# Modélisation incrémentale des réseaux biologiques

Anastasia Yartseva Smidtas

Thèse soutenue le 12/12/2007

devant le jury composé de :

| | |
|---|---|
| Hanna Klaudel | Directrice de thèse |
| François Képès | Co-directeur de thèse |
| Hidde de Jong | Rapporteur |
| Olivier Gandrillon | Rapporteur |
| Dominique de Vienne | Examinateur |
| Raymond Devillers | Examinateur |
| Jacques Demongeot | Examinateur |

Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

# Table des matières

# Chapitre 1

# Introduction

Le dogme central de la biologie cellulaire établit la liaison qui existe entre le matériel génétique contenu dans la cellule et les protéines que cette cellule synthétise : ADN → ARN → Protéine. Ce passage du gène à la protéine se fait en deux étapes : tout d'abord un segment de la molécule d'ADN, correspondant à un gène, est copié sur un brin d'ARN, que l'on appelle ARN messager (ou ARNm), c'est la *transcription*. Puis ce brin d'ARNm est à son tour recopié, mais dans un langage différent, celui des acides aminés qui constituent les protéines, pour donner la séquence correspondant à la protéine synthétisée, c'est la *traduction*. La protéine ainsi formée pourra être utilisée directement quelque part dans la cellule (par exemple, pour la reconstruction de la membrane cellulaire) ; mais elle peut tout aussi bien servir de médium à la transcription d'un autre gène ailleurs sur l'ADN, en se fixant dessus par exemple. Seulement, une protéine a une durée de vie limitée, et si sa dégradation est plus rapide que sa synthèse, alors sa quantité dans le milieu diminuera ; moins il y a de cette protéine dans le milieu, moins elle a de chance de se retrouver fixée sur l'autre gène. Le premier gène peut dès lors accélérer, ou au contraire freiner, le processus de transcription de l'autre gène sur lequel sa protéine nouvellement créée se fixera, par la multiplication ou la diminution du nombre de protéines synthétisées dans le milieu (on dit

alors que le gène s'exprime). Cette interaction, ou plutôt cette influence du premier gène sur ce second gène au travers de la quantité de protéine produite, s'appelle la *régulation* (de gènes). Seulement, tout ne s'arrête pas forcément ici : en effet, ce second gène sera sans doute également traduit en une protéine, et cette protéine peut elle aussi se fixer pour réguler la transcription d'un autre gène, et ainsi de suite. L'enchaînement de ces régulations constitue un *réseau de régulation*.

Au cours des dernières décennies, la biologie moléculaire a accumulé une somme de connaissances sur les détails des mécanismes moléculaires dans les organismes. Depuis de nombreuses années les expérimentations biologiques ont permis de découvrir nombreuses interactions entre les gènes et les protéines, mais depuis le séquencage complets d'organismes et les génomes connus différentes techniques expérimentales plus ou moins automatisées et industrialisées ont permis de découvrir les interactions entre les espèces biochimiques qu'il peut y avoir dans la cellule. On parle de protéome, d'interactome, de transcriptome ou encore de réactome.

Récemment, la contribution de l'informatique a permis un saut important dans l'acquisition et l'interprétation des données génomiques. Cependant, ces avancées n'ont pas encore permis d'obtenir une compréhension globale des modules fonctionnels et des réseaux de régulation impliqués dans la physiologie cellulaire. Au début de ce travail de thèse, en 2002, de nombreuses questions se posaient encore : Comment sont structurés les réseaux biologiques ? Quelle est la fonction de l'architecture de réseaux de régulations et de leurs modules ? Quelles sont leurs propriétés dynamiques ? Quels sont les principes sous-jacents d'organisation des systèmes biologiques ? Comment l'environnement interagit avec ces réseaux et conduit à des états pathologiques homéostatiques ? Comment extrapoler les résultats obtenus par des modèles à d'autres cas ?

Pour aborder ces questions, les scientifiques de divers domaines mettent en commun leurs compétences repsectives : des biologistes, des physiciens, des informaticiens, des

mathématiciens et des ingénieurs se retrouvent impliqués dans la résolution des problèmes biologiques. Le chemin d'une nouvelle approche des sciences de la vie appelée, Biologie des Systèmes *(Systems Biology)* , se met en place et nécessite la création de nouveaux outils mieux adaptés aux questions posées et aux méthodes d'analyse développées.

La biologie des systèmes a deux buts. Le premier est d'obtenir de grandes quantités d'informations sur des systèmes. Ceci se fait habituellement via des expériences biologiques à haut débit qui produisent les données relativement superficielles et avec beaucoup de bruit. Cette accumulation de données peut être observée dans la transcriptomique en ce qui concerne les gènes activement transcrits, dans la protéomique (collection des protéines) ou bien dans la métabolomique (collection de tous les métabolites). La bioinformatique est une discipline en pleine croissance qui permet de traiter et d'analyser ces données des "omiques". Un autre but de la biologie des systèmes est de construire avec ces données une science traitant des principes d'opération des systèmes biologiques, basée sur les interactions entre les composantes. Manifestement, les systèmes biologiques sont bien organisés : ils sont très complexes mais hautement structurées et robustes. Cependant, leur organisation n'est souvent pas facilement compréhensible [23].

A travers la science, l'industrie, l'administration et le commerce, on observe des efforts gigantesques pour assembler des données dans des bases de données. La plupart de ces efforts sont basés sur la foi que rassembler et organiser les données en vaut la peine en soi. Des investissements importants ont été consentis pour décoder le génome humain et le fournir aux chercheurs en biologie. On s'attend à ce que ces données vont mener à la compréhension de l'expression des protéines et puis de la biologie et de la biochimie sous-jacentes. Maintenant la science a besoin de fournir les moyens d'exploiter ces grands volumes de données.

La biologie des systèmes cherche à comprendre les voies métaboliques ou génétiques en étudiant les interrelations (organisation ou structure) et les interactions (dynamique ou

comportement) des gènes, protéines ou métabolites. En croisant plusieurs échelle,s allant des molécules aux organismes, nous pouvons constater que les organismes, cellules, gènes et protéines sont définis comme des structures complexes interdépendantes et subordonnées. Cette définition rejoint la définition la plus générale d'un système en tant qu'ensemble de composants ou d'objets et des relations entre eux. Ainsi, la biologie des systèmes peut être vue comme une application de la théorie des systèmes à la biologie [183].

Bork et Serano [17] soulignent que la Biologie des Systèmes cherche à comprendre les systèmes biologiques de point de vue quantitatif, avec des activités qui vont de la collecte des données physiologiques (avec les détails quantitatifs sur les composants moléculaires du systèmes) jusqu'à la modélisation mathématique abstraite des processus biologiques.

La biologie des systèmes en est encore au stade où sa signification et son objet sont encore discutés. On peut s'en faire une idée en observant les titres d'articles clés publiés récemment. Ils parlent de l'ambiguïté innée de la catégorisation : "La signification de la Biologie des Systèmes" [103], "La Biologie des Systèmes au sens le plus large du terme" [176] et "Y a-t-il une recherche biologique après la Biologie des Systèmes ?" [16] ; de l'objet du domaine : "Les questions fondamentales de la Biologie des Systèmes" [141], "La Biologie des Systèmes : Ses pratiques et ses défis" [1] et "Biologie des systèmes, biologie intégrative et biologie prédictive" [123] ; de son avenir : "Où est la biologie des systèmes en 2005 ?" [120], "La biologie des systèmes : va-t-elle marcher ?" [155], "La biologie des systèmes pourra fonctionner quand on apprendra à comprendre les parties en termes de l'entier " [32] et "Vers les systèmes cellulaires en 4D" [17].

Pour atteindre le but que se fixe la biologie des systèmes, il est nécessaire d'établir des méthodologies et des techniques qui permettent de comprendre les systèmes biologiques en tant que tels, c'est à dire leur structure et leur dynamique, des méthodes pour les contrôler et des méthodes pour les concevoir ou les modifier afin de satisfaire les propriétés désirées [105]. Aderem [1] définit la biologie des systèmes comme étant basée sur des hypothèses

de nature globale, quantitative, itérative, intégrative et dynamique. Ces caractéristiques peuvent également très bien s'appliquer à des domaines voisins tels que la bioinformatique ou la biologie computationnelle. La biologie des systèmes cherche à modéliser, simuler et analyser les voies biochimiques [40, 91, 104]. Il a été espéré que l'approche au niveau des systèmes aiderait à dégager des vérités plus profondes sur la biologie cellulaire. La bioinformatique s'appuie sur des recherches pré-existantes sur la composition et les fonctions cellulaires et utilise des idées issues de la statistique et de l'apprentissage automatique afin de traiter et d'analyser ces données.

Les données sont généralement utilisées par les statisticiens sous la forme d'une matrice rectangulaire avec $N$ lignes et $D$ colonnes. Les lignes représentent les individus ou les différentes observations, et les colonnes – les attributs ou les variables. L'étude des maladies cardiaques dans *Framingham Heart study* [80] en fournit un bon exemple, dans lequel $N = 25000$ enregistrements sur différentes personnes avec $D = 100$ variables mesurées. Un autre exemple typique inclu l'étude des $D = 3800$ gènes de *S. cerevisiae* dans $N = 365$ expériences concernant des modifications génétiques, des expositions à divers produits chimiques et des conditions de croissance [9]. Le but est d'apprendre quels sont les gènes associés aux divers états et conditions expérimentales.

Ainsi, la méthodologie de base utilisée dans l'analyse de données classique (regroupement, classification, régression ou analyse des variables cachées) n'est plus applicable telle quelle car basée sur l'hypothèse que $D < N$ et que $N \to \infty$. En même temps, le cas où $D > N$ n'a rien d'anormal et devient maintenant un cas générique. Par exemple, pour un grand nombre des gènes, il peut y avoir peu de patients avec une maladie génétique donnée. Ce phénomène a été caractérisé par Richard Bellman en tant que "malédiction de la dimension" [11, 12]. Cependant, l'augmentation du nombre des dimensions peut aussi aider l'analyse mathématique, surtout dans le cadre de la théorie des probabilités. Par exemple, l'existence de plusieurs dimensions "identiques" selon lesquelles on pourrait

moyenner est un des outils fondamentaux. Il peut y avoir également des résultats, obtenus lorsque le nombre de dimensions tend vers l'infini, qui peut révéler l'existence d'une limite indépendante du nombre des dimensions. Il peut aussi y avoir des cas où les données à grand nombre de dimensions représentent un phénomène continu dans l'espace ou dans le temps comme une image ou une courbe, ce qui facilite l'analyse avec l'augmentation du nombre des dimensions considérées [41].

Imaginons maintenant la situation où parmi toutes les variables observées, il n'y a que peu de variables en rapport avec le phénomène étudié, mais on ne sait pas lesquelles : c'est un le problème typique de la fouille des données. Si on laisse dans le modèle trop de variables non pertinentes, on risque d'obtenir de mauvais résultats. Dans ce cas, l'idée est d'essayer de déterminer automatiquement les variables pertinentes et d'imposer un coût sur les modèles trop larges, souvent logarithmique en fonction du nombre des variables.

L'objectif de ce travail de thèse est de définir un cadre de modélisation incrémentale des réseaux biologiques qui pourrait aider les biologistes à comprendre la structure et la dynamique des systèmes modélisés.

Chi, Felovich et Glaser [28], qui travaillent dans le domaine de la formulation des problèmes, ont montré que le raisonnement causal d'un expert implique l'utilisation de connaissances spécifiques aux domaines pour choisir la formulation correcte du problème, qui mène ensuite à des solutions directes.

La biologie des systèmes est difficile à aborder et à comprendre. De nombreuses méthodes de modélisation ont été utilisées afin de représenter ses mécanismes et leur fonctionnement (modèles discrets, continus, quantitatifs, qualitatifs, bases de données). Toutes ces approches de modélisation ont toutefois quelques limites connues.

Afin de modéliser et simuler les mécanismes, de nombreux paramètres biologiques doivent être connus, et la modélisation est très gourmande en données expérimentales. Toutefois, les expériences biologiques pour les obtenir sont longues et coûteuses. Un système

d'équations différentielles incomplet par exemple ne sera pas d'une grande utilité, et l'ajout d'une réaction dans un système peut changer totalement son comportement. Ainsi, de petits modèles de mécanismes biologiques locaux existent, mais les modèles incomplets ne peuvent pas toujours profiter de nouvelles connaissances biologiques afin d'obtenir des résultats cohérents, à cause de la structure même de ces modèles.

Tous ces modèles supposent la consistance des données, ce qui permet des analyses rigoureuses et performantes. La garantie d'absence de données contradictoires dans les modèles demande au biologiste plus de travail avant de pouvoir intégrer les informations dans une base de connaissances. Cependant, en biologie, l'existence de nombreuses variables cachées est un fait indiscutable vu le nombre d'acteurs (molécules) impliqués dans le fonctionnement d'un système biologique. Deux expériences peuvent être réalisées par des équipes différentes, dans des conditions qui paraissent semblables, mais dont les résultats peuvent être différents, et cela de manière reproductible. Il ne s'agit donc pas de phénomènes stochastiques. Une condition expérimentale non décrite est souvent la cause de cette différence. Le bioinformaticien et le modélisateur doivent pouvoir travailler avec de telles données, sans devoir demander au biologiste de refaire toutes les expériences. Les données a priori incohérentes doivent être intégrées au modèle sans plus de formalité. Elles seront utiles de plusieurs manières. Les conditions non décrites des expériences permettront de faire des hypothèses sur ces dernières afin de choisir, suivant les cas d'études, les résultats pertinents. Générer plusieurs modèles, à partir d'un modèle ou d'un ensemble de modèles en faisant des hypothèses supplémentaires sur des variables cachées par exemple, est certainement une voie à explorer.

Souvent, les différentes approches de la modélisation des systèmes biologiques essayent de faire entrer les données dans le lit de Procruste[1] d'un formalisme donné. Cependant, si

[1]Dans la mythologie grecque, Procruste offre l'hospitalité aux voyageurs qu'il capturait pour les torturer ainsi : il les attache sur un lit, où ils doivent tenir exactement ; s'ils sont trop grands, il coupe les membres qui dépassent ; s'ils sont trop petits, il les étire jusqu'à ce qu'ils atteignent la taille requise. Procuste est

la question biologique change, le processus de modélisation du même système biologique doit alors être recommencé dans un autre formalisme, plus adapté à de nouvelles questions auxquelles on cherche à répondre. Ainsi, le choix du formalisme de modélisation peut être crucial pour la description efficace des systèmes biologiques afin d'éviter les changements inutiles du langage les décrivant et afin de permettre la réutilisation des modèles déjà construits.

Notre objectif était de construire un formalisme de modélisation pour les biologistes qui permet l'expression formelle des divers types de connaissances biologiques et la traduction de ces connaissances vers d'autres formalismes pour l'analyse ou la simulation. Nous avons cherché à construire un formalisme qui satisfasse les critères suivants :

- universalité : l'intégration des divers types de données biologiques disponibles aujourd'hui doit être possible ;
- parcimonie : la représentation des données doit être la plus simple possible ;
- incrémentalité : la construction de modèles plus complexes doit pouvoir se faire à partir de modèles simples ;
- précision : l'expression des relations doit se faire de manière (mathématiquement) non-ambiguë ;
- transposabilité : Il doit être possible de définir des règles formelles pour la traduction de l'information contenue dans le modèle vers d'autres formalismes communément utilisés en biologie.

Dans un tel formalisme, le modèle peut être vu plutôt comme une base des connaissances bien organisée qui contient l'information disponible sur un système biologique. Chaque unité d'information contenue dans le modèle, c'est à dire une information qui n'a plus de sens biologique si on la décompose encore, peut être appelée "une donnée". Dans cette approche, nous considérons qu'il n'y a ni données contradictoires, ni "mauvaises" données.

devenu le symbole du conformisme et de l'uniformisation. On parle couramment de « lit de Procuste » pour désigner toute tentative de réduire les hommes à un seul modèle, une seule façon de penser ou d'agir.

En d'autres termes, chaque mesure, chaque observation peut être révélatrice et intéressante en prenant en compte le contexte.

L'approche principale développée dans cette thèse et appelée MIN pour *Modular Interaction Network* – le réseau des interactions modulaires, est un formalisme conçu pour représenter des données biologiques. MIN possède la structure d'un graphe biparti et une représentation graphique associée, même si ce n'est pas son aspect primordial. MIN permet l'intégration de données microscopiques (les interactions moléculaires) et macroscopiques (les états observables du système), permettant ainsi de se positionner au niveau d'abstraction voulu. Cette abstraction permet d'éviter le problème d'explosion de la complexité du modèle. MIN a un nombre restreint de types de noeuds et d'arcs, ce qui permet de représenter des réseaux biologiques de manière simple, même si l'ensemble des informations détaillées peut également être stocké et retrouvé. MIN est adapté à la représentation des réseaux de régulation génétique ainsi que des réseaux métaboliques, avec leurs processus biologiques multi-moléculaires, et ceci d'un manière naturelle et incrémentale. MIN permet une traduction naturelle dans le formalisme de la modélisation logique de R. Thomas, dans les équations différentielles et dans les réseaux de Petri, par exemple. Ces traductions peuvent être effectuées à chaque étape de la modélisation.

Dans le chapitre deux, nous présenteons l'état de l'art des modèles statiques et dynamiques d'étude des réseaux biologiques.

Au cours de cette thèse, nous avons effectué des recherches basées sur les méthodologies existantes, telles que l'utilisation de graphes pour l'étude des réseaux biologiques, puis nous avons proposé des nouvelles techniques de modélisation regroupées au sein du formalisme MIN (Modular Interaction Network).

Tout d'abord dans le chapitre trois, nous avons utilisé des graphes simples pour étudier les interactions de deux réseaux biologiques partageant des noeuds communs. Nous avons utilisé les noeuds à l'interface des deux réseaux, appelé racines, pour structurer hiérarchiquement

les noeuds des réseaux en fonction de leurs distances aux racines. Cette approche a permis d' appréhender la manière dont un réseau biologique peut interagir avec son environnement, lui-même modélisé par un autre réseau.

Puis, au chapitre quatre, pour pallier les difficultés révélées lors de l'analyse des réseaux biologiques hétérogènes à l'aide des graphes simples, nous avons défini le formalisme MIB (Model of Interactions in Biology). MIB est basé sur les graphes bipartis et permet de définir, rechercher et étudier les motifs hétérogènes, composés des régulations génétiques et des interactions protéine-protéine. Afin d'approfondir l'étude de la structure et de la dynamique des réseaux biologiques, nous avons ensuite proposé le formalisme MIN au chapitre cinq Celui-ci possède la structure bipartie de MIB, mais permet d'avoir des annotations beaucoup plus riches des noeuds et des arcs du réseau. Ceci augmente son expressivité par rapport au MIB. De plus, MIN permet la représentation des données macroscopiques relatives au système biologique, telles que la description de ses différents états observés par les biologistes. Un modèle MIN permet de générer, pour divers formalismes, l'ensemble des modèles compatibles, à l'aide d'algorithmes de traduction automatique. Les modèles obtenus peuvent ensuite être analysés par des outils standards, propres au formalisme de destination. Une extension permet également de doter MIN d'une structure hiérarchique, permettant d'associer à un problème biologique donné un niveau d'abstraction adéquat, et de réutiliser les modèles ainsi créés de manière efficace.

Cette approche sera mise en application tout au long de la thèse, sur des exemples de processus biologiques. Enfin au chapitre six, nous conclurons, et présenterons les perspectives de ce travail.

.



Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

# Chapitre 2

# Etat de l'art

De nombreux formalismes sont utilisés pour étudier les réseaux d'interactions biologiques. La modélisation de ces systèmes est le premier pas vers la compréhension, le contrôle, la conception et la modification des systèmes biologiques afin de leur donner des propriétés voulues [105].

## 2.1 Modèles statiques

### 2.1.1 Bases de données en bioinformatique

Un progrès fantastique a été réalisé durant ces dernières années dans l'assemblage des données sur le génome humain [151]. Ceci n'est cependant que l'avant-garde d'une longue série de découvertes. Le génome n'est lié que indirectement aux fonctions des protéines, et les fonctions des protéines ne sont que indirectement en rapport avec le fonctionnement de la cellule entière. Donc, l'attention des chercheurs se déplace de la génomique vers la protéomique, et au-delà. A chaque étape, des bases de données de plus en plus grandes vont être compilées.

Les bases de données en biologie ont une organisation particulière qui leur permet

d'intégrer des données cliniques et biologiques produites expérimentalement et des données issues du séquençage des génomes (les annotations structurales et fonctionnelles). Ces données sont disponibles dans des bases de données publiques afin de les rendre disponibles à l'ensemble de la communauté scientifique. Les objectifs essentiels des chercheurs qui utilisent ces bases de données, souvent via des interfaces Web, sont d'une part de prédire les fonctions des protéines à partir de leurs séquences d'acides aminés, d'autre part de construire, à l'aide de modèles mathématiques, des réseaux fonctionnels décrivant les rôles et les interactions de centaines de molécules dans les cellules.

Les bases de données et de connaissances actuelles contenant de l'information sur les interactions peuvent être vues, en général, comme s'appuyant sur une représentation de graphe annotée fortement [35]. De tels exemples sont aMaze [179], BioCyc [99] ou Kegg [95].

Les graphes sont des structures abstraites utilisés pour modéliser de grands réseaux d'interactions. Un graphe est un tuple $< V, E >$ où $V$ est l'ensemble des noeuds, et $E$ l'ensemble des arêtes. Une arête dirigée (arc) est une paire $< i, j >$ où $i$ dénote une extrémité de l'arête et $j$ l'autre. Les noeuds correspondent aux gènes ou à tout autre entité biologique d'interêt. Les arêtes représentent les interactions entre ces entités. Différents types de graphes, comme les graphes colorés, étiquetés ou valués, sont utilisés pour traiter une variété d'interactions entre les noeuds dans le même graphe. Les hypergraphes sont une variante des graphes permettant de modéliser des situations dans lesquelles par exemple plusieurs protéines régulent coopérativement un gène en formant un hétérodimère.

### Kegg

Parmi les bases de données hétérogènes, c'est-à-dire celles qui intègrent plusieurs types de données, et parmi les plus populaires, on trouve, Kegg – l'encyclopédie de gènes et génomes de Kyoto [140, 96]. C'est la base de données de réactions biochimiques la plus

utilisée. Elle stocke des données génomiques, biochimiques et de réseaux métaboliques dans trois sous-bases de données distinctes : GENES, LIGAND et PATHWAY. Actuellement, Kegg fournit 30 000 voies construites à partir de 269 voies de référence intégrant 6400 réactions pour 212 bactéries, 21 archéobactéries et 77 eucaryotes. En 2005, Kegg s'est ouvert considérablement pour les développeurs. Des APIs et des webservices qui permettent l'interrogation (requêtage) de la base à distance via Internet ont été écrits. Kegg repose toutefois toujours sur des cartes dessinées à la main, ce qui l'empêche de grandir véritablement, d'intégrer rapidement de nouvelles réactions et d'être modifié par l'utilisateur (Figure 2.1).

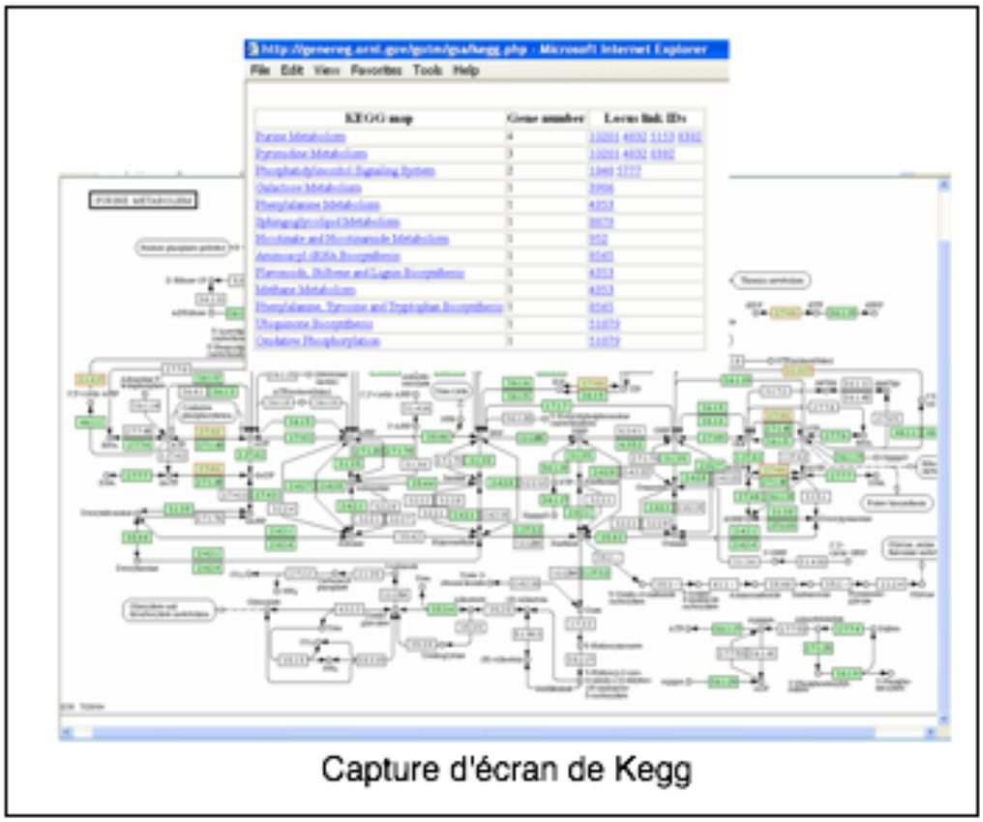



FIG. 2.1 – Capture d'écran de Kegg. Kegg est basé sur des images statiques de chemins de réactions, (reproduit de Kegg.com).

**BioCyc**

BioCyc [100, 75] est une collection de descriptions des voies de réactions et d'informations génomiques pour plus de 300 organismes, crée et gérée par P. Karp et ses collaborateurs au NIH (National Center for Research Resources). Chaque base pour un organisme donné décrit le génome (gènes et promoteurs), le réseau métabolique, les complexes protéiques, les différentes formes actives de ces complexes, les voies de signalisation, les réactions de transport et le réseau de régulation transcriptionnelle (Figure 2.2).

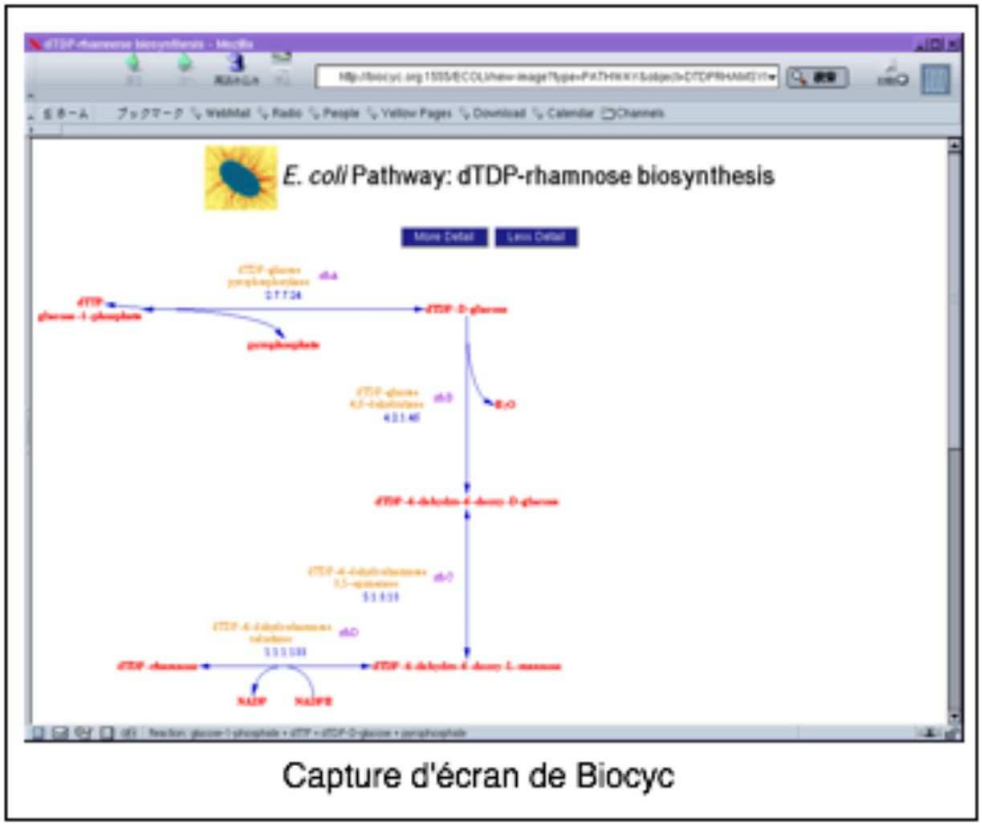


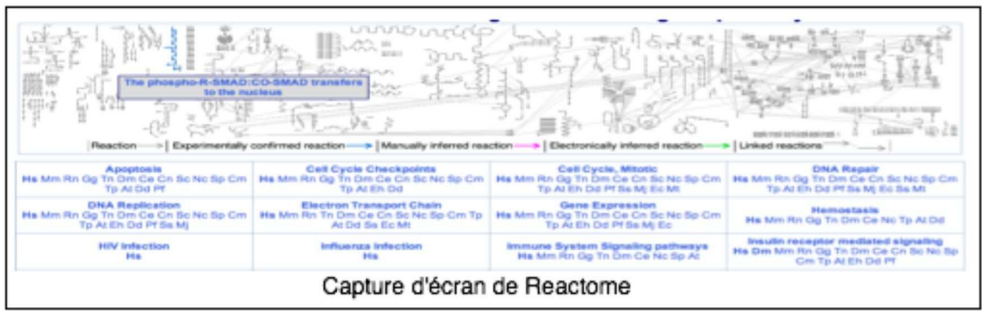

FIG. 2.2 – Capture d'écran de Biocyc. Biocyc est visualisable à partir du web ou d'un logiciel spécifique ; les chemins réactionnels sont dessinés automatiquement, (reproduit de Biocyc.org).

De plus, il existe une base de donnée supplémentaire, MetaCyc, qui est composée de voies métaboliques non-redondantes sur plus de 450 organismes provenant de résultats

d'expériences biologiques. MetaCyc contient actuellement 601 voies et 5000 réactions référencées dans plus de 6500 articles. Les bases de données BioCyc sont divisées en différentes catégories suivant le soin avec lequel les données ont été vérifiées expérimentalement. EcoCyc pour Escherichia coli K12 et MetaCyc contiennent uniquement des données confirmées par des expériences manuelles. 17 organismes, dont la levure, ont subi une légère annotation manuelle. Les bases de données pour les autres organismes ont été générées automatiquement par inférence, sans annotation. BioCyc repose sur un format de données propriétaire, qui n'est pas accessible sans passer par l'outil d'interrogation appelé PathwayTools, très spécifique, qui est fourni avec la base de données. Biocyc est l'outil existant le plus diversifié en terme de données intégrées (hétérogènes) et le plus riche en quantité d'informations contenues. Il a l'inconvénient d'être un logiciel propriétaire, ce qui rend son utilisation difficile par des développeurs externes. Les analyses à grande échelle des données qui y sont contenues sont donc difficiles.

**Reactome**

Reactome [86, 89] est une base de données de processus biologiques développée par une collaboration entre Cold Spring Harbor Laboratory, The European Bioinformatics Institute, et The Gene Ontology Consortium. L'organisme ciblé est l'humain Le code source est librement accessible. Reactome couvre toutes les informations associées, des réactions biochimiques aux processus de plus haut niveau, comme les voies de signalisation des hormones. Quelques informations parcellaires pour d'autres organismes sont maintenant aussi intégrées. Chez l'homme, plus de 400 voies de réactions sont décrites. Parmi les bases de données présentées, seul Reactome fournit librement toutes les données ainsi que le logiciel d'interface (Figure 2.3).



Capture d'écran de Reactome

FIG. 2.3 – Capture d'écran de Reactome ; comme dans Kegg, la carte du métabolisme est dessinée manuellement une seule fois pour tous les organismes, (reproduit de reacome.org).

### aMAZE

aMAZE est dédié à la représentation des interactions moléculaires et processus cellulaires [178, 19]. aMaze a été développé par J. van Helden et son équipe au Laboratoire de Bioinformatique des Génomes et des Réseaux en Belgique. Sa force provient de son modèle de données conçu avec soin. Le modèle permet d'intégrer des données sur les voies métaboliques, les interactions protéiques, la régulation des gènes, le transport et les voies de signalisation. Actuellement aMAZE contient des données provenant principalement de Kegg pour trois organismes (l'humain, la levure et *E. coli* ). Le domaine métabolique contient 100 voies et plus de 5000 réactions. Toutefois cette base de données est propriétaire, ce qui rend difficile son accès et son utilisation. Son développement semble actuellement au point mort.

Les bases de données présentées, comportent toutes des inconvénients qui empêche leur utilisation pour lancer des analyses statistiques sur des données hétérogènes qu'elles contiendraient. Kegg comporte principalement des données du métabolisme, et le modèle utilisé dérive de dessins de voies métabolique sans représentation d'objets véritablement, même si ces deux dernières années, Kegg s'est ouvert grâce à la fourniture d'une API à la communauté. BioCyc est un logiciel propriétaire et les données ne sont accessible qu'en LISP au moyen d'interfaces très spécifiques et lourdes. BioCyc est toutefois la base de

donnée la plus 'hétérogène', mais de nombreuses informations biologiques ne peuvent y être intégrées comme les données résultantes de l'expression des gènes, ou des phénotypes. Reactome est focalisé sur l'humain et un projet européen devrait prochainement s'intéresser à la production d'un "Bacterial Reactome". aMaze semblait la plus prometteuse lorsqu'elle a été lancée, mais les développeurs se sont trop focalisés sur les techniques d'implémentations modernes et son originalité scientiique, n'a pu être poussé assez loin.

### 2.1.2 Integration de données

Dans cette partie, nous présentons des bases de données utiles non seulement pour le stockage des données, mais aussi pour l'analyse et la manipulation des réseaux d'interactions biologiques hétérogènes. On peut distinguer plusieurs catégories de bases de données [8] en fonction du type principal de données contenues dans chacune d'elles, leur format de données et le centre d'intérêt biologique. Nous ne présenterons ici que les bases de données, souvent couplées à des logiciels d'interrogation, qui intègrent plusieurs types de données biologiques. Peu de bases de données sont libres d'accès, et elles ne respectent que partiellement des formats d'échanges comme PSIMI [69], BioPAX [76], SBML [48, 88] ou CellML [124].

#### Formats d'échange de données

SBML [88, 48] est un format d'échange de type XML pour représenter des modèles biochimiques de réseaux de réactions. Le format est utilisé pour les réseaux métaboliques, les voies de signalisation, les réseaux de régulation, et bien d'autres. SBML est utilisé par de nombreuses bases de données et par des outils de simulation. Le schéma modélise la structure statiques des réseaux mais certains paramètres sur la dynamique, comme les équations différentielles modélisant la dynamiques de certaines entités, peuvent y être encodées.

BioPAX [76] est un format d'échange pour les données portant sur les voies biochimiques. BioPAX se focalise sur les données présentes dans une sélection de bases de données modèles. Le format représente les données communes entre ces différentes bases. Chaque base de données participante (aMaze, BioCyc, Reactome, Patika,...) exporte une partie de leurs données, au format BioPax, chacune avec leur propre identifiant.

**BioWarehouse**

BioWarehouse [118] est une base de données publique qui intègre un ensemble de bases de données biologiques dans une unique base (mySQL ou Oracle) pour en faciliter l'exploitation, l'analyse et l'exploration. La structure de BioWharehouse repose sur un schéma de données relationnel qui modélise les types de données bioinformatiques. De nombreux programmes d'importation de données existent, du moins pour une partie des données de ces bases (notamment BioCyc, BioPAX, ENZYME, GO, KEGG, UniProt). Le schéma comporte 16 relations ; il est simple et unique pour l'ensemble des données ; toutefois, les entités communes importées des différentes bases ne sont pas fusionnées en entités uniques. BioWarehouse est en quelque sorte une vue partielle des principales bases de données existantes, analogue au BioPAX.

**Patika**

PATIKA [38, 37] est une nouvelle base de données intégrative construite à partir de plusieurs sources de données (Entrez [84], UniProt[79], PubChem[73], GO[81], IntAct[77], HPRD[83] et Reactome[86]). Elle se focalise sur l'humain et contient plusieurs centaines d'états d'entités biologiques et quelques milliers de réactions. Les données sont interrogeables au moyen d'Internet. Cet outil est construit à partir d'autres technologies informatiques modernes : XML et Hibernate. Le système est compatible avec les formats BioPax level 2 et SBML.

**Modèles basés sur UML**

BioUML [113, 112] est un cadre de programmation en Java pour la biologie des systèmes. Il facilite l'accès aux bases de données, il fournit des outils de visualisation et de simulation (au moyen de diagrammes de blocs), et un formalisme de description des systèmes biologiques qui permet de manipuler des graphes. Chacun de ces objets (noeuds, arêtes), constitue un élément du diagramme auquel un rôle dynamique peut être associé, et relié soit à une variable, soit à une équation d'un système d'équations différentielles .

Des différents formats de données qui permettent d'échanger et intégrer des informations biologiques, le SBML est celui qui est le plus populaire parmi les scientifiques. Le format XML de SBML se prête bien à l'extension du format, et de nombreuses extensions permettent d'annoter les modèles de différentes informations. L'inconvénient est que chaque groupe interprète à sa manière le format, et seules certains types d'interactions sont représentées de manière standard en SBML. Historiquement SBML se prettait bien aux modèles comportant des équations différentielles. L'UML devrait cependant un jour prendre le dessus sur l'XML, car il permet d'exprimer beaucoup plus de connaissances, notamment sur less fonctions et les rôles des objets biologiques manipulés. Un modèle UML, probablement différent de BioUML qui peine a grandir, devrait voir le jour au cours des prochaines années.

### 2.1.3 Modèles graphiques

Les cartes métaboliques [106, 107, 74, 53, 136, 109, 110, 111, 145, 31, 72] fournissent une représentation claire et concise des flux de métabolites dans la cellule. Cependant, il reste difficile de représenter les réseaux biologiques impliqués dans les voies de signalisation qui régulent les fonctions cellulaires du fait de la quantité importante d'informations en cause, de leurs différentes natures, et des liens qui existent entre voies métaboliques. Les cartes

métaboliques sont toutefois importantes pour comprendre comment les réseaux biologiques fonctionnent. Par ailleurs, les sytèmes de régulation dans ces réseaux sont si complexes (interconnections, boucles) qu'ils sont difficilement compréhensibles intuitivement, et que l'informatique est nécessaire pour en étudier le comportement.

### Diagrammes de processus

Hiroaki Kitano et ses collaborateurs [106, 107] ont décrit un formalisme qui permet de représenter les diagrammes de réseaux biologiques sous la forme des "diagrammes de processus" lisibles par l'homme et par l'ordinateur. Un modèle généralisé a été proposé plus tard sous le nom de Systems Biology Graphical Notation (SBGN) [74]. L'apport important des diagrammes de processus est qu'ils constituent une représentation des séquences d'événements ou de voies biochimiques d'un réseau.

Un diagramme de processus comporte des 'noeuds d'état' et des 'noeuds de transition' [107]. Les noeuds d'état représentent les entités dans le processus biologique, comme les protéines, l'ARN ou les gènes, et les noeuds de transition représentent les différentes réactions comme les associations, dissociations, activations ou inhibitions. Les 'arcs' lient les noeuds d'état aux noeuds de transition. Cette approche conduit à une matrice de connectivité qui définit le réseau d'une manière lisible par ordinateur. Visuellement, les diagrammes de processus font apparaître chacune des occurrences des espèces, définie par les noms de chacune de ses composantes monomoléculaires et ses états d'activation. Cela permet d'interpréter aisément chaque réaction indépendamment.

CellDesigner [53] est un outil qui repose sur le même type de représentation de diagrammes de processus. Il s'agit en plus d'un outil qui permet de créer, d'éditer et de visualiser des diagrammes de processus, qui peuvent ensuite être échangés au format d'échange SBML [139].

Une autre représentation (notation dite 'Edinburgh') [136] est plus compacte qu'une

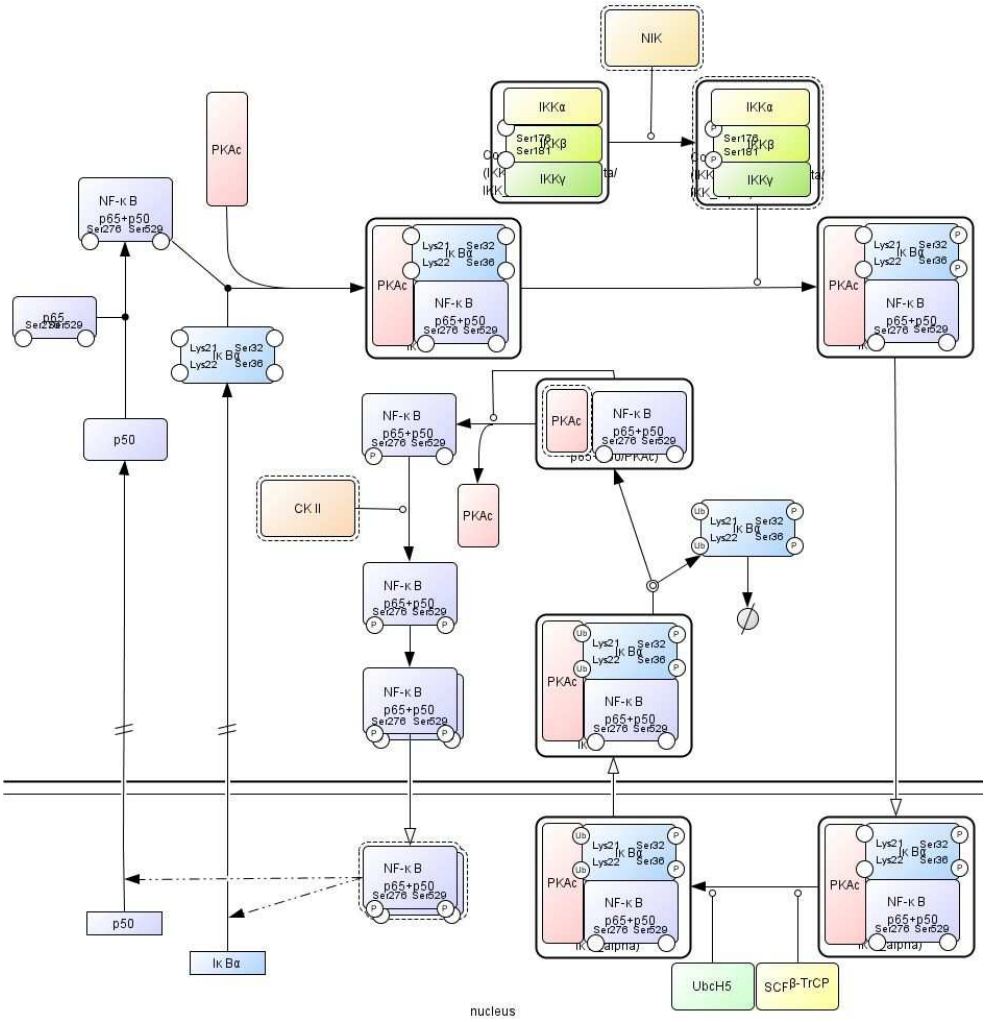



FIG. 2.4 – CellDesigner, (reproduit de sbgn.org).

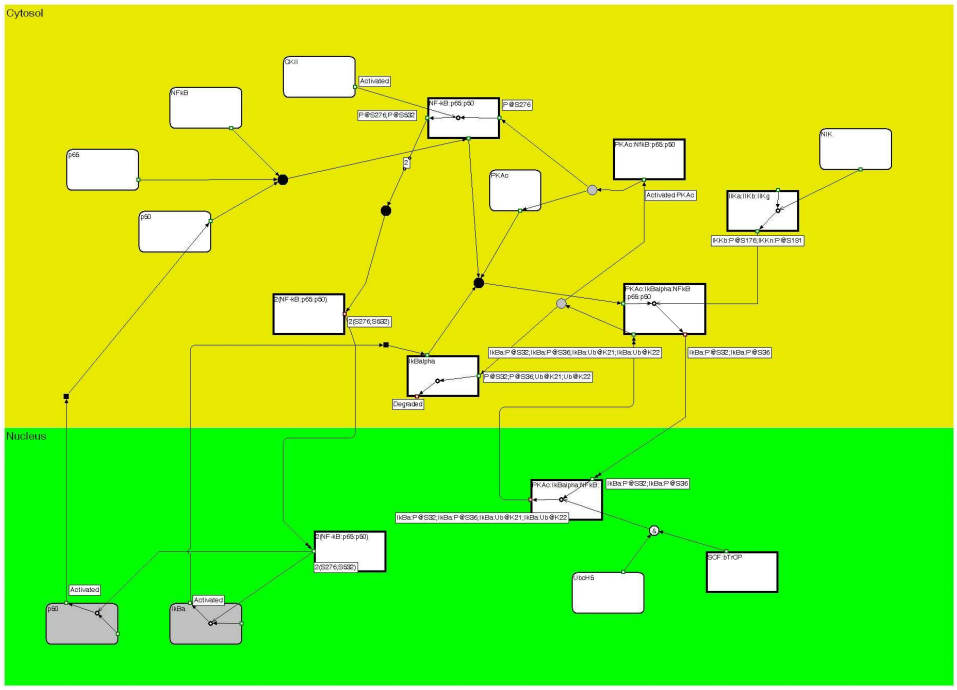



FIG. 2.5 – Edinbourg notation (reproduit de sbgn.org).

représentation en diagrammes de processus et s'oriente vers des interactions logiques, la localisation subcellulaire, et les états des espèces. Cela est possible en limitant la représentation d'espèce à une seule par compartiment cellulaire, en distinguant les complexes comme des espèces à part entière et en décrivant les processus par des diagrammes de transition. Une autre idée clé de ce formalisme est d'utiliser une notion de hiérarchie qui permet de cacher certaines parties d'une voie, que l'on peut rendre visible à la demande pour faire apparaître tout le mécanisme en détail.

### Cartes des voies métaboliques

Une approche différente de la représentation des réseaux biologiques utilise le modèle entité-relation. Elle a été introduite et développée par Kohn et ses collaborateurs [109, 110, 111] sous le nom de Molecular Interaction Maps (MIM). Les cartes de Kohn possèdent

deux grandes différences avec les diagrammes de processus. Ils ne représentent une espèce chimique nommée qu'une seule fois sur la carte, et ils n'imposent pas d'ordre dans les événements, mais à la place ils indiquent toutes les réactions potentielles possibles si les réactifs se retrouvent au même endroit, au même moment. Chaque type de représentation a ses avantages et, d'après Kohn [108], une représentation standardisée reste à définir pour être aussi utile pour la biologie des systèmes que le sont les schémas de circuits intégrés en électronique.

Pirson et ses collaborateurs [145] ont proposé une représentation originale des informations sur des réseaux biologiques, qui s'appuie sur des cartes métaboliques telles que les *"Biochemical Pathways"* de Boehringer Mannheim [78] . Cependant, les auteurs se focalisent plutôt sur des relations logiques (de régulation) entre les noeuds du réseaux, plutôt que sur des mécanismes des réactions. De plus, ils accordent une grande importance à annoter le type d'interaction en fonction du temps qu'elle prend dans le système biologique : ils font la distinction entre l'action immédiate et rapide, l'action lente et l'action après un délai. Les éléments de cette représentation sont énumérés sur la Figure 2.7.

Cook et ses collaborateurs [31] proposent eux BioD, un langage de description visuelle des problèmes biologiques, basé sur Internet. Leur objectif est de créer un langage CAD (conception assistée par ordinateur) pour simuler et analyser les systèmes biologiques. Les mots de ce langage sont présentés sur la Figure 2.8. Même si chaque brique est sensée être comprise implicitement, elle peut posséder des attributs explicites tels que la concentration pour les atomes ou molécules ou le volume ou la surface pour un compartiment cellulaire. De nouvelles briques peuvent être dérivées des briques de base. La structure du système biologique peut être représentée avec BioD, ainsi que les séquences des événements associées.

VitaPad [72] est un outil utilisé dans le laboratoire de Zhao à Yale pour la présentation des données biologiques. VitaPad utilise une structure de base de données qui sépare les

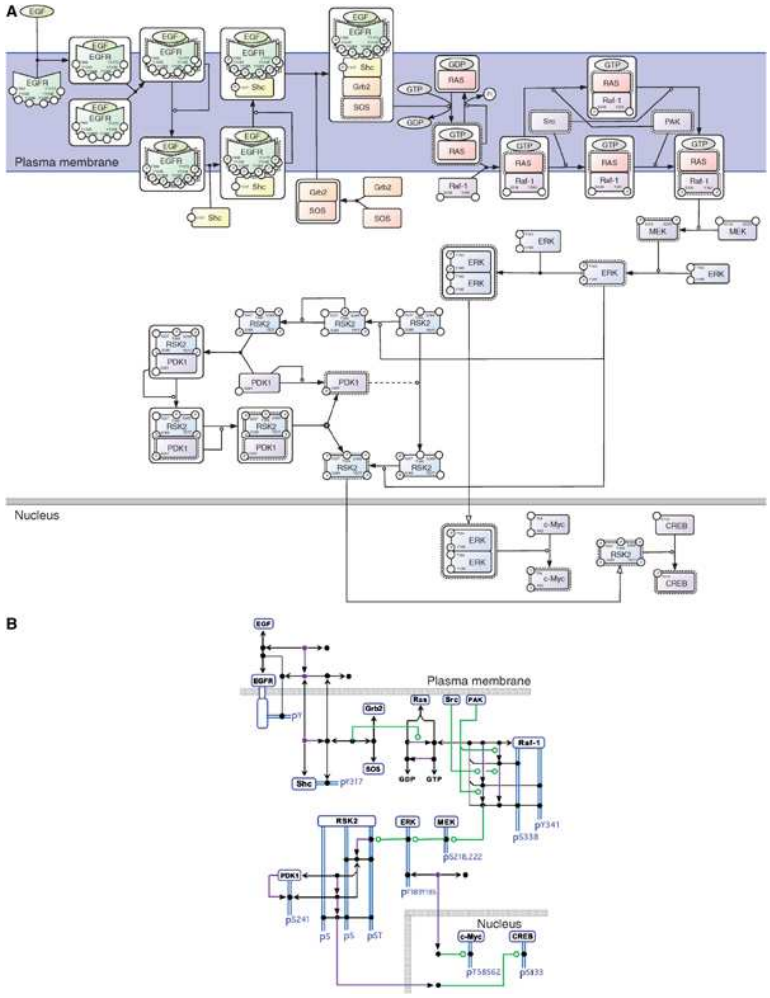



FIG. 2.6 – Diagrammes de Kitano (A) et de Kohn (B) (reproduit de sbgn.org).

