Serge SMIDTAS

mi-Avril à mi-Septembre 2002
Rapport

Modélisation du réseau de régulation transcriptionelle de la levure et étude de ses propriétés structurales

Responsable de stage : Vincent Schachter

Hybrigenics
3-5 impasse Reille
75014 Paris - France

Résumé

L'accumulation de données post-génomiques permet de commencer à appréhender globalement les mécanismes cellulaires de certains organismes modèles. En particulier, plusieurs études à grande échelle tant sur le protéome que sur le transcriptome de *Saccharomyces cerevisiae* ont récemment produit une masse critique d'informations sur ses réseaux de régulation et d'interaction. Les premiers travaux d'analyse ont fourni des résultats théoriques sur la structure topologique globale de ces réseaux. Toutefois pour appréhender réellement leur fonctionnement, le découpage en sous-réseaux ayant des rôles de contrôle et de transmission d'information clairement identifiés semble une étape indispensable.

L'objectif de notre travail est d'étudier les propriétés structurelles de la régulation transcriptionnelle de la levure, à partir d'un réseau plus complet obtenu en intégrant des informations d'interactions entre facteur de transcription et ADN avec d'autres données, notamment d'interactions protéine-protéine.

Cette étude nécessite la définition d'un *modèle formel* permettant la description de ce réseau et la constitution d'une base constituée autour d'un *modèle de données*, intégrant les données nécessaires à sa construction. Ces deux piliers doivent permettre d'aborder des analyses de propriétés structurelles de complexité croissante.

Introduction

Depuis le séquençage complet d'organismes comme la levure, l'ère de ce qu'on appelle le "post-génome" est entamée. De nombreuses expériences à grande échelle sont menées fournissant un très grand nombre de données qui permettent de décrire le réseau global d'interactions de l'organisme. L'étude du transcriptome[20] et du protéome devient un chapitre obligé de la biologie. L'analyse *in silico* du réseau reconstitué d'interaction autorise la détection de mécanismes fonctionnels nouveaux, mais permet également de tester d'anciennes hypothèses biologiques jusqu'ici non vérifiables. L'approche traditionnelle de la biologie moléculaire qui consiste à étudier les propriétés structurelles des gènes et protéines pris individuellement est fondamentale pour comprendre le détail des mécanismes biologiques, mais il ne suffit pas pour appréhender ces nouvelles données, et résoudre des problèmes de biologie, physiologie et pathologie qu'il était impossible d'aborder préalablement.

Une première approche de l'étude de ce réseau a été de s'intéresser à leurs propriétés topologiques globales, comme par exemple la connectivité. Les premiers travaux dans cette voie se sont appuyés sur des cartes d'interactions entre protéines d'une part, et des cartes d'interactions entre facteurs de transcription et gènes d'autre part[1]. Ils ont montré que de tels réseaux sont très hétérogènes avec des propriétés d'échelles particulières en loi de puissance[21]. Cette approche est riche d'informations, en termes d'évolution par exemple, mais ne permet pas une analyse plus fine du réseau ni d'en comprendre le fonctionnement. Pour cela, il faut découper le réseau en modules fonctionnels.
Les premiers travaux très récents, pionniers dans ce domaine, ont permis de mettre en valeur de petits sous réseaux locaux intéressants tant par la fonction qu'ils opèrent dans l'organisme que par leur surreprésentation dans le réseau. Ces motifs sont par exemple des petites boucles de régulation génétique directes, positives ou négatives faisant intervenir 2 à 3 gènes[1] ou encore des voies concurrentes cohérentes ou incohérentes de régulation d'un gène ne faisant intervenir que 3 gènes.[2]
La levure, qui est un organisme modèle, 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 telle analyse. Par ailleurs, c'est un eucaryote modèle très bien décrit dans la littérature.

L'objectif de ce travail est d'étudier ce réseau global de régulation transcriptionnelle de la levure en complétant le réseau d'interactions de facteur de transcription gène[1] par les cartes d'interactions diverses. Ces interactions sont des interactions, entre protéines[7][8][12][13], de complexes[9][10][16], des voies de signalisation [15], ou encore d'expression d'ARN messager. Cette intégration de données diverses est la base de l'étude de propriétés structurelles, intéressantes pour comprendre la dynamique locale et globale du réseau. Une fois l'étude locale de motifs établi, on pourra imaginer dans un second temps tenter de comprendre le type de comportements résultant de l'assemblage de ces motifs constitutifs.
Ce modèle doit permettre une représentation visuelle de ces diverses interactions avec une bonne lisibilité et répondre à des problèmes posés spécifiques par cette fusion de cartes d'interactions. Certaines données sont des interactions physiques entre molécules, d'autres, des données d'interaction entre complexes, ensemble de molécules. D'autres données encore sont des données d'interactions non physiques. Les différentes données traitées sont à des niveaux différents. Certaines données se recouvrent également, comme les données d'expression qui se décomposent en fait en de nombreuses interactions physiques plus élémentaires. Le deuxième problème posé est que les motifs d'intérêts ne font pas intervenir qu'un seul type d'interaction. Ils font appels à plusieurs types d'interactions à des niveaux différents. Il faut être capable de travailler avec ces motifs. Ces problèmes particuliers demandent de travailler avec un modèle approprié à la modélisation de ces données très différentes, et d'avoir un outil approprié à l'étude de motifs composites de différentes interactions. Ce modèle est mis en place dans la première partie.
L'intégration de données de différents types, demande une gestion unique commune de ces données. C'est pourquoi il est nécessaire de recueillir dans un premier temps toutes les données traitées qui se trouvent dans des formats différents afin de les insérer dans une base de données commune. La grande quantité d'information traitée (plus de 50 000 interactions) demande la mise en place et la gestion d'une base de données appropriée avec un modèle relationnel apte à conserver les conditions et résultats expérimentaux des technologies qui ont servi à fournir les interactions introduites, c'est ce qui fait l'objet de la deuxième partie.
L'analyse des propriétés structurelles du réseau de régulation ainsi reconstitué commence par la recherche de motifs remarquables dans le réseau de la levure, et la recherche de motifs fonctionnellement intéressants. . Il s'agit d'un travail en cours présenté dans la troisième partie. Le modèle formel et la base permettra par la suite d'autres études en fonction des résultats obtenus allant de complexité croissante.

1ère partie : Modélisation

L'objectif de ce modèle est d'étendre les cartes existantes de régulations transcriptionnelles en intégrant, en plus de données d'interaction protéine-ADN, d'autres données d'interaction entre molécules, notamment des données d'interaction protéine-prtéine afin d'étudier les propriétés structurelles du réseau obtenu. L'objectif est donc de pouvoir représenter visuellement ces différents types d'interactions, les stocker et les analyser.

L'objectif primaire est de décrire plus finement les cartes de régulations transcriptionnelles génétiques en leur ajoutant des données descriptives supplémentaires composées d'interactions entre protéines.
Cet enrichissement ayant pour but de faire apparaître de nouvelles influences régulatrices et de nouveaux schémas d'influences d'intérêt.
L'objectif secondaire est de pouvoir modéliser des types de données très différentes les unes des autres parfois se recouvrant et d'avoir une intuition de fonctionnement dynamique pour son analyse.
L'objectif tertiaire est de mettre en évidence des schémas dépourvus d'échelle artificielle prédéfinie où des propriétés d'émergence du réseau d'interaction pourront apparaître.

