Exploitation à large échelle de bases de données de connaissance sur les réactions et régulations pour trouver des régulateurs clés de groupe de gènes

Pierre Blavy, Florence Gondret, Sandrine Lagarrigue, Jaap J. van Milgen, Anne Siegel

Using large scale knowledge database about reactions and regulations to find key regulators of sets of genes

Pierre BLAVY[1,2,3], Florence GONDRET[1,2], Sandrine LAGARRIGUE[1,2],
Jaap VAN MILGEN[1,2]and Anne SIEGEL[3]

[1] INRA, UMR 1348 PEGASE, Domaine de la Prise, 35590 Saint-Gilles
[2] AgroCampus-Ouest, UMR 1348 PEGASE, 74 rue de Saint Brieuc, 35000 Rennes
[3] CNRS-Université de Rennes 1-INRIA, UMR 6074 IRISA, Campus de Beaulieu, 35042 RENNES Cedex

Abstract *Many experimental observations and known cellular mechanisms are available but large-scale analyses of such information remain difficult due to the heterogeneity of the biological mechanisms. In this context, we introduce a new automatic method to integrate this information in order to find potential key regulators of a set of molecules.*
First, we introduce the concept of "regulated reaction" to gather in a unified formalism information about reactions (consumption and prediction of molecules) and causalities (effect of the variation of a molecule on the variation of another molecule). This formalism is then modeled as a causality graph describing in a predictive way consequences of the variations of fluxes and molecules quantities. Finally, we use graph-based algorithms and biologically relevant scores to extract key regulators of a set of molecules. We validate this method by recovering among the causal graph derived from the Transpath database the main regulators of 190 groups of genes which are known to share a transcription factor according to the TRED database.

Keywords Large scale, supervised analysis, integration, key regulator

Exploitation à large échelle de bases de données de connaissances sur les réactions et les régulations afin de trouver des régulateurs clefs de groupes de gènes

Résumé *Bien que de nombreuses données expérimentales et connaissances relatives aux mécanismes cellulaires soient disponibles, leur analyse à large échelle reste difficile, notamment en raison de la diversité des mécanismes biologiques. Dans ce contexte, nous proposons une nouvelle méthode pour intégrer ces mécanismes afin d'identifier des régulateurs clefs potentiels d'ensembles de molécules.*
Dans un premier temps, nous proposons un formalisme unifié représentant des "réactions régulées" afin de réunir les informations relatives aux réactions (consommant et produisant des molécules) et aux causalités (conséquences de la variation d'une quantité de molécule sur une autre molécule). Dans un second temps, nous interprétons les régulations régulées sous la forme d'un graphe de causalités qui représente, de manière prédictive, l'effet des variations de quantités ou de flux sur les différents acteurs métaboliques et génétiques du système. Au final, nous développons une approche combinant des parcours de graphe et des scores biologiquement adéquats pour identifier les régulateurs clefs d'un ensemble de sommets connus pour être co-régulés. Nous validons notre approche en construisant un graphe de causalité à partir de Transpath, puis en recherchant les régulateurs clefs de 190 ensembles de molécules extraits de la base de données TRED dont le principal facteur de transcription est connu.

Mots-clés Large échelle, analyse supervisée, intégration, régulateur clef

1 Introduction

À grande échelle, les systèmes biologiques sont notamment décrits par deux types d'informations : d'une part, des données expérimentales sur l'expression des gènes (puce ADN, RNA-seq) [7] qui sont souvent qualitatives (ou quantitatives analysées qualitativement) et obtenues en comparant un faible nombre de conditions. Quelques études cinétiques ou concernant de nombreuses conditions expérimentales existent mais elles sont relativement rares. D'autre part, de nombreuses connaissances relatives aux mécanismes cellulaires sont présentes dans des bases de données de connaissances [17], dans la littérature ou peuvent aussi être obtenues sous forme synthétique, sur des problèmes particuliers, en questionnant les experts.

Pourtant, il est actuellement difficile d'interpréter à grande échelle les données d'expression dans le cadre des connaissances sur les mécanismes afin, par exemple, d'identifier les régulateurs d'un ensemble de molécules ou de gènes. Plusieurs raisons sont généralement invoquées pour expliquer ces difficultés : les systèmes observés par les données d'expression comportent plusieurs milliers de molécules, ce qui nécessite des approches qualitatives et à grande échelle sur un formalisme unifié des mécanismes moléculaires. Or, l'information contenue dans les bases de données n'est pas disponible de manière uniforme : les bases combinent souvent des interactions agissant à des échelles très différentes (de la transcription au métabolisme), et contiennent à la fois des mécanismes de type réactionnels (transformation d'un produit en un autre) et des mécanismes de type causaux (effet d'un produit sur un autre, sans consommation du produit initial).

Classiquement, l'analyse haut débit des données d'expression comparant quelques conditions [14,4] commence par l'identification des ensembles de gènes différentiellement exprimés entre conditions. Les groupes de gènes corrélés sont construits puis analysés à l'aide d'annotations issues d'ontologies. On obtient au final des groupes de gènes caractérisés par un ensemble d'attributs souvent relatifs à leur fonction. Ces étapes sont généralement réalisées automatiquement, y compris à l'échelle d'une puce pangénomique. Dans un second temps, une analyse plus fine est réalisée par le biologiste, soit en sélectionnant un groupe de gènes puis en faisant la bibliographie de ses éléments, soit en se concentrant sur un ensemble de gènes relatifs à son champ d'expertise. Cette partie permet de prendre en compte les connaissances précises relatives à la régulation ou à l'effet des éléments sélectionnés. Cependant, cette dernière étape est réalisée manuellement, et ne peut donc pas être appliquée à l'ensemble des éléments d'intérêts. Le travail présenté ici vise à automatiser en partie cette dernière étape.

