette démarche a été appliquée (section 4) à l'étude du métabolisme des lipides chez le poulet sur un modèle de taille significative : plus de 500 variables et plus de 1 000 interactions.

1 Système d'équations qualitatives décrivant un réseau biologique

Dans cette partie, nous détaillons les formalismes mathématiques qui vont permettre sous certaines hypothèses de traduire en terme d'équations un réseau d'interactions biologiques.

L'approche mathématique suppose que les quantités impliquées dans le modèle obéissent à un système d'équations différentielles. Les informations qualitatives extraites de la littérature sont interprétées comme des informations de signe sur ces équations, et codées dans un graphe d'interaction. Deux situations expérimentales sont considérées comme les réalisations de deux états d'équilibre du système différentiel qui sont partiellement observés à travers les mesures. Ces hypothèses permettent de dériver un système d'équations qualitatives liant les variations des grandeurs du modèles entre les deux points d'équilibre. Un tel système se calcule facilement à partir du graphe d'interaction.

Les données expérimentales qualitatives permettent d'instancier les variables correspondantes du système qualitatif. Ceci est le point de départ de la vérification de la cohérence entre le modèle et les données. À partir du même modèle, il est possible d'identifier d'autres phénomènes biologiques tels que les compétitions entre voies.

1.1 Modélisation différentielle d'un réseau biologique

Soit M un réseau biologique dans lequel interagissent des produits (gènes, ARN transcrits, protéines, métabolites, etc.) qui sont désignés par les indices $1, \dots, n$. Notons X le vecteur dont les coordonnées sont les concentrations de ces produits.

Nous faisons l'hypothèse que le réseau est régi par une dynamique différentielle, c'est-à-dire que X satisfait une relation du type $\frac{dX}{dt} = F(X, P)$, où P désigne un ensemble de paramètres modélisant les influences extérieures.

Un état d'équilibre est caractérisé par le système d'équations non linéaires $F(X, P) = 0$.

La modification des paramètres entraîne un déplacement de l'état d'équilibre. Nous supposons que, aux états d'équilibre, les concentrations X_i sont des fonctions différentiables par morceaux des paramètres P . Notons que ceci n'est pas incompatible avec l'existence de multi-stationnarité : entre les valeurs critiques qui induisent un saut d'états d'équilibre, la dépendance de ces états peut être considérée comme lisse.

1.2 Matrice jacobienne et graphe d'interaction

En général, très peu d'information est disponible sur la fonction non linéaire F . Le but des paragraphes suivants est de détailler comment exploiter l'information fournie par le jacobien de F , c'est-à-dire la matrice des dérivées partielles $\frac{\partial F_i}{\partial X_j}$.

La matrice jacobienne permet d'introduire de manière naturelle un graphe orienté appelé graphe d'interaction du réseau. Ses nœuds sont les acteurs du réseau, désignés par les entiers $\{1, \dots, n\}$, et correspondent aux variables X_i du système différentiel. Un arc relie le sommet j au sommet i si et seulement si j agit sur i , c'est-à-dire si et seulement si $\frac{\partial F_i}{\partial X_j} \neq 0$. Le signe de cette dérivée est une étiquette de l'arc.

Nous supposons que chaque produit agit sur lui-même (processus de dégradation, auto-régulation...), impliquant qu'il existe un arc de i vers i pour tout nœud i du graphe d'interaction.

Puisque F est non linéaire, le signe de ses dérivées partielles peut changer ; il dépend de l'état X du système. Néanmoins, il est concevable qu'un système soumis à certains changements d'états ne passe pas par la

variété $\frac{\partial F_i}{\partial X_j} = 0$, ce qui signifie que le graphe d'interaction a une certaine stabilité. Cette hypothèse justifie l'étiquetage de l'arc de j vers i par le signe $s(j, i)$ de $\frac{\partial F_i}{\partial X_j}$.

D'un point de vue biologique, un arc relie j à i dès que le produit j a une influence directe sur la dynamique de i . Par exemple, j peut être un facteur de transcription régulant l'expression d'un gène i , une protéine impliquée dans la phosphorylation de i , une enzyme catalysant la production de i , etc. La mention d'une telle influence dans la littérature est interprétée en un arc du graphe d'interaction.

1.3 Données expérimentales

Supposons que nous disposons de données relatives à deux états d'équilibre distincts du système. Formellement, il faut aussi supposer que le changement d'état est dû à l'influence des paramètres P , ces paramètres prenant des valeurs continues. Ceci est difficilement vérifiable en réalité, mais il est assez naturel d'imaginer un ensemble continu d'états d'équilibre intermédiaires entre les deux situations observées. Cette approximation permet d'interpréter dans le modèle des données expérimentales comme la variation totale, entre deux états d'équilibre, des variables représentant les nœuds observés.

Une telle hypothèse n'est pas toujours vérifiée, en particulier pour des observations faites à des instants relativement proches ; il est alors inexact de dire que le système a atteint un état d'équilibre à tous les instants où il est observé.

Dérivons l'équation $F(X, P) = 0$ pour des petites variations des paramètres [15]. Si la production et la consommation d'une variable X_i ne dépendent pas directement des paramètres (c'est-à-dire si $\frac{\partial F_i}{\partial P_k} = 0$ pour tous les paramètres P_k), et si X_i exerce une régulation sur lui-même (c'est-à-dire si $\frac{\partial F_i}{\partial X_i} \neq 0$), alors les variations de concentration du produit i peuvent être calculées à partir des variations des prédécesseurs de i dans le graphe d'interaction :

$$\delta X_i = - \left(\frac{\partial F_i}{\partial X_i} \right)^{-1} \sum_{k \in \text{pred}(i)} \frac{\partial F_i}{\partial X_k} \delta X_k. \quad (1)$$

Ici, $\text{pred}(i)$ désigne l'ensemble des prédécesseurs de i distincts de lui-même. Cette relation est analogue au formalisme développé pour les réseaux électriques : $\chi_i = - \left(\frac{\partial F_i}{\partial X_i} \right)$ s'assimile à une impédance, c'est-à-dire le rapport entre la force produite par les prédécesseurs de i et les variations de i [13].