Le modèle doit permettre de représenter les données telles qu'on les connaît, c'est-à-dire des données très incomplètes, plutôt qualitatives, où les rôles ne sont pas toujours prouvés être directs ; permettre une description de plus en plus fine avec la progression des connaissances. Le modèle doit aussi permettre une analyse du réseau, tel que la recherche de motifs plus ou moins définis, la reconnaissance de fonction aux protéines de par leur contexte comme les facteur de transcription, cofacteurs, gènes régulés ou régulateurs.

L'organisme modèle choisi pour une première étude doit être assez simple, mais tout de même intéressant : il doit être bien décrit, avec de nombreuses données aussi bien de double-hybride, de complexes, ou d'expression. De plus, entièrement séquencée, la levure s'impose.

Les premières données exploitées sont par conséquent celles de la levure et de provenance multiple :

- Carte d'interactions double hybride
- Carte d'interactions directes entre facteur de transcription et gènes
- Des données de complexe
- Des données de modèles prédictifs
- Des données introduites à partir de cas particulier de la littérature, comme des voies de signalisation principalement.
- Des données de transcriptome obtenues par des expériences de microarray / puce à ADN.

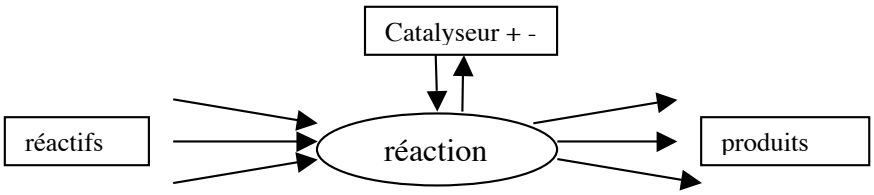
I.1 Modèles existants

Il existe de nombreux modèles qui ont chacun leurs objectifs propres. Certains vont jusqu'à s'attaquer au problème à la représentation et la visualisation de cartes d'interactions de types multiples [4].
La plupart considèrent des interactions binaires entre deux espèces: protéine-protéine, protéine-ADN. Les arêtes sont des interactions éventuellement orientées, et leur nœud des entités biologiques : (protéines, gène…). Leur inconvénient majeur est qu'il est impossible avec de tels modèles de modéliser un complexe (sa formation, sa fonction…) ou tout autre réaction avec plus de 2 partenaires.
Kohn [5] introduit un formalisme, étendu [6], qui permet de travailler avec des complexes, avec une représentation qui est principalement axée sur la représentation spatiale visuelle des données, et complète le peu d'expressivité du modèle par des annotations sur les arêtes.
La visualisation des données est de plus en plus difficile avec l'amoncellement des données en tout genre. Il faut représenter de plus en plus d'informations sur un espace réduit, ce qui aboutit à des cartes touffues peu compréhensibles en leur cœur. C'est en effet un tour de force que de présenter plus de 2358 interactions entre protéines chez la levure sur une page[4]. Une des solutions est de réduire le graphe par exemple en regroupant les protéines d'une classe de fonction particulière[4], mais cette réduction arbitraire fait perdre beaucoup d'informations sur le fonctionnement du réseau d'interaction.
Intégrer dans le modèle les données de transcriptome est également un challenge, car ces données se situent à une autre échelle et sont de plus en plus nombreuses. Un modèle [4] tente une intégration de ces données au moyen de couleurs superposées au graphe, qui est en fait la superposition d'une expérience d'hybridation avec le graphe d'interaction protéine-protéine.
Les modèles appropriés à la description de mécanismes biochimiques sont inaptes à mixer données biochimiques avec des interactions non-physiques comme les données d'expression. Car cela suppose d'avoir différents niveaux prédéfinis pour chaque type d'interaction.

La base de connaissance actuelle fournit de nombreuses données biochimiques d'interaction physique. Ces types de données ont leurs propres modèles adaptés à la représentation de chemins et mettent l'accent sur des réactions de transformation de petites molécules et sur les enzymes qui les catalysent [19]. D'autres types de données d'interaction physique à plus grande échelle apparaissent (Doubles hybrides, complexes,…) avec leurs modèles propres [4]. De même que les données d'expression qui ont leurs propres modèles.
Un méta modèle doit permettre une représentation unifiée de toutes ces données pour leur exploitation commune.

I.2 Modèle

Le réseau de régulation transcriptionnel fait appel à de nombreuses réactions de types différents : liaison à l'ADN, interaction entre protéines, transcription... Toutes ces interactions peuvent être représentées comme des réactions transformant des réactifs en produits.



Elles peuvent représenter aussi bien l'expression d'un gène en protéine, qu'une phosphorilation d'une protéine.
Les produits et les réactifs sont des objets de tous types : gène, protéine, sucre … qui peuvent participer bien entendu à plusieurs réactions.

Ce modèle ou tout est représenté sous forme de réactions et d' espèces est très général et permet décrire tout type d'interaction et de traitement de signaux.
Cependant il n'est pas conçu pour être résolu par un modèle dynamique particulier, mais il est fait pour permettre une analyse topologique fine du réseau statique y compris certaines propriétés dynamiques. Un modèle en réseau de pétri en considérant les réactions comme des transitions, et les espèces comme des places, colle parfaitement comme on va le voir plus loin. Mais il n'est pas question dans un premier temps d'analyser le réseau global dynamiquement comme on le ferait avec un réseau de pétri. Seul le signe des réactifs et produits par exemple est représenté et non les coefficients stoechiométriques des réactions. La constante de réaction intervient par une notion de vitesse qui peut être uniquement qualitative.

I.2.1 Eléments de base du modèle

Toutes les interactions physiques ou d'un point de vue informationnel plus abstrait sont vues comme des réactions entre des espèces dans le modèle.

Les espèces

Les *espèces* du modèle représentent toutes les ressources et les produits des réactions de l'organisme. Une *espèce* représente aussi bien une espèce biochimique dans une localisation donnée que d'autres ressources et produits moins matériels comme une plage de température ou un état environnemental comme un phénotype. Les *espèces* incluent donc toutes les conditions environnementales que l'on souhaite modéliser pour qu'une réaction ait lieu.

Le modèle est conçu pour représenter la compartimentation cellulaire, en introduisant des réactions de transport, et en considérant des espèces chimiques dans des compartiments différents comme des espèces différentes. Une réaction de transport est nécessaire pour transporter une espèce biochimique d'une localisation à une autre, ce qui n'est par conséquent ni plus ni moins qu'une réaction de transformation d'espèce.

Définition : Un objet du modèle est défini par un *Nom*,une *Localisation* et un *Type*.
Le *Nom* est le nom de l'espèce chimique représentée.
La *Localisation* indique le compartiment cellulaire dans lequel on trouve l'espèce modélisée. Par défaut, la localisation est cellulaire.
Le *Type* donne une indication sur la nature de l'espèce biochimique
Les types peuvent ne pas être spécifiées. Ils sont classés hiérarchiquement.

