Modélisation dynamique de la signalisation cellulaire : aspects différentiels et discrets; application à la signalisation du facteur de croissance TGF-beta dans le cancer

Geoffroy Andrieux

N^o d'ordre : 00000 ANNÉE 2013

UNIVERSITÉ DE RENNES 1

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

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Biologie

École doctorale Vie Agro Santé

présentée par

Geoffroy ANDRIEUX

préparée à l'unité de recherche UMR INSERM U1085 – IRSET
Institut de Recherche en Santé, Environnement et Travail
en collaboration avec l'équipe Dyliss IRISA/INRIA
UFR SVE

Modélisation dynamique de la signalisation cellulaire : Aspects différentiels et discrets ; application à la signalisation du facteur de croissance TGF-β dans le cancer.

Thèse soutenue à Rennes
le 18 Juillet 2013

devant le jury composé de :

François FAGES
Directeur de recherche INRIA / *Rapporteur*

Laurent TRILLING
Professeur à l'Université Joseph Fourier Grenoble I / *Rapporteur*

Jérôme FERET
Chargé de recherche INRIA / *Examinateur*

Germain GILLET
Professeur à l'Université Claude Bernard Lyon I / *Examinateur*

Sophie PINCHINAT
Professeur à l'Université de Rennes 1 / *Examinatrice*

Nathalie THÉRET
Directeur de recherche INSERM / *Directrice de thèse*

Michel LE BORGNE
Maître de conférence à l'Université Rennes 1 / *Co-directeur de thèse*

Table des matières

Table des figures

Liste des tableaux

Première partie

Introduction

Chapitre 1

De la standardisation à la modélisation

Est complexe ce qui ne peut se résumer en un maître mot, ce qui ne peut se ramener à une loi, ce qui ne peut se réduire à une idée simple. C'est en ces mots qu'Edgard Morin définit la complexité dans son Introduction à la pensée complexe [100]. Cette définition sied parfaitement au vivant, et ce à toutes ses échelles : de la cellule aux populations en passant par les individus et les tissus qui les composent. La complexité émane non seulement du nombre de composants mais surtout de leurs interactions. De la complexité de ces réseaux d'interactions naît la notion de propriétés émergentes, qui ne peuvent être déduites de la simple connaissance de leurs composants et interactions locales. Depuis 20 ans la biologie a franchit un cap technologique qui a permit d'accéder à des quantités de données toujours plus importantes : l'intégralité du génome, mais aussi du transcriptome, du protéome et du métabolome. L'afflux de ces données confronte le biologiste à de nouveaux challenges pour stocker, intégrer et modéliser ces données afin de comprendre la complexité du vivant.

Génome Un réel changement a eu lieu suite au séquençage du génome Humain, aboutissement du Human genome project [144]. Malgré cette magnifique avancée dans la recherche en Biologie, la connaissance du génome ne répond pas à toutes les question posées. La complexité d'une cellule, *a fortiori* d'un individu, dépasse celle du génome et la majorité des processus biologiques ne sont pas associés à des variations du code génétique.

Transcriptome Le transcriptome englobe l'intégralité des ARN produits suite à la transcription des gènes d'une cellule. Si le génome permet de décrire la séquence génétique d'un individu, le transcriptome lui est spécifique à une cellule donnée dans une condition donnée. En effet le niveau d'expression d'un gène varie, au cours du développement ou d'une pathologie. Il est donc intéressant de connaître l'expression des gènes dans de multiples conditions pour ainsi inférer des relations de cause à effets entre ces variations et le

comportement phénotypique. Pour accéder à ces données, différentes techniques ont été développées à commencer par les microarray ou puce à ADN. Cette technologie permet de quantifier les ARN présents dans un échantillon (*i.e.* une population de cellule) par hybridation avec des séquences d'ADN. Plus récemment la technique RNA-seq, basé sur le séquençage haut débit, à supplanté l'utilisation des microarray [93, 151] notamment grâce à sa plus grande précision et la non-nécessité de connaître d'avance la séquence des ARN.

Protéome Le protéome représente l'ensemble des protéines, qui émanent de la traduction des ARN messagers et sont les entités clefs d'une cellule. Elles participent à tous les mécanismes essentiels, de la régulation des gènes, aux transports, en passant par la catalyse de réactions métaboliques. La diversité du protéome repose sur la traduction des ARN dont l'épissage alternatif permet la synthèse de variants pour une même protéine, ainsi que sur les régulations post-traductionelles (phosphorylation, ubiquitination, protéolyse, etc). L'identification haut débit de la composition protéique d'un échantillon a débuté avec la spectrométrie de masse [9] qui discrimine les protéines en fonction de leur masse. La méthode de purification par affinité en tandem permet d'extraire les interactants d'une protéine pour ensuite les identifier par spectrometrie de masse. Les différentes techniques permettant l'identification des interactions protéines-protéines et protéines-ADN sont présentées dans [148]. Les techniques d'analyse du protéome sont, au même titre que les autres, en constante évolution et leurs utilisations future sont discutées dans [2].

Métabolome Le métabolome regroupe l'ensemble des métabolites, molécules de petit poids moléculaire comme le glucose. A l'instar du trancriptome et du protéome, les études sur le métabolome ont pour but d'identifier et de mesurer quantitativement tout les métabolites d'une cellule et de déterminer la façon dont ils interagissent. L'obtention de ces informations peut se faire grâce à la spectrometrie de masse ou encore avec la technique de résonance magnétique nucléaire (NMR) [47].

L'acquisition de ces données haut débit, dans différentes conditions (ex : physiologiques *Vs.* pathologiques) et en fonction du temps, a réellement changé la vision de la biologie. Les études ne sont plus limitées à l'observation d'un gène donné, mais permettent aujourd'hui de considérer un système global avec de nombreux composants, interagissant de façon dynamique et entraînant un certain comportement. Ce changement de paradigme dans la biologie s'est concrétisé par l'apparition de la biologie des systèmes (ou biologie systémique) qui vise à étudier la complexité des interactions entre les composants biologiques prenant en compte les échelles de temps et d'espace.

De nombreuses définitions sont proposées, si bien que Christopher Wanjek à récemment écrit dans un rapport interne du National Institute of Health (NIH) :
Demandez à cinq astrophysiciens de définir un trou noir et vous obtiendrez cinq définitions différentes. Mais demandez à cinq chercheurs du domaine biomédical de définir la biologie des systèmes et vous obtiendrez dix définitions différentes... voir plus.

Ainsi Kirschner présente la biologie des systèmes comme *l'étude du comportement de processus biologiques complexes en terme de composants moléculaire* [73]. Pour Cassman, l'objectif de la biologie systémique est *la compréhension du comportement d'un réseau et en particulier de ses aspects dynamiques, ce qui requiert l'utilisation de modèles mathématique intimement liés aux expérimentations* [16]. Ces différentes définitions se regroupent sur certains aspects tel que la vision moléculaire ou l'aspect dynamique. Ainsi pour Kitano, 4 points sont essentiels aux approches systémiques [74] :

- **la structure du système** de manière à mettre en avant les différents acteurs, et la façon dont ils interagissent.
- **la dynamique du système**, pour observer comment le système se comporte au cours du temps.
- **le contrôle du système**, afin d'identifier les acteurs impliqués dans tel ou tel rôles (*e.g.* gènes responsable d'une pathologie).
- **la conception du système** dans le but de reproduire un système biologie à l'identique. Cette partie conduit à des réflexions sur la biologie synthétique [22], en dehors de la thématique de cette thèse.

Le vivant est hiérarchisé en différents niveau (ADN, ARN, protéines, ..., cellules, tissus, organismes), les composants sont capables d'interagir au sein d'un même niveau, et entre différents niveaux. Ainsi une partie de la biologie des systèmes, appelée biologie intégrative a vocation à intégrer les informations recueillies à ces différents niveaux.

Le vivant en tant que système complexe est généralement décomposé en trois réseaux : les réseaux de gènes, les réseaux métaboliques et les réseaux de signalisation [91] (Figure 1.1). Bien qu'ayant des caractéristiques propres, ces réseaux ne sont pas autonomes et de nombreuses relations existent entre eux : la signalisation régule l'expression des gènes, les protéines issues de l'expression des gènes interviennent dans les réactions métaboliques, etc.

La biologie des systèmes requiert l'utilisation de formalismes mathématiques et informatiques. Cependant, avant de pouvoir utiliser ces formalismes, il est nécessaire de trouver la ou les bonnes abstractions, c'est à dire un point de vue bien adapté à ce qui est modélisé dans le but d'utiliser des méthodes mathématique et/ou informatique [115]. Dans le domaine de l'analyse de séquences ADN, chaque acide nucléique est abstrait à une lettre

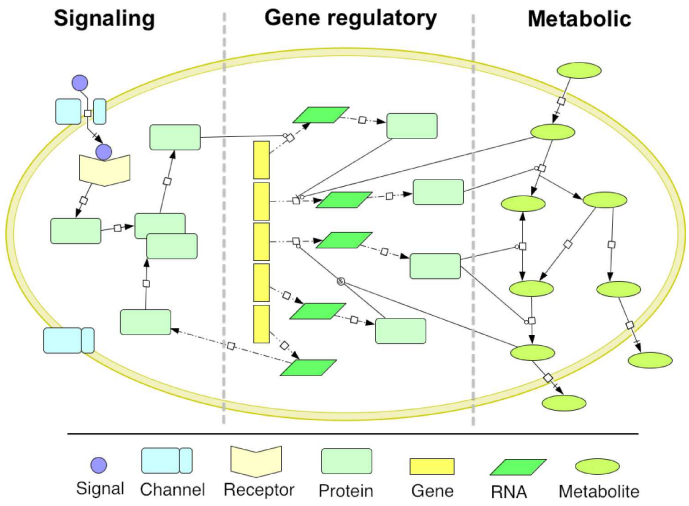


FIG. 1.1: Schéma des différents réseaux biologique d'après Machado *et al* [91]. Les réseaux de signalisation, décrivent l'ensemble des mécanismes permettant aux cellules de transmettre l'information et notamment de répondre à leur environnement. La propagation du signal fait intervenir différents types de réactions dont l'activation de récepteurs membranaire, la formation de complexe ou encore la régulation de gènes. La signalisation est généralement présentée sous forme de voies (TGF-β, EGF, NFKB, etc.) : enchainement quasi linéaire de réactions. Les réseaux de gènes regroupent les régulations d'expressions des gènes. Ces derniers peuvent être activés ou réprimés par des facteurs de transcription, eux même codés par des gènes. Les réseaux métaboliques comprennent l'ensemble des réactions biochimiques (transformation moléculaire et énergétique). Les réactions métaboliques sont régulées par des enzymes, protéines générées par le réseaux de gènes. Les réseaux métaboliques sont découpés en voies métaboliques (glycolyse, cycle de Krebs, etc.).

en fonction de sa base azotée : A pour Adénine, T pour Thymine, G pour Guanine et C pour Cytosine. Ce simple fait d'abstraire une molécule complexe à une lettre à permis de développer et d'appliquer de puissants algorithmes pour analyser des séquences d'ADN en les comparant, les alignant, etc.

En biologie des systèmes, différentes abstractions sont utilisées pour représenter les connaissances et en tirer de nouvelles informations. Nous présentons ici les étapes essentielles aux approches systémiques, permettant le stockage (section 1.1) dans des bases de données et surtout l'interprétation (section 1.2) des données dans des modèles (section 1.3).

1.1 Stockage de l'information

Les bases de données ont pour but de stocker les données de façon à les rendre accessibles à la communauté. Ces données ont vocation à être utilisées pour des analyses statistiques, ou pour concevoir des modèles. Au cours de ces dernières années, la production de données à connue une forte inflation, conduisant à des données parfois redondantes. Pour faire face à cela, les bases ont dû s'adapter et proposer un contenu et une interface permettant aux utilisateurs de naviguer dans cette masse de données. En fonction du contenu et de la façon dont sont structurées les données, il est possible de différencier des catégories de bases de données.

1.1.1 Bases de données moléculaires

Ces bases de données émergent directement des suites du séquençage et autres techniques haut débit. Elles ont pour but de stocker les noms, les séquences et les annotations fonctionnelles associées à chaque gène ou protéine. Les plus référencées sont celles hébergées par le NCBI (National Center for Biotechnology Information), telles que Genbank [11] et par l'EBI (European Bioinformatics Institute), telles que Ensembl [40] ou Uniprot [6]. Ces bases sont dédiées soit à des séquences nucléiques, protéiques, des structures tridimensionnelle de protéines ou encore des informations sur des molécules de petits poids moléculaires [30]. Ces bases sont spécialisées sur un type de donnée (ADN, proteine,...) mais sont généralistes du point du vue des espèces, dans le sens où elles regroupent des informations de tout organisme vivant étudié. D'autres sont en revanche dédié à une espèce, comme les bases Human Protein Reference Database (HPRD) [69] ou Human Metabolome DataBase (HMDB) [152] respectivement pour les protéines et les métabolites chez l'humain. Les données d'expression ont également leurs bases dédiées comme Gene Expression Omnibus (GEO) [10] ou GEMMA [159], de même que les données d'interactions avec Chip Enrichment Analysis (ChEA) [83]. Certaines initiatives proposent de faire le lien entre ces données de différentes échelles. C'est le cas de la base Genecards [126] qui dresse une carte d'identité pour chaque gène avec des informations diversifiées telles que la sequence, les fonctions ou encore les données d'expression. Genecards permet également de diriger l'utilisateur vers les bases de données présentées précédemment. L'état actuel des base de données moléculaire est présenté dans [39] où plus de 1500 bases sont répertoriées.

1.1.2 Bases de données d'interactions moléculaires

En biologie des systèmes, nous avons besoin de connaître la manière dont les protéines, et autres molécules interagissent entre elles pour former un tout cohérent. Ce type d'in-

formation n'est pas décrit dans les bases précédentes. Pour combler ce manque, des bases de données dédiées aux interactions entre différentes molécules ont vu le jour et sont aujourd'hui en plein essor. En effet en 2006, Bader *et al* répertoriaient 190 bases de ce type sur leur site pathguide (`http://www.pathguide.org`) [8], il en existe aujourd'hui plus de 320 présentées sur ce même site.

Ces interactions binaires ou multiples concernent majoritairement des interactions entre protéines. Cependant on y retrouve également des interactions protéines-ARN, et protéines-métabolites. Ces informations sont généralement obtenues par des manipulations type double hybride ou purification par affinité suivi d'une spectrométrie de masse pour les interactions protéines-protéines, et par immunoprécipitation de la chromatine pour les interactions protéines-ADN. Certaines bases de données se veulent assez généralistes et réunissent des interactions chez différentes espèces, entre différentes molécules (Mint [17], Intact [68], BioGrid [135], DIP[128]...).

D'autres en revanche sont beaucoup plus spécialisées, sur un compartiment particulier comme MatrixDB [19] qui répertorie les interactions uniquement à l'extérieur des cellules, ou PIN [89] pour les interactions dans le noyau. Il existe également des bases spécialisées sur un type d'interactions tel que PhophoPOINT [155] qui répertorie des phosphorylations. Le Tableau 1.1 présente les caractéristiques des bases de données d'interactions les plus utilisées.

Base de données	MINT	IntAct	BioGRID	DIP	MatrixDB	PhosphoPOINT
Nombre d'interactions	241458	305970	638453	75019	2283	15738
Nombre d'interactants	35606	64569	47972	25388	8679	4195
Nombre de publications	5427	6177	37954	5857	-	-
Types d'interactions	PP	PP, PG, PM	PP, PG	PP	PP, PM	PP (phosphorylation)
Formats	PSI-MI	PSI-MI, BioPAX ...	PSI-MI	PSI-MI	PSI-MI, MITAB	Tabular
Organismes	> 30	> 10	> 30	541	Mammifères	Humain

TAB. 1.1: Tableau comparatifs des principales bases de données d'interactions moléculaires. PP : protéine-protéine, PG : protéine-ADN, PM : Protéine-métabolite.

1.1.3 Bases de données de réactions

Faire l'inventaire de tout les composants d'une cellule, et de leurs interactions locales ne suffit généralement pas à comprendre le fonctionnement intrinsèque de la cellule. Un parallèle peut être fait dans le domaine de la mécanique où lister toutes les pièces d'un moteur ne permet pas de comprendre la manière dont elles interagissent à l'échelle du moteur. Savoir que deux protéines forment un complexe n'explique pas un comportement, il faut replacer le complexe formé dans son contexte et se demander quel est son rôle et son devenir. C'est pourquoi, des concepts plus élaborés doivent permettre de donner une signification biologique. En ce sens, le terme réaction regroupe aussi bien les réactions biochimiques, métaboliques ou biologiques. Ainsi des bases spécialisées dans certains types de réactions ont vu le jour parmi lesquelles : KEGG [67] et HumanCyc [121] pour les réactions métaboliques, PID [131] et NetPath [66] pour la signalisation. D'autre comme Reactome [27] regroupe tout type de réactions. Le contenu de ces bases est résumé dans le Tableau 1.2.

Base de données	KEGG	Reactome	PID	NetPath	HumanCyc
Nombre d'interactions	9622	9060	9248	2567	2443
Nombre d'interactants	6326	15782	27876	1469	27019
Types de réactions	Biochimiques	Biologiques	Biologiques	Biologiques	Métabolique
Formats	-	BioPAX, SBML, PSI-MI ...	BioPAX, PID xml	BioPAX, SBML, PSI-MI ...	BioPAX, SBML, Tabular ...

Tab. 1.2: Tableau comparatifs des principales bases de données de réactions.

Ces bases présentent généralement un contenu moins conséquent que les bases d'interactions moléculaire car l'annotation d'une réaction est plus complexe à rentrer dans une base de donnée que l'interaction entre 2 composants. La base de données KEGG fut l'une des premières à regrouper les informations sous forme de voies. Cependant des réactions très précises comme la phosphorylation y côtoient des concepts plus flous comme l'inhibition. Plus récemment des bases comme PID ou Reactome se sont efforcées d'homogénéiser la représentation de leurs contenus en définissant des concepts de réactions plus précisément et utilisant des standards de représentation qui seront détaillés par la

suite.

1.1.4 Bases de données agglomératives : Méta bases de données

Afin de tirer parti de la richesse d'information contenu dans les bases de données d'interactions et de réactions, de nouvelles bases de données ont été développées pour intégrer les données des bases spécialisées. C'est le cas des bases de donnée telles que Search Tool for the Retrieval of Interacting Genes (STRING) [138], GeneMania ou encore VISANT [59]. En rassemblant les connaissances de dizaine de bases, et des prédictions d'interactions, STRING couvre actuellement plus de 1000 espèces et plus de 5 millions de protéines ayant des millions d'associations. En utilisant la base de données STRING, il est possible de retrouver facilement les associations les plus significatives à partir d'une protéine ou d'un ensemble de protéines. Le terme association regroupe à la fois les interactions physiques directes comme la complexation mais aussi les interactions fonctionnelles indirectes (co-expression, analyse textuelle, etc.). La Figure 1.2 illustre les réseaux obtenus en recherchant les associations liées à la protéine de signalisation TGFB1 (facteur de croissance transformant). Les preuves sur lesquelles sont basées les associations peuvent être affichées (Figure 1.2A). Ces preuves regroupent les différents moyens utilisés pour identifier une association, des données de co-expression à l'analyse de données textuelles en passant par la proximité des gènes sur le génome. Pour ce même jeux de protéines associée au TGFB1, le type d'interaction est illustré en (Figure 1.2B). Là encore les actions sont multiples : activation, inhibition, catalyse...Les associations sont délivrées en fonction de leurs scores, il est possible d'étendre le réseau obtenu avec un score moins stringent. Des études récentes ont utilisé STRING pour l'analyse fonctionnelle des protéines cibles de l'acétylation [23] ou encore pour l'analyse inter-espèces de connectivité d'un réseau [80].

Dans la même volonté d'intégrer des informations dispersées dans les bases actuelles, des bases comme Pathway Commons [18], hiPathDB [156] et ConsensusPathDB [65] regroupe des réactions provenant de différentes bases. ConsensusPathDB regroupe 30 bases de données orientées réactions telles que Reactome [27], PID [131], KEGG [67], Biocarta, mais aussi les bases d'interaction comme MINT, INTACT. ConsensusPathDB comptabilise ainsi 360 308 interactions entre 154 540 molécules. Les identifiants sont homogénéisés entre les différentes bases de façon à pouvoir assembler le contenu. Les protéines sont désignées par leur identifiant dans la base de donnée Uniprot, les gènes et transcrits par leur identifiant Ensembl, et les métabolites par leur numéro d'accession dans KEGG/ChEBI [30]. Les réactions sont elles aussi assemblées, si elles ont les mêmes molécules en entrée et en sortie. En revanche ni la localisation ni les modifications post-traductionelles pour les protéines, ne sont prises en compte pour définir si 2 protéines issues de 2 bases différentes

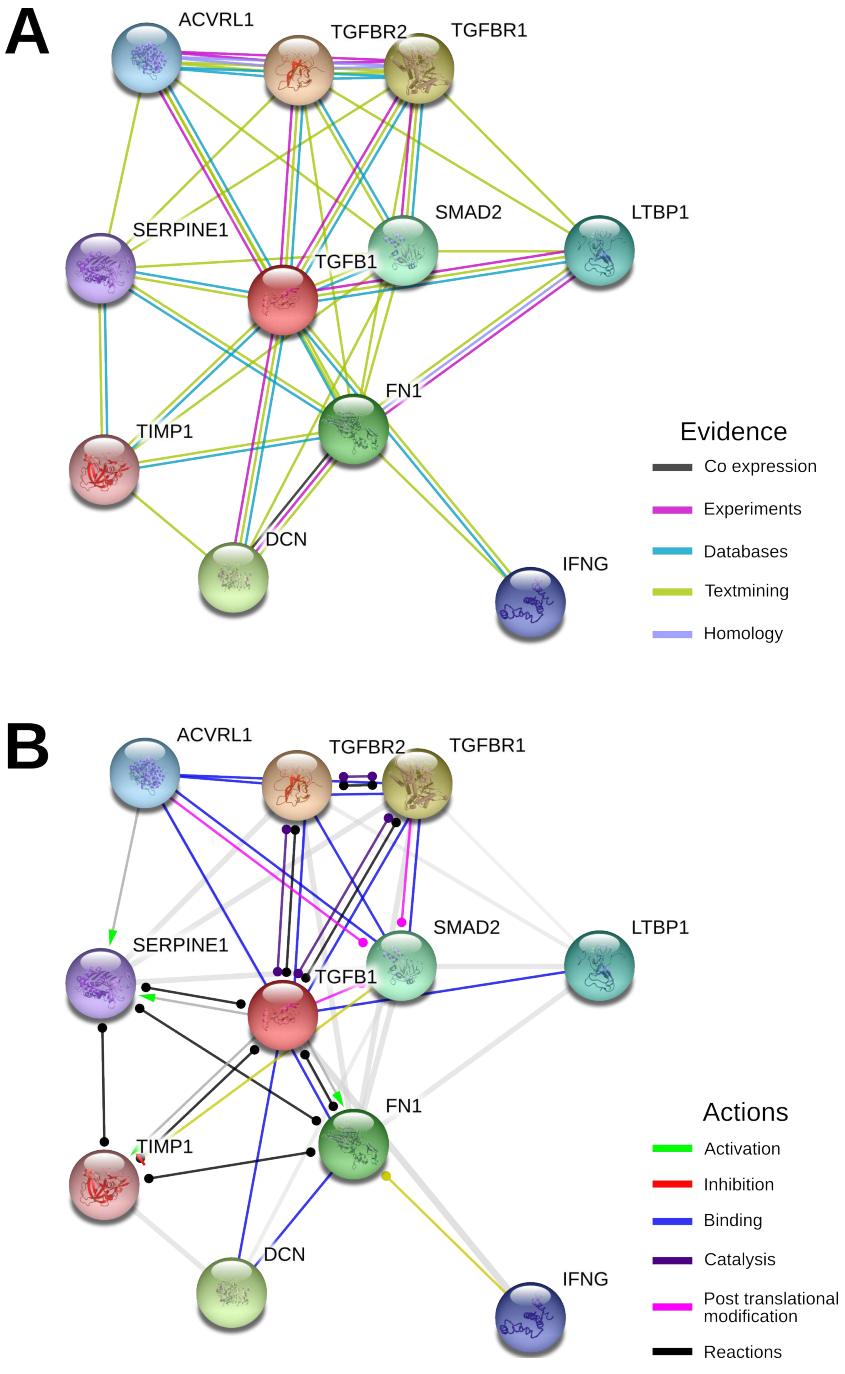


FIG. 1.2: Utilisation de la base STRING pour identifier les protéines les plus liées au TGFB1. Les sphères représentent les protéines liées au TGFB1. Les arcs peuvent figurer les évidences qui ont permis d'établir le lien (**A**) ou le type de liens qui unit ces protéines (**B**). Ces arcs sont distingués par un code couleur et peuvent être filtrés pour cibler le type d'évidence et/ou d'actions à prendre en compte.

font référence à une unique entité. Une faible proportion de molécules et réactions est partagée entre plusieurs bases de données, si bien que 75% des réactions ne sont pas retrouvées dans au moins 2 bases de données. Ce chiffre montre par conséquence une forte complémentarité du contenu des bases de données et l'utilité de travaux permettant de regrouper l'information.

1.2 Standardisation des connaissances

La représentation ambiguë des connaissances en biologie est un frein pour les approches systémiques. Dans un premier temps, les annotations des molécules ont été générées de façon quasi-spécifique à une base de donnée, avec par exemple un identifiant interne à chaque base. La multiplication des différents identifiants tend à complexifier la compréhension et le partage des données. Afin d'homogénéiser ces connaissances, l'organisation internationale HUman Genome Organisation (HUGO) a défini une nomenclature pour les noms et symboles de plus de 34000 genes [48]. En parallèle, des convertisseurs ont été développés pour permettre une meilleure compatibilité des données issues de différentes bases (Babelomics [96], DAVID id converter [60]).

Au delà des noms et identifiants, les annotations des molécules concernent également leurs fonctions, leurs localisations. Il était important de clairement définir une hiérarchie entre ces annotations. L'initiative Gene Ontology (GO) a répondu à ce besoin en proposant une ontologie des processus biologiques, fonctions moléculaires et composants cellulaires [7]. Un produit de gène est alors associé à un ou plusieurs termes GO. Exemple du facteur de croissance transformant (TGFB1) qui est associé à 165 termes GO chez l'humain, parmi lesquels l'arrêt du cycle cellulaire dont l'ontologie est illustrée en Figure 1.3. Les termes sont reliés par des relations de type "est un", "fait partie de" permettant ainsi de structurer les annotations. Annoter les molécules permet de retrouver facilement toutes celles qui sont associées à un terme donné. Par exemple 1032 produits de gènes sont associés au termes *arrêt du cycle cellulaire*.

Notons également les efforts pour standardiser les informations dès l'acquisition des données avec notamment : Minimum Information About A Proteomics Experiment (MIAPE) [140] pour la génération de données autour du protéome, Metabolomics Standards Initiative (MSI) [88] ou encore Minimum Information required to report a Molecular Interaction Experiment (MIMIx) [108] pour les interactions. De même le consortium International Molecular Exchange consortium (IMEx) [107] a proposé de standardiser le niveau de détail des interactions présentes dans les bases et ses recommandations sont actuellement suivies par de nombreux projets.

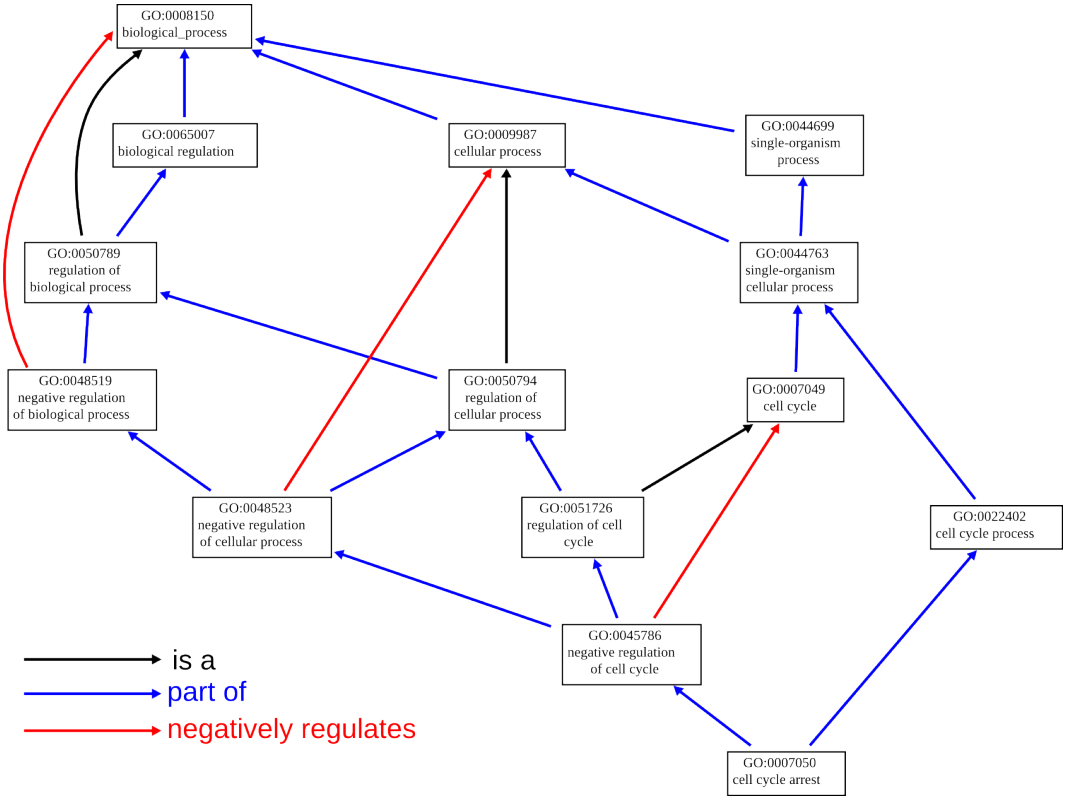


FIG. 1.3: Hiérarchie des ontologies du terme *cell cycle arrest* d'après Gene Ontology. Chaque noeud représente un terme Gene Ontology avec son nom et son identifiant. Les noeuds sont reliés par des arcs de couleurs différentes suivant le type de relation qui les unissent parmi : *is a* (est un), *part of* (fait partie de) et *negatively regulates* (regule négativement). Cette organisation structure la connaissance autour des termes utilisés pour annoter les protéines.

Afin de partager les informations sur les interactions, il faut aussi standardiser la façon de les représenter. Lorsque des interactions entre des molécules sont décrites schématiquement, une simple flèche peut être interprétée de multiple façons (activation, complexation, modification, etc.). L'interprétation est subjective et dépend énormément des connaissances de l'utilisateur. Des représentations graphiques ont vu le jour pour homogénéiser la façon dont sont présentées les connaissances sur ces interactions et réactions : Molecular Interaction Map (MIM) [78, 79], Process Diagram [76] ou encore SBGN [84]. Une représentation graphique doit pouvoir (i) différencier les processus biologiques, (ii) être interprétée sans ambiguïté, (iii) être soutenue par des logiciels, facilitant la libre création/édition par la communauté.

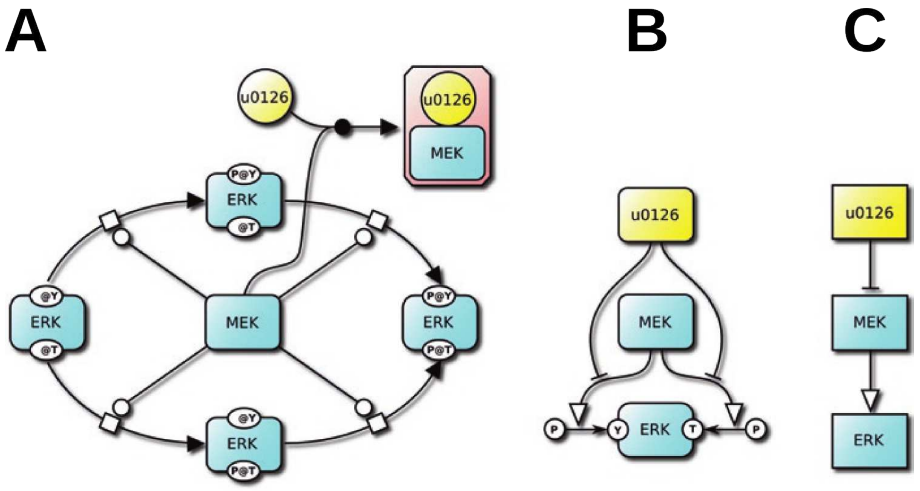


FIG. 1.4: Illustration d'un ensemble de réactions sous les 3 vues de SBGN (d'après [84]). Ces réactions représentent une partie de la signalisation des MAPK (Mitogen-activated protein kinases). La protéine ERK peut être phosphorylée sur 2 sites différents par MEK. MEK peut être inhibée par u0126. (**A**) la vue diagramme de processus explicite tout les états de ERK suivant la phosphorylation de ses sites. Le mécanisme d'inhibition par captation de MEK est clairement représenté. (**B**) la vue diagramme relations d'entités ne représente qu'une seule protéine ERK et MEK agit comme activateur de la phosphorylation des 2 sites de ERK. Le mécanisme d'inhibition n'est pas représenté, u0126 inhibe la phosphorylation de ERK par MEK, représenté par une flèche. (**C**) la vue diagramme flux d'activité est la plus abstraite. La phosphorylation est abstrait à l'activation de ERK par MEK. De la même façon, u0126 n'inhibe plus la phosphorylation mais la protéine MEK elle même, l'empêchant ainsi d'activer (phosphoryler) ERK.

Le *Diagramme de Processus* proposé par Kitano *et al* [76] rempli ces conditions. Il propose de représenter les connaissances sous forme de réactions. Pour cela différents symboles sont disponibles pour représenter différentes molécules (protéines, récepteurs, gènes, ions, etc.) et réactions (changement d'états, association, translocation, etc.). Le niveau de détail offert par cette représentation permet notamment de définir des résidus sur des molécules, en distinguant par exemple les différents niveaux d'activation d'une protéine en fonction de ses sites phosphorylés ou non.

Le diagramme de processus est utilisé par différents logiciels tel que CellDesigner [41] qui permet la conception de réseaux de réactions sous ce format. Cette représentation a été appliquée à de grand réseaux biologiques tels que la signalisation des récepteurs de type TOLL (TLR : Toll-Like Receptors) [106] et le facteur de croissance épidermique (EGF : Epidermial Growth Factor) [105].

Cependant le diagramme de processus ne constitue pas une représentation universelle des processus biologiques et un consortium de chercheurs a récemment proposé une approche plus adapté à la diversité des mécanisme biologiques. Ainsi le projet *Systems Biology Graphical Notation* (SBGN) a défini la notion de "vues" d'un système biologique [84]. Un même système peut être perçu de différents points de "vues", en fonction des données, de l'interprétation et du degré d'abstraction. Forts de leurs réflexions passées sur les précédentes représentations, Le Novere *et al* proposent 3 vues orthogonales et complémentaires :

- Le diagramme de processus (Figure 1.4A) décrit précisément les différentes entités, les sites de réactions et les localisations. Cette vue reprend les caractéristiques proposées par Kitano *et al* [76] décrites précédemment.
- Le diagramme relations d'entités (Figure 1.4B) dépeint les relations entre les entités et leurs conditions de réalisation. Les molécules n'apparaissent qu'une seule fois, leurs différents états sont implicites. Les mécanismes de régulation sont abstraits à des relations.
- Le diagramme flux d'activité (Figure 1.4C) représente de façon plus abstraite les influences entre les entités du système. Cette représentation demande d'avantage d'interprétation pour être générée. Les détails des réactions biochimiques n'apparaissent plus. Le système obtenu est plus petit, mais en contrepartie les mécanismes fins des réactions sont perdus.

SBGN a rapidement été adopté par la communauté de la biologie des systèmes. Cette représentation est en effet utilisé par différentes bases tel que Reactome, Panther [98] ou BioModels [85] pour visualiser leurs contenus.

En complément de ces représentations graphique, des formats textuels de type XML ont été développés pour permettre l'échange des données structurées. Comme indiqué dans le Tableau1.2, la plupart des bases de données récentes proposent d'exporter leur contenu dans des formats communs tels que BioPAX [31] et *Systems Biology Markup Language* (SBML)[61] pour les réactions et *Proteomics Standards Initiative Interaction* (PSI-MI) [57] pour les interactions. Cependant ces formats ne sont pas équivalents et ne permettent pas toujours de traduire les mêmes faits biologiques. Bien qu'il existe des convertisseur entre ces formats, la conversion n'est pas parfaite et une perte de donnée est souvent observée. Les Figure 1.5 et 1.6 comparent les données structurées suivant le format BioPAX ou SBML pour le complexe entre les protéines SMAD2/3 et SMAD4 (Figure 1.5) et la représentation de l'inhibition de la transcription du gène MYC (Figure 1.6). A partir d'une même source de données, les informations peuvent être représentées de manière différente dans ces deux formats. Dans cet exemple, BioPAX semble mieux adapté à la description

de voie de signalisation, en permettant la représentation de concept comme l'inhibition.

Le format PSI-MI est utilisé pour les interactions locales, individuelles entre protéines. Il n'a pas vocation à retranscrire des réactions à l'échelle d'un système. Contrairement à SBML qui à l'origine a été conçu pour représenter des systèmes différentiels. Il permet ainsi d'y stocker des données quantitatives avec des coefficient de réactions et des concentrations. Depuis ces dernières révisions, SMBL offre également la possibilité de représenter des réactions qualitatives. Cependant des concepts importants tel que l'inhibition reste plus délicat à décrire dans ce formalisme. A l'inverse dans BioPAX le concept de réaction biologique peut y être représenté plus intuitivement. BioPAX utilise une hiérarchie de classes pour représenter entités et réactions. Cette hiérarchie permet de bien structurer le rôle de chaque entité dans une réaction, y compris pour les régulateurs, et d'y ajouter bon nombre d'informations importantes pour le détail des réactions (localisation, état d'activation des protéines...). Néanmoins les formats BioPAX et SBML ne permettent pas encore une représentation consensuelle des données. Malgré une sémantique assez bien définie, surtout pour BioPAX, il est toujours possible de représenter les mêmes données de différentes façons, compliquant l'analyse ultérieure des fichiers. Dans leur article, Strömbäck *et al* font la comparaison entre les formats BioPAX, SBML et PSI-MI [137]. Ces formats doivent avant tout être vus comme différents moyens d'échanges au sein de la communauté, en ayant conscience des forces et faiblesses de chacun. L'utilisateur doit interpréter le contenu de la base de donnée de façon identique à celui qui a généré le fichier.

BioPAX

```
<bp:Complex rdf:ID="Complex1">
    <bp:displayName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">p-2S-SMAD2/3:SMAD4</
bp:displayName>
    <bp:name rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Phospho-R-SMAD:CO-SMAD complex</
bp:name>
    <bp:cellularLocation rdf:resource="#CellularLocationVocabulary1" />
    <bp:componentStoichiometry rdf:resource="#Stoichiometry2" />
    <bp:component rdf:resource="#Protein2" />
    <bp:componentStoichiometry rdf:resource="#Stoichiometry3" />
    <bp:component rdf:resource="#Protein1" />
    <bp:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Reactome DB_ID: 171175</
bp:comment>
    <bp:xref rdf:resource="#UnificationXref15" />
    <bp:xref rdf:resource="#UnificationXref16" />
    <bp:dataSource rdf:resource="#Provenance1" />
  </bp:Complex>
```

SBML

```
<species id="species_171175" name="p-2S-SMAD2/3:SMAD4 [cytosol]" metaid="metaid_36"
sboTerm="SBO:0000253" compartment="compartment_70101">
<notes><p xmlns="http://www.w3.org/1999/xhtml">Derived from a Reactome Complex.  Reactome uses a
nested structure for complexes, which cannot be fully represented in SBML.</p></notes>
        <annotation>
          <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:bqbiol="http://
biomodels.net/biology-qualifiers/" xmlns:bqmodel="http://biomodels.net/model-qualifiers/">
            <rdf:Description rdf:about="#metaid_36">
              <bqbiol:is>
                <rdf:Bag>
                  <rdf:li rdf:resource="urn:miriam:reactome:REACT_7344"/>
                </rdf:Bag>
              </bqbiol:is>
              <bqbiol:hasPart>
                <rdf:Bag>
                  <rdf:li rdf:resource="urn:miriam:uniprot:Q15796"/>
                  <rdf:li rdf:resource="urn:miriam:uniprot:P84022"/>
                  <rdf:li rdf:resource="urn:miriam:uniprot:Q13485"/>
                </rdf:Bag>
              </bqbiol:hasPart>
            </rdf:Description>
          </rdf:RDF>
        </annotation>
      </species>
```

FIG. 1.5: Représentation du complexe p-2S-SMAD2/3 :SMAD4 dans la base de données Reactome. Les formats BioPAX et SBML sont de type XML et permettent de structurer les informations par des balises pour faciliter le traitement automatique de ces fichiers. Le nom du complexe est encadré en rouge, l'identifiant du compartiment cellulaire en bleu, et les identifiants des composants en jaune. A noter que des différences existent entre ces fichiers représentant le même complexe. Les noms ne sont pas les mêmes et les composant ne sont pas décrit de la même manière. Le format BioPAX permet de spécifier qu'il s'agit d'un complexe contrairement au format SBML.

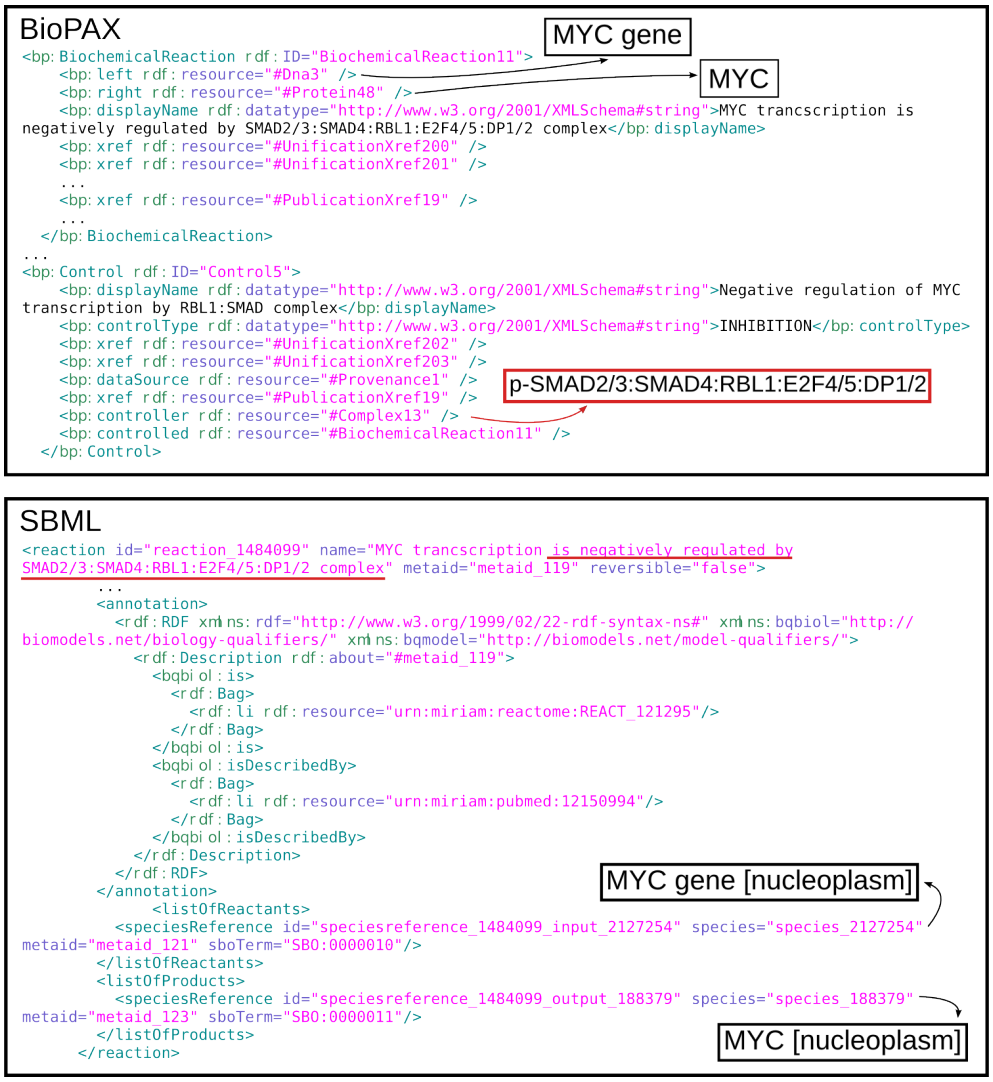


FIG. 1.6: Représentation de l'inhibition de la transcription du gène MYC par le complexe p-SMAD2/3 :SMAD4 :RBL1 :E2F4/5 :DP1/2. Les réactions se distinguent des réactants par des balises xml spécifiques. L'inhibition de la réaction n'est mentionnée que dans le nom de la réaction (soulignée en rouge) pour le format SBML. Dans le fichier BioPAx il est possible de retrouver l'information (encadrée en rouge) grâce à la balise *Control*. Les noms des réactants et produits sont encadrés en noir.

1.3 Modélisation

Un membre éminent de la communauté de la modélisation des systèmes biologiques, Trey Ideker à écrit : *Un modèle biologique commence dans l'esprit d'un chercheur, comme un mécanisme proposé pour expliquer des observations expérimentales* [62].

Ainsi un modèle est une description formelle, mathématique, d'un système permettant de faciliter sa compréhension dans l'optique d'apporter de nouvelles connaissances. Pour être efficace, un modèle doit être adapté à la biologie, aux données disponibles et aux questions à poser. Il existe deux grandes catégories de modèles : les modèles statiques basés sur l'analyse topologique et les modèles dynamiques permettant de représenter l'évolution des systèmes. La représentation sous forme de graphe est la plus couramment utilisée pour représenter ces modèles du fait de sa notoriété en informatique et de son nombre élevé de méthodes d'analyses dédiées. Avant d'expliciter les différentes classes de modèles, il convient de définir le vocabulaire utilisé en théorie des graphes.

1.3.1 Théorie des graphes en biologie

Un graphe G est une paire (V, E) où V désigne l'ensemble des sommets (ou noeuds) du graph, et E désigne l'ensemble des arêtes (ou arcs) représentant les relations entre les noeuds. Les noeuds représentent généralement les entités biologiques telles que : les protéines, les ARN, les gènes...mais aussi les réactions. Les relations peuvent désigner aussi bien des interactions physiques, telles que l'association de 2 protéines pour former un complexe, que des liens plus abstraits (indirects) comme la co-expression, l'appartenance à un même type cellulaire. Il existe différents types de graphes, illustrés en Figure 1.7. Les graphes non dirigés (ou non orientés), où les arêtes ne sont pas orientées, sont typiquement utilisés pour représenter des interactions protéines-protéines (Figure 1.7A). Les graphes dirigés (ou orientés) où les arêtes sont orientées. On parle alors d'arcs ou flèches d'un noeud A vers un noeud B. Les graphes dirigés sont utilisés pour représenter les influences ou les réactions (Figure 1.7B). Les arêtes des graphes dirigés ou non, peuvent porter des poids, généralement utilisés pour exprimer la pertinence de la relation, mais également le signe d'une influence (positive ou négative) ou le coefficient stœchiométrique d'une réaction (Figure 1.7C). Les graphes peuvent également posséder plusieurs ensembles de noeuds distincts, on parle alors de graphe bi-partite, ou un ensemble de noeuds désignent des biomolécules et l'autre l'ensemble des réactions (Figure 1.7D). En plus de permettre une description souvent attrayante et condensée des connaissances, la représentation sous forme de graphe permet de rechercher bon nombre de propriétés basées sur leurs topologies.

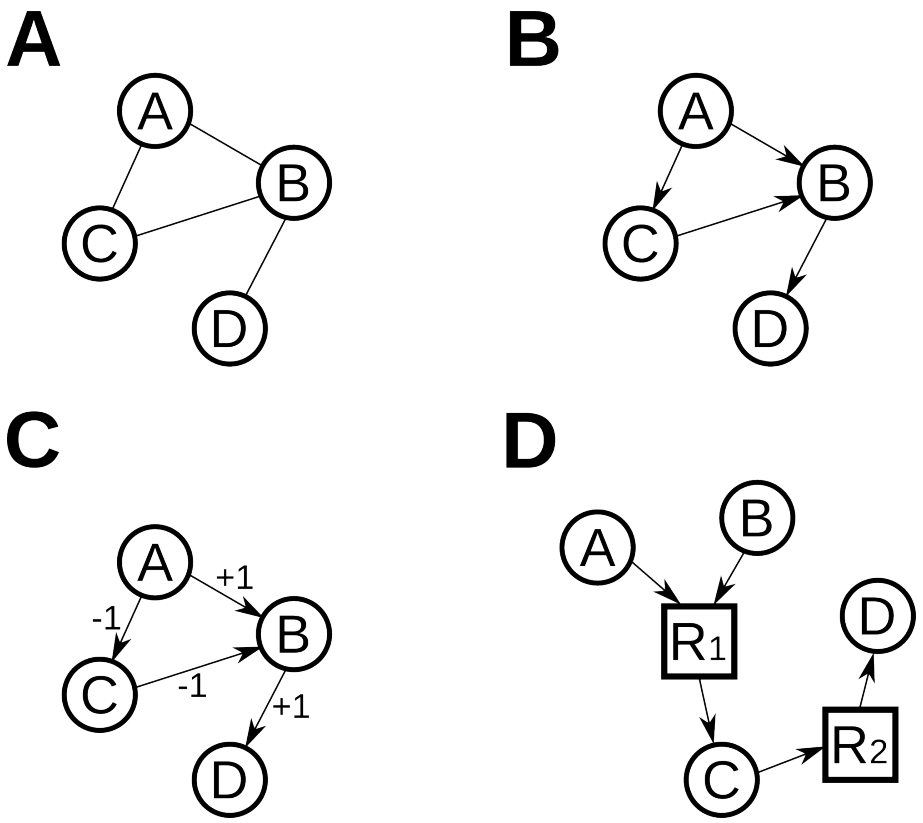


FIG. 1.7: Différents types de graphes retrouvés en biologie des systèmes. (**A**) Graphe non-orienté. L'information entre A et B est la même que celle entre B et A. Ce genre de graphe permet de représenter notamment des co-expressions entre les gènes. (**B**) Graphe orienté. L'information portée par les arc est orientée d'un noeud dit origine vers un noeud cible. Cette représentation est utilisé pour représenter des influences. Dans la figure le noeud A influence les noeuds B et C. (**C**) Graphe orienté avec étiquettes. Les arcs portent une étiquette qui suivant son utilisation peut être interprétée comme l'intensité d'une influence, un coefficient stœchiométrique d'une réaction. Si les arcs représentent des influences, le noeud A influence positivement (+1) le noeud B, au contraire du noeud C qui l'influence négativement (-1). (**D**) Graphe bi-partite orienté. Ce graphe permet de distinguer 2 types de noeuds pouvant servir par exemple à différencier molécules et réactions. Ici les noeuds A, B, C et D s'apparentent à des molécules, et R_1, R_2 à des réactions. La réaction R_1 consomme les molécules A et B et produit la molécule C.

1.3.2 Modèle statique : comprendre l'organisation du système

La plupart des bases citées précédemment proposent de représenter leurs données sous forme de graphes où les noeuds sont les molécules et les arcs entre ces noeuds peuvent représenter toute sortes de relations entre ces entités.

Les analyses statiques sont basées sur les propriétés topologiques des graphes. Les analyses et propriétés recherchées dépendent évidemment du sens donné aux noeuds et aux arcs. Les possibilités sont nombreuses et dépendent des paramètres précédemment cités. Nous dressons ici la liste non exhaustive des principales propriétés topologiques recherchées sur les graphes.

Calcul de la centralité La centralité est une mesure locale (pour chaque noeud) et permet de mettre en évidence les noeuds important d'un graphe (d'un point de vue topologique). On distingue différentes centralités (Figure 1.8) dont :

- Centralité de degré : défini pour un noeud v_i le nombre de noeud auquel v_i est directement relié. Dans les graphes orientés, il est possible de distinguer le degré entrant (nombre d'arcs entrants) et le degré sortant (nombre d'arcs sortants).
- Centralité de proximité (closeness centrality) : calcul pour un noeud v_i la somme des plus courts chemins entre v_i et tous les autres noeuds du graphe.
- Centralité d'intermédiarité (betweenness centrality) : calcul pour un noeud v_i le nombre de plus courts chemins entre toutes les combinaisons de noeuds possibles, passant par v_i. Une forte centralité d'intermédiarité est liée à un noeud trés utilisé dans le graphe.

Ces techniques sont appliqués pour identifier de nouveaux complexes proteiques [87] ou encore mettre en évidence une corrélation entre la centralité des protéines et leur caractère essentiel chez la levure [64].

Recherche de motifs Un motif est un profil retrouvé avec une fréquence significativement plus élevée que la fréquence prévue dans un graphe aléatoire. Ces motifs ont une réelle signification en biologie et leurs identifications peut permettre de mettre en évidence des ensembles fonctionnels (boucle de rétro-contrôle, protéines régulatrices) [1, 99, 153].

Recherche de cluster Les techniques de clustering sur les graphes permettent de mettre en évidence les groupes extrêmement connectés. Dans les graphes d'interactions protéine-protéine, la recherche de cluster a notamment permis la prédiction de macro-complexes protéiques [90].

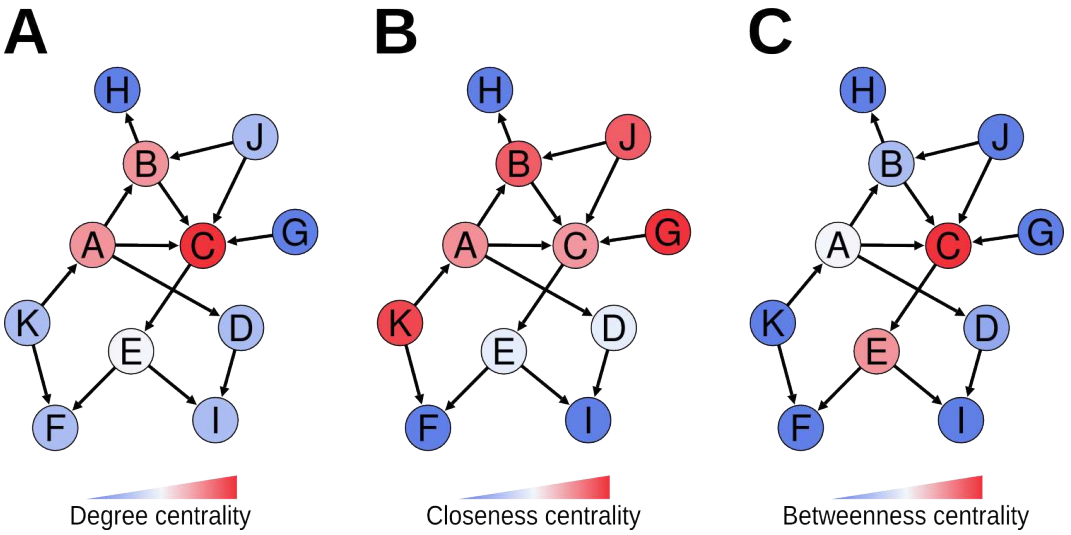


FIG. 1.8: Mesures de centralités sur un graphe. (**A**) Centralité de degré, permet de mettre en évidence les noeuds ayant le plus de voisins dans le graphe. (**B**) Centralité de proximité, dépend pour un noeud de la somme des plus court chemin entre ce noeud et tous les autres noeuds. Les noeuds sans arc sortant ont une centralité de proximité nulle. (**C**) Centralité d'intermediarité, augmente en fonction du nombre de plus courts chemins qui traversent le noeud.

En plus des propriétés topologiques précédentes qui sont dites locales, car elles sont calculées pour un noeud ou un sous ensemble du graphe, il existe des propriétés topologiques dites globales qui sont calculées sur le graphe entier. Par exemple, le diamètre d'un graphe connexe est défini comme étant le plus long des plus courts chemins entre tous les couples de noeuds possibles. Cet attribut permet d'avoir une idée sur la proximité des noeuds, et donc des entitées biologiques (molécules, réactions, etc.) qu'ils représentent.

Graphe et contexte biologique Les données biologiques peuvent être utilisées pour améliorer la représentation des graphes, en ajoutant des attributs ayant une signification biologique [63]. Par exemple, les arcs des graphes peuvent être pondérés par un nombre représentant l'intensité des interactions observées. C'est à dire que deux protéines ayant une forte affinité seront reliées par un arc avec un poids supérieur à celui reliant des protéines de faibles affinités [24, 86].

Un effort récent s'est porté sur la visualisation de ces modèles statiques [44, 112], dont l'accès est facilité par les formats d'export présentés précédemment. Des logiciels tel que VisANT [59], ARCADIA [147], VISIBIOweb [32] ou encore Cytoscape [134] proposent une

interface graphique attractive et des méthodes d'analyses statiques. Précurseur, Cytoscape est aujourd'hui soutenu par la communauté et par plus de 150 plugins venant compléter ses fonctionnalités [127] et est cité dans plus de 1000 publications. Cytoscape s'inscrit parfaitement dans la volonté actuelle d'intégrer des données de différents types et propose de recouvrir les graphes avec des données génomique, d'expressions, des ontologies. Les fonctionnalités sont multiples parmi lesquelles :

- la gestion de graphes à partir des formats BioPAX, SBML, PSI-MI, etc.
- la visualisation et la personnalisation très poussée des graphes (couleur, forme, disposition...)
- La connexion avec de nombreuses bases de données
- L'exploration de graphes (recherche, filtre sur critère topologique, etc.)

1.3.3 Modèle dynamique : comprendre le comportement du système

Dans [75], Kitano fait l'analogie entre l'étude des interactions et une carte routière, dont la description est un premier pas essentiel mais au delà de cette information figée, ce qui importe le plus c'est le trafic, comment il se caractérise et comment le réguler. En biologie tout est en perpétuel évolution, au cours du temps les molécules changent d'état, de localisation, de concentration. Les interactions/réactions possibles à un instant donné sont elles aussi amenées à évoluer. Tout ces changements doivent être pris en compte de façon à comprendre et reproduire un comportement observé. Les modèles dynamiques permettent de tester une hypothèse sur le comportement d'un système.

Est considéré comme modèle dynamique, tout modèle dont les entités peuvent prendre des valeurs qui évoluent au cours du temps. La valeur associée à chaque entité qualifie l'état du système et peut dépendre de l'état précédent du système et de paramètres externes. Les modèles dynamiques regroupent les modèles continus et discrets.

Modèles continus Ces modèles sont basés sur des valeurs physiques d'entités et de temps. Les entités sont généralement décrites par leur concentration, et le temps par une unité adéquate en fonction du modèle (seconde, minute, heure, etc.). Les modèles continus requièrent des données quantitatives, souvent très difficiles à obtenir, voir impossible à grande échelle. C'est pourquoi les modèles continus sont généralement utilisés pour étudier de petits systèmes, de l'ordre de 20 entités. En contrepartie ces modèles offrent une étude très fine de la dynamique sur un temps continu. Les modèles continus peuvent être déterministes (*équations différentielles ordinaires* (ODE) et *équations aux dérivées partielles* (PDE)) si l'évolution est fixé par les conditions initiales et les paramètres du modèle. Il existe également des modèles non-déterministes (*équations différentielles sto-*

chastiques (SDE)) avec l'ajout de probabilité dans le modèle. Le formalisme le plus utilisé est celui des *équations différentielles ordinaires* (ODE) [3]. Dans ces modèles, chaque entité x évolue en fonction d'autre entités $(x_1, x2, \ldots, x_n)$ suivant une équation différentielle de la forme :

$$\frac{dx}{dt} = f_x(X)$$

où $X = \{x_1, x_2, \ldots, x_n\}$ et f_x est la fonction d'évolution de x.

Ces modèles ont été utilisés pour analyser de nombreux mécanismes biologiques comme les différents phases du cycle cellulaire chez la levure [20], la réponse à la voie de signalisation ERK [130], ou encore la régulation du métabolisme [114].

Modèles discrets Les valeurs des entités et du temps peuvent être discrétisées. Ainsi les entités ne sont plus représentées par leur concentration mais par des valeurs discrètes comme les valeurs booléennes (0 ou 1, absent ou présent) ou multivaluées pour représenter différents niveaux de concentrations (absent, faiblement concentré, fortement concentré). Il existe deux approches pour discrétiser le temps : l'échantillonnage et les systèmes à événement discrets.

L'échantillonnage consiste à observer un système à des intervalles de temps réguliers. La longueur de cet intervalle de temps est appelée échantillonnage et doit être choisie avec précaution. Une bonne approximation d'un système en temps continu ne doit pas entraîner de perte de données, et doit pour cela vérifier la condition de Shannon :

$$f_{ech} \geq 2\ f_{max}$$

où f_{ech} est la fréquence d'échantillonnage et f_{max} la fréquence maximale des signaux émis par le système, également appelée longueur d'onde. Cette condition assure que l'échantillonnage est suffisamment fréquent pour être représentatif du système continu. Cette représentation du temps est analogue à certaines manipulations expérimentales visant à faire des prélèvements toutes les x minutes, heures,... Ces relevés doivent être fait à des intervalles de temps réalistes compte tenu de ce qui doit être mesuré.

Dans les systèmes à événements discrets, le temps est symbolisé par les entrelacements d'occurrences d'événements. Le système est observé uniquement lorsqu'il se passe quelque chose : une molécule change d'état, une réaction a lieu. Ce temps logique est généré par le modèle suivant différents schémas temporels et ne nécessite pas de mesure de temps physique contrairement à l'échantillonnage. En pratique les systèmes à événements discrets sont les plus utilisés pour représenter le temps discret dans les modèles biologiques et seront

détaillés dans le chapitre suivant.

Les caractéristiques des modèles discrets apportent l'énorme avantage de demander moins de connaissance au préalable sur les concentrations et les taux de réactions. Les informations nécessaires sont d'ordre qualitatif. L'évolution de ces modèles discrets (déterministes) est exprimée par la relation :

$$X_t = f(X_{t-1})$$

où X_t caractérise l'état du système au pas de temps t, c'est à dire la valeur associée à chaque entitée du système au pas de temps t.

Il existe de nombreux formalismes discrets appliqués en biologie comme les modèles booléens, les réseaux de Petri et les automates cellulaires. Ces formalismes se distinguent sur le niveau d'abstraction, l'expressivité et les méthodes d'analyses associées. et s'adaptent à la question posée. Aucun formalisme n'est universel, le choix de ce dernier est une étape cruciale dans une approche de modélisation. Il faut cependant noter que des interactions sont possibles entre les formalismes, ainsi qu'entre les approches discrètes et continues.

La facilité d'utilisation de ces méthodes d'analyse de la dynamique fait l'objet d'un intérêt croissant [45]. C'est pourquoi la plupart de ces approches sont mises en avant à travers des logiciels : GinSim pour les modèles logiques [103], BioNetSim pour les réseaux de Petri [42], Biocham [15] pour les modèles basés sur des règles.

Chapitre 2

Modéliser la signalisation cellulaire

La signalisation cellulaire regroupe l'ensemble des processus de communication permettant à la cellule d'adapter son comportement à son microenvironnement. La signalisation régit des mécanismes essentiels à la vie de la cellule comme la prolifération et la différentiation cellulaire, l'apoptose et la survie cellulaire. Bien que la signalisation ait longtemps été étudiée comme un processus linéaire, où les voies (pathway) étaient décrites de façon isolées, il est maintenant reconnu que ces voies sont inter-connectées. Cet entrelacement conduit à une importante combinatoire moléculaire faisant de la signalisation un système complexe par excellence. La compréhension de la réponse d'une cellule à un (ou plusieurs) ligand(s), en fonction de son contexte intrinsèque, et de son microenvironnement est non triviale. En conséquence, les approches de modélisation se sont multipliées, faisant de la signalisation le réseau biologique décrit par le plus grand nombre de formalisme [91], en comparaison avec les réseaux de régulations de gènes et les réseaux métaboliques. Sont présentées ici les principales caractéristiques de la signalisation, ainsi que les grandes approches de modélisation appliquées.

2.1 La signalisation cellulaire

2.1.1 Composants de la signalisation

La transduction du signal est souvent schématisée selon un axe extracellulaire → intracellulaire en impliquant des cascades d'événements impliquant des protéines qui transmettent le signal de la membrane au noyau (Figure 2.1).

L'initiation du signal est engendrée par la fixation d'un ligand sur un récepteur membranaire. Ces ligands regroupent différents types de biomolécules comme les hormones, facteurs de croissances, les cytokines et les composants matriciel. La diffusion du message de l'espace membranaire vers le noyau passe par un ensemble de réactions d'activation, désactivation et de transports protéiques. Si la cible ultime des voies de signalisation est

souvent présentée comme la régulation transcriptionelle de gènes, d'autres mécanismes cellulaire peuvent aussi être impactés : l'organisation du cytosquelette, le trafic des vésicules ou encore le métabolisme.

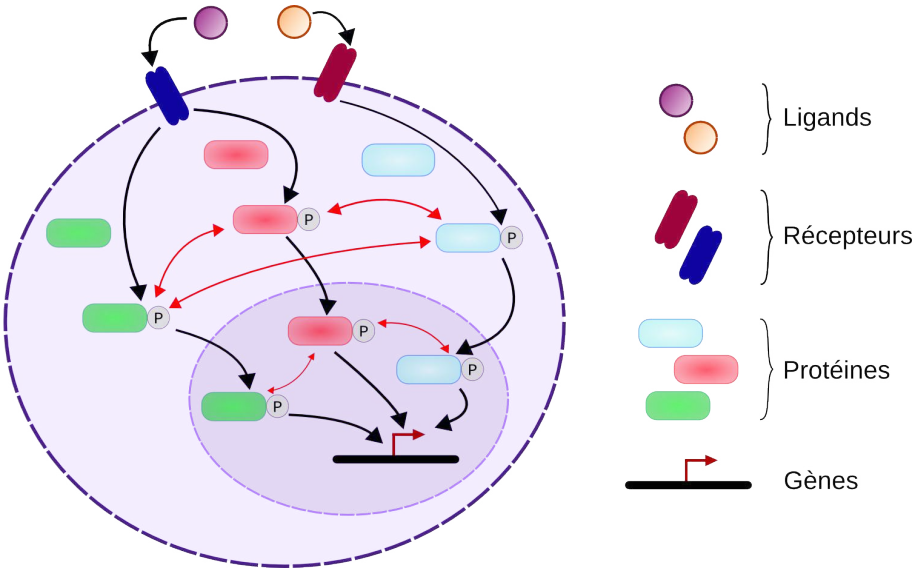


FIG. 2.1: Représentation schématique de la signalisation cellulaire. Les récepteurs membranaires peuvent être "activés" suite à la fixation de différents ligands extracellulaires. Le signal est ensuite transmis jusqu'au noyau par une succession de réactions impliquant de nombreuses protéines pour réguler des gènes cibles. Les modifications de l'expression des gènes va conduire à un changement phénotypique permettant d'adapter la cellule aux modifications du microenvironnement. De multiple interactions (symbolisés par des flèches rouges) existent entre les voies, conduisant ainsi à considérer la signalisation comme un réseau.

2.1.2 Caractéristiques

La signalisation cellulaire se différencie des autres réseaux biologiques par différents aspects. Tout d'abord, les réactions sont extrêmement hétérogènes. En effet, les réseaux de régulation de gènes ne comportent que des régulations transcriptionelles, et les réseaux métaboliques sont essentiellement formés de réactions enzymatiques. Les réseaux de signalisation regroupent aussi bien des modifications post-traductionelles, des translocations, des formations et des dissociations de complexes. Chacune de ces réactions pouvant être régulée par des activateurs et des inhibiteurs. Cette diversité dans la compositions des réactions complique l'interprétation des observations et leur formalisation qui doit être suffisamment souple et abstraite, pour permettre d'intégrer ces différents mécanismes dans

une seule représentation.