Il existe plusieurs approches permettant d'intégrer à large échelle les connaissances relatives aux mécanismes moléculaires afin d'analyser des données qualitatives. Un premier groupe de méthodes, composé des réseaux booléens [15], de leur généralisation asynchrone [5] ou stochastique [21], des réseaux logiques généralisés [3] et systèmes d'équations différentielles linéaires par morceaux [8,20] permet d'étudier des trajectoires (*i.e.,* une succession d'états au cours du temps). Cet ensemble de méthodes est basé sur la modélisation qualitative de la dynamique des systèmes. Il s'appuie donc sur des données cinétiques portant sur les variations de quantités de produits par rapport à un ou des seuils, données qui sont encore difficiles à produire. De plus, les espaces des trajectoires possibles croissent de manière exponentielle en fonction de la taille du système, rendant impossible des études automatiques à grande échelle.

Une alternative à ces méthodes de modélisation des propriétés temporelles des systèmes biologiques consiste à se concentrer sur les causalités, c'est-à-dire les effets des variations de quantité des molécules sur le système en ne modélisant pas les phénomènes de consommation. Dans ce cas, les études se concentrent sur la propagation des effets des régulations le long d'un réseau et sur leur compatibilité avec des observations différentielles entre deux états d'un système (par exemple l'effet d'un stress environnemental ou d'une perturbation génétique). Ces approches causales sont principalement appliquées à la reconstruction de réseaux biologiques avec des méthodes statistiques, par exemple bayésiennes

[11], ou à la prédiction du comportement d'un réseau par des approches formelles [10,9]. Par construction, ces approches modélisent les causalités (*i.e.,* les conséquences de la variation d'une quantité de molécules sur la quantité d'autres molécules) ce qui les rend parfaitement adaptées à la modélisation des régulations génétiques, souvent exprimées en terme d'activation ou d'inhibition. Cependant, elles sont inadaptées pour modéliser directement les effets des réactions (*i.e.,* les consommations et productions de molécules). Pour dépasser cette limite, nous avons développé une interprétation des réactions dans le cadre des causalités.

Brièvement, nous introduisons d'abord le concept de "réaction régulée", pour représenter de manière unifiée les connaissances relatives aux réactions et aux causalités. Ce formalisme permet d'intégrer les connaissances classiquement disponibles dans les différentes bases de connaissances telles que Transpath [2], KEGG [18] ou Pathway Commons [1] [1] (c.f. partie 2.1). Dans un second temps, nous associons à toute réaction régulée un graphe de causalité. Ceci est rendu possible, sous une hypothèse de quasi-stationnarité, en introduisant différents types de sommets représentant les flux ou les quantités de molécules (c.f. partie 2.2). Cette démarche est résumée Fig. 1. Enfin, nous développons une approche s'appuyant sur des parcours du graphe de causalités pour identifier les régulateurs clefs d'un ensemble de gènes coexprimés. Cette démarche est validée en construisant l'ensemble des "réactions régulées" et le graphe de causalité associé à partir de Transpath, puis en utilisant ce dernier graphe pour rechercher les régulateurs clefs de 190 groupes de gènes connus pour être régulés par un facteur de transcription. Ces groupes sont extraits de TRED [2] [13] (c.f. partie 2.3).

2 Matériel et méthodes

2.1 Formalisme unifié des réactions et des régulations

Les formalismes existants pour représenter les mécanismes de régulation sont de deux types : d'une part, les réactions permettent de décrire les consommations et productions de molécules ; d'autre part, les causalités permettent de décrire l'effet d'une modification de la quantité de la source d'une interaction sur ses cibles sans consommer la source de l'interaction. Ces deux points de vue (réaction *Vs.* causalité) coexistent dans les banques de données portant sur les interactions biologiques.

Pour unifier ces deux points de vue, nous introduisons le concept de "réactions régulées". Ces objets sont constitués de substrats, de produits, d'activateurs, d'inhibiteurs, de modulateurs (*i.e.,* une molécule étant soit activateur, soit inhibiteur) et d'un booléen indiquant si la réaction est réversible ou irréversible. Ce formalisme très simple permet de représenter l'information utile à notre contexte à partir de la plupart des bases de connaissances existantes, et fait la distinction entre les flux de matière (substrats et produits) et les régulateurs (activateurs, inhibiteurs, modulateurs) qui influencent les vitesses des réactions sans être consommés. Sur la Fig. 1, nous décrivons les règles de transformations d'interactions causales en réactions régulées.

En pratique, l'ensemble des réactions décrites dans Transpath [16] a été extrait et codé dans ce formalisme. Les molécules ATP, ADP, *protein remanants*, NDP, NTP, sp1, phosphate, Coenzyme-A, eau et H+ ont été considérées comme jamais limitantes *a priori* et ont été retirées. De plus, afin de prendre en compte l'aspect multi-espèces des connaissances de Transpath, les réactions régulées ayant les mêmes ensembles de substrats et produits ont été fusionnées en faisant l'union de leurs régulateurs. Si au moins une réaction régulée est réversible, la fusion est considérée comme réversible. Cette règle permet de décrire l'ensemble des réactions présentes quelque soit l'espèce dans laquelle elles ont été observées. Nous avons obtenu un graphe de grande taille (c.f. Table 1) fortement connexe. Ce graphe a une structure *scale-free* [12] : un petit nombre de molécules très connectées (*i.e.,* les *hubs*) est relié à un grand nombre

[1] Pour une revue des principales bases, voir [19] et http://pathguide.org
[2] Tred est disponible sur http://rulai.cshl.edu/TRED/

de molécules très peu connectées. Privé des 1000 molécules intervenant dans le plus de réactions (*i.e.,* les principaux *hubs*), le graphe devient très faiblement connecté.

