2007EVRY0024

Modélisation et analyse, globale et locale, de réseaux d'interactions biologiques hétérogènes (RIBH) appliqué à la Levure.

Serge SMIDTAS

Thèse de doctorat

Université d'Evry, Val d'Essonne

Travail d'Avril 2002 - Octobre 2007

Thèse soutenue et le 15 / 11 / 2007 devant le jury composé de :

François KEPES, CoDirecteur
Vincent SCHACHTER, CoDirecteur
Jean Pierre MAZAT, Président du jury
Olivier MARTIN, Rapporteur
Olivier POCKES, Rapporteur
Paul BOURGINE, Examinateur
Hanna KLAUDEL, Examinateur

Le jury a apprécié l'exposé oral structuré, clair, pédagogique qui éclairait la cohérence de l'ensemble des travaux du candidat. Il a montré sa maîtrise à la fois des questions biologiques, de modélisation et d'outils informatiques, pendant l'oral aussi que pendant la réponse au jury. Son exposé a aussi révélé la créativité que l'on trouvait dans son travail.

Admis
Très honorable.
J.P. MAZAT

Modélisation et analyse, globale et locale, de Réseaux d'Interactions Biologiques Hétérogènes (RIBH) appliqué à la Levure.

Table des matières

CHAPITRE I – INTRODUCTION GENERALE **7**
1) Introduction *7*
2) Présentation de la thèse *10*

CHAPITRE II - ÉTAT DE L'ART **13**
1) Introduction *13*
2) Méthodologie: modélisation des RIBH *13*
A) Organismes 13
B) Obtention des RIBH 17
Prédictions de Réseaux homogènes 26
C) Intégration de données 26
3) Études des RIBH *30*
A) Topologie 31
B) Dynamique des RIBH 42
4) Conclusion sur l'état de l'art *54*

CHAPITRE III – INTEGRATION DE DONNEES POUR LA CONSTRUCTION DE RIBH **57**

CHAPITRE IV – CORRELATIONS GLOBALES DE RIBH **65**

CHAPITRE V – MODELE BIPARTITE & TOPOLOGIE DES RIBH **83**

CHAPITRE VI – DYNAMIQUE DE RIBH : EXEMPLE DU GALACTOSE **93**

CHAPITRE VII – CONCLUSION ET PERSPECTIVES **107**
1) Conclusion *107*
2) Discussion *109*
3) Comparaison avec l'état de l'art de 2006 ..... *111*
4) Perspectives *112*

CHAPITRE VIII – REFERENCES **115**
1) Bibliographie *105*
1) Quelques liens *116*

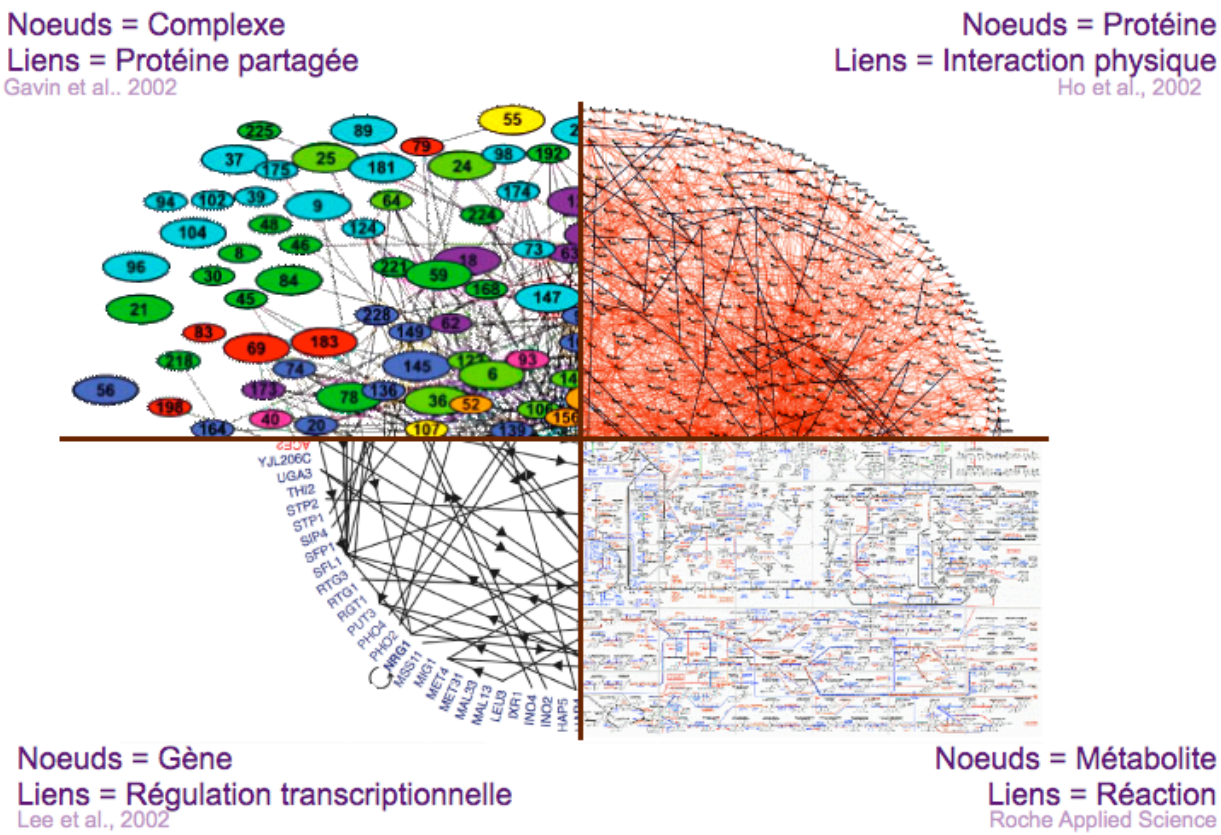
INFORMATIONS SUR L'AUTEUR **129**
SELECTION D'AUTRES PUBLICATIONS DE L'AUTEUR **131**
REMERCIEMENTS **133**
ANNEXES (DISPONIBLES EN LIGNE) **135**

Chapitre I – Introduction générale

Introduction

La biologie moléculaire a conduit à une bonne connaissance des détails des mécanismes moléculaires dans les organismes. 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 régulations impliqués dans la physiologie cellulaire. La biologie a conduit la majorité des biologistes moléculaires à se focaliser sur des gènes en particulier, plutôt que sur les systèmes. De nombreuses questions se posent: Comment sont structurés les réseaux? Quelle est la fonction de ces architectures de réseaux de régulations et de ces modules, quelles sont leurs propriétés dynamiques? Quels sont les principes sous-jacents d'organisation des systèmes biologiques? À quel point les comportements émergents d'événements moléculaires sont-ils robustes? Comment la cellule s'adapte à son environnement et contrôle le bruit dans ses réseaux de régulation? 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 compétences de biologistes, physiciens, informaticiens, mathématiciens, et ingénieurs sont nécessaires. Le chemin d'une nouvelle approche des sciences de la vie appelée Biologie des Systèmes (*Systems Biology*) se met en place. Pour parvenir au développement d'une telle approche, de nouvelles techniques et des outils conceptuels sont nécessaires.



La densité des réseaux d'interactions entre entités biochimiques dans la cellule rend l'exploration de ces réseaux difficile. Depuis le séquençage complet d'organismes, de nombreuses interactions ont été mises en évidence dans la cellule. La densité des réseaux construits à partir de ces interactions pose de nombreux problèmes, tant pour caractériser les réseaux et les distinguer, que pour les modéliser afin comprendre comment ces réseaux fonctionnent de concert.

L'objectif de ce travail de thèse est l'étude de la dynamique de réseaux hétérogènes, c'est-à-dire des réseaux composés de différents types d'interactions biologiques.

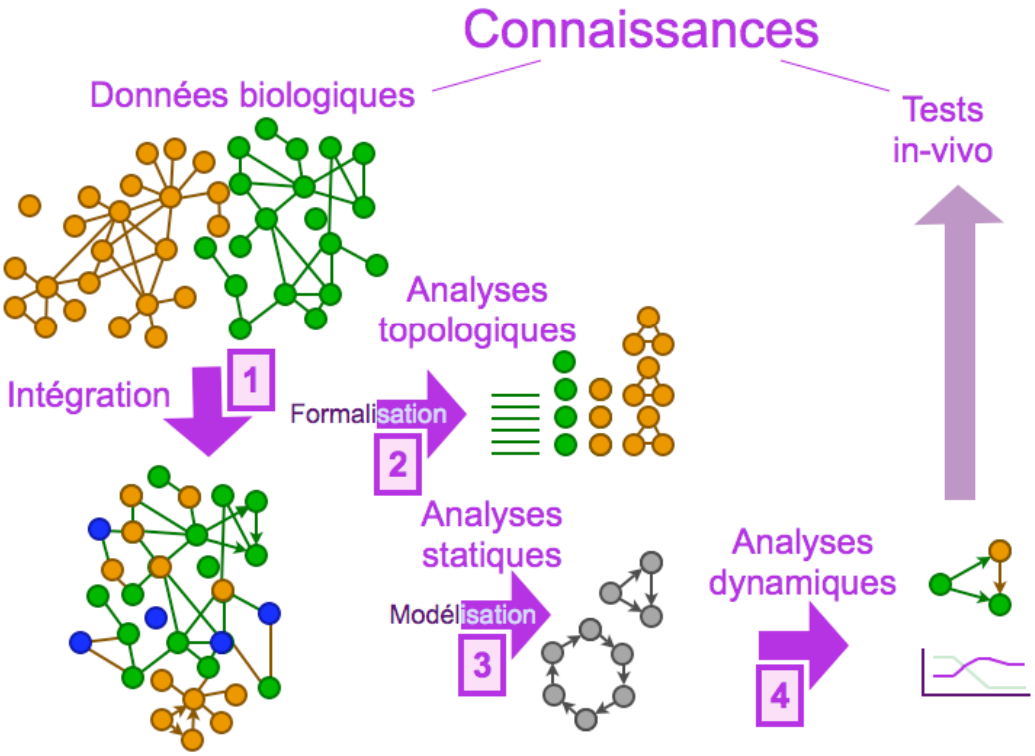
La structure topologique globale de réseaux dits *homogènes*, constitués d'un seul type d'interaction biologique donné, a été beaucoup étudiée. De nombreux réseaux réels, comme le réseau des interactions entre protéines, se sont révélés posséder une distribution du nombre d'interactions par protéine qui suit une loi de puissance, contrairement à des réseaux aléatoires pour lesquelles la loi suivie est exponentielle. Cette propriété, sur laquelle nous reviendrons, permet par exemple de distinguer les réseaux, de proposer des hypothèses sur leur mise en place au cours de l'évolution, ou

d'analyser leur robustesse à différentes modifications. L'analyse de la structure topologique locale a montré que, comparés à des modèles de réseaux aléatoires, le réseau de régulation transcriptionnel, constitué des liens de régulations entre gènes, contient plus d'ensembles de 3 gènes dont les liens de régulation forment un triangle. De tels modules peuvent être généralisés à d'autres topologies. Leur comportement, sur lequel porte la pression de sélection, doit aussi être décortiqué.

Toutefois, il est connu que la dynamique du vivant utilise des processus biologiques faisant intervenir plusieurs types d'interactions biologiques et nous devons donc appréhender des réseaux hétérogènes. Par exemple, le métabolisme (ensemble d'interaction de transformation biochimiques) est couplé à la régulation de la transcription des gènes et aux interactions physiques entre protéines: les enzymes sont des complexes protéiques assemblés grâce à plusieurs interactions protéines-protéines, les enzymes régulent les réactions métaboliques, enfin les facteurs de transcription qui régulent l'expression des protéines des enzymes sont aussi des complexes. Peut-on étudier le réseau de chacun des types d'interactions indépendamment? De quelle manière, tous ces types d'interactions s'influence globalement? Ces interactions sont-elle corrélées entre elles? et quantitativement comment? Quel est l'intérêt de chaque type de ces interactions dans le schéma global? Comment la cellule fait intervenir plusieurs types d'interactions pour implémenter une fonction?

Mon travail a donc pour objet l'étude statique et dynamique des réseaux biologiques hétérogènes. Il a été construit principalement autour de l'organisme modèle *Saccharomyces cerevisiae* (levure). Cet organisme a été la source de très nombreuses données expérimentales, tant sur son transcriptome que son protéome, et qui convient bien de ce fait à une analyse de son réseau hétérogène d'interactions. Par ailleurs, c'est un eucaryote modèle très bien décrit dans la littérature.

Présentation de la thèse



Boucle d'analyse et plan de thèse. Nous partons des connaissances biologiques que l'on considère ici sous la forme d'ensembles de données biologiques de types différents. Dans une première partie, nous verrons comment on peut intégrer tous ces types de données ensemble, et les problèmes que cela pose. Puis, une fois le réseau hétérogène construit, nous étudierons la topologie de ce réseau hétérogène afin notamment de déterminer si les différentes interactions coopèrent au sein du réseau. Nous rechercherons des motifs fonctionnels dans le réseau au moyen d'un modèle parfaitement adapté aux réseaux hétérogènes. Il s'agit là d'une approche d'analyse statique, c'est-à-dire qui consiste à tirer des conclusions sur des comportements dynamiques sans faire de simulation au cours du temps. Enfin, nous étudierons la dynamique d'un des modules trouvés, et notamment, comment modéliser dynamiquement les différents types d'interactions qui les composent. Dans la dernière partie, dite de conclusion et perspectives, nous verrons comment il est possible de revenir en fin de compte aux connaissances biologiques, en étudiant les propriétés de modules in-vivo.

L'idée au début de mon travail de doctorat, en 2002, était de rechercher de petits modules hétérogènes avec une fonction particulièrement intéressante, dans l'amas de données d'interaction, pour servir d'exemples comme modules de base pour construire des circuits artificiels de protéines qui pourraient être implémentés dans une cellule. La biologie synthétique se réduisait alors aux seuls travaux de Ron Weiss ou de Leibler, Gardner, Collins et Serrano et de leurs équipes.

Dans le chapitre II, nous commencerons par présenter l'état de l'art portant sur l'étude des réseaux biologiques hétérogènes, puis nous présenterons le travail, et l'apport de cette thèse dans ce domaine.

Pour étudier les réseaux hétérogènes, pour élargir et généraliser l'approche de Guelzim et al. et de Alon et al. de 2002 afin d'intégrer plus d'hétérogénéités dans les réseaux, une stratégie globale en quatre étapes a été mise en place (Smidtas et al., 2004a). L'analyse de réseaux hétérogènes comporte une première étape d'intégration de données. Puis un cadre de modélisation adéquat est établi. Ce cadre permet d'étudier dans le réseau reconstruit, les structures topologiques macroscopiques (globales) et microscopiques (locales). Enfin, la dynamique locale de petits modules peut être approfondie. Ces quatre étapes constituent les quatre chapitres suivants :

Pour combler le manque de logiciel d'intégration (regroupement) de données biologiques et remédier à la difficulté d'accès aux données stockées dans des formats propriétaires, nous avons conçu et développé un outil appelé Cyclone. Cet outil s'interface avec BioCyc, la collection de base de données génomiques et de voie métabolique la plus complète à ce jour. Cyclone permet notamment d'interroger ces connaissances biologiques à grande échelle, facilement. Le code source écrit en langage Java est public. Cet outil d'intégration sera présenté au chapitre III qui intègre le papier publié dans le journal *Bioinformatics* sous le titre *Cyclone: a Java workbench designed to manipulate Pathway Genome Databases* (Le Fèvre et al., 2005, 2006).

Une fois les données d'interactions intégrées, nous avons pu étudier l'interaction entre les réseaux biologiques homogènes dans le réseau hétérogène. La recherche de structures de niveau supérieur (macroscopiques) a pu être conduite. Une analyse de statistiques globales est présentée au chapitre IV. Elle montre en effet qu'en observant certaines propriétés topologiques comme la mesure de distribution de distances dans le réseau, les interactions protéines-protéines et la régulation transcriptionnel travaillent globalement de concert. Les résultats ont été publiés par ailleurs sous le titre *Property-*

driven statistics of biological networks (Schachter et al., 2005; Bourguignon et al., 2006).

Pour analyser les modules hétérogènes, à des niveaux intermédiaires (mésoscopiques), le Biological Interaction Browser, 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 motif spécifique au modèle sous-jacent, basé sur un graphe bipartite avec des nœuds et des liens typés. Ces travaux constituent le chapitre V, travaux publiés par ailleurs sous le titre *Model of Interactions in Biology and Application to Heterogeneous Network in Yeast* (Smidtas et al., 2004b, 2006). 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.

Pour illustrer l'utilité biologique de notre approche de modélisation nous avons étudié plus en détail le mécanisme d'un exemple concret d'un des motifs identifié lors des analyses topologiques. 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 adaptation de la levure à ce sucre. Ce travail, présenté au Chapitre VI, a été publié dans *The adaptative filter of the yeast galactose pathway* (Smidtas et al., 2006).

Enfin, au chapitre VII, nous conclurons sur les apports au cours de ma thèse, et replacerons ces travaux dans le contexte de la recherche devenue aujourd'hui très concurrentielle sur les réseaux biologiques. La perspective d'utiliser les résultats obtenus dans cette thèse pour construire des réseaux artificiels sera abordée.

Chapitre II - État de l'art

Introduction

Dans ce chapitre, nous allons présenter l'état de l'art de l'analyse des réseaux d'interaction biologique hétérogènes (RIBH) en 2002 et au cours de mon doctorat qui a suivi. Dans une première partie, nous aborderons les méthodes, en commençant par présenter les organismes étudiés. Nous définirons les *réseaux homogènes* qui ne font intervenir qu'un seul type d'interaction biologique et la manière d'intégrer et de construire des réseaux *hétérogènes*, composés d'au moins deux types d'interactions biologiques différents. Dans la partie suivante, nous présenterons les résultats des études sur les RIBH, tout d'abord portant sur leurs topologies, puis sur leurs dynamiques. À chaque fois nous commencerons par présenter les principales études conduites sur les réseaux homogènes puis nous aborderons le cas des réseaux hétérogènes.

Méthodologie: modélisation des RIBH

A) Organismes

On présente ici les organismes sur lesquels la plus grande majorité des études biologiques portent et sur lesquelles notre travail a porté.

La levure

C'est la levure qui est l'organisme modèle de référence, sur lequel le plus d'expériences ont été conduites, et donc pour lequel le plus de données ont été fournies. En 1996, c'est le premier organisme eucaryote dont le génome est entièrement séquencé (Goffeau et al., 1996). *Saccharomyces cerevisiae* que l'on appelle couramment levure, (nom créé pour une souche de levure observée sur du malt en 1837), est probablement l'espèce domestiquée la plus ancienne. Elle vit sur des sucres et a été employée pour le brassage

de bière dans Sumeria et Babylone, il y a déjà plus de huit millénaires. En parallèle, *S. cerevisiae* a été employée dans la culture du raisin en Géorgie et pour lever la pâte en Egypte. Le nom de Levure vient du latin *levare*, faisant référence à la capacité de faire lever le pain en produisant du CO_2 en condition anaérobique et de fermenter le sucre. En 1680 pour la première fois elle est observée au microscope (voir figure ci-dessous), et en 1857 Louis Pasteur (qui étudie la bière) corrèle le processus de fermentation au métabolisme de la levure. En 1877, le mot inspiré du grec *enzyme* apparaît, qui signifie littéralement 'dans la levure'. Son étude depuis des siècles, et son potentiel qui a été exploité durant des milliers d'années font de la levure un des organismes dont les processus cellulaires sont les mieux connus. Par ailleurs, *S. cerevisiae* et d'autres levures sont des organismes modèle pour comprendre d'autres organismes et elles sont à l'origine d'une grande quantité d'applications industrielles et médicales bénéficiant à la vie humaine comme support à la production de vaccin contre l'Hépatite B dès 1980.

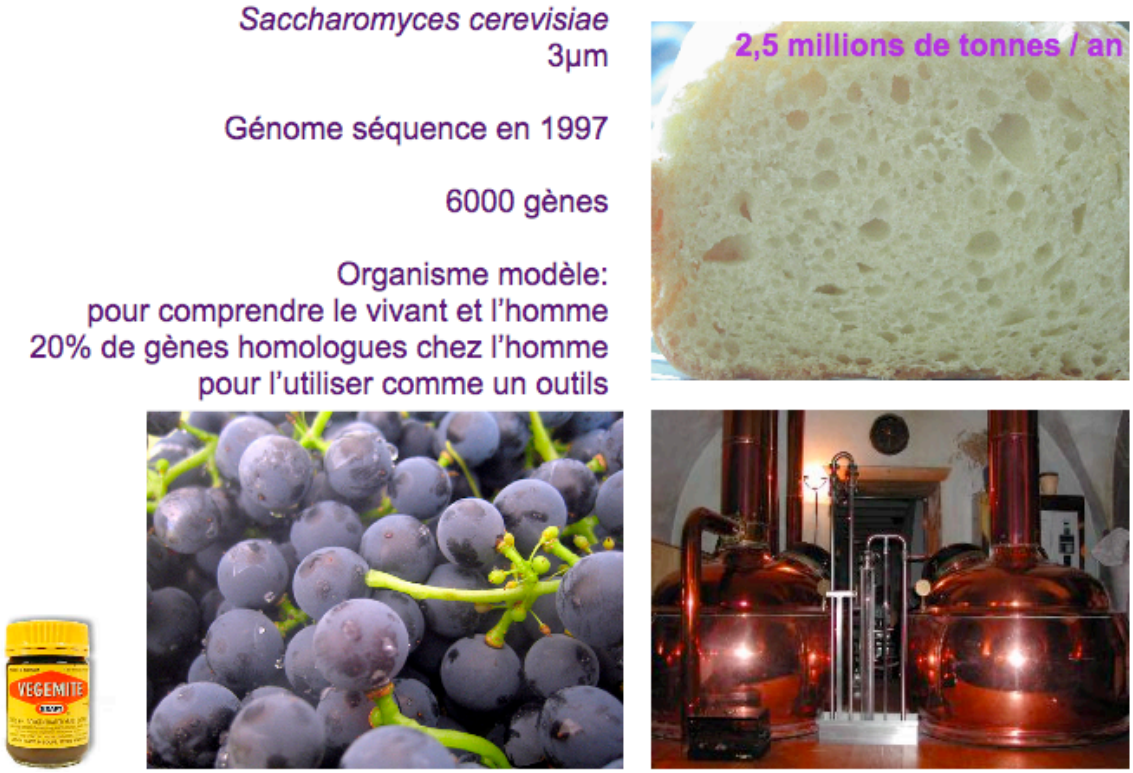


Illustration de levure, et de son utilisation source : wikipedia

Le nom latin scientifique de la levure est *Saccharomyces cerevisiae*. L'organisme mesure 3µm. Pour donner une idée de la taille, dans $1cm^3$ de levure, si on met bout à bout à bout toutes les cellules, on peut faire le tour de la terre. Le génome complet de la levure a été

séquencé en 1997. Depuis, de nombreuses innovations, l'automatisation, et la robotisation ont permis de mettre en places des méthodes expérimentales qui permettent de produire à haut débit, c'est-à-dire, à l'échelle de l'organisme, des connaissances biologiques qui portent sur l'existence d'interactions entre entités biologiques. L'organisme comporte 6000 gènes, et ce sera le nombre moyen des différentes entités manipulées dans la suite de cette thèse dans les réseaux biologiques. L'habitat naturel de la levure est sur la peau de raisin. Sa présence est révélée par une pellicule blanche que l'on peut voir sur le raison parfois. La levure est intéressante car c'est un organisme modèle pour comprendre comment fonctionnent les cellules vivantes, et l'homme, en effet 20% des gènes de la levure ont une homologue chez l'homme. Enfin la levure est un outils pour l'homme. 2,5 millions de tonnes de levure de panifaction sont produites chaque année. La levure sert aussi à produire les alcools consommés par l'homme, notemment la bière. Enfin, la levure est un complément alimentaire pour l'homme et l'animal, qui contient 50% de protéines. On en trouve dans des petits pots dans les pays anglosaxons, et cela se consomme comme de la confiture.

Escherichia coli

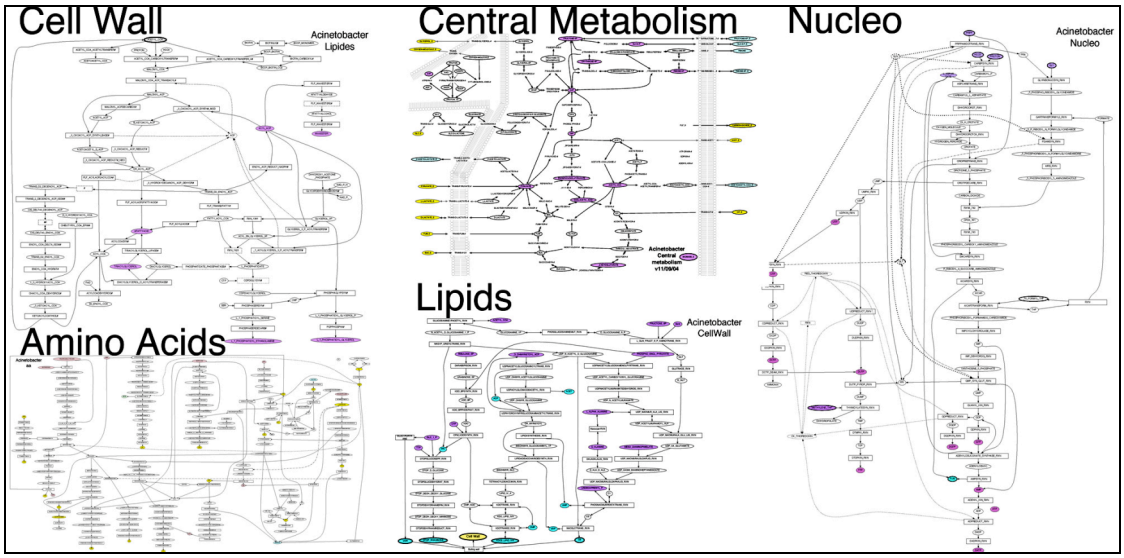
Escherichia coli, autrement appelé colibacille ou *E. coli*., est une bactérie intestinale des mammifères très commune chez l'être humain. Elle a été découverte en 1885 par Théodore Escherich, dans des selles de nourrissons. *E. coli* constitue tout au long de la vie de l'hôte, l'espèce bactérienne dominante de la microflore anaérobie facultative de l'intestin. Cependant, certaines souches d´*E.coli* peuvent être pathogènes, certaines souches de *E. coli.* sont associées à des pathologies très diverses (y compris extra-intestinales), diarrhées, gastro-entérites, infections du tractus urinaire etc. Mis à part cela, la plupart des souches d´ *E. coli* sont bénéfiques et même essentielles. *E. coli* est un des organismes vivant le plus étudié à ce jour : en effet, l'ancienneté de sa découverte et sa culture facile (division cellulaire toutes les 20 minutes à 37 °C dans un milieu riche) en font un outil d'étude aisé. C'est un organisme procaryote modèle qui permet d'étudier de nombreux mécanismes de biologie moléculaire, et sur lequel de nombreuses données sont disponibles. Ci-dessous, figure une illustration des composants principaux d'*E. coli.*

Modèle physique d'E.coli. En vert, l'ADN, en noir, l'ARN, en Jaune clair, les protéines, en blanc les ribosomes. (Sources : Yartseva et al., 2006).

Acinetobacter baylyi

Acinetobacter baylyi est une bactérie dont la découverte ne remonte qu'en 1911 et qui a été redécouverte plus tard en 1954. C'est en 1960 que la souche ADP1 est découverte et qui est en fait un mutant de laboratoire obtenu à partir d'une autre souche BD4 de la bactérie *Acinetobacter* caractérisée pour la finesse de sa capsule et sa compétence naturelle pour être transformée, c'est-à-dire sa compétence à intégrer de l'ADN synthétique qu'on lui présente. Ces transformations permettent d'en étudier les phénotypes, et les mécanismes biochimiques. Chez l'homme les bactéries *Acinetobacter* sont principalement responsables d'infections nosocomiales (infections contractées à l'hôpital), notamment chez des patients affaiblis (traumatismes multiples, cancer, immunodépression...). Les acinetobactéries sont ainsi responsables de septicémies, de méningites, d'endocardites, de suppurations (abcès du cerveau, abcès du poumon, abcès de la thyroïde, surinfections des plaies d'origine traumatique ou chirurgicale, lésions purulentes de l'œil...), de pneumopathies, d'infections urinaires... En France, l'espèce la plus souvent incriminée est la *Acinetobacter baumannii* qui à elle seule représente plus de 90 pour cent des souches d'origine hospitalière. Nous reviendrons sur les mécanismes

biochimiques de la bactérie *Acinetobacter* que nous avons prédit (voir figure ci-dessous).



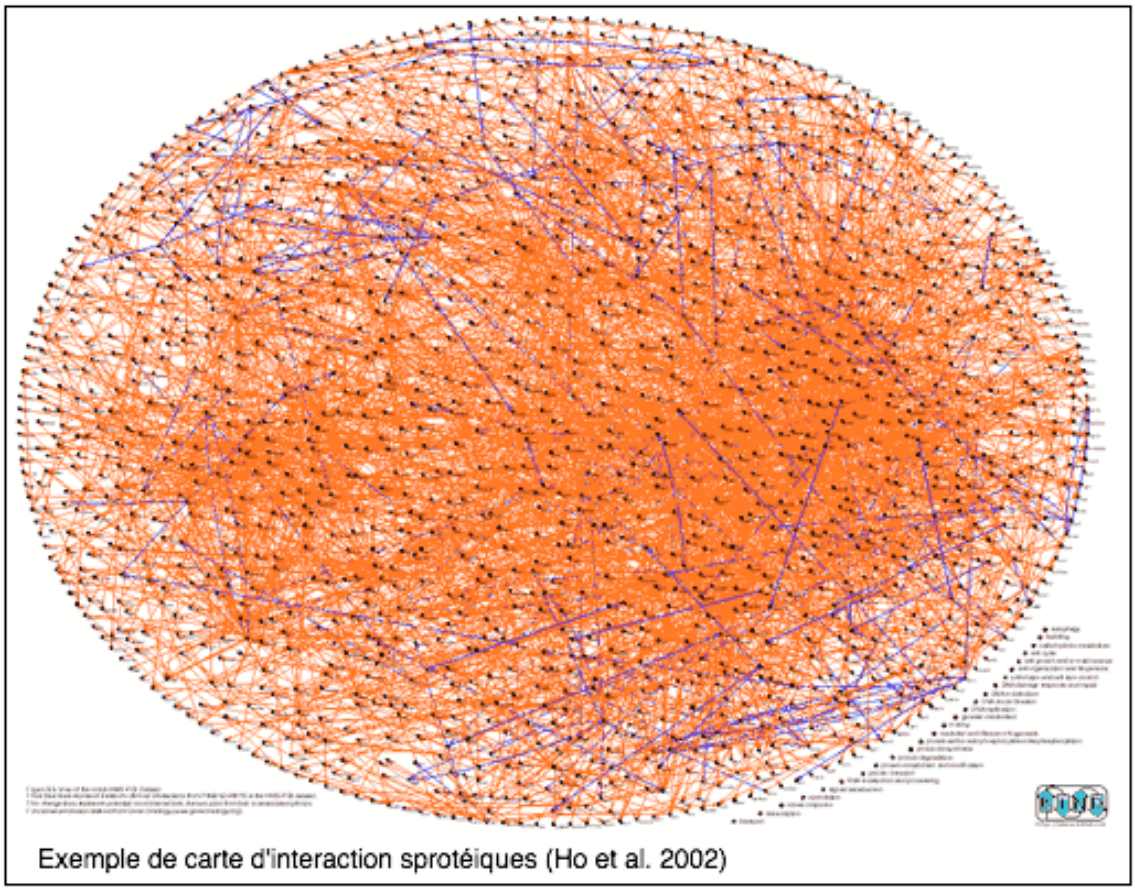
Métabolisme de la bactérie *Acinetobacter baylyi* ADP1. La carte est décomposée en différentes parties : 'Cell Wall' métabolisme des parois cellulaires, 'Amino Acids', métabolisme des acides aminés, 'Central Metabolism' , métabolisme central, 'Lipids', métabolisme des lipides, 'Nucleo', métabolisme des nucléosides et des nucléotides. D'après (Barbe et al., 2004) et (Durot et al., 2005)

B) Obtention des RIBH

Nous allons brièvement présenter ci-dessous un par un les différents type d'interactions biologiques. À chaque fois, nous commencerons par définir le type d'interaction et présenter les principaux types d'expériences biologiques qui permettent de caractériser ces interactions, puis nous présenterons la méthode de représentation associée. Chaque réseau homogène correspondant, constitué des interactions d'un type, est modélisé en général par un graphe (voire variants ou extensions). Les interactions protéine-protéine. Pour plus de sources de données biologiques, voir (Cary, Bader et al. 2005; Bader, Cary et al. 2006).

Les réseaux d'interactions protéine-protéine

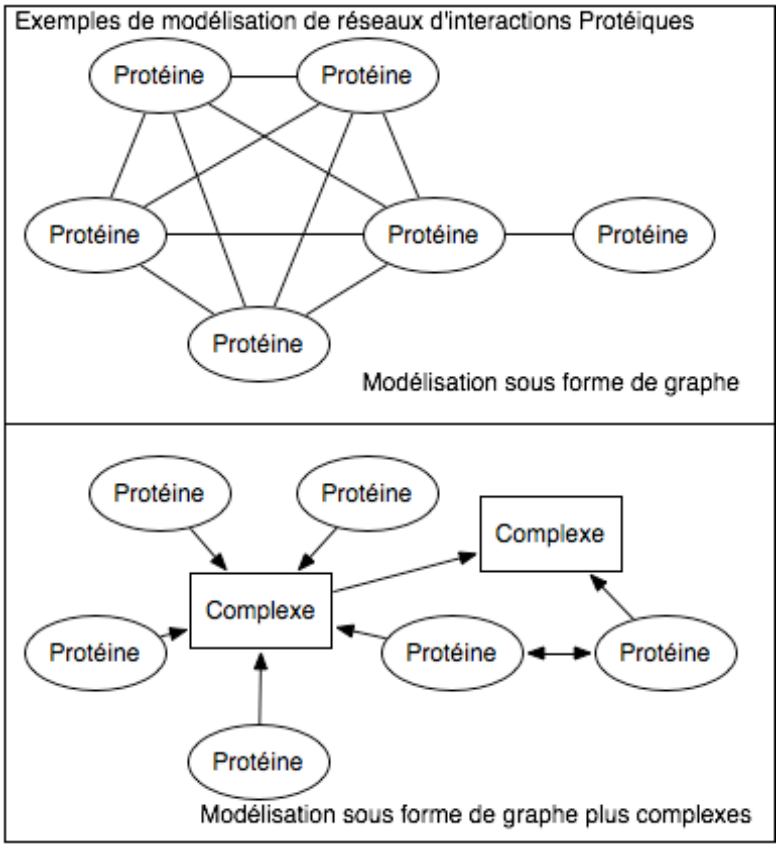
Les *interactions protéine-protéine* représentent les associations stables ou transitoires entre les protéines dans un environnement cellulaire donné. Il s'agit d'interactions entre deux protéines, ou entre plusieurs protéines qui forment des complexes. Les réseaux d'interactions protéine-protéine ont été construits pour plusieurs organismes, y compris pour des virus (McCraith et al., 2000), des procaryotes comme *H. Pylori* (Rain et al., 2001) et des eucaryotes comme la levure à partir de compilation de la littérature (Mewes, 2002), de systèmes de double hybride (Uetz et al., 2000; Chiba et al. 2001; Ito et al., 2001), ou de purification de complexes et de spectrométrie de masse (Gavin et al., 2002; Ho et al., 2002). Récemment, des cartes d'organismes multicellulaires ont été publiées (Giot et al., 2003; Barabasi et Oltvai, 2004). Les versions actuelles de ces cartes sont incomplètes et ont un fort taux de faux positifs (Bork, 2002; von Mering et al., 2002; Hoffmann et valencia, 2003; Bork et al. 2004). La couverture, la fiabilité et les biais de ces technologies à haut-débit ont été discutés dans plusieurs articles (Schachter 2002; von Mering, Krause et al. 2002).



Carte d'interaction protéique. Les nœuds représentent des gènes. Les liens rouges représentent des interactions entre gènes. La représentation visuelle est difficile à cause de la densité des interactions. Reproduit de Ho et al, 2002.

Les interactions protéine-protéine sont typiquement représentées comme des graphes non dirigés. Dans les graphes d'interactions protéiques, les nœuds représentent les protéines et deux nœuds sont connectés par une arête non dirigée si les deux protéines se lient. Ci-dessus, un exemple de réseau d'interactions protéine-protéine illustre la densité de tels réseaux. La représentation sous forme de graphe possède cependant deux limites principales (voir figure suivante). Il est difficile de représenter correctement un complexe constitué de plus de deux protéines au moyen d'arêtes binaires. Il n'est pas possible de représenter la structure hiérarchique de complexes, c'est-à-dire représenter les complexes qui sont constitués de l'imbrication de sous-complexes. Plusieurs choix, non sans conséquences sur les analyses ultérieures, sont à prendre lors de la représentation pour construire le réseau. La représentation sous forme de simple graphe est en général

insuffisante.

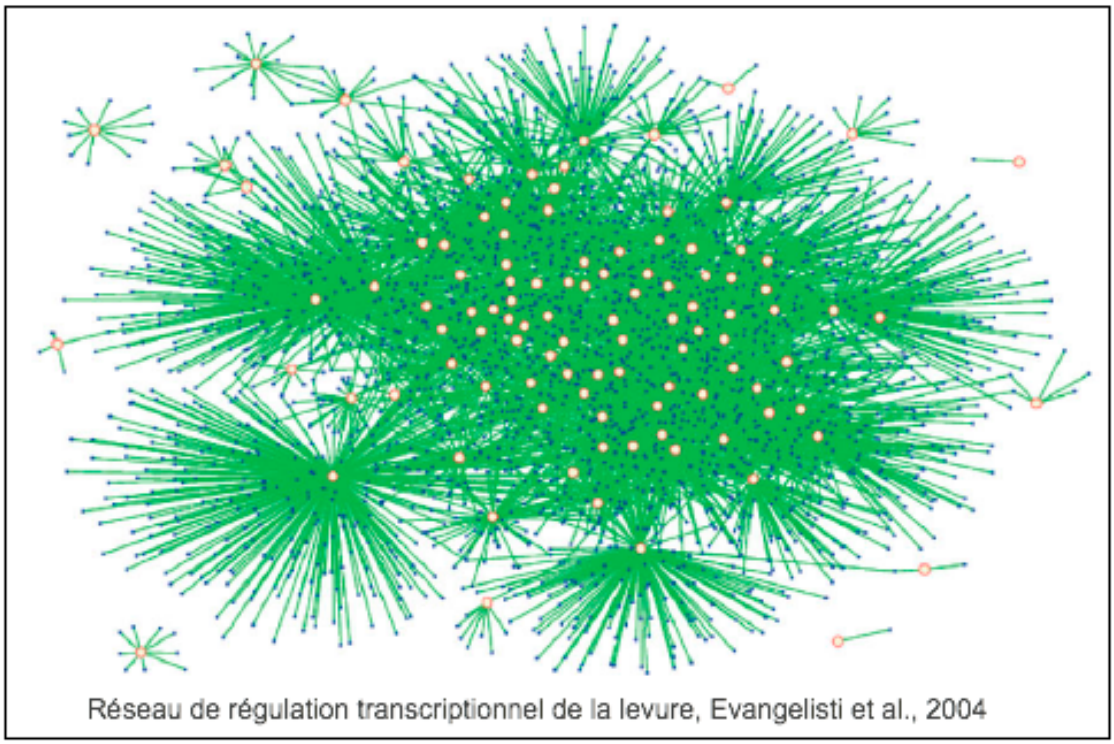


Exemples de modélisation de réseaux d'interactions protéiques. Différents choix de modélisation sont illustrées avec une expressivité variée.

Les réseaux de régulation transcriptionnelle

La régulation transcriptionnelle est un mécanisme important de l'influence d'une protéine ou d'un complexe protéique, appelé facteur de transcription, sur l'expression d'un gène sous la forme d'une protéine. Les grands ensembles d'interactions transcriptionnelles directes, c'est-à-dire les interactions entre un facteur de transcription et un site de régulation sur le gène régulé, ont été compilés à partir de la littérature (Guelzim et al. 2002), de l'immunoprécipitation de chromatine et des expériences avec des puces à ADN (Lee 2002; Harbison et al. 2004). Ces approches directes sont complétées par une famille de méthodes indirectes dites de *reverse engineering* (Segal, Shapira et al. 2003; Friedman 2004; Hartemink 2005) destinée à inférer les interactions à

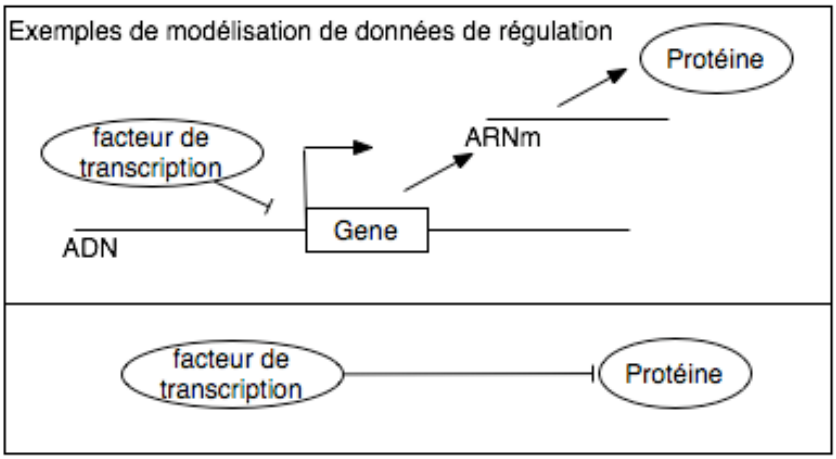
partir d'autres mesures de l'état de la cellule, généralement à partir de mesures du niveau d'expression des ARNm (Cho et al., 1998; Spellman et al. 1998; Causton et al. 2001). De tels réseaux ont été construits pour *E.coli* (Shen-Orr et al., 2002) et la levure (Guelzim et al., 2002; Lee et al., 2002; Luscombe et al., 2004).



Réseau de régulation transcriptionnel de la levure. Le réseau est trop dense pour être dessiné convenablement, toutefois cette représentation laisse entrevoir la nature 'scale-free' du réseau. Reproduit de Evangelisti et al., 2004.