1.4 Système d'équations qualitatives

Pour un certain nombre d'expérimentations, comme les puces à ADN, les seules données disponibles sont de type qualitatif : ces données indiquent simplement le signe des variations des concentrations de produits. Pour les traiter, le système fourni par l'équation 1 est transposé dans l'algèbre des signes ; nous pouvons ainsi exprimer des relations entre les signes des variations des variables et les signes des coefficients d'interaction. Le terme « algèbre des signes » désigne l'ensemble $\{+, -, ?\}$, où $?$ = $\{+, -\}$ représente l'indétermination. Les lois de composition commutatives naturelles s'appliquent :

$$++- = ? \quad +++ = + \quad -+- = - \quad + \times - = - \quad + \times + = + \quad - \times - = +$$

et la présence d'un opérande indéterminé entraîne l'indétermination du résultat. Nous pouvons injecter la valeur nulle ($\mathbf{0}$) dans l'algèbre des signes avec les règles intuitives : $\pm + \mathbf{0} = \pm$, $? + \mathbf{0} = ?$, $\{\pm, ?\} \times \mathbf{0} = \mathbf{0}$. Une équation dans l'algèbre des signes est *satisfaite* si et seulement si un des membres est égal à $?$ ou si les deux membres sont égaux et différents de $?$ [6, 17].

L'équation 1 est étendue à des variations importantes de concentration en supposant que les coefficients d'interaction sont constants sur le domaine de validité de l'expérimentation. Dans ce cadre, nous avons démontré le théorème suivant :

Théorème 1.1 ([15]) *Soit M un réseau biologique de n variables dont les concentrations (X_1, \dots, X_n) satisfont la dynamique $\frac{dX}{dt} = F(X, P)$, avec les propriétés suivantes :*

- les dérivées partielles $\frac{\partial F_i}{\partial X_i}$ sont strictement négatives sur le domaine de variation des paramètres ;
- les signes des coefficients d'interaction $\frac{\partial F_i}{\partial X_j}$ sont constants sur le domaine de variation des paramètres.

Soit X_i une variable pour laquelle on suppose que l'influence des paramètres P sur X_i est transmise par les prédécesseurs de i dans le graphe d'interaction, c'est-à-dire $\frac{\partial F_i}{\partial P} = 0$. Alors le signe des variations de X_i entre deux états d'équilibre vérifie la relation suivante dans l'algèbre des signes :

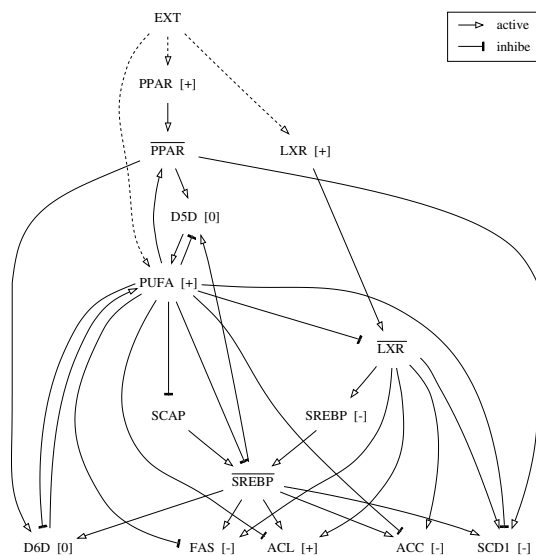
$$s(\delta X_i) = \sum_{k \in \text{pred}(i)} s(k, i) s(\delta X_k). \quad (2)$$

1.5 Exemple

Considérons à titre illustratif un modèle simplifié de la régulation génétique de la synthèse des acides gras dans le foie. Deux voies de production d'acides gras y coexistent : les acides gras saturés sont produits à partir des citrates au moyen d'une voie métabolique composée de quatre enzymes (ACL, ACC, FAS, SCD1) ; les acides polyinsaturés, tels que l'acide arachidonique, sont pour leur part synthétisés à partir d'acides gras essentiels apportés par l'alimentation. D5D et D6D catalysent les étapes clés de cette synthèse. Entre autres fonctions, les acides gras polyinsaturés (notés PUFA) régulent l'expression de gènes qui influent sur le métabolisme des lipides, des carbohydrates et des protéines. Les acides gras polyinsaturés agissent soit directement en se liant avec divers récepteurs nucléaires (PPAR α , LXR α , HNF-4 α), ce qui induit des changements de l'activité de leurs protéines associées, soit indirectement en régulant la quantité de certains facteurs de transcriptions (SREBP-1c, ChREBP, etc.) [5]. Notons $\overline{\text{LXR}}$ et $\overline{\text{PPAR}}$ les complexes actifs LXR:RXR de LXR et PPAR:RXR de PPAR, et $\overline{\text{SREBP}}$ la forme active (clivée) de SREBP.

La figure 1 représente un graphe d'interaction obtenu en rassemblant les informations disponibles dans la littérature au sujet des relations entre les produits cités ci-dessus (sommets, arcs et signes des arcs). Nous nous plaçons dans la situation d'un protocole de mise à jeun. Divers résultats expérimentaux indiquent que l'on devrait observer (signes des sommets) une baisse des quantités de SREBP, ACL, ACC, FAS, SCD1, une augmentation des quantités de PUFA et LXR (ARN) et la stabilité de D5D et D6D [8, 10].

Dans l'exemple détaillé sur la figure 1, nous supposons que, dans nos données expérimentales, ACL augmente au lieu de diminuer. Le système d'équations qualitatives obtenu avec ce jeu de données n'admet alors pas de solution. Un changement de la donnée portant sur ACL permet de retrouver un système compatible [15].