2.2 Interprétation des réactions régulées en graphe de causalité

L'ensemble des réactions régulées permet de représenter la connaissance mais il n'exprime pas directement les effets des variations de quantité de molécules. Dans la Table 2, nous introduisons des règles d'interprétation de réactions régulées en graphes de causalité à l'aide de noeuds distincts pour les flux (noeuds : "*disponibilité en substrat*" et "*vitesse de réaction*") et les quantités de molécules (noeuds de "*quantité*"). Ces règles sont une interprétation qualitative des coefficients d'élasticité [6] sous l'hypothèse quasi-stationnaire. Par la suite elles sont utilisées pour propager les effets des régulations sans prendre en compte les effets globaux de la dynamique du système. On obtient ainsi une surapproximation des comportements possibles du système.

Lors de la construction du graphe de causalités associé à Transpath, nous avons fait l'hypothèse de considérer toutes les réactions comme irréversibles. En effet, la majeure partie des réactions réversibles de Transpath correspond à des inhibitions par formation de complexes pour lesquelles le sens complexe→substrats peut être négligé, car aucune autre réaction ne produit le complexe.

Afin de pouvoir faire facilement le lien entre les observations sur les molécules et les noeuds du graphe de causalité, on associe à chaque noeud de type "*quantité*" et à chaque noeud de type "*disponibilité en substrat*", la molécule à laquelle ils se rapportent.

Cette étape induit une augmentation dans la taille du graphe (nombre de noeuds $\times$ 1,4, nombre d'arêtes $\times$ 4,9) en raison de l'introduction des différents types de noeuds et des nombreux liens lorsque plusieurs réactions concernent la même molécule (c.f. Table 1).

2.3 Recherche de régulateurs clefs

Dans le cadre de cette étude, nous avons développé une méthode pour proposer un ensemble de molécules qui sont ordonnées par leur capacité à réguler (directement ou indirectement) un autre ensemble de molécules cibles. Ces cibles sont définies *a priori* en prenant par exemple un *cluster* de gènes coexprimés issu d'une analyse à haut débit par puces à ADN, ou un groupe de métabolites d'intérêt issu d'une synthèse bibliographique.

La méthode prend comme entrée un ensemble de réactions régulées $\mathcal{R}$ et un ensemble de *molécules* cibles $\mathcal{S}$. Elle se base sur le calcul sous contraintes de la fermeture transitive du graphe de causalité, afin d'évaluer pour chaque noeud, les noeuds dont il peut expliquer la variation. Elle est décrite par les étapes ci-dessous, dans lesquelles le terme *molécule* désigne toujours une molécule dans un ensemble de réactions régulées et le terme *noeud* un sommet d'un graphe de causalité.
(**1**) Le but de cette étape est de sélectionner dans l'ensemble des réactions régulées $\mathcal{R}$ un sous-ensemble en lien avec les cibles. Pour cela, nous sélectionnons les réactions régulées de $\mathcal{R}$ situées à moins de 3 réactions de l'ensemble des *molécules* cibles $\mathcal{S}$ sans passer par les *hub* de $\mathcal{R}$. Le graphe de réactions régulées obtenu est noté $\mathcal{R}(\mathcal{S})$. Comme indiqué partie 2.1, une *molécule* est considérée comme un *hub* si elle fait partie des 1000 premières *molécules* impliquées dans le plus de réactions de $\mathcal{R}$.
(**2**) L'ensemble des réactions régulées sélectionnées $\mathcal{R}(\mathcal{S})$ est converti en graphe de causalité noté $\mathcal{C}(\mathcal{S})$ à l'aide des règles décrites Fig. 2. Au cours de cette conversion, une relation d'*association* est définie entre chaque *molécule* M_1 de $\mathcal{R}(\mathcal{S})$ et les *noeuds* : *disponibilité en* M_1 et *quantité de* M_1 dans $\mathcal{C}(\mathcal{S})$.
(**3**) A chaque *noeud* N_1 du graphe $\mathcal{C}(\mathcal{S})$, nous associons deux ensembles de *molécules* $\mathcal{U}^+_{N_1}$ et $\mathcal{U}^-_{N_1}$. L'ensemble $\mathcal{U}^+_{N_1}$ contient la molécule M_2 si la propagation de l'effet d'une augmentation de N_1 au travers d'un chemin entre N_1 et un noeud N_2 *associé* à M_2 est cohérente avec les variations observées le long

de ce chemin. Ces ensembles sont calculés par la procédure[3] ci-dessous illustrée Fig. 2.

- **Initialisation** Pour toute *molécule* M_1 et tout *noeud* N_1 *associé* à M_1, si M_1 augmente ou n'est pas observée alors $\mathcal{U}^+_{N_1}$=$\{M_1\}$, sinon $\mathcal{U}^+_{N_1}$=$\emptyset$.
- **Chemins cohérents** Pour tout *noeud* N_1 et toute molécule *molécule* M_3, M_3 est ajoutée à $\mathcal{U}^+_{N_1}$ si a) il y a cohérence avec les observations relatives à N_1 et b) il y a cohérence avec au moins un *noeud* N_2 sucesseur de N_1.
  - a) est vraie si : la molécule M_1 *associée* à N_1 augmente ou si M_1 n'est pas observée.
  - b) est vraie pour N_2 un sucesseur de N_1 :
    - si : $N_1 \overset{+}{\longmapsto} N_2$ et $\mathcal{U}^+_{N_2}$ contient M_3
    - ou si : $N_1 \overset{-}{\longmapsto} N_2$ et $\mathcal{U}^-_{N_2}$ contient M_3
    - ou si : $N_1 \overset{?}{\longmapsto} N_2$ et ($\mathcal{U}^+_{N_2}$ contient M_3 ou $\mathcal{U}^-_{N_2}$ contient M_3)
- **Propagation** Tant qu'au moins un ensemble $\mathcal{U}^+_{N_1}$ ou $\mathcal{U}^-_{N_1}$ a été mis à jour pour un sommet N_1, le calcul des chemins cohérents est répété. Ceci permet de prendre en compte les régulations indirectes.