Les réseaux de régulation composés de l'ensemble de ces interactions sont représentés généralement comme des graphes dirigés où les noeuds symbolisent soit les gènes, soit les facteurs de transcription (voir figure ci-dessus qui illustre la densité de tels réseaux), et les arcs dirigés connectent les régulateurs aux gènes régulés. Pour simplifier, dans ces réseaux, on assimile le gène, l'ARNm et la protéine en un seul nœud (voir figure ci-dessous). Les liens représentent les activations ou inhibitions d'un gène (facteur de transcription) sur un autre gène. L'objet mathématique de graph simple est insuffisant pour exprimer l'ensemble des connaissances recueillies à grande échelle sur ces réseaux et

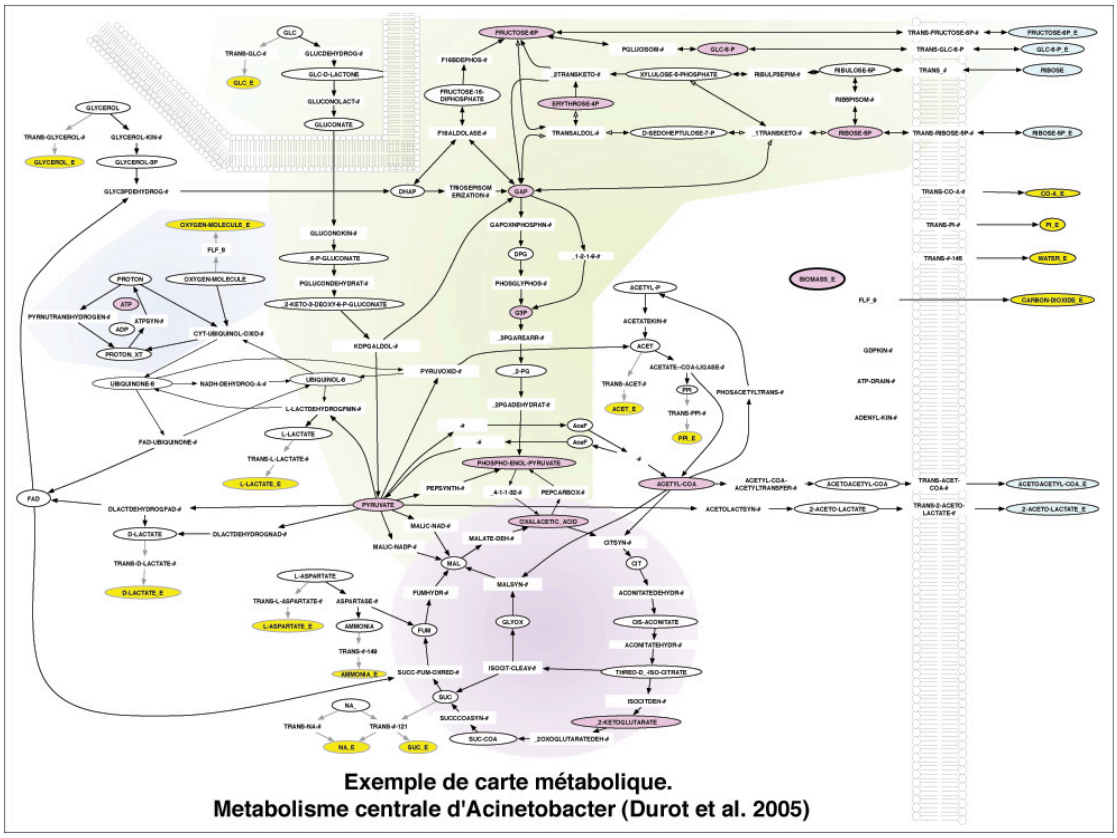
notamment pour l'expression distinguer l'inhibition de l'activation.



Exemples de modélisation de données de régulation. Deux choix de modélisation sont illustrés avec un niveau de détail et une expressivité différente.

Les réseaux métaboliques

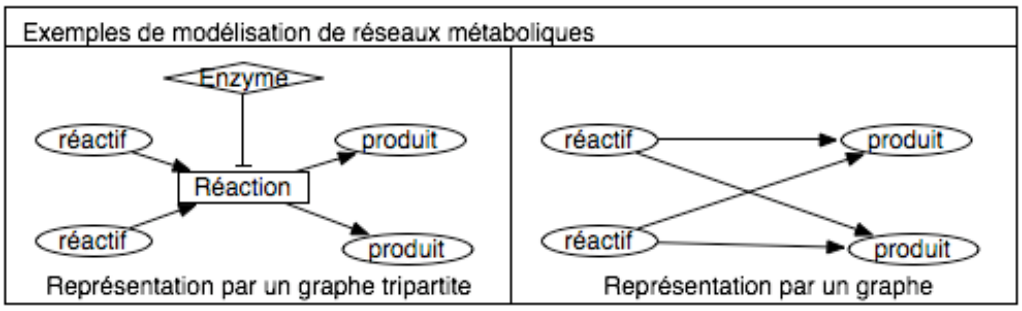
Une réaction du métabolisme est une réaction biochimique entre des métabolites réactifs et produits et un complexe enzymatique (constitué notamment de protéines) qui catalyse la réaction. Les réseaux métaboliques sont composés des réactions biochimiques d'un organisme. Les réactions métaboliques peuvent soit provenir de résultats expérimentaux soir être inférées d'un organisme à un autre au moyen des liens d'orthologie entre les gènes des enzymes des réactions. Les cartes métaboliques de plus de 200 organismes sont décrites à ce jour (biocyc.org, kegg.org).



Exemple de carte métabolique. Extrait de la carte du métabolisme central de la bactérie *Acinetobacter*. Reproduit de Durot et al., 2005.

Les réseaux métaboliques peuvent être modélisés par des graphes composés de trois types de nœuds: les métabolites, les enzymes et les réactions, et deux type d'arêtes: les arêtes stoechiométriques et la régulation catalytique (voir figure ci-dessous). Les arêtes stoechiométriques connectent les réactants aux réactions et les réactions aux produits et sont marquées par le coefficient stoechiométrique du métabolite dans la réaction (Feinberg 1980). Les enzymes qui catalysent la réaction sont reliées par une arête de type de régulation catalytique à la réaction (Jeong et al., 2000). Plusieurs représentations simplifiées ont aussi été étudiées comme le graphe de substrats dont les nœuds sont les réactants, joints par une arête s'ils participent à une réaction commune (Wagner et Fell, 2001), ou encore le graphe de réaction dont les nœuds sont des réactions et deux nœuds sont joints si les deux réactions partagent un même métabolite. Sur la figure ci-dessus, les réactions et certains métabolites choisis figurent uniquement. Plusieurs choix de représentation sont donc possible avec des implications sur les analyses possibles

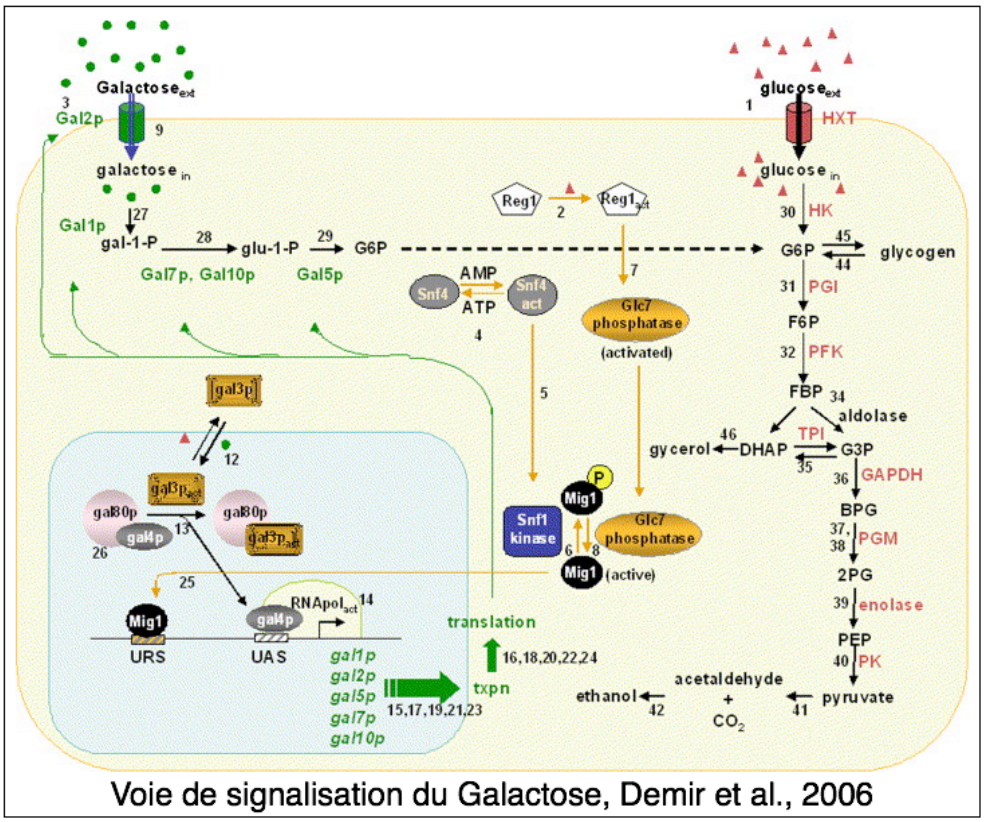
ensuite.



Exemples de modélisation de réseaux métaboliques. Deux choix de représentation sont illustrées avec des expressivités différentes.

Voies de transduction de signaux

Les voies de transduction de signaux représentent l'ensemble des mécanismes qui connectent les signaux extracellulaires aux facteurs de transcription. À proprement parler, ce sont des réseaux qui comportent plusieurs types d'interactions: ils comportent des interactions protéine-protéine et parfois des réactions biochimiques. En fait il existe des recouvrements entre les cartes de régulation transcriptionnelle, d'interactions protéiques, du métabolisme et de voies de transductions de signaux. Toutefois, ces cartes n'ont été produites qu'à la main sur réseaux relativement petits, il est toutefois important de les mentionner ici. Leur représentation se fait en général par des schémas savamment dessinés comme illustré par la figure ci-dessous qui intègre aussi souvent une notion de localisation cellulaire. Différents formalismes permetent de les représenter formellement comme des systèmes d'équations différentielles ou des algèbres de procesus (Cardelli, 2003; Loew et al., 2001). La voie de signalisation la plus étendue vient d'être reconstruite. Elle comporte toutefois 1259 interactions faisant participer 545 composants cellulaires et a été obtenue à partir de la lecture de 1200 articles de la littérature (Ma'ayan et al., 2005).



Voie de signalisation du Galactose, Demir et al., 2006

Voie de signalisation du Galactose. Cette représentation est un dessin qui laisse apparaître des interactions de différente nature. Reproduit de Demir et al., 2006.

Interactions de létalité synthétique

Les réseaux de *létalité synthétique* sont construits à partir de l'ensemble des couples de gènes pour lesquels les mutations dans les deux gènes non-essentiels ne sont létales qu'une fois combinées. Les interactions génétiques ont été identifiées par l'observation de mutants, mais des études récentes ont appliqué des méthodes à haut débit (Tong 2001; Tong 2004 ; Ooi, Pan et al. 2006) pour identifier 4000 des interactions la levure.

Les réseaux de coexpression

Finalement, on va aussi considérer les *réseaux de coexpression*. Les liens de coexpression lient les gènes qui ont des profils d'expression similaires dans un ensemble de conditions expérimentales, c'est-à-dire que leur expression varie de manière corrélée ou anticorrélée suivant les conditions expérimentales (Cho 1998; Hughes 2000; Stuart et

al., 2003; Valencia and Pazos, 2002). Ces interactions ne présupposent donc aucun mécanisme biochimique sous-jacent. Les mesures d'expression sont en général obtenues à partir d'expériences de puces à ADN (microarray) qui permettent de mesurer la quantité relative d'ARN dans la cellule. La rapidité et le faible coût de ces expériences ont permis de construire de tels réseaux y compris chez l'homme.

Prédictions de Réseaux homogènes

Pratiquer des expériences biologiques pour déterminer les interactions afin de construire des réseaux à grande échelle demande un travail long et coûteux. Pour cette raison, de nombreuses méthodes informatiques ont été établies pour proposer des réseaux candidats d'un type d'interaction donné. À titre d'exemple, Jansen et al (Jansen, Yu et al. 2003) prédisent le graphe d'interactions protéine-protéine à partir d'une combinaison de données biologiques (coexpression, co-essentialité, et colocalisation) au moyen de réseaux bayesiens pour combiner et de pondérer les différents indices dont ils disposent. Ces réseaux sont, bien entendu, bien moins fiables que ceux qui sont issus directement d'expériences biologiques. Ils sont toutefois suffisants pour nous, pour mettre en place des méthodes d'analyse, étudier leurs propriétés statistiques et proposer des expériences biologiques sur une partie du réseau. Ces réseaux sont aussi très utiles pour prédire des interactions chez l'homme (Rhodes, Tomlins et al. 2005) où les expériences sont plus complexes à mettre en place que dans les organismes modèles.

C) Intégration 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.

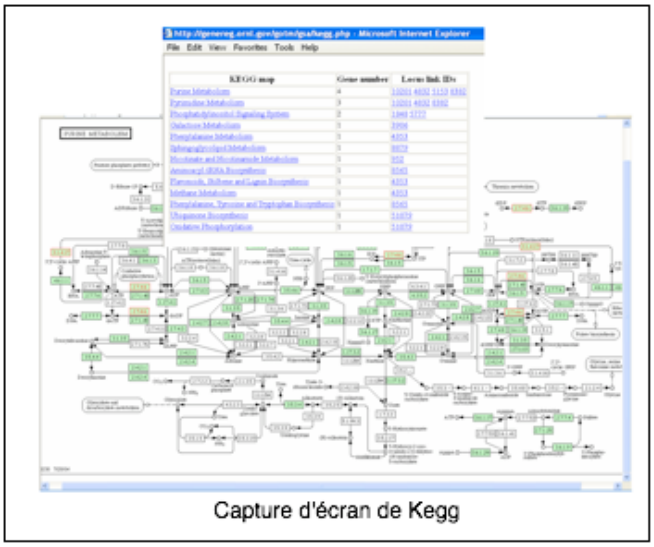
On peut distinguer plusieurs catégories de bases de données (Bader et al., 2006) en fonction du type principal de données contenues dans chacune d'elles, leur format de données et le centre d'intérêt biologique. Nous ne présenterons ici que les bases de données, souvent couplées à des logiciels d'interrogation, 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 (Hermjakob et al., 2004), BioPAX (www.biopax.org), SBML (hucka et al., 2003) ou CellML (Lloyd et al., 2004).

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 (www.genome.jp/kegg) l'encyclopédie de gènes et génomes de Kyoto. C'est la base de données de réactions biochimiques la plus utilisée. Elle stocke les données génomiques, biochimiques et de réseau 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. Kegg repose en fait sur un ensemble de cartes qui sont en fait des images ce qui en rend son utilisation difficile pour des analyses informatiques.

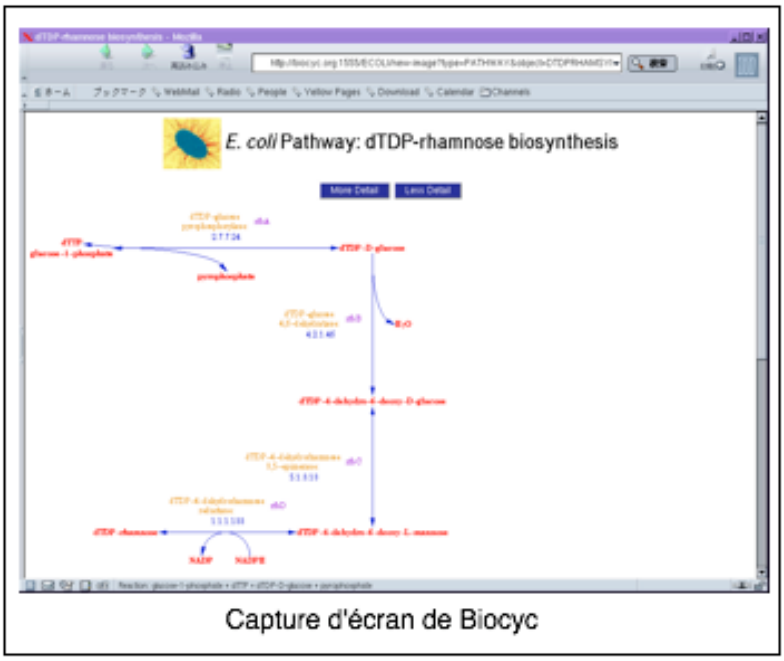
En 2005, Kegg s'est ouvert considérablement pour les développeurs. Des APIs et des *webservices* qui permettent l'interrogation (requêtage) de la base à distance via internet ont été écrits. Kegg repose toutefois toujours sur des cartes dessinées à la main ce qui l'empêche de grandir véritablement, d'intégrer rapidement de nouvelles réactions et d'être modifié par l'utilisateur.


Capture d'écran de Kegg

Capture d'écran de Kegg. Kegg est basé sur des images statiques de chemins de réactions. Reproduit de Kegg.com.

BioCyc

BioCyc est une collection de voies de réactions et d'information génomiques pour plus de 300 organismes. Chaque base 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. 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. 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 curation. Les autres bases de données pour les autres organismes ont été générées automatiquement par inférence sans curation. BioCyc repose sur un format de donnée propriétaire qui n'est pas accessible sans passer par l'outil d'interrogation appelé PathwayTools très spécifique qui est fourni avec la base de donnée. Biocyc est l'outil existant le plus diversifié dans le type de données intégrées (hétérogènes) et le plus riche en quantité d'informations contenues. Il a l'inconvénient d'être un logiciel propriétaire qui en rend son utilisation difficile par des développeurs externes. Les analyses à grande échelle des données contenues y sont donc difficiles.



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

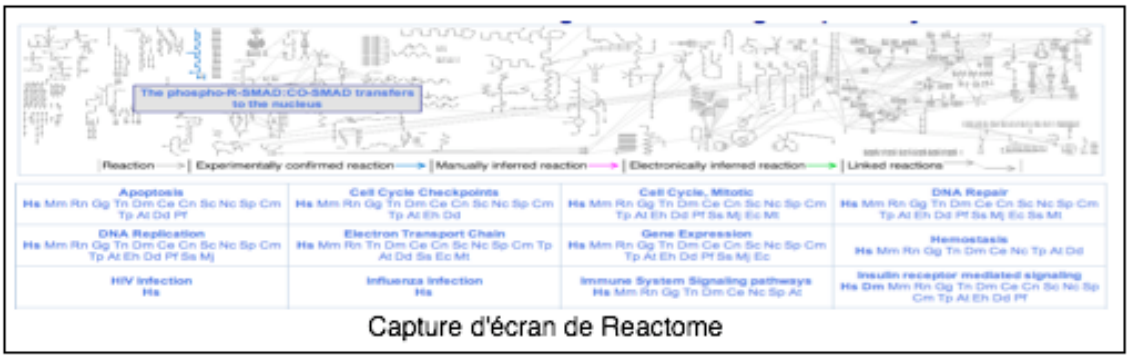
aMAZE, Reactome et Patika

aMAZE est dédié à la représentation des interactions moléculaires et processus cellulaires (Helden et al., 2001). Sa force provient de son modèle de données conçu avec soin. Le modèle permet d'intégrer des données des 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.

Depuis le début de mon travail de thèse, deux solutions récemment développées sont apparues et sont toutes les deux orientées sur l'Homme.

Reactome est une base de données de processus biologiques développée pour l'humain et dont le code source est librement accessible. Reactome couvre toutes informations 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 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.



Capture d'écran de Reactome

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.

Patika est une nouvelle base de données intégrative construite à partir de plusieurs sources de données (Entrez, UniProt, PubChem, GO, IntAct, HPRD et Reactome). La base de donnée 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 dans une certaine mesure compatible avec les formats BioPax et SBML.

Études des RIBH

L'étude des propriétés topologiques des RIBH permet de décrire ces réseaux, de les distinguer dans différents organismes, et de comprendre comment ils fonctionnent. Les propriétés topologiques qui ont été étudiées dans ce contexte et sur des réseaux homogènes sont la distribution des degrés des noeuds, la distribution des coefficients de *clusterisation* ainsi que d'autres notions de densité de distribution de distances entre nœuds ou de distribution d'apparitions de motifs du réseau. Nous rappellerons brièvement ces résultats obtenus principalement pour des réseaux homogènes (Barabasi

and Oltvai 2004 et Milo 2002). Nous présenterons ensuite les études topologiques des réseaux hétérogènes conduites au cours de mon travail de doctorat. Ce sont principalement des études de corrélation entre réseaux homogènes ou des analyses modulaires de ces réseaux. Toutes montrent que les différents types d'interactions ne sont pas indépendantes et qu'elle doivent fonctionner de concert.

Dans une deuxième partie, nous aborderons justement la dynamique des RIBH, là encore, tout d'abord brièvement dans les modules homogènes, puis hétérogènes. L'objectif est de comprendre d'un point de vue dynamique comment les différents types d'interactions coopèrent.

A) Topologie

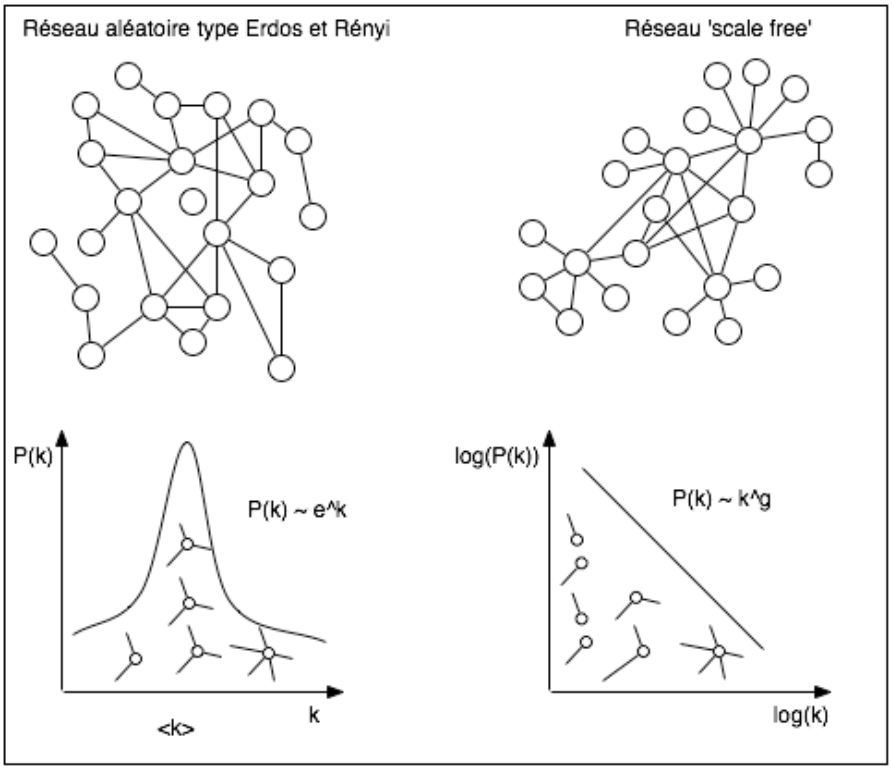
(1) Réseaux homogènes

Statistiques globales

Plusieurs propriétés de la structure globale des réseaux d'interactions ont été étudiées depuis la fin des années 90s afin d'en établir leurs classifications (Wagner, 2001; Watts et Strogatz, 1998). Parmi les propriétés les plus couramment évaluées, on peut citer la distribution de connectivité (Bader et Hogue, 2003) (i.e. la probabilité P(k) d'un nœud d'être entouré par k voisins) qui permet de distinguer les graphes entre plusieurs catégories et de proposer des hypothèses sur leur mise en place au cours de l'évolution.

L'étude des réseaux réels est faite par comparaison avec des réseaux dits aléatoires, notamment en utilisant le modèle introduit par Erdos et Rényi (Bollobas, 1985; Erdös et Rényi, 1960). Dans ce modèle, pour un réseau de taille N donné et un nombre n de liens dans ce réseau, chaque lien est choisi de façon équiprobable parmi les N(N-1)/2 liens possibles. Chaque paire de nœuds est donc reliée avec une probabilité p. Ce modèle a l'avantage de pouvoir être analysé analytiquement très simplement dans la limite où N est grand. Cette distribution est répartie autour de la connectivité moyenne<k> (voir figure ci-dessous), ce qui signifie qu'un tel réseau est relativement uniforme : tous les nœuds du réseau ont à peu près le même nombre de voisins les nœuds fortement ou faiblement connectés sont très rares. Cependant, la distribution de connectivité des réseaux biologiques est très différente de celle de ce modèle (Barabasi et Oltvai, 2004).

La distribution de connectivité de la plupart des réseaux biologiques observés suit loi de puissance $P(k) = k^{-\gamma}$. Cette structure avait déjà été observée auparavant dans certains réseaux non biologiques (Barabasi et Albert, 1999) sous le nom de structure *scale-free* (invariant d'échelle). C'est le cas par exemple des réseaux métaboliques (Jeong et al., 2000) des réseaux d'interactions protéine-protéine (Maslov et Sneppen, 2002). Dans les réseaux *scale-free*, il y a quelques rares nœuds très connectés reliés à de très nombreux autres nœuds très peu connectés, si bien qu'une valeur moyenne n'est pas caractéristique. Par ailleurs, les liens entre ces *hubs* très connectés sont surabondants (Ozier et al., 2003).



Distribution des degrés de nœuds de différents graphes. Les réseaux aléatoires ont surtout des nœuds de degré moyen alors que les réseaux 'scale free' ont surtout des nœuds de faible degrés, et très peu de nœuds avec plein de liens.

Une façon de construire un modèle de réseau *scale-free* est d'ajouter un à un des nœuds en les connectant de façon préférentielle aux nœuds les plus connectés (les riches deviennent plus riches). Ce type de modèle a conduit à émettre différentes hypothèses sur la mise en place de réseaux moléculaires au cours de l'évolution de la vie (Aloy et

Russel, 2004). Les réseaux *scale-free* sont dits robustes : l'inactivation d'un gène au hasard aura peu d'influence sur les autres gènes car elle a de grandes chances de toucher un gène peu connecté étant donné leur nombre important (Albert et al., 2000). Cependant, aucune autre conséquence biologique claire n'a été montrée à partir de ces observations malgré un intérêt populaire très important pour ce type de réseau.

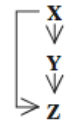
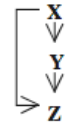
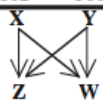
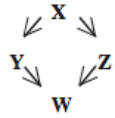
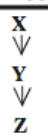
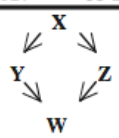
Guelzim *et al*. ont montré par la suite que la distribution de connectivité est en fait plus complexe (Guelzim et al, 2002). En tenant compte des orientations des liens dans un réseau de régulation génétique, il semble en effet que les distributions de degré entrant (nombre de facteurs de transcription d'un gène) et sortant (nombre de gènes régulés) suivent des lois différentes. La distribution de connectivité sortante suit plutôt une loi de puissance alors que la distribution entrante de connectivité suit plutôt une loi exponentielle $P(k) \sim e^{-k}$.

Analyse locales : Modules

Un *module* peut être approximativement défini comme un groupe de molécules liées physiquement ou fonctionnellement qui travaillent ensemble pour accomplir une fonction (Hartwell et al., 1999, Barabasi et Oltvai 2004). Le regroupement des noeuds peut s'opérer par observation de la topologie. Pour identifier de tels modules, plusieurs techniques dites de *clustering* sont utilisées. Les modules permettent de révéler une organisation de niveau supérieure dans le réseau (Jeong et al., 2001; Schwikowski et al., 2000; Snel et al., 2002; Vazquez et al., 2003). De tels groupes de gènes ont été recherchés dans le réseau métabolique (Ravasz et al., 2002), dans les données d'interactions à grande échelle (Bader et Hogue, 2003; Krause et al., 2003; Rives et Galitski, 2003; Spirin et Mirny, 2003; Pereira-Leal et al., 2004) ou dans des réseaux prédits in silico (Snel et al., 2002; von Mering et al., 2003). Ces modules sont aussi intéressants pour qualifier des gènes de rôle inconnu qui feraient partie de modules trouvés topologiquement ou pour suivre le changement et la mise en place des réseaux au cours de l'évolution.

Un *motif de réseau* est un patron d'interactions significativement surreprésenté dans le réseau, par rapport à une famille de réseaux aléatoires de même taille. Cette notion a été

introduite en biologie par Uri Alon et ses collaborateurs dans le contexte d'analyse topologique du réseau de la régulation transcriptionnelle chez E.coli (Milo et al., 2002, Shen-Orr et al., 2002). La figure suivante illustre la surreprésentation de quelques motifs de réseaux dans différents réseaux. Puis au cours de mon travail de doctorat, la méthode a été appliquée à d'autres réseaux homogènes (Alon 2003; Milo, Itzkovitz et al. 2004), même si dans le graphe d'interaction protéique (Wuchty et al., 2003) ils ont eu moins de succès. Une des explications de la surreprésentation d'un motif est que le comportement du motif est sélectionné positivement dans l'évolution.

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			

Plusieurs motifs ont été identifiés dans les réseaux biologiques. Dans différents types de réseaux, les motifs surreprésentés ne sont pas forcément les mêmes. Reproduit de Milo et al., 2002.

De nombreux travaux semblent montrer que la structure des réseaux, est organisée en modules (Hartwell et al., 1999; Wolf et Arkin, 2003), comparé à des réseaux aléatoires. Le traitement de nombreux signaux extérieurs se ferait à l'aide de ces modules/*programmes génétiques*, conformément à l'intuition de Jacob et Monod il y a plus de 40 ans (Monod et Jacob, 1961).

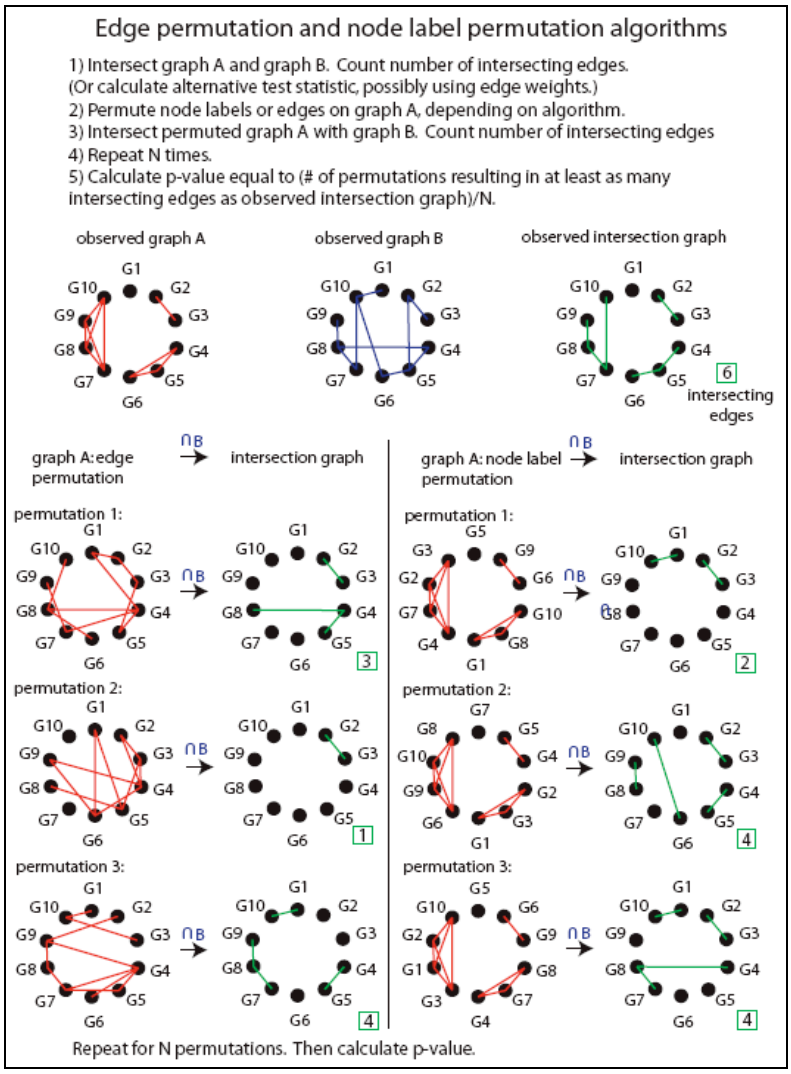
(2) Topologie des RIBH

Lorsque j'ai commencé mon travail de thèse, étant donné le manque de résultats biologiques expérimentaux à grande échelle, aucune étude n'a porté véritablement sur la topologie des RIBH.

Corrélations ente deux réseaux homogènes

Balasubramanian et al (Balasubramanian, LaFramboise et al. 2004) présentent une méthode d'analyse de deux réseaux homogènes ensemble. Ils étudient la corrélation entre l'existence d'une interaction entre deux mêmes gènes dans les deux réseaux étudiés. La procédure de *permutation des arêtes* introduite par ces auteurs est illustrée à la figure ci-dessous. Elle consiste à permuter les arêtes d'un réseau et construire par là un réseau permuté. La corrélation entre l'existence d'une interaction dans le réseau permuté et le deuxième réseau intact est alors calculée. La proportion des permutations pour laquelle la corrélation est au moins aussi grande que la valeur observée avec les réseaux réels est une mesure qui permet de tester l'hypothèse nulle d'indépendance entre les deux graphes.

Cette méthode est appliquée aux associations deux à deux entre 3 réseaux homogènes: le réseau construit à partir des corrélations entre profiles d'expression génétique, celui construit sur les corrélations de phénotypes de croissance de mutants et le réseaux ayant des annotations fonctionnelles de GO (Ontologie Génétique) partagées. Les résultats montrent des corrélations qui ne sont pas bien fortes et demandent plus d'investigations d'après les auteurs.



Tests de permutations pour évaluer les interactions deux à deux dans les associations de graphes Reproduit de Balasubramanian et al. (2004).

Plusieurs études ont analysé la corrélation entre le réseau métabolique et le réseau de coexpression de gènes. Une corrélation a été montrée (Kharchenko et al., 2005) entre le niveau de coexpression de gènes et la distance entre les enzymes correspondantes dans le réseau métabolique de la levure. La coexpression positive est la plus forte entre les gènes d'enzymes qui interviennent successivement dans le réseau métabolique et diminue de façon monotone avec la distance dans le réseau.

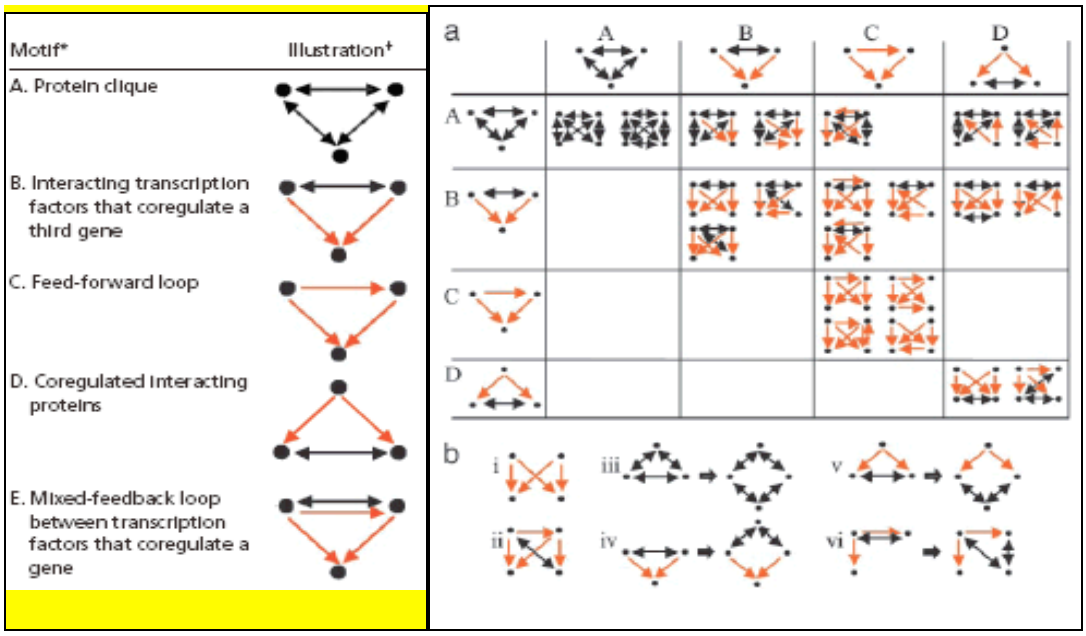
Chez la levure, une forte corrélation existe entre les sous-unités de complexes (réseau

d'interaction protéine-protéine) et le niveau de coexpression des protéines composantes (réseau de coexpression) (Jansen et al., 2002). La corrélation est beaucoup plus faible pour les complexes transitoires, ainsi que pour les interactions protéine-protéine binaires venant notamment d'expériences double-hybride.

Modules dans les RIBH

Cette partie se concentre sur la valeur ajoutée apportée par les réseaux hétérogènes, c'est-à-dire la possibilité de chercher des modules/motifs composés d'interactions de types différents.

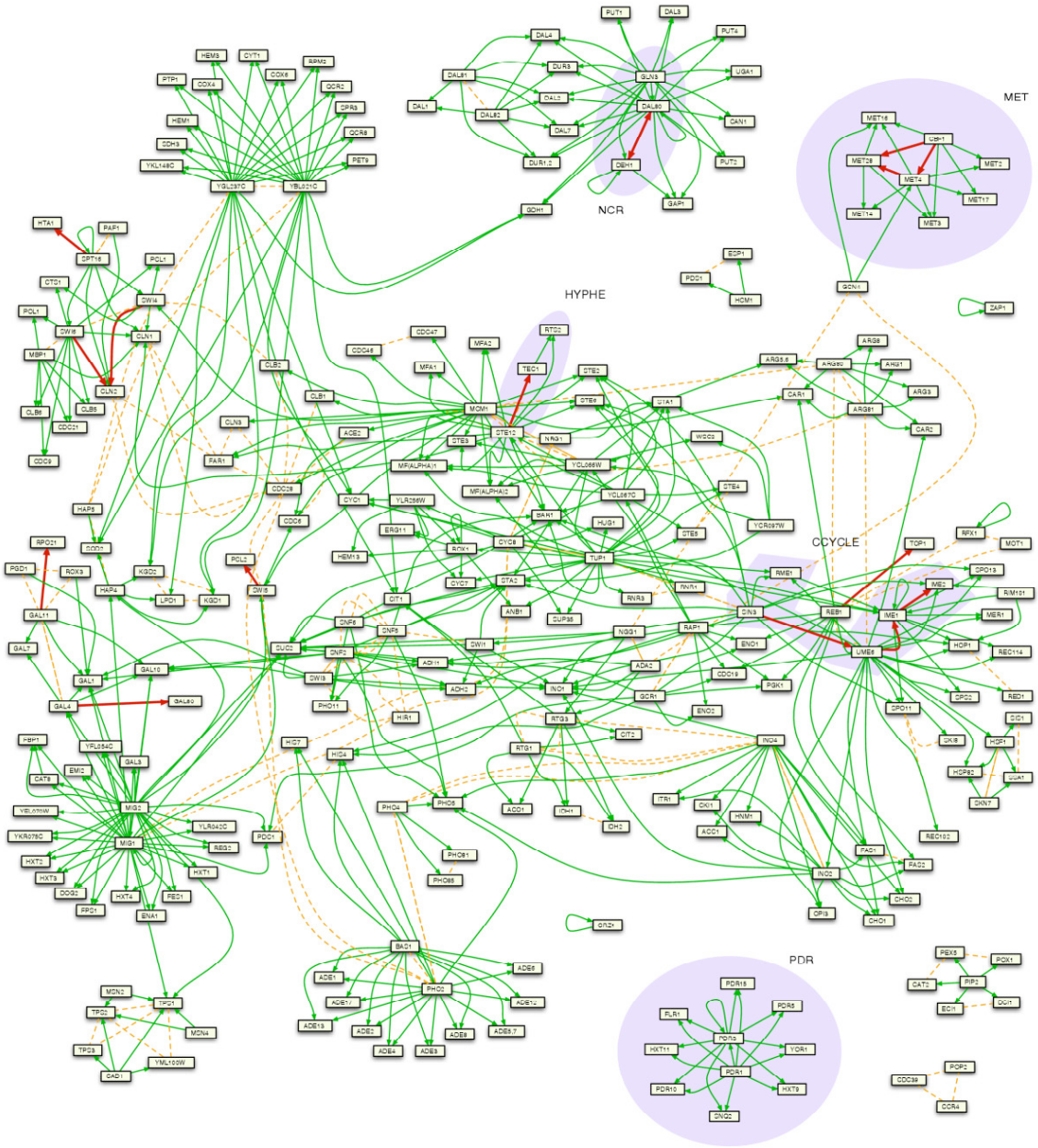
En 2004, les motifs définis par Uri Alon et son équipe ont été généralisés aux motifs sur le réseau hétérogène constitué de deux types d'interactions: les interactions protéiques et la régulation transcriptionnelle chez la levure (Yeger-Lotem et al., 2004). Ont été recherchés les motifs composés de deux à quatre noeuds surreprésentés de façon statistiquement significative par comparaison avec à un ensemble de réseaux obtenus par permutation du réseau étudié. La procédure de permutation échange les arêtes itérativement de manière à préserver le degré du noeud (le nombre des arêtes entrantes et sortantes de chacun des types). Les auteurs ont identifié deux motifs surreprésentés à deux protéines dont une boucle de rétroaction mixte qui inclut les deux types d'interaction. Trois types de motifs hétérogènes à trois protéines sont aussi identifiés (voir Figure ci-dessous). 63 motifs à quatre protéines ont été identifiés, et 57 contenaient un ou plusieurs motifs à trois protéines. 36 sont des motifs à trois protéines avec un noeud supplémentaire, et 21 peuvent être vus comme une combinaison plusieurs motifs plus petits. Cette étude leur a permis de retrouver des modules hétérogènes déjà connus, mais aucun nouveau module n'a été étudié plus en détail suite à cette étude.



(Gauche) Motifs hétérogènes composés de régulation transcriptionnelle (TRI) et d'interactions protéine-protéine (PPI) Ces motifs ont été significativement surreprésentés dans le réseau comparé à la moyenne de 1000 réseaux *randomisés*.