D'autre part, la signalisation est un processus transitoire. C'est à dire que l'on observe généralement une propagation du signal comme une succession d'événements biologiques. Les réseaux de signalisation possèdent une structure particulière permettant de discerner des entrées (cytokine, stimuli) et des sorties (gènes, phénotype).

Cette vision est différente des réseaux métaboliques par exemple qui sont fortement ancrés dans la notion de production/consommation et donc d'évolution de concentrations. Dans ce cas on parle généralement de flux de concentration et on recherche des conditions d'équilibre où ce flux est équilibré. Même si à une certaine échelle ce sont bien les réactions biochimiques qui se déroulent, les réseaux de signalisation ne sont généralement pas vus sous cet angle, mais sous celui de diffusion d'information (ou flux d'information). L'analogie est parfois faite avec un circuit électrique dont on observe la diffusion du courant.

2.1.3 Les questions autour de la signalisation

La signalisation cellulaire pose à l'heure actuelle de nombreuses questions tant sur son comportement global que sur ses mécanismes fins. D'une part, l'identification et la compréhension du rôle de certains acteurs de la signalisation demeurent floues. D'autre part, la prédiction d'un comportement en fonction de conditions données doit prendre en compte la complexité du circuit.

La signalisation permettant d'induire différents comportements, de nombreuses questions portent sur la compréhension de ces comportements résultant de la plasticité de la signalisation cellulaire. Quelles sont les molécules/réactions nécessaires pour induire un comportement donné ? Dans quel mesure la quantité d'un composant, ou son affinité avec ses partenaires influence la transduction du signal ?

L'impact du croisement entre les différentes voies de signalisation, généralement étudiées de façon isolée, suscite également un grand intérêt. Puisque ces observations sur chaque voie sont obtenues lors d'expérimentations indépendantes, les comportements résultant du croisement de ces observations restent pour la plupart méconnues.

Un fait important réside dans l'altération des voies de signalisation dans de très nombreuses pathologies, suscitant un intérêt croissant dans le ciblage des voies de signalisation pour le développement d'approches thérapeutiques. Comment un même ligand/signal peut-il entraîner différents comportement, parfois antagonistes dans des contextes physiopathologiques différents ?

Pour répondre à ces questions de nature diverses, de nombreuses méthodes de modélisation ont été développées et sont présentées dans la section suivante.

2.2 Les approches de modélisation

De part sa complexité et l'hétérogénéité des processus biologiques qui la composent, la signalisation cellulaire à été étudiée par de nombreuses approches de modélisation [91]. Il n'existe pas d'approche idéale permettant de répondre à toutes les questions sur les mécanismes de la signalisation. Il n'y a pas non plus de consensus sur la classification de ces approches. Il serait trop réducteur de les classer en fonction d'un unique critère en se basant sur le formalisme utilisé par exemple. Un même formalisme pouvant être utilisé de façon différentes, avec différentes interprétations de la biologie. Ces approches se distinguent en réalité sur bien d'autres critères, notamment sur la façon d'interpréter les faits biologiques. Afin de mettre en exergue leurs particularités, nous avons choisi de les discriminer suivant différents aspects qui nous semblent pertinents. La classification proposée ici ne veut en aucun cas imposer un système de castes, mais souhaite illustrer les grandes tendances actuelles en terme de modélisation de la signalisation cellulaire. Le Tableau 2.1 résume cette classification.

2.2.1 Les approches réactions centrées

Les approches réactions centrées considèrent la signalisation comme un ensemble de réactions entre molécules. Ces réactions peuvent être de nature variée en fonction du niveau d'abstraction et sont les éléments de base autour desquels s'articulent les modèles. Ces approches permettent de prendre en compte les caractéristiques de la signalisation : hétérogénéité des réactions et propagation du signal.

La volonté de ces approches est de pouvoir représenter de façon intuitive les réactions connues. C'est à dire, être proche de la biologie. Suivant le point de vue, le niveau d'abstraction, différents formalismes peuvent être utilisés comme les modèles différentiels ou les réseaux de Petri.

Les modèles différentiels

Les modèles différentiels sont basés sur la notion de réaction biochimique et donc de vitesse de réactions. Ces modèles se représentent sous la forme d'un système d'équations différentielles où les espèces sont réparties entre les réactants (membres gauches de l'équation) et les produits (membres droits de l'équation).

Ils permettent une étude très fine de l'évolution de chaque entité au cours du temps. La plupart des modèles ODE (Ordinary Differential Equation) sont basés sur la loi d'action de masse. D'autres lois en dérivent, comme la loi de Hill, utilisée pour la régulation d'expression de gènes ou encore la loi de Michaelis-Menten utilisée pour les réactions enzy-

Formalisme	Booléen, Logique	Réseaux de Petri	Basé sur des Règles	Équations différentielles	Automate cellulaire	Basé sur des Agents
Point de vue	molécule	réaction	réaction	réaction	multi-échelles	multi-échelles
Données principales	qualitatives	qualitatives	qualitatives	quantitatives	qualitatives	qualitatives
Taille des modèles	100	100	500	50	100	100
Dynamique	**discrète**, stochastique	**discrète**, stochastique, continue	**stochastique**, continue, discrète	continue	**discrète**, continue	**discrète**, continue
Analyses	états stable, cycles	statique (matrice), dynamique (simulation)	simulation, logique temporelle	simulation, LTL	Simulation	Simulation
Avantages	Représentation des influences, simulation de mutants	Représentation du flux de concentration, système concurrent	Description très précises des interactions, modélisation de la concurrence	Dynamique quantitative	Représentation de l'espace	Représentation de cellules ayant un comportement spécifique
Inconvénients	Calcul exhaustif du graphe de transition d'états	Nombre de jetons non borné dans la sémantique de base	Concept d'inhibition difficile à représenter quand le mécanisme n'est pas connu	Besoin de données quantitatives	Manque de méthodes d'analyse	Manque de méthodes d'analyse
Logiciels	Bio-Logic Builder [56], GinSim [103], CellNetAnalyzer [77]	Snoopy [120], BioNetSim [42]	Biocham [15], Ksim, BioNetGen[13], Pathway logic [139]	Mathlab	-	-

TAB. 2.1: Tableau des principales approches appliquées pour modéliser la signalisation cellulaire.

matiques. L'application de la loi d'action de masse repose sur 2 principales hypothèses : (i) les réactions ont lieu dans un espace parfaitement mélangé, et les réactants sont répartis de façon homogène ; (ii) le nombre de molécules est suffisamment grand (tend vers l'infini pour être optimal). Cette approximation est bien sur très loin de la réalité biologique. En effet la cellule ne constitue pas un récipient bien mélangé où les molécules sont uniformément réparties, cependant cette approximation est très largement utilisée dans la littérature.

Ce formalisme mathématique permet d'associer une vitesse de réaction à chaque réaction. Ainsi les équations décrivant la vitesse des réactions de phosphorylation d'une protéine A par une kinase K et de complexation entre une protéine A phosphorylée (A_p) et une protéine B sont formulées comme suit :

$$v1 = k_{phos}\ [K]\ [A]$$
$$v2 = k_{on}\ [A_p]\ [B]$$

où $v1$ et $v2$ sont respectivement les vitesses des réaction de phosphorylation et de complexation. k_{phos} est le coefficient de phosphorylation et k_{on} est le coefficient de complexation. $[A], [A_p], [B] et [K]$ sont les concentration respectives de A, A_p, B et K où A_p est la protéine phosphorylée.

Les vitesses de réactions sont proportionnelles aux concentrations des réactants. Les espèces qui composent le modèle évoluent donc en fonction de ces vitesses de réaction, chaque espèce est associée à une équation différentielle comme indiqué :

$$\frac{d[A]}{dt} = -(k_{phos}\ [K]\ [A])$$
$$\frac{d[A_p]}{dt} = (k_{phos}\ [K]\ [A]) - (k_{on}\ [A_p]\ [B])$$
$$\frac{d[B]}{dt} = -(k_{on}\ [A_p]\ [B])$$
$$\frac{d[AB]}{dt} = k_{on}\ [A_p]\ [B]$$
$$\frac{d[K]}{dt} = 0$$

En intégrant ce système d'équations, il est possible de suivre l'évolution déterministe de chaque entité au cours du temps. Ces modèles permettent de tester des hypothèses quantitatives, telle que l'augmentation de l'affinité entre 2 composants, la sur-expression ou l'extinction d'un gène en faisant varier les concentrations initiales. Il faut également noter que les données utilisées pour inférer les valeurs des constantes et concentrations des entités sont obtenues sur des populations de cellules, les données correspondent à une moyenne intégrant à la fois des variations individuelles et l'imprécision des mesures. Dans ce contexte la loi d'action de masse est applicable, et arrive à retranscrire le comportement

observé.

Ces modèles sont actuellement les plus utilisés et les principales voies de signalisation ont ainsi été modélisée grâce à ces approches : les voies induites par le TGF ou l'EGF [118], les voies MAPK [133], NFKb [58] mais aussi des phénomènes plus complexes comme l'apoptose [116]. Cette utilisation est contradictoire avec le fait que ces modèles demandent les données les plus précises, et donc les plus compliquées à obtenir.

Ces modèles différentiels comportent généralement un nombre restreint d'entités, de l'ordre de quelques dizaines. La difficulté majeure du développement de ces modèles vient de la nécessité d'avoir des données quantitatives (vitesse de réaction et concentration) car l'obtention de ces données quantitatives précises demeure un facteur limitant en biologie. Quand bien même les données seraient disponibles pour concevoir un modèle différentiel à large échelle, l'interprétation des résultats de simulation serait une tache des plus ardue compte tenu de la complexité combinatoire.

Les modèles basés sur des règles

Les modèles à base de règles regroupent différents formalismes basés sur des spécifications de type "si - alors". Les réactions sont représentées par un ensemble de règles ayant lieu sous certaines conditions ("si" ou membre gauche de la règle) et qui entraînent différentes actions ("alors" ou membre droit de la règle). Les principaux acteurs de ces règles sont des molécules : protéines, gènes, métabolites... D'un point de vue biochimique, une règle fixe les contraintes que doivent respecter les réactants afin d'interagir ensemble dans le but de générer un effet donné. Ces modèles sont donc conçus en précisant les règles locales des interactions, et le comportement global émane de la réalisation de ces règles dans une certaine dynamique.

Afin d'illustrer cette approche, nous décrivons ici deux langages : Biocham [15] et Kappa [29] qui permettent de formaliser des réactions biologiques en modèles à base de règles.

Biocham L'idée de départ de la machine abstraite biochimique Biocham [15] est d'offrir un langage de description intuitif permettant la simulation et l'analyse de réseaux de réactions biochimiques. La syntaxe de Biocham permet de représenter différents types de réactions, formalisées par des règles, entre différents type de d'objet biochimiques. Par exemple, les règles suivantes définissent les réactions de phosphorylation et de complexation :

$$A = [K] \Rightarrow \ A \sim \{p_1\}$$
$$A \sim \{p_1\} \ + \ B \Rightarrow A \sim \{p_1\} - B$$

où A,K et B sont des objets biochimiques, ici des protéines. La catalyse de la phosphorylation de A par K est symbolisée par $= [K] \Rightarrow$, ce qui est équivalent à placer K dans le membre gauche et le membre droit de la règle. Les symboles $-$ et $\sim$ représentent respectivement le complexe (ici entre $A \sim \{p_1\}$ et B) et un site modifié (ici le site p_1 de A est phosphorylée).

L'interprétation de la dynamique peut se faire suivant 3 sémantiques différentes : continue, stochastique et booléenne. Chacune de ces sémantiques permet l'utilisation de la simulation et de méthodes spécifiques de vérification de propriétés temporelles, implémentées dans le logiciel Biocham. La sémantique continue associe une valeur réelle à chaque objet biochimique, représentant leur concentration, et un coefficient à chaque réaction. Les règles sont interprétées en un système d'équations différentielles qui peut ensuite être simulé par des méthodes d'intégration classique comme Runge-Kutta. Cette sémantique est la seule déterministe sur les 3 proposées, C'est à dire qu'il n'existe qu'un enchaînement d'états (vecteur des valeurs de chaque entités pendant la simulation) pour un état initial et des paramètres donnés. La sémantique stochastique permet de représenter les variations observées sur une population, et représente généralement une version bruitée des simulations obtenues avec un système d'équations différentielles. Cependant une différence significative peut être observée pour de petites quantités de molécules, là où les critères des modèles continus ne s'appliquent pas. En stochastique, les objets biochimiques sont représentés par un nombre entier qui s'apparente à la quantité de chaque objet. Les coefficients sont transformés en taux de transition, en fonction du type de réaction. Différentes méthodes de simulations existent pour la sémantique stochastique comme celle de Gillespie. Enfin la sémantique booléenne introduit beaucoup d'indéterminisme, où les valeurs associées aux objets biochimiques sont booléennes (0 ou 1). Faute d'informations précises sur la dynamique des interactions, les règles sont interprétées avec plusieurs transitions, en fonction de la consommation totale ou partielle des entités. Par exemple la règle de complexation entre $A \sim \{p_1\}$ et B est interprétée en 4 transitions :

$$
\begin{aligned}
A \sim \{p_1\} \ \wedge \ B &\longrightarrow \ A \sim \{p_1\} \wedge \ B \ \wedge \ A \sim \{p_1\} - B \\
A \sim \{p_1\} \ \wedge \ B &\longrightarrow \ \neg A \sim \{p_1\} \wedge \ \neg B \ \wedge \ A \sim \{p_1\} - B \\
A \sim \{p_1\} \ \wedge \ B &\longrightarrow \ \neg A \sim \{p_1\} \wedge \ B \ \wedge \ A \sim \{p_1\} - B \\
A \sim \{p_1\} \ \wedge \ B &\longrightarrow \ A \sim \{p_1\} \wedge \ \neg B \ \wedge \ A \sim \{p_1\} - B
\end{aligned}
$$

De la même façon, la règle de phosphorylation donne lieu à 2 transitions suivant qu'il reste ou non de l'entité A (non phosphorylée) après la réaction. Le modèle est simulé sur un temps discret et à chaque pas une seule transition est appliquée. La simulation génère une trace représentant l'évolution des valeurs des entités au cours des pas de temps de

simulation.

Compte tenu de l'indéterminisme, une simulation n'est pas réellement informative puisque d'autres comportements peuvent être observés en simulant à d'autres reprises. Pour permettre l'étude des comportements de manière globale, BIOCHAM propose un ensemble de méthodes de vérifications de propriétés temporelles adaptées à la sémantique de la dynamique choisie. Pour la sémantique booléenne, la logique CTL (Computational Tree Logic) permet de représenter le graphe d'évolution d'états sous forme d'arbre pour explorer les trajectoires (suite d'états) possibles. La logique CTL est une extension de la logique propositionnelle avec l'ajout de quantificateur de chemin permettant de spécifier si il existe un chemin, ou si tout les chemins vérifient une condition qui porte sur les objets biochimiques. Par exemple, est ce qu'une trajectoire du modèle permet la phosphorylation de A. Des comportements plus complexes peuvent être recherchés, comme par exemple l'oscillation. Pour la vérification de modèles stochastique la logique PCTL (Probabilistic Computational Tree Logic) est utilisée, les quantificateurs de chemins sont remplacés par des probabilités de réalisations. Enfin la vérification de modèles basés sur la sémantique continue se fait grâce à la logique LTL (Linear time logic) et permet de faire des vérifications en précisant des valeurs quantitatives. Par exemple vérifier si la concentration d'une espèce x est supérieur à celle de l'espèce y, ou si x est toujours inférieur à une valeur seuil.

La vérification des formules de logiques temporelles est réalisée par un vérificateur de modèle (model checker) appelé NuSMV. En plus de pouvoir répondre à ces questions sur le comportement du modèle, le logiciel BIOCHAM propose également des méthodes pour : inférer des valeurs de paramètres dans la sémantique continue, réduire un modèle en conservant les propriétés logiques ou encore inférer de nouvelles propriétés logiques.

Notons que BIOCHAM propose également d'interpréter un système de règles de manière plus abstraite sous forme de graphe d'influence, qui s'apparente alors au modèle booléen et logiques présentés en section 2.2.2.

Le représentation sous forme de règles avec le langage Biocham à été appliquée en signalisation, sur la transduction des voies MAPK [53]. Ce modèle présente les phénomènes de compétions entre les protéines G et la β-arrestin qui sont capable de transmettre le signal issu de différents récepteurs transmembranaires. L'utilisation de la logique LTL implémentée dans Biocham a permit de contraindre le modèle pour obtenir des jeux de paramètres qui coïncident avec les données expérimentales.

Dans une autre étude, Biocham a été utilisé pour coupler la description du cycle cellulaire et de l'apoptose dans un même modèle et ainsi étudier la réponse à l'irinotecan utilisé dans le traitement de certain cancer [92]. Le modèle couplé permet d'investiguer

l'effet de l'irinotecan en fonction des phase du cycle en recherchant à maximiser son effet anti-tumoral tout en minimisant sa toxicité.

Kappa Kappa [29] est un langage à base de règles où les molécules sont représentées par des *Agents* possédant des *sites* pouvant être liés ou non, et dans un certain état (ex : phosphorylé). L'objectif Kappa est de permettre au modélisateur de ne représenter dans une règle que les critères qui importent. Ainsi la combinatoire de la combinaison des sites, en pratique très élevée, n'a pas besoin d'être énoncée. Un autre grand intérêt du langage Kappa est de permettre d'expliciter la topologie des complexes, c'est à dire de spécifier les sites réactionnels qui interagissent. Par exemple les réactions de phosphorylation et complexation sont représentée comme suit :

Complexation entre A et K	$K(a), A(k) \rightarrow K(a!1).A(k!1)$
Phosphorylation de A	$K(a!1).A(k\sim u!1) \rightarrow K(a!1).A(k\sim p!1)$
Complexation entre A et B	$A(k\sim p, b), B(a) \rightarrow A(k\sim p, b!2).B(a!2)$

La règle $K(a), A(k) \rightarrow K(a!1).A(k!1)$ spécifie que le site a de l'agent K est lié au site k de l'agent A puisque l'identifiant est le même (!1). Un site ne peut être lié qu'a un seul autre site à la fois, et les phénomènes de compétition peuvent ainsi être représentés lorsque qu'un Agent a deux partenaires potentiels pour un même site.

Les modèles à base de règles en Kappa peuvent être représentés graphiquement par une carte de contact. Cette dernière permet de visualiser les interactions possible entre les molécules. La Figure 2.2 illustre la carte de contact des réactions de phosphorylation et complexation. Le site k de l'agent A et a de l'agent K sont susceptible d'être liés, de même que le site b de l'agent A et le site a de l'agent B. Il faut noter que cette représentation perd toutefois les conditions d'interactions entre ces sites ainsi que les états possibles des différents sites. C'est donc une vue synthétique qui ne suffit pas à elle seule à représenter un modèle à base de règles.

Le logiciel KaSim permet d'utiliser des modèles Kappa pour réaliser des simulations stochastiques ou continues, et des analyses sur la dynamique. Parmi les analyses proposées, la recherche des *histoires* permet de visualiser les enchainements des règles observés au cours des différentes simulations stochastique. Ces histoires s'apparentent à une notion de voie en biologie. Il est donc possible de retrouver les règles importantes, les plus fréquemment utilisées, pour réaliser une règles donnée.

Parmi les applications utilisant le langage Kappa, la conception d'un modèle de signalisation EGFR a permis d'expliquer différents comportements et a montré que la structure

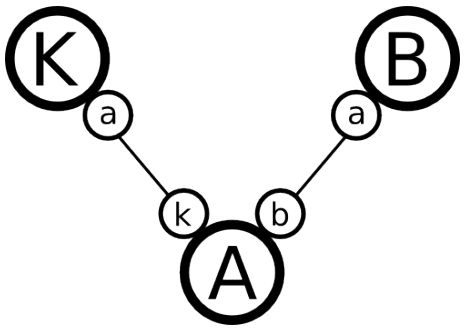


FIG. 2.2: Carte de contact des réactions de phosphorylation et complexation. Les agent A, K et B sont représentés par des grands cercle, leurs sites sont représentés par de petits cercle. Les arc entre les sites représentent une interactions possibles entre ces sites.

des règles pouvait avoir un impact supérieur sur la dynamique par rapport aux taux réactionnels [28].

Les modèles à base de règles sont très performants pour représenter les réactions biochimiques, en particulier les interactions entre molécules. Les logiciels développés autour de ces approches permettent de nombreuses simulations et analyses sur la dynamique. Ils demandent cependant de bonnes connaissances pour fixer les taux de réactions de chaque règles, et dans le cas contraire nécessitent d'inférer ces paramètres. De plus les modèles à base de règles sont bien adaptés pour modéliser l'inhibition par compétition, en revanche lorsque le mécanisme d'inhibition est méconnu, le représentation est moins intuitive.

Les Réseaux de Petri

Créé par Carl Adam Petri dans le but de modéliser des systèmes concurrents, les réseaux de Petri sont des graphes bipartites avec un premier type de noeuds représentant des places, et un second type de noeuds représentant des transitions. Les arcs ne peuvent relier qu'une place à une transition ou une transition à une place.

Actuellement un grand nombre de variantes des réseaux de Petri existent, nous présentons ici la sémantique originelle. Formellement, un réseaux de Petri N se représente sous la forme d'un 4-tuple $N = (P, T, F, W)$ où P est l'ensemble des places $P = \{p_1, p_2, \ldots, p_n\}$, T est l'ensemble des transitions $T = \{t_1, t_2, \ldots, t_m\}$, F l'ensemble des arcs orientés $F \subseteq (P \times T) \cup (T \times P)$ et W est une fonction de poids $W : F \to \{1, 2, 3, \ldots\}$. Les places peuvent contenir un nombre entier positif borné (ou non) de jetons, et les transitions permettent de déplacer ces jetons lorsqu'elles sont franchies, tirées. L'état du système

est caractérisé par la composition de chaque place en jetons et est appelé la marque M. La fonction de poids s'applique aux arcs (peut rendre compte de la stœchiométrie pour des réactions biochimiques) exprimant la quantité de jetons nécessaires avant la transition, et générés après la transition.

En biologie, les places s'apparentent aux molécules, regroupant les protéines, les gènes, les ARN, les métabolites... C'est pourquoi les approches sur les réseaux de Petri en signalisation, et plus généralement en biologie, utilise une version restreinte du formalisme de base des réseaux de Petri en considérant un nombre borné de jetons par place. Ceci s'explique par le fait qu'il existe un nombre fini de molécules dans une cellule. Cette restriction permet d'utiliser des méthodes d'analyse applicables uniquement sur des systèmes bornés. Les transitions sont utilisées pour représenter des réactions. Si la relation entre les places et les molécules semble simple, l'interprétation d'une réaction en transitions est une tâche plus ardue. Dans ce contexte une étude propose un schéma d'interprétation pour différents type de réactions couramment rencontrées en signalisation comme la phosphorylation ou la complexation [21].

La dynamique des réseaux de Petri, basée sur la notion de consommation/production de jetons, se fait grâce aux tirage des transitions qui va faire évoluer l'état du système. Pour qu'une transition soit tirée, il faut que chaque place entrante (place ayant un arc vers la transition) possède au minimum le nombre de jetons indiqués sur l'arc de cette place vers la transition. Si cette condition est vérifiée, la transition est tirée entraînant le déplacement du nombre de jetons indiqués sur chaque arc depuis les places entrantes vers les places sortantes. Dans la sémantique de base des réseaux de Petri, la dynamique est asynchrone, c'est a dire qu'une seule transition est franchie par pas de temps.

Dans la sémantique originelle, les transitions possèdent des places entrantes et sortantes exprimant les réactants et les produits d'une réaction mais il n'y a pas d'information sur la régulation de ces réactions. Cependant les régulations sont des phénomènes importants en biologie et différentes extensions au formalisme ont vu le jour pour augmenter l'expressivité de la sémantique pour permettre la représentation des régulations. En effet, les réseaux de Petri étendus (extended Petri nets) ajoutent de nouveaux types d'arcs afin d'exprimer des activations et des inhibitions de réactions [38]. L'introduction d'arc de lecture (read arc) et d'arc inhibiteur ajoute des conditions supplémentaires pour franchir une transition. Les places entrantes des arcs de lecture doivent contenir le nombre minimal de jetons indiqué sur l'arc et à l'inverse les places entrantes des arc inhibiteurs doivent contenir moins de jetons que le nombre inscrit comme poids de l'arc. Contrairement aux jetons des places entrantes représentant les réactants, les jetons des places entrantes des arc de lecture et des arc inhibiteurs ne sont pas consommés.

L'interprétation des réactions de phosphorylation et de complexation est présentée dans le réseau de Petri en Figure 2.3. Les réactions correspondent à des noeuds transitions, les molécules sont elles représentées par des places. L'utilisation d'arc de lecture permet de représenter l'effet de la kinase qui n'est pas consommée lors de la réaction.

En signalisation, les réseaux de Petri ont été appliqués à différentes voies telles que l'apoptose [52, 21], EGFR, réponse au phéromones [123]. Les réseaux de Petri offrent un large panel d'analyses basées à la fois sur leurs propriétés structurelles et dynamiques.

La recherche de propriétés structurelles permet d'obtenir des informations qualitatives sur la dynamique sans avoir à simuler le modèle et à générer le graphe de transition d'états, représentant l'enchaînement des marquages possibles. Les p-invariants et t-invariants sont deux propriétés basées sur une représentation matricielle du réseau de Petri. Leur étude porte sur la réponse aux phéromones à l'aide d'un réseau de Petri de 42 places et 48 transitions [123]. La recherche de p-invariants consiste à identifier les ensembles de places dont la somme des jetons est constante indépendamment de la dynamique. En biologie, ce concept se rapproche de la conservation des espèces (molécules) où la somme de toutes les formes d'une molécule est constante (sous l'hypothèse qu'elle ne soit ni dégradée ni produite). Cette propriété est également observée dans les systèmes différentiels. Dans [123], les p-invariants mettent en évidence la conservation de jetons entre des formes actives/inactives ou monomères/complexes. La recherche de t-invariants quand à elle permet d'identifier un ensemble de transitions dont le tirage successif n'a pas d'effet sur le marquage du modèle. C'est à dire qu'après avoir tiré chacune de ces transitions, les jetons retrouvent leur configuration initiale. D'un point de vue biologique, après le tirage de ces transitions le système retrouve la même répartition de ses molécules qu'au départ. Dans [123], chaque t-invariant caractérise un phénomène biologique tel que des cascades de signalisation avec rétro-contrôle. Comme la conception de modèle depuis la littérature peut parfois sembler empirique, un autre intérêt des p-invariants et t-invariants est de pouvoir être expliqués par des concepts biologiques dans le but de valider le modèle [52, 123].

D'autres approches statiques permettent de rechercher un sous réseau d'intérêt par rapport à une place (*i.e.* une molécule) donnée [33]. Cette étude propose également de rechercher une séquence de tirage des transitions pour définir une voie de signalisation. Une autre façon d'extraire des voies de signalisation est proposée par Li *et al*, en simulant leur modèle de l'apoptose en réseaux de Petri [21]. En fixant les conditions de départ à partir de leurs connaissances, ils retrouvent par simulation l'ordonnancement des molécules clefs telle que les caspases. De plus ils discernent les différentes voies capables d'induire l'apoptose dans leur modèle à savoir l'induction par Fas ou des dommages de l'ADN mitochondrial. La visualisation des différentes voies est bien plus intuitive dans le formalisme

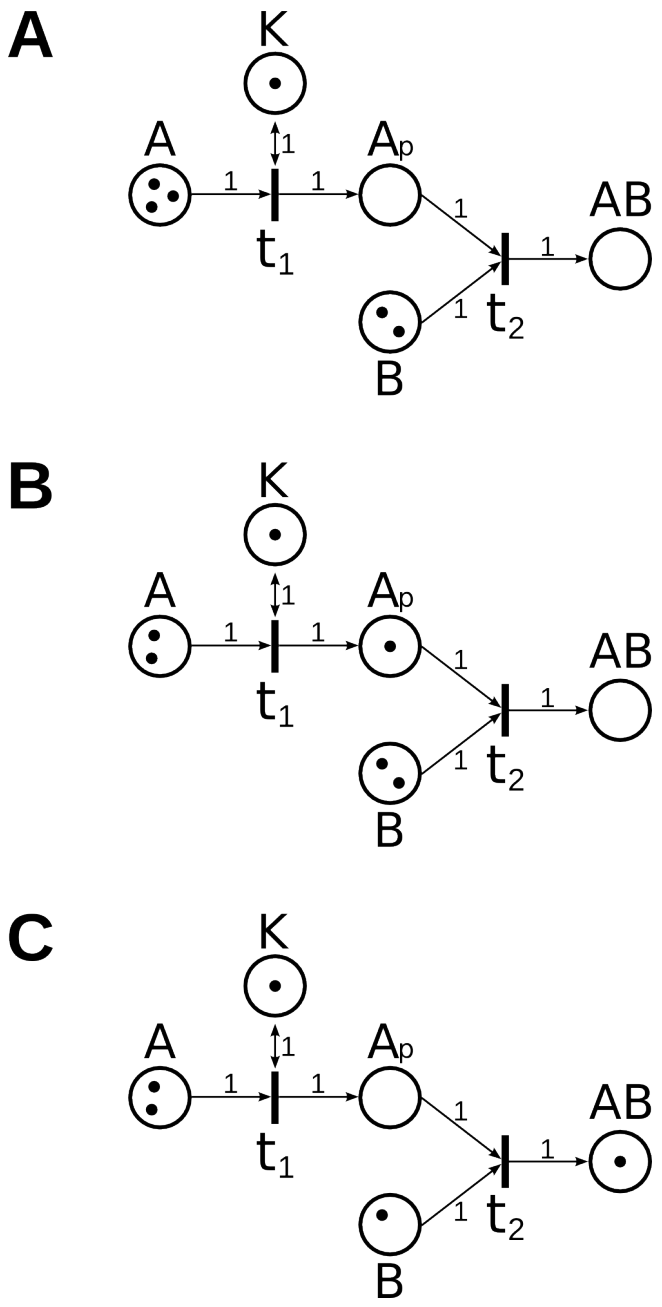


FIG. 2.3: Représentation des réactions de phosphorylation et de complexation en réseaux de Petri. À l'initialisation, les places A, B et K possèdent respectivement 3, 2 et 1 jetons. Les autres en sont démunies (**A**). Au premier pas, la transition t_1 est tirée, 1 jeton de A est consommé au profit de la place A_p (**B**). Au second pas, il est possible de tirer la transition t_1 ou la transition t_2. Ici la transition t_2 est tirée, 1 jeton est placé sur la place AB suite à la consommation d'un jeton de A_p et d'un jeton de B (**C**).

des réseaux de Petri qu'avec les approches différentielles où les cinétiques obtenues sont visualisées pour chaque espèce sans forcément pouvoir relier leurs comportements.

D'autres variantes du formalisme ont été appliquées à la signalisation parmi lesquelles :

- les réseaux de Petri continus qui sont définis en donnant des valeurs réelles positives au lieu des valeurs entières aux places, ainsi que des constantes cinétiques sur les transitions. De cette manière ces réseaux de Petri peuvent être perçus comme une description structurée d'un système différentiel classique [45].
- Les réseaux de Petri hybrides utilisent à la fois des valeurs continues pour certaines places et transitions, et des valeurs discrètes pour d'autres [143, 34]. Ceci permet d'abstraire certaines partie du modèle comme des mécanismes simples de type interrupteur (on/off) tel que l'expression de gène. D'autres mécanismes comme les réactions enzymatiques sont eux décrits par des places et transitions continues.
- Les réseaux de Petri stochastiques ajoutent des probabilités pour choisir l'ordre dans lequel les transitions sont tirées.

Les réseaux de Petri ont l'avantage d'offrir une représentation graphique intuitive tout en ayant une solide base mathématique, permettant des analyses aussi bien sur la structure que sur la dynamique. Cependant, leur conception repose sur le principe de consommation/production qui, bien qu'adapté aux processus métaboliques, l'est moins pour la signalisation en dépit des efforts pour étendre l'expressivité du formalisme pour les régulations. De plus si l'on souhaite utiliser les possibilités offerte par ce formalisme, il faut notamment des informations précises pour pondérer les arcs, ou encore ajouter des probabilités de franchissement de transitions. Ces limitations se rapprochent alors de celles des modèles différentiels.

2.2.2 Les approches molécules centrées

Contrairement aux approches réactions centrées, la vision molécule centrée ne s'intéresse pas à la nature des réactions mais uniquement aux concentrations des molécules, et à la façon dont elles s'influencent. Cette représentation se concentre sur le fait de savoir qui agit sur qui, mais pas comment. Le système est considéré comme un ensemble de molécules ayant des influences positives et négatives entre elles. Chaque molécule peut prendre un nombre fini de valeurs, souvent booléennes (0 ou 1). Les approches molécules centrées peuvent être perçues comme ayant un niveau d'abstraction plus élevé car elles nécessitent d'abstraire l'effet des différents acteurs sans considérer les mécanismes sous-jacents.

Les modèles Booléens et logiques

Basé sur les travaux de Kaufman et Thomas, ces modèles considèrent les effets entre les molécules, sans prendre en compte "comment se produit cet effet" [46, 141]. Ce point de vue émerge de l'abstraction des modèles différentiels type ODE, où l'on observe pour bon nombre d'interactions en biologie qu'un régulateur n'a d'effet qu'au dessus d'un certain seuil. Pour l'expression d'un gène par exemple, la concentration de la protéine exprimée peut être visualisée en fonction de son régulateur. La même relation peut être observée en signalisation entre la concentration du signal et l'intensité de la réponse. La sigmoïde obtenue se caractérise notamment par une valeur seuil θ au dessus de laquelle le régulateur a un effet sur la cible. Cette sigmoïde peut être approximée, idéalisée, par une fonction logique (Figure 2.4).

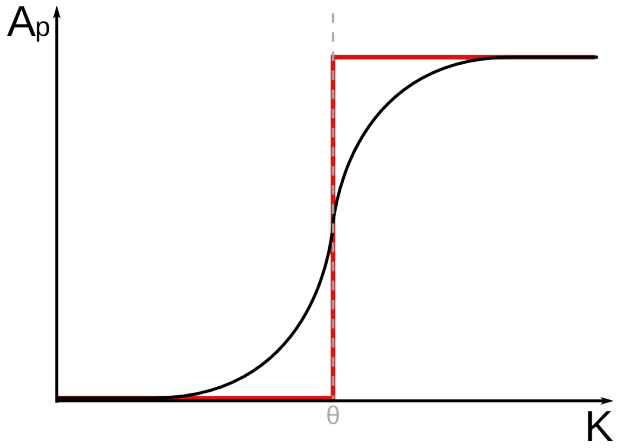


FIG. 2.4: Augmentation du signal (A_p) en fonction de la quantité d'activateur (K). La sigmoide (en noir) caractérise la dynamique de A_p suivant un système différentiel. Son comportement est approximé en fonction de K (courbe rouge).

Ainsi discrétisée, la molécule cible prend la valeur 1 (présente) au dessus du seuil θ (régulateur = 1), 0 sinon (absente) et régulateur = 0. Cette description s'adapte bien avec les observations biologiques de type : "en présence/en l'absence de", "à forte/faible concentration". Un modèle booléen se représente donc sous la forme d'un graphe orienté $G = (V, E)$ où $V = \{v_1, v_2, \ldots, v_n\}$ est l'ensemble des noeuds et E est l'ensemble des arc entre ces noeuds. Ces arcs dénotent les influences entre les noeuds et permettent d'établir les fonction de transferts $F = \{f_1, f_2, \ldots, f_n\}$. Les arcs portent des poids, typiquement +1 ou -1 dans le cadre des modèles booléens pour dénoter une influence positive ou négative. Chaque noeud est associé à une fonction de transfert qui dépend de ses régulateurs positifs et négatifs.

Pour représenter la phosphorylation de A par une kinase K, il est possible d'abstraire la phosphorylation à une activation [129]. Ainsi l'état actif de A ($A = 1$) dépend uniquement de la présence de K ($K = 1$). Cette dépendance est représentée graphiquement par un arc, portant un poids de 1, du noeud K vers le noeud A. De la même façon la modélisation de la formation du complexe entre A et B est représentée par des arcs de poids 1 de A vers AB et de B vers AB. La représentation de ces réactions sous forme de modèle booléen est illustrée en Figure 2.5, les fonctions de transfert permettent d'actualiser les valeurs de chaque noeud en fonction des noeuds qui l'influencent (Tableau 2.2).

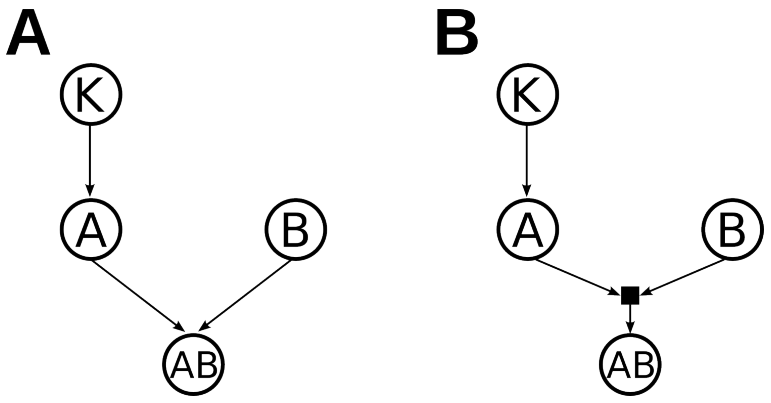


FIG. 2.5: Représentation des réactions de phosphorylation et complexation avec un point de vue molécule centré. (A) Les noeuds symbolisent les molécules, les arcs les influences entre ces molécules de telle manière que l'arc entre K et A signifie que K influence positivement A. (B) La nécessité de la présence simultané de A et B pour former le complexe AB peut être représentée en utilisant un graph bi-partite où deux noeuds servent de connecteurs logique. Le carré noir symbolise ici un "et" logique.

Noeud	**Fonction de transfert**
A	$A^* = K$
AB	$AB^* = A \; et \; B$

TAB. 2.2: Fonction de transfert des noeuds du modèle booléen présenté en Figure 2.5

La dynamique du modèle se fait donc on observant l'état du système sur des pas de temps discret. L'état de système étant un vecteur composé de la valeur de chaque noeud. L'évolution du système dépend du schéma dynamique, c'est à dire l'ordre dans lequel les valeurs des noeuds sont actualisées. Les plus utilisés sont (i) la dynamique synchrone, où toutes les valeurs sont actualisées à chaque pas de temps, (ii) la dynamique asynchrone où une seule valeur est actualisée à chaque pas. Ces dynamiques sont présentées plus en

détail dans la partie dédié aux temps dans les modèles.

Pour représenter des niveaux plus fins de relation entre différentes molécules, les modèles booléens peuvent être étendus aux modèle logiques où les entités peuvent prendre des valeurs entières. Ceci permet de préciser des influences à différents niveaux, protéine absente (= 0), faiblement présente (= 1), fortement présente (= 2). Ce niveau de détail est notamment utilisé pour représenter des molécules qui ont différents effets en fonction de leur concentration.

Les modèles Booléens permettent un certain nombre d'analyses, basées pour certaines sur le graphe de transition d'états. Dans ce graphe chaque noeud représente un état du modèle, le graphe est donc de taille 2^n où n est le nombre de noeuds du modèle. Cette définition en fait un graphe explosif en terme de taille, il devient très rapidement compliqué à visualiser, voire même à générer dès une vingtaine de noeuds donc une vingtaine de molécules. Sur ce graphique il est possible de rechercher différentes propriétés issues de la dynamique du modèle : états stables, cycles, bassins d'attraction, etc.

A la différence des modèles réactionnels, différent type de réaction sont ici représentées de la même façon. L'activation de A par une kinase K et la formation de complexe de A en AB, sont toutes deux représentées par des arcs. Dans le premier cas, il s'agit d'une réaction enzymatique où K n'est pas consommée. A l'inverse la consommation du monomère A semble naturelle lors de la formation d'un complexe mais ici non représentée. Dans un modèle booléen, il n'y a pas réellement de notion de consommation ni même de réaction. Un arc d'un noeud x vers un noeud y n'implique pas la consommation de x suite à l'activation de y. Ce point de vue est bien différent de la vision réactionnelle des réseaux de Petri. La notion de consommation peut éventuellement être représentée par le biais d'arc inhibiteur (poids = -1), mais le formalisme n'étant pas conçu pour cette notion, cela alourdi la représentation et reste difficile à mettre en place en conservant la réalité biologique. En pratique cette possibilité n'est pas employée.

Les modèles booléens/logiques ont été appliqués en signalisation. Par simulation, ces modèles permettent de prédire l'impact de telle ou telle molécule sur un critère défini. A l'aide d'un modèle booléen comportant 58 noeuds, Zhang *et al* ont étudié la survie des lymphocytes T [157]. Pour cela, ils ont représenté l'apoptose sous la forme d'un noeud, qui est activé en présence de caspase. En observant la valeur du noeud *apoptose* en fonction de différentes initialisations et différentes valeurs de noeuds fixées, ils ont mis en évidence des molécules essentielles à la survie. En effet, en forçant l'inactivation d'un noeud il est possible de mimer l'inhibition d'une protéine. Les noeuds induisant l'apoptose lorsqu'ils sont inactifs, sont donc potentiellement responsables de la survie. Leurs prédictions ont de plus été validées expérimentalement. Dans une approche similaire, Mendoza s'est intéressé

à la différentiation des lymphocytes T auxiliaires à l'aide d'un modèle logique de 17 noeuds [97].

Les modèles logiques sont aussi utilisés pour étudier la sensibilité des voies de signalisation à différents signaux. Une étude récente a montré que l'environnement extérieur, c'est à dire la combinaison de différentes entrées (ligands), avait un impact sur la réponse et pouvait changer les voies empruntées lors de la propagation du signal [35]. Pour cela les auteurs ont analysé la sensibilité des noeuds au nombre de ligands différents dans l'environnement tel que l'EGF, le calcium ou encore le stress. De façon intéressante ces résultats ont mis en avant le fait que les modèles de signalisation étaient sensibles à cette combinaison d'input, à l'inverse de modèles dont les influences sont réparties de façon aléatoire.

D'autres approches comme celle proposées par Samaga *et al* [129] sont d'avantage liées aux données expérimentales. En confrontant leur modèle de signalisation de EGF, ils ont été en mesure de vérifier la majorité des comportements observés sur différents jeux de données haut débits. Plus récemment cette approche de comparaison a permis d'affiner les connaissances sur les voies de signalisation en condition physiologique et pathologiques [125]. En comparant des données issues de cellules saines et tumorales, ils ont conçu différents modèles de signalisation qui permettent d'expliquer les résultats obtenus. Ces modèles se distinguent par certaines influences, et voies de signalisation qui peuvent être activées uniquement dans l'état physiologique ou pathologique.

Si ces modèles permettent d'expliquer un comportement observé, ils ne sont en revanche généralement pas utilisés pour en prédire. De plus, en pratique les modèles logiques sont plus abstraits que les modèles de réactions. Par exemple la formation de complexe est rarement explicitée, de même que les mécanismes d'activation ou d'inhibition. Il peut être difficile de retrouver les liens de causalité entre les différents noeuds. Les modèles logiques sont plus adaptés pour l'étude des réseaux de régulation de gène, où le type de réaction est unique : la régulation transcriptionelle. De ce fait il est intuitif de représenter les influences de la même façon. A l'inverse la signalisation regroupe différents types de réactions qui sont interprétées de la même façon. Ce niveau d'abstraction offert par les modèles logiques semble moins intuitif lorsqu'il est appliqué à la signalisation.

Les modèles booléens et logiques sont supportés par différents logiciels permettant la conception ainsi que l'analyse statique et/ou dynamique de ces modèles : GinSim [103], CellNetAnalyzer [77].

A noter également que les formalismes présentés dans cette introduction peuvent être employés avec différents point de vues, à l'instar des réseaux de Petri qui peuvent être utilisés avec une vision centrée sur les molécules. En ce sens, différents travaux proposent

une interprétation des modèles logiques en réseaux de Petri dans le cadre de la signalisation cellulaire [117, 122].

2.2.3 Les approches multi-échelles

La biologie couvre un spectre extrêmement large en terme d'échelle de temps (de la milliseconde à plusieurs mois pour le développement des tissus) et d'échelle de taille (de l'atome à la molécule, de la cellule au tissu, de l'individu à la population). Les approches précédentes comme les réseaux booléens et autres modèles se concentrent sur une échelle particulière, le plus souvent moléculaire. A de rares exceptions près, ces formalismes ne peuvent être appliqués dans des études multi-échelles. C'est le cas des réseaux de Petri pour étudier la polarité planaire cellulaire qui définit l'orientation d'une cellule en fonction des cellules voisines [43]. Les formalismes suivants sont plus adaptés à modéliser différentes échelles, comme notamment l'interaction entre différentes cellules.

Automate cellulaire

Historiquement les automates cellulaires sont nés d'une collaboration entre John von Neumann et Stanislaw Ulam dans les années 40 [104]. A la différence des formalismes proposés précédemment, les automates cellulaires permettent de représenter explicitement l'espace. En effet, dans les formalismes précédents (modèle booléen, réseaux de Pétri, etc.) la localisation n'est décrite au mieux que dans le nom de la molécule de façon "vague", les automates cellulaires distribuent les composants sur une matrice (généralement bi-dimentionelle). Cette matrice, d'une taille définie par le modélisateur, est divisée en *cellules*. Afin d'éviter toute confusion, nous utiliserons *cellule* écrit en italique pour définir l'unité fondamentale des automates cellulaires, qui n'a aucun lien avec la cellule en biologie. A un pas de temps donné, chaque *cellule* est dans un état donné : libre ou occupée. Une *cellule* peut être occupée par une des entités définies dans le modèle, typiquement une molécule. L'ensemble des états des *cellules* caractérise la configuration de l'automate cellulaire à chaque pas de temps. L'état d'une *cellule* au temps $t + 1$ dépend de son état et de ceux de ses voisins au temps t. Le contenu d'une *cellule* peut se déplacer sur cette matrice, avec une certaine probabilité de mouvement, fixée par le modélisateur. En plus de la possibilité de se déplacer, le contenu d'une *cellule* peut évoluer en fonction de règles définies dans l'automate. Ces règles s'apparentent à des réactions, ne pouvant avoir lieu que si les réactants sont "voisins", c'est à dire si ils sont dans des *cellules* adjacentes.

A titre d'exemple, les règles suivantes définissent les réactions de phosphorylation et complexation :

$$A + K \xrightarrow{P_{kphos}} A_p + K$$

$$A_p + B \xrightarrow{P_{kon}} AB$$

Des probabilités peuvent être associées aux réactions : P_{kphos} et P_{kon} respectivement pour la phosphorylation et la complexation. Compte tenu des probabilités de déplacement et d'interactions, ces modèles ne sont pas déterministes. De nombreuse simulations sont généralement réalisées pour être moyennées. La Figure 2.6 illustre un exemple de simulation sur la phosphorylation de A par K. Un automate cellulaire est initialisé en plaçant chaque *cellule* dans un état donné, ici 2 cellules sont occupées par A et K. Ces entités ont une certaine probabilité de mouvement, et d'interaction (si une cellule adjacente contient un potentiel interactant défini par les règles). L'évolution présentée dans la figure n'est qu'une des dynamiques possibles parmi d'autres. Lorsque les entités A et K se retrouvent dans des cellules adjacentes, elles peuvent interagir. L'entité A est alors changée en A_p.

Compte tenu de leur faculté à retranscrire l'essentiel de la dynamique et à pouvoir être simulé un grand nombre de fois, les automates cellulaires ont été utilisés pour modéliser les mécanismes de signalisation dans le but d'analyser :

- (i) la sensibilité du modèle à la variation de paramètres (quantité de molécules initiales, probabilités...) [70, 4]. Appliqués à la voie MAPK, les automates cellulaires ont permi de déterminer quelle enzyme, et donc quelle réaction de la voie, maximise l'amplification de la propagation du signal [70]. Cette étude a également ciblé des actions qualitatives pour obtenir des effets sur la signalisation, comme inhiber une enzyme spécifique pour augmenter la quantité de telle ou telle molécule de la voie MAPK.

- (ii) l'impact de la topologie sur la dynamique du modèle [5]. Cette étude a classé différents motifs de boucle de régulation en fonction de leur efficacité à propager le signal d'une molécule entrante (e.g. cytokine) vers une molécule sortante (e.g. synthèse d'une protéine cible) (Figure 2.7).

Modèles basés sur des agents (Agent-based models)

Comme le micro-environnement comporte généralement la présence d'autres cellules, il est intéressant de pouvoir prendre en compte plusieurs cellules capables d'interagir entre elles. Les modèles basés sur des agents permettent de modéliser les interactions entre les cellules, ainsi que le comportement de celles ci en réponse à des changements de l'envi-

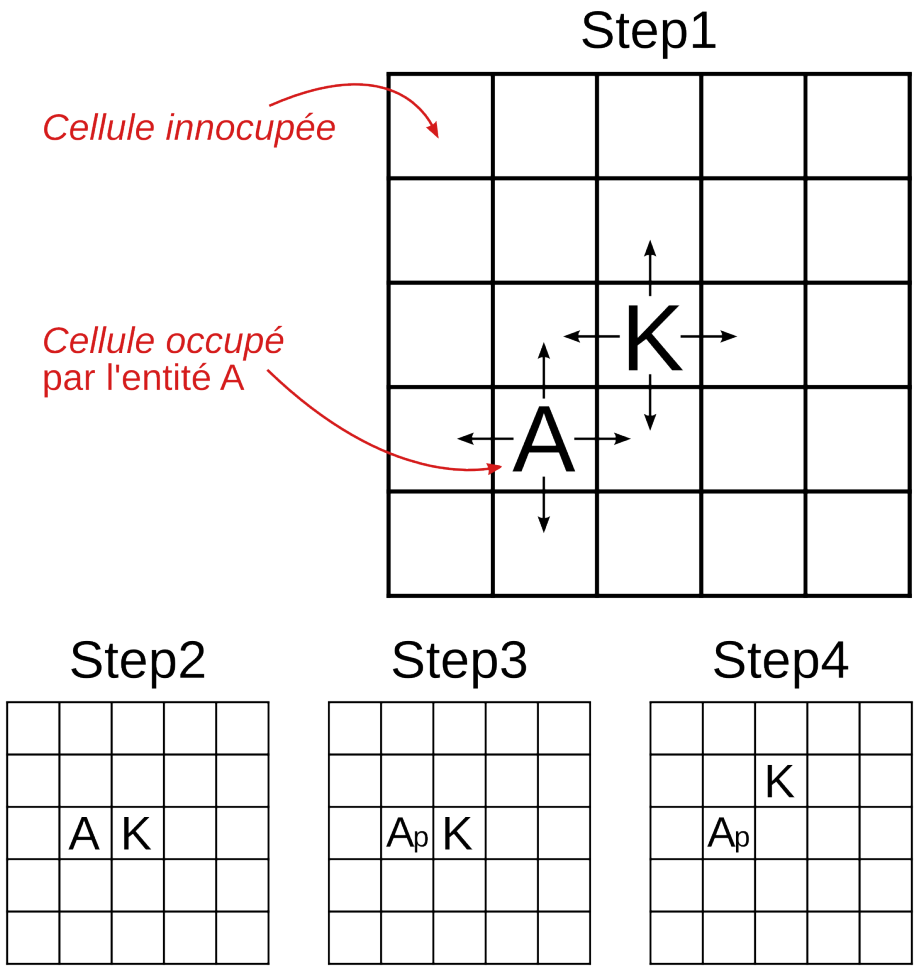


FIG. 2.6: Modélisation de la phosphorylation en automate cellulaire. L'automate est représenté par une matrice 2D comportant 25 *cellules*. Chaque *cellule* peut être occupée ou non, ce qui définie la configuration de l'automate. A chaque pas cette configuration est susceptible de changer en fonction des mouvements et interactions possibles entre les *cellules*. Au premier pas, 1 *cellule* est occupée par l'entité A et une autre par l'entité K. Ces entités peuvent se déplacer sur les *cellules* adjacentes avec une certaine probabilité fixée par l'utilisateur. Ces mouvements possibles sont représentés par les flèches noires. Au deuxième pas, l'entité A s'est déplacée et se retrouve sur une *cellule* adjacente à celle occupée par K, rendant l'interaction entre A et K possible. Au troisième pas, suite à l'interaction définie par la règle $A + K \xrightarrow{P_{kphos}} A_p + K$, l'entité A est changée en entité A_p. Au quatrième pas, l'entité K se déplace sur une *cellule* adjacente.

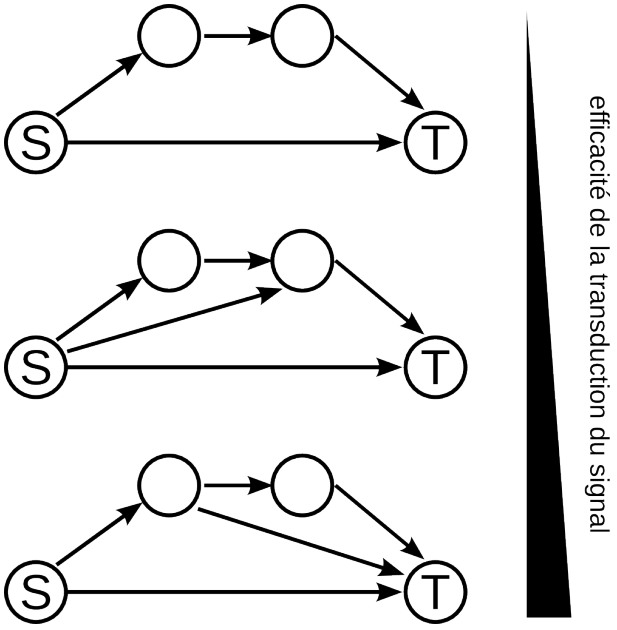


FIG. 2.7: Impact de la topologie sur la transduction du signal d'après [5]. Différents motifs de 4 molécules pouvant interagir sont comparés sur leur faculté à transmettre le signal de S vers T. Un motif est d'autant plus efficace qu'il diminue le temps nécessaire à la conversion de la molécule entrante S en molécule sortante T. L'introduction d'une seconde boucle de contrôle améliore la transmission du signal. La position de cette seconde boucle est également un critère important pour faciliter la conversion du signal.

ronnement. Comme les automates cellulaires, les modèles à base d'agents permettent la spatialisation de leurs composants sur une matrice généralement bi-dimensionnelle. Mais les composant ne sont pas ici des cellules avec simplement 2 états, on parle ici d'agents, qui ont un comportement plus complexe. En effet, chaque agent possède un système de règles lui permettant d'avoir un comportement autonome. Ces règles décrivent un ensemble de réactions, comme une voie de signalisation. De cette façon un agent est souvent utilisé pour représenter, en biologie une cellule. L'avantage est de pouvoir représenter différents types d'agents, c'est à dire différents systèmes de règles pour décrire différents types de cellules.

Grâce à ce formalisme, Walker *et al* ont pu reproduire le comportement observé pour la prolifération des cellules épithéliales [150]. Différents comportements ont été modélisés tels que la prolifération, la migration, l'apoptose. Les agents peuvent émettre et recevoir un signal (ex : cytokine) et ainsi adapter leur comportement en fonction des règles qui les composent et du signal reçu. Cette approche est fondamentalement multi-échelle,

étant capable de décrire les réactions à l'intérieur des cellules, et à plus haut niveau les comportements de ces cellules.

Les modèles à base d'agents ont également été appliqué à la signalisation de NFKB [111]. Plus récemment, le croisement entre les voies TGF et EGF à été exploré avec cette approche [136].

2.2.4 Les approches temporelles

Différentes échelles de temps sont rencontrées en signalisation : de l'activation d'une protéine en quelque seconde, à la régulation d'un gène en plusieurs heures. Le développement de modèles nécessite d'intégrer ces réactions dans un système unique avec une dynamique cohérente. Que ce soit dans les approches centrées sur les réactions ou sur les noeuds, le choix d'un modèle de temps est crucial. Dans les modèles basés sur les réactions, la question est de savoir quelle réaction a lieu à quel moment. Pour les modèles basés sur les molécules, il faut savoir quand les molécules changent d'états. Là encore il n'y a pas de réponse absolue, tout modèle dynamique est basé sur un certain schéma temporel, du temps physique quantitatif, au temps logique qualitatif.

Les schémas de base

Dans les modèles différentiels le temps est explicite, il est intégré dans les constantes des modèles. Il s'agit d'un temps physique, continu et il implique l'existence d'une horloge universelle. C'est à dire que toutes les constantes cinétiques basées sur les expérimentations sont toutes basées sur le même chronomètre, avec un référentiel universel.

Dans les approches discrètes le temps est logique, on s'intéresse surtout à l'ordonnancement des événements. La dynamique dépend de la façon dont sont actualisées les variables (molécules/réactions). Dans les approches molécules centrées, les deux schémas les plus utilisés sont le synchrone et l'asynchrone et la question est de savoir quels sont les noeuds qui sont susceptibles de changer d'état. C'est à dire, les noeuds sur lesquels la règle d'évolution va être appliquée. Le schéma synchrone actualise toutes les variables à chaque pas de temps. Ce schéma est donc déterministe, chaque état du modèle au temps t ne pouvant conduire qu'a un seul état au temps $t+1$. Pour cette raison le schéma dynamique synchrone peut s'avérer trop contraignant et les dynamiques observées *in-vivo* ne sont pas toujours reproduites. A l'inverse, le schéma asynchrone considère qu'une et une seule variable peut être actualisée à chaque pas. Cette dynamique est par définition non déterministe, car plusieurs variables sont potentiellement actualisables, il est nécessaire de choisir une de ces variables ou de tester les multiples possibilités. Ainsi un état au temps t peut entraîner plusieurs états au temps $t+1$. Cette indéterminisme entraîne une augmen-

tation des trajectoires dans le graphe de transition d'états. Pour réduire les possibilités, il est possible de fixer l'ordre dans lequel les variables sont actualisées, rendant ainsi le modèle déterministe. Plusieurs ordres peuvent être simuler, et l'impact sur la dynamique est ensuite analysé. La dynamique asynchrone génère d'avantage de comportements que le schéma synchrone, cependant toutes les possibilités ne peuvent être testées en pratique.

A partir du modèle Booléen présenté en Figure 2.5, le graphe de transition d'états peut être obtenu avec la dynamique synchrone (Figure 2.8A) ou asynchrone (Figure 2.8B). La dynamique asynchrone offre plus de possibilités, cependant pour être complètement exhaustif, il faudrait pouvoir tester toutes les combinaisons possibles où à chaque pas $1, 2, \ldots, n$ variables sont actualisées (n étant le nombre de variables du modèle). Une telle dynamique n'est cependant pas applicable en pratique, le nombre de combinaisons étant bien trop grand pour être simulé.

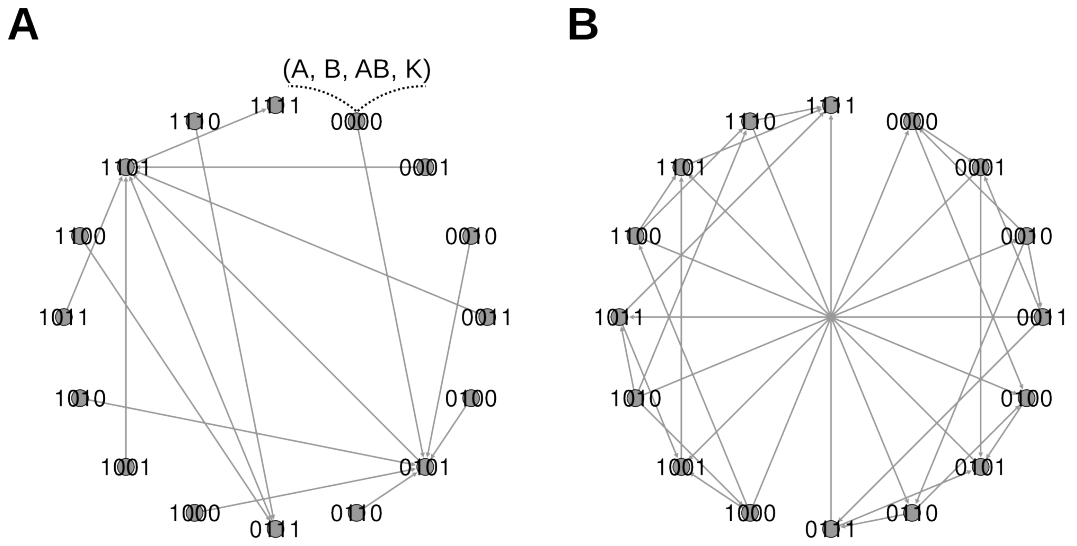


FIG. 2.8: Graphe de transition d'état obtenue avec le modèle présenté en Figure 2.5 suivant les règles du Tableau 2.2. L'état basal des noeuds K et B est égal à 1, les autres étant nuls. (**A**) Dynamique synchrone, toutes les valeurs sont actualisées à chaque pas. La dynamique étant donc déterministe, chaque état possède un unique successeur (potentiellement lui même si c'est un état stable). A l'état 0000, B et K peuvent passer de 0 à 1. L'état suivant est donc 0101. (**B**) Dynamique asynchrone, un seul noeud est actualisé par pas de temps. Lorsque plusieurs noeud peuvent changer de valeurs, il existe plusieurs successeurs à l'état courant. La dynamique est ainsi indéterministe. A l'état 0000, B et K peuvent passer de 0 à 1. L'état suivant peut donc être 0100 si B est actualisé, ou 0001 si K est actualisé. Les deux dynamiques possèdent les mêmes états stables. Ici l'unique état stable (sans arc sortant) est 1111 où tous les noeuds sont à 1.

Pour les modèles basés sur des réactions comme les réseaux de Petri, les conditions nécessaires au tirage des transitions influencent le temps. Contrairement aux modèles molécules centrées, seul un sous-ensemble des valeurs est actualisable à chaque pas. Ce sous-ensemble est déterminé par les transitions tirables ou non. En effet, si une transition n'a pas les conditions requises pour être tirée (nombre de jetons d'une place entrante insuffisant...) alors il n'est pas nécessaire de s'intéresser à la valeur de la place sortante de cette transition. Cependant il faut tout de même choisir parmi les solutions de ce sous-ensemble celle(s) qui seront tirées à chaque pas. En pratique une seule transition est tirée à chaque pas, de cette façon il est possible de rechercher une séquence de transitions conduisant à un état particulier du système.

Affiner le temps

Différentes approches utilisent ces schémas d'évolution en proposant d'affiner la dynamique à l'aide de connaissances biologiques. En connaissant les vitesses de réaction ou au minimum l'ordre de celles-ci, il est possible d'ajouter des probabilités sur chaque transitions. Ces probabilités sont comparables en biologie à des vitesses de réaction. Une réaction rapide possède une probabilité élevée, à l'inverse une réaction lente a une probabilité faible.

Certaines approches proposent l'ajout de noeud *nul*, sans signification biologique, pour différencier les événements précoces des événements tardifs dans la signalisation. Considérons par exemple une molécule M dépendant à la fois d'un activateur A et d'un inhibiteur I. Le noeud I_nul jouera le rôle d'inhibiteur, comme il dépend de I, il mettra un pas de temps de plus à être activé et à empêcher l'activation de M (Figure 2.9). Cette méthode est notamment utilisée pour modéliser l'inhibition tardive de ErbB1 dans la voie EGF [129]. Si cette méthode s'applique sur des modèles de taille raisonnable, de l'ordre de quelques dizaines de molécules, son utilisation sur des modèles large échelle n'est pas envisageable compte tenu des informations requises pour classer l'ordre des réactions.

La durée des réactions, ou le temps de vie de molécules peut être représenté par des modèles stochastiques. Pour déterminer l'ordre dans lequel sont franchies les transitions, ces dernières peuvent être associées à une probabilité en fonction des durées des réactions qu'elles représentent. Ainsi une réaction rapide aura une plus forte probabilité qu'une réaction lente et aura donc lieu le plus souvent en premier.

Il est également possible de calibrer les modèles en connaissant la durée de phénomènes connus [136]. Ainsi si une cascade de signalisation est réalisée au bout de 60 pas, et dure en réalité 1 heure, la durée d'un pas peut être approximée à une minute. Le calibrage facilite la comparaison avec les expérimentations puisqu'il est en mesure d'indiquer un temps physique bien qu'approximatif.

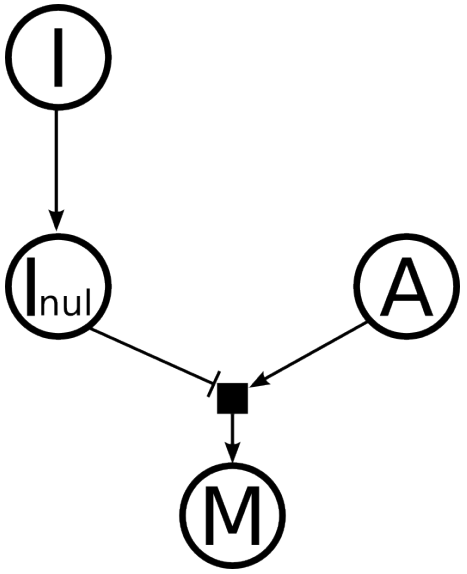


FIG. 2.9: Introduction de noeud "nul" pour différencier des événements précoces et tardifs. La présence de A et l'absence de I sont nécessaires à l'activation de M. L'inhibition par I est un phénomène tardif par rapport à l'activation. Le noeud I_nul est activé par I et le remplace en tant qu'inhibiteur. L'inhibition de M par I demande plus de pas de temps que l'activation par A et reflète bien le retard observé.

2.3 Données et conception des modèles

Tout modèle est basé sur des données, elles sont le fondement d'une approche de modélisation. Ce sont elles qui conditionnent les limites de la modélisation, celles sur quoi repose souvent la réussite d'une étude. Les données se distinguent d'une part par leur type, et d'autre part par leur moyen d'acquisition.

Certains formalismes utilisent des données quantitatives (modèles différentiels, rule based, etc.) telles que des concentrations et des constantes cinétiques. D'autres formalismes utilisés dans les modèles Booléen et les réseaux de Petri se basent sur des données qualitatives.

La majorité des approches de modélisation actuelles autour de la signalisation sont basées sur une acquisition manuelle des données. En effet, la littérature demeure la source la plus utilisée pour concevoir un modèle. Ceci s'explique de plusieurs façon : l'acquisition manuelle permet une meilleur curation et donc un contrôle plus strict de la qualité des données. Une meilleure interprétation des résultats est aussi obtenue.

En contrepartie, cette acquisition demande énormément de temps, en comparaison avec une acquisition automatique puisque qu'elle dépend de ressources humaines. En effet,

certains modèles sont basés sur près d'un millier de publications, faisant de l'interprétation de résultats vers la conception de modèle une activité des plus chronophage [55]. En lien avec ce recueil de données à partir de la littérature, la définition des limites du modèle constitue une difficulté importante. Les modèles créés par l'interprétation manuelle de la littérature sont généralement conçus autour d'une voie de signalisation ou d'une question spécifique. Dans ce contexte il faut définir les limites d'un modèle, c'est à dire quels sont ses composants et ses réactions. La décision d'ajouter ou non une réaction dans une voie de signalisation est un critère hautement "auteur spécifique" [72] et conduit à des modèles différents pour un même processus biologique. De plus l'homogénéité de l'interprétation n'est pas garantie, à l'inverse d'une acquisition automatique où plusieurs réactions du même type (phosphorylation, translocation) seront traitées avec le même pattern défini au préalable par le modélisateur. Dans le même esprit, tout modélisateur s'est déjà demandé si la connaissance dirigeait la conception des modèles. Ce qui signifie qu'en connaissant à la fois les données et la question posée, il est possible, de façon tout à fait involontaire de concevoir un modèle spécifique dont l'interprétation sera orientée vers la réponse. L'interprétation demeure un processus subjectif et il est évident que différentes personnes, avec les même données et le même formalisme, peuvent concevoir des modèles différents suivant leur connaissance de la question posée.

