HAL open science

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

Anne Siegel, Carito Guziolowski, Philippe Veber, Ovidiu Radulescu, Michel Le Borgne

To cite this version:

Anne Siegel, Carito Guziolowski, Philippe Veber, Ovidiu Radulescu, Michel Le Borgne. Optimiser un plan d'expérience à partir de modèles qualitatifs?. Biofutur, 2007, 275, pp.27-31. inria-00178791

HAL Id: inria-00178791

https://inria.hal.science/inria-00178791v1

Submitted on 12 Oct 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Dossier

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

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

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

Plaçons dans le contexte où un biologiste dispose de données de type transcriptôme, et souhaite faire de nouvelles expériences pour préciser son modèle ou valider des prédictions. Comment augmenter l'efficacité de ses expérimentations ? La démarche que nous présentons cherche à évaluer l'intérêt que présente l'observation d'un composant par rapport à un autre, en prenant en compte l'ensemble des données et connaissances déjà disponibles.

La « biologie systémique » développe des méthodes pour interpréter et exploiter les données (souvent massives) produites par l'observation d'une cellule, *via* une approche globale et analytique de son fonctionnement. Pour ce faire, il faut d'abord construire un modèle des interactions au sein de la cellule. Le modélisateur cherche ensuite à comprendre le comportement du système à l'aide de simulations et de prédictions. Le biologiste confronte enfin les prédictions aux données disponibles, ce qui permet de valider le modèle, ou de le corriger (processus d'inférence).

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

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 le suggérait 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 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. A partir d'une modélisation par graphe d'interaction, 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 bon sens. Les équations qualitatives sont le résultat de la formalisation mathématique du bon sens. Cette formalisation se fera en plusieurs étapes décrites par la suite.

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 $1, \ldots, i, \ldots, n$. La production d'une molécule i dépend des concentrations des autres molécules $X_1, X_2, \ldots, X_n$. Dans un modèle quantitatif différentiel, pour traduire cette information, on exprime les taux de production dX_i/dt sous la forme d'équations différentielles $dX_i/dt = F_i(X_1,\ldots, X_n)$, où F_i est une fonction de $X_1, \ldots, X_n$.

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. L'article de J.-P. Mazat dans ce numéro en est une excellente illustration. 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 cAMP. Cette dépendance est représentée par une relation graphique (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 ne varient 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 plutôt que 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 de 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'interaction 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 y a 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 256 (soit 2^8) valeurs possibles pour le jeu de variations (*Le, LacI, A, LacZ, Li, G, cAMP, 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 écrira *LacI = [+]* et *A = [+]*, et on constate que l'équation (E1) du tableau, *LacI* ≈ *-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 = [+], cAMP = [-], 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)**).

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, suivie 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.

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, qui sont detaillées dans la base *RegulonDB* **(9)** et portent 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)**.

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,cAMP)* 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* (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 expérimental. Ceci revient à propager dans le graphe d'interaction 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 transcriptômes (**figure 2**) **(10)**. Les 30% de prédictions non validées indiquent des défauts du modèle qui doit être précisé.

Conclusion

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 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.

(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
(5) Tyson JJ *et al.* (2001) *Nat Rev Mol Cell Biol* 2, 908-13
(6) Siegel A *et al.* (2006) *BioSystems* 84, 153-74
(7) Radulescu O *et al.* (2006) *J Roy Soc Interface* 3(6), 185-96
(9) Salgado *et al.* (2006) *Nucleic Acids Res* 34, 394-7
(8) Veber P *et al.* (2004/5) *Complexus* 2, 140-51
(10) Guziolowski C *et al.* (2006) à paraître
(11) Barabasi AL, Albert R (1999) *Science* 286, 509-12

Figure 1 : **Graphe d'interaction et contraintes qualitatives d'un modèle de l'opéron lactose.**
Sur les flèches se terminant par ▶**,** la production du produit d'arrivée augment. Sur celles se terminant par |, le produit de départ diminue. Les boucles de rétrorégulations négatives 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 (**cAMP**). 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 (noeud externe).
© DR

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.
© DR

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

Tableau 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.
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 droite est une solution du système d'équations.

Tableau 3 :
Pouvoir de validation en fonction du groupe de sommets observés sur le réseau de l'opéron lactose.
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'interaction, 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.

