Étude exhaustive de voies de signalisation de grande taille par clustering des trajectoires et caractérisation par analyse sémantique

Jean Coquet

ANNÉE 2017

UNIVERSITÉ DE RENNES 1

UNIVERSITE BRETAGNE LOIRE

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

Ecole doctorale MathSTIC

présentée par

Jean COQUET

préparée à l'unité de recherche IRISA – UMR6074 et IRSET – UMR1085
Institut de Recherche en Informatique et Systèmes Aléatoires
Institut de Recherche en Santé, Environnement et Travail
ISTIC

Étude exhaustive de voies de signalisation de grande taille par clustering des trajectoires et caractérisation par analyse sémantique

Thèse soutenue à Rennes le 20 décembre 2017

devant le jury composé de :

Élisabeth REMY
Chargée de Recherche CNRS (Institut de mathématiques de Luminy) / rapporteure

Jean-Paul COMET
Professeur (Université de Nice-Sophia Antipolis) / rapporteur

Carito GUZIOLOWSKI
Maîtresse de conférences (École centrale de Nantes) / examinatrice

Denis MICHEL
Professeur (Université de Rennes 1) / examinateur

Jérôme FERET
Chargé de Recherche INRIA (École Normale Supérieure Ulm) / examinateur

Ulrich VALCOURT
Professeur (Université de Lyon 1) / examinateur

Olivier DAMERON
Maître de conférences (Université de Rennes 1) / co-directeur de thèse

Nathalie THÉRET
Directrice de recherche INSERM (IRSET Rennes) / co-directrice de thèse

À mes parents

Remerciements

Je tiens à remercier en premier lieu mes deux directeurs de thèse, Nathalie Théret et Olivier Dameron. Merci de m'avoir permis de réaliser ce doctorat avec vous, je pense avoir beaucoup appris durant ces trois années. Merci aussi à Michael Houle de m'avoir encadré pendant mon séjour au Japon.

Merci également à Élisabeth Remy et Jean-Paul Comet d'avoir accepté la charge de rapporteur. De plus merci à Carito Guziolowski, Denis Michel, Jérôme Feret et Ulrich Valcourt pour leur participation au jury.

Je remercie aussi tous les membres des équipes Dyliss, GenOuest et GenScale avec qui j'ai passé trois années très sympathiques. Merci aussi aux membres de l'équipe 5 de l'IRSET.

Merci à toutes les personnes qui ont participé au projet « Sciences-en-cour[t]s », festival de courts métrages de vulgarisation scientifique. Une pensée toute particulière est destinée à Victorien Delannée et Aymeric Antoine-Lorquin pour la réalisation de notre film « Une rencontre percutante » (disponible gratuitement sur le site web du festival [1]).

En vrac, pour ces trois années, merci à :

— Coraline, Nathalie et Charlotte pour mes petites pauses passées en votre compagnie au service communication ;

— Fannie, Sébastien C. et Cédric pour les fous rires lors de « Sciences-en-cour[t]s » ;

— Cyril, Claudia, Matéo, Efflam, Joseph et Xavier pour les bonnes soirées à rigoler ensemble ;

— toute l'équipe d'enseignement avec qui ça a été un vrai plaisir de travailler ;

— tous mes étudiants, spécialement ceux qui ont assisté à mon cours le vendredi matin à 8h ;

— Miquel, Simone, Ksenia et Myriana pour les bons moments passés dans Tokyo ;

— Solène, Tora, Thomas, Armelle et Nicolas pour ce trek dans les montagnes du Kirghizistan ;

— la Depress team (Arnaud, Lucas, Nathan et Florian) pour nos rêves les plus fous ;

— Samantha et toute sa famille pour m'avoir accueilli partout au Mexique ;

— Aymeric, Charles, Julie et Sébastien L. pour les soirées jeux de rôle et de société ;

1. `http://sciences-en-courts.fr`

— Siva, Jeff, Cyril, Claudia, Guillaume, Coline, Nicolas, Gardouille, Emillie, Delphine et Sandie pour ce voyage en Inde décidé à la dernière minute ;
— tous les membres du club de volley (en particulier Elisabetta, Victorien, Mathilde, Bryan et Evgueni) pour avoir accepté de jouer avec moi malgré mon niveau ;
— Pamela pour être une amie toujours présente même avec les milliers de kilomètres qui nous séparent ;
— Lakshmi pour nos vacances en Argentine qui s'annoncent inoubliables.

Enfin, merci du fond du cœur à toute ma famille : ma mère et mon père qui ont toujours cru en moi et m'ont apporté leur aide dés qu'ils le pouvaient (même pour la relecture de ce rapport), ma sœur avec qui j'ai partagé l'expérience du doctorat (bon courage pour la fin de ta thèse) et mon frère qui restera pour moi le meilleur compagnon de jeu de tous les temps.

Table des matières

Contexte

Les voies biologiques (*Biologic Pathways*) orchestrent les processus biologiques (différentiation, apoptose, transcription, etc.). Elles sont constituées de réseaux de réactions biochimiques et trois types sont couramment décrits dans la littérature :

— un réseau métabolique est un ensemble de réactions se produisant dans une cellule qui explique un aspect de son fonctionnement global ;

— un réseau de signalisation est un ensemble de réactions impliquées dans la réaction d'une cellule à un stimulus externe ;

— un réseau de régulation est un ensemble de réactions affectant l'expression des gènes.

La représentation des voies biologiques doit être de bonne qualité (précise, exhaustive et dans un format prenant en charge le traitement automatique) et être compatible avec les observations biologiques. Cependant, l'analyse des réseaux et de leur dynamique est un défi informatique majeur nécessitant de nouvelles stratégies de raisonnement.

Dans le cadre des systèmes de signalisation, nous sommes confrontés à des modèles multi-échelles présentant des dynamiques très différentes. Les cellules répondent à différents stimuli extra-cellulaires par des ensembles de réactions qui partagent de nombreux éléments. La combinatoire des stimuli extra-cellulaires et des réactions conduit à une grande plasticité de la réponse cellulaire. Un exemple typique est le TGF-β, une protéine qui contrôle la prolifération et la différenciation cellulaire [Hiroaki Ikushima and Kohei Miyazono, 2011] dans un contexte normal, mais qu'est aussi impliquée dans la progression tumorale [Maozhen Tian et al., 2011]. La présence de TGF-β à la surface d'une cellule peut influencer l'expression de nombreux gènes dans le noyau. Une étude basée sur la base de données PID a identifié plus d'une centaine de ces gènes et plus de 16 000 chaînes de réactions reliant TGF-β à au moins un des gènes cibles dans le noyau [Andrieux et al., 2014]. La taille et la complexité des voies de signalisation dépendantes du TGF-β rendent leur analyse difficile et ne permettent pas de prédire de façon satisfaisante les évènements biologiques.

Une limitation majeure dans l'analyse des grands réseaux de signalisation vient du nombre considérable de trajectoires possibles pour expliquer un phénomène biologique. Ces « solutions » sont toutes compatibles avec les contraintes topologiques du réseau de réactions et c'est pourquoi il est essentiel de trouver des méthodes de classification et d'analyse afin de proposer un sens à ces solutions.

Objectif de la thèse

Le but de ma thèse est de proposer un ensemble de stratégies visant à analyser les solutions générées par les simulations de systèmes biologiques de grande échelle, et

en particulier les modèles booléens de signalisation cellulaire. Pour ce faire, je présenterai d'abord les principes de la modélisation de systèmes biologiques, ainsi que les bases de données de voies de signalisation et les différentes stratégies d'analyse dans ce domaine. Ceci me permettra d'expliquer pourquoi les méthodes non supervisées constituent une approche pertinente pour classifier les trajectoires de signalisation (solutions) associées à l'expression de gènes.

Ma thèse a vocation à être générique, mes travaux ont consisté dans un premier temps à proposer une analyse grande échelle sur une application biomédicale : l'analyse des voies de signalisation de la protéine TGF-β. Puis dans un deuxième temps, j'ai réalisé une étude préliminaire pour généraliser cette méthode sur plusieurs bases de données.

Chapitre 1

Introduction

Après une présentation générale de la biologie des systèmes, je détaillerai dans ce chapitre les différentes approches de modélisation des systèmes biologiques. Je discuterai ensuite des stratégies possibles pour l'analyse statique ou dynamique d'un modèle. Enfin je présenterai les techniques de fouille de données (*data-mining*) pour analyser l'ensemble de solutions générées par la dynamique d'un modèle de signalisation cellulaire. Je conclurai ce chapitre en présentant l'annotation par *Gene Ontology* qui facilite l'analyse fonctionnelle des prédictions issues du modèle.

1.1 État de l'art de la modélisation des systèmes biologiques

1.1.1 Définition des systèmes biologiques

Depuis un peu plus d'une vingtaine d'années, le développement des technologies à haut débit en biologie, dites « omiques », a multiplié la quantité de données biologiques [Palsson, 2002]. Ces données concernent différents niveaux de granularité. Il est désormais possible de considérer un système dans sa globalité avec ses nombreux composants et leurs différentes interactions [Kitano, 2002]. Cette nouvelle stratégie a conduit à l'émergence d'un nouveau domaine d'étude la « biologie des systèmes » (ou biologie systémique) qui a pour but d'étudier le fonctionnement des interactions entre les composants biologiques du système en fonction du temps et de l'espace [Bree B. Aldridge et al., 2006]. La biologie des systèmes prend en compte les données du génome (ensemble des gènes de l'ADN), du transcriptome (ensemble des ARNs résultant de la transcription des gènes), du protéome (ensemble des protéines résultant de la traduction des ARNs messagers) et du métabolome (ensemble des métabolites issus de réactions biochimiques) [Oltvai and Barabási, 2002]. En fonction de la question posée, une étude en biologie des systèmes intègre plus ou moins de granularité dans ces données (ADNs, ARNs, protéines), et sera plus ou moins spécifique (espèce, organisme, tissu, cellule) (figure 1). La biologie des systèmes admet pour principe qu'il n'est pas possible de caractériser un système dans sa globalité en étudiant séparément chacune de ses parties. Elle repose sur l'idée que le tout vaut plus que la somme des parties. Pour travailler sur les modèles biologiques de façon systémique Kitano propose quatre étapes [Kitano, 2002] :

1. Concevoir la structure du système, c'est-à-dire déterminer les différents acteurs et leurs interactions ;
2. Établir la dynamique du système, c'est-à-dire décrire la façon dont le système se comporte au fil du temps dans différentes conditions ;
3. Identifier les acteurs contrôlant le système, c'est-à-dire mettre en avant les acteurs qui contrôlent systématiquement un état du système ;
4. Concevoir des systèmes, c'est-à-dire être capable de construire physiquement des systèmes biologiques ayant une fonction biologique souhaitée. C'est le but de la biologie synthétique.

Les voies (*pathways*) biologiques sont en général classées en trois groupes majeurs : les réseaux de signalisation, les réseaux métaboliques et les réseaux de régulation de gènes [Machado et al., 2011].

— Les voies de signalisation représentent la réponse cellulaire à un stimulus extérieur provenant de l'environnement (cellules voisines, matrice extra-cellulaire, cytokines, facteurs de croissance, etc.). Elles régulent tous les processus cellulaires comme la différenciation et l'apoptose. La liaison d'un facteur extracellulaire à un récepteur à la surface de la cellule entraîne une cascade de réactions dans la cellule influençant au final la régulation des gènes dans le noyau.

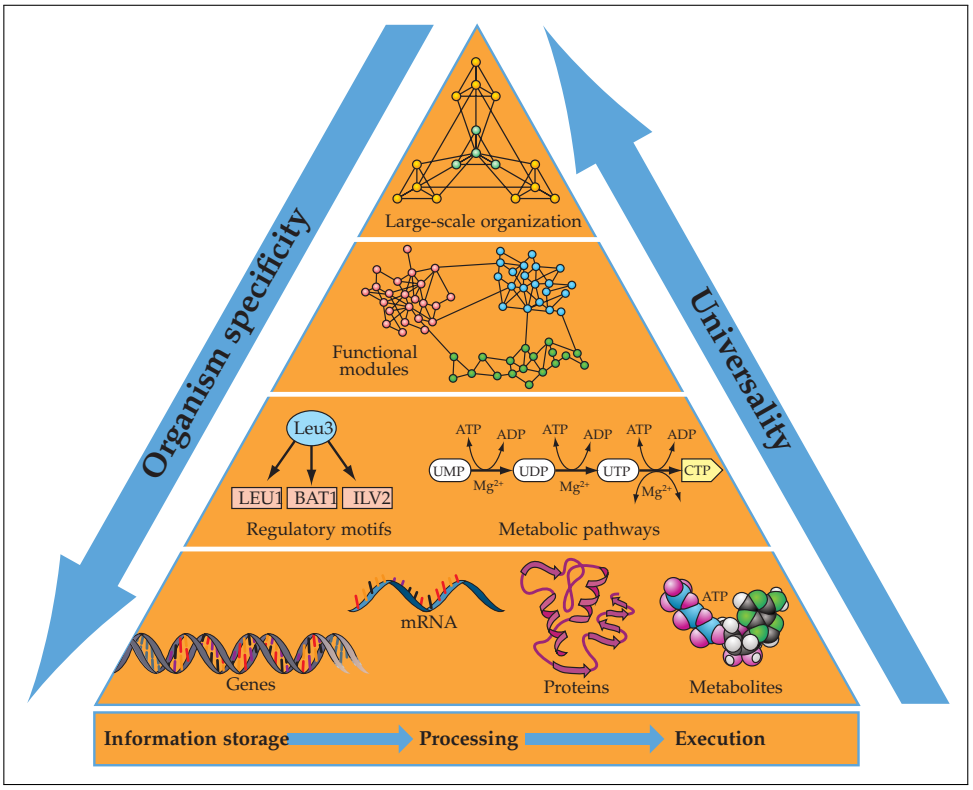


Figure 1 – ***Pyramide des différents niveaux de la biologie des systèmes proposée par [Oltvai and Barabási, 2002].***

Cette pyramide représente l'ensemble des types d'études en biologie des systèmes. Plus l'étude prend en compte divers attributs de la cellule (gènes, protéines, métabolismes, etc.) plus cette étude sera universelle et aura un niveau d'abstraction important. Cette pyramide peut aussi se lire de gauche à droite, du stockage de l'information par les gènes à l'exécution de cette information par les protéines et les métabolites.

— Les réseaux métaboliques décrivent des cascades de réactions biochimiques grâce auxquelles la cellule satisfait ses besoins énergétiques. Ces réactions sont catalysées par des enzymes qui utilisent des substrats et les transforment en énergies nécessaires aux fonctions de cellulaires (croissance, prolifération, etc.)

— Les réseaux de régulation de gènes représentent le contrôle de l'expression des gènes codant pour toutes les protéines, par conséquent affectant toutes les fonctions cellulaires. L'expression d'un gène implique qu'il soit transcrit en ARN messager, puis que cet ARN soit traduit en protéine. Le processus de transcription est contrôlé par des facteurs de transcription qui peuvent fonctionner comme activateurs ou inhibiteurs.

Même si ces trois niveaux d'informations décrivent des systèmes différents, ils sont étroitement connectés. Ainsi les réseaux de régulation de l'expression des gènes sont stimulés par la réponse cellulaire médiée par le réseau de signalisation et codent pour les composants à la fois des réseaux de signalisation et métaboliques. Les réseaux métaboliques déterminent les fonctionnalités de la cellule et donc de ses réseaux de signalisation et régulation génique.

Les comportements des systèmes biologiques relèvent de l'interaction entre ses composants. L'augmentation du nombre de composés et le fait que leurs interactions ne sont pas linaires dans le temps et dans l'espace nécessitent de développer des approches mathématiques relevant de la théorie des systèmes dynamiques. Pour aider à la compréhension des phénomènes biologiques, les approches de modélisation mathématique sont donc utilisées depuis de nombreuses années pour effectuer des simulations et des prédictions [Papin et al., 2003].

Nous pouvons définir un « système » comme une collection d'objets interdépendants et un « objet » est une unité élémentaire sur laquelle des observations peuvent être faites, mais dont la structure interne n'existe pas ou est ignorée. Un « modèle » est la description du fonctionnement d'un système. Enfin, une « description » est un signal qui peut être décodé ou interprété par des humains. Les modèles sont donc des outils facilitant l'étude des systèmes.

La taille de modèle de systèmes biologiques varie considérablement. Certains modèles possèdent quelques dizaines de composants comme la voie de signalisation EGF et TNF-α composée de 28 biomolécules [Chaouiya et al., 2013]. D'autres modèles décrivent des systèmes ayant des centaines ou des milliers de composants. Par exemple, [Ryall et al., 2012] proposent un modèle de 106 éléments et 193 réactions incluant 14 voies de signalisation différentes afin de décrire la signalisation hypertrophique des myocytes cardiaques. On peut aussi noter le projet d'envergure « Recon » [Duarte et al., 2007, Thiele et al., 2013] visant à proposer la représentation la plus complète du métabolisme humain. La version 2 est composée de 1 789 gènes, 7 440 réactions et 2 626 métabolites. On définit un système complexe comme un système composé d'un grand nombre d'entités et d'intéractions rendant la prédiction de son comportement difficile.

1.1.2 Le cycle de modélisation

Avec la biologie des systèmes, est apparue la notion de cycle de modélisation (figure 2). Dans une première étape, celui-ci consiste à construire la structure du modèle, en choisissant les composants biologiques et le formalisme mathématique qui va caractériser les relations entre ces composants. Ensuite, il est possible de calibrer le modèle en identifiant par exemple les valeurs de paramètres (composants et interactions) et les conditions initiales pour effectuer des simulations numériques qui peuvent à la fois reproduire des observations et conduire à des prédictions.

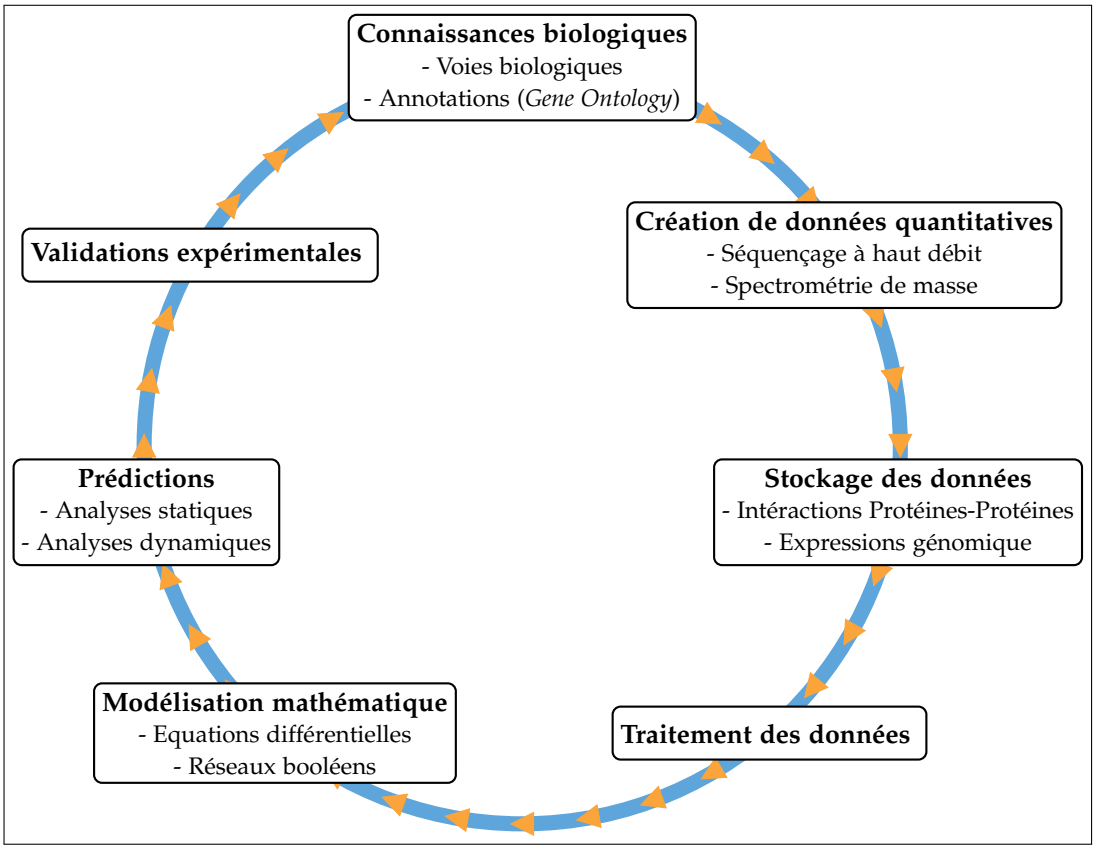


Figure 2 – ***Représentation du cycle de modélisation en biologie des systèmes inspirée par [Schilling et al., 2008].***

Les problématiques biologiques sont étudiées à partir de cycles d'analyses, de traitement de données quantitatives, de modélisation mathématique, de prévisions in silico et de validations expérimentales. Pour chaque processus de recherche réalisé dans ce cycle, quelques exemples sont notés en italique.

Collecte et standardisation des composants du modèle

Décrire de façon précise un processus biologique peut permettre de créer un modèle avec précision dans un langage contrôlé. Il peut servir de dispositif d'organisation de la pensée. Très souvent l'élaboration du modèle nécessite une réflexion approfondie qui déjà apporte une meilleure compréhension du système biologique. En effet, ce travail nécessite d'identifier les composants essentiels du système et l'ensemble des interactions entre ces composants. Ces données proviennent d'articles, des bases de données ou d'expériences en laboratoire réalisées directement pour faire ce modèle. Si ces données proviennent de différentes sources, le modèle permet de les formaliser de la même manière et donc de les rendre plus cohérentes les unes par rapport aux autres. Une bonne modélisation suscite d'autres questions sur le comportement du système et, à long terme, l'applicabilité à d'autres systèmes de tout principe nouvelle-

ment découvert [Chowdhury and Sarkar, 2015].

Il existe une grande quantité de connaissances biologiques disponibles dans les publications et qui sont maintenant organisées dans des bases de données. Posséder une carte décrivant de manière détaillée les différents types de modifications chimiques et les cascades de réactions est nécessaire afin de découvrir les éléments essentiels d'un système biologique. L'importance de ces bases de données ne se limite pas à l'accumulation des données expérimentales, mais est également précieuse pour les développeurs de modèles afin d'interpréter les propriétés émergentes de ces connaissances organisées en réseau. Depuis 1995, le nombre de bases de données de systèmes biologiques a fortement augmenté [Soh et al., 2010]. D'après [Galperin et al., 2017] il en existe 166 à l'heure d'écriture de ce manuscrit [1]. Une description détaillée de 24 bases de données de signalisation est présentée dans la section 1.1.5.

Selon les systèmes d'architecture et de stockage de données utilisés par les différentes bases de données, les procédures d'accès, de visualisation et d'analyse des données diffèrent considérablement. En conséquence, il faut beaucoup de temps aux utilisateurs pour extraire les données de ces bases. Pour surmonter ce problème, presque toutes les bases de données académiques (par exemple REACTOME [Croft et al., 2014, Fabregat et al., 2016], PANTHER [Mi and Thomas, 2009] ou NCI-PID [Carl F. Schaefer et al., 2009]) ont développé des strutuctures de données standards interrogeables automatiquement et facilement accessibles, tels que le *Systems Biology Markup Language* (SBML) [Chaouiya et al., 2013], le *Biological Pathway Exchange* (BioPAX) [Demir et al., 2010] ou le *System Biology Graphical Notations* (SBGN) [Novère et al., 2009].

Simulations et concordance avec les données expérimentales

Une fois le modèle décrit dans un formalisme choisi, il est possible d'analyser sa dynamique c'est-à-dire d'analyser le comportement du système en fonction du temps. L'état d'un modèle se définit comme l'ensemble des valeurs de ses variables à un instant t. Étudier le comportement du système nécessite de calculer l'évolution des valeurs des variables de manière continue ou discrète dans le temps. L'évolution temporelle des variables du modèle (par exemple, les concentrations de protéines) est affectée par les valeurs d'autres variables et par des paramètres tels que les constantes de dissociation, les constantes de taux cinétique et les ordres de réaction.

Une simulation dépend des conditions initiales des variables du modèle. S'il existe des données expérimentales sur l'évolution du système, alors il est admis qu'un modèle correctement réalisé doit avoir une dynamique concordante avec les données expérimentales.

De telles observations sont utiles d'une part pour étudier l'ensemble des valeurs de paramètres compatibles avec les observations et d'autre part pour étudier les familles de modèles.

1. `https://www.oxfordjournals.org/nar/database/cat/6`

Prédictions

Les études de simulations du modèle permettent de mettre en évidence certaines lacunes et proposent des hypothèses sur le comportement non décrit initialement. Si le modèle est en accord avec les observations expérimentales, une stratégie est d'effectuer des modifications (mutations dans le cas de gènes) des valeurs des variables ou des constantes afin de perturber le système et de prédire de nouveaux comportements. Ce sont ces nouvelles données qui permettront d'enrichir le modèle en ajoutant des connaissances. La boucle est bouclée et de nouvelles simulations peuvent être relancées.

1.1.3 Les graphes en biologie des systèmes

La représentation des données sous forme de réseaux est au cœur de la biologie des systèmes, et derrière chaque modèle mathématique de processus biologiques se cache un réseau.

Un graphe est défini de façon générale comme un ensemble de nœuds V (pour *vertices* en anglais) et un ensemble d'arêtes E (pour *edges*), soit le graphe $G = (V, E)$. Un nœud est une entité élémentaire représentant un objet dans le modèle, la structure de cet objet n'est pas prise en compte dans le réseau. Par exemple, si un nœud représente une molécule alors sa structure chimique ne sera pas prise en compte. Une arête est une relation entre deux nœuds, par exemple la liaison possible entre deux molécules. Il est possible que les arêtes soient « orientées » dans ce cas la relation représentée par l'arête a un sens de lecture, par exemple la protéine a active le gène g. Enfin, les nœuds et les arêtes peuvent posséder n'importe quel attribut (nom, type, propriétés physiques, etc.) et ce sont ces attributs qui vont définir la représentation du modèle.

Ces représentations sont plus que de simples illustrations, elles imposent une sémantique spécifique au processus biologique que l'on veut modéliser. Les représentations proposent une abstraction plus ou moins importante et supportent différents types de raisonnement. Il est donc important de choisir la représentation appropriée en fonction de la question posée et des données disponibles. Ce choix se répercute dans la sélection des méthodes de modélisation et de simulation, ainsi que le traitement des données utilisées pour la validation des modèles.

Il existe de nombreuses représentations de réseau, et [Le Novère, 2015] a récemment proposé une classification en quatre familles (figure 3) :

— **Réseaux d'interaction**
Les réseaux d'interaction (figure 3a) sont utilisés pour représenter des interactions physiques ou fonctionnelles entre des protéines ou des gènes. Ces réseaux sont souvent non dirigés, si une protéine p_1 interagit avec p_2 alors p_2 interagit aussi avec p_1. De plus ils ne sont pas séquentiels, c'est-àdire que un chemin entre plusieurs arêtes ne permet pas de décrire un méchanisme biologique, s'il existe une interaction entre p_1 et p_2 et une interaction p_2 et p_3 alors il n'y a pas forcément une interaction entre p_1 et p_3. Les réseaux d'interactions génétiques et protéiques ont été construits pour obtenir une vue globale de la régulation du

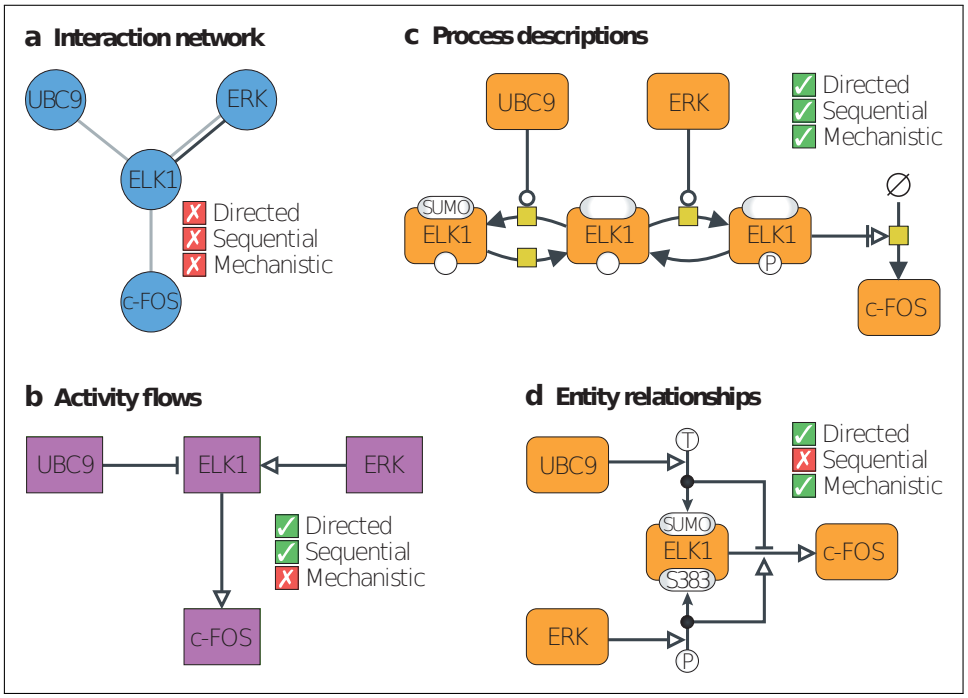


Figure 3 – ***Quatre types de graphes de la biologie des systèmes proposés par [Le Novère, 2015]***

Cette figure regroupe le même système biologique représenté en quatre types de réseaux. (a) Un réseau d'interaction peut être utilisé pour représenter les interactions physiques (ligne noire) et les interactions fonctionnelles (lignes grises). (b) Un flux d'activité peut être utilisé pour montrer la stimulation ou l'inhibition d'une activité d'un élément par une autre activité, par exemple l'activité C-FOS est stimulée par l'activité ELK1. (c) Une description détaillée du processus peut être utilisée pour montrer des interactions précises, comme la phosphorylation (P) de ELK1 ou l'expression de c-FOS. (d) Les relations d'entités peuvent être utilisées pour décrire les différentes interactions entre les entités, mais sans montrer la séquentialité de ces interactions.

génome ou pour comprendre des processus de régulation spécifiques.

— **Flux d'activité**
Les flux d'activité (figure 3b) sont des diagrammes d'influence permettant de représenter les effets d'inhibition ou d'activation d'un élément (une molécule par exemple) sur un autre. Les effets chimiques (dans le cas des molécules) ne sont pas représentés, nous savons seulement que l'activité de la molécule m_1 stimule l'activité de la molécule m_2. Les flux d'activité sont utilisés lorsque le détail d'une réaction chimique n'est pas connu ou n'est pas considéré comme essentiel pour comprendre le processus modélisé. Les réseaux sont dirigés et séquentiels, mais les mécanismes chimiques ne sont pas décrits.

— **Description des processus**

Les descriptions de processus (figure 3c) sont des graphes bipartis, c'est-à-dire avec deux types de nœuds : les variables du modèle (quantités des molécules par exemple) et les réactions qui diminuent ou augmentent (consomment ou produisent) les valeurs de ces variables. Ces réseaux sont dirigés et séquentiels, et grâce au niveau de granularité qu'ils proposent, ils permettent de décrire les mécanismes impliqués dans les réactions. Malheureusement, cette granularité a un coût, les processus ne sont pas indépendants et entraînent une explosion de la combinatoire. Plus les processus sont décrits finement, plus leur combinatoire augmente, ce qui constitue un frein à l'analyse de grands réseaux très connectés.

— **Relations d'entités**
Les réseaux de relations d'entités (figure 3d) représentent les entités, les états de ces entités (par exemple la méthylation) et l'influence des entités sur ces états. Ce sont des réseaux dirigés, expliquant les mécanismes impliqués mais non séquentiels.

Certains types de réseaux seront plus adaptés à certains systèmes biologiques. En effet les systèmes de régulation de gènes ne comportent que des régulations transcriptionnelles, la description des mécanismes internes n'est donc pas essentielle. On aura donc tendance à se tourner vers une représentation en réseaux d'interactions. Quant aux réseaux métaboliques, les modélisateurs ont plus une vision de production ou consommation des éléments. Ils s'intéressent à l'évolution de concentrations à l'échelle enzymatique. La mécanique des réactions et leur séquentialité constituent une information importante pour ce type de réseaux. Enfin pour les systèmes de signalisation, la notion de propagation du signal comme une succession d'évènements biologiques est très importante. Même si les réactions biochimiques ont bien lieu dans un système de signalisation, la notion importante est le flux d'informations.

Le type de réseaux peut aussi impacter le formalisme choisi. Par exemple, les réseaux de flux d'activités décrivent une information qualitative et sont donc utilisés en général dans les modèles logiques. Alors que les réseaux de relations d'entités sont un ensemble de relations indépendantes et sont donc faciles à transcrire dans un formalisme basé sur des règles. Dans la section suivante, je vais justement décrire les différents formalismes de modélisation de systèmes biologiques.

1.1.4 Différents formalismes de modélisation

Comme expliqué précédemment, la visualisation et l'analyse de modèles de systèmes biologiques nécessitent l'utilisation de formalismes permettant de décrire le système de façon précise et non ambiguë. Toutes les méthodes de modélisation se caractérisent par la représentation du temps et le type de variables utilisées (figure 4).

Le choix d'un formalisme de modélisation est directement lié à la nature des données biologiques et à la quantité de données disponibles. En effet si des expériences quantitatives sur la variation du niveau de concentration des molécules ou du niveau d'expression des gènes sont accessibles alors on pourra utiliser un formalisme très

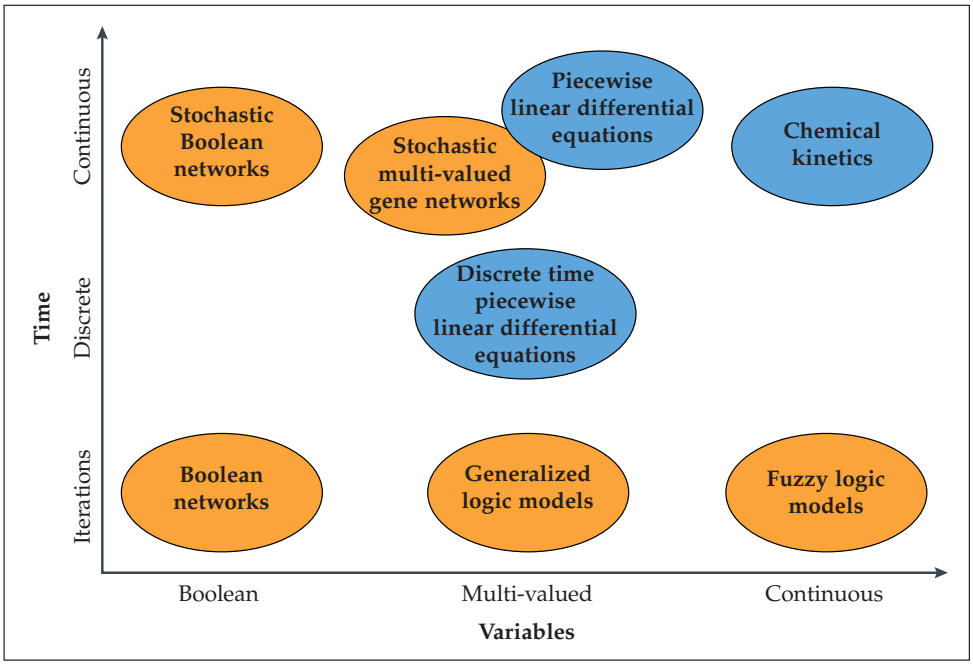


Figure 4 – ***Granularité de la représentation du temps et des valeurs des variables pour diverses approches de modélisation proposée par [Le Novère, 2015].***

Les variables dans un modèle peuvent prendre des valeurs continues (par exemple, des concentrations ou un nombre de molécules), des valeurs multiples c'est-à-dire un ensemble limité de valeurs (par exemple, nulles, moyennes ou élevées) ou des valeurs booléennes (présentes ou absentes, actives ou inactives). La progression des variables pendant les simulations peut être représentée à l'aide du temps continu, de manière discrète (avec les mises à jour effectuées après des durées de temps spécifiées), ou en utilisant des itérations (qui ne représentent pas nécessairement une durée spécifique). Les méthodes oranges sont mises à jour selon les règles logiques, tandis que les méthodes bleues calculent les nouvelles valeurs des variables à l'aide de fonctions quantitatives.

précis et plus fidèle à la réalité. En opposition si nous n'avons connaissance que des mécanismes du système de façon qualitative alors un formalisme plus abstrait devra être utilisé. De plus, la taille et la complexité d'un système auront une influence sur le choix du formalisme. En effet pour des raisons de performance, il pourra être compliqué d'analyser la dynamique de modèles de très grandes tailles basés sur des données numériques.

Modèles continus

Équations différentielles comme outil de modélisation

Pour une représentation très fine du processus, le modélisateur va chercher à représenter les variables et le temps de manière continue. Ainsi plusieurs types de modèles utilisent des données quantitatives et font appel à un formalisme mathématique

basé sur les équations différentielles (*Ordinary Differential Equations* – ODE), afin de décrire la variation de la quantité des éléments dans le système modélisé en fonction du temps et des autres éléments [Tyson et al., 2003]. Ils ont été appliqués à toutes sortes de voies biologiques comme pour le métabolisme central du carbone chez *Escherichia Coli* [Chassagnole et al., 2002] ou les cascades de signalisation des protéines kinases [Markevich et al., 2004]. La construction de modèles ODE nécessite de nombreuses données expérimentales pour identifier les lois de vitesse appropriées et pour estimer les valeurs des paramètres cinétiques. Les modèles à équation différentielle s'appliquent donc essentiellement aux petits réseaux [Bree B. Aldridge et al., 2006]. Les approches différentielles sont déterministes et ne permettent pas de traiter la variabilité des individus, mais seulement le comportement moyen d'une population. C'est pourquoi les modélisateurs se sont intéressés à prendre en compte la variabilité des individus simulant plusieurs évolutions temporelles des variables afin d'analyser leur distribution ou leur probabilité en fonction de temps [Wilkinson, 2009]. Ces modèles basés sur des systèmes d'équations différentielles stochastiques (*Stochastic Differential Equations* – SDE) ont été utilisés avec succès, par exemple pour modéliser l'excitation des cellules à granules [Saarinen et al., 2008].

[Smallbone and Mendes, 2013] démontrent qu'il est tout de même possible d'utiliser les équations différentielles sur des réseaux à plus de 700 éléments en utilisant la FBA (*Flux Balance Analysis*) pour estimer les valeurs des paramètres. La FBA est une recherche de la distribution du flux dans les métabolites à l'état stationnaire et dynamiquement faisable, dans le but d'activer une réaction.

Modèles discrets

Modèles booléens

Il est possible de réaliser une abstraction plus importante en considérant que le temps n'a pas de régularité mais correspond seulement à un changement d'état, c'est-à-dire que deux réactions peuvent avoir une vitesse différente dans la réalité mais ne seront pris en compte que leur occurrence dans le modèle. Il est aussi possible d'attribuer seulement deux valeurs aux variables (présent/absent, actif/inactif, etc.), la quantité d'une molécule n'est alors plus prise en compte. Ces deux abstractions sont en général réalisées s'il n'y a pas assez de données nécessaires ou si l'on cherche à simuler de très grands modèles. Les réseaux booléens ont été introduits par Kauffman en 1969 pour modéliser les réseaux de régulation des gènes [Kauffman, 1969], où à chaque pas de temps, l'état de chaque gène est déterminé par une règle logique en fonction de l'état de ses régulateurs. En effet, la valeur de chaque composant d'un modèle logique est définie par une fonction logique qui peut être en partie déduite du graphe. La valeur des composants de sortie d'une réaction dépendra de la valeur des composants d'entrée, des activateurs et des inhibiteurs. Par exemple la formation d'un complexe AB catalysé par une molécule C ou la molécule D et inhibé par la molécule E peut être décrite par la fonction logique suivante :

$$X_{AB} = x_A \wedge x_B \wedge (x_C \vee x_D) \wedge \neg x_E$$

où x_Y correspond à la valeur booléenne de Y au pas de temps t et X_Y correspond au pas de temps $t+1$. Les symboles $\wedge$, $\vee$ et $\neg$ correspondent respectivement aux fonctions logiques *ET*, *OU inclusif* et *NÉGATION*.

De cette manière, un modèle logique est défini par l'ensemble des fonctions des composants et il est possible de calculer la dynamique discrète d'un tel modèle [Samaga and Klamt, 2013]. Les réseaux booléens ont été utilisés pour modéliser les voies de signalisation, comme la voie de signalisation des neurotransmetteurs [Gupta et al., 2007] ou pour étudier les altérations génétiques dans la tumorigénése des cancers de la vessie [Remy et al., 2015]. D'autres études, comme [Videla et al., 2015], proposent d'identifier les fonctions logiques booléennes à partir de données expérimentales. Tout comme pour les approches différentielles, il est possible de réaliser des réseaux booléens probabilistes, cette méthode a été proposée par [Shmulevich et al., 2002].

Réseaux bayésiens

Les réseaux bayésiens ont été introduits par Pearl [Pearl, 1988]. Ce sont des réseaux dirigés probabilistes, où les nœuds représentent des variables aléatoires (discrètes ou continues) et les arêtes représentent des conditions. Chaque nœud contient une fonction probabiliste qui dépend des valeurs des conditions d'entrée. La dynamique d'un réseau bayésien consiste à calculer l'évolution des variables aléatoires en fonction d'une séquence discrète de pas de temps. Ils ont été notamment utilisés pour inférer et représenter la régulation des gènes [Grzegorczyk et al., 2008] et les réseaux de signalisation [Sachs et al., 2005].

Formalismes compacts

Modèles basés sur des règles

Les formalismes basés sur des règles permettent de modéliser de manière compacte des systèmes biologiques comme des ensembles de règles, à l'échelle moléculaire, décrivant comment le système peut évoluer dans le temps [Chylek et al., 2011]. Les règles sont basées sur la notion de causalité : si certaines conditions ont lieu, alors différentes actions sont provoquées. Pour une réaction biochimique, une règle décrit les contraintes nécessaires à l'interaction des réactants afin de provoquer la réaction. Le comportement global du système est donc calculé à partir des descriptions locales des règles. Les règles spécifiées peuvent être utilisées pour générer automatiquement un système d'équations différentielles [Smith et al., 2012]. De plus une règle peut posséder certaines propriétés comme le taux d'affinité, la vitesse d'application, etc. Un avantage important de ce type de formalisme est sa modularité qui facilite la modification ou l'extension du modèle, lorsque de nouvelles connaissances deviennent disponibles. Ce type de formalisme est utilisé pour le langage BioCham [Chabrier-Rivier et al., 2004] et Kappa [Danos et al., 2008]. La figure 5 détaille en Kappa la règle permettant la formation du complexe AB à partir des protéines A et B.

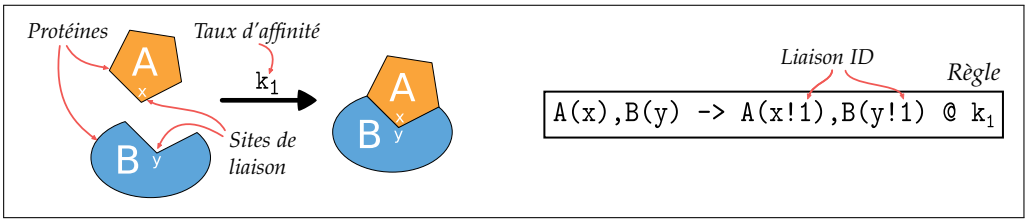


Figure 5 – *Exemple d'une règle de modélisation en Kappa*

Cette figure représente l'écriture en Kappa de la formation du complexe AB. La partie gauche décrit la liaison de A et B avec leur site de liaison respectif x et y, et le taux d'affinité de la réaction k_1. La partie droite correspond à la même liaison décrite en Kappa. À noter la présence du signe « ! » suivi d'un identifiant qui correspond à la liaison de deux sites, si ce signe n'est pas présent cela veut dire que le site est libre.

Modèles de contraintes

Les modèles de contraintes [Reed and Palsson, 2003] sont basés sur les contraintes de stoechiométrie, de thermodynamique et de capacité enzymatique. Ils définissent un espace de solutions possibles représentant différents phénotypes conformes aux contraintes. La simplicité de leur formulation permet de les appliquer à des modèles comprenant des milliers de réactions. C'est par exemple le cas pour la reconstruction du réseau métabolique d'*Escherichia coli* [Orth et al., 2011].

1.1.5 Modélisation de réseaux de signalisation cellulaire

Durant ma thèse, je me suis intéressé plus précisément aux réseaux de signalisation et à l'analyse des trajectoires issues de ces réseaux, c'est pourquoi dans la suite de ce manuscrit je me focalise sur ce type de réseau.

Signalisation moléculaire

Les réseaux de signalisation expliquent comment les cellules vivantes adaptent leur phénotype aux modifications du micro-environnement [Kholodenko, 2006], on parle de réponse cellulaire. La signalisation contrôle des mécanismes fondamentaux de la cellule comme la prolifération cellulaire, la différenciation cellulaire ou encore l'apoptose. Le processus de transduction du signal commence par la liaison d'une molécule de signalisation extra-cellulaire (par exemple les facteurs de croissance ou les hormones) à un récepteur à la surface de la cellule. Le signal est ensuite propagé et amplifié à l'intérieur de la cellule à travers des cascades de signalisation qui impliquent une série de réactions très hétérogènes telles que la phosphorylation des protéines, la formation de complexes, le transport protéique, etc. Ces cascades de réactions débouchent sur la régulation de la transcription des gènes dans le noyau qui vont coder de nouvelles protéines impliquées dans l'adaptation de la cellule à son environnement (figure 6). Une trajectoire correspond à l'ensemble des molécules activés

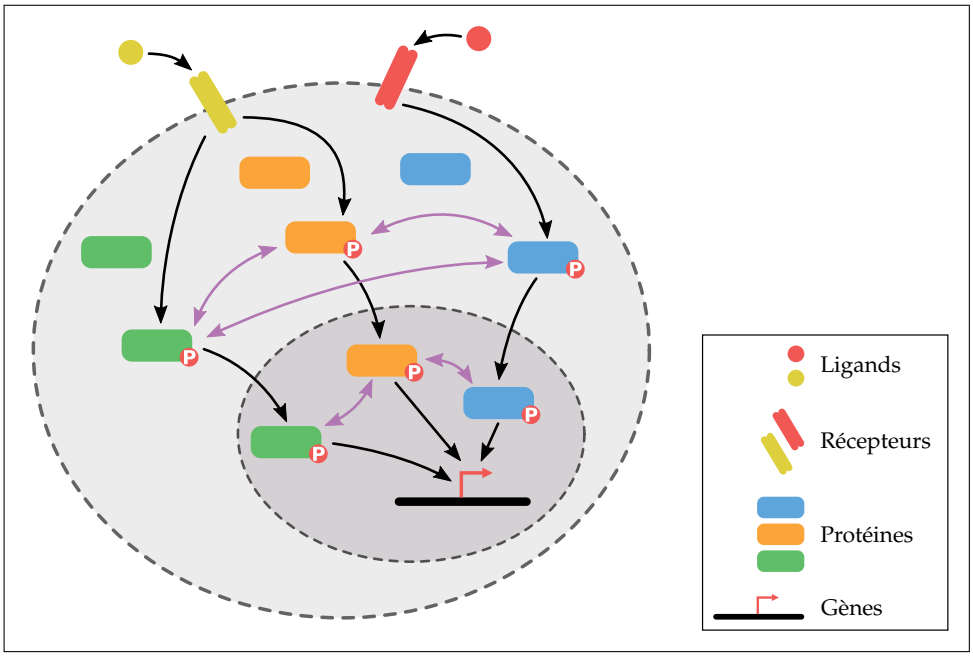


Figure 6 – *Représentation schématique d'une voie de signalisation cellulaire.*
Les ligands extra-cellulaire peuvent se fixer aux récepteurs membranaires. Le signal est ensuite transmis jusqu'au noyau par l'enchaînement de réactions biochimiques, afin de réguler l'expression des gènes. Les flèches noires indiquent le transport des biomolécules et les flèches violettes correspondent aux interactions entre les protéines. Le signe « P » indique la phosphorylation d'une protéine.

lors de la propagation du signal jusqu'aux gènes.