Notation : E(N,L,Type)est une espèce avec un nom,une localisation et un type.
E : ensemble des espèces

Exemple de hiérarchie pour le Type :
Molécule
- Gène
  - Régulateur
  - Régulé
- Protéine
  - Facteur-Transcription
  - Cofacteur
  - Enzyme
- Petite molécule
  - Minéral
  - Sucre

Environnement
- Température
- Phénotype
- État de l'environnement

Représentation graphique : Une espèce est représentée graphiquement par un rectangle. Il est possible de représenter à l'intérieur ses différents paramètres, principalement son nom.

Remarques sur la localisation : La localisation permet uniquement de distinguer deux espèces qui ont le même nom, une hiérarchie peut être ajoutée correspondant à l'imbrication des compartiments cellulaires.

Exemples d'espèces:
E(H , Universel , Atome)
E(Mg2+ , Extracellulaire , Minéral)
E(Adénosine , Cellulaire , Base)
E(RanGTP , Protéine , Nucléaire)
E(YPR158W , Cellulaire , Gène)
E(Apoptose , Cellulaire , Phénotype)

Les réactions

Les transformations et interactions entre objets sont représentées dans le modèle par des réactions. Les réactions représentent les relations qui peuvent se produire entre les objets, comme des assemblages de protéines, ou bien l'expression de gènes.

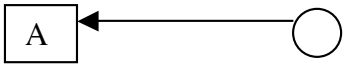
Définition : Une *réaction* est définie par un *identifiant,* une *vitesse* et une *crédibilité.*
L'*identifiant* permet d'identifier les réactions.
La *vitesse* donne une indication sur le temps que met la réaction. Typiquement, la vitesse permet de distinguer des réactions lentes comme la transcription, de réactions rapides, comme l'interaction entre protéine. On pourra choisir une expression qualitative ou quantitative de la vitesse.
D'autres paramètres peuvent être attribuées à la réaction, notamment pour pouvoir retourner à la source expérimentale de sa mise en évidence.
La *crédibilité* qui indique la pertinence de la connaissance effective d'une réaction, et qui sert à distinguer les parties du réseau de régulation qui sont connues de manière certaines de celles sur lesquelles un grand doute subsiste et qui demandent à être vérifiées expérimentalement. Typiquement, la crédibilité d'une interaction produit d'un faisceau d'expériences biochimiques et génétiques sera par exemple plus crédible qu'une seule interaction montrée uniquement au moyen d'expériences de puce à ADN. Elle pourra prendre des valeurs qualitatives ou quantitatives.
La *Source* fait référence aux expériences qui ont permis de mettre en évidence la réaction. C'est la source qui permet de donner un indice de crédibilité à la réaction.
Crédibilité = f(Source)

Notation : R(I,V,Nom,Crédibilité,Source) est une réaction particulière
R : ensemble des réactions.

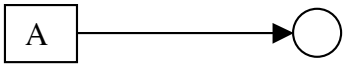
Représentation graphique : Une réaction est représentée graphiquement par un rond ou un point. Il est possible de représenter à l'intérieur ses différents paramètres, principalement sa vitesse.

Les Arêtes

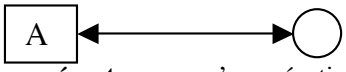
Les arêtes symbolisent les liens qui existent entre espèces et réaction.
Définition: Les arêtes A(R(),E(),p) sont des triplés constitués d'une réaction R(), d'une espèce E() et d'un paramétrés p.
Les arêtes du graphe relient une réaction aux espèces, et ne peuvent donc pas joindre deux réactions entre elles, ni deux espèces directement entre elles.
Le *Paramètre* peut prendre 4 valeurs, et colorer ainsi différement l'arête.
Lorsqu'une réaction **produit une** espèce, on représente l'arête par une flêche pointant de la réaction vers l'espèce.



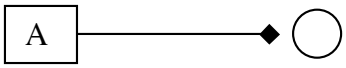
Lorsqu'une réaction **consomme** une espèce, on représente l'arête par une flêche pointant de l'espèce vers la réaction.



Lorsqu'une réaction consomme une espèce et la produit à la fois, on représente l'arête par un double flêche, pointant à la fois vers l'espèce et vers la réaction. Typiquement c'est le cas des catalyseurs. La signification est que l'espèce est **nécessaire** à la réaction.

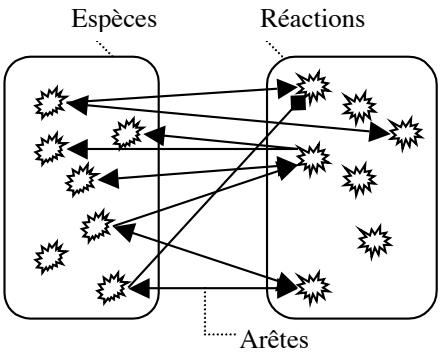


Lorsqu'une espèce **ne doit pas être présente** pour qu'une réaction ait lieu, on représente l'arête par une flèche avec un bout droit pointant vers la réaction. C'est une inhibition.



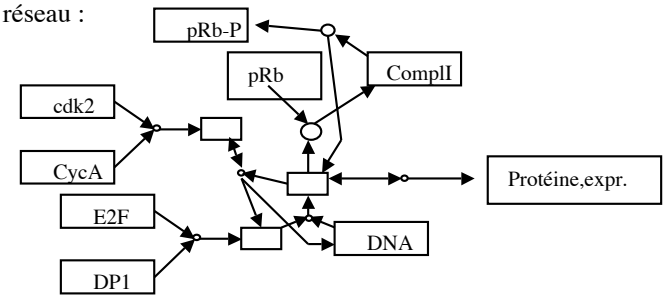
Règle : Lorsque 2 arêtes fléchées de sens opposés rejoignent le même couple (Espèce(),Réaction()), le graphe peut être allégé en représentant une seule arête avec double flèche. Voir plus loin une interprétation en terme de motifs.

L'ensemble des arêtes est noté A.



Réseaux

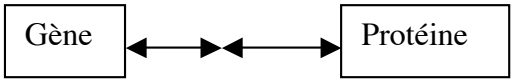
Définition : Le réseau est un ensemble d'espèces, de réactions et d'arêtes : (E,R,A)
Exemple de réseau :



A.2.2 Motifs

Un motif est un sous-ensemble des l'ensembles des sous reseau du reseau tout entier. Un motif peut être vu comme un sous-réseau dont les objets qui le constituent ne sont pas parfaitement définis. Formellement, il s'agit d'une classe d'équivalence sur les sous-réseaux.
Un motif est défini par les propriétés structurelles que doivent partager les sous-réseaux pour appartenir à la même classe.

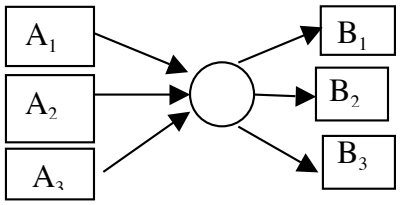
Exemple:
{ (R1,E(,,Gène),active),(R1,E(,,Protéine),produit) } est un motif représenté ci-dessous.