(Droite) Motifs protéiques de taille inférieure à 4 comme combinaison ou extension de motifs plus petits. (a) Motifs qui peuvent être représentés comme une combinaison de motifs à trois protéines. (b) Motifs qui ne peuvent pas être construits à partir de motifs à trois protéines. i, le motif bi-fan, ii, le motif contenant la boucle 'feed-forward' . Un noeud représente un gène et son produit (protéine), une arête rouge dirigée représente une TRI, une arête noire bidirectionnelle représente une PPI. Reproduit de: Yeger-Lotem et al. (2004).

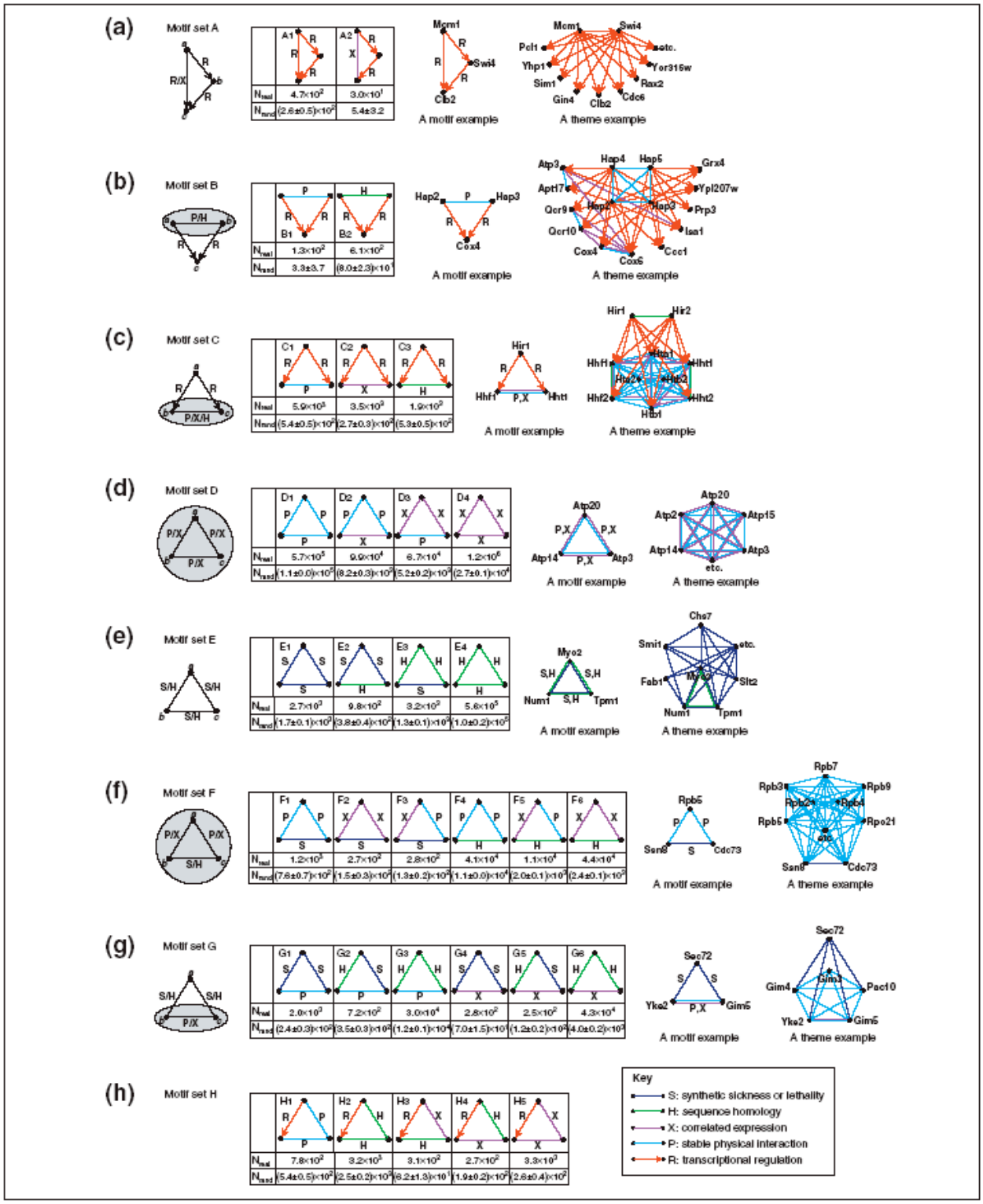
L'interprétation de ces motifs a été mise à mal par une étude qui a suivi (Mazurie et al, 2005) et qui montre que les motifs du type de ceux précédemment mis en valeur étaient difficilement interprétable car ils étaient tous imbriqués les uns aux autres et au sein de structures plus importantes (voir figure ci dessous).



Occurrences de motifs dans la levure. Le réseau des occurrences de motifs dans *S. cerevisiae* illustre le fait que la plupart des motifs ne sont pas isolés et font partis de plus gros agrégats. Vert, régulation transcriptionnelle. Rouge, régulation transcriptionnelle et interaction protéine-protéine. En pointillé, interactions protéines-protéines. Figure reproduite de Mazurie et al. (2005).

En 2005, un réseau hétérogène a été construit à partir de cinq types de liens chez la levure: des liens de coexpression, des interactions de régulation, des interactions

protéine-protéine, génétiques et d'homologie de séquence et a permi de rechercher des motifs toujours de taille trois ou quatre (Herrgard et al., 2005; Zhang et al., 2005). Plus de 5000 motifs à 4 nœuds sont trouvés. Le nombre important de ces motifs rend leur exploitation et compréhension difficile. La notion *de thèmes de réseau* définie approximativement comme des patrons d'interconnections de haut niveau englobant les occurrences multiples de motifs est introduite. La Figure ci-dessous résume leur travail sur les motifs à 3 noeuds et thèmes correspondants avec quelques exemples biologiques.



Motifs à 3 noeuds et thèmes correspondants dans le réseau hétérogène *de S. cerevisiae* (a) Motif correspondant au thème *feed-forward*; (b) Motif correspondant au thème *co-pointage*; (c) Motif correspondant au thème *complexe de régulation*; (d) Motif

correspondant au thème *complexe protéique*; (e) Motif correspondant au thème de clusterisation de voisinage dans le réseau SSL/homologie; (f) Motif correspondant au thème des *membres de complexe compensatoires*; (g) Motif correspondant au thème *protéine et complexe/processus compensatoires*; (h) Autres motifs non classés. Chaque lien coloré représente un des cinq types d'interactions. Pour un motif donné, Nreal est le nombre des sous-graphes correspondants dans le réseau, et Nrand décrit le nombre de sous-graphes correspondants dans le réseau *randomisé*. Reproduit de (Zhang et al., 2005).

Il faut noter que toutes ces études reposent sur un modèle sous-jacent de graphe qui rend impossible la prise en compte de relations n-aires qui sont pourtant nécessaires afin de modéliser par exemples les réactions métaboliques ou les complexes protéiques. Ces modèles ne permettent pas non plus de modéliser la dynamique, même qualitativement. Aucune distinction entre l'activation ou l'inhibition, la production ou la consommation n'est faite. Parmi les 5000 motifs trouvés, nombreux sont en fait des sous-motifs de complexes, et sont en quelque sorte des artefacts liés au modèle de représentation utilisé qui est trop pauvre.

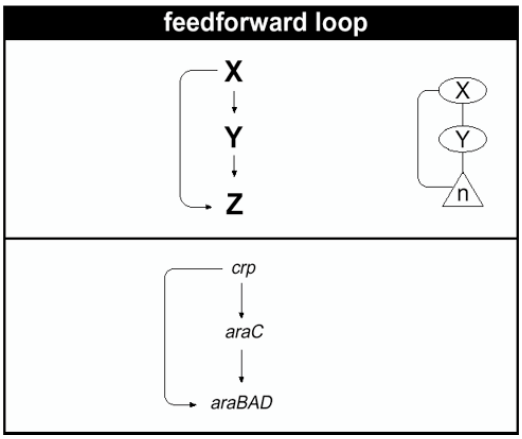
B) Dynamique des RIBH

La modélisation de la dynamique de systèmes biologiques a été à l'origine de l'utilisation d'une grande variété de méthodes dont certaines font appel aux équations différentielles (Savageau 1976), à la théorie des flux métaboliques (Heinrich et Schuster 1996; Fell 1997) ou plus récemment à différents types de modèles qualitatifs (Sima et al., 2005; Siegel et al., 2006). Le problème le plus difficile est d'intégrer dans un modèle unique plusieurs types d'interactions qui ne sont pas habituellement décrits dynamiquement avec les mêmes outils. La plupart des travaux d'analyse des RIBH sont concentrés sur des systèmes relativement petits. Les petits réseaux d'interactions sont analysés au moyen d'équations différentielles étant donné que cet outil permet de décrire une grande variété de type d'interaction de la biologie. Les analyses dites dynamiques à grande échelle portant sur des réseaux hétérogènes sont en fait des analyses statiques sur

lesquelles des conclusions dans le domaine de la dynamique sont astucieusement tirées.

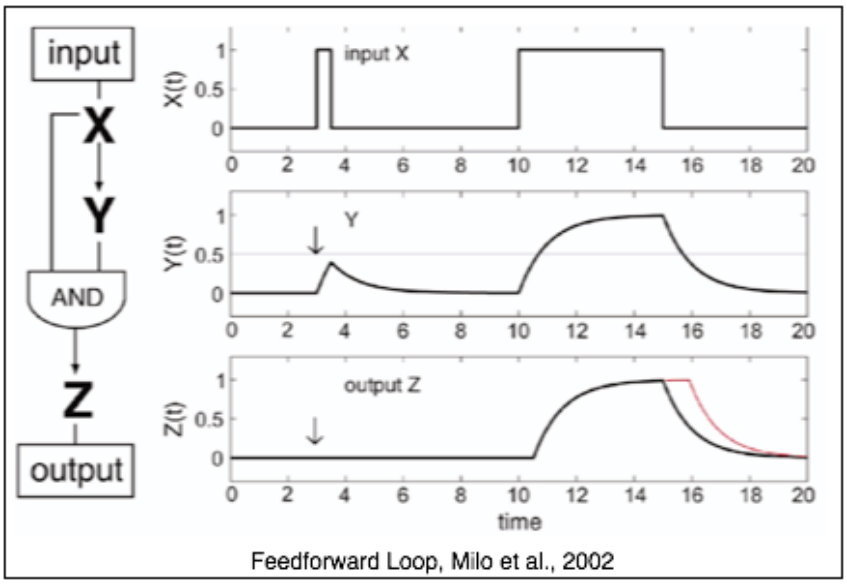
(1) Dynamique de modules homogènes

La dynamique des motifs de réseau trouvés par les méthodes présentées précédemment peut être étudiées. Un des sous-réseaux particulièrement intéressant, près de huit fois plus représenté dans le réseau transcriptionnel d'*E.coli* que dans les réseaux aléatoires est le *FeedForward Loop* (FFL). Ce motif a été retrouvé dans divers organismes: Saccharomyces cerevisiae (Lee et al, 2002; Milo et al, 2002), Bacillus subtilis (Eichenberger et al, 2004), Caenorhabditis elegans (Mangan et al, 2003) et l'homme (Mangan et al, 2003; Odom et al, 2004). Ce module est composé de trois protéines *X*,*Y*,*Z*. *X* régulant *Y*, *X* et *Y* régulant *Z* par coordination. Selon l'effet des régulations (positives ou négatives) il existe huit type de FFL. Le module le plus fréquent chez *E.coli* et la levure est le modèle composé de trois activations (Mangan et Alon, 2003). Une simple modélisation permet alors de comprendre le rôle potentiel de la FFL en tant que *détecteur de persistence*. Si nous considérons la concentration de *X* comme un signal d'entrée et celle de Z comme un signal de sortie, pour induire la production de protéine *Z*, il faut maintenir une concentration de *X* élevée pendant une longue durée. Si la concentration de *X* n'est pas maintenue assez longtemps, *X* n'a pas le temps d'activer la production de *Y* et *Z* ne peur donc être produit.



Motif de boucle *FeedForward*. Ce motif est composé de trois gènes X, Y, Z. X régule Y, et X et Y régulent Z. En bas, une instance de ce motif est représentée dans *E.coli* Figure reproduite de (Shen-Orr et al., 2002).

Depuis 2003, sa dynamique a été étudiée (Mangan et al., 2003) in-vivo lorsqu'une fonction logique 'et' (Setty et al., 2003) est implémentée sur le promoteur du gène Z. Dans ce cas, le module se comporte comme un retardateur. Dans le cas où une fonction logique 'ou' est implémentée sur le promoteur du gène Z (Kalir et al., 2004,2005), le module permet de maintenir le signal après son interruption. Dans le cas de FFL incohérent, c'est-à-dire que le gène Z est activé directement par X ou inhibé via Y par X (ou inversement), la boucle peut générer des impulsions et fournir une réponse rapide (Mangan et al., 2003; Basu et al., 2004).



Modélisation de la boucle 'Feedforward'. Le module filtre les entrées transitoires. Reproduit de Milo et al., 2002.

D'autres modules purement transcriptionnels ont été étudiés et modélisés. Un module ayant une unique entrée et qui génère en sortie des signaux d'expression '*uniquement lorsque c'est nécessaire'*, implémentant ainsi un programme temporel ont aussi été beaucoup étudiés (Laub et al, 2000; Ronen et al, 2002; Shen-Orr et al, 2002; McAdams and Shapiro, 2003; Zaslaver et al, 2004). La boucle de rétroaction (Tyson et al., 2003) négative permet d'accélérer la réponse (Savageau, 1974; Rosenfeld et al, 2002; Krishna et al., 2006) à une variation de signal. La cascade de régulation est un module qui permet la temporisation de signaux (Rosenfeld and Alon, 2003). Des modules

comportant uniquement des interactions protéiques ont été étudiées in-vivo comme une boucle de phosphorylation déphosphorylation (Batchelor et al., 2003).

(2) Dynamique dans les RIBH

Comme nous l'avons vu au sous-chapitre précédent, l'étude de la topologie des RIBH est un domaine tout à fait naissant. L'analyse de la dynamique de RIBH est un problème encore bien plus difficile.

Dynamique globale

Pour étudier la dynamique globale, modéliser tous les processus est complexe. Toutefois, des astuces récentes ont été trouvées afin d'aborder ce problème sans modélisation dynamique à proprement parler.

Couplage dynamique entre Métabolisme et régulation

La dynamique du réseau métabolique couplé au réseau de régulation transcriptionnelle ont été modélisés ensemble (Covert et al., 2001; Covert et Palsson 2002) sous le nom de *regulated Flux Balanced Analysis* (rFBA). Dans ce modèle, le réseau de régulation transcriptionnelle de la cellule est restreint à des règles booléennes. L'hypothèse est faite que les constantes de temps caractéristiques de la régulation transcriptionnelle sont généralement de l'ordre de quelques minutes et qu'elles sont considérablement plus lentes que les constantes de temps associées au métabolisme. Par conséquent la dynamique lié au réseau de régulation est modélisée comme une succession de périodes (succession d'états stationnaires) au cours desquels l'état transcriptionnel est constant. À la fin de chaque pas de temps, l'état du réseau de régulation est calculé à partir de l'état du métabolisme prédit puis le modèle métabolique contraint par l'état du réseau de régulation est ensuite utilisé pour prédire la nouvelle distribution de flux métaboliques.

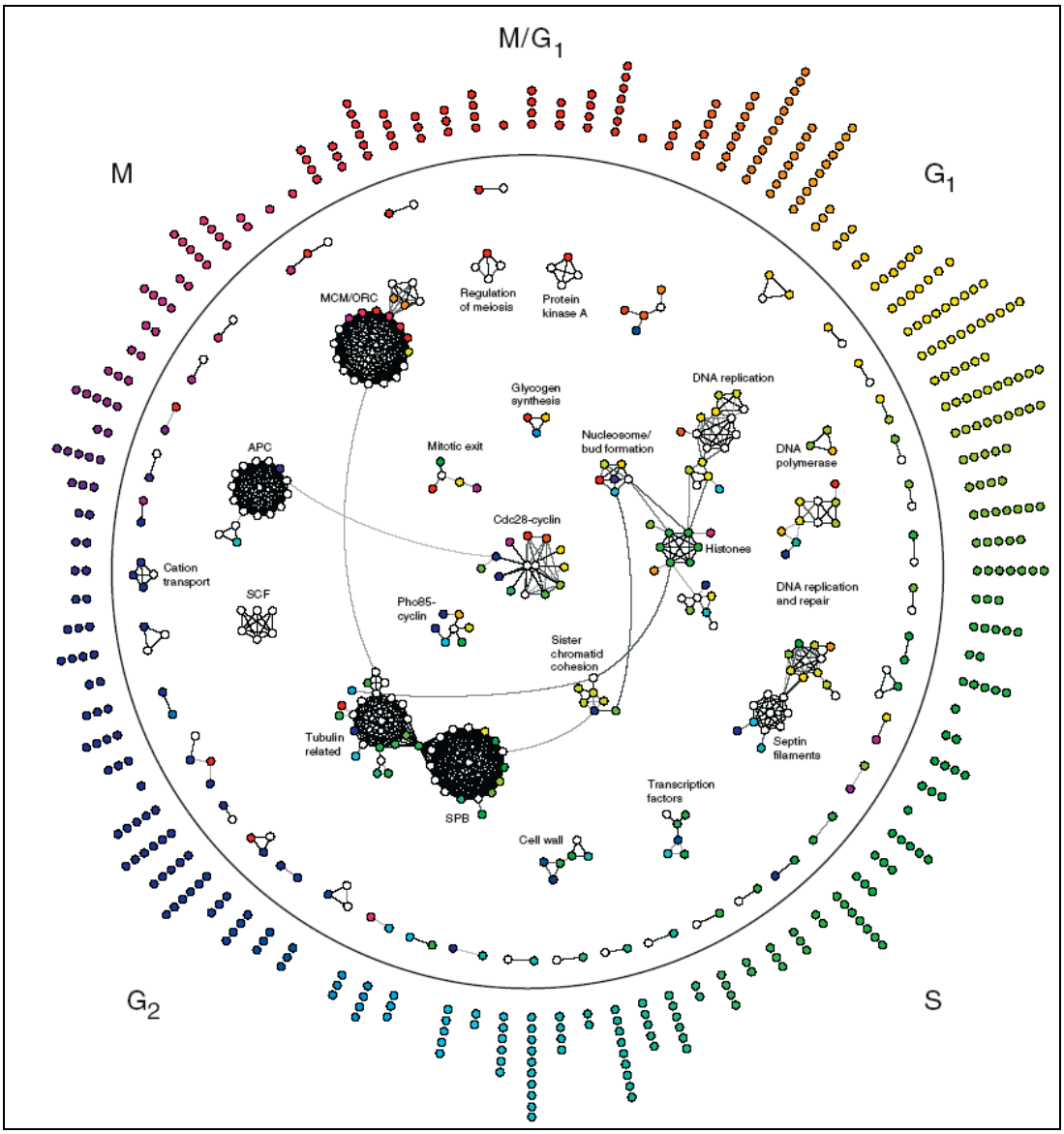
En 2005, cette méthode est appliquée à *E.coli* et afin de montrer que le réseau de régulation favorise le glucose comme source de carbone (Barrett et al., 2005). Le rFBA est plus généralement utilisé pour simuler la croissance cellulaire dans un environnement

dynamique donné.

Dynamique entre Régulation et interactions PPI

Ici nous présentons un autre travail récent qui permet de coupler le graphe d'interactions protéine-protéine avec celui de co-expression. Cette approche permet d'identifier des complexes qui se forment dynamiquement dans la levure au cours du cycle cellulaire (de Lichtenberg et al. 2005). Le réseau comporte 304 gènes dont 188 sont liés par des interactions des deux types (coexpression et interaction protéine-protéine). Cette procédure a abouti à l'identification de 29 modules denses, représentant les complexes ou des groupes de complexes, chacun identifié à un instant spécifique du cycle cellulaire (voir Figure ci-dessous).

Les auteurs proposent un concept d'*assemblage juste à temps*, où uniquement quelques sous-unités du complexe sont régulées transcriptionnellement et contrôlent le moment de l'assemblage.

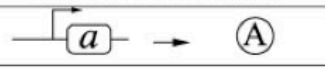
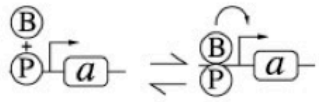
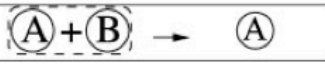


Réseau temporaire d'interaction protéique dans le cycle cellulaire mitotique de la levure. Les protéines du cycle cellulaire qui sont des parties des complexes ou autres interactions physiques sont montrées dans le cercle. Pour les protéines dynamiques, l'instant où l'expression est maximale est indiqué par la couleur du noeud; les protéines statiques sont représentées par les noeuds blancs. En dehors du cercle, les protéines dynamiques sans interactions sont situées et servent de légende de couleur. Reproduit de (de Lichtenberg, Jensen et al. 2005)

Dynamiques de modules dans les RIBH

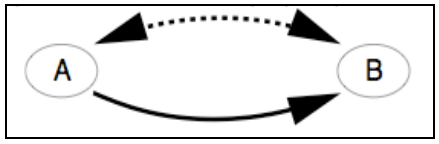
Analyse de modules in-silico

L'espace des dynamiques possibles peut être observé en construisant de petits réseaux aléatoires et en les traduisant automatiquement sous la forme d'un système d'équations différentielles (François et Hakim, 2004). Inversement, les auteurs ont aussi cherché pour une dynamique donnée (oscillation, interrupteur bistable), les différentes implémentations de réseaux obtenus grâce à un algorithme génétique de génération de réseau. Les circuits obtenus montrent une grande variété d'implémentations possible pour une fonction donnée. Tous les réseaux trouvés comportent au moins une interaction protéine-protéine et une régulation génétique, montrant par là leur intérêt. Par ailleurs, certains réseaux trouvés correspondent à des motifs trouvés in-vivo.

#	REPRESENTATION	EQUATIONS
i)		$\frac{d}{dt}[A] = \tau_A[a] - \delta_A[A]$
ii)		$\frac{d}{dt}[a:P] = \theta[a:P:B] - \gamma[a:P][B]$ $\frac{d}{dt}[a:P:B] = \gamma[a:P][B] - \theta[a:P:B]$ $\frac{d}{dt}[A] = \tau_A[a:P] + \tau'_A[a:P:B]$
iii)		$\frac{d}{dt}[A] = -\tau_M[A]$ $\frac{d}{dt}[A^*] = \tau_M[A]$
iv)		$\frac{d}{dt}[A:B] = -\delta[A:B]$ $\frac{d}{dt}[A] = \delta[A:B]$
v)		$\frac{d}{dt}[A] = -\gamma[A][B]$ $\frac{d}{dt}[B] = -\gamma[A][B]$ $\frac{d}{dt}[A:B] = \gamma[A][B]$
vi)		$\frac{d}{dt}[B] = -\delta[A][B]$
vii)		$\frac{d}{dt}[A:B] = -\delta[A:B][C]$ $\frac{d}{dt}[C] = -\delta[A:B][C]$ $\frac{d}{dt}[A] = \delta[A:B][C]$

Liste des réactions possibles dans le modèle. A :B représente le complexe formé de A et B. a:P dénote le gène a avec la protéine P liée à son promoteur. La réaction ii illustre la réaction où la protéine B se lie au promoteur de a sur lequel une protéine P est déjà fixée. La même réaction existe et est possible avec un gène dont le promoteur n'a fixé aucune protéine. Seuls les termes tracés dans la simulation sont mentionnés dans la partie droite du tableau, donnant la concentration des protéines. Pour un réseau de réaction donné, tous les termes à droite des équations doivent être ajoutés pour obtenir l'évolution d'une variable donnée (e.g. une protéine A produite par un gène a et qui subit une modification post-traductionnelle est obtenue par la l'ajout des membres de droite des équations de i et iii. Figure reproduite de (François et Hakim, 2004).

Comme module hétérogène, la boucle de rétroaction hétérogène dans son comportement oscillatoire a été la plus étudiée. Il s'agit d'un motif comportant deux gènes dont les produits forment un dimère et dont l'un des produits régule la transcription de l'autre gène. En modélisant mathématiquement ce module, il a été montré qu'il peut servir d'interrupteur bistable, ou d'oscillateur en fonction des paramètres du système (François et Hakim, 2006). Le diagramme de phase et la description du régime non linéaire oscillant montrent l'intérêt des interactions protéine-protéine dans les modules de régulation génétique.



Représentation schématique du motif de boucle de rétroaction hétérogène. La flèche noire représente une régulation transcriptionnelle. La flèche en trait discontinue représente une interaction protéine-protéine entre deux gènes/protéines A et B. Figure reproduite de (François et Hakim, 2006).

Le motif de boucle de rétroaction hétérogène a été décrit dans plusieurs organismes et avec plusieurs fonctions (Goldbeter, 2002). Le module peut servir d'horloge (Lahav et al, 2004) ou d'interrupteur bistable (François et Hakim, 2004,2005). Différentes

implémentations de la boucle hétérogène ont été répertoriées, y compris un mécanisme avec transport nucleo-cytoplasmique (Nelson et al, 2004). Le module plus complexe de l'opéron Sin de *Bacillus subtilis* (Voigt et al., 2005) a été modélisé. Il s'agit d'une boucle de rétroaction négative faisant intervenir différentes formes d'ARNm. Suivant le paramétrage, le module peut se comporter comme un interrupteur bistable, un oscillateur, un générateur d'impulsion ou un amplificateur.

Construction de modules in-vivo (Synthetic Biology)

Pour valider expérimentalement la compréhension théorique de ces modules étudiés in-silico, certains modules ont été synthétisés, c'est à dire construits par assemblage de fragments d'ADN, au sein d'une cellule (Fu, 2006). Les premiers modules n'ont comporté que des régulations transcriptionnelles. Toutefois, depuis peu, des modules hétérogènes sont également implémentés in vivo. L'étude de réseaux hétérogènes doit permettre de mieux comprendre comment construire de tels petits réseaux hétérogènes. On présente ici les premiers modules ainsi construits historiquement. Pour une liste plus importante de module, on laisse le lecteur se reporter aux sites web : http://www.biobricks.org et http://www.syntheticbiology.org.

Le 'Toggle switch'

Un type important de traitement du signal de l'environnement ou intracellulaire pour une cellule est la discrétisation du signal, et notamment rendre le signal binaire. Ces processus contrôlent notamment la différentiation cellulaire. La mémorisation du signal est aussi importante pour maintenir le signal de sortie et poursuivre la différention commencée. Le système est donc bistable (phénomène d'hystérésis). De nombreux exemples biologiques ont ainsi été décrits, notamment la maturation de l'œuf chez le Xenopus (Ferrell, 1999; Xiong et Ferrell, 2003), et l'opéron lactose (Ozbudak et al., 2004 ; Yartseva et al., 2006). Des boucles de rétroaction positives directes ou indirectes composent toujours ces systèmes, conformément à la conjecture de René Thomas récemment prouvée mathématiquement (Soulé, 2003). Ces systèmes peuvent être étudiés in-vivo, mais il est aussi possible d'en synthétiser artificiellement c'est-à-dire les

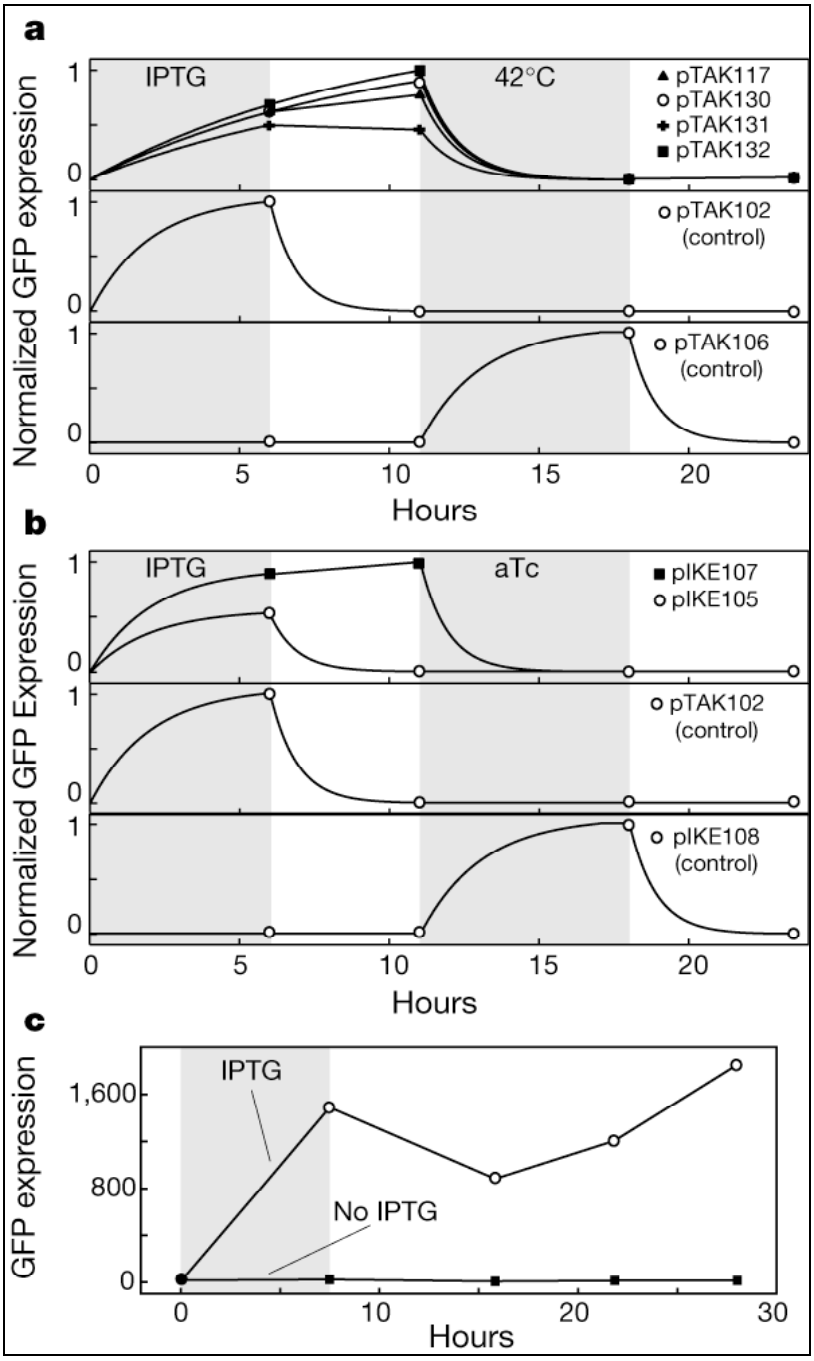
implémenter dans la cellule pour les étudier plus en détail, de manière plus isolé, avec peut être un jour des applications biotechnologiques (Kobayashi et al., 2004).
Un module est un commutateur (ou *switch*) si le module est bistable et si les deux états stables du système peuvent être caractérisés par une forte quantité d'une espèce et une faible quantité d'une autre espèce pour un des états et inversement pour l'autre état.
Le commutateur génétique le plus simple est un réseau de deux gènes qui se répriment mutuellement. Ce module a été implémenté (Gardner et al., 2000) dans *E.coli*.

Le lecteur pourra se référer à une étude plus complète du module (Cherry et Adler, 2000). Le module est modélisé qualitativement par le système d'équations différentielles:

$$\frac{dA}{dt} = \frac{\rho_A}{1 + (B / B_0)^{\nu}} - \delta_A A$$

$$\frac{dB}{dt} = \frac{\rho_B}{1 + (A / A_0)^{\mu}} - \delta_B B$$

A et B représentent les concentrations des protéines. δ_A et δ_B sont les constantes de dégradation des protéines. L'inhibition génétique est modélisée par une fonction de Hill où ρ_A et ρ_B sont les taux de production des protéines A et B en l'absence de leur répresseur, ν et μ sont les coefficients de Hill qui modélisent la coopération des facteurs de transcription. Lorsque l'on analyse le système, on s'apercoit que la coopération est nécessaire pour la présence des deux états stables.
Des plasmides ont été construits dans *E.coli*. Ils font intervenir le promoteur/répresseur Lac (lacI) de l'opéron lactose et le promoteur Ptrc2, le promoteur et la protéine λCI, du bactériophage λ. Pour rendre visible le fonctionnement du module, la GFP ('Green Fluorescent Protein' : protéine fluorescente verte) est aussi utilisée. Il est possible de contôler les inhibiteurs qui interviennent dans ce module par des inducteurs que l'on peut introduire ou retirer du milieu de culture cellulaire: l'IPTG (isopropyl-ß-D-thiogalactopyranoiside) qui empêche la répression de Lac, et la température désactive λ CI. Deux signaux d'entrée peuvent donc être mis à 0 ou à 1 et influencer l'état du commutateur dont la sortie est visible (couleur verte ou non).



Démonstration de fonctionnement. Les zones grises indiquent les périodes d'induction chimiques ou thermiques durant laquelle les signaux d'entrée sont stables. A & b. Différents plasmides de configuration légèrement différentes sont essayés et comparés aux plasmides de contrôle. c, Démonstration de la stabilité à long terme de l'état stationnaire de la boucle. Reproduit de (Gardner et al., 2000).

Le ‘Répressilateur’

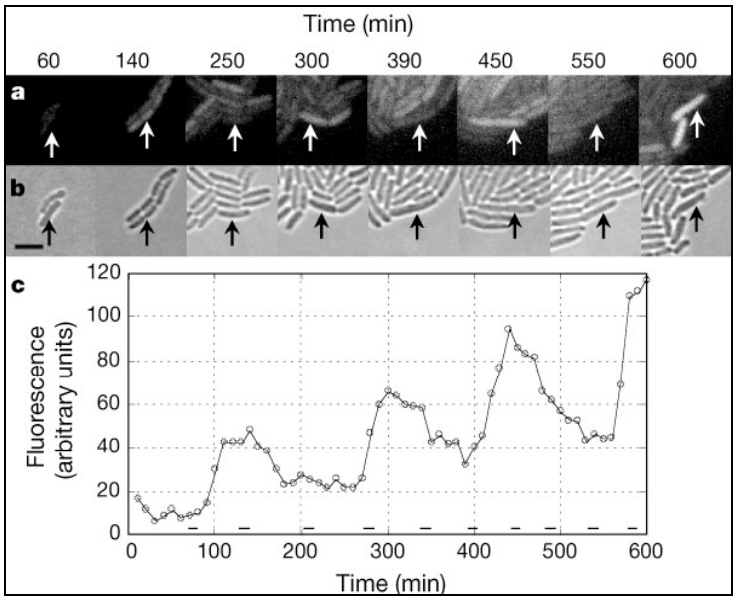
Un autre exemple de module dynamique important dans les organismes vivants est l’oscillateur. Il intervient dans des mécanismes de régulation cellulaires ou de synchronisation des cellules entre elles ou avec des stimuli extérieurs. Un *oscillateur* est un module dans lequel les concentrations intracellulaires de certaines protéines oscillent au cours du temps. De nombreux oscillateurs biologiques isolés ont été décrits: le cycle cellulaire (Chen et al., 1999; Tyson, 1991), les oscillateurs circadiens (Daan et Pittendrigh, 1976; Dunlap, 1998; Young, 2002), ou encore l’oscillateur p53/Mdm2, ou des oscillateurs couplés entre eux : oscillateurs circadiens et du cycle cellulaire (Matsuo et al., 2003). Ces modules comportent tous une boucle de rétroaction négative (Tyson et al., 2003). Un des oscillateurs les plus simples qui peuvent être synthétisés est une boucle de rétroaction négative comportant trois gènes A, B, C où A inhibe B, B inhibe C, et C inhibe A. Un tel module a été modélisé au moyen d’un système de six équations différentielles et a été synthétisé dans *E.coli* (Elowitz et Leibler, 2000)

$$\frac{dm_i}{dt} = \alpha_0 + \frac{\alpha}{1 + p_j^n} - m_i$$

$$\frac{dp_i}{dt} = \beta(m_i - p_i)$$

Le gène i=1,2,3 inhibe respectivement le gène j=3,1,2 dans ces équations. α_0 représente le taux de production des protéines par cellules i en présence de la protéine j, $\alpha_0+\alpha$ représente ce même taux en absence de cette protéine j, n est le coefficient de Hill correspondant à la répression. ß représente le rapport entre le taux de dégradation des protéine et celui de l’ARN. Là encore, une coopération importante, ce qui introduit une forte non linéarité, est important pour le fonctionnement de l’oscillateur.

Les gènes qui interviennent sont lacI, tetR et la protéine λCI et la GFP est utilisée comme rapporteur.



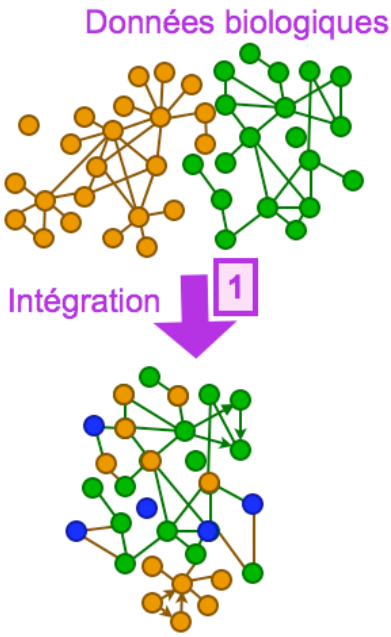
Repressilateur implémenté dans la bactérie. a, b, Croissance temporelle rapportée par l'expression de GFP pour une cellule unique transformée d' *E.coli*. (Fig. 1a). Des photos de croissance des micro-colonies ont été obtenues périodiquement en fluorescence (a) et lumière naturelle (b). (c) Fluorescence des cellules. Reproduit d'après Elowitz and Leibler, 2000.

Conclusion sur l'état de l'art

Fin 2002, plusieurs sources de données d'interactions biologiques deviennent disponibles (compilation de la littérature, expériences à haut débit, prédiction d'interactions). Peu de logiciels permettant de les intégrer existent et chaque type d'interaction est traité puis analysé séparément. Les travaux de Alon et al et Guelzim et al ont été précurseurs dans l'analyse des réseaux homogènes en terme de motif de réseau, mais il restait encore beaucoup à découvrir pour comprendre comment les différents type d'interactions jouent un rôle ensemble. Un grand enthousiasme s'empare des chercheurs au sein du domaine appelé *Biologie des Systèmes*.

Depuis, dans le domaine des logiciels d'intégration de données, il faut surtout noter l'essor d'un format d'échange commun entre ces logiciels, BioPax, mais il encore bien incomplet. De très nombreuses études ont essayé d'étendre les premiers travaux de Uri

Alon. La dynamique des modules trouvés a été étudiée, certains de ces modules ont été synthétisés, et des modules plus riches ont été recherchés. Le cadre de formalisation commun à l'ensemble de ces travaux est un graphe simple, trop pauvre pour modéliser la richesse des interactions biologiques comme les complexes protéiques ou le métabolisme. Une certaine déception s'est ensuite dégagée de la recherche de motifs de réseaux définis uniquement comme recherche de motifs surreprésentés dans les réseaux. Ces motifs ne sont pas génériques, leur nombre est important, et leur enchevêtrement dans les réseaux en complique l'interprétation. Une définition de module dit fonctionnel, c'est-à-dire pas uniquement basé sur la topologie de graphe est nécessaire. Enfin, chez bon nombre de chercheurs, l'espoir est toujours intact de trouver, d'isoler, et de manipuler de tels modules avec une fonction particulièrement intéressante dans ces réseaux que ce soit par analyse du vivant ou au moyen de la biologie synthétique, tout comme de nombreux chimistes cherchent en permanance de nouvelles molécules d'intérêt aussi bien par exploration du vivant, que par exploration combinatoire de synthèse.



Chapitre III – Intégration de données pour la construction de RIBH

Comme nous avons vu dans l'état de l'art, il existe de nombreuses bases de données qui listent les connaissances sur des entités biochimiques, ou sur des interactions entre les entités. Il existe aussi depuis quelques temps, des bases de données qui comportent des informations sur des interactions qui peuvent influencer d'autre interactions, et qui intègrent différents types de données.

Dans l'étape d'intégration de données, afin de construire des réseaux d'interactions biologiques hétérogènes, le logiciel Biocyc, dont les données sont difficiles d'accès, a été choisi. En effet, c'est le logiciel qui comporte à la fois un modèle de données riche ave différents types d'interactions, mais aussi des données sur plus de 300 organismes. Pour pallier au manque d'accessibilité au code source des logiciels et base de données intégratives (en l'occurrence de BioCyc), nous avons développé Cyclone. Cyclone facilite l'utilisation des données de BioCyc (collection de voies de réactions et d'information génétiques) pour le bioinformaticien, en fournissant un programme complet en Java qui va bien plus loin que les interfaces fournies par BioCyc, l'*API Lisp* (Interface d'Accès au Programme écrit en langage Lisp) ou ses dérivés (PerlCyc, JavaCyc). BioCyc est une collection de base de donnée de voies de transformation et d'information génomique (PGDBs). Leur contenu peut être interrogé et visualisé au moyen d'une interface graphique web ou d'un logiciel dédié appelé *Pathway Tools*. Les bases de données BioCyc contiennent des données de très bonne qualité, mais leur contenu est difficilement accessible et manipulable pour des développeurs extérieures ne connaissant pas le Lisp. Cyclone utilise le modèle de données de BioCyc (le *Frame Representation System* (FRS)) pour construire un modèle *objet* qui permet la

manipulation et l'analyse aisées des données. Cyclone peut lire et écrire les données de BioCyc, écrire ses propres données au format CycloneML défini par un fichier au format XML et appelé schéma XSD. Ce schéma est automatiquement généré à partir de l'ensemble structuré de concepts (ontologie) de BioCyc par Cyclone, assurant ainsi la compatibilité entre les deux formats de stockage : celle de BioCyc et de Cyclone. Les objets de Cyclone peuvent être sauvegardés dans une base de données relationnelle, CycloneDB. Les requêtes peuvent être écrites en language SQL, mais plus intéressant, en HQL (langage de requêtes orienté objet) intuitif et concis (hibernate.org). De plus, Cyclone s'interface facilement avec différents outils pour l'édition de données, la manipulation de graphes, pour leur visualisation.

Le code source du programme écrit ainsi que les différents schémas de données de Cyclone sont disponibles sur http://nemo-cyclone.sourceforge.net.
Cyclone a été rendu public le 1er Mai 2006. À la date du 1er Janvier 2007, plus de 1000 visiteurs ont découvert le site web suite à une publicité très limitée au cours des deux premiers mois. Il a été téléchargé par 150 à 200 groupes de bioinformatique dans le monde. 15 nous ont fait un retour encourageant sur son utilisation. Une description de Cyclone a été publiée dans la revue *Bioinformatics*. Ma contribution principale dans ce travail a été d'une part de comprendre et ouvrir le format de données de BioCyc, afin d'en obtenir une image en terme d'objets Java, et d'autre part, valoriser et rendre l'outil public.

BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 10 2007, pages 1299–1300
doi:10.1093/bioinformatics/btm107

Systems biology

Cyclone: java-based querying and computing with Pathway/Genome databases

François Le Fèvre, Serge Smidtas and Vincent Schächter*

Computational Systems Biology Group, Genoscope/CNRS-UMR8030, 2 rue Gaston Crémieux, 91057 Evry Cedex, France

Received on November 3, 2006; revised on March 12, 2007; accepted on March 13, 2007

Advance Access publication March 28, 2007

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Cyclone aims at facilitating the use of BioCyc, a collection of Pathway/Genome Databases (PGDBs). Cyclone provides a fully extensible Java Object API to analyze and visualize these data. Cyclone can read and write PGDBs, and can write its own data in the CycloneML format. This format is automatically generated from the BioCyc ontology by Cyclone itself, ensuring continued compatibility. Cyclone objects can also be stored in a relational database CycloneDB. Queries can be written in SQL, and in an intuitive and concise object-oriented query language, Hibernate Query Language (HQL). In addition, Cyclone interfaces easily with Java software including the Eclipse IDE for HQL edition, the Jung API for graph algorithms or Cytoscape for graph visualization.

Availability: Cyclone is freely available under an open source license at: http://sourceforge.net/projects/nemo-cyclone

Contact: cyclone@genoscope.cns.fr

Supplementary information: For download and installation instructions, tutorials, use cases and examples, see http://nemo-cyclone.sourceforge.net

1 INTRODUCTION

The availability of *usable* biological pathways information is both a key enabler and a bottleneck in systems biology research. Usability implies not only high quality, but also the possibility to query and access the information in a format suitable for a variety of modeling and analytical tasks. The two most prominent general-purpose metabolic pathways databases are Kyoto Encyclopedia of Genes and Genome (KEGG) (Kanehisa and Goto, 1999) and BioCyc (Karp *et al*., 2002). BioCyc is a collection of 205 species-specific 'Pathway/Genome Databases' (PGDB) managed by the proprietary software called 'Pathway Tools'. These PGDBs include automated reconstructions of metabolic networks from genome annotation and also curated sets of metabolic pathways, regulatory networks and chemical information for model organisms. The MetaCyc database recapitulates a set of non-redundant, experimentally elucidated metabolic pathways from 900 organisms.

*To whom correspondence should be addressed.

MetaCyc and some PGDBs have been very successful, in particular among microbiologists, as reference datasets. Whereas they can be queried and visualized through the Pathways Tools interface, their use for advanced querying, model building or computation is more problematic.

Non-standard querying requires the writing of LISP code that targets BioCyc's native frame-based representation scheme (Karp *et al*., 1995). An alternative solution is the use of the JavaCyc (arabidopsis.org) or the PerlCyc APIs, which provides full access to BioCyc data by encapsulating calls of Pathway-Tools Lisp functions in Java or Perl, but do not create native objects for each biological entity. The latest version of BioCyc uses a relational database to store utility classes, but not the biological objects. Finally, the recently developed BioWarehouse (Lee *et al*., 2006) integrates a set of biological databases, including PGDBs, into a single platform. Its use for querying and computation is limited, however, only part of Biocyc's model is included, and data from PGDBs can only be read but not updated by BioWarehouse.

Cyclone addresses some of these limitations by providing a Java Object-Oriented API aimed at accessing, manipulating and computing with BioCyc information in an intuitive manner.

2 INFORMATION FLOW IN CYCLONE

Cyclone maps BioCyc objects on Java objects. Using an extension of JavaCyc, Cyclone extracts the BioCyc data model from Pathway Tools and converts it into an XML Schema (Fig. 1a), defining an object model which is Cyclone's pivotal representation.

Cyclone uses JAXB (Java Architecture for XML Binding) in order to define Java classes corresponding to this schema (Fig. 1b). HyperJAXB is used in order to define an adequate correspondence in the object-relational mapping software Hibernate (Eliott *et al*., 2004), which implements persistence of the classes using any compatible relational database management system (CycloneDB) (Fig. 1c). Overall, this mechanism ensures the automatic adjustment of the Cyclone representation model to update in the BioCyc model.

Once the Cyclone representation has been defined and instantiated, e.g. using data from a PGDB, the resulting 'biological' objects can be queried using HQL (Fig. 1d).

© The Author 2007. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

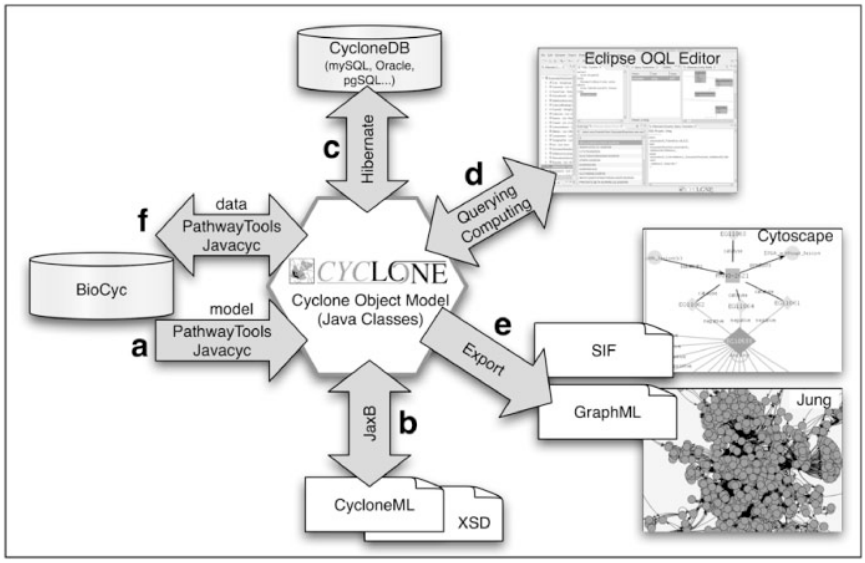


Fig. 1. Information flow in Cyclone from model and data extraction to querying computing and visualization.

Query results can be further manipulated as Java objects. These objects can be stored in CycloneDB using Hibernate (Fig. 1c). They also can be exported to/imported from Cyclone Markup Language files using JAXB (Fig. 1b). Partial exports to other formats (SIF, GraphML) are possible (Fig. 1e). All changes made in Cyclone, such as adding or editing a pathway, can be committed back to BioCyc (Fig. 1f).

3 FUNCTIONALITIES

Cyclone can load an entire PGDB from BioCyc, modify it, for instance by adding user-specific information, and save it back into BioCyc. Cyclone can export and import data in CycloneML, allowing easy interface with other XML tools. The current distribution is fully compatible with BioCyc v9.0 and above.

The Cyclone API allows the extraction of data from BioCyc in order to build biological networks (e.g. bipartite metabolic graph or transcriptional regulatory network). The resulting networks can be manipulated as graphs using the Jung graph library (jung.sourceforge.net), bundled with the Cyclone installation package. They can also be exported in the GraphML exchange format, or in the SIF format, readable by Cytoscape. Cytoscape is a popular and intuitive software tool dedicated to biological networks visualization (Shannon *et al*., 2003).

Cyclone queries can be written in SQL and, more interestingly, in the Hibernate Query Language, (HQL), an object-oriented Query Language HQL. HQL queries are very concise and can express notions such as inheritance, polymorphism and association.

Below is a simple query example: 'Find all enzymes of *Escherichia coli* for which ATP is an inhibitor' (Krummenacker *et al*., 2005) can be written as follows:

```
SELECT er.enzyme
FROM EnzymaticReactions er
WHERE er.Organism = 'Ecoli'
AND er.InhibitorsAll.Value like 'ATP'
```

In conclusion, via its use of mainstream technologies such as Java and XML, Cyclone facilitates the access to PGDBs for a broad community of computational biologists and bioinformaticians. The structured and curated pathways information from these databases thus becomes more readily usable for a variety of exploratory and computational goals.

ACKNOWLEDGEMENTS

We are grateful to the NeMo group for beta testing, to A. Yartseva for the 'Cyclone' name and to the P.Karp group at SRI, for its help with BioCyc. This work was supported by BioSapiens, an EU FP6 Network of Excellence (contract number LSHG-CT-2003-503265).

Conflict of Interest: none declared.

REFERENCES

Eliott,J. (2004) *Hibernate: a developer's notebook*. Sebastopol, CA, USA: O'Reilly Media. ISBN 0596006969.

Lee,T.J. *et al*. (2006) BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics*, **7**, 170.

Kanehisa,M. and Goto,S. (1999) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acid Res*., **27**, 29–34.

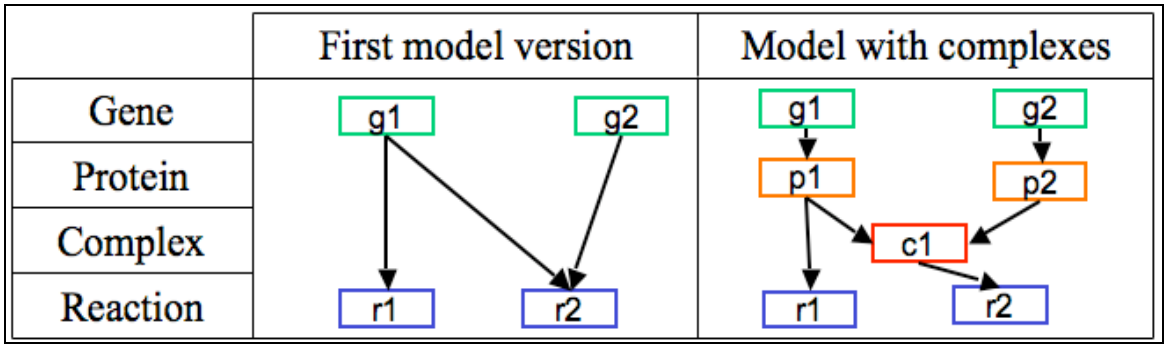
Krummenacker,M. *et al*. (2005) Querying and computing with BioCyc databases. *Bioinformatics*, **16**, 3454–3455.

Karp,P.D. *et al*. (1995) The Generic Frame Protocol. In *Proceedings of the 1995 International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, San Francisco, USA, pp. 768–774.

Karp,P. *et al*. (2002) The pathway tools software. *Bioinformatics*, **18**, S225–S232.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.

Nous avons utilisé Cyclone pour prédire le réseau d'interactions protéiques, et notamment les complexes et leurs structures en sous-complexes dans la bactérie *Acinetobacter*. La méthode de prédiction est basée sur l'orthologie de séquence des protéines des complexes d'*E.coli* avec les protéines de la bactérie *Acinetobacter*. Nous avons prédit ainsi (Durot et al, 2005) 297 complexes enzymatiques. Ils ont été inférés automatiquement en utilisant les données d'EcoCyc. Pour valider ces résultats, nous avons quantifié l'influence de la prise en compte de ces complexes dans les simulations du modèle du métabolisme de l'organisme. Le modèle du métabolisme de la bactérie *Acinetobacter* (Durot et al., 2005) permet par simulation de prédire la croissance ou non de l'organisme sur différents milieux de culture. Les complexes sont importants pour lier les gènes de l'organisme aux enzymes qui catalysent les réactions du métabolisme (voir figure ci-dessous). Les complexes que nous avons prédits permettent d'améliorer de 66% à 85% la ressemblance entre les résultats de croissance obtenus expérimentalement et les résultats issus de la simulation du métabolisme de la bactérie.



Incorporation des complexes prédits dans le modèle du métabolisme de la bactérie *Acinetobacter*. La modélisation de complexes permet de détailler les relations qui lient les gènes aux réactions du métabolisme. Sur l'exemple à gauche, sans complexe, on ne distingue pas si g1 ET g2 ou si g1 OU g2 catalysent la réaction r2. À droite, on modélise que g1 ET g2 sont nécessaires pour former le complexe c1 qui a l'activité enzymatique pour la réaction r2. Reproduit de (Durot et al., 2005).

Cet exemple illustre l'intérêt de travailler avec différents types d'interactions (ici ajouter les informations sur les complexes à l'étude du métabolisme) d'une part, et d'autre part, comment CyClone permet des avancées importantes et automatiques pour le modélisateur.

Reconstruction of a Genome-scale Model of *Acinetobacter ADP1 sp.* Metabolism and Analysis of Phenotypic Profiles

M. Durot, F. Le Fèvre, B. Pinaud, A. Kreimeyer, A. Perret, V. De Berardinis, S. Smidtas, J. Weissenbach, V. Schachter,
Genoscope / UMR-CNRS 8030, Evry, France

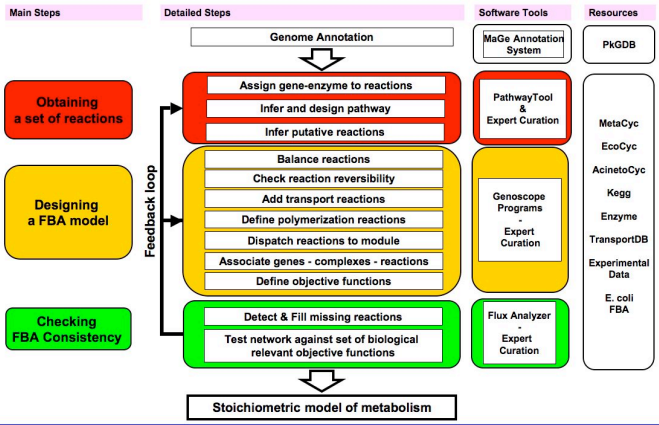
ABSTRACT

In this poster, we present the reconstruction of a genome scale steady-state metabolic model of Acinetobacter sp ADP1. The initial model was derived from genome annotations, biochemical and physiological data. It was then refined using FBA consistency checks, as well as an ongoing process of expert curation. It was successfully validated against large-scale mutant phenotype data (see poster: *De Berardinis, V. et al., "A collection of gene replacement mutants of Acinetobacter sp. ADP1", session Genomics of agriculturally and environmentally important prokaryotes*). Conflicts were used to further refine the model and increase its predictive accuracy.

The current model includes 535 reactions, 498 compounds, 468 genes and 297 enzymatic complexes. This model provides a solid foundation for the interpretation of additional types of high-throughput experimental data, as well as a testbed for the development of machine learning methods.

The modeling framework itself is being improved in order to include regulatory information and to allow the exploitation of metabolite concentration data.

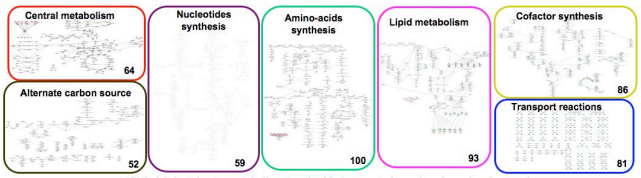
BUILDING THE METABOLIC NETWORK : from genome annotation to stoichiometric model



OUR INITIAL FLUX MODEL:

535 reactions - 498 compounds - 468 genes - 297 enzymatic complexes

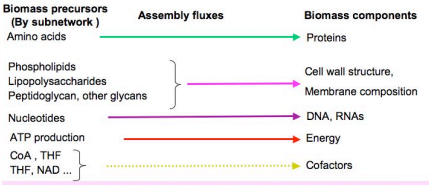
Acinetobacter metabolism has been divided into 7 subnetworks representing principal metabolic functions and a "transport subnetwork":
→ Benefit: facilitate reconstruction (consistency checks, expert curation),
→ Need for methodology & software infrastructure to manage subnetworks ↔ global network .



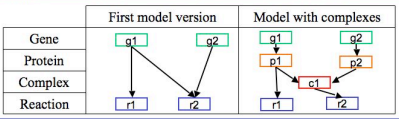
Global Acinetobacter metabolic network with the count of reactions in each subnetwork.

From subnetworks to the biomass production

Proportion of metabolites in each objective function are based on : Acinetobacter experimental data, and E. coli experimental data when necessary.



Model upgrade: refining gene-reaction correspondance using complexes predictions



297 enzymatic complexes were inferred by automatic comparison with E. coli complexes annoted in EcoCyc.

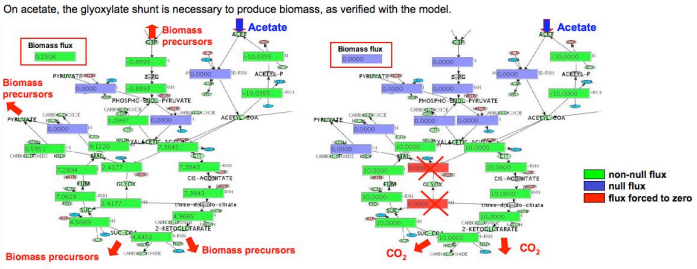
Mutant phenotype prediction accuracy increases from 66% to 85% by adding complexes in our model.

CONSISTENCY CHECKS

The reconstructed stoichiometric model can be checked against known basic metabolic behaviour to verify consistency. The goal is not to fine-tune the model but rather to correct modelling mistakes that show through global analyses. These global checks can also be used to scale some quantitative parameters such as flux yields.

Use of the glyoxylate shunt on Acetate carbon source

On acetate, the glyoxylate shunt is necessary to produce biomass, as verified with the model.



Normal pathway through glyoxylate — *Carbon can only be transformed to CO_2 and cannot yield biomass*

Maximal ATP production from several carbon sources

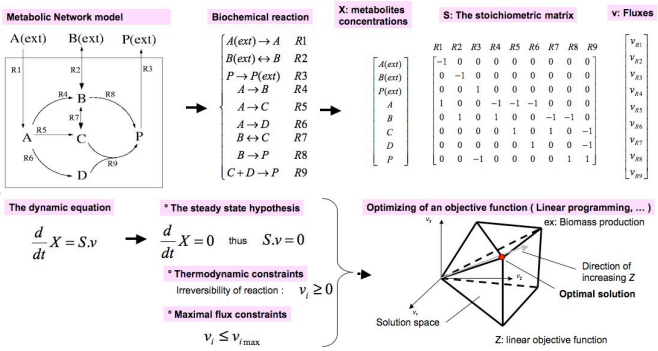
By setting ATP production as an objective function, we can assess the global efficiency of respiration for ATP synthesis. These results were used to help fine-tune the stoichiometric coefficients of the respiratory fluxes.

	akGlut.	Acetate	Glucose	Lactate	Pyruv.	Sucu.
02 consumption	8	4	11	6	5	7
ATP production	13	5,5	18	8,5	8	10,5

INTRODUCTION TO STOICHIOMETRIC MODELS AND FLUX BALANCE ANALYSIS (FBA)

Steady-state assumption.

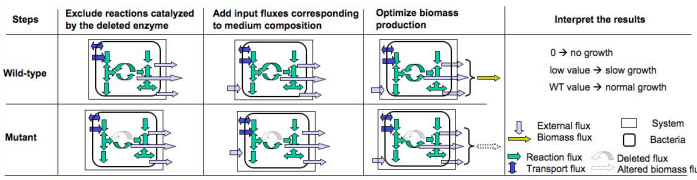
Variables of interest : metabolic fluxes
State of the system : distribution of fluxes through all reactions
Key ideas :
- Reason on sets of metabolic states rather than on unique « solution » states
  Allows for incompleteness or inaccuracies in both the reaction network structure and the experimental constants
  The feasible solution space can be progressively reduced using constraints (thermodynamic, input/output, regulatory…
- Decompose metabolic activity on a set of elementary modes / extreme pathways
  Elementary modes/ extreme pathways enjoy a natural mathematical & biochemical interpretation
- Predict optimal (or good) flux distributions given a defined metabolic objective (set by an objective function)
  Extensible to sets of objectives



MODEL VALIDATION & REFINEMENT : PREDICTING GROWTH PHENOTYPES

Principle of mutant phenotype prediction

Without considering the isozymes, every reactions linked to the deleted genes are excluded from the wild-type stoichiometric model, designing a new network configuration. Using biomass production optimization, each new model is tested on the selected media by allowing or not the external compounds to enter the flux network. The flux value of biomass production determines the mutant's phenotype.



Prediction accuracy for a set of 182 mutants

Comparison with experimental phenotypes on Acinetobacter ADP1 sp collection.

	akGlutara.	Acetate	Glucose	Lactate	Pyruvate	Succinate
Match	84,1%	84,1%	83,5%	85,2%	85,2%	84,1%
Perfect match	72,0 %	64,3 %	72,5 %	73,6 %	75,3 %	73,6 %
Approx match	12,1 %	19,8 %	11,0 %	11,5 %	9,9 %	10,4 %
False +	9,3 %	9,3 %	11,5 %	9,3 %	9,9 %	8,8 %
False -	6,6 %	6,6 %	5,0 %	5,5 %	4,9 %	7,1 %

An approximated match is a match between normal growth and slow growth.

Many of the false negative can be explained by the presence of redundant enzymes or enzymatic complexes. The gene-to-reaction correspondence is being systemically updated.

Exploiting conflicts to refine the model

Gene	Initial enzymatic activity	akGl	Acet.	Gluc.	Lact.	Pyru.	Succ.	Possible explanations
0476	Asparaginase	0/1	0/1	0/1	0/1	0/1	0/1	Allowed to find an isozyme for the reaction
2879	Succinate dehydrogenase	0/1	0/1	1/1	1/1	1/1	0/1	Detection of experimental error: gene was not deleted
0272	Glutamate--tRNA ligase	1/1	1/1	1/0	1/1	1/0	1/0	Hypothesis: pair of isozymes, constitutive versus regulated expression
3371	Glutamate--tRNA ligase	1/0	1/0	1/0	1/0	1/0	1/0	
2875	Oxoglutarate dehydrogenase	1/0	1/0	1/0	1/0	1/0	1/1	Evidence for the existence of complex rather than isozymes
2876	Oxoglutarate dehydrogenase	1/0	1/0	1/0	1/0	1/0	1/1	
0546	G3P dehydrogenase	1/1	1/1	1/0	1/0	1/1	1/1	Detection of experimental error: gene in Entner-Doudoroff operon
2893	P-Pantetheine adenylyltransferase	1/0	1/0	1/0	1/0	1/0	1/0	CoA has to be added to biomass production

IS: *in silico* / IV: *in vitro*.
0 : 0 → no growth 1 : WT value → normal growth

0/1 False negative
1/0 False positive 1/1 True positive

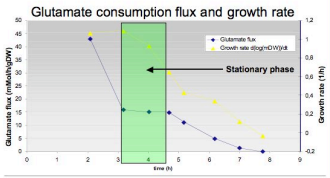
A QUANTITATIVE ANALYSIS : PREDICTING BIOMASS PRODUCTION

Biomass production flux is determined in vivo and in silico for a specific input flux of glutamate.

Glutamate concentration was assayed using commercial glutamate dehydrogenase.

The glutamate flux and growth rate are approximated from concentration and optical density values

The relationship between optical density (λ=600 nm) and dry weight was experimentally determined on the wild type strain.



For an input flux of glutamate of 15 mMol/h/gDW in areobic condition, the gap between the observed growth rate and the one predicted by our model is **less than 20%**

	Growth rate
In Vivo	1,19 h^{-1}
In Silico	1,03 h^{-1}

CONCLUSION

Outlines

- Initial reconstruction of a metabolic flux model of *Acinetobacter* sp. ADP1.
- Development of a methodology and a software infrastructure for the reconstruction of metabolic models

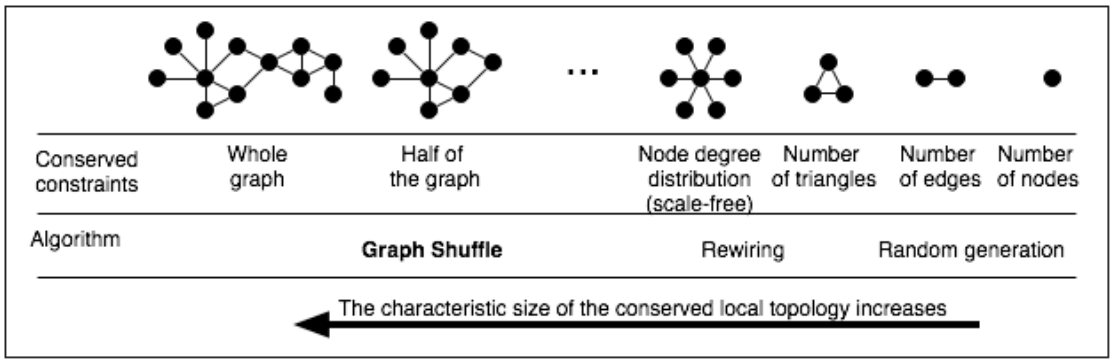
Ongoing work

1. Refinement of the model using biological expertise:
   - Complete the gene-to-reaction correspondence (isozymes and enzymatic complexes).
   - Improve assembly fluxes/objective functions to better fit with *Acinetobacter* metabolism specificities.
   - Include additional putative reactions in the network.
   - Annotate and include a more complete set of transporters.
   - Refine biomass production fluxes (cofactors and alternate carbon sources subnetworks).
2. Design of an integrated qualitative modeling framework for metabolism and regulation.
3. Design of automated refinement methodology using experimental data on metabolite fluxes and concentrations (MS), and mRNA expression.

Chapitre IV – Corrélations globales de RIBH

Pour comprendre comment les réseaux homogènes interagissent ensemble en formant un réseau hétérogène, nous avons étudié la corrélation entre les propriétés des réseaux homogènes composants. L'originalité de l'approche consiste à se référer à des graphes rendus aléatoires partageant comme invariant une composante géante.

Nous nous intéressons ici à l'analyse de réseaux hétérogènes s'appuyant sur la comparaison avec des réseaux aléatoires qui préservent la structure macroscopique. La méthode introduite de *graph shuffle* est appliquée au réseau d'interactions protéine-protéine couplé au réseau de régulation transcriptionnelle.



Algorithme *Graph Shuffle* parmi l'ensemble des stratégies de perturbation de graphes.

Il y a plusieurs types de propriétés qui peuvent être conservés dans des graphes. La taille caractéristique de la partie rendue aléatoire croît de la gauche vers la droite sur la figure ci-dessus. Ici nous proposons une méthode qui conserve une des propriétés des parties géantes du réseau hétérogène (voir figure ci-dessus). Chacun des sous-graphes homogènes sont conservées dans leur totalité mais pas la manière dont les sous-graphes interagissent. En permutant ces réseaux, sous la contrainte de conserver les sous-graphes d'interaction protéine-protéine d'une part et de régulation transcriptionnelle

d'autre part, nous cherchons des propriétés statistiques non triviales, dans le sens où elles sont différentes des propriétés de réseaux aléatoires.

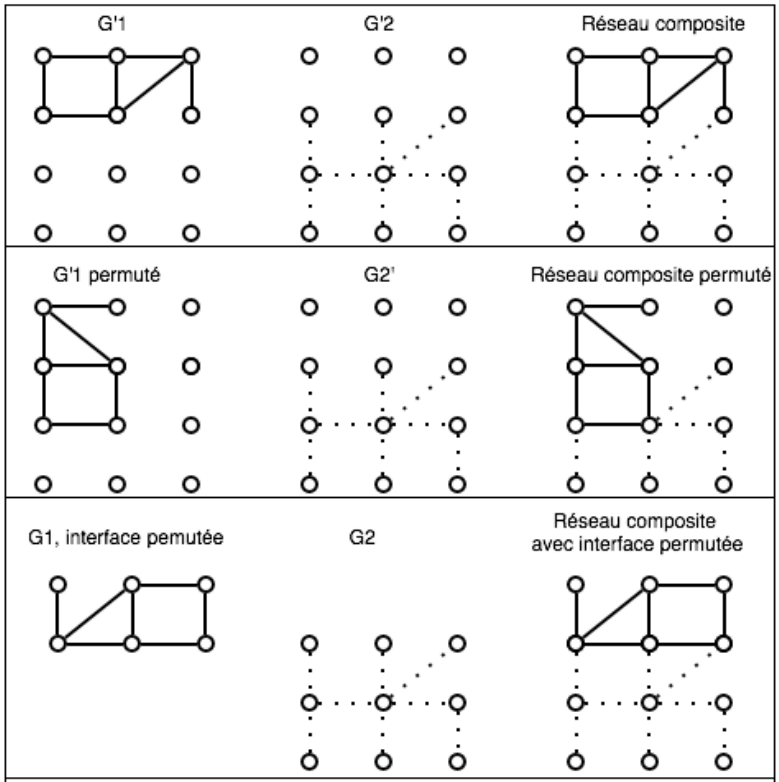


Illustration de la procédure de permutation (*Graph shuffle*) entre deux graphes G1, G2

Illustration de la procédure de permutation entre deux graphes G1 et G2. En haut figure le réseau simplement composé de G1 et G2. Au milieu figure le réseau composé du réseau G2 et du réseau G'1 qui est le réseau G1 permuté dans l'ensemble des protéines. En bas, figure le réseau composé de G2 et du réseau G''1 qui est le réseau G1 permuté dans l'ensemble des protéines liées à G1.

Parmi l'ensemble des réseaux qui conservent les deux sous réseaux, le réseau réel présente à la fois une bi-connectivité (le nombre de paires de nœuds connectés par les deux sous réseaux) supérieure (nombre de paire de nœuds connectés dans les deux sous graphes indépendamment), et des distances plus grandes que la moyenne. De plus, au moyen de formes plus restrictives de permutation, qui conservent l'interface entre les deux sous réseaux, nous montrons que ces propriétés sont indépendantes : les permutations restreintes tendent à produire des réseaux plus compacts, sans perdre en bi-connectivité. Enfin, nous proposons une interprétation des propriétés ci-dessus en termes de capacité de voie de signalisation dans le réseau. L'étude conjointe de ces

réseaux révèle une coopération entre ces réseaux et valide l'intérêt de l'intégration de données.

Les résultats suivants ont été présentés à Computational Methods in Systems Biology CMSB 2005, le 5 Avril 2005 à Edinburgh en Scotland, et publiés dans LNCS Transactions on Computational Systems Biology, Springer. Le travail suivant est un travail d'équipe dans lequel ma contribution a été principalement de proposer et mettre place la procédure *Graph Shuffle*, et tester cette procédure sur des données préliminaires de la levure.

Property-driven statistics of biological networks

Pierre-Yves Bourguignon[1], Vincent Danos[2], François Képes[3], Serge Smidtas[1], and Vincent Schächter[1]

[1] Genoscope
[2] CNRS & Université Paris VII
[3] CNRS

Abstract. An analysis of heterogeneous biological networks based on randomizations that preserve the structure of component subgraphs is introduced and applied to the yeast protein-protein interaction and transcriptional regulation network. Shuffling this network, under the constraint that the transcriptional and protein-protein interaction subnetworks are preserved reveals statistically significant properties with potential biological relevance. Within the population of networks which embed the same two original component networks, the real one exhibits simultaneously higher bi-connectivity (the number of pairs of nodes which are connected using both subnetworks), and higher distances. Moreover, using restricted forms of shuffling that preserve the interface between component networks, we show that these two properties are independent: restricted shuffles tend to be more compact, yet do not lose any bi-connectivity.
Finally, we propose an interpretation of the above properties in terms of the signalling capabilities of the underlying network.

1 Introduction

The availability of genome-scale metabolic, protein-protein interaction and regulatory networks [25, 7, 3, 5, 21] —following closely the availability of large graphs derived from the Internet hardware and software network structure, from social or collaborative relationships— has spurred considerable interest in the empirical study of the statistical properties of these 'real-world' networks. As part of a wider effort to reverse-engineer biological networks, recent studies have focused on identifying *salient* graph properties that can be interpreted as 'traces' of underlying biological mechanisms, shedding light either on their dynamics [23, 11, 6, 28] (*i.e.*, how the connectivity structure of the biological process reflects its dynamics), on their evolution [10, 30, 27] (*i.e.*, likely scenarios for the evolution of a network exhibiting the observed property or properties), or both [9, 14, 15]. The statistical graph properties that have been studied in this context include the distribution of vertex degrees [10, 9], the distribution of the clustering coefficient and other notions of density [17–19, 22, 4], the distribution of vertex-vertex distances [22], and more recently the distribution of network motifs occurrences [15].

Identification of a salient property in an empirical graph —for example the fact that the graph exhibits a unexpectedly skewed vertex degree distribution— requires a prior notion of the distribution of that property in a class of graphs relatively to which saliency is determined. The approach chosen by most authors so far has been to use a *random graph model*, typically given by a probabilistic graph generation algorithm that constructs graphs by local addition of vertices and edges [20, 1, 24]. For the simplest random graph models, such as the classical Erdös-Rényi model (where each pair of vertices is connected with constant probability p, [2]), analytical derivations of the simplest of the above graph properties are known [20, 1].

In the general case, however, analytical derivation is beyond the reach of current mathematical knowledge and one has to retort to numerical simulation. The random graph model is used

to generate a sample of the corresponding class of graphs and the distribution of the graph property of interest is evaluated on that sample, providing a standard against which the bias of the studied graph can be measured [23, 14, 29]. Perhaps because of the local nature of the random graph generation process, it is mostly simple *local* network properties that have been successfully reproduced in that fashion. Another, somewhat more empirical, category of approaches reverses the process: variants are generated from the network of interest using a random rewiring procedure. The procedure selects and moves edges randomly, preserving the global number of edges, and optionally their type, as well as local properties such as the degree of each vertex. Rewirings are thus heuristic procedures which perform a sequence of local modifications on the structure of the network.

The specific focus of the present paper is on measuring the degree of cooperation between the two subgraphs of the yeast graph of interactions induced by the natural partition of edges as corresponding either to transcriptional interaction (directed) or to protein protein interaction (undirected). To evaluate a potential deviation with respect to such a measure, one needs as a first ingredient a suitable notion of random variation of the original graph. The goal is here, as in many other cases, to contrast values of a given observable on the real graph, against the distribution of those same values in the population of variants. We define *shuffles* of the original graph as those graphs that are composed exactly of the original two subgraphs of interest, the variable part being the way these are 'glued' together.

From the probabilistic point of view, this notion of randomisation coincides with a traditional Erdös-Renyi statistics, except that it is conditioned by the preservation of the original subgraphs. Designing a generative random graph model that would only yield networks preserving this very precise property seems to be a hard endeavor ; it is not as easy as in the unconditional Erdös-Renyi model to draw edges step by step yet ensure that component subgraphs will be obtained in the end. Shuffling might also be seen as rewiring, except the invariant is large-scale and extremely precise: it is not edges that are moved around but entire subgraphs. Moving edges independently would break the structure of the subnetworks, and designing a sequential rewiring procedure that eventually recovers that structure is not an obvious task. Moreover, it would be in general difficult to ensure the uniformity of the sample ; see [16] for a thorough analysis of rewiring procedures. This choice of an invariant seems rather natural in that one is interested in qualifying the interplay between the original subgraphs in the original graph. Now, it is not enough to have a sensible notion of randomisation, it is also crucial to have a computational handle on it. Indeed, whatever the observable one wants to use to mark cooperation is, there is little hope of obtaining an analytic expression for its distribution, hence one needs sampling. Fortunately, it turns out it is easy to generate shuffles uniformly, since these can be described by pairs of permutations over nodes, so that one can always sample this distribution for want of an exact expression. As explained below in more details, the analysis will use two different notions of subgraph-preserving sampling: *general* shuffles, and *equatorial* ones that also preserve the interface between our two subgraphs. Equatorial shuffles are feasible as well, and in both cases the algorithms for sampling and evaluating our measures turn out to be fast enough so that one can sweep over a not so small subset of the total population of samples.

Regarding the second necessary ingredient, namely which observable to use to measure in a meaningful way the otherwise quite vague notion of cooperation, there are again various possibilities. We use two such observables in the present study: the *connectivity*, defined as the percentage of disconnected pairs of nodes, and a refined quantitative version of connectivity, namely the full distance distribution between pairs of nodes. The latter is costlier, requiring about three hours of computation for each sample on a standard personal computer.

Once we have both our notion of randomisations and our observables in place, together with a feasible way of sampling the distribution of the latter, we can start. Specifically we run four

experiments, using general or equatorial shuffling, and crude or refined connectivity measures. The sampling process allows us to compare the values of these measures for the original graph with the mean value for the sample, and, based on the assumption that those values follow a normal distribution over the sample, one can also provide a p-value that gives a rough estimate of the statistical deviation of the observable in the given graph.

The general shuffle based experiments show with significant statistical confidence that shuffling reduces connectivity (1), and at the same time contracts distances (2). More precisely, both bi-connectivity (the amount of pairs of nodes which are connected using both subgraphs) and distances are higher than average in the real network. A first interpretation might be that the real graph is trading off compactness for better bi-connectivity. In order to obtain a clearer picture and test this interpretation, we perform two other experiments using equatorial shuffles. Surprisingly, under equatorial shuffles connectivity hardly changes, while the global shift to shorter distances is still manifest. It seems therefore there is actually no trade-off, and both properties (1) and (2) have to be thought of as being independently captured by the real graph. With appropriate caution, we may try to provide a biological interpretation of this phenomenon. Since all notions of connectivity and distances are understood as directed, we propose to relate this to signalling, and interpret bi-connectivity as a measure of the capability to convey a signal between subgraphs. With this interpretation, the above properties may be read as:(1) signal flows better than average and (2) signal is more specific than average. The second point requires explanation. At constant bi-connectivity, longer average distances imply that upon receipt of a signal, the receiver has a better chance of guessing the emitter. In other words, contraction of distances (which can be easily achieved by using hubs) will anonymise signals, clearly not a desirable feature in a regulatory network. Of course this is only part of the story, since some hubs will also have an active role in signal integration and decision making. The latter is probably an incentive for compactness. If our reading of the results is on track, we then may think of the above experiments as showing that the tropism to compactness due to the need for signal integration, is weaker than the one needed for signal specificity.

Beyond the particular example we chose to develop here because of the wealth of knowledge available on the yeast regulatory and protein interaction networks, one can think of many other applications of the shuffling methodology for heterogeneous networks. The analyses performed here rely on edges corresponding to different types of experimental measurements, but edges could also represent different types of predicted functional links. Indeed, there are many situations where a biological network of interactions can be naturally seen as heterogeneous. Besides, the notions of shuffle we propose can also accomodate the case where one would use a partition of nodes, perhaps given by clustering, or localisation, or indeed any relevant biological information, and they may therefore prove useful in other scenarios.

The paper is organised as follows: first, we set up the definitions of edge-based general and equatorial shuffles based, and also consider briefly node-based shuffles though these are not used in the sequel; then we describe the interaction network of interest and the way it was obtained; finally we define our observables and experiments, and interpret them. In the conclusion, we discuss generalization and potential applications of the method. The paper ends with an appendix on the algorithmical aspects of the experiments, and a brief recall of the elementary notions of statistics we use to assert their significance.

2 Shuffles

Let $G = (V, E)$ be a directed graph, where V is a finite set of nodes, and E is a finite set of directed edges over V. We write M for the incidence matrix associated to G. Since G is directed,

M may not be symmetric. In the absence of parallel edges M has coefficients in $\{0,1\}$, where parallel edges are allowed.

Given such a matrix M and a permutation σ over V, one writes $M\sigma$ for the matrix defined as for all u, v in V:

$$M\sigma(u,v) := M(\sigma^{-1}u, \sigma^{-1}v)$$

Note that $M\sigma$ defines the same abstract graph as M does, since all σ does is changing the nodes names.

2.1 Shuffles Induced by Properties on Edges

We consider first shuffles induced by properties on edges. Suppose given a partition of $E = \sum E_i$; this is equivalent to giving a map $\kappa : E \to \{1, \ldots, p\}$ which one can think of as colouring edges.

Define M_i as the incidence matrix over V containing the edges in E_i (of colour i).

Define also V_I, where $I \subseteq \{1, \ldots, p\}$, as the subset of nodes v having for each $i \in I$ at least one edge incident to v with colour i, and no incident edge coloured j for $j \notin I$. We abuse notation and still write $\kappa(u) = I$ when $u \in V_I$. This represents the set of colours seen by the nodes.

Clearly $V = \sum V_I$, $V_\varnothing$ is the set of isolated nodes of G, and the set of nodes of G_i is the union of the graphs generated by V_I, for $i \in I$.

Given $\sigma_1, \ldots, \sigma_p$ permutations over V, define the *global shuffle* of M as:

$$M(\sigma_1, \ldots, \sigma_p) := \textstyle\sum_i M_i\sigma_i$$

The preceding definition of $M\sigma$ is the particular case where $p = 1$ (one has only one colour common to all edges). Each G_i (the abstract graph associated to M_i) is preserved up to isomorphism under this transformation. However the way the G_is are glued together is not, since one uses a different local shuffle on each.

For moral comfort, we can check that any means of glueing together the G_is is obtainable using a general shuffle in the following sense: given G' and $\sum q_i : \sum G_i \to G'$ where the disjoint sum $\sum_i q_i$ is an isomorphism on edges, one has that G' is a general shuffle of G. To see this, define $\sigma_i(u) := q_i p_i^{-1}(u)$ if $u \in \kappa^{-1}(i)$, $\sigma_i(u) = u$ else (we have written p_i for the inclusion of G_i in G), one then has $G' = \sum G_i\sigma_i = G(\sigma_1, \ldots, \sigma_p)$.

Note also that $(M(\sigma_1, \ldots, \sigma_p))\tau := \sum_i M_i(\tau\sigma_i)$, and so in particular, without loss of generality one can take any the σ_i's to be the identity (just take $\tau = \sigma_i^{-1}$). This is useful when doing actual computations, and avoids some redundancy in generating samples.

An additional definition will help us refine the typology of shuffles. One says a shuffle $M\sigma$ is *equatorial* if in addition for all I, and all i, V_I is closed under σ_i. Equivalently, one can ask that $\kappa \circ \sigma_i = \kappa$. An *equatorial shuffle* preserves the set of colours associated with each node and in particular preserves for a given pair of nodes (u, v) the fact that (u, v) is heterochromatic, *i.e.*, $\kappa(u) \cap \kappa(v) = \varnothing$. This in turn implies that the distance between u and v must be realised by a path which uses edges of different colours. In the application such paths are mixing different types of interaction, and are therefore of particular interest; without preserving this attribute, an observable based on path with different colours would not make sense. In the particular case of two colours, nodes at the 'equator', having both colours, will be globally preserved, hence the name.

2.2 Shuffles Induced by Properties on Vertices

One can also consider briefly shuffles induced by properties on nodes. Suppose then given a partition of nodes $V = \sum_i V_i$, again that can be thought of as a colouring of nodes $\kappa : V \to \{1, \ldots, p\}$, and extended naturally to the assignment of one or two colours to each edge.

A node shuffle is defined as a shuffle associated to σ which can be decomposed as $\sum_i \sigma_i$, σ_i being a permutation over each cluster V_i. Clearly each graph G_i generated by V_i is invariant under the transformation: only the inter-cluster connectivity is modified.