Bases de données de signalisation

Les avancées en biologie cellulaire et moléculaire, les études de génomique ou de protéomique à haut débit et l'achèvement du projet de séquençage du génome Humain [J. Craig Venter et al., 2001] ont rendu possible une forte accumulation des données génomiques et protéomiques. De nombreuses communautés de recherche et développeurs de bases de données ont participé à la reconstruction des voies biochimiques en compilant les observations expérimentales à partir de la littérature [Galperin et al., 2017].

En 1993, EcoCyc a lancé la première représentation formelle des voies métaboliques d'*Escherichia coli* [Karp and Riley, 1993]. Quelques années plus tard, la base de données « *Kyoto Encyclopedia of Genes and Genomes* » (KEGG) est lancée par Kanehisa [Kanehisa and Goto, 2000], ce qui représente le premier hébergement web de voies biologiques créées manuellement. Le classement des données de réseaux s'est organisé selon les trois catégories : réseaux de signalisation, réseaux métaboliques et réseaux de régulation de gènes. En parallèle sont apparues des bases de données

créées à des fins commerciales par des entreprises, comme BIOCARTA [Nishimura, 2001], en complément des bases de données académiques.

D'après le site web *Pathguide* [Bader et al., 2006], depuis 1994 le nombre de bases de données a fortement augmenté. En effet il existe à l'heure de l'écriture de ce manuscrit plus de 166 bases de données actives des réseaux de signalisation et de réseaux métaboliques.

Il existe trois types de bases de données :

— Les bases de données proposant des données qu'elles ont elles-mêmes organisées, comme la majorité des bases de données académiques (KEGG [Kanehisa and Goto, 2000], SPAD [Gerstmann, 2002], DOQCS [Sivakumaran et al., 2003], NetPath [Kandasamy et al., 2010], REACTOME [Croft et al., 2014, Fabregat et al., 2016], SignaLink [Fazekas et al., 2013], SPIKE [Elkon et al., 2008], BioModels [Chelliah et al., 2015], INOH [Yamamoto et al., 2011] et PANTHER [Mi and Thomas, 2009]) et des bases de données commerciales (BIOCARTA [Nishimura, 2001], GENEGO / METACORE, Cell Signaling TECHNOLOGY, PROTEIN LOUNGE [Besaw, 2013], MILLIPORE, Applied Biosystems et INVITROGEN).

— Les bases de données agrégeant plusieurs autres bases de données, comme Pathway Commons [Cerami et al., 2011] et hiPathDB [Yu et al., 2012].

— Les bases de données hybrides qui possèdent à la fois des données autoorganisées et des données provenant d'autres bases, comme WikiPathways [Kutmon et al., 2016], NCI-PID [Carl F. Schaefer et al., 2009], GOLD.db [Hackl et al., 2004], CPDB [Kandasamy et al., 2010] et InnateDB [Breuer et al., 2013].

[Chowdhury and Sarkar, 2015] ont identifié une liste de 24 bases de données de signalisation sur la base de deux critères : (1) la base de données doit fournir uniquement ou partiellement les données relatives aux signalisations cellulaires de l'espèce humaine et (2) le lien HTTP des bases de données doit être actif, fonctionnel et accessible aux utilisateurs sans frais d'accès. En termes de données de signalisation cellulaire, ils répertorient 19 catégories de voies de signalisation. La figure 7 présente les différentes catégories de voies décrites dans chacune des 24 bases de données. On peut voir que NCI-PID est la base la plus hétérogène proposant le plus de types de voies de signalisation, alors que BioModels et DOQCS sont très spécifiques. Les principales bases de données telles que la BDCP, KEGG, REACTOME, GOLD.db, PROTEIN LOUNGE, MILLIPORE et PANTHER contiennent un large éventail de données, ce qui en fait des bases généralistes. De plus, il possible de voir que les voies de signalisation induites par des facteurs de croissance (TGF, EGF, IGF, etc.) sont trouvées dans presque toutes les bases de données. En revanche, les voies de signalisation activées par les médicaments ou les voies lipidiques sont rarement trouvées dans les bases de données généralistes.

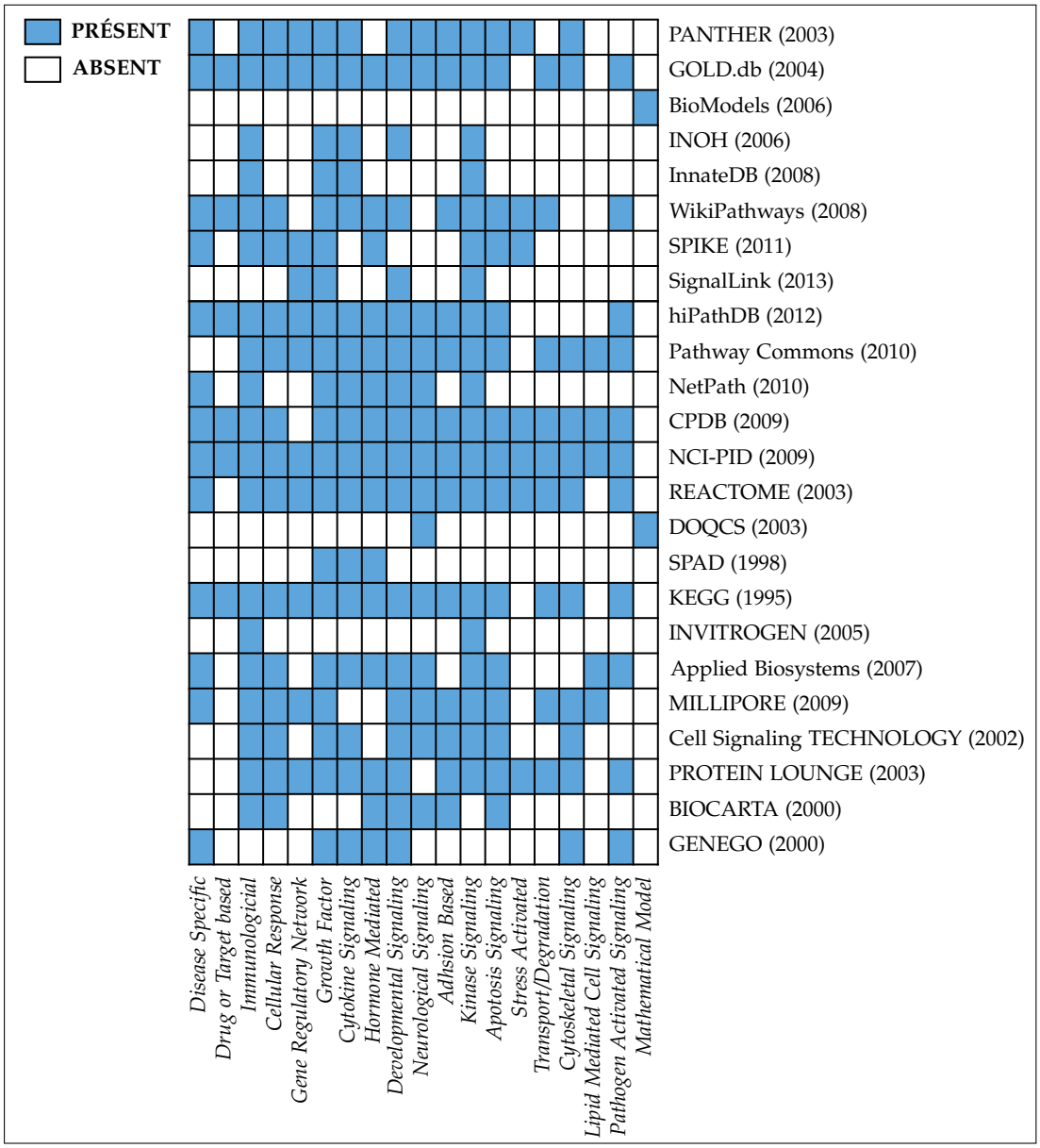


Figure 7 – ***Liste des bases de données de voies de signalisation et leurs types de données disponibles, inspiré par [Chowdhury and Sarkar, 2015].***

Cette figure illustre une matrice représentant les différents types de voies de signalisation (axe X) disponibles dans différentes bases de données (axe Y). La couleur bleue est utilisée pour représenter la présence d'une des voies de signalisation (19 au total) disponible dans l'une des 24 bases de données sélectionnées.

Modélisation de la signalisation, le formalisme Cadbiom

Afin de déchiffrer la dynamique des voies de signalisation, les modèles mathématiques ont été développés en utilisant différentes stratégies [Hans A. Kestler et al., 2008, Nesma ElKalaawy and Amr Wassal, 2015]. Pratiquement tous les types de formalismes, décrits dans la section 1.1.4 en ont été utilisés pour modéliser des voies de signalisation (la signalisation du récepteur des cellules T a été modélisée avec un réseau booléen [Saez-Rodriguez et al., 2007], les équations différentielles ont permis de créer un modèle de la signalisation du facteur de croissance épidermique [Chen et al., 2009], il existe un modèle basé sur des règles de la signalisation des tyrosines kinases JAK [Barua et al., 2009], etc.)

[Andrieux et al., 2014] ont développé un formalisme discret qualitatif spécialement orienté pour la signalisation cellulaire à grande échelle (des milliers d'objets) [Andrieux et al., 2014]. Le langage Cadbiom [1] (*Computer Aided Design of Biological Models*) est un formalisme basé sur les transitions gardées [Antoine B. Rauzy, 2008].

Cadbiom est un formalisme booléen où les nœuds représentent des biomolécules (protéines, complexes, gènes, etc.) et peuvent avoir deux états possibles : actif ou inactif. Une réaction biologique est formalisée par une ou plusieurs transitions représentant la propagation du signal. Une transition gardée est composée d'une biomolécule d'entrée vers une biomolécule de sortie en fonction de certaines conditions appelée garde. Les conditions sont des expressions logiques impliquant des biomolécules en tant qu'activateurs ou inhibiteurs et la réaction biologique n'a lieu que si les entrées sont présentes et que la condition soit vraie. Par exemple, si la formation du complexe AB est activée par la protéine C alors cette réaction sera représentée par deux transitions, l'une liant A vers AB et l'autre liant B à AB. Le signal ne pourra se propager que si A, B et C sont actifs (figure 8).

À chaque étape de la simulation, certaines transitions seront déclenchées en fonction des conditions remplies par l'état du modèle (l'ensemble des nœuds actifs). Une réaction biologique est considérée comme la propagation de l'information de biomolécule en biomolécule. Effectuer une transition en Cadbiom implique donc de désactiver le nœud entrant et d'activer le nœud sortant, les activateurs eux par contre restent actifs (figure 9).

Pour surmonter les limites des modèles synchrones ou asynchrones, Cadbiom intègre un système d'événements permettant d'éviter les effets de concurrence et de gérer les coopérations. En effet, étant donné qu'il n'y a pas de notion de temps mais seulement le déclenchement de transitions, une transition n'est pas plus ou moins rapide qu'une autre. Comme le montre la figure 9, un chemin plus court qu'un autre sera plus « rapide » (il y aura moins d'étapes à parcourir) et dans le cas de deux transitions concurrentes le chemin le plus court amenant à une de ces transitions aura toujours la priorité. C'est pour cette raison que [Andrieux et al., 2014] ont intégré la notion d'évènement dans le modèle. Un événement est un concept mathématique qui permet d'introduire des retards dans une ou plusieurs transitions au fil du temps. Toutes les transitions sont associées à un évènement h_i qui permet de synchroniser ou

1. `http://cadbiom.genouest.org`

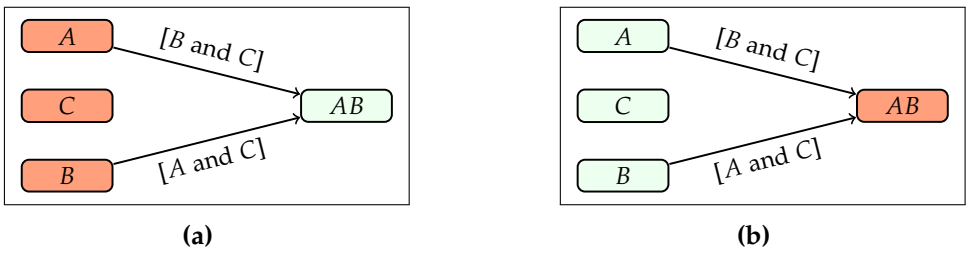


Figure 8 – *Exemple simple d'un modèle Cadbiom.*

Exemple de la formalisation et de la simulation en Cadbiom de la formation du complexe AB activée par C. (a) L'état intial du modèle, les protéines A, B et C sont présentes ; (b) Les conditions sont remplises pour que les deux transitions propagent le signal.

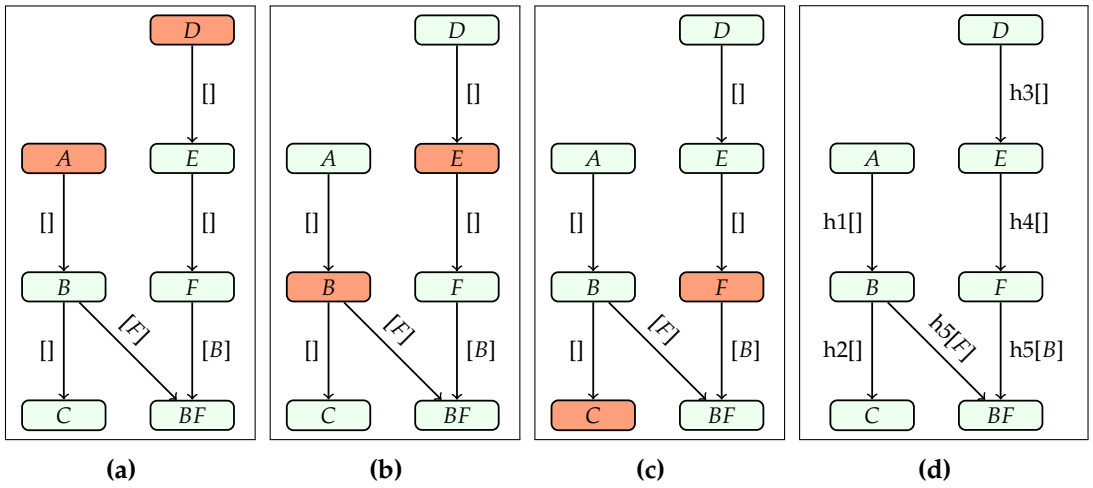


Figure 9 – *Exemple d'un modèle Cadbiom possédant une limitation temporelle.*

Exemple d'un modèle où la molécule BF n'est jamais atteignable sans la notion d'évènements. (a) L'état initial du modèle, les protéines A et D sont présentes ; (b) Première propagation du signal, B et E sont actifs ; (c) Deuxième propagation du signal, C et F sont actifs, il n'est pas possible d'atteindre BF ; (d) Représentation du même modèle avec des évènements, si la transition h_1 est retardée alors il est possible d'activer le nœud BF.

désynchroniser les transitions. Dans l'exemple de la figure 9d, si l'évènement h_1 est retardé alors il devient possible d'activer le nœud BF.

Cadbiom est aussi une application qui permet une analyse basée sur des méthodes de vérification de modèles. Il est possible de calculer l'atteignabilité d'une ou plusieurs propriétés, c'est-à-dire de trouver l'ensemble minimal de biomolécules nécessaires à l'activation ou à la non-activation d'un nœud qui peut être un gène. De plus, Cadbiom nous informe sur l'ensemble des biomolécules activées lors de la propagation du signal. Par exemple dans la figure 9d, pour activer la biomolécule BF, les biomolécules nécessaires sont A et D, et les biomolécules activées lors de la propagation du signal

sont B, E et F.

Cadbiom permet aussi de faire des simulations dès lors que les paramètres d'initialisation sont connus, c'est-à-dire les places actives et l'état de tous les évènements h. À partir de conditions initiales, comme par exemple la présence d'une protéine extra-cellulaire, Cadbiom permet de calculer la chaîne de propagation du signal au sein du modèle. Il est donc possible de voir quelles biomolécules sont activées lors de la propagation du signal dans des conditions initiales données.

1.2 Analyse des réseaux et de leurs simulations

Les données expérimentales disponibles peuvent fournir des connaissances quantitatives détaillées (par exemple les cinétiques des réactions ou la stoechiométrie des molécules) ou simplement une vue qualitative du réseau. Dans cette section, je présente les différentes manières d'analyser soit le graphe biologique soit le résultat des simulations dynamiques. Il n'est pas question ici de savoir comment formaliser et simuler les données mais de savoir que faire ensuite avec ces résultats.

1.2.1 Analyse statique de graphe

Les approches de modélisation statiques reposent principalement sur la structure du réseau et ne nécessitent pas d'informations sur les paramètres cinétiques. Cela les rend généralement applicables à des réseaux à grande échelle. Elles reposent uniquement sur la topologie du réseau, mais elles permettent d'analyser les propriétés fonctionnelles importantes telles que les relations entrée-sortie ou les boucles de rétroaction [Pujol et al., 2010, Vidal et al., 2011]. Ces méthodes topologiques proposent des mesures locales ou globales, c'est-à-dire qui fournissent des informations sur des nœuds individuels ou sur l'ensemble du réseau [Aittokallio and Schwikowski, 2006]. Il existe différentes méthodes qui dépendent évidemment du type de graphe utilisé (voir sous-section 1.1.3). Voici quatre approches couramment utilisées pour les réseaux biologiques :

— **Recherche de chemins**
En théorie des graphes, un chemin entre deux nœuds (u, v) correspond à un ensemble de nœuds et d'arêtes permettant de relier u à v. Par exemple, dans un réseau de signalisation, un chemin entre un stimulus de la matrice extra-cellulaire et l'activation d'un gène correspondra à l'ensemble des réactions (transports, formations de complexes, etc.). La présence de plusieurs chemins entre la même paire de nœuds est une propriété importante qui est considérée comme l'une des raisons de la robustesse de nombreux réseaux cellulaires [Zhou et al., 2002]. Il est possible de trouver l'existence de chemins alternatifs aux voies principales.

— **Centralité**
La centralité est une valeur numérique calculée pour chacun des nœuds du

graphe. Elle permet de trier les éléments du système en fonction de leur importance topologique, et donc d'identifier les acteurs clés d'un processus biologique. Par exemple, il a été démontré que les nœuds hautement connectés dans un réseau d'interactions de protéines sont souvent fonctionnellement importants [Jeong et al., 2001]. Il existe différentes mesures de centralités [Koschützki and Schreiber, 2008], par exemple :

- Centralité de degré : cette mesure correspond au nombre de nœuds auquel le nœud d'intérêt est directement relié. Pour les réseaux dirigés, la centralité à deux degrés, correspondant aux arêtes entrantes et aux arêtes sortantes. Cette mesure est une mesure locale, car seuls les voisins immédiats du nœud d'intérêt sont considérés.
- Centralité de proximité : cette mesure correspond à la somme des distances des plus courts chemins du nœud d'intérêt à tous les autres nœuds. Dans ce cas, cette mesure est une mesure globale.
- Centralité intermédiaire : Cette mesure correspond à la somme des longeurs plus courts chemins entre toutes les combinaisons possibles de nœuds passant par le nœud d'intérêt. Plus la centralité est haute, plus le nœud d'intérêt est un carrefour important du graphe.

— **Recherche de motifs**
La recherche de motifs dans un graphe consiste à trouver certains types d'interconnexions entre les nœuds ou de sous-graphes qui sont répétés de manière significative par rapport à un réseau construit aléatoirement. Il a été prouvé que de tels motifs étaient associés à une fonction biologique [Milo et al., 2002, Alon, 2007]. Différents types de réseaux présentent différents profils de motifs, ce qui suggère un moyen de classification des réseaux [Milo et al., 2004]. Parmi les différents motifs possibles, on peut retrouver les composantes fortement connexes (*Strongly Connected Component* – SCC) qui, au sein d'un graphe orienté, correspondent à un sous-graphe où pour tous couples de nœuds (u, v) il existe un chemin de u vers v (figure 10). Les SCC en biologie sont des boucles de rétroaction, c'est-à-dire des chaînes de réactions par lesquelles les composants peuvent influencer leur propre niveau d'activation. Elles sont présentes dans presque toutes les voies de signalisation connues et ont montré des impacts majeurs sur la dynamique du réseau et la médiation de fonctions biologiques importantes [Brandman and Meyer, 2008].

— ***Clusters* de graphes**
Cette approche consiste à découvrir des sous-graphes de nœuds semblables ou fortement connectés, et susceptibles d'être impliqués dans des fonctions biologiques communes [Dunn et al., 2005]. Elle repose sur l'hypothèse qu'un groupe de nœuds possédant des fonctions biologiques semblables sont susceptibles d'interagir fortement les uns avec les autres dans le réseau. Les *clusters* ne sont pas en général isolés du reste du réseau mais sont connectés avec d'autres *clusters*.

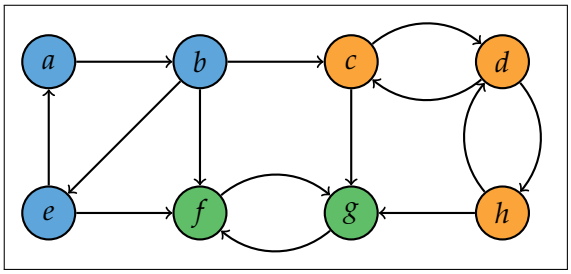


Figure 10 – *Exemple de composantes fortement connexes*

Cette figure représente un graphe dirigé composé de 3 composantes fortement connexes (SCC). Les couleurs des nœuds représentent les SCC : $\{a, b, e\}$, $\{c, d, h\}$ et $\{f, g\}$. Pour tous couples de nœuds (u, v) d'une SCC il existe un chemin de u vers v.

1.2.2 Analyse dynamique

La dynamique d'un modèle est définie comme étant son changement d'état en fonction du temps. Ce changement d'état peut être caractérisé par le changement de concentration des molécules, le déplacement spatial, l'activation de molécules, etc. Nous avons vu dans la section précédente que le temps peut être modélisé de manière continue, discrète ou par itérations. De plus, les variables peuvent posséder des valeurs continues, multi-valuées ou booléennes. Voici trois approches classiquement utilisées pour analyser la dynamique des réseaux biologiques :

— **Recherche d'états stationnaires ou d'attracteurs**
En général, la première analyse de la dynamique d'un modèle est la recherche d'états stationnaires. Il y a deux types d'états stationnaires possibles, soit les valeurs des variables ne changent plus au court du temps, soit les valeurs changent de manières cycliques. En fonction des valeurs initiales des variables, le modèle peut atteindre différents états stationnaires. De plus, il a été prouvé que la présence de plusieurs états stationnaires était liée à l'existence d'un circuit positif dans le graphe [Richard and Comet, 2007, Didier and Remy, 2012]. On définit un attracteur comme étant un état où le système tend à s'approcher au court du temps. La recherche d'attracteurs a été appliquée à la transition de l'état des cellules souches et celui des cellules différenciées [Furusawa and Kaneko, 2012].

— **Évaluation des comportements du modèle**
Calculer la dynamique d'un modèle permet de trouver les conditions initiales influençant le comportement du modèle. Cette stratégie peut être utilisée pour optimiser la production de certaines molécules ou pour trouver les différents chemins possibles. Si des données expérimentales sont disponibles, il est aussi possible de comparer les résultats de simulation avec ces données expérimentales et d'ajuster en conséquence le modèle afin qu'il décrive au mieux la réalité. Cette méthode appelée *multi-experiment-fitting* a été implémentée dans un outil MATLAB : *PottersWheel* [Maiwald and Timmer, 2008, Maiwald et al., 2012b].

Cette approche a été utilisée pour étudier le temps de demi-vie des molécules STAT dans la signalisation de la voie JACK-STAT [Maiwald et al., 2012a].

— **Analyse des perturbations**
Une fois que le modèle est correctement ajusté, une stratégie consiste à créer différentes perturbations dans le modèle pour comparer la dynamique du modèle perturbé avec celle du modèle d'origine. Par exemple, l'inhibition d'une molécule par un médicament peut être modélisée en forçant la valeur de cette molécule à rester à 0. En modifiant les fonctions de régulation, des perturbations plus subtiles peuvent être définies comme la mutation d'une région d'un promoteur le rendant insensible à un régulateur donné. Par exemple, [Traynard et al., 2016] ont analysé la conservation des propriétés dynamiques en fonction de diverses perturbations pour un modèle logique du cycle cellulaire des mammifères.

Ces stratégies d'analyse souffrent cependant de la multiplication du nombre de solutions dans les grands modèles. Plus un modèle est grand (milliers de réactions), plus la combinatoire des valeurs des variables est élevée et plus les états possibles du modèle sont nombreux et le nombre de solutions sera élevé. Il sera alors plus dur d'évaluer le comportement du modèle Dans le cas de la signalisation cellulaire, le nombre de trajectoires déclenchées par un stimulus sera d'autant plus élevé qu'il peut induire différentes voies et se combiner avec d'autres stimulus.

Une solution abordée par beaucoup d'études est de réduire la taille du modèle pour diminuer cette combinatoire. Toute réduction peut modifier la dynamique d'un modèle. Mais, lorsque les impacts de la réduction sur la dynamique sont bien maîtrisés, l'analyse d'un modèle réduit peut être utilisée pour déduire des propriétés intéressantes du modèle d'origine [Snowden et al., 2017].

Cette stratégie a été appliquée avec succès pour la reconstruction de réseau de signalisation Syk dans les cellules du cancer du sein où [Naldi et al., 2017] ont réduit le réseau en extrayant plusieurs plus petits sous réseaux. Afin de diminuer le nombre d'états possibles, [Steinway et al., 2014] ont réduit le modèle de transition épithéliale-mésenchymateuse en supprimant les nœuds influencés par un unique nœud et influençant un autre unique nœud. Ce qui leur a permis de passer de 70 nœuds et 135 arêtes à un réseau de 19 nœuds et 70 arêtes.

Cependant réduire le modèle peut s'avérer inefficace si on cherche à analyser toutes les solutions de manière exhaustive et sans *a priori*, car la réduction du modèle peut entraîner une perte d'information. C'est pourquoi il est utile de trouver des méthodes d'analyse de milliers de solutions permettant de les regrouper ou de les classer et de leur donner du sens biologique.

1.3 Méthodes de *data-mining* pour traiter les nombreuses solutions

La modélisation des systèmes biologiques complexes est associée à la description de comportements multiples et le nombre de solutions pour répondre à une question biologique est très élevé dans les grands réseaux.

Caractériser un ensemble de solutions nécessite de développer des méthodes appropriées. Il existe plusieurs techniques de fouille de données (*data-mining*) [Liao et al., 2012] consistant à extraire les données importantes ou des tendances particulières d'un grand jeu de données. Les techniques de classification supervisées ou *machine learning* cherchent à associer les données à différentes catégories grâce à l'apprentissage préalable de connaissances existantes [Kotsiantis, 2007]. Les méthodes de classification non supervisées ou méthodes de *clustering* cherchent à regrouper ou classer les données en différents ensembles sans connaissance préalable [Jain, 2010]. Il existe d'autres méthodes de *data-mining* telles que la visualisation de données, les analyses topologiques ou les méthodes de régressions, mais qui sont moins utilisées pour les grands jeux de données et que nous n'évoquerons pas dans ce chapitre.

1.3.1 Méthodes de clustering

L'analyse par *clusters* est un processus de regroupement d'objets (physiques ou abstraits). D'une manière générale, une méthode de clustering cherche (1) à maximiser l'homogénéité de chaque *cluster* (c'est-à-dire de regrouper les objets les plus similaires entre eux) et (2) à minimiser la similarité entre les clusters.

Il existe une différence entre la méthode de *clustering* et l'algorithme de *clustering*. Une méthode de *clustering* est une stratégie générale appliquée pour résoudre un problème de regroupement, alors qu'un algorithme de *clustering* est simplement une instance d'une méthode [Jain, 2010]. Les méthodes de *clustering* peuvent être classées en deux grandes catégories : partitionnement et hiérarchique, en fonction des propriétés des *clusters* générés (table 1).

Méthodes hiérarchiques

Les méthodes de *clustering* hiérarchiques tentent de décomposer l'ensemble des objets en une hiérarchie de groupes. Cette décomposition hiérarchique peut être représentée par un diagramme arborescent appelé dendrogramme (figure 11). Il se présente souvent comme un arbre binaire dont les feuilles sont les objets alignés sur l'axe des abscisses. Chaque nœud est un groupe contenant l'ensemble des objets du sous-graphe dont il est le sommet. Les longueurs des arêtes peuvent représenter la distance entre deux nœuds. Enfin, les *clusters* peuvent être obtenus en coupant le dendrogramme à différents niveaux. Il existe deux approches générales pour les méthodes hiérarchiques : agglomérative (de bas en haut) et divisive (de haut en bas).

Une méthode agglomérative va chercher à regrouper les *clusters* les plus similaires

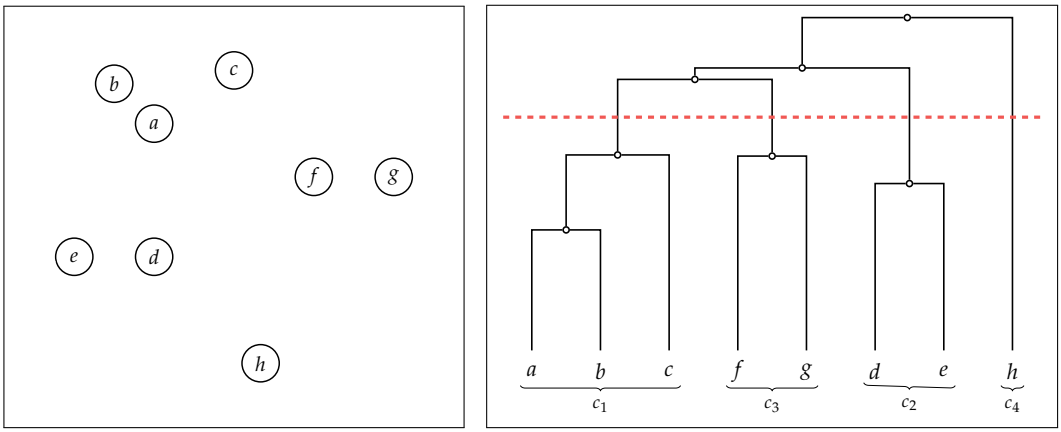


Figure 11 – *Exemple de clustering hiérarchique.*

(a) Données d'exemple : a, b, c, d, e, f, g et h sont des objets avec des coordonnées, plus un objet est proche d'un autre plus il est similaire à celui-ci ; (b) Dendogramme résultant d'un clustering hiérarchique, chaque nœud est un groupe contenant l'ensemble des objets du sous graphe dont il est le sommet. Les longueurs des arètes peuvent représenter la distance entre deux nœuds. La ligne rouge correspond à un délimitation possible et permet d'obtenir les clusters $C_1 = \{a, b, c\}$, $C_2 = \{d, e\}$, $C_3 = \{f, g\}$ *et* $C_4 = \{h\}$.

en un seul *cluster* et construire peu à peu le dendrogramme. Elle va donc débuter avec autant de *clusters* qu'il y a d'objets puis les fusionner de façon successive afin d'atteindre la racine, c'est-à-dire un *cluster* regroupant tous les objets. L'opération de fusion est basée sur la distance entre deux clusters.

Une méthode divisive, opposée à celle agglomérative, commence par un nœud racine qui regroupe tous les objets dans un *cluster* unique et, à partir d'étapes successives, tente de diviser les *clusters* jusqu'à atteindre un nœud unique par objet.

Partitionnement de données

Principe

Les méthodes de partitionnement de données tentent de diviser les données en sous-ensembles ou partition en fonction de certains critères d'évaluation. Il a été prouvé que tester tous les sous-ensembles de données possibles est un problème NP-difficile [Drineas et al., 2004], la plupart des approches cherchent donc une solution approximative. La méthode par partitionnement la plus connue est la méthode *K-means* qui cherche à attribuer les objets au centre d'un des *k clusters* de telle sorte que les distances du *cluster* soient minimisées, pour une valeur *k* donnée (figure 12).

Les méthodes basées sur une grille cherchent à diviser l'espace d'objets en un nombre fini de cellules qui forment une grille séparant les *clusters* (STING, Wave Cluster et CLIQUE). D'autres méthodes sont basées sur la notion de densité, l'idée est

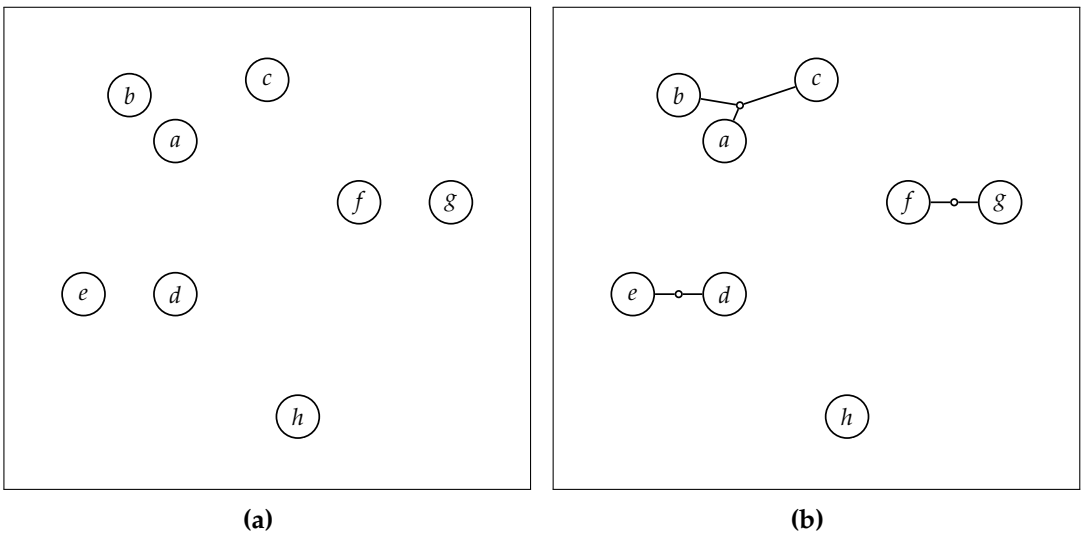


Figure 12 – *Exemple de clustering K-means.*

(a) Données d'exemple : a, b, c, d, e, f, g et h sont des objets avec des coordonnées, plus un objet est proche d'un autre plus il est similaire à celui-ci ; (b) Résultat du clustering K-means avec $k = 4$, les clusters sont représentés par des points et les arrêtres représentent l'inclusion de l'objet au cluster. Les clusters possédant un objet n'ont pas été affichés. Cette méthode permet d'obtenir les clusters $C_1 = \{a, b, c\}$, $C_2 = \{d, e\}$, $C_3 = \{f, g\}$ et $C_4 = \{h\}$.

d'étendre un *cluster* tant que le voisinage d'un rayon donné peut contenir au moins un nombre minimum d'objets (DBSCAN, OPTICS, DBCLASD, DENCLUE et SNN) (table 1).

SNN *clustering*

Certaines méthodes, comme K-means et ses variantes [Kanungo et al., 2002], nécessitent l'utilisation de mesures spécifiques de similarité. D'autres méthodes, comme les méthodes hiérarchiques BIRCH [Zhang et al., 1996] et CURE [Guha et al., 1998], deviennent de plus en plus coûteuses en calcul en fonction de l'augmentation de la taille du jeu de données. Pour ces raisons, il peut être compliqué d'utiliser ces méthodes pour de larges jeux de données fortement hétérogènes. Une approche intéressante pour résoudre ce problème est la méthode de *clustering* basée sur les voisins les plus proches (*Shared Nearest Neighbor* – SNN) [Hamzaoui et al., 2011].

Principe de base L'idée du *clustering* basé sur les voisins les plus proches (*Shared nearest neighbor clustering*, SNN) [Hofman and Jarvis, 1998] est de regrouper les objets en fonction de leurs voisins les plus proches en tenant seulement compte de l'ordre des objets similaires calculé à partir d'un score. Il n'y a pas de notion numérique de distance ou de similarité, mais seulement la notion d'un ordre de similarité entre les objets. Par exemple l'objet *a* a plus de similarité avec l'objet *b* qu'avec l'objet *c*.

	Nom	Principe	Algorithmes	Avantages	Inconvénients
Méthodes hiérarchiques	Agglomérative	Partitionnement d'échantillons	CURE [Guha et al., 1998]	Robuste aux *outliers* et applicable à un grand ensemble de données	Ne prend pas en compte l'inter-connectivité des objets
		Multidimensionnel	BIRCH [Zhang et al., 1996]	Applicable à un grand ensemble de données et sa complexité augmente linéairement	Ne gère que des données numériques
		Basé sur les liens entre les objets	ROCK [Guha et al., 1999]	Robuste et applicable à un grand ensemble de données	Sa complexité dépend de l'initialisation des paramatères
		Basé sur les plus proches paires de points	S-Link [Sibson, 1973]	Il n'est pas nécessaire de spécifier le nombre de *clusters*	Sensible aux *outliers*
	Divisive	Basé sur les barycentres	DIANA [Kaufman and Rousseeuw, 1990b]	Applicable à un grand ensemble de données	
Méthodes de partitionnement	Basée sur des réallocations itératives	Basé sur les barycentres	K-means [Kanungo et al., 2002]	Simple	Sensible aux *outliers*
		Basé sur les médoïdes	PAM [Zhang et al., 2012]	Robuste aux *outliers*	
			CLARA [Kaufman and Rousseeuw, 1990a]	Applicable à un grand ensemble de données	Sensible aux *outliers*
			CLARANS [Ng and Han, 2002]	Robuste aux *outliers*	Coût élevé en calcul
	Basée sur la densité	Taille fixe	DBSCAN [Ester et al., 1996]	Résistant au bruit et manipule des *clusters* de différentes formes et tailles	Ne peut pas gérer les densités variables des *clusters*
		Taille variable	OPTICS [Ankerst et al., 1999]	Résistant au bruit et rapide	Nécessite un grand nombre de paramètres
			DENCLUE [Hinneburg and Gabriel, 2007]	Rapide	Nécessite un grand nombre de paramètres
			RDBC [Santoso and Nisa, 2016]	Manipule des *clusters* de différentes formes et Résistant au bruit	Coût variable en calcul
	Basée sur une grille	Multiple grilles	STING [Wang et al., 1997]	Permet la parallélisation	
			WaveCluster [Sheikholeslami et al., 2000]	Résistant au bruit	Coût variable en calcul
		Basé sur la densité des grilles	CLIQUE [Santhisree and Damodaram, 2011]	Résistant au bruit	

Table 1 – *Résumé des différents types de méthode de clustering et leurs algorithmes, inspiré par [Popat and Emmanuel, 2014].*

Les méthodes de SNN vont chercher à regrouper le maximum d'objets en suivant le principe que si un objet partage les mêmes plus proches voisins qu'un autre objet alors ils doivent être dans le même groupe.

Prenons un exemple d'objets représentés en deux dimensions, $S = \{a, b, c, d, e, f, g, h\}$ (figure 13a). Leur similarité peut par exemple correspondre à la distance qui les sépare, l'objet a est plus proche (similaire) de l'objet b que de l'objet c. D'après cette figure, nous pouvons voir que les meilleurs *clusters* possibles seraient $C_1 = \{a, b, c\}$, $C_2 = \{d, e\}$, $C_3 = \{f, g\}$ et $C_4 = \{h\}$. Partons du principe que pour trouver ces quatre *clusters*, il faudrait que les voisins les plus proches d'un groupe de k objets soient ces mêmes objets, donc forment une clique [1]. Par exemple, les voisins les plus proches du groupe $\{a, b, c\}$ sont $\{a, b, c\}$ pour a, pour b et pour c. À noter qu'en général le voisin le plus proche d'un objet est lui même. Si on analyse les $k = 2$ plus proches voisins de chacun des objets (figure 13b), on peut former les *clusters* $C_1 = \{a, b\}$, $C_2 = \{d, e\}$ et $C_3 = \{f, g\}$. Par contre si on analyse les $k = 3$ plus proches voisins alors on récupère le *cluster* $C_1 = \{a, b, c\}$ (figure 13c). Dans notre cas la meilleure solution est de trouver un k variable en fonction des *clusters* que l'on veut trouver (figure 13d).

De telles méthodes sont connues pour être applicables à un grand ensemble de données, robustes aux données bruitées et elles ne nécessitent pas de représentation spécifique des données. Ces propriétés les rendent applicables à tous types de données, quels que soient les objets ciblés et les mesures de similarité requises [Hamzaoui et al., 2011]. L'algorithme hiérarchique ROCK [Guha et al., 1999] est une méthode SNN, car il se base sur le degré de chevauchement entre les ensembles voisins d'éléments des *clusters* pour fusionner ces *clusters*. Cependant ce critère de fusion peut donner lieu à des *clusters* composés de longues chaînes de sous-*clusters* dans lesquels les éléments d'une extrémité de la chaîne sont très différents de ceux de l'autre extrémité. De plus, ROCK nécessite de fixer la taille k du voisinage. Il existe aussi une variante de l'algorithme DBSCAN [Ester et al., 1996] avec une notion de plus proches voisins [Ertöz et al., 2003] qui semble détecter des *clusters* localement plus denses qu'avec l'algorithme DBSCAN classique. Néanmoins, cette méthode nécessite elle aussi de fixer la taille k du voisinage.

Méthode de Houle [Houle, 2008] a proposé une méthode de clustering SNN qui prend en compte la variation de ce nombre de voisins les plus proches k. Le *Relevant Set Correlation (RSC) model* n'exige pas que l'utilisateur choisisse le nombre de voisins les plus proches ou spécifie un nombre de *clusters* en particulier. La méthode de regroupement n'est pas guidée par un critère d'optimisation globale mais par un critère d'optimisation locale des membres des clusters. Pour chaque objet, le modèle définit un rayon optimal maximisant la qualité du *cluster*. Puis un algorithme glouton (*greedy*) est ensuite appliqué pour garder les *clusters* selon leurs qualités et leur chevauchement avec les autres clusters.

1. Une clique est un sous-graphe complet, c'est-à-dire un sous-ensemble de sommets tels que chacun est connecté à tous les autres.

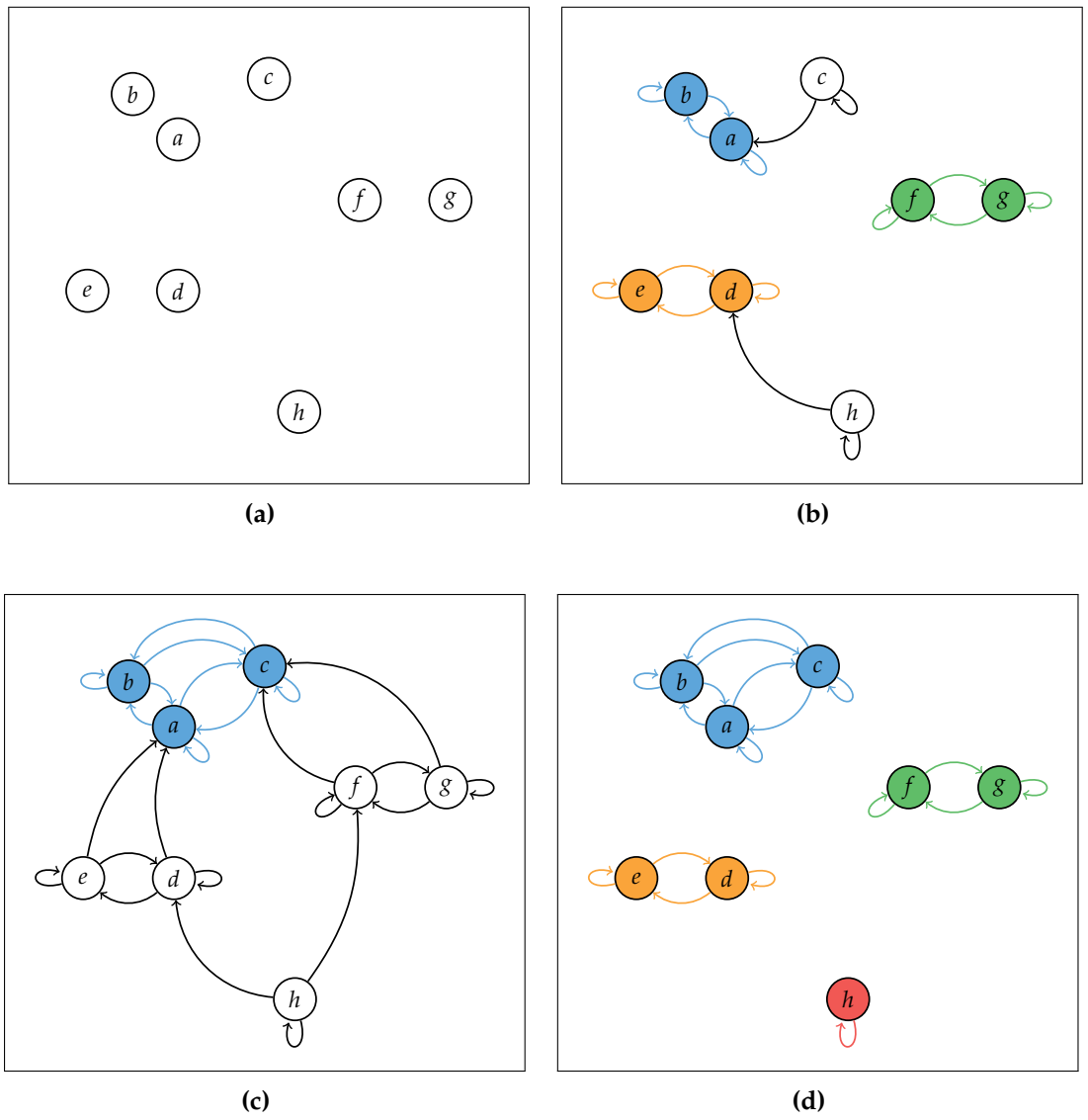


Figure 13 – *Exemple de Shared nearest neighbor clustering.*

Données d'exemples : a, b, c, d, e, f, g et h sont des objets avec des coordonnées. (a) Les voisins ne sont pas représentés ; (b) Représentation des $k = 2$ voisins les plus proches pour chacun des objets ; (c) Représentation des $k = 3$ voisins les plus proches pour chacun des objets. (d) Représentation des regroupements des objets, le k est variable en fonction des groupes. Quand les plus proches voisins forment une clique, les nœuds et arrêtes en question possèdent une couleur et forment un cluster.

	Q							
	1	2	3	4	5	6	7	8
a	a	b	c	d	e	f	g	h
b	b	a	c	e	d	f	g	h
c	c	a	b	f	g	d	e	h
d	d	e	a	h	f	b	c	g
e	e	d	a	b	h	c	f	g
f	f	g	c	a	d	h	b	e
g	g	f	c	h	a	d	b	e
h	h	d	f	e	g	a	c	b

Table 2 – *Exemple de fonction $Q(o)$.*

Le modèle RSC a besoin d'une fonction $Q(o)$ qui pour un objet o retourne une liste composée de tous les objets du domaine triés en fonction de leur similarité avec o. On notera l'ensemble des k plus proches voisins de o comme $Q(o,k)$. Par exemple, dans la figure 13, la fonction $Q(o)$ serait définie par la table 2 où la similarité serait la distance entre chaque paire d'objets.

Comme nous l'avons vu, nous avons besoin de trouver les k plus proches voisins d'un groupe d'objets. Houle définit donc une fonction de corrélation inter-ensembles $R(A,B)$ permettant de calculer un score de corrélation entre deux ensembles d'objets, cette fonction correspond à la généralisation de la corrélation de Pearson pour deux ensembles d'objets :

$$R(A,B) = \frac{|S|\left(\frac{|A\cap B|}{\sqrt{|A||B|}} - \frac{\sqrt{|A||B|}}{|S|}\right)}{\sqrt{(|S|-|A|)(|S|-|B|)}} \tag{1.1}$$

où A et B sont des ensembles d'objets et S est l'ensemble du domaine.