I.3 Discussion

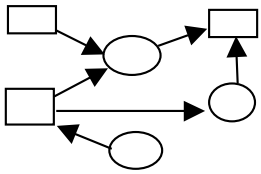
Réactions biochimiques

Les réactions biochimiques se modélisent intuitivement très bien par le modèle. Les réactifs et les produits sont représentées par des espèces, dont le rôle dans la réaction est modélisée par le paramètre des arêtes.
Ainsi une réaction écrite sous la forme ∑ai.Ai -> ∑bi.Bi se modélise en faisant intervenir les coefficients stoechiométriques que par leur signe, c'est à dire en distinguant les produits des réactifs. Ce qui donne:



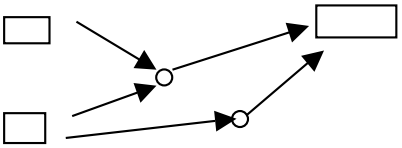
Structures mathématiques semblables

Graphe
Les réactions et les espèces peuvent êtres vues comme des nœuds du graphe. Les arêtes relient ces 2 types de nœud ce sont des couples d'une espèce et d'une réaction. Les arêtes sont de 4 types différents.



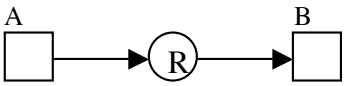
Hypergraphe
Les réactions avec l'ensemble des arêtes qui les entourent peuvent être perçues comme des relations n-aires admettant 4 types de rôles :

- Entrée (ressource pour la réaction)
- Sortie (produit de la réaction)
- Nécessité (nécessaire à la réaction)
- Interdiction (interdit la réaction)

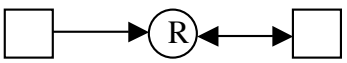


Le modèle comporte des espèces, comme plus haut, et des réactions qui sont des ensembles de couples de paramètre d'arête et d'espèces. $\{ (E,p)_i \}$

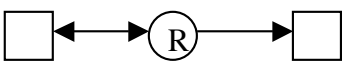
Interprétations de relation binaires autour d'une réaction

Elément de réseau	de gauche à droite (Rôle de A sur B)	de droite à gauche (Rôle de B sur A)
	Transformation	X

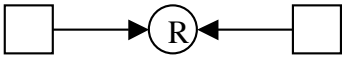
La réaction consomme A et produit B.

	Dégradation catalysée	Régulation négative

La réaction qui consomme A a besoin de B comme catalyseur.
En présence de peu de B, A est consommée. En absence de B, A n'est pas consommé.

	Formation catalysée	X

B est produit, et catalysé par A

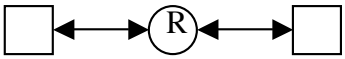
	Dégradation coopérative ou régulation négative

A et B sont consommée au cours de la réaction.
En présence de A, B est consommé. En présence de B, A est consommé.

En absence de A, B est inchangée, En absence de B, A est inchangée.

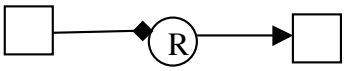
Co Formation

A et B sont produits par la même réaction.



interaction

A et B interagissent ensemble. L'expressivité est très pauvre ici.

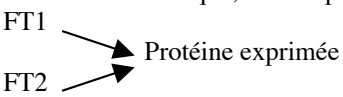


Inhibition de réaction

On a plusieurs cas de régulations négatives. Une symétrique, une asymétrique, et une inhibition.
La régulation négative peut avoir lieu par réaction concurrente comme on va le voir plus loin, ou par dégradation d'une espèce, comme on vient de le voir.
Dans le cas symétrique de la régulation négative, A et B jouent des rôles identiques l'un sur l'autre. C'est-à-dire que A régule négativement B. Et B régule négativement A. La symétrie n'est levée que par le contexte, c'est-à-dire les réactions qui sont autour. Dans le cas asymétrique, B a également un rôle de catalyseur dans le sens où il est nécessaire à la réaction mais n'est pas consommé, ce qui brise la symétrie.
L'inhibition de réaction est un peu différente, dans le sens ou l'inhibition joue sur la réaction toute entière et pas seulement sur la régulation de la formation d'une espèce.

Sur la complexité des arêtes, comparaison avec un modèle de graphe.

Dans les modèles de graphe, l'arête exprime la liaison directement entre deux entités biologiques. Tous ces modèles ont par construction l'inconvénient de n'exprimer que des 'liaisons' binaires. Il ne suffit certes que d'une 'flèche' pour dire qu'un gène code pour une protéine, la où il nous en faut 2 (on est en effet obligé de faire intervenir une réaction entre le gène et la protéine), mais cette complexité apparente est bien plus expressive et laisse plus de liberté. En effet, lorsque les réactions ne sont que des arêtes, il faut distinguer un grand nombre de flèches qui peuvent avoir pour signification : « code pour une protéine », « interaction physique », « transforme chimiquement une molécule », « se lie avec une molécule»… pour toutes les réactions possibles.
La fonction même est plus expressive. Par exemple, Deux facteurs de transcription permettent d'exprimer un gène. Avec un modèle binaire classique, on le représenterait ainsi :

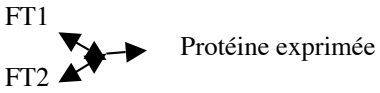


Avec un modèle avec réactions explicites on le représente comme ceci :

Prolongeons l'exemple, en fait pour être exprimé, le gène a besoin de FT1 et FT2, par exemple parce qu'ils se dimérisent.
Avec un modèle binaire on représente:

(*) avec une annotation quelque part.

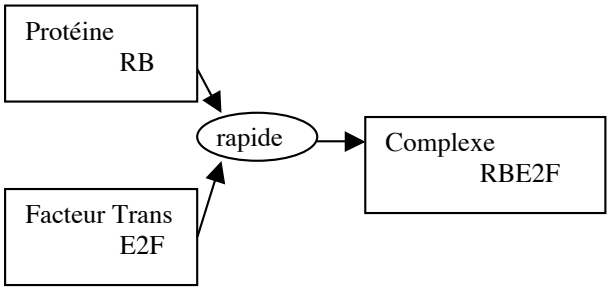
Avec un modèle avec réactions explicites , cela donne:



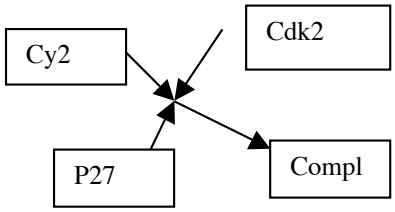
On arrive donc à distinguer les fonctions logiques : le ET du OU.

Exemples

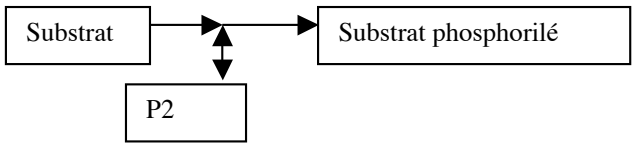
Interaction protéine-protéine
RB et E2F interagissent pour donner un complexe
On peut indiquer graphiquement
la vitesse de la réaction et le type des espèces.