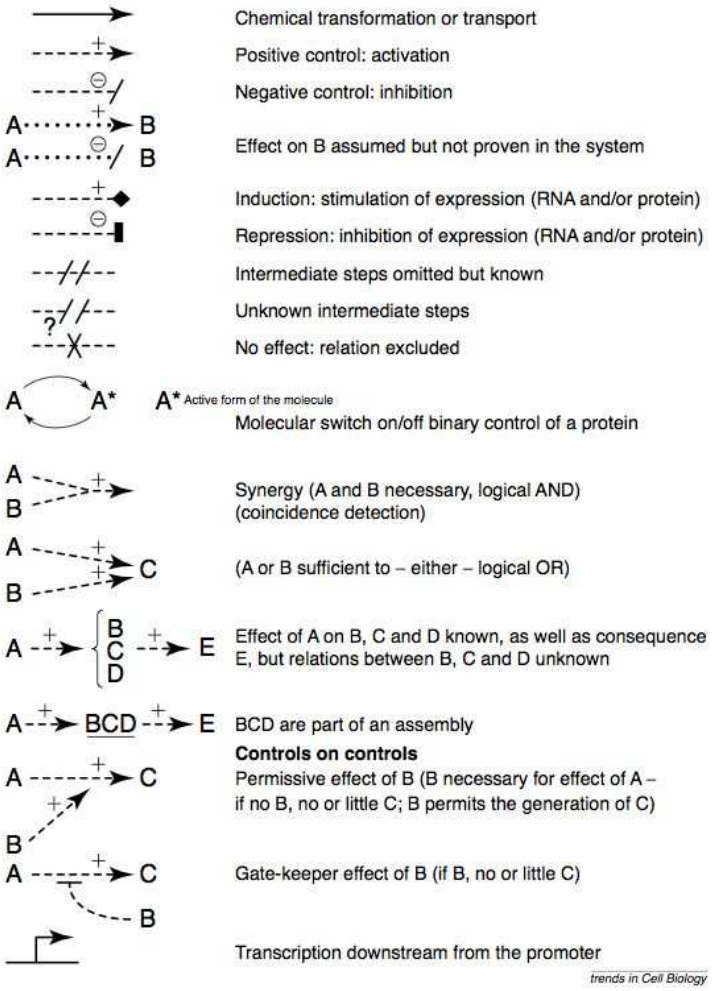



FIG. 2.7 – Liste des arcs pour la représentation graphique des réseaux de régulation et leurs interprétation biologique. Tous les arcs et les noeuds peuvent être dessinés avec un code couleur, comme par exemple le rouge pour la régulation négative, le vert pour la régulation positive pour les arcs, la couleur bleue pour les protéines, la couleur noire pour les autres espèces chimiques. Des étiquettes $s$, $l$ ou $d$ peuvent être utilisées sur les arcs pour une régulation rapide, lente ou après un certain délai, respectivement, (reproduit de [145]).

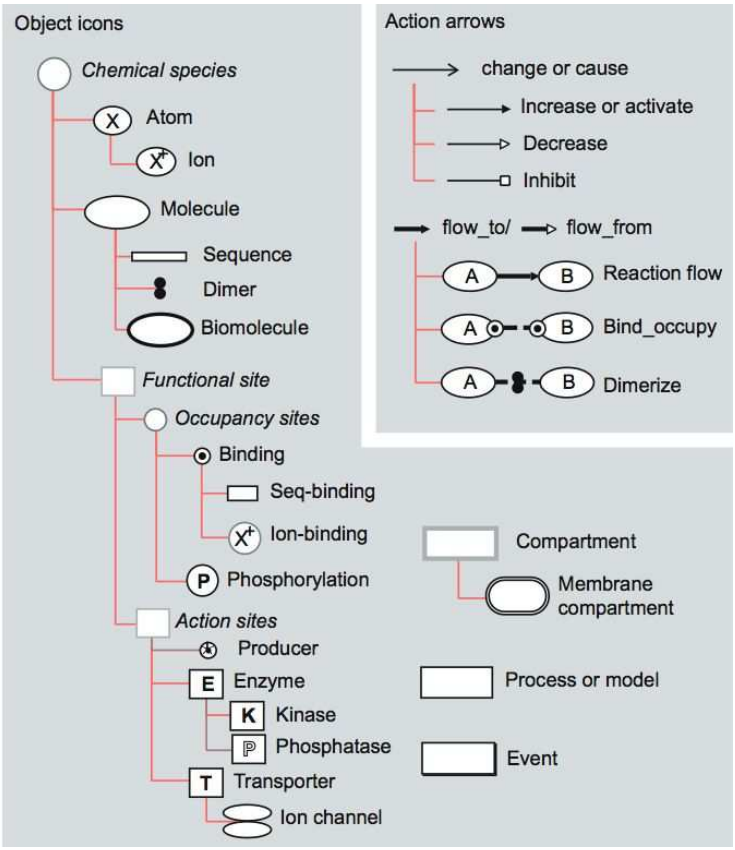



FIG. 2.8 – Ensemble des symboles de base pour la description des systèmes biologiques avec BioD, (reproduit de [31]).

aspects biologiques de la représentation graphique des voies biologiques. Le but est de pouvoir représenter les données des types nouveaux au fur et à mesure qu'elles sont ajoutées dans le système. Il peut importer et exporter l'information dans un format basé sur XML et utilise Graphviz [82]pour dessiner les réseaux biologiques.

Tous ces modèles graphiques placent au centre de leur intérêt la visualisation statique des données qu'ils contiennent. Ils formalisent, certes, les types des interactions, des entités, des états, des processus biologiques. Cependant, leur richesse qui garantie leur expressivité permet de modéliser le même processus biologique par des modèles bien différents, ce qui rend difficile leur comparaison ou combinaison. Les modèles graphiques, même s'ils sont centrés sur la visualisation des données, n'acceptent pas des données contradictoires, c'est à dire celles qui représentent des issus alternatives d'une même expérience biologique. Ainsi, un choix de modélisation implicite doit être fait au moment de construction du modèle pour intégrer un comportement du système et pas l'autre. Cependant, ces différences et ces contradictions peuvent provenir non seulement d'une erreur expérimentale, mais également d'une différence dans une des variables cachées du système. Les modèles présentés dans cette section sont orientés sur la représentation des réactions biochimiques, et même s'ils permettent parfois la simulation dynamique, elle est souvent de niveau trop détaillé (niveau des molécules) pour répondre à des questions biologiques divers.

## 2.2 Modèles dynamiques

Le problème de la modélisation, de la simulation et de l'analyse des processus biochimiques est, sans doute, un des points majeurs de la biologie des systèmes. Les différents acteurs et processus biochimiques peuvent être représentés à différents niveaux d'abstraction [35, 90]. Quelques-unes de ces approches sont présentées plus loin.

L'état d'une espèce chimique est exprimé à travers un nombre fini d'états abstraits, où les niveaux d'activité intermédiaires sont supposés avoir le même comportement. Des fonctions sont utilisées pour décrire les nouveaux états (les intervalles de concentration probables) des espèces chimiques étant donné leurs états précédents. Les transitions entre les états sont supposées se produire simultanément ou, mieux, de manière asynchrone. Dans le cas le plus simple, seuls deux états ("on" et "off") sont utilisés et l'algèbre booléenne décrite la dynamique. Les systèmes de transitions concurrentes [26, 25] et la logique de réécriture des voies [45] sont de bons exemple de la modélisation logique. Kappler et collaborateurs [97] ont montré comment élargir les réseaux booléens simples en utilisant des équations différentielles pour capturer la concentration, tandis que les fonctions booléennes servent à déterminer les taux de réactions. La probabilité d'être dans un état donné est parfois la quantité plus intéressante à estimer, comme dans le cas de Sachs et ses collaborateurs [152] qui utilisent les réseaux bayésiens pour modéliser les voies de signalisation cellulaires. De la même manière, Shmulevich et ses collaborateurs [158] décrivent l'utilisation des réseaux booléens stochastiques pour modéliser les réseaux de régulation génétiques et déterminer le comportement probabiliste à long terme des gènes choisis. Platzer et ses collaborateurs [146] simulent le développement embryonnaire de *C. elegans* en associant des états booléens aux gènes et en les mettant à jour de manière synchrone selon la matrice des interactions. Batt et ses collaborateurs [10] ont appliqué la théorie du model-checking sur des systèmes biochimiques à travers une simulation qualitative.

Si les concentrations sont représentées exactement par des fonctions réelles continues, les équations différentielles de la dynamique sont déductibles directement de la loi cinétique d'action de masse. En tant que compromis entre les représentations discrètes et continues, les équations différentielles qualitatives peuvent être utilisées, avec les états qualitatifs correspondant à des plages de concentrations différentes [10, 35].

Un système biochimique peut également être vu comme un système stochastique. Dans ce cas, la probabilité que le système soit dans un état donné est estimée, plutôt que l'évolution de l'état du système dans le temps [42]. A chaque pas de temps de la simulation stochastique une réaction suivante est tirée au sort, et le temps qu'elle prend est estimé. Puis, l'état du système est mis à jour en fonction du bilan de la réaction.

Chen et Hofestaedt [27] donnent un bon résumé des réseaux de Petri hybrides et illustrent leur application à la modélisation quantitative et la simulation des réseaux métaboliques régulés génétiquement. Une version plus étendue, les réseaux de Petri fonctionnels hybrides, a été utilisée par Nagasaki et ses collaborateurs [137] pour capturer les événements et transitions continues et discrètes. Regev et ses collaborateurs [149] et Curti et ses collaborateurs [33] proposent d'utiliser le $\pi$-calcul pour la modélisation biochimique, tandis que Phillips et Cardelli ont proposé le $\pi$-calcul stochastique [144]. Weimar [181] utilise les automates cellulaires pour modéliser les réactions catalysées par les enzymes. Bockmayr et ses collaborateurs ont proposé d'utiliser la programmation par contraintes hybride et concurrente pour la modélisation biologique [46, 15]. Faeder et ses collaborateurs [47] ont détaillé l'approche générale de modélisation à base des règles, où les interactions sont spécifiées entre les espèces chimiques (ou leurs parties), ce qui conduit à l'émergence d'un comportement global. Tous les modèles ci-dessus impliquent une approximation à des niveaux d'abstraction divers. On va revoir maintenant certaines de ces approches plus en détail.

### 2.2.1 Equations différentielles

Les équations différentielles se situent parmi les formalismes les plus utilisés pour modéliser des systèmes dynamiques en sciences et en ingénierie. Elles ont été largement utilisées en biologie également, et de nombreuses variantes et extensions de ce formalisme ont été développées pour la biologie. Ce formalisme modélise les concentrations d'ARNs, de protéines, et autres molécules par des variables dépendant du temps, avec des valeurs réelles positives. Les interactions sont représentées par des fonctions et relations différentielles entre les variables de concentration. Des retards peuvent être modélisés par une extension spécifique du formalisme [162]. Une variante a aussi été proposée pour la modélisation de réseaux de régulation qui se restreint à des équations linéaires par morceaux (PLDE) [60, 131, 170] avec de nombreuses applications [4, 36, 56]. Cette simplification repose sur la modélisation des fonctions sigmoïdes, ou fonctions de Hill, par une marche d'escalier.

Par exemple, Lee et ses collaborateurs [117] ont développé un modèle mathématique pour la voie de signalisation Wnt en prenant en compte la cinétique des interactions protéine-protéine, de la synthèse et de la dégradation des protéines, ainsi que la phosphorylation et la déphosphorylation. Ils ont été capables dans l'exemple de modélisation du cycle circadien de prédire et de vérifier expérimentalement que axin et APC agissent de manière différente. Ils ont également effectué l'analyse théorique afin d'obtenir une expression pour la suppression de tumeur ou la cancérogenèse.

Collier et ses collaborateurs [30] ont proposé un modèle simple de la voie de signalisation Delta-Notch pour montrer que l'inhibition latérale via la rétroaction est suffisante pour expliquer l'apparition des patrons. Cependant, au lieu d'intégrer les détails des mécanismes biochimiques, ils se sont appuyés sur les observations sur la production des facteurs Delta et Notch au niveau cellulaire. Ensuite, ils ont montré qu'un état stationnaire homogène n'est pas stable, tandis que l'état stationnaire hétérogène peut l'être sous certaines conditions.

Les équations différentielles supposent souvent un système spatialement homogène. Ce-

pendant, il est parfois possible d'introduire des notions de compartimentation cellulaire discrète en considérant des entités dans des compartiments différents comme incapables d'interagir. De cette manière, il est possible de modéliser des phénomènes de gradients de molécules, ou de modéliser des phénomènes de diffusion. De tels systèmes sont largement étudiés en biologie [58, 59, 61, 62, 101].

Des paramètres stochastiques peuvent aussi être introduits dans les équations afin de modéliser le fait que les concentrations d'entités ne varient pas continûment et de manière déterministe. La première hypothèse est d'autant plus importante que les quantités de molécules considérées sont souvent faibles et qu'il est alors difficile de modéliser les phénomènes en suivant des concentrations continues [57, 42, 128]. Le nombre de facteurs de transcription dans une cellule est souvent de l'ordre de la dizaine, et le nombre de molécules d'ADN est de un. Gepasi [129] est un outil informatique basé sur cette approche à la modélisation des systèmes biochimiques : il a pour but d'aider à la traduction des réactions chimiques en matrices et équations différentielles.

### 2.2.2 Réseaux de neurones

Marnellos et Mjolsness [126] ont proposé de modéliser les gènes de la voie de signalisation Delta-Notch en tant que noeuds dans des réseaux de neurones récursifs, avec des poids correspondants à la force d'interaction. Leur modèle n'intègre pas de connaissances a priori sur cette voie, mais essaie de reproduire le mécanisme d'inhibition latérale. Chaque gène est supposé additionner les entrées de la même cellule et des cellules voisines à chaque instant $t$. L'idée de cette approche est d'optimiser les forces d'interactions entre les gènes pour obtenir, par recuit simulé, les patrons d'expression connus. Il est supposé que l'optimisation va fournir les paramètres du modèle qui permettent au système biologique d'arriver à l'état final à partir de l'état initial. Cette méthode a permis de découvrir les combinaisons de paramètres du système robustes à des perturbations et prédit des patrons d'expression des

gènes en fonction de la position de la cellule dans un regroupement des cellules du système nerveux en développement. Ceci montre que les informations macroscopiques sur les changements d'état du système peuvent servir à décortiquer les interactions sous-jacentes au phénomène biologique étudié.

### 2.2.3 Réseaux bayésiens

Dans les réseaux bayésiens [52, 142], la structure des systèmes de régulation génétique est modélisée par un graphe dirigé acyclique $G =< V, E >$. Les noeuds $i$ de $V$ représentent les gènes ou d'autres entités, et correspondent aux variables aléatoires $X_i$. Si $i$ est un gène, alors $X_i$ représente le niveau d'expression du gène $i$. Pour chaque $X_i$, une distribution conditionnelle $p(X_i|parents(X_i))$ est définie où $parents(X_i)$ correspond aux variables définies dans le graphe $G$ par les régulateurs directs de $i$. Le graphe décrit les relations de Markov et la distribution de probabilité jointe peut se décomposer ainsi $p(X) = \prod p(X_i|parents(X_i))$. Friedman et al., [52] ont proposé un algorithme pour induire un réseau bayésien à partir de données d'expression de gènes. Des extensions ont été proposées afin de prendre en compte les mutations génétiques, distinguer l'activation et l'inhibition [143]. La modélisation par les réseaux bayésiens est intéressante pour les réseaux de régulation car elle repose sur de solides bases de statistiques, ce qui lui permet de modéliser les aspects stochastiques ou de prendre en compte des mesures expérimentales bruitées. De plus, de tels réseaux sont utilisables même si une partie de la description des interactions est manquante. Dans ces réseaux, la dynamique n'est décrite qu'implicitement. En effet, un réseau bayésien est un modèle statique. En fait, la notion de temps n'intervient pas dans un réseau bayésien classique. Mais pour modéliser un processus stochastique, on utilise un réseau bayésien dynamique qui est créé en répétant dans le temps un réseau classique et en reliant ces réseaux classiques par des liens causaux d'un pas de temps à l'autre. Un réseau bayésien dynamique est donc une chaîne du même réseau bayésien répété autant de fois que nécessaire (suivant

la longueur de la séquence d'observations). Chaque répétition est un pas de temps permettant de représenter l'évolution d'un processus stochastique. Ils contiennent chacun un certain nombre de variables aléatoires représentant les observations et les états (cachés) du processus.

### 2.2.4 Modèles hybrides

Les états discrets des systèmes hybrides décrivent naturellement les régimes du comportement du système qui sont qualitativement différents, en termes des espèces et des réactions prédominantes. Les gardes et les restaurations sur les transitions discrètes permettent la description des conditions biochimiques dans lesquelles l'état du système change. L'utilisation des automates hybrides pour la modélisation des réseaux biomoléculaires a été décrite par Alur et ses collaborateurs [4] et Mishra [135]. Amonlirdviman et ses collaborateurs [5] on montré l'utilité des systèmes hybrides par la modélisation de la polarité des cellules de la drosophile. En commençant avec la définition des S-systèmes, formulée par Savageau et Voit [180], Antoniotti et ses collaborateurs [7] ont utilisé un automate supplémentaire pour augmenter l'ensemble des systèmes qui peuvent être représentés, en utilisant ensuite un automate hybride à part entière [6]. Lincoln and Tiwari [122] ont détaillé la modélisation des réseaux biochimiques avec les automates hybrides, tandis que Hu et ses collaborateurs [87] décrivent un système hybride stochastique pour la modélisation de la production de la subtiline chez *Bacillus subtilis*.

#### Automates hybrides

Ghosh et Tomlin [56] ont proposé un modèle hybride simplifié pour capturer la formation des motifs via l'inhibition latérale. Chaque cellule biologique est modélisée par un automate hybride linéaire par morceaux, avec quatre états qualitatifs qui permettent de simplifier les termes non-linéaires du modèle de Collier et ses collaborateurs [30]. Ces quatre

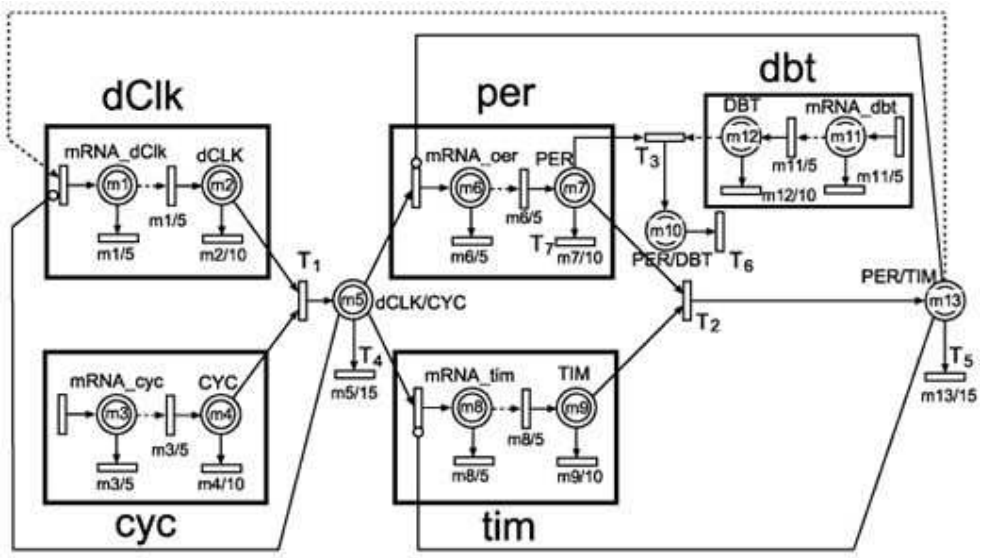



FIG. 2.9 – Exemple de modélisation du cycle circadien avec HFPN (reproduit de www.GenomicObject.Net).

états sont accessibles les uns à partir des autres, avec une garde sur chaque transition qui correspond à l'invariant de l'état destinataire. Les modèles multicellulaires sont obtenus par la composition d'automates représentant chacun une seule cellule.

### Réseaux de Pétri Fonctionnels Hybrides (HFPN)

Matsuno et ses collaborateurs [127], décrivent une nouvelle variante de réseaux de Petri hybrides, les réseaux fonctionnels, et présentent un outil pour la modélisation et l'analyse des HFPN - Genomic Object Net (GON). Ce formalisme élargit les notions des réseaux de Petri hybrides [3] et les réseaux de Petri fonctionnels [177] afin de les adapter à la modélisation des voies biologiques. Les réseaux fonctionnels hybrides contiennent des noeuds et des transitions de deux types : discrets et continus (Figure 2.9). Le formalisme HFPN a été appliqué pour modéliser le rythme circadien de la Drosophile et l'apoptose induite par le Fas ligand.

### 2.2.5 Modèles qualitatifs

#### Raisonnement qualitatif

La théorie des processus qualitatifs, créée par Kennet D. Forbus pendant son travail de thèse et développé ensuite par son groupe de recherche [51, 50], a beaucoup influencé le domaine de l'intelligence artificielle. Cette théorie introduit quelques idées clés pour la physique qualitative, qui peuvent s'appliquer également au raisonnement biologique. Tout d'abord, elle organise les connaissances physiques autour de la notion de processus, tel que le flux, le mouvement, la transition des phases. Ensuite, elle représente les valeurs numériques via des relations d'ordre. En physique et en biologie, de nombreux processus se déclenchent, s'arrêtent ou progressent en fonction des valeurs relatives des paramètres du système. Puis, tout événement qui a lieu dans le système est considéré comme causé par un ou plusieurs processus du système. Ainsi, cela permet de construire une hiérarchie des processus qui révèle les liens de cause à effets entre eux. De plus, la mathématique qualitative compositionnelle permet d'utiliser l'information partielle sur le système et de la combiner dès que possible. Par exemple, si on sait qu'un processus dépend de deux variables, nous pouvons les utiliser pour influencer le système, même sans connaïtre ni le mécanisme exact de cette dépendance, et ni les détails de l'interaction entre ces variables. Finalement, la théorie des processus qualitatifs représente explicitement les conditions et les hypothèses qui sont faites au cours de la modélisation et qui délimitent son applicabilité.

#### Equations différentielles qualitatives

Les équations différentielles linéaires par morceaux peuvent être analysées qualitativement en analysant des états qualitatifs qui correspondent à des domaines dans l'espace des phases. Lorsqu'une trajectoire passe d'un domaine à un autre, il y a une transition entre les états qualitatifs correspondants. Il en résulte un graphe de transition qui décrit qualitative-

ment la dynamique du système [43, 60]. Snoussi a démontré que le formalisme de Thomas généralisé peut être vu comme une abstraction d'un cas particulier [163]. L'idée d'abstraire une description discrète d'un modèle continu et d'analyser les équations discrètes au lieu des équations continues pour en tirer des conclusions sur la dynamique du système est courant en intelligence artificielle pour travailler sur des raisonnements qualitatifs. Un des formalismes les plus connus développés pour cela sont les équations différentielles qualitatives utilisées par les méthodes de simulation QSim [116] qui calculent les comportements qualitatifs. Contrairement aux ODEs, une variable $x$ ou $y$ prend des valeurs qualitatives composées d'une valeur discrète et d'une direction (signe de la dérivée). Tout comportement qualitatif des ODE correspond à un comportement généré par les QDE, mais la réciproque est fausse. Cette modélisation a été utilisée en biologie, notamment pour modéliser le réseau de contrôle de croissance du phage lambda [67] ou d'autres réseaux [2, 36, 174]. La difficulté de contraindre le système qualitativement conduit à une explosion de l'arbre des possibles et limite l'approche à l'étude de petits réseaux.

**Simulation qualitative**

Le but des mathématiques qualitatives est de prédire le comportement du système à partir d'équations de contraintes qualitatives [114]. Forbus [51] et Kuipers [115] définissent "l'espace des quantités" en tant qu'ensemble partiellement ordonnée des valeurs repères. Ainsi, une quantité est décrite par rapport à ces relations d'ordre avec les autres repères. L'approche de Kuipers [115] permet aussi la création de nouvelles valeurs repères au cours de la simulation qualitative.

Tous les systèmes de simulation qualitative produisent l'ensemble des comportements possibles par la génération et le filtrage des ensembles des transitions possibles d'un état qualitatifs vers ses successeurs. Les critères de filtrage sont locaux et dépendent des quantités dans les descriptions des états, ainsi que des contraintes structurales du système. A

cause de cela, la simulation qualitative peut également prédire des comportements incorrects, qui ne correspondent à aucun mécanisme sous-jacent satisfaisant la description du système.

La simulation qualitative commence par la description de la structure connue du système et d'un état initial, et produit un graphe dirigé qui contient tous les états possibles du système dans le futur, ainsi que les successeurs possibles pour tous les états. Les comportements du système sont des chemins dans ce graphe qui commencent à l'état initial. Si un état possède plus d'un successeur possible, la simulation diverge.

La structure du système est décrite par un ensemble de symboles qui représentent les "paramètres physiques" du système, c'est à dire les fonctions continues et différentiables, et un ensemble d' "équations de contraintes" qui décrivent les relations entre les paramètres physiques. Le temps est représenté par un ensemble complètement ordonné des moments symboliques qui sont générés au cours de la simulation. Le "domaine opératoire" définit l'intervalle des valeurs des paramètres pour lesquelles la contrainte donnée a un sens. L'état initial est également décrit par un ensemble de contraintes.

L'algorithme QSim [114] décrit les quantités comme un ensemble de repères ordonné linéairement. Dans le simulateur QSim, l'information quantitative obtenue peut être utilisée pour une simulation semi-quantitative ou numérique.

L'approche utilisée dans BioSim [68] permet de construire des modèles structurés en utilisant des objets et des processus, ainsi que l'application de contraintes fonctionnelles pour caractériser le comportement du modèle pendant un certain temps. Les équations qualitatives dans les objets et les processus peuvent être directement interprétés en termes biologiques, ce qui permet d'examiner leur consistance biologique, et la représentation graphique des résultats de la simulation qualitative en facilite la tâche. Cette approche reprend les idées de base du raisonnement qualitatif [51] telles que la vision du système modélisé à travers les objets et les processus. La manière de contraindre les variables de

façon qualitative est empruntée à [114]. Dans BioSim, les paramètres sont des fonctions du temps et ils prennent leurs valeurs dans un ensemble ordonné des repères. Les valeurs successives des paramètres sont déterminées par l'ensemble des transitions possibles.

Gensim [98] est un outil de simulation qualitative qui est destiné à représenter qualitativement la biochimie et à effectuer la simulation qualitative des systèmes biochimiques. Gensim utilise des schémas pour représenter les objets biochimiques qui correspondent aux populations homogènes des molécules. Cette représentation décrite la composition des objets complexes en leurs composants. Gensim utilise des bases de connaissances des schémas pour représenter les objets qui font partie du système à l'état initial dans les diverses expériences. Il utilise les schémas des processus pour représenter les réactions biochimiques. Les réactions sont arrangées dans une hiérarchie qui permet aux objets d'hériter une partie de leurs définitions à partir de classes plus générales de réactions. Le simulateur Gensim utilise l'information dans la base des connaissances des processus pour déterminer les réactions qui peuvent avoir lieu parmi les objets à travers d'un expérience, et de prédire quels seront les objets biologiques présents dans le système à la fin de l'expérience. Puisque les réactions biochimiques sont des événements probabilistes, Gensim partage à chaque pas de simulation toutes les molécules de même type en deux catégories : celles qui participent à la réaction, et celles qui restent inertes.

Le système de simulation de la réparation de l'ADN, basée sur l'environnement d'ingénierie des connaissances KEE, est présenté dans [18]. Dans cette simulation, l'information quantitative disponible est retranscrite dans les plages des valeurs symboliques pour le raisonnement qualitatif. Toute activité dans le modèle doit être spécifiée par de règles de type "IF-THEN-ELSE", ce qui rend compliqué l'ajout des règles contenant des relations arithmétiques. Les auteurs utilisent le raisonnement non-monotone, et les actions de simulation ne peuvent pas toujours être expliquées a posteriori puisqu'elles peuvent modifier la base des connaissances.

### 2.2.6 Modèles à états

#### Automates cellulaires

Les automates cellulaires sont des structures abstraites qui permettent d'étudier des univers virtuels dont on maîtrise l'ensemble des lois des interactions locales entre les éléments. Ce type de modélisation a été abondamment utilisé en biologie, car ils permettent de simuler des phénomènes d'émergence, comme la vie serait un phénomène émergent de l'ensemble des règles d'interaction entre les entités qui la composent. Les travaux tels que [182] peuvent fournir une revue plus détaillé de la modélisation par les automates cellulaires. Par exemple, dans le modèle Immune System Modeling and Simulation [147], un automate cellulaire est utilisé pour simuler les effets d'interactions entre les cellules et entre les cellules et les molécules dans le système lymphoïde. Basé sur le nombre des composants qui participent à la réponse immunitaire, le simulateur applique un ensemble spécifique de règles pour déterminer l'état suivant. Ce modèle possède une interface graphique conviviale (voir Figure 2.10) qui permet d'utiliser le modèle à des fins de recherche ainsi qu'en tant qu'outils pédagogique.

#### Réseaux booléens et réseaux logiques généralisés

Les réseaux booléens reposent sur l'approximation selon laquelle l'état d'activité d'un gène est discret (actif ou inactif) et que par conséquent son produit est présent ou absent. L'état d'un gène est modélisé par une fonction à valeur booléenne qui dépend de l'état des autres gènes. L'état d'un élément à l'instant $t + 1$ est calulé en fonction de l'état de $k$ autres gènes à l'instant $t$. Les réseaux booléens ont été les premiers pour lesquelles des méthodes d'inférences aient été proposées [121, 125, 138, 168]. Les réseaux booléens permettent d'analyser de grands réseaux d'interactions au prix de fortes simplifications. L'état d'activité transitoire des entités n'est pas modélisé, et les transitions sont modélisées

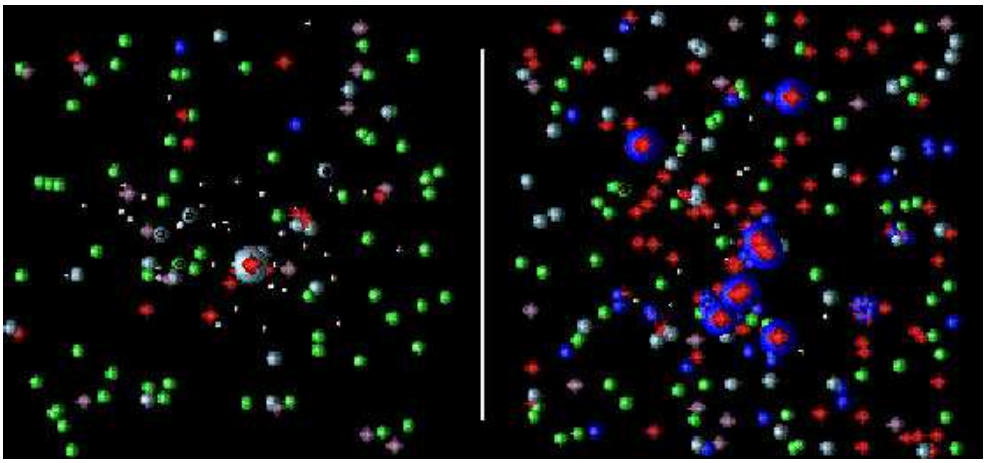


FIG. 2.10 – Capture d'écran d'une simulation du système immunitaire après l'infection primaire (à gauche) et secondaire (à droite) par un antigène (reproduit de http ://www.cs.princeton.edu/immsim/immsim.html).

de manière synchrone.

Une approche développée par R. Thomas et ses collaborateurs [169, 170, 173] est une généralisation dans laquelle le nombre d'états d'une entité peut être supérieur à deux, et dans laquelle les transitions peuvent être asynchrones. Les variables peuvent prendre cette fois des valeurs entières qui représentent de façon abstraite les concentrations des entités biologiques. Les valeurs possibles des $x_i$ sont définies par comparaison à des seuils de concentrations des $x_i$ ayant des rôles différents sur les autres états des entités du réseau. Etant donné que le nombre d'états abstraits est fini à cause de la discrétisation introduite par le formalisme, il est possible dans certains cas de tester tous les états stables du système. De tels réseaux ont été implémentés et utilisés dans plusieurs cas biologiques [165, 167, 170, 171, 153, 154, 130].

L'approche booléenne de René Thomas a été justifiée comme une discrétisation d'un système des équations différentielles continues [163]. Puis, il a été confronté à l'analyse plus classique en termes d'équations différentielles [102]. Ensuite, Thomas et Snoussi ont montré que tous les états stables du système peuvent être retrouvés via l'approche discrète [164].

Plus récemment, Thomas et Kaufman ont montré que cette description discrète génère les ensembles de paramètres qualitatifs des équations différentielles avec un nombre limité de combinaisons possibles pour leurs valeurs [172]. Plus généralement, les travaux de René Thomas et de ses collaborateurs fournissent une base pour le développement d'un cadre formel pour le calcul et l'analyse de la régulation génétique [29].

Les réseaux de régulation biologiques BRN (pour Biological Regulatory Networks) modélisent également les interactions entre les entités biologiques : ARNs ou protéines. Cependant, ce modèle permet au gène d'être activateur ou inhibiteur d'un autre gène en fonction de sa concentration. Il permet également l'application des outils de model-cheking pour l'analyse des modèles. La logique temporelle CTL peut exprimer les informations qualitatives sur la dynamique des modèles des réseaux biologiques modélisés à l'aide des BRN [13].

**Statecharts**

Un autre formalisme populaire est celui des Statecharts [65], où les états et les événements qui induisent les transitions entre les états sont utilisés pour capturer visuellement la dynamique du système. Les Statecharts permettent la modélisation concurrente, hiérarchique et multi-échelle [44, 49, 93]. Kam et ses collaborateurs [94] ont modélisé l'activation des lymphocytes T avec les Statecharts [65]. Cette approche a ensuite été enrichie par Efroni et ses collaborateurs [44]. Les grands ensembles de données sur la migration et la différentiation cellulaire, sur l'histologie et la microscopie électronique, la biochimie et la biologie moléculaire ont été intégrés dans un modèle à deux niveaux de la maturation thymique basé sur les Statecharts. L'exécution de simulations et leur analyse peuvent se faire à l'aide d'un outil performant de visualisation de l'information (voir Figure 2.11).

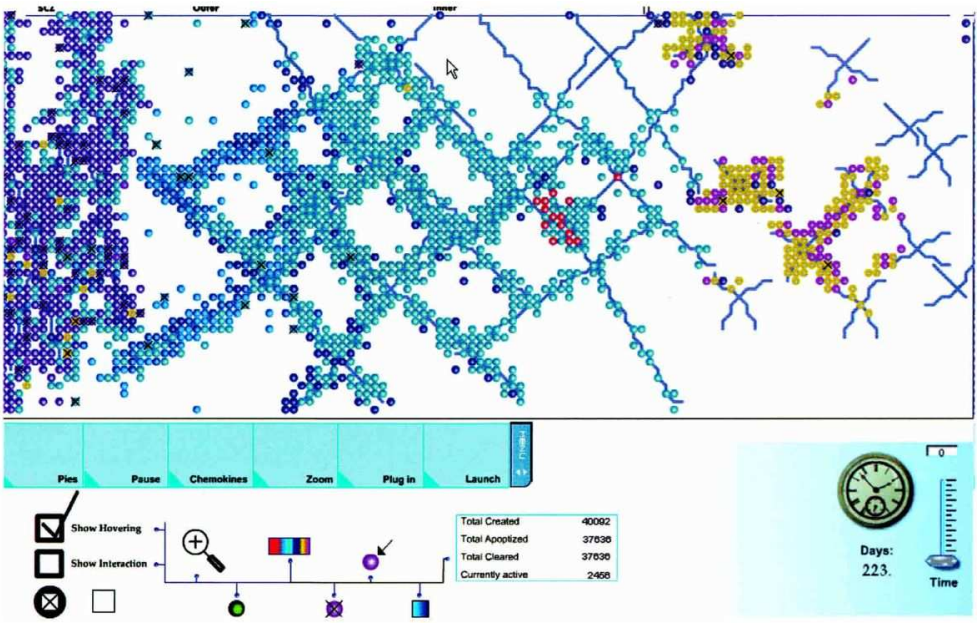



FIG. 2.11 – Capture d'écran d'une simulation de la maturation lymphocytaire, (reproduit de [44]).

### Formalismes à base de règles

Les formalismes que nous avons présentés jusque là étaient centrés sur une notion d'état d'entités biochimiques représentées par leur concentration à un instant donné. Les formalismes à base de règles, développés dans le cadre de l'intelligence artificelle et de l'informatique théorique, permettent de modéliser une plus grande variété de connaissances biologiques dans un formalisme unique. De tels formalismes sont constitués à partir d'un ensemble de faits et de règles, listés dans une base de connaissances [66]. Les règles sont composées d'une partie décrivant les conditions et d'une partie décrivant les actions. Une simulation consiste à répéter le processus de confrontations de la base de faits avec les conditions, et à opérer les actions dont les conditions sont satisfaites. De tels modèles ont été appliqués à la biologie, du phage lambda [132, 157].

L'avantage majeur des modèles à base de règles est la capacité de modéliser une grande variété de connaissances biologiques de manière intuitive. Leurs deux principaux

inconvénients sont qu'il est difficile de maintenir la consistance d'une base de connaissances (révisée) et il est difficile d'y incorporer des valeurs quantitatives [71], et donc quasi impossible de traiter des formalismes continus.

BioNetGen est une approche de modélisation à base de règles qui a été développée par Blinov et ses collaborateurs [14] pour aborder le problème de la complexité combinatoire des systèmes biologiques. Dans cette approche, tous les états possibles des domaines moléculaires sont spécifiés, ainsi que les règles pour les activités et les interactions de ces domaines. Quelques exemples de telles spécifications sont reproduits à la Figure 2.12. Les règles sont ensuite utilisées dans un outils informatique pour générer le réseau des réactions composé des différentes espèces et des réactions qui résultent des propriétés des domaines d'interaction de ces espèces. Chaque réaction peut être paramétrée par la le taux de réaction associé à une classe de réactions semblables définie par une règle spécifique. Un outil informatique, BioNetGen, a été développée pour faciliter l'utilisation de cette approche. Il permet également de traduire le réseau de réaction généré à base des règles en équations différentielles et de les sauvegarder en SBML. Ceci permet d'utiliser sur les mêmes modèles des algorithmes de simulation stochastique. Par exemple [14], un modèle qui contient 95 règles d'interaction génère 3680 réactions chimiques après la traduction en SBML.

Le développement de langages formels pour modéliser les systèmes biologiques ouvre la voie à la conception de nouveaux outils de raisonnement automatique destinés au biologiste modélisateur. La machine abstraite biochimique BIOCHAM est un environnement logiciel qui offre un langage simple de règles pour modéliser des interactions biomoléculaires, et un langage puissant fondé sur la logique temporelle pour formaliser les propriétés biologiques du système. En s'appuyant sur ces deux langages formels, il devient possible d'utiliser des techniques d'apprentissage automatique pour inférer de nouvelles règles de réaction, estimer les valeurs des paramètres cinétiques, et corriger ou compléter les modèles semi-automatiquement. BIOCHAM permet l'analyse et la simulation des modèles stochastiques,

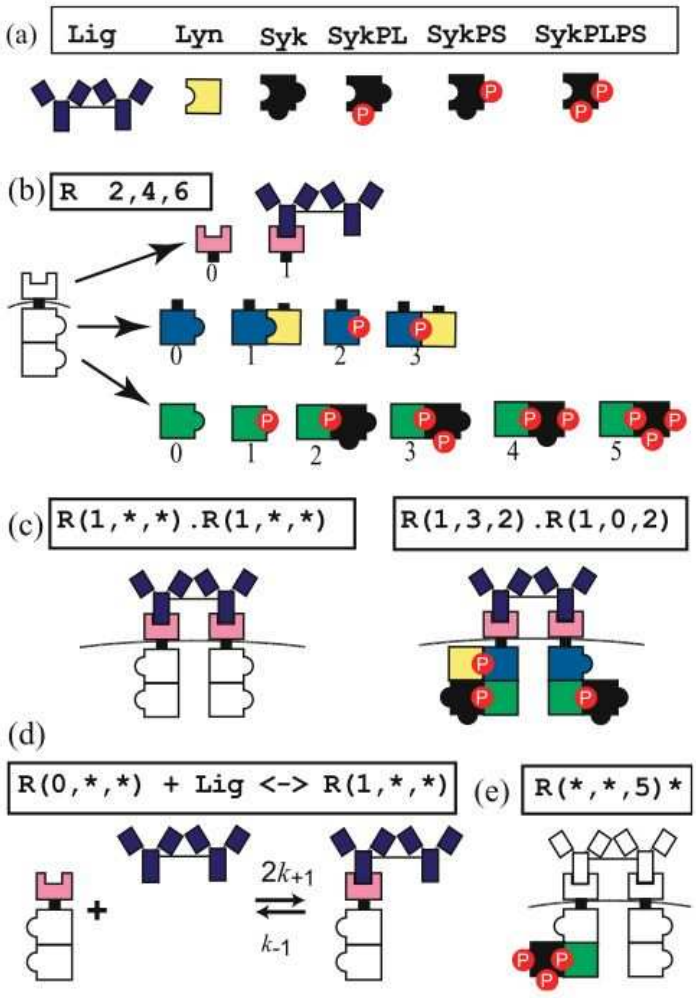



FIG. 2.12 – Les déclarations du fichier d'entrée pour BioNetGen et ses illustrations : le texte du fichier est montré dans les boîtes. (a) Déclarations de six espèces chimiques indépendantes. (b) Déclaration d'une espèce qui possède 48 états différents. (c) Déclaration de complexes qui contiennent deux récepteurs (à gauche) et une référence à une des 300 espèces moléculaires individuelles de même classe (à droite). (d) La règle pour la réaction de liaison d'un ligand au récepteur qui nécessite 24 réactions réversibles. Toutes les réactions ont le taux ($k_{+1}$ ou $k_{-1}$) associé. (e) Déclaration d'une fonction de sortie : la concentration du produit est fonction de la somme pondérée de 98 concentrations des espèces impliquées dans la réaction (reproduit de [14]).

cinétiques et booléens. Il intègre également la possibilité de fournir des contraintes sur le fonctionnement du système en tant que formules de logique temporelle. Pour les modèles cinétiques, il est possible d'explorer l'espace des paramètres des modèles afin de trouver ceux qui fournissent des comportements cohérents avec les spécifications contenues dans les formules de la logique temporelle. Par exemple, BIOCHAM vérifie automatiquement que le modélisateur ne se trompe pas à différentes étapes de modélisation : il vérifie que si une molécule ou une interaction est ajoutée dans le diagramme, les propriétés globales du système, exprimés par les formules de logique temporelle, sont conservées [21].

### 2.2.7 Algèbres des processus

Une algèbre de processus est un langage défini par une syntaxe et une sémantique. La syntaxe comprend un nombre réduit d'opérateurs algébriques primitifs (composition séquentielle, choix non déterministe, composition parallèle, etc.) qui, par assemblage, permettent de décrire des comportements complexes. Ainsi, un système asynchrone est-il décrit par un terme algébrique. La sémantique est définie formellement, de manière axiomatique ou opérationnelle. Une sémantique axiomatique consiste en un ensemble de lois algébriques (commutativité, associativité, distributivité des opérateurs...) qui permettent de démontrer l'équivalence de termes. Une sémantique opérationnelle consiste en une relation de transition $B \xrightarrow{L} B'$ exprimant le fait qu'un terme $B$ peut effectuer l'action $L$ puis évoluer et se transformer en un terme $B'$. Cette relation de transition est généralement définie par induction structurelle sur la syntaxe des termes en utilisant des formats de règles standards qui garantissent par construction que la sémantique est correcte. Elle détermine implicitement une correspondance entre un terme $B$ et un automate qui décrit les évolutions futures de $B$ (les transitions de cet automate étant étiquetées par les actions $L$ effectuées par $B$) ; il est ainsi possible d'exécuter les termes algébriques et de vérifier leur correction en analysant l'automate qui leur correspond [54].

L'expressivité des algèbres de processus permet de modéliser simplement les caractéristiques essentielles des systèmes asynchrones. Leur grande capacité de modélisation permet de représenter le modèle avec précision, et l'utilisation des opérateurs de l'algèbre permet de représenter la structure de l'objet de la modélisation. L'exécutabilité assure que les modélisations ne servent pas uniquement de documentation mais peuvent aussi être traitées par des outils de simulation, ou de prototypage rapide.

### $\pi$-calcul

Le $\pi$-calcul [133] est une algèbre de processus dont le niveau d'abstraction permet des applications très variées. Par exemple, Aviv Regev et ses collaborateurs l'ont appliqué à la description de la voie de signalisation RTK/MAPK [149]. Ils ont également développé un système de simulation des systèmes biologiques qui utilise le $\pi$-calcul pour exprimer les propriétés d'évolution des objets biologiques constituant le système étudié. Le $\pi$-calcul rend la structure des réseaux biologiques souple et capable d'évoluer dans le temps.

### Bio-ambients

En informatique, le calcul des ambients, créé par Luca Cardelli et Andrew D. Gordon en 1998 [24], a été utilisé pour décrire et étudier les systèmes concurrents mobiles. Ensuite, en considérant que les compartiments servent à organiser des systèmes biomoléculaires, composés de réseaux des protéines, Aviv Regev et ses collaborateurs [148] ont étendu leur variante de $\pi-$calcul pour la biologie afin de représenter ces compartiments. Quelques règles de calcul pour ce formalisme, appelé BioAmbients, sont présentées sur la Figure 2.13. Il est particulièrement adapté à la représentation des divers aspects de la localisation et de la compartimentation moléculaires, ainsi que du mouvement des molécules entre les compartiments, les modifications dynamiques des compartiments et les interactions des molécules dans l'environnement compartimenté. Ce calcul est intégré dans le système de

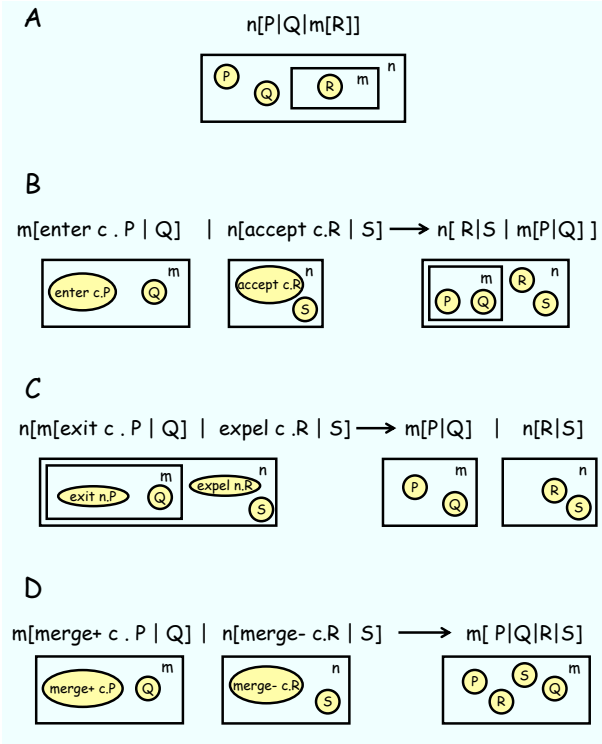



FIG. 2.13 – Les mouvements des ambients. A. Ambient $m$ (enfant) à l'intérieur de l'ambient $n$ (parent). B. Entrée de l'ambient $m$ dans l'ambient $n$. C. Sortie de l'ambient $m$ de son parent $n$ (reproduit de [148]).

simulation BioSpi [150].

### Calcul des Membranes

Le langage de membranes *Projective Brane Calculus* [34] permet de décrire le comportement des membranes de manière plus proche de la biologie que son inspirateur, le "Brane Calculus" de Luca Cardelli [22], en étendant la notion de Bio-ambient, avec la différence que le calcul a lieu sur la membrane, plutôt qu'à l'intérieur d'une membrane. Le Projective Brane Calculus (PBC) prend en compte le fait que les membranes biologique opèrent sous une forte contrainte : toute les implémentations moléculaires des actions effectuées sont dirigées. Alors, le PBC est enrichi par rapport au Brane-Calcul afin de représenter ces contraintes et d'obtenir une description plus précise des membranes biologiques. De plus,

ceci rend le calcul des membranes plus simple structurellement. Ce type de calcule permet de modéliser des processus biologiques où il est important de pouvoir distinguer et suivre les objets biologiques protéines qui se trouvent à l'intérieur, à l'extérieur ou sur la frontière des compartiments tels que les organelles intracellulaires ou les espaces intercellulaires. Ceci est une manière astucieuse de prendre en compte l'espace dans les modèles biologiques.

### 2.2.8 Conclusion sur les modèles dynamiques

Différentes méthodes de modélisation sont utilisées en biologies. Il n'est pas possible de trancher et de n'en retenir que quelques unes. Certaines s'appliquent à de grands systèmes, d'autres à de petits. Certaines sont qualitatives, d'autres sont quantitatives. Les hypothèses sur lesquelles chacune reposent sont différentes. Un modèle de réseaux biologiques devrait permettre de travailler avec toutes (ou chacune) de ces modélisations de dynamique. Les même connaissances biologiques, (on parle de modèle biologiques) contenues dans un modèle de réseau biologique, doivent pouvoir s'analyser en terme de dynamique avec ces différentes approche, de manière indépendante du formalisme choisi pour le modèle de réseau.

Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

# Chapitre 3

# Graph Rooting : étude de graphes partageant des noeuds

## 3.1 Présentation

De nombreux phénomènes que l'on peut étudier par modélisation à l'aide de réseaux ne sont pas isolés mais demandent d'être étudiés avec leur environnement, c'est à dire leur contexte. Il est possible de modéliser un tel phénomène comme un graphe interagissant avec un autre graphe. Nous caractérisons ici l'interaction entre deux graphes qui partagent des noeuds, et pour chaque graphe une structure en couche est définie en fonction de l'interface avec un autre graphe. Nous appliquons cette procédure, appelée enracinement d'un graphe par un autre graphe, au réseau d'interactions de la levure.

Nous démontrons que la procédure d'enracinement est intéressante pour étudier les réseaux hétérogènes d'interactions. Nous définissions des co-facteurs de transcription, et des co-co-facteurs de transcription, en fonction de la position du noeud, étant donné que les couches définies sont corrélées avec la localisation intra-cellulaire des protéines qui les composent. Nous montrons que la structure topologique de l'interface du réseau de régulation génétique et du réseau d'interaction protéine-protéine de la levure suggèrent que les interactions entre voies sont implémentées par des interactions protéiques qui interviennent préférentiellement au niveau des co-facteurs plutôt qu'au niveau des facteurs

de transcription ou des co-co-facteurs de transcription. Le réseau biologique étudié est disponible à l'adresse http ://magicalwebsite.com/bib. Cette étude a fait l'objet d'un article [160] publié à la conférence de la Société Francophone de Biologie Théorique à Winnipeg, 2007.

# Rooting a Graph by the Environment Interface Applied to Heterogeneous Interaction Network of the Yeast

Serge Smidtas*and Anastasia Yarsteva†

April 15, 2007

**Abstract**

**Motivation:** Many complex phenomena in natural and social sciences, finance and technology are not isolated but should be studied together with their environment. Thus, such phenomenon may be modeled as a graph interacting with another graph. Here, we characterize the interaction between two graphs sharing common nodes, and for each graph, a layered structure is defined respectively to the interface with another graph. We apply this analysis method, called rooting of a graph by another graph, to the biological interaction network of the yeast.

**Results:** We demonstrated that the graph rooting procedure is an interesting tool to study interacting networks. We defined Co-TFs and Co-CoTFs as a node position within the rooted network layers was correlated with the intracellular localization of the corresponding protein. We showed that topological structure on the interface of the genetic regulatory network and the protein-protein interaction network of the yeast suggests that 'crosstalk' between signaling pathways is mostly implemented by protein interactions that occurs on the level of Co-TF rather than between TFs or Co-CoTFs.

**Availability:** The studied biological network is available for browsing at http://magicalwebsite.com/bib. The rooting and shuffle scripts are available upon request.