1. $\overline{\text{PPAR}} = \text{PPAR} + \text{PUFA}$
2. $\overline{\text{LXR}} = -\text{PUFA} + \text{LXR}$
3. $\text{SREBP} = \overline{\text{LXR}}$
4. $\overline{\text{SREBP}} = \text{SREBP} + \text{SCAP} - \text{PUFA}$
5. $\text{ACL} = \overline{\text{LXR}} + \overline{\text{SREBP}} - \text{PUFA}$
6. $\text{ACC} = \overline{\text{LXR}} + \overline{\text{SREBP}} - \text{PUFA}$
7. $\text{FAS} = \overline{\text{LXR}} + \overline{\text{SREBP}} - \text{PUFA}$
8. $\text{SCD1} = \overline{\text{LXR}} + \overline{\text{SREBP}} - \text{PUFA} + \overline{\text{PPAR}}$
9. $\text{SCAP} = -\text{PUFA}$
10. $\text{D5D} = \overline{\text{PPAR}} + \overline{\text{SREBP}} - \text{PUFA}$
11. $\text{D6D} = \overline{\text{PPAR}} + \overline{\text{SREBP}} - \text{PUFA}$

PPAR = +, PUFA = +, LXR = +, SREBP = -, SCD1 = -, FAS = -,
ACL = +, ACC = -, D5D = 0, D6D = 0

FIG. 1 – Graphe d'interaction et système d'équations qualitatives associé, pour un modèle abstrait de la régulation de la synthèse des acides gras dans le foie. Le nœud EXT représente l'influence des paramètres extérieurs. Les boucles d'autorégulation négatives sont omises. Dans le système d'équations, A représente le signe $s(\delta A)$ des variations du produit A.

2 Base de données Gardon

La section précédente a montré comment utiliser les signes du graphe d'interaction d'un réseau biologique pour définir un système d'équations vérifiées par des données expérimentales. Dans cette partie, nous décrivons une base de données qui permet, à partir d'informations issues de la littérature, de construire le graphe d'interaction.

2.1 Modélisation conceptuelle d'une interaction

Nous modélisons les informations bibliographiques autour de la notion d'interaction entre produits.

- Un produit peut être un gène, un produit de gène, une substance, etc.
- Une interaction est une fonction opérant sur des acteurs dans un contexte. Deux niveaux de description sont autorisés :
 - une interaction comportementale, qui décrit l'effet d'une variation de concentration d'un produit sur la concentration d'une cible (« activation », « inhibition », « aucun effet »). Cette interaction peut être directe (s'il n'y a pas de mécanisme intermédiaire) ou indirecte (en particulier, lorsque le mécanisme est inconnu).
 - un niveau biochimique (« transcription d'un gène », « clivage », etc.). Il s'agit du même type d'interactions que dans les bases de données existantes. Ces interactions ne renseignent pas directement sur les variations de quantités impliquées mais décrivent des mécanismes. Dans la suite, il faudra donc interpréter l'effet de ces interactions sur les concentrations des produits.

Suivant le type d'interaction, les acteurs seront des produits ou d'autres interactions ; ils prendront différents rôles, comme « cible », « facteur », ou « résultat ». Le contexte prend en compte les divers paramètres décrivant l'expérimentation. Une interaction est aussi liée à l'article PubMed qui la décrit.

2.2 Base de données du métabolisme des lipides

La liste des types d'interactions que nous avons utilisés pour le métabolisme des lipides est présentée dans le tableau 1. Une famille de types d'interactions peut être codée de deux manières : soit comme un seul type prenant des paramètres, soit comme des types distincts. Par exemple, « action sur une interaction » rassemble trois possibilités (actions positive, négative ou sans effet) ; à l'inverse, « transcription » et « inhibition de la transcription » sont séparés. Une interaction biochimique peut être reliée explicitement à son interprétation comportementale. Par exemple, une « transcription » se traduit qualitativement en une « activation directe ». Ces liens remplacent si nécessaire la traduction automatique.

Interactions comportementales	Interactions biochimiques	
activation	clivage	liaison protéique
inhibition	dégradation	transcription
aucun effet	glycolisation	inhibition de la transcription
action sur une interaction	phosphorylation	ubiquitination

TAB. 1 – Types d'interactions

La modélisation décrite se retrouve dans le diagramme de classes UML de la figure 2¹. Chaque interaction (Action) est d'un certain type (Action_type), lui-même d'une certaine catégorie, comportementale ou biochimique (Action_category). Le type d'interaction détermine les types d'acteurs possibles (Action_type_actor). Chaque type d'acteur indique un label, s'il s'agit d'un produit ou d'une action², un sens entrant ou sortant, et des cardinalités minimale et maximale.

Différentes classes décrivent le contexte expérimental de l'interaction. L'espèce (Taxon), qui utilise la nomenclature du NCBI, est liée à l'interaction ou à chacun de ses acteurs, suivant que l'interaction est mono-espèce ou non. Le tissu (Tissue) est lui toujours précisé au niveau de l'interaction. Le contexte

1. Les différences de pointillés des arcs sont des artefacts d'impression.

2. Un type d'acteur peut en fait indiquer « product », « action_type » ou « any ». Dans ce dernier cas, chaque acteur individuel de ce type sera soit « product », soit « action_type » ; c'est pourquoi le lien exclusif vers Product ou vers Action_type est dans la classe Actor. Rétrospectivement, un tel degré de liberté n'est pas nécessaire.