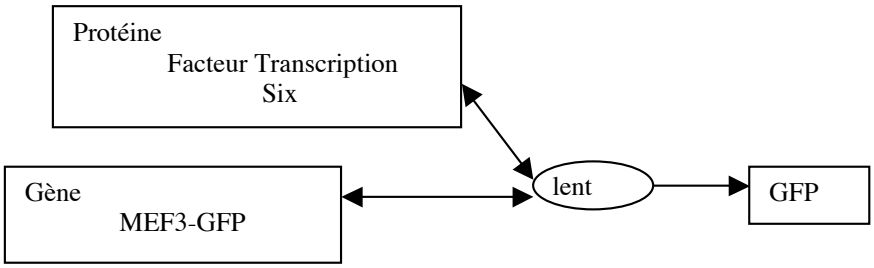
Formation de complexes
Le complexe est formé d'une kinase,d'une cycline et de P27. On ne représente pas l'ordre de formation du complexe. Il est directement le produit des 3 molécules.



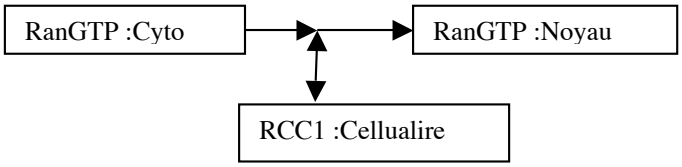
Réaction enzymatique
Le substrat est phosphorylé par l'enzyme. Chaque état (phosphorylation, méthylation …) d'une espèce biochimique est représenté par un espèce distincte.



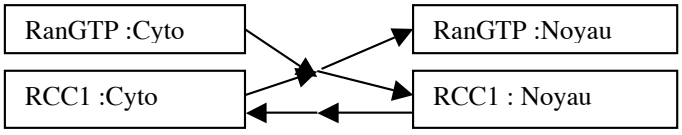
Régulation
Un facteur de transcription, Six, se lie à la boite MEF3 d'un gène codant pour la GFP, il en résulte la GFP. Au cours de la réaction, ni le gène, ni le facteur de transcription n'est consommé. Mais ils sont bien entendu nécessaires.



Transport
RanGDP peut être transporté du cytoplame au noyau une fois phosphaté par le transporteur RCC1. On distingue RanGTP dans le noyau et dans le cytoplasme par l'indication de sa localisation. Le transporteur a pour localisation la cellule toute entière car on n'a pas choisi de modéliser son transport, et en particulier à son retour dans le cytoplasme.



Si on veut représenter le cycle du transporteur, on modélise ainsi :



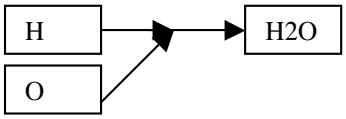
Dans ce cas, le détail du mécanisme de retour dans le cytoplasme de RCC1 n'est pas détaillé.

La localisation ne doit intervenir que lorsque les espèces ont des propriétés différentes suivant leur localisation. Un gène par exemple qui n'a de fonction que dans le noyau n'aura qu'une localisation. Il est inutile a priori de distinguer le gène nucléaire du gène cytoplasmique par exemple.

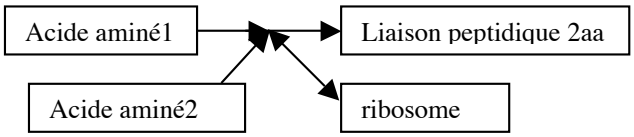
Différentes échelles de représentation

Le modèle permet de représenter des réactions à toute échelle, de la plus petite, comme les réactions élémentaires, aux plus complexes, comme un phénotype.

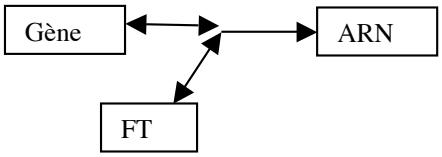
Réaction élémentaire:



Réaction simple : comportant peu de réactions élémentaires:

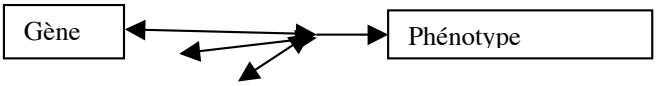


Réaction complexe :



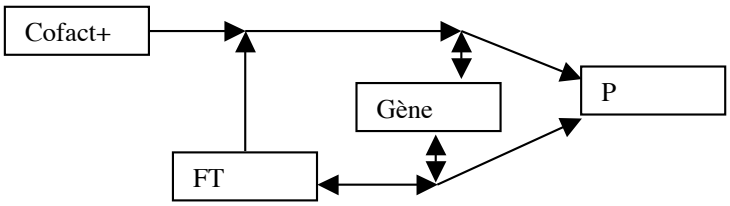
Réaction information

Phénotype



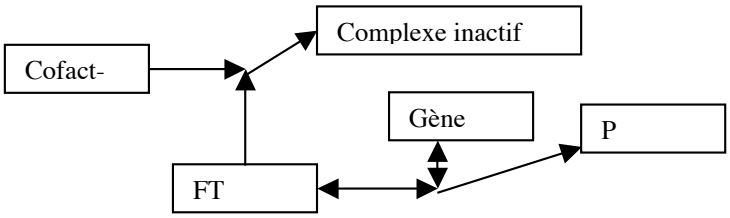
Co facteur activateur

Avec hypothèse que la voie d'activation par le cofacteur est plus rapide qu'avec le Facteur de transcription seul.



Ce motif s'interprete ainsi :grâce au cofacteur, il existe un 2ème chemin pour produire la protéine P à partir du gène et du facteur de transcription. Plus la voie de formation de P par le cofacteur est rapide, plus le cofacteur sera actif.

Exemple de Cofacteur inhibiteur



Le Cofacteur consomme les ressources nécessaires à la production de la protéine P. Ici le cofacteur utilise FT dans une autre réaction qui ne sert pas à produire P. Il s'agit d'une inhibition par réaction concurrente.

I.4 Dynamique

Le modèle est conçu pour être étendu par la définition d'une dynamique, c'est-à-dire avec des états du système et des règles de transition. L'objectif est ici d'en étendre son analyse pour l'étude de phénomènes biologiques. Il est important de rappeler que les résultats trouvés ne doivent pas dépendre du modèle choisi pour représenter la dynamique. Il n'est pas question ici de simuler entièrement le comportement du réseau. En effet, la simulation demande à avoir une bonne connaissance de tous les paramètres. Le modèle dynamique est conçu ici comme un moyen d'analyse plus formel pour trouver l'influence d'une espèce sur le réseau, trouver des chemins concurrents, ou encore étudier de manière locale le fonctionnement de sous réseaux. Il est plus question de montrer des fonctionnements possibles qualitativement que d'établir une dynamique quantitative précise de l'organisme.