Afin de trouver les groupes d'objets partageant un maximum de plus proches voisins, il définit la fonction de corrélation intra-ensemble $SR_1(A)$ comme la moyenne des scores de corrélation inter-ensembles définie ci-dessus entre l'ensemble A et les $|A|$ voisins les plus proches de chaque $o \in A$:

$$SR_1(A) = \frac{1}{|A|}\sum_{o\in A} R(A, Q(o,|A|)) \tag{1.2}$$

Afin de pouvoir interpréter la fonction de corrélation intra-ensemble $SR_1(A)$, il est nécessaire de la comparer avec la corrélation moyenne de données choisies de façon aléatoire. Houle définit donc un score de corrélation significative intra-ensemble

$Z_1(A)$ comme le zScore de $SR_1(A)$ et un score de corrélation significative inter-ensembles $Z_1(A, B)$ comme le zScore de $R(A, B)$:

$$Z_1(A) = \sqrt{|A|(|S| - 1)}SR_1(A) \tag{1.3}$$

$$Z_1(A, B) = \sqrt{(|S| - 1)}R(A, B) \tag{1.4}$$

Houle utilise aussi d'autres fonctions qui ne seront pas expliquées ici, comme par exemple la valeur de contribution d'un objet au sein d'un cluster. Les détails de calcul sont dans la publication.

La méthode GreedyRSC est un algorithme heuristique appliquant le modèle RSC [Houle, 2008]. Il effectue un *soft clustering*, où les *clusters* peuvent se chevaucher et ne couvrent pas nécessairement l'ensemble des données. En plus de la fonction $Q(t)$, il nécessite quatre paramètres :

— x_1 : taille minimale du *cluster*.

— x_2 : taille maximale du *cluster*.

— x_3 : score maximum de corrélation significative inter-ensembles entre deux *clusters*.

— x_4 : score minimum de corrélation significative intra-ensemble.

L'algorithme GreedyRSC cherchera donc à maximiser la fonction $Z_1(A)$ et à minimiser la fonction $Z_1(A, B)$.

J'ai proposé une méthode permettant de réduire le nombre de paramètres de la méthode GreedyRSC à trois paramètres qui sera détaillée dans le chapitre suivant.

En conclusion la méthode RSC a de nombreux avantages, elle :

— est applicable à de grands jeux de données ;

— n'est pas sensible au bruit ;

— est utilisable pour des données très hétérogènes ;

— ne nécessite pas un nombre fixe de *clusters* ;

— prend en compte la variation du nombre de voisins les plus proches ;

— ne nécessite pas un grand nombre de paramètres ;

— permet aux *clusters* de se chevaucher (pour l'algorithme GreedyRSC).

	groupe	actif	drogue	français	masculin	féminin
Supertramp	1				1	
Santana		1			1	
Jimi Hendrix			1		1	
Janis Joplin			1			1
The Who	1		1		1	
M		1		1	1	

Table 3 – *Exemple de relation binaire I de rock stars G et d'attributs M qui les définissent.*

Les méthodes de *clustering* ont pour but de classifier les données en une hiérarchie ou dans différents groupes. Elles ne permettent pas à un élément de faire partie de plusieurs *clusters*, ce qui est peut-être contraignant face à la nature pléiotropique des protéines. Les méthodes de *soft-clustering*, comme le modèle RSC, résolvent cette limitation.

Il existe d'autres méthodes permettant le « chevauchement » des *clusters* et qui sont basés sur la recherche de concepts formels.

1.3.2 Analyse de concepts formels

L'analyse de concepts formels (*Formal Concept Analysis* – FCA) est une méthode de recherche exhaustive d'ensembles maximaux d'éléments partageant les mêmes attributs. Le treillis que génère cette analyse permet de tenir compte de tous les niveaux de précision : depuis les ensembles de quelques éléments partageant de nombreux attributs en commun jusqu'aux plus grands ensembles d'éléments avec peu d'attributs en commun [Wille, 1982, Poelmans et al., 2013].

Avant de définir un concept, il faut définir un contexte formel $K = (G, M, I)$ où G est un ensemble d'objets, M est un ensemble d'attributs qui caractérise les objets et I est la relation binaire $I \subseteq G \times M$ spécifiant quels objets sont spécifiés par quels attributs.

Prenons l'exemple du contexte des « rock stars », G correspond aux célébrités et M aux attributs : « est un groupe », « est encore en activité », « a eu un problème d'addiction connue aux drogues », « est français », « est masculin » et « est féminin » (table 3).

À partir de ce contexte, il est possible de définir les fonctions de dérivation $f(A)$ et $g(B)$ pour $A \subseteq G$ et $B \subseteq M$ tel que $f(A)$ soit l'ensemble des attributs communs à tous les objets de A et $g(B)$ est l'ensemble d'objets partageant tous les attributs de B. Traditionnellement, elles sont notées :

$$A' = f(A) = \{m \in M \mid \forall g \in A : gIm\} \tag{1.5}$$

$$B' = g(B) = \{g \in G \mid \forall m \in B : gIm\} \tag{1.6}$$

Les concepts formels peuvent être définis comme un ensemble maximal d'objets et d'attributs (propriétés, significations, etc.). Un concept (A, B) est donc une paire d'objets $A \subseteq G$ et d'attributs $B \subseteq M$, tel que $f(A) = B$ et $g(B) = A$. A est appelé extension du concept, et B est appelée intention du concept.

Dans notre exemple, les concepts formels de nos rock stars, sont les suivants :

0. $\{\} \times \{\text{français, actif, groupe, masculin, féminin, drogue}\}$
1. $\{\text{The Who}\} \times \{\text{masculin, groupe, drogue}\}$
2. $\{\text{Supertramp, The Who}\} \times \{\text{masculin, groupe}\}$
3. $\{\text{Jimi Hendrix, The Who}\} \times \{\text{masculin, drogue}\}$
4. $\{\text{Supertramp, Santana, Jimi Hendrix, The Who, M}\} \times \{\text{masculin}\}$
5. $\{\text{M}\} \times \{\text{français, actif, masculin}\}$
6. $\{\text{Santana, M}\} \times \{\text{actif, masculin}\}$
7. $\{\text{Janis Joplin}\} \times \{\text{drogue, féminin}\}$
8. $\{\text{Janis Joplin, Jimi Hendrix, The Who}\} \times \{\text{drogue}\}$
9. $\{\text{Santana, The Who, Supertramp, Janis Joplin, Jimi Hendrix, M}\} \times \{\}$

Les concepts formels peuvent être considérés comme des *bi-clusters*, c'est-à-dire des groupes de deux types d'éléments (objets et attributs). Mais à l'inverse des autres méthodes de *bi-clustering*, les concepts formels proposent une détermination exhaustive des groupes possibles de telle sorte qu'il est possible de retrouver les données originales à partir des concepts.

Enfin, les concepts peuvent être partiellement triés par inclusion. Si l'extension A d'un concept (A, B) est incluse dans l'extension C d'un concept (C, D) alors l'intention D sera incluse dans l'intention B. On dira que le concept (A, B) est un « sous-concept » de (C, D) et (C, D) est un « super-concept » de (A, B), ils suivent donc la relation $(A, B) \leq (C, D)$. La relation d'inclusion est notée comme suit :

$$(A, B) \leq (C, D) \Leftrightarrow A \subseteq C \Leftrightarrow B \supseteq D \tag{1.7}$$

En se basant sur l'inclusion des concepts formels, il est possible d'inférer des relations d'implication et d'exclusion entre les groupes d'objets.

L'ensemble de tous les concepts ordonnés par la relation d'inclusion forme un treillis, appelé treillis conceptuel. Le super-concept incluant tous les autres est en haut. S'il existe la relation $(A, B) \leq (C, D)$ et qu'il n'existe pas de concept (E, F) tel que $(A, B) < (E, F) < (C, D)$, alors il existe une arête entre (A, B) et (C, D). La figure 14 représente le treillis des rock stars, chaque nœud est un concept. Elle utilise la notation condensée où pour chaque concept on affiche que ce qu'il le différencie de ses super-concepts et de ses sous-concepts. On peut déduire l'ensemble des objets d'un concept à partir de tous les concepts de celui-ci à la racine et inversement on peut déduire l'ensemble des attributs de ce concept à partir de tous les concepts de

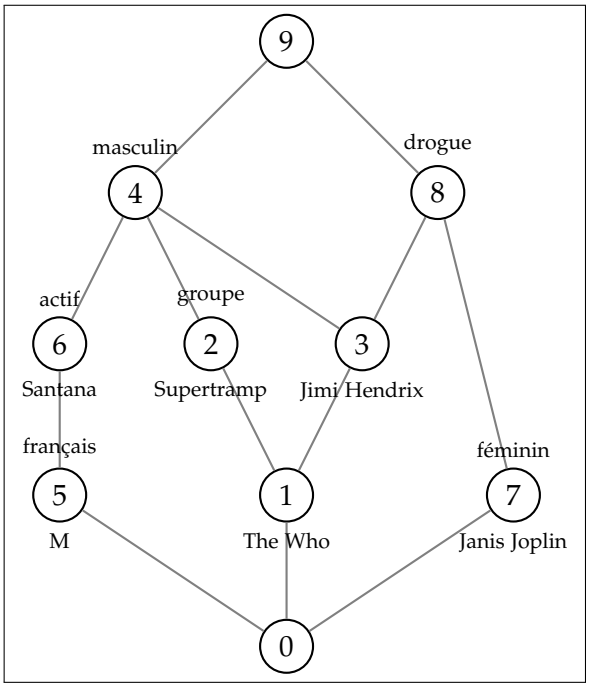


Figure 14 – *Exemple de treillis de concepts formels*

Cette figure représente le treillis de concepts généré à partir du contexte des « rock stars ». Le super-concept incluant tous les autres est en haut. Chaque nœud est un concept. S'il existe la relation $(A, B) \leq (C, D)$ et qu'il n'existe pas de concept (E, F) tel que $(A, B) < (E, F) < (C, D)$, alors il existe une arête entre (A, B) et (C, D). Pour chaque nœud, on affiche au-dessus les attributs qu'il partage avec tous ses sous-concepts et en dessous les objets qu'il partage avec tous ses super-concepts. Par exemple le concept n° 2 correspond à $\{Supertramp, The\ Who\} \times \{groupe, masculin\}$.

celui-ci au concept le plus haut. Le treillis rend ainsi explicite les relations d'inclusion et d'exclusion.

L'avantage est qu'à partir d'un treillis de concepts il est possible de retrouver le contexte formel correspondant. Alors que cela est impossible lorsque l'on utilise des méthodes de *clustering*.

[Blachon et al., 2007] ont appliqué la FCA à l'étude de l'expression des gènes, ce qui a permis d'identifié des groupes de gènes avec des profils d'expressions similaires. Cette méthode a aussi été appliquée à la modélisation de réseaux de signalisation [Videla et al., 2015] dans le but d'apprendre, à partir d'un réseau de connaissances préalables, des modèles logiques booléens.

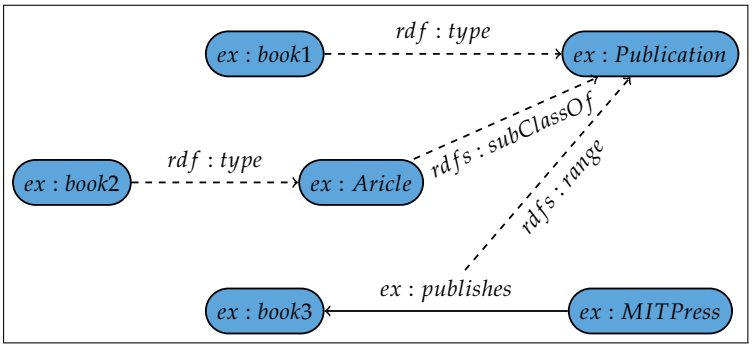


Figure 15 – *Exemple de graphe RDF inspiré par la documentation du W3C*

Représentation graphique du graphe RDF proposé dans la documentation de SPARQL du W3C (`https://www.w3.org`) où les nœuds représentent les concepts à décrire, les arêtes noires indiquent les relations dans les données alors que les arêtes en pointillées indiquent les relations RDF. RDF (Resource Description Framework) est un modèle de données et RDFS (RDF Schema) est un langage représentant des vocabulaires RDF.

1.4 Analyse de la pertinence biologique des solutions grâce à leurs annotations

Afin de valider les prédictions du modèle, il est possible d'analyser leur pertinence biologique en les confrontant avec des données déjà connues. Si une partie des résultats générés représentent des connaissances biologiques absentes lors de l'élaboration du modèle alors cela valide la méthode utilisée. Les résultats qui n'ont pas encore été documentés constituent des hypothèses ou éventuellement des faux-positifs pour un problème donné.

Depuis l'avènement des techniques à haut débit qui génèrent d'énormes quantités de données, l'interprétation de ces données est devenue de plus en plus complexe. C'est pourquoi il est nécessaire d'annoter ces données. On peut définir les données comme étant des collections de faits qui nécessitent une interprétation.

Le modèle de données de RDF (Resource Description Framework) [McBride, 2004] a été conçu pour permettre le traitement automatique des données. Un exemple de graphe RDF est présenté dans la figure 15. Les ressources RDF peuvent être des objets physiques ou des concepts abstraits. De plus, un langage de requêtes nommé SPARQL [Prud'hommeaux and Seaborne, 2008] permet d'interroger une source de données RDF. La description de données en RDF et l'interrogation de ces données en SPARQL permettent d'interpréter ces données de manière automatique.

La signification des annotations constitue l'ensemble des connaissances. La connaissance est générique, elle est donc différente des données « anecdotiques ». On appelle la « gestion des connaissances » (*Knowledge Management*) [Antezana et al., 2009] la façon de structurer de façon systématique ces connaissances.

L'intégration de données est une notion importante de la gestion de connaissances,

car en combinant des données provenant de diverses sources, il est possible d'extraire l'ensemble maximal de connaissance de ces données. La manière dont les données sont stockées dans une base de données est définie par un schéma de base de données. Le problème est que ce schéma doit être assez précis pour décrire au mieux les données mais assez universel pour en recevoir de nouvelles et pour interroger de multiples bases de données simultanément.

Par exemple, les bases de données de systèmes biologiques proposent une gestion des connaissances, car elles apportent une signification aux données. Nous l'avons vu dans la section 1.1.2, des efforts ont été faits pour proposer des schémas (représentations) de données universels (SBML [Chaouiya et al., 2013], BioPAX [Demir et al., 2010], SBGN [Novère et al., 2009]) pour décrire les systèmes.

Les ontologies font partie intégrante de la gestion des connaissances. Une ontologie est une manière formelle de représenter un domaine de connaissances spécifiant les caractéristiques des termes (ou concepts) et leurs relations les uns aux autres [Bard and Rhee, 2004]. Par exemple les termes « chat » et « mammifère » sont liés par la relation « est un ». Les ontologies vont donc proposer un schéma dans un vocabulaire contrôlé pour l'intégration et l'interrogation des données. Il existe différentes ontologies liées à des domaines d'étude particuliers, comme par exemple *Gene Ontology* (GO) [Ashburner et al., 2000, Consortium, 2017] qui propose des schémas de connaissances liés à la biologie moléculaire (voir plus bas), *Sequence Ontology* [Eilbeck et al., 2005] qui définit un ensemble de termes et de concepts décrivant les caractéristiques des séquences biologiques ou encore *OntoBiotope* [Bossy et al., 2015] qui concerne les habitats bactériens. Il existe un portail web regroupant toutes les ontologies biomédicales nommé *BioPortal* [Whetzel et al., 2011].

Avec l'avènement des technologies de Web sémantique, le format OWL (*Web Ontology Language*) [Antoniou and Harmelen, 2004] s'est imposé comme un moyen de décrire des ontologies. OWL est un langage de représentation des connaissances construit sur le modèle de données de RDF (*Resource Description Framework*) [McBride, 2004]. Il permet de définir des ontologies web structurées.

1.4.1 Caractérisation des solutions en identifiant leurs annotations significatives

Gene Ontology (GO) [Ashburner et al., 2000, Consortium, 2017] est une ontologie développée par un consortium mondial visant à proposer une description cohérente des fonctions génétiques et à attribuer à chaque produit de gène des termes biologiques. Il existe trois ontologies qui décrivent « les processus biologiques », « les composants cellulaires » et « les fonctions moléculaires ». Le consortium de *Gene Ontology* a donc répertorié un ensemble de termes biologiques et a défini des relations entre ces derniers (figure 16). Chaque produit de gène est annoté par des termes GO définissant sa fonction, sa localisation cellulaire et son appartenance à des processus biologiques.

Afin de valider les résultats des simulations de modèles biologiques, j'ai décidé d'utiliser l'analyse d'enrichissement d'annotations de termes GO sur des ensembles de gènes. Cela consiste à comparer l'annotation d'un sous-ensemble de gènes par

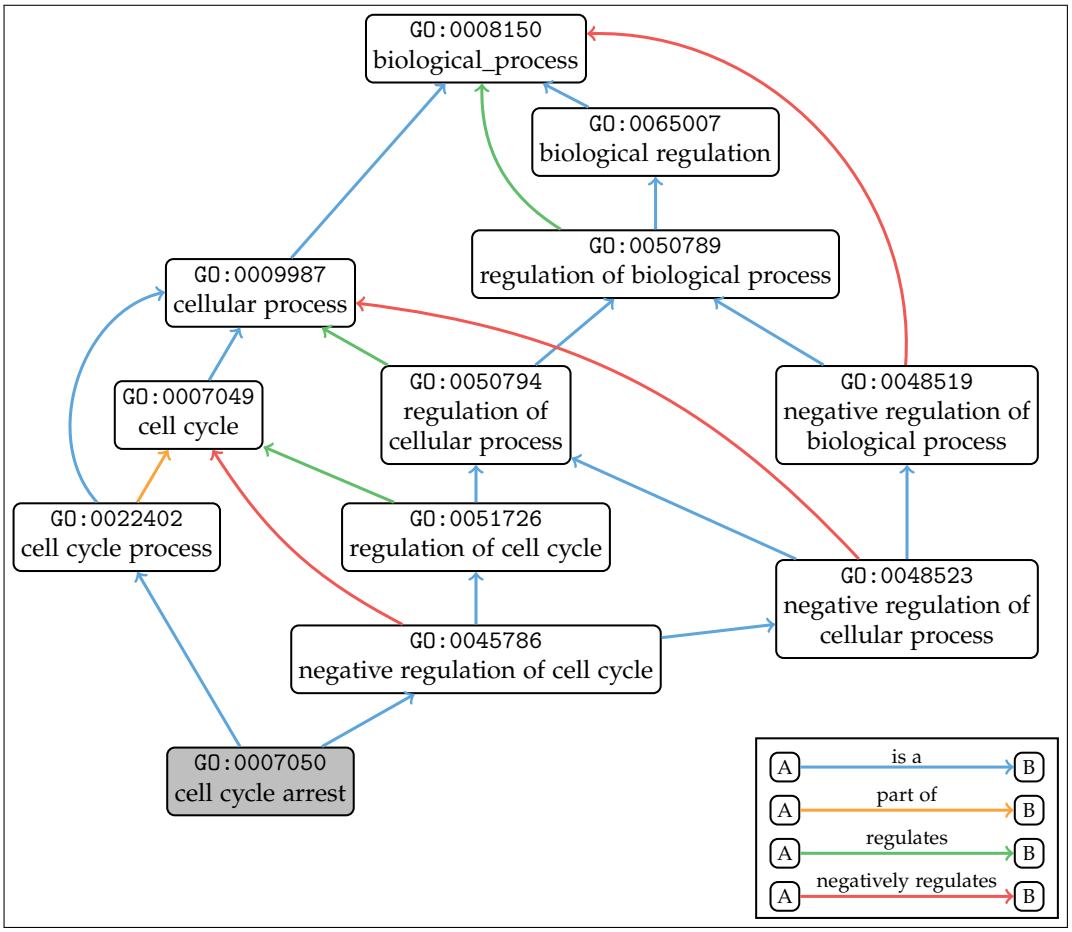


Figure 16 – *Exemple de hiérarchie des termes GO parents de « cell cycle arrest ».*

Représentation de la hiérarchie des termes parents du terme « cell cycle arrest » d'après Gene Ontology, inspiré de la représentation faite par l'outil QuickGO [Binns et al., 2009]. Les nœuds sont des termes GO avec leur identifiant et leur nom, les arêtes correspondent aux concepts qui lient deux termes GO. Chaque couleur correspond à un concept : « is a » (est un) en bleu, « part of » (fait partie de) en orange, « regulates » (régule) en vert et « negatively regulates » (régule négativement) en rouge.

rapport à l'annotation d'un ensemble plus grand, par exemple une liste de gènes surexprimés dans le foie par rapport à l'ensemble du génome. On dira que l'ensemble de gènes « foie » est enrichi en termes GO si ces termes sont plus représentés que dans l'ensemble du génome.

Le protocole de l'outil PANTHER [Mi et al., 2013] cite l'exemple suivant. Si dans le génome humain, composé de 20 000 gènes, 440 gènes sont annotés avec le terme « *induction of apotosis* », alors 2.2% (440 divisé par 20 000) des gènes du génome humain sont impliqués dans l'induction de l'apotose. Dans une liste de 500 gènes, 11 gènes (500 multiplié par 2.2%) sont censés être impliqués dans l'induction de l'apotose. S'il y a plus de gènes impliqués dans cette fonction alors le terme GO est surreprésenté dans cette liste. Une *p-value* est ensuite calculée pour déterminer si la surreprésentation ou la sous-représentation est significative ou non. PANTHER utilise le test binomial pour calculer cette *p-value*.

Une *p-value* inférieure à 0.05 est considérée comme suffisante pour indiquer l'enrichissement d'un terme GO, c'est-à-dire que l'on considère acceptable d'avoir 5 % de chance de considérer un terme GO sur-représenté alors qu'il ne l'est pas. Étant donné que nous calculons la *p-value* de chacun des termes GO, nous multiplions le nombre de fois la possibilité de créer un faux-positif. C'est pourquoi il existe des tests statistiques qui corrigent la *p-value*, comme la correction de Bonferroni utilisée par PANTHER qui multiplie les *p-value* par le nombre de termes GO testés.

1.4.2 Comparaison de plusieurs solutions grâce à la similarité de leurs annotations

Une autre stratégie pour juger de la pertinence biologique d'un ensemble de gènes ou de protéines est d'utiliser une mesure de similarité sémantique des annotations GO de chacun des gènes ou protéines [Pesquita et al., 2009]. Étant donné qu'une ontologie peut être considérée comme un graphe hiérarchique, chaque terme possède une profondeur qui correspond au nombre maximal de parents entre celui-ci et le terme le plus haut de la hiérarchie. Par exemple, le terme « *cell cycle arrest* » a une profondeur de 6 (figure 16). Plus le terme a une profondeur élevée plus il sera bas dans l'ontologie, et inversement plus la profondeur sera faible plus il sera haut dans l'ontologie. Dans le calcul de similarité, il est possible de déterminer les chemins entre deux termes ou de prendre en compte le terme parent le plus bas commun entre deux termes (*Lowest Common Ancestor* – LCA).

Il existe plusieurs mesures de similarité comme la mesure de [Nagar and Al-Mubaid, 2008] dépendant de la profondeur des termes et de la longueur des chemins dans l'ontologie ou la mesure de [Wang et al., 2007] qui tient compte de la longueur des chemins par rapport à tous les termes ancêtres communs.

Dans cette thèse, j'ai décidé d'utiliser la mesure SPBHM (*Shortest Path Based Hybrid Measure*) [Bandyopadhyay and Mallick, 2014] qui calcule la similarité entre deux termes en prenant en compte le chemin le plus court entre ces termes et le LCA et en normalisant le nombre d'annotations de ces termes. Cette mesure semble présenter de très bons résultats, comparée aux autres mesures [Bandyopadhyay and Mallick, 2014].

SPBHM se base sur deux intuitions :

1. Plus le LCA est bas dans l'ontologie (plus il est spécifique), plus la similarité entre les deux termes est élevée.
2. Inversement, plus la distance entre les termes et le LCA est élevée, plus les deux termes sont dissimilaires.

La fonction $sim_{SPBHM}(t_1, t_2)$ (où t_1 et t_2 sont deux termes) est donc composée de deux parties, la partie « similarité » qui prend en compte la spécificité du LCA et la partie « disimilarité » qui prend en compte la distance entre les deux termes et le LCA.

Pour calculer la similarité sémantique entre deux gènes ou deux protéines, il existe différentes stratégies. Prenons deux gènes g_1 et g_2, et les annotations respectives de ces termes $annot(g_1) = \{t_1^1, t_2^1, t_3^1, ..., t_m^1\}$ et $annot(g_2) = \{t_1^2, t_2^2, t_3^2, ..., t_n^2\}$, où m et n sont le nombre de termes annotés pour g_1 et g_2 respectivement. Il est donc possible de calculer la matrice de similarité $SimMat$ de taille $m \times n$ correspondant à la similarité $sim_{SPBHM}(t_i^1, t_j^2)$ entre de tous les termes de g_1 et tous les termes de g_2.

La première stratégie consiste à calculer la similarité moyenne $\frac{1}{m*n} \sum_{\forall(i,j)} SimMat_{i,j}$ mais il suffit que certaines fonctions des gènes soient peu annotées pour sous-estimer la réelle similarité entre ces gènes. La seconde stratégie, proposée par [Bandyopadhyay and Mallick, 2014] et celle que j'ai choisie, est de considérer toutes les valeurs de similarité maximale entre chaque terme de g_1 par rapport à tous les termes de g_2 et vice versa. Le score est la moyenne entre $RowScore$ et $ColScore$, deux fonctions correspondant aux moyennes des valeurs maximales de toutes les lignes et de toutes les colonnes. Elles sont définies de la façon suivante :

$$RowScore = \frac{1}{m} \sum_{i=1}^{m} max_{1 \leqslant j \leqslant n}(SimMat_{i,j}) \tag{1.8}$$

$$ColScore = \frac{1}{n} \sum_{j=1}^{n} max_{1 \leqslant j \leqslant m}(SimMat_{i,j}) \tag{1.9}$$

Cette méthode permet de prendre en compte les fonctions multiples de ces gènes. Définissons $sim_{SPBHM}(A)$ la similarité sémantique des annotations GO d'un ensemble de gènes ou protéines A.

Enfin, une dernière chose à prendre en compte est la significativité de ce score de similarité. Comment savoir si un score de 6 est un score élevé ou non par rapport à nos données. Pour calculer cela, si S est un ensemble gène et $A \in S$ alors j'utilise le zScore de la similarité A en fonction de la similarité moyenne et de l'écart-type :

$$zSim_{SPBHM}(A) = \frac{sim_{SPBHM}(A) - \mathbf{E}[sim_{SPBHM}(A)]}{\sqrt{\mathbf{Var}[sim_{SPBHM}(A)]}} \tag{1.10}$$

où $\mathbf{E}[sim_{SPBHM}(A)]$ et $\mathbf{Var}[sim_{SPBHM}(A)]$ correspondent à la moyenne et la variance des similarités de gènes choisies de façon aléatoire dans S.

Si $zSim_{SPBHM}(A)$ n'est pas proche de 0 alors la similarité ou la disimilarité de A est significative.

Conclusion

Pour étudier les solutions de la dynamique des modèles biologique, il existe différentes stratégies. Ainsi, il est possible de rechercher les conditions permettant d'atteindre les différents états stationnaires et de trouver des cibles thérapeutiques par exemple. Il est aussi possible de réduire le modèle afin de ne garder que les éléments les plus importants. Cependant plus le nombre de solutions est grand plus il sera difficile d'appliquer ces stratégies. C'est pourquoi il existe des méthodes de *data-mining*, comme le *clustering* ou l'analyse en concepts formels, pour regrouper les ensembles de solutions en différentes familles. Une fois que cette classification est réalisée, la deuxième étape consiste à étudier la pertinence biologique de ces familles et pour ce faire il existe des techniques de web sémantique applicable à *Gene Ontology*.

Avec les outils de *data-mining* et de web sémantique, j'ai proposé lors de ma thèse une étude exhaustive sur les trajectoires de la dynamique de signalisation du TGF-β.

Chapitre 2

Un cas pratique le TGF-β

Afin d'analyser les solutions d'un modèle biologique, j'ai travaillé lors de ma thèse sur les trajectoires issues de la signalisation de la protéine TGF-β. Les trajectoires ont été générées par [Andrieux et al., 2014] à partir d'un modèle Cadbiom. Dans un premier temps, j'identifierai différentes familles de trajectoires TGF-β-dépendantes en fonction de leur composition en protéines ayant des annotations de fonctions biologiques significativement enrichies. Cette partie est inspirée de l'article « *Identifying Functional Families of Trajectories in Biological Pathways by Soft Clustering : Application to TGF-β Signaling* » publié lors de la conférence 2017 « *Computational Methods in Systems Biology* » (CMSB) [Coquet et al., 2017]. Puis dans un second temps, je regrouperai les gènes TGF-β-dépendants en fonction des trajectoires qui les activent et je déterminerai quels sont les groupes ayant une pertinence biologique.

2.1 Signalisation du TGF-β

La signalisation induite par le polypeptide *Transforming Growth Factor* (TGF-β) constitue un réseau de signalisation très intéressant du point de vue de son profil multifonctionnel. Le TGF-β a d'abord été décrit comme un inhibiteur puissant de la croissance pour une grande variété de cellules affectant l'apoptose et la différenciation et contrôlant ainsi l'homéostasie des tissus [Hiroaki Ikushima and Kohei Miyazono, 2011]. À l'opposé, la régulation positive et l'activation du TGF-β ont été liés à diverses maladies, y compris la fibrose et le cancer par la stimulation de la prolifération et l'invasion cellulaire [Maozhen Tian et al., 2011] (figure 17). Les effets pléiotropes du TGF-β sont associés à la diversité des voies de signalisation qui dépendent du contexte biologique [Joan Massagué, 2012] (figure 18). La liaison du TGF-β à ses récepteurs induit la phosphorylation de substrats intracellulaires (les protéines R-Smad) qui s'hétérodimerisent avec la protéine Smad4. Les complexes R-Smad/Smad4 se déplacent dans le noyau et régulent la transcription des gènes cibles du TGF-β. Alternativement, les voies non-Smad sont activées par fixation du TGF-β sur ses récepteurs pour activer d'autres réponses cellulaires en aval [Yabing Mu et al., 2012]. Ces voies non-Smad comprennent les voies des protéines kinases activées par les mitogènes (MAPK), telles que JNK et p38, les voies de signalisation GTPase de type Rho et la voie phosphatidylinositol-3-kinase/protéine kinase B (PKB/AKT). Les combinaisons de voies Smad et non-Smad contribuent à la haute hétérogénéité des réponses cellulaires au TGF-β. En outre, de nombreuses molécules de ces voies sont impliquées dans d'autres voies de signalisation activées par d'autres stimuli du microenvironnement, ce qui conduit à une hétérogénéité complexe des voies de signalisation [Kunxin Luo, 2017].

2.2 Présentation des données et du projet

Des approches numériques utilisant des modèles différentiels ont été développées pour décrire le comportement de la voie canonique TGF-β impliquant des protéines Smad [Zhike Zi et al., 2012]. En raison des nombreux composants et du manque de données quantitatives, les voies non canoniques n'ont jamais été incluses dans ces modèles TGF-β. Sur la base du formalisme Cadbiom, [Andrieux et al., 2014] ont intégré les 137 voies de signalisation de la base de données PID [Carl F. Schaefer et al., 2009] en un modèle unique qui intègre notamment toutes les voies dépendantes du TGF-β (voies canoniques et non-Smad). À l'aide de ce modèle, ils ont identifié 15 934 trajectoires de signalisation régulant 159 gènes cibles du TGF-β et ont trouvé des signatures spécifiques pour l'activation de gènes dépendants de TGF-β.

Notez qu'un gène influencé par TGF-β peut coder pour une protéine impliquée ailleurs dans la voie, de sorte que les protéines et les gènes forment des ensembles non disjoints. Les trajectoires sont d'abord soumises à une étape de pré-traitement pour générer un ensemble non redondant de voies de signalisation.

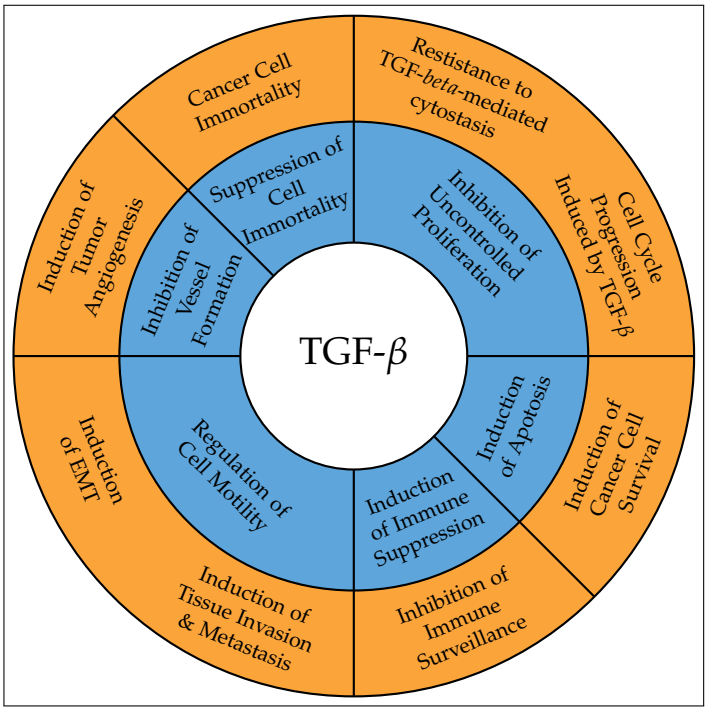


Figure 17 – ***Schéma des différentes fonctions du TGF-β dans un tissu tumoral, adapté de [Maozhen Tian et al., 2011].***

Le TGF-β est un puissant suppresseur de tumeur dans les cellules normales (zone bleue) mais est un promoteur tumoral dans les phases avancées du cancer permettant aux cellules cancéreuses d'acquérir les différents phénotypes associés au cancer (zone orange).

Une trajectoire de signalisation est définie comme un ensemble de protéines requises pour l'activation de gènes dépendants de TGF-β (figure 19A). Chaque trajectoire originale T_k était composée du TGF-β, de protéines de signalisation et d'un seul gène cible (figure 19B). Il y a 321 protéines de signalisation (identifiées par leur identifiant Uniprot) impliquées dans au moins une des 15 934 trajectoires de signalisation. Et ces trajectoires influencent 144 gènes (ce chiffre diffère des 159 gènes décrits dans l'article de [Andrieux et al., 2014], car j'ai identifié les gènes par leur identifiant Uniprot et non par leur identifiant Cadbiom dont certains correspondaient aux mêmes uniprots).

Pour comparer les trajectoires en fonction de leur composition en protéines, nous avons d'abord retiré le TGF-β qui appartenait à toutes les trajectoires. Ensuite, nous avons observé que plusieurs trajectoires étaient composées des mêmes protéines de signalisation, mais ne différaient que par les gènes cibles. Nous avons décidé de séparer les gènes cibles des trajectoires et de représenter l'influence d'une trajectoire sur l'expression d'un ou de plusieurs gènes (figure 19C). Les motivations étaient (1) d'éviter la duplication artificielle des trajectoires, et (2) d'avoir un modèle qui représente explicitement le fait qu'une seule chaîne de réactions peut influencer plusieurs gènes.

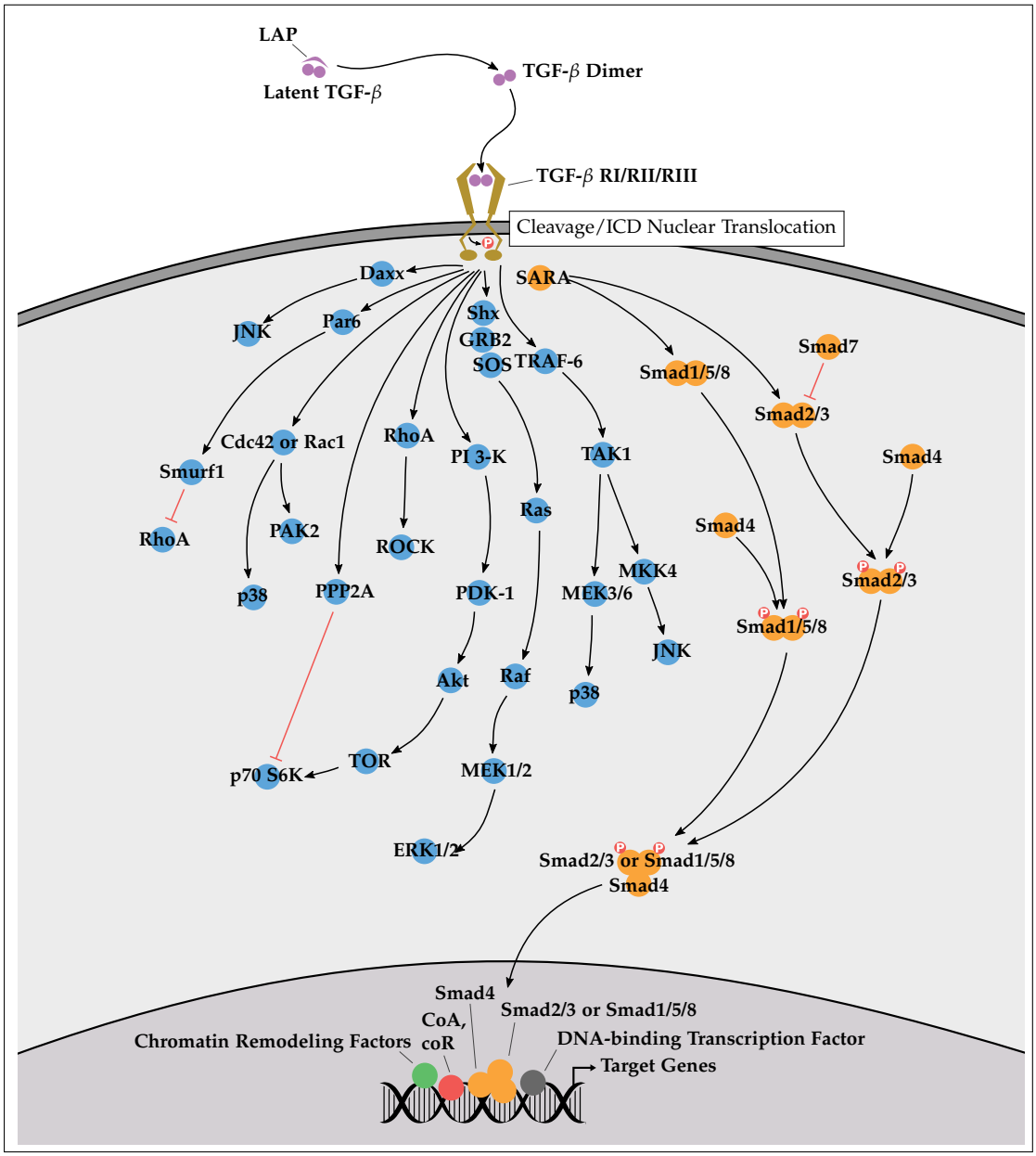


Figure 18 – *Carte des voies de signalisation du TGF-β.*

Figure représentant la signalisation du TGF-β inspirée de la figure de R&D systems[1]*. Le TGF-β n'est pas fonctionnel dès sa traduction. La protéine est synthétisée sous forme latente, constituée de la région Latency Associated Peptide (LAP) et du peptide mature. Une fois le TGF-β activé, c'est-à-dire la libération du peptide actif, il peut interagir avec un récepteur situé à la membrane de la cellule. En orange est représentée la cascade de signalisation canonique et en bleu les voies de signalisation non Smad.*

1. `https://www.rndsystems.com/pathways/tgf-beta-signaling-pathways`

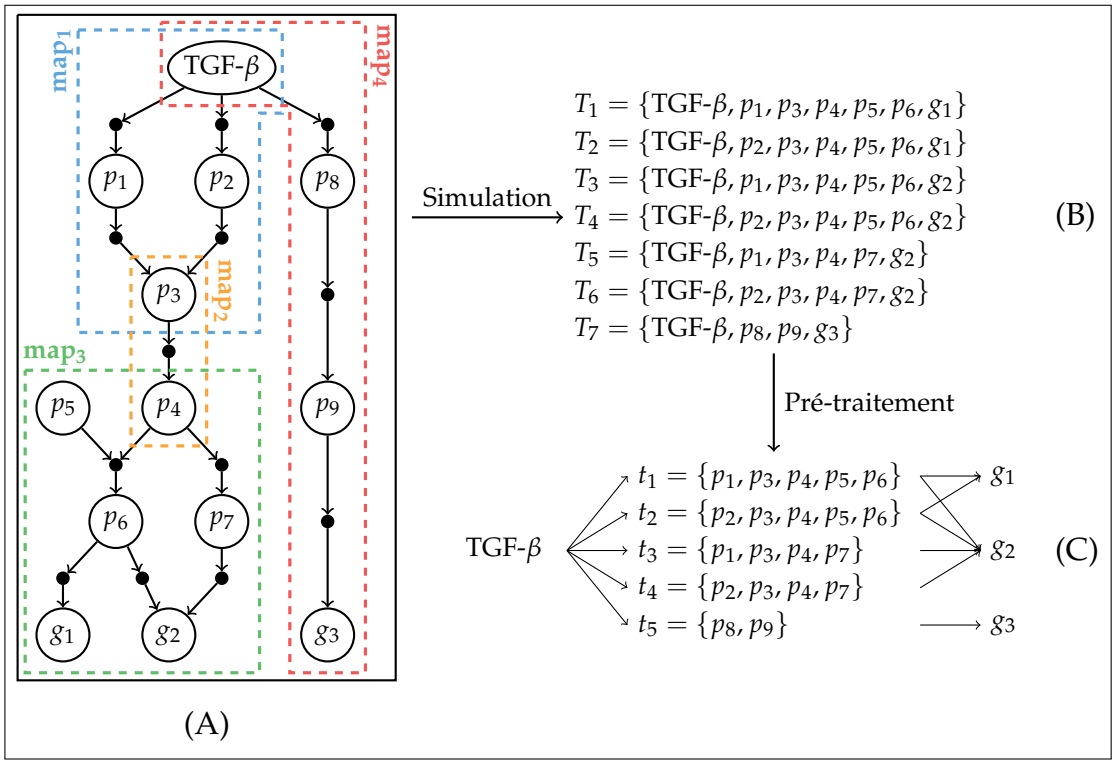


Figure 19 – *Exemple de génération de trajectoires et de leur pré-traitement.*

(A) Le réseau de signalisation est constitué de 4 cartes contenant des protéines, le TGF-β et des gènes. (B) Les trajectoires, identifiées par Cadbiom, sont définies par un ensemble de protéines contenant le TGF-β, des protéines de signalisation (p_i) et des gènes cibles (g_i). (C) Les trajectoires pré-traitées sont limitées aux protéines de signalisation. Après le prétraitement, les trajectoires T_1 et T_3 sont représentées par la trajectoire t_1 ; T_2 et T_4 sont représentées par t_2 ; T_5 est représentée par t_3 ; T_6 est représentée par t_4 ; T_7 est représenté par t_5.

Cette étape de pré-traitement a réduit l'ensemble des 15934 trajectoires à un ensemble de 6017 trajectoires composées de 321 protéines différentes.

À partir de ces données, j'ai réalisé deux études. La première correspond à l'analyse et la classification des trajectoires en fonction de leur composition, afin d'identifier des familles de trajectoires et des modules fonctionnels ayant des fonctions biologiques communes. Et la seconde correspond à l'analyse en concept formel des gènes en fonction des trajectoires qui les influencent, et d'analyser si des gènes influencés par les mêmes trajectoires ont des fonctions biologiques communes. Ces travaux avaient pour but de caractériser de manière exhaustive les voies de signalisations dépendantes du TGF-β.

Tout au long de ces études, nous utiliserons les notations suivantes :

— t_i : une trajectoire post-traitée.
— S : l'ensemble des trajectoires t_i, $\forall(i) \in [1, |S|]$.
— g_j : un gène cible TGF-β dépendant.
— G : l'ensemble des gènes cibles g_j, $\forall(j) \in [1, |G|]$.
— p_k : une protéine faisant partie du réseau de signalisation dépendant du TGF-β.
— P : l'ensemble des protéines p_k, $\forall(k) \in [1, |P|]$.
— I : la relation binaire $I \subseteq G \times S$ spécifiant l'influence d'une trajectoire $t_i \in S$ sur un gène cible $g_j \in G$.

2.3 Analyse des trajectoires de signalisation

Afin de déchiffrer la complexité de la signalisation des réseaux dépendants de TGF-β, nous avons recherché à caractériser ces 6017 trajectoires de signalisation sur la base des protéines impliquées dans les réactions (réactifs, produits et catalyseurs). Cette tâche reste difficile, car elles sont principalement composées de molécules de signalisation dont la modularité et la combinaison sont la base de la plasticité et de l'adaptabilité de la cellule [Nadav Kashtan and Uri Alon, 2005, Sergio G. Peisajovich et al., 2010, John D. Scott and Tony Pawson, 2009]. Dans cette étude, nous avons développé une approche méthodologique pour identifier les familles de trajectoires avec une signature biologique fonctionnelle basée sur leur contenu protéique. La difficulté majeure était la complexité interne du réseau et le fait que certaines protéines peuvent être impliquées dans de multiples familles, comme l'ont suggéré les différents rôles du TGF-β en fonction du contexte. Pour remédier à ces défis, nous avons utilisé la méthode de *clustering* non supervisée de [Houle, 2008]. Les *clusters* correspondent à des familles de trajectoires et peuvent partager des protéines communes. Notre analyse ne repose pas sur une connaissance *a priori* du nombre de *clusters* ni sur l'appartenance d'une protéine à un *cluster*. Sur la base de cette approche, nous avons identifié cinq familles de trajectoires de signalisation. De plus, notre étude a montré que ces cinq groupes sont associés à des fonctions biologiques spécifiques, démontrant ainsi la pertinence du *soft-clustering* pour déchiffrer les réseaux de signalisation cellulaire.

2.3.1 Les trajectoires de signalisation TGF-β sont fortement connectées

Comme illustré dans la figure 20, le nombre de protéines par trajectoire varie de 1 à 50, avec plus de 90% des trajectoires contenant au moins 10 protéines. L'analyse de la répartition de chaque protéine dans toutes les trajectoires a montré une grande hétérogénéité. Plus de 70 protéines sont présentes dans au moins 500 trajectoires et 6

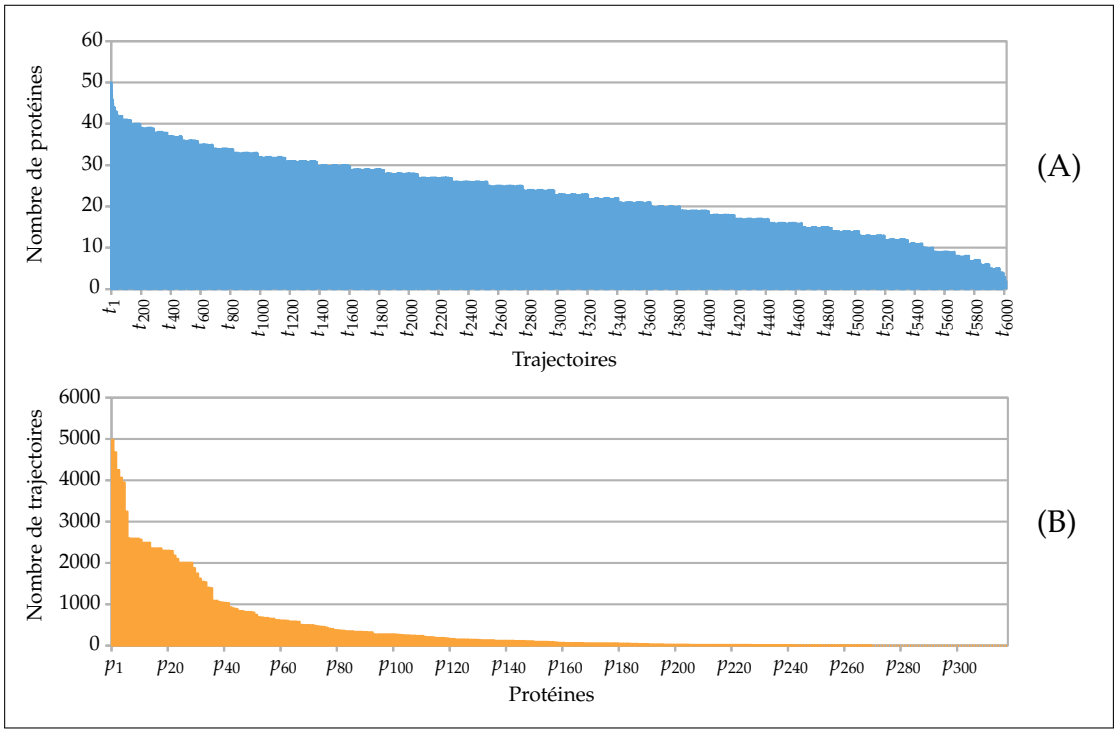


Figure 20 – *Statistiques de la composition des trajectoires.*

Répartition (A) du nombre de protéines pour chaque trajectoire et (B) du nombre de trajectoires impliquant chaque protéine. Ces résultats ont montré que la plupart des protéines sont partagées par de nombreuses trajectoires suggérant un haut degré de connectivité des voies de signalisation dépendantes du TGF-β.

protéines présentes dans plus de 3000 trajectoires (FOS, JUN, ATF2, MAP2K4, ELK1, JAK2). À l'inverse, 75 protéines apparaissent dans seulement moins de 10 trajectoires. Ensemble, ces résultats ont montré que de nombreuses protéines sont partagées par de nombreuses trajectoires suggérant un haut degré de connectivité des voies de signalisation dépendantes du TGF-β.