(**4**) Afin d'étendre le concept précédent aux molécules du graphe de réactions, nous calculons pour chaque *molécule* M_1 de $\mathcal{R}(\mathcal{S})$ *associée* à l'ensemble de noeuds N deux ensembles : $\mathcal{V}^+_{M_1} = \bigcup_{\forall a \in N} U^+_a$ et $\mathcal{V}^-_{M_1} = \bigcup_{\forall a \in N} U^-_a$. Les ensembles $\mathcal{V}^+_{M_1}$ et $\mathcal{V}^-_{M_1}$ contiennent les molécules cibles (directes et indirectes) de M_1.

(**5**) Pour chaque molécule Ma de $\mathcal{R}(\mathcal{S})$, et chaque cas c (c=+ si Ma augmente, - sinon) nous calculons trois scores :

- le score de **couverture** égal au nombre de molécules appartenant à la fois à $\mathcal{V}^c_{Ma}$ et aux cibles $\mathcal{S}$.
- le score de **spécificité** égal à un moins la probabilité d'obtenir un nombe de molécules cibles régulées $\geq$ à ce qui est attendu au hasard. Cette probalité est estimée par un test hypergéométrique.
- le score de **couverture spécifique** égal au produit des deux scores précédents.

En ordonnant les différentes molécules en fonction de ces scores, nous sommes capables de fournir au biologiste une liste ordonnée de candidats ; les meilleurs étant ceux de score maximal.

2.4 Validation de la méthode

TRED est une base de connaissances décrivant un ensemble de facteurs de transcription associés à leur cibles, obtenue chez l'homme, la souris et le rat. Un script a été développé pour extraire à partir de l'interface web des couples "facteurs de transcription – cibles de ce facteur". Seuls les couples avec un nombre de cibles supérieur à 5 ont été retenus. Moins de 5 cibles est considéré comme une information bien trop faible pour justifier une étude à haut débit. 190 couples ont été extraits.

Ces couples sont ensuite utilisés pour valider notre méthode. Nous utilisons la méthode de recherche de régulateurs clefs décrite précédemment afin de proposer un ensemble de candidats. Nous considérons ensuite un test comme un succès si le facteur de transcription connu se retrouve parmi les 50 premiers gènes candidats. Le nombre de 50 a été choisi car il correspond à ce qu'un biologiste peut facilement analyser manuellement dans un temps raisonnable.

Dans un second temps, nous avons analysé manuellement les candidats proposés par notre méthode pour 19 cibles de PPARα choisies par expertise manuelle (ACSL1, ACSL3, ACSL4, ACSL5, ACSL6, CD36, ACADVL, ACADL, ACAA2, CPT1A, CPT1B, 3HCDH, 3KACT, HADHA, ETFDH, HMGCS2, FADS1, FADS2, SCD).

[3] La procédure est décrite pour $\mathcal{U}^+_{N_1}$. La procédure pour $\mathcal{U}^-_{N_1}$ est similaire, il suffit d'échanger dans la description $\mathcal{U}^+_{N_1}$ avec $\mathcal{U}^-_{N_1}$ et de remplacer augmentation par diminution.

3 Résultats et discussion

Lorsque notre méthode est appliquée entièrement : le score de couverture spécifique a un taux de succès de 58%. Ce résultat démontre l'efficacité de notre méthode pour proposer automatiquement au biologiste des régulateurs clefs potentiels à partir d'un ensemble de cibles. Lorsque le résultat des tests est calculé aléatoirement, mais que les sous-ensembles du graphe de réaction sont calculés à partir des bonnes cibles, 9% des tests sont des succès, ce taux est dû à la relative pertinence de notre méthode de sélection de sous-graphes. Lorsque les noms des cibles décrites dans TRED sont échangés au hasard, ce score tombe à 2%, ce qui nous assure que notre méthode produit bien des candidats spécifiques des cibles, et non une liste de régulateurs ubiquitaires.

Pris séparément, les scores de couverture et de spécificité donnent des résultats moins bons que leur produit. Ceci signifie qu'un bon candidat doit à la fois bien couvrir les molécules cibles, sans en réguler de nombreuses autres à coté. Dans l'avenir, on peut imaginer optimiser la méthode en testant différentes pondérations des scores de spécificité et de couverture.

Parmi les 80 échecs, 7 s'expliquent car le facteur de transcription connu n'est pas présent dans le graphe total des réactions régulées, un problème dû soit à un mauvais lien entre les identifiants de Transpath et ceux de TRED, soit à un manque de connaissance dans Transpath.

Une analyse manuelle de quelques échecs nous a permis de mettre en évidence des ensembles de cibles très vastes (plus de 100 cibles) contenant un grand nombre de cibles peu spécifiques du facteur de transcription auxquelles elles sont associées dans TRED. Il n'est donc pas du tout surprenant de retrouver parmi les régulateurs candidats de ces cibles de nombreuses molécules ayant une pertinence biologique mais différentes de la solution du test. De plus, de nombreux facteurs de transcription ont un très vaste nombre de cibles (dont la plupart sont inconnues ou absentes des bases), ce qui rend la réalisation de tests difficile et explique probablement certains échecs. Au final, la méthode développée reste pertinente pour rechercher des régulateurs clefs potentiels.