La compréhension du traitement de l'information de l'organisme est très liée à sa dynamique. Le nombre immense de molécules dans une cellule rend difficile voire impossible sa simulation globale. De nombreuses propriétés dynamiques du réseau peuvent êtres déterminées sans avoir à faire tourner le modèle dynamiquement. Une étude statique du réseau permet déjà de fournir beaucoup d'information, comme la présence ou non de boucles rétroactives comme on va le voir plus loin. Il existe tout un panel de modèles du continu au discret en passant par le semi quantitatif [17] ou le semi-discret qui ont l'ambition de simuler l'intégralité de l'évolution dans le temps des espèces considérées. D'autre part, il existe des modèles basés sur des règles d'évolution et sur l'algèbre des processus qui permettent de montrer certaines propriétés dynamiques de manière statique. Tous ces modèles dynamiques peuvent s'appliquer au modèle décrit, comme les exemples présentés plus haut, et retrouver leur fonction.
Une solution pour comprendre le mécanisme global consiste à tirer des renseignements sur le comportement d'une partie du réseau obtenu à une échelle détaillée pour les faire passer à une échelle supérieure en ne conservant que les caractéristiques principales. Pour cela, on peut reconnaître des motifs de manière statique ou dynamique. Pour le cas dynamique, un modèle en réseau de pétri s'adapte très bien au modèle, mais d'autres modèles dynamiques peuvent également être utilisés. Il existe de nombreuses variantes autour des réseaux de pétri appliqués à la biologie alliant temporalité, stochasticité, et même mélange de continu et discret [14].
Les *places* sont identifiées aux *espèces* du modèle, Les *transitions* sont les *réactions*. Les 4 types d'arêtes consommation, production, activation, inhibition deviennent respectivement input des transitions, output, input et output, et un input particulier d'inhibition qui empêche la transition d'avoir lieu. La vitesse est modélisée par un retard associé à chaque transition.
Il est possible d'introduire un *marquage* sur les exemples présentés plus haut, et d'étudier l'évolution des jetons. Des équations différentielles permettent d'obtenir les même fonctions.

I.5 Réduction Expansion Renormalisation

Le modèle doit permettre de ne pas favoriser une échelle d'interaction, c'est-à-dire qu'il doit permettre de travailler avec un ensemble d'informations à plusieur niveaux de détail : les réactions qui peuvent intervenir entre des espèces, peuvent être simples chimiquement, ou être composées de nombreuses étapes intermédiaires, voire font l'état de processus longs et peu reproductibles
Il n'est pas concevable de considérer que le traitement d'information effectué par un organisme ne soit restreint à une échelle donnée préalablement établie : Il ne faut pas étudier que les interactions des protéines entre elles et oublier les connaissances biochimiques. Il ne faut pas étudier les réseaux d'expression de gènes et oublier les étapes de leur expression qui peuvent jouer un rôle important. La construction du réseau n'est pas modulaire par construction, mais le résultat d'une évolution sélective longue et surtout aléatoire. C'est pourquoi le traitement de l'information ne peut pas être analysé uniquement à une couche donnée.
Le modèle doit être accompagné d'outils pour permettre la mise en valeur du traitement de l'information lui même, exprimé au travers de sa carte, et cela sans privilégier un niveau.

Par exemple, dans un réseau de régulation génétique, une interaction connue entre protéines non modélisées peut faire apparaître de nouvelles fonctions comme la connectivité entre différentes voies. Interaction catalysée elle-même par une enzyme phosphorylée et qui nécessite de l'énergie. Le rôle de la phosphorylation est directement corrélé à l'étude du réseau de régulation génétique puisqu'elle fait intervenir d'autres molécules, d'autres protéines, et d'autres gènes et permet ainsi d'expliquer le comportement d'ensemble.

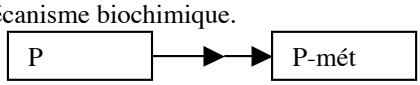
Lorsque l'on regarde le réseau de très loin, seules certaines interactions entre gènes sont visibles faisant intervenir des entrées ou sorties du système. Lorsque l'on zoom, plus de gènes ayant une interaction avec le gène sur lequel on zoom apparaissent, puis les interactions entre protéines, les réactions catalysées, biochimiques autour du point sur lequel on zoom. Ce modèle permet de représenter visuellement un nombre constant maximal d'information sur une surface donnée, en zoomant sur des points particuliers, en laissant accessibles les différentes échelles de données.

L'avantage est de pouvoir intégrer des données de différents types, de différentes échelles, de pouvoir les représenter. Les données du transcriptome comparée aux données d'interactions physiques sont à une autre échelle.

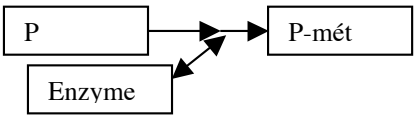
Qu'est ce que l'échelle ?
On veut classer intuitivement les réactions du modèle, en les ordonnant, des réactions simples chimiquement, rapides, faisant intervenir de petites espèces, aux réactions complexes, lentes faisant intervenir de nombreux états intermédiaires, avec des espèces complexes. Il s'agit de munir l'ensemble des réactions d'un ordre partiel permettant de retrouver le classement intuitif. Une réaction A sera dite à une échelle plus élevée qu'une B si elle fait intervenir dans son mécanisme la réaction B. Par exemple l'expression d'un gène fait intervenir la traduction, qui elle même fait intervenir la formation de liaisons peptidiques. Cela signifie qu'il est ainsi impossible de comparer le niveau de deux réactions totalement distinctes.
La renormalisation est un outil pour l'exploration du graphe, et jouer entre les différentes échelles.

I.5.1 Expansion

L'expansion permet de faire face à la précision des informations que l'on a sur un organisme.
Au fur et à mesure que l'on acquiert des connaissances sur les mécanismes cellulaires, on veut les rajouter au modèle. Par exemple si l'on trouve deux protéines identiques dans un organisme, dont une est sous sa forme méthylée, on peut se douter qu'il existe une réaction qui transforme sa forme déméthylée en version méthylée, sans pour autant connaître le mécanisme biochimique.



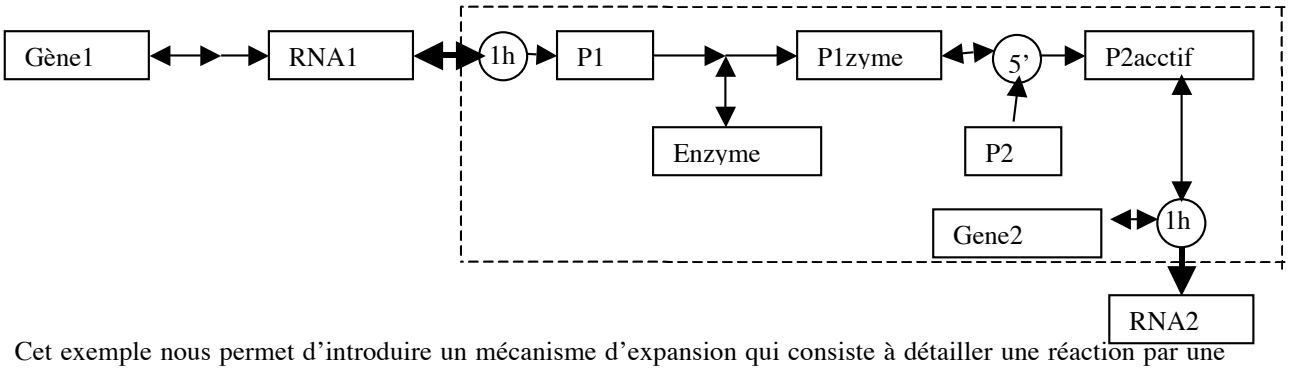
Lorsque les connaissances nous apprennent comment on passe d'une forme à l'autre, il est intéressant de rajouter ces informations au modèle. Sur notre exemple, imaginons que l'on découvre un jour une enzyme qui effectue cette méthylation mystérieuse. Il suffit de corriger le modèle localement par expansion.