Pour éliminer ces potentiels défauts dus à une interprétation manuelle, il est possible de concevoir un modèle de façon automatique, à partir d'une base de données dont le recueil des connaissances aura été effectué sans a priori par rapport au modèle à construire. Pour que la conception automatique de modèle soit efficace, la base de données doit contenir suffisamment de données dont la qualité doit être vérifiée afin de permettre une interprétation identique et non ambigue du contenu da la base. Cette difficulté est évoquée dans les travaux de Heiner *et al* [52] où plusieurs interprétations de la représentation de la base de donnée KEGG sont présentées(Figure 2.10). Sans connaissance sur les données, la représentation adoptée par KEGG ne permet pas de trancher entre une activation de CASP2 et CASP3 par CASP8, ou une réaction consommant CASP8 pour produire CASP2 et CASP3. Cette exemple souligne l'ambiguïté de cette représentation et la nécessité de travailler avec des données clairement définies.

L'intérêt de la démarche automatique réside dans le fait que les données n'ont pas été structurées dans le but de concevoir un modèle et n'oriente pas la construction du modèle dans le sens de la réponse à une question. L'interprétation automatique permet de prendre en compte toute l'information sans a priori, et de générer des modèles plus homogènes et formalisés.

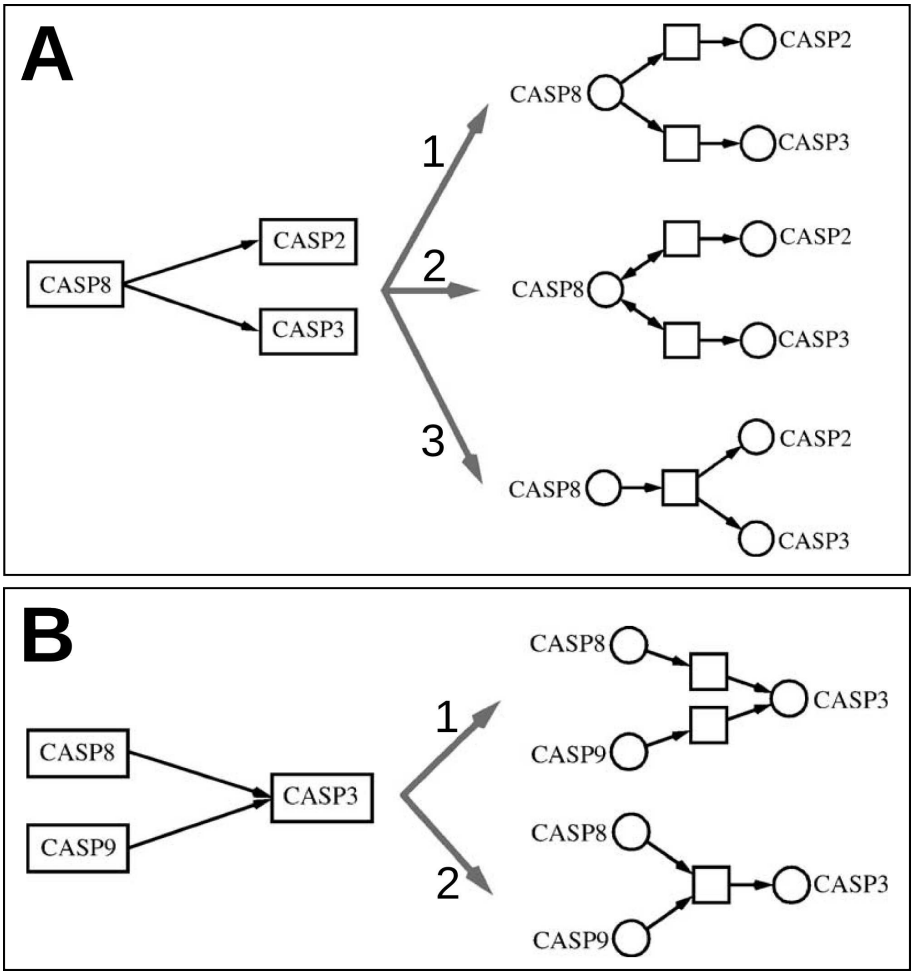


FIG. 2.10: Interprétations multiples de schéma KEGG en réseau de Petri d'après [52]. (**A**) L'activation de CASP2 et CASP3 par CASP8 symbolisée par des flèches peut être interprétée de différentes façons. CASP8 permet l'activation de CASP2 ou CASP3 de façon exclusive (cas 1), ou permettre l'action de CASP2 et CASP3 à différents moments (cas 2). Enfin l'activation peut être simultanée (cas 3). (**B**) Les activations de CASP3 par CASP8 et CASP9 peuvent etre considérées comme indépendantes (cas 1) ou conjointes (cas 2). Si elles sont indépendantes, la présence de CASP8 ou CASP9 suffit à activer CASP3. Si elles sont conjointes, la présence de CASP8 et CASP9 est nécessaire pour activer CASP3. Actuellement la notation graphique utilisée par la base de donnée KEGG ne permet pas de choisir entre les différents cas.

2.4 Contributions de la thèse

Cette thèse s'inscrit dans la volonté d'intégrer des connaissances biologique et de les représenter sous divers formalismes de façon non ambiguë. La formalisation des connaissances permet non seulement de reproduire, mais surtout de prédire un comportement. Le choix de la représentation étant guidé par la question posée, nous présentons ici deux approches pour répondre à des questions de granularités différentes. Nos approches de modélisation se sont portées sur le processus de signalisation cellulaire et notamment sur la voie de signalisation de facteur de croissance TGF-β.

Dans une premier étude, nous avons développé des modèles différentiels pour étudier le rôle d'un nouveau régulateur transcriptionel TIF1γ dans la signalisation canonique du TGF-β. En effet le rôle de TIF1γ fait l'objet de controverses dans la littérature et nous avons formalisé les observations divergentes dans des modèles différentiels. La confrontation des dynamiques des différents modèles avec les données expérimentales nous a permis de discriminer les différentes hypothèses et de proposer un modèle unique expliquant la fonction de TIF1γ.

Dans un second temps, la recherche d'un modèle intégrant la complexité de la signalisation du TGF-β nous a conduit à développer une approche globale prenant en compte l'ensemble des données sur la voies de signalisation cellulaire. Dans ce contexte nous avons opté pour une approche discrète et dynamique, qui offre la possibilité de représenter des modèles de grandes tailles. Pour cela nous nous sommes focalisé sur (i) le développement d'une méthodologie, (ii) sa concrétisation à travers un logiciel et (iii) l'application de cette méthodologie à l'étude de régulations de processus complexes par la signalisation cellulaire.

Nos efforts se sont donc tout d'abord portés sur le choix d'un formalisme adapté aux caractéristiques de la signalisation et compatible avec des modèles composés de milliers de molécules. Nous avons choisi une approche basée sur les réactions, permettant de modéliser la propagation du signal/de l'information de la membrane aux gènes. Un grand intérêt a été porté à la gestion du temps dans notre système dynamique, de façon a pouvoir représenter les croisements entre les voies de signalisation de façon non biaisée, sans être contraint par le manque d'information sur la dynamique du système biologique d'origine. Nous avons ainsi développé le formalisme Cadbiom pour *Computer Aided Design of BIOlogical Models*, basé sur les transitions gardées en y introduisant la notion d'événements pour palier à l'absence de représentation du temps. Pour concevoir nos modèles nous avons utilisé les connaissances stockées dans les bases de données. Notre interprétation de la biologie est fondée sur la notion de réaction biologique, ce qui nous a permis d'établir un

schéma de traduction des bases PID et Reactome en modèles Cadbiom.

Pour promouvoir notre méthodologie et faciliter son utilisation, nous avons développé le logiciel Cadbiom. L'interface graphique propose un accompagnement complet de la conception à l'analyse de nos modèles. Les analyses des modèles dynamiques s'inspirent des vérifications formelles étudiant des propriétés telles que l'atteignabilité et l'invariance.

Afin d'éprouver notre modèle généré à partir de la base de données PID, nous avons exploré la dynamique du signal permettant la régulation du cycle cellulaire, processus fondamental en biologie. Dans un deuxième temps nous avons pratiqué par une approche globale la relation entre la régulation des gènes et la complexité des mécanismes de signalisation.

Deuxième partie

Modélisation de la régulation du signal canonique du facteur de croissance TGF-β par le facteur de transcription TIF1γ

Chapitre 1

Présentation

Cette première partie est consacrée à la modélisation quantitative de la régulation de la voie de signalisation canonique du TGF-β par un nouvel interactant : TIF1γ(également appelé TRIM33 ou ectodermin). Ce dernier appartient à la famille des protéines TRI partite Motif (TRIM), et a été récemment impliqué dans la régulation du signal TGF-β. Cependant les observations sont en partie contradictoires et le rôle de TIF1γ au sein de la signalisation TGF-β ne fait pas l'objet d'un consensus. Dans ce contexte, nous avons développé plusieurs modèles pour comprendre l'impact de TIF1γet analyser ses effets en modifiants les paramètres tels que la concentration de TIF1γ, de TGF-β ou encore de la durée d'exposition au TGF-β.

Dans une première partie, nous présentons la voie de signalisation du TGF-β dite canonique et dépendante des protéines SMAD, et les principaux résultats issus de la littérature, autour des relations entre TIF1γ et la signalisation TGF-β qui ont permis d'établir les hypothèses de modélisations. Dans une seconde partie les résultats de notre approche sont présentés sous forme d'article.

1.1 La signalisation du TGF-β

Le facteur de croissance transformant (TGF-β1) est une cytokine appartenant à la super-famille des facteurs morphogénétiques, regroupant les Bone Morphogenetic Proteins (BMP), les Growth Differentiation Factors (GDF), les Müllerian Inhibiting Substance (MIS) et les Transforming Growth Factor (TGF-β1,2,3). Cette famille régule une large gamme de processus biologiques tel que la prolifération et la différentiation cellulaire, l'apoptose et la motilité. Par commodité nous utiliserons le terme générique TGF-β pour parler du TGF-β1. Au sein des tissus, le TGF-β régule l'homéostasie cellulaire en contrôlant la prolifération et la mort cellulaire. Par contre dans des contextes pathologiques comme le cancer, le TGF-β se comporte comme un agent pro-tumoral, induisant la prolifération et l'invasivité des cellules [94]. Cet effet pléiotropique est associé à la diversité

de ses mécanismes de signalisation (Figure 1.1). Son rôle anti- ou pro-tumoral suivant le contexte à la fois intra- et extracellulaire, lui a valu le surnom de *Jekyll and Hyde protein* [12].

1.1.1 La voie canonique de la signalisation TGF-β

En condition physiologique, le rôle majeur du TGF-β est de bloquer la prolifération cellulaire. Pour cela il induit un signal par l'intermédiaire des protéines SMAD. Cette voie,appelée voie canonique du TGF-β, est décrite dans les paragraphes suivant et illustrée par la Figure 1.1.

Activation des récepteurs transmembranaires La signalisation du TGF-β est initialisée par la fixation du TGF-β à son récepteur de type II (TβRII). Les TβRII sont organisés en homodimères, constitutivement actifs. La fixation du TGF-β entraîne le recrutement des récepteurs de type I (TβRI), et leur activation par transphosphorylation des résidus sérine/thréonine par l'activité kinase des TβRII. Le complexe actif des récepteurs aux TGF-β se présente sous la forme d'un hétérotétramère composé d'un dimère TβRII de d'un dimère TβRI. En plus de ces récepteurs directement impliqués dans la signalisation du TGF-β, d'autres récepteurs peuvent réguler la signalisation TGF-β, tel que le betaglycan (également appelé TβRIII) et l'endogline facilitant la présentation du TGF-β au TβRII. Le récepteur I activé va ensuite recruter et phosphoryler les protéines SMAD.

Transduction du signal Les SMAD sont les messagers secondaire utilisés pour propager le signal TGF-β. Cette famille de protéines comporte 8 membres, qui peuvent être regroupées en trois catégories, sur des critères fonctionnels et structuraux : les SMAD activées par les récepteurs ou R-SMAD (SMAD1, 2, 3, 5 et 8), la Co-SMAD (SMAD4) et les SMADS inhibitrices ou I-SMAD (SMAD7 et 8). La voie canonique du TGF-β de type 1 fait intervenir les SMAD 2, 3, 4 et 7 qui possèdent toutes les quatre un domaine MH1 et un domaine MH2 relié par un linker. L'activation par phosphorylation en C-terminal des SMAD2 et 3 par TβRII se fait au niveau du domaine MH2. Le domaine MH1 est responsable de la fixation aux facteurs de transcription. Le linker est important pour les processus de régulation, sa phophorylation par la voie MAPK pouvant induire l'inhibition de la voie TGF-β canonique [14]. Suite à leur activation, les SMAD2 et 3 peuvent former des complexes hétérodimerique avec SMAD4 et ces complexes actifs sont ensuite transportés dans le noyau.

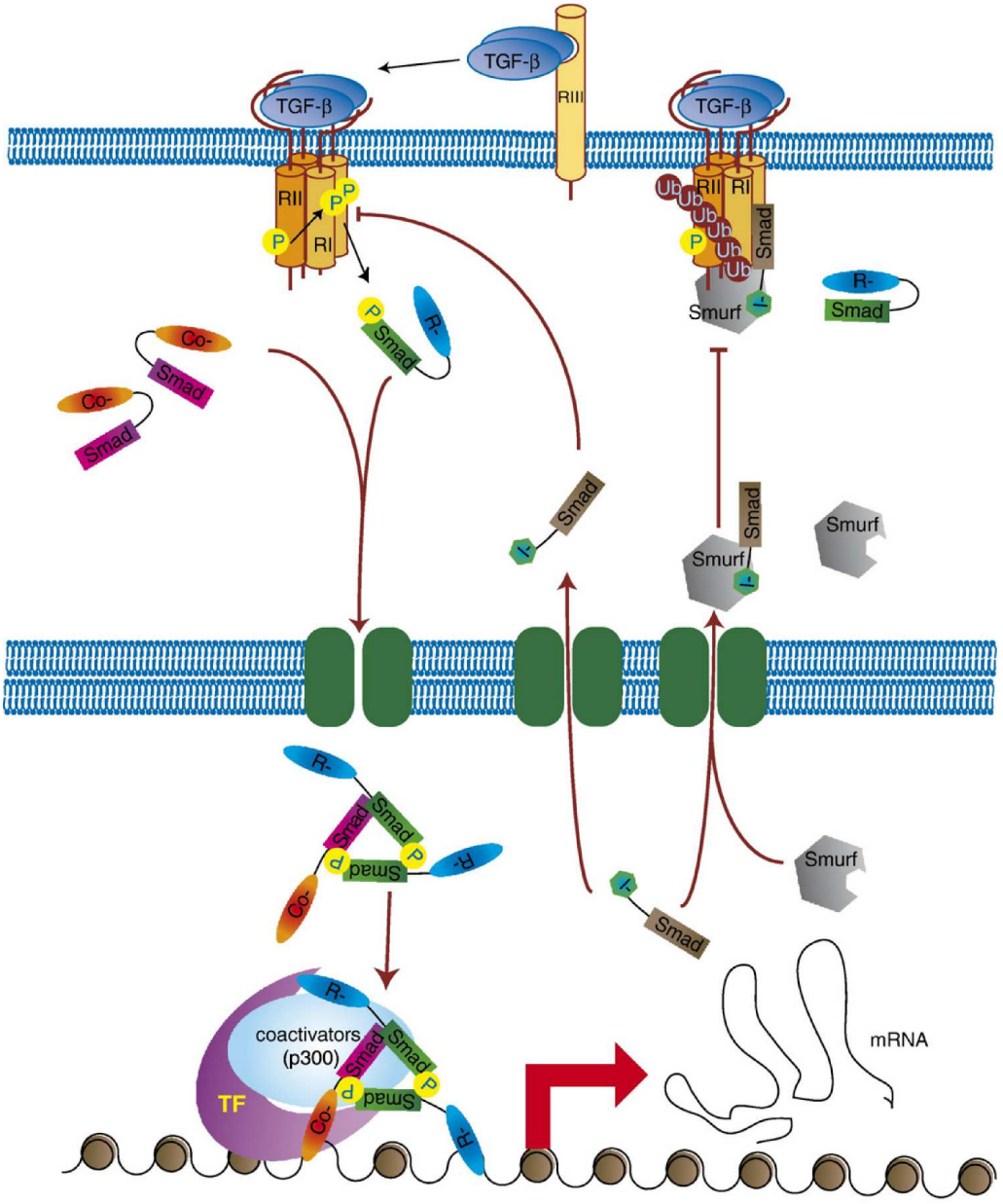


FIG. 1.1: Représentation schématique de la voie canonique du TGF-β d'après [109]. Le bétaglycan (RIII) facilité la présentation du TGF-β avec le récepteur de type 2 (RII). Le récepteur de type 1 (RI) est activé par transphosphorylation et phosphoryle ensuite les R-SMAD sur leurs domaine MH2. Les R-SMAD activées forment un complexe avec la Co-SMAD, qui est ensuite transporté vers le noyau et recrute de facteurs de transcription (TF) et de co-activateurs pour induire l'expression des gènes cibles. L'expression de I-SMAD permet le rétro-contrôle négatif de la voie canonique, et la dégradation des récepteurs suite au recrutement de l'ubiquitine ligase Smurf.

Recrutement de facteurs de transcription et régulation de gènes cibles Pour réguler l'expression des gènes cibles du TGF-β, les complexes SMAD s'associent avec différents facteurs de transcription. Ainsi pour induire le blocage du cycle cellulaire, le facteur de transcription SP1, peut être recruté pour activer la la trancription de la protéine p21 (également appelée CDKN1A) appartenant à la famille des *Cyclin Dependant Kinase Inhibitor* [54, 101]. La voie canonique permet aussi la répression de c-Myc, l'empêchant ainsi d'inhiber p21 [54, 154]. Le transcription de SMAD7 est également activées, responsable d'un rétrocontrôle négatif en participant à la dégradation des récepteurs activé, avec l'aide de de la protéine Smurf2.

1.1.2 Voies dites "non SMAD" de la signalisation TGF-β

La pléiotropie des effets du TGF-βrepose notamment sur la diversité de ses mécanismes de signalisation non canonique. Les récepteurs du TGF-β sont en mesure d'activer d'autres protéines que les SMAD, entraînant des effets différents. Ces voies sont regroupées sous le terme de voies non-SMAD parmi lesquelles : (i) les voies des MAP Kinases (p38, JNK, ERK), (ii) les voies des Rho-like GTPase et (iii) la voie PI3K.

Ces voies non-SMAD ne sont pas totalement indépendantes de la voie SMAD et il existe des régulations entre toutes ces voies. Ainsi les SMAD peuvent être phosphorylées au niveau du linker par la voie MAPK [81], ce qui supprime l'activité des SMAD et empêche leur rétention dans le noyau. La pléiotropie des effets du TGF-β en conditions physiologiques et pathologiques (Figure 1.2) s'explique par la complexité de ces voies SMAD et non SMAD et est détaillée dans de récentes revues [109, 82, 95].

1.2 Contexte biologique autour de TIF1γ

L'implication du facteur transcriptionel TIF1γ dans la signalisation canonique du TGF-β repose sur les travaux de deux équipes [37, 51, 36] et ont abouti à considérer TIF1γ soit comme un inhibiteur de la voie canonique, soit comme un inducteur d'une voie alternative.

1.2.1 TIF1γ : un inhibiteur de la voie canonique

C'est en recherchant les protéines responsable de la différentiation de l'endoderme que l'équipe de Piccolo, a identifié TIF1γ comme étant un répresseur de la voie canonique du TGF-β [37]. Dans cette étude Dupont *et al* ont démontré que TIF1γ interagit avec SMAD4 et que cette association conduit à l'ubiquitination de SMAD4 par TIF1γ l'empêchant ainsi de se fixer à SMAD2 phosphorylé (pS2), allant même jusqu'à dissocier les complexes

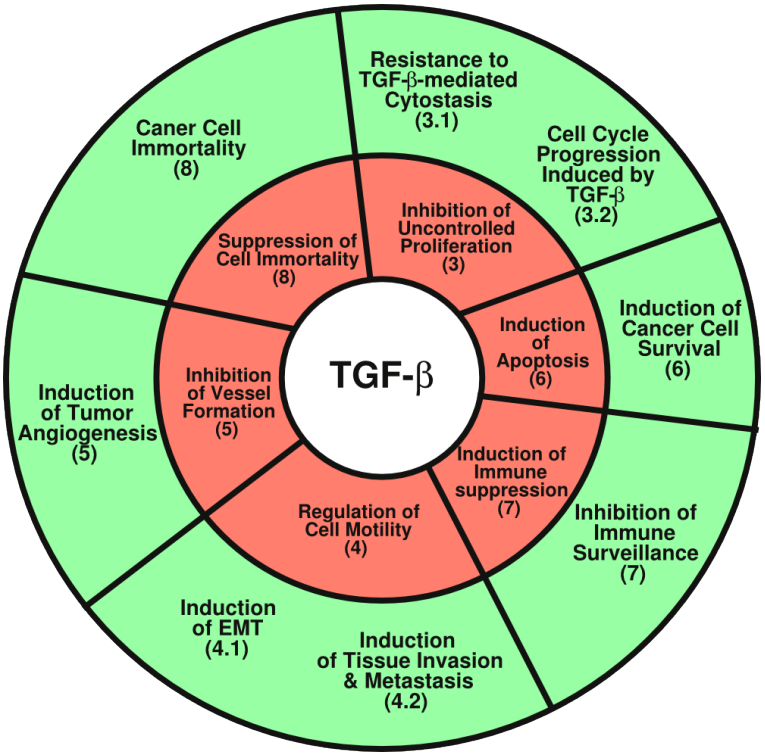


FIG. 1.2: Les différents effets du TGF-β en conditions physiologiques (en rouge) et pathologiques (en vert) d'après [142]

SMAD2 phosphorylé/SMAD4 (pS24). Par conséquent la diminution de complexe pS24 aboutit à une diminution du signal TGF-β dans le noyau et traduit un rôle fortement inhibiteur de TIF1γ sur la signalisation TGF-β. Ces résultats sont complétés dans une autre étude par de nouvelles analyses montrant que l'ubiquitination de SMAD4 n'est pas un processus entraînant la dégradation de la co-SMAD, mais qu'il existe un signal d'exportation nucléaire, la protéine étant dé-ubiquitiniée dans le cytoplasme par la protéine FAM afin de régénérer la protéine SMAD4 [36].

1.2.2 TIF1γ : vers une voie de régulation alternative à la voie canonique

Dans le même temps, l'équipe de Massagué a identifié une interaction entre TIF1γ et les protéines SMAD2 (et SMAD3) phosphorylé conduisant à un nouveau signal régulant la différentiation cellulaire [51]. Leurs travaux démontrent que l'association TIF1γ / SMAD2 phosphorylé n'inhibe pas le signal TGF-β et ne trouve aucune ubiquitination de SMAD4 par TIF1γ contrairement au résultats de [37, 36]. Cependant une faible association entre

TIF1γ et SMAD4 est observée lorsque ce dernier est surexprimé suggérant que l'association de TIF1γ avec SMAD2-3 n'est pas incompatible avec une association avec SMAD4.

Si ces résultats semblent antagonistes, les deux équipes s'accordent sur certains points. Le rôle de TIF1γ est dépendant de la signalisation TGF-β et les effets ne sont observés qu'après fixation de la cytokine. La diminution de l'effet TGF-β dépendant de SMAD4 est observée, soit parce que SMAD4 est en compétition avec TIF1γ pour le pool de SMAD2 phosphorylée [51], soit parce que SMAD4 est ubiquitinée par TIF1γ [37, 36]. De plus TIF1γ est toujours dans le noyau et ses effets sont fonction de sa concentration. Afin d'explorer différentes hypothèses nous avons composé un modèle différentiel de la voie canonique du TGF-β pour y ajouter les informations relatives à TIF1γ.

1.3 Modèles quantitatifs de la voie canonique du TGF-β

Depuis plusieurs années la voie canonique du TGF-β à fait l'objet d'approches de modélisation différentielle. Ainsi d'un coté le trafic des récepteurs membranaires a été modélisé par Vilar *et al* [146], d'un autre l'importance des régulations sur les protéines SMAD tel que le transport [132], la complexation [102] ou la phosphorylation [26]. Des modèles plus récent regroupent l'ensemble de la voie et observe le comportement de leurs modèles en fonction de la concentration et du temps d'exposition au TGF-β[25, 158]. D'autres approches ont exploré l'interface entre la voie canonique du TGF-β et celle de l'EGF [71], ou encore la régulation par ADAM12 [49].

Pour implémenter la régulation du signal TGF-β par TIF1γ, nous avons utilisé le modèle du trafic des récepteurs proposé par Vilar *et al* [146] ainsi que le modèle du transport des protéines SMAD de Schmierer *et al* [132] afin de représenter le plus fidèlement possible la voie canonique de signalisation du TGF-β.

1.3.1 Trafic des récepteurs

Ce modèle couvre le début de la signalisation intracellulaire du TGF-β, après la fixation de ce dernier sur les récepteurs membranaires puis décrit les mécanismes d'internalisation des récepteurs, activation, désactivation, production et dégradation. Les constantes ont été déterminées de façon expérimentale et le modèle vérifié à posteriori. Le tableau 1.1 reprend les 6 entités, et 8 paramètres du modèle, dont l'organisation est illustrés par la Figure 1.3A

1.3.2 Transport des SMADs

Publié par l'équipe de Hill en 2008 [132], ce modèle apporte une description précise des mécanismes de transport des protéines SMAD entre cytoplasme et noyau, forme monomèrique et complexé, en présence ou en absence de TGF-β. Les récepteurs sont abstraits à une seule entité, qui peut être activée en présence de TGF-β. Le transport des SMAD est une part importante de la cinétique du signal TGF-β permettant une réponse adaptée, à la concentration du ligand. Le modèle comporte 15 entités et 8 paramètres présentés dans le tableau 1.1. La Figure 1.3B montre schématiquement les réactions présentes dans ce modèle.

1.3.3 Modèle composé

Pour offrir une modélisation la plus réaliste possible, nous avons pris en compte les deux mécanismes importants que sont le trafic des récepteurs et le transport des SMAD pour composer un unique modèle de la voie canonique de la signalisation TGF-β qui servira de support à l'intégration des données sur le nouveau composant TIF1γ. Pour se faire, nous avons remplacé les entités relatives à l'activation/inactivation des récepteurs dans le modèle de Hill (R, $Ract$, $Rinact$, SB), par le modèle complet de Vilar. Réunir deux modèles différentiels en un seul requiert d'homogénéiser les concentrations initiales des entités, les paramètres et les constantes.

Nous avons choisi d'exprimer les concentrations en nano molaire (nM) et le temps en seconde (s). Pour cela, nous avons calculé une concentration initiale de récepteurs dans une cellule, en nous basant sur un nombre de 10000 récepteurs [149] (5000 RI et 5000 RII). Ce nombre à été converti en mole en le multipliant par le nombre d'Avogadro. Puis en nano molaire, en considérant le volume équivalent à celui du cytoplasme ($2.27 * 10^{-12} L$). De la même manière, les taux de production pRI et $pRII$ exprimés en $unité.min^{-1}$ ont été convertis en $nM.s^{-1}$.

1.4 Hypothèses sur le rôle de TIF1γ

Nous avons développé 3 modèles, en considérant les résultats de l'une ou l'autre des deux équipes soit de façon séparé (hypothèse 1 et 2) soit au sein d'un même modèle (hypothèse 3). Nous avons utilisé le module Numpy du langage de programmation python pour intégrer les système d'équations différentielles. Pour la visualisation des courbes, nous utilisons Matplotlib, un autre module python dédié à l'affichage de données numériques complexes.

Trafic des récepteurs	
Symbol	**Definition**
RI	TGF type I receptor
RII	TGF type II receptor
LR	ligand receptor I receptor II complex
RIe	endosomal TGF type I receptor
RIIe	endosomal TGF type II receptor
LRe	endosomal ligand receptor I receptor II complex
k_a	Ligand receptor association rate
k_{cd}	constitutive degradation rate
k_{lid}	ligand induces degradation rate
k_i	internalization rate
k_r	recycling rate
pRI	receptors I production rate
pRII	receptors II production rate
alpha	efficiency of recycling of active receptors

Transport des SMAD	
Symbol	**Definition**
S2c	cytoplasmic SMAD2
S2n	nuclear SMAD2
S4c	cytoplasmic SMAD4
S4n	nuclear SMAD4
pS2c	cytoplasmic phospho SMAD2
pS2n	nuclear phospho SMAD2
pS24c	cytoplasmic SMAD2 SMAD4 complex
pS24n	nuclear SMAD2 SMAD4 complex
pS22c	cytoplasmic SMAD2 SMAD2 complex
pS22n	nuclear SMAD2 SMAD2 complex
TGFβ	transforming growth factor beta
PPase	Phosphatase
R	TGFβ receptors
Ract	activated TGFβ receptors
Rinact	inactivated TGFβ receptors
SB	SB-431542 (inhibitor of activated receptors)
k_{in}	import rate
k_{ex}	export rate
k_{phos}	phosphorylation rate
k_{dephos}	dephosphorylation rate
CIF	complex import factor
k_{on}	Smad complex association rate
k_{off}	Smad complex separation rate
k_{TGF}	receptors activation rate

Tab. 1.1: Entités et constantes des modèles différentiels du trafic des récepteurs d'après [146] et du transport des SMAD d'après [132]

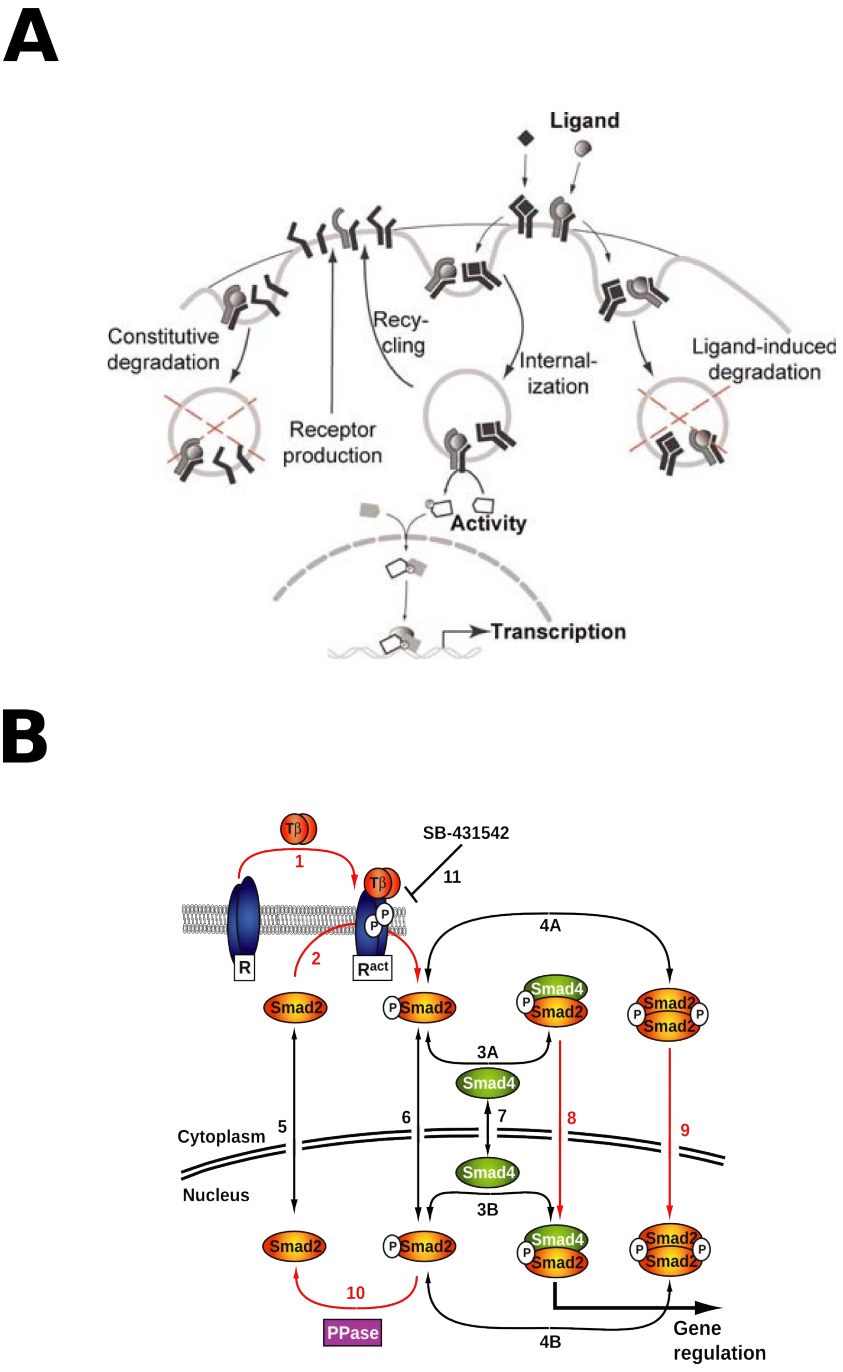


FIG. 1.3: Représentation schématique des réactions entre les entités présentées dans le Tableau 1.1 pour le modèle du trafic des récepteurs (**A**) d'après [146] et le modèle du transport des SMAD (**B**) d'après [132].

1.4.1 Hypothèse 1 : Ubiquitination de SMAD4 par TIF1γ

En accord avec les résultats [37, 36] nous proposons un modèle où TIF1γ est capable de dissocier les complexes pS24n et d'induire dans le même temps l'ubiquitination de SMAD4 dans le noyau (S4ub_n). SMAD4 ubiquitinée est ensuite exportée vers le cytoplasme (S4ub_c) où FAM est capable de dé-ubiquitiner SMAD4. Les espèces suivantes ont donc été ajoutées au système différentiel précédent : S4ub_c, S4ub_n, TIF1γ, FAM.

En utilisant les données expérimentales publiées par ces auteurs, nous avons été en mesure d'inférer les paramètres des nouvelles réactions. Ainsi l'association entre TIF1γ et le complexe SMAD2 phosphorylée/SMAD4 (pS24n) présente la même cinétique que l'association entre SMAD2 phosphorylée et SMAD4. De la même manière la cinétique du processus d'ubiquitination/dé-ubiquitination est considéré comme similaire à la phosphorylation/dé-phosphorylation de SMAD2. Enfin l'export de SMAD4 ubiquitinée étant décrit comme plus rapide que celui de SMAD4 nous avons choisi d'appliquer un facteur 2 au coefficient d'export.

1.4.2 Hypothèse 2 : Association TIF1γ/pS2n

Basé sur les résultats de l'équipe de Massagué [51], ce deuxième modèle décrit l'association préférentielle entre TIF1γ et SMAD2 phosphorylée. Dans ce modèle TIF1γ entre donc en compétition avec SMAD4 pour le pool de SMAD2 phosphorylée. Comme mentionné dans [51], une association entre TIF1γ et SMAD4 est tout de même possible mais avec une affinité moins grande.

Comme pour l'hypothèse 1, nous avons inféré les paramètres des nouvelles réactions à l'aide des données expérimentales issues des publications. Ainsi l'association / dissociation entre TIF1γ et pS2n est considérée comme identique à celle entre pS2n et S4n.

1.4.3 Hypothèse 3 : Modèle hybride

Ce modèle a pour but de réconcilier les observations en apparence divergentes des précédentes publications [51, 37, 36]. Nous avons ici proposé la formation d'un complexe ternaire (pS24nTIF1γ) entre TIF1γ, pS2n et S4n. En se dissociant, le complexe génère d'une part des complexes pS2n/TIF1γ (conformément à [51]) mais aussi SMAD4 ubiquitinée (conformément à [36]). Cette hypothèse est fondée sur la description d'une association préférentielle de TIF1γ avec le complexe pS24n [37]. La possible existence de ce complexe étant également suggérée dans les données supplémentaire de [36]. Les valeurs des constantes d'association/dissociation entre TIF1γ et pS24n sont les mêmes que celles décrites dans les hypothèses précédentes. De même pour les constantes liées à l'ubiquiti-

nation/dé-ubiquitination de SMAD4.

Ces 3 hypothèses ont été analysées par simulation de façon à déterminer quel modèle et quels paramètres sont à même de rendre compte de toutes les observations biologiques autour de TIF1γ. Les résultats de notre étude ont été publié dans la revue PLOS ONE sous le titre *Dynamical regulation of TGF-β signaling by TIF1γ : a computational approach.*

Chapitre 2

Article : Dynamic regulation of TGF-β signaling by TIF1γ : a computational approach. – PLOS ONE 2012

OPEN ACCESS Freely available online

PLoS one

Dynamic Regulation of Tgf-B Signaling by Tif1γ: A Computational Approach

Geoffroy Andrieux[1,2], Laurent Fattet[3], Michel Le Borgne[2], Ruth Rimokh[3], Nathalie Théret[1]*

1 Inserm U1085-IRSET, Université de Rennes 1, Rennes, France, **2** Université de Rennes 1, IRISA, Rennes, France, **3** Inserm U1052/CNRS 5286, Centre de Recherche en Cancérologie de Lyon, Lyon, France

Abstract

TIF1γ (Transcriptional Intermediary Factor 1 γ) has been implicated in Smad-dependent signaling by Transforming Growth Factor beta (TGF-β). Paradoxically, TIF1γ functions both as a transcriptional repressor or as an alternative transcription factor that promotes TGF-β signaling. Using ordinary differential-equation models, we have investigated the effect of TIF1γ on the dynamics of TGF-β signaling. An integrative model that includes the formation of transient TIF1γ-Smad2-Smad4 ternary complexes is the only one that can account for TGF-β signaling compatible with the different observations reported for TIF1γ. In addition, our model predicts that varying TIF1γ/Smad4 ratios play a critical role in the modulation of the transcriptional signal induced by TGF-β, especially for short stimulation times that mediate higher threshold responses. Chromatin immunoprecipitation analyses and quantification of the expression of TGF-β target genes as a function TIF1γ/Smad4 ratios fully validate this hypothesis. Our integrative model, which successfully unifies the seemingly opposite roles of TIF1γ, also reveals how changing TIF1γ/Smad4 ratios affect the cellular response to stimulation by TGF-β, accounting for a highly graded determination of cell fate.

Citation: Andrieux G, Fattet L, Le Borgne M, Rimokh R, Théret N (2012) Dynamic Regulation of Tgf-B Signaling by Tif1γ: A Computational Approach. PLoS ONE 7(3): e33761. doi:10.1371/journal.pone.0033761

Editor: Dipankar Chatterji, Indian Institute of Science, India

Received December 13, 2011; **Accepted** February 21, 2012; **Published** March 23, 2012

Funding: This work was supported by the Institut National de la Santé et de la Recherche Médicale (www.inserm.fr) and the Ligue Nationale Contre le Cancer (http://www.ligue-cancer.net). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nathalie.theret@univ-rennes1.fr

Introduction

Complex signaling by transforming growth factor β(TGF-β) forms a pivotal network that plays an essential role in tissue homeostasis and morphogenesis. At the same time, up-regulation and activity of TGF-β has been linked to various diseases, including fibrosis and cancer, by promoting cell proliferation and invasion and the epithelial-mesenchymal transition [1]. TGF-β signaling occurs through association with a heteromeric complex of two types of transmembrane serine/threonine kinases, the type I (TβRI) and type II (TβRII) receptors. TGF-β binding to TβRII induces recruitment and phosphorylation of TβRI, which in turn transmits the signal through phosphorylation of the receptor-bound R-Smad transcription factors, Smad2 or Smad3. Once phosphorylated, the R-Smads hetero-dimerize with their common partner, Smad4. The resulting complexes then migrate to the nucleus, where they regulate the transcription of TGF-β-target genes in conjunction with other transcription factors [2].

Nuclear Transcriptional Intermediary Factor 1 γ, TIF1γ (also known as tripartite motif protein TRIM33), is a member of the transcriptional intermediary factor 1 family [3] and was recently identified as a new partner of the Smad-dependent TGF-β signaling pathway. A screen for molecules involved in the specification of the embryonic endoderm first revealed TIF1γ as a Smad4-binding protein and as a negative regulator of TGF-β signaling [4]. TIF1γ mono-ubiquitinates Smad4, inducing its nuclear export to the cytoplasm, where the FAM/UPS9x deubiquitinating enzyme was recently shown to allow Smad4 recycling [5]. The role of TIF1γ as a repressor was also reported in the control of Smad activity during embryogenesis [6]. In contrast, TIF1γ was identified as a protein partner for receptor-activated Smad2/3, resulting in an alternative positive regulatory Smad4-independent TGF-β signaling pathway [7].

Whether TIF1γ down-regulates or promotes alternative TGF-β signaling may be linked to the cellular context. TIF1γ is a ubiquitous protein and its mRNA has been detected in all tissues [8]. Its loss of expression has been shown to favor $Kras^{G12D}$-dependent precancerous pancreatic lesions [9], induce cell-autonomous myeloproliferative disorders in mice [10] and potentiate TIF1α-induced murine hepatocellular carcinoma [11], thereby supporting a protective role of TIF1γ in cancer. Consistent with this view, a decrease in TIF1γ expression in human pancreatic cancer and human chronic myelomonocytic leukemia has been reported [9,11] and TIF1γ silencing in human mammary epithelial cell lines was shown to lead to a strong epithelial-mesenchymal transition mediated by TGF-β1 [12]. In contrast, a pro-tumorigenesis role for TIF1γ has been suggested by the observation that its expression prevents Smad4-mediated growth inhibition in response to TGF-β [4]. In line with the uncertain role of TGF-β in cancer, TIF1γ may differentially affect TGF-β signaling according to the cellular context by acting either as tumor suppressor or promoter.

Several mathematical models have been developed to predict the dynamic behavior of TGF-β signaling. In particular, initial differential models that couple signaling with receptor trafficking have significantly improved our understanding of the plasticity of the TGF-β signaling pathway [13]. Models focusing on Smad phosphorylation [14], Smad nucleocytoplasmic shuttling [15,16]

and Smad oligodimerization [17] have also been developed to understand the dynamics and flexibility of Smad-dependent pathways, while integrative models have coupled receptor trafficking to Smad pathways [18–20]. As the latter models recapitulate the essential components of the canonical Smad-dependent TGF-β signaling pathway, they constitute useful tools to investigate the role of new regulatory components of TGF-β signaling.

We have used an integrative modeling approach to explore the impact of TIF1γ on the outcome of TGF-β signaling. Taking advantage of mathematical models of receptor trafficking [13] and Smad shuttling [16], we have developed a new TGF-β signaling model that includes TIF1γ and FAM/UPS9x. Our model, which is based on the transient formation of a ternary complex containing TIF1γ Smad4 and Smad2/3, successfully reconciles the different observations reported for TIF1γ-Smad4 [4] and TIF1γ-Smad2/3 [7] interactions. We show that TGF-β signaling is highly sensitive to the TIF1γ/Smad4 ratio, suggesting a critical role for the FAM/UPS9x deubiquitinase. This model also predicts how varying TIF1γ/Smad4 ratios can modulate the cellular response to transient and sustained TGF-β stimulation, accounting for a highly graded TGF-β response. We discuss how the seemingly opposite roles of TIF1γ may be resolved by taking into account the dynamic balance of interactions involving Smad4 and Smad2/3.

Materials and Methods

Mathematical modeling

The model consists of a system of nonlinear, ordinary differential equations that merge the ODE models of receptor trafficking [13] and Smad shuttling [16]. Briefly, the receptors described in the Smad shuttling model were replaced by those of the receptor trafficking model using unit conversion in a cell volume of 2.27×10^{-12}L. Model building, parameters, system ordinary equations and description of the model in Systems Biology Markup Language (SBML) are detailed in Tables S1 and S2 and Model S1. Model simulations were implemented with the mathematical Scipy library of Python language programming and the Matplotlib Python 2D plotting library was used to visualize the simulation curves.

Cell culture and siRNA transfection

Human mammary epithelial (HMEC) cells infected with a retrovirus carrying hTERT and the oncogenic H-RasV12 (HMEC-TR) allele were provided by R. A. Weinberg [21] and cultured as previously described [12]. Cells were transfected with 5 nM siRNA and 0.5 µl/ml lipofectamine RNAiMax (Invitrogen) and further cultured in the presence or absence of 10 ng/ml TGF-β1 (Peprotech) for the indicated times.

Chromatin immunoprecipitation (ChIP)

Assays were carried out on cells transfected with the PAI-1 p800-Luc construct, as previously described [12], using the kit from Upstate Biotechnology. Briefly, cell lysates were subjected to anti-Smad4 (SantaCruz) or anti-TIF1γ (Bethyl) immunoprecipitation. Smad4- or TIF1γ-precipitated genomic DNA was subjected to PCR. The 351-bp PAI-1 promoter region harboring the Smad-binding elements was amplified with primers 5′-AGCCAGACAAGGTTGTTG-3′ and 5′-GACCACCTCCAGGAAAG-3′. An unrelated genomic DNA sequence (actin) was amplified with primers 5′-AGCCATGTACGTTGCTATCCAG-3′ and 5′-CTTCTCCTTAATGTCACGCACG-3′.

Relative quantification of mRNA by real-time PCR

Real-time quantitative PCR was performed using the qPCRTM Core Kit for SybrTM Green I from Eurogentec and the ABI Prism 7700 thermocycler (Perkin-Elmer, Foster city, CA, USA). Primer pairs for target genes were: sense CDH11 (OB-Cadherin), 5′CCC TGA AAT CAT TCA CAA TCC3′, antisense 5′AGT CCT GCT TCT GCC GAC T3′; CDH2 (N-Cadherin), sense: 5′GTG CAT GAA GGA CAG CCT CT3′, antisense: 5′ATG CCA TCT TCA TCC ACC TT3′; HPRT, sense: 5′TGA CCT TGA TTT ATT TTG CAT ACC3′, antisense: 5′CGA GCA AGA CGT TCA GTC CT3′.

Western blot analysis

Cell lysates were subjected to SDS-polyacrylamide gel electrophoresis and transferred onto PVDF membranes. The blots were incubated for 1 hr in Tris-buffered saline containing 0.1% Tween 20 and 5% non-fat dry milk and further incubated for 1 hr with specific primary antibodies (anti-Smad4, SantaCruz biotechnology; anti-TIF1γ, Euromedex). The bound antibodies were visualized with horseradish peroxidase–conjugated antibodies using the ECL-Plus reagent (Roche).

Results and Discussion

Quantitative models for TIF1γ-dependent TGF-β signaling

Merging receptor trafficking [13] and Smad cytonucleoplasmic shuttling [16] models through their common receptor-ligand complex in the endosome (LRe), we developed new models that integrate TIF1γ. Kinetic parameters were estimated according to the experimental data from [5] and [7] and are detailed in Table S1. We first constructed two separate models, each taking into account the different hypotheses regarding Smad/TIF1γ interactions. The first model is based on the TIF1γ-dependent negative regulation associated with the ubiquitination of Smad4 ([4,5]; Figure 1A). In this model, TIF1γ interacts preferentially with Smad4 within phosphorylated Smad2-Smad4 complexes in response to TGF-β, leading to a rapid dissociation of complexes and formation of ubiquitined Smad4 (Smad4ub) that is exported from the nucleus. Similar to the transient interaction of the phosphatase (PPase) with phosphorylated Smad2 [15,16], the formation of TIF1γ-Smad complexes was neglected because of fast reaction rates. In the cytoplasm, ubiquitinated Smad4 undergoes deubiquitination by FAM/UPS9x (FAM), thereby recycling Smad4 for TGF-β signaling (Figure 1A). We set the same kinetic parameters for association between TIF1γ and phosphorylated Smad2-Smad4 complexes in the nucleus (pS2S4n) and association between phosphorylated Smad2 and Smad4. Ubiquitination/deubiquitination and phosphorylation/dephosphorylation kinetics were considered to be similar, as previously described [5]. Export of ubiquitinated Smad4 from the nucleus to the cytoplasm was assumed to be 2-fold higher than entry of Smad4 in the nucleus, based on the observation suggesting that ubiquitined Smad4 is less efficiently retained in the nucleus [4,5].

Our second model is based on results from He *et al.* [7], who proposed that TGF-β induces a competing interaction between TIF1γ and phosphorylated Smad2, although an association of TIF1γ with Smad4 was also detected in the nucleus (Figure 1B). In the absence of conclusive experimental data, we considered the kinetic parameters for association between TIF1γ and either phosphorylated Smad2 or Smad4 in the nucleus to be similar to those for phosphorylated Smad2 with Smad4. To test this hypothesis, we analyzed the effect of a 2-fold decrease in k_{on}/k_{off}

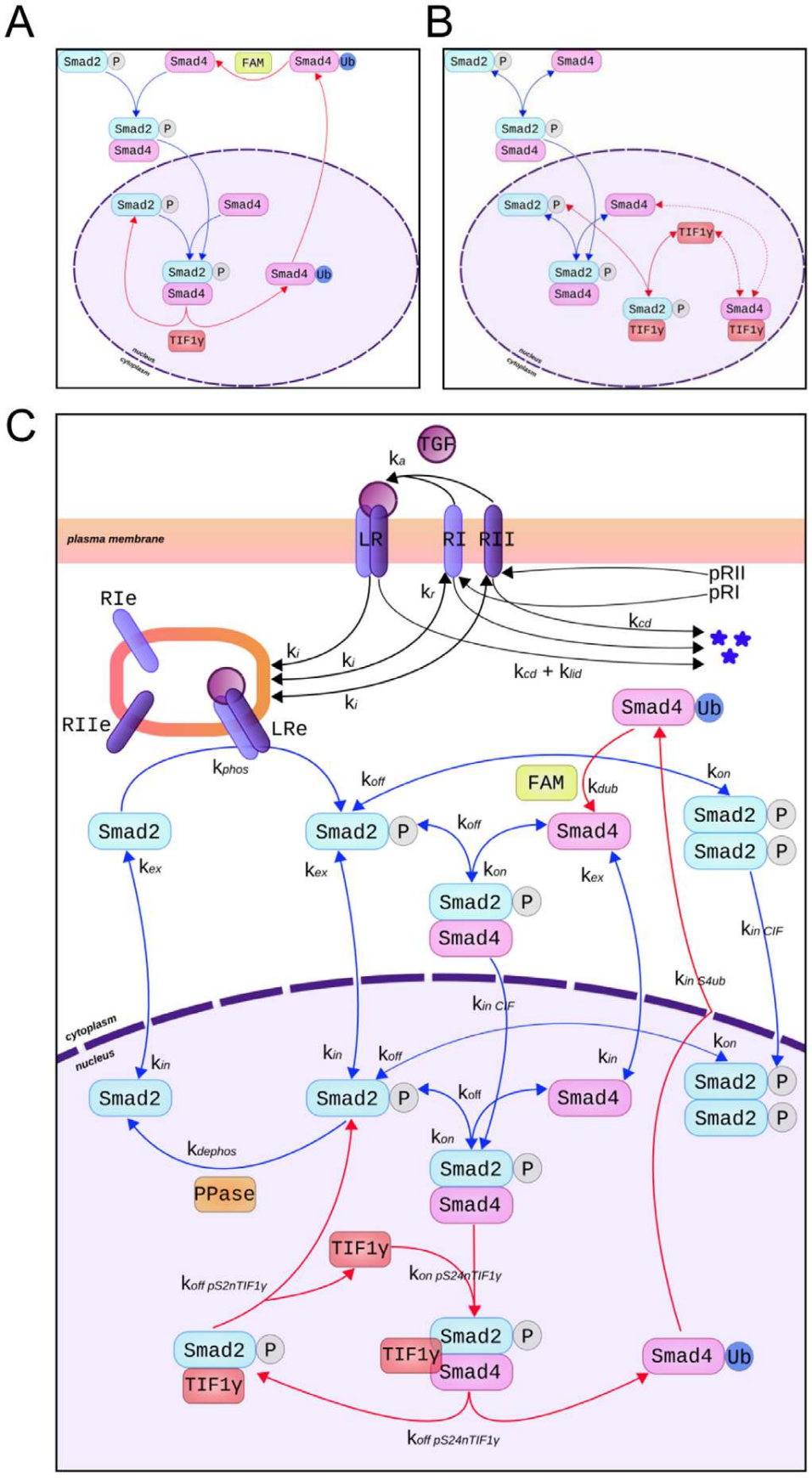

A
Smad2 P
Smad4
FAM
Smad4 Ub
Smad2 P
Smad4
Smad2 P
Smad4
Smad2 P
Smad4 Ub
Smad4
TIF1γ
nucleus
cytoplasm
B
Smad2 P
Smad4
Smad2 P
Smad4
Smad2 P
Smad4
TIF1γ
Smad2 P
Smad4
Smad2 P
TIF1γ
Smad4
TIF1γ
nucleus
cytoplasm
C
TGF
ka
plasma membrane
LR
RI
RII
pRII
pRI
kr
kcd
ki
ki
kcd + klid
ki
RIe
RIIe
LRe
kphos
Smad4 Ub
FAM
kdub
koff
kon
Smad2
Smad2 P
koff
Smad4
Smad2 P
Smad2 P
kex
kex
kon
kex
Smad2 P
Smad4
kin CIF
kin S4ub
kin CIF
cytoplasm
nucleus
kin
kin
koff
kin
kon
Smad2
Smad2 P
koff
Smad4
Smad2 P
Smad2 P
kon
kdephos
Smad2 P
Smad4
PPase
TIF1γ
koff pS2nTIF1γ
kon pS24nTIF1γ
Smad2 P
TIF1γ
TIF1γ
Smad2 P
Smad4
Smad4 Ub
koff pS24nTIF1γ

Figure 1. Schematic representation of the models. Detailed information on parameters and entities are given in Tables S1 and S2. A) Model hypothesis from [4]. B) Model hypothesis from [7]. C) Integrated model including TIF1γ (red rectangle) and FAM (green rectangle).
doi:10.1371/journal.pone.0033761.g001

for the association between TIF1γ and Smad2 or Smad4, which did not modify the TGF-β response in simulation studies.

Finally, we integrated the TIF1γ and FAM/UPS9x modulators into a unique model that merges all experimental observations (Figure 1C). Unlike the model depicted in Figure 1A, we considered TIF1γ binding to Smad4 as part of a ternary complex, in which phosphorylated Smad2, Smad4 and TIF1γ are associated in the nucleus (pS24nTIF1γ). In this case, note that the interaction of TIF1γ with Smad2 occurs within phosphorylated Smad2-TIF1γ (pS2nTIF1γ) complexes that are generated by dissociation of the ternary complexes in the nucleus. We set the same kinetic parameters for the formation/dissociation of the ternary pS24nTIF1γ complexes and the formation/dissociation of the phosphorylated Smad2-Smad4 complexes.

Model analysis and simulation

We next performed computational experiments to investigate the dynamics of TGF-β signaling according to each model. TGF-β signaling was expressed as the amount of phosphorylated Smad2-Smad4 complexes in the nucleus (pS24n) because TGF-β target genes are regulated by these heterodimeric complexes. To explore the functional effect of TIF1γ on the TGF-β transcriptional signal, simulation studies were performed using different concentrations of TIF1γ varying from 0 to 50 nM, the latter corresponding to the initial concentration of Smad4 (Figure 2, Table S1). These prediction studies showed that each model was either too sensitive, with total inhibition of signaling at low concentrations of TIF1γ according to the first model (Figure 2A), or too insensitive, with only a slight variation of signaling at higher TIF1γ concentrations according to the second model (Figure 2B). Each predictive model hence yielded a significant mismatch with the experimental data derived from the other. The strict negative regulatory role of TIF1γ proposed by Dupont *et al.* [4] is not compatible with the lack of sensitivity of the second model adapted from He *et al.* [7]. Similarly, He *et al.* observed a moderate TIF1γ effect on TGF-β transcriptional activity that did not agree with the high sensitivity of the first model adapted from Dupont *et al.* In contrast, our integrative model that includes all observations yielded a graded effect of TIF1γ on pS24n complex formation that is in agreement with the relative abundance of TIF1γ-Smad complexes reported in both studies, leading to a graded regulation of TGF-β signaling (Figure 2C).

To further explore the robustness of our integrative model, we evaluated the sensitivity of TGF-β signaling to variations in kinetic parameters. As shown in Figure 3, varying the rate of formation (Figure 3A) or dissociation (Figure 3B) of complexes containing TIF1γ and pS24n had little effect on TGF-β signaling. Similarly, varying the kinetic parameters for the dissociation of phosphorylated Smad2-TIF1γ complexes (pS2nTIF1γ) induced only few changes in the concentration of pS24n (Figure 3C). In contrast, TGF-β signaling was highly sensitive to the variation of k_{in}-Smad4ub (Figure 3D), suggesting that the export rate of ubiquitinated Smad4 is a critical component of the regulation of TGF-β transcriptional activity. In addition, the slight alteration in TGF-β signaling induced by changes in the deubiquitination rate of Smad4 (Figure 3E) disappeared with increasing concentrations of the FAM deubiquitinase (Figure 3F), suggesting that changes in FAM expression might be a sensitive marker to predict modulation of TGF-β signaling. Taken together, the results of our simulation studies reveal a new pivotal role of the Smad4 ubiquitination/

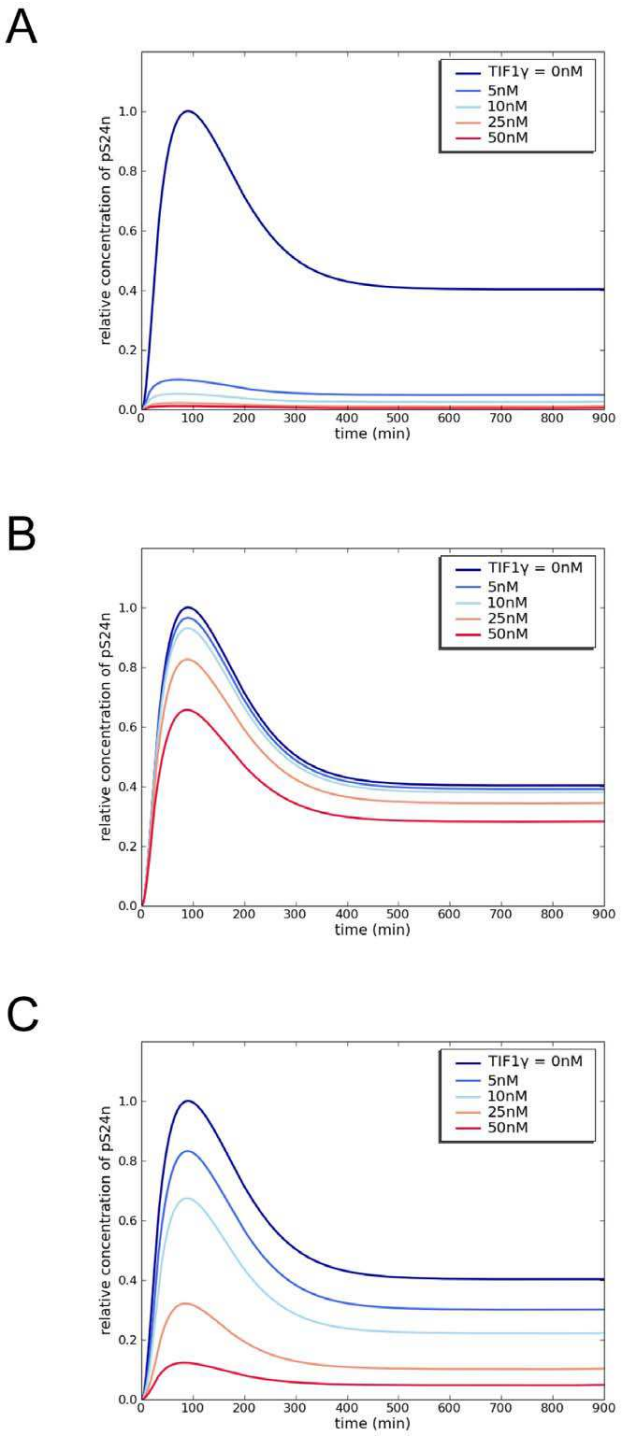


Figure 2. Effect of TIF1γ on TGF-β signaling. Modeling analysis of the pS24n response to increasing TIF1γ concentrations at a 10 nM TGF-β input. A) Model according to [4]; B) model according to [7] and C) integrated model.
doi:10.1371/journal.pone.0033761.g002

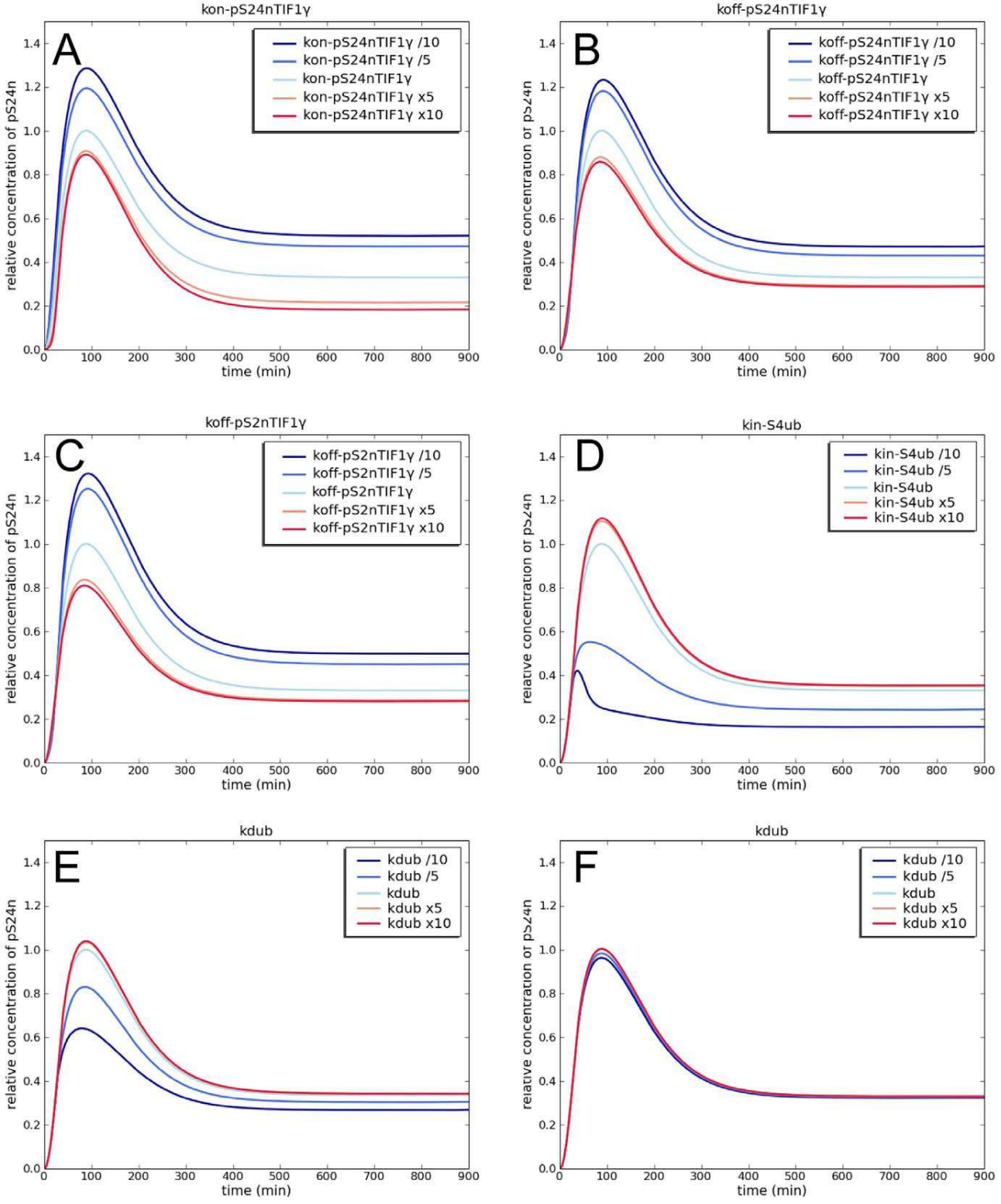


Figure 3. Parameter sensitivity analysis. Modeling analysis of pS24n response to variations of kinetic constants at a 10 nM TGF-β input. A) k_{on}-pS24nTIF1γ, binding of TIF1γ to phosphorylated-Smad2/Smad4 complexes; B) k_{off}-pS24nTIF1γ, dissociation of phosphorylated-Smad2/Smad4/TIF1γ complexes in the nucleus; C) k_{off}-pS2nTIF1γ, dissociation of phosphorylated Smad2-TIF1γ complexes in the nucleus; D) k_{in}-S4ub, nuclear export of ubiquitinated Smad4 in the cytoplasm; E) and F) k_{dub}, deubiquitination of Smad4 according to relative FAM concentrations of 1 nM (E) and 10 nM (F).
doi:10.1371/journal.pone.0033761.g003

deubiquitination cycle in the regulation of the dynamics of TGF-β signaling. Of note is the predicted critical regulatory role of FAM in TGF-β signaling through Smad4 recycling.

Experimental validation of the model

A key component of our model is based on the hypothesis that a transient ternary complex is formed, associating Smad4, TIF1γ and Smad2. To investigate the reality of such an interaction, we performed chromatin immunoprecipation (ChIP) assays as previously described [12]. As shown in Figure 4, stimulation of cells with TGF-β induced the recruitment of Smad proteins on the promoter sequence of PAI-1, a TGF-β target gene. In the absence of TGF-β stimulation, TIF1γ showed a significant association with DNA while Smad2/3 was not detected. A faint Smad4 signal could be detected under these conditions. TGF-β stimulation led to the detection of a strong Smad2/3 ChIP signal. Between 30 and 90 min of TGF-β stimulation, the association of all three proteins with DNA appears consistent with the hypothesis that a ternary complex containing Smad4, Smad2/3 and TIF1γ transiently forms. After 120 min, Smad4 dissociated from DNA whereas Smad2/3 and TIF1γ remained present on the PAI-1 promoter. This observation is in agreement with our hypothesis that Smad2-TIF1γ complexes are released from the ternary complexes. Importantly, Dupont *et al.* [4], using a double-immunoprecipitation approach for TIF1γ and Smad4, previously reported formation of these ternary complexes. More recently TIF1γ was shown to be present at the promoter region of PAI-1 gene in uninduced cells, whereas an increase in TIF1γ association with the Smad-binding region of the promoter was also observed upon TGF-β stimulation [22].

We next devised an experimental approach that could be used to evaluate TGF-β transcriptional activity as a function of variable TIF1γ/Smad4 ratios. Cells were transiently transfected with siRNAs to silence Smad4 or TIF1γ expression and were further stimulated or not with TGF-β for the indicated times (Figure 5A).