Par exemple, en utilisant le score de couverture, l'analyse des gènes candidats pour expliquer les cibles de PPARα (c.f. partie 2.4) fait ressortir un ensemble de 28 gènes qui expliquent 9 ou 10 cibles. Parmi eux, on retrouve comme attendu des gènes liés à la voie de signalisation des PPAR (PPARα,PPARγ, PPARδ, RXRα, NCOA1, PPARGC1A, FADS1 et ALOX15). La présence de RXRα et de ALOX15 s'explique par l'implication des acides gras dans la voie de signalisation de PPARα (divers acides gras ont un score de couverture de 9), ce qui illustre l'intérêt d'inclure dans le système les métabolites. De plus, on retrouve de nombreux régulateurs ubiquitaires du métabolisme de l'énergie (insuline et NRIP1) de la réponse hormonale (CREB1, VDR, NR2F1, NRIP1) et du cycle cellulaire (SP1,SP3,SRF). Parmi les gènes restants, 3 sont impliqués dans le métabolisme des lipides mais le lien avec les cibles n'est pas trivial et 6 ont un lien qui reste à élucider.

Conclusion et perspectives

Nous avons développé un formalisme permettant de représenter de manière simple et unifiée les informations courantes décrivant les mécanismes cellulaires impliqués dans les voies de régulation et dans le métabolisme. Nous avons ensuite proposé une méthode pour interpréter ce formalisme sous la forme d'un graphe de causalité en distinguant, par des noeuds différents, les concepts de flux, de vitesse de réaction et de quantités de molécules. Au final, nous avons démontré sur un jeu de test que ce formalisme est pertinent pour proposer automatiquement au biologiste un ensemble de régulateurs clefs d'un groupe de molécules d'intérêt.

Dans l'avenir, nous souhaitons tout d'abord améliorer la méthode de recherche de régulateurs candidats en développant de nouveaux scores, ou en prenant en compte de manière fine l'effet combiné de plusieurs molécules, puis utiliser les scores développés afin d'analyser des jeux de données réels tels que les clusters de gènes identifiés dans les études transcriptionnelles haut débit.

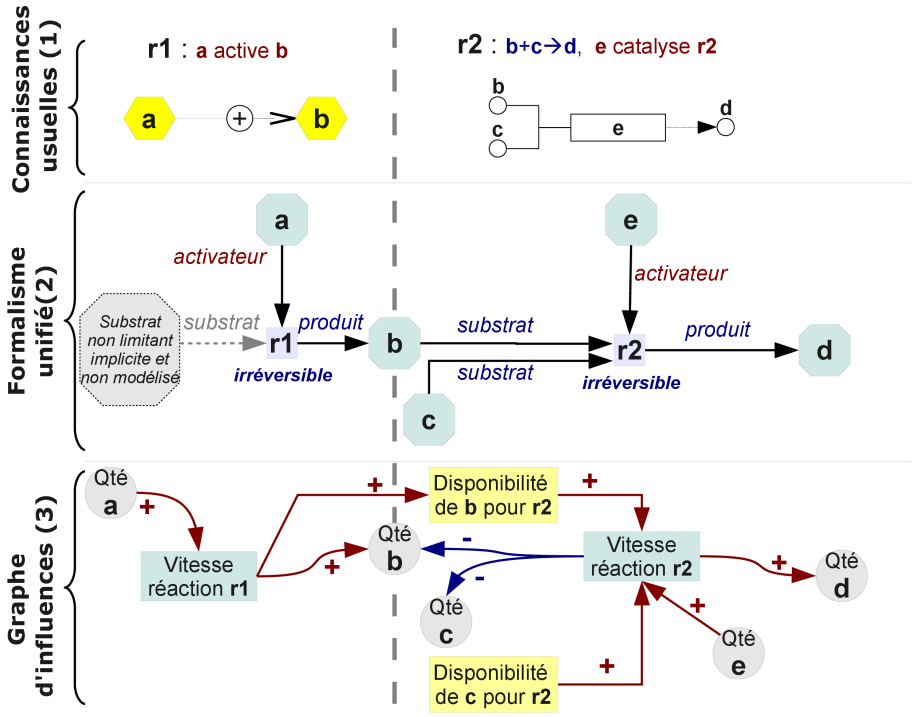


Figure 1. Formalisme unifié pour les réactions et les régulations : interprétation en graphe de causalité

(**1**) Les connaissances usuelles relatives au métabolisme et à sa régulation modélisent deux concepts différents : les *reactions* (b+c→d) qui consomment ou produisent des molécules et les *effets* (**a** active **b** et **e** catalyzes **r2**) qui décrivent les conséquences de la variation d'une quantité de molécule sans consommer cette dernière.

(**2**) Ces deux concepts sont unifiés dans un formalisme commun : les *réactions régulées*, en supposant que les effets (comme **a** active **b**) sont des réactions dont la vitesse est régulée par le régulateur (**a**) qui consomment un substrat non limitant non modélisé et qui produisent l'élément régulé (**b**). Ce formalisme fait clairement la distinction entre les flux ($\overset{substrats}{\longmapsto}$ et $\overset{produit}{\longmapsto}$) et les régulations de la vitesse de réaction par des molécules non consommées ($\overset{activateur}{\longmapsto}$, $\overset{inhibiteur}{\longmapsto}$ et $\overset{modulateur}{\longmapsto}$).

(**3**) Sous l'hypothèse que la vitesse de réaction est une fonction croissante de la disponibilité de chacun des substrats, une fonction croissante de la quantité d'activateur, une fonction décroissante de la quantité d'inhibiteur et une fonction monotone de la quantité de modulateur, les *réactions régulées* peuvent être interprétées sous forme de graphe de causalités. Ce dernier modélise les conséquences ($\overset{+}{\longmapsto}$,$\overset{-}{\longmapsto}$,$\overset{?}{\longmapsto}$) des variations des flux (rectangles : vitesses de réactions et disponibilité) et des quantités de molécules (cercles).