**Contact:** iartseva@gmail.com

## 1 Introduction

Graph theory and statistical techniques for the analysis of networks provide a substantial background for studying complex network structures. For the homogeneous networks (with only one type of links), describing the interactions between agents represented by a network's vertices, previous works already tried to characterize their complexity [Berwanger *et al.*, 2005]. Other studies characterized network nodes or edges in terms of their importance. If there is no outside reference for a network, various definitions of its node importance exist, such as *centrality*, *closeness* or *betweenness* [Costa *et al.*, 2005]. For example, the centrality definition assumes that the greater the number of paths in which a vertex or edge takes part,

*ISI Foundation, Viale S. Severo 65, I-10133 Torino – Italy, CEA CNRS UMR8030, 2 rue Gaston Cremieux, 91000 Evry

†IBISC UMR 8042 CNRS - Universite d'Evry Val d'Essonne, Genopole, Tour Evry 2, 523 place des terrasses de l'Agora, F-91000 Evry, France

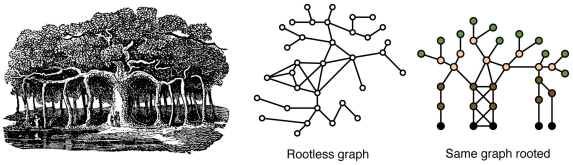



Figure 1: **Rooting of the Banyan Tree representation.** On the left, a banyan tree is represented. Only its aerial part is drawn, including leaves, branches, and stems, as well as the interface with the ground. In the middle, a rootless graph representing the aerial part of the tree is shown. In this representation, it is impossible to distinguish between branches and stems. On the right, the same graph is rooted with the graph depicting the roots of the tree. The rooted graph allows the localization of stems, branches, and leaves. Roots, represented in black, are joined. Stems (the first layer above the interface) are shown in brown, leaves (the last layer above the interface) in green, and branches (the intermediate layer) in orange.

the greater the importance of this vertex or edge is for the network [Costa *et al.*, 2005]. Assuming that interactions tend to follow the shortest paths between two vertices, it is possible to quantify the importance of a vertex or an edge in this sense by its *betweenness centrality* [Costa *et al.*, 2005]. The notion of centrality can be generalized to introduce sets, or circles, of nodes. They represent a collection of nodes with the same centrality. This has been studied extensively in social network analysis [Batagelj *et al.*, 1999].

However, systems modeled by networks are usually not isolated but interact with their environment. To tackle this problem, we modeled the environment of the system by another graph. Therefore, to capture the environmental influences on the network, we need to understand how corresponding graphs interact and influence one another.

In this work, we defined the *rooting* procedure (presented in Figure 1) which allows the organization of a graph nodes into *layers*, according to the distance between it and the graph *interface* with another graph. These layers are analogous to the circles of nodes with the same centrality for only one network. Our definition of rooting generalizes the common one [Weisstein, 2006] for multiple roots which are the nodes of the two networks interface. The analogous approach was already used in quantum mechanics to characterize interactions between system elements and to predict their dynamics [Giorda *et al.*, 2003].

In this paper, we analyzed the interaction between two interconnected networks, thus using one network as an outside reference for the other. The developed method was applied to the heterogeneous interaction network of yeast. First we found that the layer structure of rooted protein interaction network is strongly correlated with the intracellular localization (annotated in MIPS (Mewes et al., 2004)) of network nodes (proteins).The more far from the Transcriptional Factors we go into the interaction network, the more far the molecules are localized from the nucleus. This validates the method. Second, the topology of the layers is significantly different from the layer structure of the randomized networks with the same statistical properties. The connectivity between Co-Transtricptional Factors is a decade higher than among Transcriptional Factors suggesting that they implement where the integration of information occurs from the whole proteins to the small set of Transcriptional Factors.

# 2 Methods

## 2.1 Graph Rooting

The goal of the graph rooting procedure is to organise nodes of a graph in layers according to their distance from the interface with another graph. Such an interface provides "roots", an external reference for the study of the topology of a graph. To give the intuition of the rooting procedure, let us consider a graph $G$, representing a tree with its branches, leaves, roots and stems. We will call $G_2$ the *roots*, that section of the tree that remains underground, and $G_1$ its *aerial part*. Thus, the intersection of the vertices of $G_1$ and $G_2$ constitutes the *interface* between them. The rooting procedure consists of representing the initial graph with the interface placed horizontally, with roots below it and the aerial part above. After this, the vertices of the initial graph can be organized into *layers*, depending upon their *distance* from the interface. For example, a banyan tree (Ficus benghalensis) [Rodney Goke *et al.*, 1973] and its connectivity graph, $G$, is represented in Figure 1, left and center. The rooting produces a tree standing up above its roots. This rooted view, shown in Figure 1, right, allows distinguishing stems (the first layers above the interface) from leaves (the last layer above the interface) and branches (the intermediate layers).

Now, let us give a formal definition of the graph rooting procedure. The studied network is represented as a graph $G(V, E)$ with $V$ being the set of vertices and $E$ the set of graph edges.

**Definition 1 (Interface)** *Considering a pair of graphs* $(G_1(V_1, E_1), G_2(V_2, E_2))$*, their* interface*,* $I$*, is the set of nodes that are common to* $G_1$ *and* $G_2 : I = V_1 \cap V_2$*.*

We can define the layers in a graph $G_1$ as a set of its nodes at the same distance from the $G_1$ interface with its environment $G_2$.

**Definition 2 (Layers)** *Considering a pair of graphs* $(G_1(V_1, E_1), G_2(V_2, E_2))$*: a* layer$_1$-k *is defined in* $G_1$ *by* $layer_1 - k = \{n \in V_1 \mid dist(n, V_2) = k\}$ *, where* $dist(n, V_2) = min(card\{(n_i, n_{i+1})|1 \leq i < j, n_1 = n, n_j \in I, (n_i, n_{i+1}) \in E_1\})$ *is the distance between the node* $n$ *of graph* $G_1$ *and graph* $G_2$*, thereby defined as the length of the shortest path from* $n$ *to the interface,* $I$*.*

In a same way, the layers of the graph $G_2$ can be defined using $G_1$ as a reference..

## 2.2 Control networks for comparative study of layer topology

The shuffled and randomized graphs, obtained from the original ones and preserving their local and global topological properties, respectively, were used as a control for the rooting of the real biological networks.

To construct a *shuffled* control pair of graphs for $G_1$ and $G_2$, one of them ($G_2$) was left unchanged. The second graph, $G_1$, was *shuffled* [Bourguignon *et al.*, 2006], which means that from graph $G_1(V_1, E_1)$ we built a randomly shuffled graph $G'_1(V_1, E'_1)$ which is an isomporphism of the graph $G_1$ according to the permutation of nodes $\pi : V_1 \rightarrow V_1$ such that $E'_1 = \{(\pi(u), \pi(v))|(u, v) \in E\}$ (see Figure 2 for an illustration of the network shuffling).

To construct a *random* control pair of graphs for $G_1$ and $G_2$, one of them, $G_2$, was left unchanged. The links of the second graph, $E_1$, was removed and then randomly distributed between the nodes of both graphs, $V_1 \cup V_2$ (see Figure 2 for an illustration).

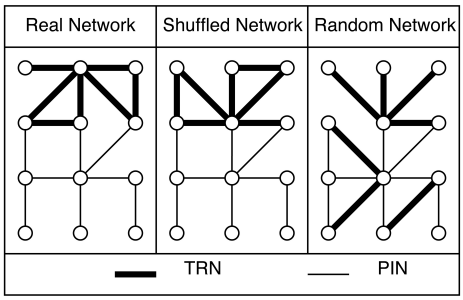



Figure 2: **Shuffled and random networks.** In the shuffled network, the names of the TRN nodes have been randomly permuted, but the links between them were preserved. In the random network, the links of the TRN have been removed and placed randomly between the whole set of nodes.

### 2.3 Biological Data and Model Networks

In this section, we describe the data used for demonstration of the rooting procedure, namely the biological interaction network of *Saccharomyces cerevisiae* (the yeast). We considered the yeast network of proteins, linked via the physical protein-protein interactions reported by Yeast Two Hybrid experiments [Ito *et al.*, 2001, Uetz *et al.*, 2000] or by complex purification [Gavin *et al.*, 2002, Ho *et al.*, 2002], thereby forming the Protein-protein Interaction Network (*PIN*).

The Transcriptional Regulation Network (*TRN*) was constructed with transcriptional factors, *TFs*, that regulate the transcription of other proteins (regulated genes). Regulated genes may be TFs themselves. The transcriptional regulation data comes from [Guelzim *et al.*, 2002, Lee *et al.*, 2002].

The two resulting biological graphs comprise 4488 vertices ($V_1 \cup V_2$) denoting yeast proteins and $E_1$ contains 24377 undirected edges, accounting for the presence of an interaction between proteins or for the presence of a complex comprised of at least two proteins, and $E_2$ contains 7412 directed edges accounting for the presence of transcriptional activation or inhibition of one protein by another. The topology of the PIN graph exhibits scale-free properties. The degree distribution follows a power law with an exponent $\gamma = -2.5\,(R^2 = 0.94)$. For more details see [Przujl *et al.*, 2004]. The TRN in-degree distribution follows an exponential with a coefficient $\alpha = -0.50\,(R^2 = 0.990)$; the out-degree also with a coefficient of $\alpha = -0.02\,(R^2 = 0.8)$.

The intracellular localization data associated to different proteins were extracted from the MIPS database [Mewes *et al.*, 2004].

## 3 Results

In this section, we present the results of the rooting procedure of the biological interaction network. We studied the structure of layers of TRN rand PIN rooted on each other.

Table 1: **Connectivity and node number for TRN and PIN layers.** For each layer, obtained by mutual rooting of TRN and PIN, the number of nodes and average degree of nodes are reported. The degree of the nodes at the interface is obtained by counting all the edges of the same type. Results are shown for the real, shuffled and random graphs.

| | TRN | | | | PIN | | | |
|---|---|---|---|---|---|---|---|---|
| | L3 | L2 | L1 | Interface | Interface | L1 | L2 | L3 |
| Real | | | | | | | | |
| degree | 0 | 1.2 | 5.4 | 4.9 | 11.3 | 6.9 | 1.3 | 1.4 |
| nodes | 0 | 685 | 1301 | 1401 | 1401 | 831 | 190 | 15 |
| Shuffled | | | | | | | | |
| degree | 0 | 1.4 ± 0.1 | 5.1 ± 0.9 | 4.5± 0.5 | 8.7± 0.1 | 10.9± 0.3 | 1.2 ± 0.1 | 0.8 ± 0.7 |
| nodes | 0 | 655 ± 130 | 1330 ± 130 | 1401 | 1401 | 851 ± 20 | 176 ± 6 | 1.4 ± 1 |
| Random | | | | | | | | |
| degree | 0.4 ± 0.4 | 0.8 ± 0.7 | 1.1 ± 0.1 | 4.1 ± 0.1 | 5.9 ± 0.1 | 1.3 ± 0.1 | 2.1 ± 1.6 | 0.3 ± 0.3 |
| nodes | 0.4 ± 0.4 | 2 ± 2 | 280 ± 25 | 3563 ± 40 | 3576 ± 30 | 613 ± 30 | 5 ± 4 | 0.3 ± 0.3 |

## 3.1 Structure of the PIN &TRN interface and layers

To study the layers obtained by mutual rooting of TRN and PIN, we first counted the number of nodes in each layer (Table 1). This structure was compared to shuffled and random networks. The biological network we examined had a relatively small interface (1400 vs. 3570 nodes for random one). Only 57% of PIN nodes were found to be also involved in genetic regulatory process. 41% of all regulated genes or transcriptional factors are also involved in protein-protein interaction.

To characterize the topological structure of each layer, we measured the average node degrees in each layer. The results were compared to those for shuffled control networks, maintaining the local and global topological properties of the original network, such as scale free behavior. In the PIN graph the degree of nodes at the interface was greater than in the shuffled graph (11.3 vs 8.7), and for the TRN graph it was exactly the same (4.9) (see Table 1). The difference of the average node degrees between the real network and the random one is even bigger then in a previous case (11.3 vs 5.9 for PIN and 4.9 vs 4.1 for TRN).

## 3.2 The PIN nodes upstream of TFs as transcriptional co-regulators

To explore the layer structure of PIN upstream of the transcriptional regulation, we studied the PIN rooted on a subnetwork of TRN composed of TFs only. TRN contains 3387 nodes, and only 157 of them are TFs, others being regulated and non regulator genes. So, restraining the roots of the PIN to TFs only enables us to examine the protein-protein interactions of transcriptional factors in order to gain comprehension on how the transcriptional activity of TFs is regulated in yeast.

In a PIN graph rooted on TFs, TFs constitute the interface layer (layer-0). Any node corresponding to the protein interacting with a TF and not being a TF by itself constituted the layer-1, and the protein was annotated as *CoTF*. A node corresponding to the protein interacting with a CoTF and not being a CoTF or TF, was attributed to the layer-2, and the protein classified as *Co-CoTF*. Thus, a CoTF (a Co-CoTF) is at the distance 1 (respectively,

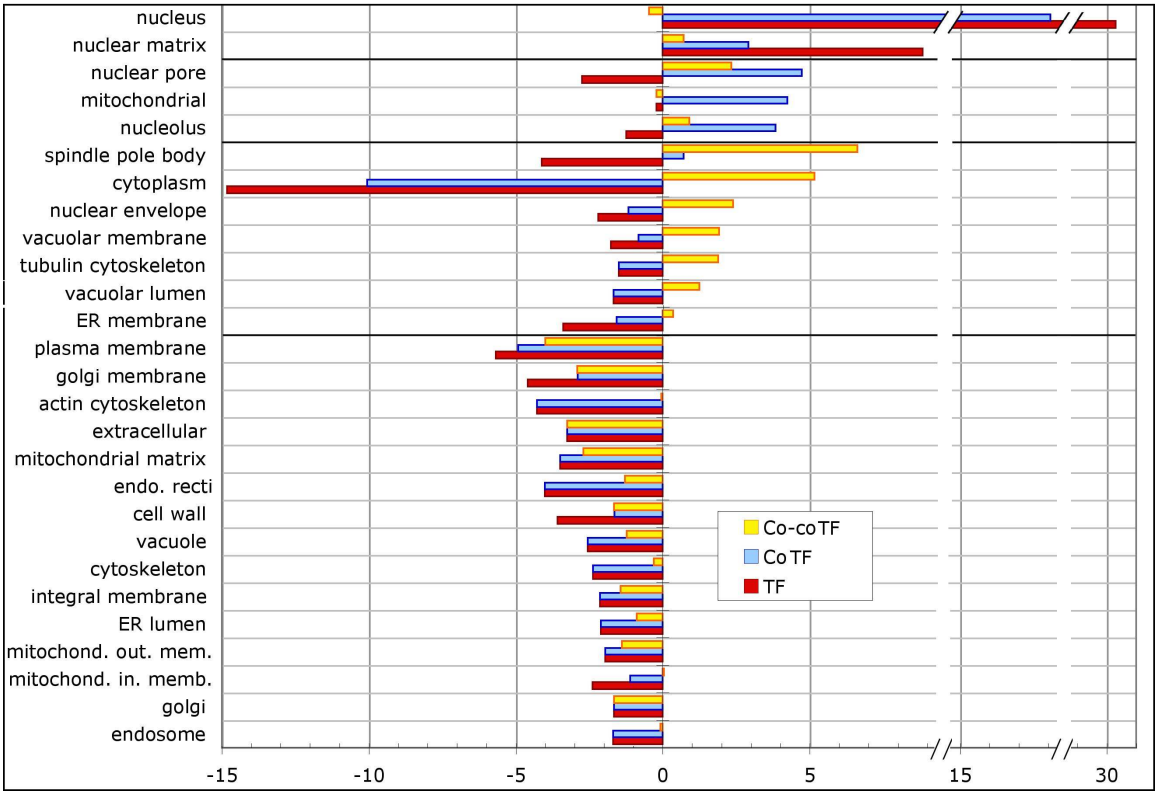



Figure 3: **The cellular localization of TFs, CoTFs and Co-CoTFs.** Number of proteins in each layer of PIN rooted on TFs was calculated for each MIPS localization category and compared with results for the random selection from overall yeast proteins. Only categories with significant difference of results are represented.

2) from a TF (see Figure 4).

There were 157 TFs, 186 CoTFs and 824 Co-CoTFs found in the PIN rooted on TFs, with average node degrees of 3.4, 13.8 and 13.3, respectively.

## 3.3 Cellular localization validates the node position in PIN layers

The proteins forming different layers in TRN and PIN networks may be annotated with their localization from the MIPS database [Mewes *et al.*, 2004]. To capture the spatial unfolding of the topology of the PIN rooted on transcriptional factors, the distribution of cell compartment localization for proteins from each layer was compared to the localization distribution of all yeast proteins (Figure 3).

The Figure 3 shows that TFs were significantly surrepresented in the nucleus and nuclear matrix. CoTFs were surrepresented in the nucleus, nuclear pore, mitochondria, nucleolus, nuclear matrix and spindle pole body. Co-CoTFs were surrepresented in the spindle pole body, cytoplasm, nuclear envelope, nuclear pore, nucleolus, nuclear matrix, vacuolar membrane and lumen, tubulin cytoskeleton, ER membrane and mitochondrial inner membrane.

Thus, the position of a node in the PIN rooted on TFs layer is correlated with the localization of the corresponding protein in the cell: the distance from the nucleus increases with the layer number.

### 3.4 The topology of PIN layers revealed

To characterize the topology of the layers of PIN rooted on TFs we computed the probability of protein-protein interaction for each pair of proteins within the same layer and between different layers. To compute the probability of interactions we divided the overall number of links between the nodes of the given types (for example, between a TF and a CoTF) and divided by the number of all possible pairs of nodes of these types (TFs number multiplied by CoTFs number). The results are reported in Figure 4 by the width of corresponding links between different nodes and link labels.

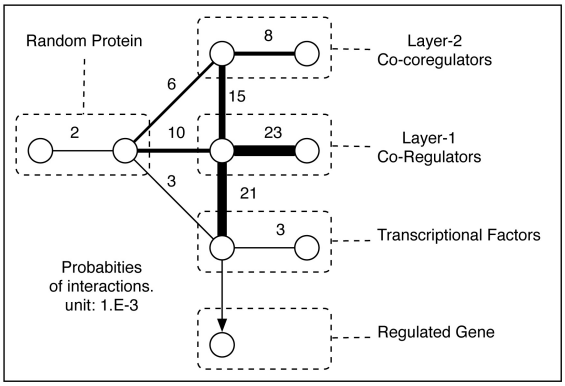



Figure 4: **Probability of interactions among proteins within and between layers.** The width and the label of the link between two nodes show the probability of protein-protein interaction between such nodes (for example, from the same layer or adjacent layers). The probability of interaction with any PIN node is indicated as a reference for each layer.

## 4 Discussion

From the Table 1 we saw that TRN and PIN contained mostly different nodes: 59% of TRN nodes did not have any known protein-protein interactions, and 43% of PIN nodes did not participate in transcriptional regulation. Most probably this phenomenon is related to the incompleteness of the biological interaction data.

The real network showed a higher average degree for the interface nodes than for the shuffled or random ones, especially for the PIN (11.3 vs 8.7±0.1 and 5.9±0.1, see Table 1). This result suggests that either the PIN/TRN interface proteins interact more than others, or that their interactions are better studied from the transcriptional regulation and protein-protein interaction point of view.

Shuffled graphs showed slightly higher number of nodes in the layer-1 (851±20 vs 831 in the real graph, see Table 1). We found it surprising, knowing that shuffled networks

presented lesser average degree of the nodes in the interface. This means that in the real network, compared to the shuffled one, the interface nodes interacted more in between them then with the nodes of the layer-1. For a node, being on the interface of TRN and PIN means that it participates in transcriptional regulation and in protein-protein interaction. The fact that the interface proteins were densely connected by protein-protein interactions could mean that they formed complexes, and that the proteins of these complexes are transcriptionally regulated or regulators themselves.

For the TRN, the topological structure of different layers, obviously, was not changed by shuffling (as shuffling is a sort of "rotation" of the graph nodes, see Figure 2), but was slightly different from the random graph, and the average node connectivity was significantly bigger for the layer-1 (5.4 for real graph vs $1.1 \pm 0.1$ for the random graph in Table 1). It means that transcriptional regulation includes longer regulatory cascades (two levels and more) then expected for a random network.

The distributions of the TFs, CoTFs and Co-CoTFs (see Figure 3) were consistent with the putative roles of proteins in each layer. TFs stay most of the time near promoter genes; CoTFs are also localized in the nucleus; for instance, they play an important role in DNA opening [Benecke *et al.*, 2003]. They can also constitute nuclear pores through which TFs or CoTFs can move from the cytoplasm into the nucleus. Finally, Co-CoTFs are less often confined within small compartments, as they are also involved in communication processes and may contribute to CoTF transport. From the perspective of the genetic regulation process, upstream network traffic flows away from receptor molecules through TFs, towards gene regulatory elements. Conversely, downstream traffic flows away from regulated genes towards the set of molecules that are under the influence of their expression. One of the well-known biological paths through layers is the galactose transcriptional regulation path [Smidtas *et al.*, 2005] (see Figure 5). Gal4 is a TF, therefore, Gal4 is at the interface with the PIN. Upstream, Gal80 can move from the cytoplasm to the nucleus going throw the nuclear pore. In the nucleus, Gal80 modifies the activity of Gal4. Gal80 also interacts with Gal3 in the cytoplasm. Gal3 binds to galactose that enter the cell through the Gal2 transporter. Each protein in this cascade represents one of the upstream layers. Gal80 is a CoTF. Gal3 is the Co-CoTF. Gal2 is at layer-3. Gal1, a galactokinase, also can bind to the CoTF Gal80. Gal1 plays his own role into the transcriptional activity of the Gal4 [Bhat *et al.*, 1990, Platt *et al.*, 2000, Timson *et al.*, 2002, Bhat *et al.*, 2004] and the Long Term Adaptation Pathway. The CoTF Gal80 integrates the two signals coming from Gal1 and Gal3 with an OR function. Gal4 is transcriptionnaly activated if Gal1 or Gal3 is present. As illustrated by this example, the higher the layer is, the farther the proteins are from the place where transcription occurs. Figure 5 shows cell compartments with colors characteristics of each layers. These colors also correspond to the spatial distance of cell compartments from the nucleus. Thus, the results produced by rooting procedure were consistent with other biological observations and enabled the validation of our approach.

From the topological point of view, in the PIN rooted by TFs each layer had its own distinctive characteristics (see Figure 4). For example, the layer-1, corresponding to Co-TFs, was very dense, and its connectivity with TFs was very strong, as well as with Co-CoTFs. However, TFs themselves were found to participate in barely more protein-protein interaction with one another then with a random protein. We found that instead TFs interacts about 10 times more through their CoTFs. This is in line with results that the average path length between two TFs is of three steps in PIN graph [Manke *et al.*, 2003, Chen, 1999], thus passing

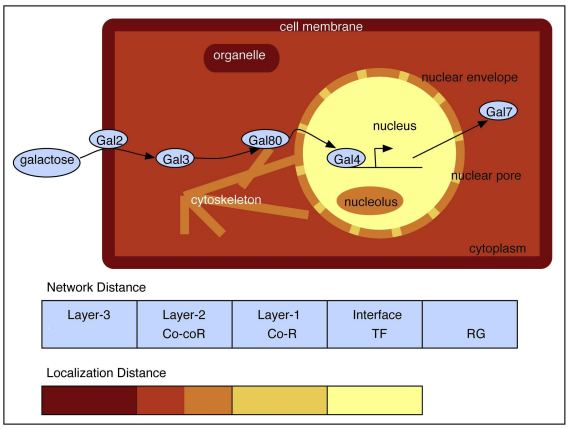



Figure 5: **Galactose loop localization layers and network distance**. The cell is illustrated in this figure. Colors correspond to spatial distance from the nucleus to the rest of the cell. The galactose regulatory pathway is also illustrated. Galactose enters the cell throw the Gal2 transporter in the cytoplasm. It then binds Gal3. This interaction free Gal80 that can then goes to the nucleus. There it activates Gal4 that regulates the transcription of Gal7. Gal1 that has galactokinase also binds to the CoTF Gal80.

from one TF to its CoTF, between two CoTFs and from the second CoTF to the second TF.

One of the disadvantages of our approach is its sensitivity to the input sets: rooting a graph by irrelevant roots may give results that are difficult to interpret. However, the rooting procedure could be very useful to study the relative structure of two interacting graphs, representing, for example, a system and its environment, especially in biology where the frontier between both of them is very difficult to define precisely.

Earlier attempts to study the topology of the PIN and the TRN together, focused on local structures such as motifs made of the two types of interactions [Yeger-Lotem *et al.*, 2004], were not able to highlight macroscopic topological structures such as layers. A similar approach was recently developed independently, that focus more specifically on the TRN [Yu *et al.*, 2006] defining level of hierarchy that represent a different macroscopic topological property.

# 5 Conclusion

As we knew, networks are not isolated as for example the PIN network in yeast. We searched to determine how it interacts with its environment, modeled by the TRN. We added the TRN environment to the PIN introducing a new point of view on this network leading to the annotation of proteins. This automatic functional annotation was validated with localization data. Knowing that the network which represent real systems are highly heterogeneous, we showed that the focusing on a homogeneous graph may be patched by considering the environment of the modeled system as another graph. The methodology presented here is general and powerful so we are now applying it to other pairs of networks in economy and social science.

We have considered the specific example of biological interaction network, where it was possible to appreciate the importance of the correlations between corresponding pairs of networks composed of protein-protein interactions and transcriptional regulation interactions, and the importance of topological studies in the characterization of the real network properties. Indeed, the analysis of the relative topologies of pairs of networks functioning together provide a complementary perspective on the structural organization of a global network, a perspective that might remain undetected by quantitative analysis based only on topological information derived from each graph independently. Consequently, our study offers a quantitative and general approach to understanding the complex architecture of real networks that are not isolated, but that interact with their environment represented as another network.

## References


[Yu *et al.*, 2006] Yu H., Gerstein M. (2006) Colloquium Papers: Genomic analysis of the hierarchical structure of regulatory networks, *PNAS 103; 14724-31*

[Yeger-Lotem *et al.*, 2004] Yeger-Lotem E., Sattath E., Kashtan N., Itzkovitz S., Milo R., Pinter R.Y., Alon U., Margalit H., (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction, *PNAS 101 (16) 5934-9*

[Platt *et al.*, 2000] Platt A., Ross H., Hankin S., Reece R. (2000) The insertion of two amino acids into a transcriptinal induces converts it into a galactokinase, *PNAS 97 (7) 3154-9*

[Bhat *et al.*, 2004] Bhat P., Venkatesh K.V. (2004) Stochastic variation in the concentration of a repressor activates GAL genetic switch: implications in evolution of regulatory network, *FEBS Letters 29191 579-603*

[Bhat *et al.*, 1990] Bhat J., Oh D., Hopper J. (1990) Analysis of the GAL3 Signal Transduction Pathway Activating GAL4 Protein-Dependent transcription in Saccharomyces cerevisiae, *Genetics 125 281-91*

[Timson *et al.*, 2002] Timson D., Ross H., Reece R. (2002) Gal3p and Gal1p interact with the transcriptional repressor Gal80p to form a complex of 1:1 stoichiometry, *Biochem. J. 363 515-20*

[Batagelj *et al.*, 1999] Batagelj, V., Mrvar, A., Zaversnik, M. (1999). Partitioning Approach to Visualization of Large Networks, *Castle Stirin, Czech Republic, LNCS 1731, In Graph Drawing'99.*

[Benecke *et al.*, 2003] Benecke, A., (2003) Genomic Plasticity and Information Processing by Transcription Coregulators, *Complexus*, **1**, 65-76.

[Berwanger *et al.*, 2005] Berwanger, D., Grdel, E. (2005) Entanglement ? A Measure for the Complexity of Directed Graphs with Applications to Logic and Games, *Logic for Programming, Artificial Intelligence, and Reasoning: 11th International Conference, LPAR 2004, Montevideo, Uruguay. Proceedings, LNCS*, **3452/2005**, 209.

[Chen, 1999] Chen, L. (1999). Combinatorial gene regulation by eukaryotic transcription factors, *Curr. Opin. Struct. Biol.* , **9**, 48-55.

[Costa *et al.*, 2005] Costa, L.F., Rodrigues, F.A., Travieso, G., Villas Boas, P. R. (2005) Characterization of complex networks: A survey of measurements, *Preprint cond-mat/0505185.*

[Gavin *et al.*, 2002] Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, **415**, 141-7.

[Giorda *et al.*, 2003] Giorda, P., Zanardi, P. (2003) Mode entanglement and entangling power in bosonic graphs, *Phys. Rev. A*, **68**, 062108.

[Guelzim *et al.*, 2002] Guelzim, N., Bottani, S., Bourgine, P., Kepes, F. (2002) Topological and causal structure of the yeast transcriptional regulatory network, *Nat Genet*, **31**, 60-3.

[Ho *et al.*, 2002] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., Tyers, M. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry, *Nature*, **415**, 180-3.

[Ito *et al.*, 2001] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc Natl Acad Sci U S A*, **98**, 4569-74.

[Lee *et al.*, 2002] Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K.,Young, R. A. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae, *Science*, **298**, 799-804.

[Manke *et al.*, 2003] Manke, T., Bringas, R., Vingron, M. (2003) Correlating Protein ? DNA and Protein ? Protein Interaction Networks, *J. Mol. Biol.*, **333**, 75-85.

[Mewes *et al.*, 2004] Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J., Ruepp, A. (2004) MIPS: analysis and annotation of proteins from whole genomes, *Nucleic Acids Research*, **32 Database issue**, D41-4.

[Przujl *et al.*, 2004] Przulj N., Corneil, D.G., Jurisica, I. (2004) Modeling interactome: scale-free or geometric? *Bioinformatics*, **20(18)**, 3508-15.

[Rodney Goke *et al.*, 1973] Rodney Goke, L., Lipovski, G. J. (1973) Banyan networks for partitioning multiprocessor systems, *Proceedings of the 1st annual symposium on Computer architecture*, 21-28.

[Bourguignon *et al.*, 2006] Bourguignon, P.Y., Danos, V., Kepes, F., Schachter, V., Smidtas, S., (2006) Property driven statistics of biological networks, *LNCS Transactionson Computational Systems Biology (2006) in press.*

[Smidtas *et al.*, 2005] Smidtas, S., Schachter, V., Kepes, F. (2005) The adaptive filter of the yeast galactose pathway, *J. Theor. Biol. (in press)* doi:10.1016/j.jtbi. 2006.03.005.

[Uetz *et al.*, 2000] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J. M. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, *Nature*, **403**, 623-7.

[Weisstein, 2006] Weisstein, E.W. (2006) Rooted Graph, *From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/RootedGraph.html*

## 3.2 Conclusion

Ce travail nous a permis de nous intéresser à la représentation des graphes hétérogènes représentant les interactions biologiques. La méthode d'enracinement des graphes consiste à définir dans deux graphes qui partagent des nœuds communs, des ensembles de noeuds appelés 'couches' selon la distance des nœuds qui les composent aux noeuds communs aux deux graphes constituant l'interface. Nous avons constaté que la régulation de l'activité des facteurs de transcription dans la levure passe en grande partie via les interactions entre les co-facteurs. Ceci est en accord avec les observations biologiques qui montrent la formation dynamique de complexes régulateurs entre les facteurs de transcriptions et leurs co-facteurs chez les eukaryotes. Pour pouvoir aller plus loin, la représentation des réseaux biologiques sous la forme de graphes simples est insuffisante, car cette représentation ne permet pas de représenter les complexes protéiques régulateurs qui sont des relations $n$-aires avec $n > 2$. D'où notre effort pour construire une représentation des réseaux biologiques sous la forme de graphes bipartis qui permettent non seulement d'exprimer les relations $n$-aires entre les objets biologiques mais également de préciser leurs rôles respectifs dans ces relations. Cette représentation, appelée MIB pour *Model of Interactions in Biology*, est présentée dans le chapitre suivant.

Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

# Chapitre 4

# MIB : Un modèle biparti de réseaux biologiques

Ce chapitre est composé de deux parties : la première introduit le formalisme MIB, et la seconde présente l'outil informatique BIB, basé sur MIB, qui permet d'effectuer des recherches de motifs hétérogènes dans les réseaux biologiques.

## 4.1 Modèle MIB

Un défi important pour la bioinformatique et la biologie théorique est de construire un modèle unifié qui intègre de nombreuses connaissances biologiques issues notamment d'expériences à haut débit, mais qui permette aussi leur analyse. Des travaux antérieurs ont analysé des données homogènes indépendamment les unes des autres (interactions protéiques, régulation génétique, métabolisme, synexpression) en les modélisant par des graphes [175, 92, 63, 119, 70, 55]. Toutefois ces modèles ne permettent pas de comprendre comment ces différentes interactions implémentent ensemble une fonction biologique. Plusieurs études indépendantes conduites en même temps ont tenté d'agréger plusieurs types

de données biologiques, la plupart en essayant d'étendre l'approche de Uri Alon, basée sur la recherche de motifs dans des graphes sous- ou sur-représentés [156], uniquement en considérant des propriétés topologiques de graphes biologiques. Toutes ces études sont malheureusement basées sur un modèle de graphes trop pauvre pour permettre des analyses intégrant plusieurs types de données. Cependant, les études précédentes ont montré que les données phénotypiques peuvent être combinées avec succès avec des données de l'interactome et des données d'expression pour générer un réseau de relations fonctionelles pour l'embryogénèse précoce de *C. elegans* [64].

Pour cette raison, nous avons établi un modèle dérivé d'un graphe biparti pour modéliser les réseaux hétérogènes d'interactions biologiques. Ce modèle représente la dynamique qualitative des réactions biochimiques, et modélise les interactions $n$-aires. Il comprend des interactions protéiques, des complexes, des liens de régulation transcriptionnelle, des réactions métaboliques, de liens de *synthetic lethality* ou de coexpression. Le modèle a été implémenté et s'accompagne d'une interface web graphique permettant de saisir et de rechercher des motifs hétérogènes. Le modèle est illustré par des exemples. Nous proposons notamment des mécanismes moléculaires sous-jacents à la coexpression de gènes. Dans le modèle, il est par exemple possible de rechercher des instances du motif composé de deux complexes constitués de dix protéines. Si un modèle de graphe simple était utilisé à la place, une unique instance de ce motif serait comptabilisée 2025 fois ! En effet, chacun des complexes sera modélisé par 45 interactions binaires, qu'il faut ensuite combiner deux à deux.

Les résultats suivants ont été publiés dans les Comptes Rendus de l'Académie des Sciences de Biologies en 2006 [161].

Available online at www.sciencedirect.com
ScienceDirect
C. R. Biologies 329 (2006) 945–952
COMPTES RENDUS BIOLOGIES
http://france.elsevier.com/direct/CRASS3/

Biological modelling / Biomodélisation

# Model of interactions in biology and application to heterogeneous network in yeast

Serge Smidtas [a], Anastasia Yartseva [b,c,*], Vincent Schächter [a], François Képès [d]

[a] *Genoscope and CNRS UMR 8030, 91057 Évry cedex, France*
[b] *IBISC–université d'Évry-Val-d'Essonne, tour Évry 2, 523, place des Terrasses de l'Agora, 91000 Évry, France*
[c] *ISI Foundation, Viale S. Severo 65, 10133 Torino, Italy*
[d] *Epigenomics Project, and Atelier de génomique cognitive (ATGC), CNRS UMR 8071, Génopole, 523, Terrasses de l'Agora, 91000 Évry, France*

Received 23 March 2006; accepted after revision 27 June 2006

Available online 7 August 2006

Presented by Michel Thellier

**Abstract**

A major challenge for bioinformatics and theoretical biology is to build and analyse a unified model of biological knowledge resulting from high-throughput experiment data. Former work analyzed heterogeneous data (protein–protein interactions, genetic regulation, metabolism, synexpression) by modelling them by graphs. These models are unable to represent the qualitative dynamics of the reactions or to model the $n$-ary interactions. Here, MIB, the Model of Interactions in Biology, a bipartite model of biological networks, is introduced, and its use for topological analysis of the heterogeneous network is presented. Heterogeneous loops and links between synexpression pattern and underlying molecular mechanisms are proposed. ***To cite this article: S. Smidtas et al., C. R. Biologies 329 (2006).***


**Résumé**

**Modèle de réseaux d'interactions biologiques.** Un défi important pour la bioinformatique et la biologie théorique est de construire un modèle unifié qui intégre de nombreuses connaissances biologiques, issues notamment d'expériences haut débit, et qui permette leur analyse. Des travaux antérieurs ont analysé des données hétérogènes (interactions protéiques, régulation génétique, métabolisme, synexpression), en les modélisant par des graphes. Toutefois, ces modèles ne sont capables, ni de représenter la dynamique qualitative des réactions biochimiques, ni de modéliser les interactions $n$-aires. Un modèle bipartite des réseaux hétérogènes MIB (modèle d'interactions biologiques), est présenté et illustré par les résultats d'analyse des boucles régulatoires hétérogènes ainsi que des mécanismes moléculaires sous-jacents à la synexpression des gènes. ***Pour citer cet article : S. Smidtas et al., C. R. Biologies 329 (2006).***


*Keywords:* Formal model; Biological network; Heterogeneous data

*Mots-clés :* Modèle formel ; Réseau biologique ; Données hétérogènes

* Corresponding author.
*E-mail addresses:* sergi@sergi5.com (S. Smidtas), anastasia.yartseva@sergi5.com (A. Yartseva), vs@genoscope.cns.fr (V. Schächter), Francois.Kepes@genopole.cnrs.fr (F. Képès).

1631-0691/$ – see front matter 
doi:10.1016/j.crvi.2006.06.010

## 1. Introduction

The last few years have seen the advent of high-throughput technologies to analyze various properties of the transcriptome and proteome of several organisms. The congruency of these different data sources, or lack thereof, can shed light on the mechanisms that govern cellular function. A central challenge for bioinformatics research is to develop a unified framework for combining the multiple sources of functional genomics information, thus obtaining a robust and integrated view of the underlying biological phenomena.

Since the complete DNA sequence of *S. cerevisiae* became available in 1996 [1], a variety of large-scale, high-throughput experimental studies have provided partial, potentially complementary insights into the structure of the yeast regulatory network and, indirectly, into its dynamics.

A major challenge of the post genomic research is to understand how cellular phenomena arise from the interaction of genes, proteins and metabolites. Investigations into the structure of these molecular interaction networks include studies on their global topological properties [2,3], such as connectivity distribution [4] or scale-free nature [5] have been performed. The local properties such as clustering proteins within the network into functional subnets using combinations of attributes and local connectivity properties to uncover a higher level of network organization [4,6–9] were also studied on each homogeneous network separately.

Several studies [8,10,11] have already tried to aggregate many types of data, mostly extending the approach of [31], based on the research of under- or over-expressed static graph motifs, only in order to understand the topological properties of biological graphs.

In previous work, gene expression data in *Saccharomyces cerevisiae* have already been combined with gene ontology-derived predictions [8] and phenotypic experiments [12]. Recent studies assembled an integrated *S. cerevisiae* network, in which nodes represent genes (or their protein products) and differently coloured links represent five types of biological interactions: protein–protein interaction, genetic interaction, transcriptional regulation, sequence homology, and expression correlation [10,11].

However, most of these studies rely on the graph-theoretic approach, which fails to represent $n$-ary relations between biological objects, for example in metabolic networks or complexes, as well as qualitative dynamics of the interaction: for example, the distinction between activation and inhibition, production and consumption.

In this work, we present a bipartite graph model of heterogeneous biological network that comprises directed transcriptional regulation, protein–protein interaction, the complexes, the metabolic networks, synthetic lethality experiments and micro-array expression results.

This type of models allows searching for complex heterogeneous network motifs with qualitative dynamics and biologically relevant properties.

Based on this model, the *S. cerevisiae* dataset was represented as a global database including the aforementioned data types.

## 2. The MIB model

The main model-constructing principle that we used is made to apprehend the organization of the complex system that constitutes the cell with its distributed control (see Fig. 1). Here we proposed a qualitative modelling framework, Model of Interactions in Biology (MIB), a bipartite graph model of heterogeneous biological network. MIB is designed to fill the gap between, on the one hand, existing techniques for quantitative modelling of biological systems [13–16], and, on the other hand, techniques for analysis of the network structure mostly based on graph theory [2,3,5]. Our approach is largely inspired by the Structured Analysis and Design Technique [17].

A biological system can be seen as an emergent [18] phenomenon of the chemical reactions set, including protein–protein interaction (PPI) and transcriptional regulation interactions (TRI). This set may be modelled by a composite reactions network and it should satisfy the following constraints:

- to include information about chemical species and chemical reactions of the biological system;
- to consider biological interactions that are not binary, like in the case of a complex of several proteins;
- to distinguish between undirected and directed (positive or negative) interactions of species;
- the representation should be simple enough to allow the study of global structural properties of the network and the search for sub-networks in the composite network.

Thus, the set of biochemical reactions composing the biological system is represented in MIB as a network that comprises nodes, either *entities* (chemical species) or *transformations* (chemical reactions), and links between nodes, divided in four *roles*: *consumed, produced,*

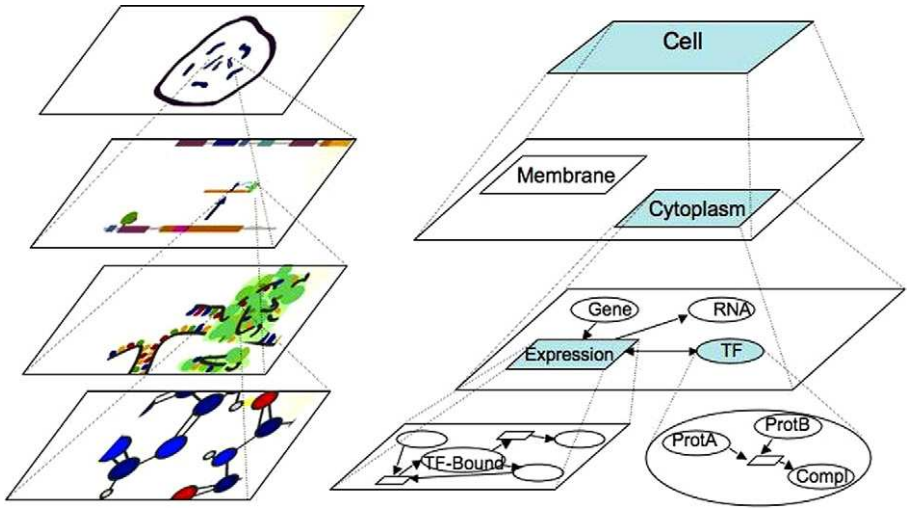



Fig. 1. Representation of biological systems seen as a set of chemical reactions. The top layer represents the most general view of the hierarchy. The bottom layer is the most detailed view of the system structure. Two intermediate layers are presented, showing the topological or functional structure of the system. On the left side (top down), the artistic view of a cell with chromosomes is shown, followed by the gene regulatory network scheme, the translation and ribosomal machinery layer and interacting molecules and atoms layer (captures of the artistic MIB movie: http://sergi5.com/bio/MIB). On the right side (top down), the artistic view of the biological system is modelled in MIB. The first layer box represents the cell that contains membrane and cytoplasm (second layer). Zooming out the cytoplasm (third layer), gene expression, involving a transcriptional factor, is represented. At the bottom layer, the transcriptional factor is magnified into a complex made of two proteins, and gene expression is symbolized by the transient TF/DNA complex.

*activates, inhibits*. The same chemical species may have different properties and participate in different reactions depending on intracellular localization. In this case, such a species may be represented by more than one entity in the MIB model. The next paragraph presents the formal definition of the MIB model.

**Definition 1** *(MIB model).* The MIB network $N$ is a tuple $(\{X, Y\}, E)$ where:

- $X$ is a set of *entities* $x = (n, l, t)$ where $n$ is a *name*, $l$ is a *localization*, and $t$ is a *type* of the entity;
- $Y$ is a set of *transformations* $y = (n, s, t)$ where $n$ is a *name*, $s$ is a *speed* (kinetic rate) and $t$ is a *type* (e.g., inversible or not, protein–protein or DNA–protein, etc.) of the transformation;
- $E$ is a set of *links* $(x, y, r)$ or $(y, x, r)$ where $x \in X$ is an entity and $y \in Y$ is a transformation and $r$ is one of four possible *roles* (production, consumption, activation, and inhibition) of an entity $x$ in a transformation $y$.

Kinetic rates can be dependent on the biological context. The above definition does not make any restriction on it.

The MIB network $(\{X, Y\}, E)$ can be represented graphically as a bipartite graph (as shown in Fig. 2) where elliptic nodes represent entities $X$ and rectangular ones represent transformations $Y$. Nodes are labelled with the attributes of related entities and transformations. Edges of this graph represent links $E$ between an entity and a transformation. There are four arrow types to express four possible roles of an entity in a transformation: production ($\square \rightarrow \bigcirc$) or consumption ($\bigcirc \rightarrow \square$) of an entity by a transformation and activation ($\bigcirc \leftrightarrow \square$) or inhibition ($\bigcirc \dashv \square$ ) of a transformation by an entity.

In the following paragraphs, two examples of MIB model of common biochemical reactions will be presented. The first example is catalytic. The second is stoichiometric.

**Example 1** *(Transcriptional regulation).* One of the important properties of the reaction *transcriptional regulation* is that the participating species are not consumed (this type of reaction can be also called *gene expression regulation*). This type of reaction (the expression of Gal3 protein) is shown in Fig. 2A. The *GAL*3 gene and transcriptional factor Gal4p are needed for the reaction (they activate it), but are not consumed [19].

More generally speaking, the *information transfer reaction* represents the production of a biological macromolecule using the informational template (DNA for transcription or RNA for translation reaction). The template is not consumed in such a reaction.

**Example 2** *(Association reaction).* In Fig. 2B, the complexation of Gal3 and Gal80 proteins and of galactose

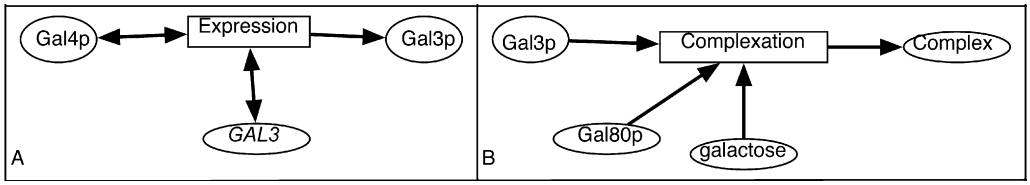



Fig. 2. Examples of representation of a biological system. **A.** In yeast, Gal4p is the transcriptional factor that regulates the GAL3 gene. **B.** Gal3p, Gal80p and galactose constitute a complex.

is represented [19]. This is an example of a chemical reaction that can not be represented with a simple graph because it involves three different entities. It may be labelled with the kinetic rate. The association reactions are generally reversible, and the corresponding reverse transformation could also exist and encoded in a distinct reaction.

The *topology* of the MIB or its parts can be described by *motifs*, thus characterizing the number of reactions, species and roles of the species in the system.

**Definition 2** *(Motif of MIB and its occurrence).* A *motif* $M$ on MIB is a tuple $\{(X_M, Y_M), E_M\}$ where:

- $X_M$ is a set of entities;
- $Y_M$ is a set of transformations;
- $E_M$ is a set of links between entities and transformations of the motif.

An *occurrence* of a motif $M$ in the MIB model $N = \{(X_N, Y_N), E_N\}$ is a sub network $O = \{(X_O \subset X_N, Y_O \subset Y_N), E_O \subset E_N\}$ and two bijections $B_X : X_O \to X_M$ and $B_Y : Y_O \to Y_M$ can be established between nodes of both graphs such that, if $x_M = B_X(x_O)$, $l_{x_M} \in l_{x_O}$, $t_{x_M} \in t_{x_O}$ and $y_M = B_Y(y_O)$, $s_{y_O} \in s_{y_M}$, $t_{y_O} \in t_{y_M}$, then $\forall (x_M, y_M, r_M) \in E_M \; \exists r'_M$: $\exists (x_O, y_O, r'_M) \in E_O \wedge \exists (x_M, y_M, r'_M) \in E_M$.

A motif can have several occurrences in the network, in which case they are distinguished by their labels. Fig. 3 represents the MIB motifs used to represent every type of biological data included into the database. Motif A illustrates a transcriptional factor that inhibits (or activates) the expression of a protein. Reactions involving two proteins that form a complex were represented by motifs D, and PPIs by motif B. Two more transformations represent indirect and even unknown mechanisms: synexpression data (correlated expression of a couple of proteins) are represented by motif E, and synthetic lethality by motif C. So long-distance and short-distance interactions can be mixed during the analysis as we studied for synexpression and its molecular mechanism (Fig. 5).

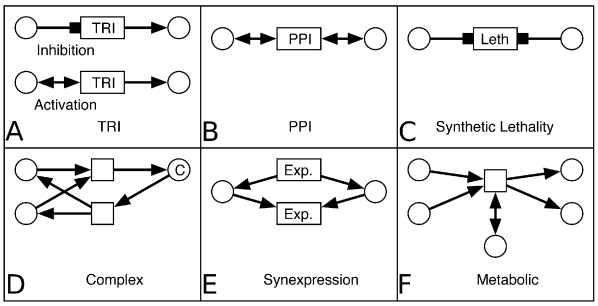



Fig. 3. Motifs used for biological data representation in MIB. **A.** Two motifs representing *TRIs*: inhibition (top) and activation (bottom) of the production of the entity (macromolecule) (right) by another entity (transcription factor) (left). **B.** A motif representing *physical interaction*: two entities activate a transformation (PPI). **C.** The *synthetic lethality* is represented by a motif with two entities inhibiting a transformation 'Leth' (for *lethality phenotype*). **D.** A motif representing *association* transformation (top) that consumes two entities and produces a complex C. The reverse transformation (*dissociation*) is represented in the bottom of the panel. **E.** The *synexpression* of a couple of entities is represented by a motif with two transformations in which they are produced (top) and consumed (bottom) together. **F.** A motif representing a *metabolic reaction*. Two entities are consumed by a transformation, one entity activates it and two entities are produced.

Finally, a metabolic reaction catalysed by an enzyme is illustrated by motif F, where two reactants are consumed, two other molecules are produced, and one enzyme is needed by the transformation.

## 3. Application to the heterogeneous network of *S. cerevisiae*

Modelled data, coming from various sources, were integrated in the *Biological Interaction Browser* (BIB) (http://www.genoscope.cns.fr/biopathways/bib/). We integrated the following datasets: protein–protein interaction (PPI) data, generated using high-throughput variants of the yeast two-hybrid method to identify binary interactions [20,21] or using techniques to isolate multiprotein complexes based on mass-spectrometry such as HMS-PCI [22], TAP [23] and compilation from the literature [24]. The data include also direct transcriptional interactions (TRI) compiled from the literature [25] and from ChIP-Chip experiments [26]. The synexpression results come from microarrays experiments [27] representing pairs of genes with a correlated expression.

Table 1
Number of feedback loops as a function of loop size (column 1): loops including only TRIs (column 2), TRIs and one PPI (column 3), TRIs and two PPIs that are not adjacent (column 3)

| Loop size | TRIs + 0 PPI | TRIs + 1 PPI | TRIs + 2 PPIs |
|---|---|---|---|
| 2 | 5 | 17 | – |
| 3 | 4 | 32 | – |
| 4 | 5 | 71 | 125 |
| 5 | 4 | 144 | 529 |
| 6 | 9 | 222 | 1372 |
| 7 | 6 | 390 | 3140 |
| 8 | 12 | 740 | 8464 |
| 9 | 22 | 1197 | 14 863 |
| 10 | 41 | 1987 | 30 444 |

The synthetic lethality results [27] represent pairs of yeast genes whose joint disruption is lethal. Finally, the metabolic network data were taken from Biocyc [28] using Cyclone [35]. The complete network contains 6513 proteins, 1440 complexes, two phenotypes. The interactions include 7455 cases of DNA–protein interactions, 8531 protein–protein interactions, 16496 synexpressions, 886 synthetic lethality cases. Feedback loops and synexpression patterns were searched in this entire heterogeneous network.

### 3.1. Feedback loops

Feedback loops are a basic example of a static motif, from which dynamical properties such as homeostasis and differentiation can be inferred. The dynamical behaviour of regulatory loops has been studied by several authors using a variety of techniques [16], mostly in the context of transcriptional networks and abstract networks of regulatory influences. Here, we searched for the first time for feedback loops that include both TRI and PPI.