The equivalent of the equatorial constraint would be to require in addition $\sigma(u) \in \partial V_i$ if $u \in \partial V_i$, where ∂V_i is defined as those nodes of V_i with an edge to some V_j, $i \neq j$. Other variants are possible and the choice of the specific variant will likely depend on the particular case study. We now turn to the description of the network the shuffle experiments will be applied to.

3 A Combined Network of Regulatory and Protein-Protein Interactions in Yeast

With our definitions in place, we can now illustrate the approach on a heterogeneous network obtained by glueing two component networks.

It is known that regulatory influences, including those inferred from expression data analysis or genetic experiments, are implemented by the cell through a combination of direct regulatory interactions and protein-protein interactions, which propagate signals and modulate the activity level of transcription factors. The detailed principles underlying that implementation are not well understood, but one guiding property is the fact that protein interaction and transcriptional regulation events take place in the regulatory network at different time-scales.

In order to clarify the interplay between these two types of interactions, we have combined protein-protein (PPI) and protein-DNA (TRI, for 'transcriptional regulation interaction') interaction data coming from various sources into a heterogeneous network by glueing together these two networks on the underlying set of yeast proteins.

The data from which the composite network was built includes: 1440 protein complexes identified from the literature, through HMS-PCI or TAP [3, 5], 8531 physical interactions generated using high-throughput Y2H assays [26], and 7455 direct regulatory interactions compiled from literature and from ChIP-Chip experiments [4, 12], connecting a total of 6541 yeast proteins. A subnetwork of high-reliability interactions was selected, using a threshold on the confidence levels associated to each inferred interaction. For the ChIP-Chip data produced by Lee et al. [12], interactions with a p-value inferior to 3.10^{-2} were conserved ; for the Y2H data produced by Ito et al. [26], a threshold of 4.5 on the Interaction Sequence Tag was used (see [8]). The PPI network was built by connecting two proteins, in both directions, whenever there was a protein-protein or a complex interaction between the two corresponding proteins. In the case of the TRI network, an edge connects a regulator protein with its regulatee. To simplify the discussion, we will refer in the rest of the paper to the TRI graph as TRI, and to the PPI graph as PPI. With some more precision, define G as the real graph, TRI as the subgraph induced by the set of TRI nodes, *i.e.,* nodes such that $TRI \in \kappa(u)$, and PPI as the subgraph induced by the set of PPI nodes.

Their respective sizes are:

$$TRI = 3387,\ PPI = 2517,\ TRI \cup PPI = 4489,\ TRI \cap PPI = 1415$$

The set of nodes $TRI \cap PPI$ of both colours is also referred to in the sequel as the *equator* or the *interface*. Since the object of the following is to discuss the interplay between the TRI and PPI subgraphs, the interface naturally plays an important role. A qualitative measure of the connectivity between TRI and PPI which will be useful later in the discussion, is the number of bi-connected pairs in G (these are the pairs which are connected in G, but not connected in either TRI or PPI), which is roughly $p_{bi} = 23\%$. To complete this statistical portrait of the data, we provide in figure 1 the histograms of degree distributions in the PPI and TRI networks, with in and out degrees pictured separately for the latter. Figure 1 also shows the hub size distribution

for the TRI network (the PPI network has no non-trivial hubs). Note that hubs are defined as sets of nodes connected to a single node. The TRI network (here considered as unoriented) has 124 such hubs ; the histogram of the distribution of their sizes is given in figure 1.

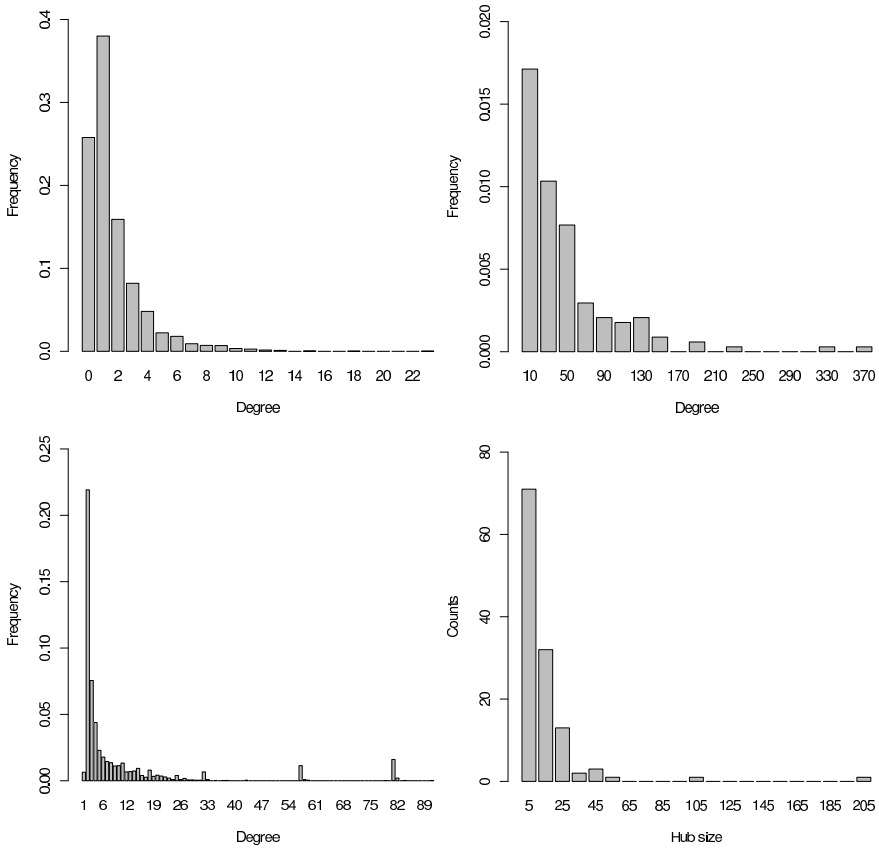


Fig. 1. First row: Histograms of the in and out degree distributions of the TRI network. Second row: Histogram of the degree distribution of the PPI network and of the distribution of the hub size in the TRI network.

4 Results and Interpretations

Hereafter, notions of connectivity, distance, etc. should be understood as *directed* unless explicitly stated otherwise. We now turn to the various shuffle experiments and consecutive observations.

4.1 General Shuffle vs Connectivity

We take here as a rough measure of the connectivity of a graph the percentage of unconnected pairs. Comparing first the real graph with the randomised versions under the general shuffle, one

finds that in the average 4% of the population pairs are disconnected under shuffle. So general shuffle disconnects, or in other words G maximises bi-connectivity.

Clearly mono-connected pairs (pairs connected in either PPI or TRI) cannot be disconnected under general shuffle; a pair is 'breakable' only if bi-connected in G; therefore a more accurate measure of the connectivity loss under general shuffle is that about 17.5% of the breakable pairs are actually broken (this obtained by dividing by p_{bi}), a rather strong deviation with a p-value below 10^{-11}.

Inasmuch as a directed path can be thought of as a signal-carrying pathway, one can interpret the above as saying that the real graph connects PPI and TRI so as to maximise the bandwidth between the subgraphs.

4.2 Equatorial Shuffle vs Connectivity

Keeping with the same observable, we now restrict to equatorial shuffles. One sees in this case that no disconnection happens, and actually about 1% more pairs are connected *after* shuffling. The default of connected pairs of the real graph has a far less significant p-value of 3%. However the point is that equatorial shuffles leaves bi-connectivity rather the same.

This complements the first observation and essentially says that the connectivity maximisation seen above is a property of the set of equatorial nodes ({TRI,PPI} nodes) itself, and not of the precise way TRI and PPI edges meet at the equator.

Both observations can be understood as saying that the restriction of G to the equator is a much denser subgraph than its complement (as evidenced by the connectivity loss under general shuffle), and dense enough so that equatorial shuffling does not impact connectivity.

Note that so far the observable is somewhat qualitative, being only about whether a pair is connected or not. Using a refined and quantitative version of connectivity, namely the distribution of distances (meaning for each n the proportion of pairs at distance n), will reveal more.

4.3 Impact of Equatorial Shuffles on Distance Distribution

Using this refined observable, one sees that the whole histogram shifts to the left, so equatorial shuffle contracts the graph (Fig. 2). This is confirmed by the equality between the number of lost pairs at distance 7 to 9 and the number of new ones at distance 3 to 5. In accordance with the preceding experiment, one also does not see any disconnection under equatorial shuffle.

This is to be compared with the general shuffle version (Fig. 3) where both effects are mixed, and the cumulated excess of short pairs does not account for the loss of long pairs (indeed we know 4% are broken, *i.e.*, disappear at infinity and are not shown on the histogram).

To summarize the distance distribution results in a single number, one can compute the deviation of the real graph mean distance under both shuffles. As expected the mean distance is higher in the real graph with respective p-values of 0.2% and 2% in the general and equatorial shuffles (see Appendix for details). We conclude that while the real graph does maximise bi-connectivity, it does not try to minimise the associated distances.

To provide an intuition on the potential interpretation of the above result, let us again consider paths as rough approximations of signalling pathways. Now compare a completely linear chain-shaped graph and star-shaped one, with the same number of nodes and edges. In the star case, any two nodes are close, at constant distance 2, while in the chain distances are longer. As said, compactness comes with a price, namely that in a star graph all signals go through the hub and are anonymised, *i.e.*, there may be a signal, but there is no information whatsoever in the signal about where the signal originated from. Quite the opposite happens in a linear graph. Of course this is an idealized version of the real situation; nevertheless it is tempting to interpret this last

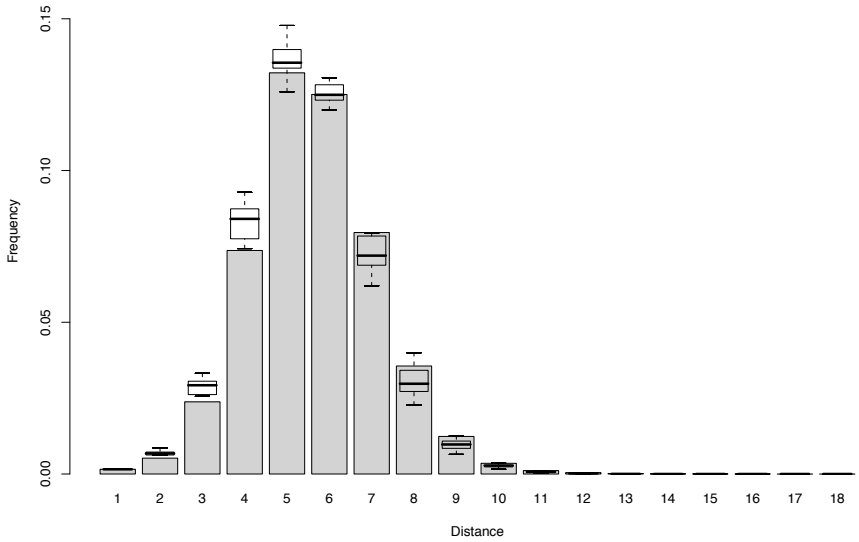


Fig. 2. Equatorial shuffle distance histogram: grey boxes stand for the real graph; one sees that shuffles have more pairs at shorter distance, and consequently (because the number of connected pairs is about the same) less such at higher distances.

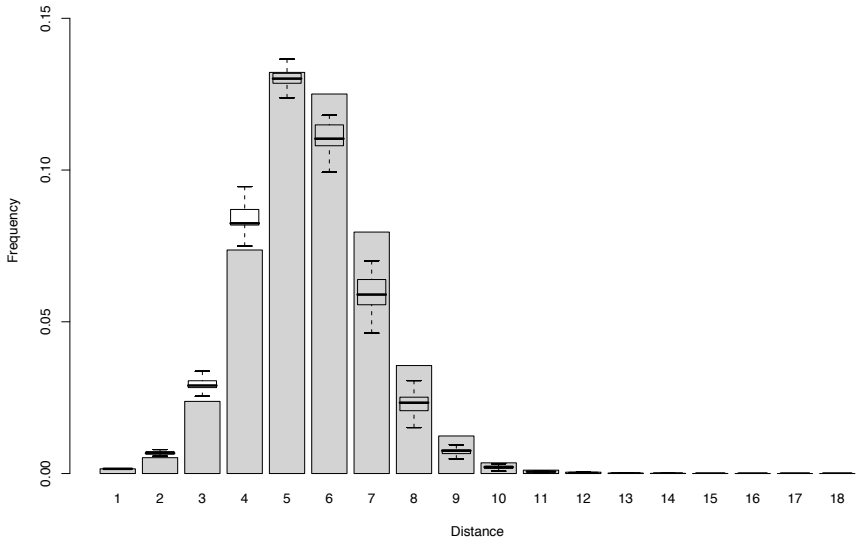


Fig. 3. Global shuffle distance histogram

observation as an indication that the real graph is trading off fast connectivity against specificity of signals. The heterogeneous network is likely to result from a trade-off between causality and signal integration.

As suggested in the introduction, finer observables would have to be developed to further refine this interpretation. Furthermore, there are intrinsic limits on the nature of properties that can be identified using pure topology; deeper, reliable insights about signal transmission in the joint network will ultimately require a dynamical view of signaling with corresponding experimental data.

5 Conclusion

In order to assess the cooperation between the network of protein-protein interactions and the regulatory network in yeast, we have defined two notions of shuffle, *i.e.* tractable randomisations of the original network that preserve global invariants. While general shuffles preserve the entire structure of the component subnetworks, equatorial shuffles also preserve the interface between the networks. We assessed cooperation between the subnetworks using two observables: the percentage of connected node pairs, and the distribution of distances between nodes. For each shuffle-observable pair, the observable in the real network was assessed against the distribution of observables in the set of network variants generated by the respective shuffle.

To summarise the results of this case study, we can say that the statistical analysis of G shuffles under the constraint of preserving its component subnetworks suggests the existence of two *independent* properties of G regarding the cooperation of its components:

- bi-connectivity, i.e. the proportion of node pairs connected only by paths using both types of edges, is higher in the actual network than in the shuffles;
- distances between pairs of nodes are higher in the actual network;

The first property can be given an interpretation in terms of bandwidth: signals flow better between the two networks than would be expected if they were connected randomly. The second property can be interpreted as favoring signal specificity: for cellular interaction networks (in contrast with telecommunication networks, for instance, where each packet carries significant intrinsic information) the information borne by a signal is very much related to the path it has followed. Longer paths thus provide more opportunity for specific signals. Note that the fact that we worked with directed notions (and not with undirected ones as we did in a first version of this paper) makes the interpretation of paths as potential signaling pathways somewhat more convincing.

We have been careful in the discussion of the results of our statistical experiments in terms of signalling capacities, and this needs to be thrashed out in subsequent work. To do so one would first need refined and yet feasible observables pertaining to the dynamics of the network of interest. A recent paper equips the subgraph induced by the major molecular players in the budding yeast cell cycle (cyclins, their inhibitors, and major complexes) with a discrete Boolean dynamics [13], and obtains a dynamics with a stable state corresponding to the G_1 phase, which is attracting a significantly higher number of states than a random graph (with the same number of nodes and edges) would. It seems therefore possible to explicitly construct signal-related observables. However there are several problems: first, this analysis relies on sorting positive and negative regulation edges, and that is an information which one doesn't have for the full graph; second it also critically relies on the rather small size of the subgraph; finally the model only handles a limited number of signals (corresponding to the various cell cycle phases). Nevertheless, a comparable study, using shuffles as a means of randomising, and confined to a

well-chosen subgraph could help in qualifying our speculative interpretation of the contraction phenomenon we have observed.

On the methodological front, both the general notion of shuffle and the restricted notion of equatorial shuffle proved useful: they reveal different properties and complement one another. The same holds for the pair of observables: both the qualitative connectivity observable and its refined distance-based version are useful, and yield different and complementary insights on the cooperation between the two component networks.

We believe that the shuffling methodology developed for this case study has general applicability to the study of heterogeneous biological networks, i.e. networks that can be seen as the "glueing" of two or more component networks. Shuffles preserve global invariant properties (the structure of component networks), and define rigorously and unambiguously the class of networks which obey these properties. They are also easily computable and can be generated uniformly, by drawing from a set of acceptable permutations. Note that the latter property is in contrast with randomizations based on sequential rewiring strategies, where each rewiring step perturbs the structure while preserving one or more local invariants. While these approaches may prove to be asymptotically equivalent in some cases, they typically do not provide a direct definition nor the means to uniformly sample the set of randomizations which preserve the invariant, since the order of the rewiring steps matters.

Given an interaction network between biochemical species, any biological property on edges (type of interaction, degree of confidence, localization of interaction...) or on nodes (type of entity, functional annotation, inclusion within clusters generated using a given data type and methodology, etc...) with discrete values can be used to define a heterogeneous version of that network. Then, either the type of edge shuffles used above, or shuffles preserving other categories of top-down invariants, such as the projection of a network onto a given network of abstract clusters, could be explored. Likewise, a variety of observable properties may be used to investigate cooperation between component subnetworks. Perhaps the foremost promise of the shuffling approach resides in the interplay between different shuffle-observable pairs, which allows an exploratory assessment of cooperation adapted to the heterogeneous network at hand.

References

1. William Aiello, Fan Chung, and Linyuan Lu. A random graph model for massive graphs. In *32nd Annual ACM Symposium on Theory of Computing*, pages 171–180, 2000.
2. P. Erdös and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:1761, 1960.
3. AC Gavin, M Bosche, R Krause, P Grandi, M Marzioch, A Bauer, J Schultz, JM Rick, AM Michon, CM Cruciat, M Remor, C Hofert, M Schelder, M Brajenovic, H Ruffner, A Merino, K Klein, M Hudak, D Dickson, T Rudi, V Gnau, A Bauch, S Bastuck, Huhse, C Leutwein, MA Heurtier, RR 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)(Jan 10):141–7., 2002.
4. N. Guelzim, S. Bottani, P. Bourgine, and F. Kepes. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, 31(1):60–3, 2002.
5. Y Ho, A Gruhler, A Heilbut, GD Bader, L Moore, SL 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, AR Willems, H Sassi, PA Nielsen, KJ Rasmussen, JR Andersen, LE Johansen, LH Hansen, H Jespersen, A Podtelejnikov, E Nielsen, J Crawford, V Poulsen, BD Sorensen, J Matthiesen, RC Hendrickson, F Gleeson, T Pawson, MF Moran, D Durocher, M Mann, CW Hogue, D Figeys, and M Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868)(Jan 10):180–3, 2002.

6. J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nat Genet*, 31(4):370–7, 2002.
7. 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.
8. Takashi Ito, Tomoko Chiba, Ritsuko Ozawa, Masahira Yoshida, Mikio a nd Hattori, and Yoshiyuki Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein inter actome. *PNAS*, 98(8):4569–4574, 2001.
9. H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–2, 2001.
10. H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.
11. N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 2004.
12. T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.
13. Fangting Li, Tao Long, Ying Lu, Qi Ouyang, and Chao Tang. The yeast cell-cycle network is robustly designed. *PNAS*, 101(14):11250–11255, April 2004.
14. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–42, 2004.
15. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–7, 2002.
16. R. Milo, S. S. Shen-Orr, S. Itzkowitz, N. Kashtan, D. Chklovskii, and U. Alon. On the uniform generation of random graphs with prescribed degree sequence. *ArXiv*, 2003.
17. M. E. Newman. Assortative mixing in networks. *Phys Rev Lett*, 89(20):208701, 2002.
18. M. E. Newman. Properties of highly clustered networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68(2 Pt 2):026121, 2003.
19. M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(2 Pt 2):026113, 2004.
20. M. E. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys Rev E Stat Nonlin Soft Matter Phys*, 64(2 Pt 2):026118, 2001.
21. N. D. Price, J. A. Papin, C. H. Schilling, and B. O. Palsson. Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol*, 21(4):162–9, 2003.
22. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–5, 2002.
23. S. S. 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, 2002.
24. S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–76, 2001.
25. 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.
26. C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.
27. A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, 18(7):1283–92, 2001.
28. D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2, 1998.
29. 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):5934–9, 2004.
30. S. H. Yook, H. Jeong, A. L. Barabasi, and Y. Tu. Weighted evolving networks. *Phys Rev Lett*, 86(25):5835–8, 2001.

A Computation of the Shortest Paths Length Distribution

This section is devoted to a brief description of the algorithms and methods used to derive the various statistics used in the study of the yeast regulation and protein interaction networks.

Clearly the (i, j) coefficient of M^n is the number of oriented paths of length n connecting i to j in the graph underlying M. Since we are only interested in knowing whether two nodes are connected by an oriented path of a given length we may use a simplified matrix product defined as:

$$M^n(i,j) = \begin{cases} 1 & \text{if } \exists k \in V : M^{n-1}(i,k) = M(k,j) = 1 \\ 0 & \text{otherwise} \end{cases}$$

which is just forgetting the numbers of connecting paths, only to remember whether there is at least one.

Furthermore, the addition of the identity matrix I to the adjacency matrix before the computation of the products gives an immediate access to the value of the cumulative distribution function of the oriented, shortest path length distances in the network. Indeed, writing $\widehat{M} = M + I$:

$$\widehat{M}^n(i,j) = \begin{cases} 1 & \text{if } \exists k \in V, M^{n-1}(i,k) = M(k,j) = 1 \\ & \text{or } \widehat{M}^{n-1}(i,j) = 1 \\ 0 & \text{otherwise} \end{cases}$$

Thus the number of 1s in $\widehat{M}(G)^n$ is the number of ordered pairs connected by at least one path of length $\leq n$, and the entire distribution is obtained when the computation reaches a fixpoint. Computing the distribution on the real PPI-TRI graph takes about 180' on a recent PC ; the distribution for the 100 shuffles were computed down in less than 10 hours on a cluster of 41 computers hosted by Genoscope.

B Statistics

This section details the definition and computation of p-values shown in the statistical results, concerning both the amount of connected pairs and the average distance.

In order to compute p-values for the deviation of the observable on the real graph from its disribution over the set of shuffled ones, we need to approximate this distribution by a Gaussian one, with mean and standard deviation fixed to the empirical values computed on the sample. This is necessary, since the rather low amount of shuffled networks (100) prevents a direct estimation of the p-value as the proportion of shuffled networks with a larger observable.

Concerning the amount of disconnected pairs, which is the first observable considered in the results, the empirical mean over the set of general shuffles is $m_g = 0.574$, and the standard deviation $s_g = 0.005$. In the case of the equatorial shuffle, the mean falls to $m_e = 0.534$, with a standard deviation of 0.002.

Assuming this average proportion is a Gaussian random variable A with those parameters, the p-value of the deviation of the average proportion of disconnected pairs in the real network from its distribution over the sample of general shuffled networks is defined as:

$$p_g = \mathbb{P}(A < m_G), \quad \textit{with } A \sim \mathcal{N}(m_g, s_g)$$

where $m_G = 0.538$ is the observed proportion of disconnected pairs in the real network. In this case, this yields $p_g = 9 \times 10^{-12}$.

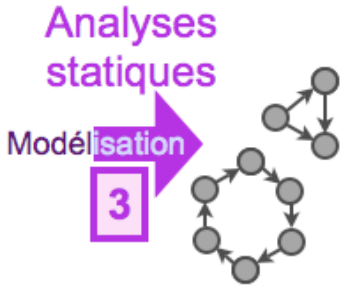
Since the proportion of disconnected pairs in the real graph is higher than the average amount of disconnected pairs in the equatorially shuffled ones, one computes the p-value p_e using the upper tail of the distribution instead of the lower one:

$$p_e = \mathbb{P}(A > m_G), \quad with\ A \sim \mathcal{N}(m_e, s_e)$$

so that $p_e = 0.03$.

The computation of the p-value for the deviation of the mean distance from its value on shuffled networks follows the same scheme. The mean distance in the real graph is $m_G^d = 5.66$, while its average over the set of shuffled graphs is $m_g^d = 5.38$ for the general shuffle, and $m_e^d = 5.5$ for the equatorial one. Standard deviation is $s_g^d = 0.09$ with the general shuffle, and $s_e^d = 0.08$ with the equatorial one. The p-values for these deviations are $p_g^d = 0.002$ and $p_e^d = 0.02$, respectively.

Chapitre V – Modèle bipartite & Topologie des RIBH



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 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 (Uetz et al., 2000; Ito et al., 2001; Guelzim et al., 2002; Lee et al., 2002; Ho et al., 2002; Gavin et al., 2002). Toutefois ces modèles ne permettent pas de comprendre comment ces différentes interactions implémentent ensemble une fonction. 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 de graphes sous- ou sur-représentés (Shen-Orr et al., 2002), uniquement par des considérations de propriétés topologiques de graphes biologiques. Toutes ces études sont basées sur un modèle de graphe trop pauvre pour permettre des analyses intégrant plusieurs types de données.

Pour cette raison, j'ai établi un modèle dérivé d'un graphe bipartite 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 rechercher des motifs hétérogènes. Le *Biological Interaction Browser* est disponible à l'adresse suivante : http://www.genoscope.cns.fr/bioinfo. Le modèle est illustré par des exemples. Nous proposons notamment des mécanismes moléculaires sous-jacents à la synexpression 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 et présentés partiellement à ISMB - ECCB 2004 du 31 Juillet au 4 Août, à Glasgow UK. Une démonstration d'une heure a été effectuée, ainsi qu'à ICSB 2004, en Octobre 9 -13 à Heidelberg GE

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).***
© 2006 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

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).***
© 2006 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

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 © 2006 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.
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,*

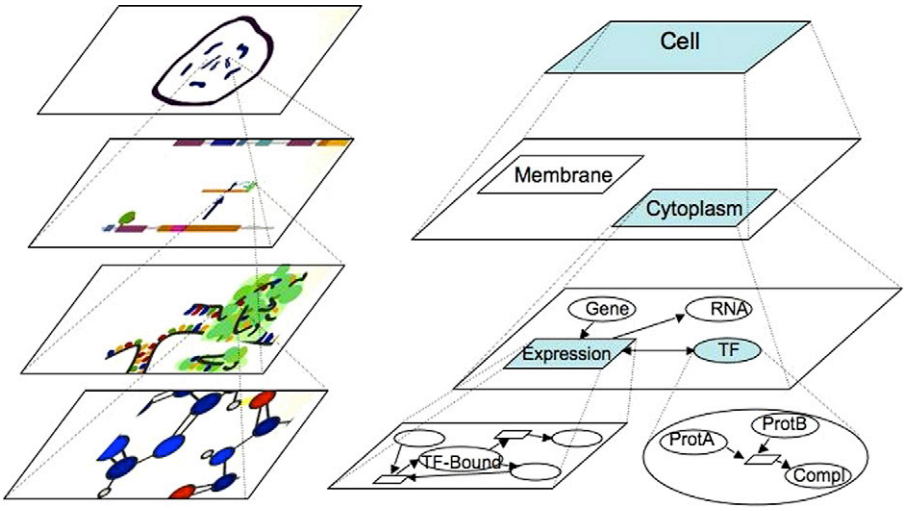


Fig. 1. Representation of biological systems seen as a set of chemical reactions. The top layer represents the most general view of the hierarchy. The bottom layer is the most detailed view of the system structure. Two intermediate layers are presented, showing the topological or functional structure 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

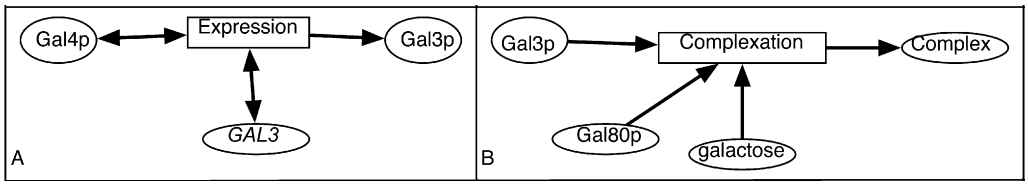


Fig. 2. Examples of representation of a biological system. **A.** In yeast, Gal4p is the transcriptional factor that regulates the GAL3 gene. **B.** Gal3p, Gal80p and galactose constitute a complex.

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).

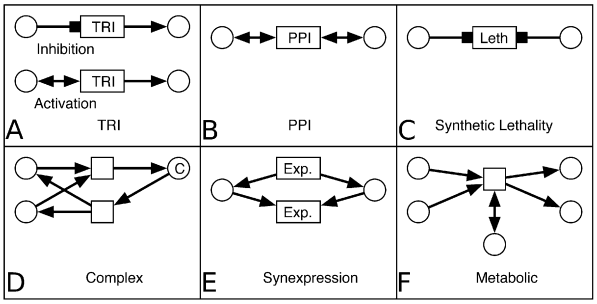


Fig. 3. Motifs used for biological data representation in MIB. **A.** Two motifs representing *TRIs*: inhibition (top) and activation (bottom) of the production of the entity (macromolecule) (right) by another entity (transcription factor) (left). **B.** A motif representing *physical interaction*: two 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.

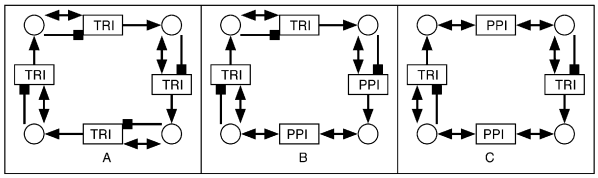


Fig. 4. Feedback-loop motifs made of TRIs only (**A**), with one PPI (**B**) or with 2 PPIs (**C**). Each motif contains four transformations (rectangular shapes), four entities (circles), and possible roles of entities in transformations are represented by arcs.

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.

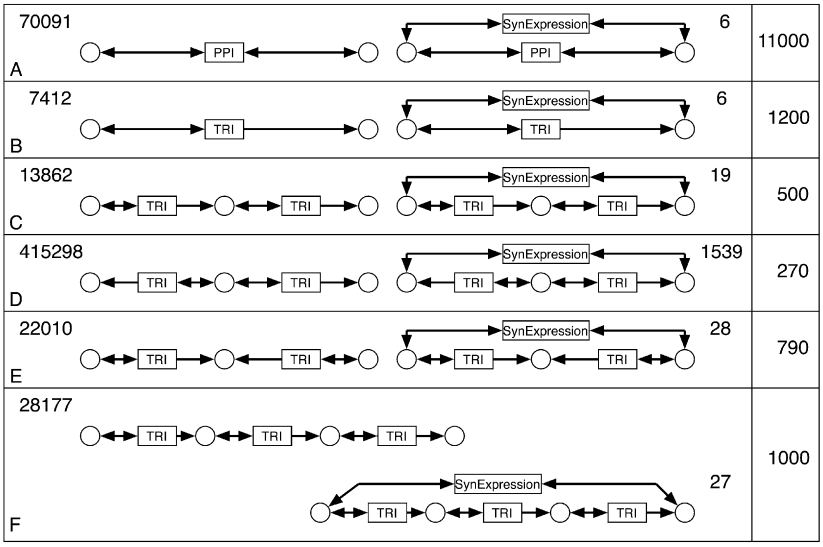


Fig. 5. Correlation between synexpression data and underlying biochemical mechanisms. Six motifs were proposed to be candidates for the synexpression mechanisms (**A**–**F**, left). For each motif, the number of occurrences in the BIB database is indicated on the side. The motifs combining the regulatory 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.

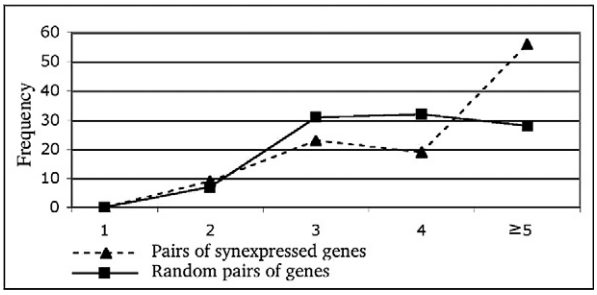


Fig. 6. Shortest path length distribution between all synexpressed pairs of proteins (dashed line) versus all possible pairs of proteins (plain line). The shortest paths of length 1 to 4 have been searched. The value of 5 on the x-axis indicates that no shorter path than five has been found.

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

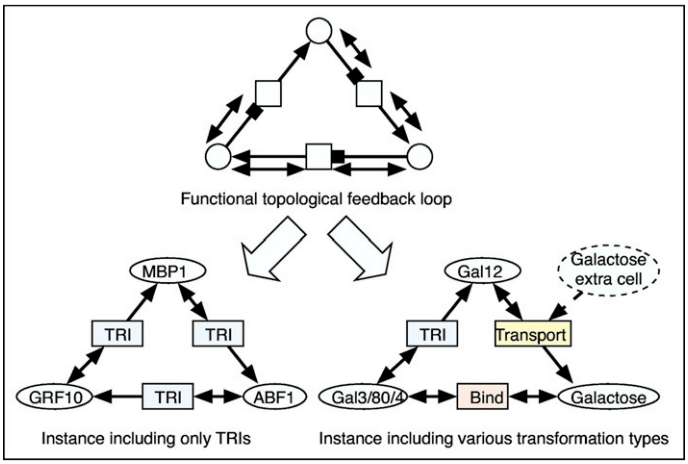


Fig. 7. A MIB motif (with a specified dynamics; top) allows searching for both TRI only subnets (left) and mixed TRI/Metabolism/PPI subnets (right). This is illustrated here with a feedback loop.

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.

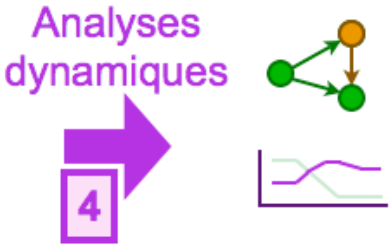
[31] S. Shen-Orr, R. Milo, S. Mangan, U. Alon, Network motifs in the transcriptional regulation network of *Escherichia coli*, Nat. Genet. 31 (1) (2002) 64–68.

[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.

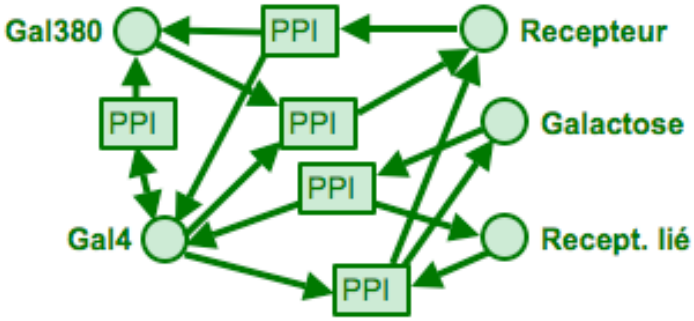
[34] U. de Lichtenberg, L. Jensen, S. Brunak, P. Bork, Dynamic complex formation during the yeast cell cycle, Science 307 (5710) (2005) 724–727.

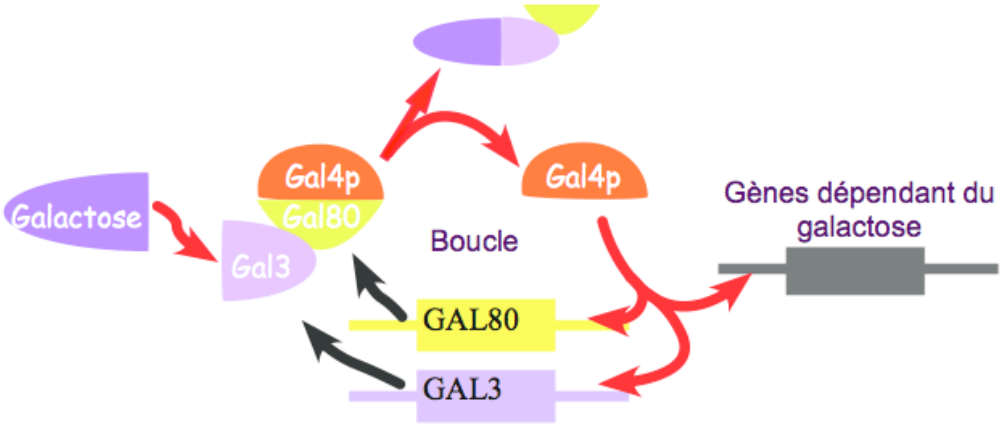
[35] http://nemo-cyclone.sourceforge.net.



Chapitre VI – Dynamique de RIBH : exemple du galactose

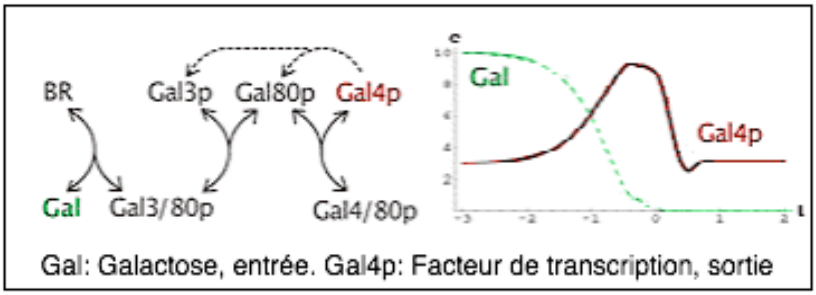
Un des modules hétérogènes découvert au cours de notre travail présenté ci-dessus d'étude des RIBH et qui est intéressant pour ses propriétés dynamiques, est la boucle de régulation galactose. La dynamique de petits réseaux peut être modélisée au moyen d'équations différentielles.





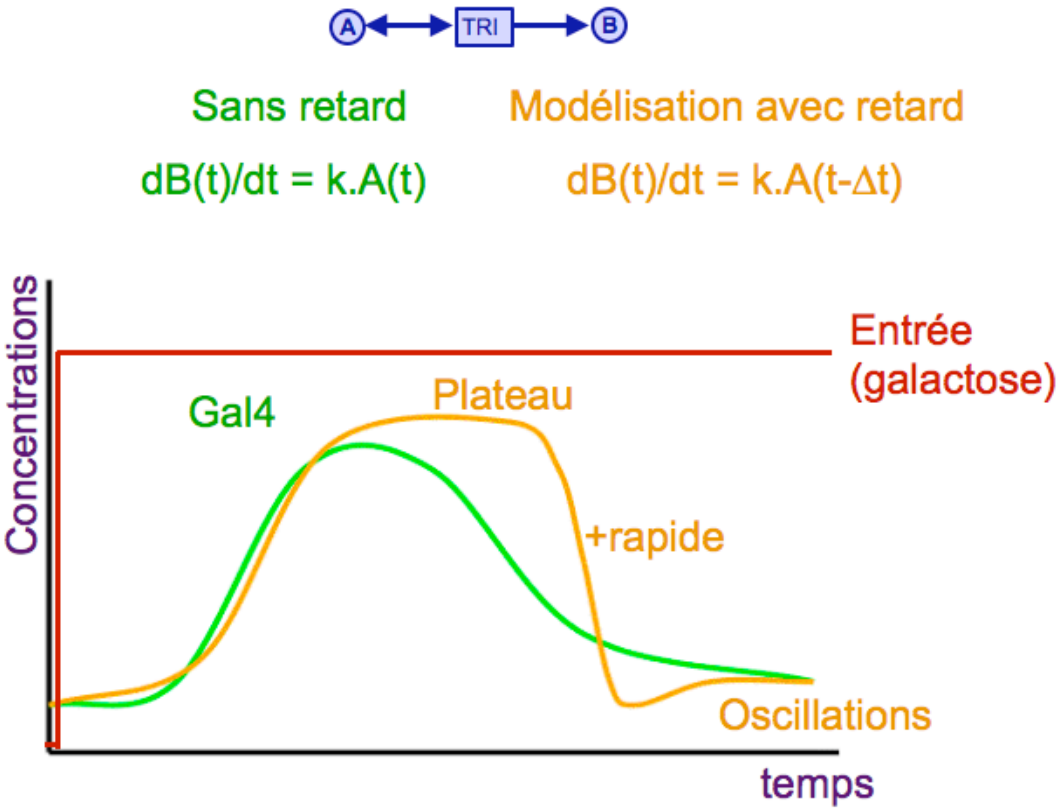
Boucle hétérogène trouvée au chapitre V, dont on peut étudier la dynamique. Au dessus, le module est représenté dans le formalisme MIB. En dessous, un shéma qui illustre le même système.

Les interactions entre le galactose, Gal3p, Gal80p et Gal4p déterminent l'état transcriptionnel des gènes requis pour l'assimilation du galactose (voir figure ci-dessous). Après une augmentation de la concentration de galactose, les molécules du sucre se lient à Gal3p. Gal3p active Gal4p via Gal80p. Gal4p activé induit la transcription des gènes GAL3 et GAL80. La boucle de rétroaction est fermée par la capture de Gal4p par les protéines Gal3p et Gal80p nouvellement synthétisées, ceci conduit à la baisse de l'activité transcriptionnelle de Gal4p. Si l'on considère le galactose comme le signal d'entrée, l'activité de Gal4p comme la sortie, le système se comporte comme un *filtre dérivateur* en terme de traitement du signal.



Résumé du fonctionnement de la boucle de régulation du galactose. À gauche sont représentées les différentes interactions de la boucle. À droite, les concentration de l'entrée et de la sortie du système sont représentées.

L'avantage pour la cellule d'un tel système est important. Les enzymes de dégradation du galactose ne sont produites qu'en proportion avec le galactose détecté. Cela permet aussi à la cellule d'être aussi sensible à de faibles qu'à de fortes concentrations de Galactose.



L'introduction d'un retard pour modéliser les interactions (régulation transcriptionnelle), permet à la cellule d'avoir une réponse plus rapide (ce qui est contre intuitif). En vert, la réponse sans retard modélisée, en jaune, la réponse de Gal4p en ayant introduit un retard. Ce résultat illustre l'intérêt de modéliser les différntes interactions différemment.

Les résultats suivant ont été publiés dans le *Journal of Theoretical Biology* en 2005.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 242 (2006) 372–381
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

The adaptive filter of the yeast galactose pathway

Serge Smidtas[a], Vincent Schächter[a], François Képès[b,*]

[a]Genoscope-Centre National de Séquençage, CNRS UMR8030, 2 rue Gaston Crémieux, 91000, Evry, France
[b]Epigenomics Project, Genopole®, 523 Terrasses de l'Agora, 91000 Evry, France & ATGC, CNRS UMR8071/Genopole®, Evry, France

Received 12 September 2005; received in revised form 20 February 2006; accepted 10 March 2006
Available online 27 April 2006

Abstract

In the yeast *Saccharomyces cerevisiae*, the interplay between galactose, Gal3p, Gal80p and Gal4p determines the transcriptional status of the genes required for galactose utilization. After an increase in galactose concentration, galactose molecules bind onto Gal3p. This event leads via Gal80p to the activation of Gal4p, which then induces *GAL3* and *GAL80* gene transcription. Here we propose a qualitative dynamical model, whereby these molecular interaction events represent the first two stages of a functional feedback loop that closes with the capture of activated Gal4p by newly synthesized Gal3p and Gal80p, decreasing transcriptional activation and creating again the protein complex that can bind incoming galactose molecules. Based on the differential time-scales of faster protein interactions versus slower biosynthetic steps, this feedback loop functions as a derivative filter where galactose is the input step signal, and released Gal4p is the output derivative signal. One advantage of such a derivative filter is that *GAL* genes are expressed in proportion to cellular requirements. Furthermore, this filter adaptively protects the cellular receptors from saturation by galactose, allowing cells to remain sensitive to variations in galactose concentrations rather than to absolute concentrations. Finally, this feedback loop, by allowing phosphorylation of some active Gal4p, may be essential to initiate the subsequent long-term response.
© 2006 Elsevier Ltd. All rights reserved.