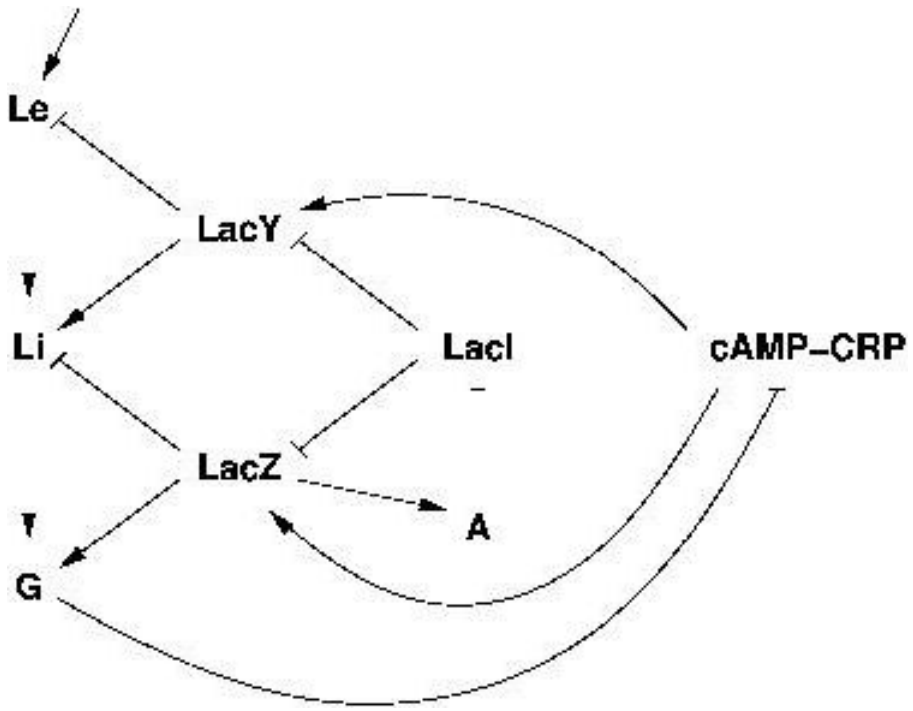

Figure 1

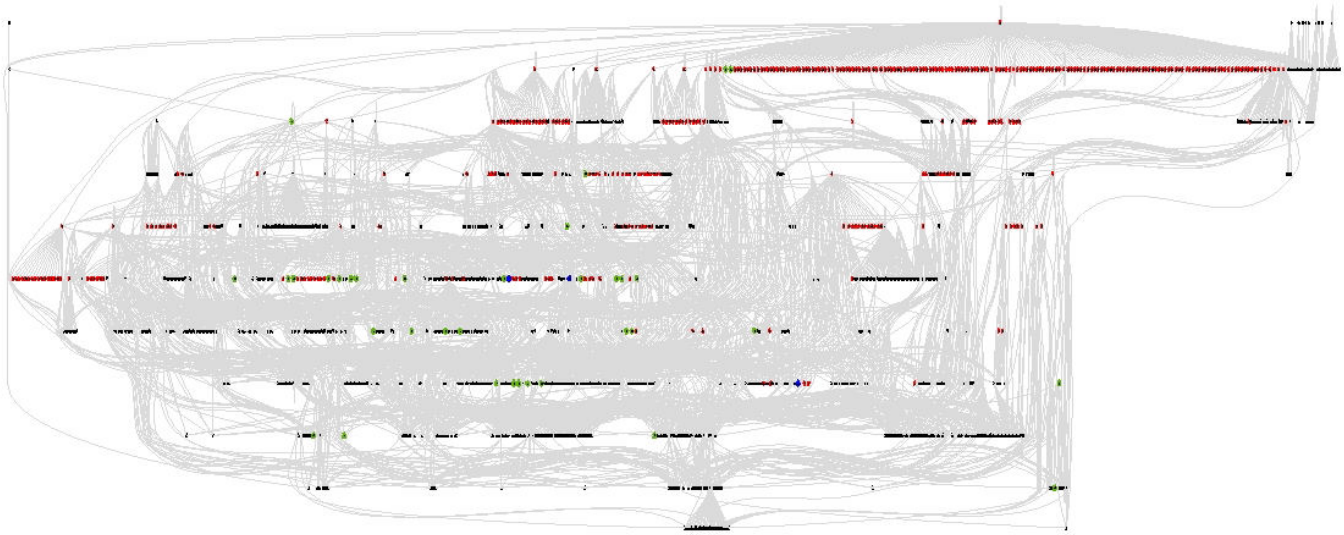

Figure 2

[+] + [-] = [?]	[+] + [+] = [+]	[-] +[-] = [-]
[?] + [-] = [?]	[?] + [+] = [?]	[?] + [?] = [?]
[+] * [-] = [-]	[+] * [+] = [+]	[-] * [-] = [+]
[?] * [-] = [?]	[?] * [+] = [?]	[?] * [?] = [?]

Table 1

LacI	≈	***-A***	*(E1)*
A	≈	***LacZ***	*(E2)*
LacZ	≈	***cAMP - LacI***	*(E3)*
Li	≈	***Le + LacY – LacZ***	*(E4)*
G	≈	***Li + LacZ***	*(E5)*
cAMP	≈	***-G***	*(E6)*
LacY	≈	***cAMP – LacI***	*(E7)*

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

Table 2

Observation	Pouvoir de validation
(***LacZ,LacI,G,A***)	0.75
(***cAMP,LacI,G,A***)	0.75
(***cAMP,LacZ,G,A***)	0.75
(***LacZ,LacY,LacI,A***)	0.75
(***Le,LacZ,LacI,A***)	0.75
(***cAMP,,LacZ,LacI,G***)	0.75

Observation	Pouvoir de validation
(***Li,Le,LacY,G***)	0
(***cAMP,Li,Le,LacY***)	0
(***Li,Le,LacY,A***)	0.125
(***Li,Le,LacY,LacI***)	0.125
(***Li,Le,LacZ,LacY***)	0.125
(***cAMP,Le,LacY,A***)	0.25

Table 3