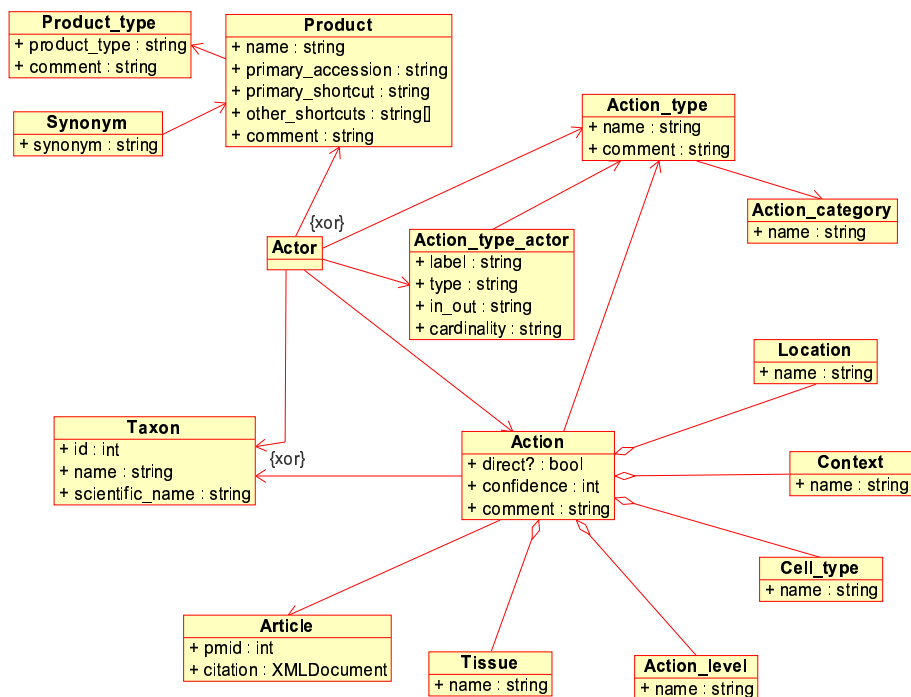


FIG. 2 – Diagramme de classes

expérimental (Context) signale si l'expérience a été faite *in vivo*, *in vitro* ou *ex vivo*; le type cellulaire (Cell_type) peut aussi être renseigné, notamment dans le cas de cultures cellulaires où la lignée est mentionnée. Le niveau d'observation de l'interaction (Action_level) indique si elle est observée au niveau de l'ADN, de l'ARN ou de la protéine. Sa localisation (Location) peut être le noyau, le cytoplasme, ... Chaque produit (Product) est caractérisé par un type (Product_type), et décrit par un nom, un code d'accèsion, un ensemble de raccourcis et des synonymes (Synonym). Par exemple, le « récepteur nucléaire, sous-famille 1, groupe H, membre 3 » a pour raccourcis NR1H3, LXRA, RLD1 et pour synonyme son ancien nom de « récepteur X du foie, alpha ».

2.3 Mise en œuvre

Le modèle qui vient d'être détaillé est matérialisé sous forme relationnelle dans une base de données PostgreSQL³. D'autres éléments sont stockés dans la base de données : des vues, des procédures et fonctions pour en faciliter l'utilisation, des tables d'authentification et autorisation, une table d'audit, et une copie de la taxinomie du NCBI.

L'interface utilisateur a été réalisée avec le serveur d'applications Zope⁴. Celui-ci permet de combiner des pages en HTML ou XML (contenant des directives spécifiques sous forme d'attributs XML) avec des scripts Python, des requêtes SQL, etc. Une partie de l'interface est gérée par les clients par l'intermédiaire de scripts Javascript : par exemple, les citations PubMed sont transmises sous forme XML au navigateur qui les affiche grâce à des feuilles de style XSLT et CSS.

2.4 Alimentation de la base de données

Alimentation manuelle Le remplissage de la base de données se fait à partir d'une bibliographie. Il faut donc : 1. déterminer les articles à utiliser ; 2. en extraire les informations sous forme d'interactions ; 3. saisir ces interactions.

Les articles « intéressants » d'un domaine peuvent soit être déjà connus, soit à déterminer : dans le premier cas, le spécialiste du domaine dispose d'une liste. Dans le second cas, un noyau de publications est déterminé grâce à des outils comme BiblioSphere puis étendu avec Entrez.

3. <http://www.postgresql.org>

4. <http://www.zope.org>

L'extraction des interactions demande un apprentissage de la part du lecteur : il doit savoir déterminer quelles sont les interactions décrites dans une publication, quels sont les produits impliqués, si l'éventuel caractère direct d'une interaction est suffisamment démontré, etc. Notons que le résumé d'une publication ne suffit pas à décrire les interactions. D'abord parce qu'il ne contient pas toutes les informations souhaitées ; ensuite parce que la distinction entre les résultats de manipulations effectuées et les conclusions qu'en tirent les auteurs n'y est pas forcément claire.

La saisie des interactions se fait en plusieurs étapes. L'utilisateur entre d'abord l'identifiant de la publication (PMID) qu'il va traiter. Le système en importe les caractéristiques depuis le serveur Entrez si elle n'est pas déjà connue (ces caractéristiques sont celles qu'on voit quand on recherche manuellement une publication sur Entrez : titre, auteurs, journal, résumé, etc.). Pour accélérer la saisie, l'utilisateur peut indiquer le produit principal de l'article. Il entre ensuite chaque interaction. Le principal problème à cette étape est le choix des noms de produits. En effet, nous voulons donner un seul nom à un même gène ou produit de gène quelle que soit l'espèce (nous mémorisons cette espèce à part). De même, des substances chimiques identiques peuvent avoir des noms différents suivant leurs fournisseurs. Un exemple de méthodologie est de suivre la nomenclature d'Hugo⁵ si le gène existe chez l'homme. Sinon, le même nom que dans la publication est utilisé et un commentaire le signale. Pour les autres substances (acides, alcools, sucres, etc.), plusieurs catalogues de fournisseurs sont disponibles sur le web et indiquent formules chimiques et synonymes. Un moteur de recherche interne est utilisé pour savoir si un produit existe déjà dans la base de données.

Alimentation semi-automatique de la base Comme indiqué plus haut, nous ne négligeons pas les données de nature biochimiques ; nous pouvons donc utiliser les informations disponibles dans des bases de données comme Bind [1], IntAct [4], Amaze [9], KEGG [11] ou TransPath [14]. Ces informations structurées seront soit intégrées dans la base de données, soit récupérées au moment de la construction du graphe d'interaction. Le problème récurrent (qui justifie l'utilisation intermédiaire de la base de données) est à nouveau celui de l'identification des produits.