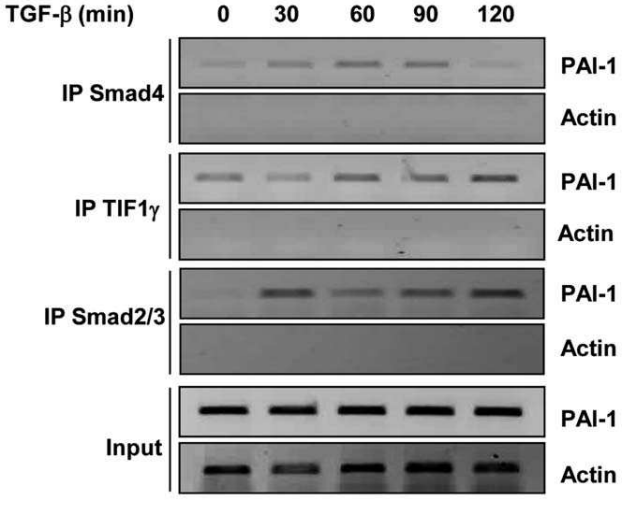


Figure 4. TIF1γ, Smad2 and Smad4 bind to the PAI-1 promoter. ChIP assays were performed on HMEC cells treated with TGF-β for the indicated times. Cell lysates were subjected to anti-Smad4 (IP Smad4), or anti-TIF1γ (IP TIF1γ), or anti-Smad2/3 (IP Smad2/3) chromatin immunoprecipitation. PCR amplification of the endogenous PAI-1 promoter (733/484) was performed to detect protein bound DNA. Primers specific to actin were used as controls.
doi:10.1371/journal.pone.0033761.g004

The expression of Smad4 and TIF1γ was efficiently inhibited since no proteins were detected at day 3 post-transfection compared with cell transfected with non-targeted siRNAs (scr). The efficacy of RNA interference was confirmed at the mRNA level (Figure S1). This effect decreased with time according to siRNA availability and mRNA turnover, leading to the recovery of protein basal levels after several days (Figure 5A, upper panel). Note that silencing Smad4 and TIF1γ affected the amounts of TIF1γ and Smad4 proteins, respectively, detected at day 3. The time courses shown in Figure 5 finally allowed us to analyze cells containing variable amounts of endogenous Smad4 and TIF1γ proteins. For each time point, cell extracts were used for western blot analyses and TIF1γ/Smad ratios were evaluated by densitometric scanning of blots (Figure 5A, bottom panel). To perform this experimental verification, we quantified the mRNA levels of endogenous TGF-β target genes instead of using the over-expression of reporter genes to estimate transcriptional activities. We selected the CDH2 and CDH11 cadherin genes as they are up-regulated by TGF-β through Smad4- and TIF1γ-dependent pathways in our cell model (Figure S2). Using the same cell extracts used for western blotting (Figure 5A), the mRNA levels of CDH2 and CDH11 were quantified and TGF-β transcriptional activity was evaluated as the ratio of mRNA levels observed in the presence or absence of TGF-β (Figure 5B). TGF-β-induced expression of CDH2 and CDH11 was correlated with the amount of Smad4 and TIF1γ proteins. Compared to control cells (scr), low Smad4 expression (Day3) prevented TGF-β-dependent expression of CDH2 and CDH11 while the absence of TIF1γ led to up-regulation of CDH2 and CDH11.

We then compared these experimental data with results predicted by our integrative model. As shown in Figure 5C, our observations could be fitted to the simulation curves of TGF-β transcriptional signaling, a validation reinforced by the use of physiological parameters. We conclude from these results that TIF1γ is a new regulator that plays a pivotal role in the control of Smad4-dependent TGF-β transcriptional activity. These data also show that TIF1γ/Smad4 ratios can determine TGF-β-dependent transcriptional activity. Accordingly, our model supports the hypothesis of fast binding of TIF1γ to phosphorylated Smad2/Smad4 complexes and the release of both ubiquitinated Smad4 and phosphorylated Smad2-TIF1γ complexes.

TGF-β dose- and time-dependent responses

The concentration of TGF-β in the cellular microenvironment is highly variable and its increased expression has been reported in numerous pathologies, including inflammation, fibrosis and cancer [23]. However the determination of TGF-β concentrations at the cellular level within tissues remains a difficult task since TGF-β is stocked as a latent form in the extracellular matrix [24]. In addition, its conversion from latent to biologically active forms involves numerous protease- and non protease-dependent mechanisms that differ according to cell type and the physiological context, leading to a complex non-linear delivery [25]. All previous mathematical models are based on biological data obtained from *in vitro* experiments using either TGF-β concentrations (in the nM range) or on/off signal inputs. However, Zi and al. [19] recently developed an integrative model that includes a ligand depletion parameter and demonstrated that cell-fate decision in response to TGF-β stimulation depends not only on its concentration but also on the time course of its delivery. Because we did not integrate ligand depletion in our model, response predictions were insensitive to TGF-β concentration except for concentrations as low as 0.1 nM (Figure 6A) and we routinely used concentrations of 10 nM as the TGF-β input.

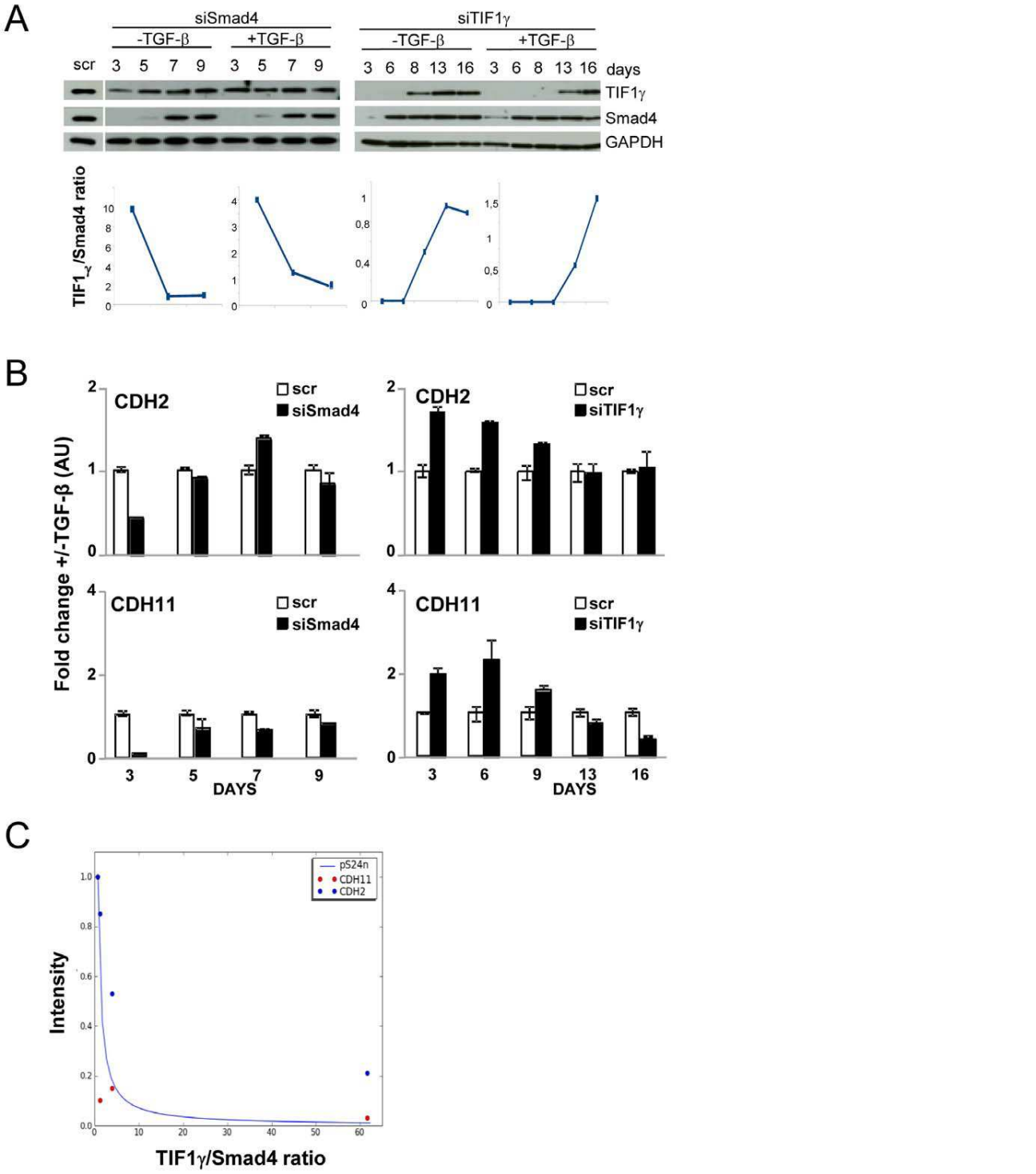


Figure 5. Expression of the CDH2 and CDH11 TGF-β target genes is sensitive to TIF1γ/Smad4 ratios. HMEC cells were transfected with Smad4 (siSmad4) or TIF1γ (siTIF1γ) siRNAs and cultured in the presence (+) or absence (−) of TGF-β for the indicated times (days). Controls were cells transfected with non-targeted siRNA (scr). A) Smad4 and TIF1γ protein levels were analyzed by immunoblotting (upper panels) and quantified by densitometric scanning (lower panels). B) TGF-β-induced fold changes in CDH2 and CDH11 expression were analyzed by RT–qPCR. All values were normalized to the amount of HPRT mRNA and expressed relative to the value obtained for TGF-β-untreated controls in arbitrary units (AU). Results are expressed as the mean ± SD of 3 independent experiments. C) mRNA levels of CDH2 (red circles) and CDH11(blue circles) were plotted against TIF1γ/Smad4 ratios and were fitted to the predictive equation curve of pS24n relative concentrations.
doi:10.1371/journal.pone.0033761.g005

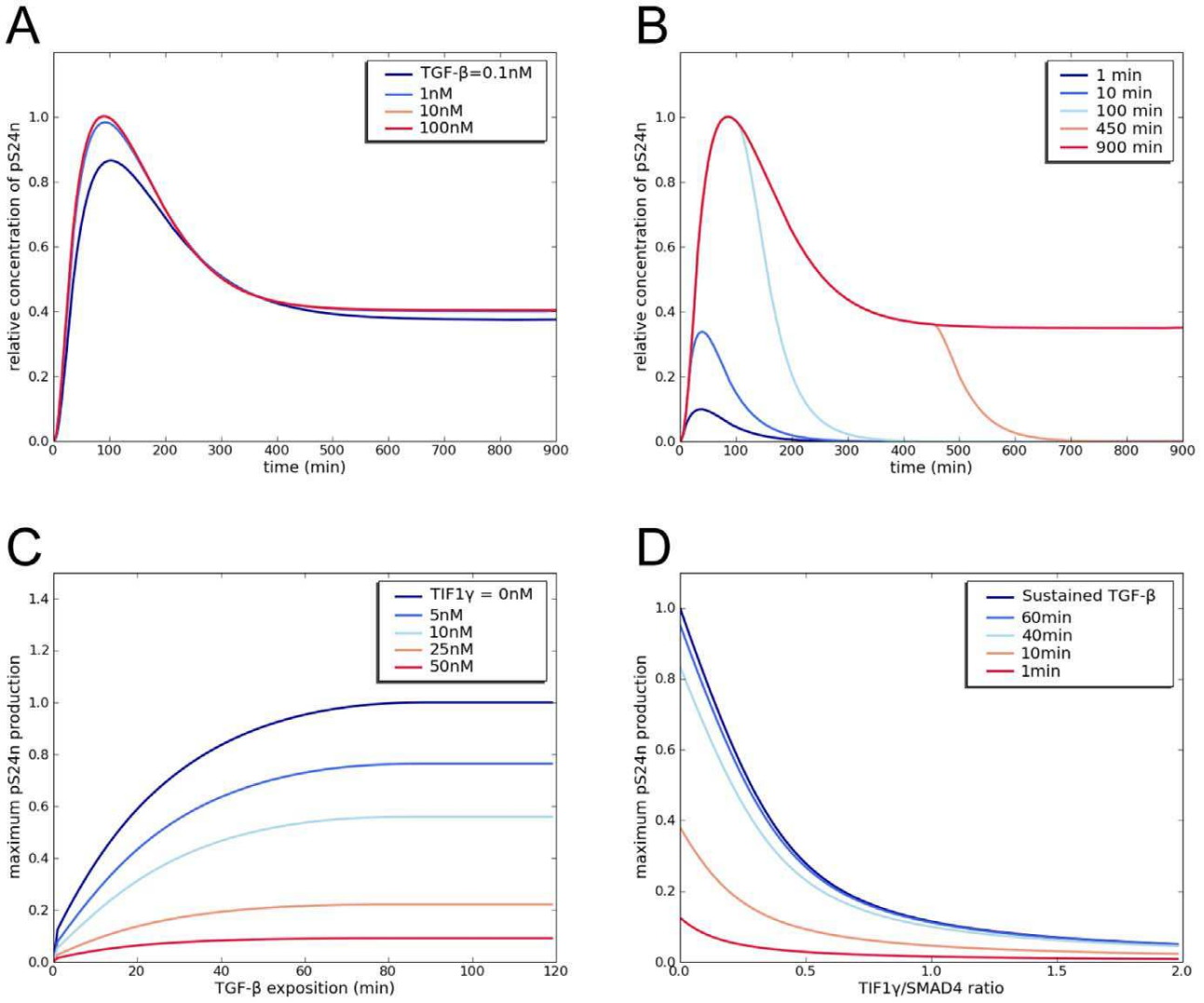


Figure 6. Concentration and time dependence of TGF-β signaling. A) and B) Modeling analysis of the pS24n response to increasing concentrations of TGF-β (A) and duration of stimulation with 10 nM TGF-β (B). C) and D) Modeling analysis of the maximum pS24n response as a function of TGF-β duration of exposure (C) or increasing TIF1γ/Smad4 ratios (D).
doi:10.1371/journal.pone.0033761.g006

When TGF-β depletion was included in our model, both graded short-term and switch-like long-term responses to TGF-β were conserved as reported by Zi *et al.* [19]. However, they were attenuated, suggesting that the presence of TIF1γ does not affect the signal shape, but only the amplitude of TGF-β signaling (Figure S3).

In contrast, we observed that, in our model, the length of stimulation modified the cell response. This was particularly true for short times (Figure 6B), maximum pS24n complex formation being highly dependent on TIF1γ concentration (Figure 6C and 6D). This indicates that the magnitude of the cellular response to TGF-β depends on both TIF1γ/Smad4 ratios and time-dependent stimulation, predicting a broad range of responses according to TGF-β cellular content and availability in the microenvironment (Figure 7A). Note that the alternative pS2TIF1γ transcription complexes proposed by He *et al.* [7] displayed an opposite profile that required high TIF1γ/Smad4 ratios and longer stimulation times to be fully active (Figure 7B).

In agreement with Zi *et al.* [19], our model showed that periodic short pulses of ligand stimulation yielded an outcome similar to that produced by sustained ligand simulation, whereas an increase in the duration between pulses prevented a continuous response. These observations support the memory concept of ligand-receptor complex (LCR) activity (Figure S4A). When TIF1γ was added, the shape of the response was similar, albeit attenuated, suggesting that, in our model, TIF1γ does not affect LCR recycling (Figure S4B).

Conclusions

Taking into account the seemingly contradictory observations of Smad4-TIF1γ and Smad2/3-TIF1γ interactions, we propose an integrative model based on the formation of Smad2-Smad4-TIF1γ ternary complexes. Validation of our hypotheses by *a posteriori* biological experiments provides strong support for our model, which shows that the TIF1γ/Smad4 ratio serves as a regulator of TGF-β signaling that may affect determination of cell

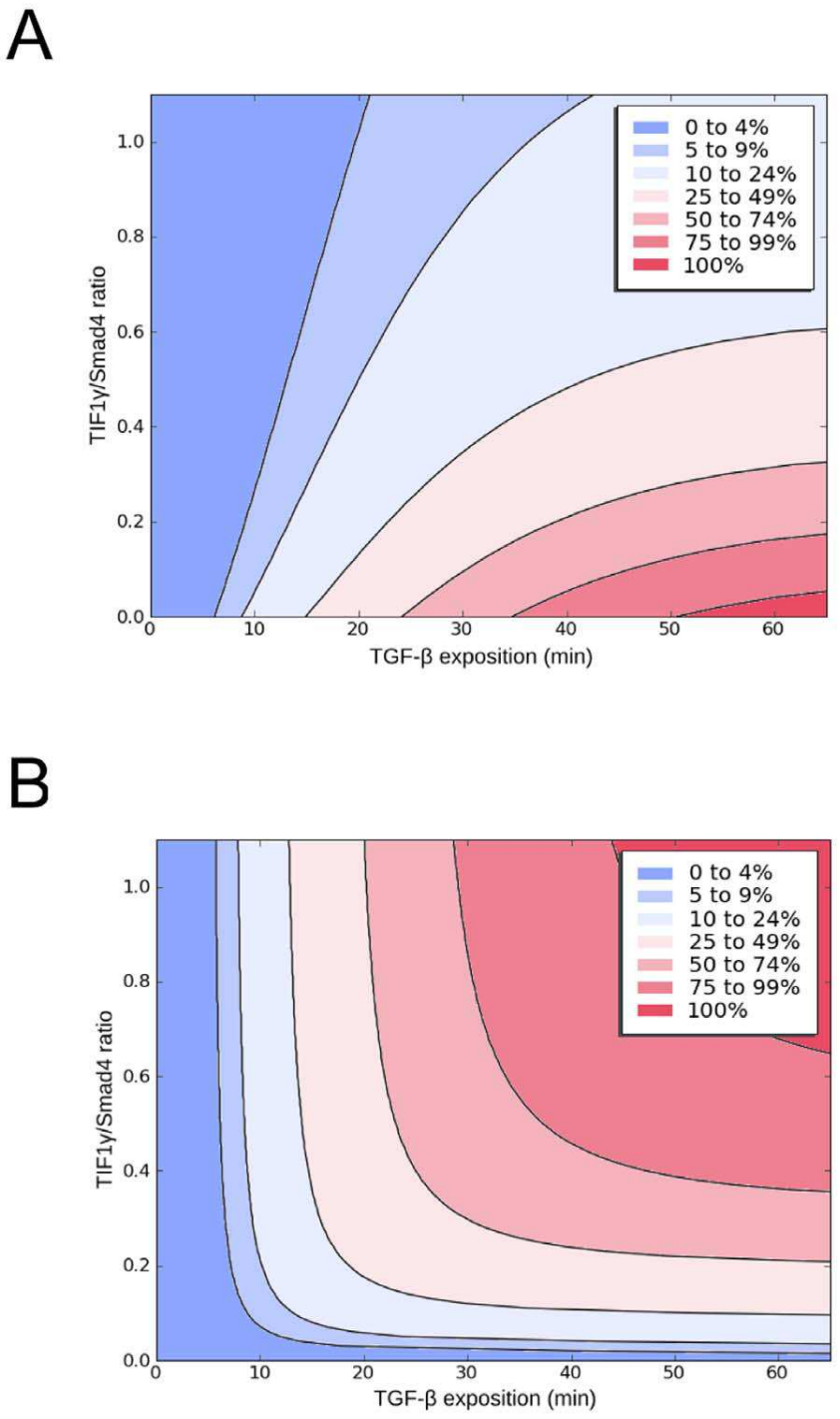


Figure 7. TGF-β time-dependent pS24n and pS2nTIF1γ response profiles as a function of TIF1γ/Smad4 ratios. Results are expressed as percentage of the maximum production of pS24n (A) or pS2nTIF1γ (B).
doi:10.1371/journal.pone.0033761.g007

fate. We demonstrate that the response to TGF-β signaling is highly sensitive to TIF1γ/Smad4 ratios, especially for short stimulation times that mediate higher threshold responses. A critical role for the TIF1γ/Smad4 ratio in the regulation of TGF-β signaling is supported by the antagonistic role of TIF1γ and Smad4 in the epithelio-mesenchymal cell transition [12], embryonic patterning and trophoblast stem-cell differentiation [6], suggesting that TIF1γ acts as a negative regulator of higher TGF-β threshold responses.

Our results emphasize the significance of TIF1γ in orchestrating the pleiotropic effects of TGF-β signaling according to the cellular context. Its sensitivity to Smad4 levels and stimulation times suggests that TIF1γ helps define a broad landscape of TGF-β responses. We note that Agricola *et al.* recently proposed a new

model for TIF1γ ubiquitin ligase activity that requires binding to histones [22], thus implicating chromatin dynamics in the control of Smad localization at the promoter of TGF-β target genes. According to these results, epigenetic events contribute to the transcriptional regulation of TGF-β target genes *via* acetylation and methylation processes [26–28]. In order to understand the complexity of TGF-β-dependent gene regulation and to predict cellular responses, we believe that future models will need to integrate not only the Smad canonical pathway but also Smad-independent pathways and epigenetic events. Because of the lack of quantitative data, such an ambitious goal will require the development of different modeling-based approaches that utilize discrete models [29,30].

Supporting Information

Figure S1 Effects of Smad4 and TIF1γ knockdown on gene expression. HMEC cells were transfected with Smad4 (siSmad4) or TIF1γ (siTIF1γ) siRNAs and cultured in the presence (+) or absence (−) of TGF-β for the indicated times (days). Controls were cells transfected with non-targeted siRNA (Scr). Smad4 and TIF1γ gene expression was quantified by RT-qPCR. All values were normalized to the amount of HPRT mRNA and expressed in arbitrary units (AU). Results are expressed as the mean+SD of 3 independent experiments.
(PDF)

Figure S2 Expression of the CDH2 and CDH11 is induced by TGF-β through TIF1γ- and Smad4-dependent pathways. HMEC cells were transfected with Smad4 (siSmad4) or TIF1γ siRNAs (si TIF1γ) and cultured in the presence (+) or absence (−) of TGF-β for 2 days. Control cells transfected with non-targeted siRNA (Scr). CH2 and CDH11 gene expression was quantified by RT-qPCR. Results are normalized to the amount of mRNA in untreated cells and expressed as the mean+SD of 3 independent experiments.
(PDF)

Figure S3 TIF1γ does not affect short-term and switch-like long-term responses to TGF-β. TGF-β depletion was added to the integrated model and modeling analysis of the pS24n response was performed using either increasing concentrations of TGF-β (A) or increasing concentrations of TIF1γ in the presence of 1 nM (B) 5 nM (C) and 10 nM TGF-β (D).
(PDF)

Figure S4 TIF1γ does not modify the pS24n response to a pulsed exposure to TGF-β. Model prediction of the pS24n response in the absence (A) or presence (B) of 10 nM TIF1γ to sustained TGF-β (10 nM) stimulation (blue curve), continuous short pulses at 30-minute intervals (green curve) or 3-hour intervals (red curve), as previously described experimentally (Zi et al 2011). We use 10 nM TIF1γ as an average dose of tested concentrations. Concentrations up to 50 nM TIF1γ did not modify the behavior of the signal but only reduced the signal range.
(PDF)

Table S1 System parameters.
(PDF)

Table S2 System of ordinary differential equations. Equations in black are from Vilar *et al.*, 2006 and Schmierer *et al.*, 2008; equations in red are estimated from biological experiments from Dupont *et al.*, 2005, 09 and He *et al.*, 2006.
(PDF)

Model S1 Description of the model in Systems Biology Markup Language (SBML).
(PDF)

Acknowledgments

The authors would like to thank Dr. E.mmanuel Käs (LBME, CNRS/ Université Paul Sabatier) for help in writing this manuscript.

Author Contributions

Conceived and designed the experiments: GA MLB RR NT. Performed the experiments: GA LF. Analyzed the data: GA MLB RR NT. Contributed reagents/materials/analysis tools: MLB RR NT. Wrote the paper: NT.

References

1. Massague J (2008) TGFbeta in Cancer. Cell 134: 215–230.
2. Schmierer B, Hill CS (2007) TGFbeta-Smad signal transduction: molecular specificity and functional flexibility. Nat Rev Mol Cell Biol 8: 970–982.
3. Venturini L, You J, Stadler M, Galien R, Lallemand V, et al. (1999) TIF1gamma, a novel member of the transcriptional intermediary factor 1 family. Oncogene 18: 1209–1217.
4. Dupont S, Zacchigna L, Cordenonsi M, Soligo S, Adorno M, et al. (2005) Germ-layer specification and control of cell growth by Ectodermin, a Smad4 ubiquitin ligase. Cell 121: 87–99.
5. Dupont S, Mamidi A, Cordenonsi M, Montagner M, Zacchigna L, et al. (2009) FAM/USP9x, a deubiquitinating enzyme essential for TGFbeta signaling, controls Smad4 monoubiquitination. Cell 136: 123–135.
6. Morsut L, Yan KP, Enzo E, Aragona M, Soligo SM, et al. (2010) Negative control of Smad activity by ectodermin/TIF1gamma patterns the mammalian embryo. Development 137: 2571–2578.
7. He W, Dorn DC, Erdjument-Bromage H, Tempst P, Moore MA, et al. (2006) Hematopoiesis controlled by distinct TIF1gamma and Smad4 branches of the TGFbeta pathway. Cell 125: 929–941.
8. Yan KP, Dolle P, Mark M, Lerouge T, Wendling O, et al. (2004) Molecular cloning, genomic structure, and expression analysis of the mouse transcriptional intermediary factor 1 gamma gene. Gene 334: 3–13.
9. Vincent DF, Yan KP, Treilleux I, Gay F, Arfi V, et al. (2009) Inactivation of TIF1gamma cooperates with Kras to induce cystic tumors of the pancreas. PLoS Genet 5: e1000575.
10. Aucagne R, Droin N, Paggetti J, Lagrange B, Largeot A, et al. (2011) Transcription intermediary factor 1gamma is a tumor suppressor in mouse and human chronic myelomonocytic leukemia. J Clin Invest 121: 2361–2370.
11. Herquel B, Ouararhni K, Khetchoumian K, Ignat M, Teletin M, et al. (2011) Transcription cofactors TRIM24, TRIM28, and TRIM33 associate to form regulatory complexes that suppress murine hepatocellular carcinoma. Proc Natl Acad Sci U S A 108: 8212–8217.
12. Hesling C, Fattet L, Teyre G, Jury D, Gonzalo P, et al. (2011) Antagonistic regulation of EMT by TIF1gamma and Smad4 in mammary epithelial cells. EMBO Rep 12: 665–672.
13. Vilar JM, Jansen R, Sander C (2006) Signal processing in the TGF-beta superfamily ligand-receptor network. PLoS Comput Biol 2: e3.
14. Clarke DC, Betterton MD, Liu X (2006) Systems theory of Smad signalling. Syst Biol (Stevenage) 153: 412–424.
15. Melke P, Jonsson H, Pardali E, ten Dijke P, Peterson C (2006) A rate equation approach to elucidate the kinetics and robustness of the TGF-beta pathway. Biophys J 91: 4368–4380.
16. Schmierer B, Tournier AL, Bates PA, Hill CS (2008) Mathematical modeling identifies Smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system. Proc Natl Acad Sci U S A 105: 6608–6613.
17. Nakabayashi J, Sasaki A (2009) A mathematical model of the stoichiometric control of Smad complex formation in TGF-beta signal transduction pathway. J Theor Biol 259: 389–403.
18. Chung SW, Miles FL, Sikes RA, Cooper CR, Farach-Carson MC, et al. (2009) Quantitative modeling and analysis of the transforming growth factor beta signaling pathway. Biophys J 96: 1733–1750.
19. Zi Z, Feng Z, Chapnick DA, Dahl M, Deng D, et al. (2011) Quantitative analysis of transient and sustained transforming growth factor-beta signaling dynamics. Mol Syst Biol 7: 492.
20. Zi Z, Klipp E (2007) Constraint-based modeling and kinetic analysis of the Smad dependent TGF-beta signaling pathway. PLoS One 2: e936.
21. Elenbaas B, Spirio L, Koerner F, Fleming MD, Zimonjic DB, et al. (2001) Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. Genes Dev 15: 50–65.

22. Agricola E, Randall RA, Gaarenstroom T, Dupont S, Hill CS (2011) Recruitment of TIF1gamma to chromatin via its PHD finger-bromodomain activates its ubiquitin ligase and transcriptional repressor activities. Mol Cell 43: 85–96.
23. Bierie B, Moses HL (2006) Tumour microenvironment: TGFbeta: the molecular Jekyll and Hyde of cancer. Nat Rev Cancer 6: 506–520.
24. Hyytiainen M, Penttinen C, Keski-Oja J (2004) Latent TGF-beta binding proteins: extracellular matrix association and roles in TGF-beta activation. Crit Rev Clin Lab Sci 41: 233–264.
25. Annes JP, Munger JS, Rifkin DB (2003) Making sense of latent TGFbeta activation. J Cell Sci 116: 217–224.
26. Bruna A, Darken RS, Rojo F, Ocana A, Penuelas S, et al. (2007) High TGFbeta-Smad activity confers poor prognosis in glioma patients and promotes cell proliferation depending on the methylation of the PDGF-B gene. Cancer Cell 11: 147–160.
27. Barter MJ, Pybus L, Litherland GJ, Rowan AD, Clark IM, et al. (2010) HDAC-mediated control of ERK- and PI3K-dependent TGF-beta-induced extracellular matrix-regulating genes. Matrix Biol 29: 602–612.
28. Hannigan A, Smith P, Kalna G, Lo Nigro C, Orange C, et al. (2010) Epigenetic downregulation of human disabled homolog 2 switches TGF-beta from a tumor suppressor to a tumor promoter. J Clin Invest 120: 2842–2857.
29. Assmann SM, Albert R (2009) Discrete dynamic modeling with asynchronous update, or how to model complex systems in the absence of quantitative information. Methods Mol Biol 553: 207–225.
30. Sreenath SN, Cho KH, Wellstead P (2008) Modelling the dynamics of signalling pathways. Essays Biochem 45: 1–28.

Supporting Information

Table S1 System parameters.

Table S2 System of ordinary differential equations. Equations in black are from Vilar *et al.*, 2006 and Schmierer *et al.*, 2008 ; equations in red are estimated from biological experiments from Dupont *et al.*, 2005, 09 and He *et al.*, 2006.

Figure S1 Effects of Smad4 and TIF1γ knockdown on gene expression. HMEC cells were transfected with Smad4 (siSmad4) or TIF1γ (siTIF1γ) siRNAs and cultured in the presence (+) or absence () of TGF-β for the indicated times (days). Controls were cells transfected with non-targeted siRNA (Scr). Smad4 and TIF1γ gene expression was quantified by RT-qPCR. All values were normalized to the amount of HPR mRNA and expressed in arbitrary units (AU). Results are expressed as the mean+SD of 3 independent experiments.

Figure S2 Expression of the CDH2 and CDH11 is induced by TGF-β through TIF1γ- and Smad4-dependent pathways. HMEC cells were transfected with Smad4 (siSmad4) or TIF1γ siRNAs (si TIF1γ and cultured in the presence (+) or absence () of TGF-β for 2 days. Control cells transfected with non-targeted siRNA (Scr). CH2 and CDH11 gene expression was quantified by RT-qPCR. Results are normalized to the amount of mRNA in untreated cells and expressed as the mean+SD of 3 independent experiments.

Figure S3 TIF1γ does not affect short-term and switch-like long-term responses to TGF-β. TGF-βdepletion was added to the integrated model and modeling analysis of the pS24n response was performed using either increasing concentrations of TGF-β (A) or increasing concentrations of TIF1γ in the presence of 1 nM (B) 5 nM (C) and 10 nM TGF-β (D).

Figure S4 TIF1γdoes not modify the pS24n response to a pulsed exposure to TGF-β. Model prediction of the pS24n response in the absence (A) or presence (B) of 10 nM TIF1γ to sustained TGF-β (10 nM) stimulation (blue curve), continuous short pulses at 30-minute intervals (green curve) or 3-hour intervals (red curve), as previously described experimentally (Zi *et al* 2011). We use 10 nM TIF1γ as an average dose of tested concentrations. Concentrations up to 50 nM TIF1γ did not modify the behavior of the signal but only reduced the signal range.

Table S1

Symbol	Definition	Value	Reference
S2c	cytoplasmic SMAD2	121.2nM	Schmierer et al, PNAS, 2008
S2n	nuclear SMAD2	57nM	Schmierer et al, PNAS, 2008
S4c	cytoplasmic SMAD4	50.8nM	Schmierer et al, PNAS, 2008
S4n	nuclear SMAD4	50.8nM	Schmierer et al, PNAS, 2008
pS2c	cytoplasmic phospho SMAD2	0nM	Schmierer et al, PNAS, 2008
pS2n	nuclear phospho SMAD2	0nM	Schmierer et al, PNAS, 2008
pS24c	cytoplasmic SMAD2 SMAD4 complex	0nM	Schmierer et al, PNAS, 2008
pS24n	nuclear SMAD2 SMAD4 complex	0nM	Schmierer et al, PNAS, 2008
pS22c	cytoplasmic SMAD2 SMAD2 complex	0nM	Schmierer et al, PNAS, 2008
pS22n	nuclear SMAD2 SMAD2 complex	0nM	Schmierer et al, PNAS, 2008
TGFβ	transforming growth factor beta	0 or 10nM	Schmierer et al, PNAS, 2008
PPase	Phosphatase	1nM	Schmierer et al, PNAS, 2008
RI	TGF type I receptor	3.66nM	Vilar et al, Plos Computational Biology, 2006
RII	TGF type II receptor	3.66nM	Vilar et al, Plos Computational Biology, 2006
LR	ligand receptor I receptor II complex	0nM	Vilar et al, Plos Computational Biology, 2006
RIe	endosomal TGF type I receptor	0nM	Vilar et al, Plos Computational Biology, 2006
RIIe	endosomal TGF type II receptor	0nM	Vilar et al, Plos Computational Biology, 2006
LRe	endosomal ligand receptor I receptor II complex	0nM	Vilar et al, Plos Computational Biology, 2006
TIF1γ	Transcriptional Intermediary Factor 1γ	from 0 to 50nM	Dupont et al, Cell, 2009
FAM	deubiquitinase	10nM	Dupont et al, Cell, 2009
pS24nTIF1γ	nuclear SMAD2 SMAD4 TIF1γ complex	0nM	Dupont et al, Cell, 2009 and He et al, Cell, 2006
pS2nTIF1γ	nuclear phospho SMAD2 TIF1γ complex	0nM	Dupont et al, Cell, 2009 and He et al, Cell, 2006
S4ub c	cytoplasmic ubiquitinate SMAD4	0nM	Dupont et al, Cell, 2009
S4ub n	nuclear ubiquitinate SMAD4	0nM	Dupont et al, Cell, 2009
k_{in}	import rate	$2.6*10^{-3}s^{-1}$	Schmierer et al, PNAS, 2008
k_{ex}	export rate	$5.6*10^{-2}s^{-1}$	Schmierer et al, PNAS, 2008
k_{phos}	phosphorylation rate	$4.04*10^{-4}nM^{-1}.s^{-1}$	Schmierer et al, PNAS, 2008
k_{dephos}	dephosphorylation rate	$7*10^{-3}nM^{-1}.s^{-1}$	Schmierer et al, PNAS, 2008
CIF	complex import factor	5.672 no unit	Schmierer et al, PNAS, 2008
k_{on}	Smad complex association rate	$2*10^{-3}nM^{-1}.s^{-1}$	Schmierer et al, PNAS, 2008
k_{off}	Smad complex separation rate	$1.6*10^{-2}s^{-1}$	Schmierer et al, PNAS, 2008
k_a	Ligand receptor association rate	$1nM^{-2}.s^{-1}$	Vilar et al, Plos Computational Biology, 2006
k_{cd}	constitutive degradation rate	$4.68*10^{-4}s^{-1}$	Vilar et al, Plos Computational Biology, 2006
k_{lid}	ligand induces degradation rate	$4.16*10^{-3}s^{-1}$	Vilar et al, Plos Computational Biology, 2006
k_i	internalization rate	$5.55*10^{-3}s^{-1}$	Vilar et al, Plos Computational Biology, 2006
k_r	recycling rate	$5.55*10^{-4}s^{-1}$	Vilar et al, Plos Computational Biology, 2006
pRI	receptors I production rate	$9.75*10^{-5}nM.s^{-1}$	Vilar et al, Plos Computational Biology, 2006
pRII	receptors II production rate	$4.87*10^{-5}nM.s^{-1}$	Vilar et al, Plos Computational Biology, 2006
alpha	efficiency of recycling of active receptors	1 no unit	Vilar et al, Plos Computational Biology, 2006
$k_{on\,pS24nTIF1\gamma}$	pS24nTIF1γ complex association rate	$2*10^{-3}nM^{-1}.s^{-1}$	Dupont et al, Cell, 2009
$k_{off\,pS24nTIF1\gamma}$	pS24nTIF1γ complex dissociation rate	$1.6*10^{-2}s^{-1}$	Dupont et al, Cell, 2009
$k_{off\,pS2nTIF1\gamma}$	pS2nTIF1γ complex dissociation rate	$1.6*10^{-2}s^{-1}$	He et al, Cell, 2006
$k_{in\,S4ub}$	S4ub import rate	$5.2*10^{-3}s^{-1}$	Dupont et al, Cell, 2009
K_{dub}	S4ub c deubiquitination rate	$7*10^{-3}nM^{-1}.s^{-1}$	Dupont et al, Cell, 2009

Table S2

$$
\begin{cases}
\frac{d}{dt}[S4c] = k_{in}[S4n] - k_{in}[S4c] - k_{on}[S4n][pS2n] + k_{off}[pS24c] + k_{dub}[S4ubc][FAM] \\
\frac{d}{dt}[S4n] = k_{in}[S4c] - k_{in}[S4n] - k_{on}[S4n][pS2n] + k_{off}[pS24n] \\
\frac{d}{dt}[S2c] = k_{ex}[S2n] - k_{in}[S2c] - k_{phos}[S2c][LRe] \\
\frac{d}{dt}[S2n] = k_{in}[S2c] - k_{ex}[S2n] + k_{dephos}[S24n][PPase] \\
\frac{d}{dt}[pS2c] = k_{ex}[pS2n] - k_{in}[pS2c] + k_{phos}[S2c][LRe] - k_{on}[pS2c]([S4c] + 2[pS2c]) + k_{off}([pS24c] + 2[pS22c]) \\
\frac{d}{dt}[pS2n] = k_{in}[pS2c] - k_{ex}[pS2n] - k_{dephos}[pS2n][PPase] - k_{on}[pS2n]([S4n] + 2[pS2n]) + k_{off}([pS24n] + 2[pS22n]) \\
+ k_{offpS2nTIF1\gamma}[pS2nTIF1\gamma] \\
\frac{d}{dt}[pS24c] = k_{on}[pS2c][S4c] - k_{off}[pS24c] - k_{in} * CIF * [pS24c] \\
\frac{d}{dt}[pS24n] = k_{on}[pS2n][S4n] - k_{off}[pS24n] + k_{in} * CIF * [pS24c] - k_{onpS24nTIF1\gamma}[pS24n][TIF1\gamma] \\
\frac{d}{dt}[pS22c] = k_{on} * 2[pS2c] - k_{off}[pS22c] - k_{in} * CIF * [pS22c] \\
\frac{d}{dt}[pS22n] = k_{on} * 2[pS2n] - k_{off}[pS22n] + k_{in} * CIF * [pS22c] \\
\frac{d}{dt}[TGF\beta] = 0 \\
\frac{d}{dt}[PPase] = 0 \\
\frac{d}{dt}[LR] = k_a[TGF\beta][RII] - (k_{cd} + k_{lid} + k_i)[LR] \\
\frac{d}{dt}[RI] = pRI - k_a[TGF\beta][RI][RII] - (k_{cd} + k_i)[RI] + k_r[RIe] + alpha * k_r[LRe] \\
\frac{d}{dt}[RII] = pRII - k_a[TGF\beta][RI][RII] - (k_{cd} + k_i)[RII] + k_r[RIIe] + alpha * k_r[LRe] \\
\frac{d}{dt}[RIe] = k_i[RI] - k_r[RIe] \\
\frac{d}{dt}[RIIe] = k_i[RII] - k_r[RIIe] \\
\frac{d}{dt}[LRe] = k_i[LR] - k_r[LRe] \\
\frac{d}{dt}[TIF1\gamma] = -k_{onpS24nTIF1\gamma}[pS24n][TIF1\gamma] + k_{offpS2nTIF1\gamma}[pS2nTIF1\gamma] \\
\frac{d}{dt}[FAM] = 0 \\
\frac{d}{dt}[pS24nTIF1\gamma] = k_{onpS24nTIF1\gamma}[pS24n][TIF1\gamma] - k_{offpS24nTIF1\gamma}[pS24nTIF1\gamma] \\
\frac{d}{dt}[pS2nTIF1\gamma] = k_{offpS24nTIF1\gamma}[pS24nTIF1\gamma] - k_{offpS2nTIF1\gamma}[pS2nTIF1\gamma] \\
\frac{d}{dt}[S4ubc] = k_{inS4ub}[S4ubn] - k_{dub}[S4ubc][FAM] \\
\frac{d}{dt}[S4ubn] = k_{offpS24nTIF1\gamma}[pS24nTIF1\gamma] - k_{inS4ub}[S4ubn]
\end{cases}
$$

Figure S1

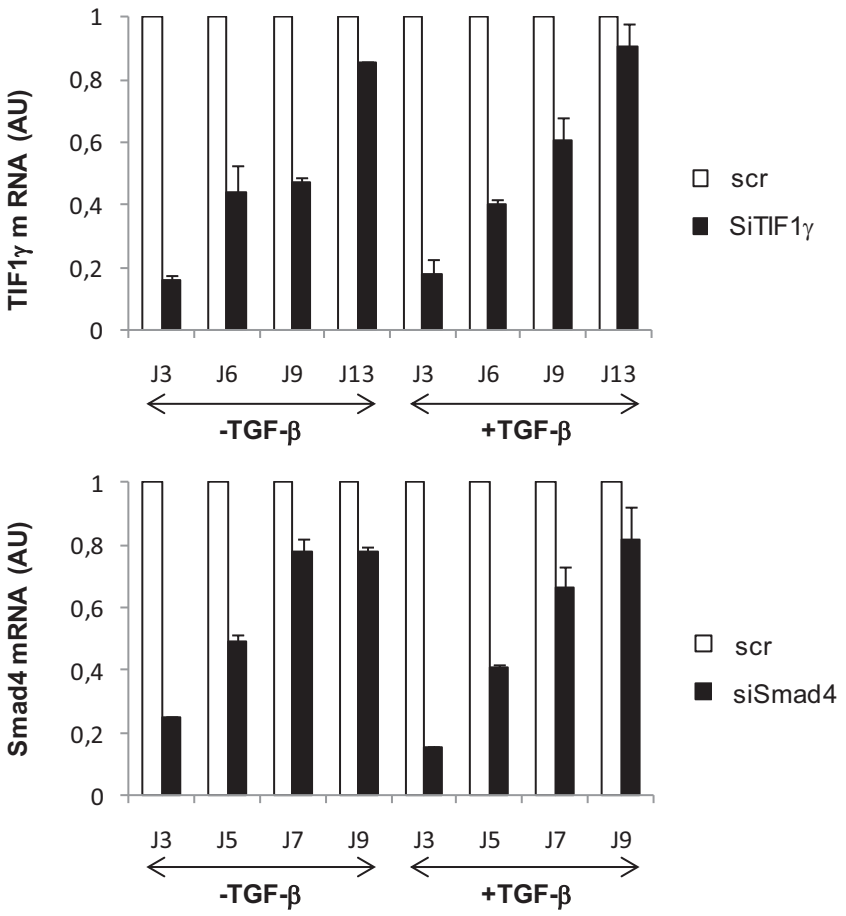

TIF1γ m RNA (AU)
1
0,8
0,6
0,4
0,2
0
J3
J6
J9
J13
J3
J6
J9
J13
-TGF-β
+TGF-β
scr
SiTIF1γ
Smad4 mRNA (AU)
1
0,8
0,6
0,4
0,2
0
J3
J5
J7
J9
J3
J5
J7
J9
-TGF-β
+TGF-β
scr
siSmad4

Figure S2

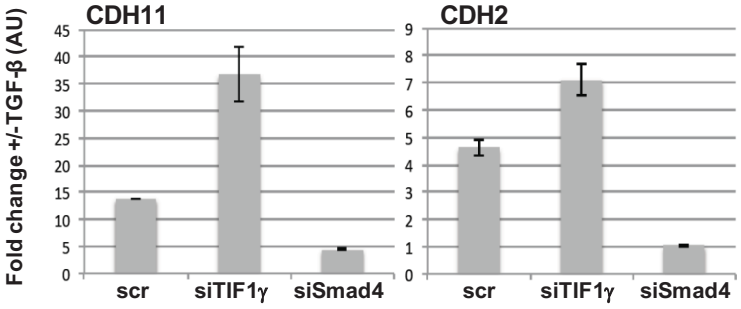

CDH11
CDH2
Fold change +/-TGF-β (AU)
45
40
35
30
25
20
15
10
5
0
9
8
7
6
5
4
3
2
1
0
scr
siTIF1γ
siSmad4
scr
siTIF1γ
siSmad4

Figure S3

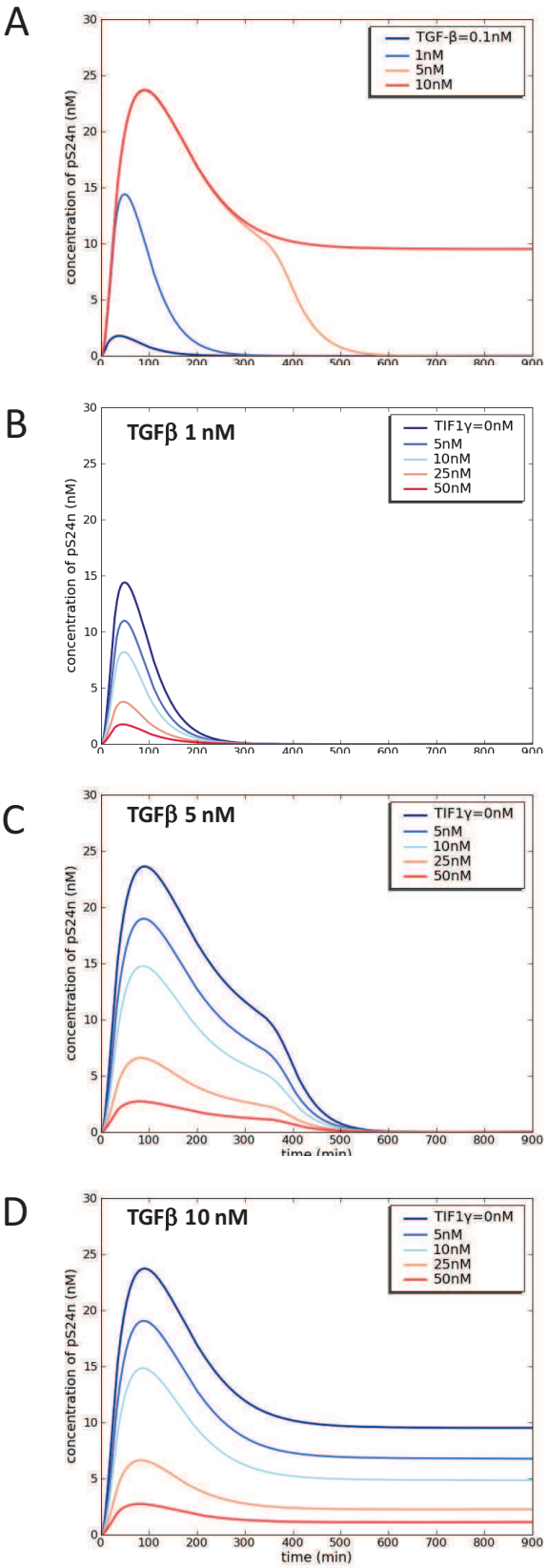
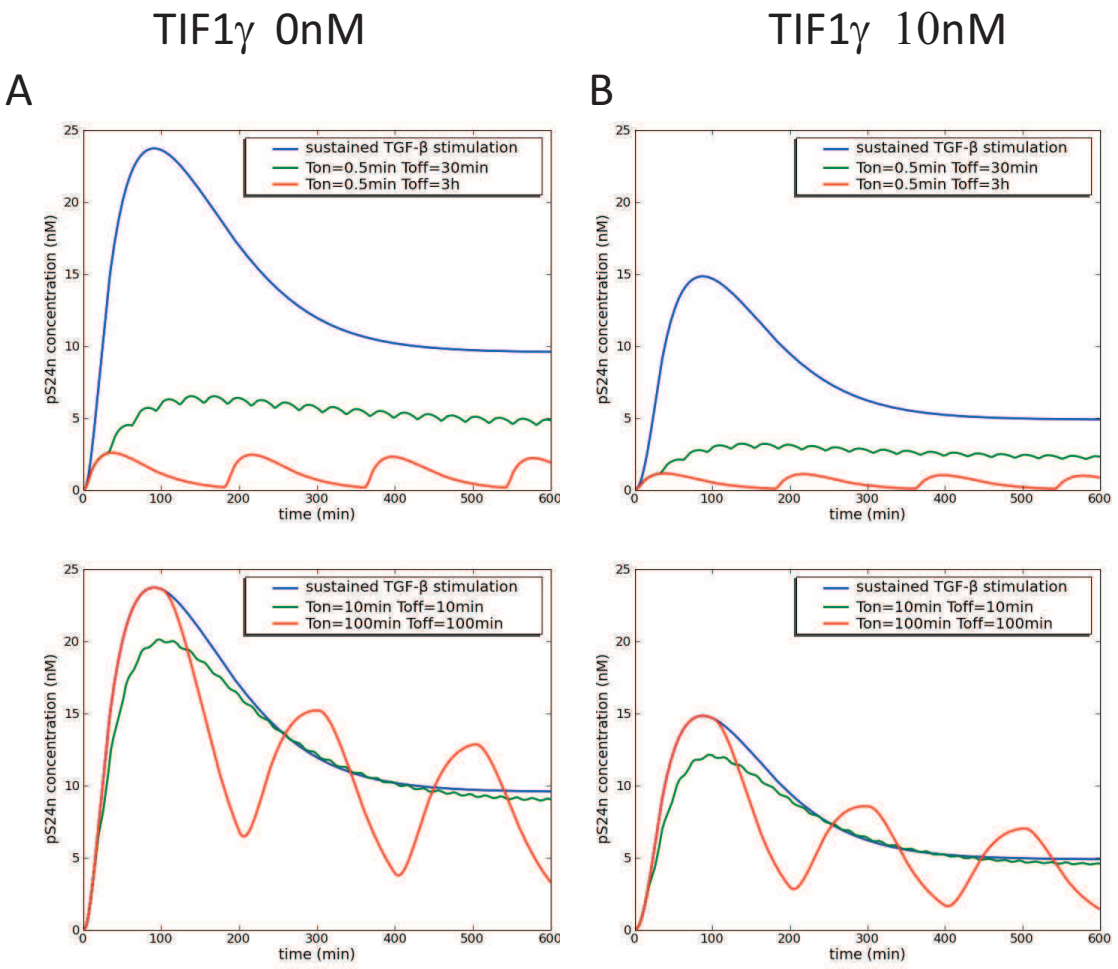

A
TGF-β=0.1nM
1nM
5nM
10nM
concentration of pS24n (nM)
B
TGFβ 1 nM
TIF1γ=0nM
5nM
10nM
25nM
50nM
concentration of pS24n (nM)
C
TGFβ 5 nM
TIF1γ=0nM
5nM
10nM
25nM
50nM
concentration of pS24n (nM)
time (min)
D
TGFβ 10 nM
TIF1γ=0nM
5nM
10nM
25nM
50nM
concentration of pS24n (nM)
time (min)

Figure S4


TIF1γ 0nM
TIF1γ 10nM
A
B
sustained TGF-β stimulation
Ton=0.5min Toff=30min
Ton=0.5min Toff=3h
Ton=10min Toff=10min
Ton=100min Toff=100min
pS24n concentration (nM)
time (min)

Chapitre 3

Conclusion et Perspectives

Cette étude avait pour but de caractériser le rôle d'un nouveau régulateur du signal TGF-β. Des effets en apparence contradictoires ayant été observés dans différentes études, nous avons testé différentes hypothèses fondées sur ces observations. A savoir, l'association préférentielle de TIF1γ avec SMAD4 et/ou SMAD2. Pour cela nous avons conçu 3 modèles différentiels : (i) interaction entre TIF1γ et SMAD4, (ii) interaction entre TIF1γ et SMAD2 phosphorylée, (iii) interaction entre TIF1γ et le complexe SMAD2 phosphorylée/SMAD4 (modèle hybride). Ces modèles sont basés sur l'intégration de 2 modèles différentiels issus de la littérature [146, 132]. Pour inférer les paramètres concernant TIF1γ, nous avons utilisé les données qualitatives issues de 3 publications [37, 36, 51]. Ce qui nous a permis de déterminer des paramètres cinétiques réalistes, en sachant par exemple que l'affinité entre TIF1γ et les SMAD est décrite comme identique à celle entre SMAD2 et SMAD4.

Les modèles obtenus ont été simulés afin d'observer le comportement du signal TGF-β, abstrait à la concentration du complexe pS24n, en fonction du ratio entre TIF1γ et SMAD4. Les deux premiers modèles parviennent à reproduire le comportement supposé par leurs hypothèses respectives, mais ne peuvent en revanche expliquer les observation de l'autre équipe. Le modèle hybride est en revanche capable de reproduire les effets observés par les différentes équipes et est en mesure d'expliquer les résultats obtenus sur TIF1γ dans les 3 publications [37, 36, 51]. La modularité de cet effet est basée sur le ratio TIF1γ/SMAD4 qui peut ou non inhiber le signal canonique du TGF-β (*i.e.* diminuer la concentration de pS24n). Nos prédictions par simulation ont été validées à posteriori par des manipulations expérimentales visant à observer l'expression de gènes cibles du TGF-β en fonction de différents ratio TIF1γ/SMAD4. D'autre part, l'hypothèse de l'existence d'un trimère entre TIF1γ et p24n fut également confortée par le fait que ces 3 entités ont été observées sur le promoteur du gène PAI-1, cible du TGF-β.

L'utilisation d'un modèle différentiel nous à permis de tester et de valider des hypothèses sur le mécanisme de régulation de la voie canonique du TGF-β par TIF1γ. Ce-

pendant le modèle existant décrit uniquement la voie canonique du TGF-β et ne permet donc pas d'étudier l'impact de ce nouveau régulateur sur les voies non SMAD. L'extension de ce modèle différentiel à un modèle intégrant les voies non SMAD n'est pas envisageable. En effet nous avons ici eu besoin d'un très grand nombre d'expérimentations recueillies dans 3 publications pour ajouter le nouveau composant TIF1γ et les paramètres cinétiques associés à ces réactions. En raison de la nécessité d'avoir des données quantitatives, cette méthode ne peut être étendue à un modèle avec un grand nombre de nouveaux interactants. Bien que robuste et informative, cette approche n'est pas applicable à de grands modèles, par manque de données biologiques.

Récemment ce modèle à fait l'objet d'autres utilisations suite à différents partenariats. Dans le cadre du projet ANR Biotempo (`http://biotempo.genouest.org`), ce modèle est aujourd'hui utilisé pour tester des techniques de réduction de modèles basées sur la tropicalisation. Cette réduction est liée a une analyse de sensibilité afin de déterminer quels sont les réactions et les paramètres les plus influents dans le modèle. Ces travaux ont également vocation à intégrer le modèle quantitatif de la voie canonique du TGF-β avec un autre modèle différentiel existant, celui de la voie EGFR. Le but étant de représenter les interactions entre ces deux voies très étudiées et d'analyser quantitativement l'impact de chacune de ces voies sur l'autre.

Troisième partie

Modéliser la dynamique de propagation du signal : approche basée sur le formalisme des transitions gardées

Chapitre 1

Présentation

Comme présenté dans la partie précédente, l'utilisation de modèles différentiels peut être très utile dans la compréhension de mécanismes fins de régulation du signal. Cependant, les limites de ces approches sont rapidement atteintes quand on augmente le nombre de composants du modèle. D'une part les données nécessaires sont inaccessibles, et d'autre part le point de vue biochimique sur lequel sont basées les approches différentielles n'est pas adapté à la modélisation des différents mécanismes impliqués dans la propagation du signal. Enfin les résultats obtenus par simulation de ces modèles n'offrent pas explicitement d'information sur le flux de signal et la causalité entre les entités du système n'est pas directement explicitée par les simulations.

La deuxième partie des résultats de cette thèse est consacrée au développement d'une approche de modélisation mieux adaptée à la conception ainsi qu'a l'analyse de modèles large échelle de la signalisation cellulaire. Considérant l'état de l'art actuel dans ce domaine, les approches discrètes sont encore limitées en pratique par la taille des modèles dynamiques représentés qui ne dépasse généralement pas la centaine d'interactants. Une autre limite récurrente est la gestion de la dynamique, qui par manque d'information ne parvient généralement pas à retranscrire la propagation du signal de façon réaliste. Au travers des résultats présentés dans cette partie, nous tacherons de contribuer à l'avancée de ces réflexions actuelles. Nous présentons dans un premier temps la démarche scientifique qui à conduit à notre approche, puis dans un second temps, l'élaboration et l'application du formalisme Cadbiom.

1.1 Qu'est ce que le signal ? Quelle abstraction ?

Indépendamment de tout formalisme, il est nécessaire de définir la question biologique afin de fixer un cadre de travail. L'idée de départ visait à développer un vaste modèle intégrant l'ensemble des voies TGF-β : voie canonique et voies non SMAD. Étudier l'ensemble de la signalisation dépendante du TGF-β permet de prendre en compte la com-

plexité de la réponse au TGF-β et ainsi de pouvoir répondre à des questions plus globales. Afin de recueillir les informations nécessaires à la conception d'un modèle, nous avions le choix entre une interprétation manuelle de la littérature, ou une interprétation automatique d'une base de données.

Nous avons d'abord exploré la première solution qui s'est révélé être inapplicable en pratique pour différentes raisons. Tout d'abord le coût en temps, du au choix et à la lecture des publications pour concevoir un modèle large échelle, était bien trop grand pour une équipe restreinte. L'interprétation de la littérature est de plus trop subjective et peut biaiser le modèle dès lors qu'elle est réalisée par le modélisateur.

Nous avons donc opté pour la deuxième solution : la conception automatique de nos modèles en choisissant pour référence la base de données Pathway Interaction Database (PID), pour la richesse et la qualité de son contenu manuellement annoté et orienté sur la signalisation cellulaire chez l'homme. L'utilisation d'une base de donnée diminue les problèmes liés à l'exploitation directe de la littérature. Les données sont en effet déjà interprétées par un ensemble d'experts du domaine et n'ont pas pour vocation première de concevoir un modèle particulier. Elles sont donc moins biaisée qu'une interprétation manuelle par un modélisateur unique. Nous avons donc développé un schéma de traduction pour interpréter les données de PID. Comme dans la plupart des bases de données, les grandes voies de signalisation comme celles du TGF-β, sont décrites dans PID. Cependant ces voies se limitent généralement à la description de la partie canonique, et ne prennent pas en compte les nombreuses interactions observées avec d'autres voies, pourtant ces informations sont contenues dans la base de données. Nous avons donc recherché les termes qui dépendaient du TGF-β dans cette base en analysant le graphe d'activation issu de PID (voir chapitre suivant). A notre grande surprise nous avons constaté que la majorité des données était reliées plus ou moins directement à la voie du TGF-β. Dans ce contexte, nous avons fais le choix d'utiliser toutes les connaissances à notre disposition dans la base PID afin de modéliser la signalisation cellulaire dans son ensemble en intégrant toutes les voies de signalisation. Dans cet optique, un changement de paradigme, entraînant également un changement de formalisme, a été nécessaire. A l'inverse des méthodes actuelles, nous n'avons pas cherché à modéliser une voie de signalisation spécifique. Compte tenu de l'extrême connectivité des données sur la signalisation, nous avons choisi d'intégrer l'ensemble des connaissances autour de la signalisation cellulaire pour concevoir notre modèle. De cette manière, nous souhaitons obtenir un modèle qui ne soit pas biaisé par une notion restrictive et subjective de voie, mais obtenir au contraire un modèle représentatif de la signalisation c'est à dire un réseau extrêmement bien connecté.

Cependant l'utilisation de cette grande quantité de données dans le but de concevoir

un modèle large échelle de la signalisation cellulaire a également des inconvénients qu'il nous faut dépasser. D'une part la plupart des modèles dynamiques se limite à une centaine de molécules notamment pour une question de temps de calcul. La complexité qui émane d'un modèle de plusieurs milliers de molécules va sans aucun doute entraîner des temps de calcul important. D'autre part la signalisation du TGF-β se retrouve noyée dans un système où certaines molécules n'ont aucun effet sur le signal TGF-β. Nous allons donc avoir besoin d'extraire le sous réseaux représentatif de la propagation du signal TGF-β.

Après avoir défini la limite de ce que l'on souhaite représenter, il faut ensuite s'intéresser à la façon dont on perçoit la biologie, c'est à dire avec quelle abstraction interpréter ces données. Nous avons choisi de ne pas utiliser des notions biochimique, tel que la stœchiométrie, mais de nous focaliser sur la notion de propagation du signal, c'est à dire de l'information biologique. Le but étant d'observer quelles sont les voies empruntées pour transmettre le signal, et produire tel ou tel effet. Nous représentons "qui agit avec qui", sous quelles conditions et dans quels buts. Pour toutes ces raisons, notre interprétation de la biologie est basé sur la notion de *réaction biologique* entre *biomolécules* (Figure 1.1).

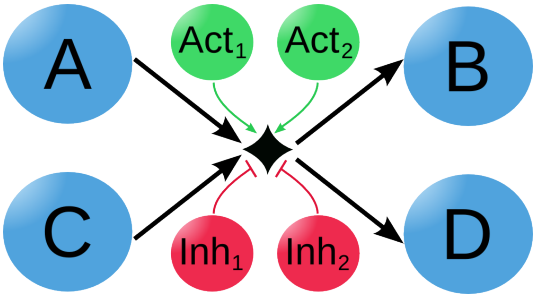


FIG. 1.1: Illustration du concept de réaction biologique. La réaction est composée de deux biomolécules entrantes A et C, ainsi que de deux biomolécules sortantes B et D. Cette réaction est régulée positivement par des activateurs (Act1 et Act2) et négativement par des inhibiteurs (Inh1 et Inh2).

Réaction biologique Une réaction biologique consomme des biomolécules dites entrantes, et produit des biomolécules dites sortantes. Elle peut dépendre de régulateurs qui sont des biomolécules ayant soit un rôle positif : activateur, soit un rôle négatif : inhibiteur. Les régulateurs conditionnent la réaction mais ne sont ni produits, ni consommés par celle-ci. Cette représentation de la biologie est en adéquation avec la pensée du biologiste et permet de représenter la plupart des observations faites à partir des expérimentations, celles ci étant reportées ensuite dans la littérature et les bases de données. A ce niveau

d'abstraction, l'information biologique se transmet des biomolécules entrantes vers les biomolécules sortantes.

Biomolécule Les réactions biologiques font généralement intervenir des protéines modifiées (ex : phosphorylée) ou non, sous forme de monomère ou de complexe. Cependant d'autres acteurs peuvent jouer un rôle dans une réaction biologique : gènes, ARN, ions ou encore métabolites. Ces différents acteurs sont ainsi regroupés sous le terme de biomolécules. D'un point de vue biologique, ces biomolécules sont sujet à différents changements au cours de la signalisation cellulaire. Elles peuvent changer d'état d'activation (ex : par phosphorylation), de compartiment sub-cellulaire (ex : du cytoplasme vers le noyau), être dégradées, etc. Tous ces changements observés doivent être représentés car ils influencent le rôle des biomolécules. En effet une protéine phosphorylée ou non n'intervient pas nécessairement dans les mêmes réactions. C'est pourquoi cette protéine est représenté par 2 biomolécules : une annoté "phosphorylé", l'autre pas. De la même façon une protéine située dans le cytoplasme ou le noyau sera représentée par 2 biomolécules différentes.

1.2 Cadbiom : un formalisme discret pour modéliser la signalisation cellulaire

Pour analyser la dynamique de propagation du signal dans les réseaux de signalisation, il est nécessaire de représenter formellement la notion de réaction biologique. Pour cela, nous nous sommes focalisés sur 3 points qui nous semblent essentiels : (i) le développement d'un formalisme adéquat, (ii) la mise au point de méthodes permettant d'explorer la dynamique du système généré, (iii) la restitution des résultats de manière compréhensible. Ces 3 points sont brièvement introduits dans les paragraphes suivants.

Développement d'un formalisme adéquat Compte tenu du fait que nous souhaitions modéliser la signalisation cellulaire dans son ensemble, il nous fallait un formalisme discret capable de représenter la notion de réaction biologique, et surtout de pouvoir traiter des réseaux de grande échelle. Pour ces raisons,nous avons développé le formalisme Cadbiom (Computer Aided Design of Biological Models), basé sur les transitions gardées [50, 113].

Un modèle Cadbiom se représente par un ensemble de places et de transitions. Les places représentent les biomolécules, leur domaine de valeur est booléen (0 ou 1). Une place avec une valeur égale à 0 est dite inactive, active si égale à 1. A notre niveau d'abstraction de la biologie, une place active signifie que la biomolécule qu'elle représente est en concentration suffisante pour jouer son rôle quel qu'il soit. Au cours de l'évolution

du système dynamique, les places changent d'état d'activation, peuvent être activés ou inactivées en fonction du franchissement des transitions (cf. paragraphe sur les transitions). L'ensemble des valeurs des places définit l'état du système à un pas de temps donné.

Une transition représente un flux d'information d'une biomolécule entrante (source de la transition) vers une biomolécule sortante (cible de la transition). Une transition est ainsi orientée d'une place entrante vers une place sortante. Le franchissement des transitions est responsable de la dynamique du modèle. C'est pourquoi nous y avons porté un grand intérêt de façon à ce que la propagation du signal ne soit pas biaisée par des critères topologiques, ou le niveau de détails choisi pour représenter une voie. Pour cela nous avons utilisé la notion d'événement qui va permettre d'affiner la dynamique du modèle en comparaison avec les approches usuelles synchrone ou asynchrone.

Afin de concevoir un modèle automatiquement à partir de la base de données PID, nous avons développé une interprétation des réactions biologiques stockées dans PID, dans notre formalisme Cadbiom. Pour cela nous avons développé un schéma de traduction capable d'intégrer, de façon cohérente, l'ensemble des réactions contenues dans cette base pour ainsi générer un unique modèle à partir des 137 voies indépendantes décrites dans PID.

Exploration de la dynamique du système Le conception de modèle n'étant pas une fin en soi, il est nécessaire d'analyser le système pour faire ressortir de nouvelles informations biologiques. C'est pourquoi nous avons implémenté un simulateur pour visualiser la propagation du signal à travers l'enchaînement de l'activation/inactivation des places. Nous avons également travaillé sur l'interrogation du système dynamique autour des concepts classiques de vérification de modèle que sont l'atteignabilité et l'invariance. Ces méthodes permettent de répondre à des questions qualitatives sur la dynamique du signal, de façon à retrouver par exemple les biomolécules nécessaires à l'activation ou à la répression d'un gène.

Restitution des résultats La difficulté majeure dans l'analyse de grands réseaux est la formulation des résultats. La principale forme de résultats générés est une liste de places (*i.e.* biomolécules) qui, si elles sont activées, vont induire un comportement recherché (ex : activation d'un gène) que nous définissons par le terme de propriété. En fonction du modèle et de la question posée, la quantité de résultats peut rapidement s'avérer trop importante pour être traitée manuellement. Le signal pouvant se propager à travers différentes voies, il n'est en pratique pas rare d'obtenir des dizaines, des centaines voir des milliers d'ensemble de biomolécules participant à la régulation d'un gène. Cette abondance de conditions permettant d'obtenir un comportement unique d'un système dynamique complexe s'explique par au moins 2 phénomènes :

1. La complexité combinatoire intrinsèque au modèle qui résulte des embranchements entre les voies de signalisation mais aussi d'une certaine redondance dans la description des faits. Par exemple, pour transporter le complexe SMAD3/SMAD4 du cytoplasme vers le noyau, 4 activateurs sont décrits. En considérant ces activateurs comme indépendants, ce transport donne lieu à 4 régulations possibles et multiplie ainsi les conditions possibles par 4 uniquement par la présence de l'un de ces activateurs.
2. Le bruit que peut entraîner la présence de biomolécules indépendantes du comportement recherché. C'est à dire des biomolécules qui ne sont ni nécessaires ni conséquentes de la propriété désirée.

L'élimination de ces comportements parasités a été obtenue en introduisant le concept de conditions d'activation minimale. Ce concept se décline de différentes manières et nous nous sommes focalisés sur la minimalité des conditions initiales : le nombre de places qui doivent être activées lors de l'initialisation du modèle pour parvenir à une propriété désirée. Des algorithmes spécifiques ont été développés pour calculer ces conditions. Malgré ce critère de minimalité, nous avons obtenu de très grands ensemble de solutions nous obligeant à nous pencher sur le problème de la restitution des résultats en cherchant à regrouper les conditions obtenues. Les techniques de classification classique se révélant être inefficaces, nous avons développé nos propres méthodes qui s'appuient sur différents critères pour ordonner et regrouper les solutions.