Keywords: Galactose switch; Yeast; Adaptive filter; Feedback loop; Qualitative modeling; Interaction networks

1. Background

Living organisms constantly adapt to fluctuations in their intra- and extra-cellular environments, in part by regulating the expression of their genes. Gene expression can be controlled at many levels that involve protein–DNA (transcriptional), protein–protein and protein–small molecule interactions. The process of galactose (GAL) utilization in the common yeast *Saccharomyces cerevisiae* has been thoroughly studied; yeast is known to exhibit sophisticated responses to the presence of different types of sugar in its environment. The GAL pathway is a classical example of a genetic regulatory switch, in which enzymes specifically required for the transport and catabolism of galactose are expressed only when galactose is present and repressing sugars such as glucose are absent in the cellular environment (Biggar and Crabtree, 2001).

The permease encoded by the *GAL2* gene, and possibly other hexose transporters (HXTs) transport galactose across the cell membrane. Other genes encode the enzymes required for conversion of intracellular galactose, including galactokinase (*GAL1*), uridyltransferase (*GAL7*), epimerase (*GAL10*), and phosphoglucomutase (*GAL5*/*PGM2*). Galactose activates the transcription of *GAL* genes from undetectable or low basal levels to high levels. The activated genes include *GAL1*, *GAL2*, *GAL3*, *GAL5*, *GAL7* and *GAL80* (Sakurai et al., 1994), but not *GAL4* (Ren et al., 2000; Ideker et al., 2001). The complex interplay of Gal4p, Gal80p, and Gal3p determines the transcriptional status of these *GAL* genes (Platt and Reece, 1998). Gal4p is a DNA-binding transcriptional activator that can bind to upstream activating sequences in the promoter regions of target *GAL* genes, thereby strongly activating their transcription. However, in the absence of galactose, Gal4p is sequestered by Gal80p and is unable to

*Corresponding author. Tel.: +33 1 60 87 40 94.
E-mail addresses: sergi@sergi5.com (S. Smidtas), vs@genoscope.cns.fr (V. Schächter), francois.kepes@genopole.cnrs.fr (F. Képès).

0022-5193/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2006.03.005

activate transcription of the *GAL* genes, although this Gal4p/80p complex appears to bind DNA (Parthun and Jaehning, 1992). The interaction between Gal4p and Gal80p is weaker in the presence of galactose (Sil et al., 1999). Gal80p and Gal3p may also form a complex, which in contrast is stabilized in the presence of galactose (Yano and Fukasawa, 1997). Gal3p overproduction, presumably by sequestering Gal80p away from Gal4p, causes galactose-independent activation of the GAL pathway (Bhat and Hopper, 1992; Peng and Hopper, 2000).

Gal3 mutant cells are still able to activate the GAL pathway in response to galactose. However, induction requires several days rather than a few minutes in wild-type yeast, a phenomenon called long-term adaptation (LTA) (Winge and Roberts, 1948; Bhat and Murthy, 2001). It was proposed (Rohde et al., 2000) that the LTA of the GAL pathway is mediated by Gal4p phosphorylation. Indeed, when Gal4p is bound to DNA and interacts with the RNA-polymerase II holoenzyme, its serine at position 699 (S699) becomes phosphorylated by Srb10p/Cdk8p, a component of the 'Mediator' subcomplex of the holoenzyme (Hirst et al., 1999; Bhaumik and Green, 2001; Larschan and Winston, 2001). Gal4p S699 phosphorylation is necessary to amplify and maintain full *GAL* gene induction (Sadowski et al., 1996; Yano and Fukasawa, 1997; Rohde et al., 2000).

The above set of experimental observations raises two main questions. Firstly, the system responds to galactose increases rather than to absolute galactose concentration: how is this achieved (Rohde et al., 2000)? Secondly, several authors have observed that Gal4p does not become phosphorylated unless it activates transcription, yet that it is not fully active unless it is phosphorylated (Sadowski et al., 1991; Sadowski et al., 1996; Hirst et al., 1999). A satisfactory explanation for this 'chicken and egg' enigma is lacking. In this paper we propose a mathematical model of the early response to galactose and we analyse its dependence upon time delays, protein degradation rate and initial conditions. The model accounts for the above-mentioned sensitivity to galactose fluctuations. It also proposes a solution to the apparent paradox described above by showing that a feedback loop brings active Gal4p onto gene promoters, thus allowing its phosphorylation and consequent maintenance of transcriptional activation.

2. Qualitative modeling of the galactose response

2.1. Assumptions

The present model deals with the early steps of galactose induction; it does not consider the events occurring after Gal4p phosphorylation. It does not emphasize the details of signal transmission from galactose to Gal4p (except in Appendix A). Thus, Gal4p appears in this model either bound to DNA, or bound to DNA and to Gal80p. Gal80p is either bound to DNA and to Gal4p, or bound to Gal3p, or unbound. An equilibrium between nuclear and cytoplasmic forms of Gal80p has been considered by other authors (Peng and Hopper, 2000, 2002; Verma et al., 2003), but is not relevant here given the scope of our model. The order in which galactose, ATP, Gal3p and Gal80p bind together is not fully known but should have no effect on the conclusions reached with our model, which simply considers Gal80p consumption upon galactose addition.

The *GAL1* gene is a paralogue of the *GAL3* gene (Wolfe and Shields, 1997) that encodes a galactokinase, while Gal3p does not have galactokinase activity (Platt et al., 2000). Galactokinase activity is irrelevant to the present model which does not address galactose catabolism. Therefore, Gal1p and Gal3p are taken to play a similar role in GAL pathway activation, averaged over their respective abundances and inducing properties. In the model, they will be lumped together under the name of Gal3p.

2.2. Dynamical description of the GAL system

Fig. 1 shows the different states of the system. Fig. 2A illustrates the core regulatory mechanism. In the absence of galactose, Gal4p can bind to Gal80p and has no transcriptional activity. Following a step increase in galactose, Gal3p rapidly binds galactose, and Gal80p is consumed by being recruited in a complex with Gal3p. As the concentration of unbound Gal80p decreases, the

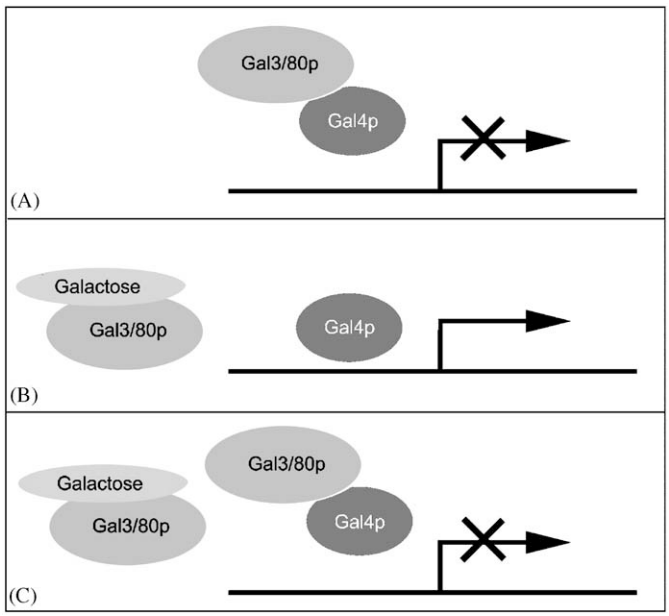


Fig. 1. Diagrammatic representation of the galactose induction loop. (A) In the absence of galactose, the transcriptional activity of Gal4p is inhibited by Gal3/80p. (B) The association of galactose with Gal3/80p allows Gal4p to be freed from Gal80p inhibition and to activate transcription of new Gal3/80p. (C) Newly synthesized Gal3/80p inhibits the transcriptional activity of Gal4p.

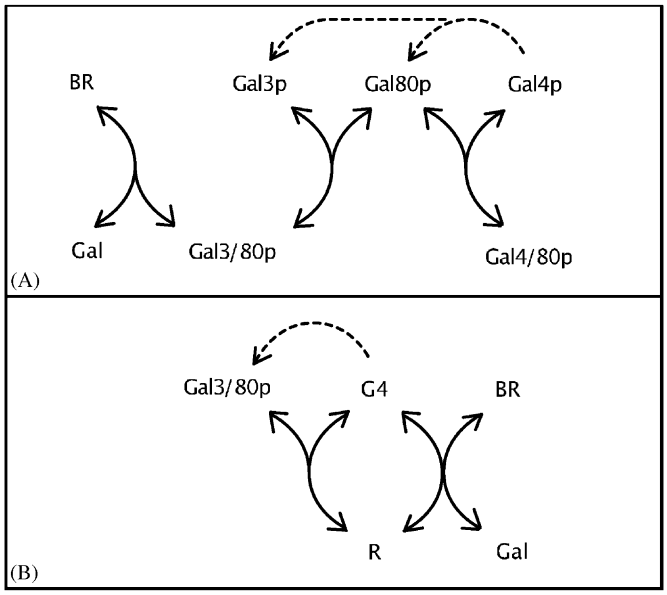


Fig. 2. GAL core regulatory pathway. (A) Detailed model. In the presence of galactose, Gal3p, Gal80p and galactose ('Gal') bind together, thus decreasing the binding of Gal80p to Gal4p. Without Gal80p, Gal4p becomes active and induces transcription (dashed arrows) of the *GAL3* and *GAL80* genes. This closes the feedback loop, as newly synthesized Gal3p and Gal80p shift the equilibrium back towards Gal4p inactivation (Gal4p/80p). (B) Simplified model. Receptor ('R') denotes Gal4/3/80p; Bound Receptor ('BR') denotes the Gal3/80p/galactose complex. In the absence of galactose, Gal3/80p sequesters Gal4p ('G4') into the Receptor, thus preventing its transcriptional activity. In the presence of galactose, Gal4p is released and induces transcription of the *GAL3* and *GAL80* genes. Newly synthesized Gal3/80p shift the equilibrium back towards Gal4p inactivation.

Gal4p/80p complex is destabilized, which activates Gal4p. Activated Gal4p then initiates the slower biosynthetic reactions, transcription of the *GAL* genes including *GAL3* and *GAL80*, followed by translation into their protein products.

Following Gal4p activation, and consequent *GAL* gene expression, newly synthesized Gal3p and Gal80p shift the equilibrium back towards Gal4p inactivation. As a result, *GAL* transcriptional activity decreases back. Newly formed proteins can bind incoming galactose molecules, thus restoring sensitivity to any further galactose input. This effectively closes the feedback loop, the central point of this model.

2.3. Model simplification

The detailed model shown in Fig. 2A is needlessly complex with respect to the focus of our study: the role of the feedback loop. In this section, we show how the model could be simplified, yielding a reduced model (Fig. 2B) that features the feedback-loop and preserves the qualitative dynamics of the detailed model, while allowing deeper analysis and understanding.

The *GAL3* and *GAL80* genes are both transcriptionally regulated by Gal4p. However, the *GAL3* gene is activated about five-fold stronger than *GAL80* (Peng and Hopper, 2002). This fact may amplify or accelerate the response. Indeed, a relative increase of Gal3p with respect to Gal80p will shift complex formation towards additional Gal80p consumption, further increasing the concentration of activated Gal4p. Even though this may bring about changes in the exact response kinetics, it does not change the qualitative behavior of the system. Furthermore, concomitant overexpression of *GAL3* and *GAL80* was shown to suppress the constitutive *GAL* gene expression elicited by overexpression of *GAL3* alone (Suzuki-Fujimoto et al., 1996), suggesting that their two products play a complementary role in the reaction cascade. Accordingly, Gal3p and Gal80p are lumped together in this model as the 'Gal3/80p' complex. This simplification is relaxed in a more complex model described in Appendix A. This latter model, while useful for simulation purposes, is much harder to understand analytically, as it involves six equations and eight parameters. We show in Appendix A that the qualitative behavior is preserved by the simplification.

In the absence of galactose, Gal3p, Gal80p and Gal4p form an inactive complex Gal4/3/80p called 'receptor' ('R'; Fig. 2B). This simplifies the model, lumping a cascade of reactions into one. A 'bound receptor' ('BR') comprising Gal3p, Gal80p and galactose remains inactive or may be degraded. Finally, the total Gal4p concentration is assumed to be constant during the GAL response, as suggested by transcriptomics data (Ren et al., 2000; Ideker et al., 2001).

The three main transformations of the simplified model corresponding to the three reactions of Fig. 2B are shown below (Eqs. (1)–(3)). We represent a gene, its encoded mRNA and protein as a single entity. 'G4' denotes an active Gal4p protein. The first equation pertains to the slow biosynthetic steps of transcription and translation, comprising the binding of Gal4p to the *GAL3* and *GAL80* gene promoters, and all subsequent actions until Gal3/80p molecules are newly synthesized one at a time, and Gal4p stays activated. The second equation represents the inactivation of Gal4p into its inactive form called the receptor *R*. The third equation expresses the activation of Gal4p due to the binding of galactose ('Gal') to the receptor, yielding the bound receptor ('BR').

$$\mathrm{G4} \rightarrow \mathrm{G4} + \mathrm{Gal3/80}, \tag{1}$$

$$\mathrm{G4} + \mathrm{Gal3/80} \leftrightarrow R, \tag{2}$$

$$R + \mathrm{Gal} \leftrightarrow \mathrm{G4} + \mathrm{BR}. \tag{3}$$

To facilitate analyses and simulations, this model can be further reduced to a two equation, two variable system readily amenable to phase plane analysis. It emphasizes the role played by the negative feedback loop in the *GAL* pathway dynamics. Combining transformations (1) and (2) into (2a), gives the following set of two equations, with

three kinetic constants K_1, K_2, K_3.

$$R + \mathrm{Gal} \underset{K3}{\overset{K2}{\rightleftarrows}} \mathrm{G4} + \mathrm{BR}, \tag{1a}$$

$$\mathrm{G4} \xrightarrow{K1} R. \tag{2a}$$

S_0 (for "Sugar") is the initial quantity of galactose, and R_0 (for "Receptor") the total amount of Gal4p. These values are constant, as galactose is initially provided in a finite amount, and total Gal4p has been assumed to be a constant amount (Johnston et al., 1994; Ren et al., 2000; Ideker et al., 2001).

$$\mathrm{Gal} + \mathrm{BR} = S_0, \tag{1b}$$

$$R + \mathrm{G4} = R_0. \tag{2b}$$

This leads to two nonlinear differential equations, with kinetic constant K_1 normalized to 1. Assuming a homogeneous spatial distribution and an average cell volume of 70 μm^3 (Ruhela et al., 2004), the concentration of galactose and proteins involved is of the order of 10^{-6} M. This leads to S_0 and R_0 of the order of 10^{-6} M and concentration ratios $K_1 = K_2 = K_3 = 1$. The typical time response of the galactose switch is 1–10 h, implying that the order of magnitude of a time unit in this context is 5×10^2 s.

$$\frac{\mathrm{d}(R)}{\mathrm{d}t} = (K_3 - K_2)\,\mathrm{Gal}\,R - K_3\,R_0\,\mathrm{Gal} - (K_3\,S_0 + K_1)\,R + R_0\,(K_3\,S_0 + 1), \tag{1c}$$

$$\frac{\mathrm{d}(\mathrm{Gal})}{\mathrm{d}t} = (K_3 - K_2)\,\mathrm{Gal}\,R + K_3\,S_0\,R_0 - K_3\,S_0\,R - K_3\,R_0\,\mathrm{Gal}. \tag{2c}$$

In the special case where K_2 is set equal to K_3, these equations can be simplified into linear equations:

$$\frac{\mathrm{d}(R)}{\mathrm{d}t} = (K_2\,S_0 + 1)R_0 - (K_2\,S_0 + 1)R - K_2\,R_0\,\mathrm{Gal}, \tag{1d}$$

$$\frac{\mathrm{d}(\mathrm{Gal})}{\mathrm{d}t} = K_2\,S_0\,R_0 - K_2\,S_0\,R - K_2\,R_0\,\mathrm{Gal}. \tag{2d}$$

3. Results

3.1. Steady states and phase plane portrait

The phase plane for Eqs. (1c) and (2c) is the Cartesian coordinate system representing the two variables, receptor and free galactose concentrations (Fig. 3). Two nullclines are represented, that correspond to the situations where: (1) the receptor synthesis and consumption by galactose fixation are balanced, or (2) free galactose addition and galactose fixation are balanced. Thus, the nullclines correspond to steady states, where the derivatives of Eqs. (1c) and (2c) are null (Fig. 3). The two nullclines cross each other at a stable point which corresponds to the situation where (1) and (2) simultaneously hold true. This stable point turns out to correspond to low free galactose and high receptor concentrations (Fig. 3). When K_1 is normalized to 1, and K_2 and K_3 are equal (Eqs. (1d) and (2d)), the nullclines are linear, thus facilitating further analysis (Fig. 3A and B). The two nullclines (plain lines) delimit three domains in the plane. From whichever of these three domains the starting point is in, trajectories (Fig. 3B, broken line) return to the unique steady state. This provides robustness with respect to fluctuations in receptor concentration.

3.2. Influence of parameters values

To assess the influence of the variations of parameters K_1, K_2 and K_3, we studied how the phase plane was modified by varying each of the parameters separately (Smolen et al., 2001; Morohashi et al., 2002; Sriram and Gopinathan, 2004). Three different sets of values for K_1, K_2 and K_3 are explored in Fig. 3, giving different pairs of nullclines. K_2 range was explored over two decades (Fig. 3C) and only affects the concavity of the nullclines without modifying the qualitative dynamics. Then, keeping the ratio of K_2 and K_3 constant, we explored two decades of K_3 variations and this modifies the area between the two nullclines (Fig. 3D). This does not change the qualitative dynamics of the system either. K_1 variations were also studied. For K_1 greater than 1, the two nullclines cross each other at the stable point that corresponds to an equilibrium between the reactants (Fig. 3E); this does not otherwise alter the qualitative dynamics. This qualitative behavior resembles the one of the model with degradation (see below and Fig. 5). For K_2 lesser than 1 (Fig. 3F), the two nullclines cross each other outside of the reachable concentration area but still near the bottom right corner of the phase plane. As in all cases, the stable point is kept in the same area, the trajectories are not qualitatively different and the corresponding behavior appears to be robust.

3.3. Functional interpretation as a 'derivative filter'

The 'chicken and egg' paradox can be readily explained by this feedback loop model. In cells expressing a mutated Gal4p A699S (where an alanine replaces the serine 699) that cannot be phosphorylated, *GAL* gene activation can initiate but cannot be maintained (Yano and Fukasawa, 1997). Fig. 4A shows that in response to the addition of galactose (dashed line), free Gal4p (bold line) increases and initiates transcription of the *GAL3/80* genes. As a consequence, Gal4p binds to new Gal3/80p to produce receptors (plain line) and its free concentration decreases, hence Gal4p cannot maintain transcriptional activity. During the early response phase modelled here, this system acts as a derivative filter (Lauffenburger, 2000; Basu et al., 2004), where galactose is the input step signal and Gal4p is the output peak signal, an approximate derivative of the step input. Furthermore, the model is consistent with the

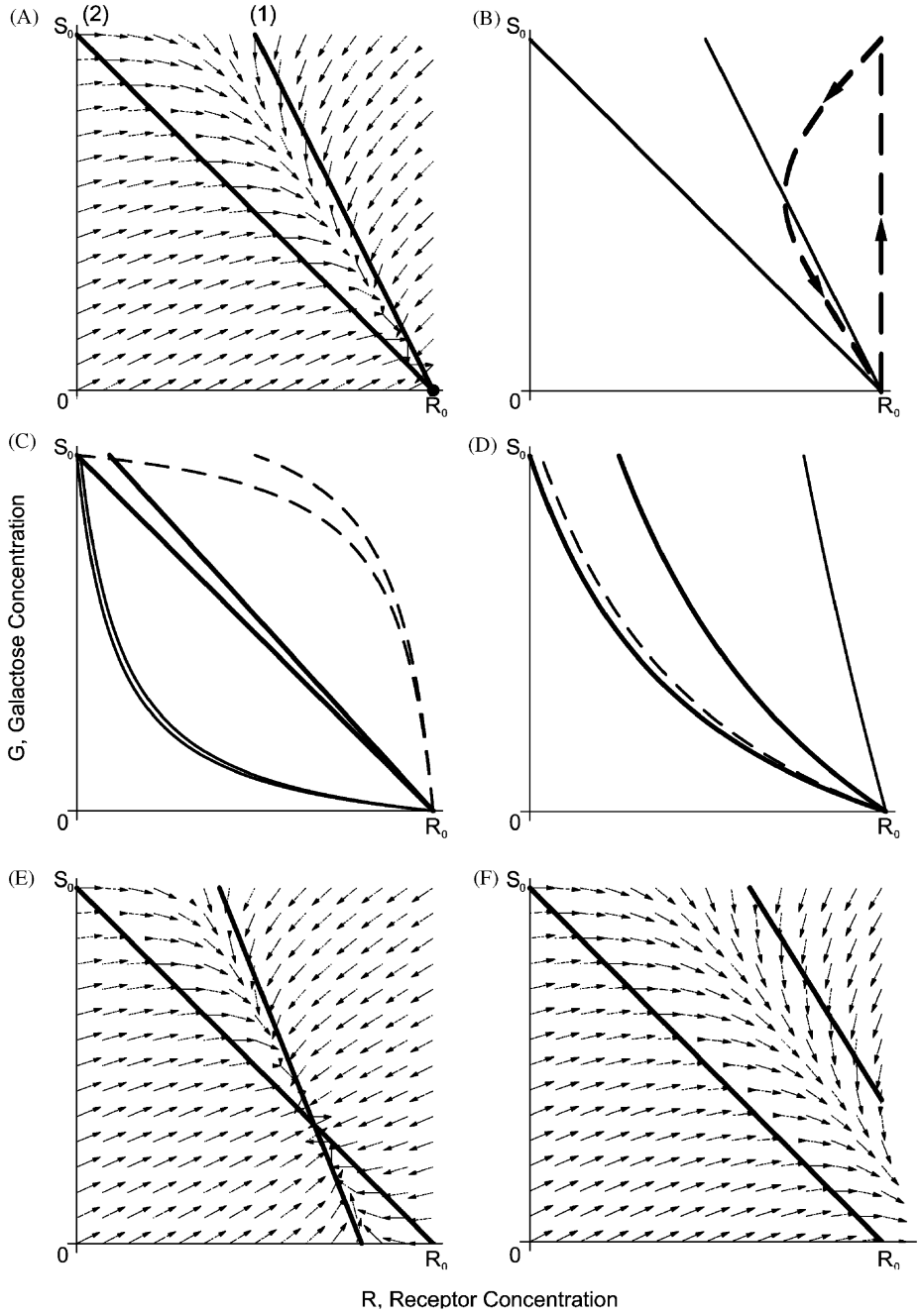


Fig. 3. Phase portrait and parameter variation study of the two-equation system. The curves labeled 1 or 2 are nullclines that correspond to the situations where: (1) receptor synthesis and consumption by galactose fixation are balanced, or (2) free galactose addition and galactose fixation are balanced. There is a single stable point where the two curves cross each other. (A,B) Phase portrait for Eqs. (1d) and (2d). Simplified phase portrait and trajectory for Eqs. (1d) and (2d). The nullclines (plain lines) are represented for $K_1 = 1$ and $K_2 = K_3$. Arrow pairs represent the slopes in each of the three domains defined by the two nullclines. (B) A typical trajectory is represented by a broken line. In the absence of galactose, the system is at the stable point in the lower right corner. When galactose is introduced, all galactose is in its free form and the state moves to the upper right corner. Then, galactose binds to the receptor and causes free galactose and receptor concentrations to decrease. This binding causes Gal4p to activate transcription, thereby producing new receptor whose concentration increases back to the stable point. (C–F) Phase portrait for Eqs. (1c) and (2c). These curve pairs are represented for various values of K_1, K_2 and K_3. (C) $K_1 = 1$, $K_2 = 10$, $K_3 = 10$ (bold line), $K_1 = 1$, $K_2 = 100$, $K_3 = 10$ (plain line), $K_1 = 1$, $K_2 = 1$, $K_3 = 10$ (dashed line). (D) $K_1 = 1$, $K_2 = 3$, $K_3 = 1$ (bold line), $K_1 = 1$, $K_2 = 30$, $K_3 = 10$ (plain line), $K_1 = 1$, $K_2 = 0.3$, $K_3 = 0.1$ (dashed line). (E) $K_1 = 1.5$, $K_2 = 1$, $K_3 = 1$. (F) $K_1 = 0.6$, $K_2 = 1$, $K_3 = 1$.

observation that transcription is permanently activated in *gal80Δ* mutant cells (Ideker et al., 2001). Indeed, within the present model, the absence of Gal80p breaks the loop; Gal80p can no longer sequester Gal4p which remains permanently active.

3.4. Influence of time delays

As transcriptional interactions generally occur at longer time-scales than protein–protein interactions, we experimented with the introduction of time delays (de Jong,

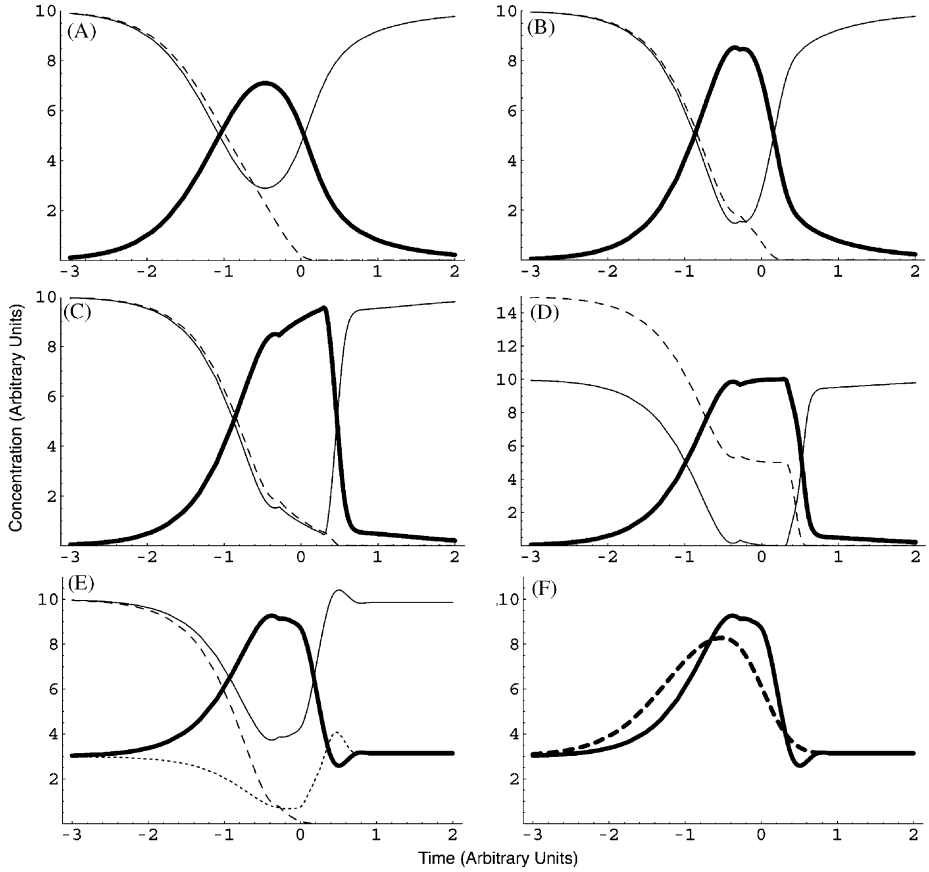


Fig. 4. Simulations of the GAL system time response depicted in Eqs. (1)–(3). The initial state consists of ten units of receptor (plain line), no free Gal3/80p away from the receptor, and no free Gal4p/DNA (bold line) unless otherwise stated. Ten units of galactose (dashed line) are initially introduced, unless otherwise stated. Note the log scale on the time axis. The Mathematica function NDelayDSolve written by Alan Hayes was used for computation. Curves A–E show computed time responses for different sets of parameters, as follows: (A) No delay in transformations. (B) Short transcriptional delay of 0.5 arbitrary units of time (a.u.). (C) Long transcriptional delay of 2 a.u. (D) Long transcriptional delay. Fifteen units of galactose are initially introduced, and the concentration scale has been changed to reflect this change. (E) Long transcriptional delay, and Gal3/80p (dotted line) degradation (half-life of 1 a.u.). Although no additional free Gal4p has been introduced, Gal3/80p decay results in excess Gal4p that is not sequestered in the receptor, hence the nonzero value of initial and final Gal4p. (F) Long transcriptional delay (bold line), compared to the absence of delay (dashed line), and Gal3/80p degradation as in (E). Only Gal4p is represented here (same remark as in E).

2002) in the differential equations. We introduced a delay in the transcription reaction that led to Eqs. (1e) and (2e).

$$\frac{\mathrm{d}(R)}{\mathrm{d}t} = (K_2\,S_0 + 1)R_0 - K_2\,S_0\,R - R(t-\tau) - K_2\,R_0\,\mathrm{Gal}, \tag{1e}$$

$$\frac{\mathrm{d}(\mathrm{Gal})}{\mathrm{d}t} = K_2\,S_0\,R_0 - K_2\,S_0\,R - K_2\,S_0\,\mathrm{Gal}. \tag{2e}$$

In the simulations shown in Fig. 4A–C, ten units of galactose are initially introduced at time $t = 0$ in the presence of ten units of receptor. These three simulations correspond respectively to no, short, or long biosynthetic delay. As protein concentrations shift away from equilibrium during the biosynthetic steps, potential chemical energy is stored in the form of Gal3/80p undergoing synthesis. In the absence of delay, this energy is immediately released by the consumption of Gal4p and of the remaining galactose. As the delay increases, more Gal3/80p are being synthesized simultaneously, and therefore more potential energy will suddenly be released at the end of the delay. Thus, the differential time-scales of fast protein interactions and slower biosynthetic processes are responsible for the observed feedback acceleration and output signal sharpening.

3.5. Adaptation of receptor concentration

To investigate how the system adapts to various galactose inputs, galactose was initially introduced in larger quantity units (15 units) than the receptor (10 units) (Fig. 4D). Compared with an initial dose of ten units of galactose in an otherwise identical experiment (Fig. 4C), a three-step response is observed in Fig. 4D. Firstly, the receptor is entirely consumed and excess galactose remains

as a plateau. Secondly, newly formed receptors rapidly bind the remaining galactose. Thirdly, the initial receptor and Gal4p concentrations are restored. Thus, the receptor concentration is maintained at a constant level, independent of the initial galactose concentration. The *GAL* genes are expressed in proportion to the cellular requirement, at each increase of galactose concentration.

3.6. Effect of protein degradation

To investigate the effect of protein degradation, Gal3/80p half-life was set to 1 arbitrary time unit in the presence of a long transcriptional delay of 2 units. It appears that Gal3/80p degradation elicits dampened oscillations of the receptor, Gal4p and Gal3/80p concentrations (Fig. 4E). These oscillations depend on the presence of a transcriptional delay, as shown in Fig. 4F, where Gal4p concentration is monitored in the absence (dashed bold line) or in the presence (bold line) of a long transcriptional delay. As was already shown in panels 3A and 3C, introducing a delay sharpens the time evolution of Gal4p. However, Gal3/80p degradation makes the signal more realistic in that, for instance, Gal4p becomes in excess to Gal3/80p and never reaches zero values.

When Gal3p and Gal80p are distinguished according to the more complete model shown in Fig. 2A, while including protein degradation, the outcome is similar (Appendix A).

4. Discussion

Earlier attempts to model the galactose pathway did not emphasize the feedback loop involving Gal4p activation and induction of Gal3p and Gal80p synthesis. Here we propose that the crucial mechanism generating the ill-understood behavior of the system prior to Gal4p phosphorylation is this feedback loop. A similar approach was recently developed independently, that emphasizes the importance of autoregulation (Ruhela et al., 2004), but with no discussion of the underlying mechanism.

To emphasize the role that feedback may play in the GAL pathway, we simplified our model of the molecular machinery by lumping together Gal3p and Gal80p to the point where its behavior could be described by two differential equations. This simplification has been justified but a posteriori it can be also verified. We have shown that the simple model and simulations of the more detailed one have the same qualitative behaviors. In both case, the system is adaptive and acts as a derivative filter (Figs. 3–6). Degradation modifies moderately the dynamic preventing concentrations to go back down to zero but led to an equilibrium (Figs. 4E-F and 5). Both models are robust to parameter variations (Figs. 3 and 6). The simple two-equation model has several advantages.

Firstly, it captures the key qualitative features of the system dynamics. For instance, phase plane analysis provides useful insight into the poorly understood mechanism of adaptation and high sensitivity to galactose fluctuations. A consequence of the existence of a feedback loop is that receptor concentration is maintained at a constant level, independent of the amount of galactose captured by the cell. Hence, the receptor cannot be saturated by galactose and cells remain sensitive to galactose variations, not to the absolute galactose concentration, without requiring a high number of receptors. Another advantage is that Gal4p-regulated genes are expressed in proportion to cellular requirements, i.e. the galactose flux or the time derivative of the galactose concentration.

Secondly, predictions obtained with this model are consistent with previous biological investigations, which were mostly based on the study of one branch of the loop, signal transduction from galactose to Gal4p (Ideker et al., 2001).

Thirdly, the model is rich enough to allow predictions that could be tested at the bench. For instance, consider a mutant strain expressing a Gal4p A699S version to suppress the onlocking of the GAL activation, and Gal2p under the control of a constitutive promoter to avoid permease-dependence of the response to galactose. In these mutant cells, each successive galactose step increase should result in an increase of a Gal1p-GFP chimeric reporter. Furthermore, if the reporter protein is destabilized, its increase should be followed by a decrease that would constitute the signature of the feedback loop.

The question arises whether the feedback network studied in this paper has some generality beyond the galactose case. Indeed, topological studies (Shen-Orr et al., 2002) on mixed networks that include protein–protein and transcriptional interactions (Yeger-Lotem and Margalit, 2003; Yeger-Lotem et al., 2004) point to the existence of other known cases. A well-known example is the Dig1-Ste12p feedback loop (Bardwell et al., 1998), and longer loops have also been detected, such as Met28-Met4-Cbf1, Mot2-Set1-Ccr4, Ngg1-Ada2-Rtg3-A1-Spt7 or a 13-size feedback loop Sin3-Adh2-Ccr4-Pri2-Swi6-Cdc6-Sfp1-Tec1-Ste12-Tup1-Sps1-Ime1-Rgr1 (Smidtas et al., 2006). Since interactions may be missing in some of these loops, the interpretation of their roles often requires prior knowledge.

In the future, the model could be refined by considering the relationship between the short-term adaptation described here and the long-term response achieved through Gal4p phosphorylation, or by immersing this system into the wider web of other sugar regulatory pathways and of other types of interaction.

Acknowledgements

We are really grateful to Anastasia Yartseva for helpful comments and discussion and K. Sriram and S. Bottani for critically reading this paper. This work was supported by funding from CNRS and Genopole®.

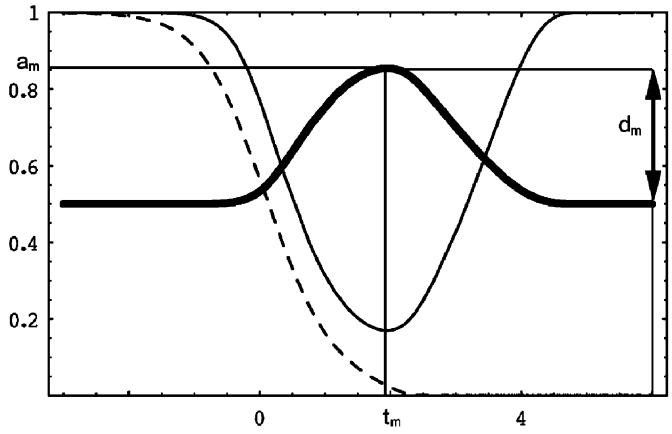


Fig. 5. Time response of the GAL pathway depicted in Fig. 2A. The initial state consists of 0.1 unit of Gal4p/DNA (bold line), and 1 unit of Gal80p (plain line). Two units of galactose (dashed line) are initially introduced, unless otherwise stated. Note the log scale on the time axis. Along the ordinate axis, the Gal4p scale has been amplified five-fold, and the galactose scale has been decreased two-fold. This simulation involves no delay in transformations. In this respect, the result may be compared with Fig. 4A.

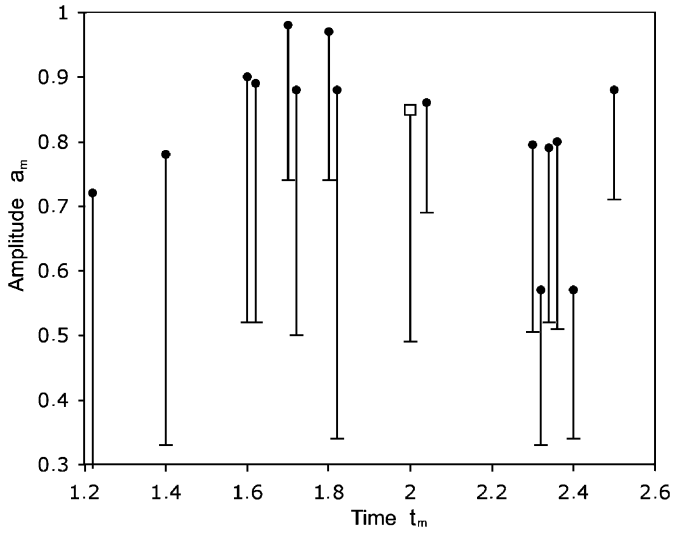


Fig. 6. Robustness of the nonreduced model to parameter variations. The scatter plot displays the amplitude a_m versus time t_m of the maximal signal response of Gal4p (as defined in Fig. 5) for 10-fold increase or decrease of parameter $k1-k7$ and $\pm 50\%$ variation in parameter c. The square represents the model with no parameter variation. Each vertical bar indicates how much the Gal4p response decreases after the maximal signal response until time $t = 6$ (d_m distance in Fig. 5).

Appendix A

The detailed model includes the following equations corresponding to Fig. 2A. This model includes slow Gal3/80p degradation (Ruhela et al., 2004) and distinguishes Gal3p and Gal80p. Altogether, it involves more parameters than the reduced model of the main text. Several of these additional parameters add very little to the analysis of the feedback loop, which is the main focus of this study. They do allow for a slightly more realistic simulation, however. It is possible to introduce time delays corresponding to the biosynthetic reactions.

$$c \times \text{Gal4p/DNA} \xrightarrow{k_1} c \times \text{Gal4p/DNA} + \text{Gal3p} + \text{Gal80p},$$

$$\text{Gal80p} + \text{Gal4p/DNA} \underset{k_3}{\overset{k_2}{\longleftrightarrow}} \text{Gal80p/Gal4p/DNA},$$

$$\text{Gal3p} + \text{Gal80p} \underset{k_5}{\overset{k_4}{\longleftrightarrow}} \text{Gal3p/Gal80p},$$

$$\text{Gal3p/Gal80p} + \text{Galactose} \xrightarrow{k_6} \text{Galactose/Gal3p/Gal80p},$$

$$\text{Gal3p} \xrightarrow{k_7} \text{degradation},$$

$$\text{Gal80p} \xrightarrow{k_7} \text{degradation}.$$

The model has been simulated in Mathematica (Fig. 5) with the following equations and parameters.

$$\begin{aligned}
\mathrm{d}(\text{Gal4p|DNA})/\mathrm{d}t = &-k_2\,\text{Gal4p|DNA}\,\text{Gal80p} \\
&+k_3\,\text{Gal80p|Gal4p|DNA}, \\
\mathrm{d}(\text{Gal3p})/\mathrm{d}t = &\,k_1\,\text{Gal4p|DNA}^{c} \\
&-k_4\,\text{Gal3p}\,\text{Gal80p} \\
&+k_5\,\text{Gal3p|Gal80p} \\
&-k_7\,\text{Gal3p}, \\
\mathrm{d}(\text{Gal80p})/\mathrm{d}t = &\,k_1\,\text{Gal4p|DNA}^{c} \\
&-k_4\,\text{Gal3p}\,\text{Gal80p} \\
&+k_5\,\text{Gal3p|Gal80p} \\
&-k_7\,\text{Gal80p} \\
&-k_2\,\text{Gal4p|DNA}\,\text{Gal80p} \\
&+k_3\,\text{Gal80p|Gal4p|DNA}, \\
\mathrm{d}(\text{Gal3p|Gal80p})/\mathrm{d}t = &-k_5\,\text{Gal3}p\text{|Gal80p} \\
&+k_4\,\text{Gal3p}\,\text{Gal80p} \\
&-k_6\,\text{Gal3p|Gal80p}\,\text{Galactose} \\
&+k_7\,\text{Galactose|Gal3p|Gal80p} \\
\mathrm{d}(\text{Gal80p|Gal4p|DNA})/\mathrm{d}t = &\,k_2\,\text{Gal80p}\,\text{Gal4p|DNA} \\
&-k_3\,\text{Gal80p|Gal4p|DNA}, \\
\mathrm{d}(\text{Galactose})/\mathrm{d}t = &-k_6\,\text{Gal3p|Gal80p}\,\text{Galactose}.
\end{aligned}$$

The various parameters of the complete dynamic model used in the simulations are: $k_1 = 1$, $k_2 = 1$, $k_3 = 1$, $k_4 = 1$, $k_5 = 1$, $k_6 = 1$, $k_7 = 1\text{E}{-4}$, $c = 4$.

Initial concentrations are: Gal4p/DNA(0) = 0.1, Gal3p(0) = 1, Gal80(0) = 1, Gal3p/Gal80p(0) = 1, Gal80p/Gal4p/DNA(0) = 0.1, Galactose(0) = 2.