2.5 Utilisation de la base de données pour construire un graphe d'interaction

L'objectif de la base de données est de permettre la construction d'un graphe d'interaction à partir des connaissances de la littérature. Dans ce but, chaque interaction inscrite dans la base qui est en relation avec une expérimentation donnée est interprétée qualitativement, et le graphe est généré grâce au logiciel Garmen⁶. Ainsi, une interaction comportementale de j vers i est traduite directement dans le graphe ($s(j, i) = +$ si « activation », $s(j, i) = -$ si « inhibition », $s(j, i) = 0$ si « aucun effet »). Les interactions biochimiques sans liaisons explicites, et les interactions agissant sur d'autres interactions, sont actuellement traduites sous une forme délibérément simplifiées. Si la cible d'une interaction est un produit, un arc allant de chaque facteur vers cette cible est ajouté au graphe. Le signe de l'arc dépend du type d'interaction. Si la cible de l'interaction A est une autre interaction B , nous créons des arcs depuis les facteurs de A vers la cible de B . Par exemple, si I inhibe la phosphorylation de P en \bar{P} , nous créerons l'arc $I \xrightarrow{-} \bar{P}$.

Le graphe d'interaction est complété par l'introduction de boucles de rétro-régulation négatives sur chaque sommet du graphe, impliquant $s(i, i) = -$ pour tout i . Comme expliqué précédemment, ces rétro-régulations résultent par exemple de processus de dégradation. Elles sont nécessaires pour exprimer la variation d'un sommet en fonction de ses prédécesseurs.

Il est bien évident que le graphe d'interaction extrait des publications risque d'être incomplet. En fait, nous estimons qu'il n'est pas nécessaire de rassembler toutes les connaissances biologiques dans un modèle. Les méthodes que nous développons ont en particulier pour but de décider si le modèle donné par la littérature est cohérent avec les données expérimentales disponibles, ou si des interactions sont manquantes.

5. <http://www.gene.ucl.ac.uk/nomenclature/>

6. Logiciel d'analyse graphique des réseaux biologiques, en cours de développement à l'IRISA.

3 Résolution des équations qualitatives décrivant le réseau

3.1 État de l'art

Les systèmes qualitatifs ont d'abord été utilisés en économie pour une raison similaire à celle rencontrée en biologie : absence de données quantitatives en nombre suffisant et suffisamment fiables. Leur étude a été relancée dans les années mille neuf cent quatre-vingt avec le développement de la physique qualitative, branche de l'intelligence artificielle visant à imiter les raisonnements des ingénieurs face à un système technologique modélisé de manière incomplète. L'absence de données chiffrées est compensée par le raisonnement sur les signes des variations ou plus finement sur leurs ordres de grandeur [6, 17].

La résolution d'équations linéaires qualitatives dans l'algèbre des signes a été un des premiers problèmes abordés. Ces systèmes ne peuvent se résoudre directement, en éliminant successivement les variables du système, par analogie avec l'élimination de Gauss : un obstacle essentiel est l'absence de transitivité de l'« égalité » qualitative.

Plusieurs heuristiques ont été proposées pour la résolution des systèmes qualitatifs, avec deux objectifs : d'abord, la résolution combinatoire du système d'équations en tant que problème mathématique ; ensuite, l'élaboration d'une explication qualitative au comportement du système physique (à l'aide du cheminement de la résolution), cette explication reflétant l'explication que pourrait donner un expert du système.

Les caractéristiques essentielles des systèmes d'équations qualitatives linéaires sont établies ainsi par Dormoy [3] : la résolution d'un problème qualitatif est un problème NP-complet ; il existe un système complet (on obtient toutes les solutions) et bien fondé (les solutions déduites sont correctes) de règles de résolution.

Les règles de résolution s'articulent autour d'une adaptation de la règle d'élimination de Gauss au contexte qualitatif. Des heuristiques permettent de choisir à tout moment quelle règle utiliser et à quelle équation l'appliquer pour progresser vers la solution. L'application systématique, comme dans l'élimination de Gauss classique, n'étant plus possible, la mise en œuvre de la résolution de systèmes qualitatifs en utilisant ces règles est délicate et de complexité exponentielle.

3.2 Une nouvelle approche : fonctions polynomiales sur un corps fini

Notre objectif est ici de vérifier de cohérence des données d'interaction issues de la littérature, d'abord entre elles, ensuite vis-à-vis de données expérimentales. Pour cela, il nous faut résoudre le système donné par l'équation 2.

Le caractère contradictoire d'un modèle ou son inadéquation aux données expérimentales se traduira par l'absence de solutions pour le système qualitatif. En effet, toute solution quantitative donne naissance à une solution qualitative ; l'absence de solution qualitative implique donc l'absence de solution quantitative. La clé de la méthode de résolution proposée réside en un codage des équations qualitatives en équations polynomiales sur un corps fini. Nous nous limitons aux corps finis de la forme $\mathbb{Z}/p\mathbb{Z}$, où p est un nombre premier. Un tel corps a pour éléments $\{0, 1, \dots, p-1\}$ avec les opérations de somme et produits calculées modulo p . L'application se fera avec $p = 3$; ainsi, dans $\mathbb{Z}/3\mathbb{Z}$, $1 + 1 = 2$; $1 + 2 = 0$; $2 + 2 = 1$; $2 \times 2 = 1$; etc. Les propriétés fondamentales suivantes des corps finis vont être utilisées :

- Dans un corps fini, toute fonction polynomiale est représentée par un polynôme où chaque variable est au plus de degré $p-1$. À l'aide d'une base de l'espace des fonctions polynomiales formée de polynômes de Lagrange, on montre que toute fonction $f : (\mathbb{Z}/p\mathbb{Z})^n \rightarrow \mathbb{Z}/p\mathbb{Z}$ à n variables sur $\mathbb{Z}/p\mathbb{Z}$ est égale à une fonction polynomiale.
- Si \oplus désigne l'opération suivante sur les fonctions polynomiales : $f \oplus g = f^{(p-1)} + g^{(p-1)}$, alors tout système d'équation $p_1(X) = 0, \dots, p_k(X) = 0$ admet le même ensemble de solutions que l'équation unique $p_1 \oplus p_2 \oplus \dots \oplus p_k(X) = 0$.