Pour mettre en avant cette nouvelle méthodologie et faciliter son accès à la communauté des biologistes/modélisateurs, nous avons développé le logiciel Cadbiom. Ce logiciel offre une aide complète, permettant ainsi de concevoir des modèles à l'aide de l'interface graphique, ou directement en important un modèle depuis la base de données PID. Les modèles générés sont ensuite totalement modifiables via l'interface. Différentes analyses sont proposées, notamment des analyses statiques, la simulation des modèles ou encore la recherche de propriétés sur la dynamique des modèles. Une version du manuel d'utilisation est présentée en annexe.

La description du formalisme Cadbiom, le schéma d'interprétation des réactions biologiques et l'application de notre méthode sont présentées en détails dans le chapitre 2.

En complément du développement de cette méthodologie, nous avons travaillé sur différents points nous permettant d'éprouver notre approche. La conception d'un modèle dynamique de la signalisation cellulaire nous permet d'étudier des comportements très vastes tels que la régulation des gènes présents dans le modèle. Dans ce contexte nous

avons réalisé une étude d'identification des conditions d'activation des 787 gènes de la base PID afin de caractériser les trajectoires de régulation de ces gènes et identifier un lien potentiel avec leur corrélation d'expression. Nous avons également décidé d'interpréter une autre source de donnée pour ainsi disposer d'un autre modèle de la signalisation. Nous avons choisi la base de données Reactome qui utilise également la notion de réaction biologique. Enfin, grace aux partenariats de l'ANR Biotempo, il nous à été possible de comparer notre méthode avec un autre formalisme : le Process Hitting. Ces 3 études supplémentaires sont présentés dans le chapitre 3.

Chapitre 2

Article : A guarded transition approach to integrate the human cell signaling pathways into a single unified dynamic model.

A guarded transition approach to integrate the human cell signaling pathways in a unique dynamic model

Geoffroy Andrieux[1], Michel Le Borgne[2], Nathalie Théret[1]

1. INSERM U1085, IRSET, Université de Rennes 1, 2 avenue Pr Léon Bernard, 35043 Rennes, France
2. Université de Rennes 1, IRISA, 263 avenue du général Leclerc, 35042 Rennes, France

Abstract

Large dynamic models are necessary to decipher the complexity of the cell signaling networks that orchestrate cell life. The formal representation of biological knowledge and its interpretation into mathematical models constitute major challenges. Here we address these difficulties by proposing a new non-ambiguous formal interpretation of signaling pathways into discrete dynamic models. Based on guarded transitions, the CADBIOM language describes dynamic behavior by introducing temporal parameters to manage competition and cooperation processes. Using this new formalism, we have built the first single unified model of cell signaling by integrating the 137 human signaling maps from the Pathway Interaction Database. Exploration of this model yields new insights, as illustrated by temporal property-checking analyses of signaling-dependent regulation of the cell-cycle. Indeed, such an exploration extends to the identification of conditions for activation or inhibition of p21, a major regulator of the cell cycle. CADBIOM is composed of applications for model design that combine a graphical interface, tools for simulation and checking of temporal properties and methods for the exploration of solutions through visualization of graphs.

Introduction

The life of a cell is regulated by a permanent dialogue with its microenvironment and with other cells. This communication is supported by the signal transduction network containing all the molecules involved in detection, transmission and translation of information from the extracellular space to the nucleus. Over the last decade, most of the models that aim to capture the dynamics of a signal transduction pathway have used models based on Ordinary Differential Equations (ODE) with a limited number of molecules (see Biomodel Database (28)). However, the explosion in the number of variables in complex networks requires the development of qualitative modeling approaches. This results in two major challenges: how to collect and represent data from biological knowledge and how to interpret them into mathematical models. Recent advances in high-throughput technologies have helped accumulate data and considerable effort has gone into building cell signaling maps. While much information about the molecular components of these pathways is widely collected in databases such as Kegg (20), Ingenuity Pathway (16) or Biocarta, the output remains largely limited to descriptive graph-based representations. Indeed, this distributed knowledge mixes various biological concepts such as biochemical reactions and functional processes and includes heterogeneous levels of details that do not permit the automatic generation of mathematical models. More recent databases such as Reactome (9) and the Pathway Interaction Database (PID) (41) use a more homogeneous concept of biological reactions which might facilitate further formal interpretations. However, no dynamic integrative models emerge from such repositories. In addition, the development of formal model specifications -or formalisms- is required to describe biological events in a language suitable for mathematical analysis and simulation.

Interpreting biological processes in a formal representation is far from trivial and many attempts have been made over the past several years to propose standard visual languages that mimic the biologist's view. Wiring diagrams are the most popular representation of signaling pathways. Accordingly, molecular interaction maps (24) and process description diagrams (23; 33; 37) have formed the basis for formal description by providing symbols, syntax and grammars to describe interactions and relationships between biomolecules. More recently, the System Biology Graphical Notation (SBGN) standard language has been developed by a consortium of scientists to offer a uniform representation and visualization of biological networks (27). SBGN consists of three formal description types, the activity flow, the entity and the process diagrams, leading to an information-rich representation of biological networks. The SBGN specification has been rapidly adopted by the software designers' community for the development of visualization tools such as Celldesigner (15), VISIBIOweb (12) and IPAVS (42). In addition, the Biological Connection Markup Language has been created to manage the representation of signaling pathways with SBGN specifications (3), thereby completing the previously available machine-readable formats such as Systems Biology Markup Language (SBML) (18) and Biological Pathways eXchange (BioPAX)(11). While all of these systems greatly improve graph-based exploration and our understanding of network complexity, they do not tell us how a signal flows through a given circuit.

To analyze the dynamics of signaling networks (simulation, property search), we need to identify the class of mathematical models that is best suited to the formalism used to describe biological processes (29). Using the concept of event is naturally associated with qualitative data and discrete time. Many discrete formalisms have been proposed for biological network modeling such as Boolean networks (21), petri nets (6) and rule-based methods (10; 13). These formalisms differ essentially in the modeling paradigms they use, that is, in the mental framework that represents biological processes. Accordingly, Boolean networks are based on switching functions and consider gene networks as logical circuits, while rule-based models rely on multi-state component description.

To address the difficulties that arise from signaling complexity, we have developed a new non-ambiguous formal interpretation of signaling pathways as discrete dynamic models. This interpretation differs from previous approaches in that it models the signal rather than the molecular biochemical network that transmits the signal. The resulting Computer-Aided Design for BIOlogical Models (CADBIOM

) language is based on a simplified version of guarded transitions, a state/events formalism with a long history in computer science (17; 36). CADBIOM describes the dynamic behavior of this state transition-based system by introducing temporal parameters to manage competition and cooperation between parts of the models. To facilitate the use of CADBIOM language by biologists, we provide a graphical interface that lends itself to intuitive model design. This is complemented by tools for the simulation and checking of temporal properties. Finally, CADBIOM software offers simple methods for the exploration of solutions through visualization of reduced graphs. As an example of CADBIOM applications, we have automatically integrated the 137 available signaling maps from the PID database into a single unified CADBIOM model to build the first global model of cell signaling networks, whose model-checking temporal property is illustrated by the analysis of signaling-dependent regulation of the cell cycle.

As described below, we propose a new semantic approach that is supported by a language for modeling signaling networks. This semantics is highly appropriate for the integration of biological observations collected from heterogeneous sources. We also provide easy-to-use tools to create CADBIOM models and explore their dynamic behavior using computational methods.

Results

In this section we first describe the new CADBIOM formalism based on guarded transitions and its use to formalize biological reactions into dynamic models. We next propose methods to analyze the CADBIOM models using formal verification techniques. Finally, we use this formalism to build a large-scale dynamic model of cell signaling in its entirety and we further explore the properties of signals that regulate cell proliferation. Model design and analysis tools are implemented into an open-source CADBIOM application package, available at `cadbiom.genouest.org`.

Modeling biological reactions as guarded transitions

From biological reactions to guarded transitions The data from biological experiments are chiefly described according to the concept of a biological reaction. A biological reaction has inputs, outputs and is modulated by inhibitors and activators. From the perspective of information propagation, we consider that the occurrence of a biological reaction induces information propagation from each input to each output. These givens lead to models based on guarded transitions.

A guarded transition is initially represented graphically by: $A \overset{[Cond]}{\longrightarrow} B$. A and B are the origin and target places, respectively, of the transition. The condition (or guard) is called *Cond* and is a logical formula with places as variables. Places represent biomolecules which are *activated* or *inactivated*. A biomolecule is activated when it is present in sufficient quantity to play its role in the signaling network. The transition represents the information transfer from the origin place to the target place under conditions described in the guard. The top part of Figure 1A illustrates the scheme for translating a simple biological reaction into a guarded transition. The transition requires the presence of activator (*Act*) and the absence of inhibitor (*not Inh*). Generalization to a biological reaction with several inputs and outputs is illustrated in the bottom part of Figure 1A. Because all the inputs are required for the reaction to occur, the input C appears in the $Guard1$ of the transition $A \longrightarrow B$ and A appears in the $Guard2$ of the transition $C \longrightarrow B$. In addition, the inputs must be combined with other conditions within the guards. Combining inhibitors and activators in a transition guard requires detailed biological information. The combinatorics of inhibitors/activators are rarely well documented, if at all, in the literature. In the absence of information, we can combine activators and inhibitors either with the conjunction "and" or the disjunction "or". Using the "or" option in the example from the bottom part of Figure 1A, the $Guard1$ combines the input C with activators and inhibitors as follows: C *and* $(act1$ *or* $act2)$ *and not* $(Inh1$ *or* $Inh2)$.

As described in Figure 1B, this general translation rule can be adapted to different types of biological reactions such as binding/dissociation or synthesis/degradation. Note that to represent protein synthesis

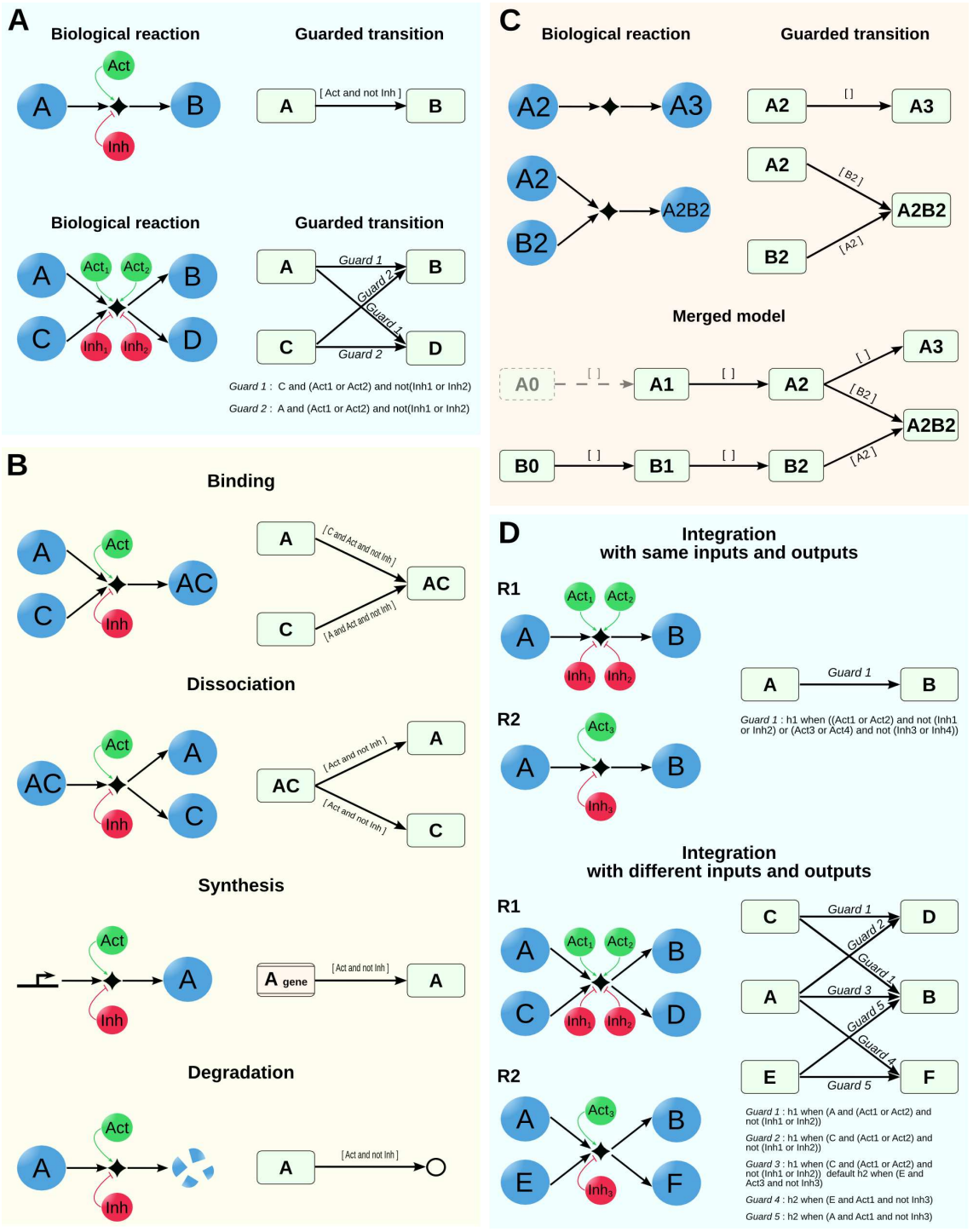


Figure 1: Translation scheme for biological reactions into guarded transitions.

we introduce a permanent place that symbolizes the corresponding coding gene and which, unlike other places, is never inactivated by an outgoing transition. On the contrary, trap places are introduced for degradation processes and do not have outgoing transitions.

The two rules for the dynamics of the system are: 1) a transition is fired when its origin is activated and its guard condition is true (it is fireable); 2) after transition firing, the origin is inactivated and the target is activated.

The simplest way to compose guarded transitions is to connect them by places: the target of a transition becomes the origin of the following transition. The resulting network of guarded transitions is turned into a dynamic system according to the choice of transitions which are fired at each step. This can be done in several different ways: for example, asynchronous approaches allow at most one transition whereas data-flow approaches support a transition every time it is fireable. The asynchronous scheme is not well adapted to the representation of signal propagation since, as demonstrated for complex formation, transitions do not fire one by one, making the data-flow scheme more realistic. Nevertheless, difficulties arise from reactions that share the same input and/or output biomolecules. As shown in Figure 1C (top), $A2$ is involved in two reactions and can either be transformed into $A3$ or make a complex with $B2$. In this case, the transition $A2 \xrightarrow{[\,]} A3$ is fired when $A2$ alone is activated while the transition $A2 \xrightarrow{[B2]} A2B2$ is fired when $A2$ and $B2$ are both activated.

A composition rule is applied to produce a merged model (bottom of Figure 1C in the absence of $A0$) in which it is assumed that $A1$ and $B0$ are both activated. Using the data-flow evolution rule, the transition $A2 \xrightarrow{[B2]} A2B2$ does not occur as it formally requires two transitions, ($B0 \xrightarrow{[\,]} B1$, $B1 \xrightarrow{[\,]} B2$), whereas only one transition, ($A1 \xrightarrow{[\,]} A2$), is required to render $A2 \xrightarrow{[\,]} A3$ fireable. In more simple terms, according to the data-flow scheme, $A2$ is consumed before $B2$ is available.

Adding a transition $A0 \xrightarrow{[\,]} A1$ (dotted line at the bottom of Figure 1C) allows for the formation of the $A2B2$ complex if the model has been initialized with $A0$ activated instead of $A1$. Such a postulate illustrates how the schedule of the model depends on the number and configuration of the transitions. It should also be noted that there is a competition between activation of $A3$ and formation of the $A2B2$ complex while the model must be capable of accounting for all possible outcomes: $A3$ activation alone, simultaneous $A3$ and $A2B2$ activation or $A2B2$ activation alone. An intrinsic limitation then is that the data-flow evolution rule only allows activation of $A3$ in the absence of $A0$ and activation of $A3$ and $A2B2$ in the presence of $A0$.

Introducing events to assemble guarded transitions To overcome such limits we introduce events to allow for delays in transition firing. Events are discrete signals that guide/restrain the choice of fireable transitions. Events only take one conventional "top" value denoted by $\top$. Relative occurrences of events

Figure 1: (**A**) Top: General case of a biological reaction where biomolecule A gives rise to biomolecule B. The regulation of the reaction is symbolized by activators Act and inhibitors Inh. The interpretation into a guarded transition (top right) from place A to place B is symbolized by rounded rectangles; the Act and Inh regulators form the guard of the transition. Bottom: a general representation of the case of multiple inputs, outputs and regulators. (**B**) Translation scheme for different categories of biological reactions into guarded transitions. Synthesis of a protein from its coding gene is symbolized by a double-lined rectangle, degradation is symbolized by a black circle. (**C**) Representation of the specific case where two reactions share the same biomolecule $A2$ (top). Merging of the reactions (bottom) leads to competition and the data-flow semantics do not allow the activation of place $A2B2$. (**D**) Representation of the integration of reactions that share the same biomolecules. The regulators of reactions that share all their inputs and outputs are merged into the guard transition using "or" operators (top). The guards for a transition resulting from different reactions are merged using "default" event operators (bottom).

are represented by a realization concept. The absence of an event at a realization step is symbolized by the $\perp$ symbol. A realization, then, is a sequence of N-tuples in $\{\perp, \top\}^N$ that excludes the N-tuples whose components are equal to $\perp$. For example, the realization of a pair h_1, h_2 of events can be represented as:

$$\begin{array}{rccc} h_1 & : & \top & \top & \perp \\ h_2 & : & \perp & \top & \top \end{array}$$

where each column corresponds to a discrete step and each line is an event. At the first step of the realization, the event h_1 occurs ($\top$) while the event h_2 does not ($\perp$). At the second step h_1 and h_2 occur ($\top$), and at the third step h_2 occurs ($\top$) but h_1 does not ($\perp$). This example can be generalized to N events. If at some step of the realization of a set of events, an event h occurs, we say that this event is present at this step and absent at the others. The potential firing of reactions must be dependent on the presence/absence of events. This is made possible by extending the syntax of guards to take these events into account. A guard is now of the form $h[Cond]$ where h is an event and $Cond$ is a logical condition, and the reaction is fireable if the origin is activated, the condition is satisfied and -this is new- the event is present. However, integrating many biological reactions requires that events be combined. To do this, two operators **default** and **when** are introduced to combine events with one another and with logical conditions. The **default** operator represents a merge of its operands and the **when** operator represents a selection of event occurrences that are based on model states. Intuitively, the **default** operator represents a kind of temporal extension of the *or* logical operator. h_1 **default** h_2 is present if h_1 is present or h_2 is present. Similarly, the **when** operator extends the *and* operator: h_1 **when** C is present if h_1 is present and C is true.

According to this event concept, an event expression h is introduced in the transition guard, graphically represented as $A \xrightarrow{h[C]} B$. The firing conditions now are:

1. The origin is activated
2. The condition is $True$
3. The event is present

The introduction of h increases the flexibility of the model by allowing different behaviors. Events allow for the selection of a subset of transitions among those enabled as follows: a missing event disables the firing of a group of transitions.

As shown in Figure 1C (Merged model), one way to fire the transition $A2 \xrightarrow{[B2]} A2B2$ is to modify the timing by adding an event h in each transition guard as follows: $B0 \xrightarrow{h_1[\,]} B1$, $B1 \xrightarrow{h_2[\,]} B2$, $B2 \xrightarrow{h_3[A2]} A2B2$, $A2 \xrightarrow{h_3[B2]} A2B2$, $A1 \xrightarrow{h_4[\,]} A2$, $A2 \xrightarrow{h_5[\,]} A3$. We use the same event h_3 for both $A2 \xrightarrow{h_3[B2]} A2B2$ and $B2 \xrightarrow{h_3[A2]} A2B2$ transitions since complex formation implies simultaneous firing of the two transitions. The introduction of events permits the activation of $A2B2$ when $A1$ and $B0$ are activated at the initial step if the following realization of events is enforced:

$$\begin{array}{rccc} Steps & 1 & 2 & 3 \\ h_1 & : & \top & \perp & \perp \\ h_2 & : & \perp & \top & \perp \\ h_3 & : & \perp & \perp & \top \\ h_4 & : & \top & \perp & \perp \\ h_5 & : & \perp & \perp & \perp \end{array}$$

Using this schedule, the following transitions are fired and $A2B2$ is activated:

$$\begin{array}{ccccc} 1 & & 2 & & 3 \\ A1 \xrightarrow{h_4[\,]} A2 & & B1 \xrightarrow{h_2[\,]} B2 & & A2 \xrightarrow{h_3[B2]} A2B2 \\ B0 \xrightarrow{h_1[\,]} B1 & & & & B2 \xrightarrow{h_3[A2]} A2B2 \end{array}$$

A complete mathematical definition of the semantics of guarded transitions and events is given in Supplementary Information.

Assembling guarded transitions with events into a dynamic model Combining reactions sharing the same biomolecules in order to integrate numerous reactions into a single model is a challenging task. As shown next, CADBIOM formalism is well adapted to the formulation of such combinations. To integrate $R1$ and $R2$ biological reactions that share the same inputs and outputs but have different regulators (inhibitors and activators), we generate two conditions: $cond1 = (Act1\ or\ Act2)\ and\ notInh1\ or\ Inh2)$ for $R1$ and $cond2 = (Act3\ and\ not\ Inh3)$ for $R2$ (Figure 1D, top). The new condition for the transition $A \rightarrow B$ resulting from the two reactions is then $cond1\ or\ cond2$. The event of the transition is a variable $h1$ since an occurrence of $R1$ cannot be distinguished from an occurrence of $R2$.

$R1$ and $R2$ are considered as two distinct biological reactions when at least one input or one output differs between reactions (Figure 1D, bottom). According to previous definitions, we generate pairs $(h1, cond1)$ and $(h2, cond2)$ for each reaction $R1$ and $R2$, where $h1$ and $h2$ are event variables and $cond1$ and $cond2$ are condition expressions for $R1$ and $R2$, respectively. Since information is transferred from A to B when $R1$ or/and $R2$ occurs, the event of the A transition is $(h1$ **when** $cond1)\ default\ (h2$ **when** $cond2)$ where $cond1 = C\ and\ (Act1\ or\ Act2)\ and\ not(Inh1\ or\ Inh2)$ and $cond2 = E\ and\ Act3\ and\ not\ Inh3$ (*Guard3* in Figure 1D, bottom).

Model analysis

Methods based on this new formalism can then be developed to explore the temporal properties of models. To do this, it is first necessary to define the concept of frontiers in guarded transition models.

A frontier is the set of places which cannot be activated from inside the model. In guarded transition models, places without input transitions belong to the frontier. As shown in Figure 2A, the place A is easily identified as a frontier place; however, other places that cannot be activated from inside the model are not as intuitively identified. To identify these places, one can consider the *transition graph* where places are the vertices and transitions are the edges and then search for isolated strongly connected components. For a strongly connected component, a transition is defined as an entering transition if its origin lies outside the component and its target is within the component. A strongly connected component is *isolated* if no entering transition can be fired when all the places in the component are inactivated. Figure 2A shows two isolated components: (C, D, E) and (F, G). Because these isolated components cannot be activated from inside the model, at least one place in each component must belong to the frontier for this component to be activated.

According to these definitions of frontiers, we can then analyze models using two major temporal properties, reachability and invariance, taking into account the initial conditions and events that occur along a trajectory.

Reachability of a property

Searching the reachability of a biological property such as "how to express a given gene?" requires that the dynamic scenarios that lead to the activation of the place symbolizing the gene be identified. These scenarios correspond both to the list of starting places to be activated and to the timing of activation of the places across the network. Choosing the places to be activated at the initialization step is not trivial, especially when considering large systems: forcing the activation of the frontier places is recommended when the signaling circuitry is investigated. All the other places are initialized in an inactivated state. Accordingly, the solutions to a reachability request are scenarios expressed as $(F, \mathscr{T})$ pairs, where F is a set of places which are activated at the initialization step and $\mathscr{T}$ is a sequence of sets of events h. At each $0, 1, \ldots, n-1$ step, the events h which are present are combined in a set H, leading to

the sequence $H_0, H_1, \ldots, H_{n-1}$ called the schedule or timing condition of the scenario. As shown in Figure 2B, $(\{A,B\},[\{h_2,h_4\},\{h_3\},\{h_0,h_1\},\{h_5\}])$ is a scenario to reach activation of the place P in 4 steps, where A and B are places activated at initialization and $[\{h_2,h_4\},\{h_3\},\{h_0,h_1\},\{h_5\}]$ is the sequence H_0,H_1,H_2,H_3. The events h which are not in subset H_i means that they are absent ($\perp$) at step i of the scenario.

This reachability property is then used to search for minimal scenarios such that the property is not achieved as soon as one component is removed in the initialization places or as soon as a component of event timing is disabled. As a result, *minimal activation conditions* (MAC) are defined as the set of scenarios which cannot be reduced, either at the level of the set of activated places at initialization, or at level of the set of events h. Note that in no case do we try to reduce the number of steps. Formally, a condition of activation is said to be minimal if, for any place $A \in F$, $(F \setminus \{A\}, \mathscr{T})$ is not an activation condition and for any $i < n$ and any $h \in H_i$, $(F,(H_1,\ldots,H_i \setminus \{h\},\ldots,H_{n-1}))$ is not an activation condition. In the model illustrated in Figure 2B, the activation condition required to reach activation of P, $(\{A,B\},[\{h_2\},\{h_3\},\{h_0,h_1\},\{\}])$, is minimal since no place and no event can be discarded to reach the property P. In addition to the concept of minimal activation conditions, we also define *strong activation conditions*. Scenarios with the same initialization condition but different timing of events can either reach or not reach a property P. A strong activation condition leading to P is defined as an activation condition such that there are no overplaying timing sets to not reach the property P. A schedule $\mathscr{T}' = (H'_0,H'_1,\ldots,H'_{n-1})$ overplays the schedule $\mathscr{T}$ if there is at least one step i and an event $h \in H'_i$ which is not an element of $ev(\mathscr{T})$, where $ev(\mathscr{T})$ is the event appearing at least once in one of the H_i. If an activation condition is not strong, it will be considered weak, meaning that there is at least a timing $\mathscr{T}'$ that overplays $\mathscr{T}$ such that P is never reached for the $(F,\mathscr{T}')$ scenario. As shown in Figure 2B, the minimal activation condition $(\{A,B\},[\{h_2\},\{h_3\},\{h_0,h_1\},\{\}])$ is weak since the condition $(\{A,B\},\ [\{h_2,h_4\},\{h_3,h_5\},\{h_0,h_1\},\{\}])$ does not activate place P.

Invariance of a property

Invariance, or the fact that a property P is always satisfied, is a useful and powerful tool to identify biological inhibitors. Identifying inhibitors of a biological property such as gene expression corresponds to the identification of dynamic conditions that lead to the non-activation of the gene: the inactivated state of the place is retained. This means that, using activation conditions for property P, we search for conditions that keep the inactivated place (*not* P) invariant. The same number of steps is always used for both activation and inhibition conditions. As shown in Figure 3A, the strong activation condition $(\{A_0\},[\{h_1\},\{h_2\}])$ has no inhibitor whereas the condition $(\{A_0\},[\{h_1\},\{\},\{\},\{h_2\}])$ has an inhibitor defined as $(\{A_0,B_0\},[\{h_1,h_3\},\{h_4\},\{h_5\},\{h_2\}])$ because the activation of P depends on the event h_2 and the absence of $B3$ but the activation of $B3$ depends on the event h_5 that occurs before h_2.

Similarly, we define minimal inhibition conditions as inhibition conditions that cannot be reduced. We consider that a strong inhibition condition must inhibit the activation condition whatever the unknown of schedules of the model. In the model shown in Figure 3B, the activation condition $(\{A_0\},[\{h_1\},\{h_2\}])$ of P has a strong inhibitor $(\{A_0,B_0\},[\{h_1\},\{h_2\}])$ because of the synchronization of the first two transitions that lead to the activation of $B1$ during the first step, thereby preventing the transition of $A1$ towards P, which requires the absence of $B1$ (*not* $B1$).

Integration of 137 cell signaling pathways into a single unified model

As an extreme test of the power of CADBIOM formalism, we next describe the integration of the entire set of large and complex cell signaling networks extracted from the PID database (41). This results in the creation of a single unified model for cell signaling.

The Pathway Interaction Database (PID) developed by the National Cancer Institute, describes 137 human cellular signaling pathways using homogeneous concepts that allow for automatic and non-

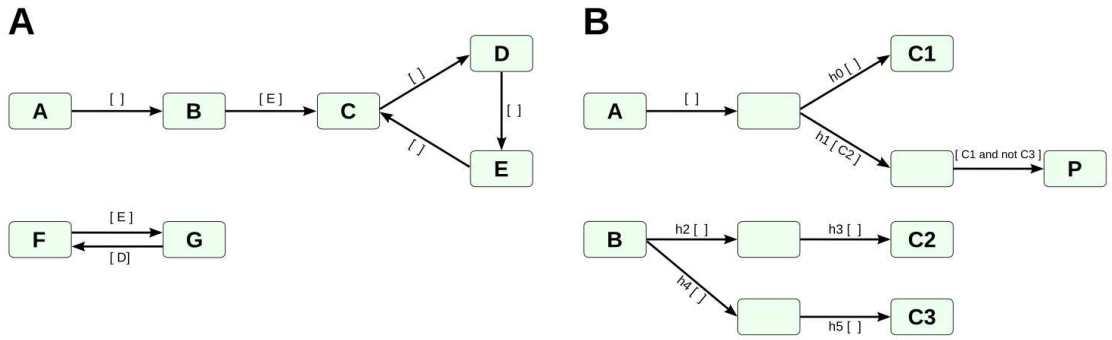


Figure 2: Frontier concept (**A**) and reachability of a property P (**B**). (**A**) Place A belongs to the frontier and at least one place of each of the isolated strongly connected components (C,D,E) and (F,G) belongs to the frontier. (**B**) When the system is initialized by activation of places A and B, the scenario $(\{A,B\},[\{h_2\},\{h_3\},\{h_0,h_1\},\{h_5\}])$ is a solution, that is, conditions that allow reaching property P in 4 steps. In this model, the scenario $(\{A,B\},[\{h_2\},\{h_3\},\{h_0,h_1\},\{\}])$ is a minimal activation condition for the reachability of the place P, in 4 steps. In addition, this activation condition is weak since overplaying the condition $(\{A,B\},\ [\{h_2,h_4\},\{h_3,h_5\},\{h_0,h_1\},\{\}])$ does not activate place P.

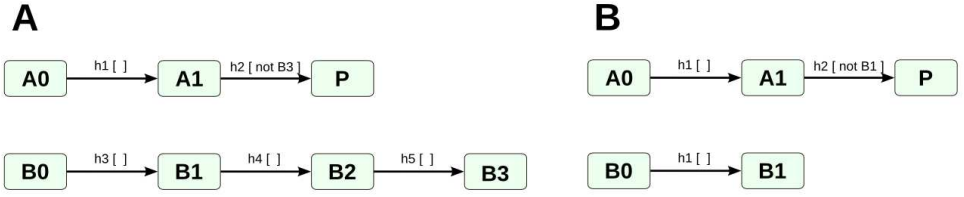


Figure 3: Invariance of a property P. A strong activation condition $(F,\mathscr{T})$ of a property P has an inhibition condition $(I,\mathscr{T}')$ if the scenario $(F\cup I,\ \mathscr{T}')$ keeps $(not\ P)$ invariant and where $\mathscr{T}'$ is a schedule that overplays $\mathscr{T}$. An inhibitor $(I,\mathscr{T}')$ of the activation condition $(F,\mathscr{T})$ is strong if there is no schedule $\mathscr{T}'$ such that P is reached during a scenario resulting from the condition $(F\cup I,\mathscr{T}')$. (**A**) The condition $(\{A_0,B_0\},[\{h_1,h_3\},\{h_4\},\{h_5\},\{h_2\}])$ is an inhibitor of the activation condition $(\{A_0\},[\{h_1\},\{\},\{\},\{h_2\}])$ because the activation of P depends on the event h_2 and the absence of $B3$ $(not\ B3)$, but the activation of $B3$ depends on the event h_5 that occurs before h_2. (**B**) The condition $(\{A_0,B_0\},[\{h_1\},\{h_2\}])$ is a strong minimal inhibitor of the activation condition $(\{A_0\},[\{h_1\},\{h_2\}])$ since $B1$ is always activated at step 1, preventing the transition of $A1$ towards P.

ambiguous translation. Based on the concept of biomolecules and biological reactions, knowledge is formalized so as to permit translation in terms of guarded transitions. We have developed a program that automatically extracts and translates XML files from the PID database into CADBIOM language. The translation scheme is described in Supplementary Figure 1. According to the integrative rules described in Figure 1, we could build a single CADBIOM model for cell signaling (Figure 4). Due to its size, this representation of the transition graph, in and of itself, is not informative. However, the graph can be explored using several tools for viewing, searching and extraction of subgraphs (see manual at `cadbiom.genouest.org`). The 9248 reactions from the 137 PID pathways are integrated into 9264 CADBIOM transitions. Note that transitions and reactions are not equivalent since a reaction for the formation of a complex between two biomolecules must be translated into two transitions. In contrast, two different reactions described in PID and sharing the same inputs and outputs can be translated into one or several CADBIOM transitions. Taken together, these observations show that the number of transitions generated by the translation of pathways depends on the structure of PID reaction graphs. The efficiency of integration is illustrated by the reduction of the 27876 biomolecules from the 137 PID pathways to 9177 non-redundant places from the CADBIOM model. Indeed, one PID biomolecule can be implicated in several reactions and/or pathways, while a CADBIOM biomolecule is represented by a unique place and all of its reactions are either incoming or outgoing transitions. Importantly, the translation and integration of the 137 PID pathways into one CADBIOM model does not alter the distribution of the ontology terms associated with biomolecules (Figure 5).

To further explore the CADBIOM cell signaling model, we next investigate several features of the graph. The analysis of connectivity shows that the largest connected component contains 5340 places (58%). This demonstrates that the inputs and outputs of reactions are well connected in spite of the heterogeneous sources of information that document the 137 pathways. While the average node degree of the transition graph is 2, some biomolecules have much higher degrees, as can be seen for metabolites (GTP: 111; GDP: 98) and signaling proteins (GRB2: 25; beta-catenin: 24). Accordingly, these biomolecules are involved in many reactions as input and/or output, strengthening the non-linear aspect of the signaling networks. An important observation relates to the abundance (43%) of frontier places, suggesting that there exists a huge number of possible combinations for the solutions of reachability and invariance analyses described above.

The model can also be explored by building the *activation graph* (Supplementary Figure 2). Because conditions for guarded transitions influence the output node of the transition, edges are added to the transition graph, connecting condition places to output places. This graph shows all the relations between the biomolecules of the database, not only in terms of information propagation (as in the transition graph), but also in terms of regulation of reactions. Quite spectacularly, information from the PID database is highly connected since 8986 nodes (97%) belong to the largest connected component, suggesting significant cross-talk between signaling pathways.

Taken together, our results demonstrate that CADBIOM can automatically generate a single unified model from 137 independent pathways. Obviously, each PID pathway can be translated separately into a CADBIOM model as shown on `cadbiom.genouest.org`. All generated models are totally editable by adding or removing places and transitions, thereby facilitating the generation of new models and the test of new hypotheses. In conclusion, our CADBIOM model generates the first complete human cell signaling network that integrates all reported pathways. It also yields the first formal view of the complex signaling circuitry that exists in the cell.

Application to cell cycle regulation We next show that the formal verification-based tools of CADBIOM described above can be applied to complex biological questions. This is illustrated here by the analysis of the regulatory pathways involved in control of the cell cycle.

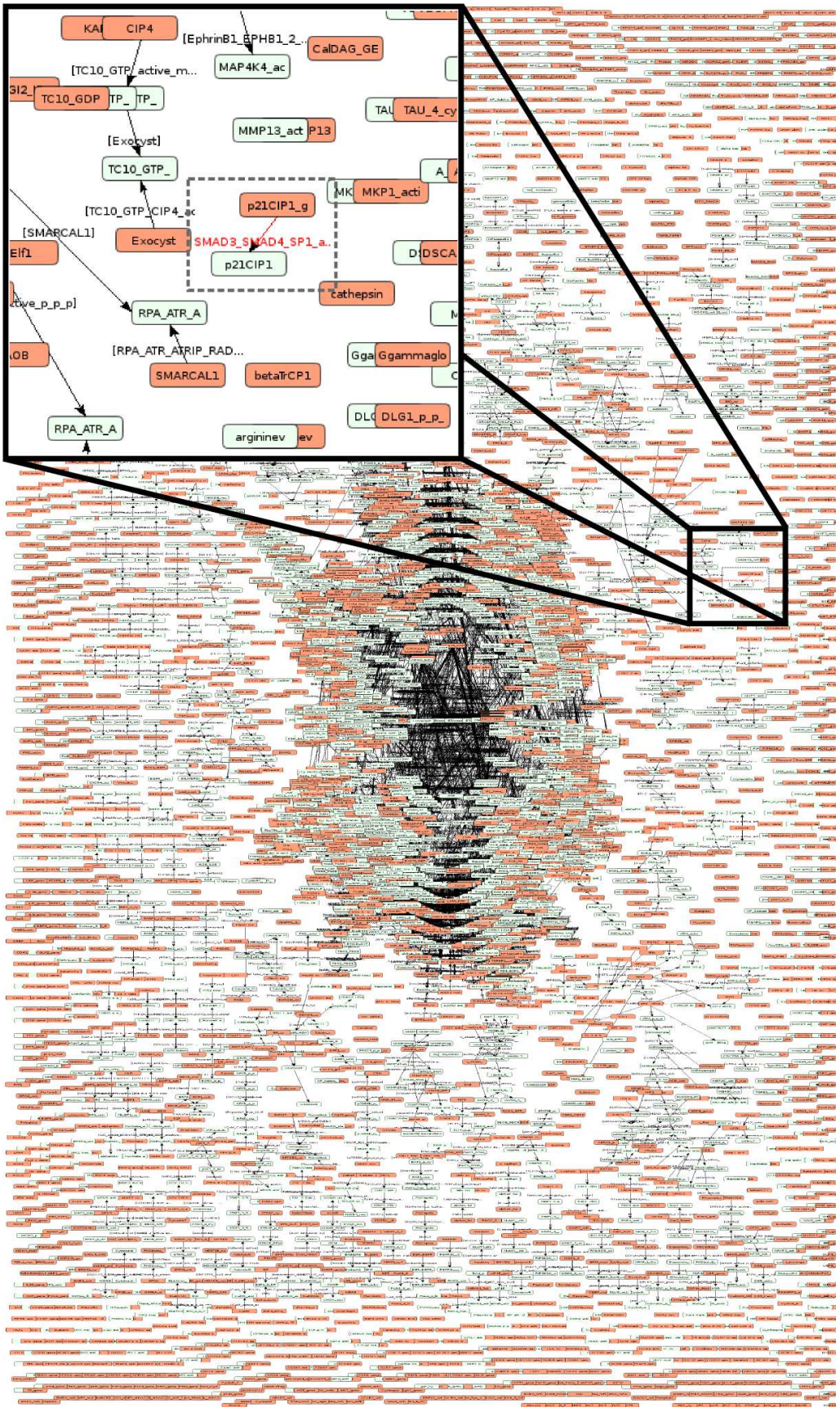


Figure 4: A single unified model of all known human cell signaling pathways.

The mammalian cell cycle is a fundamental physiological process that orchestrates cell growth and division. Loss of cell-cycle control is associated with numerous diseases, including cancer (30; 2). Among the numerous signals that control the cell cycle, the cyclin-dependent kinase inhibitor p21 plays a pivotal role as a sensor of multiple anti-proliferative signals (1) (Figure 6A). How the cell signaling network controls the expression of p21 is a critical question that can be probed using CADBIOM . According to the CADBIOM cell signaling model, p21 is described as a place since it is present as a biomolecule in the PID database (Figure 4). The biological question "how to stop the cell cycle by activating the expression of p21?" is translated into "how to activate the place p21 in the model?" and, more formally, by "what are the activation conditions for the reachability of property $p21$?".

The CADBIOM user interface (check button) proposes a step-by-step formulation of the defined property (Figure 6B). In the example shown, the reachability property is specified with a horizon of 10 steps. Empirical computations (not shown) enforced by graph metrics such as graph diameter of the activation graph show that 10 steps correspond to a reasonable choice. As listed in Supplementary File 1, 35 minimal activation conditions are identified for $p21$ reachability. To facilitate biological interpretation, we have developed methods to group these solutions according to their common frontier places.

Classical clustering methods do not yield significant results (not shown). To overcome this difficulty, we chose to use the frequency of places in solutions and the distance of the places to the target property as new parameters to sort the places within each set of solutions. The places are first ordered from most to least frequent and then from farthest to closest to the property. Methods inspired by automata minimization are then used to discover the places shared by the set of solutions. The result is a Direct Acyclic Graph (DAG) that represents the solutions. A DAG has several roots (nodes without antecedent) and several leaves (nodes without successor) and a path from a root to a leaf represents a solution. Notice that these graphs do not represent pathways but sets of frontier places. Using this method, two major subgraphs are identified from the set of minimal conditions associated with the $p21$ property (Figure 7A). Place analysis clearly demonstrates that these subgraphs are related to the Notch and TGF-β pathways. In this example the different connected DAG components discriminate between different related pathways. However, these results cannot be generalized and, on the contrary, different related pathways could belong to the same connected component. Focusing on the TGF-β part of the graph, we can identify a linear path from SP1 to importin alpha that corresponds to places common to all solutions of the subset, suggesting that they play an essential role in p21 activation. In addition, we show that the plasticity of cell responses can be illustrated by alternative subpaths that share the same predecessors and successors. For example, the complexes of TGFBR1 with either FKBP12 or TRAP1 both mediate TGF-β-dependent pathways as illustrated by the presence of alternative paths after importin alpha. Another advantage of DAG representations is that they can serve to highlight poorly annotated solutions, characterized by small connected components. This is for instance the case for (Miz_1_nucl),(RB1_nucl and MITF_nucl), which directly activate the transcription of p21 without being linked to a PID signaling pathway.

Importantly, each solution described by frontier conditions and event schedule can be used in a simulator to visualize signal propagation. Using one of the 35 minimal solutions described in Figure 7A, the dynamics of the scenario are analyzed through a chronogram (Figure 7B). As a further demonstration of the validity of CADBIOM representations, the main known phases of TGF-β signal propagation are chronologically depicted with great accuracy: activation of TGF-β receptors

Figure 4: The 137 pathways from the PID database containing 27876 biomolecules are integrated into a single unified dynamic model containing 9177 places. Frontier places are colored in orange. The insert focuses on the transition $p21CIP1\_g \rightarrow p21CIP1$, which represents activation of the p21 gene. The condition for the transition contains an extensive logical formula and is not fully displayed.

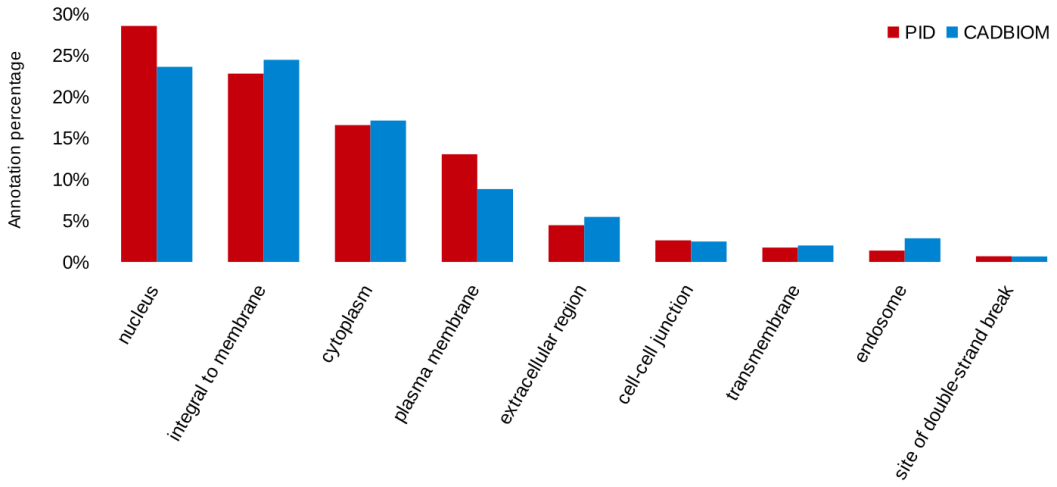


Figure 5: Conservation of biological annotations. The cellular component of GO ontologies of biomolecules (PID) and places (CADBIOM) are analyzed according to their frequencies. The ten most represented location terms in PID and CADBIOM are equally distributed.

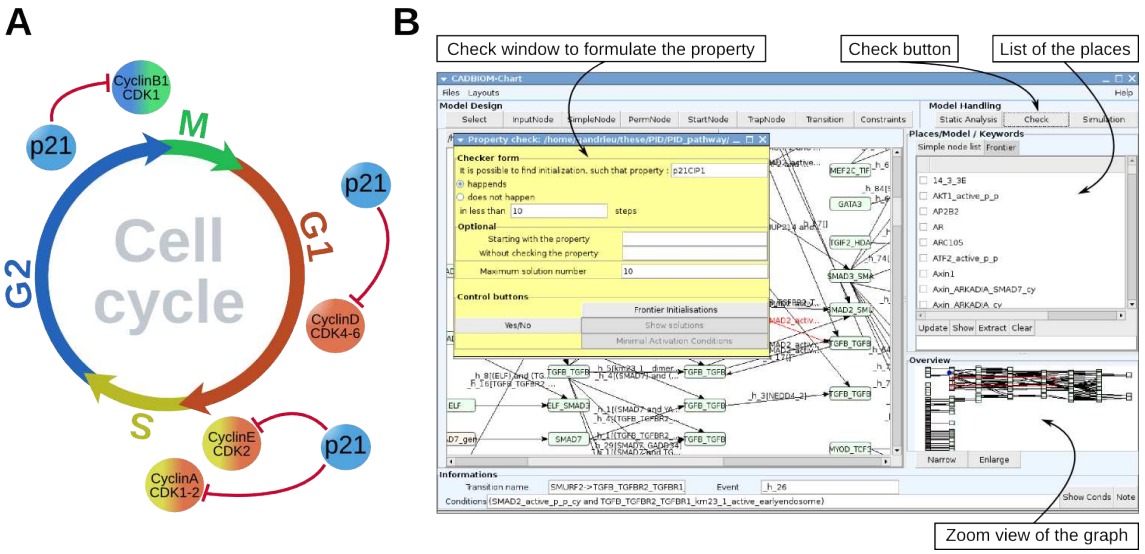


Figure 6: Regulation of p21 expression. (**A**) Schematic representation of p21 effects on cell-cycle phases through cyclin/cdk inhibition (bar-headed lines). M = Mitosis, G1 = Gap1, G2 = Gap2 and S = Synthesis. (**B**) A view of the CADBIOM graphical interface used to check the reachability of p21.

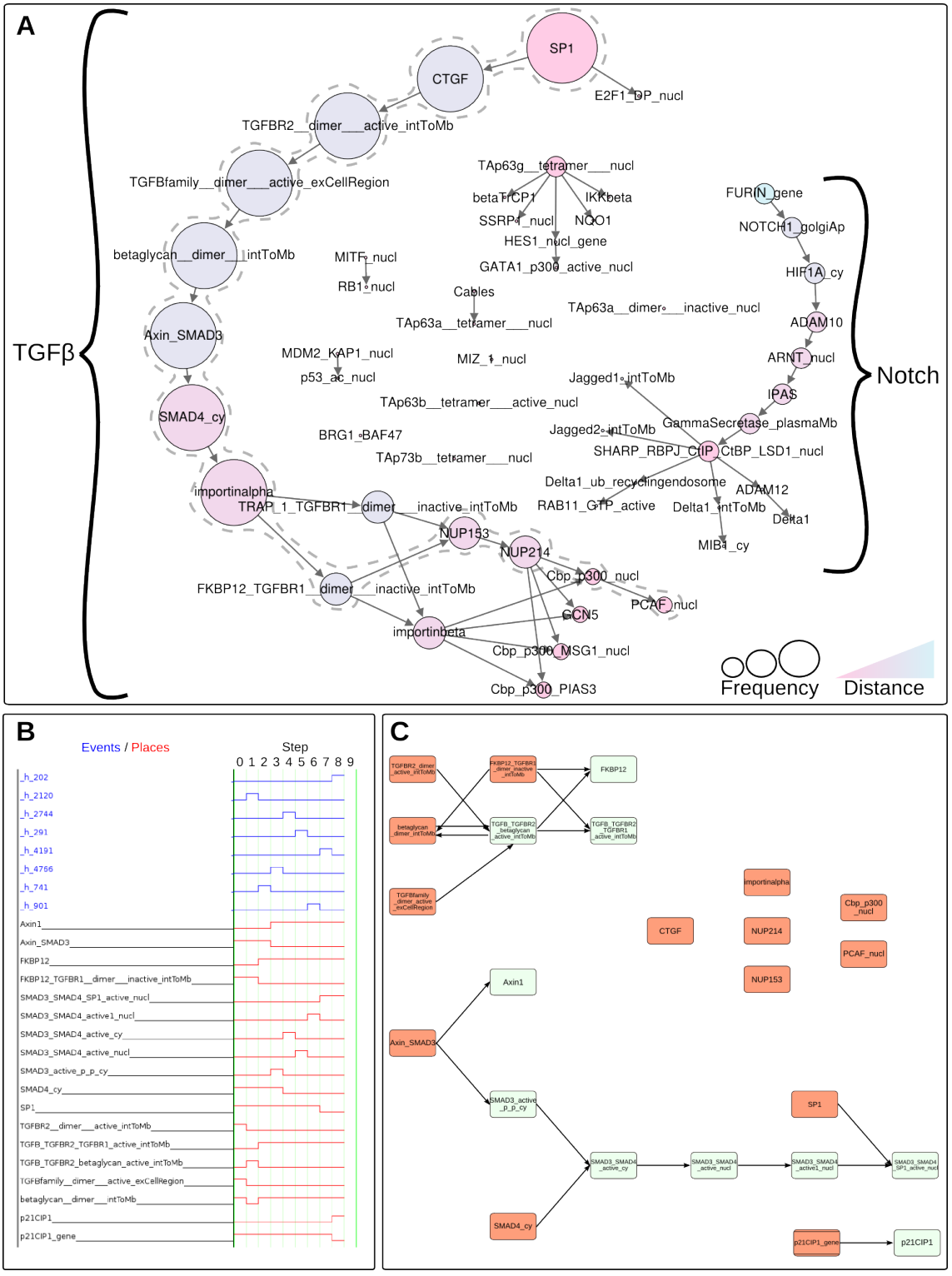


Figure 7: Analysis of solutions for reachability of the *p*21 property. (**A**) DAG representation of the 35 minimal conditions for p21 activation. The size of the nodes depends on the frequencies of places in the set of solutions. The color gradient from red to blue represents increasing distance to the property. (**B** and **C**) A solution delineated by the dotted lines in panel A is further analyzed using simulation studies illustrated by the chronogram (**B**) and the extraction of the corresponding sub-model (**C**). Frontier places (solution places) are depicted in red.

(TGFB_TGFBR2_TGFBR1_betaglycan_active_intToMb at step 2), SMAD activation and complex formation (SMAD3_SMAD4_active_cy at step 4), association with transcription factors (SMAD3_SMAD4_SP1_active_nucl at step 7) and p21 gene regulation (p21CIP1 at step 8). In addition, a sub-guarded transition model can be extracted, which only contains places used during propagation of the signal (Figure 7C). Remarkably, the biomolecules represented by these places are not found in a unique pathway as originally described. Rather, they distribute into 3 different PID pathways, illustrating how the model can be enriched by integrating the whole database.

Not only do these results confirm the known role of the TGFβ pathway in p21 transcriptional regulation (34), they also demonstrate that the Notch pathway is likely to be a positive regulator of cell-cycle arrest. Such a conclusion is of interest since the effect of the Notch pathway on regulation of the cell cycle remains controversial (4).

Because the cell cycle is deregulated in many diseases, searching for putative inhibitors is an important task. In accordance with this objective, we next explored the conditions where p21 is inhibited, potentially leading to the non-inhibition of the cell cycle and of cell proliferation. For this purpose, we first initialized the model in a condition of p21 activation (a solution for p21 reachability) and then searched for places that need to be activated to prevent p21 activation. Note that searching for direct solutions to not activate p21 is not appropriate because only an empty list is returned as a unique minimal solution. Indeed, if no places are activated upon initialization of the system, the signal does not spread and p21 remains inactivated. Searching for the inhibitors of each solution previously identified for the property $p21$ leads to a list of solutions that correspond to places and schedule. In this case as well, DAG representations permit the analysis of these solutions, as illustrated in Supplementary Figure 3. This identifies different groups of inhibitors that differ according to their biological role and shows that the three major potential targets for inhibiting the activation of p21 are SMAD4, SP1 and TGF-β receptors. This result can be rationalized in a simple way. Numerous inhibitors activate SMAD1, 5 or 8, which trap SMAD4, required for TGF-β signaling. Similarly, the (p53, MDM2_KAP1_nucl) complex traps the SP1 transcription factor involved in TGF-β-mediated effects. These observations provide supporting evidence for the negative regulation of signaling through trapping of activators, a mechanism that does not require destruction of proteins. These results illustrate yet another powerful use of CADBIOM : overcoming the complexity of signaling networks to identify potential targets that could be used for new therapeutic strategies.

Discussion

Signaling pathway networks orchestrate all cell life through a unique complex molecular circuitry for signal propagation. Because of the huge number of biological reactions that are implicated, modeling cell signaling to predict cell responses is a major challenge that requires new approaches to aggregate all available information.

We have developed a new formalism based on guarded transitions and combined discrete abstraction to propose, for the first time, a fully integrated cell signaling model. Major issues in modeling biological large-scale phenomena are the collection of information and the identification of a mathematical formalism that can be used to model these data. While numerous databases describe cell signaling pathways, a recent report demonstrated a high degree of inconstancy between databases (22). Based on the Jaccard similarity coefficient, the authors compared four well known pathways involving the cytokines EGF (Epidermal growth factor), TGF-β (Transforming growth factor), TNFα (Tumor necrosis factor) and the signaling protein WNT (wingless-type) in six databases, including GeneGo (`www.genego.com`), KEGG (20), NCI-PID (41), NetPath (19), PANTHER (31) and Reactome (9). Only 10% similarity was found, suggesting that the description of each pathway is database- or even curator-specific. To overcome such a major problem, we sought to extract information from all pathways using one database to integrate them into a single model. Accordingly, we extracted the 137 available curated human pathways

from the PID database. Unlike other database, PID formalizes signal propagation instead of describing biochemical reactions. This means that an interaction is described as a biological event that includes its participating molecules and conditions, an interaction consuming its inputs and producing its outputs (41). Such a formalism is close to that used for guarded transition systems (36), a state/event formalism that can represent both flow circulation with transitions and natural composition rules, and remote influences with transition guards. As reported in this work, we developed an automatic translation of the complete XML-formatted database content into CADBIOM formalism, thereby allowing for the integration of the 137 signaling pathways and automatically creating the first dynamic model of cell signaling. An important point to consider in automatic translation from a database is the continuous upgrading of the generated model. PID upgrading was recently stopped and closing of the PID project, a collaborative effort between the National Cancer Institute (NCI, Besthesda) and the Nature Publishing Group, was announced for the end of 2013. Fortunately, the Reactome project from the EBI consortium shares information with PID, which already imported biological data from Reactome's BioPAX2. While some limitations in the specifications of Reactome's BioPAX2 were previously reported by PID's authors (41), the recent upgrade of BioPAX 2 to BioPAX3 (11) should facilitate data extraction from Reactome into CADBIOM . It will be necessary to complete cell signaling models by adding biological data from the extracellular environment, the major regulator of cell signaling which thereby controls cellular responses (38). To this end, the MatrixDataBase (7), which describes extracellular biomolecule networks, might prove to be a valuable source of new inputs for simulation analysis of CADBIOM models.

Regardless of the biological source, we need to describe the dynamics of signal propagation in a non-ambiguous way. We demonstrate here the usefulness of a state/event formalism based on guarded transitions that differ from traditional ones by the use of an event algebra. The formalisms used in UML (Unified Modeling Language) or in state-charts (17), include an event in the transition guard but lack an operation for combining events into new events. The introduction of the **default** and **when** operators, borrowed from the Signal language (26), supports a clean translation scheme from biological data to guarded transition models. A parallel can be drawn between the way we build transition guards combining the effects of several biological reactions and the way the expression of concentration variations are built by summing up different reaction contributions in a differential model. The **default** operator plays a role similar to that of the *sum* operator in differential systems, and the **when** operator combines conditions for a transition similar to the concentrations combined in a reaction rate.

Handling logical time based on partial ordering of events is standard in computer science (25), but systems biology and computer science lack a universal time reference. In computer science, this is due to the problem of communications and processor speed. In modeling biological systems, this is due to the integration of information from different and disconnected sources. For example, in a differential model, a competition between two reactions $R1$ and $R2$ for a species A is ruled by the difference between the two reaction rates, and refers to a universal time reference, physical time. In discrete models, we ignore the reaction rates and we lose the universal time reference. In Boolean models, this problem is addressed by imposing a uniform asynchronous time model (43). In this approach, only one component of the state is allowed to change at each step of model evolution. A similar approach is applied in Petri net models where one transition is fired at each step. Using CADBIOM formalism, the timing of model evolution is directly linked to the biological reactions by associating an event to each. Consequently, we obtaint a richer modeling framework since simultaneous reactions are not excluded. Importantly, our approach permits the extraction of a reaction pathway from any computed scenario of a CADBIOM model. This greatly facilitates the biological interpretation of computation results. Obviously, for large models, the introduction of events prevents any simulation or state-space exploration. This limitation is also well known for discrete models using asynchronous timing. However, since the timing of reactions is part of the answer to queries, it is always possible to simulate a scenario computed by the CADBIOM checker as demonstrated in Figure 7B.

Numerous modeling methods consider that any variable can be activated at initialization (39). This is

not suitable for signaling models since many molecules depend on upstream signals and have no reason to be activated. In a guarded transition model, it is natural to consider a place without an incoming transition as representing an upstream signal. The frontier concept is derived from this simple notion. Frontier places represent biomolecules that can either be signals such as growth factors at membrane receptors or signal-circuitry components such as cytoplasmic kinases. Importantly, the frontier concept lends itself to the design of more specific models, thanks to the possibility of removing or adding frontier places in the conditions. These conditions can reflect specific biological criteria such as pathological parameters.

Our current single unified CADBIOM cell signaling model is a highly connected graph containing 9177 places and 9264 transitions. The analysis of discrete models has been widely reported in the literature but mainly focuses on the search for steady states (35). Even if such approaches make sense for gene regulatory or metabolic networks, they are not appropriate for the modeling of signal propagation. Based on model-checking methods, we have developed tools to investigate the reachability and invariance of significant properties (8; 32). These analyses usually require the computation of state transition diagrams, which contain approximately 2^{9000} ($\approx 10^{2709}$) states in our model. To avoid graph calculus, we use propositional logic and SAT solver-based approaches that have been demonstrated to be efficient for solving biochemical networks (5).

The verification of a temporal property, as proposed by standard model-checking methods, generally yields a yes/no answer that is not sufficient for studying cell signaling. Instead, we need to find the conditions that lead to a property such as gene activation. The answer to these problems is most often not unique, making interpretation of solutions a daunting task. Matrix representations (40), sometime complemented by clustering methods, are the most common way to display a set of results. Recently, elegant approaches using boolean representation and integer programming have been used to search input sets, as we do, for large signaling networks (14). This approach generates sets of solutions but the readout is difficult to interpret in a biological context without a global analysis. To classify solutions, we propose an original DAG representation based on the frequency of appearance of places in solution sets and on their distance to the property in the activation graph. Using the reachability of the p21 property as a paradigm, we demonstrate that both criteria provide an adequate representation and exhibit a biologically meaningful pattern within a set of solutions.

In conclusion, one major contribution of our work is the creation of a new state-event formalism to integrate biological reactions into a dynamic model for signal propagation. Because it is a rule-based-like language, CADBIOM is likely to be useful for the integration of other models. Another important contribution is the development of a user-friendly interface that is readily accessible to biologists to import or design discrete models and to perform analyses using methods based on model checking. Finally, the creation of the first single unified model of the human cell signaling network, which integrates 137 known signaling pathways, constitutes an important landmark and provides a unique and powerful tool for the exploration of signaling behavior. This new computational approach for modeling signaling networks should improve our understanding of cellular responses to complex stimuli.

Materials and Methods

Methods for solution analysis

The algorithm below is used to order places in a set of solutions according to their relative frequency and the distance to a property.

Algorithm 1 Class MacTree (*placeName*, *listOfMac*)

```
name ← placeName
childList ←new empty list
while listOfMac is not Empty do
  mostFreqPlaces ← get the most frequent places from listOfMac
  place ← get the farthest place to the property in mostFreqPlaces
  newListOfMac ←new empty list
  for mac in listOfMac do
    if place in cam then
      cam ← cam \ {place}
      if cam is not null then
        add cam to newListOfMac
      end if
    else
      add cam to newListOfMac
    end if
  end for
  child ← MacTree(place, newListOfMac)
  add child to childList
  listOfMac ← newListOfMac
end while
```

Compilers

CADBIOM software includes several compilers sharing the same back-end. The front-ends compile PID XML files, CADBIOM XML files and CADLANG files into an intermediate representation of guarded transition models. The common back-end generates logical constraints in propositional clause form. During this step, constant propagation and common sub-expression elimination optimizations are performed. Combined with many peephole optimizations, these techniques allow the back-end to generate reduced sets of constraints that facilitate the work of the SAT solver.

Resources

The cell signaling network is a translation of the PID database content (`http://pid.nci.nih.gov/`). For the complete translation scheme, see Supplementary Information. Model checking analysis is performed using the cryptominisat SAT solver (`http://www.msoos.org/cryptominisat2/`). Graph representations of solutions are displayed using Gephi, an open-source platform for graph visualization (`https://gephi.org/`).

Availability

CADBIOM software is freely available at `cadbiom.genouest.org`.

Supplementary Information

Supplementary Information is available at the *Molecular Systems Biology* website (`www.nature.com/msb`).

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported by the Institut National de la Santé et de la Recherche Medicale (INSERM) and the National Research Agency (ANR). The authors would like to thank Dr. Jérôme Feret (Ecole Normale Supérieure and INRIA Paris-Rocquencourt) and Dr. Emmanuel Käs (CNRS UMR 5099, Toulouse, France) for help in writing this manuscript and the GenOuest Bioinformatics Platform for hosting the CADBIOM website.

Author Contributions

GA conceived the study, developed the software and drafted the manuscript. MLB conceived the study, developed the software and drafted the manuscript. NT conceived the study and drafted the manuscript. All authors have read and approved the final manuscript.

References

[1] T. Abbas and A. Dutta. p21 in cancer: intricate networks and multiple activities. *Nat. Rev. Cancer*, 9(6):400–414, Jun 2009.

[2] E. Bazigou and C. Rallis. Cell signaling and cancer. *Genome Biology*, 8(7):310, 2007.

[3] L. Beltrame, E. Calura, R. R. Popovici, et al. The Biological Connection Markup Language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways. *Bioinformatics*, 27(15):2127–2133, Aug 2011.

[4] S. J. Bray. Notch signalling: a simple pathway becomes complex. *Nat. Rev. Mol. Cell Biol.*, 7(9):678–689, Sep 2006.

[5] M. Carrillo, P. A. Gongora, and D. A. Rosenblueth. An overview of existing modeling tools making use of model checking in the analysis of biochemical networks. *Front Plant Sci*, 3:155, 2012.

[6] C. Chaouiya. Petri net modelling of biological networks. *Briefings in Bioinformatics*, 8(4):210–219, July 2007.

[7] E. Chautard, M. Fatoux-Ardore, L. Ballut, N. Thierry-Mieg, and S. Ricard-Blum. MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Res.*, 39(Database issue):D235–240, Jan 2011.

[8] E.M. Clarke, J. Faeder, C. Langmead, et al. Statistical model checking in biolab: Applications to the automated analysis of t-cell receptor signaling pathway. In Monika Heiner and AdelindeM. Uhrmacher, editors, *Computational Methods in Systems Biology*, volume 5307 of *Lecture Notes in Computer Science*, pages 231–250. Springer Berlin Heidelberg, 2008.

[9] D. Croft, G. O'Kelly, G. Wu, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39(Database issue):D691–697, Jan 2011.

[10] V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. Rule-based modelling of cellular signalling. In *Proceedings of the 18th international conference on Concurrency Theory*, CONCUR'07, pages 17–41, Berlin, Heidelberg, 2007. Springer-Verlag.

[11] E. Demir, M. P. Cary, S. Paley, et al. The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, 28(9):935–942, Sep 2010.

[12] A. Dilek, M. E. Belviranli, and U. Dogrusoz. VISIBIOweb: visualization and layout services for BioPAX pathway models. *Nucleic Acids Res.*, 38(Web Server issue):W150–154, Jul 2010.

[13] J. R. Faeder, M. L. Blinov, and W. S. Hlavacek. Rule-based modeling of biochemical systems with BioNetGen. *Methods Mol. Biol.*, 500:113–167, 2009.

[14] L. G. Fearnley and L. K. Nielsen. PATHLOGIC-S: a scalable Boolean framework for modelling cellular signalling. *PLoS ONE*, 7(8):e41977, 2012.

[15] A. Funahashi, Y. Matsuoka, A. Jouraku, et al. Celldesigner 3.5: A versatile modeling tool for biochemical networks. *Proceedings of the IEEE*, 96(8):1254 –1265, aug. 2008.

[16] C Guerra. Ingenuity pathways analysis: software for discovering and modelling pathways and networks in your systems data. *Comparative Biochemistry and Physiology Part A Molecular Integrative Physiology*, 150(3):S50–S50, 2008.

[17] D. Harel. Statecharts: A visual formalism for complex systems. *Sci. Comput. Program.*, 8(3):231–274, June 1987.

[18] M. Hucka, A. Finney, H. M. Sauro, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, Mar 2003.

[19] K. Kandasamy, S. S. Mohan, R. Raju, et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biol.*, 11(1):R3, 2010.

[20] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, Jan 2000.

[21] S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, 22(3):437–467, Mar 1969.

[22] D. C. Kirouac, J. Saez-Rodriguez, J. Swantek, et al. Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. *BMC Syst Biol*, 6:29, 2012.

[23] H. Kitano, A. Funahashi, Y. Matsuoka, and K. Oda. Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.*, 23(8):961–966, Aug 2005.

[24] K. W. Kohn, M. I. Aladjem, J. N. Weinstein, and Y. Pommier. Molecular interaction maps of bioregulatory networks: a general rubric for systems biology. *Mol. Biol. Cell*, 17(1):1–13, Jan 2006.

[25] L. Lamport. Time, clocks, and the ordering of events in a distributed system. *Communications of the ACM*, 1978.

[26] P. Le Guernic, T. Gautier, M. Le Borgne, and C. Le Maire. Programming real-time applications with signal. *Proceedings of the IEEE*, 79(9):1321–1336, September 1991.

[27] N. Le Novere, M. Hucka, H. Mi, et al. The Systems Biology Graphical Notation. *Nat. Biotechnol.*, 27(8):735–741, Aug 2009.

[28] C. Li, M. Donizelli, N. Rodriguez, et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol*, 4:92, 2010.

[29] D. Machado, R. S. Costa, M. Rocha, et al. Modeling formalisms in Systems Biology. *AMB Express*, 1:45, 2011.

[30] G. S. Martin. Cell signaling and cancer. *Cancer Cell*, 4(3):167 – 174, 2003.

[31] H. Mi, B. Lazareva-Ulitsky, R. Loo, et al. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.*, 33(Database issue):D284–288, Jan 2005.