Méthode de clustering

Nous avons utilisé le modèle RSC pour identifier les *clusters* de trajectoires [Houle, 2008]. Ce modèle utilise comme entrée une fonction $Q(t)$ qui retourne pour chaque trajectoire $t \in S$ une liste de toutes les autres trajectoires triées par leur corrélation décroissante avec t. C'est pourquoi il faut d'abord trouver une manière de définir cette fonction $Q(t)$.

2.3.2 Définition de la fonction $Q(t)$

Une trajectoire $t_i \in S$ est représentée par un vecteur binaire v_i dont la dimension est égale à 321, c'est-à-dire égale au nombre total de protéines contenues dans le modèle. La valeur "1" indique que la trajectoire contient la protéine, et la valeur "0" l'inverse (table 4).

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
t_1	1	0	1	1	1	1	0	0	0
t_2	0	1	1	1	1	1	0	0	0
t_3	1	0	1	1	0	0	1	0	0
t_4	0	1	1	1	0	0	1	0	0
t_5	0	0	0	0	0	0	0	1	1

Table 4 – *Exemple de matrice binaire représentant la composition protéique des trajectoires. Si une protéine p_j est présente dans une trajectoire t_i alors la valeur de la cellule vaut "1" autrement "0".*

Sur la base des vecteurs binaires, nous appliquons la formule de corrélation de Pearson et construisons une matrice de similarité (table 5) :

$$r(t_i, t_j) = \frac{\sum_{k=1}^{n}(t_{i,k} - \overline{t_i})(t_{j,k} - \overline{t_j})}{\sqrt{\sum_{k=1}^{n}(t_{i,k} - \overline{t_i})^2 \sum_{k=1}^{n}(t_{j,k} - \overline{t_j})^2}} \tag{2.1}$$

où $(t_{i,1}, t_{i,2}, ..., t_{i,n})$ et $(t_{j,1}, t_{j,2}, ..., T_{j,n})$ sont les vecteurs des trajectoires t_i et t_j avec leur moyenne $\overline{t_i}$ et $\overline{t_j}$ respective.

	t_1	t_2	t_3	t_4	t_5
t_1	1.000	0.550	0.350	-0.100	-0.598
t_2	0.550	1.000	-0.100	0.350	-0.598
t_3	0.350	-0.100	1.000	0.550	-0.478
t_4	-0.100	0.350	0.550	1.000	-0.478
t_5	-0.598	-0.598	-0.478	-0.478	1.000

Table 5 – *Exemple d'une matrice de corrélation des trajectoires t_i obtenue à partir de la composition en protéines (table 4). Si deux trajectoires t_i, t_j ont exactement les mêmes protéines alors la valeur de la cellule (i, j) est de* 1.0*. Si les trajectoires ne partagent aucune protéine, la valeur est de* 0.0*.*

Pour chaque trajectoire $t_k \in S$, la corrélation de Pearson donne un ordre partiel $< t_i >_{i=1}^{|S|}$ de trajectoires où la relation $i < j$ implique que $r(t_k, t_i) \geqslant r(t_k, t_j)$ (table 6). Si deux trajectoires ont le même score de corrélation, elles sont classées par ordre alphabétique.

Pour rappel, fonction $Q(t)$ retourne pour chaque trajectoire t une liste composée de toutes les trajectoires du triés en fonction de leur corrélation avec t. Nous définissons la fonction $Q(t)$ comme suit :

$$Q(t_k) =< t_i >_{i=1}^{|S|} \qquad \forall (i,j) \in [1,|S|]^2, i < j \Rightarrow r(t_k,t_i) \geqslant r(t_k,t_j) \tag{2.2}$$

	Q				
	1	2	3	4	5
t_1	t_1	t_2	t_3	t_4	t_5
t_2	t_2	t_1	t_4	t_3	t_5
t_3	t_3	t_4	t_1	t_2	t_5
t_4	t_4	t_3	t_2	t_1	t_5
t_5	t_5	t_3	t_4	t_1	t_2

Table 6 – *Exemple d'ordre partiel de toutes les trajectoires pour chaque trajectoire t_i. Toutes les trajectoires sont triées en fonction de leur score de corrélation Pearson.*

Dans la section 1.3.1, nous avons vu que l'algorithme GreedyRSC nécessite quatre paramètres :

— x_1 : taille minimale du *cluster*.

— x_2 : taille maximale du *cluster*.

— x_3 : score maximum de corrélation significative inter-ensembles entre deux *clusters*.

— x_4 : score minimum de corrélation significative intra-ensemble.

Cependant il est possible de réduire ce nombre à trois paramètres de la façon suivante.

La taille minimale x_1 d'un *cluster* signifie que tous les *clusters* doivent être composés d'au moins x_1 éléments. Pour respecter cette contrainte, nous devons choisir le score minimum de signification $x_4 = \sqrt{x_1(|S|-1)}$ où $|S|$ est le nombre d'éléments.

Nous pouvons prouver le calcul du score minimum de signification comme suit :

Soit A un *cluster*,

$$\begin{aligned} |A| &\geqslant x_1 \geqslant 0 \\ SR_1(A)\sqrt{|A|(|S|-1)} &\geqslant SR_1(A)\sqrt{x_1(|S|-1)} \\ Z_1(A) &\geqslant SR_1(A)\sqrt{x_1(|S|-1)} \end{aligned}$$

où $SR_1(A)$ est la mesure de corrélation intra-ensemble. Une valeur de 1.0 indique une identité parfaite entre les trajectoires de A, alors qu'une valeur approchant 0 indique une différence totale. Parce que $0 \leqslant SR_1(A) \leqslant 1$, nous avons besoin d'un score minimum de signification x_4 égal à $\sqrt{x_1(|S|-1)}$ pour s'assurer que tous les *clusters* ont un minimum de x_1 éléments.

Regroupement des trajectoires

Pour étudier la robustesse du modèle RSC, nous avons effectué 64 analyses ($= 4 \times 4 \times 4$) avec quatre valeurs différentes pour les paramètres de l'algorithme greedyRSC x_1, x_2 et x_3 :

— $x_1 = [2, 5, 10, 50]$

— $x_2 = [1500, 2000, 3000, 6000]$

— $x_3 = [0.1, 0.5, 1.0, 2.0]$

Étant donné que l'algorithme greedyRSC est une méthode de clustering non déterministe, nous avons effectué cinq réplicats de chacune des 64 analyses de clustering. Nous avons donc réalisé 320 *clusterings* sur les 6017 trajectoires. Chaque *clustering* a généré 3, 4 ou 5 *clusters* correspondant en tout à 1139 *clusters* différents de trajectoires.

Chaque *clustering* a généré des *clusters* plus ou moins similaires avec les autres analyses. Afin de comparer la similitude de ces *clusters*, nous avons calculé l'indice Jaccard pour toutes les combinaisons de paires de *clusters* en fonction du nombre de trajectoires partagées entre eux. En utilisant une classification hiérarchique de cette similarité entre les *clusters*, nous avons identifié cinq groupes de *clusters* (figure 21). Chacun de ces groupes possède des *clusters* provenant de différents *clusterings*, il n'y a pas deux *clusters* générés par le même *clustering* dans un même groupe.

Pour caractériser les cinq groupes de clusters, nous avons analysé pour chaque groupe le nombre de *clusters*, le nombre de trajectoires associées à ces *clusters* (taille moyenne du cluster) et la redondance entre les *clusters* (union et intersection). Comme décrit dans le tableau 7, les groupes 1 et 2 ont été caractérisés par des *clusters* générés à partir de 320 et 319 regroupements respectivement. Donc presque la totalité des 320 *clusterings* a généré au moins deux clusters catégorisés dans ces deux groupes, ce qui suggère une classification robuste des trajectoires. Les trois autres groupes 3, 4 et 5 contiennent des *clusters* générés à partir de 160 regroupements suggérant une plus grande sensibilité aux paramètres. La taille moyenne du groupe exprimée en nombre moyen de trajectoires contenues dans les *clusters* varie de 202 dans le groupe 4 à 2170 dans le groupe 1.

Nous définissons le noyau d'un groupe comme l'intersection des *clusters* d'un groupe. C'est l'ensemble des trajectoires qui appartiennent à tous les clusters du groupe, de sorte qu'il permet de se concentrer sur les trajectoires les plus stables du groupe. Les noyaux des groupes 1 et 2 contiennent respectivement 1485 (57%) et 1458 (67%), tandis que la taille des groupes 3, 4 et 5 était identique ou très similaire à l'union des *clusters*. Pour caractériser davantage ces noyaux, nous avons déterminé le nombre de protéines impliquées dans les trajectoires et le nombre de gènes cibles activés par ces trajectoires de signalisation. Alors que le nombre total de protéines impliquées dans les trajectoires de chaque groupe était presque similaire, le nombre de gènes cibles était très variable. Les 1485 trajectoires du noyau 1 ont été caractérisées par 114 protéines, mais seulement 3 gènes cibles suggérant des combinaisons complexes de signalisation pour ces gènes. Au contraire, les trajectoires du noyau 4 qui

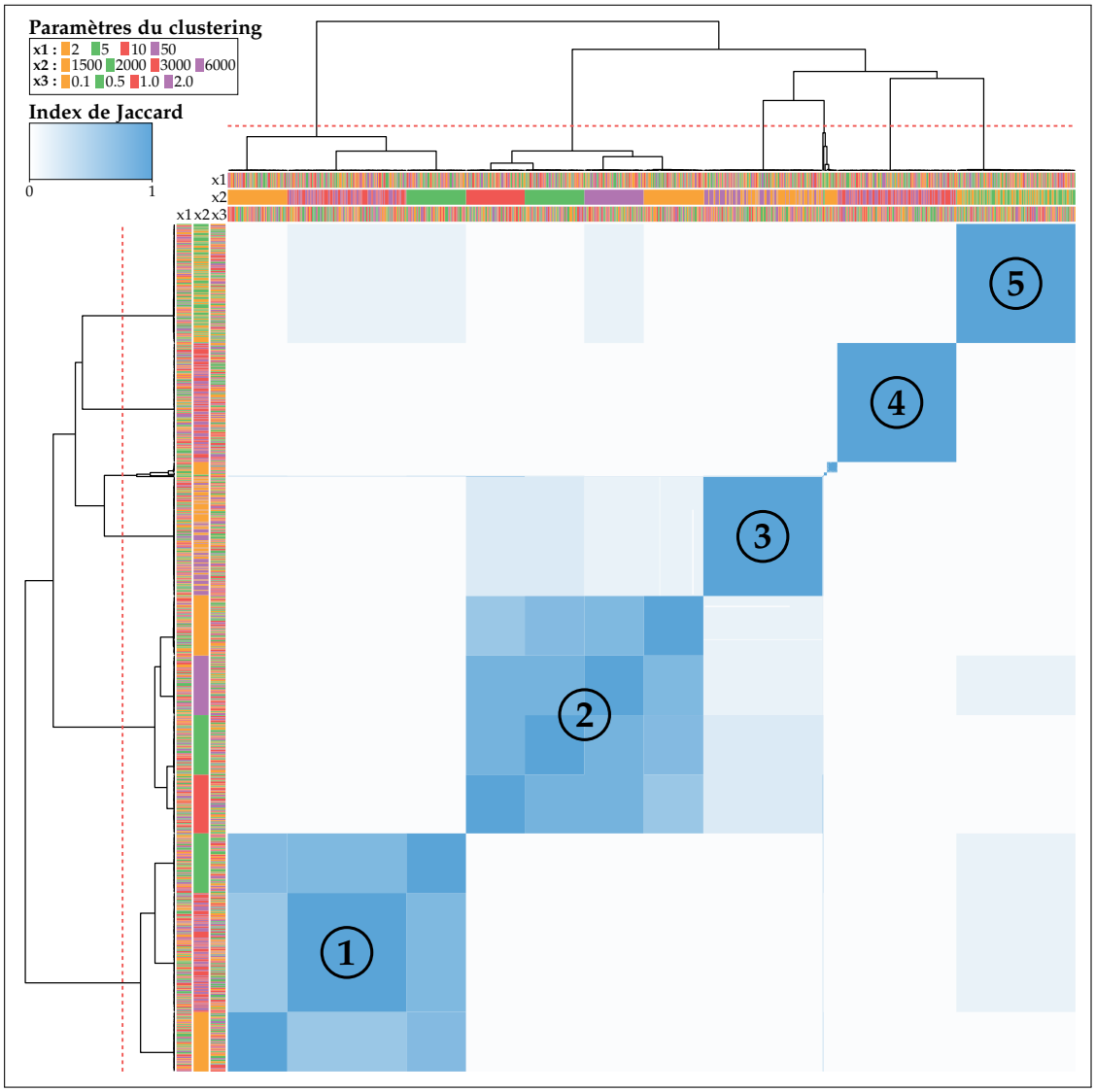


Figure 21 – *Classification hiérarchique des clusters de trajectoires.*

Classification hiérarchique des clusters générés par les 320 clusterings à l'aide de paramètres (x_1, x_2 et x_3) selon leurs similitudes (indice Jaccard). Les valeurs des paramètres sont indiquées par quatre couleurs différentes. Chaque cluster est le résultat d'un clustering caractérisé par une combinaison des trois paramètres. Les cinq groupes de clusters identifiés sont numérotés de 1 à 5 et l'intensité de couleur bleue indique l'indice Jaccard entre deux clusters.

	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
Nombre de *clusters*	320	319	160	160	160
Taille moyenne des clusters (Nombre de trajectoires)	2170.0	1905.58	899.62	202.0	877.12
Union des *clusters* (Nombre de trajectoires)	2590	2289	904	202	888
Taille des noyau = Intersection des *clusters* (Nombre de trajectoires)	1485	1458	894	202	870
Nombre total de protéines de chaque noyau	114	188	110	156	151
Nombre total de gènes de chaque noyau	3	68	58	19	16

Table 7 – *Statistiques de chaque groupe de clusters de trajectoires*

ne contiennent que 202 trajectoires sont caractérisées par 156 protéines qui activent 19 gènes. Nous reviendrons sur ce point dans la section 2.4.

2.3.3 Identification des protéines sur-représentées dans chaque noyau

Le regroupement des trajectoires a été effectué en utilisant un score de corrélation basé sur la présence et l'absence de protéines. Le noyau de chaque groupe peut être caractérisé par un ensemble de protéines sur-représentées, c'est-à-dire les protéines qui apparaissent plus souvent dans les trajectoires du noyau que ce à quoi nous nous attendions si nous avions choisi le même nombre de trajectoires au hasard (figure 22).

Nous pouvons calculer le niveau de représentation de la protéine pour chaque groupe de trajectoires avec un zScore de la fréquence des protéines :

$$Z_A(p) = \frac{N_A(p) - F_S(p)|A|}{\sqrt{F_S(p)|A|(1 - F_S(p))}} \tag{2.3}$$

où p est une protéine et A est un groupe de trajectoires, $N_A(p)$ est le nombre de trajectoires de A impliquant p, $F_S(p)$ est la fréquence de p dans toutes les trajectoires S et $|A|$ est la taille du groupe.

Le zScore permet de normaliser la fréquence des protéines dans un groupe de trajectoires par rapport à toutes les trajectoires. Pour chaque noyau, nous avons calculé le zScore de toutes les protéines. Un zScore positif indique que la protéine est sur-représentée dans le groupe de trajectoire, un zScore négatif indique qu'elle est sous-représentée et un zScore proche de zéro indique une représentation non significative. Nous avons ensuite identifié une liste de protéines sur-représentées avec un zScore élevé. Cette liste est considérée comme la signature en protéines de chacun des noyaux.

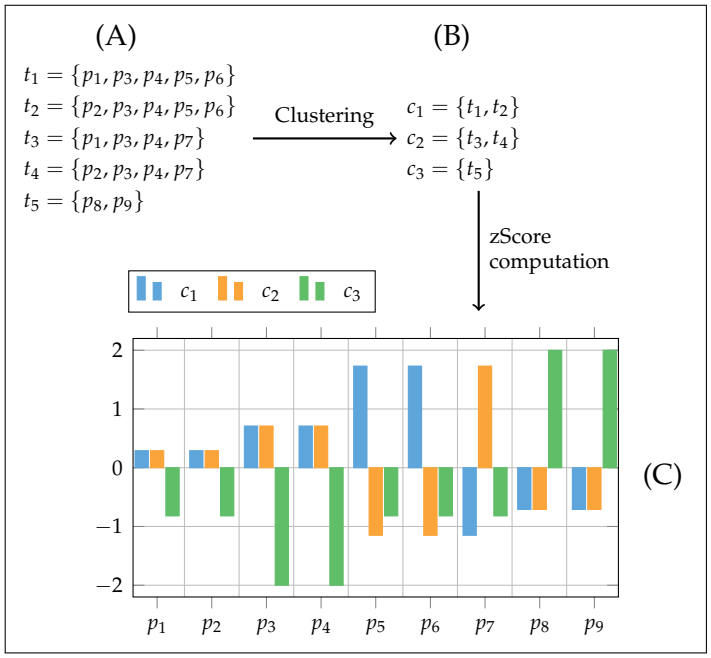


Figure 22 – ***Exemple de calcul pour déterminer les protéines surreprésentées entre trois noyaux de trajectoires***

(A) t_1, t_2, t_3, t_4 et t_5 sont cinq trajectoires contenant des protéines p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8 et p_9. (B) la méthode de cluster identifie trois noyaux c_1, c_2 et c_3. (C) répartition du niveau de représentation des protéines dans les noyaux c_1, c_2 et c_3. Par exemple, p_1 et p_2 sont légèrement surreprésentées dans les noyaux c_1 et c_2 mais pas surreprésentées en c_3, contrairement à p_9. Le noyau c_3 peut être caractérisé par p_8 et p_9.

Les noyaux sont caractérisés par des signatures de protéines sur-représentées, spécifiques de processus biologiques

Comme le montre la figure 23, la distribution du zScore des 321 protéines des trajectoires de chacun des 5 noyaux est très hétérogène. La liste des zScores de protéines pour chaque noyau a été fournie en annexe 15. Il est intéressant de noter que la distribution des valeurs de zScore du noyau 1 est inversement corrélée à celle du noyau 2 suggérant différentes fonctions biologiques associées à ces trajectoires. Ensemble, ces observations suggèrent que chaque noyau de trajectoires est caractérisé par des signatures spécifiques de protéines.

Nous savons que la probabilité de trouver au hasard une protéine dans un groupe de trajectoires avec un zScore supérieur à 4.0 est inférieure à 0.006%. En conséquence, nous avons décidé de sélectionner les protéines avec un zScore supérieur à 4.0 comme étant la signature de chacun des cinq noyaux de trajectoires.

Le nombre de protéines sur-représentées varie entre 40 et 66 (table 8), et les noyaux partagent très peu de protéines sur-représentées (en moyenne 7.7 %) (table 9). Les

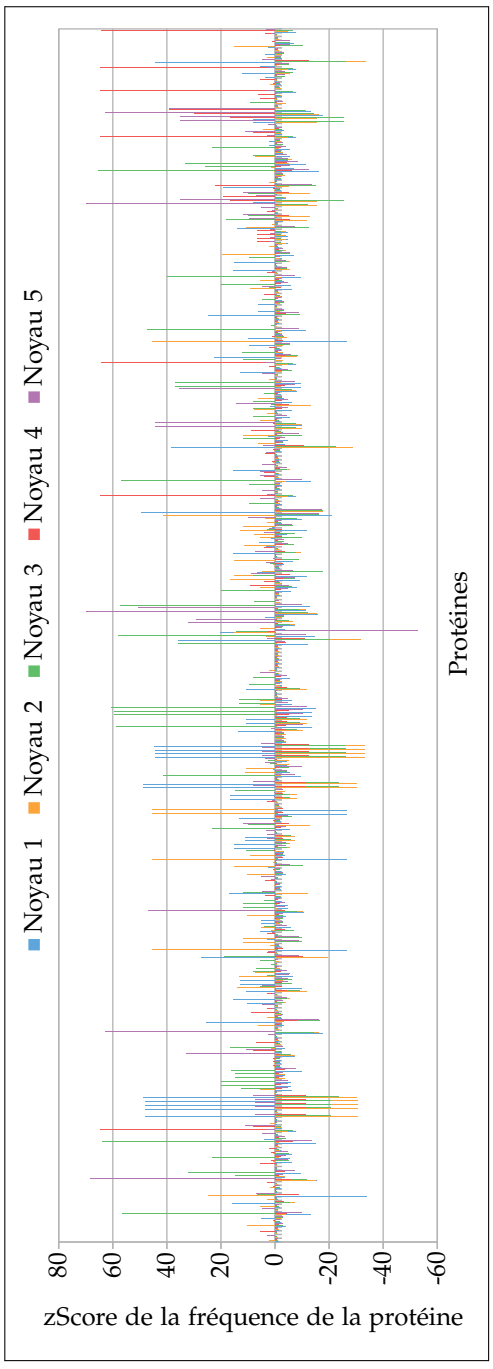


Figure 23 – ***Répartition des valeurs de zScore des fréquences des protéines des trajectoires de chaque noyaux***

Cette figure représente le zScore des 321 protéines des trajectoires de chaque noyau (intersection des cinq groupes de clusters). Pour chaque protéine, il y a donc cinq valeurs associées. La distribution du zScore des protéines est très hétérogène. Ces observations ont suggéré que chaque noyau de trajectoires était caractérisé par des signatures spécifiques de protéines.

Noyau 1		Noyau 2		Noyau 3		Noyau 4		Noyau 5	
Protéine	**zScore**	**Protéine**	**zScore**	**Protéine**	**zScore**	**Protéine**	**zScore**	**Protéine**	**zScore**
MAPK8	49.36	EP300	45.18	SRF	65.21	MDM2	64.70	KPNA2	69.80
CD4	48.76	GATA1	45.18	CALM2	63.65	TP53	64.70	SMAD3	69.61
HLA-DRA	48.76	HES1	45.18	IL2RG	60.35	TRIM28	64.70	AXIN1	68.43
HLA-DRB1	48.76	HEY1	45.18	IL2RA	59.35	CAV1	64.48	CTGF	62.53
CD247	47.84	PRKCA	45.18	IL2RB	59.35	TFDP1	64.48	TGFBR3	62.53
CD3D	47.84	MAPK14	41.33	IL2	58.63	PML	64.06	KPNB1	50.45
CD3E	47.84	ATF2	24.56	JAK3	58.02	ZFYVE9	64.06	FKBP1A	46.67
CD3G	47.84	RIPK1	19.46	LCK	57.08	TGFBRAP1	38.92	NUP214	44.34
IL12RB2	44.67	MAP2K3	16.62	MTOR	56.82	CTGF	29.67	NUP153	44.09
IL12A	44.30	MAP3K10	14.89	AKT1	56.48	TGFBR3	29.67	TGFBRAP1	39.09
IL12B	44.30	FZD2	14.83	PRKCZ	47.02	SP1	22.06	PIAS3	35.33
IL12RB1	44.30	WNT5A	14.83	HRAS	41.10	SMAD7	18.95	SMAD4	35.04
TYK2	44.21	JUN	14.82	RAF1	39.79	SMAD4	16.60	TGFBR1	35.04
NOS2	38.44	MAP2K4	14.82	PIK3CA	36.79	TGFBR1	16.60	TGFBR2	35.04
JAK2	35.62	EFNB1	13.84	PIK3R1	36.79	TGFBR2	16.60	CITED1	32.69
ELK1	27.35	EGR1	13.10	JAK1	35.82	MAP2K1	9.26	KAT2A	31.89
DAB2	25.23	MAP3K7	12.67	STAT3	33.21	DCC	8.63	KAT2B	28.93
PTEN	24.54	MAP3K8	11.83	BAD	32.07	NTN1	8.63	PDGFRA	14.10
PPAP2A	22.26	ERBB2	11.66	STAT1	25.77	CBFB	8.03	SHC1	11.59
JUN	20.37	ERBB3	11.66	BLNK	23.16	TFE3	8.03	SOS1	11.59
SP1	19.01	NRG1	11.57	SYK	23.16	CREB1	7.83	GRB2	11.58
FOS	16.96	MAP3K3	11.13	GRAP2	23.02	CSF2	6.95	CBFB	11.04
HIF1A	16.49	HSF2	11.04	CD79A	19.97	SDC1	6.31	TFE3	11.04
HIF3A	16.49	HSP90AA1	10.71	CD79B	19.97	SDC2	6.31	CREB1	10.74
ARNT	15.74	SGMS1	10.60	LYN	19.97	SDC3	6.31	MAPK12	9.73
MAP3K12	15.53	AGTR1	10.17	PTPRC	19.97	SDC4	6.31	SHC1	9.56
DOK1	15.40	FGF2	10.17	ELK1	18.65	TNFSF10	6.13	MAP2K4	8.78
NCK1	15.40	FOXO1	10.17	SHC1	17.88	ADAM17	5.27	CD4	7.80
RASA1	15.40	GATA2	9.07	CREBBP	16.46	BAG4	5.27	HLA-DRA	7.80
GDNF	15.08	PTPN13	9.01	CEBPB	15.97	TNFRSF1A	5.27	HLA-DRB1	7.80
GFRA1	15.08	MAP3K6	7.68	B2M	14.53	TRADD	5.27	CD247	7.22
RET	15.08	PDGFB	7.50	CD8A	14.53	FOXH1	5.09	CD3D	7.22
SGMS1	13.86	PDGFRB	7.50	CD8B	14.53	SMAD2	5.09	CD3E	7.22
IL1B	13.70	EGR2	7.37	HLA-A	14.53	MAP2K2	4.07	CD3G	7.22
GSN	12.99	IRF4	7.37	JUN	14.23	PTP4A3	4.07	MAP3K12	7.19
PLCG1	12.87	PARP14	7.37	IL4	13.11	IRF7	4.01	ATF2	6.96
EGF	12.64	STAT6	7.37	IL4R	13.11	MAP3K14	4.01	SMAD7	6.92
EGFR	12.64	NR3C1	6.21	CD40LG	12.60	MAX	4.01	IRF7	5.51
TRAF6	12.09	CXCR3	6.04	PPARG	12.07	MYC	4.01	MAX	5.51
GNB1	10.76	PF4	6.04	FOS	11.69	MYOD1	4.01	MYC	5.51
GNG2	10.76	ARHGDIA	5.56	POU2F1	11.65			MYOD1	5.51
EBI3	10.49	PAK1	5.56	FKBP4	11.51			IL12RB2	4.99
IL27	10.49	MAP2K1	5.40	FKBP5	11.51			IL12A	4.74
IL27RA	10.49	MAP3K5	5.33	NR3C1	11.51			IL12B	4.74
IL6ST	10.46	CD40LG	5.32	GATA3	10.70			IL12RB1	4.74
DOCK7	10.27	RAC1	5.26	SHC1	9.94			TYK2	4.68
PRKCB	9.95	IL4	5.21	SOS1	9.94			PTPN13	4.55
PRKACA	9.43	IL4R	5.21	GRB2	9.84			AR	4.50
SMAD4	7.81	JUNB	5.21	ILK	9.57			CARM1	4.50
TGFBR1	7.81	ETS1	4.84	MAPKAP1	9.57			DNAJA1	4.50
TGFBR2	7.81	MAP2K6	4.44	MLST8	9.57			FOXH1	4.50
MAP2K6	6.42	TGFB2	4.10	RICTOR	9.57			MEF2C	4.50
PTGDR	6.30			TNF	9.12			NCOA2	4.50
PTGIR	5.98			EGR2	8.11			PKN1	4.50
MAP3K4	5.80			IRF4	8.11			SMAD2	4.50
ESR1	5.25			PARP14	8.11			EFNB1	4.50
TXN	5.24			STAT6	8.11			FOS	4.47
AKAP1	5.13			PDGFB	8.09			NOS2	4.15
FAS	4.93			PDGFRB	8.09				
FASLG	4.93			MAP2K4	8.07				
CAMK2B	4.01			LIMS1	7.66				
				EGR4	6.91				
				ATF2	6.43				
				ELF1	5.41				
				EFNB1	5.34				

Table 8 – *Liste des protéines sur-représentées et leur zScore dans chacun des noyaux.*

	Noyau 1	Noyau 2	Noyau 3	Noyau 4	Noyau 5
Noyau 1	100.00 %	2.73 %	2.44 %	4.12 %	18.00 %
Noyau 2	2.73 %	100.00 %	13.60 %	1.10 %	3.80 %
Noyau 3	2.44 %	13.60 %	100.00 %	0.00 %	6.09 %
Noyau 4	4.12 %	1.10 %	0.00 %	100.00 %	19.75 %
Noyau 5	18.00 %	3.80 %	6.09 %	19.75 %	100.00 %

Table 9 – *Pourcentages des protéines sur-représentées en pourcentage communes entre les cinq noyaux de trajectoires.*

noyaux 2 et 3 ont des signatures similaires à 13.6% et le noyau 5 paraît être le moins spécifique puisqu'il a une signature similaire à 18% avec le noyau 1 et à 19.75% avec le noyau 4.

2.3.4 Caractérisation fonctionnelle des regroupements de trajectoires

À partir des scores de sur-représentation des protéines de chaque noyau, nous avons ensuite cherché la signification biologique qui caractérise les noyaux.

Le principal problème est que les 321 protéines sont déjà en rapport avec le TGF-β et donc les processus biologiques sont déjà très spécifiques, il faut donc trouver des fonctions encore plus précises pour chacun des noyaux. Les analyses ont été effectuées à l'aide de l'outil GSEA (*Gene Set Enrichment Analysis*) développé par le Broad Institute [Aravind Subramanian et al., 2005]. C'est une méthode qui permet d'identifier des classes de gènes significativement enrichies par rapport à un grand ensemble de gènes ou de protéines associées à des fonctions biologiques spécifiques.

Les listes de protéines et leur fréquence respective de zScore ont été utilisées comme *input* et les processus biologiques provenant de la base de données de *Gene Ontology* ont été sélectionnés comme base de références. Les résultats sont des ensembles de termes GO de « processus biologiques » significativement enrichis pour chaque noyau par rapport aux autres noyaux. La liste complète est disponible en annexe table 17.

Afin d'identifier les différentes familles de processus biologiques respectifs, nous avons utilisé Revigo (*Reduce and Visualize Gene Ontology*) [Fran Supek et al., 2011] qui réduit la liste des termes GO en fonction de leur similarité sémantique et propose un score de singularité pour chaque terme GO. Ce score indique si le terme GO est singulier par rapport à la liste complète des termes GO. Calculé comme 1 moins la similitude sémantique moyenne d'un terme par rapport à tous les autres termes. Plus la singularité est élevée, plus le terme GO tend à être indispensable.

Comme le montre la figure 24, chaque noyau a été caractérisé par un ensemble spécifique de fonctions biologiques puisque 25 (57 %), 54 (90 %), 84 (80 %), 46 (81 %) et 54 (88 %) de termes GO étaient spécifiques au noyau 1, noyau 2, noyau 3, noyau 4 et noyau 5 respectivement. Par conséquent, les trajectoires du noyau 1 sont associées

à la signalisation induite par les récepteurs antigéniques alors que les trajectoires du noyau 2 sont principalement associées à l'activité des kinases serine/threonine (figure 24). Les annotations fonctionnelles des noyaux 3 et 4 sont plus hétérogènes alors que les trajectoires du noyau 5 sont clairement impliquées dans la réponse immunitaire.

Une conclusion importante à partir de ces résultats est que, même si les trajectoires de signalisation partagent de nombreuses protéines, notre analyse a révélé des groupes de trajectoires qui correspondent à différentes familles fonctionnelles.

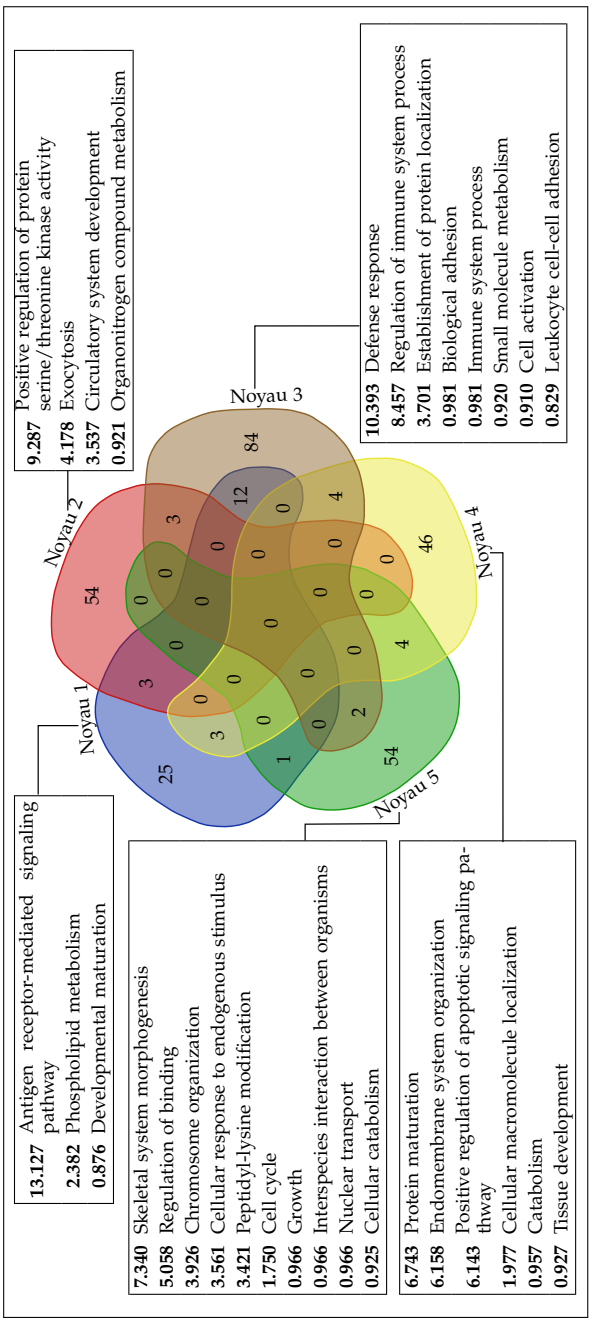


Figure 24 – ***Analyse des processus biologiques de chaque noyau.***

Les listes de protéines et leur fréquence de zScore respective pour chaque noyau ont été analysées par GSEA. En découle une liste de termes GO enrichis associés aux processus biologiques et une p-value associée. Ces termes GO pour chaque noyau sont comparés à l'aide du diagramme de Venn. Le résumé des processus biologiques a été effectué par l'outil Revigo qui fournit une liste de processus et un score de singularité associé.

2.3.5 Visualisation Web des voies de signalisation influencées par le TGF-β

L'analyse des fonctions biologiques associées à chaque noyau de trajectoires est une analyse globale. Pour faciliter une exploration locale des trajectoires de signalisation regroupées dans chaque noyau, nous avons développé une interface Web :
`http://www.irisa.fr/dyliss/public/tgfbVisualization`

L'interface est basée sur la bibliothèque JavaScript Cytoscape [1]. Les grands nœuds sont des protéines et leur taille est corrélée au nombre de trajectoires impliquant cette protéine. La couleur du nœud indique l'apparition de la protéine dans les trajectoires du noyau par rapport à son apparition dans les 6017 trajectoires. Cette nuance de couleur est donc basée sur le zScore de la fréquence des protéines (bleu pour zScore < 0 et rouge pour zScore > 0) et le curseur de zScore permet de filtrer l'information. Les petits nœuds noirs illustrent les réactions biologiques (association, dissociation, phosphorylation, dégradation, migration, etc.) comme décrites dans [Andrieux et al., 2014]. Les arêtes noires relient les protéines d'entrée et de sortie d'une réaction, les arêtes vertes relient les protéines qui régulent positivement la réaction et les arêtes rouges relient les protéines qui régulent négativement la réaction (figure 25). L'exploration des réseaux est facilitée par le repositionnement manuel des nœuds dans le graphe. Le graphique peut être exporté au format JSON.

2.4 Regroupement des gènes influencés par le TGF-β

Dans cette partie, je m'intéresserai à classifier les gènes en fonction des trajectoires qui les influencent. Les questions que je me suis posées sont :

— Y a-t-il des groupes de gènes influencés par les mêmes trajectoires et ces groupes sont-ils biologiquement significatifs ?

— Peut-on déduire les fonctions biologiques des trajectoires par rapport aux gènes qu'elles influencent ?

Dans le cas du TGF-β, un gène est influencé en moyenne par 70.18 trajectoires et comme le montre la figure 26 certains gènes sont influencés par des centaines de trajectoires. Les trois gènes influencés par le plus de trajectoires sont SMAD7 (1617 trajectoires), TGF-β1 (1362 trajectoires) et JUN (790 trajectoires).

2.4.1 Analyse topologique du graphe de gènes

Nous savons que les voies de signalisations influencent (activent ou inhibent) l'expression de gènes, dans notre cas cela se traduit par la relation binaire I entre les 6017 trajectoires t_k et les 144 gènes g_k. En réalisant un graphe où les nœuds représentent les gènes influencés par le TGF-β et en prenant comme hypothèse que deux gènes sont

1. `http://js.cytoscape.org`

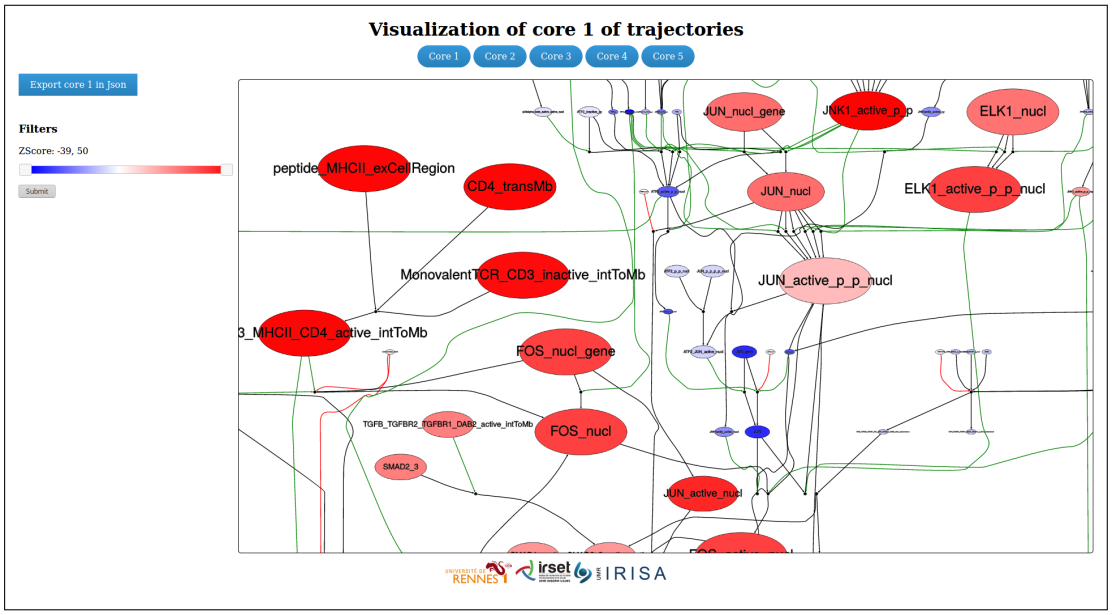


Figure 25 – ***Capture d'écran de la visualisation Web du noyau 1.***

Un nœud est une protéine, la taille du nœud correspond au nombre de trajectoires impliquant cette protéine et la couleur du nœud correspond à la représentation de la protéine dans le noyau par rapport à toutes les trajectoires (bleu la protéine est sous-représentée et rouge elle est surreprésentée).

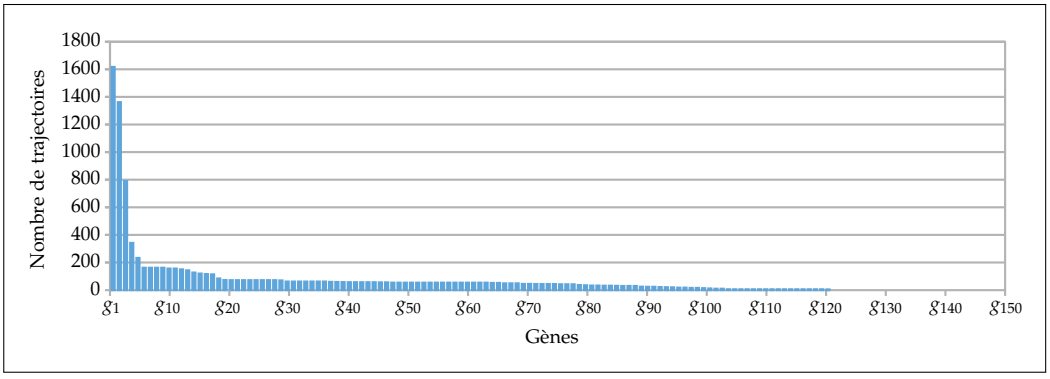


Figure 26 – ***Statistiques de la répartition du nombre de trajectoires en fonction des gènes qu'elles influencent***

Répartition du nombre de trajectoires influençant chaque gène. Ces résultats ont montré que la plupart des gènes sont influencés par de nombreuses trajectoires communes.

liés s'ils sont au moins influencés par une même trajectoire, on peut faire apparaître 19 cliques maximales de gènes partageant les mêmes trajectoires (figure 27). Dans cette figure seulement les 95 gènes partageant des trajectoires ont été représentés, les 49 gènes ne partageant aucune trajectoire avec d'autres gènes n'ont pas été affichés. Les gènes sont donc influencés par de mêmes voies de signalisation et forment différents groupes spécifiques.

De plus, 11 gènes se trouvent à la fois dans plusieurs cliques différentes, ils sont donc influencés par plusieurs groupes de trajectoires (table 10) et peuvent potentiellement représenter des gènes co-exprimés dans des processus biologiques impliquant le TGF-β.

Gène	**Cliques**	**Nom**
SOCS3	n° 10 et n° 13	*Granzyme B*
IRF1	n° 5, n° 8 et n° 11	*Interferon Gamma*
PRF1	n° 2, n° 8, n° 9 et n° 11	*Perforin 1*
BCL2L1	n° 2, n° 4, n° 5, n° 11 et n° 19	*Chemokine Ligand 8*
CCNA2	n° 2 et n° 3	*Cyclin A2*
MMP2	n° 1 et n° 3	*Interferon Regulatory Factor 1*
GZMB	n° 7 et n° 9	*Fas Ligand*
IL5	n° 6 et n° 17	*Bcl2-Like 1*
FASLG	n° 6 et n° 7	*Suppressor Of Cytokine Signaling 3*
IFNG	n° 1 et n° 4	*Matrix Metallopeptidase 2*
CXCL8	n° 1 et n° 6	*Interleukin 5*

Table 10 – *Liste des gènes présents dans plusieurs cliques.*

Afin de trouver les fonctions biologiques spécifiques à chaque clique de gènes, il est nécessaire de comparer les annotations des gènes au sein de la clique par rapport à l'ensemble de 144 gènes. L'extraction des termes GO et le calcul de sur-représentation ont été réalisés par l'outil Panther [Mi et al., 2013], qui à partir d'une liste de gènes et une liste de référence retourne une *p-value* de sur-représentation pour chaque terme GO associé aux gènes. Malheureusement sur les 18 cliques, il n'a pas été possible d'identifier une ou plusieurs fonctions biologiques avec une *p-value* inférieure à 0.05 et donc pertinentes. L'absence de différences significatives entre les cliques peut être liée à la proximité des gènes qui sont tous influencés par le TGF-β.

Cependant le calcul de similarité sémantique $zSim_{SPBHM}$ a permis de mettre en évidence certaines cliques de gènes significativement similaires par rapport aux 144 gènes, comme la clique n° 6 et la clique n° 18, mais cela reste peu informatif.

Cette approche reste limitée pour plusieurs raisons. Premièrement, il est impossible de savoir si au sein d'une clique certains gènes partagent des trajectoires spécifiques. Par exemple au sein d'une clique de 10 gènes influencés par 100 trajectoires en commun, 5 gènes pourraient être influencés par 130 trajectoires communes. L'étude topologique de ce type de graphe ne permet pas de proposer différents niveaux de précision.

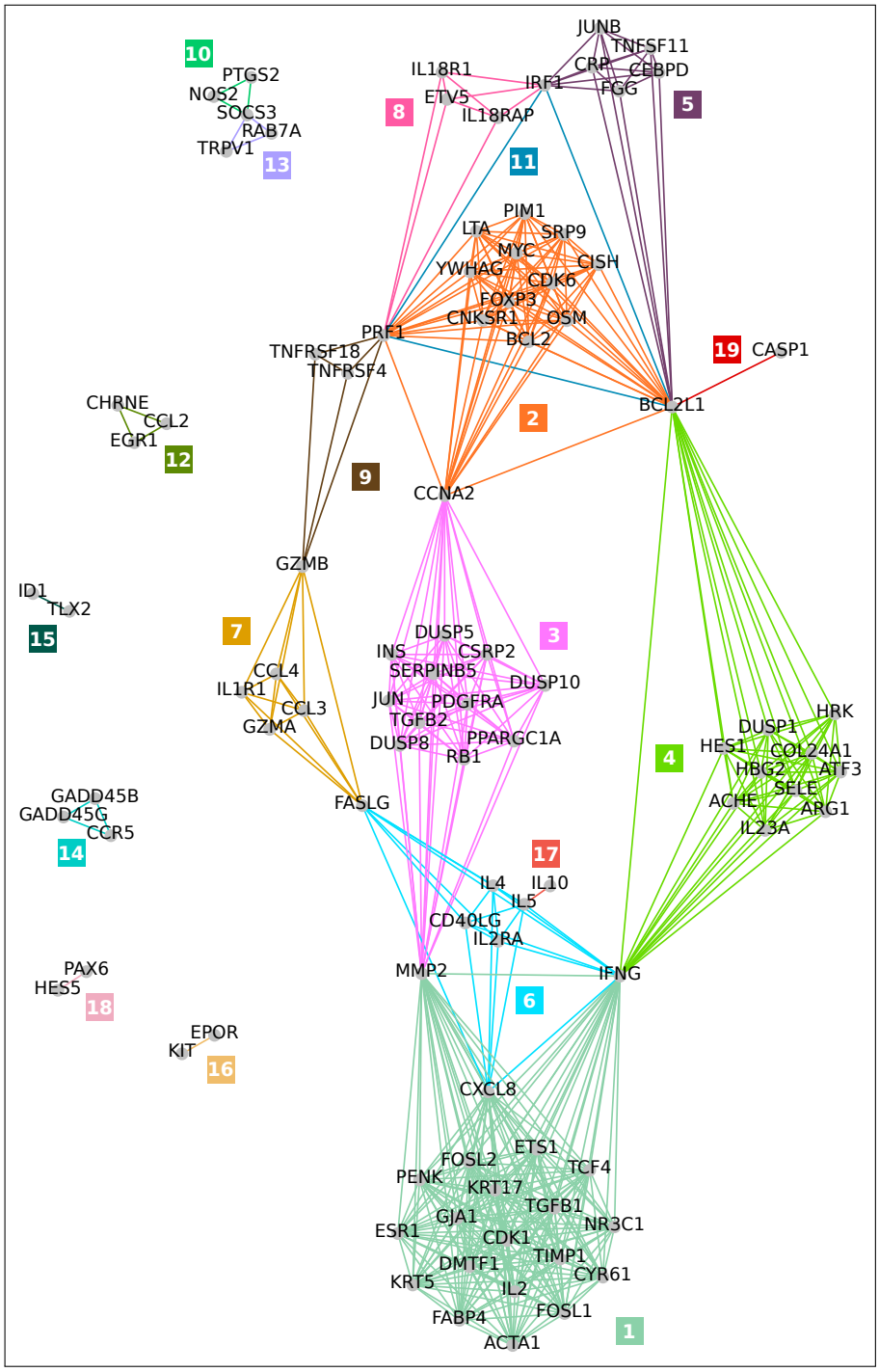


Figure 27 – *Représentation des gènes au moins influencés par une même trajectoire.*

Dans cette figure, un nœud représente un gène influencé par le TGF-β et une arête indique que les deux gènes liés par cette arête sont au moins influencés par au moins une même trajectoire. Chaque couleur correspond à une clique maximale de gènes, et est numérotée de 1 à 19. Les gènes ne partageant aucune trajectoire avec d'autres gènes n'ont pas été affichés.