Before studying heterogeneous motifs, TRI-only loops were searched. One hundred and eight TRI-only feedback loops were found in the entire network, with lengths ranging from 2 to 10 (see Table 1, columns 1 and 2).

Then, one TRI at a time was replaced by a PPI. Fig. 4 shows feedback loops, each comprising four entities (circles) and the following sets of transformations (squares): TRI only (A), 3 TRIs + 1PPI (B) and 2TRIs + 2 PPIs (C). For example, the motif (B) illustrates a feedback loop made of four entities, one PPI and three TRIs. All TRIs are oriented in the same direction and can represent either an activation (double arrows) or an inhibition (squared arrows).

We compared the number of TRI-only loops with the number of loops where a TRI had been replaced by a PPI (Table 1, columns 2 and 3). Depending on the loop size, 3–50 times more loops with one PPI were found. If two non-adjacent TRIs are replaced by two PPIs, the number of loops increases up to three orders of magnitude, depending on the loop size (Table 1, columns 2 and 4). Thus, adding a second PPI in a motif that already included one PPI increases the number of matching subnets from 2 to 15 times.

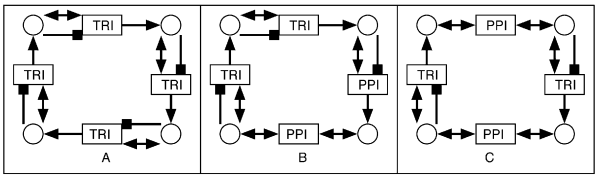



Fig. 4. Feedback-loop motifs made of TRIs only (**A**), with one PPI (**B**) or with 2 PPIs (**C**). Each motif contains four transformations (rectangular shapes), four entities (circles), and possible roles of entities in transformations are represented by arcs.

### 3.2. Micro-arrays

Synexpression may involve various underlying molecular mechanisms, thus being a biological result at an intermediate level between molecular physical mechanisms and phenotypes (see Fig. 1). To evaluate the correlation between the molecular knowledge integrated in the BIB and synexpression data, we searched for possible mechanisms accounting for each synexpressed couple of genes.

We used BIB to find the correlation between the micro-array data on the synexpression of gene pairs, and the biochemical reactions in which these two genes participate. Thus, a molecular mechanism underlying the synexpression of two genes, based on the PPI and TRI graphs, could be proposed. These molecular mechanisms, symbolized by candidate motifs, are presented in Fig. 5, together with the number of observed occurrences of each motif type. To determine which motifs are under- or over-represented, the ratio of motif occurrences with and without synexpression was calculated for six candidate mechanisms (last column in Fig. 5).

We looked for modules comprising one gene that regulates the transcription of another gene (Fig. 5B, left) and where the two genes are synexpressed (Fig. 5B, right). Six occurrences of such a module were found with synexpression, and 7412 occurrences were observed without synexpression, which makes the difference of 1200 times. A more complex motif would include one (Fig. 5C, right) or two (Fig. 5F, right) additional genes between the two initial ones. Such motifs were found 19 and 27 times, respectively, with a ratio of 500 and 1000 times less compared to the same motif without synexpression.

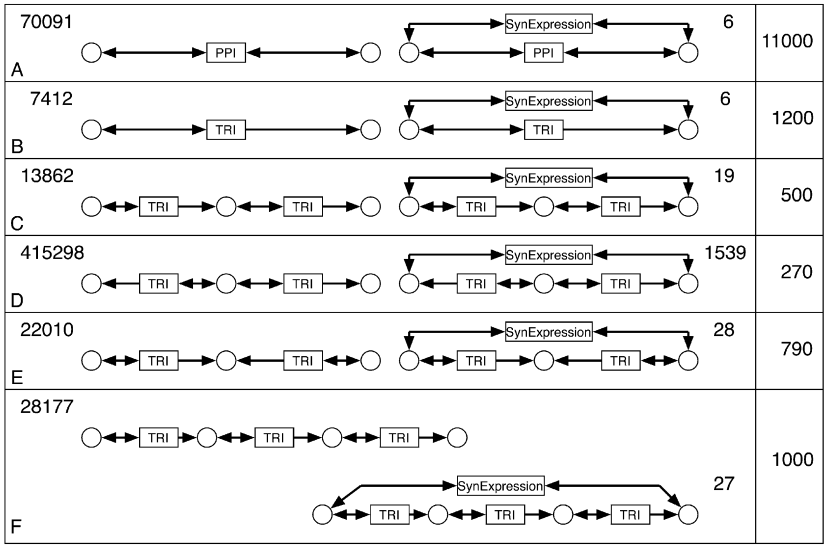



Fig. 5. Correlation between synexpression data and underlying biochemical mechanisms. Six motifs were proposed to be candidates for the synexpression mechanisms (**A**–**F**, left). For each motif, the number of occurrences in the BIB database is indicated on the side. The motifs combining the regulatory mechanism and the synexpression data (**A**–**F**, right) were searched, and the number of encountered occurrences of such subnets is indicated. The last column shows the ratio between occurrences of each motif without or with synexpression condition.

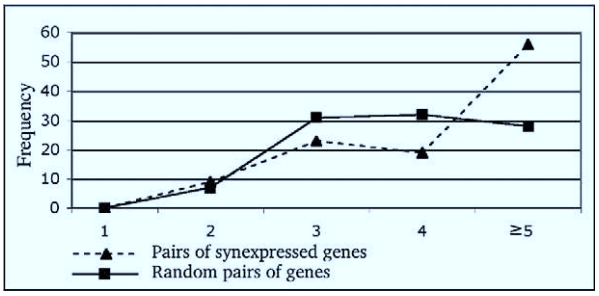



Fig. 6. Shortest path length distribution between all synexpressed pairs of proteins (dashed line) versus all possible pairs of proteins (plain line). The shortest paths of length 1 to 4 have been searched. The value of 5 on the $x$-axis indicates that no shorter path than five has been found.

A different candidate motif that accounts for synexpression of two genes could involve a third gene that regulates these two genes (Fig. 5D, right). This motif is found 1539 times in yeast, 270 times less than without synexpression constraint. It is interesting to see that the inverse situation, when two synexpressed genes regulate a third one (Fig. 5E, right) is much less frequent (28 cases, 790 times less than without synexpression). As for the synexpression motif A, it was strongly underrepresented (6 cases, 11 000 times underrepresented), meaning that synexpressed genes are seldom participating in a PPI.

For further analysis of the link between synexpression phenotype and the physical interaction network structure, we analyzed the shortest path-length distribution between synexpressed genes compared to that of any pair of genes. The results are shown in Fig. 6. There is little difference between the two distributions, except for long paths ($\geqslant 5$ steps). The average path length between two synexpressed genes is significantly different from that between random pairs of genes for long paths only, in contrast with previous results [12].

## 4. Discussion

Most studies involving heterogeneous networks thus far have focused either on network topology, either local or global. However, most important biological processes such as signal transduction, cell-fate regulation, transcription and translation involve more than four but much fewer than hundreds of proteins. MIB is slightly more complex than a simple graph representation, but has greater expressiveness. One of the great advantages of this approach is that this model enables various static and dynamic analysis. It directly represents $n$-ary relations that are essential for the representation of complexes and of metabolic reactions. The added expressiveness is also related to the assumption that each modelled transformation occurring in the biological system may be broken down into elementary parts [29]. Our model is more abstract than the one proposed in [30], so we can deal with different types of biological objects and processes uniformly. MIB enables the semi-automatic translation in other modelling formalisms such as, for example, Petri Nets, Ordinary

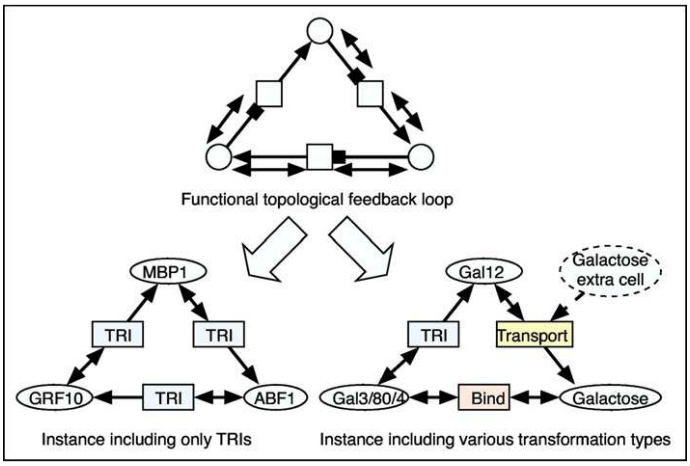



Fig. 7. A MIB motif (with a specified dynamics; top) allows searching for both TRI only subnets (left) and mixed TRI/Metabolism/PPI subnets (right). This is illustrated here with a feedback loop.

Differential Equations, or Pi-calculus (Yartseva et al., in prep.). The BIB tool adapts some of the algorithms available for graphs (e.g., motif search) to the case of bipartite graphs. It can be used to analyse how various data types complement each other in the full heterogeneous network. As most biologically interesting features concern the dynamics of biological functions implemented by molecules, reactions or pathways, biologically meaningful queries are better expressed at the level of functions and the objects that support these functions. A simple graph representation does not allow this type of query formulation. Fig. 7 provides an example of how the MIB formalism allows to search for instances of a function, independently of the precise 'implementation' of this function in a cell. Both subnetworks at the bottom of Fig. 7 can fulfil the specified dynamics depicted by the motif at the top. The subnetwork on the left is implemented by TRIs only, and the one on the right by one TRI, one metabolic reaction (transport) and one physical interaction (binding).

TRI only feedback loops have already been studied [25]. In the present study, we searched for such loops in larger datasets, and therefore we found more loops in the larger size range. We also provide a new perspective on these feedback loops studies by relaxing previous constraints [31] to allow PPI anywhere in the loops. Some of the modules found are well known, such as the Ste12–Fus3 feedback circuit [32], others are unknown.

The analysis of synexpression data relations between 1625 pairs of genes allowed us to propose for each pair a biologically relevant circuit with a parsimonious topology. This result illustrates how an interaction of higher-level order than biochemical reactions may be modelled in MIB, thus enabling the study of the whole set of yeast interactions.

We have found that the paths between synexpressed genes were longer than for random pairs of proteins (see Fig. 6). We will further investigate synexpressed gene paths. However, the situation is opposite for transcriptional factors: the paths between pairs of them are shorter than between random pairs of proteins [33]. This difference could mean that the genes that are not close in the biological interaction network need to be synexpressed in order to synchronize their biological activity. Our explanation is in line with the results on just-in-time assembly regulation of various complexes [34].

All the interactions integrated in the model come from experimental results, but the context in which a given interaction effectively takes place is not known and may vary among experiments. Therefore, the validation step consists in finding the conditions in which the modules are functional, either by calling on an expert, or if prior knowledge is unavailable, by bench experimentation, as has been done in the case of the galactose feedback loop [18].

These preliminary studies represent a proof of concept for the MIB as a useful tool for future investigations involving regulation, protein interactions, and metabolic networks together with higher-level types of interactions, like synthetic lethality or synexpression.

## Acknowledgements

We are grateful to P. Bourgine for discussions. This work was financially supported by CO3 European Project, ISI Foundation, CNRS, Genopole, Genoscope, and S. Smidtas.

## References


[1] A. Goffeau, B. Barrell, H. Bussey, R. Davis, B. Dujon, H. Feldmann, F. Galibert, J. Hoheisel, C. Jacq, M. Johnston, E. Louis, H. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, S. Oliver, Life with 6000 genes, Science 274 (5287) (1996) 546, 563–567.

[2] D. Watts, S. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440–442.

[3] A. Wagner, The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, Mol. Biol. Evol. 18 (7) (2001) 1283–1292.

[4] G. Bader, C. Hogue, An automated method for finding molecular complexes in large protein interaction networks, BMC Bioinformatics 4 (2) (2003).

[5] A. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.

[6] B. Schwikowski, P. Uetz, S. Fields, A network of protein–protein interactions in yeast, Nat. Biotechnol. 18 (12) (2000) 1257–1261.

[7] H. Jeong, S. Mason, A. Barabási, Z. Oltvai, Lethality and centrality in protein networks, Nature 411 (6833) (2001) 41–42.

[8] B. Snel, P. Bork, M. Huynen, The identification of functional modules from the genomic association of genes, Proc. Natl Acad. Sci. USA 99 (9) (2002) 5890–5895.

[9] A. Vazquez, A. Flammini, A. Maritan, A. Vespignani, Global protein function prediction from protein–protein interaction networks, Nat. Biotechnol. 21 (6) (2003) 697–700.

[10] M. Herrgård, B. Palsson, Untangling the web of functional and physical interactions in yeast, J. Biol. 4 (5) (2005).

[11] L. Zhang, O. King, S. Wong, D.S. Goldberg, A. Tong, G. Lesage, B. Andrews, H. Bussey, C. Boone, F. Roth, Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network, J. Biol. 4 (6) (2005), doi:10.1186/jbiol23.

[12] R. Balasubramanian, T. LaFramboise, D. Scholtens, R. Gentleman, A graph-theoretic approach to testing associations between disparate sources of functional genomics data, Bioinformatics 20 (18) (2004) 3353–3362.

[13] S. Troncale, D. Campard, J. Guespin, J.-P. Vannier, F. Tahi, Modelisation of interleukin-6 system in early hematopoiesis with hybrid functional petri nets, in: Genopole (Ed.), Modélisation de systèmes biologiques complexes dans le contexte de la génomique, Montpellier, 4–8 avril 2005.

[14] A. Doi, S. Fujita, H. Matsuno, M. Nagasaki, S. Miyano, Constructing biological pathway models with hybrid functional Petri nets, In Silico Biol. 4 (0023) (2004).

[15] H. Matsuno, A. Doi, M. Nagasaki, S. Miyano, Hybrid petri net representation of gene regulatory network, in: Pac. Symp. Biocomput. 2000, pp. 341–352.

[16] H. de Jong, Modeling and simulation of genetic regulatory systems: a literature review, J. Comput. Biol. 9 (1) (2002) 67–103.

[17] D. Ross, A. Schoman, Structured analysis for requirements definition, in: Requirements analysis, IEEE Trans. Softw. Eng. 3 (1) (1977) 6–15 (special issue).

[18] Y. Louzoun, S. Solomon, H. Atlan, I. Cohen, The emergence of spatial complexity in the immune system, Physica A 297 (1–2) (2001) 242–252.

[19] S. Smidtas, V. Schächter, F. Képès, The adaptive filter of the yeast galactose pathway, J. Theor. Biol. (in press), doi:10.1016/j.jtbi.2006.03.005.

[20] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, Y. Sakaki, A comprehensive two-hybrid analysis to explore the yeast protein interactome, Proc. Natl Acad. Sci. USA 98 (8) (2001) 1569–1574.

[21] P. Uetz, L. Giot, G. Cagney, T. Mansfield, R. Judson, J. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, J. Rothberg, A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae, Nature 403 (6770) (2000) 623–627.

[22] Y. Ho, A. Gruhler, A. Heilbut, G. Bader, L. Moore, S. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. Willems, H. Sassi, P. Nielsen, K. Rasmussen, J. Andersen, L. Johansen, L. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. Sørensen, J. Matthiesen, R. Hendrickson, F. Gleeson, T. Pawson, M. Moran, D. Durocher, M. Mann, C. Hogue, D. Figeys, M. Tyers, Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry, Nature 415 (6868) (2002) 180–183.

[23] A. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. Rick, A. Michon, C. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. Heurtier, R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, G. Superti-Furga, Functional organization of the yeast proteome by systematic analysis of protein complexes, Nature 415 (6868) (2002) 141–147.

[24] http://mips.gsf.de/proj/yeast/catalogues/complexes/.

[25] N. Guelzim, S. Bottani, P. Bourgine, F. Képès, Topological and causal structure of the yeast transcriptional regulatory network, Nat. Genet. 31 (1) (2002) 60–63.

[26] T. Lee, N. Rinaldi, F. Robert, D. Odom, Z. Bar-Joseph, G. Gerber, N. Hannett, C. Harbison, C. Thompson, I. Simon, J. Zeitlinger, E. Jennings, H. Murray, D. Gordon, B. Ren, J. Wyrick, J. Tagne, T. Volkert, E. Fraenkel, D. Gifford, R. Young, Transcriptional regulatory networks in *Saccharomyces cerevisiae*, Science 298 (5594) (2002) 799–804.

[27] C. von Mering, R. Krause, B. Snel, M. Cornell, S. Oliver, S. Fields, P. Bork, Comparative assessment of large-scale data sets of protein–protein interactions, Nature 417 (6887) (2002) 399–403.

[28] http://www.biocyc.com.

[29] P. Maziere, C. Granier, F. Molina, A description scheme of biological processes based on elementary bricks of action, J. Mol. Biol. 339 (1) (2004) 77–88.

[30] J. van Helden, A. Nairn, C. Lemer, R. Mancuso, M. Eldridge, S. Wodak, From molecular activities and processes to biological function, Briefings in Bioinformatics 2 (1) (2001) 81–93.

[31] S. Shen-Orr, R. Milo, S. Mangan, U. Alon, Network motifs in the transcriptional regulation network of *Escherichia coli*, Nat. Genet. 31 (1) (2002) 64–68.

[32] L. Bardwell, J. Cook, J. Zhu-Shimoni, D. Voora, J. Thorner, Differential regulation of transcription: repression by unactivated mitogen-activated protein kinase kss1 requires the dig1 and dig2 proteins, Proc. Natl Acad. Sci. USA 95 (26) (1998) 15400.

[33] T. Manke, R. Bringas, M. Vingron, Correlating protein–DNA and protein–protein interactions, J. Mol. Biol. 333 (1) (2003) 75–85.

[34] U. de Lichtenberg, L. Jensen, S. Brunak, P. Bork, Dynamic complex formation during the yeast cell cycle, Science 307 (5710) (2005) 724–727.

[35] http://nemo-cyclone.sourceforge.net.

## 4.2 Explorateur des réseaux biologiques BIB

Pour analyser des modules hétérogènes à des niveaux intermédiaires (mésoscopiques), le Biological Interaction Browser (BIB), outil d'analyse de module, a été développé. Il est basé sur un modèle mathématique original qui permet une représentation de la dynamique qualitative des interactions hétérogènes biologiques. Cet outil s'appuie sur une définition formelle de motifs spécifique au modèle sous-jacent, basé sur un graphe biparti avec des noeuds et des liens typés. Ces travaux ont été publiés sous le titre Model of Interactions in Biology and Application to Heterogeneous Network in Yeast [161]. Le modèle présenté a l'originalité d'être suffisamment simple pour se prêter à des analyses statistiques globales portant sur des comportements dynamiques qualitatifs locaux. S'appuyant sur cette modélisation, nous avons proposé de nombreux motifs topologiques dignes d'intérêt.

Le Biological Interaction Browser permet de naviguer dans l'ensemble du réseau d'interaction. Il a fallu développer un outil ad-hoc car le degré de connectivité des noeuds du réseau est bien trop élevé pour qu'une représentation graphique du réseau puisse visuellement conduire à autre chose qu'une surface pleine : il n'est pas possible de représenter en une fois l'intégralité du réseau comme le montre la Figure 4.1. Le browser doit donc être un browser local, qui permet de représenter un environement local du réseau ; il est possible de visualiser l'environnement d'une protéine, soit en partant d'un noeud, soit en partant d'un motif, comme nous allons le voir. Enfin, ce programme sert d'Interface Home Machine pour que l'utilisateur puisse lancé de manière graphique ses propres études et utiliser les algorithmes développés.

Cet outil permet, à partir de données expérimentales et de connaissances biologiques, d'étudier la structure des données,de rechercher des motifs d'intérêt, puis de convertir ces motifs en modules dynamiques, amenant à proposer de nouvelles connaissances biologiques décrivant leur dynamique.

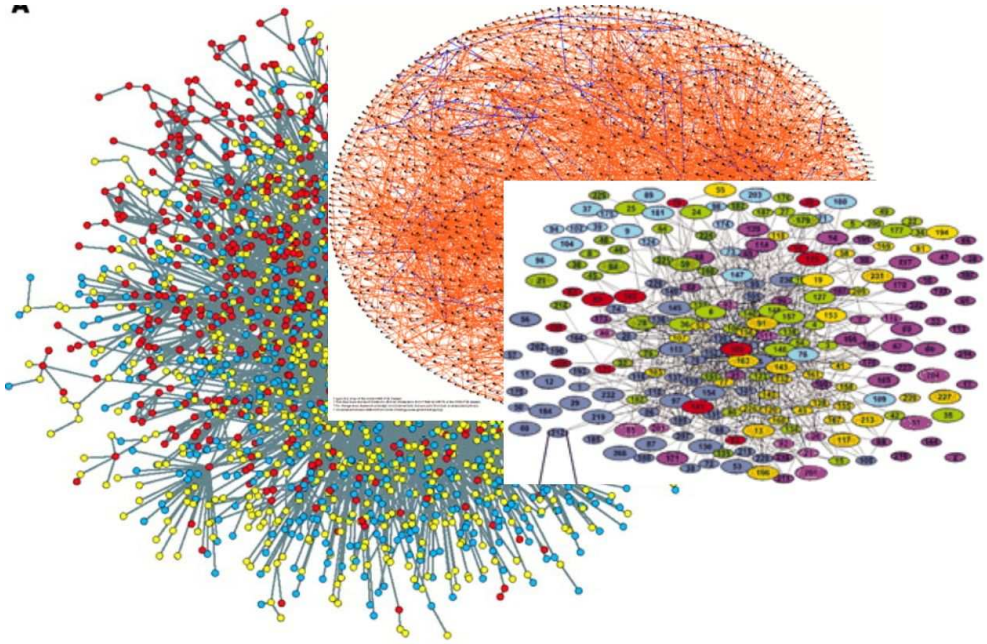


FIG. 4.1 – Représentations spatiales antérieures de données homogènes. Ici sont représentés des réseaux d'interactions tel qu'ils ont illustré des articles lors de leur publication. Ces représentations partagent toutes le fait que la quasi totalité de l'espace est trop dense pour être déchiffrable. Il paraït donc vain de vouloir intégrer toutes ces données et de vouloir les représenter spatialement. Un système de navigation local et de requêtage original doit être mis en place. Au fond, nous avons représenté une interaction protéine-protéine dans la drosophile [20] , les noeuds représentent des protéines. Au dessus, nous avons des interactions de régulations de gènes [119]. Devant, relations entre complexes qui partagent des protéines [55] .

Le Biological Interaction Browser permet d'intégrer des données hétérogènes en un réseau. Des modules topologiques définis y seront recherchés puis traduits en système dynamique sous forme d'un système d'équations pouvant être analysé ou simulé permettant (ou non) de caractériser le module comme fonctionnel (ou non).

Pour illustrer l'utilité biologique de notre approche de modélisation, le mécanisme d'un exemple concret d'un des motifs identifié lors des analyses topologiques a été étudié plus en détail. Il s'agit du module de boucle de régulation de la voie de dégradation du galactose dans la levure. L'analyse de sa dynamique montre que le module permet une grande stabilité et une adaptation de la levure à ce sucre. Ce travail a été publié dans [159].

### 4.2.1 Implémentation de l'Explorateur de graphes BIB

L'implémentation est réalisée en PHP mySQL. Ce programme a été rendu public à l'occasion d'ISMB-ECCB en juillet 2004 à Glasgow. BIB s'appuie sur la base de données qui lui est propre [161].

Une première façon de retrouver l'information de la base est de rechercher un noeud du réseau. Une recherche systématique par nom ou synonyme le permet de réaliser cela. La liste des interactions autour de ce noeud est représentée sous forme graphique et textuelle, comme le montre la Figure 4.2.

#### Algorithme de recherche de modules topologiques

Pour rechercher des groupes de gènes formant des modules topologiques hétérogènes dans BIB, un algorithme spécifique a été développé. En effet, des travaux préalables [134] portant uniquement sur le réseau de régulation d'E.coli, et ne représentant qu'une faible fraction du réseau hétérogène reconstruit ici, ont seulement permis à leurs auteurs de rechercher des modules de taille inférieure à 4 interactions. Notre algorithme procède comme suit :

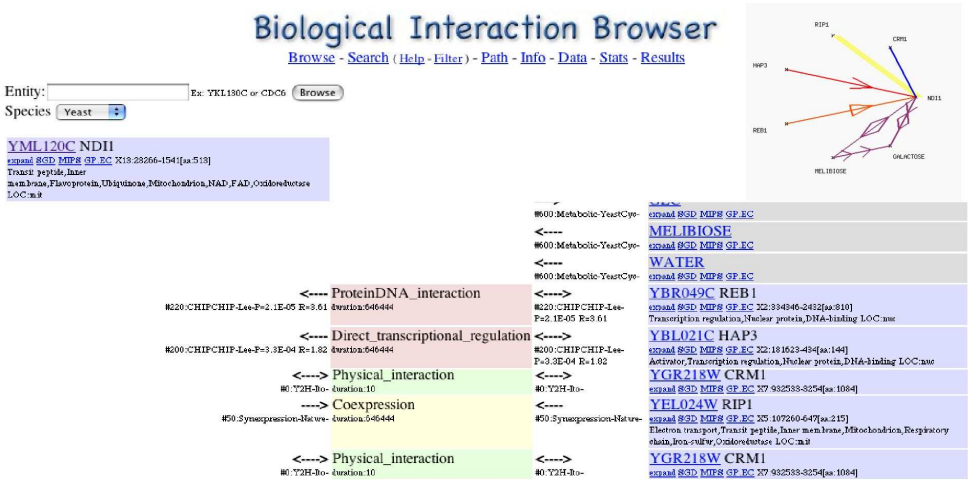



FIG. 4.2 – Capture d'écran du browser BIB. Cette page permet de naviguer dans l'ensemble des données à partir d'un nom de gène ou d'un synonyme. La liste des interactions auxquelles participe ce gène est affichée, accompagnée d'une représentation graphique locale de ces interactions. Il est alors possible, de proche en proche, de se déplacer dans le réseau.

L'algorithme de recherche des motifs hétérogènes dans le modèle MIB a été écrit pour tirer profit de l'optimisation et des performances de jointures de tables de MySQL.

La recherche d'un motif commence par trouver séparément toute les arêtes susceptibles de composer ce motif joignant un couple de protéine via une transformation, ou uniquement une arête entre une transformation et une protéine. Ainsi, s'il est composé de $n$ arrêtes, nous allons nous retrouver avec $n$ tables contenant des arêtes candidates. Ensuite, selon le schéma de motif, les tables sont rassemblées deux par deux par une jointure sur les nœuds qui doivent être les mêmes. Par exemple, si le motif qu'on recherche contient des arêtes $(P1, P2, t_1), (P1, P3, t_2), (P1, P4, t_3)$ et $(P3, P4, t_4)$ où $Pi$ sont des variables gènes ou protéines, et $t_i$ sont des types des arêtes en question, nous allons faire une jointure des tables $(P1, P2, t_1)$ et $(P1, P3, t_2)$, ainsi que la jointure des tables $(P1, P4, t_3)$ et $(P3, P4, t_4)$. Nous nous retrouvons à cet étape avec les tables qui contiennent des lignes de la longueur deux fois supérieure à des tables d'origine, sauf peut être une si le nombre des arêtes dans le motif est impaire. La procédure de la jointure des tables qui partagent les mêmes nœuds

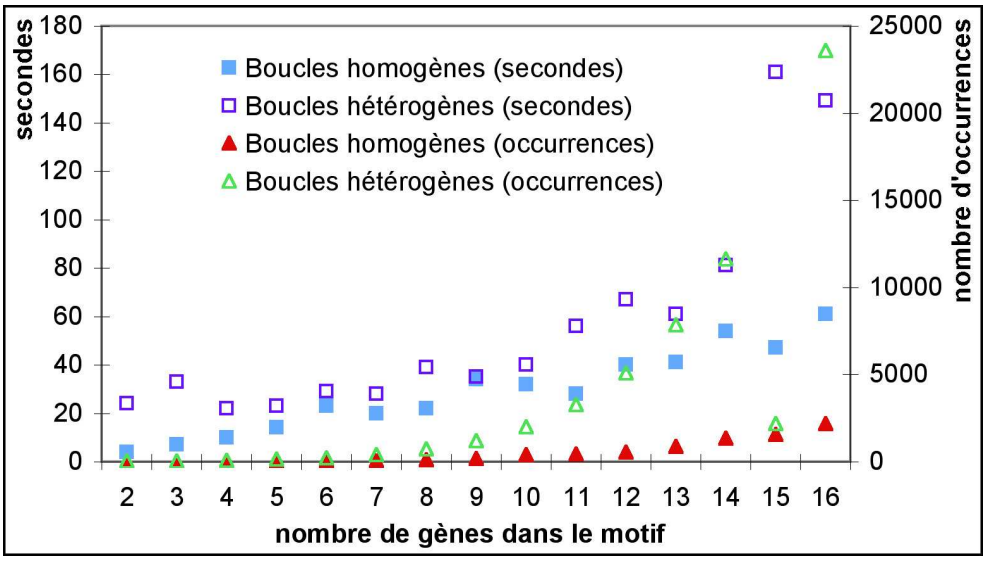



FIG. 4.3 – Analyse de performance de l'explorateur de BIB. Icones pleines : temps et résultats de recherche d'occurrences de boucles de régulation transcriptionnelle homogènes (incluant uniquement des TRI), en fonction de la taille de ces boucles. Icones vides : temps et résultats de recherche d'occurrences de boucles hétérogènes (incluant une interaction protéique PPI), en fonction de la taille de ces boucles.

peut être poursuivie jusqu'à ce qu'il ne reste qu'une table avec le résultat de recherche du motif d'origine. Pour notre exemple, nous obtenons une table qui contient quatre arêtes demandées déjà à la deuxième itération de l'algorithme.

Cet algorithme ainsi implémenté est extrêmement performant, en terme de temps de calcul, comparé aux études précédentes [184] : 99,8% des requêtes saisies par les utilisateurs à ce jour s'exécutent en moins d'une heure et leur très grande majorité s'exécute en moins de 10 minutes sur un ordinateur portable. De plus, le temps de recherche n'est pas une fonction croissante de la taille du module recherché (voir Figure 4.3) : les gros modules ne sont pas toujours recherchés en plus de temps que les petits. Enfin cet algorithme a l'avantage d'être parallélisable facilement, en transformant en un nouveau processus indépendant tout passage par l'étape (1).

### 4.2.2 Interface de BIB

L'interface de saisie des requêtes de recherche de module (ici motif de réseau) est simple et intuitif. Une module est défini par un ensemble de réactions. Le formulaire de saisie reprend le découpage en réactions. Chaque ligne du formulaire apparaissant sur la Figure 4.4 permet de rechercher une interaction entre entités. Un ensemble de lignes compose la requête, et donc par conséquent un motif à rechercher. Un nom générique (tel que P1, P2 etc) peut être attribué aux noeuds, ce qui permet de faire référence à un même noeud dans plusieurs interactions. Le type de l'entité (Protéine, Complex, Phénotype) permet de caractériser l'entité. Le type de l'interaction dans laquelle l'entité participe peut être spécifié. Enfin, une ou plusieurs autres entités qui participent à la même transformation sont saisies. Une ligne du formulaire correspond donc à saisir par exemple la recherche d'une protéine P1 qui active ou inhibe l'expression transcriptionnelle d'une autre protéine P2. Les modules recherchés s'écrivent sur plusieurs lignes de la sorte. Exemple d'un boucle de rétroaction : une protéine P1 active ou inhibe l'expression transcriptionnelle d'une autre protéine P2 et la protéine P2 active ou inhibe l'expression transcriptionnelle de la protéine P1. Un autre exemple est la formation d'un complexe à 3 protéines : une protéine P1 est consommée dans la formation d'un complexe C1 et une protéine P2 est consommée dans la formation du complexe C1 et une protéine P3 est consommée dans la formation du complexe C1. Ce système très souple permet donc de rechercher des modules dont les transformations font participer plus 2 protéines comme dans l'exemple précédent.

## 4.3 Résultats d'une recherche de module avec BIB

Les résultats d'une recherche de module sont listés graphiquement et d'une manière textuelle. Chaque module trouvé peut être détaillé comme illustré à la Figure 4.5.

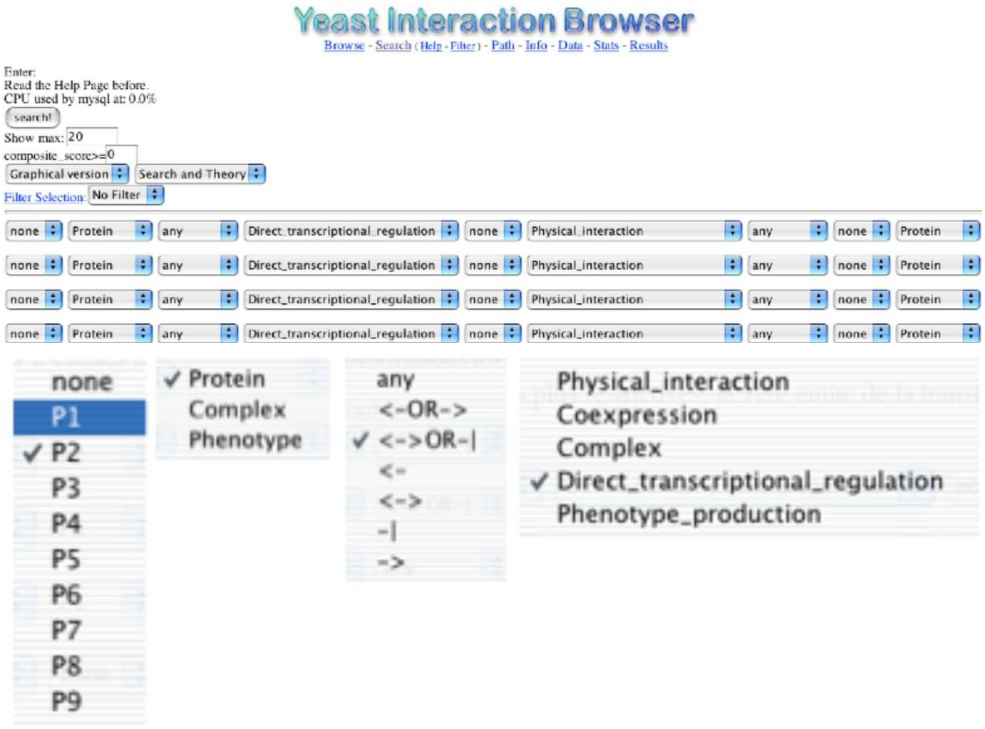
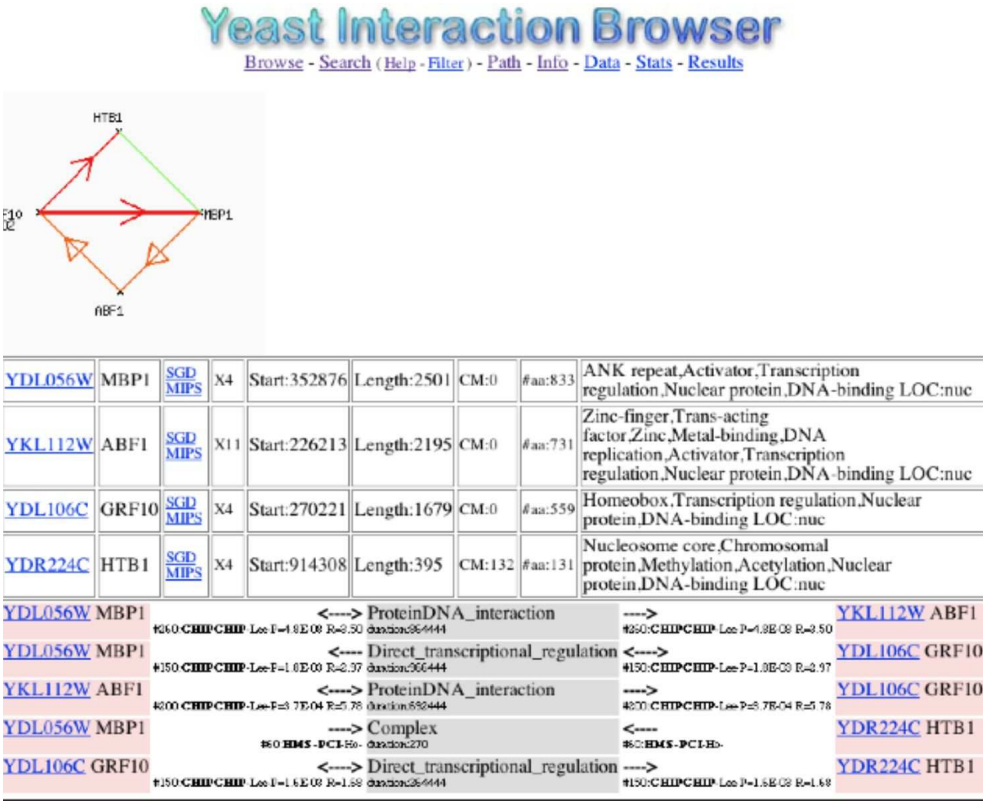



FIG. 4.4 – Interface de requête. Cette page permet de construire facilement des requêtes. Chaque ligne permet de rechercher une interaction entre entités. Un ensemble de lignes compose la requête et par conséquent, un motif à rechercher. Un nom générique (tel que P1, P2) peut être donné aux noeuds, ce qui permet de faire aisément référence à un même noeud dans plusieurs interactions.




| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| YDL056W | MBP1 | SGD MIPS | X4 | Start:352876 | Length:2501 | CM:0 | #aa:833 | ANK repeat,Activator,Transcription regulation,Nuclear protein,DNA-binding LOC:nuc |
| YKL112W | ABF1 | SGD MIPS | X11 | Start:226213 | Length:2195 | CM:0 | #aa:731 | Zinc-finger,Trans-acting factor,Zinc,Metal-binding,DNA replication,Activator,Transcription regulation,Nuclear protein,DNA-binding LOC:nuc |
| YDL106C | GRF10 | SGD MIPS | X4 | Start:270221 | Length:1679 | CM:0 | #aa:559 | Homeobox,Transcription regulation,Nuclear protein,DNA-binding LOC:nuc |
| YDR224C | HTB1 | SGD MIPS | X4 | Start:914308 | Length:395 | CM:132 | #aa:131 | Nucleosome core,Chromosomal protein,Methylation,Acetylation,Nuclear protein,DNA-binding LOC:nuc |

| | | | | |
|---|---|---|---|---|
| YDL056W MBP1 | <----> ProteinDNA_interaction | | ----> | YKL112W ABF1 |
| YDL056W MBP1 | <---- Direct_transcriptional_regulation | | <----> | YDL106C GRF10 |
| YKL112W ABF1 | <----> ProteinDNA_interaction | | ----> | YDL106C GRF10 |
| YDL056W MBP1 | ----> Complex | | <---- | YDR224C HTB1 |
| YDL106C GRF10 | <----> Direct_transcriptional_regulation | | ----> | YDR224C HTB1 |

FIG. 4.5 – Détail d'un module. Cette capture d'écran donne les informations disponibles sur un module c'est-à-dire sur ses entités et sur ses relations : Type d'interaction, vraisemblance, durée de réaction, nom des entités, synonymes, position des gènes, commentaires etc.

### 4.3.1 Recherche de modules dynamiques

Les modules recherchés et trouvés sont statiques. Aucune information sur leur dynamique n'a été précisée jusqu'ici. Un module, une fois trouvé, peut être partagé en SBML comme l'illustre le texte ci-dessous (comme indiqué plus haut, SBML est un format d'échange XML standardisé qui peut être ensuite importé dans un simulateur biologique).

```
<xml version="1.0" encoding="UTF-8">
<sbml xmlns="http://www.sbml.org/sbml/level2" level="2" version="1">
<model id="Module" name="Module">
<notes>
<body />
</notes>
<listOfFunctionDefinitions />
<listOfUnitDefinitions />
<listOfCompartments>
<compartment id="Cell" name="Cell" size="1" volume="1" />
</listOfCompartments>
<listOfSpecies>
<species id="UME6" name="UME6" compartment="Cell" initialAmount="1" boundaryCondition="true" />
<species id="HSF1" name="HSF1" compartment="Cell" initialAmount="1" boundaryCondition="true" />

</listOfSpecies>

<listOfParameters>
<parameter id="delta_t" value="1" />
</listOfParameters>

<listOfReactions>
<reaction id="TRI_2_1_0" name="TRI_2_1_0" reversible="false">
<listOfReactants>
   <species name="UME6" compartment="Cell" initialAmount="1" boundaryCondition="true" />
   </listOfReactants>
<listOfProducts>
   <species name="HSF1" compartment="Cell" initialAmount="1" boundaryCondition="true" />
   </listOfProducts>
<kineticLaw formula="k1*HSF1/(k2+HSF1)">
<listOfParameters>
   <parameter name="k1" value=".1" />
   <parameter name="k2" value="1" />
   </listOfParameters>
   </kineticLaw>
</reaction>

</listOfReactions>
</model>
</sbml>
</xml>
```

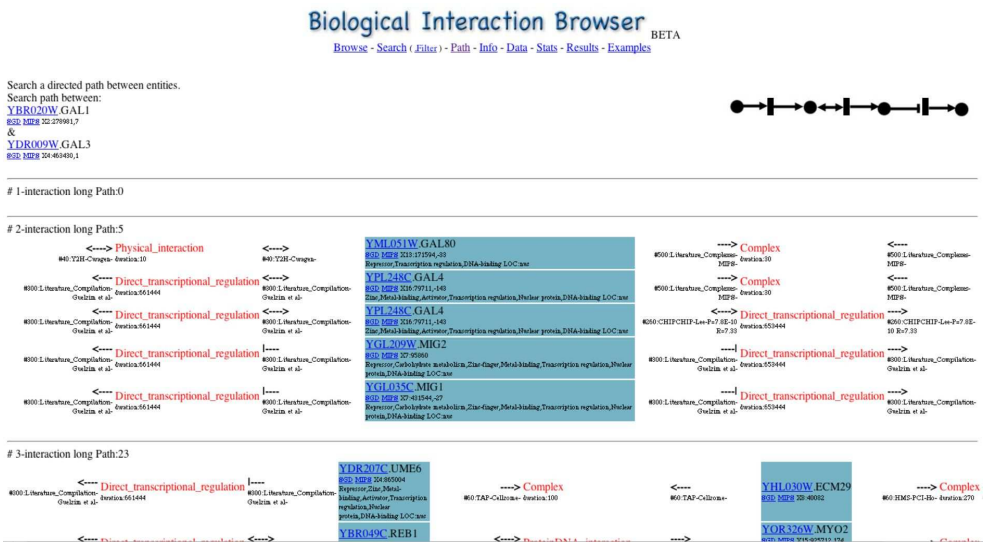



FIG. 4.6 – Recherche de chemin. Les n plus courts chemins entre 2 protéines sont recherchés et affichés au moyen de cet interface.

### 4.3.2 Autres éléments d'interface

La recherche de chemin entre protéines constitue la recherche d'un motif particulier et possède sa propre interface (Figure 4.6). En effet, l'algorithme dans ce cas particulier est étendu afin de ne pas avoir à spécifer une longueur de chemin à rechercher comme indiqué dans les perspectives de l'algorithme de recherche exposé plus haut.

Le browser est accompagné d'un plug-in d'affichage qui permet d'afficher tout module, comme la Figure 4.7 en donne quelques exemples.

### 4.3.3 Perspectives

Comme perspectives et amélioration, il serait intéressant de pouvoir chercher des motifs sans avoir à en préciser l'échelle. Il serait intéressant de travailler aussi bien avec de petits motifs élémentaires que des motifs plus grands constitués de 'boïtes noires'. C'est à dire spécifier qu'entre 2 protéines, on souhaite non pas une interaction d'un certain type, mais une chaîne de réactions d'un certain type par exemple.

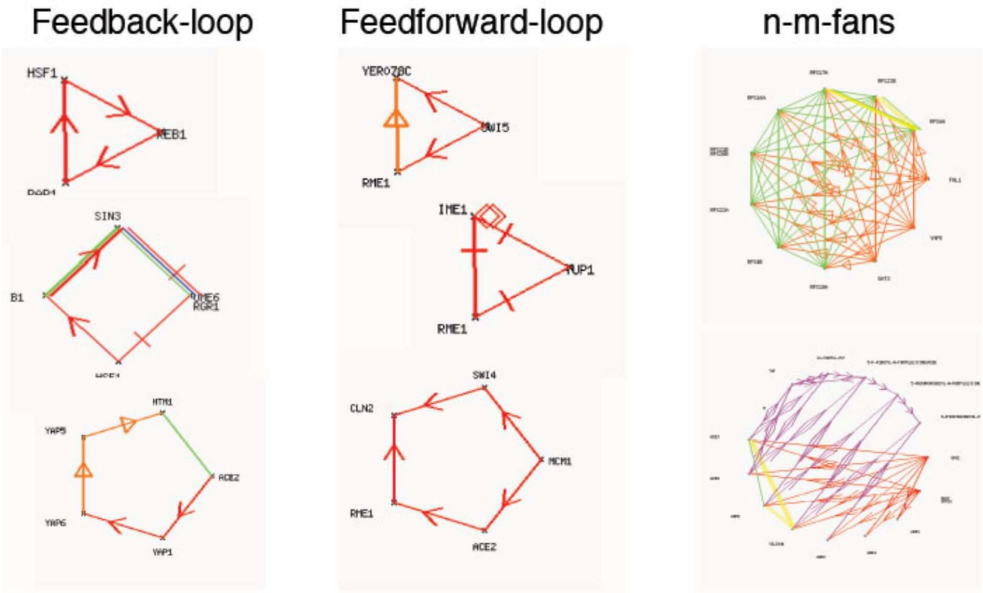



FIG. 4.7 – Exemple d'affichage de modules trouvés. Chaque entité est identifié par son nom court et positionnée sur un cercle, puis toutes les interactions reliant des noeuds du module sont tracées avec un code de couleur : rouge : pour une régulation transcriptionnelle dirigée, orange pour une non dirigée, vert pour une interaction double hybride, vert foncé pour un complexe, jaune pour une synexpression, mauve pour le métabolisme. Les autorégulations sont également représentées.

Il est facile d'imaginer à partir de là, de filtrer les instances de module trouvés suivant leur comportement dynamique après analyse de leur représentation en SBML. La voie serait ainsi ouverte à la rechercher de boucles de rétroaction homéostatique, ou bien la recherche de filtre passe-haut, par exemple. BIB doit être adossé à une structure plus importante, comme cytoscape, afin de profiter par exemple de sa grande implantation et de ses algorithmes de dessin de graphe notamment.

### 4.3.4 Résultats obtenus sur la corrélation des réseaux métaboliques et des réseaux de régulation génétique

Le problème que nous allons considérer ici est de décortiquer la relation de régulation entre des enzymes impliquées dans des voies métaboliques dans des configurations topologiques différentes. Pour cela, nous allons utiliser le modèle MIB pour représenter les données et l'outil BIB pour rechercher ces configurations. Afin de trouver la corrélation entre les voies métaboliques et le réseau de régulation transcriptionnel, nous avons recherché onze motifs topologiques à l'aide de BIB. Chacun de ces motifs est composé de deux briques de bases parmi celles représentées sur la Figure 4.8 ; la première brique est choisie parmi : 'A' (réactions alternatives), 'B' (réactions convergentes), 'C' (réactions consécutives) ou 'D' (réactions indépendantes) et la seconde parmi 'a' (enzymes régulés indépendamment), 'b' (corégulés par le même facteur de transcription F) ou 'c' (corrégulés par un couple de facteurs de transcription F et F'). Une combinaison de 'B' et 'b' est présentée sur la Figure 4.9(A) et une occurrence trouvée pour le motif composé de 'C' et 'c' est présentée sur la Figure 4.9(B).

De nombreuses instances de motifs de chaque type ont été retrouvées (voir Tableau 4.1).

Pour analyser la sur- ou sous-représentation de chacun des motifs, le rapport entre le

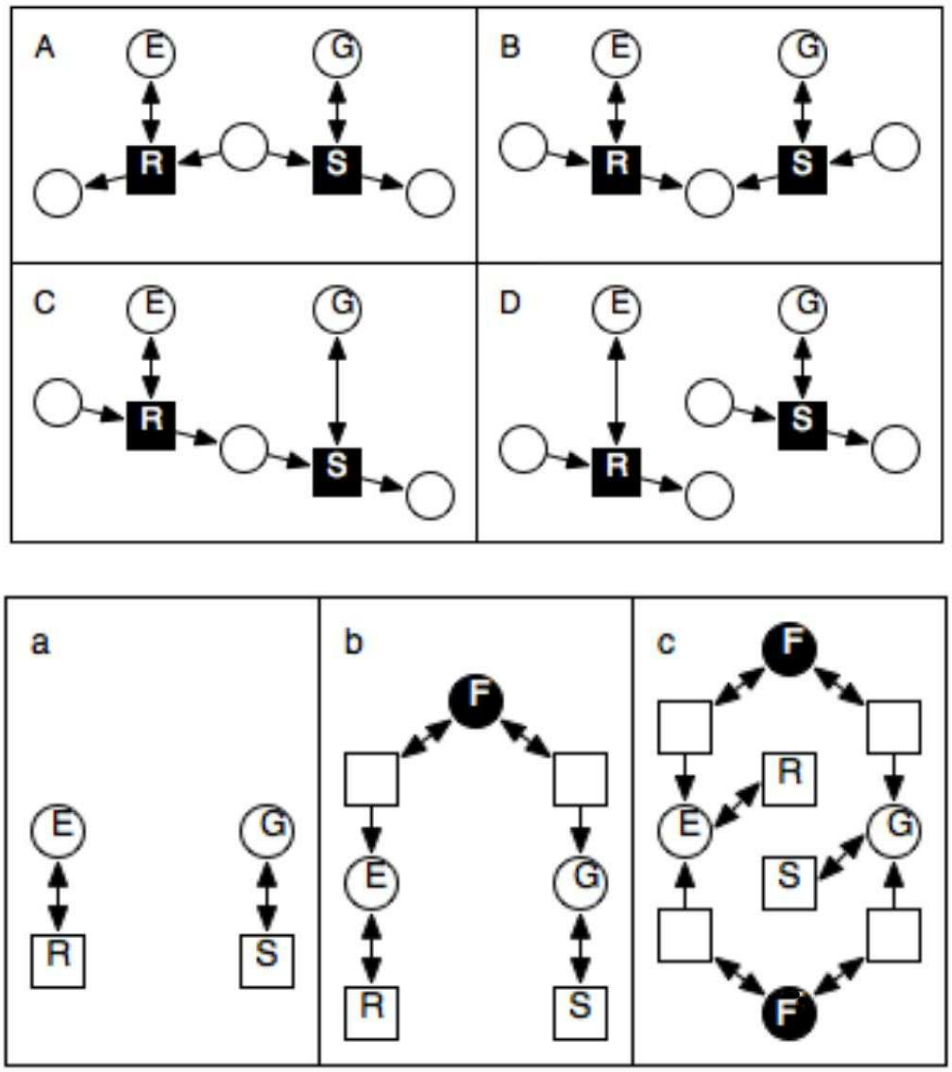
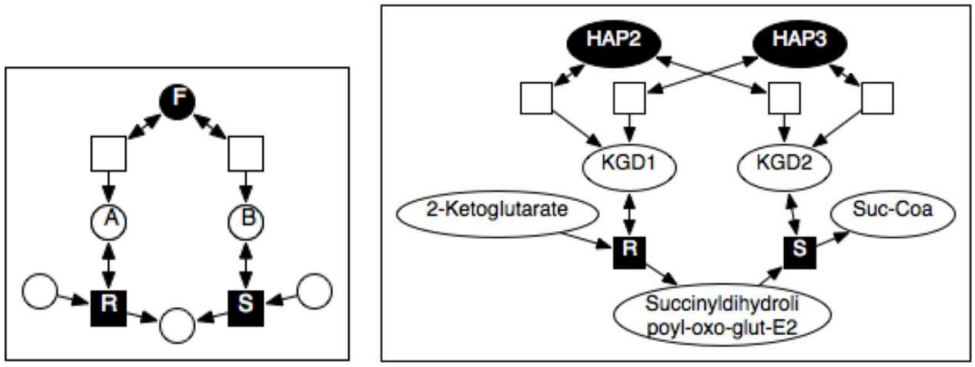



FIG. 4.8 – Briques de régulation métabolique des motifs recherchés. Chaque motif est composé d'une combinaison des quatre motifs A,B,C ou D et d'un des trois motifs a, b ou c.