[32] P. T. Monteiro, D. Ropers, R. Mateescu, A. T. Freitas, and H. de Jong. Temporal logic patterns for querying dynamic models of cellular interaction networks. *Bioinformatics*, 24(16):i227–233, Aug 2008.

[33] Stuart L. Moodie, Anatoly Sorokin, Igor Groyanin, and Peter Ghazal. A Graphical Notation to describe the Logical Interactions of Biological Pathways. *Journal of Integrative Bioinformatics*, 3(2), 2006.

[34] A. Moustakas and D. Kardassis. Regulation of the human p21/WAF1/Cip1 promoter in hepatic cells by functional interactions between Sp1 and Smad family members. *Proc. Natl. Acad. Sci. U.S.A.*, 95(12):6733–6738, Jun 1998.

[35] A. Naldi, D. Berenguier, A. Faure, et al. Logical modelling of regulatory networks with GINsim 2.3. *BioSystems*, 97(2):134–139, Aug 2009.

[36] A. Rauzy. Guarded transition systems: a new states/events formalism for reliability studies. *Journal of Risk and Reliability*, 222(4):495–505, 2008.

[37] S. Raza, K. A. Robertson, P. A. Lacaze, et al. A logic-based diagram of signalling pathways central to macrophage activation. *BMC Syst Biol*, 2:36, 2008.

[38] T. Rozario and D. W. DeSimone. The extracellular matrix in development and morphogenesis: a dynamic view. *Dev. Biol.*, 341(1):126–140, May 2010.

[39] D. Ruths, M. Muller, J. T. Tseng, L. Nakhleh, and P. T. Ram. The signaling petri net-based simulator: a non-parametric strategy for characterizing the dynamics of cell-specific signaling networks. *PLoS Comput. Biol.*, 4(2):e1000005, Feb 2008.

[40] R. Samaga, J. Saez-Rodriguez, L. G. Alexopoulos, P. K. Sorger, and S. Klamt. The logic of EGFR/ErbB signaling: theoretical properties and analysis of high-throughput data. *PLoS Comput. Biol.*, 5(8):e1000438, Aug 2009.

[41] C. F. Schaefer, K. Anthony, S. Krupa, et al. PID: the Pathway Interaction Database. *Nucleic Acids Res.*, 37(Database issue):D674–679, Jan 2009.

[42] P. K. Sreenivasaiah, S. Rani, J. Cayetano, N. Arul, and d. o. H. Kim. IPAVS: Integrated Pathway Resources, Analysis and Visualization System. *Nucleic Acids Res.*, 40(Database issue):D803–808, Jan 2012.

[43] R. Thomas. Regulatory networks seen as asynchronous automata: A logical description. *Journal of Theoretical Biology*, 153(1):1–23, November 1991.

Supplementary Information

Geoffroy Andrieux, Michel Le Borgne and Nathalie Théret

Table of contents for Supplementary Information:

Supplementary method: Mathematical semantics of guarded transitions

Events and states

Event realization

An event is the mathematical concept that denotes a from-time-to-time occurrence. In the absence of a universal time reference, an event alone is of no interest. It requires at least another event to exist, such as an observer or a clock.

An event has a name h and a singleton domain to denote the occurrence of the event. We define $\top$ as the unique element of the event domain[1]. Occurrences of different events are equivalent as far as the domain is concerned. It is comparable to the "tick" of a clock. In general, when considering only time, we don't make any distinction between the sounds emitted by different clocks. Accordingly, it is legitimate to consider that all events have the same value.

Definition 0.1
A realization of a finite set of events $(h_i)_{i\in I}$ is a sequence of elements of $\{\top,\bot\}^I \setminus \{\bot^I\}$.

The sequence may be finite or infinite. The symbol $\bot$ denotes the absence of an event relative to another event. The simultaneous absence of all the events cannot occur since at least one observer must be present to mark this fact. For this reason, the multiple $\bot^I$ is excluded. The following example shows a realization of the events e_1, e_2 and e_3:

$$\begin{array}{lccccccc} \mathbf{e}_1: & \top & \top & \bot & \bot & \top & \top & \top & \dots \\ \mathbf{e}_2: & \bot & \top & \top & \top & \bot & \bot & \top & \dots \\ \mathbf{e}_3: & \bot & \bot & \top & \bot & \bot & \bot & \top & \dots \end{array}$$

State variables

An event is well adapted to model a transient phenomenon. A new type of variable is required to model something that is persistent. This type of variable is called a **state variable**. The domain of values for state variables is any useful domain. For guarded transition systems, we will consider boolean and finite domains.

In a realization, the value of a state variable may change. However and contrary to events, a state variable is always available. At any step or for any index of the realization, the value of a state variable is either *True* or *False*.

The following example shows a realization of the events e_1, e_2 and the state variable A:

$$\begin{array}{lccccccc} \mathbf{e}_1: & \top & \top & \bot & \bot & \top & \top & \top & \dots \\ \mathbf{e}_2: & \bot & \top & \top & \top & \bot & \bot & \top & \dots \\ \mathbf{A}: & F & F & F & F & T & T & T & \dots \end{array}$$

where A is a state variable and F and T stand for false and true, respectively.

This leads to the following definition:

Definition 0.2
A realization of a finite set of events and states $(h_i)_{i\in I}, (S_j)_{j\in J}$ is a finite or infinite sequence of elements of $(\{\top,\bot\}^I \setminus \{\bot^I\}) \times \{T,F\}^J$.

[1] $\top$ must not be confused with *True* although it will be assimilated to this boolean later on.

Basic operations on events and states

In many models, we need to combine events and, more generally, events and states. Operations on events are well known in computer science. The two basic operators are a merge that corresponds to multiplexing and a selection of occurrences corresponds to under-sampling. For CADBIOM we have borrowed the **default** operator and the **when** operator from the SIGNAL language (Le Guernic *et al*, 1991) with semantics close to SIGNAL semantics.

The default operator

The **default** operator merges the two events h_1 and h_2 or, more precisely, the occurrences of the events in any realization . In the following example, we assume that there are more events. This legitimizes the case where h_1 and h_2 are simultaneously absent.

$$\begin{array}{rccccccccc} \mathbf{h_1}: & \top & \top & \bot & \bot & \top & \bot & \top & \dots \\ \mathbf{h_2}: & \bot & \top & \top & \top & \bot & \bot & \top & \dots \\ \mathbf{h_1\ default\ h_2}: & \top & \top & \top & \top & \top & \bot & \top & \dots \end{array}$$

This operator on events is commutative.

For the mathematically inclined reader, we provide a rigorous definition. We first define an operator $\uparrow$ on $\{\top,\bot\}$ by following the rules:

$$\begin{array}{rcl} \top\uparrow\top & = & \top \\ \top\uparrow\bot & = & \top \\ \bot\uparrow\top & = & \top \\ \bot\uparrow\bot & = & \bot \end{array}$$

We then define **default** by its semantics:

Definition 0.3
*Given two events h_1 and h_2, $h = h_1$ **default** h_2 if and only if, for any realizations s of (h_1,h_2,h), the relation $s_i^h = s1_i^{h_1} \uparrow s2_i^{h_2}$ is satisfied.*

The when operator The **when** operator is an operator between events and logical combinations of state variables. Since state variables have a boolean domain, it is possible to write propositional logic formulas with state variables. At each instant of a realization, the formula can be evaluated since state variables always have values.

The **when** operator selects occurrences of an event when the propositional formula evaluates to *True* on its right hand side. For example:

$$\begin{array}{rccccccccc} \mathbf{h_1}: & \top & \top & \bot & \bot & \top & \top & \top & \dots \\ \mathbf{B}: & F & T & F & T & F & T & F & \dots \\ \mathbf{h_1\ when\ B}: & \bot & \top & \bot & \bot & \bot & \top & \bot & \dots \end{array}$$

where B is a state variable.

For a more mathematical definition, we need to introduce the operator $\downarrow$ with the rules:

$$\begin{array}{rcl} \top\downarrow True & = & \top \\ \top\downarrow False & = & \bot \\ \bot\downarrow True & = & \bot \\ \bot\downarrow False & = & \bot \end{array}$$

We then define **when** by its semantics:

Definition 0.4
Given an event h_1 and a state propositional formula $B(X)$ where X represents a multiple of state variables, $h = h_1$ **when** *$B(X)$ if and only if for all realizations s of (h_1, X, h), the relation $s_i^h = s1_i^{h_1} \downarrow B(x_i)$ is satisfied.*

The logical operators $\vee$, $\wedge$ and $\neg$ are implemented in CADBIOM . They can be used in the condition part of a transition guard with place names as state variable names. The **default** and **when** operators are also implemented and can be used in the event of a transition guard. The right hand side operand of a **when** must be a propositional formula with state variables (and $True, False$ constants).

Extension to signals

The event concept naturally generalizes to the signal concept. Signals are useful to define the mathematical semantic of guarded transition-based models. With the extension of events to signals, it becomes possible to write the evolution equation of the dynamic system into relatively simple mathematical formulas.

A signal is a generalization of an event. Contrary to an event, a signal can have an arbitrary domain of values. An event is simply a signal with $\{\top\}$ as domain. Given a family $(Z_i)_{i\in I}$ of signals with domains $(D_i)_{i\in I}$, a realization is a sequence of multiples $(z_i^n)_{i\in I}$ such that $z_i^n \in D_i \cup \{\bot\}$. Again, the multiple $\bot^I$ is excluded.

Extending the **default** operator to signals, we face a new problem. If, in a realization of Z_1 and Z_2, we have for some instant $z_1 \neq \bot$ and $z_2 \neq \bot$, we have to decide which value to choose. Here, we decided to keep the value of the left hand side operand. Note that, in general, the domain of a signal built with the default operator is the union of the domains of the operands. In most cases, **default** is used only with operand signals that have the same domain. We will only consider boolean domains.

The domain of an event is an arbitrary singleton. To combine events and boolean signals in a mathematical formula, we will identify the $\top$ value with $True$. With this identification, an event is assimilated to a boolean signal with constant value $True$. It is also possible to extend boolean operations to boolean signals with different semantics. Since general operations are not required on boolean signals, they will not be discussed and only the negation of an event which is assimilated to a boolean signal with constant value $False$ need be considered.

A state variable can be assimilated to a signal. In any realization and at any instant, the value associated with a state variable is different from $\bot$. A state is present in any realization, it acts as memory in a computer. A state, an event and the negation of an event are limit cases of signals. They are either always present (state) or always have the same value (event or event negation). This special feature is essential to obtain a simple encoding of guarded-transition model dynamics into logical clauses.

Semantic of guarded transitions

The firing of a transition is an event. Given a guarded transition $A \xrightarrow{h[C]} B$, we define a new event, called the transition event, as:

$$h_{tr} = h \textbf{ when } (A \wedge C)$$

The transition event h_{tr} is present if and only if the three conditions for transition firing are satisfied.

If we focus on the evolution of the source and target places when a transition is fired, informal semantics state that:

- the source is inactivated
- the target is activated

The evolution function of a state B relies on the current value B_k at step k to the next value B_{k+1} at step $k+1$. Traditionally the current value is denoted as B and the next value B'. With this notation, the evolution of the target, upon possible firing of one transition, is described by:

$$B' \;=\; h_{tr} \;\textbf{default}\; B$$

which must be interpreted with the extension of the operator **default** to signals. The state variables are considered as signals and the events are considered as signals with value $True$. The priority of the left operand of the **default** operator is essential for defining correct semantics. When the transition is fired (h_{tr} is present), the state takes the value of the **default** left hand side operator ($True$) regardless of the preceding value. When it isn't fired, the value of the state remains unchanged.

Using the same conventional notations, the evolution of the source is formalized by:

$$A' \;=\; \neg h_{tr} \;\textbf{default}\; A$$

reflecting the inactivation of the source when the transition is fired. Again, we use the extensions introduced at the end of the preceding section.

In general, several transitions are adjacent to a place. For a place, we call an in-transition a transition having the place as target. The set of in-transitions of a place A is denoted $T_{in}(A)$. A transition having the place as source will be called an out-transition and the set of out-transitions of a place A is denoted $T_{out}(A)$. When there is no ambiguity, the name of the place is dropped.

The state of a place changes if either an in-transition or an out-transition is fired. In case of simultaneous firing we apply the rule:

- activation prevails overs inactivation.

We define two events associated with a place A by:

$$\begin{array}{rcl} h_{in} & = & \textbf{default}_{\,t_r \in T_{in}} h_{tr} \\ h_{out} & = & \textbf{default}_{\,t_r \in T_{out}} h_{tr} \end{array}$$

The h_{in} event is present if at least one of the in-transitions is fired. The h_{out} event is present if at least one of the out-transitions is fired. The mathematical formalization of the rules is given by:

$$A' \;=\; (h_{in} \;\textbf{default}\; \neg h_{out}) \;\textbf{default}\; A$$

The semantics of **default** on signals are again essential to obtain a correct formalization of the priority rule.

Guarded transition systems

With the rigorous definition of the semantics of a transition, it is straightforward to verify that the two guarded transitions $A \overset{h[C]}{\rightarrow} B$ and $A \overset{(h \textbf{when } C)[]}{\longrightarrow} B$ are semantically equivalent. They induce the same dynamics on the state variables. Taking advantage of this, we will consider guarded transitions without condition.

Definition 0.5
A guarded transition system (GTS) is a triplet $(\mathscr{P},\ \mathscr{H},\ \mathscr{T})$ where:

- *$\mathscr{P}$ is a finite set of places*
- *$\mathscr{H}$ is a finite set of free events*
- *$\mathscr{T}$ is a finite set of triplets (A,B,h) where $A,B \in \mathscr{P}$ and h is an event built on $\mathscr{P}$ and $\mathscr{H}$*

The state of a guard transition system is a boolean multiple in $\{T,F\}^{|\mathscr{P}|}$. This state is equivalently described by the set $\mathscr{A}$ of activated places. The evolution of the state depends also on the events H present at the current evolution step. This motivates the following definition:

Definition 0.6
A configuration of a guarded transition system $(\mathscr{P},\ \mathscr{H},\ \mathscr{T})$ *is a couple* $(\mathscr{A},H)$ *where* $\mathscr{A}$ *is a subset of* $\mathscr{P}$ *representing activated places, and* H *is a subset of free events which are present at the evolution step.*

Because of the semantics of guarded transitions, the evolution of a guarded transition system is completely determined by the initial activated places and a sequence of free events.

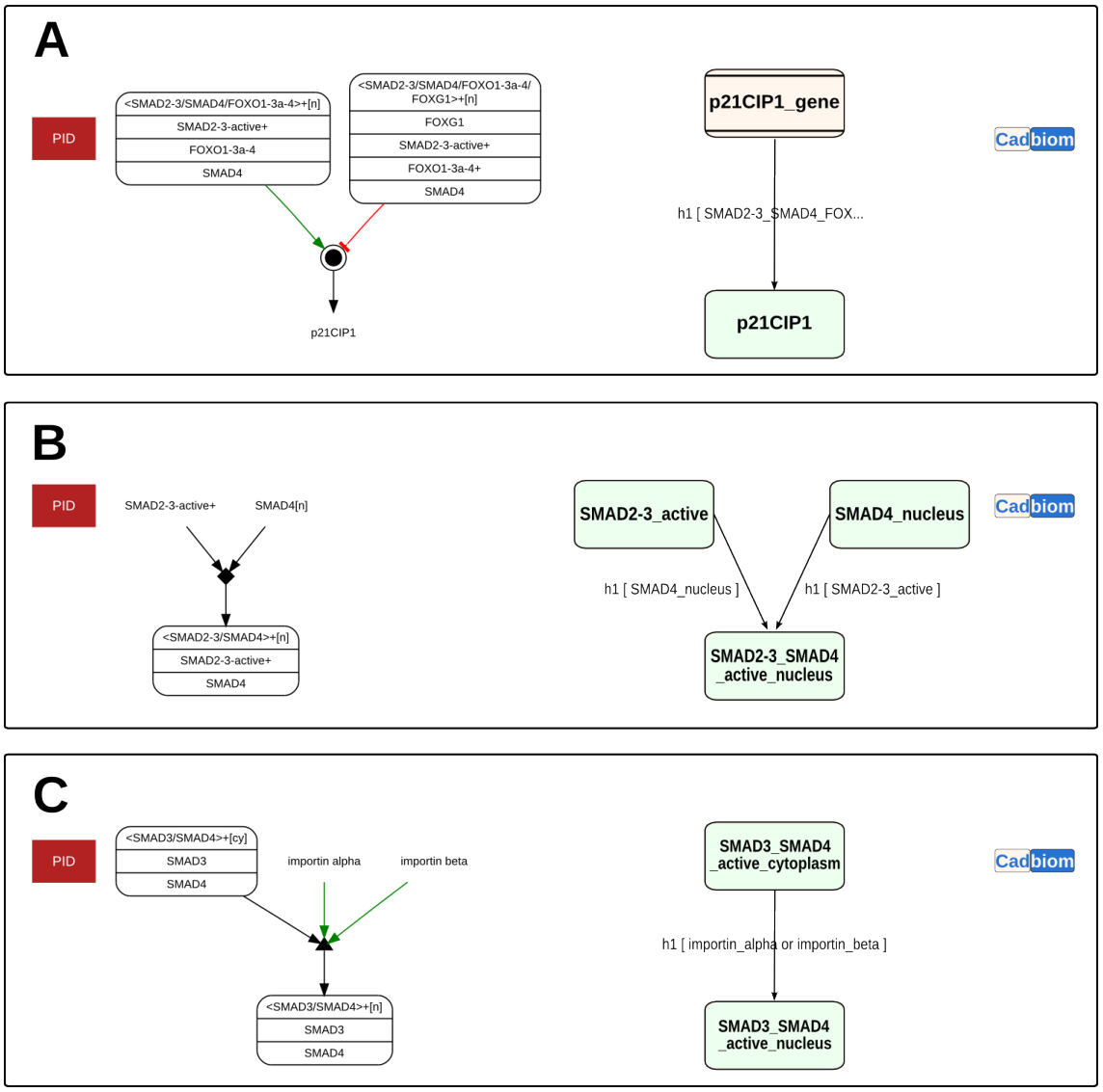
Definition 0.7
A scenario is a couple $(\mathscr{A}_0,(\mathscr{E}_0,\mathscr{E}_1,\ldots,\mathscr{E}_{n-1}))$ *where*

- $\mathscr{A}_0$ *is the set of activated places at initialization*
- $(\mathscr{E}_0,\mathscr{E}_1,\ldots,\mathscr{E}_{n-1})$ *is a sequence of subset of free events.*

n *is the length or the horizon of the scenario.*

A sequence of configurations $(\mathscr{A}_k,\mathscr{E}_k)$ is associated with a scenario where $\mathscr{A}_k$ is the set of activated places after the kth step. $\mathscr{A}_n$ is the set of activated places reached by the scenario. A state property represented by a logical formula f on places is reached by the scenario if for some $k \leq n$, $\mathscr{A}_k$ is a model of the formula f.

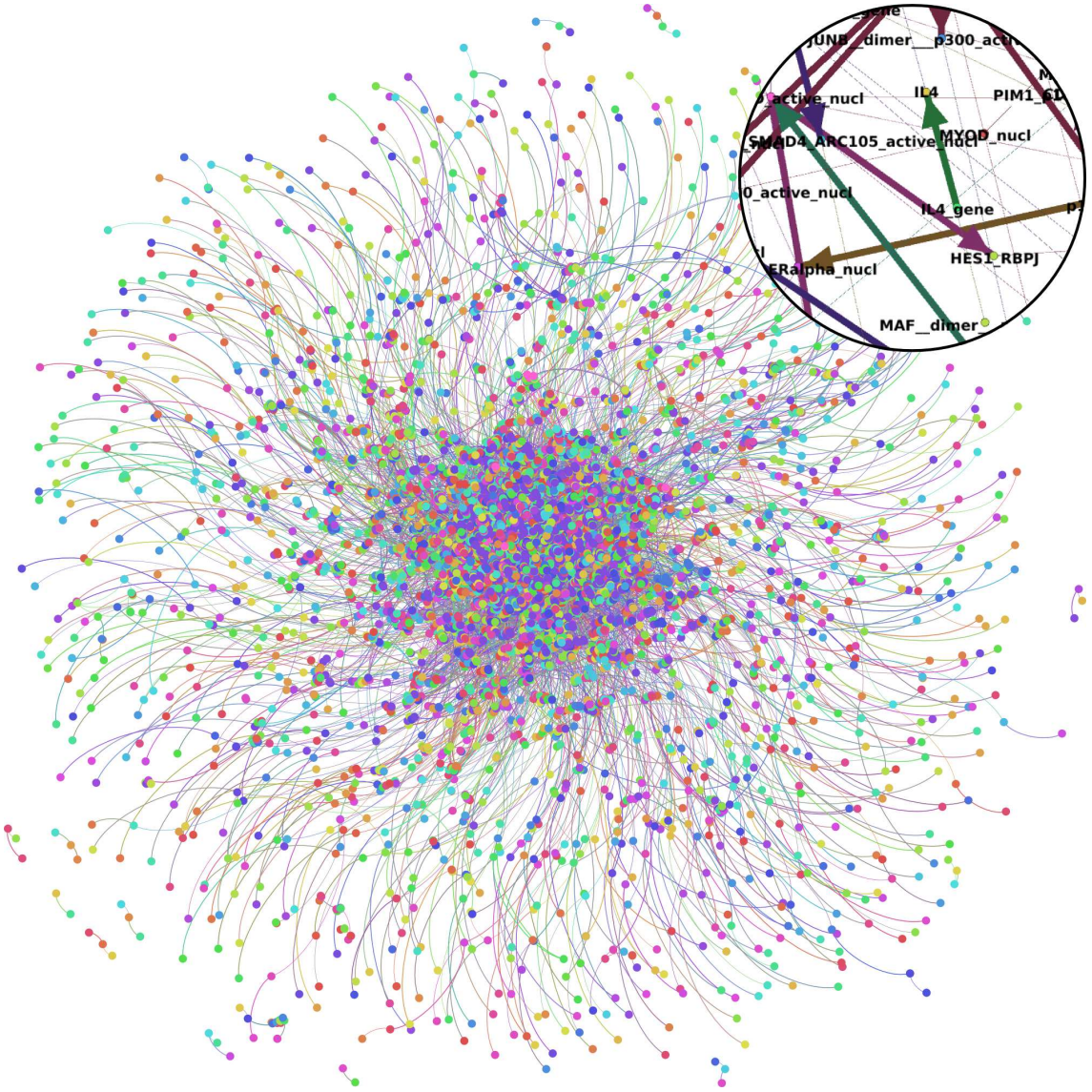
Supplementary Figure 1: PID translation



Supplementary Figure 1: PID translation
Elementary translation of the three categories of biological reactions from the PID database (left) into CADBIOM representation (right). (**A**) Transcription. The graphical representation of p21CIP1 transcription in the PID database is described by a dark circle for the transcription process and the p21CIP1 name for the protein. The different transcription factors are listed in boxes with green arrow for the activators and red bar-headed lines for the inhibitors. The CADBIOM representation uses a permanent place for the gene (salmon rectangle) linked by an arrow to the protein place (green rectangle). The activators and inhibitors are merged in the guard of the transition (written across the arrow) as a logical formula (SMAD23_SMAD4_FOXO1-3a-4_nucleus and not (SMAD23_SMAD4_FOXO1-3a-4_FOXG1_nucleus) . The event $h1$ indicates the timing of the transition. (**B**) Modification (interaction). The graphical representation of interactions between SMAD2-3-active+ and SMAD4[n] in the PID database is denoted by a black diamond for the interaction process with the two components as input and by a box with the list of components and the complex as output (black arrows). The CADBIOM representation uses three places for the two components and the complex. Each transition from a component towards

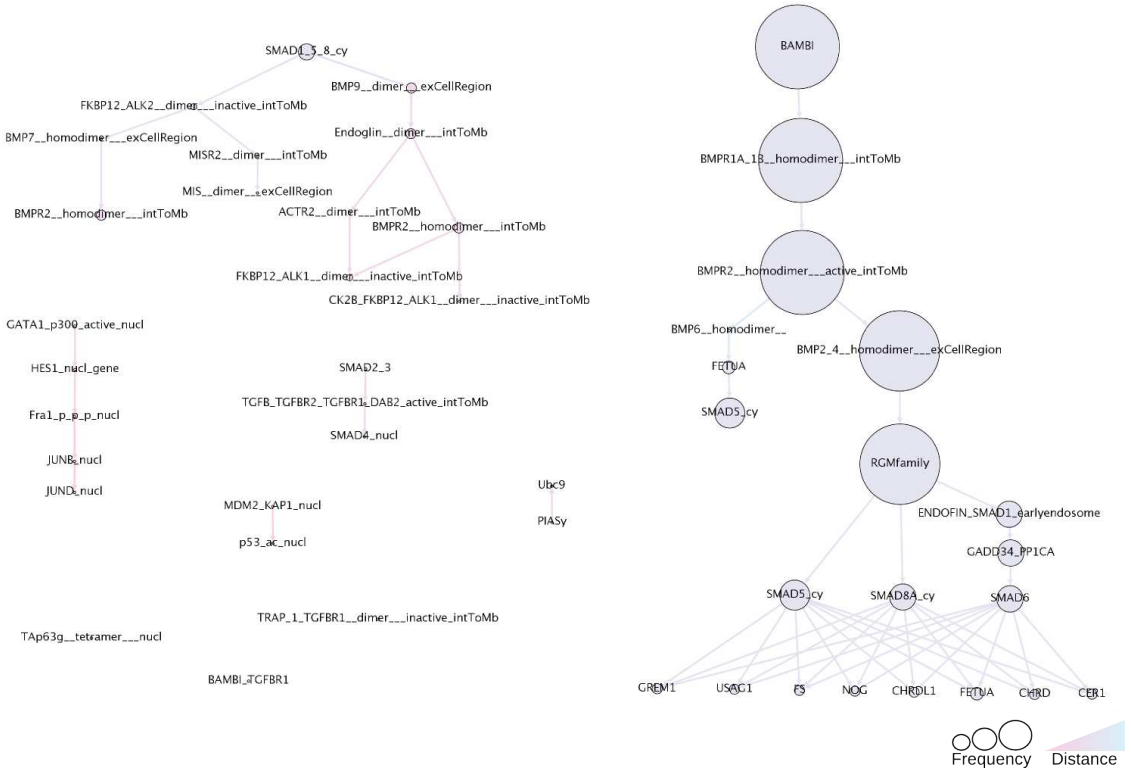
the complex is conditioned by the presence of the other component (formalized in the guard), leading to the synchronization of the transition. Note that the same event $h1$ is associated with the two transitions that occur simultaneously. (**C**) Translocation. The PID graphical representation of translocation of the complex SMAD3_SMAD4 from the cytoplasm to the nucleus is described by a black triangle for the translocation process with the complex SMAD3/SMAD4 +[cy] as input and the complex SMAD3/SMAD4 +[n] as output. The role of the two activators, importin alpha and importin beta, is indicated by green arrows. The CADBIOM representation is described by a transition between the two places SMAD3_SMAD4_active_cytoplasm and SMAD3_SMAD4_active_nucleus conditioned by the presence of the importins (formalized in the guard using the operand "or"). The event $h1$ indicates the timing of the transition.

Supplementary Figure 2: Activation graph of the cell signaling model



Supplementary Figure 2: Activation graph of the cell signaling model
Activation graph obtained from the CADBIOM model of the entire signaling network. The 137 PID pathways are integrated into a single unified model according to the CADBIOM rules. The activation graph is built using places as nodes and the dependencies between places as edges. The dependency relationships include the transitions and the influence of conditions on the output of transitions. The resulting activation graph contains 9077 nodes and 15499 edges. The intensity of the coloration is associated with the connectivity of components. In the zoom view (insert), thin edges denote condition dependencies and thick edges transition dependencies.

Supplementary Figure 3: DAG representation of inhibitor sets



Supplementary Figure 3: DAG representation of inhibitor sets
DAG representation of inhibitor sets for the p21 property. The reachability analysis is performed for the p21 property and 35 minimal activation conditions are identified (listed in Supplementary File 1). Using the solution ($NUP153$ *and* $TGFBfamily\_dimer\_active\_exCellRegion$ *and* $NUP214$ *and* $Axin\_SMAD3$ *and* $betaglycan\_dimer\_intToMb$ and $GCN5$ *and* $p21CIP1\_gene$ *and* $SP1$ *and* $CTGF$ *and* $FKBP12\_TGFBR1\_dimer\_inactive\_intToMb$ *and* $TGFBR2\_dimer\_active\_intToMb$ *and* $SMAD4\_cy$ *and* $importinalpha$) as the condition to initialize the model, we search for the invariant property $notp21$. 37 minimal inhibition conditions are identified and used for DAG representation according to the frequency of places and their distance to the property. Nine distinct DAGs illustrate the solutions.

References

Le Guernic P, Gautier T, Le Borgne M, Le Maire C (1991) Programming Real-Time Applications with Signal. *Proceedings of the IEEE* **79**: 1321–1336

Chapitre 3

Résultats supplémentaires

Ce chapitre est consacré aux résultats obtenus autour du formalisme Cadbiom en complément de l'article présenté dans le chapitre précédent. Dans ce chapitre, nous allons détailler les 3 études que nous avons réalisé pour (i) caractériser les profils de signalisation impliqués dans la régulation des 787 gènes de PID, (ii) utiliser la base de données Reactome comme source de données pour le modèle de signalisation cellulaire, (iii) affiner les connaissances sur nos modèles grâce à une collaboration avec les auteurs du formalisme du Process Hitting.

3.1 Analyse des profils de signalisation impliqués dans l'expression des gènes dans le modèle Cadbiom issu de PID

En générant un modèle de la signalisation cellulaire à partir du PID, nous avons accès à toutes les connaissances actuelles de cette base dans un système dynamique. Dans notre article nous avons illustré l'utilisation de notre modèle en étudiant la régulation du gène p21CIP1, inhibiteur du cycle cellulaire. La recherche des solutions minimales conduisant à l'activation des gènes peut être étendue à l'ensemble des gènes du modèle.

Dans ce type d'analyse haut débit nous recherchons les conditions d'activation de chacun des gènes présents dans notre modèle afin de comparer les solutions obtenues et identifier des groupes de gènes présentant des signaux de régulations similaires. En comparant nos résultats avec des données expérimentales d'expression, nous évaluerons la corrélation entre la co-expression des gènes et des mécanismes de signalisation similaires. Cette analyse permettra en outre d'explorer le contenu du modèle, à savoir la qualité de l'annotation décrivant la régulation des gènes.

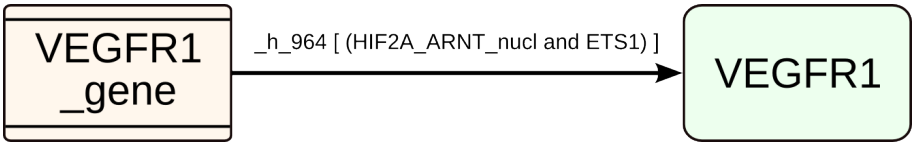


FIG. 3.1: Représentation de l'expression d'un gène dans un modèle CADBIOM. Le gène VEGFR1 est représenté par une place permanente (qui n'est jamais désactivée) $VEGFR1\_gene$. Cette place possède une transition sortante portant le nom de la protéine exprimée $VEGFR1$. L'événement $\_h\_964$ est spécifique de cette transition, et la condition est composée des régulateurs de la transcription du gène. Dans cet exemple l'activation de la place $VEGFR1$ ne peut se faire qu'en présence de $HIF2A\_ARNT\_nucl$ et de $ETS1$.

3.1.1 Description du modèle

Le modèle est le même que celui utilisé dans le chapitre précédent. Il a été généré à partir de la dernière version de la base de données PID datant du 23 Septembre 2012. Ce modèle comporte 787 places annotées par le terme "gene", et possédant une transition sortante vers la protéine synthétisée par le gène en question comme illustré en Figure 3.1. Ces 787 places représentent 744 gènes codant pour des protéines différentes d'après leur symbole HUGO. Cette différence s'explique par le fait que certaines places "gene" peuvent faire référence à la même protéine mais annotée différemment. Par exemple, les places $IL4\_exCellRegion$ et $IL4$ représentent toutes deux la protéine IL4, de même que pour $EGR1\_nucl$ et $EGR1$ qui font référence à EGR1. Puisque nous n'avons pas d'information supplémentaire sur ces places, nous recherchons les solutions des 787 et nous commenterons les différences obtenues pour les doublons.

Notre protocole expérimental vise à rechercher toutes les solutions minimales permettant l'activation des 787 gènes indépendamment les un des autres. Chaque gène code pour une protéine, et est représenté par une transition $G \rightarrow P$ où G représente le gène et P la protéine codée par le gène. Nous recherchons alors les scénario (place frontières et timing d'événements) qui vérifient l'atteignabilité de P en un maximum de 10 pas de temps. L'ensemble de solutions de chaque protéine est ensuite analysé.

3.1.2 Statistique des résultats

Les 787 protéines codées par les gènes ne sont pas toute atteignables en 10 pas de temps. Une partie d'entre elle, soit 108 n'ont obtenu aucune solution. L'absence de solution peut s'expliquer par le nombre de pas de temps trop faible pour permettre l'atteignabilité de la place représentant la protéine. Une autre possibilité est une incohérence dans le modèle

comme illustrée dans la Figure 3.2.

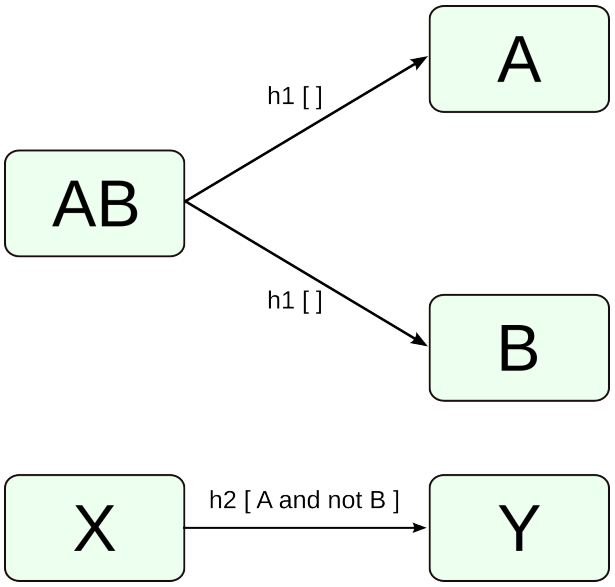


FIG. 3.2: Réactions incohérentes. L'atteignabilité de la place Y n'admet aucune solution car la formule logique conditionnant la transition $X \rightarrow Y$ ne peut être vérifiée. Quelque soit le timing des événements, il n'existe aucune initialisation des places frontière permettant d'atteindre A sans atteindre B, puisque ces deux places proviennent de la décompléxation de AB.

Sur l'ensemble de gènes possédant au moins une solution, nous avons d'abord analysé la répartition du nombre de solutions minimales par protéine (Figure 3.3). De façon surprenante un nombre important de protéines possèdent plus d'un millier de solutions minimales. En effet 99 protéines soit 14% des protéines codées par des gènes ont une régulation extrêmement complexe. Nul doute que les annotations de la base PID n'ont pas détaillé 1000 voies de signalisation. Ce nombre de solution observées résulte de la combinatoire des différentes réactions, dont l'information à été recherchée localement par les curateurs. L'agrégation de ces réactions dans un unique système dynamique entraîne une augmentation de la complexité intrinsèque du modèle.

Nous avons ensuite calculé la répartition des tailles des solutions, en terme de places frontières à activer. Chaque propriété ayant un ensemble de solution, nous avons considéré pour chaque propriété la taille de la plus grande solution (Figure 3.4). Une proportion importante de propriétés possède des solutions de grande taille contenant près de 50 places. Ces résultats reflètent là encore l'extrême complexité du réseaux de signalisation régulant ces gènes.

Pour analyser le contenu de ces solutions, nous avons souhaité extraire les solutions qui dépendaient de voies de signalisation dans le sens communément employé par les Biologistes, c'est à dire de l'extracellulaire vers l'intracellulaire. Pour cela nous avons recherché le terme "exCellRegion" qui correspond à l'annotation de la localisation extracellulaire d'une biomolécule dans les solutions. Parmi les 787 gènes, 456 ne dépendent d'aucun terme extracellulaire. En observant les tailles de solutions de ces gènes, nous constatons que ces gènes correspondent à des solutions de petites taille, et ainsi sans doute à des parties moins bien annotées de la base de données. Par exemple certaines solutions ne contiennent qu'un facteur de transcription en plus du gène, ce qui signifie qu'aucune information n'a été rentrée quand à l'activation de ce facteur de transcription par une cascade de signaux provenant d'un stimuli extracellulaire. Pour approfondir cette analyse nous avons compté le nombre de termes extracellulaires dont dépend l'expression de chaque gène en recherchant la présence de termes extracellulaires dans l'ensemble des solutions (Figure 3.5). Là encore les résultats s'avèrent très surprenants puisque certains gènes peuvent être régulés par plus de 20 biomolécules extracellulaires différentes soulignant une grande plasticité dans la régulation de l'expression des gènes.

Nous avons également regardé si dans une même solution, plusieurs termes extracellulaires sont présents, et donc nécessaires à l'expression du gène par cette solution (une trajectoire donnée parmi l'ensemble des solutions). Les solutions comportent majoritairement un terme extracellulaire correspondant à une stimulation par une biomolécule soluble (TGF-β, EGF, etc.), cependant certaines solutions en comportent 2 ou plus ce qui est extremement difficile à observer de façon expérimentale car les manipulations se font généralement en réponse à un unique ligand.

A l'inverse nous avons également recherché le nombre de gènes que peut réguler une biomolécule extracellulaire, c'est à dire le nombre de gènes pour lesquels la biomolécule est retrouvée dans les solutions (Figure 3.6). Sur les 63 biomolécules extracellulaires retrouvées dans l'ensemble des solutions des gènes, 23 sont spécifiques à un gène. Majoritairement, les biomolécules sont en mesure de réguler différents gènes, jusqu'à 158 pour la place $TRAIL\_trimer\_exCellRegion$. Ces données sont en accord avec la littérature ou de large profil d'expression de gène on été caractérisé pour ces biomolécules.

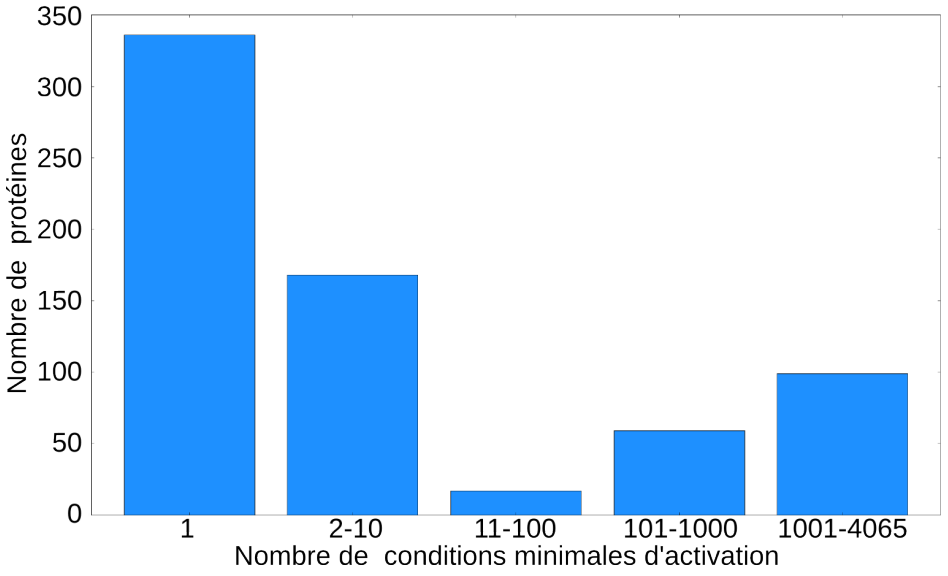


FIG. 3.3: Répartition du nombre de conditions minimales d'activation par protéine

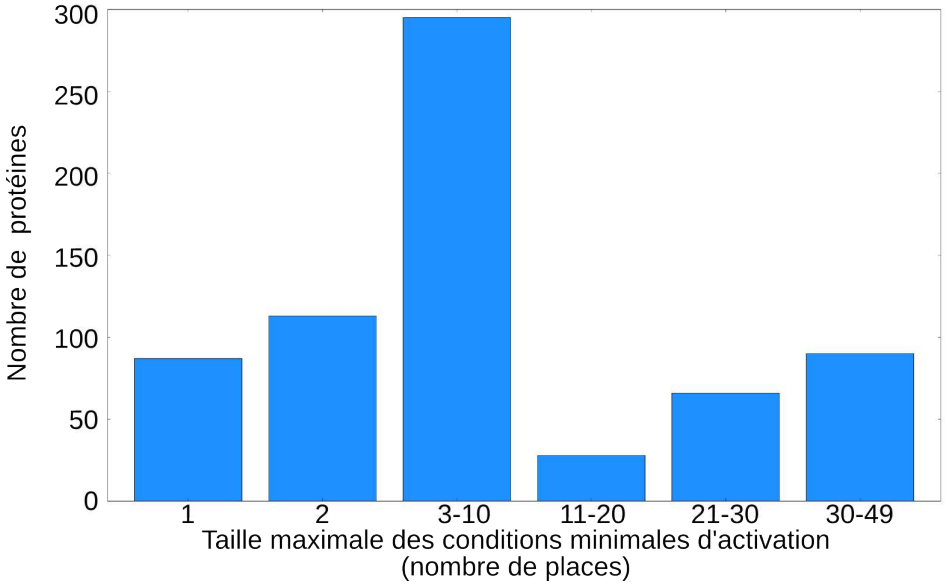


FIG. 3.4: Répartition de la taille maximale des conditions minimales d'activation par protéine

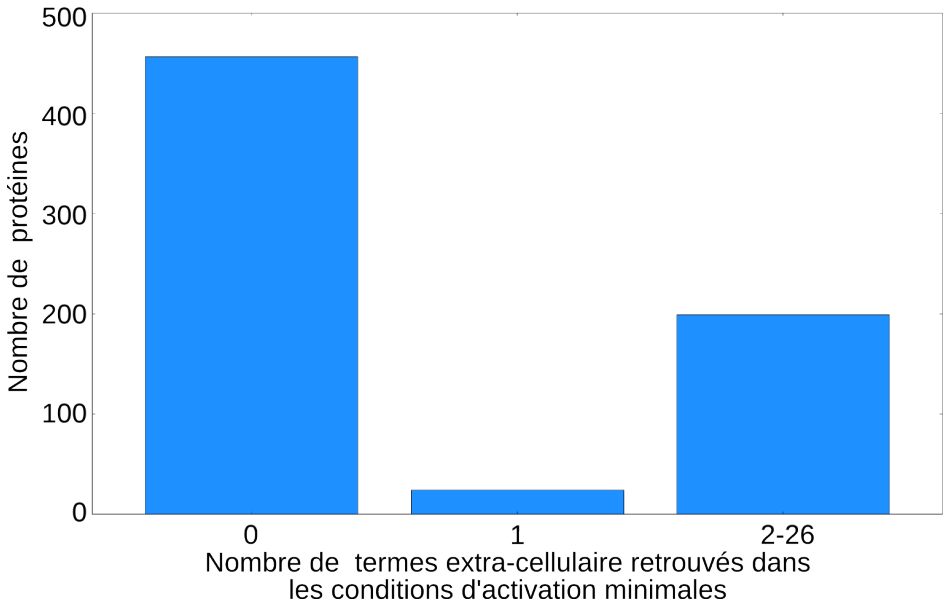


FIG. 3.5: Répartition du nombre de termes extracellulaires retrouvés dans les solutions

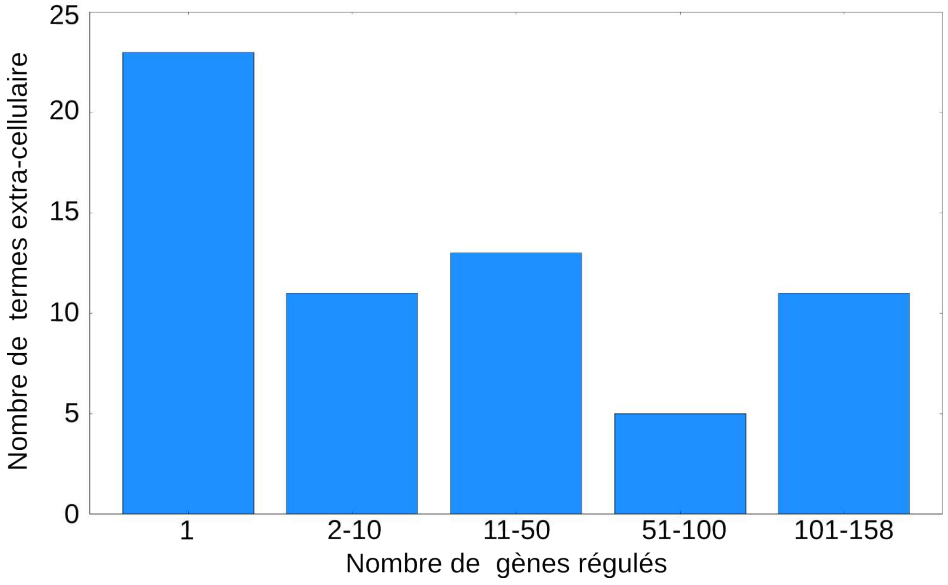


FIG. 3.6: Répartitions du nombre de gènes régulés par un terme extra-cellulaire

3.1.3 Analyse comparée avec les données de co-expression issues de la base de données GEMMA

Afin de confronter nos résultats à des données expérimentales, nous avons choisi GEMMA qui est à la fois une base de données pour les données d'expression de gènes haut débits (puce à ADN) et un ensemble d'outils pour les méta-analyses de ces données [159]. Actuellement GEMMA contient plus de 4000 jeux de données dont la moitié obtenue chez l'Homme. Nous avons donc utilisé l'outil de recherche de co-expression de GEMMA pour caractériser les 787 gènes de notre modèle. Les requêtes ont été effectuées sur l'ensemble des données de l'homme, soit 2234 expérimentations. L'outil GEMMA propose une donnée statistique du degré de corrélation (basé essentiellement sur le nombre d'expériences). Les résultats sont exprimés sous forme de tableau de corrélation où chaque gène est associé à la liste de gènes avec lesquels il a été retrouvé co-exprimé. Nous n'avons conservé que les couples de gènes présents dans notre modèle pour un total de 13210 couples. Chacun de ces couples est associé à un score GEMMA qui représente le nombre d'expérimentations dans lesquelles les deux gènes ont été trouvés co-exprimés.

L'idée de départ est de voir si notre modèle est capable de vérifier les données de co-expression et si tel est le cas, pouvoir expliquer les données expérimentales. Dans un premier temps nous avons sélectionné les couples de gènes co-exprimé dans Gemma et retrouvés dépendant d'un terme extracellulaire, c'est à dire présentant un terme contenant "exCellRegion" dans leurs solutions. Nous avons recherché l'atteignabilité des 2 protéines issues des gènes co-exprimés, c'est à dire si il existe au moins une solution capable d'atteindre P_1 et P_2.

Afin de comparer les résultats de Cadbiom nous avons besoin d'une fonction de score capable de s'apparenter au score de co-expression pour GEMMA. Pour chaque gène nous disposons de l'ensemble des solutions et donc des trajectoires de propagation du signal. Il n'existe pas de technique permettant d'exprimer la co-expression en fonction de ce type de résultats. Nous avons par conséquent développé un score qui bien que empirique, nous semblait être le plus proche d'un score de co-expression. L'idée est de comparer deux gènes, en comparant leur ensemble de solutions qui illustrent les voies de signalisation impliquées dans la régulation de leur expression. Plus deux ensembles de solutions sont proches, plus les deux gènes qu'ils régulent ont de chance d'être co-exprimés. Nous avons comparé les solutions 2 à 2 et défini une valeur qui représente leur ressemblance, c'est à dire le nombre de places (biomolécules) partagées par les 2 solutions. Ce score est ensuite pondéré par la probabilité de choisir ces 2 solutions parmi les 2 ensembles de solutions. Nous considérons une équiprobabilité entre toutes les solutions, la probabilité de choisir une solution dans

un ensemble de n solutions est donc de $\frac{1}{n}$. Cette pondération s'apparente à la probabilité que le signal passe par telle ou telle voie de signalisation. De cette manière si deux gènes ont une solution très proche dans un grand ensemble de solution, le score associé sera faible. Le score final entre 2 gènes est ainsi obtenu par la formule :

$$S_{AB} = \sum \left[\left(\frac{a_i \cap b_j}{Max(a_i, b_j)} \right) P_{a_i b_j} \right]$$

où S_{AB} représente le score CADBIOM entre les protéines A et B, a_i et b_j sont respectivement la $i^{ème}$ solution de A et la $j^{ème}$ solution de B et $P_{a_i b_j}$ la probabilité de choisir la $i^{ème}$ solution de A et la $j^{ème}$ solution de B. Un score nul entre deux protéines signifie que les ensembles de solutions de ces protéines sont distincts, sinon elles partagent une ou plusieurs places et donc potentiellement des voies de signalisation.

En plus du score obtenu comparant les ensembles de solutions de deux protéines P_1 et P_2, nous recherchons l'atteignabilité de la propriété (P_1 *et* P_2). Deux protéines P_1 et P_2 sont dit co-atteignables si il existe une solution permettant d'atteindre la propriété (P_1 *et* P_2). C'est à dire que les gènes codant pour les protéines P_1 et P_2 peuvent être activés au cour d'une même trajectoire dans notre modèle.

En utilisant à la fois le score de ressemblance, et le test d'atteignabilité de la propriété (P_1 *et* P_2), différents cas peuvent être interprétés et sont résumé dans le Tableau 3.1 en fonction de leur co-expression dans la base GEMMA.

Nous avons expérimenté cette approche en analysant les couples de gènes retrouvés co-exprimés dans GEMMA et dont les solutions dépendent d'un terme extracellulaire soit 649 couples de gènes (Tableau 3.2). Les gènes co-exprimés dans GEMMA sont presque tous co-atteignables. En effet, seul 1 couple de gènes n'a aucune solution, et les deux gènes partagent des places dans leurs solutions (score non nul) lorsque l'on recherche leur atteignabilité de façon indépendante. Parmi ces couples de gènes co-atteignables une forte majorité (81%) partagent des places dans leurs solutions et par conséquent des trajectoires communes. Ces résultats suggèrent que les gènes co-exprimés dans GEMMA dépendent majoritairement de voies de signalisation communes. Par exemple les gènes DUSP1 et FOS retrouvés fortement co-exprimés dans GEMMA, sont co-atteignables et les solutions permettant l'atteignabilité des places représentant leurs protéines respectives contiennent des places liées à la voie de signalisation de BMP.

A l'inverse nous avons recherché si des gènes non co-exprimés dans GEMMA avaient un profil de signalisation particulier. Pour cela nous avons généré une liste aléatoire de même taille que celle utilisée pour les couples co-exprimés, soit 649 couples de gènes parmi les gènes présents dans notre modèle et non co-exprimés dans GEMMA. L'analyse de ces

P_1 **et** P_2 **co-exprimés dans GEMMA**		
Score CADBIOM	(P_1 et P_2) atteignable	(P_1 et P_2) non atteignable
Score = 0	P_1 et P_2 peuvent être co-exprimés dans le modèle CADBIOM mais ne partagent pas de places donc de voie commune	P_1 et P_2 ne semblent pas pouvoir être co-exprimés dans le modèle CADBIOM. Les solutions de P_1 peuvent inhiber celles de P_2 ou inversement.
Score $\neq$ 0	P_1 et P_2 peuvent être co-exprimé dans le modèle CADBIOM et partagent potentiellement une voie de régulation	P_1 et P_2 ne semblent pas pouvoir être co-exprimés dans le modèle CADBIOM. Les solutions de P_1 peuvent inhiber celles de P_2 ou inversement. Il peut également exister une compétition entre les régulations puisqu'ils partagent des voies de régulations.
P_1 **et** P_2 **NON co-exprimés dans GEMMA**		
Score CADBIOM	(P_1 et P_2) atteignable	(P_1 et P_2) non atteignable
Score = 0	P_1 et P_2 peuvent être co-exprimés dans le modèle CADBIOM. Cependant cette co-expression n'a pas été observé observée dans la bases de données GEMMA	P_1 et P_2 ne semblent pas non plu pouvoir être co-exprimés dans le modèles CADBIOM.
Score $\neq$ 0	P_1 et P_2 peuvent être co-exprimés dans le modèle CADBIOM et partagent potentiellement une voie de régulation. Cependant cette co-expression n'a pas été observé dans les données de GEMMA.	P_1 et P_2 ne semblent pas non plus pouvoir être co-exprimés dans le modèle CADBIOM. Cependant il peut exister une compétition entre les régulations puisqu'ils partagent des voies de régulations.

TAB. 3.1: Interprétation des valeurs de score et d'atteignabilité dans le modèle CADBIOM pour les couples de gènes co-exprimés dans Gemma.

	Co-atteignable	Non co-atteignable
Score = 0	123	0
Score ≠ 0	525	1

Tab. 3.2: Test des couples de gènes co-exprimés dans GEMMA et dépendant d'un terme extracellulaire

	Co-atteignable	Non co-atteignable
Score = 0	530	0
Score ≠ 0	119	0

Tab. 3.3: Test des couples de gènes non co-exprimés dans GEMMA

couples est présenté dans le Tableau 3.3. Les profils observés sont différents de ceux issus des gènes co-exprimés. En effet même si tous les couples sont co-atteignable, les gènes ne partagent pas de places dans leurs solutions. D'après notre modèle, ces couples de gènes ne sont pas incompatibles mais dépendent donc de voies de signalisation différentes, ce qui permet d'expliquer qu'ils ne sont pas retrouvés dans les données de GEMMA.

En conclusion, cette analyse nous a permis de discriminer des profil de signalisation pour les 787 gènes du modèle issu de PID en accord avec les données de co-expression extraites de GEMMA. La recherche des solutions permettant l'activation de ces gènes a démontrée que notre modèle permet de retrouver des solutions pour activer les gènes co-exprimés et que les gènes co-exprimés partagent des voies de signalisation communes à l'inverse des gènes non co-exprimés qui sont régulés par des voies distinctes.

3.2 Interprétation de la base de données Reactome en Cadbiom

Dans le chapitre précédent, nous avons montré que le concept de réaction biologique était adopté par la base de donnée PID, nous permettant une interprétation en Cadbiom. Malheureusement depuis septembre 2012, PID n'est plus actualisée et son contenu ne sera plus disponible dès septembre 2013. Pour faire perdurer notre approche, nous avons entrepris d'interpréter une autre base de données dont le contenu est également structuré autour du concept de réaction biologique. Cependant des différences subtiles existent entre la représentation du contenu dans PID et dans Reactome. C'est pourquoi nous décrivons ici le schéma de traduction que nous avons mis en place pour traduire Reactome en modèle Cadbiom.

3.2.1 Description du contenu de Reactome

Reactome est une base de données s'inscrivant parfaitement dans la volonté actuelle qui est de réunir des informations de qualité, obtenues de façon diverses, dans le but de les visualiser, les analyser ou encore les partager. En ce sens, Reactome utilise une visualisation de son contenu basé sur la notation graphique SBGN, et propose les formats d'exports les plus répandus au sein de la communauté tel que BioPAX, SBML et PSI-MI. Le contenu est manuellement annoté par des biologistes experts dans leurs domaines. A l'instar de PID, le contenu de Reactome est bien structuré, dans des concepts proches de la notion de réaction biologique sur laquelle nous nous basons pour permettre une traduction automatique en transitions gardées.

Reactome contient actuellement près de 6000 protéines, soit plus de 25% des identifiants retrouvés dans la base de donnée SwissProt (chez l'Homme). Reactome n'est pas uniquement dédié à la signalisation mais référence également des réactions issues du Métabolisme, des réseaux de régulation de gènes et a l'ambition de réunir toutes les réactions ayant lieu dans une cellule. Son contenu se découpe en trois éléments de base : entités, réactions et attributs de réactions dont la charte graphique est illustrée en Figure 3.7.

3.2.1.1 Biomolécules

Les biomolécules sont représentées par des noeuds dans l'interface de Reactome. Reactome regroupe différents types de biomolécules :

– les petites molécules (ovale vert), regroupant des nucléotides tels que l'ATP, des ions (Ca2+,H+,etc.) ainsi que d'autres molécules (ex : H2O).

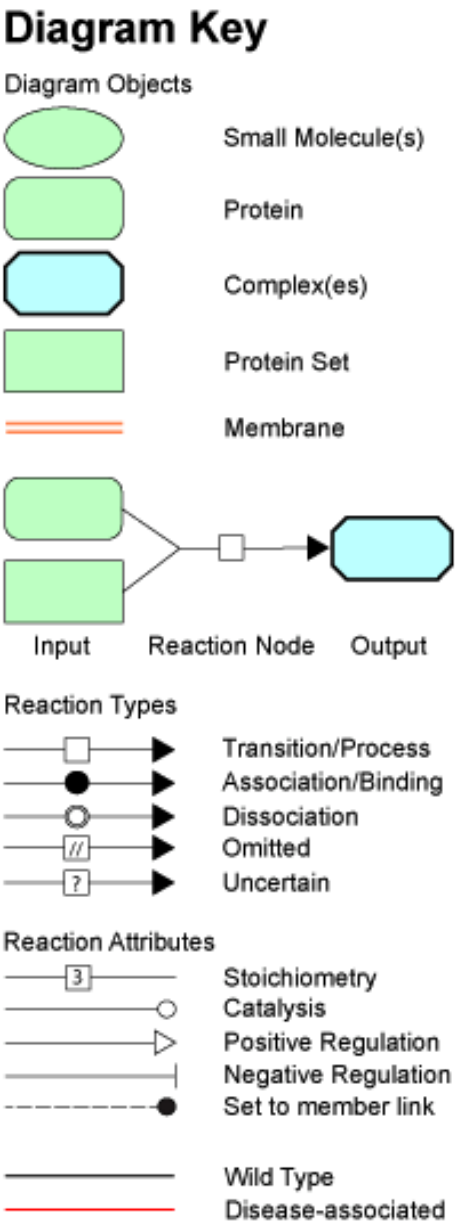


FIG. 3.7: Légende de la représentation graphique utilisée par Reactome. Les différents types de biomolécules sont représentés par des noeuds de formes différentes. Les types de réactions et de controles sont discriminés par différents arcs. Cette légende figure sur le site `www.reactome.org`.

- les protéines (rectangle vert aux bords arrondis), se distinguent en fonction des réactions dans lesquelles elles sont impliquées.
- les complexes (octogone bleu)
- les ensembles de protéines (rectangle vert), regroupent des familles, ou sous familles de protéines qui ont le même rôle dans une réaction donnée.

3.2.1.2 Réactions

Le concept de réaction est la clef de voûte de Reactome. Une réaction consomme et produits des biomolécules. Une reaction est représentée par un noeud, ayant pour entrée un arc provenant des noeuds biomolécules "entrées", et pour sortie un arc vers les biomolécules "sorties". Les réactions regroupent les processus suivant :

- La transition qui permet le transport d'une protéine d'un compartiment à un autre.
- L'association qui permet de former un complexe entre plusieurs protéines.
- La dissociation qui est le phénomène inverse.
- Réaction "omitted" dont le contenu est volontairement simplifié. Ce type de réaction est notamment utilisé pour décrire la transcription d'un gène en ne considérant que le gène et ses facteurs de transcription sans indiquer toute la machinerie de transcription. De la même façon, les réactions "omitted" peuvent être utilisées pour représenter la dégradation d'une biomolécule. Ces réactions "omitted" sont donc des abstractions de processus biologiques complexes.
- Réaction "uncertain", où le processus fin de régulation n'est pas connu, ces réactions sont donc potentiellement incomplètes.

3.2.1.3 Régulation

Les régulations sont considérées comme des attributs des réactions, elles ne portent donc que sur des réactions, ou des régulations, mais pas sur des biomolécules. Trois principaux types de régulation sont décrit dans Reactome :

- Catalyse : Régulation positive d'une réaction, potentiellement effectuée par une entrée de la réaction.
- Positive régulation : Régule positivement une réaction ou une catalyse
- Negative régulation : Régule négativement une réaction ou une catalyse

Notons qu'une régulation donnée utilise un unique régulateur (une biomolécule) et régule une unique réaction (ou catalyse). Ainsi une réaction activée par x biomolécules possédera x régulateurs.

3.2.2 Règles de traduction en Cadbiom

Grâce à ces concepts bien définis, il est possible d'établir un schéma de traduction non ambigu. Les différentes réactions décrites dans Reactome peuvent se généraliser dans le concept de réaction biologique, avec des entrés, des sorties et des régulateurs. Pour parser les données de Reactome nous avons choisi d'utiliser le format d'échange BioPAX de niveau 3. Ce format largement répandu dans la communauté de la biologie des systèmes possède une sémantique bien adaptée aux concepts de réaction représentés dans Reactome. La Figure 3.8 montre la hiérarchie des principales entités définies dans la sémantique de BioPAX.

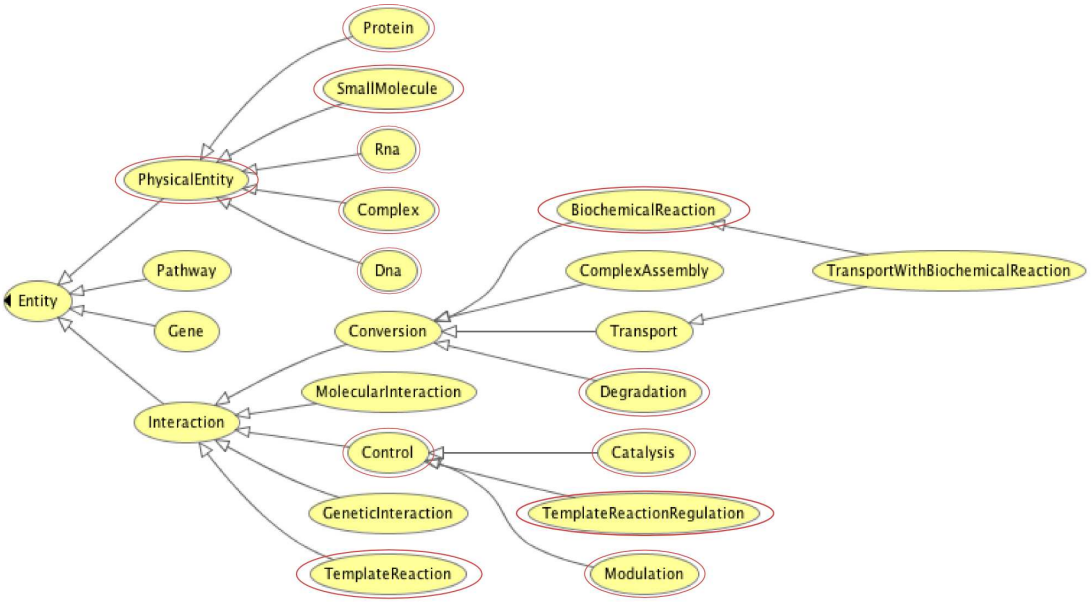


FIG. 3.8: Hiérarchie des entités du format Biopax. Chaque noeud représente une entité, les arcs sont des relations de type "est un". A titre d'exemple, l'arc de *Catalysis* vers *Control* signifie que la catalyse est un type particulier de contrôle. Les entités présentes dans Reactome et que nous avons interprétées sont entourées en rouge. Ce schéma est adapté de la documentation de BioPAX disponible sur `www.biopax.org`.

Les biomolecules présentes dans Reactome sont donc réparties dans les catégories : *Protein*, *SmallMolecule*, *Rna*, *Complex*, *Dna* et *PhysicalEntity* (utilisé notamment pour les ensembles de protéines). Pour traduire l'information issue de Reactome, nous avons tout d'abord parsé ces différentes biomolécules afin de générer les noms de leurs places. Pour chaque biomolécules (PhysicalEntity et ses sous classes), le nom de la place dans un modèle CADBIOM est obtenue par la concaténation du nom affiché dans Reactome

(balise bp :displayName) avec sa localisation (balise bp :cellularLocation). Nous utilisons l'information sur la localisation sub-cellulaire de façon à discriminer des biomolécules qui interviennent dans des réactions différentes parce qu'elles sont dans des compartiments différents.

Les réactions de Réactome sont réparties dans les classes BioPAX de la façon suivante : TemplateReaction pour les transcriptions, Degradation pour la dégradation, et les autres réactions (transport, complexation, etc.) sont définies comme BiochemicalReaction.

Le schéma de traduction de Reactome vers Cadbiom est similaire à celui utilisé pour PID. A partir d'une réaction biologique, nous modélisons la propagation du signal des entrées vers les sorties des réactions. Pour cela chaque biomolécule entrante possède une transition vers chaque biomolécule sortante. Les gardes des transitions sont formées d'un événement, spécifique à la réaction, ainsi que d'une formule logique assurant la présence de toutes les biomolécules entrantes pour pouvoir franchir la transition. En plus de cela, des régulateurs positifs et négatifs peuvent intervenir dans cette formule logique (cf. le paragraphe ci dessous sur les classes Control, Catalysis, Modulation et TemplateReaction-Regulation).

Traduction de la classe BiochemicalReaction Cette classe correspond au schéma de réactions biologiques décrites dans le paragraphe précédent. Parmi les réactions appartenant à la classe BiochemicalReaction, nous avons illustré la formation d'un complexe et son interpretation dans la Figure 3.9A.

Traduction de la classe Degradation Les dégradations sont des réactions particulières dans le sens où elles ne possède pas de biomolécules sortantes. Ceci implique également qu'une dégradation ne peut être suivie par aucune autre réaction. En utilisant une place particulière du formalisme Cadbiom il est possible de représenter ce concept. En effet les places "trap", nous permettent de représenter ce processus car elles ne peuvent avoir de transitions sortantes. En Cadbiom la place représentant la biomolécule qui va être dégradée, possède une transition sortante vers une place "trap" (Figure 3.9B). Notons que certaines dégradations peuvent avoir plusieurs entrée, généralement des protéines de la même famille. Lors de l'interprétation, chaque entrées possède une transition sortante vers une place "trap" spécifique. De cette manière, ces biomolécules sont dégradées de façon indépendantes, ce qui nous semble être le plus en accord avec l'interprétation en biologie.

Traduction de la classe TemplateReaction Cette classe regroupe les phénomènes d'expression de gène. A l'inverse de la dégradation, ce mécanisme est représenté par une réaction sans entrée, avec en sortie la protéine codée par le gène. Pour représenter ces mécanismes d'expression nous utilisons une autre place particulière du formalisme Cadbiom : la place permanente. Comme son nom l'indique, une fois activée, cette place ne peut être désactivée. Cette implémentation étant fondée sur le fait qu'un gène reste physiquement présent même si la protéine est traduite. L'expression est modélisée par une place permanente ayant une transition sortante vers la place représentant la protéine exprimée (Figure 3.9C).

L'expression de gènes d'une même famille peut être résumée en une seule interaction dans Reactome. Nous avons choisi de considérer des expressions indépendantes, en attribuant à chaque protéine exprimée, un gène spécifique.

Traduction des classes Control, Catalysis, Modulation et TemplateReactionRegulation Les régulations vont être intégrées dans les conditions des transitions en Cadbiom. Chaque régulation possède une entité régulatrice, une entité régulée et un type (activateur ou inhibiteur). Nous avons distingué 2 types de régulation : celles qui régulent des réactions, et celles qui régulent des catalyses.