	Taille	Clique de gènes	$zSim_{SPBHM}$
1	21	IFNG, CXCL8, KRT17, TIMP1, FOSL1, TCF4, IL2, FABP4, PENK, NR3C1, CYR61, MMP2, ESR1, FOSL2, KRT5, ETS1, CDK1, DMTF1, GJA1, ACTA1, TGFB1	-0.030
2	14	MYC, OSM, YWHAG, BCL2, FOXP3, SRP9, PRF1, CCNA2, PIM1, CNKSR1, BCL2L1, CISH, CDK6, LTA	-0.851
3	13	RB1, JUN, DUSP5, TGFB2, DUSP10, CCNA2, CSRP2, MMP2, PDGFRA, PPARGC1A, SERPINB5, DUSP8, INS	-0.349
4	12	IFNG, HBG2, ARG1, ATF3, HES1, DUSP1, SELE, BCL2L1, HRK, ACHE, COL24A1, IL23A	-0.932
5	7	IRF1, BCL2L1, CEBPD, TNFSF11, CRP, FGG, JUNB	0.577
6	7	IFNG, CXCL8, CD40LG, IL4, FASLG, IL2RA, IL5	3.229
7	6	GZMB, CCL3, CCL4, FASLG, GZMA, IL1R1	0.919
8	5	IRF1, ETV5, IL18R1, PRF1, IL18RAP	0.779
9	4	GZMB, PRF1, TNFRSF18, TNFRSF4	0.052
10	3	PTGS2, SOCS3, NOS2	0.964
11	3	IRF1, BCL2L1, PRF1	-0.347
12	3	EGR1, CHRNE, CCL2	0.177
13	3	RAB7A, TRPV1, SOCS3	-0.250
14	3	CCR5, GADD45B, GADD45G	-0.212
15	2	TLX2, ID1	1.179
16	2	KIT, EPOR	1.625
17	2	IL10, IL5	1.104
18	2	HES5, PAX6	2.340
19	2	CASP1, BCL2L1	1.059

Table 11 – *Liste des cliques de gènes influencés par au moins une même trajectoire.*

Deuxièmement, même au sein d'une clique, il est tout à fait possible que l'ensemble des gènes de la clique ne soit pas tous influencés par les mêmes trajectoires. Dans l'exemple de la figure 28, la trajectoire t_1 influence les gènes g_1 et g_2, la trajectoire t_4 influence les gènes g_2 et g_3, et la trajectoire t_2 influence les gènes g_1 et g_3. Étant donné que chaque gène partage une trajectoire en commun avec les autres gènes alors les trois gènes forment une clique. Pourtant aucune trajectoire n'influence les trois gènes à la fois.

Pour résoudre ces deux limitations, j'ai choisi d'analyser les mêmes données mais à partir de concepts formels.

2.4.2 Analyse des concepts formels des gènes et des trajectoires

Pour optimiser la recherche de groupes de gènes influencés par les mêmes trajectoires, nous avons fait une analyse de concepts formels. Pour cela j'ai défini le contexte

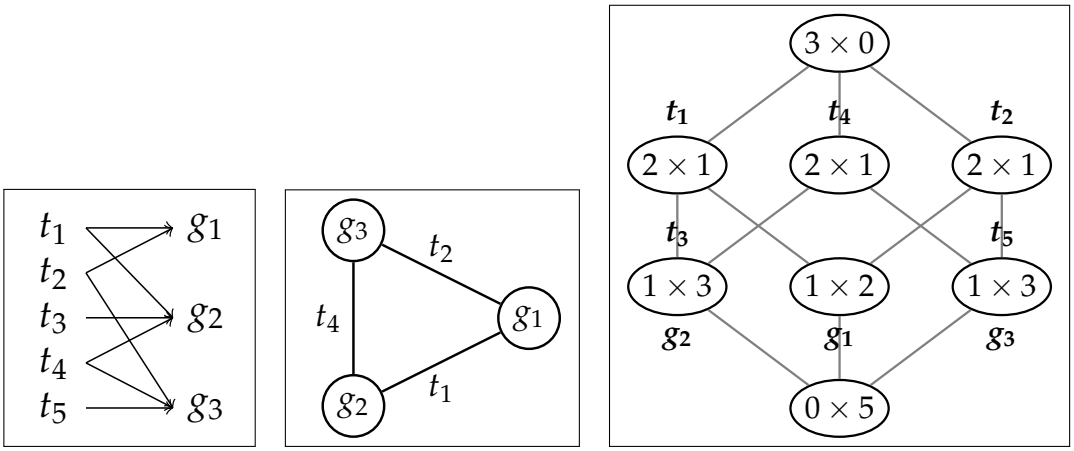


Figure 28 – ***Exemple d'analyse des groupements de gènes influencés par des trajectoires***

(a) Exemple de données, les trajectoires t_1, t_2, t_3, t_4 et t_5 influencent les gènes g_1, g_2 et g_3 ; (b) Le graphe produit à partir de ces données des gènes influencés par au moins une même trajectoire. On voit que les trois gènes forment une clique mais qu'aucune trajectoire n'influence les trois gènes à la fois ; (c) Treillis de concepts formels des gènes et des trajectoires, permettant de regrouper les gènes à différents niveaux de précision. Chaque nœud représente un concept et le label de ce nœud est sous le format « Nombre de gènes » × « Nombre de trajectoires ».

	t_1	t_2	t_3	t_4	t_5
g_1	1	0	0	0	0
g_2	1	1	1	1	0
g_3	0	1	0	1	1

Table 12 – *Exemple de matrice binaire représentant l'influence des trajectoires sur les gènes. Si une trajectoire t_j influence un gène g_i alors la valeur de la cellule vaut "1" autrement elle vaut "0".*

formel $K = (G, S, I)$ où G est un l'ensemble des gènes, S est l'ensemble des trajectoires et I est la relation binaire $I \subseteq G \times S$ spécifiant quels gènes sont influencés par quelles trajectoires (table 12).

Dans l'exemple de la figure 28c, le jeu de données a produit 8 concepts qui sont donc des pairs de gènes et de trajectoires représentés en treillis. Dans cette figure, un concept est noté sous le format « Nombre de gènes » × « Nombre de trajectoires ». On peut voir la finesse du treillis par rapport au graphe précédent. On retrouve la clique des trois gènes dans le concept le plus haut, et on peut remarquer qu'aucune trajectoire n'est commune aux trois gènes. Par contre, le treillis montre qu'il existe des sous-groupes de gènes qui partagent des trajectoires communes, et que certains gènes ont même des trajectoires spécifiques à eux-mêmes.

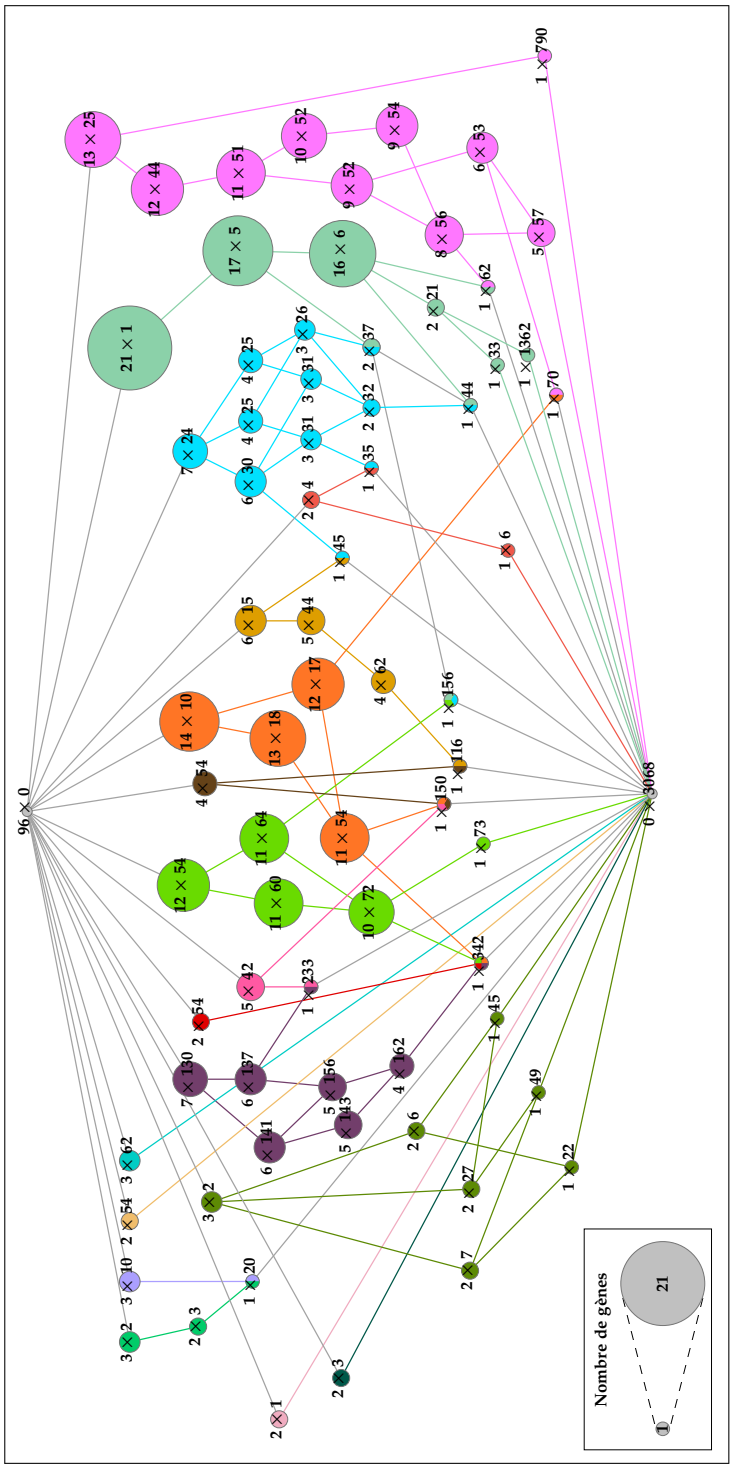


Figure 29 – *Treillis de concepts formels des gènes et des trajectoires (en fonction des cliques).*

Les nœuds sont des concepts formels, deux nœuds sont liés si l'un est inclus dans l'autre. Les concepts les plus hauts possèdent plus de gènes et moins de trajectoires communes et inversement. La taille des nœuds correspond aux nombres de gènes que possède le concept. Pour une raison de lisibilité, la taille du concept le plus haut, celui qui contient tous les gènes, n'est pas à l'échelle. Enfin la couleur des nœuds correspond aux cliques de la figure 27

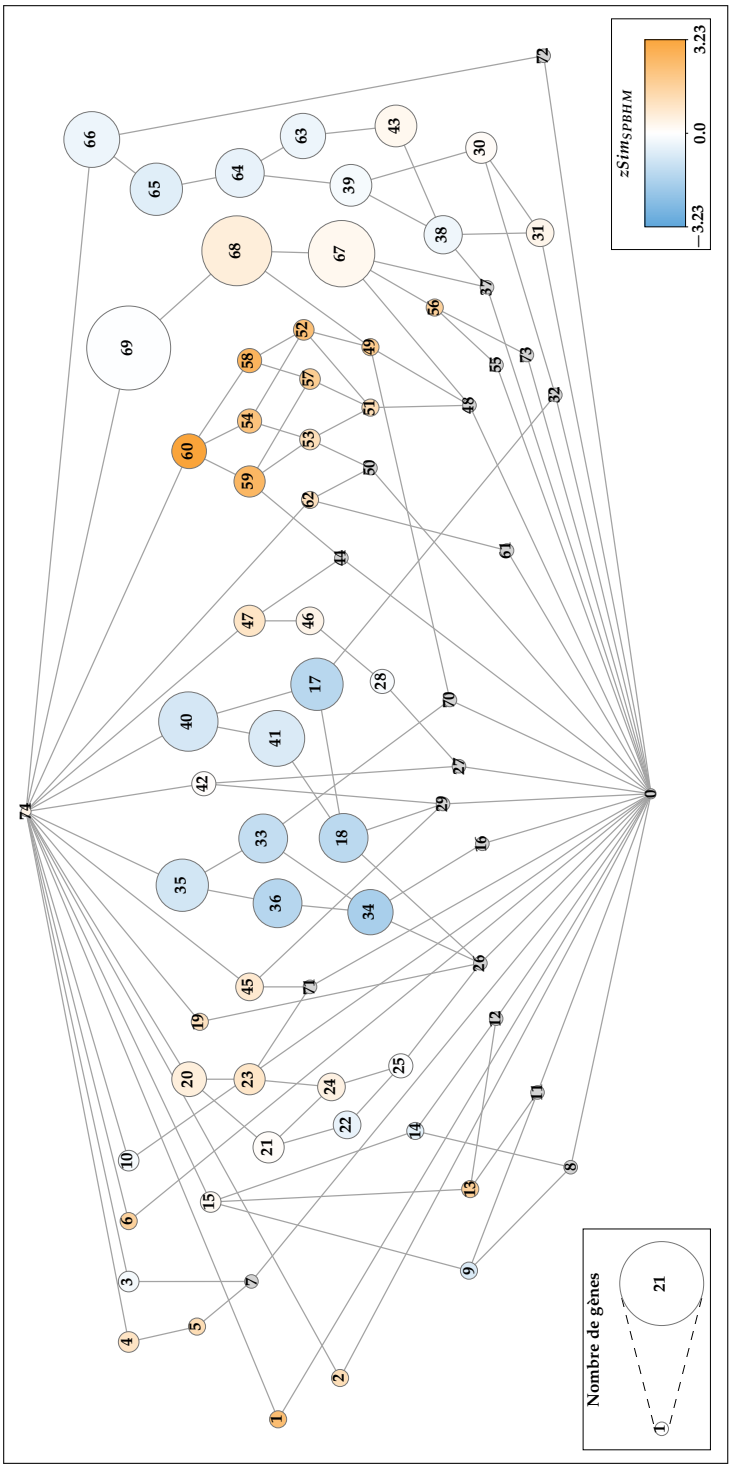


Figure 30 – *Treillis de concepts formels des gènes et des trajectoires (en fonction de SPBHM).*

Les nœuds sont des concepts formels, deux nœuds sont liés si l'un est inclus dans l'autre. Les concepts les plus hauts possèdent plus de gènes et moins de trajectoires communes et inversement. La taille des nœuds correspond aux nombres de gènes que possède le concept. Pour une raison de lisibilité, la taille du concept le plus haut, celui qui contient tous les gènes, n'est pas à l'échelle. Enfin la couleur des nœuds correspond au score de similarité $zSim_{SPBHM}$.

À partir des données du TGF-β, nous avons trouvé 74 concepts. La liste complète des concepts est disponible en annexe 19. Ces concepts sont représentés en treillis dans la figure 29. J'ai reporté les couleurs de chaque clique dans ce treillis. Les gènes interfaces possèdent plusieurs couleurs. Tout d'abord nous pouvons voir que toutes les cliques sont présentes (mise à part la clique n° 11 qui regroupait des gènes interfaces, donc déjà présents dans d'autres cliques), et qu'au moins une trajectoire influence tous les gènes de chaque clique. Le problème soulevé dans l'exemple de la figure 28 n'existe donc pas dans notre jeu de données. De plus, cette analyse révèle qu'au sein de chaque clique, plusieurs sous-groupes de gènes sont influencés par des trajectoires spécifiques. Ce qui permet de dire qu'au sein d'une clique certains gènes sont plus proches (ils sont influencés par plus de trajectoires) que d'autres.

La figure 30 présente le même treillis mais cette fois-ci les couleurs des concepts correspondent au score $zSim_{SPBHM}$ indiquant la similarité sémantique des gènes des concepts dans l'ontologie « *biological processes* » de *Gene Ontology*. Cette similarité a été calculée pour les concepts de taille supérieure ou égale à 2. On peut remarquer que les 17 gènes du concept n° 68, qui est un sous-concept du concept n° 69 correspondant à la clique numéro n° 1, sont plus similaires que les 21 gènes de la clique n° 1. Cela indique que les 4 gènes ne sont pas compris dans le concept n° 68 baissent la similarité moyenne du groupe. Cependant la grande majorité des concepts (environ 88 %) de taille supérieure ou égale à 2 ont un score compris entre -1 et 1 et ne présentent donc pas de similarité ou disimilarité significative.

2.5 Discussion

Alors que les approches qualitatives sont adaptées aux réseaux à grande échelle, l'analyse de nombreuses trajectoires de signalisation reste un réel problème. Les méthodes de réduction se concentrent sur la diminution de la taille des réseaux booléens à grande échelle [Jorge G. T. Zañudo and Réka Albert, 2013, Assieh Saadatpour et al., 2013] ou les méthodes de division dans plusieurs sous-réseaux [Yin Zhao et al., 2013]. Cependant, ces méthodes consistent généralement à effectuer une réduction avant l'analyse, alors que pour le TGF-β nous nous sommes concentrés sur une analyse exhaustive du réseau de signalisation.

En plus de l'exhaustivité, l'originalité de notre approche consiste à analyser les trajectoires de signalisation en fonction de leur composition protéique. Notre approche a été motivée par le fait que les voies de signalisation partagent un grand nombre de « domaines modulaires » dans différentes combinaisons [Wendell A. Lim, 2010]. Ces combinaisons supportent la diversité fonctionnelle des voies de signalisation.

Ces domaines modulaires fournissent une structure sous-jacente des trajectoires de signalisation. Notre objectif était d'identifier des groupes de trajectoires similaires. En considérant deux trajectoires, plus elles partagent de modules de protéines, plus elles sont similaires. Il existe de nombreuses méthodes de clustering (par exemple hiérarchique, K-means, basée sur la distribution, basée sur la densité) [Aastha Joshi and Rajneet Kaur, 2013]. Comme nous l'avons mentionné précédemment, un domaine

modulaire peut être impliqué dans plusieurs combinaisons, de sorte que leur étude nécessite des méthodes de *soft-clustering* qui permettent aux clusters de se chevaucher et de partager des éléments. Nous avons sélectionné la méthode de *clustering* des plus proches voisins (SNN), qui ont été appliqués avec succès pour gérer l'hétérogénéité de données et à grande échelle [Hamzaoui et al., 2011]. Le modèle RSC de Houle est également approprié dans le sens où il n'est pas nécessaire de définir la taille et le nombre de *clusters* attendus.

Le modèle RSC s'est révélé être une méthode de *clustering* robuste pour notre jeu de données. Toutes les 64 combinaisons de valeurs de paramètres ont généré des *clusters* qui appartenaient systématiquement au groupe 1 et au groupe 2 et à l'un des groupes 3, 4 et 5. La moitié des simulations a produit des *clusters* appartenant aux groupes 3, 4 ou 5. Dans la figure 21, l'analyse de l'influence des valeurs des paramètres pour les groupes 3, 4 et 5 a montré que x_1 et x_3 n'avaient aucune influence sur les groupes, alors que des paires de valeurs de x_2 étaient associés à différents groupes : les deux plus basses valeurs avec le groupe 5, les deux plus élevées avec le groupe 4 et une combinaison de la plus élevée et la plus basse avec le groupe 3. De manière surprenante, les deux valeurs intermédiaires de x_2 (2000 et 3000) étaient des marqueurs des groupes 4 et 5, pour lesquels ils étaient associés à leur valeur extrême la plus proche, alors que les valeurs les plus basses et les plus élevées de x_2 étaient associées au groupe 3. Ceci indique que RSC a produit soit des groupes 3 et 5 pour les valeurs basses de la taille maximale des *clusters* (x_2), ou des groupes 3 et 4 pour les valeurs élevées. À ce stade, une analyse plus approfondie serait nécessaire pour déterminer quelles sont les valeurs basses ou élevées qui sont plus adaptées à notre ensemble de données, ou si les groupes 3, 4 et 5 sont tous biologiquement. Dans l'ensemble, notre étude avec les différentes combinaisons de valeurs de paramètres a montré que (1) n'étant pas déterministe, l'exécution de plusieurs réplicats avec les mêmes valeurs de paramètres est utile, (2) le modèle RSC est une méthode de cluster robuste pour notre jeu de données, (3) les groupes 1 et 2 étaient indépendants des valeurs des paramètres, alors que les groupes 3, 4 et 5 ne l'étaient pas, et (4) des valeurs basses de la taille maximale des *clusters* ont généré des *clusters* dans les groupes 3 et 5, alors que les valeurs élevées ont produit des *clusters* dans les groupes 3 et 4.

Selon la table 17, les protéines sur-représentées dans les trajectoires des noyaux 1 et 2 discriminaient respectivement les voies canoniques associées à la réponse immune et au développement (noyau 1) et les voies non canoniques impliquant toutes les autres voies de signalisation kinase-dépendante (noyau 2). Ensemble, ces deux noyaux de *clusters* ont illustré les aspects dits « Jekyll et Hyde » du TGF-β dans le cancer [Bierie and Moses, 2006].

De plus, il est possible de faire le lien entre les fonctions biologiques du TGF-β dans les tissus tumoraux résumées par [Maozhen Tian et al., 2011] et décrites dans la figure 17 avec celles trouvées dans les différentes familles. Par exemple, la fonction « *cell activation* » a été attribuée au noyau 3, la fonction « *induction of apotosis* » au noyau 4 et la fonction « *cell cycle* » au noyau 5. Cela confirme des rôles différents pour chacune des familles de trajectoires.

Concernant le regroupement des gènes en fonction des trajectoires qui les in-

fluencent, notre étude a montré qu'il était difficile de trouver des groupes de gènes avec des fonctions significatives par rapport à l'ensemble des 144 gènes. Cependant l'analyse de concepts formels a permis de structurer les différentes influences des trajectoires sur ces gènes.

Bien qu'elle ne s'appuie pas sur une connaissance *a priori*, notre approche peut dépendre du biais d'annotation. Étant donné que les connaissances biologiques sont par nature incomplètes, certains processus de signalisation bien étudiés peuvent être décrits en détail dans les bases de données, alors que certains moins étudiés seraient décrits de manière incomplète ou avec une granularité plus grossière (habituellement les deux). Cela entraînerait une fréquence plus élevée des modules bien étudiés et donnerait une impression trompeuse d'être plus important. Il convient de noter qu'il s'agit d'un biais intrinsèque aux données sur lesquelles nous nous intéressons, et non un biais de notre méthode d'analyse. Ce biais devrait être pris en compte par les experts lors de l'analyse des résultats.

Conclusion

Nous avons proposé une méthode exhaustive et sans hypothèse préalable basée sur le *clustering* pour identifier des familles de trajectoires fonctionnellement similaires dans le réseau de signalisation. Parmi les 15 934 trajectoires impliquées dans la signalisation TGF-β, notre approche a identifié cinq groupes de trajectoires en fonction de leur composition protéique. La caractérisation fonctionnelle de ces groupes a révélé que chaque groupe est impliqué dans différents rôles de signalisation du TGF-β, ce qui a confirmé que notre approche donne des résultats biologiquement pertinents. De plus pour une exploration locale, j'ai proposé une interface permettant d'analyser en détail les réactions biologiques de chaque groupe. L'approche peut être généralisée pour explorer toutes les voies biologiques à grande échelle.

Chapitre 3

Vers une analyse des trajectoires de signalisation du TGF-β dans différentes bases de données

L'étude des voies de signalisation du chapitre 2 reposait sur le réseau établi par Geoffroy Andrieux en 2014 [Andrieux et al., 2014]. La base PID [Carl F. Schaefer et al., 2009] qu'il avait utilisée n'est plus maintenue et a été intégrée dans *Pathway Commons* [Cerami et al., 2011] en 2015 et dans NDEx [Pratt et al., 2015] en 2016. De plus, [Soh et al., 2010] ont montré que les bases de données de référence étaient complémentaires. Pour autant leur intégration en un modèle unique n'est pas triviale [Fearnley et al., 2014]. Enfin ces bases continuent d'évoluer, il faut donc réitérer le processus d'analyse périodiquement.

Ce chapitre propose une méthode pour convertir n'importe quelle base de voies de signalisation représentée au format BioPAX dans le formalisme Cadbiom développé par [Andrieux et al., 2014] permettant ainsi l'exploration de ces données. Cette méthode a été appliquée sur 5 bases de données de référence existant actuellement. Elle a permis de mettre en évidence leurs différences par rapport à la signalisation du TGF-β.

3.1 Objectif

L'objectif de cette partie est de comparer les trajectoires de signalisation dépendantes du TGF-β présentées dans le chapitre précédent avec les trajectoires d'autres bases de données. Étant donné que mon but est de présenter une étude exhaustive, il est important d'analyser le contenu d'autres bases que PID. La problématique est la suivante : par rapport au TGF-β, les bases de données de signalisation proposent-elles des données semblables ou sont-elles fortement dissimilaires ?

Il existe plusieurs formats de représentation des voies biologiques permettant l'échange et l'intégration. Les principaux sont le format *Systems Biology Markup Language* (SBML) [Chaouiya et al., 2013], le format *Biological Pathway Exchange* (BioPAX) [Demir et al., 2010] ou encore le format *System Biology Graphical Notations* (SBGN) [Novère et al., 2009]. D'après [Stromback and Lambrix, 2005], BioPAX semble fournir la représentation la plus riche et la plus générale.

De plus BioPAX est proposé par une grande majorité des bases de données [Chowdhury and Sarkar, 2015], c'est pourquoi nous avons développé un convertisseur de BioPAX à Cadbiom afin de pouvoir généraliser les analyses.

3.2 Format BioPAX

Biological Pathway Exchange (BioPAX) [1] [Demir et al., 2010] est un langage standard pour représenter les voies biologiques au niveau moléculaire et cellulaire. Le développement de BioPAX a été motivé par le grand nombre de bases de données utilisant des formats hétérogènes bien qu'elles soient complémentaires [Soh et al., 2010]. Avec un langage standard, les chercheurs ne perdent plus de temps en collectant les informations provenant de différentes sources, en comprenant les langages utilisés et en les transformant en une représentation unique. BioPAX peut représenter des voies métaboliques, des voies de signalisation, des interactions moléculaires et génétiques et des réseaux de régulation de gènes.

BioPAX a été utilisé dans de nombreuses études de systèmes biologiques, comme dans la recherche de motifs moléculaires à partir de modèles BioPAX [Babur et al., 2014] ou la création automatique de modèles cinétiques [Ruebenacker et al., 2007].

À l'heure actuelle, les principales bases de données de systèmes biologiques utilisent BioPAX, telles que Reactome, BioCYC, INOH, PID-NCI, WikiPathways, PANTHER, etc.

Le langage BioPAX fournit des termes et des descriptions pour représenter de nombreux aspects des voies biologiques et de leurs annotations. Par exemple il prend en compte les gènes, les petites molécules, les complexes et leurs états (emplacement cellulaire, modification de la protéine post-traductionnelle, variants d'épissage

1. `http://www.biopax.org`

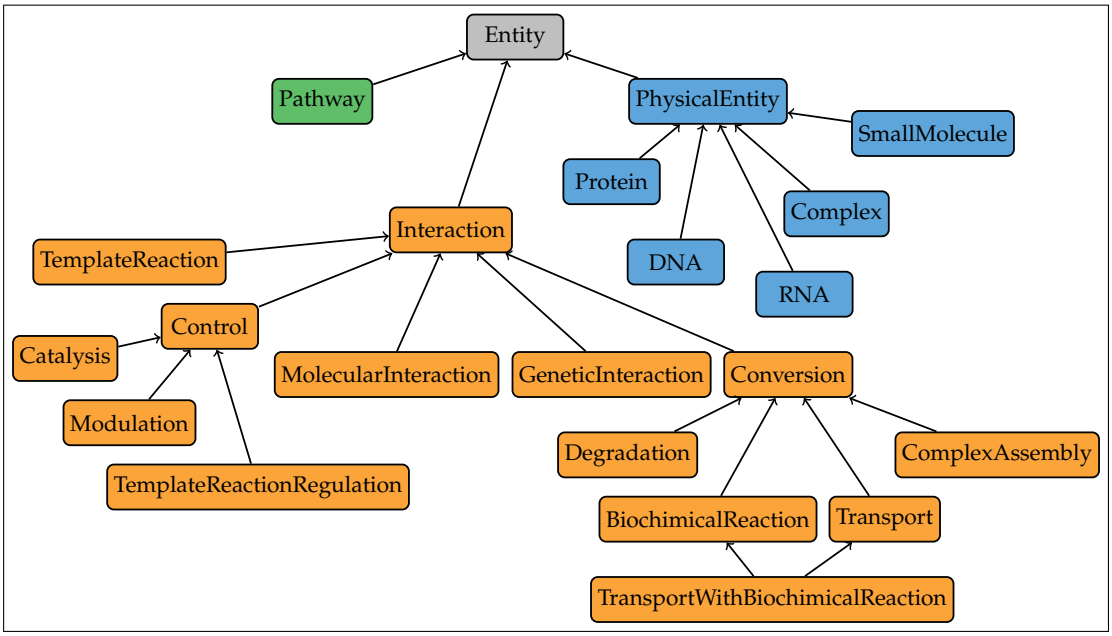


Figure 31 – *Résumé de l'ontologie BioPAX.*

Dans ce graphe les nœuds sont des classes et les arêtes représentent la relation d'héritage entre une classe fille et la classe parente. Les trois principaux types de classes en BioPAX sont Pathway (vert), Interaction (orange) et PhysicalEntity et gène (bleu).

de l'ARNm, etc.). Il est implémenté comme une ontologie qui aide à structurer les données afin qu'elles soient plus facilement traitées de façon automatique. Une voie biologique sous BioPAX sera vue comme un *Pathway* contenant des interactions et des entités physiques. Afin de décrire les différents types de voies, BioPAX propose différentes classes (figure 31).

Par exemple dans une réaction enzymatique d'une voie métabolique, les substrats et les produits pourront être représentés par la classe *SmallMolecule* et leur réaction sera décrite par la classe *Conversion*. Si une protéine catalyse la réaction, elle sera décrite avec la classe *Protein* et servira de contrôleur à une interaction de contrôle de classe *Catalysis*. Cette réaction de catalyse contrôlera la réaction de conversion.

Dans un autre exemple, une formation de complexe pourra être décrite par la classe *ComplexAssembly*. Cette réaction aura en entrée des entités de types *Protein* ou *SmallMolecule* et en sortie une entité de type *Complex*.

BioPAX prend en charge la combinaison de ces différents types de données en un seul modèle qui est utile pour obtenir une vue plus complète d'un processus cellulaire.

Chaque classe décrite par BioPAX possède à la fois un ensemble de propriétés propres et hérités de leurs classes parentes. Par exemple, toutes les classes possèdent les propriétés *name* ou *type*, la classe *Conversion* possède en plus les propriétés *letf* et *right* indiquant respectivement les réactants et les produits de la réaction, etc. Toutes

les propriétés utilisées pour la conversion de BioPAX à Cadbiom sont décrites au fur et à mesure dans la suite de ce chapitre.

BioPAX permet de créer des références croisées de bases de données et des liens vers des termes d'autres ontologies (figure 32). Par exemple, les termes de *Gene Ontology* sont utilisés pour décrire la localisation cellulaire, les vocabulaires PSI-MI (*Proteomics Standards Initiative Molecular Interaction*) [Hermjakob et al., 2004] sont utilisés pour définir des types d'interactions, des types de relations et des modifications de séquences, et *Sequence Ontology* [Eilbeck et al., 2005] est utilisé pour définir les types de régions de séquences, comme les régions du promoteur sur l'ADN impliquées dans la transcription d'un gène, etc.

La syntaxe de BioPAX est basée sur OWL (*Web Ontology Language*) [Antoniou and Harmelen, 2004]. De cette manière une ressource de données en BioPAX est facilement interrogeable en SPARQL ce qui donne la possibilité d'automatiser l'extraction de données BioPAX et d'interroger d'autres bases de données dans la même requête. Ceci est une avancée non négligeable dans le monde des données biologiques.

3.3 Conversion de données BioPAX en modèle Cadbiom

Il est possible de traduire un réseau BioPAX en un réseau Cadbiom, c'est-à-dire traduire des réactions faisant intervenir des entités en des transitions reliant des nœuds Cadbiom.

Par abus de langage, j'utiliserai le terme « entité » pour parler des sous-classes de « PhysicalEntity » et le terme « réaction » pour parler des sous-classes d'« Interaction » excepté la classe « Control », qui englobe notamment les réactions de catalyse.

3.3.1 Réactions simples à traduire

Dans le formalisme Cadbiom, la transmission du signal se traduit par des transitions d'états de biomolécules sous le contrôle d'une garde (fonction logique). Voici une liste des différentes classes BioPAX que nous traduisons en Cadbiom et comment nous les traduisons :

Classe « PhysicalEntity » et ses classes filles

Ces classes décrivent toutes les molécules possibles (protéine, complexe, gène, etc.), leur traduction n'est pas une tâche difficile. Elle revient seulement à créer un élément Cadbiom pour chaque entité et chaque localisation. BioPAX intègre une notion de localisation d'une entité. Pour traduire cette notion, nous sommes obligés de créer deux éléments Cadbiom de la même entité (figure 33b). Cependant il faut faire attention à ce que chaque entité ait un identifiant Cadbiom unique pour ne pas fusionner plusieurs entités dans le modèle Cadbiom. Pour cela nous nous basons sur la propriété « displayName » puis si elle n'est pas unique nous comparons les propriétés « name » qui sont la liste des synonymes de chaque entité et si enfin aucun synonyme n'est unique pour chaque entité

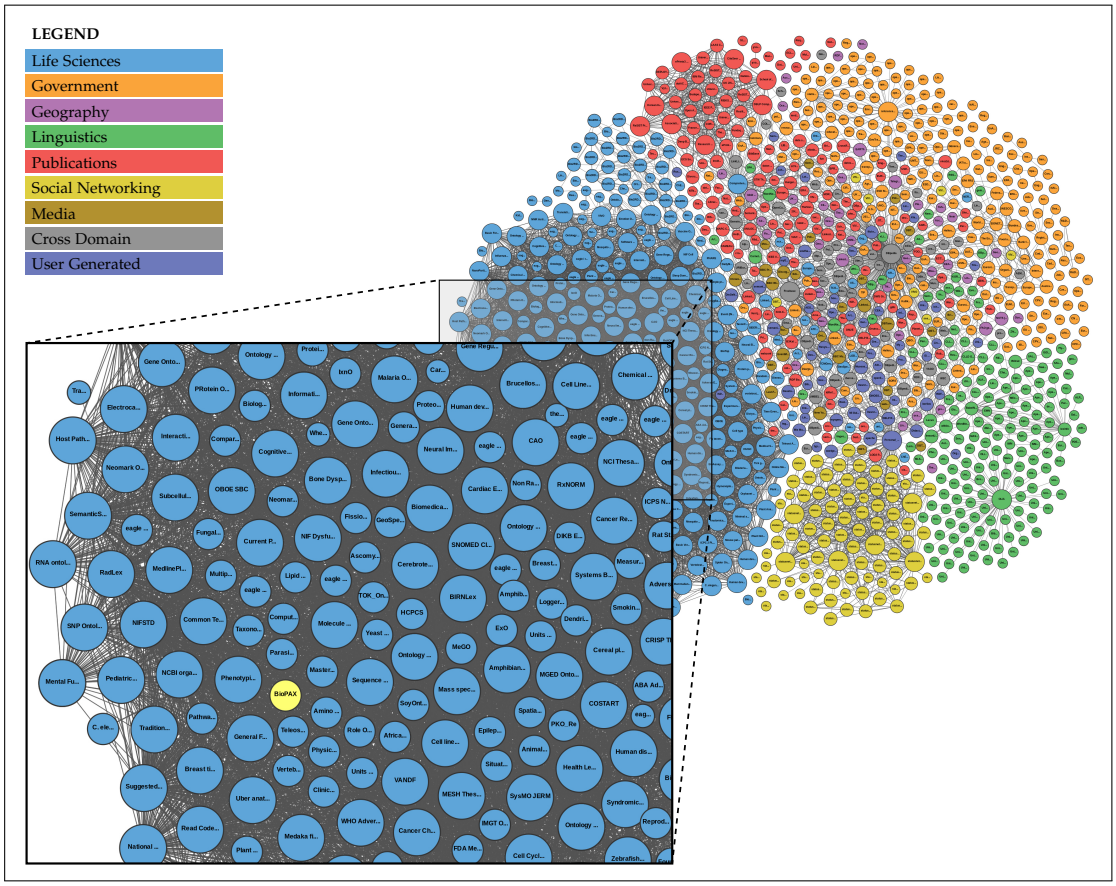


Figure 32 – *BioPAX et ses voisins dans LOD Cloud 2017.*

Cette figure représente les différentes bases de données accessibles en RDF, elle a été réalisée par les contributeurs du projet Linking Open Data (LOD) : `http://lod-cloud.net`. *Les nœuds représentent les bases, leur couleur représente le domaine de connaissance auquel elles appartiennent et BioPAX est affiché en jaune.*

alors nous créons un nom unique sous le format :
« displayName »+« _vID_UNIQUE ».

Classe « Control » et ses classes filles

Ces classes décrivent le contrôle d'une molécule sur une réaction (comme les réactions de catalyses). Ce type de classe possède trois propriétés qui nous intéressent :

— *controlled* : l'identifiant de la réaction contrôlée ;

— *controller* : l'entité contrôlant la réaction ;

— *controlType* : le type de contrôle (activation ou inhibition).

Pour traduire ces classes, il faut ajouter le « *controler* » dans la garde des tran-

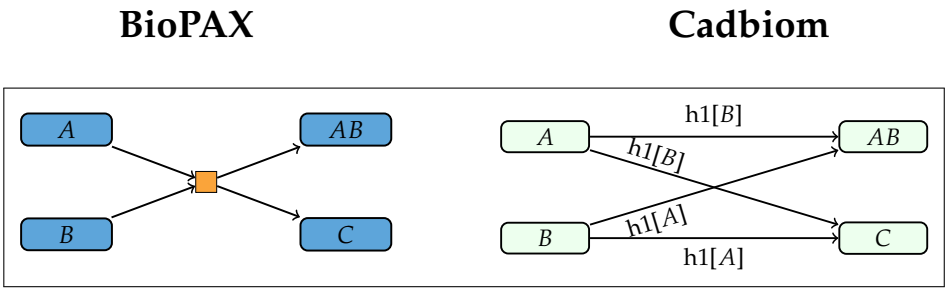
sitions traduisant la réaction contrôlée (figure 33b). Si la classe représente un contrôle négatif, comme par exemple une inhibition, alors la condition sera composée d'un « not ». Si deux entités activent une réaction alors ces deux entités seront séparées par une fonction logique OR ou AND dans la garde (figure 33d). Le modélisateur doit décider si ces deux entités doivent être toutes les deux présentes pour activer la réaction ou si l'une d'elles suffit, ce choix sera discuté plus bas dans le manuscrit.

Classes « BiochimicalReaction », « Transport » et « ComplexAssembly »
Ces classes décrivent les interactions possibles entre les molécules. À l'inverse des « TemplateReaction », elles correspondent à des réactions chimiques. Elles possèdent à la fois des entités d'entrée et des entités de sortie. Une réaction peut être traduite en plusieurs transitions. En règle générale le nombre de transitions traduisant cette réaction est égal au nombre de combinaisons possibles de ses entrées et de ses sorties. On attribue le même évènement à l'ensemble des transitions traduisant la même réaction afin de transmettre le signal en même temps. Nous verrons dans la section suivante les exceptions à cette règle (figure 33a).

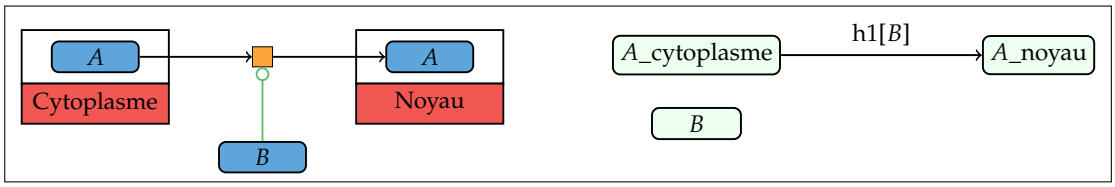
Classe « TemplateReaction »
Une « TemplateReaction » est une réaction particulière de BioPAX. Elle correspond à une réaction de polymérisation où les coefficients stœchiométriques ne sont pas précisément connus. La réaction n'obéit pas à la loi de la conservation de masse. Une « TemplateReaction » peut correspondre à la transcription d'un gène en un ARN messager, à la traduction d'un ARN messager en une protéine, ou même aux deux à la fois. L'expression d'un gène est en général décrite par ce type de réaction qui dans ce cas ne possédera pas d'entrée, mais seulement une sortie : la protéine traduite. Pour exprimer la même notion en Cadbiom, nous devons créer un élément « gène » qui n'existe pas dans cette réaction, cet élément a pour nom la protéine suivie de la chaîne de caractères « _gene ». Nous traduisons alors cette réaction par une transition du gène à la protéine, activée ou inhibée par des éléments de « Control » (figure 33c).

Classe « Degradation »
Une « Degradation » est l'inverse d'une « TemplateReaction », c'est-à-dire qu'elle a une entrée (l'entité à dégrader), mais qu'elle n'a pas de sortie. Cela correspond à l'élément « TrapNode » de Cadbiom qui permet de décrire cette dégradation. Une dégradation BioPAX est traduite par une transition de l'entité à un élément « TrapNode » (figure 33d).

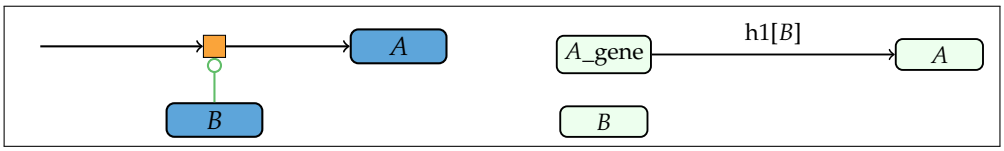
Classe « Pathway »
En BioPAX comme dans beaucoup de bases de données (Reactome, KEGG, PID), il y a une notion de *pathway* et même de sous-*pathway*. Un *pathway* représente la carte d'une voie métabolique et regroupe un ensemble de réactions. Donc au sein de la base de données il existe un ensemble de *pathways* contenant des ensembles de réactions. Il n'y a pas cette notion de *pathway* en Cadbiom, mais seulement de modèle. Un *pathway* sera traduit en un seul modèle *Cadbiom* qui comprendra tous ses sous-*pathway*.



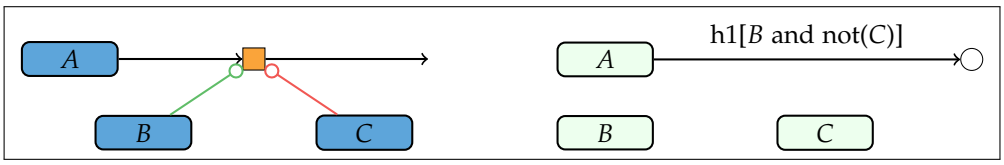
(a) *Formation d'un complexe et d'une entité.*



(b) *Réaction de transport d'une entité.*



(c) *Traduction d'une protiéine.*



(d) *Dégradation d'une entité.*

Figure 33 – ***Exemples de traduction des différentes classes BioPAX en Cadbiom.***

Cette figure présente différentes réactions BioPAX (à gauche) traduites en Cadbiom (à droite). Les entités BioPAX sont représentées par des nœuds bleus, les localisations par un cadre rouge et les réactions par un rectangle orange. Dans les réseaux, le même évènement est attribué à l'ensemble des transitions traduisant la même réaction et une entité présente dans plusieurs localisations est traduite en plusieurs éléments différents. (a) Représentation de la formation d'un complexe AB et d'une autre molécule C à partir des entités A et B. (b) Représentation d'une réaction de transport de la protéine A du cytoplasme au noyau activée par l'entité B. (c) Traduction ou expression de la protéine A activée par l'entité B. (d) Dégradation de l'entité A activée par l'entité B et inhibée par l'entité C.

3.3.2 Gestion des entités parentes de BioPAX

N'importe quelle entité en BioPAX (*Protein*, *DNA*, *SmallMolecule*, etc.) peut posséder plusieurs entités membres désignées par la propriété « memberPhysicalEntity ». Cette propriété permet de définir une entité générique qui regroupe d'autres entités.

Elle peut être utilisée de plusieurs manières. Par exemple si tous les isoformes d'une protéine, comme TGF-β1, TGF-β2 et TGF-β3, permettent d'activer une réaction, alors au lieu décrire trois fois le même contrôle avec une isoforme différente, il est possible de créer une entité générique TGF-β qui regroupera grâce à la propriété « memberPhysicalEntity » les trois isoformes. Puis il suffira de créer un seul contrôle de la réaction avec l'entité générique TGF-β.

Une autre manière d'utiliser cette propriété est de créer une famille d'entités qui possèdent un point commun. Par exemple, supposons que la protéine p_1 active la synthèse d'une dizaine de protéines. Au lieu de créer 10 « TemplateReactions » chacune contrôlée par p_1, il est possible de créer une entité générique de ces dix protéines qu'on nommerait « Protéines influencées par p_1 » et de créer seulement une « TemplateReaction » contrôlée par p_1. Dans ce cas, il n'y a aucune réalité biologique mais la lecture du réseau est simplifiée.

Cette propriété nous intéresse fortement, car Cadbiom n'a pas de notion d'élément générique regroupant d'autres éléments. Dans Cadbiom, un élément est un nœud élémentaire sans hiérarchie avec les autres éléments.

Une réaction faisant intervenir une ou plusieurs entités génériques ne sera pas traduite en Cadbiom de la même façon selon que les membres des entités génériques représentent une même molécule ou non. Pour expliquer ce concept, je vais m'appuyer sur l'exemple (a) et (b) de la figure 34. Dans l'exemple (a), la réaction BioPAX représente la transformation de l'entité générique A en l'entité générique B. Chacune de ces deux entités possède des « memberPhysicalEntity », A_1, A_2 et A_3 pour A et B_1 et B_2 pour B. Aucun membre de A ne représente la même molécule qu'aucun membre de B, il est possible pour A_1 d'être transformé autant en B_1 qu'en B_2, idem pour A_2 et A_3. Dans ce cas la traduction de cette réaction correspond à 6 transitions toutes possibles (nombre de membres de A multiplié par le nombre de membres de B). Par contre l'exemple (b) représente le transport de l'entité A du cytoplasme au noyau, A_1 du cytoplasme représente la même molécule que A_1 du noyau, idem pour A_2 et A_3. Dans ce cas seulement 3 transitions traduiront la réaction. Évidemment dans des données réelles, une réaction peut relever des deux situations précédentes, des entités qui représentent les mêmes molécules et d'autres qui n'ont aucun lien. Pour lier les entités entre elles, nous utilisons la propriété « entityRef » qui permet de définir une entité de références pour plusieurs entités ou la propriété « xref » qui permet de définir le lien avec des identifiants d'autres bases de données (par exemple les identifiants Uniprot).