3.3 Codage des équations qualitatives

Le codage des équations qualitatives se fonde sur un plongement de l'algèbre des signes $\{+, -, ?\}$ dans le corps $\mathbb{Z}/3\mathbb{Z}$. Un codage similaire de $\{+, -, \mathbf{0}, ?\}$ se fait dans $\mathbb{Z}/5\mathbb{Z}$; il ne sera pas détaillé ici.

Nous représentons $+$ par 1, $-$ par 2 (= -1) et $?$ par 0. Pour coder les équations linéaires, il suffit de coder la somme et l'égalité qualitative à partir de leurs tableaux de valeurs. La somme de deux termes est représentée

par la fonction polynomiale $sq(X, Y) \stackrel{\text{def}}{=} XY(X + Y)$ et l'égalité qualitative par une équation algébrique $eq(X, Y) \stackrel{\text{def}}{=} XY(X - Y) = 0$.

La représentation naïve des fonctions polynomiales conduit à un encombrement mémoire rédhibitoire (l'espace des fonctions polynomiales est de dimension p^n , soit 3^n ici). La clé d'une représentation efficace réside dans la décomposition suivante :

$$p(X_1, X) = (1 - X_1^2)p_{[X_1=0]}(X) + X_1(-X_1 - X_1^2)p_{[X_1=1]}(X) + X_1(X_1 - X_1^2)p_{[X_1=2]}(X)$$

où p désigne un polynôme à n variables, $X = (X_2, \dots, X_n)$ désigne les $n - 1$ dernières variables et $p_{[X_1=x_1]}(X)$ désigne le polynôme sur $n - 1$ variables obtenu en substituant dans p la valeur x_1 à la première variable X_1 . Notons que l'existence d'une telle décomposition est généralisable à tout corps fini. Un ordre des variables étant fixé, la formule précédente conduit à une représentation de toute fonction polynomiale comme un arbre dont chaque nœud non terminal a 3 fils. Une représentation plus compacte s'obtient en remarquant la redondance de l'arbre ; ainsi, en ne représentant chaque nœud d'un type donné qu'une seule fois, une fonction polynomiale est représentée par un graphe acyclique.

Cette représentation est inspirée de celle utilisée dans les BDD (*binary decision diagrams*) [2] qui sont les supports des systèmes de *model-checking* performants comme NuSMV⁷. Quoique potentiellement exponentielle, cette représentation s'est avérée efficace en pratique.

Pour obtenir les résultats présentés dans la section suivante, nous nous sommes appuyés sur une implémentation des TDD (*ternary decision diagrams*) développée dans le cadre d'un logiciel de vérification de programmes de systèmes temps réel. SIGALI (*signal algebraic interpreter*) est un système de calcul formel dédié aux fonctions polynomiales sur $\mathbb{Z}/3\mathbb{Z}$ [7]. Le choix de ce corps est lié à une interprétation abstraite des programmes qui permet d'en vérifier des propriétés logiques. Seule la partie de calcul formel de SIGALI a été utilisée. L'extension de ce travail à des propriétés dynamiques donnera ultérieurement l'occasion d'employer plus complètement ce logiciel.

3.4 Exploitation du codage

Nous avons développé un programme qui traduit le graphe d'interaction associé à un réseau en équations qualitatives codées comme des équations algébriques sur $\mathbb{Z}/3\mathbb{Z}$.

Cohérence du modèle Afin de tester la cohérence du modèle qualitatif représenté par le graphe, il faut rechercher l'existence de solutions qualitatives autres que la solution indéterminée. Pour cela, nous calculons le polynôme p équivalent au système d'équations comprenant les équations extraites du graphe qualitatif et les équations $X_i^2 = 1$ permettant d'exclure les valeurs indéterminées des variations qualitatives. Un système est incohérent si ce polynôme équivalent est égal à 1, indiquant l'absence de solution. Dans le cas où le système n'est pas incohérent, ce polynôme donne toutes les variations qualitatives compatibles avec les contraintes.

Cohérence entre données et modèle La démarche suivie pour vérifier cette cohérence est identique ; nous partons cette fois des équations qualitatives où les variables observées sont instanciées par les valeurs expérimentales.

Origine des incohérences En cas d'incohérence dans le modèle ou entre modèle et données expérimentales, il faut en isoler les causes. L'existence de solutions étant une propriété de l'ensemble d'équations, il faut rechercher un sous-ensemble minimal qui est incohérent. Ce sous-ensemble n'est pas unique en général, et son choix doit faire intervenir des considérations biologiques en plus des critères mathématiques. Il ne semble pas exister d'algorithme permettant de calculer à coup sûr un tel sous-ensemble d'équations. Une heuristique assez efficace consiste à séparer l'ensemble d'équations contradictoires en deux sous-ensembles et de recommencer avec un des sous-ensembles contradictoires. Si aucun des deux sous-ensembles n'est contradictoire, il faut essayer un autre partitionnement.

7. <http://nusmv.irst.itc.it>

4 Application : le métabolisme des lipides

Suivant la méthodologie décrite en section 2, une base de données portant sur les mécanismes de régulation des lipides chez le poulet a été développée⁸. Le graphe d'interaction qui en résulte recense 1 260 régulations (arcs du graphe) portant sur 598 produits (sommets). Par ailleurs ce graphe est constitué pour l'essentiel d'une seule grande composante connexe. La plupart des sommets sont de faible degré, et sont soit sans prédécesseur, soit sans successeur. À l'inverse, un faible nombre de nœuds ont un degré très élevé et parmi eux la plupart des produits clé de la régulation des lipides.

Un problème évident à l'issue de la saisie d'une base de données de cette ampleur consiste à tester la validité des données introduites. Pour ce faire, nous testons la cohérence des interactions saisies, au sens du modèle qualitatif détaillé en section 1. Le système linéaire qualitatif issu du graphe d'interaction est codé sous forme d'un système d'équations algébriques sur $\mathbb{Z}/3\mathbb{Z}$, comme expliqué en section 3. La base est validée si ce système admet des solutions (cohérence du modèle). Si ce n'est pas le cas, on isole un sous-ensemble d'équations contradictoire, ce qui revient à chercher l'origine des incohérences.