Régulation des réactions Dans le cas général, une réaction biologique peut avoir plusieurs activateurs et inhibiteurs. La combinatoire entre activateurs et inhibiteurs qui permet à la réaction d'avoir lieu n'est pas connue. Dans ce contexte et comme pour la base de donnée PID, nous avons développé 2 interprétations possibles. La première interprétation consiste à réunir les activateurs (et inhibiteurs) par des connecteurs logique "et". Ainsi la présence de tout les activateurs est requise pour que la réaction ait lieu. Et de même, la présence de tout les inhibiteurs est nécessaire pour empêcher la réaction d'avoir lieu. Alternativement, les activateurs (et inhibiteurs) peuvent être connectés par des "ou". De cette manière, La présence d'un seul activateur suffit à réaliser la réaction et celle d'un inhibiteur à l'empêcher. Ces 2 interprétations ne sont pas les seules, et d'autre combinatoires seraient envisageables, par exemple : connecter les activateur par des "et" et les inhibiteurs par des "ou". La solution optimale est de pouvoir utiliser n'importe quelle formule logique mais reste trop coûteuse compte tenu des innombrables possibilités.

Régulation des catalyses Ces régulations conditionnent l'effet d'une biomolécule qui catalyse une réaction. Du point de vue biologique, on considère que la catalyse ne peut avoir lieu qu'en présence de ces activateurs et en l'absence de ces inhibiteurs. On considérera la régulation de la catalyse indépendamment de celle de la réaction qu'elle

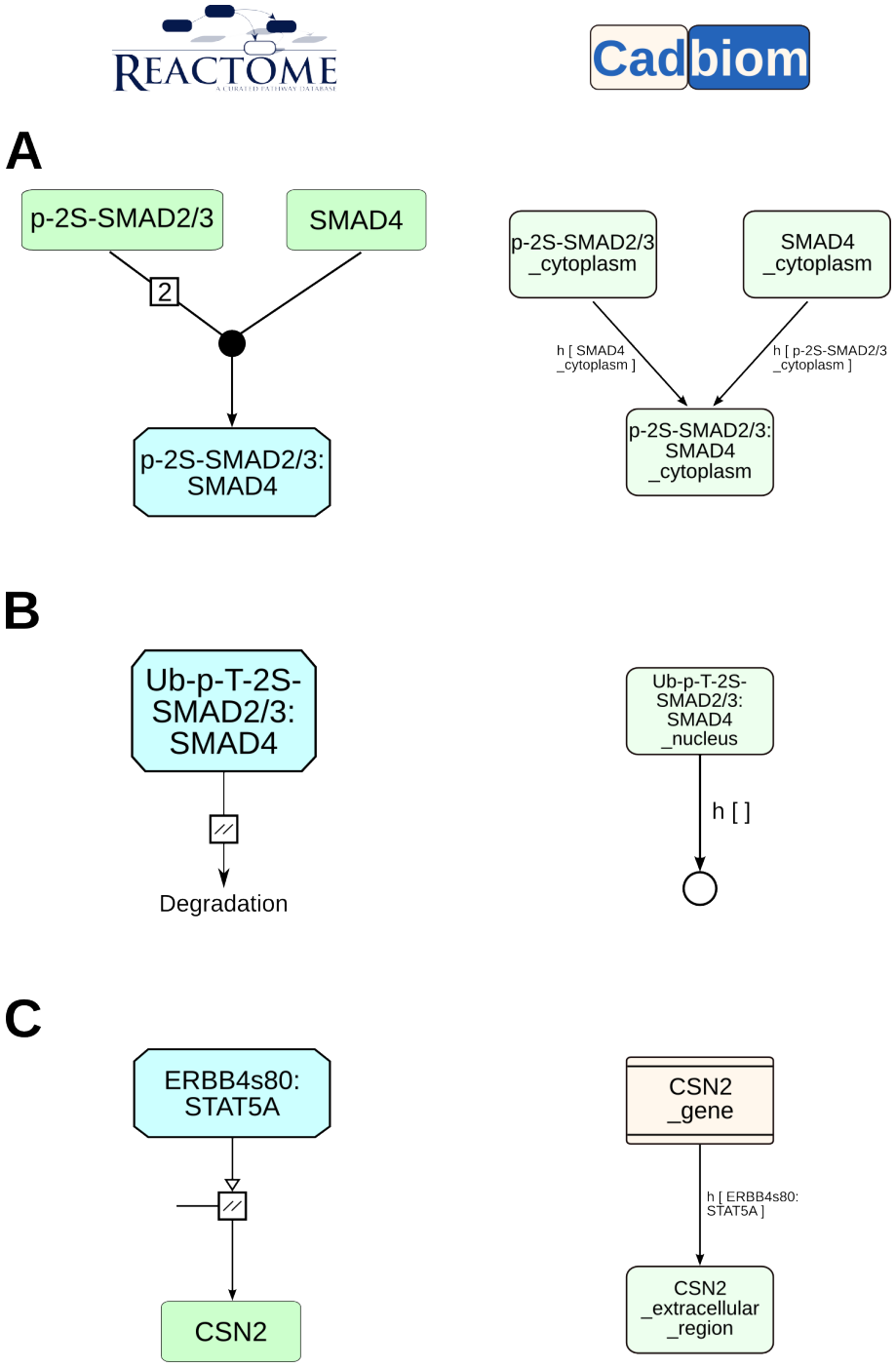


FIG. 3.9: Schéma de traduction de la base de donnée Reactome en CADBIOM.

active. Ainsi si la biomolécule A catalyse une réaction R, et que cette catalyse est activée par une biomolécule B et inhibée par une biomolecule C, l'activation de R par A sera remplacée par (A *and* B *and not* C).

Contenu non traduit Parce que nous souhaitons modéliser le contenu de Reactome avec un point de vue propagation du signal, certaines informations ne sont pas compatibles avec notre niveau d'abstraction, quelque peu éloignée des réactions biochimiques. Ainsi les informations sur la stœchiométrie des réactions ne sont pas prises en compte lors de l'interprétation en CADBIOM car nous considérons qu'une place active est en quantité suffisante pour jouer son rôle sans avoir besoin d'ajouter des informations sur la stœchiométrie. Dans le même esprit, certaines biomolécules utilisées dans de très nombreuses réactions n'apportent pas d'informations sur le signal. C'est le cas des molécules fournissant l'énergie des réactions (ATP, NAD, etc.) qui sont nécessaires à l'échelle biochimique mais dont la représentation est superflue à notre niveau d'abstraction. Ainsi ces petites molécules ne sont pas interprétées en CADBIOM, et les réactions ne comportant que des petites molécules ne sont pas présentes dans nos modèles. L'hypothèse faite ici est que ces petites molécules font parties de la machinerie cellulaire sous-jacente sans impacter sur la régulation du signal.

FIG. 3.9: (**A**) Modélisation de la classe BiochemicalReaction. La réaction symbolise la formation d'un complex entre les biomolécules p-2S-SMAD2/3 et SMAD4 (entrées de la réaction). L'arc entre p-S2-SMAD2/3 et le complexe p-2S-SMAD2/3 :SMAD4 est pondéré par le chiffre 2 pour représenter la stœchiométrie de la réaction. Dans le modèle CADBIOMgénéré, chaque biomolécule est représentée par une place. Il existe une transition de chaque entrée, vers chaque sortie de la réaction. Les conditions de ces transitions se composée des entrées complémentaire (ici l'autre membre du complexe) de façon à assurer la présence de toutes les entrées pour que la réaction ai lieu. La stœchiométrie n'est pas prise en compte à notre niveau d'abstraction. En revanche les informations disponible sur la localisation et l'état d'activation sont utilisées pour générer le nom des places. (**B**) Modélisation de la classe Degradation. Dans Reactome une biomolécule dégradée possède un arc vers l'étiquette Degradation. Dans un modèle CADBIOM la place représentant la biomolécule dégradée possède une transition sortante vers une place "trap" (représenté par un cercle noir), qui ne peut avoir de transition sortante. (**C**) Modélisation de la classe TemplateReaction. La classe TemplateReaction permet de représenter la régulation de gènes. La réaction produisant la biomolécule CSN2 n'a pas d'entrée (correspond au gène). La biomolécule ARBB4s80 :STAT5A joue ici le rôle d'activateur de la transcription, symbolisé par la flèche blanche sur la réaction. La régulation d'un gène en CADBIOMest représentée par une transition d'une place permanente symbolisant le gène, vers une place symbolisant la protéines synthétisée. La condition de cette transition est composée des régulateurs de la transcription, ici : ARBB4s80 :STAT5A.

Intégration des réactions Pour intégrer les réactions nous avons utilisé les mêmes principes que l'intégration des réactions de PID. Ainsi un événement h est lié à une réaction R. Toutes les transitions issues d'une même réaction ont le même événement. Deux réactions $R1$ et $R2$ sont identiques si et seulement si elles ont les même entrées et même sorties.

En revanche deux réactions $R1$ et $R2$ peuvent partager des entrées et sorties communes parmi leurs ensembles d'entrées/sorties. Dans ce cas, plusieurs réactions peuvent engendrer une transition gardée de A vers B. Pour chacune de ces réactions R_i, nous définissons un couple $(h, cond)$ où h est l'événement de la réaction et $cond$ est la condition (composé des activateurs et inhibiteurs de la réaction). Pour n réactions entraînant une transition de A vers B, l'événement est alors :
$(h_1\ when\ cond_1)\ default\ (h_2\ when\ cond_2)\ \ldots\ default\ (h_n\ when\ cond_n)$.

3.2.3 Analyse du modèle obtenu et comparaison avec le modèle obtenu à partir de PID

En suivant ces règles d'interprétation, l'intégralité de Reactome a été interprétée en Cadbiom. Le contenu de Reactome représentant 4383 réactions, basées sur 11691 pmids (inférence bibliographique), est ainsi réuni en un unique modèle dynamique de 8436 places et 8124 transitions. Afin de pouvoir juger du contenu quantitatif du modèle issu de Reactome, nous l'avons comparé au modèle issu de PID (Tableau 3.4). Le nombres de places et de transitions est légèrement inférieur à celui du modèle PID. Les proportions de place frontières, terminales et isolées sont sensiblement les mêmes. En revanche Reactome possède près de 4 fois moins de places permanentes, c'est à dire de gène.

Pour chaque modèle, nous construisons le graphe de transitions où les places sont les noeuds et les transitions sont les arcs. Les principales caractéristiques sont résumées dans le Tableau 3.5. La plus grande composante connexe comporte 48% des noeuds contre 58% pour celle du modèle issu de PID. De plus les degrés entrant et sortant sont plus faible dans le modèle issu de Reactome. Ces données indiquent que le modèle issu de Reactome est moins connecté que celui issu de PID.

3.3 Partenariat entre les formalismes Cadbiom et Frappes de Processus

Cette étude collaborative est née dans le cadre du projet ANR Biotempo et avait pour but d'une part de fournir un modèle biologique complexe à large échelle afin de

	Reactome	PID
Places	7733	9180
Transitions	8124	9266
Frontier places	3093	3919
Terminal places	2746	4439
Isolated places	502	757
Permanent places	232	797
Cytosol	1973	771
Nucleus	1496	1064
Membrane	2343	1637
Extracellular	664	226

Tab. 3.4: Comparaison des composants des Bases de données Reactome et PID.

	Reactome	PID
Nodes	7733	9180
Edges	8124	9266
Connected components	1413	1949
Biggest connected component	3674	5340
Average degree	1	1
Maximal input degree	33	87
Maximal output degree	62	100

Tab. 3.5: Comparaison des graphes de transitions issues de l'interprétation des données de Reactome et PID.

tester les algorithmes développés pour les frappes de processus et d'autre part d'inférer des informations en amont de nos analyses avec Cadbiom.

La recherche de propriétés dynamiques à l'aide de techniques de vérification de modèle peut s'avérer coûteuse en terme de temps de calcul. Toutefois il est possible d'extraire des connaissances à partir d'un système dynamique sans avoir à déplier la dynamique du modèle. Certains formalismes sont spécialement conçus pour ce genre d'analyses statiques sur des réseaux de grandes tailles. C'est notamment le cas du formalisme des frappes de processus (Process Hitting) qui a vocation d'analyser des réseaux à larges échelle (thèse de Loïc Paulevé [110]). Nous présentons ici succinctement le formalisme des frappes de processus puis l'interprétation du formalisme Cadbiomen frappe de processus.

3.3.1 Les frappes de processus

Le Process Hitting ou frappe de processus est un formalisme appartenant à la classe des réseaux d'automates. Un modèle est composé d'un ensemble d'automates, appelé *sortes* qui regroupe un nombre fini de processus, qui sont les différents états de l'automate. A chaque instant un et un seul processus est présent par *sorte*. Le processus actif défini ainsi l'état de cette *sorte* (ou état de l'automate). Les processus agissent entre eux par le biais d'actions. La Figure 3.10 résume les principaux composants des frappes de processus.

La structure des frappes de processus permet différentes analyses statiques basées sur le graphe de causalité locale (voir article en annexe de cette thèse : Paulevé *et al*, CAV, 2013). Il est notamment possible de rechercher des processus clefs, qui sont essentiels pour vérifier des propriétés d'atteignabilité. Si un processus clé d'une propriété n'est jamais présent alors cette propriété ne sera jamais atteinte. Cette notion peut être étendue aux ensembles de coupures minimales (ou minimal cut-sets). Un cut-set définit un ensemble de processus tel que toute trajectoire permettant d'atteindre la propriété passe au minimum par un de ces processus.

3.3.2 Sur-approximation de modèles Cadbiom en frappes de Processus

L'interprétation de modèles Cadbiom a été réalisée par Loïc Paulevé et correspond à une sur-approximation du formalisme Cadbiom sans événement. Cette sur-approximation assure que tout les processus clefs retrouvés par les analyses sur des modèles de frappes de processus, sont également essentiels dans les modèles Cadbiom. Chaque place d'un modèle Cadbiom est représentée par une sorte possèdant 2 processus : un pour l'état actif, l'autre pour l'état inactif. Pour représenter les transitions, il a fallu interpréter les conditions pour lesquelles la place sortante est activée. Considérant la transition Cadbiom $A \xrightarrow{[Cond]} B$, où $Cond$ est la formule logique composée de places du modèle conditionnant le tirage de la transition, B est actif si $(A \wedge Cond)$ est $Vrai$. La combinaison de plusieurs valeurs en frappe de processus est représentée par le biais de sortes coopératives. Afin de mieux comprendre cette démarche, la Figure 3.11 illustre la sur-approximation d'un modèle Cadbiom en frappes de processus sur un exemple biologique concret : la phosphorylation d'une protéine.

Cette interprétation à permis d'obtenir un modèle en frappes de processus représentant l'intégralité du contenu de la base de données PID, à partir du modèle Cadbiom pour ainsi y rechercher les processus clefs, ou plus généralement les ensembles de coupes minimales (minimal cut-sets). Nous avons choisi de rechercher l'atteignabilité de 3 propriété différentes, impactant des processus cellulaire fondamentaux : p15INK4b et p21CIP1, deux

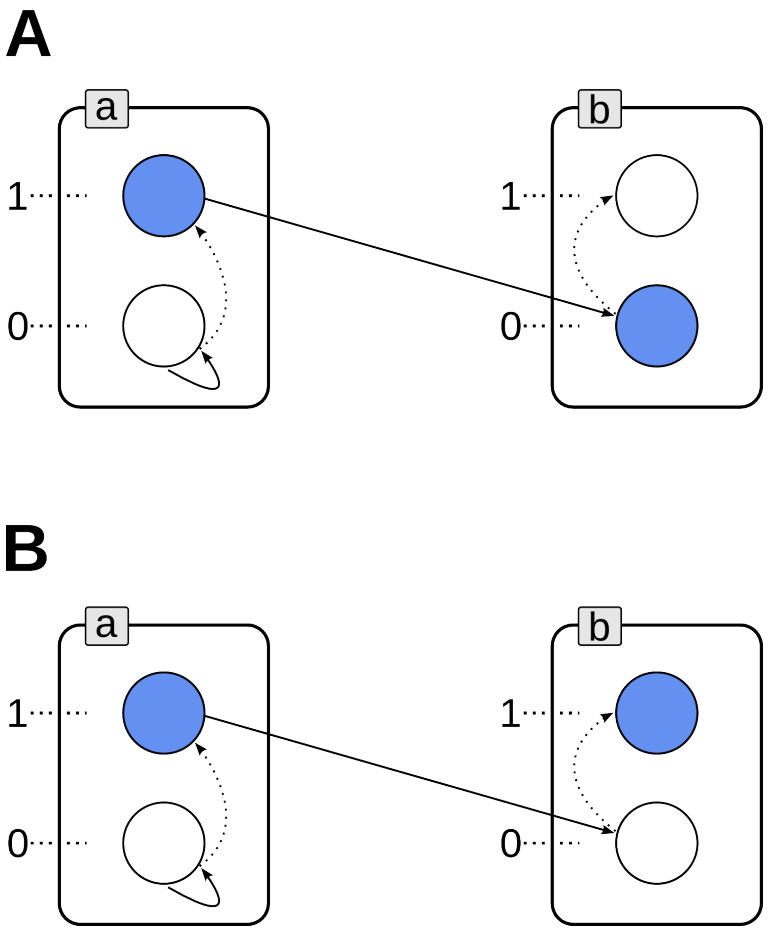


FIG. 3.10: Représentation graphique d'un modèle de frappes de processus. Le modèle est composé de 2 sortes (a et b) représentées par des rectangles qui correspondent généralement à des biomolécules. Chaque sorte possède ici 2 processus (a_0, a_1, b_0, b_1) qui s'apparentent à l'état de la biomolécules de la sorte (0 : absent, 1 : présent). Deux actions sont présentées : a_0 peut frapper a_0 pour aller en a_1 et a_1 peut frapper b_0 pour aller en b_1.Une frappe change l'état du modèle. Ainsi l'état (a_1,b_0) (A) atteint l'état (a_1,b_1) (B)quand a_1 frappe b_0.

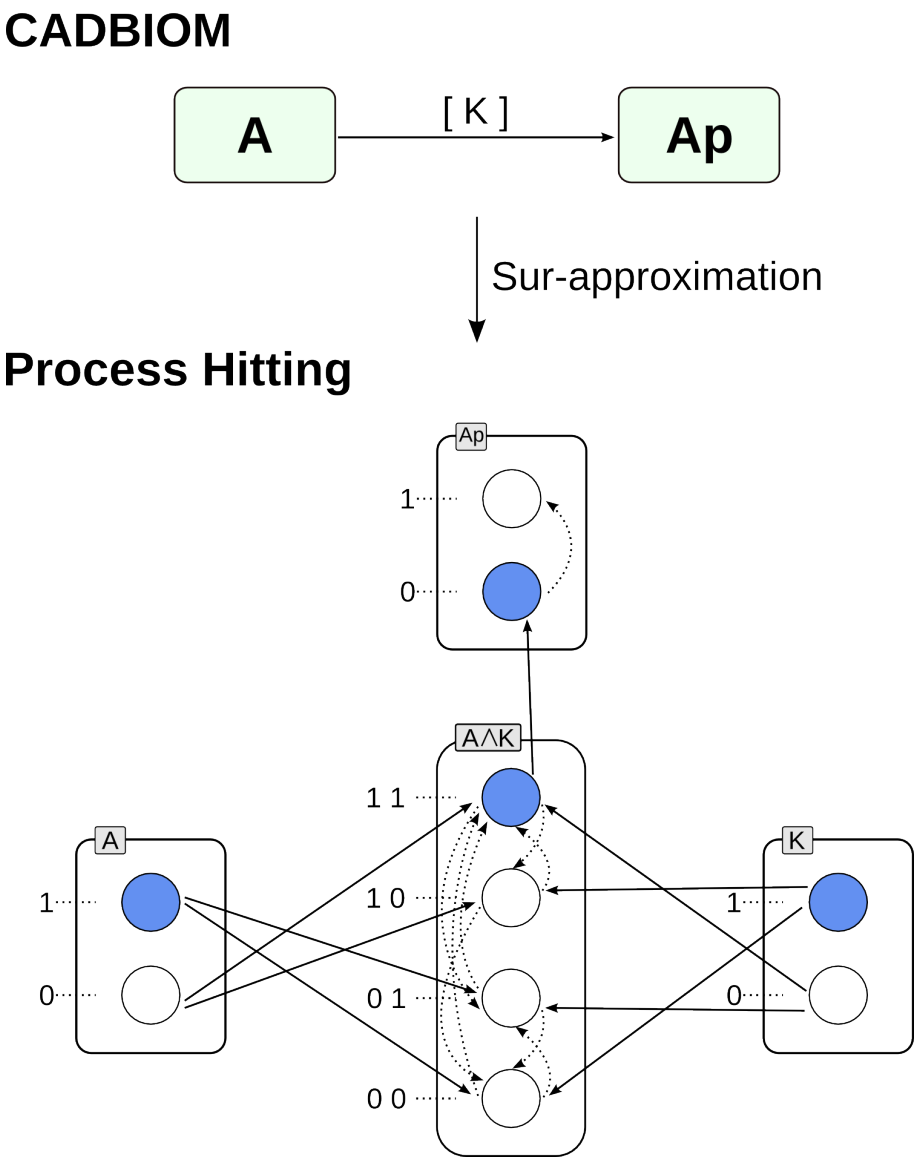


FIG. 3.11: Interpretation du formalisme CADBIOM en frappes de processus. Une réaction de phosphorylation d'une proteine A par une kinase K est représentée par une transition d'une place A vers une place Ap (A phosphorylée) conditionnée par K. Les protéines A, K et Ap sont ainsi représentées par 3 *sortes* ayant 2 processus. La nécessité de la présence de A et de K est représentée par une *sorte* coopérative (A et K) dont les processus dépendent de A et K. Le processus Ap0 est frappé par (A et K)11 pour passer en Ap1.

gènes qui régulent le cycle cellulaire et SNAIL, un gène caractérisant la transition épithélio mésenchymateuse. Les ensembles de coupures minimales de différentes tailles ont été calculés pour chacune de ces propriétés. Les résultats obtenus sont présentés dans l'article intitulé Under-approximating Cut Sets for Reachability in Large Scale Automata Networks, accepté dans la conférence Computer Aided Verification (CAV) 2013 (voir Annexes).

Dans un deuxième temps nous avons utilisé les résultats obtenus grâce aux analyses des frappes de processus pour restreindre l'espace de recherche lors de l'analyse de propriétés en CADBIOM. En effet en recherchant les solutions permettant d'atteindre une propriété dans un modèle CADBIOM, nous partons généralement sans *a priori* sur le modèle. Cependant grâce aux ensembles de coupures minimales obtenues avec le modèle frappes de processus, nous avons connaissance des processus essentiels pour l'atteignabilité d'une propriété. Nous pouvons alors les utiliser comme conditions de départ pour limiter l'espace de recherche. Puisqu'au moins un processus par cut-set doit être présent pour atteindre la propriété, nous établissons la condition C telle que :

$$C = \bigwedge_{n_i \in N} \left(\bigvee_{p_j \in n_i} \right)$$

où N est l'ensemble des minimal cut-sets.

En générant cette condition pour les coupures minimales des 3 propriétés p15INK4b, p21CIP1 et SNAIL, nous avons recherché les solutions permettant d'atteindre ces propriétés avec notre modèle CADBIOM, avec ou sans cette condition de départ. Un premier test à consisté à comparer les ensembles de solutions obtenus avec et sans la condition initiale. Conformément à nos attentes, l'utilisation de la condition initiale ne modifie pas les solutions obtenues. Nous avons ensuite testé s'il était possible d'atteindre la propriété en mettant non pas la condition générée à partir des ensembles de coupures minimales mais sa négation. Ce qui signifie qu'au moins une coupe n'a aucun processus de présent empêchant ainsi l'atteignabilité de la propriété dans le modèle de frappe de processus. Encore une fois les résultats sont conformes à nos attentes, à savoir qu'aucune solution n'est trouvée dans notre modèle CADBIOM pour atteindre les propriétés. Ces analyses démontrent que les coupes obtenue avec le modèle frappes de processus fournissent des informations sur la dynamique de nos modèles CADBIOMsans avoir besoin de déplier la dynamique.

Un autre objectif de cette collaboration était de permettre la réduction des temps de calcul du nos solutions. Cet objectif n'a en revanche pas été rempli. En effet, en pratique malgré un espace de solutions plus restreint, les temps de calcul des solutions d'atteignabilité sur nos modèle CADBIOM n'ont pas été diminués.

Quatrième partie

Discussion

Cette thèse s'inscrit dans la problématique actuelle en biologie des systèmes visant à apporter de nouvelles connaissances grâce à des modèles prédictifs. Nous avons développé deux approches de modélisation de la signalisation cellulaire pour répondre à des questions de natures différentes. Notre première approche est basée sur un modèle différentiel de régulation du signal TGF-β, la seconde est une approche discrète permettant de modéliser la signalisation cellulaire dans son ensemble. L'apport de cette thèse est ici commenté pour chacune de ces deux approches.

La signalisation cellulaire manque encore aujourd'hui d'informations sur certains mécanismes fins de régulation. C'est notamment le cas de la compréhension de la signalisation canonique du TGF-β par le facteur de transcription TIF1γ. Nous avons entrepris de réunir les informations divergentes de la littérature au sein d'un unique modèle différentiel capable d'expliquer ces observations biologiques contradictoires. Pour cela, nous avons intégré deux modèles différentiels détaillants des parties différentes de la signalisation TGF-β, pour permettre une représentation plus complète de la voie canonique du TGF-β. Le modèle ainsi obtenu à servi de clef de voûte à l'intégration des données concernant TIF1γ pour tester les hypothèses sur son rôle dans la régulation du signal. Notre modèle souligne l'importance du ratio TIF1γ/SMAD4 dans la régulation du signal TGF-β et du phénotype cellulaire. Nos résultats expliquent les observations divergentes de la littérature et ont été validées *a posteriori* par des manipulations expérimentales. Dans cette étude nous avons fait le choix d'utiliser les informations qualitatives contenues dans les publications sur TIF1γ qui nous ont permis d'estimer les paramètres. Cependant d'autres approches existent pour inférer des paramètres à un modèle différentiel à partir de données qualitatives [119]. Ainsi ces données peuvent être utilisées comme contraintes pour définir les bornes de l'espace des paramètres. Les solutions sont alors déterminées par minimisation d'une fonction de coût reflétant l'écart entre les simulations et les données

Si ce modèle s'est révélé être des plus utiles pour prédire un comportement en fonction d'un nouveau régulateur TIF1γ, il est en revanche incapable de répondre à des questions plus globales sur la propagation du signal. En effet ce modèle différentiel, comme la plupart, a été conçu dans le but de répondre à une question spécifique. Il ne représente que la voie canonique du TGF-β, et ne prend en compte ni les autres voies dépendantes du TGF-β, ni la régulation des différents gènes cibles par exemple. Ce modèle est ainsi loin de représenter la réelle plasticité de réponse d'une cellule au TGF-β. L'intégration de nouvelles données à partir d'expérimentations ou de la littérature, à l'instar du travail réalisé ici pour TIF1γ, n'est de plus pas envisageable à l'échelle d'un réseau entier de signalisation cellulaire.

L'extension de l'approche différentielle à des modèles larges échelles de quelques milliers de composants se heurte, en dehors de problèmes mathématiques et informatiques,

à un problème d'observabilité. Plus le modèle comporte de variables et plus il faut d'observations pour le valider. Les techniques haut-débit permettent cette observation simultanée de nombreuses variables, toutefois ces données sont souvent de nature binaires : présent/absent. Les données nécessaires à la conception et la validation de modèle continue ne sont pas accessibles et les questions posées ne correspondent pas aux analyses qu'il est possible de faire avec un modèle différentiel. Afin de pouvoir intégrer la complexité des voies de signalisation du TGF-β nous avons été ammené à développer une approche permettant la représentation d'un modèle large échelle de la signalisation, à l'aide d'un formalisme discret.

En développant notre approche CADBIOM, notre volonté était de pouvoir tenir compte des spécificités de la signalisation cellulaire, telle que l'hétérogénéité des réactions et l'extrême enchevêtrement des voies, dans un système dynamique permettant de visualiser la propagation du signal. Pour ce faire, nous avons utilisé un modèle centré sur les réactions, basé sur le formalisme des transitions gardées nous permettant de modéliser le concept de réaction biologique. Les autres formalismes sont généralement utilisés comme une approximation de concentrations : jetons dans les réseaux de Petri, niveaux dans les modèles booléens et logiques. Dans cette thèse nous avons pris un point de vue différent, celui de la propagation du signal à travers les réactions biologiques. Les transitions gardées nous ont semblé donner un cadre formel naturel à ce point de vue. Ce formalisme à été appliqué dans différents domaines tel que l'aéronautique avec une sémantique légèrement différente [113]. Dans notre application nous avons particulièrement apprécié la facilité de composer des modèles large échelle, l'utilisation des événements et les actions à distance symbolisées par les conditions.

Pour concevoir nos modèles nous avons définis un schéma d'interprétation non-ambigu pour les bases de données PID et Reactome qui possèdent une représentation compatible avec notre point de vue. De cette façon nous avons mis de coté les problèmes dus à l'interprétation manuelle de la littérature pouvant conduire à un biais dans la conception du modèle. Nos modèles ne sont pas orientés sur une voie de signalisation mais prennent en compte l'ensemble des connaissances actuelles de ces bases sur la signalisation. Les données présentées dans ces bases ont été interprétées par des experts, de façon plus objective car indépendante d'une volonté de modéliser ces données dans un système dynamique. C'est à dire que l'interprétation n'a pas été détournée dans le but de répondre à une question donnée.

L'interprétation d'une base de données demande de faire certains choix qui sont discutés ici. Tout d'abord, compte tenu du manque d'informations sur les conditions de régulation des réactions en présence de plusieurs activateurs et/ou inhibiteurs, nous avons

choisi d'interpréter ces régulations suivant 2 schémas. Le premier est un schéma de type "et" où tous les activateurs doivent être présents pour activer, et tout les inhibiteurs doivent être présents pour inhiber. A l'inverse dans le deuxième schéma de type "ou", 1 seul activateur suffit pour activer, et 1 seul inhibiteur suffit à inhiber. Ces schémas ne reflètent pas toute la combinatoire possible mais sans plus d'informations il est impossible de tester l'ensemble des possibilités pour chaque réaction. Différents travaux recherchent les bonnes combinaisons de régulation permettant d'expliquer un comportement, comme l'application CellNetOptimizer [124] et plus récemment une approche basée sur les méthodes ASP (Answer Set Programming) [145]. Pour cela un graphe de connaissance est entraîné avec des données de hautdébit de phospho-protéomique pour générer les modèles logiques optimaux capable d'expliquer les données. Ces approches permettent également de générer des modèles spécifiques à un type cellulaire, ce qui est extrêmement intéressant pour comprendre la diversité de réponse observée en signalisation. Ces travaux prometteurs sont appliqués à des modèles logique de la signalisation mais encore restreint à une centaine de molécules. L'utilisation de techniques similaires serait toutefois envisageable pour un sous modèle, de nos modèles dans le but d'affiner les connaissances sur les régulateurs.

Malgré la conception de modèles large échelle, la question de la complétude de ces modèles doit également être posée. Dans quelles mesures les données stockées dans ces bases représentent la réalité, ou plutôt l'ensemble des connaissances actuelles ? Ce critère est difficile à évaluer, cependant de récentes études proposent de comparer le contenu de ces bases pour évaluer leur recoupement. L'idée est de comparer le contenu de ces bases au niveau de leurs réactions et biomolécules. Les résultats de ces études vont dans le sens d'une complémentarité de ces bases, à savoir qu'une très faible proportion de données est commune à toutes les bases étudiées dont PID et Reactome [72]. Les informations issues de PID et de Reactome sont donc différentes, potentiellement complémentaires et mériteraient d'être regroupées. Cependant la réunion des données issues de ces bases n'est pas une tache triviale. Des travaux proposent d'intégrer des données de différentes bases parmi lesquelles PID et Reactome [65, 18]. Pour autant la réunion du contenu de ces bases soulève quelques problèmes. Une partie de l'information est perdue pour comparer et regrouper les informations, par exemple les annotations de localisation et des modifications ne sont pas prises en compte. De plus le niveau de détails peut être différent entre ces bases, créant ainsi de la redondance car des réactions identiques décrites avec des niveaux d'abstraction différents ne seront pas fusionnées et donc présentes en plusieurs exemplaires pour la même information. Pour ces raisons nous n'avons pas d'avantage recherché à fusionner les données issues de PID et Reactome.

Les bases de données actuelles se concentrent sur la signalisation intra-cellulaire, et

n'intègrent que très peu de données sur la régulation du signal en extra-cellulaire. Les ligands des récepteurs de surface cellulaire sont souvent considérés comme des entrées dans les modèles, leur régulation n'étant que très rarement explicitée. Cette absence est une limite pour la complétude des modèles de signalisation et très dommageable en sachant que la régulation de la signalisation fait intervenir des mécanismes complexes de la régulation extra-cellulaire. A titre d'exemple la dynamique d'interaction d'un ligand à son récepteur dépend de sa biodisponibilité au sein du réseau matriciel qui entoure les cellules.

Afin de compléter notre modèle de signalisation, une tentative a été réalisée dans le but de modéliser en Cadbiom l'activation extra-cellulaire du TGF-β (projet de Master 2 réalisé par Arnaud Le Cavorzin). Le modèle conçu à partir de la littérature parvient à reproduire les voies connues permettant l'activation du TGF-β mais ne nous a pas apporté de nouvelles informations. La majorité des réactions sont des formations de complexes dont certaines sont compétitives et d'autres non. La dynamique de ces interactions dépend du nombre de molécules et de leur affinité. Les connaissances actuelles ne sont pas suffisantes pour concevoir un modèle abstrait comme le propose Cadbiom. D'autres formalismes sont mieux adaptés à la représentation de ces interactions et permettent de tester l'effet de différentes quantités de molécules et des constantes d'affinité. C'est pourquoi un de nos projets en cours vise à développer un modèle pour la partie signalisation extracellulaire du TGF-β avec le langage basé sur des règles Kappa [29] (projet de Master 1 mené par Jean Coquet).

Le formalisme Kappa décrit les molécules avec leurs sites de réactions, qui peuvent être liés ou non, et dans un certain état, fonction de modification post-traductionelle par exemple pour les protéines (phosphorylées ou non ...). Les règles définissent la manière dont les molécules interagissent compte tenu d'une certaine configuration de leurs sites. Il est ainsi possible de définir des conditions très précises sur l'état des molécules avant et après la réaction, sans pour autant avoir à déclarer explicitement toutes les réactions possibles entre tous les états possibles des molécules. Les modèles peuvent ensuite être simulés suivants plusieurs interprétations (stochastique ou continue). En modélisant les mécanismes de l'activation extra-cellulaire du TGF-β en Kappa, nous souhaitons caractériser la dynamique extracellulaire du TGF-β, de sa sécrétion sous forme latente à sa libération sous forme active, responsable de l'initiation du signal au niveau du récepteur. Ce projet réalisé en collaboration avec Jérôme Ferret vise à plus long terme à intégrer après abstraction booléenne, le modèle Kappa au modèle Cadbiom afin de générer un modèle unique.

Le formalisme que nous avons développé est parvenu à passer à l'échelle, et ainsi gérer un modèle de la signalisation cellulaire de plus de 9000 noeuds. A notre connaissance il

s'agit du premier modèle de la signalisation, capable de rechercher des propriétés sur la dynamique du signal en prenant en compte plus de 9000 réactions. Les temps de calculs sont tout à fait acceptables pour un modèle de cette taille puisqu'une solution minimale est trouvée en environ 1 minute. Ces temps sont obtenus suite à une optimisation visant à rechercher les solutions dans un sous-modèles dont dépend la propriété. En utilisant un modèle sans a priori, il est normal que certaines transitions (donc réactions) n'aient aucun lien avec la propriété recherchée. Il est donc inutile de les considérer dans le modèle lors de la vérification de propriété. Cette optimisation est basée sur la topologie du graphe de transition et à été réalisée par Michel Le Borgne. Il s'agit d'un travail complémentaire, dont les détails ne figurent pas dans cette thèse.

En concevant un système dynamique il est important de discuter du schéma temporel choisi. L'interprétation flot de données ne permettait pas de reproduire un comportement réaliste car des réactions ne pouvaient avoir lieu à cause de la topologie du modèle. En effet la durée d'une voie ne dépendait que du nombre de réactions, donc du niveau de détails de la modélisation. Des phénomènes de compétitions naissaient de cette limitation et étaient un biais du modèle. Nous avons introduit les événements pour améliorer la dynamique sans avoir besoin de nouvelles informations. Les événements ne sont pas placés au hasard mais sont associés à chaque réaction, et ont de cette manière une signification biologique. La dynamique résultant de l'ajout des événements n'est pas forcée par un référentiel de temps universel, puisqu'à chaque pas le nombre de transitions franchies peut varier et dépend de l'occurrence des événements. En pratique l'occurrence des événements est déterminée en recherchant les solutions pour des propriétés d'atteignabilité et d'invariance. De ce fait nous ne faisons aucune hypothèse sur le timing dans nos modèles contrairement aux approches existantes, basées sur un schéma temporel contraignant la dynamique.

Concernant les analyses proposées par notre approche, nous avons souhaité être au plus proche de la pensée du biologiste en se basant sur des concepts simples comme la recherche de conditions d'activation ou d'inhibition. En obtenant les solutions, l'ensemble des régulations d'une propriété dans le modèle est alors connu, permettant d'exploiter les connaissances contenues dans la base puisque nos modèles parviennent à identifier les ensembles de réactions conduisant à une propriété souhaitée. La visualisation de ces voies de régulation est impossible dans les bases de données sans une approche de modélisation permettant d'exploiter les données.

La recherche d'inhibiteur d'une solution est un concept intéressant car il peut permettre de trouver des cibles thérapeutiques pour agir spécifiquement sur une réaction donnée. Les solutions permettant d'inhiber un comportement (par exemple l'activation d'un gène), peuvent être analysées pour identifier les différentes cibles d'inhibition possibles comme les

récepteurs, ou les messagers secondaires dans le cas de l'inhibition de p21. Les inhibiteurs retrouvés peuvent également se distinguer par les dommages collatéraux, autrement dit les réactions qu'ils impliquent. Ce point est important dans le choix de cibles thérapeutiques où l'on recherche à minimiser les réactions superflues.

Nous avons été surpris par le nombre de solutions minimales de certaines propriétés, certaines avoisinant les 4000 solutions minimales. Pour restituer ces ensembles de solutions sous une forme biologiquement intelligible, nous avons développé de nouvelles méthodes d'analyse visant à regrouper les solutions en fonction de leurs similitudes (nombre de places partagées). Cette méthode s'est révélée être utile pour des ensembles de solutions restreints (moins de 100 solutions), mais atteint ses limites pour les grands ensembles. Cependant nous sommes tout de même en mesure de retrouver les biomolécules principales (les plus fréquentes) de ces solutions. A ce jour nous travaillons à améliorer l'analyse de ces solutions pour faciliter le choix des expérimentations. Cela passe par l'analyse des événements (et donc réactions) impliqués dans ces solutions. Puisqu'une même réaction peut être régulée de différentes manière, cela crée de la diversité. Comparer les solutions sur la base de leurs réactions communes, au lieu de le faire sur les places, permettrait de mettre de coté la diversité liée aux régulations. D'autres méthodes comme l'extraction de sous-modèles sont à l'étude. Chaque solution pouvant être simulée, il est possible de récupérer toute les places activées lors de la simulation, c'est à dire lors de la propagation du signal. En faisant l'union de toutes les simulations, il serait possible d'extraire le sous-modèle de la régulation de la propriété recherchée.

La recherche de ces solutions minimales a eu un autre effet inattendu, à savoir que ces solutions donnent un réel aperçu du contenu de la base sur laquelle est fondé le modèle. En particulier l'utilisation de la recherche de propriété d'atteignabilité sur nos modèles s'est avérée être une méthode de curation de bases de données grâce à la modélisation. Par exemple, lors de la recherche de solutions permettant l'activation d'un gène, si certaines solutions représentent une voie de signalisation bien détaillée, d'autres sont moins bien annotées et montrent une faiblesse des données. Par exemple certains gènes sont exprimés en présence de leurs facteurs de transcription, mais ce dernier ne fait l'objet d'aucune régulation en amont réduisant ainsi la voie de signalisation à un gène et son ou ses régulateurs transcriptionnels. Pire encore, certaines solutions comportent des biomolécules sous forme de complexe, non relié à leurs régulateurs et se retrouvent parfois à faire doublon avec d'autre solutions. C'est le cas d'une solution permettant l'atteignabilité de p15INK4b : (p15INK4b_gene, SMAD3/SMAD4/SP1). SMAD3/SMAD4/SP1 est donc d'après cette solution suffisant pour permettre l'activation du gène p15INK4b, mais aucune information n'est donnée sur sa régulation puisqu'il s'agit d'une place frontière. Du

point de vue biologique ce complexe résulte de l'activation de SMAD3 par la voie TGF-β entraînant la complexation avec SMAD4 puis le facteur de transcription SP1. Cette information est retrouvée dans d'autres solutions conduisant à la formation du complexe SMAD3/SMAD4/SP1_nucl. Il s'agit de la même molécule, l'information est dupliquée car les annotations de localisations sont différentes. Dans un cas la biomolécule est annotée avec la localisation nucléaire (_nucl), dans l'autre aucune information n'est indiquée. Ce cas a été trouvé manuellement, il faudrait donc développer des méthodes permettant de retrouver des cas similaires pour parfaire le modèle. Pour le moment nous sommes capables de cibler les solutions ayant peu de places, qui font potentiellement référence à une partie sous annotée de la base de données. La curation de modèle est une étape importante de la modélisation, puisqu'un modèle n'est en rien figé, il demande à être constamment mis à jour à l'aide des nouvelles connaissances obtenues en l'analysant, ou de façon orthogonale avec de nouvelles données biologiques. A l'avenir nous souhaitons automatiser les méthodes d'aide à la curation de manière à faciliter l'évolution des modèles, en proposant par exemple de filtrer les solutions qui génèrent des trajectoires de petite taille.

Tout au long de cette thèse, nous avons souhaité développer une méthodologie qui soit à la fois fondée sur des concepts mathématiques solides, adaptés à la modélisation de la signalisation cellulaire, mais aussi intuitive pour permettre son utilisation par le plus grand nombre. C'est pourquoi nous avons regroupé les méthodes que nous proposons dans un logiciel qui accompagne le modélisateur de la conception à l'analyse de ses modèles. Les fonctionnalités permettent la création et l'édition de modèle depuis l'interface graphique, l'import depuis les bases PID et Reactome, ou encore les analyses statiques et dynamiques. Nous nous sommes efforcés de respecter les critères actuels en matière de logiciel ayant la vocation de modéliser des systèmes biologiques [112]. Ceci passe par différents points avec notamment la possibilité d'annoter les transitions pour y stocker les références bibliographiques ayant permis d'établir cette transition.

En conclusion, les approches présentées dans ce manuscrit participent à l'amélioration de la compréhension globale de la signalisation du TGF-β à la fois sur la régulation de sa voie canonique, mais aussi sur la façon dont les voies dépendantes du TGF-β s'inscrivent dans un réseau regroupant l'ensemble de la signalisation cellulaire. D'autres rôles du TGF-β mériteraient d'être abordés comme la régulation de processus biologique complexe tel que l'induction de la transition épithélio-mésanchymateuse, essentielle lors du développement mais aussi associée à diverses pathologies au stade adulte. Ce phénomène extrêmement complexe fait intervenir de nombreuses réactions entraînant un remodelage complet du phénotype de la cellule passant d'un état épithélial normal à un état mésenchymateux

invasif. Ce processus peut être induit par le TGF-β et fait intervenir à la fois les voies canoniques et non SMAD, et interfèrent avec les autres voies de signalisation.

Au-delà de la signalisation du TGF-β, la formalisation des connaissances biologiques dans des modèles dynamiques et prédictifs prenant en compte la complexité des mécanismes de réponses cellulaires, jouera un rôle déterminant dans l'identification de cibles thérapeutiques. L'amélioration de la conception et de l'analyse de ces modèles nécessite non seulement une collaboration entre modélisateurs pour intégrer les différents formalismes et approches mais aussi un dialogue avec les biologistes qui doivent s'approprier ces nouveaux outils.

Bibliographie

[1] U. Alon. Network motifs : theory and experimental approaches. *Nat. Rev. Genet.*, 8(6) :450–461, Jun 2007.

[2] A. F. Altelaar, J. Munoz, and A. J. Heck. Next-generation proteomics : towards an integrative view of proteome dynamics. *Nat. Rev. Genet.*, 14(1) :35–48, Jan 2013.

[3] S. S. Andrews and A. P. Arkin. Simulating cell biology. *Curr. Biol.*, 16(14) :R523–527, Jul 2006.

[4] A. Apte, D. Bonchev, and S. Fong. Cellular automata modeling of FASL-initiated apoptosis. *Chem. Biodivers.*, 7(5) :1163–1172, May 2010.

[5] A. A. Apte, J. W. Cain, D. G. Bonchev, and S. S. Fong. Cellular automata simulation of topological effects on the dynamics of feed-forward motifs. *J Biol Eng*, 2 :2, 2008.

[6] R. Apweiler, M. J. Martin, C. O'Donovan, et al. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, 38(Database issue) :D142–148, Jan 2010.

[7] M. Ashburner, C. A. Ball, J. A. Blake, et al. Gene ontology : tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1) :25–29, May 2000.

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

[9] M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster. Quantitative mass spectrometry in proteomics : critical review update from 2007 to the present. *Anal Bioanal Chem*, 404(4) :939–965, Sep 2012.

[10] T. Barrett, S. E. Wilhite, P. Ledoux, et al. NCBI GEO : archive for functional genomics data sets–update. *Nucleic Acids Res.*, 41(Database issue) :D991–995, Jan 2013.

[11] D. A. Benson, M. Cavanaugh, K. Clark, et al. GenBank. *Nucleic Acids Res.*, 41(Database issue) :36–42, Jan 2013.

[12] B. Bierie and H. L. Moses. Tumour microenvironment : TGFbeta : the molecular Jekyll and Hyde of cancer. *Nat. Rev. Cancer*, 6(7) :506–520, Jul 2006.

[13] M. L. Blinov, J. R. Faeder, B. Goldstein, and W. S. Hlavacek. BioNetGen : software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics*, 20(17) :3289–3291, Nov 2004.

[14] M. L. Burch, W. Zheng, and P. J. Little. Smad linker region phosphorylation in the regulation of extracellular matrix synthesis. *Cell. Mol. Life Sci.*, 68(1) :97–107, Jan 2011.

[15] L. Calzone, F. Fages, and S. Soliman. BIOCHAM : an environment for modeling biological systems and formalizing experimental knowledge. *Bioinformatics*, 22(14) :1805–1807, Jul 2006.

[16] M. Cassman. Barriers to progress in systems biology. *Nature*, 438(7071) :1079, Dec 2005.

[17] A. Ceol, A. Chatr Aryamontri, L. Licata, et al. MINT, the molecular interaction database : 2009 update. *Nucleic Acids Res.*, 38(Database issue) :D532–539, Jan 2010.

[18] E. G. Cerami, B. E. Gross, E. Demir, et al. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, 39(Database issue) :D685–690, Jan 2011.

[19] E. Chautard, M. Fatoux-Ardore, L. Ballut, N. Thierry-Mieg, and S. Ricard-Blum. MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Res.*, 39(Database issue) :D235–240, Jan 2011.

[20] K. C. Chen, L. Calzone, A. Csikasz-Nagy, et al. Integrative analysis of cell cycle control in budding yeast. *Mol. Biol. Cell*, 15(8) :3841–3862, Aug 2004.

[21] Li Chen, Ge Qi-Wei, Mitsuru Nakata, Hiroshi Matsuno, and Satoru Miyano. Modelling and simulation of signal transductions in an apoptosis pathway by using timed Petri nets. *Journal of Biosciences*, 32(1) :113–127, January 2007.

[22] A. A. Cheng and T. K. Lu. Synthetic biology : an emerging engineering discipline. *Annu Rev Biomed Eng*, 14 :155–178, 2012.

[23] C. Choudhary, C. Kumar, F. Gnad, et al. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, 325(5942) :834–840, Aug 2009.

[24] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, 3 :140, 2007.

[25] S. W. Chung, F. L. Miles, R. A. Sikes, et al. Quantitative modeling and analysis of the transforming growth factor beta signaling pathway. *Biophys. J.*, 96(5) :1733–1750, Mar 2009.

[26] D. C. Clarke, M. D. Betterton, and X. Liu. Systems theory of Smad signalling. *Syst Biol (Stevenage)*, 153(6) :412–424, Nov 2006.

[27] D. Croft, G. O'Kelly, G. Wu, et al. Reactome : a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39(Database issue) :D691–697, Jan 2011.

[28] V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. Rule-based modelling of cellular signalling. In *Proceedings of the 18 th International Conference on Concurrency Theory (CONCUR'07), Lecture Notes in Computer Science*, pages 17–41, 2007.

[29] V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. Rule-based modelling, symmetries, refinements. In Jasmin Fisher, editor, *Formal Methods in Systems Biology*, volume 5054 of *Lecture Notes in Computer Science*, pages 103–122. Springer Berlin Heidelberg, 2008.

[30] P. de Matos, R. Alcantara, A. Dekker, et al. Chemical Entities of Biological Interest : an update. *Nucleic Acids Res.*, 38(Database issue) :D249–254, Jan 2010.

[31] E. Demir, M. P. Cary, S. Paley, et al. The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, 28(9) :935–942, Sep 2010.

[32] A. Dilek, M. E. Belviranli, and U. Dogrusoz. VISIBIOweb : visualization and layout services for BioPAX pathway models. *Nucleic Acids Res.*, 38(Web Server issue) :W150–154, Jul 2010.

[33] DavidL. Dill, MerrillA. Knapp, Pamela Gage, et al. The pathalyzer : A tool for analysis of signal transduction pathways. In Eleazar Eskin, Trey Ideker, Ben Raphael, and Christopher Workman, editors, *Systems Biology and Regulatory Genomics*, volume 4023 of *Lecture Notes in Computer Science*, pages 11–22. Springer Berlin Heidelberg, 2006.

[34] A. Doi, S. Fujita, H. Matsuno, M. Nagasaki, and S. Miyano. Constructing biological pathway models with hybrid functional petri nets. *Stud Health Technol Inform*, 162 :92–112, 2011.

[35] N. Domedel-Puig, P. Rue, A. J. Pons, and J. Garcia-Ojalvo. Information routing driven by background chatter in a signaling network. *PLoS Comput. Biol.*, 7(12) :e1002297, Dec 2011.

[36] S. Dupont, A. Mamidi, M. Cordenonsi, et al. FAM/USP9x, a deubiquitinating enzyme essential for TGFbeta signaling, controls Smad4 monoubiquitination. *Cell*, 136 :123–135, Jan 2009.

[37] S. Dupont, L. Zacchigna, M. Cordenonsi, et al. Germ-layer specification and control of cell growth by Ectodermin, a Smad4 ubiquitin ligase. *Cell*, 121 :87–99, Apr 2005.

[38] M. Durzinsky, A. Wagler, and W. Marwan. Reconstruction of extended Petri nets from time series data and its application to signal transduction and to gene regulatory networks. *BMC Syst Biol*, 5 :113, 2011.

[39] X. M. Fernandez-Suarez and M. Y. Galperin. The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic Acids Res.*, 41(Database issue) :1–7, Jan 2013.

[40] P. Flicek, I. Ahmed, M. R. Amode, et al. Ensembl 2013. *Nucleic Acids Res.*, 41(Database issue) :48–55, Jan 2013.

[41] Akira Funahashi, Mineo Morohashi, Hiroaki Kitano, and Naoki Tanimura. Celldesigner : a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1(5) :159 – 162, 2003.

[42] J. Gao, L. Li, X. Wu, and D. Q. Wei. BioNetSim : a Petri net-based modeling tool for simulations of biochemical processes. *Protein Cell*, 3(3) :225–229, Mar 2012.

[43] Qian Gao, Fei Liu, David Gilbert, Monika Heiner, and David Tree. A multiscale approach to modelling planar cell polarity in drosophila wing using hierarchically coloured petri nets. In *Proceedings of the 9th International Conference on Computational Methods in Systems Biology*, CMSB '11, pages 209–218, New York, NY, USA, 2011. ACM.

[44] N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga, et al. Visualization of omics data for systems biology. *Nat. Methods*, 7(3 Suppl) :56–68, Mar 2010.

[45] D. Gilbert, H. Fuss, X. Gu, et al. Computational methodologies for modelling, analysis and simulation of signalling networks. *Brief. Bioinformatics*, 7(4) :339–353, Dec 2006.

[46] L. Glass and S. A. Kauffman. Co-operative components, spatial localization and oscillatory cellular dynamics. *J. Theor. Biol.*, 34(2) :219–237, Feb 1972.

[47] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell. Metabolomics by numbers : acquiring and understanding global metabolite data. *Trends Biotechnol.*, 22(5) :245–252, May 2004.

[48] K. A. Gray, L. C. Daugherty, S. M. Gordon, et al. Genenames.org : the HGNC resources in 2013. *Nucleic Acids Res.*, 41(Database issue) :D545–552, Jan 2013.

[49] J. Gruel, M. Leborgne, N. LeMeur, and N. Theret. In silico investigation of ADAM12 effect on TGF-beta receptors trafficking. *BMC Res Notes*, 2 :193, 2009.

[50] D. Harel. Statecharts : A visual formalism for complex systems. *Sci. Comput. Program.*, 8(3) :231–274, June 1987.

[51] W. He, D. C. Dorn, H. Erdjument-Bromage, et al. Hematopoiesis controlled by distinct TIF1gamma and Smad4 branches of the TGFbeta pathway. *Cell*, 125 :929–941, Jun 2006.

[52] M. Heiner, I. Koch, and J. Will. Model validation of biological pathways using Petri nets–demonstrated for apoptosis. *BioSystems*, 75(1-3) :15–28, Jul 2004.

[53] D. Heitzler, G. Durand, N. Gallay, et al. Competing G protein-coupled receptor kinases balance G protein and Î²-arrestin signaling. *Mol. Syst. Biol.*, 8 :590, 2012.

[54] C. H. Heldin, M. Landstrom, and A. Moustakas. Mechanism of TGF-beta signaling to growth arrest, apoptosis, and epithelial-mesenchymal transition. *Curr. Opin. Cell Biol.*, 21(2) :166–176, Apr 2009.

[55] T. Helikar, J. Konvalina, J. Heidel, and J. A. Rogers. Emergent decision-making in biological signal transduction networks. *Proc. Natl. Acad. Sci. U.S.A.*, 105(6) :1913–1918, Feb 2008.

[56] T. Helikar, B. Kowal, A. Madrahimov, et al. Bio-logic builder : a non-technical tool for building dynamical, qualitative models. *PLoS ONE*, 7(10) :e46417, 2012.

[57] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, et al. The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, 22(2) :177–183, Feb 2004.

[58] A. Hoffmann, A. Levchenko, M. L. Scott, and D. Baltimore. The IkappaB-NF-kappaB signaling module : temporal control and selective gene activation. *Science*, 298(5596) :1241–1245, Nov 2002.

[59] Z. Hu, J. H. Hung, Y. Wang, et al. VisANT 3.5 : multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res.*, 37(Web Server issue) :W115–121, Jul 2009.

[60] d. a. W. Huang, B. T. Sherman, R. Stephens, et al. DAVID gene ID conversion tool. *Bioinformation*, 2(10) :428–430, 2008.

[61] M. Hucka, A. Finney, H. M. Sauro, et al. The systems biology markup language (SBML) : a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4) :524–531, Mar 2003.

[62] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life : systems biology. *Annu Rev Genomics Hum Genet*, 2 :343–372, 2001.

[63] T. Ideker and N. J. Krogan. Differential network biology. *Mol. Syst. Biol.*, 8 :565, 2012.

[64] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang. High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.*, 2005(2) :96–103, Jun 2005.

[65] A. Kamburov, C. Wierling, H. Lehrach, and R. Herwig. ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Res.*, 37(Database issue) :D623–628, Jan 2009.

[66] K. Kandasamy, S. S. Mohan, R. Raju, et al. NetPath : a public resource of curated signal transduction pathways. *Genome Biol.*, 11(1) :R3, 2010.

[67] M. Kanehisa and S. Goto. KEGG : kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1) :27–30, Jan 2000.

[68] S. Kerrien, B. Aranda, L. Breuza, et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, 40(Database issue) :D841–846, Jan 2012.

[69] T. S. Keshava Prasad, R. Goel, K. Kandasamy, et al. Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, 37(Database issue) :D767–772, Jan 2009.

[70] L. B. Kier, D. Bonchev, and G. A. Buck. Modeling biochemical networks : a cellular-automata approach. *Chem. Biodivers.*, 2(2) :233–243, Feb 2005.

[71] Y. Kim and H. G. Othmer. A Hybrid Model of Tumor-Stromal Interactions in Breast Cancer. *Bull. Math. Biol.*, Jan 2013.

[72] D. C. Kirouac, J. Saez-Rodriguez, J. Swantek, et al. Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. *BMC Syst Biol*, 6 :29, 2012.

[73] M. W. Kirschner. The meaning of systems biology. *Cell*, 121(4) :503–504, May 2005.

[74] H. Kitano. *Foundations of Systems Biology*. MIT Press, October 2001.

[75] H. Kitano. Systems biology : a brief overview. *Science*, 295(5560) :1662–1664, Mar 2002.

[76] H. Kitano, A. Funahashi, Y. Matsuoka, and K. Oda. Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.*, 23(8) :961–966, Aug 2005.

[77] Steffen Klamt, Julio Saez-Rodriguez, and Ernst Gilles. Structural and functional analysis of cellular networks with cellnetanalyzer. *BMC Systems Biology*, 1(1) :2, 2007.

[78] K. W. Kohn. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell*, 10(8) :2703–2734, Aug 1999.

[79] K. W. Kohn, M. I. Aladjem, J. N. Weinstein, and Y. Pommier. Molecular interaction maps of bioregulatory networks : a general rubric for systems biology. *Mol. Biol. Cell*, 17(1) :1–13, Jan 2006.

[80] N. Kravchenko-Balasha, A. Levitzki, A. Goldstein, et al. On a fundamental structure of gene networks in living cells. *Proc. Natl. Acad. Sci. U.S.A.*, 109(12) :4702–4707, Mar 2012.

[81] M. Kretzschmar, J. Doody, I. Timokhina, and J. Massague. A mechanism of repression of TGFbeta/ Smad signaling by oncogenic Ras. *Genes Dev.*, 13(7) :804–816, Apr 1999.

[82] L. Kubiczkova, L. Sedlarikova, R. Hajek, and S. Sevcikova. TGF-Î² - an excellent servant but a bad master. *J Transl Med*, 10 :183, 2012.

[83] A. Lachmann, H. Xu, J. Krishnan, et al. ChEA : transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, 26(19) :2438–2444, Oct 2010.

[84] N. Le Novere, M. Hucka, H. Mi, et al. The Systems Biology Graphical Notation. *Nat. Biotechnol.*, 27(8) :735–741, Aug 2009.

[85] Chen Li, Marco Donizelli, Nicolas Rodriguez, et al. BioModels Database : An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 4 :92, Jun 2010.

[86] M. Li, H. Zhang, J. X. Wang, and Y. Pan. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Syst Biol*, 6 :15, 2012.

[87] Z. C. Li, Y. H. Lai, L. L. Chen, et al. Identification of human protein complexes from local sub-graphs of protein-protein interaction network based on random forest with topological structure features. *Anal. Chim. Acta*, 718 :32–41, Mar 2012.

[88] J. C. Lindon, J. K. Nicholson, E. Holmes, et al. Summary recommendations for standardization and reporting of metabolic analyses. *Nat. Biotechnol.*, 23(7) :833–838, Jul 2005.

[89] P. V. Luc and P. Tempst. PINdb : a database of nuclear protein complexes from human and yeast. *Bioinformatics*, 20(9) :1413–1415, Jun 2004.

[90] X. Ma and L. Gao. Discovering protein complexes in protein interaction networks via exploring the weak ties effect. *BMC Syst Biol*, 6 Suppl 1 :S6, Jul 2012.

[91] D. Machado, R. S. Costa, M. Rocha, et al. Modeling formalisms in Systems Biology. *AMB Express*, 1 :45, 2011.

[92] Elisabetta De Maria, François Fages, Aurélien Rizk, and Sylvain Soliman. Design, optimization and predictions of a coupled model of the cell cycle, circadian clock, {DNA} repair system, irinotecan metabolism and exposure control under temporal logic constraints. *Theoretical Computer Science*, 412(21) :2108 – 2127, 2011.

¡ce :title¿Selected Papers from the 7th International Conference on Computational Methods in Systems Biology¡/ce :title¿ ¡xocs :full-name¿7th International Conference on Computational Methods in Systems Biology¡/xocs :full-name¿.

[93] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq : an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18(9) :1509–1517, Sep 2008.

[94] J. Massague. TGFbeta in Cancer. *Cell*, 134(2) :215–230, Jul 2008.

[95] J. Massague. TGFÎ² signalling in context. *Nat. Rev. Mol. Cell Biol.*, 13(10) :616–630, Oct 2012.

[96] I. Medina, J. Carbonell, L. Pulido, et al. Babelomics : an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res.*, 38(Web Server issue) :W210–213, Jul 2010.

[97] L. Mendoza. A network model for the control of the differentiation process in Th cells. *BioSystems*, 84(2) :101–114, May 2006.

[98] H. Mi, A. Muruganujan, and P. D. Thomas. PANTHER in 2013 : modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, 41(Database issue) :D377–386, Jan 2013.

[99] R. Milo, S. Shen-Orr, S. Itzkovitz, et al. Network motifs : simple building blocks of complex networks. *Science*, 298(5594) :824–827, Oct 2002.

[100] E. Morin. *Introduction à la pensée complexe*. Collection Communication et complexité. ESF, 1990.

[101] A. Moustakas and D. Kardassis. Regulation of the human p21/WAF1/Cip1 promoter in hepatic cells by functional interactions between Sp1 and Smad family members. *Proc. Natl. Acad. Sci. U.S.A.*, 95(12) :6733–6738, Jun 1998.

[102] J. Nakabayashi and A. Sasaki. A mathematical model of the stoichiometric control of Smad complex formation in TGF-beta signal transduction pathway. *J. Theor. Biol.*, 259(2) :389–403, Jul 2009.

[103] A. Naldi, D. Berenguier, A. Faure, et al. Logical modelling of regulatory networks with GINsim 2.3. *BioSystems*, 97(2) :134–139, Aug 2009.

[104] John Von Neumann. *Theory of Self-Reproducing Automata*. University of Illinois Press, Champaign, IL, USA, 1966.

[105] K. Oda and H. Kitano. A comprehensive map of the toll-like receptor signaling network. *Mol. Syst. Biol.*, 2 :2006.0015, 2006.

[106] K. Oda, Y. Matsuoka, A. Funahashi, and H. Kitano. A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol. Syst. Biol.*, 1 :2005.0010, 2005.

[107] S. Orchard, S. Kerrien, S. Abbani, et al. Protein interaction data curation : the International Molecular Exchange (IMEx) consortium. *Nat. Methods*, 9(4) :345–350, Apr 2012.

[108] S. Orchard, L. Salwinski, S. Kerrien, et al. The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.*, 25(8) :894–898, Aug 2007.

[109] K. Pardali and A. Moustakas. Actions of TGF-beta as tumor suppressor and pro-metastatic factor in human cancer. *Biochim. Biophys. Acta*, 1775(1) :21–62, Jan 2007.

[110] L. Paulevé. *Modélisation, Simulation et Vérification des Grands Réseaux de Régulation Biologique*. PhD thesis, École centrale de Nantes, 2011.

[111] M. Pogson, M. Holcombe, R. Smallwood, and E. Qwarnstrom. Introducing spatial information into predictive NF-kappaB modelling–an agent-based approach. *PLoS ONE*, 3(6) :e2367, 2008.

[112] S. Purvi, C. North, and K. Duca. Visualizing biological pathways : requirements analysis, systems evaluation and research agenda. *Information Visualization*, 4(3) :191–205, 2005.

[113] A. Rauzy. Guarded transition systems : a new states/events formalism for reliability studies. *Journal of Risk and Reliability*, 222(4) :495–505, 2008.

[114] M. C. Reed, R. L. Thomas, J. Pavisic, et al. A mathematical model of glutathione metabolism. *Theor Biol Med Model*, 5 :8, 2008.

[115] A. Regev and E. Shapiro. Cells as computation. *Nature*, 419(6905) :343, Sep 2002.

[116] M. Rehm, H. J. Huber, H. Dussmann, and J. H. Prehn. Systems analysis of effector caspase activation and its control by X-linked inhibitor of apoptosis protein. *EMBO J.*, 25(18) :4338–4349, Sep 2006.

[117] E. Remy, P. Ruet, L. Mendoza, D. Thieffry, and C. Chaouiya. *From Logical Regulatory Graphs to Standard Petri Nets : Dynamical Roles and Functionality of Feedback Circuits*, volume 4230 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2006.

[118] H. Resat, J. A. Ewald, D. A. Dixon, and H. S. Wiley. An integrated model of epidermal growth factor receptor trafficking and signal transduction. *Biophys. J.*, 85(2) :730–743, Aug 2003.

[119] M. Rodriguez-Fernandez, P. Mendes, and J. R. Banga. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *BioSystems*, 83(2-3) :248–265, 2006.

[120] C. Rohr, W. Marwan, and M. Heiner. Snoopy–a unifying Petri net framework to investigate biomolecular networks. *Bioinformatics*, 26(7) :974–975, Apr 2010.

[121] P. Romero, J. Wagg, M. L. Green, et al. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.*, 6(1) :R2, 2005.

[122] D. Ruths, M. Muller, J. T. Tseng, L. Nakhleh, and P. T. Ram. The signaling petri net-based simulator : a non-parametric strategy for characterizing the dynamics of cell-specific signaling networks. *PLoS Comput. Biol.*, 4(2) :e1000005, Feb 2008.

[123] A. Sackmann, M. Heiner, and I. Koch. Application of Petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics*, 7 :482, 2006.

[124] J. Saez-Rodriguez, L. G. Alexopoulos, J. Epperlein, et al. Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Mol. Syst. Biol.*, 5 :331, 2009.

[125] Julio Saez-Rodriguez, Leonidas G. Alexopoulos, MingSheng Zhang, et al. Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. *Cancer Research*, 71(16) :5400–5411, 2011.

[126] M. Safran, I. Dalah, J. Alexander, et al. GeneCards Version 3 : the human gene integrator. *Database (Oxford)*, 2010 :baq020, 2010.

[127] R. Saito, M. E. Smoot, K. Ono, et al. A travel guide to Cytoscape plugins. *Nat. Methods*, 9(11) :1069–1076, Nov 2012.

[128] L. Salwinski, C. S. Miller, A. J. Smith, et al. The Database of Interacting Proteins : 2004 update. *Nucleic Acids Res.*, 32(Database issue) :D449–451, Jan 2004.

[129] R. Samaga, J. Saez-Rodriguez, L. G. Alexopoulos, P. K. Sorger, and S. Klamt. The logic of EGFR/ErbB signaling : theoretical properties and analysis of high-throughput data. *PLoS Comput. Biol.*, 5(8) :e1000438, Aug 2009.

[130] S. Sasagawa, Y. Ozaki, K. Fujita, and S. Kuroda. Prediction and validation of the distinct dynamics of transient and sustained ERK activation. *Nat. Cell Biol.*, 7(4) :365–373, Apr 2005.

[131] C. F. Schaefer, K. Anthony, S. Krupa, et al. PID : the Pathway Interaction Database. *Nucleic Acids Res.*, 37(Database issue) :D674–679, Jan 2009.

[132] B. Schmierer, A. L. Tournier, P. A. Bates, and C. S. Hill. Mathematical modeling identifies Smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system. *Proc. Natl. Acad. Sci. U.S.A.*, 105 :6608–6613, May 2008.

[133] B. Schoeberl, C. Eichler-Jonsson, E. D. Gilles, and G. Muller. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotechnol.*, 20(4) :370–375, Apr 2002.

[134] M. E. Smoot, K. Ono, J. Ruscheinski, P. L. Wang, and T. Ideker. Cytoscape 2.8 : new features for data integration and network visualization. *Bioinformatics*, 27(3) :431–432, Feb 2011.

[135] C. Stark, B. J. Breitkreutz, T. Reguly, et al. BioGRID : a general repository for interaction datasets. *Nucleic Acids Res.*, 34(Database issue) :D535–539, Jan 2006.