La difficulté de traduire en Cadbiom des données BioPAX avec des entités génériques se répercute aussi lors de la traduction des complexes d'entités. Un complexe est une molécule correspondant à la liaison de plusieurs molécules. Si un composant du complexe est une entité générique alors il multiplie le nombre de possibilités de ce complexe. Par exemple dans la figure 34c, A est une entité générique avec trois

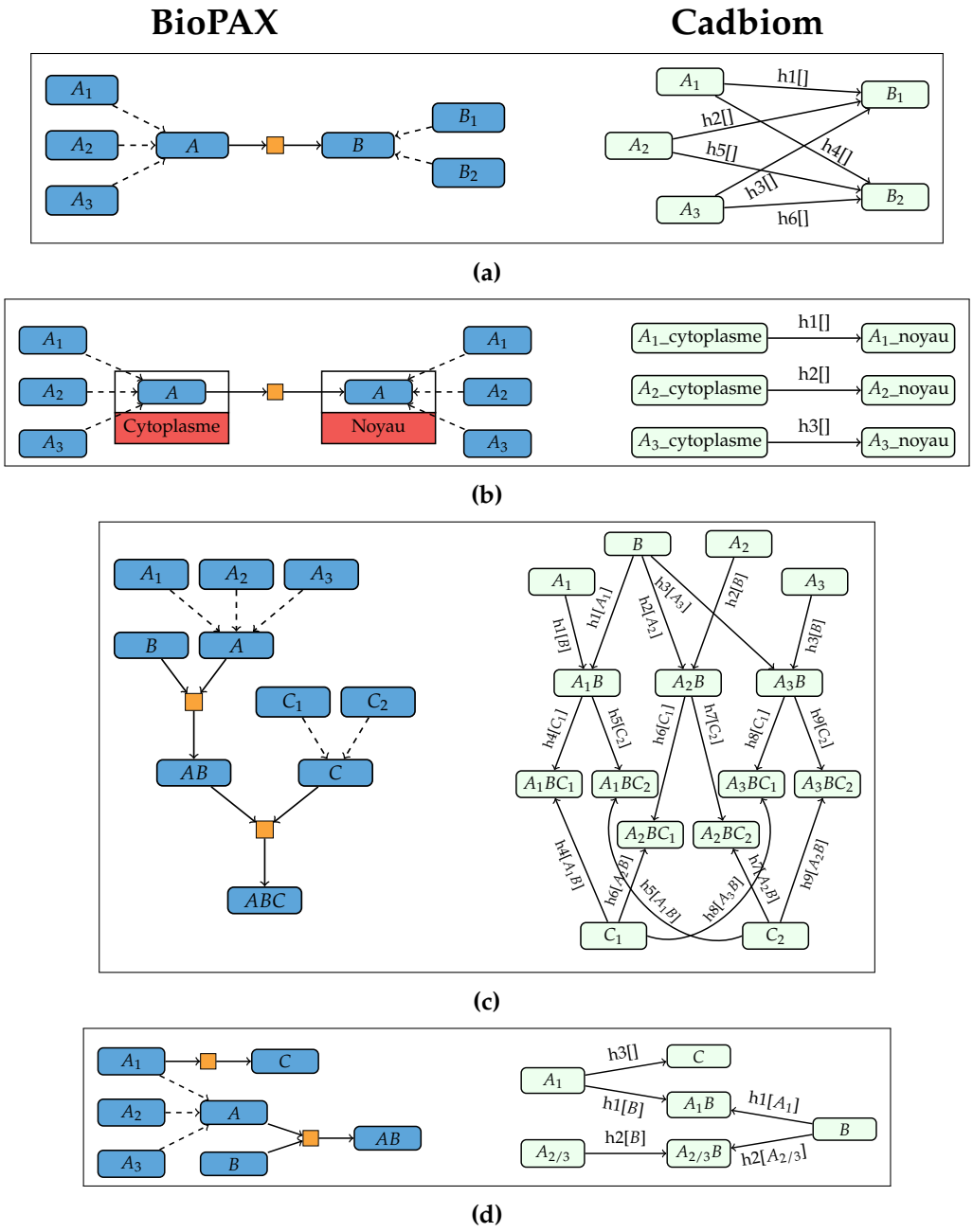


Figure 34 – *Exemples de traduction de la proprité « members » de BioPAX en Cadbiom.*

Cette figure présente différentes réactions BioPAX qui font intervenir des entités utilisant la propriété « members » (à gauche) traduites en Cadbiom (à droite). Les entités BioPAX sont représentées par des nœuds bleus, les localisations par un cadre rouge, les réactions sont représentées par un rectangle orange, et le lien entre les entités membres d'une entité générique est représenté par une arête en pointillé. (a) Représentation d'une réaction transformant l'entité A en l'entité B, étant donné que A et B sont des entités génériques le nombre de transitions possibles correspond à toutes les combinaisons de membres de A et des membres de B. (b) Représentation d'une réaction de transport de la protéine A du cytoplamse au noyau, dans ce cas les membres de A dans les deux localisations sont identiques, le nombre de transitions correspond juste aux nombres de membres de A. (c) Représentation de la formation d'un complexe ABC, si un complexe possède plusieurs entités génériques alors il est traduit en plusieurs éléments Cadbiom selon toutes les combinaisons possibles. (d) Représentation de la formation d'un complexe AB et de la transformation de A_1 en A_3, les possibilités de A_2B et A_3B ont été regroupées ensemble, car elles sont équivalentes.

membres, elle forme avec B un complexe. Le complexe AB représente en réalité trois complexes A_1B, A_2B et A_3B. C'est donc pour cette raison que l'entité AB est traduite en trois éléments Cadbiom. Un complexe se traduit en autant d'éléments Cadbiom qu'il y a de combinaisons possibles des membres de chaque entité composant le complexe. C'est pourquoi le complexe ABC correspond à 6 éléments Cadbiom (nombre de membres de A multiplié par le nombre de membres de C). Dans certaines bases de données comme Reactome, une entité générique peut regrouper des centaines d'entités et faire partie de plusieurs complexes. La complexité engendrée dans le réseau devient alors très importante.

Comme nous l'avons vu, développer toutes les entités génériques en entités élémentaires augmente considérablement la taille du réseau. Cependant il est possible de développer seulement les entités génériques nécessaires, afin d'éviter au maximum de faire grandir le réseau. Si on reprend nos exemples de la figure 34, les différentes transitions représentent différents chemins équivalents. Ils peuvent être résumés par un seul chemin sauf si un élément est utile à une autre réaction. Dans ce cas, il faut obligatoirement décrire cette possibilité. Prenons l'exemple (d) de la figure 34, les entités A_1, A_2 et A_3 font partie de l'entité générique A qui forme un complexe avec B. Comme A_1 est transformé en C par une autre réaction, il n'est pas équivalent à A_2 et A_3. C'est pourquoi il est possible de regrouper A_2 et A_3 ensemble (sous le nom de $A_{2/3}$) mais il faut décrire séparément A_1. De cette manière, nous avons développé au minimum l'entité générique.

3.3.3 Gestion des incohérences

Durant notre analyse de traduction de BioPAX en Cadbiom, nous avons trouvé certaines incohérences dans les réactions BioPAX provenant sûrement d'un problème de traduction d'un autre formalisme en BioPAX.

La première consiste en une mauvaise utilisation d'une entité catalysant une réaction. Au lieu d'utiliser la classe « Catalysis » indiquant quel est le catalyseur et quelle est la réaction catalysée, le modélisateur a directement ajouté le catalyseur à la réaction en entrée et en sortie (car l'élément n'est pas consommé). Cette erreur provoque dans le formalisme Cadbiom un élément (le catalyseur) qui a une transition pointant sur lui-même. C'est pourquoi nous corrigeons les données BioPAX avant de les traduire afin d'éviter ce type d'erreur.

La seconde incohérence que nous avons trouvée est une mauvaise façon de décrire l'expression d'un gène. Comme nous l'avons vu, la bonne solution est d'utiliser la classe « TemplateReaction », mais certaines bases de données ne sont pas formalisées de la même manière. Par exemple, Reactome décrit l'expression d'un gène à partir d'une entité gène et d'une protéine (de temps en temps il décrit même l'ARN messager). C'est pourquoi lorsque Reactome est exporté en BioPAX, il utilisera la classe « BiochimicalReaction » avec en entrée de l'ADN et en sortie une protéine. Cela ne pose pas de problème de décrire de cette manière en Cadbiom, mais il faut garder en mémoire que cette réaction correspond à l'expression d'un gène.

3.3.4 Discussion à propos de BioPAX

Nous l'avons vu, BioPAX se veut être un langage universel pour toutes les bases de données de voies biologiques. C'est d'ailleurs pourquoi une très grande majorité des bases de données sont disponibles sous ce format. De plus BioPAX se base sur le modèle RDF et est donc interrogeable en SPARQL, contrairement à SBML, ce qui facilite son intégration avec un grand nombre d'autres bases de données (figure 32).

Cependant cette universalité a un coût, car il est très permissif. Nous l'avons vu, il est possible de décrire une réaction avec une mauvaise classe. Étant donné que les classes BioPAX ont une hiérarchie, un risque est que l'utilisateur décrive toutes ses réactions ou ses entités avec les classes généralistes. Par exemple il pourrait décrire les protéines ou l'ADN avec la classe « PhysicalEntity » ou toutes les réactions avec la classe « Interaction ». Ce cas de figure a été relevé dans la base Reactome décrivant de l'ADN à partir d'une classe « PhysicalEntity ». Ce genre d'erreurs impacte toutes les méthodes automatiques qui se basent sur les différentes classes pour tirer du sens des systèmes décrits.

C'est aussi pour cette raison que l'on peut s'interroger sur la manière dont ces bases de données représentent leurs *pathways* en BioPAX. Proposent-elles une description de leurs données en BioPAX de la bonne manière ou se contentent-elles de tout traduire avec des classes très générales ? Pour s'en convaincre, une comparaison des données BioPAX par rapport aux données d'origines serait nécessaire.

De plus BioPAX est un langage d'échange de données, proposant une description qualitative. Nous avons vu que la possibilité de créer des entités génériques impactait considérablement la conversion en Cadbiom. Mais cela est vrai pour tous les langages visant à analyser un système et qui n'ont pas de notion de hiérarchie entre leurs éléments. Pourtant cette propriété facilite le travail du modélisateur et résume l'information dans le réseau. On comprend alors que la manière de représenter les données peut impacter considérablement le processus de conversion de ses données en un formalisme de modélisation. Cette information est à prendre en compte lors du choix du langage de description et du formalisme de modélisation.

3.4 Comparaison des bases de données de signalisation

Dans cette étude, je me suis concentré sur la comparaison qualitative de plusieurs bases de données en BioPAX et en Cadbiom. J'essaierai de mettre en évidence leurs différences topologiques et je proposerai un protocole d'étude pour caractériser l'apport respectif de ces différentes bases de données pour la création du réseau exhaustif de la signalisation du TGF-β. Le modèle des voies de signalisation provenant de PID, présenté dans le chapitre 2, est utilisé comme modèle de référence.

La base de données *Pathway Commons* [Cerami et al., 2011] regroupe plusieurs autres bases de données au format BioPAX que nous allons utiliser comme source de données BioPAX et traduire en langage Cadbiom.

3.4.1 *Pathway Commons*

Pathway Commons [1] est une collection libre d'accès de données de voies de signalisation, de réseaux métaboliques et de réseaux de régulation de gènes. *Pathway Commons* fournit une interface web permettant de parcourir, de rechercher des voies biologiques et de les télécharger en différents formats, comme le format BioPAX ou le format PSI–MI [Hermjakob et al., 2004]. *Pathway Commons* vise à recueillir et à normaliser toutes les données provenant de plusieurs bases. À l'heure actuelle, *Pathway Commons* contient 22 bases de données et plus de 4000 *pathways* concernant l'*Homo sapiens*.

L'intégration de données dans *Pathway Commons* comporte trois étapes :

— l'agrégation et la validation des données ;

— la normalisation des *identifiants* par ceux utilisés par *Pathway Commons* ;

— la fusion des entités physiques, si elles ont le même identifiant.

Pathway Commons ne fusionne pas les réactions ou les chaînes de réactions, il peut donc exister plusieurs réactions ou plusieurs chaînes de réactions décrivant le même processus. Nous discuterons de l'impact de cela sur le modèle Cadbiom et sur les trajectoires résultantes.

Nous utiliserons dans notre étude 5 bases de données contenues dans *Pathway Commons* et possédant des données de signalisation :

— PID [Carl F. Schaefer et al., 2009] (version finale - 27/07/2015)

— Reactome [Croft et al., 2014, Fabregat et al., 2016] (version 61 - 23/07/2017)

— Inoh [Yamamoto et al., 2011] (version 4 - 22/03/2011)

— Netpath [Kandasamy et al., 2010] (version de décembre 2011)

— Panther Pathways [Mi and Thomas, 2009] (version 3.4.1 - 04/07/2016)

J'ai décidé de ne pas inclure KEGG dans l'étude, car aucun *pathway* de signalisation n'a été traduit en BioPAX.

Le modèle PID original de [Andrieux et al., 2014] sera noté $PID_{original}$ et le modèle PID provenant de *Pathway Commons* sera noté PID afin de distinguer les réactions qui étaient dans le modèle d'origine issu de PID et celles qui ont été ajoutées à la nouvelle version en BioPAX de PID.

3.4.2 Stratégie proposée

La stratégie adoptée pour comparer la signalisation du TGF-β dans chaque base de données suit les étapes suivantes :

1. `http://www.pathwaycommons.org`

1. **Créer les modèles Cadbiom** :
   À partir des fichiers BioPAX proposés par *Pathway Commons*, j'ai créé 5 modèles Cadbiom et j'ai récupéré le modèle Cadbiom d'origine.
2. **Comparer la topologie des modèles** :
   J'ai effectué une analyse statique préliminaire des 6 modèles Cadbiom pour connaître les différences entre les bases de données concernant les nombres d'entités, le nombre de réactions, la densité du réseau, etc.
3. **Comparer les trajectoires de $PID_{original}$ et PID** :
   Afin de valider la conversion des données BioPAX en un modèle Cadbiom, il est nécessaire de comparer les trajectoires provenant du modèle $PID_{original}$ avec les trajectoires du nouveau modèle PID.
4. **Enrichir les voies de signalisation du TGF-β** :
   L'idée est de calculer l'ensemble des trajectoires contenant TGF-β et activant au moins gène dans chacun des modèles correspondant aux bases de données. Les objectifs sont : (1) comparer la composition des trajectoires et l'ensemble des gènes qu'elles influencent ; (2) enrichir de manière exhaustive les voies de signalisation du TGF-β en utilisant les trajectoires spécifiques de chaque base.

3.4.3 Création des modèles

Les 5 modèles Cadbiom ont été générés avec notre convertisseur. Cependant certaines réactions utilisent ou produisent de façon systématique des « *SmallMolecules* », comme les nucléosides triphosphate permettant le transfert d'énergie intra-cellulaire (ATP, GTP, etc.). Ces *SmallMolecules* deviennent alors des nœuds centraux dans les modèles et provoquent des trajectoires factices car elles ne représentent pas la propagation du signal. Nous avons donc décidé de faire un post-traitement pour supprimer ces *SmallMolecules* des modèles Cadbiom.

3.4.4 Comparaison topologique des modèles

Les 5 modèles provenant des bases de données et le modèle $PID_{original}$ sont très différents. La table 13 résume les différentes observations.

Premièrement, le nombre d'entités varie de 3275 pour *NetPath* à 42349 pour *Reactome*. De plus, le rapport entre le nombre d'entités et le nombre de réactions varie considérablement entre les bases de données. En effet, la base *Inoh* est composée de 17244 entités pour seulement 2756 réactions alors que la base *NetPath* est composée de 3275 entités pour 4337 réactions. Ceci pourrait indiquer de nombreuses réactions concurrentes pour les mêmes processus.

Concernant les entités *SmallMolecules* que nous supprimons des modèles, la base *NetPath* n'est composée d'aucune *SmallMolecules*, et le pourcentage maximal de *SmallMolecules* par rapport au nombre total d'entités est de 13 % pour la base *Panther*.

	$PID_{original}$	*PID*	*Reactome*	*Inoh*	*NetPath*	*Panther*
Nombre d'entités BioPAX		12017	42349	17244	3275	8064
Nombre d'entités « *SmallMolecules* »		173 (1 %)	3536 (8 %)	1034 (6 %)	0 (0 %)	1081 (13 %)
Nombre de réactions BioPAX		6504	10253	2756	4337	2983
Nombre de nœuds Cadbiom	9178	9464	16568	4571	2057	3969
Nombre de nœuds Cadbiom en place frontière	3919 (43 %)	3830 (40 %)	6105 (37 %)	2474 (54 %)	738 (36 %)	1681 (42 %)
Nombre de transitions Cadbiom	5229	6697	11131	1931	1320	2332
Degré moyen des nœuds Cadbiom	0.015	0.016	0.009	0.021	0.062	0.032

Table 13 – *Statistiques des modèles de base de données.*

Le degré moyen des nœuds d'un modèle Cadbiom correspond au nombre moyen de transitions entrantes et sortantes d'un nœud divisé par le degré maximal possible du modèle (soit $n-1$ où n est le nombre total de nœuds). Ce score permet d'avoir une notion sur la densité du réseau. Une forte densité indique que les nœuds sont fortement connectés les uns aux autres, ce qui permet différents chemins alternatifs, alors qu'une faible densité indique que les transitions ne forment que quelques chemins. On peut voir que le modèle *NetPath* a la plus haute densité (0.062) alors que celle de *Reactome* est très faible.

Les nœuds Cadbiom en place frontière correspondent aux nœuds ne possédant pas de transitions entrantes, leur pourcentage par rapport au nombre total de places varie de 36 % pour *NetPath* à 54 % pour *Inoh*.

Enfin le nombre de transitions est logiquement corrélé avec le nombre de réactions BioPAX, mis à part pour la base *NetPath* dont le nombre de transitions est 4 fois plus petit. Puisque *NetPath* semble avoir de nombreuses réactions concurrentes, cela se traduit en Cadbiom par plusieurs conditions dans la transition. Le nombre de transitions n'augmente pas pour les réactions concurrentes.

Enfin le modèle $PID_{original}$ et le modèle PID sont pratiquement semblables. Ils proviennent des mêmes données sauf que l'un a été importé à partir de données SBML et l'autre de données BioPAX. Nous pouvons tout de même noter une légère augmentation du nombre de nœuds et de transitions. Ceci peut s'expliquer par la notion d'entité parente (vue à la section 3.3.2) qui a été prise en compte dans la génération du nouveau modèle PID.

Pour illustrer la conversion des données BioPAX en un modèle Cadbiom, j'ai décidé de vous présenter le réseau de signalisation de la prolactine (*Signaling events mediated by PRL*) proposé par la base de données PID et mis à disposition en BioPAX par *Pathway Commons*. La figure 35 représente donc le *pathway* au format BioPAX et la figure 36 représente ce même *pathway* dans un modèle Cadbiom. Les *SmallMolecules* ont été supprimés du modèle Cadbiom et j'ai considéré qu'un seul activateur suffisait à activer une transition. Comme on peut le voir, les données ont été converties sans aucune erreur.

Les deux prochaines sections décrivent le protocole qu'il faudrait suivre pour finir cette analyse. Les calculs sont encore en cours.

3.4.5 Comparaison des trajectoires de $PID_{original}$ et PID

Le modèle $PID_{original}$ a été construit à partir d'une des dernières versions SBML de PID. Le fichier BioPAX de PID proposé par *Pathway Commons* devrait être très proche de la version SBML. Cependant certains points sont différents dans la façon de faire le modèle :

— La notion de classe générique, c'est-à-dire d'entité parente regroupant plusieurs entités, existe aussi en SBML. Mais elle n'a pas été prise en compte lors de la

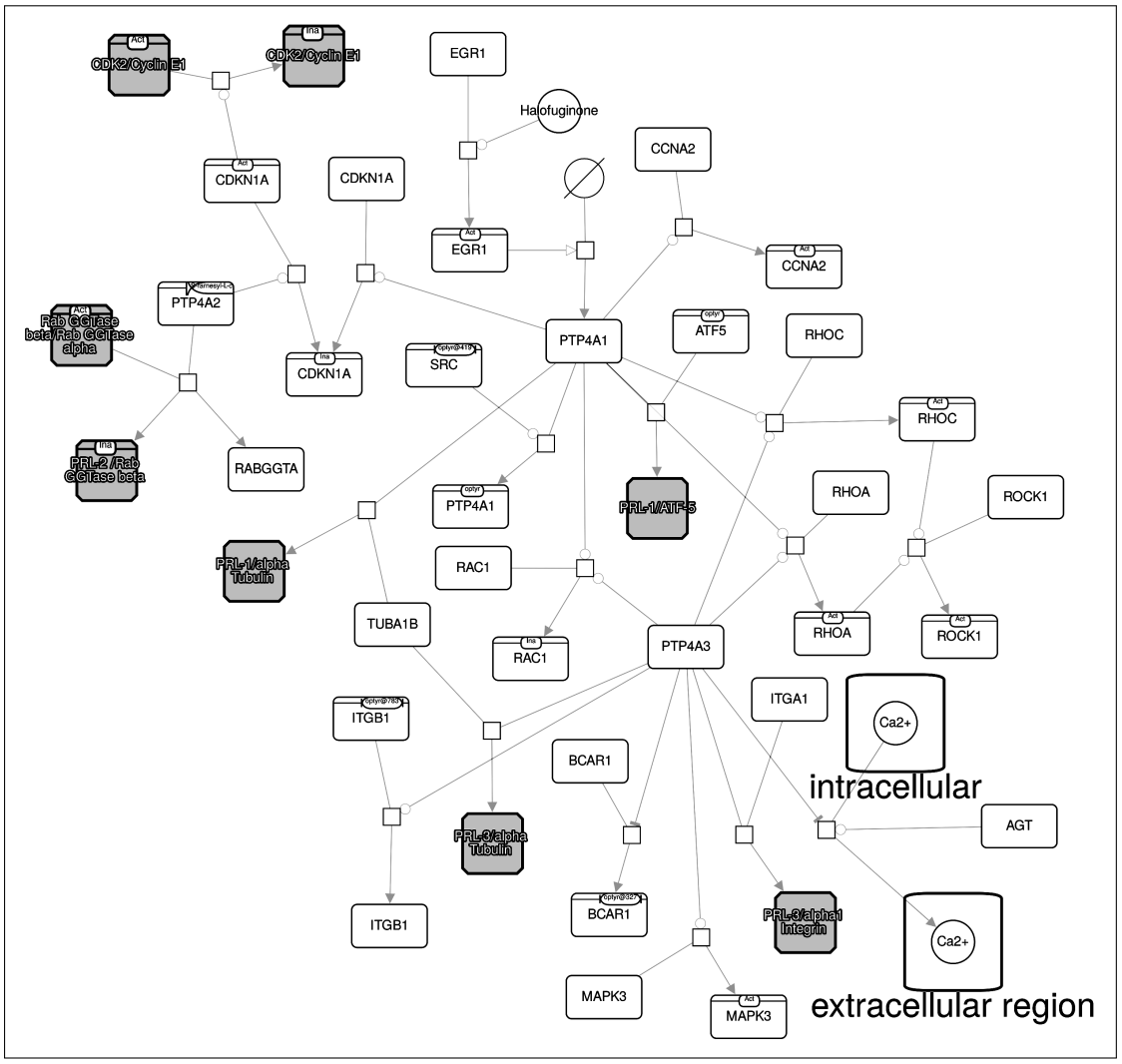


Figure 35 – ***Réseau de la voie de signalisation de la prolactine (PLR) dans la base PID***

Représentation proposée par Pathway Commons des données BioPAX concernant le pathway « Signaling events mediated by PRL » de la base PID. Les grands nœuds gris clair représentent les entités, les grands nœuds gris foncé représentent les complexes d'entités et les petits carrés nœuds représentent les réactions. Il y a cinq types d'arrêtes différentes dans ce réseau : les arrêtes sans flèche relient les réactants aux réactions ; les arrêtes avec une flèche pleine relient les réactions aux produits ; les arrêtes avec une flèche vide relient les contrôleurs positifs aux réactions ; les arrêtes avec une flèche ronde relient les catalyseurs aux réactions ; les arrêtes avec une barre relient les inhibiteurs aux réactions.

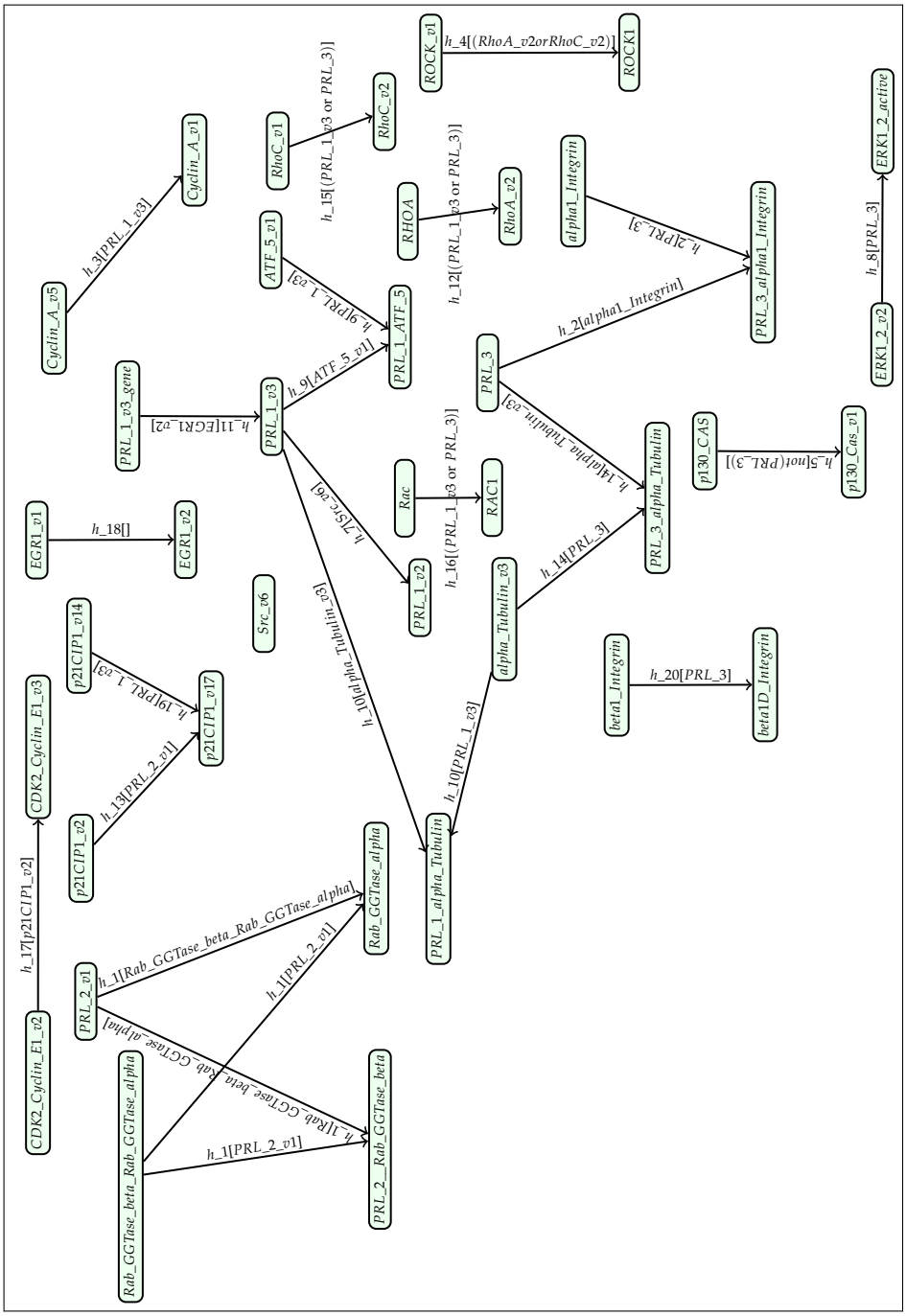


Figure 36 – ***Réseau de la voie de signalisation de la prolactine (PLR) dans le modèle Cadbiom***

Représentation de la traduction en Cadbiom du pathway « Signaling events mediated by PRL » de la base PID. Les SmallMolecules ont été supprimés du pathway. Les nœuds représentent les places Cadbiom pouvant être actives ou inactives. Les arrêtes représentent la transmission du signal et possèdent une guarde qui est une fonction logique conditionnant cette transmission.

création de $PID_{original}$.

— De plus pour des raisons de performance, [Andrieux et al., 2014] ont décidé qu'une transition ne pouvait se faire que si la totalité des activateurs de cette transition était présent. Le solveur permettant de calculer les trajectoires dans Cadbiom a été optimisé. Nous avons donc revu la manière de créer les conditions des transitions. Dorénavant une transition peut être effectuée si au moins un activateur est présent et qu'aucun inhibiteur ne l'est.

Pour analyser l'impact de ces différences, je propose l'expérience suivante :

1. Détecter l'expression de gènes dans le modèle PID. Pour cela nous pourrions utiliser trois types de réactions :
   — les *TemplateReactions* produisant une protéine.
   — tous types de réactions consommant une entité *DNA*.
   — les *BiochimicalReaction* produisant une protéine et ne consommant aucune entité. Cette description n'est pas recommandée par BioPAX mais j'ai relevé son utilisation dans certaines bases de données.
2. Calculer les trajectoires permettant d'activer l'expression de ces gènes.
3. Sélectionner seulement les trajectoires contenant la protéine TGF-β.
4. Comparer ces trajectoires avec les trajectoires de $PID_{original}$. Pour cela nous pouvons envisager différentes stratégies :
   — Comparer le nombre et la taille des trajectoires ;
   — Utiliser la méthode de *clustering* RSC sur à la fois les trajectoires originales et les nouvelles trajectoires, si les *clusters* sont composés des deux types de trajectoires sans distinctions alors la composition de celles-ci est similaire ;
   — Visualiser localement les trajectoires à partir de l'interface web pour vérifier la pertinence biologique de celles-ci.

De cette manière, il sera possible de juger des différences entre les deux modèles et d'adapter en fonction notre méthode pour permettre d'avoir les trajectoires les plus proches des connaissances actuelles.

3.4.6 Enrichissement des voies de signalisation du TGF-β

L'étape suivante consistera à calculer toutes les trajectoires de chacun des 5 modèles. Comme dans la section précédente, il faudra détecter l'expression de tous gènes des modèles, puis les trajectoires qui les influencent et enfin filtrer les trajectoires contenant la protéine TGF-β.

La méthode de *clustering* RSC pourra permettre de regrouper toutes les trajectoires dans différents *clusters*. Deux cas de figure pourront alors se produire :

— des *clusters* concernant plusieurs trajectoires de différents modèles, ce qui constituera les voies de signalisation du TGF-β communes aux différentes bases de données ;

— des *clusters* spécifiques aux modèles, c'est-à-dire des trajectoires uniques à chacune des bases de données.

Après avoir analysé chaque base séparément, il faudrait les intégrer en un modèle unique décrivant le réseau de signalisation du TGF-β le plus complet possible.

Cependant, cette stratégie se verra confrontée à un obstacle : la différence de description des processus biologiques. En effet, il est probable que les bases de données décrivent des processus similaires mais avec des niveaux de granularités différents. Les trajectoires résultantes seront alors différentes bien qu'elles décrivent la même chose. Ce problème peut s'avérer difficile à surmonter mais propose un défi intéressant.

Conclusion

J'ai proposé un convertisseur de données BioPAX en modèle Cadbiom. La subtilité de certaines notions, comme les classes parentes, et la mauvaise utilisation de BioPAX par certaines bases de données ont demandé un gros travail d'analyse pour créer ce convertisseur. Les différentes bases de données proposées par *Pathway Commons* et traduites en modèles Cadbiom ont présenté de très grandes différences de densité et de taille. Par contre les modèles $PID_{original}$ et PID étaient très semblables.

Enfin cette étude préliminaire pose les bases pour réaliser une étude plus approfondie sur la variation de données entre les différentes bases de données. J'ai établi un protocole précis sur les différentes tâches à réaliser. Même si la différence de granularité au sein des bases de données peut être une difficulté, je pense qu'il est indispensable de réaliser des comparaisons de ces bases de données à partir d'un format unique.

Chapitre 4

Conclusion et perspectives

Les réseaux de signalisation cellulaire sont essentiels à la vie. Ils permettent aux cellules de détecter et d'interpréter les changements du micro-environnement pour fournir des phénotypes adaptés tels que la différenciation, la prolifération et l'apoptose. En conséquence, la perturbation ou l'altération des réseaux de signalisation ont été associées à de nombreuses maladies telles que la fibrose et le cancer.

Ces dernières décennies ont vu s'accumuler une masse d'informations concernant les acteurs moléculaires de la signalisation (stimuli extra-cellulaires, récepteurs membranaires, facteurs transcriptionnels, etc.) et leurs réactions biochimiques associées (activation des récepteurs, phosphorylation des protéines, transport, transcription, etc.). De nombreuses bases de données ont été créées à partir des données de la littérature dans le but de fournir une vue d'ensemble sur les différentes voies de signalisation cellulaire grâce à des formats descriptifs (REACTOME [Croft et al., 2014, Fabregat et al., 2016], PANTHER [Mi and Thomas, 2009], NCI-PID [Carl F. Schaefer et al., 2009], etc.).

Ces bases de données ont permis la création de modèles mathématiques de plus en plus complets permettant de calculer la dynamique des différentes molécules [Ryall et al., 2012, Smallbone and Mendes, 2013, Thiele et al., 2013]. La compréhension de la façon dont les molécules de signalisation se combinent pour fournir des trajectoires de signalisation est une condition préalable aux futures stratégies thérapeutiques, mais les analyses des grands réseaux de signalisation restent une tâche difficile. La stratégie la plus commune consiste à réduire le modèle afin de garder seulement ses caractéristiques essentielles, mais cette stratégie prend le risque de supprimer certaines informations importantes. Si l'on souhaite étudier de grands modèles de façon exhaustive et sans réduction, il est nécessaire de proposer de nouvelles méthodes d'analyse.

En informatique, un domaine d'étude nommé « *data-mining* » consiste justement à classifier de grands jeux de données. Il existe des stratégies de *clustering* non-supervisés classifiant les données sans données d'apprentissage et donc sans *a priori*. Au sein de ces méthodes, certaines ne peuvent pas supporter des données très hétérogènes et de grandes tailles (des milliers d'objets).

Au cours de mes travaux de thèse, j'ai exploré ces différentes méthodologies et j'ai choisi le *clustering* RSC basé sur les voisins les plus proches, car elle possède de nombreux avantages, comme son applicabilité à de larges jeux de données et le fait qu'elle ne nécessite pas de fixer le nombre de *clusters*. J'ai aussi appliqué l'analyse de concepts formels (FCA), permettant de classer les données en fonctions de différents niveaux de précision. Pour valider le regroupement des solutions, j'ai quantifié la

pertinence biologique de ces solutions en utilisant les annotations de termes *Gene Ontology*.

Le TGF-β joue un rôle majeur à la fois dans les processus physiologiques et pathologiques par des voies de signalisation canoniques et non canoniques qui réagissent de manière croisée avec d'autres voies de signalisation [Joan Massagué, 2012]. La méthode de *clustering* RSC a permis de regrouper les 6017 trajectoires induites par le TGF-β de l'article [Andrieux et al., 2014] en 5 familles sur la base de leur composition protéique. Ces familles de trajectoires sont associées à des fonctions biologiques spécifiques et connues dans les effets du TGF-β, ce qui valide la méthode.

Cette méthode propose une recherche exhaustive et sans *a priori* des solutions. La classification de ces solutions à partir de méthodes non-supervisées puis la validation des résultats à partir des annotations biologiques provenant d'une source de données différente (*Gene Ontology* dans mon cas) est une approche pertinente. Néanmoins, il aurait été possible d'utiliser des connaissances existantes afin de guider la classification des solutions [Villa et al., 2009], mais avec le risque d'orienter les résultats sur des connaissances déjà connues.

Pour améliorer cette méthode, une perspective intéressante serait de prendre en considération la séquentialité des réactions qui composent les trajectoires. À l'heure actuelle, les trajectoires sont comparées en fonction de leur composition protéique en les considérant comme des ensembles de protéines. Cependant les trajectoires traduisent la propagation du signal, il y a donc une notion de séquentialité entre les différentes interactions des molécules qui n'a pas été prise en compte. Une telle démarche pourrait aboutir à une augmentation du nombre de trajectoires possibles (la séquence $[A, B, C]$ est différente de la séquence $[B, A, C]$). Il pourrait être intéressant d'utiliser un nouveau score de similarité entre les trajectoires prenant en compte cette notion à la place de la corrélation de Pearson.

Dans le même esprit, une autre perspective pourrait être d'utiliser ma méthode de classification de trajectoires mais sur des données décrivant les variations de concentration de molécules en fonction du temps. Cela demanderait de repenser ma méthode, mais ouvrirait de nouvelles possibilités dans l'analyse de modèles continus.

D'un point de vue computationnel, deux processus ont un temps d'exécution non négligeable et nécessitent d'être améliorés :

1. La génération de trajectoires : pour un modèle comprenant environ 10000 nœuds, le solveur de Cadbiom a besoin d'environ 3 à 7 jours de calcul pour générer l'ensemble des trajectoires qui influencent un gène ;
2. La recherche des *clusters* : pour plus de 6000 objets, l'algorithme GreedyRSC a besoin d'environ 30 heures de calcul pour regrouper ces objets dans différents *clusters*.

Au cours de cette thèse, j'ai été confronté à la difficulté d'homogénéisation des données de signalisation dans les différentes bases de données. La migration de la base de données PID vers *Pathway Commons* et sa transformation en BioPAX m'a conduit à mener une réflexion sur ce langage, activement développé par toutes les bases de données.

Après une étude approfondie du format BioPAX et des bases de données le proposant, j'ai été capable de traduire des données BioPAX en Cadbiom. Ce qui a permis de comparer les bases de données entre elles. Les résultats montrent que la topologie de ces bases (taille, densité, etc.) est très variable. L'objectif final serait de pouvoir intégrer toutes ses bases dans un unique modèle.

Cette stratégie d'étude exhaustive des voies de signalisation n'est pas spécifique au TGF-β. Il serait intéressant de pouvoir généraliser cette approche à d'autres stimuli de signalisation et peut-être découvrir de nouvelles familles de trajectoires non décrites dans les bases de données, mais émergentes de la fusion des *pathways* en un seul modèle.

Table des figures

Bibliographie

[Aastha Joshi and Rajneet Kaur, 2013] Aastha Joshi and Rajneet Kaur (2013). A review : Comparative study of various clustering techniques in data mining. International Journal of Advanced Research in Computer Science and Software Engineering, 3(3).

[Aittokallio and Schwikowski, 2006] Aittokallio, T. and Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. Briefings in Bioinformatics, 7(3) :243–255.

[Alon, 2007] Alon, U. (2007). Network motifs : theory and experimental approaches. Nature Reviews Genetics, 8(6) :450–461.

[Andrieux et al., 2014] Andrieux, G., Le Borgne, M., and Théret, N. (2014). An integrative modeling framework reveals plasticity of TGF-β signaling. BMC Systems Biology, 8 :30.

[Ankerst et al., 1999] Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). OPTICS : Ordering Points to Identify the Clustering Structure. In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, SIGMOD '99, pages 49–60, New York, NY, USA. ACM.

[Antezana et al., 2009] Antezana, E., Kuiper, M., and Mironov, V. (2009). Biological knowledge management : the emerging role of the Semantic Web technologies. Briefings in Bioinformatics, 10(4) :392–407.

[Antoine B. Rauzy, 2008] Antoine B. Rauzy (2008). Guarded transition systems : A new states/events formalism for reliability studies. Proceedings of the Institution of Mechanical Engineers, Part O : Journal of Risk and Reliability, 222(4) :495–505.

[Antoniou and Harmelen, 2004] Antoniou, G. and Harmelen, F. v. (2004). Web Ontology Language : OWL. In Handbook on Ontologies, International Handbooks on Information Systems, pages 67–92. Springer, Berlin, Heidelberg. DOI : 10.1007/978-3-540-24750-0_4.

[Aravind Subramanian et al., 2005] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov (2005). Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43) :15545–15550.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology : tool for the

unification of biology. The Gene Ontology Consortium. Nature Genetics, 25(1) :25–29.

[Assieh Saadatpour et al., 2013] Assieh Saadatpour, Réka Albert, and Timothy C. Reluga (2013). A Reduction Method for Boolean Network Models Proven to Conserve Attractors. SIAM Journal on Applied Dynamical Systems, 12(4) :1997–2011.

[Babur et al., 2014] Babur, O., Aksoy, B. A., Rodchenkov, I., Sümer, S. O., Sander, C., and Demir, E. (2014). Pattern search in BioPAX models. Bioinformatics, 30(1) :139–140.

[Bader et al., 2006] Bader, G. D., Cary, M. P., and Sander, C. (2006). Pathguide : a pathway resource list. Nucleic Acids Research, 34(Database issue) :D504–506.

[Bandyopadhyay and Mallick, 2014] Bandyopadhyay, S. and Mallick, K. (2014). A New Path Based Hybrid Measure for Gene Ontology Similarity. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 11(1) :116–127.

[Bard and Rhee, 2004] Bard, J. B. L. and Rhee, S. Y. (2004). Ontologies in biology : design, applications and future challenges. Nature Reviews Genetics, 5(3) :213–222.

[Barua et al., 2009] Barua, D., Faeder, J. R., and Haugh, J. M. (2009). A Bipolar Clamp Mechanism for Activation of Jak-Family Protein Tyrosine Kinases. PLOS Computational Biology, 5(4) :e1000364.

[Besaw, 2013] Besaw, M. E. (2013). Protein Lounge. Journal of the Medical Library Association : JMLA, 101(2) :164.

[Bierie and Moses, 2006] Bierie, B. and Moses, H. L. (2006). Tumour microenvironment : TGFβ : the molecular Jekyll and Hyde of cancer. Nature Reviews Cancer, 6(7) :506–520.

[Binns et al., 2009] Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., and Apweiler, R. (2009). QuickGO : a web-based tool for Gene Ontology searching. Bioinformatics, 25(22) :3045–3046.

[Blachon et al., 2007] Blachon, S., Pensa, R. G., Besson, J., Robardet, C., Boulicaut, J.-F., and Gandrillon, O. (2007). Clustering Formal Concepts to Discover Biologically Relevant Knowledge from Gene Expression Data. In Silico Biology, 7(4,5) :467–483.

[Bossy et al., 2015] Bossy, R., Golik, W., Ratkovic, Z., Valsamou, D., Bessières, P., and Nédellec, C. (2015). Overview of the gene regulation network and the bacteria biotope tasks in BioNLP'13 shared task. BMC Bioinformatics, 16(10) :S1.

[Brandman and Meyer, 2008] Brandman, O. and Meyer, T. (2008). Feedback Loops Shape Cellular Signals in Space and Time. Science, 322(5900) :390–395.

[Bree B. Aldridge et al., 2006] Bree B. Aldridge, John M. Burke, Douglas A. Lauffenburger, and Peter K. Sorger (2006). Physicochemical modelling of cell signalling pathways. Nature Cell Biology, 8(11) :1195–1203.

[Breuer et al., 2013] Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., Winsor, G. L., Hancock, R. E. W., Brinkman, F. S. L., and Lynn, D. J. (2013). InnateDB : systems biology of innate immunity and beyond—recent updates and continuing curation. Nucleic Acids Research, 41(D1) :D1228–D1233.

[Carl F. Schaefer et al., 2009] Carl F. Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H. Buetow (2009). PID : the Pathway Interaction Database. Nucleic Acids Research, 37(suppl_1) :D674–D679.

[Cerami et al., 2011] Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. Nucleic Acids Research, 39(suppl_-1) :D685–D690.

[Chabrier-Rivier et al., 2004] Chabrier-Rivier, N., Fages, F., and Soliman, S. (2004). The Biochemical Abstract Machine BIOCHAM. In Computational Methods in Systems Biology, Lecture Notes in Computer Science, pages 172–191. Springer, Berlin, Heidelberg.

[Chaouiya et al., 2013] Chaouiya, C., Bérenguier, D., Keating, S. M., Naldi, A., Van, M. I., Rodriguez, N., Dräger, A., Büchel, F., Cokelaer, T., Kowal, B., Wicks, B., Gonçalves, E., Dorier, J., Page, M., Monteiro, P. T., Von, A. K., Xenarios, I., De, H. J., Hucka, M., Klamt, S., Thieffry, D., Le, N. N., Saez-Rodriguez, J., and Helikar, T. (2013). SBML qualitative models : a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools., SBML qualitative models : a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. BMC systems biology, BMC Systems Biology, 7, 7 :135, 135–135.

[Chassagnole et al., 2002] Chassagnole, C., Noisommit-Rizzi, N., Schmid, J. W., Mauch, K., and Reuss, M. (2002). Dynamic modeling of the central carbon metabolism of Escherichia coli. Biotechnology and Bioengineering, 79(1) :53–73.

[Chelliah et al., 2015] Chelliah, V., Juty, N., Ajmera, I., Ali, R., Dumousseau, M., Glont, M., Hucka, M., Jalowicki, G., Keating, S., Knight-Schrijver, V., Lloret-Villas, A., Natarajan, K. N., Pettit, J.-B., Rodriguez, N., Schubert, M., Wimalaratne, S. M., Zhao, Y., Hermjakob, H., Le Novère, N., and Laibe, C. (2015). BioModels : ten-year anniversary. Nucleic Acids Research, 43(D1) :D542–D548.

[Chen et al., 2009] Chen, W. W., Schoeberl, B., Jasper, P. J., Niepel, M., Nielsen, U. B., Lauffenburger, D. A., and Sorger, P. K. (2009). Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data., Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. Molecular Systems Biology, Molecular Systems Biology, 5, 5 :239, 239–239.

[Chowdhury and Sarkar, 2015] Chowdhury, S. and Sarkar, R. R. (2015). Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges. Database, 2015.

[Chylek et al., 2011] Chylek, L. A., Hu, B., Blinov, M. L., Emonet, T., Faeder, J. R., Goldstein, B., Gutenkunst, R. N., Haugh, J. M., Lipniacki, T., Posner, R. G., Yang, J., and Hlavacek, W. S. (2011). Guidelines for visualizing and annotating rule-based models. Molecular BioSystems, 7(10) :2779–2795.

[Consortium, 2017] Consortium, T. G. O. (2017). Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Research, 45(D1) :D331–D338.

[Coquet et al., 2017] Coquet, J., Theret, N., Legagneux, V., and Dameron, O. (2017). Identifying Functional Families of Trajectories in Biological Pathways by Soft Clustering : Application to TGF-β Signaling. In Computational Methods in Systems Biology, Lecture Notes in Computer Science, pages 91–107. Springer, Cham.

[Croft et al., 2014] Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., and D'Eustachio, P. (2014). The Reactome pathway knowledgebase. Nucleic Acids Research, 42(Database issue) :D472–477.

[Danos et al., 2008] Danos, V., Feret, J., Fontana, W., Harmer, R., and Krivine, J. (2008). Rule-Based Modelling, Symmetries, Refinements. In Proceedings of the 1st International Workshop on Formal Methods in Systems Biology, FMSB '08, pages 103–122, Berlin, Heidelberg. Springer-Verlag.

[Demir et al., 2010] Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., Schacherer, F., Martinez-Flores, I., Hu, Z., Jimenez-Jacinto, V., Joshi-Tope, G., Kandasamy, K., Lopez-Fuentes, A. C., Mi, H., Pichler, E., Rodchenkov, I., Splendiani, A., Tkachev, S., Zucker, J., Gopinath, G., Rajasimha, H., Ramakrishnan, R., Shah, I., Syed, M., Anwar, N., Babur, O., Blinov, M., Brauner, E., Corwin, D., Donaldson, S., Gibbons, F., Goldberg, R., Hornbeck, P., Luna, A., Murray-Rust, P., Neumann, E., Ruebenacker, O., Samwald, M., van Iersel, M., Wimalaratne, S., Allen, K., Braun, B., Whirl-Carrillo, M., Cheung, K.-H., Dahlquist, K., Finney, A., Gillespie, M., Glass, E., Gong, L., Haw, R., Honig, M., Hubaut, O., Kane, D., Krupa, S., Kutmon, M., Leonard, J., Marks, D., Merberg, D., Petri, V., Pico, A., Ravenscroft, D., Ren, L., Shah, N., Sunshine, M., Tang, R., Whaley, R., Letovksy, S., Buetow, K. H., Rzhetsky, A., Schachter, V., Sobral, B. S., Dogrusoz, U., McWeeney, S., Aladjem, M., Birney, E., Collado-Vides, J., Goto, S., Hucka, M., Le Novère, N., Maltsev, N., Pandey, A., Thomas, P., Wingender, E., Karp, P. D., Sander, C., and Bader, G. D. (2010). The BioPAX community standard for pathway data sharing. Nature Biotechnology, 28(9) :935–942.