Table 1. Modélisation de Transpath en réaction régulées et en causalités.

(a) Transpath

Molécules	**158 545**
métabolites	122 951
genes	35 594
Relations	**224 080**
effect +	2 493
effect -	854
effect ?	10 251
reaction réversible	41 278
reaction irréversible	110 838
gene→protein	59 956

(b) Reactions régulées

Noeuds	**291 306**
mol ou gènes	132 752
reactions rév	40 541
reactions irr	118 013
Arêtes	**383 732**
substrats	27 418
produits	115 630
activateurs	18 813
inhibiteurs	498
modulateurs	9 307

(c) Graphe de causalité

Noeuds	**422 652**
quantité	131 065
vitesse reac.	109 408
disponibilité	181 179
Arêtes	**1 891 492**
$\overset{+}{\longmapsto}$	1 656 275
$\overset{-}{\longmapsto}$	217 279
$\overset{?}{\longmapsto}$	17 938

(a) Transpath est une base de données dévrivant des réactions et leurs régulations (b) Les réactions et les effets sont unifiés sous forme de "réactions régulées" (c.f. 2.1) (c) Les "réactions régulées" sont ensuite converties en graphe de causalité (c.f. 2.2) afin d'exprimer les conséquences des variations de quantité de molécules, de flux et de vitesse de réactions. Au final, l'introduction de nouveau noeuds crée un graphe contenant 1.5 fois plus de noeuds et 4.9 fois plus d'arêtes que le graphe d'origine. Ceci s'explique notamment par les liens créés lorsque une même molécule intervient dans de nombreuses réactions.

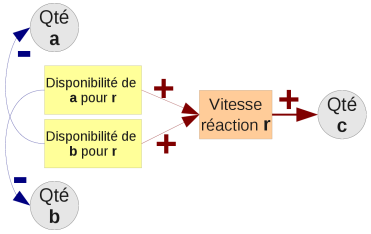
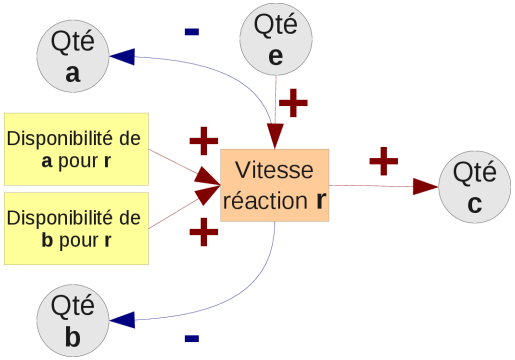
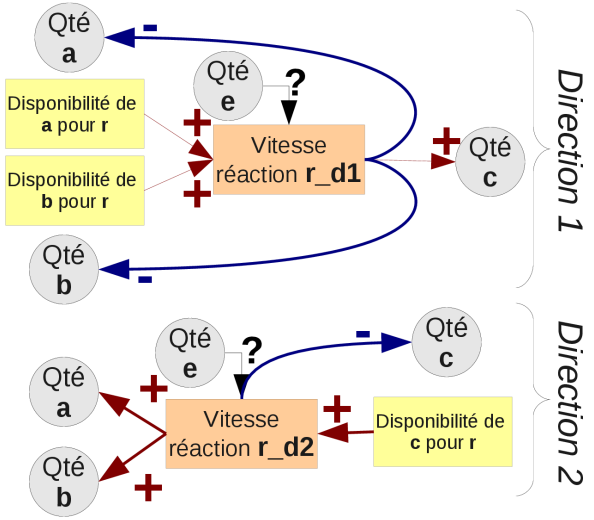
Table 2. Règles de création d'un graphe de causalités à partir de réactions régulées

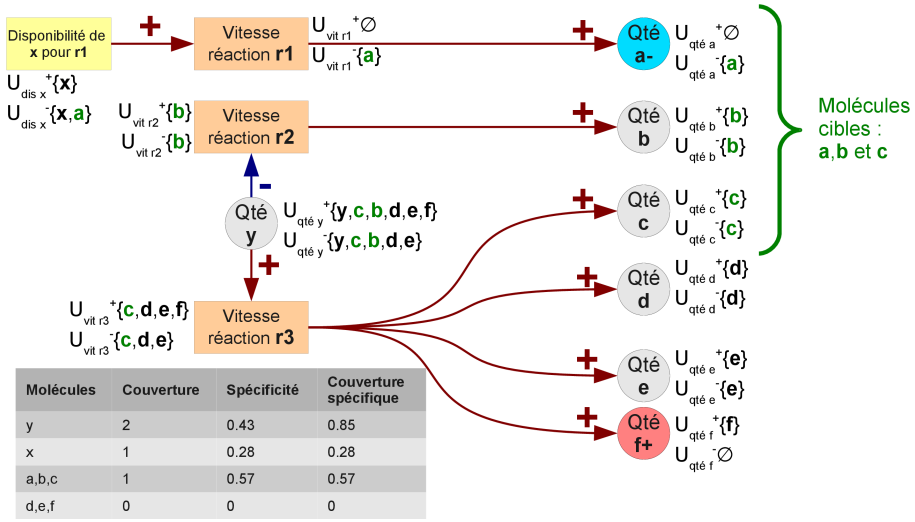
Chaque *reaction régulée* est modélisée sous la forme d'un graphe de causalités à l'aide des règles du tableau. Ensuite, chaque vitesse de réaction produisant une molécule **a** est liée par $\overset{+}{\longmapsto}$ à la disponibilité en **a** pour les autres réactions. Les lettres SPAIM représentent respectivent les ensembles de substrats, produits, activateurs, inhibiteurs et modulateurs, décrivant les *réactions régulées* ; $\emptyset$ désigne l'ensemble vide. Une réaction est considérée comme sans régulateur si elle n'a ni activateur ni inhibiteur ni modulateur. Ces règles décrivent les effets locaux des variations de flux et de quantités sous l'hypothèse quasistationnaire sans prendre en compte les effets globaux de la dynamique du système. On obtient ainsi une sur-approximation des comportements du système.