4.1 Réduction du système

En pratique, trouver des solutions pour un système de plusieurs centaines de variables requiert un temps de calcul important. Nous proposons une réduction du système telle que le système réduit est contradictoire dans les mêmes conditions que l'original.

D'abord, si le graphe d'interaction comporte plusieurs composantes connexes, le système qualitatif associé peut être décomposé en autant de sous-systèmes à résoudre indépendamment.

Ensuite, lorsque des données expérimentales sont disponibles, les variables correspondant aux sommets observés sont instanciées.

S'il reste des systèmes de taille prohibitive, Nous profitons de ce qu'un nœud sans prédécesseur n'intervient que dans l'équation qui lui est associée. Si en plus ce sommet n'est pas observé, son équation associée n'apporte aucune contrainte sur le reste du système ; elle peut donc être retirée, de même que le sommet. Soit $f : G \mapsto G'$ la fonction qui, à un graphe d'interaction G , associe le graphe G' obtenu en supprimant de G tout sommet non observé et sans successeur autre que lui-même. Comme expliqué précédemment, si $\mathcal{S}(G)$ désigne le système d'équations algébriques associé au graphe d'interaction G , alors $\mathcal{S}(G)$ admet une solution si et seulement si $\mathcal{S}(f(G))$ en admet une. De plus, la suite $f^n(G)$ est décroissante et minorée et converge vers une limite G^* . Enfin, $\mathcal{S}(G)$ admet une solution si et seulement si $\mathcal{S}(G^*)$ en admet une.

Par conséquent, il suffit, pour valider le modèle issu de la base de données, de vérifier que le système associé au graphe réduit G^* admet une solution. La réduction du nombre de variables est très importante en pratique : le système original possède près de 600 variables, le système réduit moins de 150.

4.2 Validation de la base de données

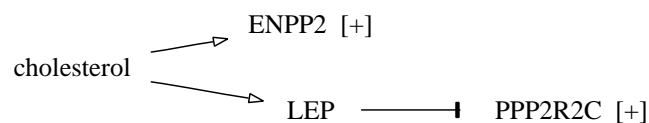


FIG. 3 – Sous-système isolé par l'algorithme de détermination d'un sous-ensemble d'équations contradictoire et minimal : les concentrations d'ENPP2 et PP2R2C ne peuvent pas augmenter simultanément. Ce système doit être résolu en remettant en cause les observations disponibles (représentées aux côtés des sommets). Selon nos données, c'est la mesure portant sur ENPP2 qui n'est pas fiable.

Cohérence du modèle et des données expérimentales L'application de cette méthode à la base de données Gardon a permis de détecter quelques imperfections mineures. Une fois corrigées, le système

8. Une version de test de cette base de données est disponible à l'adresse suivante : <http://bionix.irisa.fr/litterature/genetan>, accessible en lecture avec le nom « TestUser », mot de passe « Test05 ». Cette interface ne fonctionne pas avec Internet Explorer.

correspondant au graphe d'interaction admet des solutions. Nous disposons par ailleurs d'une trentaine de mesures obtenues *via* une expérience de puce à ADN. Parmi elles, nous distinguons un ensemble de mesures peu fiables, c'est-à-dire de support statistique faible (*P-value* élevée).

Le système obtenu est contradictoire, et une manière d'en isoler la cause consiste à déterminer un ensemble contradictoire minimal d'équations du système. Comme chaque équation est associée à un sommet du graphe et ses prédécesseurs, cette approche permet d'isoler les sous-graphes qui posent problème. Il n'existe pas à notre connaissance d'algorithme performant pour déterminer un sous-système contradictoire et minimal dans un système d'équations qualitatif, ou plus généralement dans un ensemble de formules logiques. Néanmoins, grâce à la réduction préalable du système, on peut se contenter en pratique de l'approche naïve suivante :

Entrées : un graphe d'interaction G , les équations S_i associées à chaque sommet i de G

Sorties : un système d'équation contradictoire minimal M

$R \leftarrow \{ S_i \mid i \in G \}$

$M \leftarrow \emptyset$

pour chaque sommet $i \in G$ **faire**

$R \leftarrow R \setminus \{ S_i \}$

si $M \cup R$ admet une solution **alors** $M \leftarrow M \cup \{ S_i \}$

fin

retourner M

Algorithme 1 : Détermination d'un sous-système contradictoire minimal. Au début de chaque itération, le système $M \cup R$ est contradictoire. À la fin de la boucle, R est vide, et M est minimal.

L'ensemble M retourné est un système contradictoire d'équations algébriques, minimal dans le sens qu'il suffit d'en retirer une équation pour que le système admette une solution. En appliquant cet algorithme, la comparaison du modèle aux données a révélé quelques incompatibilités, qui ont chaque fois pu être résolues en remettant en cause une observation préalablement jugée peu fiable (voir la figure 3 pour un exemple). L'intérêt principal de cette approche est qu'à l'ensemble M correspond un sous-graphe du graphe d'interaction ; elle offre donc un support très visuel pour le diagnostic.

5 Conclusion

Nous avons présenté une approche combinant l'utilisation de connaissances à priori et la production de données en masse sur des systèmes biologiques de taille significative. Le parti pris qualitatif est bien adapté à la nature des observations fournies par les techniques expérimentales actuelles. Le cadre mathématique utilisé fournit une base solide de discussion des hypothèses soutenant cette démarche et permet de cerner son domaine d'application.

Bien des phénomènes biologiques ne satisfont pas ces hypothèses. C'est en particulier le cas des phénomènes où l'essentiel du comportement du système est constitué de transitions brèves, comme cela semble le cas en signalisation. Des travaux sont en cours pour appliquer une démarche intégrative similaire à ces systèmes. Le cadre adopté pour la résolution des systèmes qualitatifs permet de traiter d'autres problèmes biologiquement significatifs. Nos travaux en cours portent sur :

Les variables uniquement déterminées Un système qualitatif linéaire a en général plusieurs solutions. Certaines ne correspondent à aucune solution du système quantitatif. Toutefois, certaines composantes des solutions prennent toujours la même valeur, quelle que soit la solution du système. Si le système quantitatif admet des solutions, alors, nécessairement, les signes de ces composantes sont parfaitement déterminés par le système qualitatif.