[Didier and Remy, 2012] Didier, G. and Remy, E. (2012). Relations between gene regulatory networks and cell dynamics in Boolean models. Discrete Applied Mathematics, 160(15) :2147–2157.

[Drineas et al., 2004] Drineas, P., Frieze, A., Kannan, R., Vempala, S., and Vinay, V. (2004). Clustering Large Graphs via the Singular Value Decomposition. Machine Learning, 56(1-3) :9–33.

[Duarte et al., 2007] Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., and Palsson, B. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences, 104(6) :1777–1782.

[Dunn et al., 2005] Dunn, R., Dudbridge, F., and Sanderson, C. M. (2005). The Use of Edge-Betweenness Clustering to Investigate Biological Function in Protein Interaction Networks. BMC Bioinformatics, 6 :39.

[Eilbeck et al., 2005] Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005). The Sequence Ontology : a tool for the unification of genome annotations. Genome Biology, 6 :R44.

[Elkon et al., 2008] Elkon, R., Vesterman, R., Amit, N., Ulitsky, I., Zohar, I., Weisz, M., Mass, G., Orlev, N., Sternberg, G., Blekhman, R., Assa, J., Shiloh, Y., and Shamir, R. (2008). SPIKE – a database, visualization and analysis tool of cellular signaling pathways. BMC Bioinformatics, 9 :110.

[Ertöz et al., 2003] Ertöz, L., Steinbach, M., and Kumar, V. (2003). Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In Proceedings of the 2003 SIAM International Conference on Data Mining, Proceedings, pages 47–58. Society for Industrial and Applied Mathematics. DOI : 10.1137/1.9781611972733.5.

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, pages 226–231, Portland, Oregon. AAAI Press.

[Fabregat et al., 2016] Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamovsky, V., Webber, M., Weiser, J., Williams, M., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (2016). The Reactome pathway Knowledgebase. Nucleic Acids Research, 44(D1) :D481–487.

[Fazekas et al., 2013] Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálfy, M., Dúl, Z., Zsákai, L., Szalay-Bekő, M., Lenti, K., Farkas, I. J., Vellai, T., Csermely, P., and Korcsmáros, T. (2013). SignaLink 2 – a signaling pathway resource with multi-layered regulatory networks. BMC Systems Biology, 7 :7.

[Fearnley et al., 2014] Fearnley, L. G., Davis, M. J., Ragan, M. A., and Nielsen, L. K. (2014). Extracting reaction networks from databases–opening Pandora's box. Briefings in Bioinformatics, 15(6) :973–983.

[Fran Supek et al., 2011] Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc (2011). REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLOS ONE, 6(7) :e21800.

[Furusawa and Kaneko, 2012] Furusawa, C. and Kaneko, K. (2012). A Dynamical-Systems View of Stem Cell Biology. Science, 338(6104) :215–217.

[Galperin et al., 2017] Galperin, M. Y., Fernández-Suárez, X. M., and Rigden, D. J. (2017). The 24th annual Nucleic Acids Research database issue : a look back and upcoming changes. Nucleic Acids Research, 45(D1) :D1–D11.

[Gerstmann, 2002] Gerstmann, S. (2002). Signaling PAthway Database (SPAD) ?ms an upcomingonline database on signal transduction. Signal Transduction, 2(1-2) :49–53.

[Grzegorczyk et al., 2008] Grzegorczyk, M., Husmeier, D., Edwards, K. D., Ghazal, P., and Millar, A. J. (2008). Modelling non-stationary gene regulatory processes with

a non-homogeneous Bayesian network and the allocation sampler. Bioinformatics, 24(18) :2071–2078.

[Guha et al., 1998] Guha, S., Rastogi, R., and Shim, K. (1998). CURE : An Efficient Clustering Algorithm for Large Databases. In Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, SIGMOD '98, pages 73–84, New York, NY, USA. ACM.

[Guha et al., 1999] Guha, S., Rastogi, R., and Shim, K. (1999). ROCK : a robust clustering algorithm for categorical attributes. In Proceedings 15th International Conference on Data Engineering (Cat. No.99CB36337), pages 512–521.

[Gupta et al., 2007] Gupta, S., Bisht, S. S., Kukreti, R., Jain, S., and Brahmachari, S. K. (2007). Boolean network analysis of a neurotransmitter signaling pathway. Journal of Theoretical Biology, 244(3) :463–469.

[Hackl et al., 2004] Hackl, H., Maurer, M., Mlecnik, B., Hartler, J., Stocker, G., Miranda-Saavedra, D., and Trajanoski, Z. (2004). GOLD.db : genomics of lipid-associated disorders database. BMC Genomics, 5(1) :93.

[Hamzaoui et al., 2011] Hamzaoui, A., Joly, A., and Boujemaa, N. (2011). Multi-source shared nearest neighbours for multi-modal image clustering. Multimedia Tools and Applications, 51(2) :479–503.

[Hans A. Kestler et al., 2008] Hans A. Kestler, Christian Wawra, Barbara Kracher, and Michael Kühl (2008). Network modeling of signal transduction : establishing the global view. BioEssays, 30(11-12) :1110–1125.

[Hermjakob et al., 2004] Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., Mering, C. v., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G. N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., and Apweiler, R. (2004). The HUPO PSI's Molecular Interaction format–a community standard for the representation of protein interaction data. Nature Biotechnology, 22(2) :177–184.

[Hinneburg and Gabriel, 2007] Hinneburg, A. and Gabriel, H.-H. (2007). DENCLUE 2.0 : Fast Clustering Based on Kernel Density Estimation. In Advances in Intelligent Data Analysis VII, Lecture Notes in Computer Science, pages 70–80. Springer, Berlin, Heidelberg.

[Hiroaki Ikushima and Kohei Miyazono, 2011] Hiroaki Ikushima and Kohei Miyazono (2011). Biology of Transforming Growth Factor-βSignaling. Current Pharmaceutical Biotechnology, 12(12) :2099–2107.

[Hofman and Jarvis, 1998] Hofman, I. and Jarvis, R. (1998). Robust and efficient cluster analysis using a shared near neighbours approach. In Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170), volume 1, pages 243–247 vol.1.

[Houle, 2008] Houle, M. (2008). The Relevant-Set Correlation Model for Data Clustering. In Proceedings of the 2008 SIAM International Conference on Data Mining,

Proceedings, pages 775–786. Society for Industrial and Applied Mathematics. DOI : 10.1137/1.9781611972788.70.

[J. Craig Venter et al., 2001] J. Craig Venter, Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The Sequence of the Human Genome. Science, 291(5507) :1304–1351.

[Jain, 2010] Jain, A. K. (2010). Data clustering : 50 years beyond K-means. Pattern

Recognition Letters, 31(8) :651–666.

[Jeong et al., 2001] Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. Nature, 411(6833) :41–42.

[Joan Massagué, 2012] Joan Massagué (2012). TGFβ signalling in context. Nature Reviews Molecular Cell Biology, 13(10) :616–630.

[John D. Scott and Tony Pawson, 2009] John D. Scott and Tony Pawson (2009). Cell Signaling in Space and Time : Where Proteins Come Together and When They're Apart. Science, 326(5957) :1220–1224.

[Jorge G. T. Zañudo and Réka Albert, 2013] Jorge G. T. Zañudo and Réka Albert (2013). An effective network reduction approach to find the dynamical repertoire of discrete dynamic networks. Chaos : An Interdisciplinary Journal of Nonlinear Science, 23(2) :025111.

[Kandasamy et al., 2010] Kandasamy, K., Mohan, S. S., Raju, R., Keerthikumar, S., Kumar, G. S. S., Venugopal, A. K., Telikicherla, D., Navarro, J. D., Mathivanan, S., Pecquet, C., Gollapudi, S. K., Tattikota, S. G., Mohan, S., Padhukasahasram, H., Subbannayya, Y., Goel, R., Jacob, H. K. C., Zhong, J., Sekhar, R., Nanjappa, V., Balakrishnan, L., Subbaiah, R., Ramachandra, Y. L., Rahiman, B. A., Prasad, T. S. K., Lin, J.-X., Houtman, J. C. D., Desiderio, S., Renauld, J.-C., Constantinescu, S. N., Ohara, O., Hirano, T., Kubo, M., Singh, S., Khatri, P., Draghici, S., Bader, G. D., Sander, C., Leonard, W. J., and Pandey, A. (2010). NetPath : a public resource of curated signal transduction pathways. Genome Biology, 11(1) :R3.

[Kanehisa and Goto, 2000] Kanehisa, M. and Goto, S. (2000). KEGG : Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 28(1) :27–30.

[Kanungo et al., 2002] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002). An efficient k-means clustering algorithm : analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7) :881–892.

[Karp and Riley, 1993] Karp, P. D. and Riley, M. (1993). Representations of Metabolic Knowledge. In Proceedings of the First International Conference on Intelligent Systems for Molecular Biology, pages 207–215. AAAI Press.

[Kauffman, 1969] Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology, 22(3) :437–467.

[Kaufman and Rousseeuw, 1990a] Kaufman, L. and Rousseeuw, P. J. (1990a). Clustering Large Applications (Program CLARA). In Finding Groups in Data, pages 126–163. John Wiley & Sons, Inc. DOI : 10.1002/9780470316801.ch3.

[Kaufman and Rousseeuw, 1990b] Kaufman, L. and Rousseeuw, P. J. (1990b). Finding Groups in Data : An Introduction to Cluster Analysis. John Wiley & Sons. Google-Books-ID : YeFQHiikNo0C.

[Kholodenko, 2006] Kholodenko, B. N. (2006). Cell-signalling dynamics in time and space. Nature reviews. Molecular cell biology, 7(3) :165–176.

[Kitano, 2002] Kitano, H. (2002). Computational systems biology. Nature, 420(6912) :206–210.

[Koschützki and Schreiber, 2008] Koschützki, D. and Schreiber, F. (2008). Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks. Gene Regulation and Systems Biology, 2 :193–201.

[Kotsiantis, 2007] Kotsiantis, S. B. (2007). Supervised Machine Learning : A Review of Classification Techniques. In Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering : Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, pages 3–24, Amsterdam, The Netherlands, The Netherlands. IOS Press.

[Kunxin Luo, 2017] Kunxin Luo (2017). Signaling Cross Talk between TGF-β/Smad and Other Signaling Pathways. Cold Spring Harbor Perspectives in Biology, 9(1) :a022137.

[Kutmon et al., 2016] Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E. L., Bohler, A., Mélius, J., Waagmeester, A., Sinha, S. R., Miller, R., Coort, S. L., Cirillo, E., Smeets, B., Evelo, C. T., and Pico, A. R. (2016). WikiPathways : capturing the full diversity of pathway knowledge. Nucleic Acids Research, 44(D1) :D488–D494.

[Le Novère, 2015] Le Novère, N. (2015). Quantitative and logic modelling of molecular and gene networks. Nature Reviews Genetics, 16(3) :146–158.

[Liao et al., 2012] Liao, S.-H., Chu, P.-H., and Hsiao, P.-Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. Expert Systems with Applications, 39(12) :11303–11311.

[Machado et al., 2011] Machado, D., Costa, R. S., Rocha, M., Ferreira, E. C., Tidor, B., and Rocha, I. (2011). Modeling formalisms in Systems Biology. AMB Express, 1 :45.

[Maiwald et al., 2012a] Maiwald, T., Blumberg, J., Raue, A., Hengl, S., Schilling, M., Sy, S. K., Becker, V., Klingmüller, U., and Timmer, J. (2012a). In silico labeling reveals the time-dependent label half-life and transit-time in dynamical systems. BMC Systems Biology, 6 :13.

[Maiwald et al., 2012b] Maiwald, T., Eberhardt, O., and Blumberg, J. (2012b). Mathematical Modeling of Biochemical Systems with PottersWheel. In Computational Modeling of Signaling Networks, Methods in Molecular Biology, pages 119–138. Humana Press, Totowa, NJ. DOI : 10.1007/978-1-61779-833-7_8.

[Maiwald and Timmer, 2008] Maiwald, T. and Timmer, J. (2008). Dynamical modeling and multi-experiment fitting with PottersWheel. Bioinformatics, 24(18) :2037–2043.

[Maozhen Tian et al., 2011] Maozhen Tian, Jason R. Neil, and William P. Schiemann (2011). Transforming growth factor-β and the hallmarks of cancer. Cellular Signalling, 23(6) :951–962.

[Markevich et al., 2004] Markevich, N. I., Hoek, J. B., and Kholodenko, B. N. (2004). Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. The Journal of Cell Biology, 164(3) :353–359.

[McBride, 2004] McBride, B. (2004). The Resource Description Framework (RDF) and its Vocabulary Description Language RDFS. In Handbook on Ontologies, International Handbooks on Information Systems, pages 51–65. Springer, Berlin, Heidelberg. DOI : 10.1007/978-3-540-24750-0_3.

[Mi et al., 2013] Mi, H., Muruganujan, A., Casagrande, J. T., and Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. Nature Protocols, 8(8) :1551–1566.

[Mi and Thomas, 2009] Mi, H. and Thomas, P. (2009). PANTHER Pathway : An Ontology-Based Pathway Database Coupled with Data Analysis Tools. In Protein Networks and Pathway Analysis, Methods in Molecular Biology, pages 123–140. Humana Press.

[Milo et al., 2004] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of Evolved and Designed Networks. Science, 303(5663) :1538–1542.

[Milo et al., 2002] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network Motifs : Simple Building Blocks of Complex Networks. Science, 298(5594) :824–827.

[Nadav Kashtan and Uri Alon, 2005] Nadav Kashtan and Uri Alon (2005). Spontaneous evolution of modularity and network motifs. Proceedings of the National Academy of Sciences of the United States of America, 102(39) :13773–13778.

[Nagar and Al-Mubaid, 2008] Nagar, A. and Al-Mubaid, H. (2008). A New Path Length Measure Based on GO for Gene Similarity with Evaluation using SGD Pathways. In 2008 21st IEEE International Symposium on Computer-Based Medical Systems, pages 590–595.

[Naldi et al., 2017] Naldi, A., Larive, R. M., Czerwinska, U., Urbach, S., Montcourrier, P., Roy, C., Solassol, J., Freiss, G., Coopman, P. J., and Radulescu, O. (2017). Reconstruction and signal propagation analysis of the Syk signaling network in breast cancer cells. PLoS Computational Biology, 13(3).

[Nesma ElKalaawy and Amr Wassal, 2015] Nesma ElKalaawy and Amr Wassal (2015). Methodologies for the modeling and simulation of biochemical networks, illustrated for signal transduction pathways : A primer. Biosystems, 129 :1–18.

[Ng and Han, 2002] Ng, R. T. and Han, J. (2002). CLARANS : a method for clustering objects for spatial data mining. IEEE Transactions on Knowledge and Data Engineering, 14(5) :1003–1016.

[Nishimura, 2001] Nishimura, D. (2001). BioCarta. Biotech Software & Internet Report, 2(3) :117–120.

[Novère et al., 2009] Novère, N. L., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., Bergman, F. T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S. E., Calzone, L., Courtot, M., Dogrusoz, U., Freeman, T. C., Funahashi, A., Ghosh, S., Jouraku, A., Kim, S., Kolpakov, F., Luna, A., Sahle, S., Schmidt, E., Watterson, S.,

Wu, G., Goryanin, I., Kell, D. B., Sander, C., Sauro, H., Snoep, J. L., Kohn, K., and Kitano, H. (2009). The Systems Biology Graphical Notation. Nature Biotechnology, 27(8) :735–741.

[Oltvai and Barabási, 2002] Oltvai, Z. N. and Barabási, A.-L. (2002). Life's Complexity Pyramid. Science, 298(5594) :763–764.

[Orth et al., 2011] Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., and Palsson, B. (2011). A comprehensive genome scale reconstruction of Escherichia coli metabolism. Molecular Systems Biology, 7(1) :535.

[Palsson, 2002] Palsson, B. (2002). In silico biology through "omics". Nature Biotechnology, 20(7) :649–650.

[Papin et al., 2003] Papin, J. A., Price, N. D., Wiback, S. J., Fell, D. A., and Palsson, B. O. (2003). Metabolic pathways in the post-genome era. Trends in Biochemical Sciences, 28(5) :250–258.

[Pearl, 1988] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference. Morgan Kaufmann.

[Pesquita et al., 2009] Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. PLOS Computational Biology, 5(7) :e1000443.

[Poelmans et al., 2013] Poelmans, J., Ignatov, D. I., Kuznetsov, S. O., and Dedene, G. (2013). Formal concept analysis in knowledge processing : A survey on applications. Expert Systems with Applications, 40(16) :6538–6560.

[Popat and Emmanuel, 2014] Popat, S. K. and Emmanuel, M. (2014). Review and comparative study of clustering techniques. International journal of computer science and information technologies, 5(1) :805–812.

[Pratt et al., 2015] Pratt, D., Chen, J., Welker, D., Rivas, R., Pillich, R., Rynkov, V., Ono, K., Miello, C., Hicks, L., Szalma, S., Stojmirovic, A., Dobrin, R., Braxenthaler, M., Kuentzer, J., Demchak, B., and Ideker, T. (2015). NDEx, the Network Data Exchange. Cell Systems, 1(4) :302–305.

[Prud'hommeaux and Seaborne, 2008] Prud'hommeaux, E. and Seaborne, A. (2008). SPARQL Query Language for RDF. Technical report.

[Pujol et al., 2010] Pujol, A., Mosca, R., Farrés, J., and Aloy, P. (2010). Unveiling the role of network and systems biology in drug discovery. Trends in Pharmacological Sciences, 31(3) :115–123.

[Reed and Palsson, 2003] Reed, J. L. and Palsson, B. (2003). Thirteen Years of Building Constraint-Based In Silico Models of Escherichia coli. Journal of Bacteriology, 185(9) :2692–2699.

[Remy et al., 2015] Remy, E., Rebouissou, S., Chaouiya, C., Zinovyev, A., Radvanyi, F., and Calzone, L. (2015). A Modeling Approach to Explain Mutually Exclusive and Co-Occurring Genetic Alterations in Bladder Tumorigenesis. Cancer Research, 75(19) :4042–4052.

[Richard and Comet, 2007] Richard, A. and Comet, J.-P. (2007). Necessary conditions for multistationarity in discrete dynamical systems. Discrete Applied Mathematics, 155(18) :2403–2413.

[Ruebenacker et al., 2007] Ruebenacker, O., Moraru, I. I., Schaff, J. C., and Blinov, M. L. (2007). Kinetic Modeling Using BioPAX Ontology. In 2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007), pages 339–348.

[Ryall et al., 2012] Ryall, K. A., Holland, D. O., Delaney, K. A., Kraeutler, M. J., Parker, A. J., and Saucerman, J. J. (2012). Network Reconstruction and Systems Analysis of Cardiac Myocyte Hypertrophy Signaling. Journal of Biological Chemistry, 287(50) :42259–42268.

[Saarinen et al., 2008] Saarinen, A., Linne, M.-L., and Yli-Harja, O. (2008). Stochastic Differential Equation Model for Cerebellar Granule Cell Excitability. PLOS Computational Biology, 4(2) :e1000004.

[Sachs et al., 2005] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. Science, 308(5721) :523–529.

[Saez-Rodriguez et al., 2007] Saez-Rodriguez, J., Simeoni, L., Lindquist, J. A., Hemenway, R., Bommhardt, U., Arndt, B., Haus, U.-U., Weismantel, R., Gilles, E. D., Klamt, S., and Schraven, B. (2007). A Logical Model Provides Insights into T Cell Receptor Signaling. PLOS Computational Biology, 3(8) :e163.

[Samaga and Klamt, 2013] Samaga, R. and Klamt, S. (2013). Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. Cell Communication and Signaling, 11 :43.

[Santhisree and Damodaram, 2011] Santhisree, K. and Damodaram, A. (2011). CLIQUE : Clustering based on density on web usage data : Experiments and test results. In 2011 3rd International Conference on Electronics Computer Technology, volume 4, pages 233–236.

[Santoso and Nisa, 2016] Santoso, A. and Nisa, K. K. (2016). Cloud Computing Application for Hotspot Clustering Using Recursive Density Based Clustering (RDBC). IOP Conference Series : Earth and Environmental Science, 31(1) :012004.

[Schilling et al., 2008] Schilling, M., Pfeifer, A. C., Bohl, S., and Klingmuller, U. (2008). Standardizing experimental protocols. Current Opinion in Biotechnology, 19(4) :354–359.

[Sergio G. Peisajovich et al., 2010] Sergio G. Peisajovich, Joan E. Garbarino, Ping Wei, and Wendell A. Lim (2010). Rapid Diversification of Cell Signaling Phenotypes by Modular Domain Recombination. Science, 328(5976) :368–372.

[Sheikholeslami et al., 2000] Sheikholeslami, G., Chatterjee, S., and Zhang, A. (2000). WaveCluster : A Wavelet-based Clustering Approach for Spatial Data in Very Large Databases. The VLDB Journal, 8(3-4) :289–304.

[Shmulevich et al., 2002] Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W. (2002). Probabilistic Boolean networks : a rule-based uncertainty model for gene regulatory networks. Bioinformatics, 18(2) :261–274.

[Sibson, 1973] Sibson, R. (1973). SLINK : An optimally efficient algorithm for the single-link cluster method. The Computer Journal, 16(1) :30–34.

[Sivakumaran et al., 2003] Sivakumaran, S., Hariharaputran, S., Mishra, J., and Bhalla, U. S. (2003). The Database of Quantitative Cellular Signaling : management and analysis of chemical kinetic models of signaling networks. Bioinformatics, 19(3) :408–415.

[Smallbone and Mendes, 2013] Smallbone, K. and Mendes, P. (2013). Large-Scale Metabolic Models : From Reconstruction to Differential Equations. Industrial Biotechnology, 9(4) :179–184.

[Smith et al., 2012] Smith, A. M., Xu, W., Sun, Y., Faeder, J. R., and Marai, G. E. (2012). RuleBender : integrated modeling, simulation and visualization for rule-based intracellular biochemistry. BMC Bioinformatics, 13(8) :S3.

[Snowden et al., 2017] Snowden, T. J., Graaf, P. H. v. d., and Tindall, M. J. (2017). Methods of Model Reduction for Large-Scale Biological Systems : A Survey of Current Methods and Trends. Bulletin of Mathematical Biology, 79(7) :1449–1486.

[Soh et al., 2010] Soh, D., Dong, D., Guo, Y., and Wong, L. (2010). Consistency, comprehensiveness, and compatibility of pathway databases. BMC bioinformatics, 11 :449.

[Steinway et al., 2014] Steinway, S. N., Zañudo, J. G. T., Ding, W., Rountree, C. B., Feith, D. J., Loughran, T. P., and Albert, R. (2014). Network modeling of TGFβ signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint Sonic hedgehog and Wnt pathway activation. Cancer research, 74(21) :5963–5977.

[Stromback and Lambrix, 2005] Stromback, L. and Lambrix, P. (2005). Representations of molecular pathways : an evaluation of SBML, PSI MI and BioPAX. Bioinformatics, 21(24) :4401–4407.

[Thiele et al., 2013] Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., Thorleifsson, S. G., Agren, R., Bölling, C., Bordel, S., Chavali, A. K., Dobson, P., Dunn, W. B., Endler, L., Hala, D., Hucka, M., Hull, D., Jameson, D., Jamshidi, N., Jonsson, J. J., Juty, N., Keating, S., Nookaew, I., Le Novère, N., Malys, N., Mazein, A., Papin, J. A., Price, N. D., Selkov Sr, E., Sigurdsson, M. I., Simeonidis, E., Sonnenschein, N., Smallbone, K., Sorokin, A., van Beek, J. H. G. M., Weichart, D., Goryanin, I., Nielsen, J., Westerhoff, H. V., Kell, D. B., Mendes, P., and Palsson, B. (2013). A community-driven global reconstruction of human metabolism. Nature Biotechnology, 31(5) :419–425.

[Traynard et al., 2016] Traynard, P., Fauré, A., Fages, F., and Thieffry, D. (2016). Logical model specification aided by model-checking techniques : application to the mammalian cell cycle regulation. Bioinformatics, 32(17) :i772–i780.

[Tyson et al., 2003] Tyson, J. J., Chen, K. C., and Novak, B. (2003). Sniffers, buzzers, toggles and blinkers : dynamics of regulatory and signaling pathways in the cell. Current Opinion in Cell Biology, 15(2) :221–231.

[Vidal et al., 2011] Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome Networks and Human Disease. Cell, 144(6) :986–998.

[Videla et al., 2015] Videla, S., Guziolowski, C., Eduati, F., Thiele, S., Gebser, M., Nicolas, J., Saez-Rodriguez, J., Schaub, T., and Siegel, A. (2015). Learning Boolean logic models of signaling networks with ASP. Theoretical Computer Science, 599(Supplement C) :79–101.

[Villa et al., 2009] Villa, F., Athanasiadis, I. N., and Rizzoli, A. E. (2009). Modelling with knowledge : A review of emerging semantic approaches to environmental modelling. Environmental Modelling & Software, 24(5) :577–587.

[Wang et al., 2007] Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics, 23(10) :1274–1281.

[Wang et al., 1997] Wang, W., Yang, J., and Muntz, R. R. (1997). STING : A Statistical Information Grid Approach to Spatial Data Mining. In Proceedings of the 23rd International Conference on Very Large Data Bases, VLDB '97, pages 186–195, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Wendell A. Lim, 2010] Wendell A. Lim (2010). Designing customized cell signalling circuits. Nature Reviews Molecular Cell Biology, 11(6) :393–403.

[Whetzel et al., 2011] Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., and Musen, M. A. (2011). BioPortal : enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Research, 39(suppl_-2) :W541–W545.

[Wilkinson, 2009] Wilkinson, D. J. (2009). Stochastic modelling for quantitative description of heterogeneous biological systems. Nature Reviews Genetics, 10(2) :122–133.

[Wille, 1982] Wille, R. (1982). Restructuring Lattice Theory : An Approach Based on Hierarchies of Concepts. In Ordered Sets, NATO Advanced Study Institutes Series, pages 445–470. Springer, Dordrecht. DOI : 10.1007/978-94-009-7798-3_15.

[Yabing Mu et al., 2012] Yabing Mu, Shyam Kumar Gudey, and Maréne Landström (2012). Non-Smad signaling pathways. Cell and Tissue Research, 347(1) :11–20.

[Yamamoto et al., 2011] Yamamoto, S., Sakai, N., Nakamura, H., Fukagawa, H., Fukuda, K., and Takagi, T. (2011). INOH : ontology-based highly structured database of signal transduction pathways. Database : The Journal of Biological Databases and Curation, 2011.

[Yin Zhao et al., 2013] Yin Zhao, Jongrae Kim, and Maurizio Filippone (2013). Aggregation Algorithm Towards Large-Scale Boolean Network Analysis. IEEE Transactions on Automatic Control, 58(8) :1976–1985.

[Yu et al., 2012] Yu, N., Seo, J., Rho, K., Jang, Y., Park, J., Kim, W. K., and Lee, S. (2012). hiPathDB : a human-integrated pathway database with facile visualization. Nucleic Acids Research, 40(D1) :D797–D802.

[Zhang et al., 2012] Zhang, L. S., Yang, M. J., and Lei, D. J. (2012). An Improved PAM Clustering Algorithm Based on Initial Clustering Centers. Applied Mechanics and Materials, 135-136 :244–249.

[Zhang et al., 1996] Zhang, T., Ramakrishnan, R., and Livny, M. (1996). BIRCH : An Efficient Data Clustering Method for Very Large Databases. In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, SIGMOD '96, pages 103–114, New York, NY, USA. ACM.

[Zhike Zi et al., 2012] Zhike Zi, Douglas A. Chapnick, and Xuedong Liu (2012). Dynamics of TGF-β/Smad signaling. FEBS Letters, 586(14) :1921–1928.

[Zhou et al., 2002] Zhou, X., Kao, M.-C. J., and Wong, W. H. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. Proceedings of the National Academy of Sciences, 99(20) :12783–12788.

Annexes

4.1 Table du niveau de représentation des protéines dans chaque noyau

Name	Uniprot	Noyau 1		Noyau 2		Noyau 3		Noyau 4		Noyau 5	
		Freq	zScore	Freq	zScore	Freq	zScore	Freq	zScore	Freq	zScore
ETS1	P14921	0.0	-5.70	0.03	4.83	0.04	3.88	0.0	-2.10	0.0	-4.36
ARNT	P27540	0.12	15.74	0.00	-7.35	0.00	-5.54	0.0	-3.02	0.04	-0.12
AGTR1	P30556	0.0	-3.80	0.03	10.17	0.0	-2.94	0.0	-1.40	0.0	-2.90
FKBP1A	P62942	0.0	-10.36	0.0	-10.27	0.0	-8.04	0.00	-3.54	0.46	46.66
IL12RB1	P42701	1.0	44.29	0.0	-33.21	0.0	-26.01	0.0	-12.36	0.51	4.73
PLD2	O14939	0.0	-1.86	0.0	-1.84	0.0	-1.44	0.00	2.23	0.00	-0.01
DOK4	Q8TEW6	0.0	-0.86	0.0	-0.85	0.0	-0.66	0.00	2.83	0.00	0.85
IFNG	P01579	0.13	-4.33	0.12	-4.87	0.19	1.78	0.15	-0.74	0.21	3.52
MYC	P01106	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.00	4.00	0.00	5.51
LPAR1	Q92633	0.0	-1.49	0.0	-1.47	0.0	-1.15	0.0	-0.55	0.0	-1.14
ITGAV	P06756	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
HIF3A	Q9Y2N7	0.12	16.48	0.0	-7.93	0.00	-5.37	0.0	-2.95	0.04	0.16
BLNK	Q8WV28	0.0	-5.20	0.01	-1.41	0.12	23.16	0.0	-1.92	0.0	-3.98
MAP2K3	P46734	0.00	-9.15	0.15	16.61	0.0	-7.19	0.00	-2.79	0.0	-7.09
FASLG	P48023	0.02	4.92	0.00	-1.53	0.0	-2.76	0.00	-0.54	0.0	-2.72
FOXA1	P55317	0.0	-1.98	0.00	0.06	0.0	-1.54	0.0	-0.73	0.0	-1.52
DAB2	P98082	0.50	25.22	0.19	-3.26	0.0	-16.47	0.0	-7.83	0.0	-16.25
SHH	Q15465	0.0	-0.86	0.0	-0.85	0.0	-0.66	0.00	2.83	0.00	0.85
NOG	Q13253	0.0	-2.16	0.00	0.18	0.0	-1.68	0.00	0.45	0.0	-1.66
BAD	Q92934	0.0	-9.56	0.08	3.63	0.30	32.07	0.0	-3.52	0.0	-7.31
WNT5A	P41221	0.12	2.57	0.21	14.83	0.0	-10.02	0.06	-1.49	0.04	-5.50
SP1	P08047	0.51	19.01	0.22	-5.10	0.06	-15.15	0.99	22.05	0.08	-13.34
TFDP1	Q14186	0.00	-7.58	0.02	-4.61	0.0	-6.55	0.99	64.48	0.06	2.93
TRADD	Q15628	0.0	-0.49	0.0	-0.49	0.0	-0.38	0.00	5.27	0.0	-0.38
TBX21	Q9UL17	0.00	2.16	0.0	-2.09	0.0	-1.63	0.0	-0.77	0.0	-1.61
NOS2	P35228	0.84	38.43	0.0	-28.78	0.0	-22.53	0.0	-10.71	0.42	4.14
GNG2	P59768	0.08	10.76	0.0	-7.13	0.0	-5.58	0.0	-2.65	0.05	2.93
MAP2K4	P45985	0.53	-11.82	0.85	14.82	0.80	8.07	0.5	-5.33	0.81	8.78
ERBB4	Q15303-1	0.0	-0.86	0.0	-0.85	0.0	-0.66	0.00	2.83	0.00	0.85
PARP14	Q460N5	0.0	-5.35	0.04	7.37	0.05	8.11	0.0	-1.97	0.0	-4.09
VEGFA	P15692-4	0.01	3.61	0.00	-2.29	0.0	-3.02	0.00	-0.73	0.0	-2.98
HGF	P14210	0.0	-1.72	0.0	-1.70	0.0	-1.33	0.0	-0.63	0.0	-1.31
CBL	P22681	0.0	-0.49	0.00	1.53	0.0	-0.38	0.0	-0.18	0.0	-0.38
PTK2	Q05397-1	0.0	-2.53	0.00	1.87	0.01	4.65	0.0	-0.93	0.0	-1.94
FOS	P01100	0.99	16.95	0.70	-12.06	0.97	11.68	0.84	0.70	0.88	4.47
MAP4K1	Q92918	0.0	-2.53	0.00	2.67	0.0	-1.96	0.0	-0.93	0.0	-1.94
IL18R1	Q13478	0.0	-3.93	0.0	-3.89	0.0	-3.05	0.0	-1.45	0.0	-3.00
FZD2	Q14332	0.12	2.57	0.21	14.83	0.0	-10.02	0.06	-1.49	0.04	-5.50
PRL	P01236	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
TFE3	P19532	0.0	-2.43	0.0	-2.41	0.0	-1.89	0.03	8.03	0.02	11.04
PRKCZ	Q05513	0.0	-11.16	0.06	-1.95	0.49	47.01	0.0	-4.11	0.0	-8.54
ELK1	P19419	0.99	27.35	0.41	-19.30	0.95	18.64	0.31	-10.18	0.51	-8.68
MAP2K7	O14733	0.0	-1.57	0.00	-0.27	0.0	-1.21	0.0	-0.57	0.0	-1.20
IL12RB2	Q99665	1.0	44.67	0.0	-32.93	0.0	-25.79	0.0	-12.25	0.51	4.99
FLT4	P35916	0.0	-1.92	0.0	-1.90	0.0	-1.49	0.01	3.52	0.00	-0.11
PF4	P02776	0.0	-3.03	0.01	6.04	0.0	-2.35	0.0	-1.11	0.0	-2.32
BMPR2	Q13873	0.0	-6.20	0.02	0.52	0.0	-4.81	0.03	1.29	0.0	-4.74
IL6ST	P40189	0.15	10.45	0.0	-11.50	0.00	-8.88	0.0	-4.28	0.08	-0.29
SGPP1	Q9BX95	0.0	-0.99	0.00	1.04	0.0	-0.77	0.0	-0.36	0.0	-0.76
TRAF6	Q9Y4K3	0.10	12.09	0.01	-5.38	0.0	-6.41	0.00	-2.36	0.01	-3.68
SPHK2	Q9NRA0	0.00	-2.02	0.00	-2.00	0.0	-1.89	0.0	-0.89	0.0	-1.86

ZBTB17	Q13105	0.0	-0.70	0.0	-0.69	0.0	-0.54	0.00	3.60	0.00	1.32
TGFB3	P10600	0.0	-0.49	0.0	-0.49	0.0	-0.38	0.0	-0.18	0.00	2.24
ABCC1	P33527	0.0	-1.21	0.00	2.11	0.0	-0.94	0.0	-0.44	0.0	-0.93
PRLR	P16471	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
TNF	P01375	0.0	-3.86	0.0	-3.83	0.04	9.11	0.00	-0.71	0.0	-2.96
SDC4	P31431	0.0	-4.77	0.00	-3.87	0.00	-2.06	0.06	6.30	0.02	1.62
NR3C1	P04150	0.0	-3.63	0.02	6.21	0.0	-2.81	0.0	-1.33	0.0	-2.78
MAP2K2	P36507	0.0	-2.22	0.0	-2.20	0.0	-1.72	0.01	4.06	0.00	-1.11
STRAP	Q9Y3F4	0.0	-2.68	0.0	-2.65	0.0	-2.08	0.0	-0.98	0.0	-2.05
ADAM17	P78536	0.0	-0.49	0.0	-0.49	0.0	-0.38	0.00	5.27	0.0	-0.38
AKAP1	Q92667	0.01	5.12	0.00	-1.90	0.0	-2.65	0.00	-0.46	0.0	-2.61
MYOD1	P15172	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.00	4.00	0.00	5.51
PLCG1	P19174	0.10	12.87	0.02	-2.34	0.0	-6.05	0.00	-2.51	0.00	-4.57
MAX	P61244	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.00	4.00	0.00	5.51
IL18RAP	O95256	0.0	-3.93	0.0	-3.89	0.0	-3.05	0.0	-1.45	0.0	-3.00
CREB1	P16220	0.0	-2.48	0.0	-2.46	0.0	-1.93	0.03	7.83	0.02	10.74
CITED1	Q99966	0.0	-7.38	0.0	-7.31	0.0	-5.72	0.01	-1.19	0.24	32.69
CAMK2B	Q13554	0.02	4.00	0.01	-0.75	0.0	-3.74	0.00	-1.21	0.0	-3.69
CABLES1	Q8TDN4	0.0	-1.40	0.0	-1.39	0.0	-1.09	0.0	-0.51	0.0	-1.07
ATF3	P18847	0.0	-0.49	0.0	-0.49	0.0	-0.38	0.0	-0.18	0.0	-0.38
KAT2B	Q92831	0.0	-6.38	0.0	-6.33	0.0	-4.95	0.01	-0.61	0.18	28.93
PRKCB	P05771	0.04	9.95	0.0	-4.23	0.0	-3.31	0.01	0.99	0.00	-2.64
PTPN13	Q12923	0.11	-6.25	0.26	9.00	0.19	1.90	0.11	-1.99	0.22	4.55
LIMS1	P48059	0.0	-1.72	0.0	-1.70	0.01	7.65	0.0	-0.63	0.0	-1.31
TAB2	Q9NYJ8	0.01	1.88	0.00	-2.19	0.0	-2.65	0.00	-0.46	0.0	-2.61
ERBB2	P04626	0.11	2.64	0.18	11.66	0.0	-9.79	0.07	-0.84	0.00	-8.85
EGR4	Q05215	0.0	-1.72	0.0	-1.70	0.01	6.90	0.0	-0.63	0.0	-1.31
JUNB	P17275	0.0	-3.59	0.02	5.20	0.00	-2.43	0.0	-1.32	0.0	-2.75
SOS1	Q07889	0.12	-0.71	0.01	-12.90	0.24	9.94	0.01	-5.01	0.27	11.58
LCK	P06239	0.0	-12.83	0.10	0.64	0.67	57.07	0.0	-4.73	0.0	-9.82
MAP2K6	P52564	0.32	6.42	0.30	4.44	0.0	-17.58	0.18	-2.40	0.16	-6.41
HSF2	Q03933	0.0	-5.49	0.06	11.03	0.0	-4.26	0.0	-2.02	0.0	-4.20
HEY1	Q9Y5J3	0.00	-26.52	0.89	45.17	0.31	-0.91	0.24	-2.73	0.31	-0.85
KPNB1	Q14974	0.0	-11.30	0.0	-11.20	0.0	-8.77	0.05	-1.04	0.54	50.45
TGFBR3	Q03167	0.00	-17.65	0.02	-15.91	0.0	-14.06	0.98	29.66	0.99	62.53
MAP3K4	Q9Y6R4	0.02	5.80	0.00	-1.55	0.0	-3.17	0.00	-0.83	0.0	-3.12
FOXG1	P55316	0.0	-0.70	0.0	-0.69	0.0	-0.54	0.00	3.60	0.00	1.32
PDGFA	P04085	0.0	-0.86	0.00	2.66	0.0	-0.66	0.0	-0.31	0.0	-0.65
MAPK12	P53778	0.00	-9.80	0.08	3.62	0.0	-7.77	0.05	-0.22	0.14	9.73
ENG	P17813-1	0.0	-0.86	0.0	-0.85	0.0	-0.66	0.00	2.83	0.00	2.37
MAPK9	P45984	0.0	-1.57	0.00	-0.27	0.0	-1.21	0.0	-0.57	0.0	-1.20
RAP1A	P62834	0.0	-0.86	0.0	-0.85	0.0	-0.66	0.00	2.83	0.00	0.85
DOK1	Q99704	0.07	15.40	0.0	-5.42	0.0	-4.24	0.00	-1.51	0.0	-4.18
BSG	P35613	0.0	-1.86	0.0	-1.84	0.0	-1.44	0.00	2.23	0.00	-0.72
CER1	O95813	0.0	-2.16	0.00	0.18	0.0	-1.68	0.00	0.45	0.0	-1.66
IFNGR1	P15260	0.13	-4.27	0.12	-4.81	0.20	2.45	0.15	-0.71	0.21	3.58
PTGDR	Q13258	0.02	6.29	0.00	-1.55	0.0	-3.17	0.00	-0.83	0.0	-3.12
MAP3K6	O95382	0.02	-4.15	0.09	7.67	0.0	-7.14	0.02	-1.52	0.08	3.75
CAV1	Q03135	0.00	-7.58	0.02	-4.61	0.0	-6.55	0.99	64.48	0.06	2.93
IL2RG	P31785	0.0	-15.10	0.15	1.99	0.81	60.35	0.0	-5.56	0.0	-11.55
NR3C1	P04150-1	0.0	-4.66	0.01	-0.67	0.06	11.50	0.0	-1.72	0.0	-3.57
ELF1	P32519	0.0	-1.21	0.0	-1.20	0.00	5.41	0.0	-0.44	0.0	-0.93
IL6	P05231	0.0	-0.86	0.0	-0.85	0.0	-0.66	0.0	-0.31	0.0	-0.65
PRKCA	P17252	0.00	-26.55	0.89	45.17	0.31	-0.94	0.24	-2.59	0.31	-0.88
PAK1	Q13153	0.0	-1.79	0.00	5.55	0.0	-1.39	0.0	-0.66	0.0	-1.37
HMGB1	P09429	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
CD79A	P11912	0.0	-5.81	0.02	-0.08	0.12	19.96	0.0	-2.14	0.0	-4.45
HES1	Q14469	0.00	-26.52	0.89	45.17	0.31	-0.91	0.24	-2.73	0.31	-0.85
TGFBR1	P36897	0.51	7.80	0.21	-15.28	0.0	-25.16	0.99	16.60	1.0	35.04
HSP90AA1	P07900	0.11	0.71	0.19	10.71	0.06	-4.84	0.07	-1.45	0.00	-9.68
EP300	Q09472	0.00	-26.52	0.89	45.17	0.31	-0.91	0.24	-2.73	0.31	-0.85
DLG4	P78352	0.0	-0.86	0.0	-0.85	0.0	-0.66	0.00	2.83	0.00	0.85
PIAS3	Q9Y6X2	0.0	-7.80	0.0	-7.73	0.0	-6.05	0.01	-1.43	0.27	35.33
IRF7	Q92985	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.00	4.00	0.00	5.51
CSNK2B	P67870	0.0	-0.49	0.0	-0.49	0.0	-0.38	0.0	-0.18	0.00	2.24
MLST8	Q9BVC4	0.0	-2.28	0.00	-1.81	0.02	9.57	0.0	-0.84	0.0	-1.74
PIK3R1	P27986	0.0	-9.35	0.04	-2.07	0.33	36.78	0.0	-3.45	0.0	-7.16

NFKB1	P19838	0.0	-1.40	0.0	-1.39	0.0	-1.09	0.0	-0.51	0.0	-1.07
CD3G	P09693	0.99	47.83	0.00	-30.53	0.05	-20.59	0.0	-11.38	0.51	7.22
ESR1	P03372	0.08	5.24	0.06	1.27	0.00	-6.96	0.0	-3.37	0.05	-0.23
CD79B	P40259	0.0	-5.81	0.02	-0.08	0.12	19.96	0.0	-2.14	0.0	-4.45
HRAS	P01112	0.0	-9.34	0.04	-2.05	0.37	41.09	0.0	-3.44	0.0	-7.15
IL1B	P01584	0.15	13.69	0.0	-10.35	0.0	-8.10	0.0	-3.85	0.08	1.39
RET	P07949	0.07	15.07	0.0	-5.49	0.0	-4.30	0.00	-1.04	0.00	-4.00
CTGF	P29279	0.00	-17.65	0.02	-15.91	0.0	-14.06	0.98	29.66	0.99	62.53
SRC	P12931	0.0	-3.66	0.0	-3.63	0.0	-2.84	0.0	-1.35	0.0	-2.80
IL4R	P24394	0.0	-6.07	0.04	5.21	0.09	13.10	0.0	-2.24	0.0	-4.65
SHC1	P29353	0.12	-2.66	0.04	-11.73	0.37	17.88	0.01	-5.28	0.27	9.55
NF1	P21359	0.0	-1.72	0.00	0.64	0.0	-1.33	0.00	0.94	0.0	-1.31
IL27RA	Q6UWB1	0.15	10.49	0.0	-11.49	0.00	-8.88	0.0	-4.27	0.08	-0.28
VTN	P04004	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
SDC1	P18827	0.0	-4.77	0.00	-3.87	0.00	-2.06	0.06	6.30	0.02	1.62
DOCK7	Q96N67	0.03	10.27	0.0	-3.76	0.0	-2.94	0.00	-0.68	0.0	-2.90
DNAJA1	P31689	0.0	-0.99	0.0	-0.98	0.0	-0.77	0.0	-0.36	0.00	4.50
TGFB2	P61812	0.0	-3.30	0.01	4.09	0.0	-2.56	0.0	-1.21	0.0	-2.53
SMAD4	Q13485	0.51	7.80	0.21	-15.28	0.0	-25.16	0.99	16.60	1.0	35.04
IL18	Q14116	0.0	-3.93	0.0	-3.89	0.0	-3.05	0.0	-1.45	0.0	-3.00
PRKCD	Q05655	0.0	-2.28	0.0	-2.25	0.0	-1.76	0.0	-0.84	0.0	-1.74
NTN1	O95631	0.0	-2.81	0.00	-2.07	0.0	-2.18	0.04	8.63	0.00	-1.22
AR	P10275	0.0	-0.99	0.0	-0.98	0.0	-0.77	0.0	-0.36	0.00	4.50
ASAH1	Q13510	0.0	-0.86	0.00	2.66	0.0	-0.66	0.0	-0.31	0.0	-0.65
MTOR	P42345	0.0	-12.97	0.07	-3.94	0.67	56.82	0.0	-4.78	0.0	-9.93
TP63	Q9H3D4-1	0.0	-1.92	0.0	-1.90	0.0	-1.49	0.0	-0.71	0.0	-1.47
RICTOR	Q6R327	0.0	-2.28	0.00	-1.81	0.02	9.57	0.0	-0.84	0.0	-1.74
EGF	P01133	0.10	12.64	0.02	-2.72	0.0	-6.10	0.00	-2.54	0.00	-4.64
TGFBRAP1	O60466	0.00	-13.26	0.02	-11.23	0.0	-10.73	0.98	38.92	0.53	39.08
CCM2	Q9BSQ5	0.0	-0.49	0.0	-0.49	0.0	-0.38	0.0	-0.18	0.0	-0.38
SDC2	P34741	0.0	-4.77	0.00	-3.87	0.00	-2.06	0.06	6.30	0.02	1.62
IL2	P60568	0.0	-13.66	0.10	-0.73	0.72	58.62	0.02	-3.92	0.00	-10.24
EGR2	P11161	0.0	-5.35	0.04	7.37	0.05	8.11	0.0	-1.97	0.0	-4.09
CBFB	Q13951	0.0	-2.43	0.0	-2.41	0.0	-1.89	0.03	8.03	0.02	11.04
PDGFRB	P09619	0.0	-6.16	0.05	7.50	0.06	8.09	0.0	-2.27	0.0	-4.71
PDGFRA	P16234	0.12	1.70	0.00	-12.95	0.12	1.26	0.0	-5.12	0.26	14.10
FOXO1	Q12778	0.0	-3.80	0.03	10.17	0.0	-2.94	0.0	-1.40	0.0	-2.90
S1PR2	O95136	0.0	-1.21	0.00	2.11	0.0	-0.94	0.0	-0.44	0.0	-0.93
FKBP4	Q02790	0.0	-4.66	0.01	-0.67	0.06	11.50	0.0	-1.72	0.0	-3.57
KAT2A	Q92830	0.0	-7.12	0.0	-7.06	0.0	-5.52	0.01	-1.05	0.22	31.89
SYK	P43405	0.0	-5.20	0.01	-1.41	0.12	23.16	0.0	-1.92	0.0	-3.98
SMAD7	O15105	0.0	-3.76	0.0	-3.73	0.0	-2.92	0.13	18.94	0.03	6.91
CARM1	Q86X55	0.0	-0.99	0.0	-0.98	0.0	-0.77	0.0	-0.36	0.00	4.50
PTP4A3	O75365	0.0	-2.22	0.0	-2.20	0.0	-1.72	0.01	4.06	0.00	-0.52
PML	P29590	0.00	-7.54	0.02	-4.57	0.0	-6.53	0.98	64.05	0.06	2.98
EGFR	P00533	0.10	12.64	0.02	-2.72	0.0	-6.10	0.00	-2.54	0.00	-4.64
EPAS1	Q99814	0.0	-1.72	0.00	1.81	0.0	-1.33	0.0	-0.63	0.0	-1.31
MAPK14	Q16539	0.00	-20.92	0.68	41.32	0.00	-16.10	0.06	-5.64	0.0	-16.20
RIPK1	Q13546	0.00	-6.81	0.12	19.46	0.00	-5.20	0.01	-1.35	0.0	-5.32
MET	P08581	0.0	-1.72	0.0	-1.70	0.0	-1.33	0.0	-0.63	0.0	-1.31
FOXH1	O75593	0.0	-0.99	0.0	-0.98	0.0	-0.77	0.00	5.09	0.00	4.50
AXIN1	O15169	0.0	-15.24	0.0	-15.10	0.0	-11.82	0.0	-5.62	0.92	68.42
MAPKAP1	Q9BPZ7	0.0	-2.28	0.00	-1.81	0.02	9.57	0.0	-0.84	0.0	-1.74
EBI3	Q14213	0.15	10.49	0.0	-11.49	0.00	-8.88	0.0	-4.27	0.08	-0.28
XIAP	P98170	0.0	-6.78	0.02	-0.59	0.0	-5.26	0.03	0.79	0.0	-5.19
TNC	P24821	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
SMAD2	Q15796	0.0	-0.99	0.0	-0.98	0.0	-0.77	0.00	5.09	0.00	4.50
ATF2	P15336	0.30	-33.83	0.99	24.56	0.80	6.42	0.43	-8.60	0.81	6.95
RASA1	P20936	0.07	15.40	0.0	-5.42	0.0	-4.24	0.00	-1.51	0.0	-4.18
KDR	P35968	0.01	3.61	0.00	-2.29	0.0	-3.02	0.00	-0.73	0.0	-2.98
PTGES2	Q9H7Z7	0.0	-0.49	0.0	-0.49	0.0	-0.38	0.0	-0.18	0.0	-0.38
TRAF2	Q12933	0.02	3.59	0.01	-0.30	0.0	-3.55	0.00	-0.49	0.0	-3.50
DCC	P43146	0.0	-2.81	0.00	-2.07	0.0	-2.18	0.04	8.63	0.00	-1.22
CHRD	Q9H2X0	0.0	-2.16	0.00	0.18	0.0	-1.68	0.00	0.45	0.0	-1.66
MAP3K5	Q99683	0.10	1.35	0.13	5.32	0.0	-9.74	0.04	-2.48	0.08	-1.33
JAK1	P23458	0.13	-12.05	0.26	-0.51	0.80	35.81	0.15	-3.71	0.21	-3.39
SHC1	P29353-2	0.12	-0.71	0.01	-12.90	0.24	9.94	0.01	-5.01	0.27	11.58