(A) (B)

FIG. 4.9 – (A). Exemple d'un motif composé des briques 'B' et 'b', recherché dans le réseau métabolique et de régulation. (B) Exemple d'instance du motif composé des briques 'C' et 'c'.

| | A | B | C | D |
|---|---|---|---|---|
| a | 1.283 | 4.239 | 2.668 | . |
| b | 107 | 218 | 247 | 13.547 |
| c | 29 | 48 | 78 | 2.429 |
| | | | | |
| a/b | 12 | **20** | 11 | . |
| b/c | 3 | 4 | 3 | 6 |

TAB. 4.1 – Résultats d'occurrences d'instances de motifs de chaque combinaison de briques A, B, C et D avec a, b et c dans la partie supérieure du tableau. Les deux dernières lignes indiquent les rapports entre les 3 lignes précédentes.

nombre d'occurrences de chaque type de motif comportant les briques 'a', 'b', ou 'c', a été calculé (Tableau 4.1). Lorsque les deux réactions sont convergentes (B), il est attendu que les deux enzymes sont environ deux fois plus souvent corrégulées (B 20) que s'il s'agit d'enzymes consécutives dans une voie métabolique (C 11) ou concurrente (A 12).

De plus, et de manière inattendue, tout couple d'enzymes (brique 'D') qui sont corrégulées par au moins un facteur de transcription, a deux fois plus de chance (6) d'être corrégulées par un second facteur de transcription que deux enzymes à une distance 1 dans le graphe métabolique (A 3, B 4, ou C 3).

## 4.4 Conclusion du chapitre

Dans ce chapitre nous avons présenté un modèle de réseaux biologiques qui permet d'intégrer les données biologiques hétérogènes et d'étudier ensuite les motifs hétérogènes qui peuvent avoir un sens biologique. Par exemple, un mécanisme moléculaire peut être proposé pour expliquer un phénomène macroscopique, tel que la létalité d'une paire de gènes ou encore l'expression simultanée des deux gènes. Le modèle MIB présenté ici bénéficie de quelques innovations par rapport à la modélisation par un graphe simple. La première consiste à utiliser le graphe biparti afin de séparer les acteurs moléculaires ou les observables plus abstraits des processus biologiques dans lesquelles ils participent, représentés chacun par un type différent de nœuds. La deuxième innovation consiste à mélanger dans le même modèle les données microscopiques, telles que les interactions entre les molécules, et les données phénotypiques, donc plus macroscopiques, sur l'expression des gènes ou sur l'effet de leur absence dans la levure. La troisième innovation consiste à charger les liens entre les nœuds du graphe par de l'information pour permettre l'expression des rôles que les entités représentées par les nœuds correspondant jouent dans le processus relié par ces liens. Ainsi, il est possible de coder dans un graphe statique les informations sur la dynamique

du système et distinguer entre les substrats, les enzymes et les produits d'une réaction biochimique donnée.

Encouragés par le succès des rôles des liens, nous sommes allés plus loin en enrichissant MIB avec des informations supplémentaires associées aux nœuds et aux liens. Le formalisme obtenus, appelé MIN pour *Modular Interaction Network*, sera présenté dans le chapitre suivant. MIN possède, en plus du graphe d'interactions, une base de données sur les états du système qui ont été observés dans le système biologique modélisé. Ainsi, il est possible d'élaborer les algorithmes de traduction des informations biologiques, essentiellement structurelles, contenues dans MIN, vers d'autres formalismes qui sont traditionnellement utilisés pour l'étude de la dynamique des systèmes biologiques.

# Chapitre 5

# MIN : Modèle de connaissances pour les réseaux biologiques

## 5.1 Présentation de MIN

Dans le domaine de la régulation biologique, les modèles d'un système obtenus à partir de la biologie expérimentale, sont habituellement des réseaux *complexes* de régulations de gènes. Ces réseaux, appelés réseaux de régulation biologique, permettent de décrire les influences entre des gènes (ou autres entités biologiques, comme les macromolécules ; on parlera alors de façon générale d'influence entre des *variables*) à l'intérieur de systèmes biologiques.

Ces réseaux ont été modélisés mathématiquement suivant différents formalismes, en particulier celui de René Thomas [166, 170, 163, 167, 164]. Ce formalisme se base tout d'abord sur une approche logique booléenne de ces réseaux (un gène est soit en position "on", soit en position "off"), avant de se généraliser avec une approche logique multivaluée (différentes interactions, différentes positions (>2) du gène) [169]. Malgré cela, l'étude de ces réseaux de régulation en est à ses débuts au cause de la complexité du traitement

algorithmique qui augmente significativement avec le nombre des variables du modèle, et l'enjeu reste tout de même de taille. Chacun des réseaux est différent et on peut dériver de chacun d'eux une évolution au cours du temps. Cette évolution s'appelle le *comportement* du réseau de régulation biologique. Ce comportement est important aux yeux des biologistes, en particulier si le réseau étudié comporte une ou plusieurs boucles, dites de rétroaction [173, 39], entre ses variables, car c'est lui qui détermine la finalité du réseau et donc son effet dans le système biologique global.

Le formalisme de modélisation de réseaux d'interactions modulaire MIN que nous avons développé décrit les réseaux biologiques à la fois au moyen d'un support de représentation graphique lisible par l'homme, et d'annotations textuelles. MIN comporte à la fois les informations nécessaires aux modélisations qualitatives et quantitatives. La définition du modèle MIN et la traduction du MIN vers le formalisme de René Thomas fait l'objet d'une publication dans la revue BMC Bioinformatics. La traduction en ODEs fait l'objet d'une publication dans JIB, Journal of Integrative Bioinformatics. Enfin, la traduction du MIN en réseaux de Petri a été sélectionnée pour une présentation à la conférence ECCS'O7.

Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

## 5.2 Modèle incrémental et unificateur pour les réseaux d'interactions biologiques

Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

# Incremental and unifying modelling formalism for biological interaction networks

ANASTASIA YARTSEVA[1], HANNA KLAUDEL[2], RAYMOND DEVILLERS[3], FRANÇOIS KÉPÈS[4]

**Abstract** We propose a new unifying and incremental formalism for the representation and modeling of biological interaction networks. This approach provides an additional level of description between the biological and mathematical ones. It yields, on the one hand, a knowledge expression in a form which is intuitive for biologists and, on the other hand, its representation in a formal and structured way. This formalism allows automated translations into other formalisms, thus enabling a thorough study of the dynamic properties of a biological system. As a first illustration, we propose a translation into the R. Thomas' multivalued logical formalism which provides a possible semantics; a methodology for constructing such models is presented on a classical benchmark: the $\lambda$ phage genetic switch. We also show how to extract from our model a classical ODE description of the dynamics of a system.

**Keywords** Abstract biological models, regulatory interaction networks, formal methods.

## 1 Introduction

Often, modeling approaches in biology try to fit the data into the Procrustean bed of a particular modeling formalism [14, 3, 15, 8, 21]. However, if the area of interest changes, the modeling process has to be continued (or even restarted) using a different modeling language, more adapted to the new area. An appropriate choice of the modeling formalism may be crucial for efficiently describing biological systems, avoiding to change the description language and permitting to reuse the previous work.

In this paper, we propose a modeling formalism for the biologists that enables the expression of various types of biological knowledge in a formal manner and its translation into target formalisms for analysis or simulation. It aims at satisfying the following requirements:

- universality: the integration of various kinds of biological data available today;
- parsimony: the simplest possible representation of the data;
- incrementality: the construction of more complex models from simpler ones;
- precision: expression of relations in a non-ambiguous (mathematical) way;
- transposability: formal rules for the translation of the information contained in the model into commonly used (target) modeling formalisms.

In such a formalism, the model can be seen rather as a well-organised knowledge base of information about the biological system. Every unit of information (which has no biological sense when divided) inside the model can be called a *data*. In this approach, we assume that there is neither contradictory nor "bad" data. In other words, every measurement, every observation may be true in some context.

[1](corresponding author) IBISC - Université d'Evry Val d'Essonne, Tour Evry 2, 523 place des Terrasses de l'Agora, F-91000 Evry, France, E-mail: aiartsev@ibisc.univ-evry.fr

[2]IBISC - Université d'Evry Val d'Essonne, Tour Evry 2, 523 place des Terrasses de l'Agora, F-91000 Evry, France

[3]Département d'Informatique, Université Libre de Bruxelles, CP212, B-1050 Bruxelles, Belgium

[4]Epigenomics Project, Genopole®, CNRS & Univ. Evry, France

Our approach, called Modular Interaction Network (MIN), is a formalism designed to represent biological data, having a bipartite network structure and admitting a graphical representation, even if not focused on it. MIN enables the integration of microscopic (molecular interactions) and macroscopic (system states) data, thus allowing to provide the desired level of abstraction. This abstraction allows to avoid the rather common problem of explosion of the model complexity [13]. MIN has a limited number of node and edges types, which enables to represent biological networks in a simple way, even if more detailed information can also be stored and recovered. MIN suits for the representation of genetic regulation as well as of metabolism with multi-molecular biological processes, in a natural and incremental manner. MIN is also provided with algorithms enabling a translation to two classical modeling formalisms: multi-level logical modeling [23] and differential equations. These translations can be performed at any stage of the modeling process.

The paper is structured as follows. After recalling the biology of the $\lambda$ phage, which will be used as a running example, the MIN model is introduced, first informally and then formally. Next, in sections 5 and 6, the multi-level logical approach is first recalled and then used as a semantics of MIN. In section 7, this translation is extensively illustrated on the $\lambda$ phage example. A translation to ordinary differential equations is then sketched in section 8. Finally, some concluding remarks and perspectives are presented.

**Biology of $\lambda$ phage** In order to illustrate our approach, we shall use as a running example a classical biological benchmark: the genetic switch of the $\lambda$ phage, which will be presented first.

The $\lambda$ phage is a virus which infects the *Escherichia coli* bacteria. It turns out that a lot of quantitative and qualitative information is now available on it, so that it has become a benchmark organism and plays a central role in modeling [17, 14, 21, 24, 15, 8, 5].

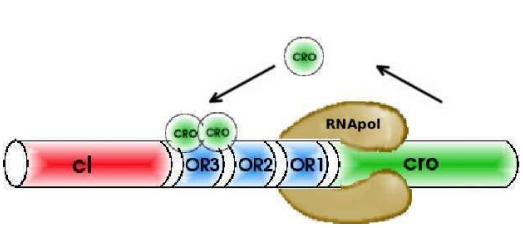
When a $\lambda$ phage encounters a bacterium, it can attach itself to specific receptors on the bacterial membrane. At this moment, the virus genome enters the bacterium. Then, two alternative pathways are possible:

- *lytic pathway*: the virus uses the host machinery in order to replicate its genetic material and create new viruses. This phase takes about 45 minutes, then the bacterium is destroyed and about one hundred viruses are released in the external media (Figure 1(a)).
- *lysogenic pathway*: the virus integrates its genetic material in the bacterial genome. There is no production of viruses. The bacterium is said to be lysogenised. The virus can stay indefinitely in the genome of its host. But there exists an escape mechanism: in some cases, the virus can extract itself from the bacterial genome and enter a lytic phase as a response to some stimuli (Figure 1(b)).

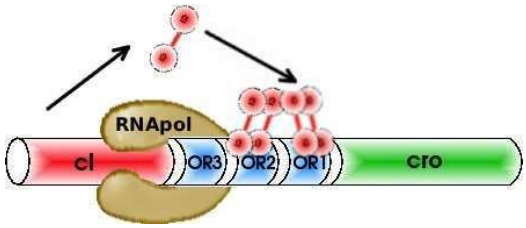
A small region of the viral genome controls the decision between lytic or lysogenic pathway. This region is composed of two genes and their two promoters (sites of regulation of the gene expression) and is referred to as the *genetic switch region* (see Figure 1). The decision results from the competition between two major proteins:

- the first one is referred to as $CRO$, encoded by gene $cro$, and expressed during lytic phase.
- the second one is called $\lambda$ repressor, referred to as $CI$. It is encoded by gene $cI$, and it can activate other genes, including itself, and repress others. $cI$ is expressed during lysogenic phase.

Note that the competition between $CI$ and $CRO$ is also influenced by the host environment. The host environment is captured through $CI$ and $CRO$ and their influence on the regulator region, i.e., the *genetic switch*.




(a) The situation in a *lytic infection*. $CRO$ protein occupies $OR3$, preventing RNA Polymerase from initiating transcription from the $cI$ promoter. RNA Polymerase transcribes the $cro$ gene, producing more $CRO$ protein, which silences $CI$ transcription.




(b) The situation in a *lysogenic cycle*. $CI$ protein induces $cI$ gene transcription and $cro$ gene silencing. The $CI$ repressor protein binds $OR2$ and $OR1$, preventing RNA Polymerase from transcribing the $cro$ gene, and promoting $cI$ transcription. Unlike $CRO$, $CI$ has an activation domain that promotes RNA Polymerase binding to its own promoter.

Figure 1: The genetic switch of the $\lambda$ phage. The $cI$ and $cro$ genes lie on opposite sides of the operator region, containing three operators ($OR1$, $OR2$, $OR3$). The two genes are transcribed in opposite directions from their respective promoters, which overlap in the middle operator, $OR2$.

## 2 Informal presentation of the unifying modeling formalism

The description of a biological system is often obtained by constructing an *interaction network*. Intuitively, as biological interactions are considered to always rely on so called *regulatory sites*, the network construction starts by their identification. Every regulatory site has a set of regulating and regulated *chemical species* and their role is expressed by *influences*. Sometimes, and in particular when the abstraction level is high, the choice of representing a set of biochemical reactions by a species or by a regulatory site is rather arbitrary. However, at the base level the chemical reactions are represented by regulatory sites and chemical species by species of MIN. In fact, both species and regulatory sites are fully characterised by their *level of activity* while describing the state of a biological system. As a consequence, regulatory sites and chemical species form the set of *variables* of the interaction network (see Table 1 for some examples of variables). Thus, two main classes of abstract entities are chosen to be components of interaction networks: *variables* and *influences* between them. We consider two kinds of *influences* between the variables of the model: *Influences of Chemical species on Regulatory sites (ICR)* and *Influences of Regulatory sites on Chemical species (IRC)*. We also assume that there is no influence between variables of the same kind. The whole representation is called Modular Interaction Network (MIN).

| biological object | role | model entity |
|---|---|---|
| gene | information storage and propagation | species |
| regulatory sequence of DNA | regulation of gene activity | regulatory site |
| protein | catalysis | species |
| phosphorylation or cleavage site | regulation of protein activity | regulatory site |
| metabolic pathway | transformation of molecules | regulatory site |
| receptor on a cell surface | detecting environmental state | regulatory site |
| ... | ... | ... |

Table 1: Examples of representations of biological objects in MIN according to their biological function, either of a catalytic or regulatory nature.

Such models may be composed. The trivial case of a composition is the union of models having no common species or sites. The union of data contained in these models is the new, composed,

model. In the case of models sharing common entities, the repeated nodes of the resulting network are collapsed.

MIN being an abstract formalism, its semantics is not intended to be defined directly, but rather as a translation into a target model. In this paper, we first define a translation of MIN into the Multivalued Logical modeling formalism (MLM) [23]. The multivalued logical approach is designed to express the interdependency between activity levels (often concentrations) of biological objects, e.g., proteins. It applies when this interdependency can be represented by a sigmoidal curve, which is approximated by a multivalued logical function. This function can distinguish between different levels of activity of a biological object, so it may be *multivalued* (see Figure 2). The multivalued logical model (MLM) consists of two parts: a directed graph of interactions and a table of *dynamic parameters*.

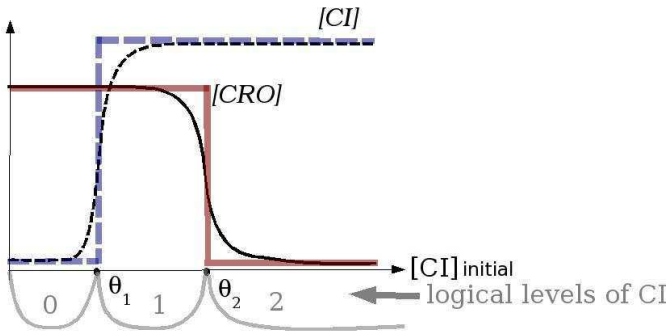



Figure 2: The multivalued logical approximation of the level of activity of biological objects. The axes represent input (abscissa) and output (ordinate) protein concentrations. The dashed thin sigmoid curve represents $[CI]$ – the measured concentration of the protein $CI$ at the equilibrium point. This curve is approximated by the thick dashed multivalued logical function with the threshold $\theta_1$. The solid curve corresponds to the influence of [CI] on $[CRO]$ and its approximation by the multivalued logical function with the threshold $\theta_2$. In this case the activity of the protein $CI$ has three logical levels: 0, 1 and 2, indicated in the bottom part and separated by the thresholds.

The multivalued logical representation of genetic regulatory networks [23] is one of the closest to the biological intuition. The major problem of this formalism is that it is not incremental, which means that updating an existing model (by adding or removing nodes or edges in the regulatory graph, for instance) leads to the situation where the set of dynamic parameters changes in an unpredictable way, as well as the dynamics of the system. In order to cope with this problem, the idea is to describe the biological system in MIN and translate it automatically, when needed, at any modeling step, into the multivalued logical formalism. This translation should preserve as many as possible of the biological properties already expressed in MIN. The dynamics of the translated MIN is then based on the information available in the attributes of its influences. The interaction graph can be obtained more or less directly from the MIN presentation of a biological regulatory network (see also Figure 11). The variables of the MLM (nodes of the graph) are obtained from the species of the MIN. The influences of MLM (edges of the graph) are obtained from pairs of $(ICR, IRC)$ present in the MIN and having a common regulatory site. The dynamic parameters of MIN indicated as attributes of its influences will serve to constrain possible dynamic parameters in the obtained multivalued logical model.

In order to further illustrate the flexibility of the MIN approach, we shall also show how to extract the dynamics of the associated chemical reactions in terms of ordinary differential equations, either directly or through a demultiplication of the regulatory sites which may represent various different reactions.

# 3 Modular Interaction Network (MIN)

As mentioned in the introduction, MIN formalism considers two types of entities: variables (chemical species and regulatory sites) and influences (IRCs and ICRs). Every model entity (site, species, influence) is characterised by its *attributes* which can be any data concerning the biological object or interaction represented by this entity; for example:

- physical attributes: size and shape for a protein, position in DNA for a genomic sequence;
- localization in space (cell compartments: nucleus, cytosol);
- expression pattern (cell types, tissues etc.);
- observable values of the activity level for the biological object;
- velocity, force, speed, amplification factor, cooperativity increase, energy of the interaction.

From the very beginning, for any bit of information added to the model, the link to the source (the set of *references* to papers, databases, etc.) of it should be specified. This will be important in later steps of the modeling, for example in order to estimate the data quality. We assume that all the data in the model has a representation which allows it to be compared (it may be, for instance, a textual "string" representation).

## 3.1 Variables

Both species and regulatory sites may represent biological objects of some abstraction level (molecules or parts of them, complex processes like regulatory pathways, complex systems like sensors, or even an entire organism). As our knowledge about biological systems is based on *observations* and *experiments*, the *observable level of activity* of biological objects can change in different states of the biological system. These objects can influence the levels of activity of the other biological objects. So, every species and site in MIN will be assumed to have a set of *observable values*, corresponding to the observable levels of activity of the corresponding biological objects.

The formal definition of a MIN variable reflects the presence of various features (attributes) in biological objects. Also, in different sources a biological object can have different names (hence the name set of a variable). Moreover, the measurement methods used to observe the activity level of this object yield a set of possible values for the variable, usually (partially) ordered.

**Definition 1** *A* variable $V$ *is an entity characterized by a tuple* $(N, W, P, L)$ *where:*

- $N$ *is a non-empty set of known* names *of the variable;*
- $W$ *is a partially ordered (by* $\prec_V$*) set of* observable values *representing the activity level of the biological object associated to the variable. We shall assume that this set has at least the default value* undef*, unordered with respect to the other values, and two defined values, meaning that the variable is not a constant;*
- $P$ *is a set of* attributes*, having a* type*, a* value *and the boolean* unique *field.* $unique = 1$ *indicates that this attribute can not be present in* $P$ *more than once. Otherwise, several attributes of the same type can have different values;*
- $L$ *is a non-empty set of* links *to (bibliographic) sources of the information about the variable. This set of attributes will always include the kind of the variable (which is unique and can be either "regulatory site" or "chemical species").*

**Chemical species** A species represents a biological object with catalytic or binding capabilities, which influence one or more regulatory sites. These influences have a chemical nature: association/dissociation reactions, electron transfers, etc. A species may have one or more influence capabilities, that will be called *affinities*.

An affinity is the ability of a biological object to interact with (potentially) a set of other biological objects through a particular regulatory site. Thus, an affinity may correspond to a protein domain for a protein or a surface molecule (receptor) for a cell.

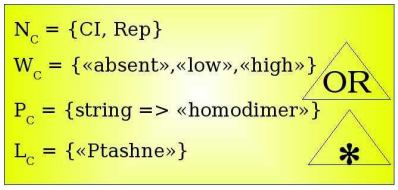
**Definition 2** *An* affinity $a$ *is a tuple* $(\ell_a, P_a, L_a)$ *where:*

- $\ell_a$ *is a* label *representing the affinity name (which is indeed the label of the binding regulatory site);*
- $P_a$ *is a set of* attributes *of the affinity, having a* type *and a* value *(not necessarily unique);*
- $L_a$ *is a non-empty set of* links *on sources of the information about this affinity (bibliographic references).*

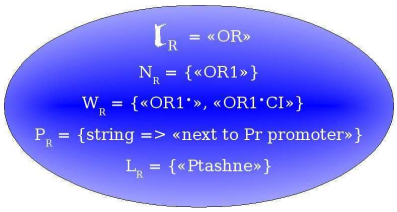
Now we are able to formally introduce chemical species:

**Definition 3** *A* chemical species $C$ *is a variable* $(N_C, W_C, P_C, L_C)$ *whose set of attributes* $P_C$ *contains* $(Kind, \text{“chemical species”}, 1)$ *and one ore more* $(Affinity, a, 0)$*, where different* $a$*'s enumerate the influence abilities of the species* $C$*.*

Chemical species are graphically represented by rectangular boxes. Various affinities can be represented inside the species (by named triangles) omitting all the details except for their label. The nature of the interaction between two biological entities can be unknown. So, a wild-card affinity, labeled “*”, may be defined for every species, standing for an unknown mechanism of regulation (see Figure 3(a) for an example of a chemical species).




(a) Chemical species named $CI$ or $Rep$, with two affinities (among which the wild-card), a bibliographic link, an attribute (besides $Kind$ and $Affinities$) and three observable values.




(b) Regulatory site named $OR1$, with a label $OR$, two observable values, an attribute (besides $Kind$ and $Label$) and a bibliographic link.

Figure 3: Representation of a chemical species and of a regulatory site.

**Regulatory sites** A *regulatory site* regulates species activity in a manner which cannot be represented by a chemical reaction, like for example by three-dimensional conformation changes in a molecule or cooperativity effects. A regulatory site may represent a genome region or a protein domain that changes its state after a chemical reaction.

A regulatory site has a *label* which characterizes its capabilities of being influenced through *affinities*. If a regulatory site and an affinity of a species have the same label, it means that the

interaction is possible between the biological objects corresponding to the site and the species. A regulatory site represents an "input" for a species and regulates its activity through integration of several influences on it.

**Definition 4** *A* regulatory site $R$ *is a variable* $(N_R, W_R, P_R, L_R)$ *with the attributes* $(Kind, \text{“regulatory site”}, 1)$ *and* $(Label, \mathfrak{l}_R, 1)$ *in the set* $P_R$*, where* $\mathfrak{l}_R$ *is a label representing the site type.*

Regulatory sites are graphically represented by ellipses containing the label $\mathfrak{l}_R$ inside a triangle. An example of a regulatory site is given on the Figure 3(b). The presented site has two different states: free $(OR1\cdot)$ and regulated $((OR1 \cdot CI))$. This means that the corresponding biological object can participate in binding with another object. The label of this site is $OR$, so it can be influenced by a species having an affinity labeled $OR$, like the one represented on Figure 3(a).

In the MIN representation, different biological objects are associated to different entities in the model. The attributes of sites and species may have types like "position", "size", "location" etc. expressing a knowledge about these biological objects. For example, if a gene has more than one regulatory site of the same type in its regulatory region, several sites will be present in the model, having the same label but with different positions (mentioned in the attribute set); clearly, in this case, the corresponding variables will not be compatible. All these sites will influence the species corresponding to the gene. However, several species with the same name may be present in MIN, if they have attributes with different values. So, we can represent a molecule of the same protein in free or dimerised state, or the same gene at its natural location and translocated in a different place in the genome.

## 3.2 Influences

Biological objects, represented by species and sites in MIN, may interact and play specific roles in these interactions. For example, they can take part in a chemical reaction, one object modifying, creating or destroying another one. We assume that every interaction happens through an affinity and a regulatory site. More formally, a chemical species $C_1$ having an affinity $a$ with a label $\mathfrak{l}_a$ can influence a chemical species $C_2$ if there is a regulatory site $R$ labeled by the same label $(\mathfrak{l}_R = \mathfrak{l}_a)$ which influences the species $C_2$. An influence is defined between two MIN variables as follows:

**Definition 5** *An* influence $I$ *between variables is a tuple* $(V, V', P, L)$ *where:*

- $V$ *is the* influencing *variable;*
- $V'$ *is the variable* influenced *by* $V$*;*
- $P$ *is the set of influence* attributes, *having a* type *and a* value *(not necessarily unique);*
- $L$ *is the set of* links *to sources of the information about the influence.*

The influence (ICR) of a species on a regulatory site of another species represents the chemical interaction between two biological objects in which the state of the regulatory site is modified by the species through an affinity. Symmetrically, a regulatory site can *influence* the value of a species, through the influence (IRC) of a regulatory site on a chemical species. In this case the interaction between corresponding biological objects cannot be represented by a chemical reaction, and there is no specific affinity associated to such an influence.

**Definition 6**

- *An* influence $ICR$ *of a Chemical species* $C_{ICR}$ *on a Regulatory site* $R_{ICR}$ *is an influence* $(C_{ICR}, R_{ICR}, P_{ICR}, L_{ICR})$ *with an attribute* $(Affinity, a_{ICR}) \in P_{ICR}$ *which is the* affinity *involved in the interaction of the species* $C_{ICR}$ *and the site* $R_{ICR}$*, hence with* $(Affinity, a_{ICR}, 0) \in P_{C_{ICR}}$ *and either* $\iota_{a_{ICR}} = \iota_{R_{ICR}}$ *or* $a_{ICR} = *$*.*
- *An* influence $IRC$ *of the regulatory site* $R_{IRC}$ *on the species* $C_{IRC}$ *is an influence* $(R_{IRC}, C_{IRC}, P_{IRC}, L_{IRC})$ *with the attribute* $(Kind, IRC) \in P_{IRC}$*.*

An influence has a set of attributes, which should describe, in particular, the relationship between the values of the species and those of the regulatory site, like the parameters of the corresponding chemical reaction: kinetic rate or speed, or stoichiometric coefficients. Several examples of the IRCs and ICRs are shown on the Figure 4, by dashed and plain arcs, respectively.

### 3.3 The network

After presenting the species and the regulatory sites, the influences between them, we can now give a formal definition of the MIN for the modeling of a biological system. The information about the possible connections between species of the system is already coded in the labels of the regulatory sites and affinities. We consider that the states of the model are expressed through observable values of species and sites, so that $\Omega_C$ denotes the set of functions associating a value of its value set to each species of the model, $\Omega_R$ is the same for the sites of the model, and $\Omega$ is the set of all possible *observable states* of the model. In the following, $\omega \in \Omega$ stands for any given observable state of the system and $\omega(V)$ will stand for the value of the variable $V$ in the state $\omega$.

In general, in a single biological experiment (an *observation*), the values of only a subset of biological objects are measured. In this case, the observable values of non observed species and sites take the special value *"undef"* and the state of the system will be considered as "partly" defined.

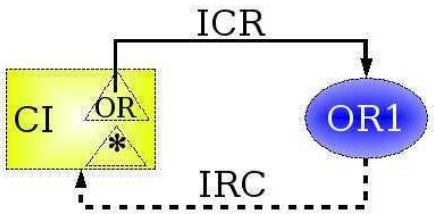
In the set $\Omega$ of observable system states a subset $\mathcal{F} \subset \Omega$ of *observed system states* will yield all the partly defined system states which were really observed in biological experiments and described by biologists. $\mathcal{F}$ plays the role of a databank from which the parameters of the dynamics of the system interactions could be inferred. If some of these parameters (as, for example, kinetic rates for biochemical reactions) are known (were measured in biology), they will be directly mentioned in the attributes of the corresponding influences (there will be some attribute of the kind $(Kinetic\_rate, 15)$ belonging to $P_{ICR}$ or $P_{IRC}$, for instance).

**Definition 7** (MIN) *A modular interaction network is a tuple* $\mathcal{M} = (\mathcal{V}, \mathcal{ICR}, \mathcal{IRC}, \mathcal{F}, \mathcal{L})$ *where:*

- $\mathcal{V} = \mathcal{C} \cup \mathcal{R}$ *is the set of variables of the model; it is partitioned in a set* $\mathcal{C} = \{C_i | i = 1..|\mathcal{C}|\}$ *of chemical species and a set* $\mathcal{R} = \{R_j | j = 1..|\mathcal{R}|\}$ *of regulatory sites;*
- $\mathcal{ICR} \subseteq \{ICR_{ija} | i = 1..|\mathcal{C}|, j = 1..|\mathcal{R}|, (Affinity, a, 0) \in P_{C_i}\}$ *is a set of influences from chemical species to regulatory sites through an affinity of the former and there is no more than one influence between such a pair of variables through the same affinity;*
- $\mathcal{IRC} \subseteq \{IRC_{jk} | j = 1..|\mathcal{R}|, k = 1..|\mathcal{C}|\}$ *is a set of influences from regulatory sites to chemical species and there is no more than one influence between such a pair of variables;*
- $\mathcal{F} \subset \Omega$ *is a set of observed partly defined states of the biological system;*

- $\mathcal{L}$ *is a set of* links *to sources of the information about those observations.*

In figures, species will be represented by boxes, affinities by triangles inside the boxes of species, regulatory sites by ellipses, influences of a species on a regulatory site by plain arcs, and influences of a regulatory site on a species by dashed arcs. A small example of an interaction network is presented in Figure 4.




**Relation $\mathcal{F}$**

| CI | OR1 |
|---|---|
| absent | free |
| low | free |
| low | CI_bound |
| high | CI_bound |

Figure 4: A small interaction network representing the chemical species $CI$ and the (regulatory) site named $OR1$. **Left**. The influence ICR links the affinity labeled $OR$ of species $CI$ with the site $OR1$, and the influence IRC links the site $OR1$ and the species $CI$. In the $\lambda$ switch, the regulatory site $OR1$ corresponds to the regulatory region in the DNA molecule coding for the protein $CI$. Thus, $CI$ can influence the regulatory site $OR1$, and the activity of $CI$ can be regulated through the regulatory site $OR1$. **Right**. The corresponding relation $\mathcal{F}$ indicating the biologically observed states of the network.

A MIN model having a highest level of detail has the property that each regulatory site corresponds to a (single) chemical reaction. We present an example of such a model in Figure 5. It illustrates the CI protein synthesis from the CI gene regulated by the OR1 regulatory site in function of the presence of CI protein dimer.

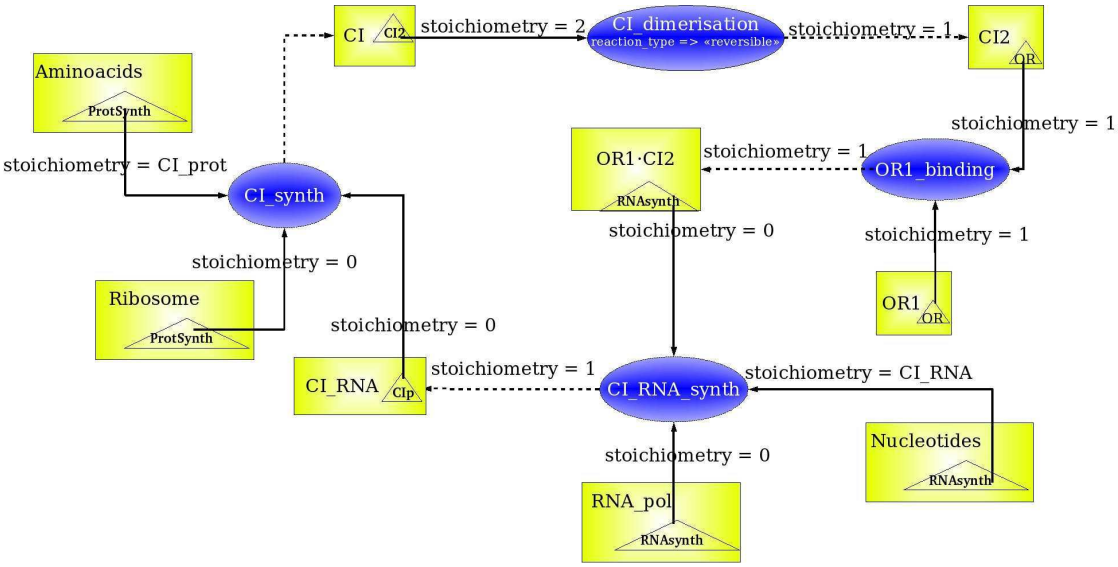



Figure 5: A MIN model representing the enzymatic reaction of CI synthesis. The reactions $CI\_dimerisation$ and $OR1\_binding$ are reversible, so they have the appropriate attribute. The reactions $CI\_RNA\_synth$ and $CI\_synth$ are non reversible and have the appropriate attribute.

The corresponding chemical species are represented by chemical species of the MIN model. The biochemical reactions of this example are represented by regulatory sites, because a reaction is possible when all the substrates are present. This reaction regulates the level of activity of a chemical species by increasing or decreasing its quantity (concentration). Each reaction has an

attribute "reversible" or "not reversible". For instance, if a reaction is reversible, this means that all the species connected to this reaction can be either products or substrates of the reaction. Another attribute of the regulatory site is a kinetic rate, which is in general a function of other mensurable parameters of the system such as concentrations of species catalyzing the reaction or even non participating directly in the reaction but influencing its kinetics. For example, such species can sequestrate one or more substrates or products or catalyze intermediate reaction steps. Another natural parameter of the kinetic rate function is the temperature: biochemical reactions go faster when the temperature increases.

On each influence adjacent to the regulatory site, an attribute corresponding to the stoichiometric coefficient is indicated. It may have 3 qualitatively different values:

- 0, which means that the corresponding species is an enzyme, i.e., it is not consumed or produced in this reaction, even if its presence is necessary for the reaction takes place;
- a numerical value, which corresponds to the number of molecules implicated in the reaction, generally one or two;
- any other label, standing for a vector of coefficients saying how many molecules of each of the 20 types of aminoacids $(a_1, a_2, \ldots, a_{20})$ or each of the 5 types of nucleotides $(n_1, n_2, n_3, n_4, n_5)$ is needed to synthesize the macromolecular product of the reaction.

For example, the stoichiometric coefficients for $Nucleotides$ and $Aminoacids$ in Figure 5 are labels, and each label represents the composition of the corresponding macromolecule: CI RNA or CI protein. In general, the opposite reaction of the biochemical synthesis is degradation, and it liberates the same quantities of the corresponding substrate residuals. The stoichiometric coefficients for $RNA\_pol$ or $Ribosome$ are 0, which means that these are enzymes in the reactions of CI RNA synthesis and of CI protein synthesis. The stoichiometric coefficient for $CI$ is 2 for the reaction of the dimerisation of CI, meaning that two molecules of CI are needed to form a dimer.

### 3.4 Compression of MINs

In order to simplify MIN models, it may be interesting to find the variables representing the same biological object and to combine them. So, the following definition introduces the syntactic *compatibility* and the *union* of variables.

**Definition 8** (Compatibility and union of variables) *Let* $\{V_i \mid i = 1, 2, \ldots, k\}$ *be the set of variables of the MIN* $\mathcal{M}$*, with* $V_i = (N_i, W_i, P_i, L_i)$*. The variables in this set will be said to be* compatible *if they have the same names* $(\forall V_i, V_j \ N_{V_i} = N_{V_j})$*, their unique attributes are compatible* $((x, y, 1) \in P_i \wedge (x, z, b) \in P_j \Rightarrow y = z \wedge b = 1)$*, if their partial orders are compatible* $((\bigcup_{i=1}^{k} <_{V_i})^*$ *is acyclic*$)$ *and their observed values are compatible* $(\forall V_i, V_j \forall (\ldots, w_i, \ldots, w_j, \ldots) \in \mathcal{F}$ *either* $w_i = undef$ *or* $w_j = undef$ *or* $w_i = w_j)$*. In such a case, their* union $\bigcup_{i=1}^{k} V_i = (\bigcup_{i=1}^{k} N_i, \bigcup_{i=1}^{k} W_i, \bigcup_{i=1}^{k} P_i, \bigcup_{i=1}^{k} L_i)$*, with* $<_{\bigcup_{i=1}^{k} V_i} = (\bigcup_{i=1}^{k} <_{V_i})^*$.

As the values of variables come from different biological experiments, in order to compare them we need to use the same approximations as generally accepted by biological science. This means that the "equality" of values $w_i = w_j$ should be confirmed by a biologist when it is not obvious. Notice also that chemical species may only be compatible with other chemical species, and similarly for regulatory sites.

This definition will sometimes allow to reduce the representation of a MIN, by replacing compatible sets of variables by their union. Moreover, the translation of MIN representation in other formalism can allow further compression of variables depending on the capability of the formalism to distinguish between different biological objects.

Thus, the *simplification* is an operation on MIN $\mathcal{M}$ which produces MIN $\mathcal{M}'$ in a following way:

- First of all, the compatible variables of the MIN $\mathcal{M}$ are combined;
- then, the ICRs (IRCs) of a variable $V_1$ on $V_2$ of the MIN $\mathcal{M}$ are linked to the variables $V_1'$ and $V_2'$ of $\mathcal{M}'$, where $V_1'$ is compatible with $V_1$ and $V_2'$ is compatible with $V_2$;
- the relation $\mathcal{F}$ is updated: the entries containing a pair of combined variables with different observed values are splitted in two entries where only one value at a time is listed for the combined variable.

The formal definition of MIN simplification is presented below.

**Definition 9** (Simplification of MIN) *If $\mathcal{M} = (\mathcal{V}, \mathcal{ICR}, \mathcal{IRC}, \mathcal{F}, \mathcal{L})$ is a MIN, $\mathcal{V}' = \mathcal{C}' \cup \mathcal{R}'$ is a partition of $\mathcal{V}$ into sets of compatible variables in $\mathcal{M}$, the* compressed *form of $\mathcal{M}$ through the partition $\mathcal{V}'$ is the MIN $\mathcal{M}' = (\mathcal{V}', \mathcal{ICR}', \mathcal{IRC}', \mathcal{F}', \mathcal{L})$ defined as follows:*

- *each variable $V' \in \mathcal{V}'$ represents the union of compatible variables composing the set $V'$ ($V' = \bigcup_{V \in V'} V$);*
- $\mathcal{ICR}' \stackrel{def}{=} \bigcup_{C' \in \mathcal{C}'} \bigcup_{a:(Affinity,a,0) \in P_{C'}} ICR'_{C',a}$ *where* $ICR'_{C',a} = \{(C', R', P', L') \mid R' \in \mathcal{R}', X = \{(C, R, P, L) \in \mathcal{ICR} \mid C \in C', R \in R', (Affinity, a) \in P_{(C,R,P,L)}\} \neq \emptyset, P' = \bigcup_{ICR \in X} P_{ICR}, L' = \bigcup_{ICR \in X} L_{ICR}\}$ .
- $\mathcal{IRC}' \stackrel{def}{=} \{(R', C', P', L') \mid R' \in \mathcal{R}', C' \in \mathcal{C}', X = \{(R, C, P, L) \in \mathcal{IRC} \mid C \in C', R \in R'\} \neq \emptyset, P' = \bigcup_{IRC \in X} P_{IRC}, L' = \bigcup_{IRC \in X} L_{IRC}\}$;
- $\mathcal{F}' \stackrel{def}{=} \{\omega' = (w'_1, \ldots, w'_{|\mathcal{V}'|}) | \exists (w_1, \ldots, w_{|\mathcal{V}|}) \in \mathcal{F}, \forall i (\forall V_j \in V'_i w'_i = w_j = undef \vee \exists V_j \in V'_i \; w'_i = w_j \neq undef)\}$.

### 3.5 Composition of MINs

One of the main characteristics of MINs is that they are modular and enable an incremental construction of models of biological systems. The operation of *composition* of two MINs includes establishing new, composed, sets of species, sites and influences. The species set of the resulting MINs is the *union* of species of the composing MINs, and the new sites set is the union of regulatory site sets of composing MINs. All the information about the interactions in composing systems must be also preserved. That means that a particular attention should be paid on the conversion of influences from composing MINs to the resulting one. If source MINs do not contain common species, there is no transformation to perform; the data from these MINs should be just put together.

**Definition 10** (Union of MINs) *If $\mathcal{M}_i = (\mathcal{C}_i, \mathcal{R}_i, \mathcal{ICR}_i, \mathcal{IRC}_i, \mathcal{F}_i)$ for $i = 1, 2$ are MINs, their* union *$\mathcal{M} = \mathcal{M}_1 \oplus \mathcal{M}_2$ is the MIN such that $\mathcal{M} \stackrel{def}{=} \{\mathcal{C}_1 \cup \mathcal{C}_2, \mathcal{R}_1 \cup \mathcal{R}_2, \mathcal{ICR}_1 \cup \mathcal{ICR}_2, \mathcal{IRC}_1 \cup \mathcal{IRC}_2, \mathcal{F}_1 \times \mathcal{U}_2 \cup \mathcal{U}_1 \times \mathcal{F}_2\}$, where $\mathcal{U}_i$ is the state of model $\mathcal{M}_i$ where all variables have the value $undef$.*

This means that MIN models can be composed from parts that share the same species or are completely independent. This can be very useful at the first construction stages of biological regulatory networks where the data is incomplete and is not necessarily connected.

In case of presence of equivalent regulatory sites or species in the resulting MIN, the union of these sites or species must replace them. In this case the influences between all sites and all the species, which were influencing one another in the source MIN, must be established (see Figure 6). If there are in the source MIN two different influences between the same affinity of a species and the same regulatory site, they must be replaced by only one influence carrying the union of all possible attributes of both connections. In a same way, if there are two different influences from a regulatory site on a given species, it must be replaced by the influence carrying the union of all possible data, using the previously defined operation of simplification of MIN.

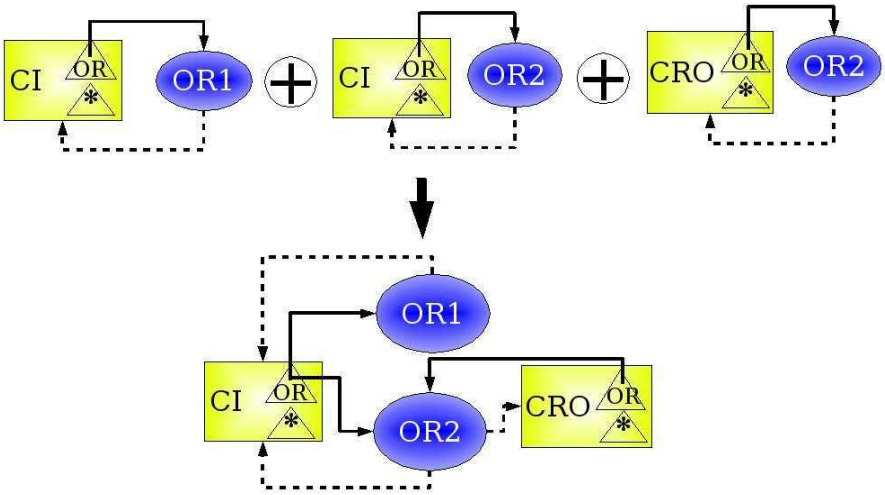



Figure 6: Union and compression of interaction networks. Three networks sharing species and regulatory sites can be combined into one by a composition and compressed by collapsing equivalent species and sites. All existing interactions are preserved.

## 4 Multivalued logical formalism (MLM): basics

The goal of modeling genetic regulatory networks in the multivalued logical formalism [23] is to obtain a state graph representing the behaviour of a biological system from a qualitative point of view. This means that an observable sequence of states of a biological system is represented by a path in the state graph of the model.

The multivalued logical formalism, which has been shown very useful for genetic networks study [22, 7], is composed of a directed labeled *regulatory graph* and a table of *dynamic parameters*. The *state* of the regulatory graph, expressed through the labels of its vertices, can evolve according to dynamic parameters. The possible traces of this evolution can be represented in the form of a *state graph*. The nodes of the state graph represent the different states of the system and the arcs of the state graph represent the possible activity modifications of the biological objects.

For dynamic systems with saturation (like genetic regulatory networks) one can approximate the sigmoid curve, representing the level of the activity of a variable as a function of the level of another one, by a multivalued logical function. This approximation is called *logical abstraction* because it allows to distinguish between only two activity states of the system: below the threshold level and above it.

The following definition describes an *instance of MLM* as introduced by R.Thomas. It is composed of a regulatory graph $(U, E)$ and a table $K$ of *dynamic parameters* (see Figure 7). Each node $u$ of the graph corresponds to a variable with integer values between 0 and the *boundary* $b_u$ of the variable, which drives the topology of the corresponding state graph. The influences between variables in MLM can be positive (*inducing*) or negative (*inhibiting*).

**Definition 11** (Instance of a Multivalued logical model) *An* instance $M$ of an MLM *of a genetic regulatory network is a pair* $(\mathcal{G}, K)$ *where:*

- $\mathcal{G} = (U, E)$ *is a labeled directed graph:*
  - *each vertex* $u \in U$ *is called a* variable *of the genetic regulatory network, and is provided with a strictly positive integer* $b_u$ *called the* boundary *of* $u$*;*
  - *each arc* $(u_1, u_2) \in E$ *is labeled by a pair* $(\theta, \varepsilon)$ *where* $\theta$*, called the* threshold*, is an integer between 1 and* $b_{u_1}$*, and* $\varepsilon$*, called the* sign*, belongs to* $\{+, -\}$*. When* $\varepsilon = +$*,* $u_1$ *is called an* inducer *of* $u_2$*. When* $\varepsilon = -$*,* $u_1$ *is called an* inhibitor *of* $u_2$*. The set of predecessors of* $u_2$ *is denoted* $\mathcal{G}^{-1}(u_2)$*.*
- $K = \{K_{u,\omega} \mid u \in U \wedge \omega \subseteq \mathcal{G}^{-1}(u)\}$ *is a family of integers such that* $0 \leq K_{u,\omega} \leq b_u$ *for any variable* $u$ *and any subset* $\omega$ *of predecessors of* $u$ *in the graph* $\mathcal{G}$*, called the* dynamic parameters *of* $u$*.*

The dynamics of an MLM instance $M$ is defined through the notion of states and transitions. A *state of* $M$ is a *mapping* $\mu : U \to \mathbb{N}$ such that, for any variable $u \in U$, $0 \leq \mu(u) \leq b_u$. The *value* $\mu(u)$ is then called the *level* of the variable $u$. For example, an MLM instance with two variables $u_1$ and $u_2$ with $b_{u_1} = b_{u_2} = 2$ has 9 states corresponding to the following mappings $\mu_1 = (0, 0), \mu_2 = (0, 1), \mu_3 = (0, 2), \ldots, \mu_7 = (2, 0), \mu_8 = (2, 1), \mu_9 = (2, 2)$. In this case the level of variable $u_2$ in state $\mu_2$ is $\mu_2(u_2) = 1$.

In order to unify the treatment of different influences between variables, the definition of *resources of a variable* is introduced in MLM. The variable $u_1$ influencing the variable $u_2$ is a resource in some state if $u_1$ *helps* the variable $u_2$ in that state, meaning that $u_1$ acts to increase the activity level of $u_2$.

**Definition 12** (Resources of a Variable) *Given a state* $\mu$ *and a variable* $u \in U$ *of a MLM* $M$*, the* set of resources *of* $u$ *is the set* $\omega_u(\mu)$ *containing all the variables* $u'$ *of* $M$ *such that:*

- $u' \in \mathcal{G}^{-1}(u)$ *is a predecessor of* $u$ *in the underlying directed graph* $G$ *of* $M$*;*
- *the arc* $(u', u)$ *is labeled by* $(\theta, \varepsilon)$ *and*
  - *if* $\varepsilon =$ "$+$" *then* $\mu(u') \geq \theta$*,*
  - *if* $\varepsilon =$ "$-$" *then* $\mu(u') \leq \theta$*.*

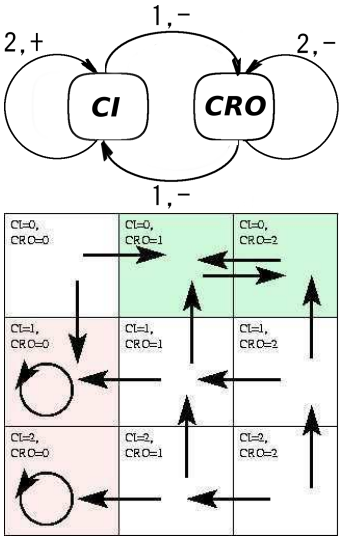
The set of variables $\omega_u(\mu)$ is consequently the subset of $\mathcal{G}^{-1}(u)$ containing both inducers of $u$ whose expression level has reached the threshold and the inhibitors of $u$ whose expression level has *not* reached the threshold.

The dynamics of the MLM reflects the dynamics of a "continuous" biological process, so the model variables cannot "skip" values: going from "1" to "3", for example, without passing by the value "2". So, the *multivalued logical function* is introduced to describe the evolution of a variable level in a given system state.

**Definition 13** (Multivalued Logical Function) *Given a state $\mu$ and a variable $u$ of an instance $M$ of MLM,* the multivalued logical function $\kappa_u(\mu)$ *is defined as follows:*

- *if* $\mu(u) < K_{u,\omega_u(\mu)}$ *then* $\kappa_u(\mu) = \mu(u) + 1$
- *if* $\mu(u) = K_{u,\omega_u(\mu)}$ *then* $\kappa_u(\mu) = \mu(u)$
- *if* $\mu(u) > K_{u,\omega_u(\mu)}$ *then* $\kappa_u(\mu) = \mu(u) - 1$

The function $\kappa_u$ represents a "step by step" evolution of the expression level of $u$ from its current expression level $\mu(u)$ to its dynamic parameters $K_{u,\omega_u(\mu)}$. The state graph of a MLM is often called *asynchronous* because only one variable can evolve at a time. Then, the evolution of the model can be represented as a *state graph*, where the system can move on a graph of system states according to its multivalued logical function.




| CI | CRO | **CI** dynamic parameter | **CRO** dynamic parameter |
|---|---|---|---|
| 0 | 0 | $K_{CI,\{CRO\}} = 1$ | $K_{CRO,\{CI,CRO\}} = 2$ |
| 0 | 1 | $K_{CI,\{\}} = 0$ | $K_{CRO,\{CI,CRO\}} = 2$ |
| 0 | 2 | $K_{CI,\{\}} = 0$ | $K_{CRO,\{CI\}} = 0$ |
| 1 | 0 | $K_{CI,\{CRO\}} = 1$ | $K_{CRO,\{CRO\}} = 0$ |
| 1 | 1 | $K_{CI,\{\}} = 0$ | $K_{CRO,\{CRO\}} = 0$ |
| 1 | 2 | $K_{CI,\{\}} = 0$ | $K_{CRO,\{\}} = 0$ |
| 2 | 0 | $K_{CI,\{CI,CRO\}} = 2$ | $K_{CRO,\{CRO\}} = 0$ |
| 2 | 1 | $K_{CI,\{CI\}} = 0$ | $K_{CRO,\{CRO\}} = 0$ |
| 2 | 2 | $K_{CI,\{CI\}} = 0$ | $K_{CRO,\{\}} = 0$ |

Figure 7: An MLM instance: its regulatory graph (left, top), the corresponding state graph (left, bottom) and the table of its dynamic parameters (right).