A.1. Influence of parameter values in the previous model

Biochemical parameters are expected to vary somewhat from cell to cell and from one member of a species to another. Furthermore, given the uncertainty that exists on parameter values, we need to test this model with parameter

variation. To assess how the qualitative behavior of this model is affected by parameter variations, we observed the main characteristic features of the Gal4p response. The two-step response, increase and decrease in Gal4p that implement adaptation (Lauffenburger, 2000), should be robust to variations in parameters. Fig. 6 plots the time t_m at which the peak of Gal4p response concentration is reached, the amplitude a_m of this response and the decrease d_m from the peak to $t = 4$. These three values are illustrated in Fig. 5. Results of parameter variations exploring two orders of magnitude for seven parameters k_1-k_7 and $\pm 50\%$ variation for c are shown in Fig. 6. The main qualitative features, perfect or partial adaptation, of the Gal4p signal are preserved in all simulations and the appearance of the response never deviated significantly from the control with no parameter change. The average amplitude (resp. time) is 0.8 (resp. 2), standard deviation of the amplitude (resp. time) is 0.1 (resp. 0.4). Dynamically, there is no qualitative difference compared to the simplified model (Fig. 4E-F) when it also includes degradation, even with a wide range of parameter variations.

References

Bardwell, L., Cook, J.G., Zhu-Shimoni, J.X., Voora, D., Thorner, J., 1998. 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), 15400–15405.

Basu, S., Mehreja, R., Thiberge, S., Chen, M.T., Weiss, R., 2004. Spatiotemporal control of gene expression with pulse-generating networks. Proc. Natl Acad. Sci. USA 101 (17), 6355–6360.

Bhat, P.J., Hopper, J.E., 1992. Overproduction of the GAL1 or GAL3 protein causes galactose-independent activation of the GAL4 protein: evidence for a new model of induction for the yeast GAL/MEL regulon. Mol. Cell. Biol. 12 (6), 2701–2707.

Bhat, P.J., Murthy, T.V., 2001. Transcriptional control of the GAL/MEL regulon of yeast *Saccharomyces cerevisiae:* mechanism of galactose-mediated signal transduction. Mol. Microbiol. 40 (5), 1059–1066.

Bhaumik, S.R., Green, M.R., 2001. SAGA is an essential in vivo target of the yeast acidic activator Gal4p. Genes Dev. 15 (15), 1935–1945.

Biggar, S.R., Crabtree, G.R., 2001. Cell signaling can direct either binary or graded transcriptional responses. EMBO J. 20 (12), 3167–3176.

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

Hirst, M., Kobor, M.S., Kuriakose, N., Greenblatt, J., Sadowski, I., 1999. GAL4 is regulated by the RNA polymerase II holoenzyme-associated cyclin-dependent protein kinase SRB10/CDK8. Mol. Cell. 3 (5), 673–678.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., Hood, L., 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292 (5518), 929–934.

Johnston, M., Flick, J.S., Pexton, T., 1994. Multiple mechanisms provide rapid and stringent glucose repression of GAL gene expression in *Saccharomyces cerevisiae*. Mol. Cell. Biol. 14 (6), 3834–3841.

Larschan, E., Winston, F., 2001. The *S. cerevisiae* SAGA complex functions in vivo as a coactivator for transcriptional activation by Gal4. Genes. Dev. 15 (15), 1946–1956.

Lauffenburger, D.A., 2000. Cell signaling pathways as control modules: Complexity for simplicity? Proc. Natl Acad. Sci. USA 97 (10), 5031–5033.

Morohashi, M., Winn, A.E., Borisuk, M.T., Bolouri, H., Doyle, J., Kitano, H., 2002. Robustness as a measure of plausibility in models of biochemical networks. J. Theor. Biol. 216 (1), 19–30.

Parthun, M.R., Jaehning, J.A., 1992. A transcriptionally active form of GAL4 is phosphorylated and associated with GAL80. Mol. Cell. Biol. 12 (11), 4981–4987.

Peng, G., Hopper, J.E., 2000. Evidence for Gal3p's cytoplasmic location and Gal80p's dual cytoplasmic-nuclear location implicates new mechanisms for controlling Gal4p activity in *Saccharomyces cerevisiae*. Mol. Cell. Biol. 20 (14), 5140–5148.

Peng, G., Hopper, J.E., 2002. Gene activation by interaction of an inhibitor with a cytoplasmic signaling protein. Proc. Natl Acad. Sci. USA 99 (13), 8548–8553.

Platt, A., Reece, R.J., 1998. The yeast galactose genetic switch is mediated by the formation of a Gal4p-Gal80p-Gal3p complex. EMBO J. 17 (14), 4086–4091.

Platt, A., Ross, H.C., Hankin, S., Reece, R.J., 2000. The insertion of two amino acids into a transcriptional inducer converts it into a galactokinase. Proc. Natl Acad. Sci. USA 97 (7), 3154–3159.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P., Young, R.A., 2000. Genome-wide location and function of DNA binding proteins. Science 290 (5500), 2306–2309.

Rohde, J.R., Trinh, J., Sadowski, I., 2000. Multiple signals regulate GAL transcription in yeast. Mol. Cell. Biol. 20 (11), 3880–3886.

Ruhela, A., Verma, M., Edwards, J.S., Bhat, P.J., Bhartiya, S., Venkatesh, K.V., 2004. Autoregulation of regulatory proteins is key for dynamic operation of GAL switch in *Saccharomyces cerevisiae*. FEBS Lett. 576 (1-2), 119–126.

Sadowski, I., Niedbala, D., Wood, K., Ptashne, M., 1991. GAL4 is phosphorylated as a consequence of transcriptional activation. Proc. Natl Acad. Sci. USA 88 (23), 10510–10514.

Sadowski, I., Costa, C., Dhanawansa, R., 1996. Phosphorylation of Ga14p at a single C-terminal residue is necessary for galactose-inducible transcription. Mol. Cell. Biol. 16 (9), 4879–4887.

Sakurai, H., Ohishi, T., Fukasawa, T., 1994. Two alternative pathways of transcription initiation in the yeast negative regulatory gene GAL80. Mol. Cell. Biol. 14 (10), 6819–6828.

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

Sil, A.K., Alam, S., Xin, P., Ma, L., Morgan, M., Lebo, C.M., Woods, M.P., Hopper, J.E., 1999. The Gal3p-Gal80p-Gal4p transcription switch of yeast: Gal3p destabilizes the Gal80p-Gal4p complex in response to galactose and ATP. Mol. Cell. Biol. 19 (11), 7828–7840.

Smidtas, S., Yartseva, A., Schächter, V., Képès, F., 2006. Model of interactions in Biological Networks, submitted for publication.

Smolen, P., Baxter, D., Byrne, J., 2001. Modeling Circadian Oscillations with Interlocking Positive and Negative Feedback Loops. J. Neurosci. 21, 6644–6656.

Sriram, K., Gopinathan, M.S., 2004. A two variable delay model for the circadian rhythm of *Neurospora crassa*. J. Theor. Biol. 231, 23–38.

Suzuki-Fujimoto, T., Fukuma, M., Yano, K.I., Sakurai, H., Vonika, A., Johnston, S.A., Fukasawa, T., 1996. Analysis of the galactose signal transduction pathway in *Saccharomyces cerevisiae*: interaction between Gal3p and Gal80p. Mol. Cell. Biol. 16 (5), 2504–2508.

Verma, M., Bhat, P.J., Kumar, R.A., Doshi, P., 2003. Quantitative analysis of GAL genetic switch of *Saccharomyces cerevisiae* reveals that nucleocytoplasmic shuttling of Gal80p results in a highly sensitive response to galactose. J. Biol. Chem. 278 (49), 48764–48769.

Winge, O., Roberts, C., 1948. Inheritance of enzymatic characters in yeast and the phenomenon of long term adaptation. CR Trav. Laboratoire Carlsberg Series Physiol. 24, 263–315.

Wolfe, K.H., Shields, D.C., 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387 (6634), 708–713.

Yano, K., Fukasawa, T., 1997. Galactose-dependent reversible interaction of Gal3p with Gal80p in the induction pathway of Gal4p-activated genes of *Saccharomyces cerevisiae*. Proc. Natl Acad. Sci. USA 94 (5), 1721–1726.

Yeger-Lotem, E., Margalit, H., 2003. Detection of regulatory circuits by integrating the cellular networks of protein–protein interactions and transcription regulation. Nucleic Acids Res. 31 (20), 6053–6061.

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. Proc. Natl Acad. Sci. USA 101 (16), 5934–5939.

Chapitre VII – Conclusion et perspectives

Nous allons tout d'abord rappeler brièvement les conclusions de ce travail. A savoir que nous avons développé un outil d'intégration Cyclone à même d'assurer un accès et une exploitation simplifiés des données présentes dans la base de données BioCyc, développé un cadre de modélisation des graphes particulièrement adapté à l'étude des réseaux d'interactions hétérogènes MIB, caractérisé et étudié la présence et le mode de connexion de sous-réseaux ou motifs à l'intérieur de réseaux plus vastes et modélisé la voie métabolique du galactose chez la levure en tant que boucle de rétroaction impliquant régulation transcriptionnelle et interaction protéine-protéine. Puis nous critiquerons l'approche, puis nous comparons les résultats à l'état de l'art fin 2006. Enfin nous aborderons les perspectives apportées par ce travail.

Conclusion

Afin de regrouper et d'intégrer les données d'interactions biologiques hétérogènes, Cyclone, inspiré de Biocyc, a été développé. Il s'agit là d'une composante logicielle dont le code source est disponible auprès de la communauté de développeurs. Biocyc est une base de chemins réactionnels et génomes déjà existante mais ne permettant pas l'extraction de donnée et donc l'analyse de donnée. Cyclone étend le modèle de données de Biocyc, permet d'y lire et écrire les données, mais aussi de les manipuler très efficacement au moyen d'objets Java et de fichiers XML. Cyclone permet la représentation des données sous la forme de graphes ou de graphes bipartite. Nous avons utilisé Cyclone afin de prédire des interactions protéiques afin d'améliorer significativement les prédictions de phénotype de croissance dans le cadre de la construction du modèle métabolique de la bactérie *Acinetobacter*.

La plupart des études sur les réseaux hétérogènes jusque-là se sont focalisés sur la topologie, locale ou globale. Toutefois, pour être véritablement compris, les processus biologiques les plus importants comme le cheminement de signaux, la régulation de destin cellulaire, la transcription et la traduction doivent être analysés avec leur contexte. Le cadre de modélisation MIB permet d'exprimer plus de natures

d'interactions biologiques que la simple représentation sous forme de graphe, tout en préservant une facilité de manipulation de réseaux de plusieurs milliers de protéines. Il permet la représentation directe des relations n-aires essentielles pour manipuler notamment les complexes et les réactions biochimiques. Il s'articule autour des notions de consommation, production, catalyse et inhibition, sans pour autant determiner une représentation dynamique. Ainsi, il est possible de traduire un module MIB dans plusieurs formalismes différents comme les réseaux de Pétri, les équations différentielles (voir aussi chapitre VI) ou le pi-calcul (Yartseva et al, 2007). MIB est particulièrement adapté à la recherche de motifs dans les réseaux hétérogènes. Les différents types de nœuds et d'arêtes (qui contiennent notamment l'information qualitative dynamique) peuvent être utilisés pour définir des catégories de motifs fonctionnels dont il est possible de rechercher des occurrences. Le formalisme du modèle permet de rechercher des instances de motifs fonctionnels, indépendamment de l'implémentation précise de cette fonction dans la cellule. Nous l'avons illustré au chapitre V par la recherche de boucles de rétroaction, de modules de régulation métabolique ou encore de modules de mécanisme moléculaire sous-jacent à la co-expression. MIB est évidemment moins précis qu'un modèle dans lequel toutes les interactions seraient décrites par des équations totalement caractérisées, mais il est adapté à l'étude de grands réseaux et il est possible d'utiliser des dynamiques plus précises si l'on se restreint à de petits sous ensembles.

Afin d'étudier de quelle manière les différents types de réseaux homogènes sont imbriqués et se complètent, nous avons introduit l'analyse de réseaux hétérogènes à partir de comparaison avec des réseaux aléatoires qui préservent la structure macroscopique de chacun des sous-réseaux homogènes. L'application au réseau d'interactions protéiques couplé à celui de la régulation transcriptionnelle a abouti à des résultats inattendus en terme de complémentarité et coopérativité de graphe observés au moyen de mesures de distances dans les réseaux. Les graphes aléatoires construits en préservant la structure topologique de chacun de ces deux graphes (et en mélangeant l'interface des deux graphes) ont des connectivités moindres que le réseau réel. Ce résultat révèle une coopération entre ces réseaux et valide l'intérêt de l'intégration de données. Nous avons défini deux notions de permutation suivant la définition de l'ensemble des éléments permutés et de l'ensemble dans lequel ils sont permutés. Les

distances de graphe entre les paires de protéines sont plus grandes dans le réseau réel que dans les réseaux permutés.

Nous avons trouvé des motifs par analyse topologique, intéressants pour leur dynamique. Nous avons approfondi un des modules hétérogènes découvert: la boucle de régulation d'assimilation du galactose chez la levure. Si l'on considère le galactose comme le signal d'entrée, l'activité de Gal4p comme la sortie, le système se comporte comme un filtre dérivateur, sensible aux variations du signal. L'avantage pour la cellule d'un tel système est de ne produire les enzymes de dégradation du galactose qu'en proportion avec le galactose disponible et avec une grande précision. De plus, cela permet de maintenir la quantité de récepteur au galactose toujours en faible quantité, mais d'être sensible à de faibles variations de concentration de galactose dans une large gamme de concentration. Nous avons proposé que le mécanisme crucial qui explique le comportement mal compris avant que n'ai lieu la phosphorylation de Gal4p est du à cette boucle de rétroaction. Il s'agit là d'un exemple fonctionnel de filtre dérivateur naturel. Cette fonction essentielle pour l'asservissement et la robustesse (Endy, 2005) des systèmes pourrait trouver sa place parmi les modules de base pour la biologie synthétique (*Synthetic Biology*). Nous y reviendrons dans les perspectives.

Discussion

Notre démarche a consisté une fois les données regroupées, à en décrire les propriétés statistiques globales, puis à rechercher des motifs dont la structure laisserait entrevoir une dynamique intéressante. Cette approche peut être appliquée à d'autres ensembles de données et à d'autres organismes afin de découvrir d'autres modules fonctionnels. Construire des réseaux aléatoires permet de proposer des modèles de référence qui permettent par exemple de mesurer les différences entre la réalité et l'ensemble des connaissances que l'on y introduit. Nous avons choisi de ne pas construire de réseaux aléatoires avec lesquels comparer le nombre de motifs trouvés pour chaque type. En effet, on part du réseau hétérogène tel qu'il est, ce qui évite de construire un générateur de graphe aléatoire dont les bases peuvent être discutables. À la place, nous avons comparé le nombre d'instances trouvées au nombre d'instances d'autres motifs comparables. Notre approche ne vise pas non plus à simuler la mise en place de réseaux au cours de l'évolution, mais elle permet de mettre le doigt sur des motifs ou des structures macroscopiques sur ou sous représentés par rapport à d'autres. Enfin, bien

que nous ayons un objectif distinct, dans MIB il est tout de même possible sans construire de réseaux aléatoires, de prédire le nombre d'instances de motifs attendu dans un réseau aléatoire et suivant un certain nombre d'hypothèses.

Cette approche est évidemment limitée dans la nature de ses prédictions sur la dynamique des motifs. En effet, celles-ci sont effectuées uniquement à partir d'une représentation qualitative qui, bien que plus riche qu'un seul modèle de graphe, laisse à priori ouverts un certain nombre de dynamiques possibles pour 1 motif donné. Toutefois on touche là non pas aux limites de la méthodologie, mais à celles de l'information expérimentale disponible. Pour chaque interprétation dynamique d'un motif identifié, des données expérimentales supplémentaires couplées à des analyses utilisant des modèles dynamiques quantitatifs seront nécessaires.

D'un certain point de vue, la recherche de motifs peut être une composante de l'approche modulaire de la biologie des systèmes, qui consiste à regrouper (*clustering*) et modulariser les interactions biologiques. Comme nous l'avons vu dans l'introduction, les modules trouvés sont imbriqués et les motifs sur-représentés sont faiblement conservés au cours de l'évolution (Mazurie et al., 2005). Par conséquent, il n'est pas possible de découper le réseau en modules séparés tout en conservant le fonctionnement du réseau dans son ensemble. Compter le nombre d'occurrences de motifs pour en mesurer l'importance biologique est aussi limité. Tout d'abord, le résultat d'une mesure-statistique dépend fortement d'un choix d'une hypothèse nulle. Dans le cas des réseaux biologiques, le modèle de réseau aléatoire pris comme hypothèse nulle et qui peut ressembler plus ou moins au réseau réel influence significativement les résultats en termes de motifs sur-représentés (Artzy-Randrup et al., 2004, Milo et al., 2004). Par ailleurs, d'autres travaux ont montré que la sur-representation de motifs peut-être déterminée par la connectivité des réseaux, et non pour une raison biologique comme une pression d'évolution (Vazquez et al., 2001). Les travaux récents montrent que toute information purement topologique sur un groupement de nœuds dans les réseaux biologiques hétérogènes, doit être complétée d'un description fonctionnel afin de produire la connaissance biologique substantielle sur ledit groupement (Meshi et al., 2007). La recherche statistique de modules n'a permis que d'emmetre quelques hypothèses sur l'évolution des réseaux, ou d'obtenir des résultats non-triviaux en termes de robustesse à des mutations aléatoires. Le manque d'applicabilité pour la biologie est l'une des principales critiques qui est faite en général à cette approche qui consiste à

vouloir découper le système. La recherche de modules est intéressante pour tenter de comprendre quelles sont les contraintes des réseaux biologiques, pour décrire ces réseaux, pour décrire la dynamique de sous parties du système dans ou en dehors de leur contexte, et pour proposer des implémentations inconnues de fonctions biologiques que l'homme ne sait pas encore reproduire. Pour y parvenir nous avons exploité au maximum les informations de dynamique qualitative fournies par les expériences biologiques, c'est-à-dire que nous ne nous sommes pas restreint à projeter les données sur de simples graphes statiques.

Les modules trouvés, si les interactions qui les composent sont bien valides, sont bien fonctionnels dans un environnement génétique particulier et dans des conditions de vie de l'organisme qui doivent aussi bien être décrites.

Comparaison avec l'état de l'art de 2006

De nombreuses études indépendantes ont été publiées sur la recherche de modules dans les réseaux entre 2003 et 2006 comme nous l'avons vu dans l'introduction.

Les outils d'intégrations se sont développés et de nouveaux sont apparus. À ce jour, quelques standards communs de publication d'interactions ressortent. BioPax est un format d'échange encore bien incomplet qui progresse assez lentement à la vue de ce qu'on en attent. On peut aussi citer SBML qui est un format d'échange plutôt adapté à la description fine de petits réseaux. Kegg, avec l'ouverture partielle aux programmeurs de sa base, et BioCyc avec Cyclone montrent une orientation vers des outils libres d'accès. BioCyc et Cyclone intègrent un schéma de donné plus riche que Kegg. Nous espérons avoir contribué à rendre plus accessible ce type outils pour tous.

En ce qui concerne la place du modèle MIN vis-à-vis de l'état de l'art actuel, il est à un niveau intermédiaire d'abstraction entre un simple graphe et un modèle de donnée adapté non pas aux analyses mais au stockage de données. MIN est plus abstrait que le modèle proposé par van Helden et al., 2001, et permet de représenter différents types d'objets biologiques et processus uniformément. Balasubramanian et al, et Yeger-Lotem et al., sont deux auteurs qui tentent de pousser l'utilisation de simples graphes à bout afin d'étudier des réseaux hétérogènes. Ces deux études utilisent un algorithme semblable à l'algorithme de Graphe Shuffle mais non-formalisé lors de leur publication originale, si bien que les ensembles de départ et d'arrivée de cette application ne sont ni mentionnés ni étudiés pour leur rôle dans les résultats obtenus. Or les résultats obtenus

dépendent fortement de ces conditions comme nous l'avons vu. Les algorithmes de recherche dans les deux cas sont limités en taille de motifs. Dans le cas de Balasubramanian et al, seuls les motifs à deux interactions sont étudiés et dans le deuxième cas, seul les motifs à trois ou quatre interactions sont étudiées. Zhang et al., ont eux ensuite tenté d'étudier plus de deux réseaux homogènes ensemble, mais le modèle sous-jacent de graphe qu'ils utilisent, les empêche cette fois d'étudier des modules à plus de quatre interactions. En effet, leur modèle implique par exemple de représenter les complexes protéiques comme l'ensemble des interactions binaires entre tous les couples de protéines qui les composent ce qui implique une dégénérescence trop grande du nombre d'instances de motifs comportant par exemple un complexe (voir chapitre V).

Enfin, plusieurs équipes ont étudié la dynamique in-vivo de petits modules, notamment des modules intégrant le motif *Feed-Forward* que l'équipe de U. Alon a mis en exergue, donnant plus de légitimité à la notion de motif de graphe. Ce type d'étude in–vivo montre la voie pour étudier plus systématiquement les autres modules que l'on trouve, ou que l'on trouvera. En ce qui concerne le cas particulier de la boucle de rétroaction du galactose, un modèle analytique a été développé récemment et indépendamment qui montre la nécessité de la boucle de rétroaction (Ruhela et al., 2004) afin que le modèle soit cohérent avec les mesures expérimentales. Toutefois les auteurs n'en discutent pas les raisons. En effet, leur modèle comporte trop d'équations et de paramètres pour pouvoir analyser véritablement le rôle de la boucle.

Perspectives

On peut estimer le nombre global de modules de taille finie, et il serait bon dans l'avenir de parvenir à décrire véritablement tous les modules pertinents présents au sein des réseaux biologiques. A cette fin, les quatre étapes de notre approche pourraient être améliorées dans un avenir à moyen terme.

L'intégration de données que nous avons pu réaliser au moyen de Cyclone et de BIB pourrait inclure plus d'informations biologiques notamment en ce qui concerne la qualité des données. La mise à jour, et la redondance des informations présentes est un point particulièrement difficile à gérer. Le format d'échange BioPax devrait, dans ses versions à venir, permettre de publier dans un format partagé tous les différents types d'interactions biologiques nécessaires.

Le modèle MIN développé comporte une hiérarchie d'interactions que nous n'avons pas prise en compte dans les analyses car relativement peu d'informations hiérarchiques existent, c'est-à-dire l'organisation de complexes en sous-unités, les mécanismes moléculaires sous-jacents à des résultats de coexpressions ou à la production de phénotypes commencent à être disponible à grande échelle (Gain et al, 2006). Les données hiérarchisées à ce jour dans le modèle sont illustrées par l'exemple suivant : un complexe est à un niveau hiérarchique supérieur aux protéines qui le composent. Une fois un module déterminé, il serait bon de le remplacer par une boîte noire équivalente, faisant abstraction de l'implémentation biologique.

La recherche de modules fonctionnels pourrait s'appuyer sur une représentation plus abstraite de la topologie sous-jacente. Ceci permettrait par exemple de rechercher des *boucles de rétroaction* sans en spécifier le mécanisme d'interactions.

Par ailleurs, les modules trouvés, comme la régulation du galactose, doivent être validés expérimentalement dans un contexte précis. Dans le cadre de la biologie synthétique, ils pourraient être reconstruits in-vivo automatiquement. En effet, la qualité des expériences biologiques à grande échelle ne permet pas d'être certain de la validité de chacune des interactions. L'analyse quantitative de ces interactions doit aussi être menée au moyen d'expériences spécifiques. La recherche et développement de la robotisation pour construire des modules automatiquement est en cours (parts.mit.edu).

La biologie synthétique est une des perspectives les plus prometteuses. Des systèmes biologiques ont été élaborés pour manipuler de l'information, synthétiser des molécules, construire des matériaux, produire de l'énergie, produire de la nourriture, et améliorer la santé et l'environnement. Les applications de la biologie synthétique peuvent trouver des utilisations pour construire des échafaudages pour les technologies à l'échelle nanoscopique, pour l'étude du contrôle de la division cellulaire, du développement animal, pour l'étude du vieillissement et du cancer au moyen de petits circuits logiques et de mémoires… Ces applications sont plausibles physiquement comme cela a été

démontré (Elowitz and Leibler, 2000; Levskaya et al., 2005), mais construire de tels modules synthétiques demandent de nombreuses compétences en biologie, en modélisation, en simulation, en traitement du signal, et beaucoup de temps afin de regrouper, concevoir et tester les parties du système indépendamment, car il n'y a pas encore de composants standardisés. En comparaison, construire un circuit électronique équivalent demanderait bien moins d'une heure à un étudiant notamment grâce à l'utilisation de filtres dérivateurs et intégrateurs pour asservir de manière fiable le signal. En 1978, Szybalksi et Skalka écrivaient : « notre travail ne permet pas seulement d'étudier des gènes individuellement, mais nous conduit de plus vers une nouvelle ère de *biologie synthétique* où non seulement des gènes existants seront décrits et étudiés, mais aussi de nouveaux arrangements de gènes seront construits et testés. » Pour cette tâche, s'inspirer de la nature est nécessaire (Blais et Dynlacht 2005), c'est pour cette raison que j'espère aussi avoir contribué à ce domaine actuellement en plein essor qu'est la biologie synthétique.

Un concours mené par le MIT chaque année permet aux étudiants de construire leurs propres modules. Cette année, pour la première fois, une équipe française a participé. En tant qe conseiller de l'équipe, je suis fier des brillants résultats de cette équipe qui a obtenu le premier prix de recherche fondamentale le 2 Novembre 2007 pour le projet intitulé : The SMB : Synthetic Multicellular Bacterium.

Concours iGEM 2007 Photo de groupe de participants au MIT le 2 Novembre 2007.

Chapitre VIII – Références

Bibliographie

1 Albert, R., (2005) Scale-free networks in cell biology Journal of Cell Science 118, 4947-4957

2 Albert, R., Jeong, H., and Barabasi, A-L., 2000, Error and attack tolerance of complex networks, Nature, 406, p378.

3 Almogy G., Stone L., Ben-Tal N., (2001) Multi-Stage Regulation, a Key to Reliable Adaptive Biochemical Pathways. Biophysical Journal 81 3016-28

4 Alon, U. (2003). "Biological networks: the tinkerer as an engineer." Science 301(5641): 1866-7.

5 Aloy, P. and R. B. Russell (2004). "Taking the mystery out of biological networks." EMBO Rep 5(4): 349-50.

5bis Artzy-Randrup Y., Fleishman S.J., Ben-Tal N., Stone L., Comment on « Network Motifs : Simple Building Blocks of Complex Network » and « Superfamilies of Evolved and Designes Networks » Science 2004, 20 :1107

6 Babu M. and Teichmann S.A., Evolution of transcription factors and the gene regulatory network in Escherichia coli, Nucleic Acids Res. 31 (2003), pp. 1234–1244

7 Bader G, Cary M., and Sander C., Pathguide: a Pathway Resource List (2006) Nucleic Acids Res. 1; 34 D504–D506.

8 Bader G. and C. Hogue. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 4:2, Jan 2003.

9 Bader, G. D., M. P. Cary, et al. (2006). "Pathguide: a pathway resource list." Nucleic Acids Res 34(Database issue): D504-6.

10 Balasubramanian R., T. LaFramboise, D. Scholtens, and R. Gentleman. A graph-theoretic approach to testing associations between disparate sources of functional genomics data. Bioinformatics, 20(18):3353–62, Dec 2004.

11 Balasubramanian, R., T. LaFramboise, et al. (2004). "A graph-theoretic approach to testing associations between disparate sources of functional genomics data." Bioinformatics 20(18): 3353-62.

12 Balázsi G., Barabási AL., Olvai ZN., (2005) Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli, PNAS | May 31, 2005 | vol. 102 | no. 22 | 7841-7846

13 Barabasi A. and R. Albert. Emergence of scaling in random networks. Science, 286(5439):509–12, Oct 1999.

14 Barabasi, A. L. and Z. N. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nat Rev Genet 5(2): 101-13.

15 Bardwell, L., Cook, J. G., Zhu-Shimoni, J. X., Voora, D., Thorner, J., 1998. Differential regulation of transcription: repression by unactivated mitogen-activated protein kinase Kss1 requires the Dig1 and Dig2 proteins. Proc Natl Acad Sci U.S.A. 95(26) 15400-5.

16 Barrett, C. L. and B. O. Palsson (2006). "Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach." PLoS Comput Biol 2(5): e52.

17 Barrett, C. L., C. D. Herring, et al. (2005). "The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states." Proc Natl Acad Sci U S A 102(52): 19103-8.

18 Basu, S., Mehreja, R., Thiberge, S., Chen, M.T., Weiss, R., 2004. Spatiotemporal control of gene expression with pulse-generating networks. Proc. Natl. Acad. Sci. U.S.A. 101 (17), 6355-60.

19 Batchelor E , Goulian M (2003) Robustness and the cycle of phosphorylation and dephosphorylation in a two-component regulatory system. Proc Natl Acad Sci USA 100: 691–696

20 Becskei A , Serrano L (2000) Engineering stability in gene networks by autoregulation. Nature 405: 590–593

21 Bhat, P.J., Hopper, J.E., 1992. Overproduction of the GAL1 or GAL3 protein causes galactose-independent activation of the GAL4 protein: evidence for a new model of induction for the yeast GAL/MEL regulon. Mol. Cell. Biol. 12 (6), 2701-7.

22 Bhat, P.J., Murthy, T.V., 2001. Transcriptional control of the GAL/MEL regulon of yeast Saccharomyces cerevisiae: mechanism of galactose-mediated signal transduction. Mol. Microbiol. 40 (5), 1059-66.

23 Bhaumik, S.R., Green, M.R., 2001. SAGA is an essential in vivo target of the yeast acidic activator Gal4p. Genes. Dev. 15 (15), 1935-45.

24 Biggar, S.R., Crabtree, G.R., 2001. Cell signaling can direct either binary or graded transcriptional responses. EMBO J. 20 (12), 3167-76.
25 Blais A and Dynlacht BD.(2005) Constructing transcriptional regulatory networks Genes & Dev., 19(13): 1499 - 1511.
26 Boelle P-Y., J-P. Comet, G. Hutzler, F. Kepes, C. Kuttler, D. Mestivier, K. Pakdaman, A. Richard, S. Smidtas and A. Yartseva. Modeling course for biology and gene regulation: applied to Lambda Switch. ISBN 978-1-4116-9545-0 (2006) (***)
27 Bollobas B., Random Graphs, Academic Press, London, UK, (1985).
28 Bork, P. (2002). "Comparative analysis of protein interaction networks." Bioinformatics 18 Suppl 2: S64.
29 Bork, P., L. J. Jensen, C. Von Mering, A. K. Ramani, I. Lee and E. M. Marcotte (2004). "Protein interaction networks from yeast to human." Curr Opin Struct Biol 14(3): 292-9.
30 Bourguignon P.Y., V. Danos, F. Kepes, V. Schachter, S. Smidtas. Property-driven statistics of biological networks : GraphShuffle LNCS Transactions on Computational Systems Biology, (2005) (***)
31 Brun, C., F. Chevenet, D. Martin, J. Wojcik, A. Guenoche and B. Jacq (2003). "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network." Genome Biol 5(1): R6.
32 Campillos M., von Mering C., Jensen LJ. and Bork P., (2006) Identification and analysis of evolutionarily cohesive functional modules in protein networks, Genome Research 16:374-382, 2006
32b Cardelli L., (2003) Brane calculi, ENTCS Proceedings of Bio-Concur, Springer
33 Cary, M. P., G. D. Bader, et al. (2005). "Pathway information for systems biology." FEBS Lett 579(8): 1815-20.
34 Causton, H. C., B. Ren, et al. (2001). "Remodeling of yeast genome expression in response to environmental changes." Mol Biol Cell 12(2): 323-37.
35 Chen. B.-C., Wang Y.-C., Wu W.-S., Li W.-H., (2005) A new measure of the robustness of biochemical networks. Bioinformatics 21(11) 2698-705
36 Cho, R. J., M. J. Campbell, et al. (1998). "A genome-wide transcriptional analysis of the mitotic cell cycle." Mol Cell 2(1): 65-73.
37 Covert, M. W. and B. O. Palsson (2002). "Transcriptional regulation in constraints-based metabolic models of Escherichia coli." J Biol Chem 277(31): 28058-64.
38 Covert, M. W. and B. O. Palsson (2003). "Constraints-based models: regulation of gene expression reduces the steady-state solution space." J Theor Biol 221(3): 309-25.
39 Covert, M. W., C. H. Schilling, et al. (2001). "Regulation of gene expression in flux balance models of metabolism." J Theor Biol 213(1): 73-88.
40 Covert, M. W., E. M. Knight, et al. (2004). "Integrating high-throughput and computational data elucidates bacterial networks." Nature 429(6987): 92-6.
41 Cusick ME., Klitgord N., Vidal M. and Hill DE. (2005) Interactome: gateway into systems biology Human Molecular Genetics 14(suppl_2):R171-R181
42 Cynthia J. Krieger, and Peter D. Karp MetaCyc: a multiorganism database of metabolic pathways and enzymes Nucleic Acids Research, 32(1):D438-42 2004.
43 Dandekar, T., B. Snel, M. Huynen and P. Bork (1998). "Conservation of gene order: a fingerprint of proteins that physically interact." Trends Biochem. Sci. 23: 324-328.
44 de Jong, H. (2002). "Modeling and simulation of genetic regulatory systems: a literature review." J Comput Biol 9(1): 67-103.
45 de Lichtenberg, U., L. J. Jensen, et al. (2005). "Dynamic complex formation during the yeast cell cycle." Science 307(5710): 724-7.
46 Demir O, Aksan Kurnaz I., (2006) An integrated model of glucose and galactose metabolism regulated by the GAL genetic switch. Comput Biol Chem. 2006 Jun;30(3):179-9
47 Doi A., S. Fujita, H. Matsuno, M. Nagasaki, and S. Miyano. Constructing biological pathway models with hybrid functional Petri nets. In Silico Biology, 4(0023), 2004.
48 Durot M., Le Fèvre F., Pinaud B., Kreimeyer A., Perret A., De Bernardinis V., Smidtas S., Weissenbach J., Schachter V., Reconstruction of a Genome-scale Model of Acinetobascter ADP1 sp. Metabolism and Analysis of Phenotypic Profiles. Poster at 2nd European Conference on Prokaryotic Genomes, Göttingen, GE, September 23-26 2005
49 Eichenberger P , Fujita M , Jensen ST , Conlon EM , Rudner DZ , Wang ST , Ferguson C , Haga K , Sato T , Liu JS , Losick R (2004) The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis. PLoS Biol 2: e328
50 Elowitz MB , Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403: 335–338

51 Endy D., (2005) Foundations for engineering biology. Nature 438, 449-453
52 Enright, A. J., I. Iliopoulos, et al. (1999). "Protein interaction maps for complete genomes based on gene fusion events." Nature 402(6757): 86-90.
53 Erdös P. and A. Rényi. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci., 5:1761, 1960.
54 Evangelisti A. M., Wagner A., (2004) Molecular evolution in the yeast transcriptional regulation network, J. of Experimental Zoology Part B: Molecular and Developmental Evolution, 302B, 4, 392 – 411.
55 Fangting Li, Tao Long, Ying Lu, Qi Ouyang, and Chao Tang. The yeast cell-cycle network is robustly designed. PNAS, 101(14):11250–11255, April 2004.
56 Feinberg, M. (1980). Chemical oscillations, multiple equilibria, and reaction network structure. In Dynamics of reactive systems, (ed. W. Stewart, Rey, W. and Conley,C.), pp. 59-130. New York: Academic Press.
57 Fell, D. (1997). Understanding the Control of Metabolism. London, Portland Press.
58 Forster, J., I. Famili, et al. (2003). "Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network." Genome Res 13(2): 244-53.
59 François P, Hakim V., Core genetic module : the Mixed Feedback Loop, Phys. Rev. E 72: 031908 (2005)
60 François P., Hakim V., Design of genetic networks with specified functions by evolution in silico ,Proc. Natl. Acad. Sci. USA, 101, 580-585 (2004).
60b François P., Thèse de doctorat de l'Université de Paris VII. Réseaux génétiques: conception, modélisation et dynamique. 16 Septembre 2006.
61 Friedman, N. (2004). "Inferring Cellular Networks Using Probabilistic Graphical Models." Science 303(5659): 799-805.
61b Fu P., A perspective of synthetic biology : Assembling blocks for novel functions (2006) Biotechnol. J. 690-699
62 Gardner TS , Cantor CR , Collins JJ (2000) Construction of a genetic toggle switch in Escherichia coli. Nature 403: 339–342
63 Gat-Viks, I., A. Tanay, et al. (2004). "Modeling and analysis of heterogeneous regulation in biological networks." J Comput Biol 11(6): 1034-49.
64 Gat-Viks, I., A. Tanay, et al. (2006). "A probabilistic methodology for integrating knowledge and experiments on biological networks." J Comput Biol 13(2): 165-81.
65 Gavin AC, M Bosche, R Krause, P Grandi, M Marzioch, A Bauer, J Schultz, JM Rick, AM Michon, CM Cruciat, M Remor, C Hofert, M Schelder, M Bra jenovic, H Ruffner, A Merino, K Klein, M Hudak, D Dickson, T Rudi, V Gnau, A Bauch, S Bastuck, Huhse, C Leutwein, MA Heurtier, RR Copley, A Edelmann, E Querfurth, V Rybin, G Drewes, M Raida, T Bouwmeester, P Bork, Seraphin, B Kuster, G Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 415(6868)(Jan 10):141–7., 2002.
65bis Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G. « Proteome survey reveals modularity of the yeast cell machinery. » Nature. 2006 Mar 30;440(7084):631-6
66 Giot, L., J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, et al. (2003). "A protein interaction map of Drosophila melanogaster." Science 302(5651): 1727-36.
67 Goffeau A., 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, and S. Oliver. Life with 6000 genes. Science, 274(5287):546, 563–7, Oct 1996.
68 Goldbeter A (2002) Computational approaches to cellular rhythms. Nature 420: 238–245
69 Guelzim N., 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.
70 Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P. et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430, 88-93
71 Harbison, C. T., D. B. Gordon, et al. (2004). "Transcriptional regulatory code of a eukaryotic genome." Nature 431(7004): 99-104.
72 Hartemink, A. J. (2005). "Reverse engineering gene regulatory networks." Nat Biotechnol 23(5): 554-5.
73 Hartwell, L. H., J. J. Hopfield, S. Leibler and A. W. Murray (1999). "From molecular to modular cell biology." Nature 402(6761 Suppl): C47-52.

74 Heinrich, R. and S. Schuster (1996). The Regulation of Cellular Systems. New York, Chapman and Hall.
75 Hermjakob H., Montecchi-Palazzi L., Bader G., Wojcik J., Salwinski L., Ceol A., Moore S., Orchard S., Sarkans U., von Mering C., et al. The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. Nat. Biotechnol. 2004;22:177–183
76 Herrgard M. and B. Palsson. Untangling the web of functional and physical interactions in yeast. Journal of Biology, 4(5), 2005.
77 Herrgard, M. J., B. S. Lee, et al. (2006). "Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae." Genome Res 16(5): 627-35.
78 Hirst, M., Kobor, M.S., Kuriakose, N., Greenblatt, J., Sadowski, I., 1999. GAL4 is regulated by the RNA polymerase II holoenzyme-associated cyclin-dependent protein kinase SRB10/CDK8. Mol. Cell. 3 (5), 673-8.
79 Ho Y, A Gruhler, A Heilbut, GD Bader, L Moore, SL 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, AR Willems, H Sassi, PA Nielsen, KJ Rasmussen, JR Andersen, LE Johansen, LH Hansen, H Jespersen, A Podtelejnikov, E Nielsen, J Crawford, V Poulsen, BD Sorensen, J Matthiesen, RC Hendrickson, F Gleeson, T Pawson, MF Moran, D Durocher, M Mann, CW Hogue, D Figeys, and M Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. Nature, 415(6868)(Jan 10):180–3, 2002.
80 Hoffmann, R. and A. Valencia (2003). "Protein interaction: same network, different hubs." Trends Genet 19(12): 681-3.
81 Hoffmann, R., M. Krallinger, et al. (2005). "Text mining for metabolic pathways, signaling cascades, and protein networks." Sci STKE 2005(283): pe21.
82 Hu Z, Mellor J, Wu J, Yamada T, Holloway D, Delisi C., (2005) VisANT: data-integrating visual framework for biological networks and modules Nucleic Acids Res.
83 Hucka M., Finney A., Sauro H.M., Bolouri H., Doyle J.C., Kitano H., Arkin A.P., Bornstein B.J., Bray D., Cornish-Bowden A., et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. 2003;19:524–531.
84 Hughes, T. R. (2000). "Functional discovery via a compendium of expression profiles." Cell 102:109-126.
85 Huynen, M. A. and P. Bork (1998). "Measuring genome evolution." Proc. Natl Acad. Sci. USA 95: 5849-5856.
86 Huynen, M., B. Snel, W. Lathe and P. Bork (2000). "Predicting protein function by genomic context: quantitative evaluation and qualitative inferences." Genome Res. 10: 1204-1210.
87 Ideker, T., O. Ozier, B. Schwikowski and A. F. Siegel (2002). "Discovering regulatory and signalling circuits in molecular interaction networks." Bioinformatics 18 Suppl 1: S233-40.
88 Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., Hood, L., 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292 (5518), 929-34.
89 Ihmels J., Bergmann S., Gerami-Nejad M., Yanai I., McClellan M., Berman J., and Barkai N. (2005) Rewiring of the Yeast Transcriptional Network Through the Evolution of Motif Usage Science, 309(5736): 938 – 940
90 Ihmels J., G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. Nat Genet, 31(4):370–7, 2002.
91 Ihmels, J., R. Levy, et al. (2004). "Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae." Nat Biotechnol 22(1): 86-92.
92 Iliopoulos, I., S. Tsoka, et al. (2003). "Evaluation of annotation strategies using an entire genome sequence." Bioinformatics 19(6): 717-26.
93 Ishihara S., Fujimoto K., and Shibata T. (2005) Cross talking of network motifs in gene regulation that generates temporal pulses and spatial stripes Genes Cells, 10(11): 1025 - 1038.
94 Ito T., 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.
95 Ito, T., T. Chiba, et al. (2001). "A comprehensive two-hybrid analysis to explore the yeast protein interactome." Proc Natl Acad Sci U S A 98(8): 4569-74.
96 Itzkovitz S , Alon U (2005) Subgraphs and network motifs in geometrical networks. Phys Rev E 71: 0261171–0261179
96bis Jacob, F. and Monod, J. 1961. “Genetic Regulatory Mechanisms in the Synthesis of Proteins.” Journal of Molecular Biology 3: 318 –356.