[136] J. R. Stern, S. Christley, O. Zaborina, J. C. Alverdy, and G. An. Integration of TGF-Î²- and EGFR-based signaling pathways using an agent-based model of epithelial restitution. *Wound Repair Regen*, 20(6) :862–871, 2012.

[137] L. Stromback, V. Jakoniene, H. Tan, and P. Lambrix. Representing, storing and accessing molecular interaction data : a review of models and tools. *Brief. Bioinformatics*, 7(4) :331–338, Dec 2006.

[138] D. Szklarczyk, A. Franceschini, M. Kuhn, et al. The STRING database in 2011 : functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, 39(Database issue) :D561–568, Jan 2011.

[139] Carolyn Talcott. Symbolic modeling of signal transduction in pathway logic. In *Proceedings of the 38th conference on Winter simulation*, WSC '06, pages 1656–1665. Winter Simulation Conference, 2006.

[140] C. F. Taylor, N. W. Paton, K. S. Lilley, et al. The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.*, 25(8) :887–893, Aug 2007.

[141] R. Thomas. Boolean formalization of genetic control circuits. *J. Theor. Biol.*, 42(3) :563–585, Dec 1973.

[142] M. Tian, J. R. Neil, and W. P. Schiemann. Transforming growth factor-Î² and the hallmarks of cancer. *Cell. Signal.*, 23(6) :951–962, Jun 2011.

[143] S. Troncale, F. Tahi, D. Campard, J. P. Vannier, and J. Guespin. Modeling and simulation with Hybrid Functional Petri Nets of the role of interleukin-6 in human early haematopoiesis. *Pac Symp Biocomput*, pages 427–438, 2006.

[144] J. C. Venter, M. D. Adams, E. W. Myers, et al. The sequence of the human genome. *Science*, 291(5507) :1304–1351, Feb 2001.

[145] Santiago Videla, Carito Guziolowski, Federica Eduati, et al. Revisiting the training of logic models of protein signaling networks with asp. In David Gilbert and Monika Heiner, editors, *Computational Methods in Systems Biology*, Lecture Notes in Computer Science, pages 342–361. Springer Berlin Heidelberg, 2012.

[146] J. M. Vilar, R. Jansen, and C. Sander. Signal processing in the TGF-beta superfamily ligand-receptor network. *PLoS Comput. Biol.*, 2(1) :e3, Jan 2006.

[147] A. C. Villeger, S. R. Pettifer, and D. B. Kell. Arcadia : a visualization tool for metabolic pathways. *Bioinformatics*, 26(11) :1470–1471, Jun 2010.

[148] C. von Mering, R. Krause, B. Snel, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887) :399–403, May 2002.

[149] L. M. Wakefield, D. M. Smith, T. Masui, C. C. Harris, and M. B. Sporn. Distribution and modulation of the cellular receptor for transforming growth factor-beta. *J. Cell Biol.*, 105(2) :965–975, Aug 1987.

[150] D. C. Walker, J. Southgate, G. Hill, et al. The epitheliome : agent-based modelling of the social behaviour of cells. *BioSystems*, 76(1-3) :89–100, 2004.

[151] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq : a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1) :57–63, Jan 2009.

[152] D. S. Wishart, C. Knox, A. C. Guo, et al. HMDB : a knowledgebase for the human metabolome. *Nucleic Acids Res.*, 37(Database issue) :D603–610, Jan 2009.

[153] E. Wong, B. Baur, S. Quader, and C. H. Huang. Biological network motif detection : principles and practice. *Brief. Bioinformatics*, 13(2) :202–215, Mar 2012.

[154] K. Yagi, M. Furuhashi, H. Aoki, et al. c-myc is a downstream target of the Smad pathway. *J. Biol. Chem.*, 277(1) :854–861, Jan 2002.

[155] C. Y. Yang, C. H. Chang, Y. L. Yu, et al. PhosphoPOINT : a comprehensive human kinase interactome and phospho-protein database. *Bioinformatics*, 24(16) :14–20, Aug 2008.

[156] N. Yu, J. Seo, K. Rho, et al. hiPathDB : a human-integrated pathway database with facile visualization. *Nucleic Acids Res.*, 40(Database issue) :797–802, Jan 2012.

[157] R. Zhang, M. V. Shah, J. Yang, et al. Network model of survival signaling in large granular lymphocyte leukemia. *Proc. Natl. Acad. Sci. U.S.A.*, 105(42) :16308–16313, Oct 2008.

[158] Z. Zi, Z. Feng, D. A. Chapnick, et al. Quantitative analysis of transient and sustained transforming growth factor-Î² signaling dynamics. *Mol. Syst. Biol.*, 7 :492, May 2011.

[159] A. Zoubarev, K. M. Hamer, K. D. Keshav, et al. Gemma : a resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics*, 28(17) :2272–2273, Sep 2012.

Cinquième partie

Annexes

Congrès international

20th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)
Lieu : Long Beach, ÉTATS-UNIS
Date : 13 au 17 Juillet 2012
Titre de la présentation : Modeling cell signaling pathways with discrete dynamical systems : Application to the Transforming Growth Factor β (TGF-β) dependent Epithelial-Mesenchymal Transition
Co-auteurs : Michel Le Borgne, Nathalie Théret
Format : poster

Congrès national

XXXI séminaire de la Société Francophone de Biologie Théorique
Lieu : Autrans, FRANCE
Date : 16 au 18 Mai 2011
Titre de la présentation : Développement de méthodes informatiques pour des simulations prédictives sur la signalisation du TGF-β
Co-auteurs : Michel Le Borgne, Nathalie Théret
Format : présentation orale

Quatrième colloque Génomique fonctionnelle du Foie
Lieu : Bordeaux, FRANCE
Date : 14 au 16 Mars 2012
Titre de la présentation : Fibrose hépatique et biologie des systèmes : modélisation de la signalisation du TGF-β
Co-auteurs : Michel Le Borgne, Nathalie Théret
Format : poster

Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM) 2012
Lieu : Rennes, FRANCE
Date : 3 au 6 Juillet 2012
Titre de la présentation : Modeling cell signaling pathways with discrete dynamical systems : Application to Transforming Growth Factor β (TGF-β) signaling
Co-auteurs : Michel Le Borgne, Nathalie Théret
Format : poster

Conférences

École d'été : Modeling Complex Biological Systems in the Context of Genomics
Date et Lieu : 23 au 27 Mai 2011, Sophia Antipolis, FRANCE
Titre de la présentation : Development of computational methods for predictive simulation of TGF-β
Co-auteurs : Michel Le Borgne, Nathalie Théret
Format : poster

Biologie Intégrative et Génomique dans le Grand Ouest (BIGOU)
Date et Lieu : 7 au 9 Novembre 2011, Auray, FRANCE
Titre de la présentation : Cadbiom un projet de modélisation des réseaux de signalisation
Co-auteurs : Michel Le Borgne, Nathalie Théret
Format : présentation orale

2e Journée Recherche Organisée par la Faculté de Pharmacie, Chimie, Biologie, Mathématiques et Physique, la recherche fondamentale au service du Médicament et de la Santé
Date et Lieu : 9 Janvier 2012, Rennes, FRANCE
Titre de la présentation : Cadbiom-Chart De la conception à l'interrogation de modèles discrets des voies de signalisations cellulaires
Co-auteurs : Michel Le Borgne, Nathalie Théret
Format : poster

Réunion Annuelle ANR BioTempo
Date et Lieu : 6 au 7 Mars 2012, Nantes, FRANCE
Titre de la présentation : Analyzing Large Models of TGF-β with Cadbiomand the Process Hitting
Co-auteurs : Loïc Paulevé, Michel Le Borgne, Nathalie Théret
Format : présentation orale

École Jeunes Chercheurs en Informatique Mathématique 2012
Date et Lieu : 19 au 23 Mars 2012, Rennes, FRANCE
Titre de la présentation : Modeling cell signaling networks
Co-auteurs : Michel Le Borgne, Nathalie Théret
Format : Présentation orale

Réunion Annuelle ANR BioTempo
Date et Lieu : 10 au 11 Avril 2013, Paris, FRANCE
Titre de la présentation : A guarded transition approach to integrate the human cell signaling pathways into a single unified dynamic model
Co-auteurs : Michel Le Borgne, Nathalie Théret
Format : présentation orale

Séminaire BCM
Date et Lieu : 23 Mai 2013, Grenoble, FRANCE
Titre de la présentation : A guarded transition approach to integrate the human cell signaling pathways into a single unified dynamic model
Co-auteurs : Michel Le Borgne, Nathalie Théret
Format : présentation orale

Enseignement et vulgarisation

Enseignement

UE Préparation au Certificat Informatique et Internet (C2I)
Niveau : Licence 1 (Biologie)
Nombre d'heures : 40
Contenu : TP éditeur de texte, tableur et outil de présentation.

UE Programmation Orienté Objet (POO)
Niveau : Master 1 (Bio-Informatique et Génomique)
Nombre d'heures : 24
Contenu : Concepts Objet, programmation JAVA, utilisation API, conception interface graphique.

UE De la génomique à la biologie intégrative
Niveau : Master 1 et Master 2 (Bio-Informatique et Génomique)
Nombre d'heures : 1
Contenu : Intervention lors de la table ronde des étudiants
du master BIG "le doctorat et les métiers de chercheur et enseignant-chercheur".

Vulgarisation

Festival de vulgarisation scientifique : "Science en courts"
Lieu : Rennes, FRANCE
Date : 21 Avril 2011 Contenu : Réalisation et projections aux lycéens et au grand public du très court métrage de vulgarisation scientifique "Bioinformaticus", réalisé avec Charles Bettembourg et Nicolas Maillet.

Journées "À la découverte de la Recherche"
Lieu : Redon et Rennes FRANCE
Date : 22 Avril 2011 et 26 Avril 2012
Contenu : Intervention dans les Lycées St Sauveur de Redon et Jean Macé de Rennes dans le but de présenter mon cursus universitaire, et le quotidien d'un doctorant.

Formations

Écoles thématiques

École d'été : Modeling Complex Biological Systems in the Context of Genomics
Lieu : Sophia Antipolis, FRANCE
Date : 23 au 27 Mai 2011

Biologie Intégrative et Génomique dans le Grand Ouest (BIGOU)
Lieu : Auray, FRANCE
Date : 7 au 9 Novembre 2011

École Jeunes Chercheurs en Informatique Mathématique 2012
Lieu : Rennes, FRANCE
Date : 19 au 23 Mars 2012

UE Rennes I

Logique
Lieu : Rennes, FRANCE
Nombre d'heures : 10
Contenu : Introduction à la logique propositionelle et logique du prémier ordre.

Autres Formations

"les nouveaux visuels : la forme au service du fond"
Lieu : Rennes, FRANCE
Nombre d'heures : 5
Contenu : Enseignement visant à savoir mettre en forme des ensembles de données complexes de façon intuitive et attractive.

Festival de vulgarisation scientifique : "Science en courts"
Lieu : Rennes, FRANCE
Nombre d'heures : 6
Contenu :
- Écriture d'un scénario (2h)
- Le droit d'auteur appliqué à l'audiovisuel (2h)
- Prise en main de la caméra (2h)

Gratifications

Prix Pierre Delattre
Prix de la meilleur présentation orale d'un doctorant, remis lors du XXXI ème séminaire de la Société Francophone de Biologie Théorique 2011 à Autrans, FRANCE.

Bourse région Bretagne
Bourse de la région Bretagne rémise pour une Communications lors du congrès ISMB 2012 à Long Beach, ÉTATS-UNIS.

Travel fellowships of International Society for Computational Biology (ISCB) student council symposium
Bourse de voyage remise aux étudiants lors de leurs participation au concil des étudiant à ISCB 2012 à Long Beach, ÉTATS-UNIS.

Autres contributions

Logiciel Cadbiom

Certificat *Inter Deposit Digital Number* (IDDN) pour le logiciel Cadbiom

Inter Deposit Digital Number

Certificat délivré par

Agence pour la Protection des Programmes

54 rue de Paradis - 75010 PARIS - FRANCE / T. +33(0)1 40 35 03 03 / F. +33(0)1 40 38 96 43

IDDN.FR.001.080005.000.S.P.2013.000.20600

(1) (2) (3) (4) (5) (6) (7) (8) (9) (10)

Pour l'œuvre : **CADBIOM-chart**

Identité du(des) titulaire(s) de droits :

UNIVERSITÉ DE RENNES 1
2 rue du Thabor
CS 46510
35065 RENNES CEDEX
Siren : 193509361

INSERM - Institut National de la Santé et de la Recherche Médicale
101 rue de Tolbiac
75654 PARIS CEDEX 13
Siren : 180036048

Ecole des Hautes Etudes en Santé Publique
Avenue du Professeur Léon Bernard
CS 74312
35043 RENNES CEDEX
Siren : 130003627

Adhérent sous le numéro : **92.35.2610**
Support utilisé : 1 CD-Rom en double exemplaire

Logibox conservée par l'adhérent : 71571
Logibox conservée par l'APP : 71572

Le titulaire *

Fait à Paris, le 18/02/2013

Le Président

AGENCE POUR LA PROTECTION DES PROGRAMMES

(1) Inter Deposit Digital Number
(2) Nationalité de l'œuvre
(3) Numéro de l'organisme d'enregistrement
(4) Numéro d'ordre de l'enregistrement
(5) Numéro de version
(6) Type d'enregistrement
(7) Type de l'œuvre
(8) Année d'enregistrement
(9) Zone réservée (Clé d'intégrité)
(10) Classe de produit

** Le titulaire s'engage à informer l'APP de toute cession ou aliénation,totale ou partielle, de ses droits de propriété intellectuelle.*
Seules les inscriptions de type S et C permettent un éventuel accès au programme source.

T. +33 (0)1 40 35 03 03 - F. +33 (0)1 40 38 96 43
54 rue de Paradis 75010 PARIS | Siren 385 385 844 – APE 9499Z
APP.ASSO.FR

950-021-L

Paulevé *et al*, CAV, 2013

Papier accepté dans la conférence Computer Aided Verification (CAV) 2013 : Under-approximating Cut Sets for Reachability in Large Scale Automata Networks

Under-approximating Cut Sets for Reachability in Large Scale Automata Networks

Loïc Paulevé[1], Geoffroy Andrieux[2], Heinz Koeppl[1]

[1] ETH Zürich, Switzerland.
[2] IRISA Rennes, France.

Abstract. In the scope of discrete finite-state models of interacting components, we present a novel algorithm for identifying sets of local states of components whose activity is necessary for the reachability of a given local state. If all the local states from such a set are disabled in the model, the concerned reachability is impossible.
Those sets are referred to as cut sets and are computed from a particular abstract causality structure, so-called Graph of Local Causality, inspired from previous work and generalised here to finite automata networks. The extracted sets of local states form an under-approximation of the complete minimal cut sets of the dynamics: there may exist smaller or additional cut sets for the given reachability.
Applied to qualitative models of biological systems, such cut sets provide potential therapeutic targets that are proven to prevent molecules of interest to become active, up to the correctness of the model. Our new method makes tractable the formal analysis of very large scale networks, as illustrated by the computation of cut sets within a Boolean model of biological pathways interactions gathering more than 9000 components.

1 Introduction

With the aim of understanding and, ultimately, controlling physical systems, one generally constructs dynamical models of the known interactions between the components of the system. Because parts of those physical processes are ignored or still unknown, dynamics of such models aim at over-approximating the real system dynamics: any (observed) behaviour of the real system has to have a matching behaviour in the abstract model, the converse being potentially false. In such a setting, a valuable contribution of formal methods on abstract models of physical systems resides in the ability to prove the impossibility of particular behaviours.

Given a discrete finite-state model of interacting components, such as an automata network, we address here the computation of sets of local states of components that are necessary for reaching a local state of interest from a partially determined initial global state. Those sets are referred to as *cut sets*. Informally, each path leading to the reachability of interest has to involve, at one point, at least one local state of a cut set. Hence, disabling in the model all the local

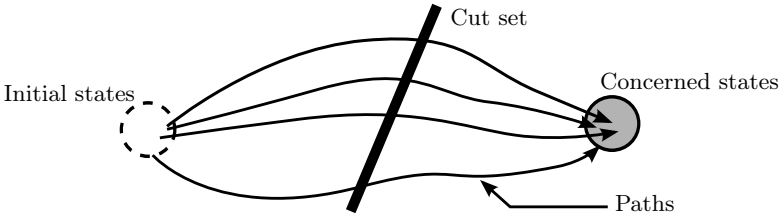


Fig. 1. A cut set is composed of local states that are involved in *all* paths from delimited initial states to concerned states. Disabling all the local states from such a cut set necessarily breaks the concerned reachability in the model.

states referenced in one cut set should prevent the occurrence of the concerned reachability from delimited initial states. This is illustrated by Fig. 1.

Applied to a model of a biological system where the reachability of interest is known to occur, such cut sets provide potential coupled therapeutic targets to control the activity of a particular molecule (for instance using gene knock-in/out). The contrary implies that the abstract model is not an over-approximation of the concrete system.

Contribution. In this paper, we present a new algorithm to extract sets of local states that are necessary to achieve the concerned reachability within a finite automata network. Those sets are referred to as cut sets, and we limit ourselves to N-sets, *i.e.* having a maximum cardinality of N.

The finite automata networks we are considering are closely related to 1-safe Petri nets [1] having mutually exclusive places. They subsume Boolean and discrete networks [9,24,17,2], synchronous or asynchronous, that are widely used for the qualitative modelling of biological interaction networks.

A naive, but complete, algorithm could enumerate all potential candidate N-sets, disable each of them in the model, and then perform model-checking to verify if the targeted reachability is still verified. If not, the candidate N-set is a cut set. This would roughly leads to m^N tests, where m is the total number of local states in the automata network. Considering that the model-checking within automata networks is PSPACE-complete [5], this makes such an approach intractable on large networks.

The proposed algorithm aims at being tractable on systems composed of a very large number of interacting components, but each of them having a small number of local states. Our method principally overcomes two challenges: prevent a complete enumeration of candidate N-sets; and prevent the use of model-checking to verify if disabling a set of local states break the concerned reachability. It inherently handles partially-determined initial states: the resulting cut N-set of local states are proven to be necessary for the reachability of the local state of interest from *any* of the supplied global initial states.

The computation of the cut N-sets takes advantage of an abstraction of the formal model which highlights some steps that are necessary to occur prior to the

verification of a given reachability property. This results in a causality structure called a *Graph of Local Causality* (GLC), which is inspired by [16], and that we generalise here to automata networks. Such a GLC has a size polynomial with the total number of local states in the automata network, and exponential with the number of local states within one automata. Given a GLC, our algorithm propagates and combines the cut N-sets of the local states referenced in this graph by computing unions or products, depending on the disjunctive or conjunctive relations between the necessary conditions for their reachability. The algorithm is proven to converge in the presence of dependence cycles.

In order to demonstrate the scalability of our approach, we have computed cut N-sets within a very large Boolean model of a biological network relating more than 9000 components. Despite the highly combinatorial dynamics, a prototype implementation manages to compute up to the cut 5-sets within a few minutes. To our knowledge, this is the first time such a formal dynamical analysis has been performed on such a large dynamical model of biological system.

Related work and limitations. Cut sets are commonly defined upon graphs as set of edges or vertices which, if removed, disconnect a given pair of nodes [21]. For our purpose, this approach could be directly applied to the global transition graph to identify local states or transitions for which the removal would disconnect initial states from the targeted states. However, the combinatorial explosion of the state space would make it intractable for large interacting systems.

The aim of the presented method is somehow similar to the generation of minimal cut sets in fault trees [13,22] used for reliability analysis, as the structure representing reachability causality contains both *and* and *or* connectors. However, the major difference is that we are here dealing with cyclic directed graphs which prevents the above mentioned methods to be straightforwardly applied.

Klamt *et al.* have developed a complete method for identifying minimal cut sets (also called intervention sets) dedicated to biochemical reactions networks, hence involving cycling dependencies [10]. This method has been later generalised to Boolean models of signalling networks [19]. Those algorithms are mainly based on the enumeration of possible candidates, with techniques to reduce the search space, for instance by exploiting symmetry of dynamics. Whereas intervention sets of [10,19] can contain either local states or reactions, our cut sets are only composed of local states.

Our method follows a different approach than [10,19] by not relying on candidate enumeration but computing the cut sets directly on an abstract structure derived statically from the model, which should make tractable the analysis of very large networks. The comparison with [19] is detailed in Subsect. 4.1.

In addition, our method is generic to any automata network, but relies on an abstract interpretation of dynamics which leads to under-approximating the cut sets for reachability: by ignoring certain dynamical constraints, the analysis can miss several cut sets and output cut sets that are not minimal for the concrete model. Finally, although we focus on finding the cut sets for the reachability of

only *one* local state, our algorithm computes the cut sets for the (independent) reachability of all local states referenced in the GLC.

Outline. Sect. 2 introduces a generic characterisation of the *Graph of Local Causality* with respect to automata networks; Sect. 3 states and sketches the proof of the algorithm for extracting a subset of N-sets of local states necessary for the reachability of a given local state. Sect. 4 discusses the application to systems biology by comparing with the related work and applying our new method to a very large scale model of biological interactions. Finally, Sect. 5 discusses the results presented and some of their possible extensions.

Notations. $\wedge$ and $\vee$ are the usual logical *and* and *or* connectors. $[1;n] = \{1, \cdots, n\}$. Given a finite set A, $\#A$ is the cardinality of A; $\wp(A)$ is the power set of A; $\wp^{\leq N}(A)$ is the set of all subsets of A with cardinality at most N. Given sets $A^1, \ldots, A^n$, $\bigcup_{i\in[1;n]} A^i$ is the union of those sets, with the empty union $\bigcup_\emptyset \triangleq \emptyset$; and $A^1 \times \cdots \times A^n$ is the usual Cartesian product. Given sets of sets $B^1, \ldots, B^n \in \wp(\wp(A))$, $\widetilde{\prod}_{i\in[1;n]} B^i \triangleq B^1 \tilde{\times} \cdots \tilde{\times} B^n \in \wp(\wp(A))$ is the *sets of sets product* where $\{e_1, \ldots, e_n\} \tilde{\times} \{e'_1, \ldots, e'_m\} \triangleq \{e_i \cup e'_j \mid i \in [1;n] \wedge j \in [1;m]\}$. In particular $\forall (i,j) \in [1;n] \times [1;m]$, $B^i \tilde{\times} B^j = B^j \tilde{\times} B^i$ and $\emptyset \tilde{\times} B^i = \emptyset$. The empty sets of sets product $\widetilde{\prod}_\emptyset \triangleq \{\emptyset\}$. If $M : A \mapsto B$ is a mapping from elements in A to elements in B, $M(a)$ is the value in B mapped to $a \in A$; $M\{a \mapsto b\}$ is the mapping M where $a \in A$ now maps to $b \in B$.

2 Graph of Local Causality

We first give basic definitions of automata networks, local state disabling, context and local state reachability; then we define the local causality of an objective (local reachability), and the *Graph of Local Causality*. A simple example is given at the end of the section.

2.1 Finite Automata Networks

We consider a network of automata $(\Sigma, S, \mathcal{L}, T)$ which relates a finite number of interacting finite state automata Σ (Def. 1). The global state of the system is the gathering of the local state of composing automata. A transition can occur if and only if all the local states sharing a common transition label $\ell \in L$ are present in the global state $s \in S$ of the system. Such networks characterize a class of 1-safe Petri Nets [1] having groups of mutually exclusive places, acting as the automata. They allow the modelling of Boolean networks and their discrete generalisation, having either synchronous or asynchronous transitions.

Definition 1 (Automata Network $(\Sigma, S, \mathcal{L}, T)$). *An automata network is defined by a tuple $(\Sigma, S, \mathcal{L}, T)$ where*

- $\Sigma = \{a, b, \ldots, z\}$ *is the finite set of automata identifiers;*
- *For any* $a \in \Sigma$, $S(a) = [1; k_a]$ *is the finite set of local states of automaton* a*;* $S = \prod_{a \in \Sigma} [1; k_a]$ *is the finite set of global states.*
- $\mathcal{L} = \{\ell_1, \ldots, \ell_m\}$ *is the finite set of transition labels;*
- $T = \{a \mapsto T_a \mid a \in \Sigma\}$*, where* $\forall a \in \Sigma, T_a \subset [1; k_a] \times \mathcal{L} \times [1; k_a]$*, is the mapping from automata to their finite set of local transitions.*

We note $i \xrightarrow{\ell} j \in T(a) \overset{\Delta}{\Leftrightarrow} (i, \ell, j) \in T_a$ *and* $a_i \xrightarrow{\ell} a_j \in T \overset{\Delta}{\Leftrightarrow} i \xrightarrow{\ell} j \in T(a)$.
$\forall \ell \in \mathcal{L}$*, we note* ${}^\bullet\ell \overset{\Delta}{=} \{a_i \mid a_i \xrightarrow{\ell} a_j \in T(a)\}$ *and* $\ell^\bullet \overset{\Delta}{=} \{a_j \mid a_i \xrightarrow{\ell} a_j \in T(a)\}$.
The set of local states is defined as $\mathbf{LS} \overset{\Delta}{=} \{a_i \mid a \in \Sigma \wedge i \in [1; k_a]\}$.
The global transition relation $\rightarrow \subset S \times S$ *is defined as:*

$$s \rightarrow s' \overset{\Delta}{\Leftrightarrow} \exists \ell \in \mathcal{L} : \forall a_i \in {}^\bullet\ell, s(a) = a_i \wedge \forall a_j \in \ell^\bullet, s'(a) = a_j \\ \wedge \forall b \in \Sigma, S(b) \cap {}^\bullet\ell = \emptyset \Rightarrow s(b) = s'(b).$$

Given an automata network $\mathcal{S}ys = (\Sigma, S, \mathcal{L}, T)$ and a subset of its local states $ls \subseteq \mathbf{LS}$, $\mathcal{S}ys \ominus ls$ refers to the system where all the local states ls have been disabled, i.e. they can not be involved in any transition (Def. 2).

Definition 2 (Local states disabling). *Given* $\mathcal{S}ys = (\Sigma, S, \mathcal{L}, T)$ *and* $ls \in \wp(\mathbf{LS})$*,* $\mathcal{S}ys \ominus ls \overset{\Delta}{=} (\Sigma, S, \mathcal{L}', T')$ *where* $\mathcal{L}' = \{\ell \in \mathcal{L} \mid ls \cap {}^\bullet\ell = \emptyset\}$ *and* $T' = \{a_i \xrightarrow{\ell} a_j \in T \mid \ell \in \mathcal{L}'\}$.

From a set of acceptable initial states delimited by a *context* ς (Def. 3), we say a given local state $a_j \in \mathbf{LS}$ is reachable if and only if there exists a finite number of transitions in $\mathcal{S}ys$ leading to a global state where a_j is present (Def. 4).

Definition 3 (Context ς). *Given a network* $(\Sigma, S, \mathcal{L}, T)$*, a context* ς *is a mapping from each automaton* $a \in \Sigma$ *to a non-empty subset of its local states:* $\forall a \in \Sigma, \varsigma(a) \in \wp(S(a)) \wedge \varsigma(a) \neq \emptyset$.

Definition 4 (Local state reachability). *Given a network* $(\Sigma, S, \mathcal{L}, T)$ *and a context* ς*, the local state* $a_j \in \mathbf{LS}$ *is* reachable *from* ς *if and only if* $\exists s_0, \ldots, s_m \in S$ *such that* $\forall a \in \Sigma, s_0(a) \in \varsigma(a)$*, and* $s_0 \rightarrow \cdots \rightarrow s_m$*, and* $s_m(a) = j$.

2.2 Local Causality

Locally reasoning within one automaton a, the global reachability of a_j from ς can be expressed as the reachability of a_j from a local state $a_i \in \varsigma(a)$. This local reachability specification is referred to as an *objective* noted $a_i \rightarrow^* a_j$ (Def. 5)

Definition 5 (Objective). *Given a network* $(\Sigma, S, \mathcal{L}, T)$*, the reachability of local state* a_j *from* a_i *is called an* objective *and is denoted* $a_i \rightarrow^* a_j$*. The set of all objectives is referred to as* $\mathbf{Obj} \overset{\Delta}{=} \{a_i \rightarrow^* a_j \mid (a_i, a_j) \in \mathbf{LS} \times \mathbf{LS}\}$.

Given an objective $P = a_i \rightarrow^* a_j \in \mathbf{Obj}$, we define $\mathsf{sol}(P)$ the *local causality* of P (Def. 6): each $ls \in \mathsf{sol}(P)$ is a set of local states that may be involved for the reachability of a_j from a_i; ls is referred to as a (local) solution for P. $\mathsf{sol}(P)$ is sound as soon as the disabling of at least one local state in *each* solution makes the reachability of a_j impossible from any global state containing a_i (Property 1). It implies that if $\mathsf{sol}(P) = \{\{a_i\} \cup ls^1, \ldots, ls^m\}$ is sound, $\mathsf{sol}'(P) = \{ls^1, \ldots, ls^m\}$ is also sound. $\mathsf{sol}(a_i \rightarrow^* a_j) = \emptyset$ implies that a_j can never be reached from a_i, and $\forall a_i \in \mathbf{LS}, \mathsf{sol}(a_i \rightarrow^* a_i) \triangleq \{\emptyset\}$.

Definition 6. $\mathsf{sol} : \mathbf{Obj} \mapsto \wp(\wp(\mathbf{LS}))$ *is a mapping from objectives to sets of sets of local states such that* $\forall P \in \mathbf{Obj}, \forall ls \in \mathsf{sol}(P), \nexists ls' \in \mathsf{sol}(P), ls \neq ls'$ *such that* $ls' \subset ls$. *The set of these mappings is noted* $\mathbf{Sol} \triangleq \{\langle P, ls\rangle \mid ls \in \mathsf{sol}(P)\}$.

Property 1 (sol *soundness).* $\mathsf{sol}(a_i \rightarrow^* a_j) = \{ls^1, \ldots, ls^n\}$ is a sound set of solutions for the network $\mathcal{S}ys = (\Sigma, S, \mathcal{L}, T)$ if and only if $\forall kls \in \widetilde{\prod}_{i\in[1;n]}\{ls^i\}$, a_j is not reachable in $\mathcal{S}ys \ominus kls$ from any state $s \in S$ such that $s(a) = i$.

In the rest of this paper we assume that Property 1 is satisfied, and consider sol computation out of the scope of this paper.

Nevertheless, we briefly describe a construction of a sound $\mathsf{sol}(a_i \rightarrow^* a_j)$ for an automata network $(\Sigma, S, \mathcal{L}, T)$; an example is given at the end of this section. This construction generalises the computation of GLC from the Process Hitting framework, a restriction of network of automata depicted in [16]. For each acyclic sequence $a_i \xrightarrow{\ell_1} \ldots \xrightarrow{\ell_m} a_j$ of local transitions in $T(a)$, and by defining $\mathsf{ext}_a(\ell) \triangleq \{b_j \in \mathbf{LS} \mid b_j \xrightarrow{\ell} b_k \in T, b \neq a\}$, we set $ls \in \widetilde{\prod}_{\ell\in\{\ell_1,\ldots,\ell_m \mid \mathsf{ext}_a(\ell)\neq\emptyset\}}\{\mathsf{ext}_a(\ell)\} \Rightarrow ls \in \mathsf{sol}(a_i \rightarrow^* a_j)$, up to supersets removing. One can easily show that Property 1 is verified with such a construction. The complexity of this construction is exponential in the number of local states within one automaton and polynomial in the number of automata. Alternative constructions may also provide sound (and not necessarily equal) sol.

2.3 Graph of Local Causality

Given a local state $a_j \in \mathbf{LS}$ and an initial context ς, the reachability of a_i is equivalent to the realization of any objective $a_i \rightarrow^* a_j$, with $a_i \in \varsigma(a)$. By definition, if a_j is reachable from ς, there exists $ls \in \mathsf{sol}(a_i \rightarrow^* a_j)$ such that, $\forall b_k \in ls$, b_k is reachable from ς.

The (directed) *Graph of Local Causality* (GLC, Def. 7) relates this recursive reasoning from a given set of local states $\omega \subseteq \mathbf{LS}$ by linking every local state a_j to all objectives $a_i \rightarrow^* a_j, a_i \in \varsigma(a)$; every objective P to its solutions $\langle P, ls\rangle \in \mathbf{Sol}$; every solution $\langle P, ls\rangle$ to its local states $b_k \in ls$. A GLC is said to be sound if sol is sound for all referenced objectives (Property 2).

Definition 7 (Graph of Local Causality). *Given a context* ς *and a set of local states* $\omega \subseteq \mathbf{LS}$, *the* Graph of Local Causality *(GLC)* $\mathcal{A}^\omega_\varsigma \triangleq (V^\omega_\varsigma, E^\omega_\varsigma)$, *with*

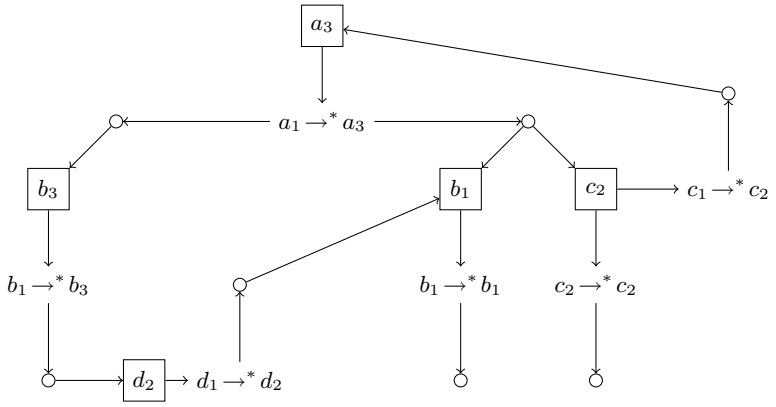


Fig. 2. Example of Graph of Local Causality that is sound for the automata network defined in Example 1

$V_\varsigma^\omega \subseteq \mathbf{LS} \cup \mathbf{Obj} \cup \mathbf{Sol}$ *and* $E_\varsigma^\omega \subseteq V_\varsigma^\omega \times V_\varsigma^\omega$*, is the smallest structure satisfying:*

$$\begin{aligned} \omega &\subseteq V_\varsigma^\omega \\ a_i \in V_\varsigma^\omega \cap \mathbf{LS} &\Leftrightarrow \{(a_i, a_j \rightarrow^* a_i) \mid a_j \in \varsigma\} \subseteq E_\varsigma^\omega \\ a_i \rightarrow^* a_j \in V_\varsigma^\omega \cap \mathbf{Obj} &\Leftrightarrow \{(a_i \rightarrow^* a_j, \langle a_i \rightarrow^* a_j, ls\rangle) \mid \langle a_i \rightarrow^* a_j, ls\rangle \in \mathbf{Sol}\} \subseteq E_\varsigma^\omega \\ \langle P, ls\rangle \in V_\varsigma^\omega \cap \mathbf{Sol} &\Leftrightarrow \{(\langle P, ls\rangle, a_i) \mid a_i \in ls\} \subseteq E_\varsigma^\omega \quad . \end{aligned}$$

Property 2 (Sound Graph of Local Causality). A GLC $\mathcal{A}_\varsigma^\omega$ is *sound* if, $\forall P \in V_\varsigma^\omega \cap \mathbf{Obj}$, $\mathsf{sol}(P)$ is sound.

This structure can be constructed starting from local states in ω and by iteratively adding the imposed children. It is worth noticing that this graph can contain cycles. In the worst case, $\#V_\varsigma^\omega = \#\mathbf{LS} + \#\mathbf{Obj} + \#\mathbf{Sol}$ and $\#E_\varsigma^\omega = \#\mathbf{Obj} + \#\mathbf{Sol} + \sum_{\langle P, ls\rangle \in \mathbf{Sol}} \#ls$.

Example 1. Fig. 2 shows an example of GLC. Local states are represented by boxed nodes and elements of **Sol** by small circles.

For instance, such a GLC is sound for the following automata network $(\Sigma, S, \mathcal{L}, T)$, with initial context $\varsigma = \{a \mapsto \{1\}; b \mapsto \{1\}; c \mapsto \{1, 2\}; d \mapsto \{2\}\}$:

$$\begin{aligned} \Sigma &= \{a, b, c, d\} & \mathcal{L} &= \{\ell_1, \ell_2, \ell_3, \ell_4, \ell_5, \ell_6\} \\ S(a) &= [1; 3] & T(a) &= \{1 \xrightarrow{\ell_2} 2; 2 \xrightarrow{\ell_3} 3; 1 \xrightarrow{\ell_1} 3; 3 \xrightarrow{\ell_4} 2\} \\ S(b) &= [1; 3] & T(b) &= \{1 \xrightarrow{\ell_2} 2; 1 \xrightarrow{\ell_5} 3; 1 \xrightarrow{\ell_6} 1; 3 \xrightarrow{\ell_1} 2\} \\ S(c) &= [1; 2] & T(c) &= \{1 \xrightarrow{\ell_4} 2; 2 \xrightarrow{\ell_3} 1\} \\ S(d) &= [1; 2] & T(d) &= \{1 \xrightarrow{\ell_6} 2; 2 \xrightarrow{\ell_5} 1\} \end{aligned}$$

For example, within automata a, there are two acyclic sequences from 1 to 3: $1 \xrightarrow{\ell_2} 2 \xrightarrow{\ell_3} 3$ and $1 \xrightarrow{\ell_1} 3$. Hence, if a_3 is reached from a_1, then necessarily, one of these two sequences has to be used (but not necessarily consecutively). For each of these transitions, the transition label is shared by exactly one local state in another automaton: b_1, c_2, b_3 for ℓ_2, ℓ_3, ℓ_1, respectively. Therefore, if a_3 is reached from a_1, then necessarily either both b_1 and c_2, or b_3 have been reached before. Hence $\mathsf{sol}(a_1 \rightarrow^* a_3) = \{\{b_1, c_2\}, \{b_3\}\}$ is sound, as disabling either b_1 and b_3, or c_2 and b_3, would remove any possibility to reach a_3 from a_1.

3 Necessary Local States for Reachability

We assume a global sound GLC $\mathcal{A}^\omega_\varsigma = (V^\omega_\varsigma, E^\omega_\varsigma)$, with the usual accessors for the direct relations of nodes:

$$\mathsf{children} : V^\omega_\varsigma \mapsto \wp(V^\omega_\varsigma) \qquad \mathsf{parents} : V^\omega_\varsigma \mapsto \wp(V^\omega_\varsigma)$$
$$\mathsf{children}(n) \triangleq \{m \in V^\omega_\varsigma \mid (n,m) \in E^\omega_\varsigma\} \quad \mathsf{parents}(n) \triangleq \{m \in V^\omega_\varsigma \mid (m,n) \in E^\omega_\varsigma\}$$

Given a set of local states $\mathcal{O}bs \subseteq \mathbf{LS}$, this section introduces an algorithm computing upon $\mathcal{A}^\omega_\varsigma$ the set $\mathbb{V}(a_i)$ of minimal cut N-sets of local states in $\mathcal{O}bs$ that are necessary for the independent reachability of each local state $a_i \in \mathbf{LS} \cap V^\omega_\varsigma$. The minimality criterion actually states that $\forall ls \in \mathbb{V}(a_i)$, there is no different $ls' \in \mathbb{V}(a_i)$ such that $ls' \subset ls$.

Assuming a first valuation $\mathbb{V}$ (Def. 8) associating to each node its cut N-sets, the cut N-sets for the node n can be refined using $\mathsf{update}(\mathbb{V}, n)$ (Def. 9):

- if n is a solution $\langle P, ls\rangle \in \mathbf{Sol}$, it is sufficient to prevent the reachability of *any* local state in ls to cut n; therefore, the cut N-sets results from the union of the cut N-sets of n children (all local states).
- If n is an objective $P \in \mathbf{Obj}$, all its solutions (in $\mathsf{sol}(P)$) have to be cut in order to ensure that P is not realizable: hence, the cut N-sets result from the product of children cut N-sets (all solutions).
- If n is a local state a_i, it is sufficient to cut all its children (all objectives) to prevent the reachability of a_i from any state in the context ς. In addition, if $a_i \in \mathcal{O}bs$, $\{a_i\}$ is added to the set of its cut N-sets.

Definition 8 (Valuation $\mathbb{V}$). *A* valuation $\mathbb{V} : V^\omega_\varsigma \mapsto \wp(\wp^{\leq N}(\mathcal{O}bs))$ *is a mapping from each node of* $\mathcal{A}^\omega_\varsigma$ *to a set of* N*-sets of local states.* $\mathbf{Val}$ *is the set of all valuations.* $\mathbb{V}_0 \in \mathbf{Val}$ *refers to the valuation such that* $\forall n \in V^\omega_\varsigma, \mathbb{V}_0(n) = \emptyset$.

Definition 9 ($\mathsf{update} : \mathbf{Val} \times V^\omega_\varsigma \mapsto \mathbf{Val}$).

$$\mathsf{update}(\mathbb{V}, n) \triangleq \begin{cases} \mathbb{V}\{n \mapsto \zeta^N(\bigcup_{m \in \mathsf{children}(n)} \mathbb{V}(m))\} & \text{if } n \in \mathbf{Sol} \\ \mathbb{V}\{n \mapsto \zeta^N(\widetilde{\prod}_{m \in \mathsf{children}(n)} \mathbb{V}(m))\} & \text{if } n \in \mathbf{Obj} \\ \mathbb{V}\{n \mapsto \zeta^N(\widetilde{\prod}_{m \in \mathsf{children}(n)} \mathbb{V}(m))\} & \text{if } n \in \mathbf{LS} \setminus \mathcal{O}bs \\ \mathbb{V}\{n \mapsto \zeta^N(\{\{a_i\}\} \cup \widetilde{\prod}_{m \in \mathsf{children}(n)} \mathbb{V}(m))\} & \text{if } n \in \mathbf{LS} \cap \mathcal{O}bs \end{cases}$$

Algorithm 1 $\mathcal{A}^{\omega}_{\varsigma}$-Minimal-Cut-NSets

```
1: M ← V_ς^ω
2: 𝕍 ← 𝕍_0
3: while M ≠ ∅ do
4:     n ← arg min_{m∈M}{rank(m)}
5:     M ← M \ {n}
6:     𝕍' ← update(𝕍, n)
7:     if 𝕍'(n) ≠ 𝕍(n) then
8:         M ← M ∪ parents(n)
9:     end if
10:    𝕍 ← 𝕍'
11: end while
12: return 𝕍
```

where $\zeta^N(\{e_1,\ldots,e_n\}) \triangleq \{e_i \mid i \in [1;n] \wedge \#e_i \leq N \wedge \nexists j \in [1;n], j \neq i, e_j \subset e_i\}$, e_i *being sets,* $\forall i \in [1;n]$.

Starting with $\mathbb{V}_0$, one can repeatedly apply update on each node of $\mathcal{A}^{\omega}_{\varsigma}$ to refine its valuation. Only nodes where one of their children value has been modified should be considered for updating.

Hence, the order of nodes updates should follow the topological order of the GLC, where children have a lower rank than their parents (i.e., children are treated before their parents). If the graph is actually acyclic, then it is sufficient to update the value of each node only once. In the general case, *i.e.* in the presence of Strongly Connected Components (SCCs) — nodes belonging to the same SCC have the same rank —, the nodes within a SCC have to be iteratively updated until the convergence of their valuation.

Algorithm 1 formalizes this procedure where $\mathsf{rank}(n)$ refers to the topological rank of n, as it can be derived from Tarjan's strongly connected components algorithm [23], for example. The node $n \in V^{\omega}_{\varsigma}$ to be updated is selected as being the one having the least rank amongst the nodes to update (delimited by $\mathcal{M}$). In the case where several nodes with the same lowest rank are in $\mathcal{M}$, they can be either arbitrarily or randomly picked. Once picked, the value of n is updated. If the new valuation of n is different from the previous, the parents of n are added to the list of nodes to update (lines 6-8 in Algorithm 1).

Lemma 1 states the convergence of Algorithm 1 and Theorem 1 its correctness: for each local state $a_i \in V^{\omega}_{\varsigma} \cap \mathbf{LS}$, each set of local states $kls \in \mathbb{V}(a_i)$ (except $\{a_i\}$ singleton) references the local states that are all necessary to reach prior to the reachability of a_i from any state in ς. Hence, if all the local states in kls are disabled in $\mathcal{S}ys$, a_i is not reachable from any state in ς.

Lemma 1. $\mathcal{A}^{\omega}_{\varsigma}$-Minimal-Cut-NSets *always terminates.*

Proof. Remarking that $\wp(\wp^{\leq N}(\mathcal{O}bs))$ is finite, defining a partial ordering such that $\forall v, v' \in \wp(\wp^{\leq N}(\mathcal{O}bs)), v \succeq v' \overset{\Delta}{\Leftrightarrow} \zeta^N(v) = \zeta^N(v \cup v')$, and noting $\mathbb{V}^k \in \mathbf{Val}$ the valuation after k iterations of the algorithm, it is sufficient to prove that

Node	rank	$\mathbb{V}$
$\langle b_1 \rightarrow^* b_1, \emptyset \rangle$	1	$\emptyset$
$b_1 \rightarrow^* b_1$	2	$\emptyset$
b_1	3	$\{\{b_1\}\}$
$\langle d_1 \rightarrow^* d_2, \{b_1\} \rangle$	4	$\{\{b_1\}\}$
$d_1 \rightarrow^* d_2$	5	$\{\{b_1\}\}$
d_2	6	$\{\{b_1\}, \{d_2\}\}$
$\langle b_1 \rightarrow^* b_3, \{d_2\} \rangle$	7	$\{\{b_1\}, \{d_2\}\}$
$b_1 \rightarrow^* b_3$	8	$\{\{b_1\}, \{d_2\}\}$
b_3	9	$\{\{b_1\}, \{b_3\}, \{d_2\}\}$
$\langle a_1 \rightarrow^* a_3, \{b_3\} \rangle$	10	$\{\{b_1\}, \{b_3\}, \{d_2\}\}$
$\langle c_2 \rightarrow^* c_2, \emptyset \rangle$	11	$\emptyset$
$c_2 \rightarrow^* c_2$	12	$\emptyset$
c_2	13	$\{\{c_2\}\}$
$\langle a_1 \rightarrow^* a_3, \{b_1, c_2\} \rangle$	13	$\{\{b_1\}, \{c_2\}\}$
$a_1 \rightarrow^* a_3$	13	$\{\{b_1\}, \{b_3, c_2\}, \{c_2, d_2\}\}$
a_3	13	$\{\{a_3\}, \{b_1\}, \{b_3, c_2\}, \{c_2, d_2\}\}$
$\langle c_1 \rightarrow^* c_2, \{a_3\} \rangle$	13	$\{\{a_3\}, \{b_1\}, \{b_3, c_2\}, \{c_2, d_2\}\}$

Table 1. Result of the execution of Algorithm 1 on the GLC in Fig. 2

$\mathbb{V}^{k+1}(n) \succeq \mathbb{V}^k(n)$. Let us define $v_1, v_2, v_1', v_2' \in \wp(\wp^{\leq N}(\mathcal{O}bs))$ such that $v_1 \succeq v_1'$ and $v_2 \succeq v_2'$. We can easily check that $v_1 \cup v_2 \succeq v_1' \cup v_2'$ (hence proving the case when $n \in \mathbf{Sol}$). As $\zeta^N(v_1) = \zeta^N(v_1 \cup v_1') \Leftrightarrow \forall e_1' \in v_1', \exists e_1 \in v_1 : e_1 \subseteq e_1'$, we obtain that $\forall (e_1', e_2') \in v_1' \times v_2', \exists (e_1, e_2) \in v_1 \times v_2 : e_1 \subseteq e_1' \wedge e_2 \subseteq e_2'$. Hence $e_1 \cup e_2 \subseteq e_1' \cup e_2'$, therefore $\zeta^N(v_1 \tilde{\times} v_2 \cup v_1' \tilde{\times} v_2') = \zeta^N(v_1 \tilde{\times} v_2)$, i.e. $v_1 \tilde{\times} v_2 \succeq v_1' \tilde{\times} v_2'$; which proves the cases when $n \in \mathbf{Obj} \cup \mathbf{LS}$.

Theorem 1. *Given a GLC $\mathcal{A}_\varsigma^\omega = (V_\varsigma^\omega, E_\varsigma^\omega)$ which is sound for the automata network $\mathcal{S}ys$, the valuation $\mathbb{V}$ computed by $\mathcal{A}_\varsigma^\omega$-*Minimal-Cut-NSets *verifies: $\forall a_i \in \mathbf{LS} \cap V_\varsigma^\omega, \forall kls \in \mathbb{V}(a_i) \setminus \{\{a_i\}\}$, a_j is not reachable from ς within $\mathcal{S}ys \ominus kls$.*

Proof. By recurrence on the valuations $\mathbb{V}$: the above property is true at each iteration of the algorithm.

Example 2. Table 1 details the result of the execution of Algorithm 1 on the GLC defined in Fig. 2. Nodes receive a topological rank, identical ranks implying the belonging to the same SCC. The (arbitrary) scheduling of the updates of nodes within a SCC follows the order in the table. In this particular case, nodes are all visited once, as $\mathbb{V}(\langle c_2 \rightarrow^* c_2, \emptyset \rangle) \tilde{\times} \mathbb{V}(\langle c_1 \rightarrow^* c_2, \{a_3\} \rangle) = \emptyset$ (hence $\mathsf{update}(\mathbb{V}, c_2)$ does not change the valuation of c_2). Note that in general, several iterations of update may be required to reach a fixed point.

It is worth noticing that the GLC abstracts several dynamical constraints in the underlying automata networks, such as the ordering of transitions, or the synchronous updates of the global state. In that sense, GLC over-approximates the dynamics of the network, and the resulting cut sets are under-approximating the complete cut sets of the concrete model: any computed cut sets is a superset of a complete cut set (potentially equal).

4 Application to Systems Biology

Automata networks, as presented in Def. 1, subsume Boolean and discrete networks, synchronous and asynchronous, that are widely used for the qualitative modelling of dynamics of biological networks [9,24,17,2,7,18,6].

A cut set, as extracted by our algorithm, informs that at least one of the component in the cut set has to be present in the specified local state in order to achieve the wanted reachability. A local state can represent, for instance, an active transcription factor or the absence of a certain protein. It provides potential therapeutic targets if the studied reachability is involved in a disease by preventing all the local states of a cut set to act, for instance using gene knock-out or knock-in techniques.

We first discuss and compare our methodology with the *intervention sets* analysis within biological models developed by S. Klamt *et al.*, and provide some benchmarks on a few examples.

Thanks to the use of the intermediate GLC and to the absence of candidate enumeration, our new method makes tractable the cut sets analysis on very large models. We present a recent application of our results to the analysis of a very large scale Boolean model of biological pathway interactions involving 9000 components. To our knowledge, this is the first attempt of a formal dynamical analysis on such a large scale model.

4.1 Related Work

The general related work having been discussed in Sect. 1, we deepen here the comparison of our method with the closest related work: the analysis of *Intervention Sets* (ISs) [19]. Cut sets and ISs have a reversed logic: an IS specifies local states to enforce in order to ensure a particular behaviour to occur; a cut set specifies local states to disable in order to prevent a particular behaviour to occur. In the scope of Boolean models of signalling networks, ISs are computed for the reachability of a given fixed point (steady state) which can be partially defined. Their method is complete: all minimal ISs are computed.

Nevertheless, the semantics and the computation of ISs have some key differences with our computed cut sets. First, they focus only on the reachability of (logical) steady states, which is a stronger condition than the transient reachability that we are considering. Then, the steady states are computed using a three-valued logic which allows to cope with undefined (initial) local states, but which is different from the notion of context that we use in this paper for specifying the initial condition.

Such differences make difficult a proper comparison of inferred cut sets. We can however expect that any cut sets found by our method has a corresponding IS in the scope of Boolean networks with a single initial state.

To give a practical insight on the relation between the two methods, we compare the results for two signalling networks, both between a model specified with CellNetAnalyser [11] to compute ISs and a model specified in the Process Hitting framework, a particular restriction of asynchronous automata networks

[15], to compute our cut sets. Process Hitting models have been built in order to over-approximate the dynamics considered for the computation of ISs[3].

Tcell. Applied to a model of the T-cell receptor signalling between 40 components [12], we are interested in preventing the activation of the transcription factor $AP1$. For an instance of initial conditions, and limiting the computations to 3-sets, 31 ISs have been identified (28 1-sets, 3 2-sets, 0 3-set), whereas our algorithm found 29 cut sets (21 1-sets and 8 2-sets), which are all matching an IS (23 are identical, 6 strictly including ISs). ISs are computed in 0.69s while our algorithm under-approximates the cut sets in 0.006s. Different initial states give comparable results.

Egfr. Applied to a model of the epidermal growth factor receptor signalling pathway of 104 components [18], we are interested in preventing the activation of the transcription factor $AP1$. For an instance of initial conditions, and limiting the computations to 3-sets, 25 ISs have been identified (19 1-sets, 3 2-sets, 3 3-sets), whereas our algorithm found 14 cut sets (14 1-sets), which are all included in the ISs. ISs are computed in 98s while our algorithm under-approximates the cut sets in 0.004s. Different initial states give comparable results.

As expected with the different semantics of models and cut sets, resulting ISs matches all the cut sets identified by our algorithm, and provides substantially more sets. The execution time is much higher for ISs as they rely on candidate enumeration in order to provide complete results, whereas our method was designed to prevent such an enumeration but under-approximates the cut sets.

In order to appreciate the under-approximation done by our method at a same level of abstraction and with identical semantics, we compare the cut sets identified by our algorithm with the cut sets obtained using a naive, but complete, computation. The naive computation enumerates all cut set candidates and, for each of them, disable the local states in the model and perform model-checking to verify if the target local state is still reachable. In the particular case of these two models, and limiting the cut sets to 3 and 2-sets respectively for the sake of tractability, no additional cut set has been uncovered by the complete method. Such a good under-approximation could be partially explained by the restrictions imposed on the causality by the Process Hitting framework, making the GLC a tight over-approximation of the dynamics [16].

4.2 Very Large Scale Application to Pathway Interactions

In order to support the scalability of our approach, we apply the proposed algorithm to a very large model of biological interactions, actually extracted from the PID database [20] referencing various influences (complex formation, inductions (activations) and inhibitions, transcriptional regulation, etc.) between more than 9000 biological components (proteins, genes, ions, etc.).

[3] Models and scripts available at `http://loicpauleve.name/cutsets.tbz2`

Amongst the numerous biological components, the activation of some of them are known to control key mechanisms of the cell dynamics. Those activations are the consequence of intertwining signalling pathways and depend on the environment of the cell (represented by the presence of certain *entry-point* molecules). Uncovering the environmental and intermediate components playing a major role in these signalling dynamics is of great biological interest.

The full PID database has been interpreted into the Process Hitting framework, a subclass of asynchronous automata networks, from which the derivation of the GLC has been addressed in previous work [16]. The obtained model gathers components representing either biological entities modelled as boolean value (absent or present), or logical complexes. When a biological component has several competing regulators, the precise cooperations are not detailed in the database, so we use of two different interpretations: all (resp. one of) the activators and none (resp. all but one of) the inhibitors have to be present in order to make the target component present. This leads to two different discrete models of PID that we refer to as `whole_PID_AND` and `whole_PID_OR`, respectively.

Focusing on `whole_PID_OR`, the Process Hitting model relates more than 21000 components, either biological or logical, containing between 2 and 4 local states. Such a system could actually generate 2^{33874} states. 3136 components act as environment specification, which in our boolean interpretation leads to 2^{3136} possible initial states, assuming all other components start in the absent state.

We focus on the (independent) reachability of active SNAIL transcription factor, involved in the epithelial to mesenchymal transition [14], and of active p15INK4b and p21CIP1 cyclin-dependent kinase inhibitors involved in cell cycle regulation [3]. The GLC relates 20045 nodes, including 5671 component local states (biological or logical); it contains 6 SCCs with at least 2 nodes, the largest being composed of 10238 nodes and the others between 20 and 150.

Table 2 shows the results of a prototype implementation[4] of Algorithm 1 for the search of up to the 6-sets of biological component local states. One can observe that the execution time grows very rapidly with N compared to the number of visited nodes. This can be explained by intermediate nodes having a large set of cut N-sets leading to a costly computation of products.

While the precise biological interpretation of identified N-sets is out of the scope of this paper, we remark that the order of magnitude of the number of cut sets can be very different (more than 1000 cut 6-sets for SNAIL; none cut 6-sets for p21CIP1, except the gene that produces this protein). It supports a notion of robustness for the reachability of components, where the less cut sets, the more robust the reachability to various perturbations.

Applied to the `whole_PID_AND` model, our algorithm find in general much more cut N-sets, due to the conjunctive interpretation. This brings a significant increase in the execution time: the search up to the cut 5-sets took 1h, and the 6-sets leads to an out-of-memory failure.

[4] Implemented as part of the PINT software – `http://process.hitting.free.fr`
Models and scripts available at `http://loicpauleve.name/cutsets.tbz2`

Model	N	Visited nodes	Exec. time	Nb. of resulting N-sets		
				$SNAIL_1$	$p15INK4b_1$	$p21CIP1_1$
whole_PID_OR	1	29022	0.9s	1	1	1
	2	36602	1.6s	+6	+6	+0
	3	44174	5.4s	+0	+92	+0
	4	54322	39s	+30	+60	+0
	5	68214	8.3m	+90	+80	+0
	6	90902	2.6h	+930	+208	+0

Table 2. Results for the computation of cut N-sets for 3 local states. For each N, only the number of additional N-sets is displayed.

The very large number of involved components makes intractable the naive exact algorithm consisting in enumerating all possible N-sets candidates and verifying the concerned reachability using model-checking. Similarly, making such a model fit into other frameworks, such as CellNetAnalyser (see previous subsection) is a challenging task, and might be considered as future work.

5 Discussion

We presented a new method to efficiently compute cut sets for the reachability of a local state of a component within networks of finite automata from any state delimited by a provided so-called context. Those cut sets are sets of automata local states such that disabling the activity of all local states of a cut set guarantees to prevent the reachability of the concerned local state. Automata networks are commonly used to represent the qualitative dynamics of interacting biological systems, such as signalling networks. The computation of cut sets can then lead to propose potential therapeutic targets that have been formally identified from the model for preventing the activation of a particular molecule.

The proposed algorithm works by propagating and combining the cut sets of local states along a *Graph of Local Causality* (GLC), that we introduce here upon automata networks. A GLC relates the local states that are necessary to occur prior to the reachability of the concerned local state. Several constructions of a GLC are generally possible and depend on the semantics of the model. We gave an example of such a construction for automata networks. That GLC has a size polynomial in the total number of local states in the network, and exponential in the number of local states within one automaton. Note that the core algorithm for computing the cut sets only requires as input a GLC satisfying a soundness property that can be easily extended to discrete systems that are more general than the automata networks considered here.

The computed cut sets are an under-approximation of the complete cut sets as the GLC abstracts several dynamical constraints from the underlying concrete model: any computed cut sets is a superset of a concrete cut set (potentially equal), some can be missed. Our algorithm prevents a costly enumeration of the potential sets of candidates, and aims at being tractable on very large networks.

A prototype implementation of our algorithm has been successfully applied to the extraction of cut sets from a Boolean model of a biological system involving more than 9000 interacting components. To our knowledge this is the first attempt of such a dynamical analysis for such large biological models. We note that most of the computation time is due to products between large sets of cut N-sets. To partially address this issue, we use of prefix trees to represent set of sets on which we have specialized operations to stick to sets of N-sets (Appendix A[5]). There is still room for improvement as our prototype does not implement any caching or variable re-ordering.

The work presented in this paper can be extended in several ways, notably with a *posterior enlarging of the cut sets*. Because the algorithm computes the cut N-sets for each node in the GLC, it is possible to construct *a posteriori* cut sets with a greater cardinality by chaining them. For instance, let $kps \in \mathbb{V}(a_i)$ be a cut N-set for the reachability of a_i, for each $b_j \in kps$ and $kps' \in \mathbb{V}(b_j)$, $(kps \setminus \{b_j\}) \cup kps'$ is a cut set for a_i. In our biological case study, this method could be recursively applied until cut sets are composed of states of automata only acting for the environmental input.

With respect to the defined computation of cut N-sets, one could also derive *static reductions* of the GLC. Indeed, some particular nodes and arcs of the GLC can be removed without affecting the final valuation of nodes. A simple example are nodes representing objectives having no solution: such nodes can be safely removed as they bring no candidate N-sets for parents processes. These reductions conduct to both speed-up of the proposed algorithm but also to more compact representations for the reachability causality.

In addition of providing potential targets to prevent the occurrence of some behaviours, it may be crucial to ensure that the modified system keep satisfying important dynamical properties, as it is tackled with constrained minimal cut sets [4,8]. Currently, such constraints could be verified a posteriori in order to filter the computed cut sets that break important dynamics requirements. Taking advantage of those constraints during the computation is hence an promising research direction for large-scale applications.

Acknowledgements. LP and GA acknowledge the partial support of the French National Agency for Research (ANR-10-BLANC-0218 BioTempo project).

References

1. Bernardinello, L., De Cindio, F.: A survey of basic net models and modular net classes. In: Rozenberg, G. (ed.) Advances in Petri Nets 1992, Lecture Notes in Computer Science, vol. 609, pp. 304–351. Springer Berlin / Heidelberg (1992)
2. Bernot, G., Cassez, F., Comet, J.P., Delaplace, F., Müller, C., Roux, O.: Semantics of biological regulatory networks. Electronic Notes in Theoretical Computer Science 180(3), 3 – 14 (2007)
3. Drabsch, Y., Ten Dijke, P.: TGF-β signalling and its role in cancer progression and metastasis. Cancer Metastasis Rev. 31(3-4), 553–568 (Dec 2012)

[5] Available at `http://loicpauleve.name/CAV2013-SI.pdf`

4. Hädicke, O., Klamt, S.: Computing complex metabolic intervention strategies using constrained minimal cut sets. Metabolic Engineering 13(2), 204 – 213 (2011)
5. Harel, D., Kupferman, O., Vardi, M.Y.: On the complexity of verifying concurrent transition systems. Information and Computation 173(2), 143 – 161 (2002)
6. Hinkelmann, F., Laubenbacher, R.: Boolean models of bistable biological systems. Discrete and Continuous Dynamical Systems - Series S 4(6), 1443 – 1456 (2011)
7. de Jong, H.: Modeling and simulation of genetic regulatory systems: A literature review. Journal of Computational Biology 9, 67–103 (2002)
8. Jungreuthmayer, C., Zanghellini, J.: Designing optimal cell factories: integer programming couples elementary mode analysis with regulation. BMC Systems Biology 6(1), 103 (2012)
9. Kauffman, S.A.: Metabolic stability and epigenesis in randomly connected nets. Journal of Theoretical Biology 22, 437–467 (1969)
10. Klamt, S., Gilles, E.D.: Minimal cut sets in biochemical reaction networks. Bioinformatics 20(2), 226–234 (2004)
11. Klamt, S., Saez-Rodriguez, J., Gilles, E.: Structural and functional analysis of cellular networks with cellnetanalyzer. BMC Systems Biology 1(1), 2 (2007)
12. Klamt, S., Saez-Rodriguez, J., Lindquist, J., Simeoni, L., Gilles, E.: A methodology for the structural and functional analysis of signaling and regulatory networks. BMC Bioinformatics 7(1), 56 (2006)
13. Lee, W.S., Grosh, D.L., Tillman, F.A., Lie, C.H.: Fault tree analysis, methods, and applications - a review. IEEE Transactions on Reliability R-34, 194–203 (1985)
14. Moustakas, A., Heldin, C.H.: Signaling networks guiding epithelial-mesenchymal transitions during embryogenesis and cancer progression. Cancer Sci. 98(10), 1512–1520 (Oct 2007)
15. Paulevé, L., Magnin, M., Roux, O.: Refining dynamics of gene regulatory networks in a stochastic π-calculus framework. In: Transactions on Computational Systems Biology XIII, Lecture Notes in Comp Sci, vol. 6575, pp. 171–191. Springer (2011)
16. Paulevé, L., Magnin, M., Roux, O.: Static analysis of biological regulatory networks dynamics using abstract interpretation. Mathematical Structures in Computer Science 22(04), 651–685 (2012)
17. Richard, A.: Negative circuits and sustained oscillations in asynchronous automata networks. Advances in Applied Mathematics 44(4), 378 – 392 (2010)
18. Samaga, R., Saez-Rodriguez, J., Alexopoulos, L.G., Sorger, P.K., Klamt, S.: The logic of egfr/erbb signaling: Theoretical properties and analysis of high-throughput data. PLoS Comput Biol 5(8), e1000438 (08 2009)
19. Samaga, R., Von Kamp, A., Klamt, S.: Computing combinatorial intervention strategies and failure modes in signaling networks. Journal of Computational Biology 17(1), 39–53 (Jan 2010)
20. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., Buetow, K.H.: PID: The Pathway Interaction Database. Nucleic Acids Res. 37, D674–9 (2009)
21. Shier, D.R., Whited, D.E.: Iterative algorithms for generating minimal cutsets in directed graphs. Networks 16(2), 133–147 (1986)
22. Tang, Z., Dugan, J.: Minimal cut set/sequence generation for dynamic fault trees. In: Reliability and Maintainability, 2004 Annual Symposium - RAMS (2004)
23. Tarjan, R.: Depth-first search and linear graph algorithms. SIAM Journal on Computing 1(2), 146–160 (1972)
24. Thomas, R.: Boolean formalization of genetic control circuits. Journal of Theoretical Biology 42(3), 563 – 585 (1973)

Résumé

La signalisation cellulaire regroupe l'ensemble des mécanismes biologiques permettant à une cellule de répondre de façon adaptée à son microenvironnement. Pour ce faire, de nombreuses réactions biologiques entrent en jeux avec un important enchevêtrement, créant ainsi un réseau dont le comportement s'apparente à un système complexe. Le compréhension de la réponse cellulaire à une stimulation passe par le développement conjoint des techniques d'acquisition de données, et des méthodes permettant de formaliser ces données dans un modèle. C'est sur ce dernier point que s'inscrivent les travaux exposés dans cette thèse. Nous présentons ici deux approches visant à répondre à des questions de natures différentes sur la signalisation cellulaire. Dans la première nous utilisons un modèle différentiel pour étudier le rôle d'un nouvel interactant dans la voie canonique du TGF-β. Dans la seconde nous avons exploré la combinatoire de la signalisation cellulaire en développant un formalisme discret basé sur les transitions gardées. Cette approche regroupe l'interprétation de la base de données Pathway Interaction Database dans un unique modèle dynamique de propagation du signal. Des méthodes de simulations et d'analyses inspirées des techniques de vérification de modèles telles que l'atteignabilité et l'invariance ont été développées. En outre, nous avons étudié la régulation du cycle cellulaire en réponse à la signalisation, ainsi que la régulation des gènes de notre modèle en comparaison avec des données d'expressions.

Mots clés : biologie systémique ; facteur de croissance transformant bêta 1 ; transduction du signal cellulaire ; systèmes dynamiques ; bioinformatique

Abstract

Cell signaling contains the whole biological mecanisms allowing the response of a cell to its microenvironnement in an adapted way. Many extremely intertwinned biological reactions are involved in a network that behaves as a complex system. The understanding of cell response requires the development of data acquisition technics and methods to formalize data into models. This point is the main drive of this thesis. We present here two approaches in order to analyse different granularities of cell signaling. In the first one, we used differential model to study the role of a new component of TGF-β canonical pathway. In the second one, we explored the combinatorial complexity of cell signaling, developing a discrete formalism based on guarded transitions. In this approach, we interpreted the whole database Pathway Interaction Database into a single unified model of signal transduction. Simulation and analysis methods, such as reachability and invariance research, have been developed. The interests are presented through an application on cell cycle regulation by cell signaling, and a global analysis on the regulation of genes compared to experimental data.

Keywords : systems biology ; transforming growth factor beta 1 ; cellular signal transduction ; dynamic system ; bioinformatics