**Definition 14** ("Asynchronous" State Graph) *The* state graph *of a MLM $M$ is the directed graph $\mathcal{SG}$ whose vertices are all the possible states of $M$ and such that there is an edge from $\mu$ to $\mu'$ if and only if there exists a variable $u$ satisfying:*

- $\mu'(u) = \kappa_u(\mu) \neq \mu(u)$*;*
- *for any variable* $u' \neq u$ *we have* $\mu'(u') = \mu(u')$*.*

An arc of the state graph from $\mu$ to $\mu'$ is usually denoted as $(\mu \to \mu')$ and is called a *transition*. This is illustrated in Figure 7(right).

# 5 Translation of a MIN into an MLM

This section presents the translation algorithm of MIN into MLM formalism. It is structured in a following way. First of all, we note that multiple translations of MIN model into MLM formalism are possible, and the impact that it has on the translation algorithm. After that, the translation itself is described, starting with the construction of the MLM regulatory graph topology, then determining the dynamic parameters. At the end, this section contains an example of a translation of a small MIN network into MLM.

The obtained by translation MLM model will be called the *translated network*. As in many cases, the values of all parameters of the MLM model cannot be deduced precisely from the experimental data; the set of all possible parametrisations consistent with biological observations must be considered as a model which can be studied and later be refined by adding other information.

The biological information presented in MIN is much richer than that of an MLM instance, so one MIN can have multiple semantics expressed through a set of MLM instances. In other words, an MLM may be assimilated to the set of its instances. The topology of the regulatory network, as well as the boundaries, will be the same for all instances (deduced from that of MIN). However, dynamic parameters, as well as arc labels can be different since an arc of an MLM regulatory graph may correspond to several arcs of a MIN (one by affinity). As the observable values of a variable of a MIN are partially ordered (see Definition 1), the different ways of enumerating values of $u$ (topological sort) will be considered as yielding different instances of the MLM. So, in the following, we will consider every combinations of possible parameters as one instance of MLM, and the translation procedure of MIN into MLM will give all these possible parameters that can be deduced from MIN data.

Now, let us introduce the construction of the MLM regulatory graph from the MIN model. First, the *translated variables* of the MLM must be defined. They are obtained from the species of the MIN, keeping only one (arbitrarily chosen[5]) name and providing it with a boundary corresponding to the number of observable values of the MIN variable.

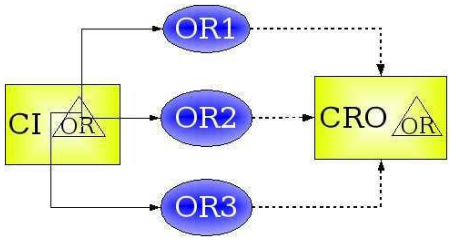
**Definition 15** (Translated variables of a MIN) *Let $C \in \mathcal{V}$ be a chemical species of the MIN $\mathcal{M}$, let $|W_C|$ be the number of different observable values of $C$ and $N \in N_C$ be a name of $C$. The* translation *of $C$ is a vertex $u \in U$ labeled with $N$ and provided with a boundary $b_u = |W_C|$. The species $C$ is then called the* original species *of $u$.*

The arcs of the regulatory graph of the MLM are deduced from the MIN structure in the following way: there is an arc between the translated variables $u_1$ and $u_2$ iff there is a pair $(ICR, IRC)$ in MIN such that $R_{ICR} = R_{IRC}$, and $C_{ICR}$ and $C_{IRC}$ are the original species of variables $u_1$ and $u_2$, respectively (see Figure 8).

The MLM regulatory graph is not complete yet, as we need to find the arc labels. These labels depend on the observed values of MIN variables. The information on the possible combinations of observed values of variables is contained in the relation $\mathcal{F}$. The same type of knowledge enables us to determine also the dynamic parameters of the MLM model. However, the influences are defined in MIN between chemical species and regulatory sites, but the MLM model encompasses the regulatory sites inside the variables representing the species, as shown in the previous definition. Thus, we need to reconstruct the parameters of influences of species on species from $\mathcal{F}$ and the MIN topology.

In order to find the arc labels of the translated regulatory graph and the corresponding dynamic parameters $K$, we introduce the relation $\Psi_{ik}$ between values of the species $C_i$ and the species $C_k$, called *interspecies regulation relation*. This relation is defined if there is a site $R_j$ such that there is an $ICR_{ij}$ with $(\mathit{Affinity}, a) \in P_{ICR_{ij}}$ and $(\mathit{Affinity}, a, 0) \in P_{C_i}$ and there is an $IRC_{jk}$ in the MIN, i.e., the species $C_i$ *regulates* the species $C_j$ through the site $R_j$. For example, on Figure 8, the species $CI$ regulates the species $CRO$ through the sites $OR1, OR2$ and $OR3$.

[5]Unless two species share a same name, due to unfortunate choices in independent sources; we shall assume it is always possible to choose those names in such a way that no two different nodes have the same name.




$\Psi_{CI,CRO}$

| | |
|---|---|
| absent | absent |
| absent | high |
| low | absent |
| high | absent |

**Relation $\mathcal{F}$**

| CI | CRO | OR1 | OR2 | OR3 |
|---|---|---|---|---|
| absent | absent | free | free | free |
| low | | CI_bound | free | free |
| low | | CI_bound | CI_bound | free |
| low | | free | free | free |
| | absent | CI_bound | | |
| low | | | free | |
| high | | CI_bound | CI_bound | CI_bound |
| | low | | | CRO_bound |
| absent | | | | CRO_bound |
| absent | | | | CI_bound |
| present | | | | free |
| | absent | CI_bound | | |
| | absent | CRO_bound | | |
| | high | free | | |
| | high | CRO_bound | CRO_bound | CRO_bound |

Figure 8: Translation of dynamic information from a MIN to an MLM model. **Top, Left** The species $CI$ regulates the species $CRO$ through the sites $OR1, OR2$ and $OR3$. **Top, Right** The relation $\Psi_{CI,CRO}$ comprises three lines characterizing the regulation of $CRO$ by $CI$ through the regulatory site $OR1$. **Bottom** The relation $\mathcal{F}$ shows $undef$ values as white spaces.

In order to translate the information about the dynamics of the biological system, contained in $\mathcal{F}$, we need to define the *choice* operation $\sigma$, which we will call a *selection*, as presented in following definition. For each pair of variables $V_i$, $V_j$, the selection $\sigma_{V_i,V_j}(\mathcal{F})$ returns the observed system states in which both values of variables $i$ and $j$ were measured.

**Definition 16** (Selection of observed states for a pair of MIN variables) *The* selection *of observed states $\mathcal{F}$ of a biological system $\mathcal{M}$ for a pair of variables $V_i, V_j$ is the subset $\sigma_{V_i,V_j} \subseteq \mathcal{F}$ such that $\omega \in \sigma_{V_i,V_j}$ if and only if $\omega(V_i)$ and $\omega(V_j)$ are both defined.*

The selection will be used in the next definition in order to formally define the *interspecies regulation relation* $\Psi_{i,k}$, which links the values of species $i$ and $k$ which could be observed experimentally at the same time. This relation lists the values coming from $\mathcal{F}$ lines where states were observed for species $i$, species $k$ and the regulatory site $R$, influenced by $i$ and influencing $k$. That means that the interaction of species $i$ and $k$ is transmitted by the regulatory site $R$.

**Definition 17** (Interspecies regulation relation) *An* interspecies regulation relation $\Psi_{i,k} \subseteq W_{C_i} \times W_{C_k}$ *is a relation between values of the species $C_i$ and $C_k$ of a MIN $\mathcal{M}$, defined when the species $C_i$ regulates the species $C_k$ :* $\Psi_{ik} \stackrel{def}{=} \{(w_1, w_2) \mid (C_i, R, P, L) \in \mathcal{ICR}, (R, C_k, P, L) \in \mathcal{IRC}, \omega_1, \omega_2 \in \mathcal{F} : w_1 = \omega_1(C_i), \omega_1(R) = \omega_2(R), \omega_2(C_k) = w_2\}$.

Thus, the $\Psi$ relation lists the pairs of values $(w_i, w_k)$ of species $C_i$ and $C_k$ such that the value $w_i$ of the species $C_i$ and the value $w_k$ of the species $C_k$ where observed simultaneously or when the regulatory site linking them was in the same state (for an example see Figure 8).

The next definition uses the interspecies regulation relation in order to add the missing labels on the arcs of MLM regulatory graph, translated from MIN. The observed values, returned by the interspecies regulation relation, are sorted by the first value, and then the algorithm tries to fit them to a sigmoid curve, an ascendant or a descendant one. If such fitting is possible, the algorithm tries to determine the threshold for this sigmoid curve. The first fact is translated by the sign, "+" or "-", in the arc label. The threshold value is also mentioned on the corresponding arc, when found.

**Definition 18** (Translated regulatory graph) *If $\mathcal{M} = (\mathcal{V} = \mathcal{C} \cup \mathcal{R}, \mathcal{ICR}, \mathcal{IRC}, \mathcal{F}, \mathcal{L})$ is a MIN, its* translated regulatory graph *$\mathcal{G} = (U, \mathcal{E})$ (representing a set of genetic regulatory graphs) is a directed graph where:*

- *$U$ is a set of translated variables of $\mathcal{M}$;*
- *$\mathcal{E}$ is the set of arcs $(u_1, u_2)$ between variables of $U$ such that:*
  - *$(u_1, u_2) \in \mathcal{E}$ if $u_i$ is a translated variable of $C_i \in \mathcal{C}, i = 1, 2$ and $\exists ICR \in \mathcal{ICR}, \exists IRC \in \mathcal{IRC}$ such that $C_{ICR} = C_1, R_{ICR} = R = R_{IRC}$ and $C_{IRC} = C_2$. For each pair $(ICR, IRC)$ satisfying these conditions we will use the notation $(ICR + IRC) \in (u_1, u_2)$.*
  - *the arc $(u_1, u_2)$ is labeled with a set of pairs $(\theta, \epsilon)$ such that:*
    - *if $\exists w_i \in W_{C_i}, i = 1, 2, (w_1, w_2) \in \Psi_{1,2}$ such that: $\exists \Psi'_{1,2} \subseteq \Psi_{1,2} : (w_1, w_2) \in \Psi'_{1,2}$ and $\forall (w'_1, w'_2) \in \Psi'_{1,2}$, if $w'_1 \preceq_{C_1} w_1 \Rightarrow w'_2 \preceq_{C_2} w_2$ and if $w_1 \preceq_{C_1} w'_1 \Rightarrow w_2 \preceq_{C_2} w'_2$, then $(w, +)$ is in the set. (In this case $w = w_1$ is a* threshold, *and $(w_1, w_2)$ is a* positive threshold pair *of MLM interaction $(u_1, u_2)$);*
    - *if $\exists w_i \in W_{C_i}, i = 1, 2, (w_1, w_2) \in \Psi_{1,2}$ such that: $\exists \Psi'_{1,2} \subseteq \Psi_{1,2} : (w_1, w_2) \in \Psi'_{1,2}$ and $\forall (w'_1, w'_2) \in \Psi'_{1,2}$, if $w'_1 \preceq_{C_1} w_1 \Rightarrow w_2 \preceq_{C_2} w'_2$ and if $w_1 \preceq_{C_1} w'_1 \Rightarrow w'_2 \preceq_{C_2} w_2$, then $(w, -)$ is in the set. ( In this case $w = w_1$ is a* threshold, *and $(w_1, w_2)$ is a* negative threshold pair *of MLM interaction $(u_1, u_2)$);*

The translated regulatory graph $\mathcal{G}$ looks very much like a MLM model, but there are still some differences. It may contain several labels by arc, and these labels contains observed values, which are not necessary numerical ones. Thus, the next definition describes how to obtain a family of well formed MLM models from $\mathcal{G}$.

**Definition 19** *The family of labeled directed graphs compatible with the translated regulatory graph $\mathcal{G} = (U, \mathcal{E})$ is the set of graphs $G = (U, E)$ constructed in the following way:*

- *$(u, u') \in E$ iff $(u, u') \in \mathcal{E}$ and it is labeled with at most one of pairs $(\theta, \epsilon)$ from the set labelling $(u, u') \in \mathcal{E}$, if any.*
- *For each node $u$ of the so constructed translated regulatory graph, let us consider the set $\Theta_u$ of all thresholds occuring on the arcs originating from $u$. The bound $b_u$ associated to $u$ will be the $|\Theta_u| + Nua$, where $Nua$ is the number of unlabeled arcs originating from $u$. For each topological sort $(\theta_1, \ldots, \theta_{b_u})$ of $\Theta$, the numerical values $1 \leq t < b_u$ are associated to the corresponding variable values $(\theta_1, \ldots, \theta_{b_u})$, and each label $(\theta, \epsilon)$ is replaced by the corresponding $(t, \epsilon)$ in arc labels.*

- *If* $(u, u') \in \mathcal{E}$ *has an empty label,* $(u, u') \in E$ *should be labeled with* $(t, \epsilon)$ *such that* $1 \leq t < b_u$ *and* $\epsilon = +$ *or* $-$.

A state $\mu$ of such a graph $G \in \mathcal{G}$ associates then to the node $u$ a numerical value in $\{0, \ldots, b_u\}$ identifying an interval between two successive thresholds.

The MIN representation of biological systems is richer than that of MLM, already because the last does not take into account states of regulatory sites. So, several states of the MIN may be represented by only one state of the MLM. In order to establish the connection between dynamic parameters of both systems, the correspondence between states of them must be introduced: one MLM state corresponds to a domain of states in MIN.

**Notation** (Translation of system states of MIN in MLM) *If* $\mathcal{M} = (\mathcal{V}, \mathcal{ICR}, \mathcal{IRC}, \mathcal{F}, \mathcal{L})$ *is a MIN, and* $G = (U, E)$ *is one of the family of labeled directed graphs compatible with the translated regulatory graph of* $\mathcal{M}$, $\mu$ *is a state of* $G$, $\mathcal{O}_\mu$ *is the set of states* $\omega \in \Omega$ *such that* $\forall u \in U$ *if* $C \in \mathcal{C}$ *is the original species of the variable* $u$ *then* $(\mu(u) = 0 \wedge \omega(C) \preceq \theta_1) \vee (0 < \mu(u) < b_u \wedge \theta_{\mu(u)} \preceq \omega(C) \wedge \omega(C) \prec \theta_{\mu(u)+1}) \vee (\mu(u) = b_u \wedge \theta_{b_u} \preceq \omega(C))$. $\mu$ *is called the* translated state *of the* domain $\mathcal{O}_\mu$, *and* $\mathcal{O}_\mu$ *is the set of* original states *of* $\mu$.

In order to obtain the MLM translation of a MIN, we still need to define the dynamic parameters $K$ associated to the possible states of the graphs $G$ compatible with $\mathcal{G}$. The dynamic parameters for a variable are composed of observed states found in $\mathcal{F}$ at lines determined by possible values of this variable's resources.

**Definition 20** (MLM translation) *If* $\mathcal{M} = (\mathcal{V}, \mathcal{ICR}, \mathcal{IRC}, \mathcal{F}, \mathcal{L})$ *is a MIN, its* MLM translation *is a family of instances* $M = (G, K)$ *such that:*

- $G$ *is one of the family of labeled directed graphs compatible with the translated regulatory graph of* $\mathcal{M}$;
- $K = \{K_{u,\omega_u(\mu)}\}$ *are the dynamic parameters of the MLM instance* $M$ *where* $K_{u,\omega_u(\mu)}$ *is a set of observable values that the variable* $u$ *(see Definition 12), the translated variable of* $C_u \in \mathcal{C}$, *can have when the MIN state of the system* $\omega$ *is an original state of the state* $\mu$ *of* $G$*: if* $C_{u'} \in \mathcal{C}$ *is the original variable of* $u' \in G^{-1}(u), K_{u,\omega_u(\mu)} \in \cup_{u' \in G^{-1}(u)} (\cup_{\omega \in \mathcal{O}(\mu)} \Psi_{C_{u'}, C_u}(\omega))$.

Numerical values are associated to dynamic parameters using the partial order on values of the original species or other information, preserving the order obtained after the threshold ordering.

The Figure 9 illustrates the dynamic parameters translation from MIN model which is presented in Figure 4.

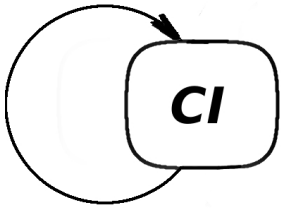
# 6 Application to the $\lambda$ phage genetic switch

**Modeling the interacting entities.** The chemical species of the model are associated to the chemically active molecules of the system: proteins $CI$ and $CRO$, which are able to bind the regulatory sites of the $\lambda$ switch. The regulatory sites named $OR1$, $OR2$ and $OR3$ can be distinguished in the regulatory region of the $\lambda$ switch. Both proteins can bind these regulatory sites. This binding capability will be represented by the affinity labeled $OR$. The regulatory sites will be labeled with the same label $OR$.

$\Psi_{CI,CI}$

| | |
|---|---|
| absent | absent |
| absent | low |
| low | absent |
| low | low |
| low | high |
| high | absent |
| high | low |
| high | high |

| $\theta$ | $\epsilon$ | threshold pairs |
|---|---|---|
| low | + | (low, low), (low, high) |
| low | − | (low, absent) |
| high | + | (high,low), (high, high) |
| high | − | (high,absent) |




| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| abs | abs | abs | abs | abs | abs | abs | abs | abs | abs | abs | abs |
| low | abs | low | abs | low | abs | low | low | **low** | **low** | low | low |
| high | abs | **high** | **low** | **high** | **high** | high | abs | high | low | high | high |
| abs | abs | abs | abs | abs | abs | abs | low | abs | low | abs | low |
| low | high | low | high | **low** | **high** | *low* | *abs* | low | abs | low | abs |
| high | abs | high | low | high | high | high | abs | high | low | high | high |
| abs | low | abs | low | abs | low | abs | low | abs | low | abs | low |
| low | low | low | low | low | low | low | high | low | high | **low** | **high** |
| *high* | *abs* | high | low | **high** | **high** | high | abs | high | low | high | high |

Figure 9: Translation of dynamic parameters from $\mathcal{F}$ to MLM. **Left** For the small network, represented on the Figure 4, the interspecies regulation relation $\Psi_{CI,CI}$ is constructed. **Right** The obtained translated regulatory graph and its labels $(\theta, \epsilon)$ with corresponding threshold pairs (shown in **bold** for positive pairs and in *italic* for negative ones in bottom tables). **Bottom** Ordering the $CI$ values as $absent \prec_{CI} low \prec_{CI} high$ enables to produce several fully ordered subset of $\Psi_{CI,CI}$.

The corresponding regulatory DNA regions $OR1, OR2$ and $OR3$, controlling the expression of $CI$ and $CRO$, are shared by two genes: $cI$ and $cro$. It means that the same regulatory site is used to control both genes, and that its state determines the activity level of both proteins simultaneously. So, the influences of $CI$ and $CRO$ on regulatory sites $OR1, OR2$ and $OR3$, and of these sites on the proteins' activity can be added into the model.

The static information about the biological system includes the information about observable values of variables. The observable states of regulatory sites $OR1, OR2$ and $OR3$ are *"CI_bound, CRO_bound"* or *"free"*. Three different observable levels of activity (concentrations) of proteins can be measured: *"absent","low","high"* for $CI$ and *"absent", "present", "high"* for $CRO$.

**Dynamics of the system.** The dynamic description of the biological system in MIN is expressed through the attributes of influences and in relation $\mathcal{F}$ (see Figure 8).

The "affinity of $CI$ for $OR1$ is tenfold higher than for $OR2$ and $OR3$" [14] can be translated in our formalism by placing the entry $(CI = low; OR1 = CI\_bound, OR2 = free, OR3 = free)$ in $\mathcal{F}$.

The property of the *cooperativity* between interacting molecules such as "$CI$ bound to $OR1$ increases the affinity of $OR2$ for another tenfold" can be represented in MIN through the refining the information about observabale states by adding the new entries $\{(CI = low, OR1 = free, OR2 = free)$ and $(CI = low, OR1 = CI\_bound; OR2 = CI\_bound)\}$ in $\mathcal{F}$.

The next type of information concerns the influence of regulatory sites on the protein activity level. The fact that the "Polymerase binding to the $CRO$ promoter is disabled if $CI$ is bound to

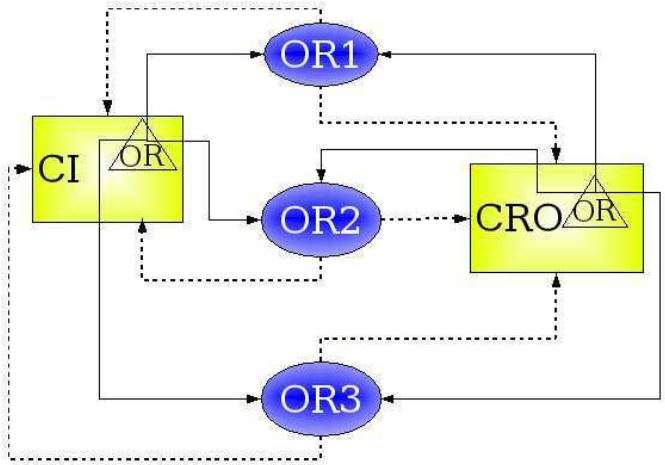
$OR1$" can be translated in our formalism by the fact that the protein $CRO$ is absent when the $OR1$ site is bound, so we add the entry $(OR1 = CI\_bound; CRO = absent)$ in $\mathcal{F}$.

In the same way the cooperativity could be represented in the expression of $CI$. Its promoter is naturally weak, but it can produce important quantities of $CI$ if the site $OR2$ is occupied. This information provides two new entries for the relation $\mathcal{F}$ : $(OR2 = free, CI = low), (OR2 = CI\_bound, CI = high)$.

The highest binding affinity of $CRO$ is for $OR3$, so that $CRO$ rapidly shuts off $CI$ production by excluding the RNA polymerase from $CI$ promoter, so, another condition for $CI$ production is that $OR3$ remains vacant. It can be represented by entries $(OR3 = CRO\_bound, CI = absent)$ and $(OR3 = free, CI = present)$ in $\mathcal{F}$.

$Pr$, the $CRO$ protein promoter, is inherently a strong one, so as soon as the site $OR1$ is vacant, $CRO$ protein is produced, which is represented in MIN by entries $(OR1 = CI\_bound, CRO = absent), (OR1 = CRO\_bound, CRO = absent)$ and $(OR1 = free, CRO = high)$ in $\mathcal{F}$.

The resulting MIN is represented in Figure 10.




Partial orders of variable values:
for $CI$ :
$absent \prec_{CI} present$,
$absent \prec_{CI} low \prec_{CI} high$;
for $CRO$ :
$absent \prec_{CRO} low \prec_{CRO} high$;
for $ORi$ :
$free$; $CI\_bound$; $CRO\_bound$.

Figure 10: A MIN representing the genetic switch of the $\lambda$ phage. Species $CRO$ and $CI$ represent proteins which bind with the affinity $OR$ to the regulatory sites $OR1, OR2$ and $OR3$. These sites are present in the regulatory regions of genes encoding both proteins, so that they influence the corresponding species $CI$ and $CRO$. The relation $\mathcal{F}$ is the same as in Figure 8.

In order to transform the MIN representation of the $\lambda$ switch in MLM we need to obtain the corresponding interaction graph and the dynamic parameters.

**Translated interaction graph.** The choice of variables of MLM is obvious: variables $CRO$ and $CI$ will represent the interacting molecular species of the MLM.

We can also follow in the MIN all described interactions between these two variables: $CI$ regulates its own expression and the expression of $CRO$ through sites $OR1, OR2$ and $OR3$. In the following, the $ICR_{iaj}$ notation means the $ICR$ from the variable $V_i$ to the variable $V_j$ of MIN through the affinity $a$, and $IRC_{ij}$ means the $IRC$ from the variable $V_j$ to $V_j$.

$$(CI, CI) = \{(ICR_{CI,OR,OR1} + IRC_{OR1,CI}), (ICR_{CI,OR,OR2} + IRC_{OR2,CI}), (ICR_{CI,OR,OR3} + IRC_{OR3,CI})\};$$

$$(CI, CRO) = \{(ICR_{CI,OR,OR1} + IRC_{OR1,CRO}), (ICR_{CI,OR,OR2} + IRC_{OR2,CRO}), (ICR_{CI,OR,OR3} + IRC_{OR3,CRO})\}.$$

$CRO$ regulates its own expression and the expression of $CI$ through the same regulatory sites:

$$(CRO, CRO) = \{(ICR_{CRO,OR,OR1} + IRC_{OR1,CRO}), (ICR_{CRO,OR,OR2} + IRC_{OR2,CRO}), (ICR_{CRO,OR,OR3} + IRC_{OR3,CRO})\};$$

$$(CRO, CI) = \{(ICR_{CRO,OR,OR1} + IRC_{OR1,CI}), (ICR_{CRO,OR,OR2} + IRC_{OR2,CI}), (ICR_{CRO,OR,OR3} + IRC_{OR3,CI})\}.$$

In order to obtain the labels of arcs of the MLM model, the corresponding $\Psi_{C_i,C_k}$ relations are calculated from the relation $\mathcal{F}$:

| $\Psi_{CI,CI}$ | $\Psi_{CI,CRO}$ | $\Psi_{CRO,CI}$ | $\Psi_{CRO,CRO}$ |
|---|---|---|---|
| (absent, absent) | (absent, absent) | (absent, absent) | (absent, absent) |
| (absent, low) | (absent, high) | (absent, low) | (absent, high) |
| (low, absent) | (low, absent) | (absent, present) | (high, absent) |
| (low, low) | (high, absent) | (absent, high) | |
| (low, high) | | (low, absent) | |
| (high, absent) | | (high, absent) | |
| (high, low) | | | |
| (high, high) | | | |

Using the Definition 18 of the translated regulatory graph, we can obtain the subsets of $\Psi_{C_i,C_k}$ relations in which the values of $C_i$ are fully ordered.

For $\Psi_{CI,CRO}$ and $\Psi_{CRO,CRO}$ two fully ordered subsets can be constructed (positive threshold pairs are shown in bold, negative threshold pairs are shown in italic):

| $\Psi^1_{CI,CRO}$ | $\Psi^2_{CI,CRO}$ |
|---|---|
| (absent, absent) | (absent, high) |
| (low, absent) | *(low, absent)* |
| (high, absent) | (high, absent) |

| $\Psi^1_{CRO,CRO}$ | $\Psi^2_{CRO,CRO}$ |
|---|---|
| (absent, absent) | (absent, high) |
| (high, absent) | *(high, absent)* |

Thus, the corresponding arcs of the translated regulatory graph will be labeled with $\theta_{CI,CRO} = low$, $\varepsilon_{CI,CRO} =$ " $-$ " and $\theta_{CRO,CRO} = high$, $\varepsilon_{CRO,CRO} =$ " $-$ ".

For the relation $\Psi_{CRO,CI}$, four fully ordered subsets can be constructed:

| $\Psi^1_{CRO,CI}$ | $\Psi^2_{CRO,CI}$ | $\Psi^3_{CRO,CI}$ | $\Psi^4_{CRO,CI}$ |
|---|---|---|---|
| (absent, absent) | (absent, present) | (absent, low) | (absent, high) |
| (low, absent) | *(low, absent)* | *(low, absent)* | *(low, absent)* |
| (high, absent) | (high, absent) | (high, absent) | (high, absent) |

Three of four cases lead to the same threshold pair, and the fourth does not have one. So, the arc $(CRO, CI)$ of the translated regulatory graph should be labeled with $\theta_{CRO,CI} = low$ and $\epsilon_{CRO,CI} = -$.

For the relation $\Psi_{CI,CI}$ 18 fully ordered subsets are possible, and they are presented in Figure 9, as well as four labels of the arc $(CI, CI)$.

Here we can take an assumption that the MLM can not distinguish between the variable values "present" and "low" and we will attribute the same numerical values to them. Replacing the MIN value *"absent"* by MLM value 0 and thresholds *"low"/"present" and "high"* by numerical values {*1 and 2* }, the family of interaction graphs of the translated MLM of the $\lambda$ switch is obtained (see the Figure 11).

**Dynamic parameters** for every instance of the obtained MLM can be derived from the relations $\Psi$ according to definition of translated parameters.

Dynamic parameters for the variable $CRO$ are the same in all three instances:

| $K_{CRO,\varnothing}$ | $K_{CRO,\{CI\}}$ | $K_{CRO,\{CRO\}}$ | $K_{CRO,\{CI,CRO\}}$ |
|---|---|---|---|
| $\Psi_{CI,CRO}(CI$ = high) $\cup$ $\Psi_{CI,CRO}(CI$ = low) $\cup$ $\Psi_{CRO,CRO}(CRO$ = high) = { absent } $\rightsquigarrow$ {0} | $\Psi_{CI,CRO}(CI$ = absent) $\cup$ $\Psi_{CRO,CRO}(CRO$ = high) = { high, absent } $\rightsquigarrow$ {0, 2} | $\Psi_{CI,CRO}(CI$ = high) $\cup$ $\Psi_{CI,CRO}(CI$ = low) $\cup$ $\Psi_{CRO,CRO}(CRO$ = absent) = { absent, high } $\rightsquigarrow$ {0,2} | $\Psi_{CI,CRO}(CI$ = absent) $\cup$ $\Psi_{CRO,CRO}(CRO$ = absent) = { high } $\rightsquigarrow$ { 2} |

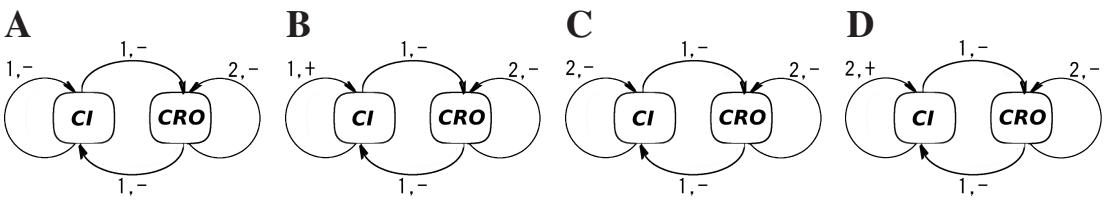
Dynamic parameters for the variable $CI$ can have different values according to the chosen MLM instance:

| | 1,- | 1,+ | 2,- | 2,+ |
|---|---|---|---|---|
| $K_{CI,\varnothing}$ | $\Psi_{CI,CI}(CI$ = low) $\cup\Psi_{CI,CI}(CI$ = high) $\cup\Psi_{CRO,CI}(CRO$ = low) $\cup$ $\Psi_{CRO,CI}$ $(CRO$ = high) = {absent, low, high } $\rightsquigarrow$ {0,1,2 } | $\Psi_{CI,CI}(CI$ = absent) $\cup\Psi_{CRO,CI}(CRO$ = low) $\cup$ $\Psi_{CRO,CI}$ $(CRO$ = high) = {absent, low} $\rightsquigarrow$ {0,1} | $\Psi_{CI,CI}(CI$ = high) $\cup\Psi_{CRO,CI}(CRO$ = low) $\cup$ $\Psi_{CRO,CI}$ $(CRO$ = high) = {absent, low, high} $\rightsquigarrow$ {0,1,2 } | $\Psi_{CI,CI}(CI$ = absent) $\cup\Psi_{CI,CI}(CI$ = low) $\cup\Psi_{CRO,CI}(CRO$ = low) $\cup\Psi_{CRO,CI}$ $(CRO$ = high) = {absent, low, high} $\rightsquigarrow$ {0,1,2 } |
| $K_{CI,\{CI\}}$ | $\Psi_{CI,CI}(CI$ = absent) $\cup\Psi_{CRO,CI}(CRO$ = low) $\cup$ $\Psi_{CRO,CI}$ $(CRO$ = high) = {absent, low} $\rightsquigarrow$ {0,1} | $\Psi_{CI,CI}(CI$ = low) $\cup\Psi_{CI,CI}(CI$ = high) $\cup\Psi_{CRO,CI}(CRO$ = low) $\cup$ $\Psi_{CRO,CI}$ $(CRO$ = high) = {absent, low, high} $\rightsquigarrow$ {0,1,2} | $\Psi_{CI,CI}(CI$ = absent) $\cup\Psi_{CI,CI}(CI$ = low) $\cup\Psi_{CRO,CI}(CRO$ = low)$\cup\Psi_{CRO,CI}$ $(CRO$ = high) = {absent, low, high} $\rightsquigarrow$ {0,1,2 } | $\Psi_{CI,CI}(CI$ = high) $\cup\Psi_{CRO,CI}(CRO$ = low) $\cup$ $\Psi_{CRO,CI}$ $(CRO$ = high) = {absent, low, high} $\rightsquigarrow$ {0,1,2} |
| $K_{CI,\{CRO\}}$ | $\Psi_{CI,CI}(CI$ = low) $\cup\Psi_{CI,CI}(CI$ = high) $\cup\Psi_{CRO,CI}(CRO$ = absent) = {absent, low, high} $\rightsquigarrow$ {0,1,2} | $\Psi_{CI,CI}(CI$ = absent) $\cup\Psi_{CRO,CI}(CRO$ = absent) = {absent, low} $\rightsquigarrow$ {0,1} | $\Psi_{CI,CI}(CI$ = high) $\cup\Psi_{CRO,CI}(CRO$ = absent) = {absent, low, high} $\rightsquigarrow$ {0,1,2} | $\Psi_{CI,CI}(CI$ = absent) $\cup\Psi_{CI,CI}(CI$ = low) $\cup\Psi_{CRO,CI}(CRO$ = absent) = {absent, low, high} $\rightsquigarrow$ {0,1,2} |
| $K_{CI,\{CI,CRO\}}$ | $\Psi_{CI,CI}(CI$ = low) $\cup\Psi_{CI,CI}(CI$ = high) $\cup\Psi_{CRO,CI}(CRO$ = low) $\cup$ $\Psi_{CRO,CI}$ $(CRO$ = high) = {absent, low} $\rightsquigarrow$ {0,1} | $\Psi_{CI,CI}(CI$ = low) $\cup\Psi_{CI,CI}(CI$ = high) $\cup\Psi_{CRO,CI}(CRO$ = absent) = {absent, low, high} $\rightsquigarrow$ {0,1,2} | $\Psi_{CI,CI}(CI$ = absent) $\cup\Psi_{CI,CI}(CI$ = low) $\cup\Psi_{CRO,CI}(CRO$ = absent) = {absent, low, high} $\rightsquigarrow$ {0,1,2} | $\Psi_{CI,CI}(CI$ = high) $\cup\Psi_{CRO,CI}(CRO$ = absent) = {absent, low, high} $\rightsquigarrow$ {0,1,2} |

This example illustrates the construction of the MIN model from the biological data and shows that this model can be automatically translated in the MLM formalism. In the worst case, the interaction graph of the MLM is constructed from the MIN representation, but no constraint is found on the dynamic parameters (as for parameters $K_{CI,\omega_\mu}$ in networks C and D, Figure 11). In the best case, only one value for each dynamic paramter will be produced (as for $K_{CI,\{CRO\}}$).

# 7 From MIN to ODEs

An important part of the biological knowledge comes from biochemistry. It covers information about the dynamics of *chemical reactions*, which are treated in the *in silico* models through the device of ordinary differential equations (ODEs).




$$K_{CRO,\varnothing} = 0, K_{CRO,\{CI\}} \in \{0,2\}, K_{CRO,\{CRO\}} \in \{0,2\}, K_{CRO,\{CI,CRO\}} = 2.$$

| | MLM **A** | MLM **B** | MLM **C** | MLM **D** |
|---|---|---|---|---|
| $K_{CI,\varnothing}$ | {0,1,2} | {0,1} | {0,1,2} | {0,1,2} |
| $K_{CI,\{CI\}}$ | {0,1} | {0,1,2} | {0,1,2} | {0,1,2} |
| $K_{CI,\{CRO\}}$ | {0,1,2} | {0,1} | {0,1,2} | {0,1,2} |
| $K_{CI,\{CI,CRO\}}$ | {0,1 } | {0,1,2} | {0,1,2} | {0,1,2} |

Figure 11: A translation of a MIN from Figure 10 into MLM. The variables $CI$ and $CRO$ of the MLM are obtained from the species $CI$ and $CRO$ of the MIN combined with the regulatory sites $OR1$, $OR2$ and $OR3$. The MLM interactions are obtained from pairs $(ICR + IRC)$ present in the MIN. For example, there is an arc $(CI, CRO)$ in the MLM because there is a pair $(ICR + IRC) = (CI, CRO)$ in the MIN presented in Figure 10. The dynamic parameters and arc labels of the MLM are calculated from the relation $\mathcal{F}$ of the MIN.

Differential equations aim at expressing the concentration of a chemical species as a function of time, knowing its production and degradation rates:

$$[\dot{P}] = \frac{d[P]}{dt} = \sum_i k_i \prod_j [S_{ij}]^{\alpha_{ij}} - \sum_l k_l \prod_j [S_{lj}]^{\alpha_{lj}}$$

where $k_i$ is the reaction rate for the $i$-th $P$-production chemical reaction, $\alpha_{ij}$ is the stoichiometric coefficient of the $j$-th substrate in this reaction, $S_{ij}$ is this substrate, $[S_{ij}]$ is the concentration of the latter, and $k_l$, $\alpha_{lj}$, $[S_{lj}]$ denote the corresponding elements for the $l$-th $P$-degradation reaction and its co-substrates.

In order to translate the MIN model in ODEs, we need to write the set of chemical reactions in the biological system, and to deduce (if possible) the reaction rates from the parameters of the influences of the MIN model. In a case where the mechanism of the reaction is unknown, it may be written in Michaelis-Menten form: $S \xrightarrow{E} P$, where $E$ is an enzyme catalyzing the reaction but not consumed in it. The translation of this reaction into differential equations is a known issue.

A MIN model detailed enough to be directly translated to ODEs is presented in Figure 5. For each chemical species in Figure 5 we can write a differential equation summing its consumption and production in chemical reactions the species is participating (see Figure 12). The stoichiometric coefficients give the $\alpha_i$ power coefficients in the formula, and the $k_j$ reaction rates come form the corresponding reaction attributes.

For example, in the third equation describing the production of the CI RNA from nucleotides, $CI\_RNA$ corresponds to the quantities of each of the four nucleotides composing the CI RNA: A, U, C and G (the last one, T, being absent from the RNAs). The RNA polymerase ($RNA\_pol$ in Figure 5) is the enzyme which catalyzes the CI RNA synthesis without being consumed in this reaction, so its concentration influences the reaction rate $k_{CI\_RNA\_synth}$ and it is taken into account in the function $f$. $OR1 \cdot CI_2$ stands for the DNA information source for the CI RNA synthesis, and it acts also as a catalyzer: without this species the CI RNA synthesis is impossible. One molecule of $CI\_RNA$ species is produced from all the necessary nucleotides on the matrix $OR1 \cdot CI_2$ and under the action of the $RNA\_pol$. The first equation describes the concentration

$$
\begin{cases}
\frac{d[CI_2]}{dt} & = k_{CI\_dimerisation}[CI]^2 - k'_{CI\_dimerisation}[CI_2] \\
\frac{d[OR1 \cdot CI_2]}{dt} & = k_{OR1\_binding}[CI_2][OR1] - k'_{OR1\_binding}[OR1 \cdot CI_2] \\
\frac{d[CI\_RNA]}{dt} & = k_{CI\_RNA\_synth}[Nucleotides], \\
 & \text{where } \; k_{CI\_RNA\_synth} = f([RNA\_pol], [OR1 \cdot CI_2]) \\
\frac{d[CI]}{dt} & = k'_{CI\_dimerisation}[CI_2] - k_{CI\_dimerisation}[CI]^2 + k_{CI\_synth}[Aminoacids], \\
 & \text{where } \; k_{CI\_synth} = g([Ribosome], [CI\_RNA]) \\
\frac{d[Ribosome]}{dt} & = 0 \\
\frac{d[RNA\_pol]}{dt} & = 0 \\
\frac{d[Nucleotides]}{dt} & = -k_{CI\_RNA\_synth}[Nucleotides] \\
\frac{d[Aminoacids]}{dt} & = -k_{CI\_synth}[Aminoacids]
\end{cases}
$$

Figure 12: Differential equations obtained by an automatic translation of the MIN model in Figure 5. Functions $f$ and $g$ come, on one hand, from the MIN topology and the information on the stoichiometry of the reaction, and on the other hand, from the reaction attribute. At this stage, the coherence of both informations should be checked by an expert. In these equations $f$ and $g$ have a definite signature reflecting the impact of the catalyzers and inhibitors on the reactions.

of the CI protein dimer $CI_2$. The right part represents the synthesis of one molecule of $CI_2$ from 2 molecules of $CI$ (first term) minus the dissociation of the $CI_2$ species on 2 $CI$ proteins (second term).

More generally, any MIN model can be translated into differential equations with an automated procedure, even if it was not explicitly constructed to represent a set of biochemical reactions. In some cases, it may be necessary to first demultiply MIN regulatory sites in order to translate the model directly as for the example in Figure 5.

While the states of a chemical species may characterize the degree of its activity, through a discrete indication like “absent”, “low”, “high”, or through a quantitative information like the concentration, leading quite directly to a representation in ODEs, the states of a regulatory site may potentially be more difficult to interpret. In the simplest case a regulatory site represents a single chemical reaction. The regulatory sites modeling to single chemical reactions, like “CI RNA synthesis”, “CI protein synthesis” or “CI dimerisation” in Figure 5, correspond to such a situation, and are easy to translate in ODEs.

However, in a more complex case, a regulatory site may encompass through its different states a family of biochemical reactions, making a direct translation difficult. Actually, the concentrations of participating species for a single chemical reaction are sufficient to find out its activity rate, thus represented by a function. For a family of reactions, the reaction rate is not always a function (but a relation) of the concentrations of each species, and this is precisely the difficulty of the translation to ODEs.

Let us consider the example in Figure 13. The MIN model looks very much like the one in Figure 4, but the IRC and ICR are provided with additional properties such as $k_i, K_{aff}$ and $production\_rate$ which reflect the kinetic properties of the corresponding biochemical reactions. If the regulatory site “OR1” in Figure 13 is in the state $OR1\cdot$, it means that neither of

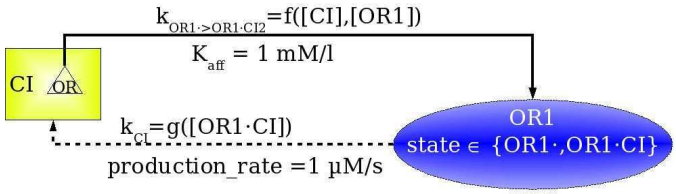



Figure 13: The same MIN model as the one used for genetic regulation modeling, enriched with complementary information allowing the translation into differential equations.

the two reactions ("CI RNA synthesis" and "CI protein synthesis") take place in the cell. When the same site is in the state $OR1 \cdot CI$, it means that both "CI RNA synthesis" and "CI protein synthesis" take place. Thus, it is possible to reduce this complexity by *demultiplicating* the regulatory sites as a first step of the translation of a MIN model in ODEs. The *demultiplication* of a regulatory site $R$ replaces it by a set of (new) species associated to the states of $R$ and a set of (new) regulatory sites associated to the chemical reactions. In other words, every regulatory state of $R$ will now give a chemical species participating in a defined set of chemical reactions, represented by newly generated regulatory sites. After the demultiplication, each regulatory site represents a single chemical reaction, which means that the species connected to it may potentially be produced or consumed, and may be automatically translated to ODEs. Some optimizations may be performed at this stage, for instance, if one knows if the species are consumed or produced, which may be indicated in the attributes (such as "stoichiometry", "production rate", "degradation rate" or "kinetic rate") of the corresponding influences ICRs and IRCs.

# 8 Conclusion and discussion

The MIN representation proposes a rich formal description of biological interaction networks. The methodology of modelling biological systems in an incremental MIN representation is illustrated by a case study on the $\lambda$ switch system. The formalisation of biological data is independent of any given modeling or simulation approach. The main goal of MIN is to contain as many different data about interacting entities as possible in order to make them accessible to any particular modeling approach. A translation into R. Thomas' formalism allows the modeler to obtain an MLM model from the available data, and the MLM is consistent with other models of the same system [22]. While the translation from MIN into MLM is rather complicated, it can be easily automated using the algorithm presented in this paper. However, without the expert intervention, the number of MLM models can be high. The modeler can act on the data put into the MIN model, changing and refining it, and this change will have an impact on the produced MLM translated models. However, there is no need for an expert to deeply understand the algorithm itself. The translation of MLM instances can be further continued into Petri nets as studied in [3] and, thus, provides an access to the available Petri net tools for analysis. Each formalism has its advantages and fits the description of a certain data type, the complete and efficient description of biological systems is possible only by combining these tools. A formalism forces an interpretation of available data in order to fit them in its framework. Some data which are incompatible with the chosen framework will inevitably be lost. Sometimes the same model represented in different formalisms can hardly be recognized [4, 14, 21, 8].

The representation of regulatory sites and affinities separately from chemical species helps to represent in a "formal" way large proteins with many functional domains, or a complex set of

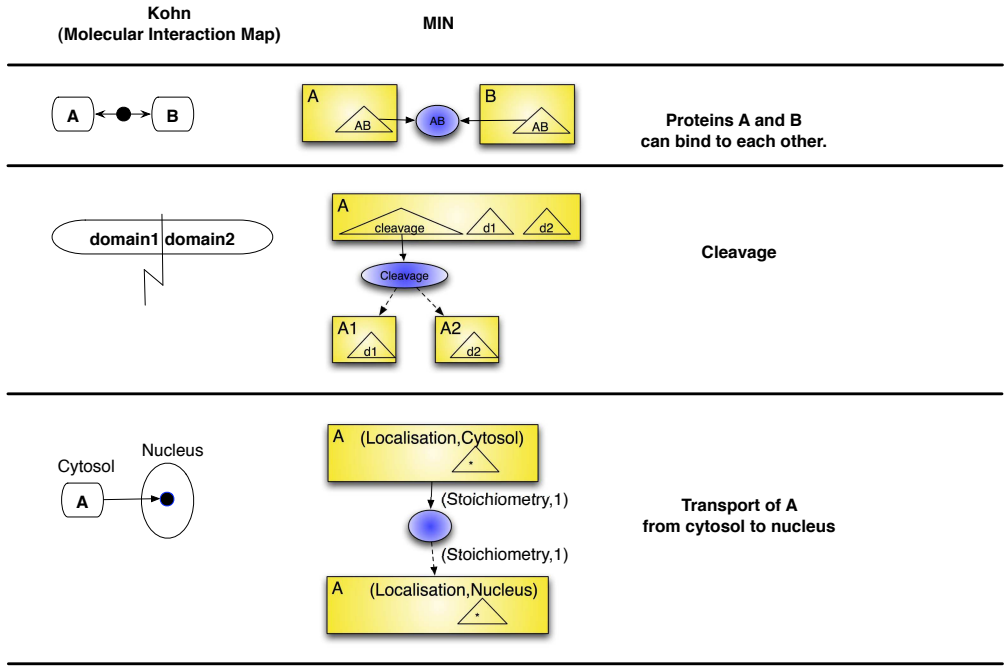



Figure 14: Examples of Kohn Maps building blocks and their MIN representations.

regulatory sites in a protein or in a gene. The specificity of the $\lambda$ phage genetic switch is that the promoter region of two different genes is represented by the same biological object (DNA region). This fact is represented in our formalism by having only one set of regulatory sites of the $\lambda$ switch which influence two different species: $CI$ and $CRO$.

MIN enables an incremental model construction through the composition of MINs and the storage (in the species affinities and regulatory site labels) of the information about possible interaction capabilities of biological entities. Thus, MIN can help in the model construction by a rational choice of new variables to be added to the model: with compatible regulatory sites or affinities.

Experimental techniques in biology collect massive amounts of information on the behavior and interaction of thousands of genes and proteins across diverse conditions. These techniques are used to question complex biological systems that use highly intricate regulatory mechanisms and control schemes. One cannot fully characterize such complex cellular systems by focusing on a single control mechanism, as measured by a single experimental technique. In MIN, the data coming from different experimental techniques are all stored in $\mathcal{F}$. To gain a deeper understanding of the system, it is pertinent to analyze heterogeneous data sources in a truly integrated fashion and to shape the analysis results into one body of knowledge [2, 20].

We proposed a new paradigm for the modeling of biological systems, in which all available experimental data are considered as a set of snapshots of the real system and stored in $\mathcal{F}$ without any interpretation. The information about the system is added and refined incrementally. The current state of knowledge in MIN can be automatically translated into a given formalism framework for the analysis of the dynamics of the system; it could also be used in the future by an inference system applying artificial intelligence techniques [9] to solve complex biological problems.

Over the last few years, some work has been carried out in the field of integration of biological and, in particular, biochemical data which includes rich but informal visualisation conventions

[11, 16]. Even if MIN is not designed as a graphical model, it provides a quite simple visualisation convention with two types of nodes and two types of links. However, combined with textual information encoded in the attributes of links and nodes, it can represent biological features encoded as Kohn Maps [11], as it is illustrated for three examples of Kohn Maps building blocks in Figure 14.

Recently, a method for representing and communicating biological networks in both human and machine readable form has been presented in [10]. The ambition of this work is obtaining a semantically and visually unambiguous diagram scheme, but this leads to a very low level representation of processes and the use of many kinds of nodes and links. Compared to this, MIN does not require an equivalent degree of details and enables to adjust the abstraction level of the model. Another approach, based on formal but not very expressive exchange formalisms, like SBML [6], attempts to standardize the expression of ODE based models of cellular systems, concentrating on chemical reactions. Obviously, existing SBML models can be wrapped in a MIN description. In the same standardisation effort more abstract and universal meta-modelling approaches [1, 19, 18, 12] tend to create a general visual language for systems biology, similar to UML. For instance, BioUML [12] provides an abstract layer to present structure of any biological system as a clustered graph. MIN should be expressed in this language to use the infrastructure based on BioUML, to access to the biological databases and to automatically generate the executable models.

Thus, the proposed new formalism, MIN, can play the role of an intermediate level between insufficiently formalized "natural language" and too specialized "mathematical descriptions" of biological systems. The MIN construction is a process of inference of the biological interaction networks from the biological observations of microscopic and macroscopic levels. Its underlying structure provides a skeleton for the understanding of "first principles" of the organisation of biological systems. A computer analysis tool to study the properties of MIN models, to perform automatically their composition and translation into different formalisms, is currently under developed and should soon become available for download.

## Acknowledgments

This work was supported by Genopole in Évry (France) and the ISI Foundation (Turin, Italy). Thanks to Sorin Solomon and anonymous referees for numerous and very useful remarks.

## References

[1] M. Beurton-Aimar, S. Pérès, N. Parisey, C. Nazaret, and J.P. Mazat. Modeling biologic networks to use them with heterogeneous treatments. In *Proceedings of Ecole Thematique "Modélisation et simulation de processus biologiques dans le contexte de la génomique - 2003 - Dieppe(France)"*, 2003.

[2] L. Cardelli. Abstract machines of systems biology. In *Transactions on Computational Systems Biology III*, Springer LNBI 3737, pages 145–168, 2005.