Règles

Irréversible sans régulateur	Irréversible avec régulateur	Réversible
$\forall \mathbf{s} \in S, \forall \mathbf{p} \in P$, disponibilité de s pour $r \overset{+}{\longmapsto}$ vitesse de la réaction r. vitesse de la réaction $r \overset{+}{\longmapsto}$ quantité de p $\forall \mathbf{s1}, \mathbf{s2} \in S^2, \mathbf{s1} \neq \mathbf{s2}$, disponibilité de $s1$ pour $r \overset{-}{\longmapsto}$ quantité de $s2$.	$\forall \mathbf{s} \in S, \forall \mathbf{p} \in P$ disponibilité de s pour $r \overset{+}{\longmapsto}$ vitesse de la réaction r, vitesse de la réaction $r \overset{-}{\longmapsto}$ quantité de s, vitesse de la réaction $r \overset{+}{\longmapsto}$ quantité de p. $\forall \mathbf{a} \in A, \forall \mathbf{i} \in I, \forall \mathbf{m} \in M$, quantité de $a \overset{+}{\longmapsto} r$ vitesse de la réaction r, quantité de $i \overset{-}{\longmapsto} r$ vitesse de la réaction r, quantité de $m \overset{?}{\longmapsto} r$ vitesse de la réaction rx.	$\forall \mathbf{s} \in S, \forall \mathbf{p} \in P$, disponibilité de s pour $r \overset{+}{\longmapsto}$ vitesse de la réaction r_{d1}, disponibilité de p pour $r \overset{+}{\longmapsto}$ vitesse de la réaction r_{d2}, vitesse de la réaction $r_{d1} \overset{-}{\longmapsto}$ quantité de s, vitesse de la réaction $r_{d2} \overset{-}{\longmapsto}$ quantité de p, vitesse de la réaction $r_{d1} \overset{+}{\longmapsto}$ quantité de p, vitesse de la réaction $r_{d2} \overset{+}{\longmapsto}$ quantité de s. $\forall \mathbf{o} \in A \bigcup I \bigcup M$, quantité de $o \overset{?}{\longmapsto}$ vitesse de la réaction r_{d1}, quantité de $o \overset{?}{\longmapsto}$ vitesse de la réaction r_{d2}.

Examples

a+b→c S={**a**,**b**},P={**c**},A=I=M= ∅,irreversible	**a+b→c**, enzyme **e** S={**a**,**b**},P={**c**}, A={**e**},I=M=∅,irreversible	**a+b↔c**, enzyme **e** S={**a**,**b**},P={**c**}, A={**e**},I=M=∅,reversible
		



Molécules	Couverture	Spécificité	Couverture spécifique
y	2	0.43	0.85
x	1	0.28	0.28
a,b,c	1	0.57	0.57
d,e,f	0	0	0

Figure 2. Exemple de recherche de régulateurs clefs dans le graphe de causalités à partir d'un ensemble de molécules cibles

Dans cet exemple, nous souhaitons identifier les régulateurs clefs de trois molécules cibles : **a**, **b** et **c** à l'aide de la procédure décrite dans la partie 2.3. Nous observons deux molécules : **a** et **f** qui augmentent. Nous supposons que la partie (1) de la procédure nous a conduit à travailler sur l'ensemble de réactions régulées suivant : la réaction irréversible **r1** de substrat **x** et de produit **a**, la réaction irréversible **r2** de produit **b** et d'inhibiteur **y** et la réaction irréversible **r3** de produits **c**, **d**, **e** et **f** et d'activateur **y**.

Dans un premier temps l'ensemble des réactions régulées est converti en graphe de causalité tel que représenté sur la figure. Les noeuds de *quantité* (Qté) et de *disponibilité en substrats* sont mis en relations avec la molécule qu'ils décrivent. Ainsi, le noeud : *disponibilité de **x** pour **r1*** est mis en relations avec la molécule **x**, le noeud *quantité de **a*** avec la molécule **a** et ainsi de suite pour toutes les autres molécules. Seuls les noeuds de *vitesse de réaction* sont associés à aucune molécule. Le sens de variation des molécules observées est représenté sur les noeuds associés à ces molécules dans le graphe de causalités (*Qté de **a*** - et *Qté de **f*** +).

Pour chaque noeud N_1 associé à une molécule M_1, les ensembles $\mathcal{U}^+_{N_1}$ et $\mathcal{U}^-_{N_1}$ sont initialisés par la molécule associée M_1 en tenannt compte des observations. Ainsi, le noeud *Qté de **a*** associé à la molécule **a** qui décroît, va être initialisé par $\mathcal{U}^+_{Qté\ de\ a}=\emptyset$ et $\mathcal{U}^-_{Qté\ de\ a}=\{\mathbf{a}\}$. Le noeud *Qté de **b***, associé à une molécule non observé sera initialisé avec : $\mathcal{U}^+_{Qté\ de\ b}=\mathcal{U}^-_{Qté\ de\ b}=\{\mathbf{b}\}$.