Le modèle permet de représenter toutes ces réactions avec un niveau de détail plus ou moins fin, soit en filtrant sur le type des molécules qui interviennent dans les réactions, par exemple en ne faisant pas apparaître les consommation d'énergie, ou bien même des enzymes.

Introduction de données d'expression
Les données d'expression produites à grande échelle, permettent de suivre dans le temps l'expression de gènes par mesure de leur ARN. Ces données temporelles permettent de corréler temporellement l'expression des gènes. Voici ci-dessous sur un exemple, la façon de prendre en compte ces résultats de corrélation.

On a peu d'information sur le lien qui existe entre la présence d'RNA1 et celle RNA2 3h après à priori. Lien qu'on pourra détailler par la suite par exemple par :

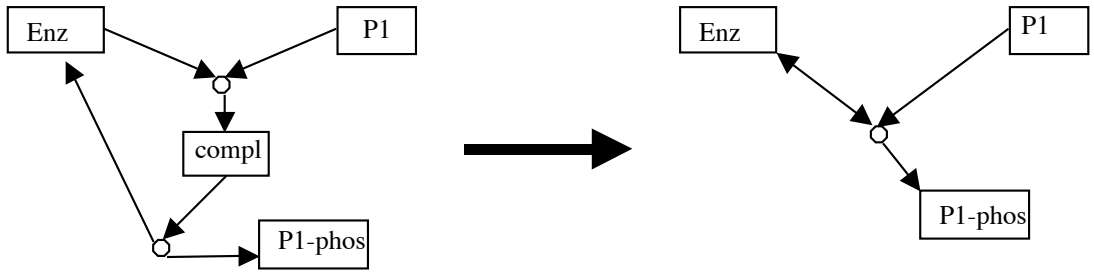


Cet exemple nous permet d'introduire un mécanisme d'expansion qui consiste à détailler une réaction par une partie d'un graphe, comportant des réactions plus élémentaires et des intermédiaires de réaction qui n'apparaissaient pas au niveau supérieur.
Avec le même processus, une réduction est possible. La réduction peut être perçue comme une renormalisation, avec toutes les connotations de celle-ci : pourra servir à faire apparaître de nouvelles propriétés.

L'analyse du réseau permet de chercher et proposer un mécanisme compatible candidat à une réaction globale, comme une donnée de transcriptome et ensuite de ne conserver l'une ou l'autre des visions de la réaction suivant le degré de finesse que l'on souhaite. Cette proposition de mécanisme candidat peut s'effectuer systématiquement pour toutes les réactions, en cherchant les chemins compatibles les plus courts. Ils doivent être pour cela compatibles en vitesse et par la nature de l'utilisation des espèces dans la réaction étudiée(consommation, production, catalyse, inhibition).

I.5.2 Réduction Renormalisation

Exemple : On sait que pour que P1 soit phosphorilé par l'enzyme, il y a formation d'un complexe. On représente donc ce mécanisme comme indiqué ci-dessous à gauche. Toutefois, le complexe intermédiaire transitoire ne participe à aucune autre réaction. Il est par conséquent souhaitable de ne pas le faire figurer au niveau de description choisi. Il est donc intéressant d'avoir une procédure de renormalisation pour passer de la représentation à gauche, à celle de droite épurée.



La question est ici de définir une renormalisation, c'est-à-dire d'expliciter la manière de court-circuiter une suite de réactions en une plus globale reprenant les caractéristiques que l'on veut conserver ou exprimer à une échelle supérieure du détail de réactions.

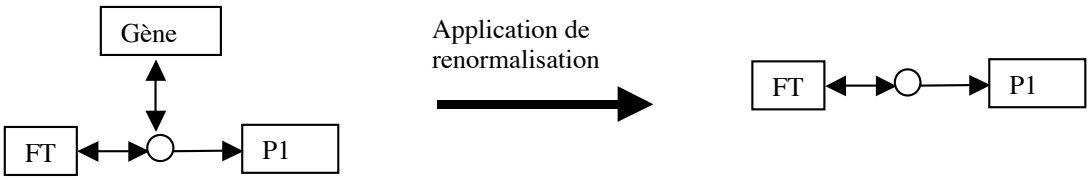
La renormalisation qui consiste à remplacer une portion du graphe par une plus petite avec de nouvelles liaisons a un grand intérêt pour simplifier le réseau, et en comprendre le fonctionnement à plus grande échelle. Le problème consiste à remplacer une portion du graphe par 'une boîte noire' dont il faut trouver les bonnes caractéristiques. Une 'boîte noire' peut recouvrir plusieurs réactions, pour ne conserver systématiquement que les entrées sorties de cette boîte, ou plus arbitrairement les principales entrées sorties seulement en négligeant par exemple les petites molécules ubiquitaires qui interviendraient également.

Définition : Une *renormalisation* est un ensemble d'applications {Application$_i$} d'un ensemble de motifs dits renormalisables, vers un ensemble de motifs dits renormalisés. La dite application est un couple de motifs Application=(motif,motif), le premier étant le motif de départ, le second, le motif d'arrivée.

Exemple d'application
Motif de départ: { (R1,E(,,Gène),nécessaire),(R1,E(,,FT),nécessaire),(R1,E(,,protéine),produit) }
Motif d'arrivée: { (R1,E(,,protéine),FT),(R1,E(,,protéine),produit) }

Illustration:

({ (R1,E(,,Gène),nécessaire)**,**(R1,E(,,FT),nécessaire)**,**(R1,E(,,protéine),produit)**}**
,{ (R1,E(,,protéine),FT)**,**(R1,E(,,protéine),produit)**}**)



Une renormalisation consiste par conséquent à rechercher un motif, et à en remplacer tous leurs représentants par leur motifs associés plus simples, c'est à dire moins détaillé.
Plusieurs renormalisations peuvent être définies sur un réseau. Par exemple une renormalisation peut consister à retirer tous les intermédiaires réactionnels qui n'ont pas de rôle dans d'autres réactions et à fusionner les 2 réactions qui en faisaient usage, comme le montre l'exemple plus haut. Ou bien encore fusionner des voies parallèles. D'autres motifs plus complexes peuvent servir à cette renormalisation.