[3] C. Chaouiya, E. Remy, and D. Thieffry. Petri net modelling of biologycal regulatory networks. In Gordon Plotkin (Ed.), editor, *Third International Workshop on Computational Methods in Systems Biology, University of Edinburgh, 2-10 April 2005*, 2005.

[4] A. Doi, H. Matsuno, and S. Miyano. Induction mechanism description of lambda phage by hybrid Petri net. *Currents in Computational Molecular Biology*, pages 26–27, 2000.

[5] H. Eisen, P. Brachet, L. Pereira da Silva, and F. Jacob. Regulation of repressor expression in $\lambda$. *Proc. natn. Acad. Sci.*, 66:855–862, 1970.

[6] M. Hucka et al. The systems biology markup language (sbml): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[7] J. Guespin-Michel, G. Bernot, J.-P. Comet, A. Mérieau, A. Richard, C. Hulen, and B. Polack. Epigenesis and dynamic similarity in two regulatory networks in pseudomonas aeruginosa. *Acta Biotheoretica*, 52(4):379–390, 2004.

[8] K.R. Heidtke and S. Schulze-Kremer. Design and implementation of a qualitative simulation model of lambda phage infection. *Bioinformatics*, 14(1):81–91, 1998.

[9] J. Keppens and Qiang Shen. On compositional modelling. *The knowledge engineering review*, 16:157–200, 2001.

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

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

[12] F.A. Kolpakov. BIOUML - framework for visual modeling and simulation biological systems. In *Proc. Int. Conf. Bioinf. of Genome Regulation and Structure (BGRS'2002)*, 2002.

[13] H. Kurata, N. Matoba, and N. Shimizu. CADLIVE for constructiong a large-scale biochemical network based on a simulation-directed notation and its application to yeast cell cycle. *Nucleic Acid Res.*, (31), 2003.

[14] C. Kuttler, J. Niehren, and R. Blossey. Gene regulation in the $\pi$-calculus: Simulating cooperativity at the lambda switch. *Bio-CONCUR*, 2004.

[15] H Matsuno, A Doi, M Nagasaki, and S Miyano. Hybrid Petri net representation of gene regulatory network. *Pac Symp Biocomput*, pages 341–52, 2000.

[16] I. Pirson, N. Fortemaison, C. Jacobs, S. Dremier, J.E. Dumont, and C. Maenhaut. The visual display of regulatory information and networks. *Trends in Cell Biology*, 10(10):404–408, 2000.

[17] M. Ptashne. *A Genetic switch*. Blackwell Science (ISBN : 978-0865422094), 1992.

[18] M. Roux-Rouquié, N. Caritey, L. Gaubert, B. Le Grand, and M. Soto. Metamodel and modeling language: towards an Unified Modeling Language (UML) profile for systems biology. In *Object-oriented Modeling in Biology and Medecine, SCI2005*, 2005.

[19] M. Roux-Rouquié and M. Soto. Virtualization in systems biology: Metamodels and modeling languages for semantic data integration. *T. Comp. Sys. Biology*, 1:28–43, 2005.

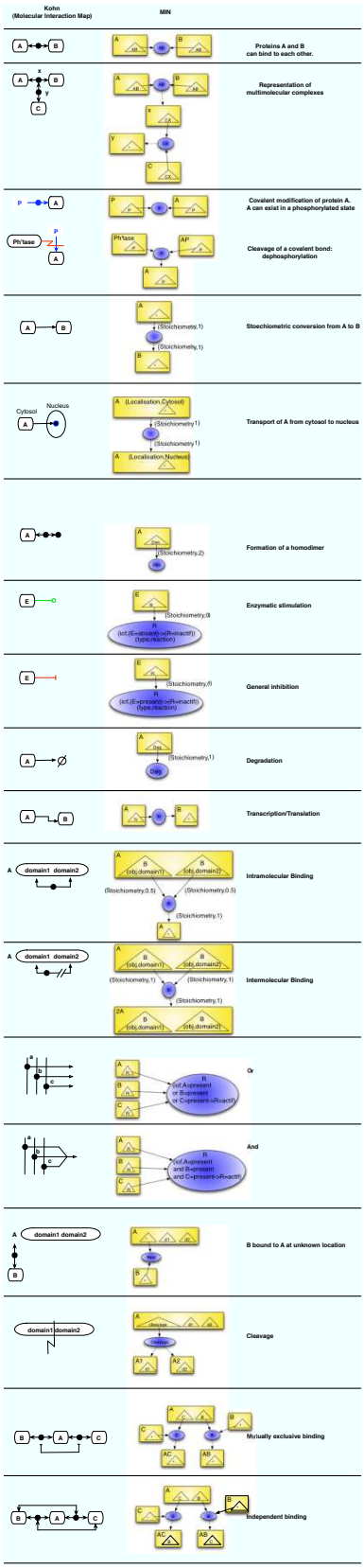
[20] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A.*, 101(9):2981–6, 2004.

[21] D. Thieffry and R. Thomas. Dynamical behaviour of biological regulatory networks - II. immunity control in bacteriophage lambda. *Bull. Math. Biol.*, 57(2):277–297, 1995.

[22] R. Thomas. Regulation of gene expression in bacteriophage $\lambda$. *Current Topics in Microbiology and Immunology*, 56:13–42, 1971.

[23] R Thomas. Regulatory networks seen as asynchronous automata : A logical description. *J. theor. Biol.*, 153, 1991.

[24] R. Thomas, A.M. Gathoye, and L. Lambert. A complex control circuit. regulation of immunity in temperate bacteriophages. *Eur. J. Biochem.*, 71(1):211–227, 1976.



## 5.3 Du modèle MIN aux équations différentielles ordinaires

# From MIN model to ordinary differential equations

**A. Yartseva[1] , R. Devillers [2], H. Klaudel [1], F. Képès [3]**

[1]IBISC - CNRS, Université d'Evry, 523 place des Terrasses de l'Agora, 91000 Evry, France

[2]Département d'Informatique, Université Libre de Bruxelles, CP212, B-1050 Bruxelles, Belgium

[3]Epigenomics Project, Genopole, CNRS & Université d'Evry, 523 place des Terrasses de l'Agora, 91000 Evry, France

**Summary**

Biological interaction networks can be modeled using the Modular Interaction Network (MIN) formalism, which provides an intermediary modeling level between the biological and mathematical ones. MIN focuses on a simple but structured and versatile representation of biological knowledge, without targeting a particular analysis or simulation technique. In this paper, we propose a translation procedure which, starting from a MIN specification of a biological system, generates its representation in ordinary differential equations (ODEs) allowing to study the dynamics of the system. The translation is illustrated on a classical benchmark: the $\lambda$ phage genetic switch.

**Keywords.** Abstract biological models, regulatory interaction networks, ODE

## 1 Introduction

The description of a biological system is often obtained by constructing an interaction network. An efficient way to represent such an interaction network is to use the Modular Interaction Network (MIN) formalism [15], which provides an intermediary modeling level between the biological and mathematical ones. MIN was designed in order to provide a structured way to maintain various biological data, taking into account their interactions, supporting incremental enrichments and several translation procedures to other formalisms currently used by modelers in biology. The translation from MIN to target modeling formalisms is crucial as it gives an access to analysis or simulation techniques allowing in particular to study the dynamics of the biological system. This has already been detailed in [15] for the R. Thomas' regulatory networks formalism [13].

In this paper, we address specifically the translation procedure which, starting from a MIN specification of a biological system, generates automatically its representation in ordinary differential equations (ODEs). This translation can be performed either directly (if some specific conditions are satisfied), or after applying an auxiliary operation of regulatory site demultiplication allowing to handle the necessary information automatically in an exhaustive way. The translation is illustrated on a classical benchmark: the $\lambda$ phage genetic switch.

The paper is structured as follows. The next section recalls MIN. Then, we present two examples of the $\lambda$ phage modeling with MIN. Our translation of MIN into ODE is introduced next and applied to those examples. Finally, we conclude with some words of discussion, related work and perspectives.

## 2 Modular Interaction Network

The MIN model can be seen abstractly as a bipartite graph involving two kinds of nodes: chemical species and regulatory sites. Every regulatory site has a set of regulating and regulated chemical species and their role is expressed by influences. Chemical species and regulatory sites together are called variables. They represent biological objects at some level of abstraction: molecules or parts of them, complex processes like regulatory pathways, complex systems like sensors, or even an entire organism.

As the knowledge about biological systems is based on observations and experiments, the observable level of activity of biological objects can change in various states of the biological system. These objects can influence the levels of activity of each other. So, every variable in MIN is assumed to have a set of observable values, corresponding to the observable levels of activity of the corresponding biological objects, such as "low", "high", or "$10\mu M$".

A chemical species represents a biological object with catalytic or binding capabilities, which can influence one or more regulatory sites. These influences have a chemical nature: association/dissociation reactions, electron transfers, etc. A species may have one or more influence capabilities, which are called affinities. An affinity is the ability of a biological object to interact with a set of other biological objects through a particular regulatory site. Thus, an affinity may correspond to a protein domain for a protein or to a surface molecule (receptor) for a cell. The nature of the interaction between two biological entities can be unknown. So, a wild-card affinity, labeled "*", may be defined for every species, standing for an unknown mechanism of regulation.

A regulatory site regulates species activity in a manner which may be assimilated to a chemical reaction or to a more abstract mechanism, like for instance three-dimensional conformation changes in a molecule or cooperativity effects. A regulatory site has a label which characterizes its capabilities of being influenced through the affinities. If a regulatory site and an affinity of a species have the same label, it means that an interaction is possible between the biological objects corresponding to the site and the species. A regulatory site represents an "input" for a species and regulates its activity through the integration of several influences on it.

The variables (chemical species and regulatory sites) can have attributes, which come from the corresponding biological objects, and may have types like "position", "size", "reaction rate", "stoichiometry" etc. expressing a knowledge about them. Several variables with the same name may thus be present in MIN, if they have attributes with different values. So, we can represent a molecule of the same protein in free or bound state, or the same gene at its natural location and translocated in a different place in the genome.

Biological objects, represented by variables in MIN, may interact and play specific roles in these interactions. It is assumed that every interaction happens through an affinity and a regulatory site and there is no influence between variables of the same kind. Thus, two kinds of influences between the variables of the model can be considered: Influences of Chemical species on Regulatory sites (ICR) and Influences of Regulatory sites on Chemical species (IRC). An influence has also a set of attributes, denoted by $P_{ICR}$ or $P_{IRC}$, which describes, in particular, the relationship between the values of the species and those of the regulatory site, like the parameters of the corresponding chemical reaction: kinetic rate, speed, ...

The dynamics of the biological system is represented in MIN by "snapshots", lines in a relation

$\mathcal{F}$. Each such line collects the measurement results for a certain number of observed variables (and '*undef*" for the others). $\mathcal{F}$ plays the role of a data bank from which the parameters of the dynamics of the system interactions could be inferred, if not yielded by parameters in $P_{ICR}$ or $P_{IRC}$.

More formally, a modular interaction network $\mathcal{M}$ is a tuple $(\mathcal{V}, \mathcal{ICR}, \mathcal{IRC}, \mathcal{F}, \mathcal{L})$ where:

- $\mathcal{V} = \mathcal{C} \cup \mathcal{R}$ is the set of variables of the model; it is partitioned in a set $\mathcal{C} = \{C_i \mid 1 \leq i \leq |\mathcal{C}|\}$ of chemical species and a set $\mathcal{R} = \{R_j \mid 1 \leq j \leq |\mathcal{R}|\}$ of regulatory sites; the name of a variable $v$ is denoted by $N_v$;
- $\mathcal{ICR}$ is a set of influences from chemical species to regulatory sites through an affinity of the former and there is at most one influence between such a pair of variables through the same affinity;
- $\mathcal{IRC}$ is a set of influences from regulatory sites to chemical species and there is at most one influence between such a pair of variables;
- $\mathcal{F}$ is a set of observed (possibly partly[1] defined) states of the biological system;
- $\mathcal{L}$ is a set of links to sources of the information (bibliography) about those observations.

Such MIN models may be composed and compressed using dedicated operations allowing to assemble incrementally and/or separately various representations of a studied biological system.

In figures, species are represented by boxes, affinities by triangles inside the boxes of species, regulatory sites by ellipses, influences of a species on a regulatory site by plain arcs, and influences of a regulatory site on a species by dashed arcs, as shown in Figures 1 or 2.

## 3 The $\lambda$ phage genetic switch and its modeling with MIN

In order to illustrate our approach, we shall use as a running example a classical biological benchmark: the genetic switch of the $\lambda$ phage. The $\lambda$ phage is a virus which infects the *Escherichia coli* bacteria. It turns out that a lot of quantitative and qualitative information is now available on it, so that it has become a benchmark organism and plays a central role in modeling [10, 7, 13, 14, 9, 4, 3, 8]. The decision between two possible (lytic or lysogenic) life phases is controlled by a region of the $\lambda$ phage genome, referred to as the genetic switch region. The decision results from the competition between two major proteins: The first one is referred to as CRO, encoded by gene cro, and expressed during the lytic phase. The second one is called $\lambda$ repressor, referred to as CI. It is encoded by gene cI, and it can activate other genes, including itself, and repress others. The gene cI is expressed during the lysogenic phase.

Various MIN models may be given for a same biological system, corresponding to various levels of abstraction or emphasizing particular aspects of it.

In figures and in the following the italic characters are used for the MIN model entities, while the ordinary roman ones for the biological objects.

[1] Some values may be "*undef*".

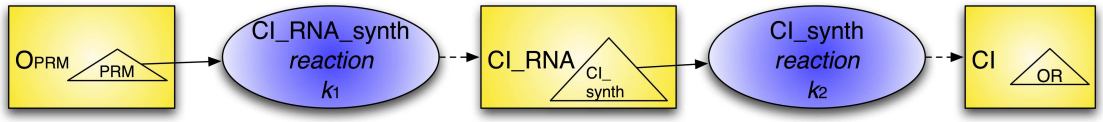



**Figure 1: A MIN model representing the CI synthesis. The regulatory sites $CI\_RNA\_synth$ and $CI\_synth$ represent non reversible reactions of CI RNA synthesis and of CI protein synthesis. They have the attributes $k_1$ and $k_2$, respectively, which represent the reaction rates measured in or calculated from biological experiments. The ICRs coming out from the species $O_{PRM}$ and $CI\_RNA$ have the attributes $stoichiometry = M$ (not shown in this figure), meaning that these species are not consumed in these reactions: they are biological matrices, or "templates" for the macromolecular synthesis.**

Figure 1 shows a possible MIN model for the CI protein synthesis from the $O_{PRM}$ promoter. Its particularity is to represent explicit biochemical reactions. The macromolecules CI RNA and CI protein are represented by corresponding MIN chemical species, and the promoter and the adjacent CI gene are represented by $O_{PRM}$. The biochemical reactions of this example are represented by regulatory sites. These reactions regulate the level of activity of a chemical species by increasing or decreasing its quantity (concentration). Each reaction possibly has an attribute "reversible" (otherwise it is non reversible). A regulatory site representing a chemical reaction cannot have observed states, since there is no experimental way in biology to observe a process; we can only follow the state change of chemical species involved in this process, such as their concentration. Another attribute of the regulatory site is a kinetic rate, which is in general a function of other parameters of the system such as concentrations of species catalyzing the reaction (enzymes) or even non participating directly in the reaction but influencing its kinetics. For example, such species can sequestrate one or more substrates or products, or catalyze intermediate reaction steps.

On the contrary to the previous example, the MIN, presented in Figure 2, contains implicit description of biochemical reactions. It describes the CI production and its regulation. Three chemical species are presented: the protein $CI$, its' gene $O_{PRM}$ and the CI protein dimer $CI2$. Two regulatory sites, $CI\_synth$ and $CI\_dim$, representing the CI protein synthesis and the CI protein dimerisation, are indicated to be reactions. The third regulatory site, $OR$, represents the regulatory region of the $\lambda$ phage DNA, and not a simple chemical reaction.

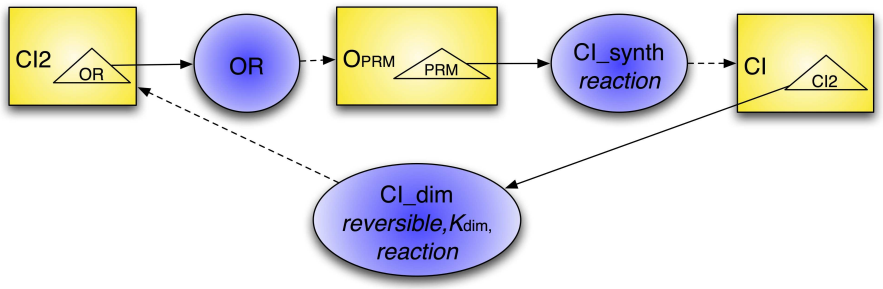



**Figure 2: A MIN model representing the CI protein synthesis from the CI gene $O_{PRM}$ together with the CI protein dimerisation and the regulation of the CI gene $O_{PRM}$ by the CI dimer $CI2$ through the regulatory site $OR$.**



Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

# 4 Translation of a MIN into ordinary differential equations (ODEs)

The ODEs are one of the most traditional mathematical approaches to modeling of biological systems, essentially because they may easily be simulated using any of the numerical integration tools. While the usual approach to construct an ODE model is to collect the needed information from literature piece by piece, which is extensively time consuming, the MIN model gathers various types of data about the structure and functioning of living systems which may be automatically translated into various modeling formalisms including ODEs. In order to perform the translation into ODEs, the chemical species implicated in reactions and the kinetic properties of these reactions should be indicated. The translation from MIN into ODEs is performed either directly (if some specific conditions are satisfied), or after applying an auxiliary operation of regulatory site demultiplication allowing to handle the necessary information automatically in an exhaustive way.

While the states of a chemical species may characterize the degree of its activity, through a discrete indication like “absent”, “low”, “high”, or through a quantitative information like the concentration, leading quite directly to a representation in ODEs as chemical species, the states of a regulatory site may potentially be more difficult to interpret. In the simplest case a regulatory site represents a single chemical reaction, like “CI RNA synthesis”, “CI protein synthesis” or “CI dimerisation” in Figures 1 and 2, are easy to translate in ODEs using the mass action law. However, in a more complex case, a regulatory site may encompass through its different states a whole family of biochemical reactions, making a direct translation difficult. Actually, the concentrations of participating species for a single chemical reaction are sufficient to find out its activity rate, thus represented by a function. For a family of reactions, the reaction rate is not always a function (but a relation) of the concentrations of each species, and this is precisely the difficulty of the translation to ODEs.

## 4.1 Direct translation from MIN into ODEs

A MIN model can be directly translated in ODEs when each regulatory site corresponds to a single chemical reaction (it has the attribute "$reaction$") which consumes no more than two molecules. This last constraint comes from the hypothesis commonly used in ODE modeling, that it is highly unlikely for more than two molecules to meet and to react, simultaneously. An obvious exception to this rule is the case of enzymatic reactions, often represented with more than two molecules participating in the chemical reaction, one of them being an enzyme. In fact, enzymes are most commonly presented on the chemical reaction arrows to say that they influence the reaction kinetics, but their quantity does not change in it. We consider that a regulatory site corresponding to any other type of reaction (representing more than one simple reaction step, and thus involving more than two species) should be transformed (demultiplied) first, in order to be translated into differential equations.

A MIN model like that presented in Figure 1 is detailed enough to be directly translated to ODEs. Indeed, each regulatory site corresponds to a simple reaction. For each chemical species we can thus write a differential equation summing its consumption and production in the chemical reactions where the species takes part. In our example, this leads to the system:

$$\begin{cases} \frac{d[CI\_RNA]}{dt} & = k_1[O_{PRM}] \\ \frac{d[CI]}{dt} & = k_2[CI\_RNA] \end{cases}$$

where the $k_j$ reaction rates come from the corresponding reaction attributes. If the attributes do not yield numerical values for the reaction rates $k_j$, they are simply kept symbolic, indexed by the reaction name.

Species $O_{PRM}$ and $CI\_RNA$ are not consumed in the reactions since the corresponding ICRs have the attribute $stoichiometry = M$, which means that they are biological matrices, i.e., they are not consumed or produced in this reaction, but bring information about reaction steps to be performed. In a more general case, on each influence adjacent to a regulatory site, an attribute corresponding to the stoichiometric coefficient can be indicated. It may have four qualitatively different values. A numerical value corresponds to the number of molecules involved in the reaction. The value "0" means that the corresponding species is an enzyme, i.e., it is not consumed or produced in this reaction, but its presence is necessary for the reaction to take place. $M$ means that the corresponding species is a biological matrix. Any other label stands for a vector of coefficients saying how many molecules of each of the 20 types of aminoacids $(a_1, a_2, \ldots, a_{20})$ or each of the 5 types of nucleotides $(n_1, n_2, n_3, n_4, n_5)$, needed to synthesize the macromolecular product of the reaction.

### 4.2 Handling multireaction sites

Any MIN model can be translated into differential equations with an automated procedure, even if some regulatory sites do not represent a single biochemical reaction. In those cases, however, it may be necessary to first demultiply MIN regulatory sites in order to transform the model into a detailed one, for which the previous translation is available.

Regulatory sites of MIN care for two main functions: to represent the regulatory regions, i.e., the physical entities which can change their states by binding to chemical species, and thus participate in different sets of chemical reactions, or to represent the chemical reactions themselves. Thus, in the first interpretation, a regulatory site stands for a set of chemical reactions, as presented in Figure 3. It shows the *demultiplication* of the regulatory site $OR$. Without any *a priori* information, each state of the regulatory site can be obtained from any other state through the influence of a chemical species regulating the original regulatory site. Each such state corresponds to a new species which should be added to the model, as well as the corresponding state transition reactions, represented by regulatory sites. These state species can regulate the activity of the "output" species of the original regulatory site. Each such state regulates these species' activity in independent chemical reactions which should be added into the model as new regulatory sites.

The automated translation of MIN into differential equations amounts to demultiply each regulatory sites of the original MIN which does not correspond to a single simple reaction (attribute $reaction$ is not present, or more than two molecules are combined). Then, the resulting system will be ready for a direct translation into differential equations.

More formally, let $R$ be a non-reaction regulatory site of a MIN $\mathcal{M}$, and $ICR_{C_i,R,a}$ denote the ICR connecting the species $C_i$ to the regulatory site $R$ through the affinity $a$. We first

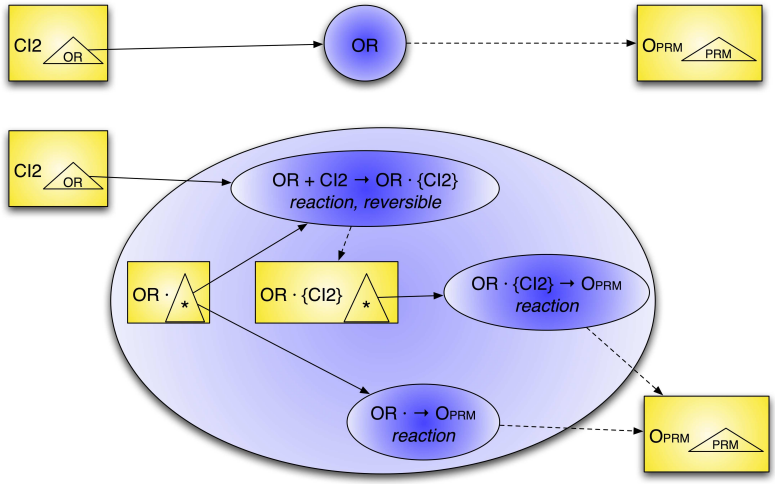



**Figure 3: Use of a regulatory site as a shortcut for a set of chemical reactions. Top. The regulatory site $OR$ (from Figure 2) is not a reaction. Bottom. The regulatory site demultiplication produces 2 new entities: $OR\cdot$ and $OR \cdot \{CI2\}$, as it has only one regulator. By definition of the regulator and regulatory site, the $OR \cdot + CI2 \rightarrow OR \cdot \{CI2\}$ reaction consumes the regulator $CI2$ and the species $OR\cdot$, creating the intermediate species $OR \cdot \{CI2\}$. The regulator $CI2$ binds to the biological site OR. If the stoichiometry is already present in the original ICR, it is added to the new ICR connecting $CI2$ and the site of the $OR \cdot + CI2 \rightarrow OR \cdot \{CI2\}$ reaction. The production reactions of the regulated species $O_{PRM}$ from $OR \cdot \{CI2\}$ or from $OR\cdot$ are automatically added, since the meaning of the regulatory site is that the production rate of the output species relies on the regulatory state of the site.**

construct the multiset $C_R \stackrel{\text{df}}{=} \{C_1, C_2, \ldots, C_n\}$ of regulators of $R$ (chemical species influencing $R$ through some affinity), where a regulator $C_i$ occurs in $C_R$ as many times as indicated by the attribute "stoichiometry" (one by default) in the ICR connecting it to $R$. The MIN $\mathcal{M}$ is then transformed by the demultiplication of $R$, replacing the site $R$ and its influences by:

- the set $\tilde{R} \stackrel{\text{df}}{=} \{R \cdot c \mid c \in \mathcal{P}(C_R)\}$ of new species which are generated by the demultiplication in order to replace $R$, where $\mathcal{P}(C_R)$ denotes the power set[2] of $C_R$;
- the set $R_{in} \stackrel{\text{df}}{=} \{r \mid N_r = R \cdot c + C_i \rightarrow R \cdot (c + \{C_i\}), reaction\_type_r = \text{“}reversible\text{”}, P_r = P_{ICR_{C_i,R,a}}$, with $R \cdot (c + \{C_i\}) \in \tilde{R}\}$ of new regulatory sites corresponding to the chemical reactions enabling the transitions between different states of the regulatory site, represented now by species from $\tilde{R}$, through reactions with their regulators $C_i \in C_R$. Hence, $C_i$ binds to $R \cdot c$ if the number of occurrences of $C_i$ in $c$ is strictly smaller than its stoichiometry coefficient. Each of these new regulatory sites inherits the attributes of the corresponding ICR (in particular, the rates $k_i$) and is connected by new ICRs to the species $R \cdot c$ and $C_i$ (with the $*$-affinity), and by an IRC to the species $R \cdot (c + \{C_i\})$, all with $stoichiometry = 1$;
- the set $E_R \stackrel{\text{df}}{=} \{C_k \mid P_{ICR_{C_k,R,a}}(stoichiometry) = 0\}$ of enzymes influencing $R$. Each $C_k$ is connected to each new regulatory site $r \in R_{in}$ by an ICR with $stoichiometry = 0$;

[2]This denotes here the set of all submultisets of $C_R$. As usual, the empty multiset will be omitted when there is no confusion.

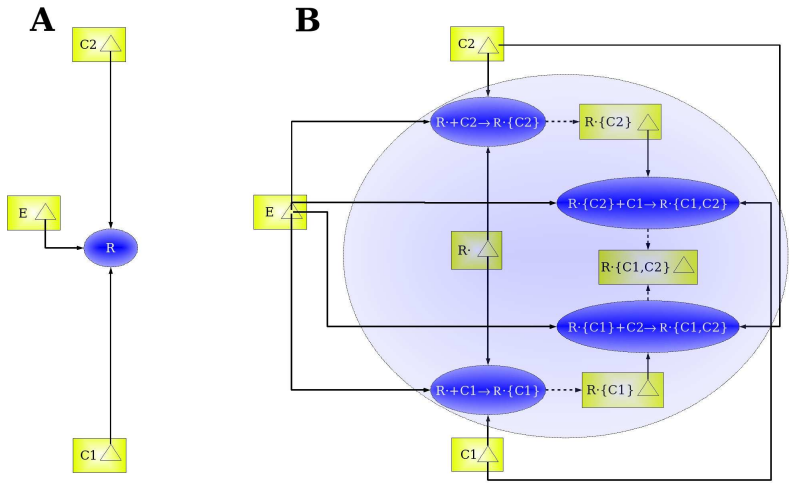



**Figure 4: Example of a regulatory site demultiplication with two regulating species and an enzyme; only ICR are illustrated (except in the gray oval).**

- the set $R_{out} \stackrel{\text{df}}{=} \{r' \mid N_{r'} = R \cdot c \rightarrow C_j, P_{r'} = P_{IRC_{R,C_j}},$ with $R \cdot c \in \tilde{R}$ and $C_j$ regulated by $R\}$ of new regulatory sites corresponding to the chemical reactions changing the activity level of species $C_j$ regulated by $R$, inheriting the attributes of the corresponding IRC in $\mathcal{M}$, and connected by IRCs to the regulatory site $R$.

An example of demultiplication of a regulatory site is presented in Figures 4 (focusing first on the translation of variables and ICRs) and 5 (focusing on the translation of IRCs). In particular, in the big gray oval, we represent the newly generated species and sites.

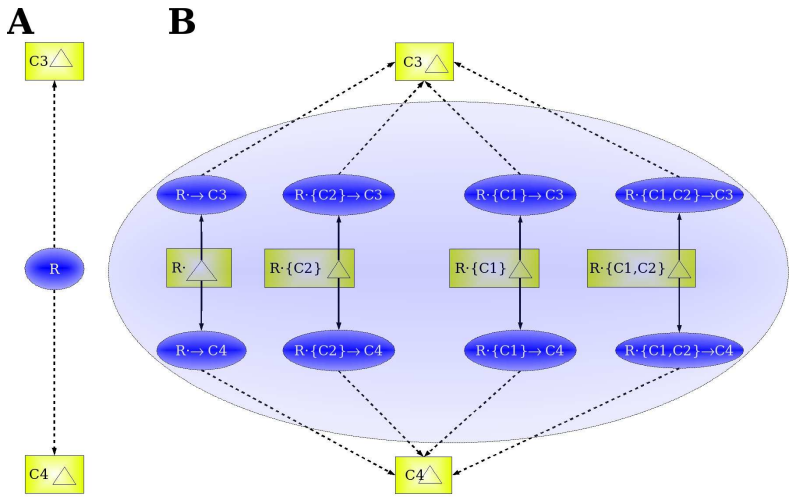



**Figure 5: Example of a regulatory site demultiplication with two regulated species (and the same regulating species as in Figure 4; only IRC are illustrated (except in the gray oval).**

The intermediate representation[3] of the biological system obtained by a simultaneous demulti-

[3]This representation has a MIN-like structure, but some elements are missing, like the relation $\mathcal{F}$. It contains however all information needed for the next translation.

plication of all regulatory sites of the original MIN may now be directly translated into differential equations.

Let us consider the example in Figure 2, where the regulatory site $OR$ (not a reaction) regulates the activity of the $O_{PRM}$ promoter and is influenced by the CI dimer $CI2$. The demultiplication of the regulatory site $OR$, as shown in Figure 3, then leads to the MIN represented in Figure 6.

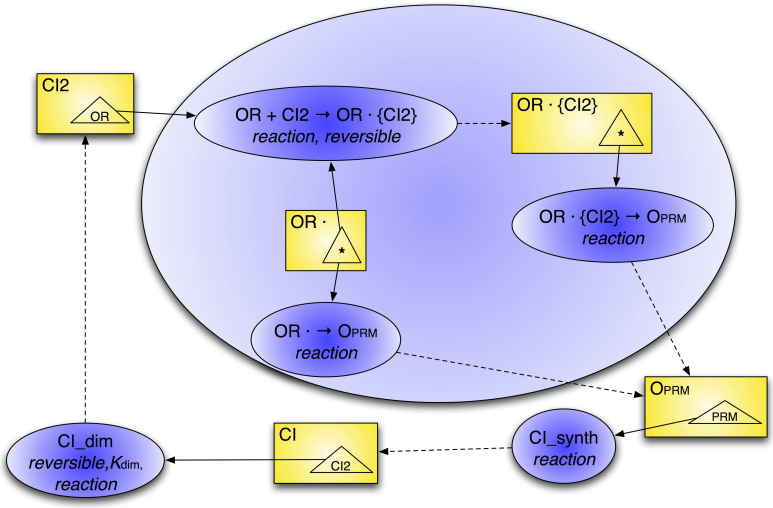



**Figure 6: Transformation of the MIN from the Figure 2, ready for the translation. The influence of the $CI$ species on the $CI\_dim$ site contains the attribute "stoichiometry = 2" (not shown) as it is a dimerisation reaction.**

The regulatory site $OR \cdot CI \rightarrow CI$ in Figure 6, represents the production of $CI$ from $OR \cdot CI$ as a function of the concentration of $OR \cdot CI$. The regulatory site $OR \cdot +CI$ represents the binding reaction that can take place in the system. The corresponding ODEs are:

$$\begin{cases} \frac{d[CI2]}{dt} & = k_{CI\_dim}[CI]^2 - k^{-1}_{CI\_dim}[CI2] \\ \frac{d[OR\cdot\{CI2\}]}{dt} & = k_{OR\cdot+CI2\rightarrow OR\cdot\{CI2\}}[CI2][OR\cdot] - k^{-1}_{OR\cdot+CI2\rightarrow OR\cdot\{CI2\}}[OR\cdot\{CI2\}] \\ \frac{d[OR\cdot]}{dt} & = k^{-1}_{OR\cdot+CI2\rightarrow OR\cdot\{CI2\}}[OR\cdot\{CI2\}] - k_{OR\cdot+CI2\rightarrow OR\cdot\{CI2\}}[CI2][OR\cdot] \\ \frac{d[O_{PRM}]}{dt} & = k_{OR\cdot\{CI2\}\rightarrow O_{PRM}}[OR\cdot\{CI2\}] + k_{OR\cdot\rightarrow O_{PRM}}[OR\cdot] \\ \frac{d[CI]}{dt} & = k_{CI\_synth}[O_{PRM}] - k_{CI\_dim}[CI]^2 \end{cases}$$

In addition to these equations, some constraints on the parameters can be found in the MIN. For example, the $K_{dim}$ attribute of the $CI\_dim$ reaction is the equilibrium constant: $K_{dim} = k_{CI\_dim}/k^{-1}_{CI\_dim}$. If the constants $k_i$ are found in the attributes of the IRCs, ICRs or regulatory sites, their possible values are enumerated.

## 5 Discussion and conclusion

In this paper, we defined and illustrated a translation from MIN models into an ODE description of the dynamics of the associated chemical reactions, but we also showed in another paper [15] how to obtain a family of R. Thomas' regulatory networks modeling the same biological system.

The major problem in modeling a genetic regulation with differential equations is that the substrate can be omitted in the model, considering that all the substrates (nucleotides, aminoacids, etc.), necessary to produce the reaction product (which is generally a protein or an RNA), are present in the cell in appropriate quantities. The mass of each type of atoms should be preserved in a chemical reaction; however, in complex biological processes small molecules (like ATP, water, etc) may be also omitted in the reaction. Sometimes, even bigger molecules are omitted in the reactions with unknown mechanism.

The biological descriptions of genetic regulation often follow the scheme:

$$Regulator + Gene \rightarrow Protein,$$

where $Regulator$ is a protein itself, possibly different from the $Protein$ in the right part of the equation. However, a more realistic equation reflecting the set of biochemical reactions of protein expression should be something like

$$Regulator + Gene + Enzymes + Resources \rightarrow Regulator + Gene + Enzymes + Protein$$

In this equation, $Enzymes$ stands for the machinery of protein synthesis (RNA polymerase, ribosomes, etc.) and $Resources$ stands for the necessary substrates to produce the $Protein$. To insure the conservation of mass in the system of biochemical equations, it is necessary to know the stoichiometric coefficients of each reaction.

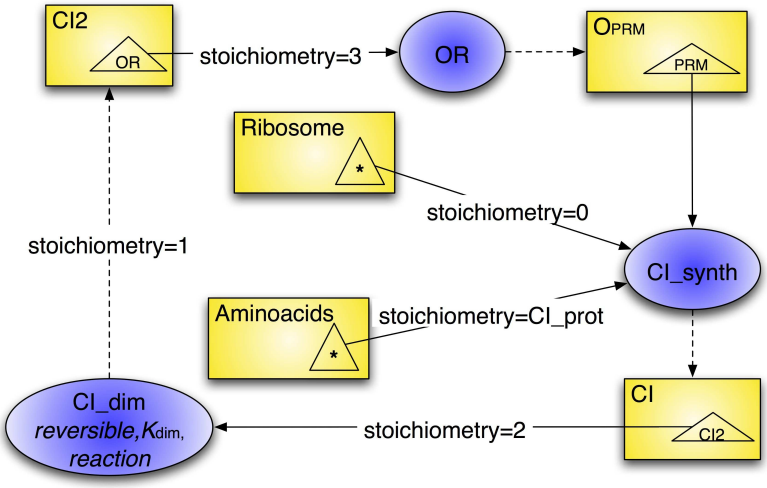



**Figure 7: A MIN model representing the CI protein synthesis, including the participation of a ribosome (which acts as an enzyme) and of aminoacids (which play the role of resources).**

To further illustrate the usage of stoichiometric coefficients in the MIN modeling, let us consider the Figure 7. The stoichiometric coefficient for $Aminoacids$ is a label. It represents

the composition of the corresponding macromolecule: CI protein. In general, the opposite reaction of the biochemical synthesis is degradation, and it releases the same quantities of the corresponding substrate residuals. The stoichiometric coefficient for the $Ribosome$ is 0, which means that these are enzymes in the reactions of the CI protein synthesis. The stoichiometric coefficient for $CI$ is 2 for the reaction of the dimerisation of CI, meaning that two molecules of CI are needed to form a dimer. The stoichiometric coefficient for the CI dimer regulating the site OR is 3 meaning that 3 dimers can bind to this site, simultaneously. The stoichiometric coefficients give the $\alpha_i$ power coefficients in the corresponding equation.

For so detailed systems, the demultiplication step during the translation into ODEs will generate a lot of intermediate reaction steps. However, this difficulty can be overcome by using the protein sequence, being possibly an attribute of the $CI$ protein species, in order to reconstruct the precise order of the protein synthesis reaction steps, instead of considering all possible aminoacid combinations.

The attributes of the ICRs and IRCs contain various types of information, such as the type of the interaction (activation, inhibition, consumed, produced), which enable to find out species being enzymes and those changing their concentration in a chemical reaction. Possible values of kinetic rates of the corresponding chemical reactions may be found in the ICR or IRC attributes. Also, to simplify the obtained model by identifying mutually exclusive regulators, or to eliminate the state changes which do not lead to the modification of the activity of the regulated species, the description of states of the regulatory site can be found in the relation $\mathcal{F}$. Another possibility is to calculate the ODE parameters based on these state description, as in [8].

The MIN formalism may play the role of an intermediate level between insufficiently precise natural language and too specialized mathematical descriptions of biological systems. The MIN construction is a process of inferring the biological interaction networks from the biological observations of microscopic and macroscopic level. The underlying structure provides a skeleton for the understanding of the organization and functioning of biological systems. Compared with some UML based models for biology [1, 11], MIN has the advantage of enabling the automatic translation in other formalisms.

Existing approaches to the modeling of biological networks using ODEs share some basic concepts with MIN, but differs from it in some points:
The CellDesigner [2] is a structured diagram editor for drawing biological networks, based on the graphical notation system proposed by Kitano [6]. These diagrams represent the biological objects, similarly as the MIN does. The CellDesigner models are stored using the Systems Biology Markup Language (SBML)[5] for the simulation with ODEs.
E-Cell [12] is an object-oriented software for modeling, simulation, and analysis of large scale complex systems. E-Cell Simulation Environment allows modeling of discrete, stochastic and continuous processes. Thus, at different steps of iterative modeling, MIN can provide quantitative models for the further analysis with E-cell.
Cell Illustrator [9] is another environment for describing biopathways with hybrid functional Petri nets (HFPN), visualizing simulation results, evaluating hypothesis and integrating data from biopathway databases. Compared to MIN, the modeling with HFPN may introduce structural elements pertinent for the model dynamics but without a direct biological interpretation. Also, the choice between discrete or continuous modeling has to be made for each entity or process during the modeling, while in MIN this decision is postponed until the analysis stage.

A specialized MIN editor including available translation algorithms to R. Thomas' regulatory networks and to ODEs is currently under development.

## References

[1] M. Beurton-Aimar, S. Pérès, N. Parisey, C. Nazaret, and J.P. Mazat. Modeling biologic networks to use them with heterogeneous treatments. In *Proceedings of Ecole Thématique "Modélisation et simulation de processus biologiques dans le contexte de la génomique - 2003 - Dieppe (France)"*, 2003.

[2] http://celldesigner.org/

[3] H. Eisen, P. Brachet, L. Pereira da Silva, and F. Jacob. Regulation of repressor expression in $\lambda$. *Proc. natn. Acad. Sci.*, 66:855–862, 1970.

[4] K.R. Heidtke and S. Schulze-Kremer. Design and implementation of a qualitative simulation model of lambda phage infection. *Bioinformatics*, 14(1):81–91, 1998.

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

[6] Kitano, et al. Using process diagrams for the graphical representation of biological networks. *Nature Biotechnology* 23(8), 961 - 966, 2005.

[7] C. Kuttler, J. Niehren, and R. Blossey. Gene regulation in the $\pi$-calculus: Simulating cooperativity at the lambda switch. *Bio-CONCUR*, 2004.

[8] C. Lou, X. Yang, X. Liu, B. He, and Q. Ouyang. A quantitative study of lambda phage SWITCH and its components. *Biophys J BioFAST*, (doi:10.1529/biophysj.106.097089), 2007.

[9] H. Matsuno, A. Doi, M. Nagasaki, and S. Miyano. Hybrid Petri net representation of gene regulatory network. *Pac Symp Biocomput*, pages 341–52, 2000.

[10] M. Ptashne. *A Genetic switch*. Blackwell Science (ISBN : 978-0865422094), 1992.

[11] M. Roux-Rouquié, N. Caritey, L. Gaubert, B. Le Grand, and M. Soto. Metamodel and modeling language: towards an Unified Modeling Language (UML) profile for systems biology. In *Object-oriented Modeling in Biology and Medecine, SCI2005*, 2005.

[12] K. Takahashi et al. E-Cell 2: multi-platform E-Cell simulation system. *Bioinformatics* 19(13):1727-9, 2003.

[13] D. Thieffry and R. Thomas. Dynamical behaviour of biological regulatory networks - II. immunity control in bacteriophage lambda. *Bull. Math. Biol.*, 57(2):277–297, 1995.

[14] R. Thomas, A.M. Gathoye, and L. Lambert. A complex control circuit. regulation of immunity in temperate bacteriophages. *Eur. J. Biochem.*, 71(1):211–227, 1976.

[15] A. Yartseva, H. Klaudel, R. Devillers, F. Képès. Incremental and unifying modelling formalism for biological interaction networks. IBISC, TR 3/07, 2007.

## 5.4 Expression des réseaux biologiques du MIN dans les réseaux de Petri

Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

# Translation into Petri Nets of Biological Networks represented in MIN formalism

Anastasia Yartseva[*], Hanna Klaudel*, Raymond Devillers[†], François Képès[‡]

## Abstract

The Modular Interaction Network (MIN) meta-modeling formalism describes biological systems in a user-friendly manner mixing both graphical (bipartite graph) and narrative (textual explanations) description features. MIN supports both the quantitative and qualitative aspects of the modeled systems. It provides also algorithms for the translation of biological knowledge into various commonly used modeling formalisms such as ordinary differential equations and René Thomas' multi-level logical formalism. In this paper, translation of MIN into Petri nets is demonstrated. A translation algorithm is described and illustrated on a biological case study: apoptosis.

## Introduction

Biological networks are commonly modeled by various sorts of graphs [6, 7, 15, 5, 1, 17], which may contain a large number of nodes and may be densely connected. The Modular Interaction Network (MIN) meta-formalism [19] enables various levels of abstraction allowing to express in a compact way detailed relations, e.g. that interactions may be mutually exclusives or may take place only in a precise order. MIN provides a user-friendly syntax mixing both graphical (bipartite graph) and narrative (textual explanations) description features. MIN additionally offers a series of translation algorithms defining automatically analyzable target semantics. In this paper, we address a translation from MIN to the Petri nets formalism, which is one such target formalisms particularly suitable for a further analysis of the biological systems. Indeed, Petri nets comprise various classes including high-level [2], hybrid [10], timed [11] or stochastic ones [14], and a wide set of structural or behavioral analysis techniques and tools [18, 1].

This paper is structured as follows. We start with a brief description of the biological example of apoptosis that serves as a case study. Then, we sketch briefly the main features of MIN and of Petri nets. A translation algorithm is introduced next and applied to the case study. Finally, the translated Petri net for apoptosis is compared with that obtained directly from the biological description in [3].

## Biological description of apoptosis

Apoptosis is an inducible intrinsic cell suicide program. Apoptosis plays an important role in the development of multi-cellular organisms, in the control of cell number and in the elimination of morbid cells. One of the apoptosis inducing signals passes by the Fas receptor in the cell membrane. The major apoptosis actors are called caspases. Caspases are present in the cell in inactive forms called procaspases that can be activated by a cleavage reaction. The extra-cellular inducer (Fas ligand) of the Fas receptor promotes its interaction with intra-cellular proteins containing the "death" domain (DD), such as FADD or Daxx. FADD activates the apoptosis through the caspase pathway and Daxx activates apoptosis through the MAPK/Jun pathway via interaction with Ask1 [13,9].

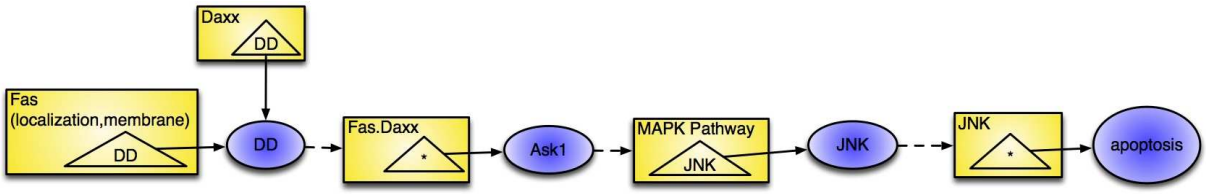
[*] IBISC, Université d'Evry, 523 place des Terrasses de l'Agora, 91000 Evry, France
[†] Département d'Informatique, Université Libre de Bruxelles, CP212, B-1050 Bruxelles, Belgium
[‡] Epigenomics Project, Genopole®, CNRS & Univ. Evry, 523 place des Terrasses de l'Agora, 91000 Evry, France

## *MIN formalism*

The MIN meta-model can be seen abstractly as a bipartite graph involving two kinds of nodes: *chemical species* and *regulatory sites*. Every regulatory site has a set of *regulating* and *regulated* chemical species and their role is expressed by *influences*. Chemical species and regulatory sites together are called *variables*. They represent biological objects at some level of abstraction. They may be, for instance, molecules or parts of them, complex processes like regulatory pathways, complex systems like sensors, or even an entire organism.

As the knowledge about biological systems is based on observations and experiments, the *observable level of activity* of biological objects can change through various states of the biological system. These objects can influence the levels of activity of each other. So, every variable in MIN is assumed to have a set of *observable values*, corresponding to the observable levels of activity of the corresponding biological objects, such as "absent", "present", "active" etc. (see the table in Figure 1 for an illustration).

Biological objects, represented by variables in MIN, may interact and play specific roles in these interactions. For example, they can take part in a chemical reaction, one object modifying, creating or destroying the other. It is assumed that every interaction happens through a particular regulatory site.




**Relation $\mathcal{F}$**

| **Fas** | **Daxx** | **DD** | **Fas.Daxx** | **Ask1** | **MAPK Pathway** | **JNK-site** | **JNK** | **Apoptosis** |
|---|---|---|---|---|---|---|---|---|
| *active* | *present* | *Undef* | *present* | *undef* | *active* | *undef* | *active* | *yes* |
| *active* | *absent* | *Undef* | *absent* | *undef* | *non active* | *undef* | *non active* | *no* |

**Figure 1** *A MIN fragment representing the apoptosis activation by the Fas receptor. The MIN chemical species are represented by boxes, their affinities by triangles inside the boxes, regulatory sites by ellipses, influences of a species on a regulatory site by plain arcs, and influences of a regulatory site on a species by dashed arcs.*

A *chemical species* represents a biological object which can influence one or more regulatory sites. These influences represent association/dissociation chemical reactions, electron transfers, etc. A species may have one or more influence capabilities, which are called *affinities*. An affinity is the ability of a biological object to take part in the interaction with a certain class of other biological objects or to catalyze a given chemical reaction. Thus, an affinity may correspond to a protein domain for a protein (as DD in Figure 1) or a surface molecule (receptor) for a cell. The nature of the interaction between two biological entities can be unknown. So, a wild-card affinity, labeled ``*", may be defined for every species, standing for an unknown mechanism of regulation.



Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

A *regulatory site* regulates species activity in a production/degradation chemical reaction, or by other mechanisms such as its structural changes or by cooperativity effects. A regulatory site may represent a chemical reaction consuming or producing chemical species, as well as a biological object such as a genome region or a protein domain that changes its state after a chemical reaction. A regulatory site has a *label* which characterizes its capabilities of being influenced through the affinities. If a regulatory site and an affinity of a species have the same label, it means that the interaction is possible between the biological objects corresponding to the site and the species. A regulatory site represents an ``input'' for a species and regulates its activity through the integration of several influences on it.

The variables (chemical species and regulatory sites) can have *attributes*, which come from the corresponding biological objects, and may have types like ``size'', ``localization'' etc. expressing the knowledge about them. For example, the attribute (localization, membrane) stands for the cell membrane localization for the Fas receptor in Figure 1. Several species with the same name may be present in MIN, if they have attributes with different values. So, we can represent copies of the same protein in free or bound state, or of the same gene at its natural location and translocated in a different place in the genome.

Two directions of *influences* between the variables of the model can be considered: *from chemical species onto regulatory sites* and *from regulatory sites onto chemical species*. It is also assumed that there is no influence between variables of the same kind. An influence has a set of attributes which describes the relationship between the values of the species and those of the regulatory site, like the parameters of the corresponding chemical reaction: kinetic rate or speed, or stoichiometric coefficients. For instance, a null stoichiometric coefficient means that the corresponding species is an enzyme.

The dynamics of the biological system is represented in MIN by “snapshots”, lines in a relation $\mathcal{F}$. In general, in a single biological experiment (an observation), the values of only a subset of biological objects are measured. In this case, the observable values of non observed species and sites take the special value “*undef*” and the state of the system is considered as “partly” defined. $\mathcal{F}$ is a relation which yields all the partly defined system states really observed in biological experiments and described by biologists (see the table in Figure 1). $\mathcal{F}$ plays the role of a databank from which the parameters of the dynamics of the system interactions could be identified. If some of these parameters (as, for example, kinetic rates for biochemical reactions) are known (were measured by bench experimentation), they will be directly mentioned in the attributes of the corresponding influences.

## Petri nets

Petri nets [16] (also known as *place/transition nets*) are one of several mathematical models of discrete distributed systems. They graphically depict the structure of a distributed system as a directed bipartite graph with annotations. A Petri net contains two kinds of nodes: *places* (depicted as rounds and representing rather passive elements of the system like conditions, resources or biological species) and *transitions* (depicted as squares and standing generally for active system elements like events or chemical reactions). The nodes are connected by directed arcs representing the causal relation between places and transitions. They are annotated with their weight (natural number), if greater than one.

A state of the system is represented by a distribution of *tokens* (dynamic elements of the system) on places of the Petri net indicating the presence of resources or fulfillment of conditions. If all input places of a transition are marked with a sufficient number of tokens

(corresponding to the arc weight), the transition may fire indicating a state change. The new state is obtained by removing from the input places of the transition the tokens accordingly to the arc weighs and producing tokens on the output places of the transition. A transition with no input place may always fire producing tokens on its output places, and symmetrically, a transition with no output place may consume tokens in its input places if sufficiently marked (modeling in this way a possible exchange with the environment). If a resource has to be present (a token on a place) but the firing of a transition does not require its consumption (like for instance in catalytic reactions), it may be modeled by two opposite direction arcs (depicted in figures as one bidirectional *read arc*). An example of a firing is shown in figure 2.