Nous projetons ainsi d'extraire les composantes d'un système d'équations polynomiales. Un algorithme utilisant l'implémentation par TDD est en cours d'élaboration et va être inclus dans SIGALI.

La pertinence d'un plan d'expérience Puisqu'un ensemble d'observations ne peut mettre en évidence qu'une contradiction avec le modèle qualitatif, un plan d'expérience n'est pertinent que s'il est susceptible de faire apparaître une telle contradiction. Sur le plan mathématique, il doit exister au moins une valeur

des variables observées X_{obs} pour laquelle il n'existe pas de valeur, parmi les non observées, compatible avec elle et avec les contraintes du modèle. Ceci peut se vérifier en calculant le polynôme des contraintes sur les variables observées.

Des travaux sont en cours pour dégager des méthodes d'aide à la conception de plans d'expérience dans le contexte de modèles qualitatifs. Des mesures de qualité d'un ensemble d'observations, ainsi que des critères autorisant la suppression d'observations sans dégrader cette qualité, sont en cours d'élaboration.

Modélisation qualitative Les interactions biochimiques sont actuellement traduites qualitativement sous forme simplifiée et parfois grossière, par exemple pour des réactions biochimiques successives d'une même voie métabolique. Dans ce genre de situation, nous pouvons améliorer les analyses en regroupant les réactions dans des modules dont nous traduiront globalement le comportement qualitatif, sans doute en utilisant les modélisations dynamiques existantes [12, 16]. Cette approche apparaît prometteuse également pour prendre en compte les différentes échelles de temps du modèle.

Remerciements Nous remercions Julien Cluchague et Patrick Lhomme pour leur travail sur les données bibliographiques, ainsi que Maud Jacquinet pour son implication dans le développement de la base de données.

Références

- [1] BADER (G. D.), BETEL (D.) et HOGUE (C. W.), « BIND : the biomolecular interaction network database », *Nucleic Acids Research*, 31, n° 1, 2003, p. 248-250.
- [2] BRYANT (R. E.), « Graph-based algorithm for boolean function manipulation », *IEEE Transactions on Computers*, 35, n° 8, August 1986, p. 677-691.
- [3] DORMOY (J. L.), « Controlling qualitative resolution », dans *Proceedings of the seventh National Conference on Artificial Intelligence, AAAI88', Saint-Paul, Minn., 1988*.
- [4] HERMIAKOB (H.), MONTECCHI-PALAZZI (L.), LEWINGTON (C.), MUDALI (S.) et autres, « IntAct – an open source molecular interaction database », *Nucleic Acids Research*, 32, Database issue, 2004, p. D452-D455.
- [5] JUMP (D. B.), « Fatty acid regulation of gene transcription », *Crit. Rev. Clin. Lab. Sci.*, 41, n° 1, 2004, p. 41-78.
- [6] KUIPERS (B. J.), *Qualitative reasoning. Modeling and simulation with incomplete knowledge*, MIT Press, 1994.
- [7] LE BORGNE (M.), DUTERTRE (B.), BENVENISTE (A.) et LE GUERNIC (P.), « Dynamical systems over Galois fields », dans *European Control Conference*, Groningen, June 1993, p. 2191-2196.
- [8] LEE (S. S.), CHAN (W. Y.), LO (C. K.), WAN (D. C.), et autres, « Requirement of pparalpha in maintaining phospholipid and triacylglycerol homeostasis during energy deprivation », *J Lipid Res.*, 45, n° 11, 2004, p. 2025-2037.
- [9] LEMER (C.), ANTEZANA (E.), COUCHE (F.), FAYS (F.), et autres, « The aMAZE LightBench : a web interface to a relational database of cellular processes », *Nucleic Acids Research*, 32, Database issue, 2004, p. D443-D448.
- [10] LIANG (G.), YANG (J.), HORTON (J. D.), HAMMER (R. E.), et autres, « Diminished hepatic response to fasting/refeeding and liver X receptor agonists in mice with selective deficiency of sterol regulatory element-binding protein-1c », *J Biol Chem*, 277, n° 15, Jan 2002, p. 9520-9528.
- [11] OGATA (H.), GOTO (S.), SATO (K.), FUJIBUCHI (W.), et autres, « KEGG : Kyoto encyclopedia of genes and genomes », *Nucleic Acids Research*, 27, n° 1, 1999, p. 29-34.
- [12] PAPIN (J. A.), STELLING (J.), PRICE (N. D.), KLAMT (S.), SCHUSTER (S.) et PALSSON (B. O.), « Comparison of network-based pathway analysis methods », *Trends in Biotechnology*, 22, 2004, p. 400-405.
- [13] RADULESCU (O.), LAGARRIGUE (S.), SIEGEL (A.), LE BORGNE (M.) et VEBER (P.), « Topology and linear response of interaction networks in molecular biology », 2005.
- [14] SCHACHERER (F.), CHOI (C.), GOTZE (U.), KRULL (M.), PISTOR (S.) et WINGENDER (E.), « The TRANSPATH signal transduction database : a knowledge base on signal transduction networks », *Bioinformatics*, 17, n° 11, 2001, p. 1053-1057.
- [15] SIEGEL (A.), RADULESCU (O.), LE BORGNE (M.), VEBER (P.), OUY (J.) et LAGARRIGUE (S.), « Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation networks », *Biosystems*, soumis (2005).
- [16] THOMAS (R.), « Regulatory networks seen as asynchronous automata : a logical description », *J. Theor. Biol.*, 153, 1991, p. 1-23.
- [17] TRAVÉ-MASSUYÉS (L.) et DAGUE (P.), *Modèles et raisonnements qualitatifs*, Hermes sciences, 2003.