Les éléments des ensembles prédécents sont propagés récursivement aux prédecesseurs de manière à identifier pour chaque sommet les molécules qu'il régule, de manière cohérente avec les observations et les signes des influences (i.e. $\overset{+}{\longmapsto}$, $\overset{-}{\longmapsto}$ et $\overset{?}{\longmapsto}$). Par exemple, lorsque la *vitesse de réaction **r1*** augmente, elle régule **a** de manière cohérente avec les influences (il y a un chemin positif entre ce noeud et le noeud *Qté de **a***), et avec les observations : **a** étant observé comme diminuant, le noeud *Qté de **a*** explique bien **a** quand il diminue ($\mathbf{a} \in \mathcal{U}^-_{vit.\ r1}$) mais pas quand il augmente ($\mathbf{a} \notin \mathcal{U}^+_{vit.\ r1}$). Le résultat final de cette étape est représenté sur la figure par les ensembles $\mathcal{U}^+_n$ et $\mathcal{U}^-_n$ associés à chaque noeud n.

Finalement, les ensembles relatifs aux noeuds $\mathcal{U}^+_n$ (et $\mathcal{U}^-_n$) qui sont en relations avec une même molécule sont fusionnés pour se rapporter aux molécules associées à ces noeuds. Ensuite les scores de *couverture* et de *specificité* sont calculés. Dans cet exemple, la molécule **y** a une bonne couverture, car elle régule de nombreuses cibles (**b** et **c**) et une mauvaise spécificité car elle régule aussi beaucoup de molécules non cibles (**y**, **d**, **e** et **f**). D'un autre coté, **a**, **b** et **c** ont un mauvais score de couverture (ils ne régulent qu'une molécule : eux-mêmes), mais une meilleure spécificité car ils ne régulent aucune molécule non cible. En faisant le produit du score de couverture et de spécificité, on obtient un compromis qui fait ressortir **y** comme la meilleure explication des cibles.

Références

[1] E.G. Cerami, B.E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G.D. Bader, and C. Sander. Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl 1) :D685–D690, 2011.

[2] C. Choi, M. Krull, A. Kel, O. Kel-Margoulis, S. Pistor, A. Potapov, N. Voss, and E. Wingender. TRANSPATH®a high quality database focused on signal transduction. *Comparative and functional genomics*, 5(2) :163–168, 2004.

[3] H. De Jong. Modeling and simulation of genetic regulatory systems : a literature review. *Journal of computational biology*, 9(1) :67–103, 2002.

[4] C. Desert, M.J. Duclos, P. Blavy, F. Lecerf, F. Moreews, C. Klopp, M. Aubry, F. Herault, P. Le Roy, C. Berri, et al. Transcriptome profiling of the feeding-to-fasting transition in chicken liver. *BMC genomics*, 9(1) :611, 2008.

[5] A. Faure, A. Naldi, C. Chaouiya, and D. Thieffry. Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. *Bioinformatics*, 22(14), 2006.

[6] David Fell. *Understanding the Control of Metabolism*. Portland Press, 1997.

[7] H. Ge, A.J.M. Walhout, and M. Vidal. Integrating [] omic'information : a bridge between genomics and systems biology. *TRENDS in Genetics*, 19(10) :551–560, 2003.

[8] J.L. Gouzé and T. Sari. A class of piecewise linear differential equations arising in biological models. *Dynamical Systems : An International Journal*, 17(4) :299–316, 2002.

[9] R.M. Gutiérrez-Ríos, D.A. Rosenblueth, J.A. Loza, A.M. Huerta, J.D. Glasner, F.R. Blattner, and J. Collado-Vides. Regulatory network of Escherichia coli : consistency between literature knowledge and microarray profiles. *Genome research*, 13(11) :2435, 2003.

[10] C. Guziolowski, A. Bourdé, F. Moreews, and A. Siegel. BioQuali Cytoscape plugin : analysing the global consistency of regulatory networks. *BMC genomics*, 10(1) :244, 2009.

[11] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano. Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *Journal of Bioinformatics and Computational Biology*, 2(1) :77, 2004.

[12] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804) :651–654, 2000.

[13] C. Jiang, Z. Xuan, F. Zhao, and MQ Zhang. Tred : a transcriptional regulatory element database, new entries and other development. *Nucleic acids research*, 35(suppl 1) :D137–D140, 2007.

[14] Y.H. Jin, P.E. Dunlap, S.J. McBride, H. Al-Refai, P.R. Bushel, and J.H. Freedman. Global transcriptome and deletome profiles of yeast exposed to transition metals. *PLoS Genetics*, 4(4), 2008.

[15] G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol*, 9(10) :770–780, 2008.

[16] M. Krull, S. Pistor, N. Voss, A. Kel, I. Reuter, D. Kronenberg, H. Michael, K. Schwarzer, A. Potapov, C. Choi, et al. Transpath® : an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic acids research*, 34(suppl 1) :D546–D551, 2006.

[17] J. Labaer. Mining the literature and large datasets. *Nature Biotechnology*, 21(9) :976–977, 2003.

[18] Kanehisa M. and S. Goto. KEGG : Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res. 28, 27-30*, 2000.

[19] S. Mathivanan, B. Periaswamy, TKB Gandhi, K. Kandasamy, S. Suresh, R. Mohmood, YL Ramachandra, and A. Pandey. An evaluation of human protein-protein interaction data in the public domain. *BMC bioinformatics*, 7(Suppl 5) :S19, 2006.

[20] EI Prokudina, RY Valeev, and RN Tchuraev. A new method for the analysis of the dynamics of the molecular genetic control systems. II : application of the method of generalized threshold models in the investigation of concrete genetic systems. *Journal of theoretical biology*, 151(1) :89–110, 1991.

[21] I. Shmulevich, E.R. Dougherty, and Wei Zhang. From Boolean to Probabilistic Boolean Networks as Models of Genetic Regulatory Networks. *Proceedings of the IEEE*, 90(11), 2002.