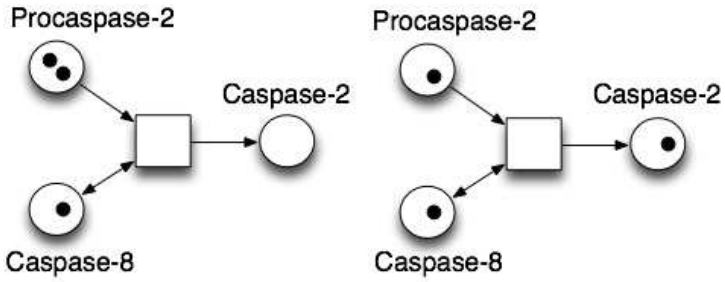



**Figure 2** *Dynamics of a Petri net*

## *Translation Algorithm*

The translation algorithm from MIN to Petri nets relies on the places and transitions reflecting the logical functions describing the evolution of biological signals from the inputs of the biological system to the outputs. In general, MIN chemical species are translated to static Petri net objects, namely places, while regulatory sites become active ones, namely transitions.

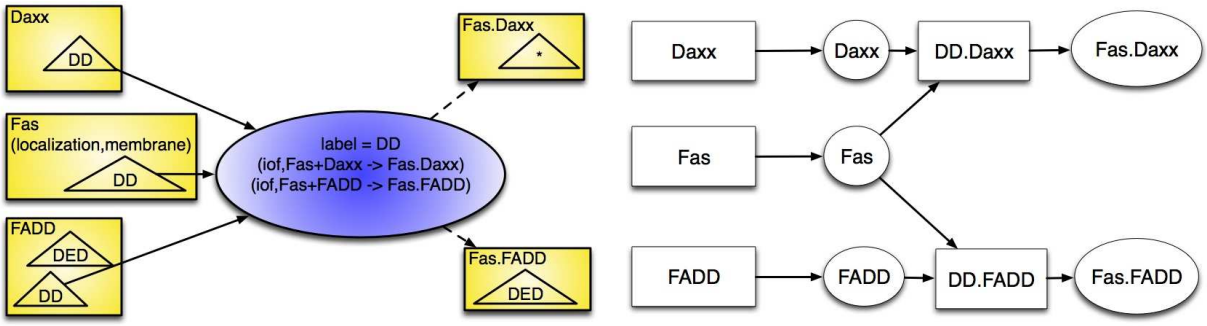



**Figure 3** *A MIN fragment from the Figure 4 and its translation in Petri net*

However, some regulatory sites (for instance, DD in Figure 3) may contain attributes "iof" (for input-output function) describing the link between the input species of this site and the output ones. During the translation, these attributes (iof, Fas + Daxx → Fas.Daxx) and (iof, Fas + FADD → Fas.FADD) are interpreted and transformed in two Petri net transitions: DD.Daxx and DD.FADD, connected to Daxx and Fas (Fas and FADD, respectively). If no information is given in the attributes of a regulatory site detailing the dynamical relations between its inputs and outputs, it is considered that all the input species are needed to produce all the output ones, simultaneously.

The chemical species that have no input regulatory site are considered to be inputs of the Petri net. As a consequence, transitions with the same name, potentially producing this chemical

species, are added to the model. The biological meaning of these transitions is that the species comes to the system from the environment. Thus, the chemical species Fas, FADD and Daxx in Figure 3, left, are represented by both transitions and places Daxx, Fas and FADD in Figure 3, right. Other species, such as Fas.Daxx and Fas.Fadd, being connected to other regulatory sites (see Figure 4, left) are translated to places with the same name. Species or regulatory sites having no outgoing influence are considered to be outputs of the translated Petri net. If the stoichiometric coefficients are indicated in the influence attributes, they are taken into account. For example, the coefficient 0 is translated by a read arc in the corresponding Petri net. An example of the translation from a MIN representation of the apoptosis induced by the Fas receptor and the resulting Petri net is shown in Figure 4.

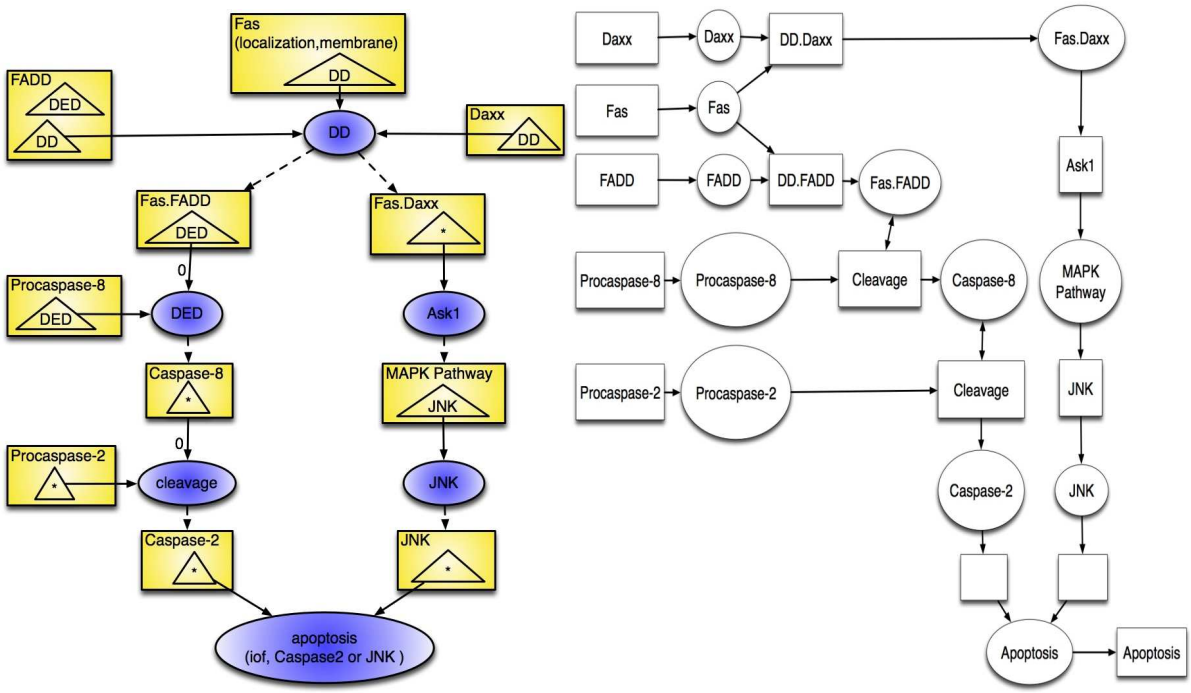



**Figure 4** *MIN representation of the Fas induced apoptosis and the translated Petri net.*

The resulting Petri net may be compared with the one obtained directly from the biological data in [3] as they belong to the same net class. It may be noticed that both nets have similar structures, while differing in minor details coming from the different level of abstraction of the models. For instance, the place Fas.Daxx and the transition Ask1 in the translated Petri net are absent from the Petri net in [3], even if Ask1 is mentioned in the biological description. This difference may come from some implicit arbitrary choices needed during any model construction, as mentioned by the authors.

## Conclusion

The presented translation from MIN to Petri nets enables to use standard Petri net tools and techniques for the analysis of the biological system represented in MIN. It may also be considered as a first step in a more complex approach allowing to exploit the biological data contained in MIN using high-level, timed or stochastic extensions of Petri nets.

Compared to other graphical representations of biological networks, MIN representations may contain textual information, allowing to reduce the graph size and useful in translation algorithms (like input-output functions for complex regulatory sites). The apoptosis example showed that modeling a biological system directly in a target formalism (like Petri nets) often

requires making implicit (undesired) choices. These kinds of problems may be avoided in MIN, which may be seen a rather well organized database enabling to use exactly the same data set for the construction of several types of target semantics. From a modeling methodological point of view, the affinities that express the interaction capabilities of chemical species enable to incrementally extend a MIN representation of the biological system, finding the most relevant candidates to play a role in the studied phenomena.

## *Bibliography*


1. Berthomieu, B.; Ribet, P-O.; Vernadat, F.: The tool TINA - Construction of abstract state spaces for Petri nets and time Petri nets; Int. J. of Production Research, 42(2004)14
2. Comet, J.-P.; Klaudel, H.; Liauzu S.: Modeling multi-valued genetic regulatory networks using high-level Petri nets. Proceedings of ICATPN, LNCS 3536, 208-227, Springer, 2005
3. Heiner, M.; Koch, I.; Will, J.: Model Validation of Biological Pathways Using Petri Nets – Demonstrated for Apoptosis. J. BioSystems, 75/1-3 (2004)15-28
4. Will, J.; Heiner, M.: Petri Nets in Biology, Chemistry, and Medicine – Bibliography; Computer Science Reports 04/02, BTU Cottbus, November 2002
5. Kitano, H.; Funahashi, A.; Matsuoka, Y.; Oda, K.: Using process diagrams for the graphical representation of biological networks. Nature Biotechnology, 23(8)961-966, 2005
6. Kohn, K. W.; Aladjem, M. I.: Circuit diagrams for biological networks; Molecular Systems Biology, 2006
7. Kohn, K.W.; Aladjem, M.I.; Weinstein, J.N.; Pommier Y.: Molecular interaction maps of bioregulatory networks: A general rubric for systems biology. Molecular Biology of the Cell, 17(1)1-13, 2006
8. Kurata, H.; Matoba, N. Shimizu, N.: CADLIVE for constructiong a large-scale biochemical network based on a simulation-directed notation and its application to yeast cell cycle. Nucleic Acid Res. 31(2003) 4071-4084
9. Moldoff, K.; Baudino, T.: Apoptosis Poster. Sigma Aldrich. http://www. sigmaaldrich.com/Area_of_Interest/Life_Science/Cell_Signaling/Apoptosis_Poster.html
10. Matsuno, H.; Tanaka, Y.; Aoshima, H.; Doi, Matsui, M.; Miyano, S.: Biopathways representation ans simulation on hybrid functional Petri net. Silico Biol. 3, 2003
11. Merlin, P.; A study of the recoverability of communication protocols; Univ. of California, Computer Science Dept., PhD thesis, Irvine, 1974
12. Murata, T., Petri Nets: Properties, Analysis and Applications. Proceedings of the IEEE, Vol. 77 No. 4, 1989
13. Nijhawan, D.; Honarpour, N.; Wang, X.: Apoptosis in Neural Development and Disease. Annual Review Neuroscience 23 (2000) 73-89
14. Peccoud, J.: Stochastic Petri nets for genetic networks. M S-Medicine, Sciences 14(1998)991-993
15. Pirson, I.; Fortemaison, N.; Jacobs, C.; Dremier, C.; Dumont, J.E.; Maenhaut, C.: The visual display of regulatory information and networks. Trends in Cell Biology, 10(10)404-408, 2000
16. Reisig, W., Petri Nets: An introduction. Springer-Verlag, 1985
17. Smidtas S., Yartseva, A.; Schachter, V.; Kepes, F.: Model of interactions in biology and application to heterogeneous network in yeast. C. R. Biologies 2006 329(2006)945-52
18. Starke, P. H.: INA - Integrated Net Analyzer. Manual, Berlin 1998.
19. Yartseva, A.; Klaudel, H.; Devillers, R.; Kepes, F.: Incremental and unifying modelling formalism for biological interaction networks; IBISC, Computer Science Reports, 3/07, 2007, submitted to BMC Bioinformatics




Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

## 5.5 Conclusion

Le formalisme MIN proposé peut jouer le rôle de niveau intermédiaire entre le "langage naturel" insuffisamment formalisé, et les descriptions mathématiques trop spécialisées pour les systèmes biologiques. La construction d'un modèle MIN est un processus d'inférence des réseaux d'interactions biologiques à partir des observations biologiques de niveau macro et microscopiques. Sa structure sous-jacente fournit un squelette pour comprendre un premier principe d'organisation des systèmes biologiques. La traduction de MIN en réseau de Petri permet d'utiliser les outils d'étude et de manipulation de ces derniers. Il s'agit là aussi d'un premier effort de traduction qui pourrait être conduit vers des réseaux de Petri stochastiques. L'exemple du cas biologique de l'apoptose, montre que la modélisation directe en réseau de Petri implique de faire des choix implicites. Ce type de problème est levé avec MIN, qui peut être vu comme une base de données utile pour la construction de plusieurs modèles en utilisant différentes sémantiques.

Anastasia Yartseva SmidtasAnastasia@Smidtas.com http://camera-contact.com

# Chapitre 6

# Conclusion et perspectives

Le travail présenté ici s'articule autour de l'étude in silico des réseaux biologiques en abordant à la fois les aspects d'intégration, de formalisation et de modélisation des réseaux et sous-réseaux biologiques. Dans ce contexte, nos travaux ont porté dans un premier temps, sur l'étude des interactions entre les sous-graphes dans les réseaux biologiques hétérogènes, puis sur le développement d'un cadre de modélisation des graphes particulièrement adapté à l'étude de réseaux d'interactions hétérogènes, MIB (pour Modèle d'Interaction Biologique) et le développement d'un outil d'intégration correspondant, BIB. Ces développements ont été mis à profit afin de caractériser et d'étudier la présence et le mode de connexion de sous-réseaux ou motifs à l'intérieur de réseaux plus vastes.

Poursuivant cet effort dans le domaine de l'étude des systèmes biologiques, nous avons introduit la notion de l'enracinement de deux graphes sur leur interface afin d'étudier la manière dont les deux graphes, représentant des interactions des types différents ( c'est à dire sous-réseau d'interaction protéine-protéine et sous-réseau de régulation transcriptionnelle), peuvent interagir via les noeuds communs, s'imbriquent et se complètent.

Le développement du modèle MIB s'est inscrit dans une des problématiques majeures de la biologie théorique, à savoir celle concernant les limites inhérentes à la représentation

par un graphe simple de réseaux d'interactions hétérogènes. Ces réseaux peuvent regrouper $n$ gènes ou protéines liés par des interactions aussi diverses que l'appartenance au même complexe protéique, à la même voie de régulation transcriptionnelle, à la même voie métabolique ou à la même voie de transduction de signal. Pour pallier ces limites, nous avons développé un cadre de modélisation basé sur des graphes bipartis qui permet de prendre en compte l'ensemble des interactions biologiques connues (interactions protéine-protéine, régulation transcriptionnelle, voie métabolique, létalité synthétique). L'apport majeur de MIB est de permettre de représenter les relations multiples ($n$-aires) existant entre les acteurs biologiques, tout en préservant leur étude dynamique. De ce point de vue, MIB est particulièrement adapté à la caractérisation et à l'étude de modules ou motifs composés d'interactions de différents types et constituant fréquemment des sous-réseaux fonctionnels répétés au sein de réseaux hétérogènes plus complexes. Dans ce contexte, MIB a été mis à profit pour analyser, au sein du réseau relationnel disponible du protéome de la levure, deux types d'interactions hétérogènes. Le premier type est représenté par les occurrences de boucles de rétroaction (Feedback loops) composées de régulation transcriptionnelle simple ou couplée à une ou plusieurs interactions protéiques. Le deuxième type est un module impliquant deux ou plusieurs gènes liés par des interactions de type transcriptionnel et vérifiant un lien de type synexpression direct ou indirect. Les résultats ont clairement montré la robustesse de l'approche dans le cadre de l'étude de réseau impliquant plusieurs milliers de gènes ou protéines et plusieurs centaines de milliers d'interactions.

L'apport majeur de l'application BIB, basée sur MIB, est de fournir un accès et un système de requête pour une bases de données de voies de réactions et d'informations génomiques hétérogènes. L'ensemble du programme a été rendu disponible à la communauté scientifique.

Enfin, dans la dernière partie, nous nous sommes basés sur l'ensemble de ces réalisations afin de créer un cadre de modélisation, MIN, permettant d'étudier les systèmes biologiques

de manière nouvelle et efficace. Ce modèle permet d'inclure la grande majorité des observations disponibles sur le système et de traduire automatiquement les données en d'autres formalismes afin de profiter des outils d'analyse disponibles dans différents domaines de modélisation.

Nous avons proposé un nouveau paradigme [1] pour la modélisation des systèmes biologiques, dans lequel toutes les données de la biologie expérimentale sont considérées comme des empreintes de l'état du système biologique à un moment donné, et stockées en tant que telles, sans aucune interprétation, c'est à dire en séparant les données biologiques des principes mathématiques de modélisation de la dynamique des systèmes, tels que les équations différentielles ou les réseaux booléens. L'information sur le système modélisé sous forme de MIN est ajoutée et raffinée incrémentalement. L'état actuel des connaissances contenues dans MIN peut être automatiquement traduit dans un formalisme donné (équations différentielles, réseaux de Petri, réseaux de René Thomas) pour l'analyse de la dynamique du système ; il pourra aussi être utilisé dans le futur par un système d'inférence via les techniques de l'intelligence artificielle afin de résoudre des problèmes biologiques complexes.

L'approche MIN présentée dans cette thèse est originale et utile pour le biologiste. En effet, il n'existait pas de modèle proposant toutes les caractéristiques citées précédemment, telles que l'absence d'interprétation des données, l'incrémentalité et la possibilité de traduction automatique vers d'autres formalismes. L'utilité de MIN provient également du fait que ce modèle permet de regrouper et réutiliser les données rassemblées dans une

[1] Un paradigme est une représentation du monde, une manière de voir les choses, un modèle cohérent de vision du monde qui repose sur une base définie (matrice disciplinaire, modèle théorique ou courant de pensée). En transposant dans l'univers informatique, un paradigme peut être comparé à un système d'exploitation (Windows, Linux, Mac). C'est en quelque sorte un rail de la pensée dont les lois ne doivent pas être confondues avec un autre paradigme. Le mot paradigme s'emploie fréquemment dans le sens de perception du monde. Par exemple, dans les sciences sociales, le terme est employé pour décrire l'ensemble d'expériences, de croyances et de valeurs qui influencent la façon dont un individu perçoit la réalité et réagit à cette perception. Ce système de représentation lui permet de définir l'environnement, de communiquer à propos de cet environnement, voire d'essayer de le comprendre ou de le prévoir.

structure unique. Les algorithmes de traductions des modèles MIN vers d'autres formalismes sont proposés afin de permettre l'utilisation des méthodes d'analyse formelle ou numérique disponibles dans différents domaines. Ceci permet à chaque étape de valider ou invalider le modèle avant de passer à une itération suivante de la construction du modèle du système biologique étudié. MIN permet de représenter très naturellement les réseaux d'interactions hétérogènes gràce à sa structure bipartite. Il n'y a pas non plus de distinction faite à l'avance (au moment de l'ajout des données) entre les processus discrets ou continus. Ce choix n'est fait qu'au moment de la traduction, ce qui permet d'étudier le même phénomène du point de vue discret ou continu.

Les perspectives de ce travail de thèse à court terme incluent l'élaboration de deux extensions du MIN : les phases du MIN et le MIN hiérarchique. Les phases du MIN représentent une couche supplémentaire qui permet de stocker les informations non seulement sur les états macroscopiques du système (appelés arbitrairement phases du MIN), mais également sur les transitions entre les phases et les changements d'états des variables qui les accompagnent. Le MIN hiérarchique est une extension qui permet d'utiliser les attributs des sites de régulation pour créer une structure hiérarchique entre plusieurs modèles MIN. En effet, la construction incrémentale du modèle peut conduire à ajouter des informations supplémentaires sur les détails d'un processus biologique, représenté par un site de régulation à l'étape précédente. Pour répondre à certaines questions biologiques, la représentation plus concise est préférable ; et pour d'autres questions, c'est la représentation détaillée qui a plus d'intérêt. Ainsi, le fait de garder les deux modèles, le modèle complet et le modèle qui le raffine, permet de naviguer aisément entre les niveaux d'abstractions pour différentes parties du modèle afin de choisir la représentation la plus appropriée. De plus, l'implémentation dun'éditeur de MIN est en cours, et elle inclura les deux extensions du MIN : les phases et le MIN hiérarchique.

Les perspectives à moyen terme incluent une application, sur un exemple biologique

étendu, des méthodes de travail développées au cours de ce travail de thèse. Les systèmes biologiques tels que l'interrupteur génétique du phage lambda, la régulation de l'hématopoïèse (hors cadre de ce manuscrit) et les réseaux d'interaction chez la levure ont déjà été traités au cours de ce travail. Une des applications possibles de MIN est la physiologie du rein, car les données extensives et multi-échelle sont déjà regroupées dans une base de données par S. Randall Thomas dans le projet Physiome [85].

# Bibliographie


[1] Alan Aderem. Systems biology : Its practice and challenges. *Cell*, 121(4) :511–513, May 2005.

[2] T. Akutsu, S. Miyano, and S. Kuhara. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16(8) :727–734, Aug 2000.

[3] H. Alla and R. David. Continuous and hybrid Petri nets. *J. Circ. Syst. Comp.*, 8 :159–188, 1998.

[4] R. Alur, C. Belta, F. Ivancic, V. Kumar, M. Mintz, G. J. Pappas, H. Rubin, and J. Schug. Hybrid modeling and simulation of biomolecular networks. In *Hybrid Systems : Computation and Control, 4th International Workshop, HSCC 2001*, pages 19–32, Rome, Italy, 2001. Springer, LNCS 2034.

[5] K. Amonlirdviman, R. Ghosh, J.D. Axelrod, and C.J. Tomlin. A hybrid systems approach to modeling and analyzing planar cell polarity. In *Proceedings of the 3rd International Conference on Systems Biology*, 2002.

[6] M. Antoniotti, B. Mishra, C. Piazza, A. Policriti, and M. Simeoni. Modelling cellular behavior with hybrid automata : Bisimulation and collapsing. In C. Priami, editor, *International workshop on Computational Methods in Systems Biology, CMSB'03*, volume 2602 of *LNCS*, pages 57–74. Springer-Verlag, 2003.

[7] M. Antoniotti, A. Policriti, N. Ugel, and B. Mishra. XS-systems : eXtended S-Systems and algebraic differential automata for modeling cellular behavior. In *Proceedings of the*

*International Conference on High Performance Computing, HiPC 2002*, Bangalore, India, December 2002.

[8] G.D. Bader, M.P. Cary, and C. Sander. Pathguide : a pathway resource list. *Nucleic Acids Res*, 34(Database issue) :D504–D506, Jan 2006.

[9] D. E. Bassett, M. B. Eisen, and M. S. Boguski. Gene expression informatics–it's all in your mine. *Nat Genet*, 21(1 Suppl) :51–55, Jan 1999.

[10] G. Batt, H. de Jong, J. Geiselmann, and M. Page. Qualitative analysis of genetic regulatory networks : A model-checking approach. In B. Bredeweg and P. Salles, editors, *Working Notes of Seventeenth International Workshop on Qualitative Reasoning, QR-03*, pages 31–38, 2003.

[11] R.E. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, NJ., 1957.

[12] R.E. Bellman. *Adaptive Control Processes.* Princeton University Press, Princeton, NJ., 1961.

[13] G. Bernot, J.-P. Comet, A. Richard, and J. Guespin. Application of formal methods to biological regulatory networks : extending Thomas' asynchronous logical approach with temporal logic. *Journal of Theoretical Biology*, 229(3) :339–347, 2004.

[14] Michael L Blinov, James R Faeder, Byron Goldstein, and William 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.

[15] Alexander Bockmayr and Arnaud Courtois. Using hybrid concurrent constraint programming to model dynamic biological systems. In *18th International Conference on Logic Programming, ICLP'02, Copenhagen,*, volume 2401 of *LNCS*, pages 85–99. Springer, July 2002.

[16] Peer Bork. Is there biological research beyond systems biology ? a comparative analysis of terms. *Mol Syst Biol*, 1, May 2005.

[17] Peer Bork and Luis Serrano. Towards cellular systems in 4d. *Cell*, 121(4) :507–509, May 2005.

[18] D. L. Brutlag, A. R. Galper, and D. H. Millis. Knowledge-based simulation of DNA metabolism : Prediction of enzyme action. *Computer Applications in Biosciences*, 7(1) :9–19, 1991.

[19] Lemer C, Antezana E, Couche F, Fays F, Santolaria X, Janky R, Deville Y, Richelle J, and Wodak SJ. The amaze lightbench : a web interface to a relational database of cellular processes. *Nucleic Acids Res*, 32(Database issue) :D443–448, 2004.

[20] Stanyon CA, Liu G, Mangiola BA, Patel N, Giot L, Kuang B, Zhang H, Zhong J, and Jr. Finley RL. A drosophila protein-interaction map centered on cell-cycle regulators. *Genome Biol*, 5(12) :R96, 2004.

[21] Laurence Calzone, François Fages, and Sylvain Soliman. Biocham : an environment for modeling biological systems and formalizing experimental knowledge. *Bioinformatics*, 22(14) :1805–1807, Jul 2006.

[22] L. Cardelli. Brane calculi. In Vincent Danos and Vincent Schächter, editors, *CMSB*, volume 3082 of *LNCS*, pages 257–278. Springer, 2004.

[23] L. Cardelli. Abstract machines of systems biology. In *Transactions on Computational Systems Biology III*, Springer LNBI 3737, pages 145–168, 2005.

[24] L. Cardelli and A. D. Gordon. Mobile ambients. *Proceedings of the First International Conference on Foundations of Software Science and Computation Structure*, March 28-April 04 :140–155, 1998.

[25] N. Chabrier, M. Chiaverini, V. Danos, F. Fages, and V. Schachter. Modeling and querying biochemical interaction networks. *Theoretical Computer Science*, 325(1) :25–44, September 2004.

[26] N. Chabrier and F. Fages. Symbolic model checking of biochemical networks. In C. Priami, editor, *Proceedings of the 1st Intern. Workshop CMSB'2003*, LNCS 2602, pages 149–162. Springer-Verlag, 2003.

[27] Ming Chen and Ralf Hofestadt. Quantitative Petri net model of gene regulated metabolic networks in the cell. *In Silico Biology*, 3(3) :347–65, 2003.

[28] M.T.H. Chi, P.J. Felovich, and R. Glaser. Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5 :121–152, 1982.

[29] O. Cinquin and J. Demongeot. Roles of positive and negative feedback in biological systems. *C.R.Biol.*, 325(11) :1085–1095, 2002.

[30] J. R. Collier, N. A. M. Monk, P. K. Maini, and J. H. Lewis. Pattern formation by lateral inhibition with feedback : a mathematical model of delta-notch intercellular signalling. *Journal of Theor. Biology*, 183 :429–446, 1996.

[31] D. L. Cook, J. F. Farley, and S. J. Tapscott. A basis for a visual language for describing, archiving and analyzing functional models of complex biological systems. *Genome Biol*, 2(4) :RESEARCH0012, 2001.

[32] A. Cornish-Bowden and M.L. Cardenas. Systems biology may work when we learn to understand the parts in terms of the whole. *Biochemical Society Transactions*, 33(3) :516–519, 2005.

[33] M. Curti, P. Degano, C. Priami, and C.T. Baldari. Modelling biochemical pathways through enhanced pi-calculus. *Theoretical Computer Science*, 325(1) :111–140, 2004.

[34] Vincent Danos and Sylvain Pradalier. Projective brane calculus. In *proceedings of CMSB'04*, 2004.

[35] H. de Jong. Modeling and simulation of genetic regulatory systems : a literature review. *J Comput Biol*, 9(1) :67–103, 2002.

[36] H. de Jong, M. Page, C. Hernandez, and J. Geiselmann. Qualitative simulation of genetic regulatory networks : Method and application. In *proceedings of IJCAI*, pages 67–73, 2001.

[37] E. Demir, O. Babur, U. Dogrusoz, A. Gursoy, A. Ayaz, G. Gulesir, G. Nisanci, and R. Cetin-Atalay. An ontology for collaborative construction and analysis of cellular pathways. *Bioinformatics*, 20(3) :349–356, Feb 2004.

[38] E. Demir, O. Babur, U. Dogrusoz, A. Gursoy, G. Nisanci, R. Cetin-Atalay, and M. Ozturk. Patika : an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics*, 18(7) :996–1003, Jul 2002.

[39] J. Demongeot, M. Kaufman, and R. Thomas. Positive feedback circuits and memory. *C.R. Acad. Sci. III.*, 323(1) :69–79, 2000.

[40] P.K. Dhar, Hao Zhu, and S.K. Mishra. Computational approach to systems biology : from fraction to integration and beyond. *NanoBioscience, IEEE Transactions*, 3(3) :144– 152, Sept. 2004.

[41] D.L. Donoho. Aide-memoire. high-dimensional data analysis : The curses and blessings of dimensionality. Department of Statistics, Stanford University, August 2000.

[42] D.T.Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Chem. Phys.*, 81(25) :2340–61, 1977.

[43] R. Edwards, H. T. Siegelmann, K. Aziza, and L. Glass. Symbolic dynamics and computation in model gene networks. *Chaos*, 11(1) :160–169, Mar 2001.

[44] S. Efroni, D. Harel, and I.R. Cohen. Toward rigorous comprehension of biological complexity : modeling, execution, and visualization of thymic T-cell maturation. *Genome Research*, (13) :2485–2497, 2003.

[45] S. Eker, M. Knapp, K. Laderoute, P. Lincoln, J. Meseguer, and K. Sonmez. Pathway logic : Symbolic analysis of biological signaling. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 400–412, January 2002.

[46] Damien Eveillard, Delphine Ropers, Hidde de Jong, Christiane Branlant, and Alexander Bockmayr. A multi-site constraint programming model of alternative splicing regulation. Technical report, INRIA, May 2003.

[47] J.R. Faeder, M.L. Blinov, B. Goldstein, and W.S. Hlavacek. Rule-based modeling of biochemical networks. *Complexity*, 10 :22–41, 2005.

[48] A. Finney. Developing sbml beyond level 2 : Proposals for development. *LNCS*, 3082, 2005.

[49] J. Fisher, N. Piterman, E.J.A. Hubbard, M. Stern, and D. Harel. Computational insights into c. elegans vulval development. *Proc. Natl. Acad. Sci. USA*, 102 :1951–1956, 2005.

[50] K. D. Forbus. Qualitative process theory : twelve years after. *Artificial Intelligence*, 59 :115–123, 1993.

[51] K.D. Forbus. Qualitative process theory. *Artificial Intelligence*, 24 :85–168, 1984.

[52] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *J Comput Biol*, 7(3-4) :601–620, 2000.

[53] A. Funahashi, M. Morohashi, and H. Kitano. Celldesigner : a process diagram editor for gene-regulatory and biochemical networks. *Biosilico*, 1 :159–162, 2003.

[54] Hubert Garavel. Défense et illustration des algèbres de processus. In *Actes de l'Ecole d'été Temps Réel ETR 2003*, September 2003.

[55] A.C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon, C.M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.A. Heurtier, R.R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868) :141–7, 2002.

[56] R. Ghosh and C.J. Tomlin. Lateral inhibition through delta-notch signaling : A piecewise affine hybrid model. *LNCS*, 2034 :232–246, 2001.

[57] M.A. Gibson and E. Mjolsness. *Computational Modeling of Genetic and Biochemical Networks.*, chapter Modeling the activity of single genes., pages 1–48. MIT Press, Cambridge, MA., 2001.

[58] A. Gierer. Generation of biological patterns and form : some physical, mathematical, and logical aspects. *Prog Biophys Mol Biol*, 37(1) :1–47, 1981.

[59] A. Gierer and H. Meinhardt. A theory of biological pattern formation. *Kybernetik*, 12(1) :30–39, Dec 1972.

[60] L. Glass. Classification of biological networks by their qualitative dynamics. *J. Theor. Biol.*, 54 :85–107, 1975.

[61] B. C. Goodwin and S. A. Kauffman. *Cell to Cell Signalling : From Experiments to Theoretical Models*, chapter in Bifurcations, harmonics, and the four color wheel model of Drosophila development., pages 213–227. Academic Press, London., 1989.

[62] B. C. Goodwin and S. A. Kauffman. Spatial harmonics and pattern specification in early drosophila development. part i. bifurcation sequences and gene expression. *J Theor Biol*, 144(3) :303–319, Jun 1990.

[63] N. Guelzim, S. Bottani, P. Bourgine, and F. Képès. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, 31(1) :60–3, May 2002.

[64] Kristin C. Gunsalus, Hui Ge, Aaron J. Schetter, Debra S. Goldberg, Jing-Dong J. Han, Tong Hao, Gabriel F. Berriz, Nicolas Bertin, Jerry Huang, Ling-Shiang Chuang, Ning Li, Ramamurthy Mani, Anthony A. Hyman, Birte Sönnichsen, Christophe J. Echeverri, Frederick P. Roth, Marc Vidal, and Fabio Piano. Predictive models of molecular machines involved in caenorhabditis elegans early embryogenesis. *Nature*, 436(7052) :861–865, 2005.

[65] D. Harel. Statecharts : A visual formalism for complex systems. *Sci. Comput. Programm.*, 8 :231–274, 1987.

[66] F. Hayes-Roth, D. A. Waterman, and D. B. Lenat. *Building expert systems.* Teknowledge Series in Knowledge Engineering, 1983.

[67] K. R. Heidtke and S. Schulze-Kremer. Design and implementation of a qualitative simulation model of lambda phage infection. *Bioinformatics*, 14(1) :81–91, 1998.

[68] K.R. Heidtke and S. Schulze-Kremer. BioSim - a new qualitative simulation environment for molecular biology, 1998.

[69] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S.G.N. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, and R. Apweiler.

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.

[70] Y. Ho, A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A.R. Willems, H. Sassi, P.A. Nielsen, K.J. Rasmussen, J.R. Andersen, L.E. Johansen, L.H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B.D. Sørensen, J. Matthiesen, R.C. Hendrickson, F. Gleeson, T. Pawson, M.F. Moran, D. Durocher, M. Mann, C.W. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868) :180–183, 2002.

[71] R. Hofestädt and F. Meineke. Interactive modelling and simulation of biochemical networks. *Comput Biol Med*, 25(3) :321–334, May 1995.

[72] M. Holford, N. Li, P. Nadkarni, and H. Zhao. Vitapad : visualization tools for the analysis of pathway data. *Bioinformatics*, 21(8) :1596–1602, Apr 2005.

[73] http ://pubchem.ncbi.nlm.nih.gov/.

[74] http ://sbgn.org.

[75] http ://www.biocyc.com.

[76] http ://www.biopax.org.

[77] http ://www.ebi.ac.uk/intact/site/index.jsf.

[78] http ://www.expasy.ch/cgi bin/show_thumbnails.pl.

[79] http ://www.expasy.uniprot.org/.

[80] http ://www.framinghamheartstudy.org/.

[81] http ://www.geneontology.org/.

[82] http ://www.graphviz.org/.

[83] http ://www.hprd.org/.

[84] http ://www.ncbi.nlm.nih.gov/sites/entrez.

[85] http ://www.physiome.org.

[86] http ://www.reactome.org/.

[87] Jianghai Hu, Wei-Chung Wu, and Shankar Sastry. Modeling subtilin production in bacillus subtilis using stochastic hybrid systems. In Rajeev Alur and George J. Pappas, editors, *Proceedings of Hybrid Systems : Computation and Control, 7th International Workshop, HSCC 2004*, volume 2993 of *LNCS*, pages 417–431, Philadelphia, PA, USA, March 2004. Springer.

[88] M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano, A.P. Arkin, B.J. Bornstein, D. Bray, A. Cornish-Bowden, A.A. Cuellar, S. Dronov, E.D. Gilles, M. Ginkel, V. Gor, I.I. Goryanin, W.J. Hedley, T.C. Hodgman, J.H. Hofmeyr, P.J. Hunter, N.S. Juty, J.L. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L.M. Loew, D. Lucio, P. Mendes, E. Minch, E.D. Mjolsness, Y. Nakayama, M.R. Nelson, P.F. Nielsen, T. Sakurada, J.C. Schaff, B.E. Shapiro, T.S. Shimizu, H.D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML) : a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4) :524–531., 2003.

[89] Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, and Stein L. Reactome : a knowledge base of biologic pathways and processes. *Genome Biology*, 8(R39), 2007.

[90] T. Ideker and D. Lauffenburger. Building with a scaffold : emerging strategies for high- to low-level cellular modeling. *Trends in Biotechnology*, 21(6) :255–62, Jun 2003.

[91] Trey Ideker, Timothy Galitski, and Leroy Hood. A new approach to decoding life : Systems biology. *Annual Review of Genomics and Human Genetics*, 2 :343–372, September 2001.

[92] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8) :4569–74, 2001.

[93] N. Kam, D. Harel, H. Kugler, R. Marelly, A. Pnueli, J. A. Hubbard, and M. J. Stern. Formal modelling of c. elegans development : A scenario-based approach. *Modelling in Molecular Biology*, pages 151–173, 2004.

[94] Na'aman Kam, Irun R. Cohen, and David Harel. The immune system as a reactive system : Modeling T cell activation with statecharts. In *IEEE 2001 Symposia on Human Centric Computing Languages and Environments (HCC'01)*, 2001.

[95] M. Kanehisa and S. Goto. Kegg : Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.*, 28(1) :27–30, 2000.

[96] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics : new developments in kegg. *Nucleic Acids Res.*, 34(D354-357), 2006.

[97] K. Kappler, R. Edwards, and L. Glass. Dynamics in high dimensional model gene networks. *Signal Processing*, 83 :789–798, 2002.

[98] P. D. Karp. *Artificial intelligence & molecular biology*, chapter 8 - A Qualitative Biochemistry and Its Application to the Regulation of the Tryptophan Operon, pages 289–324. AAAI, 1993.

[99] P. D. Karp, M. Riley, S. M. Paley, A. Pellegrini Toole, and M. Krummenacker. Ecocyc : Encyclopedia of escherichia coli genes and metabolism. *Nucl. Acids Res.*, 27(1) :55–58, 1999.

[100] P.D. Karp, C.A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. Darzentas, V. Kunin, and N. Lopez-Bigas. Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, 19 :6083–89, 2005.

[101] S.A. Kauffman. *The Origins of Order : Self-Organization and Selection in Evolution.* Oxford University Press, New York., 1993.

[102] M. Kaufman and R. Thomas. Model analysis of the bases of multistationarity in the humoral immune response. *J. Theor. Biol.*, 129(2) :141–62, 1987.

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

[104] H. Kitano. Computational systems biology. *Nature*, 420(6912) :206–10, Nov. 2002.

[105] H. Kitano. Looking beyond the details : a rise in system-oriented approaches in genetics and molecular biology. *Curr. Genet.*, 41(1) :1–10, 2002.

[106] H. Kitano. Biological robustness. *Nat Rev Genet*, 5(11) :826–837, Nov 2004.

[107] 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.

[108] K. W. Kohn and M. I. Aladjem. Circuit diagrams for biological networks. *Molecular Systems Biology*, 2 :2006.0002, 2006.

[109] K.W. Kohn. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Molecular Biology of the Cell*, 10 :2703–2734, 1999.

[110] K.W. Kohn. Molecular interaction maps as information organizers and simulation guides. *Chaos*, 11 :84–97, 2001.

[111] 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.

[112] F. Kolpakov, V. Poroikov, R. Sharipov, Y. Kondrakhin, A. Zakharov, A. Lagunin, L. Milanesi, and A. Kel. Cyclonet–an integrated database on cell cycle regulation and carcinogenesis. *Nucleic Acids Res*, 35(Database issue) :D550–D556, Jan 2007.

[113] F.A. Kolpakov. BIOUML - framework for visual modeling and simulation biological systems. In *Proc. Int. Conf. Bioinf. of Genome Regulation and Structure (BGRS'2002)*, volume http ://biouml.org/publications/pdf/bgrs_2002.pdf, 2002.

[114] B. Kuipers. Qualitative simulation. *Artificial Intelligence*, 29 :289–338, 1986.

[115] B.J. Kuipers. Commonsense reasoning about causality : deriving behaviour from structure. *Artificial Intelligence*, 24 :169 – 204, 1984.

[116] B.J. Kuipers. *Qualitative reasoning : modeling and simulation with incomplete knowledge*. MIT Press, 1994.

[117] E. Lee, A. Salic, R. Kruger, R. Heinrich, and M.W. Kirschner. The roles of apc and axin derived from experimental and theoretical analysis of the wnt pathway. *PLoS Biol*, 1(1), 2003.

[118] T. J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D.W.J. Stringer-Calvert, J.D. Tenenbaum, and P.D. Karp. Biowarehouse : a bioinformatics database warehouse toolkit. *BMC Bioinformatics*, 7 :170, 2006.

[119] TI Lee, NJ Rinaldi, F Robert, DT Odom, Z Bar-Joseph, GK Gerber, NM Hannett, CT Harbison, CM Thompson, I Simon, J Zeitlinger, EG Jennings, HL Murray, DB Gordon, B Ren, JJ Wyrick, JB Tagne, TL Volkert, E Fraenkel, DK Gifford, and RA Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594) :799–804, Oct 2002.

[120] Ben Lehner, Julia Tischler, and Andrew G. Fraser. Systems biology : where it's at in 2005. *Genome Biology,*, 6 :338, 2005.

[121] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*, pages 18–29, 1998.

[122] P. Lincoln and A. Tiwari. Symbolic systems biology : Hybrid modeling and analysis of biological networks. In R. Alur and G. Pappas, editors, *Hybrid Systems : Computation and Control HSCC*, volume 2993 of *LNCS*, pages 660–672. Springer, March 2004.

[123] Edison T. Liu. Systems biology, integrative biology, predictive biology. *Cell*, 121(4) :505–506, May 2005.

[124] C. M Lloyd, M.D.B. Halstead, and P.F. Nielsen. Cellml : its future, present and past. *Prog Biophys Mol Biol*, 85(2-3) :433–450, 2004.

[125] Y. Maki, D. Tominaga, M. Okamoto, S. Watanabe, and Y. Eguchi. Development of a system for the inference of large scale genetic networks. In R. B. Altman, A.K. Dunker, L. Hunter, K. Lauderdale, and T.E. Klein, editors, *Proc. Pac. Symp. Biocomput (PSB'01)*, volume 6, pages 446–458. Singapore, World Scientic Publishing., 2000.

[126] G. Marnellos and E. Mjolsness. A gene network approach to modeling early neurogenesis in drosophila. In *PSB'98*, volume 3, pages 30–41, 1998.

[127] H. Matsuno, Y. Tanaka, H. Aoshima, A. Doi, M. Matsui, and S. Miyano. Biopathways representation and simulation on hybrid functional Petri net. *In Silico Biol*, 3(3) :389–404, 2003.

[128] H. H. McAdams and A. Arkin. It's a noisy business ! genetic regulation at the nanomolar scale. *Trends Genet*, 15(2) :65–69, Feb 1999.

[129] P. Mendes. Gepasi : a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Comput Appl Biosci*, 9(5) :563–571, Oct 1993.

[130] L. Mendoza, D. Thieffry, and E.R. Alvarez-Buylla. Genetic control of flower morphogenesis in arabidopsis thaliana : a logical analysis. *Bioinformatics*, 15(7-8) :593–606, 1999.

[131] T. Mestl, E. Plahte, and S. W. Omholt. A mathematical framework for describing and analysing gene regulatory networks. *J Theor Biol*, 176(2) :291–300, Sep 1995.

[132] S. Meyers and P. Friedland. Knowledge-based simulation of genetic regulation in bacteriophage lambda. *Nucleic Acids Res*, 12(1 Pt 1) :1–9, Jan 1984.

[133] R. Milner. What's in a name ?

[134] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and Alon U. Network motifs : simple building blocks of complex networks. *Science*, 298(5594) :824–7, 2002.

[135] B. Mishra. A symbolic approach to modeling cellular behavior. In S. Sahni, V. K. Prasanna, and U. Shukla, editors, *High Performance Computing*, volume 2552 of *LNCS*, pages 725–732. Springer, 2002.

[136] S.L. Moodie, A. Sorokin, I. Groyanin, and P. Ghazal. A graphica notation to describe the logical interactions of biological pathways. *Journal of Integrative Bioinformatics*, 3(2) :36, 2006.

[137] M. Nagasaki, A. Doi, H. Matsuno, and S. Miyano. Genomic object net : a platform for modeling and simulating biopathways. *Applied Bioinformatics*, 2003.

[138] K. Noda, A. Shinohara, M. Takeda, S. Matsumoto, S. Miyano, and S. Kuhara. Finding genetic network from experiments by weighted network model. *Genome Informatics*, 9 :141–150, 1998.

[139] 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.

[140] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. Kegg : Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 27(1) :29–34, Jan 1999.

[141] Maureen A. O'Malley and John Dupré. Fundamental issues in systems biology. *BioEssays*, 27(12) :1270–1276, 2005.

[142] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA., 1988.

[143] D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17 Suppl 1 :S215–S224, 2001.

[144] Andrew Phillips and Luca Cardelli. A graphical representation for the stochastic pi-calculus. In *proceedings of Concurrent Models in Molecular Biology : BioConcur*, number http ://research.microsoft.com/ãphillip/, 2005.

[145] I. Pirson, N. Fortemaison, C. Jacobs, S. Dremier, J.E. Dumont, and C. Maenhaut. The visual display of regulatory information and networks. *Trends in Cell Biology*, 10(10) :404–408, 2000.

[146] Ute Platzer and Hans-Peter Meinzer. Simulation of genetic networks in multicellular context. In J.D. Kim and M.T. Polani, editors, *Fifth German workshop on Artificial Life : Abstracting and Synthesizing the Principles of Living Systems*, pages 43–51. Akad. Verl.-Ges., 2002.

[147] R. Puzone, B. Kohler, P. Seiden, and F. Celada. Immsim, a flexible model for in machina experiments on immune system responses. *Future Gener. Comput. Syst.,*, 18(7) :961–972, 2002.

[148] A. Regev, E.M. Panina, W. Silverman, L. Cardelli, and E. Shapiro. Bioambients : An abstraction for biological compartments. *Theor. Comput. Sci.*, 325(1) :141–167, Aug 2004.

[149] A. Regev, W. Silverman, and E. Shapiro. Representation and simulation of biochemical processes using the $\pi$-calculus process algebra. In *Proceedings of PSB'01*, pages 459–470, 2001.

[150] A. Regev, W. Silverman, and E. Shapiro. Representation and simulation of biochemical processes using the pi-calculus process algebra. *PSB2001*, 6 :459–470, 2001.

[151] Special Industry Report. The bioinformatics gold rush. *Scientific American*, July 2000.

[152] Karen Sachs, David Gifford, Tommi Jaakkola, Peter Sorger, and Douglas A. Lauffenburger. Bayesian network approach to cell signaling pathway modeling. In *Sci. STKE*, volume 148, 2002.

[153] L. Sánchez and D. Thieffry. A logical analysis of the drosophila gap-gene system. *J. Theor. Biol.*, 211(2) :115–141, 2001.

[154] L. Sánchez, J. van Helden, and D. Thieffry. Establishement of the dorso-ventral pattern during embryonic development of drosophila melanogaster : a logical analysis. *J. Theor. Biol.*, 189(4) :377–389, 1997.

[155] C. Sansom. Systems biology : Will it work ? *Systems Biology, IEE*, 2(1) :1–4, March 2005.

[156] SS Shen-Orr, R Milo, S Mangan, and U Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet*, 31(1) :64–8, May 2002.

[157] T. Shimada, M. Hagiya, M. Arita, S.-Y. Nishizaki, and Chew Lim Tan. Knowledge-based simulation of regulatory action in lambda phage. volume 4, pages 511–523, Washington, DC, USA, 1995. IEEE Computer Society.

[158] I. Shmulevich, I. Gluhovsky, R.F. Hashimoto, E.R. Dougherty, and W. Zhang. Steady-state analysis of genetic regulatory networks modelled by probabilistic boolean networks. *Comparative and Functional Genomics*, 4 :601–608, 2003.

[159] S. Smidtas, V. Schächter, and F. Képès. The adaptive filter of the yeast galactose pathway. *J Theor Biol*, 242(2) :372–381, Sep 2006.

[160] S. Smidtas and A. Yartseva. Rooting a graph by the environment interface applied to heterogeneous interaction network of the yeast. *Acta Biotheoretica*, in press.

[161] S. Smidtas, A. Yartseva, V. Schächter, and F. Képès. Model of interactions in biology and application to heterogeneous network in yeast. *C R Biol.*, 329(12) :945–52, 2006.

[162] P Smolen, DA Baxter, and JH Byrne. Modeling transcriptional control in gene networks - methods, recent results, and future directions. *Bull Math Biol*, 62 :247–292, 2000.

[163] E.H. Snoussi. Qualitative dynamics of a piecewise-linear differential equations : a discrete mapping approach. *Dynamics and stability of Systems*, 4 :189–207, 1989.

[164] E.H. Snoussi and R. Thomas. Logical identification of all steady states : the concept of feedback loop caracteristic states. *Bull. Math. Biol.*, 55(5) :973–991, 1993.

[165] D. Thieffry, M. Colet, and R. Thomas. Formalisation of regulatory networks : a logical method and its automatization. *Math. Modeling Scientific Comput.*, 2 :144–151, 1993.

[166] D. Thieffry and D. Romero. The modularity of biological regulatory networks. *Biosystems*, 50(1) :49–59., 1999.

[167] D. Thieffry and R. Thomas. Dynamical behaviour of biological regulatory networks - II. immunity control in bacteriophage lambda. *Bull. Math. Biol.*, 57(2) :277–297, 1995.

[168] R. Thomas. Boolean formalization of genetic control circuits. *J. Theor. Biol.*, 42 :563–585, 1973.

[169] R. Thomas. Regulatory networks seen as asynchronous automata : A logical description. *J. theor. Biol.*, 153 :1–23, 1991.

[170] R. Thomas and R. d'Ari. *Biological Feedback*. CRC Press, 1990.

[171] R. Thomas, A.M. Gathoye, and L. Lambert. A complex control circuit. regulation of immunity in temperate bacteriophages. *Eur. J. Biochem.*, 71(1) :211–227, 1976.

[172] R. Thomas and M. Kaufman. Multistationarity, the basis of cell differentiation and memory. I. & II. *Chaos*, 11 :170–195, 2001.

[173] R. Thomas, D. Thieffry, and M. Kaufman. Dynamical behaviour of biological regulatory networks - I. biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull. Math. Biol.*, 57(2) :247–276, 1995.

[174] R. B. Trelease, R. A. Henderson, and J. B. Park. A qualitative process system for modeling nf-kappab and ap-1 gene regulation in immune cell biology research. *Artif Intell Med*, 17(3) :303–321, Nov 1999.

[175] P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J.M. Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770) :623–7, 2000.

[176] David W. Ussery and Lars Juhl Jensen. Systems biology : in the broadest sense of the word. *Environmental Microbiology,*, 7(4) :482, April 2005.

[177] R. Valk. Self-modifying nets, a natural extension of Petri nets. *LNCS*, 62 (ICALP 78) :464–476, 1978.

[178] J. van Helden, A. Naim, C. Lemer, R. Mancuso, M. Eldridge, and S. J. Wodak. From molecular activities and processes to biological function. *Brief Bioinform*, 2(1) :81–93, Mar 2001.

[179] J. van Helden, A. Naim, R. Mancuso, M. Eldridge, L. Wernisch, D. Gilbert, and S. J. Wodak. Representing and analysing molecular and cellular function using the computer. *Biol. Chem.*, 381 :921–935, 2000.

[180] E. O. Voit and M. Savageau. Equivalence between S-systems and Volterra systems. *Mathematical Biosciences*, 78 :47–55, 1986.

[181] Jrg R. Weimar. Cellular automata approaches to enzymatic reaction networks. In *5th International Cp,ference on Cellular Automata for Research Industry ACRI*, pages 294–303. Springer, LNCS 2493, 2002.

[182] D.S. Wishart, R. Yang, D. Arndt, P. Tang, and J. Cruz. Dynamic cellular automata : an alternative approach to cellular simulation. *In Silico Biol.*, 5(2) :139–161, 2005.

[183] O. Wolkenhauer. Systems biology : the reincarnation of systems theory applied in biology ? *Brief Bioinform*, 2(3) :258–270, Sep 2001.

[184] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R.Y. Pinter, U. Alon, and H. Margalit. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci U S A*, 101(16) :5934–5939, Apr 2004.