97 Jansen, R., D. Greenbaum, et al. (2002). "Relating whole-genome expression data with protein-protein interactions." Genome Res 12(1): 37-46.
98 Jansen, R., H. Yu, et al. (2003). "A Bayesian networks approach for predicting protein-protein interactions from genomic data." Science 302(5644): 449-53.
99 Jeong H., B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabasi. The large-scale organization of metabolic networks. Nature, 407:651–654, 2000.
100 Jeong H., S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. Nature, 411(6833):41–2, 2001.
101 Jiang R., Tu Z., Chen T., and Sun F. (2006) Network motif identification in stochastic networks PNAS, 103(25): 9404 - 9409.
102 Johnston, M., Flick, J.S., Pexton, T., 1994. Multiple mechanisms provide rapid and stringent glucose repression of GAL gene expression in Saccharomyces cerevisiae. Mol. Cell. Biol. 14 (6), 3834-41.
103 Kalir S , Alon U (2004) Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. Cell 117: 713–720
104 Kalir S., Mangan S., Alon U., A coherent feed-forward loop with a SUM input function prolongs flagella expression in Escherichia coli, Molecular Systems Biology 1
105 Kalir, S., S. Mangan, et al. (2005). "A coherent feed-forward loop with a SUM input function prolongs flagella expression in Escherichia coli." Mol Syst Biol 1: 2005 0006.
106 Karp P., K. Myers, and T. Gruber, The Generic Frame Protocol in Proceedings of the 1995 International Joint Conference on Artificial Intelligence, pp. 768--774, 1995.
107 Karp P., S. Paley, and P. Romero, The Pathway Tools Software Bioinformatics 18:S225-32 2002.
108 Karp P.D., Arnaud M., Collado-Vides J., Ingraham J., Paulsen I.T., Saier M.H. Jr. The E. coli EcoCyc Database: No Longer Just a Metabolic Pathway Database ASM News 70(1): 25-30. 2004.
109 Karp, P. D., M. Riley, et al. (2002). "The EcoCyc Database." Nucleic Acids Res 30(1): 56-8.
110 Karp, P. D., S. Paley, et al. (2002). "The Pathway Tools software." Bioinformatics 18 Suppl 1: S225-32.
111 Kashtan N., Itzkovitz S., Milo R., Alon U., (2004) Topological Generalizations of network motifs. Phys Rev E 70, 031909
112 Kashtan N., S. Itzkovitz, R. Milo, and U. Alon. (2004) Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics.
113 Kauffman, S. A. (1993). The Origins of Order : Self Organization and Selection in Evolution. New York, Oxford University Press.
113 bis Kaufman M., Soule C., Thomas R., A new necessary condition on interaction graphs for multistationarity. Journal of Theoretical Biology Vol.248, (2007), 675-685.
114 Kelley, R. and T. Ideker (2005). "Systematic interpretation of genetic interactions using protein networks." Nat Biotechnol 23(5): 561-6.
115 Kharchenko, P., D. Vitkup, et al. (2004). "Filling gaps in a metabolic network using expression information." Bioinformatics 20 Suppl 1: I178-I185.
116 Kharchenko, P., G. M. Church, et al. (2005). "Expression dynamics of a cellular metabolic network." Mol Syst Biol 1: 2005 0016.
117 Klemm K, Bornholdt S. (2005) Topology of biological networks and reliability of information processing. Proc Natl Acad Sci U S A.102(51):18414-9.
118 Klipp, E., R. Heinrich, et al. (2002). "Prediction of temporal gene expression. Metabolic opimization by re-distribution of enzyme activities." Eur J Biochem 269(22): 5406-13.
119 Kobayashi H, Kaern M, Araki M, Chung K, Gardner (2004) Programmable cells: interfacing natural and engineered gene networks. Proc Natl Acad Sci U S A. 1;101(22):8414-9.
120 Krause, R., C. von Mering and P. Bork (2003). "A comprehensive set of protein complexes in yeast: mining large scale protein-protein interaction screens." Bioinformatics 19(15): 1901-8.
121 Krishna S., Andersson A. M. C., Semsey S., and Sneppen K. (2006) Structure and function of negative feedback loops at the interface of genetic and metabolic networks. Nucleic Acids Res. 34(8): 2455 - 2462.
122 Krummenacker M, Paley S, Mueller L, Yan T, Karp PD. Querying and computing with BioCyc databases Bioinformatics. 15;21(16):3454-5. 2005.
123 Lahav G , Rosenfeld N , Sigal A , Geva-Zatorsky N , Levine AJ , Elowitz MB , Alon U (2004) Dynamics of the p53-Mdm2 feedback loop in individual cells. Nat Genet 36: 147–150
124 Larschan, E., Winston, F., 2001. The S. cerevisiae SAGA complex functions in vivo as a coactivator for transcriptional activation by Gal4. Genes. Dev. 15 (15), 1946-56.
125 Laub MT , McAdams HH , Feldblyum T , Fraser CM , Shapiro L (2000) Global analysis of the genetic network controlling a bacterial cell cycle. Science 290: 2144–2148
126 Lauffenburger, D. A., 2000. Cell signaling pathways as control modules: Complexity for simplicity? Proc Natl Acad Sci U.S.A. 97(10) 5031-5033.

127 Le Fevre F, Smidtas S, Schachter V. Ongoing development of Cyclone Pathway Tools User Group Meeting, Geneva, Switzerland. December 1, 2005 (***)
128 Le Fevre F., Smidtas S., Schachter V. Cyclone: a Java workbench designed to manipulate Pathway Genome Databases. Bioinformatics, (accepté) http://nemo-cyclone.sourceforge.net (2006) (***)
129 Lee T., 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, and R. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. Science, 298(5594):799–804, Oct 2002.
130 Lee TJ, Pouliot Y, Wagner V, Gupta P, Stringer-Calvert DW, Tenenbaum JD, Karp PD. BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinformatics. 2006 Mar 23;7:170.
131 Lee, T. I. (2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae." Science 298: 799-804.
132 Letovsky, S. and S. Kasif (2003). "Predicting protein function from protein/protein interaction data: a probabilistic approach." Bioinformatics 19 Suppl 1: i197-204.
133 Levskaya A., Chevalier A., Tabor J., Simpson Z., Lavery L., Levy M., Davidson E., Scouras A., Ellington A., Marcotte E., Voigt C., Bacterial photography: Engineering Escherichia coli tosee light. 2005 Nature 438, 441-442
134 Li, S. (2004). "A map of the interactome network of the metazoan C. elegans." Science 303: 540-543.
135 Lloyd C.M., Halstead M.D., Nielsen P.F. CellML: its future, present and past. Prog. Biophys. Mol. Biol. 2004;85:433–450.
135b Loew, L.,M., Schaff, J.C., (2001) The Virtual Cell : a software environment for computational cell biology, Trends Biotechnol. 19 (10) 401-406
136 Luscombe, N. M., Babu, M. M., Yu, H. Y., Snyder, M., Teichmann, S. A. and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431, 308-312
137 Ma HW, Buer J., Zeng AP., (2004) Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach, BMC Bioinformatics 2004, 5:199
138 Ma HW, Kumar B , Ditges U , Gunzer F , Buer J , Zeng AP (2004) An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. Nucleic Acids Res 32: 6643–6649.
139 Ma'ayan, A., Jenkins, S. L., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, B., Eungdamrong, N. J., Weng, G., Ram, P. T., Rice, J. J. et al. (2005). Formation of regulatory patterns during signal propagation in a Mammalian cellular network. Science 309, 1078-83.
140 Mangan S , Alon U (2003) Structure and function of the feed-forward loop network motif. Proc Natl Acad Sci USA 100: 11980–11985
141 Mangan S , Zaslaver A , Alon U (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J Mol Biol 334: 197–204
142 Mangan S., Itzkovitz S., Zaslaver A. and Alon U., (2006) The Incoherent Feed-forward Loop Accelerates the Response-time of the gal System of Escherichia coli. JMB, Vol 356 pp 1073-81
143 Marcotte, E. M., M. Pellegrini, M. J. Thompson, T. O. Yeates and D. Eisenberg (1999). "A combined algorithm for genome-wide prediction of protein function." Nature 402(6757): 83-6.
144 Maslov S., Sneppen K., (2002) Specificity and Stability in Topology of Protein Networks Science, 296 (5569) 910-913
145 Matsuno H., A. Doi, M. Nagasaki, and S. Miyano. Hybrid petri net representation of gene regulatory network. Pac Symp Biocomput, pages 341–52, 2000.
146 Mazurie A., Bottani S., Vergassola M., An evolutionary and functional assessment of regulatory network motifs. Genome Biol 6 (4), R35 (2005).
147 McAdams HH , Shapiro L (2003) A bacterial cell-cycle regulatory network operating in time and space. Science 301: 1874–1877
148 McCraith, S., Holtzman, T., Moss, B. and Fields, S. (2000). Genome-wide analysis of vaccinia virus protein-protein interactions. Proc Natl Acad Sci U S A 97, 4879-84.
148bis Meshi, O., Shlomi T., and Ruppin E., (2007). Evolutionary conservation and over-represenation of funtionally enriched network patterns in yeast regulatory network. BMC Systems Biology, 1:1
149 Mewes, H. W. (2002). "MIPS: a database for genomes and protein sequences." Nucleic Acids Res. 30: 31-34.
150 Milo R., S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. Science, 303(5663):1538–42, 2004.
151 Milo R., S. S. Shen-Orr, S. Itzkowitz, N. Kashtan, D. Chklovskii, and U. Alon. On the uniform generation of random graphs with prescribed degree sequence. ArXiv, 2003.

152 Milo R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. Science, 298(5594):824–7, 2002.
153 Milo, R. (2002). "Network motifs: simple building blocks of complex networks." Science 298: 824-827.
154 Milo, R., S. Itzkovitz, et al. (2004). "Superfamilies of evolved and designed networks." Science 303(5663): 1538-42.
155 Monod J., Jacob F., Genetic Regulatory Mechanisms in the Synthesis of Proteins, Journal of Molecular Biology (1961) 3: 318-356
156 Morohashi M., Winn A.E., Borisuk M.T., Bolouri H., Doyle J., Kitano H., 2002. Robustness as a measure of plausibility in models of biochemical networks. J Theor Biol. 216 (1), 19-30.
157 Nelson DE , Ihekwaba AE , Elliott M , Johnson JR , Gibney CA , Foreman BE , Nelson G , See V , Horton CA , Spiller DG , Edwards SW , McDowell HP , Unitt JF , Sullivan E , Grimley R , Benson N , Broomhead D , Kell DB , White MR (2004) Oscillations in NF-kappaB signaling control the dynamics of gene expression. Science 306: 704–708
158 Newman M. E. and M. Girvan. Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys, 69(2 Pt 2):026113, 2004.
159 Newman M. E. Assortative mixing in networks. Phys Rev Lett, 89(20):208701, 2002.
160 Newman M. E. Properties of highly clustered networks. Phys Rev E Stat Nonlin Soft Matter Phys, 68(2 Pt 2):026121, 2003.
161 Newman M. E., S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. Phys Rev E Stat Nonlin Soft Matter Phys, 64(2 Pt 2):026118, 2001.
162 Odom D , Zizlsperger N , Gordon D , Bell G , Rinaldi N , Murray H , Volkert T , Schreiber J , Rolfe P , Gifford D , Fraenkel E , Bell G , Young R (2004) Control of pancreas and liver gene expression by HNF transcription factors. Science 303: 1378–1381
163 Ooi, S. L., X. Pan, et al. (2006). "Global synthetic-lethality analysis and yeast functional profiling." Trends Genet 22(1): 56-63.
164 Osterman, A. and R. Overbeek (2003). "Missing genes in metabolic pathways: a comparative genomics approach." Curr Opin Chem Biol 7(2): 238-51.
165 Overbeek, R., M. Fonstein, M. D'Souza, G. D. Pusch and N. Maltsev (1999). "The use of gene clusters to infer functional coupling." Proc. Natl Acad. Sci. USA 96: 2896-2901.
166 Ozier, O., N. Amin and T. Ideker (2003). "Global architecture of genetic interactions on the protein network." Nat Biotechnol 21(5): 490-1.
167 Paley S. and P. Karp Evaluation of computational metabolic-pathway predictions for H. pylori Bioinformatics 18(5):705-14 2002.
168 Papin, J. A. and Palsson, B. O. (2004). Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. J. Theor. Biol. 227, 283-297
169 Papin, J. A., Hunter, T., Palsson, B. O. and Subramaniam, S. (2005). Reconstruction of cellular signalling networks and analysis of their properties. Nat. Rev. Mol. Cell. Biol. 6, 99-111.
170 Parthun, M.R., Jaehning, J.A., 1992. A transcriptionally active form of GAL4 is phosphorylated and associated with GAL80. Mol. Cell. Biol. 12 (11), 4981-7.
171 Pastor-Satorras, R., Smith, E. and Sole, R. V. (2003). Evolving protein interaction networks through gene duplication. J. Theor. Biol. 222, 199-210
172 Patil, K. R. and J. Nielsen (2005). "Uncovering transcriptional regulation of metabolism by using metabolic network topology." Proc Natl Acad Sci U S A 102(8): 2685-9.
173 Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proc. Natl Acad. Sci. USA 96: 4285-4288.
174 Peng, G., Hopper, J.E., 2000. Evidence for Gal3p's cytoplasmic location and Gal80p's dual cytoplasmic-nuclear location implicates new mechanisms for controlling Gal4p activity in Saccharomyces cerevisiae. Mol. Cell. Biol. 20 (14), 5140-8.
175 Peng, G., Hopper, J.E., 2002. Gene activation by interaction of an inhibitor with a cytoplasmic signaling protein. Proc. Natl. Acad. Sci. U.S.A. 99 (13), 8548-53.
176 Pereira-Leal, J. B., A. J. Enright and C. A. Ouzounis (2004). "Detection of functional modules from protein interaction networks." Proteins 54(1): 49-57.
177 Petti AA, Church GM. (2005) A network of transcriptionally coordinated functional modules in Saccharomyces cerevisiae. Genome Res. 15(9):1298-306.
178 Platt, A., Reece, R.J., 1998. The yeast galactose genetic switch is mediated by the formation of a Gal4p-Gal80p-Gal3p complex. Embo J. 17 (14), 4086-91.
179 Platt, A., Ross, H.C., Hankin, S., Reece, R.J., 2000. The insertion of two amino acids into a transcriptional inducer converts it into a galactokinase. Proc. Natl. Acad. Sci. U.S.A. 97 (7), 3154-9.

180 Price N. D., J. A. Papin, C. H. Schilling, and B. O. Palsson. Genome-scale microbial in silico models: the constraints-based approach. Trends Biotechnol, 21(4):162–9, 2003.
181 Price, N. D., J. L. Reed, et al. (2004). "Genome-scale models of microbial cells: evaluating the consequences of constraints." Nat Rev Microbiol 2(11): 886-97.
182 Qin H, Lu HH, Wu WB, Li WH. (2003) Evolution of the yeast protein interaction network. Proc Natl Acad Sci U S A. 100(22):12820-4
183 Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S.,Lenzen, G., Petel, F., Wojcik, J., Schachter, V. et al. (2001). The protein-protein interaction map of Helicobacter pylori. Nature 409, 211-5.
184 Ravasz E., A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. Science, 297(5586):1551–5, 2002.
185 Reed, J. L. and B. O. Palsson (2003). "Thirteen years of building constraint-based in silico models of Escherichia coli." J Bacteriol 185(9): 2692-9.
186 Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P., Young, R.A., 2000. Genome-wide location and function of DNA binding proteins. Science 290 (5500), 2306-9.
187 Resendis-Antonio O., Freyre-González JA., Menchaca-Méndez R., Gutiérrez-Ríos RM., Martínez-Antonio A., Ávila-Sánchez C. and Collado-Vides J., (2005) Modular analysis of the transcriptional regulatory network of E. coli Trends in Genetics 21, (1) , 16-20
188 Rhodes, D. R., S. A. Tomlins, et al. (2005). "Probabilistic model of the human protein-protein interaction network." Nat Biotechnol 23(8): 951-9.
189 Rives, A. W. and T. Galitski (2003). "Modular organization of cellular networks." Proc Natl Acad Sci U S A 100(3): 1128-33.
190 Rohde, J.R., Trinh, J., Sadowski, I., 2000. Multiple signals regulate GAL transcription in yeast. Mol. Cell. Biol. 20 (11), 3880-6.
191 Romero, P. R. and P. D. Karp (2004). "Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases." Bioinformatics 20(5): 709-17.
192 Ronen M , Rosenberg R , Shraiman BI , Alon U (2002) Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. Proc Natl Acad Sci USA 99: 10555–10560
193 Rosenfeld N , Alon U (2003) Response delays and the structure of transcription networks. J Mol Biol 329: 645–654
194 Rosenfeld N , Elowitz MB , Alon U (2002) Negative autoregulation speeds the response times of transcription networks. J Mol Biol 323: 785–793
195 Ross D. and A. Schoman. Structured analysis for requirements definition. IEEE Trans Softw Eng (Special issue on requirements analysis), 3(1):6–15, 1977.
196 Ruhela, A., Verma, M., Edwards, J.S., Bhat, P.J., Bhartiya, S., Venkatesh, K.V., 2004. Autoregulation of regulatory proteins is key for dynamic operation of GAL switch in Saccharomyces cerevisiae. FEBS Lett. 576 (1-2), 119-26.
197 Sadowski, I., Costa, C., Dhanawansa, R., 1996. Phosphorylation of Ga14p at a single C-terminal residue is necessary for galactose-inducible transcription. Mol. Cell. Biol. 16 (9), 4879-87.
198 Sadowski, I., Niedbala, D., Wood, K., Ptashne, M., 1991. GAL4 is phosphorylated as a consequence of transcriptional activation. Proc. Natl. Acad. Sci. U.S.A. 88 (23), 10510-4.
199 Sakurai, H., Ohishi, T., Fukasawa, T., 1994. Two alternative pathways of transcription initiation in the yeast negative regulatory gene GAL80. Mol. Cell. Biol. 14 (10), 6819-28.
200 Salgado H, Gama-Castro S.; Martínez-Antonio A.; Díaz-Peredo E.; Sánchez-Solano F.; Peralta-Gil M.; Garcia-Alonso D.; Jiménez-Jacinto V.; Santos-Zavaleta A.; Bonavides-Martínez C.; Collado-Vides J. (2004), RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12, Nucleic Acids Res. 32 pp. D303–D306.
201 Savageau MA (1974) Comparison of classical and autogenous systems of regulation in inducible operons. Nature 252: 546–549
202 Savageau, M. A. (1976). Biochemical Systems Analysis: a Study of Function and Design in Molecular Biology. Reading, Massachusetts, Addison-Wesley.
203 Schachter V., Danos V., Smidtas S., Kepes F., Property-driven statistics of biological networks. Computational Models for Systems Biology – 3-5 Avril 2005, Edinburgh, Scotland, Long Paper, (2005). (***)
203b Schachter V., Chapitre de livre 'Heterogeneous molecular networks' à parâitre World Scientific.
204 Schwikowski B., P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. Nat Biotechnol, 18(12):1257–61, Dec 2000.

205 Segal, E., M. Shapira, et al. (2003). "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." Nat Genet 34(2): 166-76.
206 Segre, D., J. Zucker, et al. (2003). "From annotated genomes to metabolic flux models and kinetic parameter fitting." Omics 7(3): 301-16.
207 Setty, Y., Mayo, A. E., Surette, M. G. & Alon, U. (2003) Detailed map of a cis-regulatory input function, Proc. Natl. Acad. Sci. USA 100, 7702–7707
208 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498-504.
209 Sharan, R. and T. Ideker (2006). "Modeling cellular machinery through biological network comparison." Nat Biotechnol 24(4): 427-33.
210 Shen-Orr, S. S., Milo, R., Mangan, S., Alon, U., 2002. Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 31(1), 64-8.
211 Siegel, A., O. Radulescu, et al. (2006). "Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation networks." Biosystems 84(2): 153-74.
212 Sil, A. K., Alam, S., Xin, P., Ma, L., Morgan, M., Lebo, C.M., Woods, M.P., Hopper, J.E., 1999. The Gal3p-Gal80p-Gal4p transcription switch of yeast: Gal3p destabilizes the Gal80p-Gal4p complex in response to galactose and ATP. Mol. Cell. Biol. 19 (11), 7828-40.
213 Simao, E., E. Remy, et al. (2005). "Qualitative modelling of regulated metabolic pathways: application to the tryptophan biosynthesis in E.Coli." Bioinformatics 21 Suppl 2: ii190-ii196.
214 Smidtas S., Bourgine P., Kepes F., Schachter V., Complementation of the yeast transcriptional regulation network with protein-protein interactions Poster at ISMB-ECCB, Glasgow UK, 2004 and ICSB, Heidelberg GE, 2004a (***)
215 Smidtas S., The Biological Integration Browser. Demo 1h at ISMB-ECCB, Glasgow UK, 2004 and ICSB, Heidelberg GE, 2004b (***)
216 Smidtas S., Schachter V., Kepes F. The adaptative filter of the yeast galactose pathway. Journal of Theoretical Biology, 21;242(2):372-81 doi:10.1016/j.jtbi.2006.03.005 (2005). (***)
217 Smidtas S., Yartseva A., Scachter V., Kepes F. Model of Interactions in Biology and Application to Heterogeneous Network in Yeast. Compte Rendus de l'Accadémie des Sciences – Biologies 329(12):945-52 (2006). (***)
217b Smidtas S., Yartseva A., Rooting a Graph by the Environment Interface Applied to Heterogeneous Interaction Network of the Yeast, Acta biotheoretica (accepted) (2007). (***)
218 Smolen, P., Baxter, D., Byrne, J., 2001. Modeling Circadian Oscillations with Interlocking Positive and Negative Feedback Loops. J. Neurosci. 21 6644-6656.
219 Snel B., Huynen MA., (2004) Quantifying Modularity in the Evolution of Biomolecular Systems, enome Research 14:391-397
220 Snel B., P. Bork, and M. Huynen. The identification of functional modules from the genomic association of genes. Proc Natl Acad Sci U S A, 99(9):5890–5, Apr 2002.
221 Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol Biol Cell 9(12): 3273-97.
222 Spirin V. and L. A. Mirny. Protein complexes and functional modules in molecular networks. PNAS, 100(21):12123–12128, 2003.
223 Spirin V., Gelfand MS., Mironov AA., and Mirny LA. (2006) A metabolic network in the evolutionary context: Multiscale structure and modularity PNAS, June 6, 2006, vol. 103, no. 23, 8774-8779
224 Srinivasan M., 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.
225 Sriram, K., Gopinathan, M.S., 2004. A two variable delay model for the circadian rhythm of Neurospora crassa. J. Theor. Biol. 231 23-38.
226 Stelling, J., S. Klamt, et al. (2002). "Metabolic network structure determines key aspects of functionality and regulation." Nature 420(6912): 190-3.
227 Strogatz S. H.. Exploring complex networks. Nature, 410(6825):268–76, 2001.
228 Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science 302, 249-55.
229 Suzuki-Fujimoto, T., Fukuma, M., Yano, K.I., Sakurai, H., Vonika, A., Johnston, S.A., Fukasawa, T., 1996. Analysis of the galactose signal transduction pathway in Saccharomyces cerevisiae: interaction between Gal3p and Gal80p. Mol. Cell. Biol. 16 (5), 2504-8.
230 Tanaka, R. (2005). Scale-rich metabolic networks. Phys. Rev. Lett. 94, 168101

231 Tanay A, Sharan R, Kupiec M, Shamir R. (2004) 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.
232 Tanay, A., Regev, A. and Shamir, R. (2005). Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc. Natl. Acad. Sci. USA 102, 7203-7208
233 Taylor B., (2004) An Alternative Strategy for Adaptation in Bacterial Behavior. J. Bacteriology 186(12) 3671-73
234 Tomlin C., Axelrod J., (2005) Understanding biology by reverse engineering the control. PNAS 102(12) 4219-20
235 Tong, A. H. (2001). "Systematic genetic analysis with ordered arrays of yeast deletion mutants." Science 294: 2364-2368.
236 Tong, A. H., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G. F., Brost, R. L., Chang, M. et al. (2004). Global mapping of the yeast genetic interaction network. Science 303, 808-813
237 Tornow S. and H. W. Mewes (2003) Functional modules by relating protein interaction networks and gene expression Nucleic Acids Res.; 31(21): 6283 - 6289.
238 Troncale S., D. Campard, J. Guespin, J.-P. Vannier, and F. Tahi. Modelisation´ of interleukin-6 system in early hematopoiesis with hybrid functional petri nets. In Genopole, editor, Mod´elisation de syst `emes biologiques complexes dans le contexte de la g´enomique, Du 4 au 8 avril 2005, Montpellier, 2005.
239 Tyson, J. J., Chen, K. and Novak, B. (2001). Network dynamics and cell physiology. Nat. Rev. Mol. Cell. Biol. 2, 908-916.
240 Tyson, J. J., Chen, K. C. and Novak, B. (2003). Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr. Opin. Cell Biol. 15, 221-231.
241 Uetz P., 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, and J. Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. Nature, 403(6770):623–7, Feb 2000.
242 Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 12, 368-73.
243 Vazquez A., A. Flammini, A. Maritan, and A. Vespignani. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol, 21(6):697–700, Jun 2003.
244 Verma, M., Bhat, P.J., Kumar, R.A., Doshi, P., 2003. Quantitative analysis of GAL genetic switch of Saccharomyces cerevisiae reveals that nucleocytoplasmic shuttling of Gal80p results in a highly sensitive response to galactose. J. Biol. Chem. 278 (49), 48764-9.
245 Voigt CA , Wolf D , Arkin AP (2005) The B. subtilis SIN Operon: an evolvable network motif. Genetics Genetics, Vol. 169, 1187-1202
245bis van Helden J., 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.
246 von Mering C., R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. Nature, 417(6887):399-403, 2002.
247 von Mering, C., E. M. Zdobnov, et al. (2003). "Genome evolution reveals biochemical networks and functional modules." Proc Natl Acad Sci U S A 100(26): 15428-33.
248 von Mering, C., L. J. Jensen, et al. (2005). "STRING: known and predicted protein-protein associations, integrated and transferred across organisms." Nucleic Acids Res 33(Database issue): D433-7.
249 von Mering, C., R. Krause, et al. (2002). "Comparative assessment of large-scale data sets of protein-protein interactions." Nature 417(6887): 399-403.
249 bis von Mering, C., Jensen L.J., et al. (2007). "STRING 7--recent developments in the integration and prediction of protein interactions." Nucleic Acids Res. 2007 Jan;35:D358-62. Epub 2006 Nov 10.
250 Wagner A.. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol, 18(7):1283–92
251 Wagner, A. (2003). How the global structure of protein interaction networks evolves. Proc. Biol. Sci. 270, 457-466.
252 Wagner, A. and Fell, D. A. (2001). The small world inside large metabolic networks. Proceedings of the Royal Society of London Series B-Biological Sciences 268,1803-1810.
253 Wall ME , Hlavacek WS , Savageau MA (2004) Design of gene circuits: lessons from bacteria. Nat Rev Genet 5: 34–42
254 Watts D. and S. Strogatz. Collective dynamics of 'small-world' networks. Nature, 393(6684):440–2, Jun 1998.
255 Welch, G. R. and P. R. Marmillot (1991). "Metabolic "channeling" and cellular physiology." J Theor Biol 152(1): 29-33.

256 West, D. B. (1996). Introduction to Graph Theory, Prentice Hall.
257 William Aiello, Fan Chung, and Linyuan Lu. A random graph model for massive graphs. In 32nd Annual ACM Symposium on Theory of Computing, pages 171–180, 2000.
258 Winge, O., Roberts, C., 1948. Inheritance of enzymatic characters in yeast and the phenomenon of long term adaptation. CR Trav. Laboratoire Carlsberg Series Physiol. 24, 263-315.
259 Wolf D., Arkin A., (2003) Motifs, modules and games in bacteria, Current Opinion in Microbiology, 6:125–134
260 Wolfe, K.H., Shields, D.C., 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387 (6634), 708-13.
261 Wong, S. L., L. V. Zhang, et al. (2004). "Combining biological networks to predict genetic interactions." Proc Natl Acad Sci U S A 101(44): 15682-7.
262 Wuchty, S., Z. N. Oltvai and A. L. Barabasi (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat Genet 35(2): 176-9.
263 Yano, K., Fukasawa, T., 1997. Galactose-dependent reversible interaction of Gal3p with Gal80p in the induction pathway of Gal4p-activated genes of Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U.S.A. 94(5), 1721-6.
264 Yartseva, A., Smidtas S., Klaudel H., Képès F., Incremental and unifying modeling formalism for biological interaction networks. Prix du meilleur poster – ECCS 2005, Paris, France (***)
264 bis Yartseva , A., Klaudel H., Devillers R., Kepes F., Incremental and unifying modelling formalism for biological interaction networks, BMC Bioinformatics, Vol. 8 (08 November 2007), 433.
265 Ye, P., B. D. Peyser, et al. (2005). "Commensurate distances and similar motifs in genetic congruence and protein interaction networks in yeast." BMC Bioinformatics 6: 270.
266 Ye, P., B. D. Peyser, et al. (2005). "Gene function prediction from congruent synthetic lethal interactions in yeast." Mol Syst Biol 1: 2005 0026.
267 Yeger-Lotem, E., Margalit, H., 2003. Detection of regulatory circuits by integrating the cellular networks of protein-protein interactions and transcription regulation. Nucleic Acids Res 31(20), 6053-61.
268 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. Proc Natl Acad Sci U S A 101(16), 5934-9.
269 Yi T., Simon M., Doyle J., (2000) Robust perfect adaptation in bacterial chemotaxis through integral feedback control. PNAS 97 (9), 4649-53
270 Yook S. H., H. Jeong, A. L. Barabasi, and Y. Tu. Weighted evolving networks. Phys Rev Lett, 86(25):5835–8, 2001.
271 Yook, S. H., Oltvai, Z. N. and Barabási, A. L. (2004). Functional and topological characterization of protein interaction networks. Proteomics 4, 928-942.
272 Zaslaver A , Mayo AE , Rosenberg R , Bashkin P , Sberro H , Tsalyuk M , Surette MG , Alon U (2004) Just-in-time transcription program in metabolic pathways. Nat Genet 36: 486–491
273 Zhang L., O. King, S. Wong, D. S. Goldberg, A. Tong, G. Lesage, B. Andrews, H. Bussey, C. Boone, and F. Roth. Motifs, themes and thematic maps of an integrated saccharomyces cerevisiae interaction network. Journal of Biology, 4(6), 2005.
274 Zheng, Y., J. D. Szustakowski, et al. (2002). "Computational identification of operons in microbial genomes." Genome Res 12(8): 1221-30.

Quelques liens

Biological Interaction Browser - http://www.genoscope.cns.fr/biopathways/bib

BioCyc - http://www.biocyc.com

Cyclone - http://nemo-cyclone.sourceforge.net

Genopole – http://www.genopole.org

Genoscope – http://www.genoscope.cns.fr

Hybrigenics – http//:www.hybrigenics.fr

Information sur l'auteur – http://sergi5.com/bio

Logiciel Catia - http://www.3ds.com

Mips - http://mips.gsf.de/proj/yeast/catalogues/complexes

Modèle physique d'interactions hétérogène d'Escherichia coli - http://escherichia-coli.sergi5.com

NeMo (genoscope) – http://www.genoscope.cns.fr/bioinfo

Programme épigénomique – http://epigenesis.lami.univ-evry.fr

Informations sur l'auteur

Serge SMIDTAS
Serge@Smidtas.com
Tél +33 6-63 30 34 28 52 rue Botzaris
Fax +33 1 44 32 03 19 75019 Paris France

2002-2007	**Doctorat de bioinformatique – Université Evry** Réseaux d'interactions biologiques. Directeurs: F. Képès V. Schachter
2002-2003	**Dauphine, Paris - Master '104' Finance** Modélisation financière, micro et macro économie, discret, continue.
2001-2002	**Paris VII, Université d'Orsay - Master Interface Physique Biologie** Physique pour la biologie, Science de la Vie et de la Terre
2001-2002	**Paris VI, Université Pierre et Marie Curie - Master Génétique Humaine** Institut Pasteur : Cours et TP expérimentaux (Clonage, CGH, Fish, microscopie…)
1999-2001	**Ingénieur Supélec, Paris –** Ecole supérieure d'Electricité, Gif-sur-Yvette Mineur en Biologie
1996-1999	**Louis-le-Grand, Paris – Classes préparatoires** Cours de Physique, Chimie, Mathématique

2007-	**Ingénieur – CEA Evry (Commissariat à l'Energie Atomique)**
2003-2006	**Ingénieur – Genoscope Evry (Centre National de Séquençage)** et CNRS UMR 8030
2003 - 3 mois	**Stagiaire** – Plateforme de Puce à ADN, CNRS Centre de Génétique Moléculaire – Orsay Dir. Lawrence Aggerbeck & André Adoutte
2002 - 6 mois	**Ingénieur – Cortex** - Nouvelle-Calédonie
2000 - 2months	**Stagiaire – KTH (Ecole Polytechnique Royale), Stockholm Suède**
1999 – 2000	**Enseignement**, 'Colles' au lycée Louis-le-Grand en PCSI et PC* en Physique.

Sélection d'autres publications de l'auteur

Cyclone : Java-based querying and computing with Pathway/Genome Databases
F. Le Fèvre, S. Smidtas, V. Schächter, Bioinformatics 2007; doi: 10.1093/bioinformatics/btm107

The adaptive filter of the yeast galactose pathway
S. Smidtas, V. Schachter, and F. Képès. (2006), J Theor Biol. 2006 Sep 21;242(2):372-81.

Model of Interactions in Biology and Application to Heterogeneous Network in Yeast.
Smidtas S., Yartseva A., Scachter V., Kepes F (2006), Comptes Rendus Biologies. 329 (12) 945-952 doi:10.1016/j.crvi.2006.06.010

Property-driven statistics of biological networks
PY Bourguignon, V. Danos, F. Kepes, S. Smidtas, and V. Schächter (2006), Transactions in Computational Systems Biology

Graph Shuffle
V. Schachter, V. Danos, S. Smidtas, F. Kepes - Computational Methods in Systems Biology CMSB 2005 Long Paper - 3-5 April 2005, Edinburgh, Scotland

YIB: A Tool for Biological Network Motif Analysis
S. Smidtas - In Proceedings of ISMB/ECCB 2004, July 31 - August 4, Glasgow UK - 1h-Presentation

Reconstr. of a Gen.-scale Model of Acineto. ADP1 sp. Metabolism and Analysis of Phenotypic Profiles
Durot M., Le Fèvre F., Pinaud B., Kreimeyer A., Perret A., De Bernardinis V. Smidtas S. Weissenbach J., Schachter V. 2nd European Conference on Prokar. 2006

Complex Biological Networks: Gene Regulation and Protein Interaction
Smidtas, S., September 19th - October 8th 2005, EXYSTENCE Thematic Institute, Torino, It

Ongoing development of Cyclone
Le Fèvre F, Smidtas S, Schachter V. Pathway Tools User Group Meeting, Geneva, Switzerland. December 1, 2005

Completion of the yeast transcriptional regulation network by protein-protein interactions
S. Smidtas, P. Bourgine, K. Képès, V. Schachter ISMB - ECCB 2004, July 31 - August 4, Glasgow UK - ICSB 2004, October 9 - 13, Heidelberg GE

Déco: La maison de Sergi
S. Smidtas, 2005, Octobre, Le journal des femmes

The interactive repeated flip-flop game in stressful conditions : a web-based investigation of human market interactions, soumis

Clavier servant à utiliser en derivation les fonctions du téléphone.
S. Smidtas 1993 Brevet #930994000

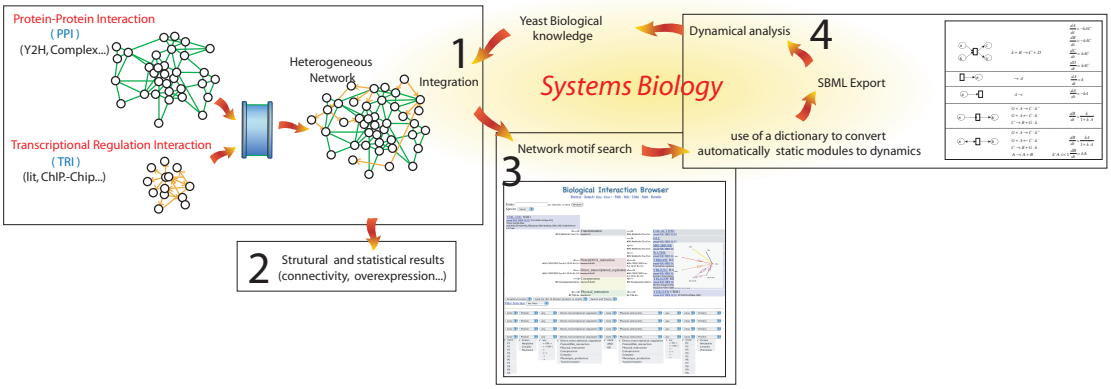
Remerciements

Annexes (disponibles en ligne)

Poster ISMB ECCB
Site web de Cyclone
Rapport de DEA sur les motifs de réseau
Code source Cyclone, Java
Code source Simulation Galactose, Matlab
Code source Biological Integration Browser, PHP
Code source Shuffle, Matlab, Perl, PHP

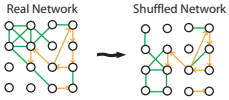
Completion of the yeast transcriptional regulation network by protein-protein interactions

Serge Smidtas, Paul Bourgine, François Képès, Vincent Schachter,
Genoscope and CNRS UMR 8030, Fr. - smidtas@genoscope.cns.fr
http://www.genoscope.cns.fr/biopathways/bib



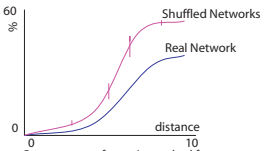
From Structural Topological properties of the heterogeneous network .....

1 Cooperativity of Networks into Distance



To assess how PPI and TRI complement each other we shuffled the interface between the two networks and build a 'shuffled network' on which statistical properties can be compared with the real network. Thus, in both networks topological properties of each subnetwork are kept the same.

From one protein, directed distance is computed to other proteins.
Compared to the real network, in the shuffled network, distance are shorter.



Pourcentage of protein reached from one protein within a distance in the network

2 Structure in Layers Upstream of Regulation

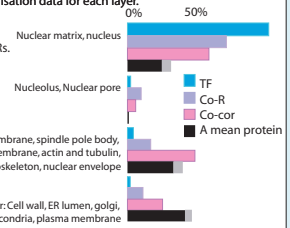
A protein is linked to about 8 other proteins in mean

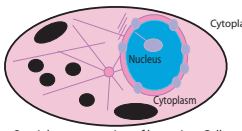
We observe an non trivial repartition of links between layers of Transcriptional Factors, putative CoRegulators, and putative Co-coregulators

Co-Coregulator
Co-Regulator — One CoR is linked to about 12 Co-cor.
Transcriptional Factor — One TF is linked to about 2 Co-r.
Regulated Gene — One Gene is regulated by about 3 TFs

To validate the structure in layers, we add protein localisation data for each layer.

-Most of TFs and Co-Rs are in the nucleus.
-Further, Nucleolus and Nuclear pores are made of CoRs.
-Further, Co-cors are next in communication parts
-Very small proportion of these first tree layers are in far from the nulceus closed in organelles





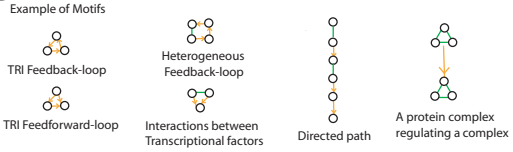
Spatial representation of layers in a Cell

Localisation of each layer is awaited and validates layers

..... To local dynamical properties and dynamical motifs

3 Heterogeneous Network Modules Discovery

Example of Motifs



Feedback-Loops — Overexpression of loops including PPI
...but sensitive to hypotheses in the underlying statistical model

Number of occurences of motifs

Loop Size	TRI only Feedback-loop	Predicted	Including 1 PPI in the loop	Predicted
2	5	2	8	1
3	4	3	6	1
4	5	5	7	1
5	4	8	22	1
6	9	15	30	1
7	6	28	58	2
8	12	53	69	6
9	22	100	120	17
10	41	190	230	35

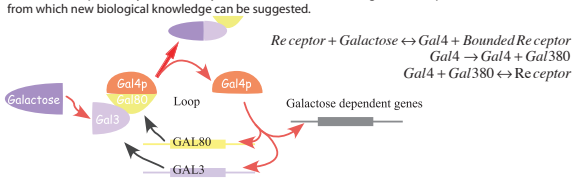
How we predicted number of motifs, assuming known:
- Number of Transfriptional Factors (TF)
- Number of Regulated Gene (RG)
- Number of TRIs (TTN)
- Number of PPI between a TF and a RG (TPS)
- Number of Proteins (PN)

and parameters:
- Motifs parameters Number (PM, SM, TTM, TPM)

$$M_L(s) \underset{\substack{TTM<<TTN \\ TPM<<TPN}}{\approx} \frac{A_{PM}^{PM}}{SM_L}\left(\frac{TTN}{TF \cdot RG}\right)^{TPM}\left(\frac{TPS}{TF \cdot RG}\right)^{TTM}, (SM_L < s)$$

4 Galactose, new dynamical properties

This is an example of a dynamical study of a miss understood heterogeneous loop found, from which new biological knowledge can be suggested.



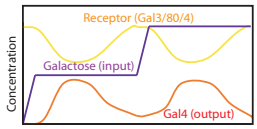
$$Re\,ceptor + Galactose \leftrightarrow Gal4 + Bounded\,Re\,ceptor$$
$$Gal4 \rightarrow Gal4 + Gal380$$
$$Gal4 + Gal380 \leftrightarrow Re\,ceptor$$

Yeast adapts its galactose receptor concentration: when receptor is consumed, new receptor is produced.

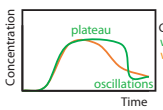
$$\frac{d(Galactose)}{dt} = K_2.CstS.CstR - K_2.CstS.Re\,ceptor - K_2.CstR.Galactose$$

$$\frac{d(Re\,ceptor)}{dt} = (K_2.CstS + 1).CstR - (K_2.CstS + 1).Re\,ceptor - K_2.CstR.Galactose$$

This loop acts as a derivator: to a step input of galactose, Gal4 answer is a pulse.



The loop includes a slow TRI and a quick TRI, but suprisingly, this makes the feedback loop faster due to the slow interaction.



Gal4 with delay for slow TRI without delay