JUN	P05412	0.99	20.36	0.93	14.82	0.97	14.22	0.84	2.36	0.03	-52.68
NCK1	P16333	0.07	15.40	0.0	-5.42	0.0	-4.24	0.00	-1.51	0.0	-4.18
SRF	P11831	0.0	-16.18	0.15	0.10	0.92	65.20	0.0	-5.96	0.0	-12.38
ARHGDIA	P52565	0.0	-1.79	0.00	5.55	0.0	-1.39	0.0	-0.66	0.0	-1.37
GATA2	P23769	0.0	-3.49	0.02	9.07	0.0	-2.70	0.0	-1.28	0.0	-2.67
FKBP5	Q13451	0.0	-4.66	0.01	-0.67	0.06	11.50	0.0	-1.72	0.0	-3.57
MAP3K3	Q99759	0.07	3.20	0.12	11.13	0.00	-7.03	0.00	-2.79	0.01	-4.69
ITGB2	P05107	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
GDNF	P39905	0.07	15.07	0.0	-5.49	0.0	-4.30	0.00	-1.04	0.00	-4.00
FOXA2	Q9Y261	0.0	-1.98	0.00	0.06	0.0	-1.54	0.0	-0.73	0.0	-1.52
STAT1	P42224	0.0	-6.92	0.03	1.27	0.18	25.77	0.00	-2.14	0.0	-5.29
JAK3	P52333	0.0	-14.50	0.15	3.01	0.76	58.01	0.0	-5.35	0.0	-11.10
IL2RA	P01589	0.0	-13.46	0.10	-0.39	0.72	59.34	0.0	-4.96	0.0	-10.30
BAG1	Q99933	0.0	-0.49	0.0	-0.49	0.0	-0.38	0.0	-0.18	0.0	-0.38
BAG4	O95429	0.0	-0.49	0.0	-0.49	0.0	-0.38	0.00	5.27	0.0	-0.38
PRKRA	O75569	0.0	-0.70	0.0	-0.69	0.00	1.28	0.0	-0.25	0.0	-0.53
GATA1	P15976	0.00	-26.52	0.89	45.17	0.31	-0.91	0.24	-2.73	0.31	-0.85
TNFSF10	P50591	0.0	-2.33	0.0	-2.31	0.0	-1.81	0.02	6.13	0.00	-0.66
IL4	P05112	0.0	-6.07	0.04	5.21	0.09	13.10	0.0	-2.24	0.0	-4.65
MDM2	Q00987	0.00	-7.59	0.02	-4.63	0.0	-6.56	1.0	64.69	0.06	2.90
RAC1	P63000	0.0	-3.15	0.01	5.25	0.0	-2.44	0.0	-1.16	0.0	-2.41
SOSTDC1	Q6X4U4	0.0	-2.16	0.00	0.18	0.0	-1.68	0.00	0.45	0.0	-1.66
KPNA2	P52292	0.0	-15.54	0.0	-15.40	0.0	-12.06	0.12	-0.45	0.96	69.80
CTNNB1	P35222	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
CD40	P25942	0.0	-0.86	0.00	0.32	0.0	-0.66	0.0	-0.31	0.0	-0.65
TNFRSF1A	P19438	0.0	-0.49	0.0	-0.49	0.0	-0.38	0.00	5.27	0.0	-0.38
TP53	P04637	0.00	-7.59	0.02	-4.63	0.0	-6.56	1.0	64.69	0.06	2.90
MAP3K14	Q99558	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.00	4.00	0.00	1.21
IL6R	P08887	0.0	-0.49	0.0	-0.49	0.0	-0.38	0.0	-0.18	0.0	-0.38
HHAT	Q5VTY9	0.0	-0.86	0.0	-0.85	0.0	-0.66	0.00	2.83	0.00	0.85
CD8A	P01732	0.0	-3.66	0.00	-3.35	0.05	14.53	0.0	-1.35	0.0	-2.80
CD4	P01730	0.99	48.75	0.0	-30.01	0.0	-23.50	0.0	-11.17	0.51	7.79
STAT3	P40763	0.0	-11.11	0.04	-4.81	0.37	33.20	0.0	-4.09	0.0	-8.50
GREM1	O60565	0.0	-2.16	0.00	0.18	0.0	-1.68	0.00	0.45	0.0	-1.66
PITPNA	Q00169	0.0	-0.70	0.00	2.17	0.0	-0.54	0.0	-0.25	0.0	-0.53
CSK	P41240	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
CD3D	P04234	0.99	47.83	0.00	-30.53	0.05	-20.59	0.0	-11.38	0.51	7.22
PIK3CA	P42336	0.0	-9.35	0.04	-2.07	0.33	36.78	0.0	-3.45	0.0	-7.16
PTGIR	P43119	0.02	5.97	0.00	-1.61	0.0	-3.07	0.00	-0.77	0.0	-3.03
PPIB	P23284	0.0	-1.86	0.0	-1.84	0.0	-1.44	0.00	2.23	0.00	-0.72
GRAP2	O75791	0.0	-5.23	0.02	1.48	0.12	23.02	0.0	-1.93	0.0	-4.00
MAP3K10	Q02779	0.03	-6.32	0.17	14.89	0.0	-8.56	0.02	-2.47	0.07	0.39
CD3E	P07766	0.99	47.83	0.00	-30.53	0.05	-20.59	0.0	-11.38	0.51	7.22
IRF1	P10914	0.0	-2.28	0.0	-2.25	0.0	-1.76	0.0	-0.84	0.0	-1.74
PPARG	P37231	0.0	-2.81	0.0	-2.79	0.03	12.06	0.0	-1.03	0.0	-2.15
ZFYVE9	O95405	0.00	-7.54	0.02	-4.57	0.0	-6.53	0.98	64.05	0.06	2.98
PKN1	Q16512	0.0	-0.99	0.0	-0.98	0.0	-0.77	0.0	-0.36	0.00	4.50
VAV2	P52735	0.0	-2.53	0.00	2.67	0.0	-1.96	0.0	-0.93	0.0	-1.94
JAK2	O60674	1.0	35.61	0.12	-31.75	0.20	-20.07	0.15	-11.00	0.58	2.84
TYK2	P29597	1.0	44.20	0.0	-33.28	0.00	-25.99	0.0	-12.38	0.51	4.67
PTEN	P60484	0.25	24.53	0.06	-1.90	0.0	-8.95	0.00	-3.99	0.00	-8.70
ACVRL1	P37023	0.0	-0.86	0.0	-0.85	0.0	-0.66	0.00	2.83	0.00	2.37
FGF2	P09038	0.0	-3.80	0.03	10.17	0.0	-2.94	0.0	-1.40	0.0	-2.90
SMAD3	P84022	0.0	-15.50	0.0	-15.35	0.0	-12.02	0.11	-0.83	0.95	69.60
SPHK1	Q9NYA1	0.0	-0.70	0.00	2.17	0.0	-0.54	0.0	-0.25	0.0	-0.53
IL12B	P29460	1.0	44.29	0.0	-33.21	0.0	-26.01	0.0	-12.36	0.51	4.73
CALM2	P62158	0.02	-14.89	0.16	-1.12	0.98	63.64	0.00	-6.36	0.0	-13.59
MAP3K12	Q12852	0.15	15.52	0.0	-9.47	0.0	-7.41	0.0	-3.52	0.11	7.18
GNB1	P62873	0.08	10.76	0.0	-7.13	0.0	-5.58	0.0	-2.65	0.05	2.93
EIF2AK2	P19525	0.0	-0.70	0.0	-0.69	0.00	1.28	0.0	-0.25	0.0	-0.53
MAPK8	P45983	0.81	49.35	0.05	-17.47	0.0	-17.49	0.20	-1.53	0.00	-17.10
FST	P19883	0.0	-2.16	0.00	0.18	0.0	-1.68	0.00	0.45	0.0	-1.66
STAT4	Q14765	0.00	-4.50	0.0	-5.97	0.00	-4.46	0.0	-2.22	0.0	-4.61
IL27	Q8NEV9	0.15	10.49	0.0	-11.49	0.00	-8.88	0.0	-4.27	0.08	-0.28
LMO4	P61968	0.0	-0.49	0.0	-0.49	0.0	-0.38	0.0	-0.18	0.0	-0.38
CREBBP	Q92793	0.0	-3.70	0.01	1.28	0.06	16.45	0.0	-1.36	0.0	-2.83
STAT6	P42226	0.0	-5.35	0.04	7.37	0.05	8.11	0.0	-1.97	0.0	-4.09

RELA	Q04206	0.0	-0.70	0.0	-0.69	0.0	-0.54	0.0	-0.25	0.0	-0.53
MAP3K7	O43318	0.20	-11.69	0.50	12.67	0.39	2.85	0.27	-2.29	0.38	2.06
CXCR3	P49682-2	0.0	-3.03	0.01	6.04	0.0	-2.35	0.0	-1.11	0.0	-2.32
AGER	Q15109	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
CD8B	P10966	0.0	-3.66	0.00	-3.35	0.05	14.53	0.0	-1.35	0.0	-2.80
NUP153	P49790	0.0	-9.73	0.0	-9.64	0.0	-7.55	0.05	-0.03	0.41	44.08
CD28	P10747	0.0	-0.86	0.0	-0.85	0.0	-0.66	0.0	-0.31	0.0	-0.65
CD40LG	P29965	0.0	-6.09	0.04	5.32	0.08	12.59	0.0	-2.24	0.0	-4.66
BAMBI	Q13145	0.0	-6.20	0.02	0.52	0.0	-4.81	0.03	1.29	0.0	-4.74
TXN	P10599	0.04	5.24	0.02	-0.45	0.0	-4.94	0.00	-1.47	0.0	-4.87
PDGFB	P01127	0.0	-6.16	0.05	7.50	0.06	8.09	0.0	-2.27	0.0	-4.71
ITGAM	P11215	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
HLA-DRB1	P04229	0.99	48.75	0.0	-30.01	0.0	-23.50	0.0	-11.17	0.51	7.79
MAP3K11	Q16584	0.0	-0.49	0.0	-0.49	0.0	-0.38	0.0	-0.18	0.0	-0.38
TP63	Q9H3D4-2	0.0	-2.16	0.00	1.58	0.0	-1.68	0.00	0.45	0.0	-1.66
AKT1	P31749	0.0	-13.04	0.07	-4.05	0.67	56.47	0.00	-4.34	0.00	-9.76
TRIM28	Q13263	0.00	-7.59	0.02	-4.63	0.0	-6.56	1.0	64.69	0.06	2.90
SDC3	O75056	0.0	-4.77	0.00	-3.87	0.00	-2.06	0.06	6.30	0.02	1.62
B2M	P61769	0.0	-3.66	0.00	-3.35	0.05	14.53	0.0	-1.35	0.0	-2.80
AHSG	P02765	0.0	-2.16	0.00	0.18	0.0	-1.68	0.00	0.45	0.0	-1.66
PTPRC	P08575	0.0	-5.81	0.02	-0.08	0.12	19.96	0.0	-2.14	0.0	-4.45
GPC1	P35052	0.0	-0.70	0.0	-0.69	0.0	-0.54	0.0	-0.25	0.00	3.18
IRF4	Q15306	0.0	-5.35	0.04	7.37	0.05	8.11	0.0	-1.97	0.0	-4.09
NCOA2	Q15596	0.0	-0.99	0.0	-0.98	0.0	-0.77	0.0	-0.36	0.00	4.50
PPAP2A	O14494	0.22	22.26	0.01	-7.94	0.0	-8.35	0.00	-3.70	0.02	-5.75
PRKACA	P17612	0.07	9.42	0.01	-2.67	0.0	-5.38	0.01	-0.94	0.0	-5.31
HLA-DRA	P01903	0.99	48.75	0.0	-30.01	0.0	-23.50	0.0	-11.17	0.51	7.79
YBX1	P67809	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
CD247	P20963	0.99	47.83	0.00	-30.53	0.05	-20.59	0.0	-11.38	0.51	7.22
IL12A	P29459	1.0	44.29	0.0	-33.21	0.0	-26.01	0.0	-12.36	0.51	4.73
ITGB3	P05106	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
PTPN1	P18031	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
EGR1	P18146	0.0	-6.64	0.08	13.10	0.04	2.62	0.0	-2.45	0.0	-5.08
RAF1	P04049	0.0	-9.59	0.04	-2.46	0.37	39.79	0.02	-2.03	0.00	-7.19
GAB1	Q13480	0.0	-0.70	0.00	0.74	0.0	-0.54	0.0	-0.25	0.0	-0.53
FLT1	P17948	0.0	-2.11	0.00	0.30	0.01	3.87	0.0	-0.77	0.0	-1.61
SGMS1	Q86VZ5	0.27	13.86	0.24	10.59	0.0	-12.44	0.05	-3.53	0.06	-7.12
PGF	P49763	0.0	-2.11	0.00	0.30	0.01	3.87	0.0	-0.77	0.0	-1.61
IL2RB	P14784	0.0	-13.46	0.10	-0.39	0.72	59.34	0.0	-4.96	0.0	-10.30
CHRDL1	Q9BU40	0.0	-2.16	0.00	0.18	0.0	-1.68	0.00	0.45	0.0	-1.66
IRS1	P35568	0.0	-0.70	0.00	2.17	0.0	-0.54	0.0	-0.25	0.0	-0.53
TGFBR2	P37173	0.51	7.80	0.21	-15.28	0.0	-25.16	0.99	16.60	1.0	35.04
MAP3K1	Q13233	0.34	-3.13	0.38	0.46	0.40	1.61	0.28	-2.74	0.43	3.18
FAS	P25445	0.02	4.92	0.00	-1.53	0.0	-2.76	0.00	-0.54	0.0	-2.72
HLA-A	P04439	0.0	-3.66	0.00	-3.35	0.05	14.53	0.0	-1.35	0.0	-2.80
CEBPB	P17676	0.0	-9.87	0.08	3.16	0.19	15.97	0.00	-3.35	0.00	-7.41
LYN	P07948	0.0	-5.81	0.02	-0.08	0.12	19.96	0.0	-2.14	0.0	-4.45
POU2F1	P14859	0.0	-2.72	0.0	-2.70	0.03	11.65	0.0	-1.00	0.0	-2.08
AURKA	O14965	0.0	-0.86	0.0	-0.85	0.0	-0.66	0.00	2.83	0.00	0.85
NUP214	P35658	0.0	-9.79	0.0	-9.70	0.0	-7.59	0.06	0.51	0.41	44.34
GFRA1	P56159	0.07	15.07	0.0	-5.49	0.0	-4.30	0.00	-1.04	0.00	-4.00
MEF2C	Q06413	0.0	-0.99	0.0	-0.98	0.0	-0.77	0.0	-0.36	0.00	4.50
ITGA9	Q13797	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
GRB2	P62993	0.12	-0.80	0.02	-12.88	0.24	9.84	0.01	-4.82	0.27	11.58
ATM	Q13315	0.0	-2.16	0.00	1.58	0.0	-1.68	0.00	0.45	0.0	-1.66
EFNB1	P98172	0.19	-9.94	0.48	13.84	0.39	5.33	0.22	-2.58	0.38	4.49
GSN	P06396	0.10	12.99	0.02	-2.56	0.0	-6.02	0.00	-2.50	0.00	-4.54
MAP3K8	P41279	0.04	0.76	0.11	11.83	0.00	-6.38	0.03	-0.41	0.00	-5.96
NKX2-5	P52952	0.0	-0.70	0.0	-0.69	0.0	-0.54	0.00	3.60	0.00	3.18
MAP2K1	Q02750	0.0	-8.09	0.07	5.39	0.00	-6.11	0.17	9.26	0.00	-5.01
CSF2	P04141	0.0	-2.11	0.0	-2.09	0.0	-1.63	0.02	6.95	0.00	-0.99
ITGB1	P05556	0.0	-1.21	0.0	-1.20	0.0	-0.94	0.0	-0.44	0.0	-0.93
NRG1	Q02297-6	0.11	2.57	0.18	11.57	0.0	-9.82	0.08	-0.63	0.00	-8.77
RIPK2	O43353	0.00	-1.62	0.0	-3.73	0.0	-2.92	0.0	-1.38	0.0	-2.88
ERBB3	P21860	0.11	2.64	0.18	11.66	0.0	-9.79	0.07	-0.84	0.00	-8.85
HIF1A	Q16665	0.12	16.48	0.0	-7.93	0.00	-5.37	0.0	-2.95	0.04	0.16
GATA3	P23771	0.0	-3.07	0.0	-3.04	0.03	10.70	0.00	0.64	0.00	0.21

ILK	Q13418	0.0	-2.28	0.00	-1.81	0.02	9.57	0.0	-0.84	0.0	-1.74
DAPP1	Q9UN19	0.0	-2.53	0.00	2.67	0.0	-1.96	0.0	-0.93	0.0	-1.94

Table 15 – *Table du niveau de représentation des protéines dans chaque noyau*

4.2 Table des termes GO significativement enrichis dans chaque noyau

Terme GO	**Nom**	***p-value* norm.**
Noyau 1		
GO :0050852	T Cell Receptor Signaling Pathway	0.000
GO :0050851	Antigen Receptor Mediated Signaling Pathway	0.000
GO :0009617	Response To Bacterium	0.000
GO :0032649	Regulation Of Interferon Gamma Production	0.001
GO :0032535	Regulation Of Cellular Component Size	0.002
GO :0035148	Tube Formation	0.004
GO :0050778	Positive Regulation Of Immune Response	0.004
GO :0032729	Positive Regulation Of Interferon Gamma Production	0.005
GO :0051345	Positive Regulation Of Hydrolase Activity	0.006
GO :0002237	Response To Molecule Of Bacterial Origin	0.008
GO :2001235	Positive Regulation Of Apoptotic Signaling Pathway	0.009
GO :0043087	Regulation Of Gtpase Activity	0.009
GO :0002768	Immune Response Regulating Cell Surface Receptor Signaling Pathway	0.009
GO :0071396	Cellular Response To Lipid	0.010
GO :0016331	Morphogenesis Of Embryonic Epithelium	0.011
GO :0009894	Regulation Of Catabolic Process	0.012
GO :0043085	Positive Regulation Of Catalytic Activity	0.015
GO :0009607	Response To Biotic Stimulus	0.016
GO :0090066	Regulation Of Anatomical Structure Size	0.017
GO :0002253	Activation Of Immune Response	0.018
GO :0002684	Positive Regulation Of Immune System Process	0.018
GO :0002821	Positive Regulation Of Adaptive Immune Response	0.019
GO :2001233	Regulation Of Apoptotic Signaling Pathway	0.022
GO :0006644	Phospholipid Metabolic Process	0.024
GO :0007173	Epidermal Growth Factor Receptor Signaling Pathway	0.024
GO :0009896	Positive Regulation Of Catabolic Process	0.026
GO :0042129	Regulation Of T Cell Proliferation	0.026
GO :0042108	Positive Regulation Of Cytokine Biosynthetic Process	0.027
GO :0071229	Cellular Response To Acid Chemical	0.028
GO :0021700	Developmental Maturation	0.029
GO :0002682	Regulation Of Immune System Process	0.032
GO :0050776	Regulation Of Immune Response	0.032
GO :0050865	Regulation Of Cell Activation	0.034
GO :0060341	Regulation Of Cellular Localization	0.037
GO :0030258	Lipid Modification	0.038
GO :0033993	Response To Lipid	0.040
GO :0042102	Positive Regulation Of T Cell Proliferation	0.041
GO :0033135	Regulation Of Peptidyl Serine Phosphorylation	0.045
GO :0019637	Organophosphate Metabolic Process	0.045
GO :0042176	Regulation Of Protein Catabolic Process	0.046
GO :0048584	Positive Regulation Of Response To Stimulus	0.047
GO :0051336	Regulation Of Hydrolase Activity	0.047
GO :0030155	Regulation Of Cell Adhesion	0.049
GO :0043623	Cellular Protein Complex Assembly	0.050
Noyau 2		
GO :0072359	Circulatory System Development	0.001
GO :0071902	Positive Regulation Of Protein Serine Threonine Kinase Activity	0.001
GO :0061448	Connective Tissue Development	0.001
GO :0051347	Positive Regulation Of Transferase Activity	0.002
GO :1902531	Regulation Of Intracellular Signal Transduction	0.002
GO :0043406	Positive Regulation Of Map Kinase Activity	0.002

GO :0043507	Positive Regulation Of Jun Kinase Activity	0.002
GO :0006887	Exocytosis	0.002
GO :0033674	Positive Regulation Of Kinase Activity	0.003
GO :0070304	Positive Regulation Of Stress Activated Protein Kinase Signaling Cascade	0.003
GO :0043549	Regulation Of Kinase Activity	0.004
GO :0044255	Cellular Lipid Metabolic Process	0.004
GO :0043506	Regulation Of Jun Kinase Activity	0.004
GO :1901564	Organonitrogen Compound Metabolic Process	0.005
GO :0010562	Positive Regulation Of Phosphorus Metabolic Process	0.006
GO :0006629	Lipid Metabolic Process	0.006
GO :0031098	Stress Activated Protein Kinase Signaling Cascade	0.006
GO :0051338	Regulation Of Transferase Activity	0.007
GO :0051174	Regulation Of Phosphorus Metabolic Process	0.009
GO :0000187	Activation Of Mapk Activity	0.009
GO :0043405	Regulation Of Map Kinase Activity	0.010
GO :1901342	Regulation Of Vasculature Development	0.010
GO :0030258	Lipid Modification	0.011
GO :1902533	Positive Regulation Of Intracellular Signal Transduction	0.011
GO :0071900	Regulation Of Protein Serine Threonine Kinase Activity	0.011
GO :0032844	Regulation Of Homeostatic Process	0.011
GO :0070302	Regulation Of Stress Activated Protein Kinase Signaling Cascade	0.011
GO :1903725	Regulation Of Phospholipid Metabolic Process	0.012
GO :0046328	Regulation Of Jnk Cascade	0.013
GO :0048878	Chemical Homeostasis	0.014
GO :0008610	Lipid Biosynthetic Process	0.016
GO :0007186	G Protein Coupled Receptor Signaling Pathway	0.018
GO :0048871	Multicellular Organismal Homeostasis	0.019
GO :0043410	Positive Regulation Of Mapk Cascade	0.019
GO :0006644	Phospholipid Metabolic Process	0.020
GO :0051480	Regulation Of Cytosolic Calcium Ion Concentration	0.020
GO :2000021	Regulation Of Ion Homeostasis	0.022
GO :0022603	Regulation Of Anatomical Structure Morphogenesis	0.023
GO :0046834	Lipid Phosphorylation	0.024
GO :0043408	Regulation Of Mapk Cascade	0.026
GO :0014066	Regulation Of Phosphatidylinositol 3 Kinase Signaling	0.031
GO :0046488	Phosphatidylinositol Metabolic Process	0.032
GO :0008544	Epidermis Development	0.032
GO :0019637	Organophosphate Metabolic Process	0.033
GO :0045787	Positive Regulation Of Cell Cycle	0.033
GO :0001890	Placenta Development	0.035
GO :0046486	Glycerolipid Metabolic Process	0.036
GO :0032147	Activation Of Protein Kinase Activity	0.036
GO :0032940	Secretion By Cell	0.036
GO :0044281	Small Molecule Metabolic Process	0.036
GO :0001944	Vasculature Development	0.037
GO :0007254	Jnk Cascade	0.039
GO :0030595	Leukocyte Chemotaxis	0.039
GO :0023014	Signal Transduction By Protein Phosphorylation	0.042
GO :0006650	Glycerophospholipid Metabolic Process	0.045
GO :0001892	Embryonic Placenta Development	0.046
GO :0009891	Positive Regulation Of Biosynthetic Process	0.048
GO :0042060	Wound Healing	0.050
GO :0080135	Regulation Of Cellular Response To Stress	0.050
GO :0030168	Platelet Activation	0.051
	Noyau 3	
GO :0030168	Platelet Activation	0.000
GO :0030098	Lymphocyte Differentiation	0.000
GO :0046649	Lymphocyte Activation	0.000
GO :0002521	Leukocyte Differentiation	0.000
GO :0098542	Defense Response To Other Organism	0.000
GO :0002683	Negative Regulation Of Immune System Process	0.000
GO :0050776	Regulation Of Immune Response	0.000
GO :0002697	Regulation Of Immune Effector Process	0.000
GO :0034110	Regulation Of Homotypic Cell Cell Adhesion	0.000
GO :0007159	Leukocyte Cell Cell Adhesion	0.000
GO :0022409	Positive Regulation Of Cell Cell Adhesion	0.000
GO :0006955	Immune Response	0.000
GO :0006952	Defense Response	0.000
GO :0002682	Regulation Of Immune System Process	0.000
GO :0022407	Regulation Of Cell Cell Adhesion	0.000
GO :0002376	Immune System Process	0.000

GO :0002684	Positive Regulation Of Immune System Process	0.000
GO :0030155	Regulation Of Cell Adhesion	0.000
GO :0050867	Positive Regulation Of Cell Activation	0.001
GO :0045321	Leukocyte Activation	0.001
GO :0001775	Cell Activation	0.001
GO :0098609	Cell Cell Adhesion	0.001
GO :0002250	Adaptive Immune Response	0.001
GO :0050900	Leukocyte Migration	0.001
GO :0050778	Positive Regulation Of Immune Response	0.002
GO :0045785	Positive Regulation Of Cell Adhesion	0.002
GO :0022610	Biological Adhesion	0.002
GO :0043086	Negative Regulation Of Catalytic Activity	0.002
GO :0002252	Immune Effector Process	0.002
GO :0007599	Hemostasis	0.002
GO :0042113	B Cell Activation	0.002
GO :0098602	Single Organism Cell Adhesion	0.003
GO :0006954	Inflammatory Response	0.003
GO :0043254	Regulation Of Protein Complex Assembly	0.003
GO :0006811	Ion Transport	0.003
GO :0033500	Carbohydrate Homeostasis	0.003
GO :0002520	Immune System Development	0.004
GO :0010563	Negative Regulation Of Phosphorus Metabolic Process	0.004
GO :0050851	Antigen Receptor Mediated Signaling Pathway	0.004
GO :0042326	Negative Regulation Of Phosphorylation	0.004
GO :0050865	Regulation Of Cell Activation	0.005
GO :0051248	Negative Regulation Of Protein Metabolic Process	0.005
GO :0042592	Homeostatic Process	0.005
GO :0031347	Regulation Of Defense Response	0.005
GO :0002703	Regulation Of Leukocyte Mediated Immunity	0.005
GO :0048878	Chemical Homeostasis	0.006
GO :0032868	Response To Insulin	0.006
GO :0030217	T Cell Differentiation	0.006
GO :0045184	Establishment Of Protein Localization	0.007
GO :0032970	Regulation Of Actin Filament Based Process	0.007
GO :0009615	Response To Virus	0.007
GO :0002768	Immune Response Regulating Cell Surface Receptor Signaling Pathway	0.008
GO :0031294	Lymphocyte Costimulation	0.008
GO :0031400	Negative Regulation Of Protein Modification Process	0.009
GO :0051346	Negative Regulation Of Hydrolase Activity	0.010
GO :0034330	Cell Junction Organization	0.010
GO :0038095	Fc Epsilon Receptor Signaling Pathway	0.010
GO :0034248	Regulation Of Cellular Amide Metabolic Process	0.012
GO :0002220	Innate Immune Response Activating Cell Surface Receptor Signaling Pathway	0.013
GO :0007169	Transmembrane Receptor Protein Tyrosine Kinase Signaling Pathway	0.015
GO :0032869	Cellular Response To Insulin Stimulus	0.015
GO :0045087	Innate Immune Response	0.016
GO :0002700	Regulation Of Production Of Molecular Mediator Of Immune Response	0.017
GO :0032102	Negative Regulation Of Response To External Stimulus	0.019
GO :0044092	Negative Regulation Of Molecular Function	0.019
GO :0034097	Response To Cytokine	0.020
GO :1903706	Regulation Of Hemopoiesis	0.021
GO :1902105	Regulation Of Leukocyte Differentiation	0.021
GO :0010608	Posttranscriptional Regulation Of Gene Expression	0.023
GO :0050688	Regulation Of Defense Response To Virus	0.024
GO :0002699	Positive Regulation Of Immune Effector Process	0.025
GO :0002705	Positive Regulation Of Leukocyte Mediated Immunity	0.025
GO :0043900	Regulation Of Multi Organism Process	0.028
GO :1901698	Response To Nitrogen Compound	0.028
GO :0009725	Response To Hormone	0.028
GO :1903035	Negative Regulation Of Response To Wounding	0.028
GO :0051649	Establishment Of Localization In Cell	0.028
GO :0002706	Regulation Of Lymphocyte Mediated Immunity	0.029
GO :0002253	Activation Of Immune Response	0.030
GO :0044087	Regulation Of Cellular Component Biogenesis	0.030
GO :0043087	Regulation Of Gtpase Activity	0.030
GO :0002831	Regulation Of Response To Biotic Stimulus	0.030
GO :0051348	Negative Regulation Of Transferase Activity	0.032
GO :0001817	Regulation Of Cytokine Production	0.033
GO :0051336	Regulation Of Hydrolase Activity	0.033
GO :0048017	Inositol Lipid Mediated Signaling	0.033
GO :1901652	Response To Peptide	0.033

GO :0050878	Regulation Of Body Fluid Levels	0.034
GO :0072594	Establishment Of Protein Localization To Organelle	0.034
GO :0010975	Regulation Of Neuron Projection Development	0.034
GO :0009607	Response To Biotic Stimulus	0.034
GO :0045834	Positive Regulation Of Lipid Metabolic Process	0.037
GO :0051170	Nuclear Import	0.037
GO :0002702	Positive Regulation Of Production Of Molecular Mediator Of Immune Response	0.038
GO :0017038	Protein Import	0.039
GO :0051495	Positive Regulation Of Cytoskeleton Organization	0.039
GO :0001819	Positive Regulation Of Cytokine Production	0.043
GO :0008104	Protein Localization	0.043
GO :0031668	Cellular Response To Extracellular Stimulus	0.045
GO :1901653	Cellular Response To Peptide	0.046
GO :0045621	Positive Regulation Of Lymphocyte Differentiation	0.046
GO :0038093	Fc Receptor Signaling Pathway	0.047
GO :0045598	Regulation Of Fat Cell Differentiation	0.048
GO :0050864	Regulation Of B Cell Activation	0.049
GO :0044281	Small Molecule Metabolic Process	0.051
	Noyau 4	
GO :0051604	Protein Maturation	0.000
GO :0070727	Cellular Macromolecule Localization	0.000
GO :0031638	Zymogen Activation	0.001
GO :0010256	Endomembrane System Organization	0.001
GO :0008104	Protein Localization	0.002
GO :0097202	Activation Of Cysteine Type Endopeptidase Activity	0.002
GO :1902580	Single Organism Cellular Localization	0.003
GO :0009057	Macromolecule Catabolic Process	0.003
GO :0006508	Proteolysis	0.004
GO :0060627	Regulation Of Vesicle Mediated Transport	0.005
GO :0018205	Peptidyl Lysine Modification	0.005
GO :0051129	Negative Regulation Of Cellular Component Organization	0.006
GO :1903649	Regulation Of Cytoplasmic Transport	0.007
GO :0051051	Negative Regulation Of Transport	0.008
GO :0010638	Positive Regulation Of Organelle Organization	0.009
GO :0045862	Positive Regulation Of Proteolysis	0.012
GO :2001235	Positive Regulation Of Apoptotic Signaling Pathway	0.012
GO :0010821	Regulation Of Mitochondrion Organization	0.012
GO :0045732	Positive Regulation Of Protein Catabolic Process	0.013
GO :1903827	Regulation Of Cellular Protein Localization	0.014
GO :0070647	Protein Modification By Small Protein Conjugation Or Removal	0.014
GO :0032386	Regulation Of Intracellular Transport	0.015
GO :1904892	Regulation Of Stat Cascade	0.015
GO :0050730	Regulation Of Peptidyl Tyrosine Phosphorylation	0.017
GO :0043086	Negative Regulation Of Catalytic Activity	0.019
GO :0034504	Protein Localization To Nucleus	0.020
GO :0007264	Small Gtpase Mediated Signal Transduction	0.021
GO :1904894	Positive Regulation Of Stat Cascade	0.021
GO :0051270	Regulation Of Cellular Component Movement	0.024
GO :0030100	Regulation Of Endocytosis	0.024
GO :0001558	Regulation Of Cell Growth	0.024
GO :0051248	Negative Regulation Of Protein Metabolic Process	0.025
GO :0033157	Regulation Of Intracellular Protein Transport	0.026
GO :0061024	Membrane Organization	0.026
GO :0001818	Negative Regulation Of Cytokine Production	0.028
GO :0009056	Catabolic Process	0.030
GO :0097191	Extrinsic Apoptotic Signaling Pathway	0.031
GO :0009888	Tissue Development	0.034
GO :0042176	Regulation Of Protein Catabolic Process	0.034
GO :2001236	Regulation Of Extrinsic Apoptotic Signaling Pathway	0.035
GO :0050770	Regulation Of Axonogenesis	0.036
GO :0010952	Positive Regulation Of Peptidase Activity	0.036
GO :0030510	Regulation Of Bmp Signaling Pathway	0.037
GO :0051259	Protein Oligomerization	0.038
GO :0042509	Regulation Of Tyrosine Phosphorylation Of Stat Protein	0.039
GO :0044248	Cellular Catabolic Process	0.039
GO :0090287	Regulation Of Cellular Response To Growth Factor Stimulus	0.039
GO :0034612	Response To Tumor Necrosis Factor	0.039
GO :0097190	Apoptotic Signaling Pathway	0.045
GO :0034599	Cellular Response To Oxidative Stress	0.045
GO :0006974	Cellular Response To Dna Damage Stimulus	0.045
GO :0050731	Positive Regulation Of Peptidyl Tyrosine Phosphorylation	0.047

GO :0051346	Negative Regulation Of Hydrolase Activity	0.047
GO :0051961	Negative Regulation Of Nervous System Development	0.047
GO :2001233	Regulation Of Apoptotic Signaling Pathway	0.048
GO :0048870	Cell Motility	0.048
GO :0034976	Response To Endoplasmic Reticulum Stress	0.049
	Noyau 5	
GO :0048705	Skeletal System Morphogenesis	0.000
GO :0051276	Chromosome Organization	0.000
GO :0071495	Cellular Response To Endogenous Stimulus	0.000
GO :0048562	Embryonic Organ Morphogenesis	0.001
GO :0048729	Tissue Morphogenesis	0.002
GO :0001501	Skeletal System Development	0.002
GO :0018205	Peptidyl Lysine Modification	0.003
GO :0090092	Regulation Of Transmembrane Receptor Protein Serine Threonine Kinase Signaling Pathway	0.003
GO :0016569	Covalent Chromatin Modification	0.004
GO :1903844	Regulation Of Cellular Response To Transforming Growth Factor Beta Stimulus	0.004
GO :0009719	Response To Endogenous Stimulus	0.004
GO :0090100	Positive Regulation Of Transmembrane Receptor Protein Serine Threonine Kinase Signaling Pathway	0.004
GO :0071559	Response To Transforming Growth Factor Beta	0.004
GO :0016569	Chromatin Modification	0.005
GO :0002009	Morphogenesis Of An Epithelium	0.005
GO :0006325	Chromatin Organization	0.006
GO :0034622	Cellular Macromolecular Complex Assembly	0.007
GO :0030522	Intracellular Receptor Signaling Pathway	0.007
GO :0048598	Embryonic Morphogenesis	0.009
GO :0044419	Interspecies Interaction Between Organisms	0.010
GO :0051098	Regulation Of Binding	0.010
GO :0009792	Embryo Development Ending In Birth Or Egg Hatching	0.010
GO :0035295	Tube Development	0.011
GO :0043623	Cellular Protein Complex Assembly	0.012
GO :0071310	Cellular Response To Organic Substance	0.012
GO :0007179	Transforming Growth Factor Beta Receptor Signaling Pathway	0.013
GO :0040007	Growth	0.014
GO :0035239	Tube Morphogenesis	0.015
GO :0007389	Pattern Specification Process	0.016
GO :0070848	Response To Growth Factor	0.016
GO :0022402	Cell Cycle Process	0.019
GO :0001701	In Utero Embryonic Development	0.020
GO :0090287	Regulation Of Cellular Response To Growth Factor Stimulus	0.020
GO :0048568	Embryonic Organ Development	0.021
GO :0065003	Macromolecular Complex Assembly	0.022
GO :0007049	Cell Cycle	0.022
GO :0009725	Response To Hormone	0.023
GO :0090288	Negative Regulation Of Cellular Response To Growth Factor Stimulus	0.027
GO :0007178	Transmembrane Receptor Protein Serine Threonine Kinase Signaling Pathway	0.028
GO :0070271	Protein Complex Biogenesis	0.028
GO :0003205	Cardiac Chamber Development	0.031
GO :0032870	Cellular Response To Hormone Stimulus	0.032
GO :2000736	Regulation Of Stem Cell Differentiation	0.034
GO :0044248	Cellular Catabolic Process	0.035
GO :0018193	Peptidyl Amino Acid Modification	0.036
GO :0009952	Anterior Posterior Pattern Specification	0.036
GO :0071822	Protein Complex Subunit Organization	0.037
GO :1903320	Regulation Of Protein Modification By Small Protein Conjugation Or Removal	0.038
GO :0070647	Protein Modification By Small Protein Conjugation Or Removal	0.038
GO :0071383	Cellular Response To Steroid Hormone Stimulus	0.040
GO :0051099	Positive Regulation Of Binding	0.040
GO :0071407	Cellular Response To Organic Cyclic Compound	0.041
GO :0022411	Cellular Component Disassembly	0.041
GO :0006357	Regulation Of Transcription From Rna Polymerase Ii Promoter	0.044
GO :0051348	Negative Regulation Of Transferase Activity	0.046
GO :0003002	Regionalization	0.046
GO :0051169	Nuclear Transport	0.046
GO :0061061	Muscle Structure Development	0.047
GO :0006366	Transcription From Rna Polymerase Ii Promoter	0.047
GO :0033673	Negative Regulation Of Kinase Activity	0.048
GO :0006109	Regulation Of Carbohydrate Metabolic Process	0.049

Table 17 – *Table des termes GO significativement enrichis dans chaque noyau*

4.3 Table des concepts formels des gènes et des trajectoires

ID concept	Nb de gènes	Gènes	Nb de trajectoires
0	0		3068
1	2	PAX6, HES5	1
2	2	ID1, TLX2	3
3	3	RAB7A, SOCS3, TRPV1	10
4	3	PTGS2, SOCS3, NOS2	2
5	2	PTGS2, SOCS3	3
6	2	KIT, EPOR	54
7	1	SOCS3	20
8	1	CHRNE	22
9	2	EGR1, CHRNE	7
10	3	GADD45B, GADD45G, CCR5	62
11	1	EGR1	49
12	1	CCL2	45
13	2	EGR1, CCL2	27
14	2	CCL2, CHRNE	6
15	3	EGR1, CCL2, CHRNE	2
16	1	DUSP1	73
17	12	CISH, YWHAG, BCL2L1, SRP9, CCNA2, CDK6, FOXP3, OSM, LTA, PIM1, CNKSR1, PRF1	17
18	11	CISH, YWHAG, BCL2L1, SRP9, CDK6, FOXP3, OSM, LTA, PIM1, CNKSR1, PRF1	54
19	2	CASP1, BCL2L1	54
20	7	TNFSF11, CEBPD, BCL2L1, CRP, FGG, IRF1, JUNB	130
21	6	TNFSF11, CEBPD, BCL2L1, CRP, FGG, JUNB	141
22	5	CRP, BCL2L1, CEBPD, FGG, JUNB	143
23	6	TNFSF11, CEBPD, BCL2L1, FGG, IRF1, JUNB	137
24	5	BCL2L1, TNFSF11, CEBPD, FGG, JUNB	156
25	4	BCL2L1, CEBPD, FGG, JUNB	162
26	1	BCL2L1	342
27	1	GZMB	116
28	4	GZMB, IL1R1, CCL4, GZMA	62
29	1	PRF1	150
30	6	PPARGC1A, CCNA2, DUSP10, DUSP8, PDGFRA, INS	53
31	5	DUSP10, PPARGC1A, INS, DUSP8, PDGFRA	57
32	1	CCNA2	70
33	11	DUSP1, ACHE, HRK, BCL2L1, COL24A1, HES1, SELE, HBG2, IFNG, ATF3, ARG1	64
34	10	DUSP1, ACHE, HRK, BCL2L1, COL24A1, HES1, SELE, HBG2, ATF3, ARG1	72
35	12	DUSP1, ACHE, IL23A, HRK, BCL2L1, COL24A1, HES1, SELE, HBG2, IFNG, ATF3, ARG1	54
36	11	DUSP1, ACHE, IL23A, HRK, BCL2L1, COL24A1, HES1, SELE, HBG2, ATF3, ARG1	60
37	1	MMP2	62
38	9	MMP2, RB1, PPARGC1A, CCNA2, INS, DUSP5, DUSP10, PDGFRA, DUSP8	52
39	8	MMP2, RB1, PPARGC1A, INS, DUSP5, DUSP10, PDGFRA, DUSP8	56
40	14	CISH, YWHAG, BCL2L1, SRP9, BCL2, CCNA2, CDK6, FOXP3, OSM, LTA, PIM1, CNKSR1, PRF1, MYC	10
41	13	CISH, YWHAG, BCL2L1, SRP9, BCL2, CDK6, FOXP3, OSM, LTA, PIM1, CNKSR1, PRF1, MYC	18
42	4	TNFRSF4, GZMB, PRF1, TNFRSF18	54
43	9	MMP2, RB1, PPARGC1A, INS, DUSP5, DUSP10, TGFB2, PDGFRA, DUSP8	54
44	1	FASLG	45
45	5	IL18R1, ETV5, IRF1, IL18RAP, PRF1	42
46	5	CCL3, GZMB, IL1R1, CCL4, GZMA	44
47	6	CCL3, FASLG, GZMB, IL1R1, CCL4, GZMA	15
48	1	CXCL8	44
49	2	IFNG, CXCL8	37
50	1	IL5	35
51	2	IL2RA, CXCL8	32
52	3	IL2RA, IFNG, CXCL8	26
53	3	IL2RA, IL5, CXCL8	31
54	4	IL2RA, IFNG, IL5, CXCL8	25
55	1	ESR1	33
56	2	ESR1, TGFB1	21
57	3	IL2RA, CD40LG, CXCL8	31
58	4	IL2RA, IFNG, CD40LG, CXCL8	25
59	6	FASLG, IL4, CXCL8, IL2RA, CD40LG, IL5	30
60	7	IL4, FASLG, CXCL8, IL2RA, CD40LG, IFNG, IL5	24

61	1	IL10	6
62	2	IL10, IL5	4
63	10	MMP2, RB1, PPARGC1A, INS, DUSP5, DUSP10, TGFB2, PDGFRA, DUSP8, CSRP2	52
64	11	MMP2, RB1, PPARGC1A, CCNA2, INS, DUSP5, DUSP10, TGFB2, PDGFRA, DUSP8, CSRP2	51
65	12	MMP2, RB1, PPARGC1A, CCNA2, INS, DUSP5, DUSP10, TGFB2, SERPINB5, PDGFRA, DUSP8, CSRP2	44
66	13	MMP2, RB1, PPARGC1A, CCNA2, INS, DUSP5, DUSP10, TGFB2, JUN, SERPINB5, PDGFRA, DUSP8, CSRP2	25
67	16	GJA1, KRT5, ESR1, TGFB1, CYR61, MMP2, TCF4, ETS1, ACTA1, FOSL2, CXCL8, TIMP1, IL2, NR3C1, PENK, KRT17	6
68	17	GJA1, KRT5, ESR1, TGFB1, CYR61, MMP2, TCF4, ETS1, ACTA1, FOSL2, CXCL8, TIMP1, IL2, NR3C1, PENK, IFNG, KRT17	5
69	21	GJA1, KRT5, DMTF1, ESR1, TGFB1, CYR61, MMP2, TCF4, CDK1, ETS1, FOSL1, ACTA1, FOSL2, CXCL8, TIMP1, IL2, NR3C1, PENK, IFNG, KRT17, FABP4	1
70	1	IFNG	156
71	1	IRF1	233
72	1	JUN	790
73	1	TGFB1	1362
74	96	GZMB, HRK, CEBPD, CASP1, BCL2L1, ACTA1, CSRP2, DUSP1, IL4, CYR61, GADD45B, IL2, CDK6, TRPV1, FABP4, PPARGC1A, FOSL1, ETS1, FOSL2, LTA, SERPINB5, IFNG, ARG1, TLX2, KIT, IL1R1, TNFSF11, TGFB1, MMP2, RB1, COL24A1, GJA1, HES5, DUSP5, CD40LG, FASLG, ETV5, KRT5, CDK1, CHRNE, CXCL8, DUSP10, CRP, IL10, FGG, CNKSR1, HBG2, IRF1, TIMP1, NOS2, CCL2, IL18R1, PENK, PRF1, TNFRSF18, ACHE, ESR1, DUSP8, IL18RAP, YWHAG, HES1, PDGFRA, EPOR, JUNB, GZMA, CISH, DMTF1, BCL2, CCL4, OSM, SELE, IL2RA, RAB7A, GADD45G, JUN, CCR5, IL5, IL23A, TNFRSF4, ID1, SRP9, TCF4, PAX6, TGFB2, NR3C1, ATF3, KRT17, CCL3, PTGS2, EGR1, CCNA2, FOXP3, PIM1, INS, SOCS3, MYC	0

Table 19 – *Table des concepts formels des gènes et des trajectoires*