2ème partie : Données

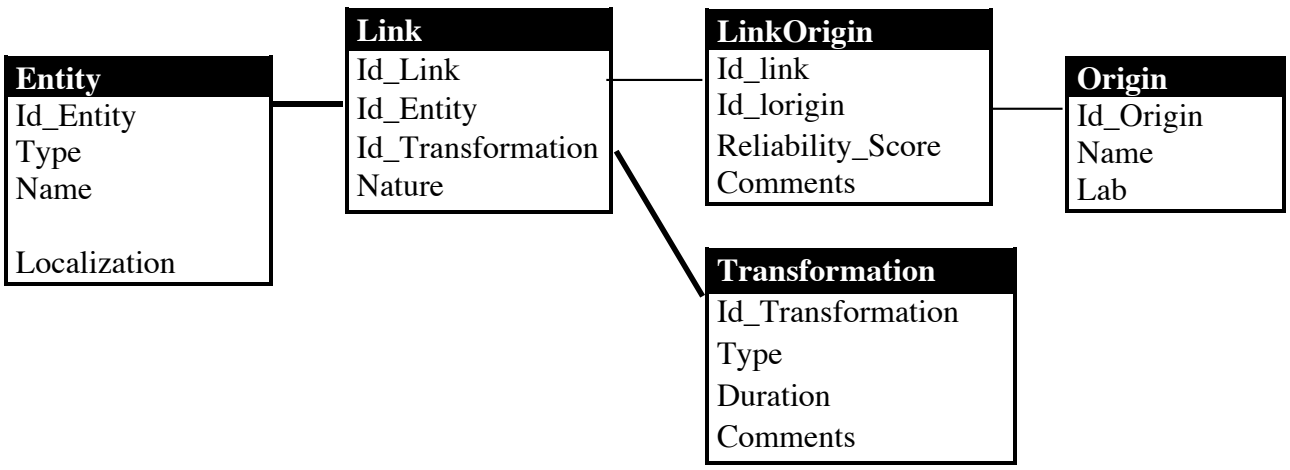
II.1 Base de Données

Le Stockage de données est une étape importante dans le traitement de la grande quantité d'information.
La base de données doit pouvoir accueillir tout types d'interactions entre espèces, produite par chacune des technologies modélisées. Elle doit permettre de conserver une tracabilité des données entrées, c'est-à-dire leur provenance afin de pouvoir y revenir à tout moment.

Le format de base de données choisi est mySQL pour sa robustesse éprouvée depuis de nombreuses années, sa simplicité suffisante pour les requêtes que l'on désire y effectuer, et son accessibilité (licence libre). De plus de nombreux outils du domaine public existent autour de mySQL. Toutefois, lors de l'exploitation de la base, il s'est déjà avéré que la vitesse était un point limitant de son utilisation. Le portage sur une autre base de donnée serait toutefois lourd à ce stade car demanderait de réécrire de nombreux scripts. La portabilité

Les scripts d'utilisation de la base de donnée sont écrits en PHP, langage choisi pour sa simplicité et rapidité d'écriture, sa bonne interface avec mySQL, et son cout puisqu'il est également du doamine public, mais qui montre également des signes de lenteurs qui lui provienne du fait qu'il ait été développé pour l'écriture de serveur web et non pour le calcul.

Le réseau décomposé au niveau élémentaire comporte des espèces, des réactions et des arêtes typées. Il sera commode par conséquent d'enregistrer les données du réseau dans une base de donnée comportant principalement 3 tables : une liste d'espèces, une liste de réactions et une liste d'arêtes.
À cela il faut ajouter d'autres tables d'annotation, principalement pour retrouver l'origine expérimentale d'une donnée de la base.
Voici schématisé les principales tables de la base suivi d'une description des différentes tables. Voir document de spécifications des tables pour leur définition formelles.



transformation_type
Cette table liste l'intégralité des types de transformations.
Voici à titre indicatif la liste utilisée à ce jour de ces types de transformation. (Coexpression | Complex | Direct_transcriptional_regulation | Phenotype_production | Physical_interaction)
Champs : type

transformation
Cette table contient la liste des réactions avec principalement leur type.

Champs :		
	id_transformation	id unique de la transformation
	type	type appartemenant à transformation_type
	duration	temps que mets la transformation
	composite_score	indice de la crédibilité de la transformation
	comments	commentaires

localization
Cette table doit permettre de distinguer des espèces localisées dans des compartiments cellulaires distincts. Elle liste l'ensemble de ces compartiments, ou région cellulaires.

Champs :		
	localization	nom du compartiment cellulaire

name
Cette table gère les synonymes des noms d'éspèce que l'on peut avoir. La gestion de listes de synonymes est importante. Lorsque les données sont introduites à partir d'autres bases de données expérimentales, le nom systématique n'est pas toujours utilisé. De plus il peut être plus agréable de visualiser les noms d'usage courrant, plus évocateurs et mnémotechniques que le nom systématique.

Champs :		
	systematic_name	nom systématique utilisé dans la base
	synonym	synonyme trouvé dans la littérature

entity_type
Tout comme transformation type, cette table contient les type d'entité (Complex| Phenotype | Protein)

Champs :		
	type	type d'entité

entity
Cette table contient la liste des entités avec principalement leur type et leur nom.

Champs :		
	id_entity	id unique
	type	type d'entity_type
	name	nom 'systématique' de l'entité s'il existe
	localization	référence à la table localization
	comments	commentaires

link
Cette table contient la liste de tous les liens entre entité et transformation ainsi que la nature du lien (produces,consumes,needs,isinhibitedby).

Champs :		
	id_link	id unique du lien reliant une
	id_entity	entité
	id_transformation	à une transformation
	nature	nature du lien

origin
Cette table assure la sauvegarde de l'origine expérimentable de tout lien.

Champs :		
	id_origin	id unique
	name	nom de l'origine
	lab	laboratoire
	comments	commentaires
	reference	référence de la publication
	date	date de publication

linkorigin
Cette table assure uniquement l'attribution d'une ou plusieurs origines à chaque lien.

Champs :		
	id_link	lien
id_origin	origine	

reliability_score indice de crédibilité fourni expérimentalement
comments commentaires

transformation_hierarchy
Cette table définit un ordre partiel dans l'ensemble des transformations
id_transformation_child l'enfant d'une transformation dans la hierarchie
id_transformation_parent le parent d'une transformation dans la hierarchie
comments commentaires

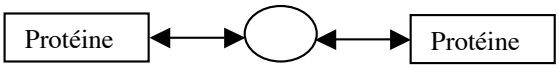
entity_hierarchy
identique à transformation_hierarchy mais appliqué aux entités.
id_entity_child
id_entity_parent
comments

link_hierarchy
identique à transformation pour les liens
id_link_child
id_link_parent
comments

II.2 Données modélisées

Les données introduites dans la base proviennent de différentes technologies expérimentales. Après avoir parsé les données, voici sous quelle forme elles sont modélisées.

Double hybride



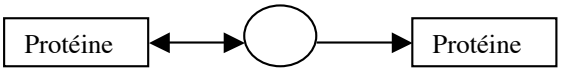
- Hybrigenics[12]

Il s'agit d'une partie des données d'Hybrigenics sur la levure. Ce sont les seules données propriétaires de la base.

- Utz Curagen[4]

Les doonées publics d'interaction double hybride publiées par Utz et al.

Régulation



- Kepes [1]

Compilation de données d'interaction entre Gènes et Facteur de transcription par Guelzim et al à partir de la littérature.

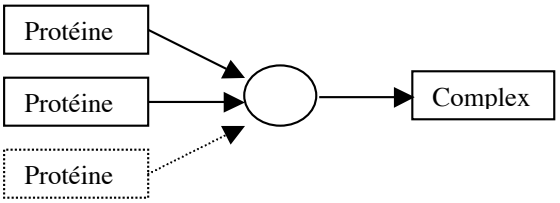
Complexes

- Ho [9]

Données publiées des complexes de la levure obtenues par spectrométrie de masse

- TAP-HMS-PCI [8]

Sélection de données commune de Ho et Cellzome, ces données ne sont plus utilisées.



- MIPS MDSP

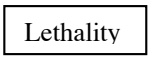
Données compilées de la littérature fournies par MIPS.

- Cellzome [16]

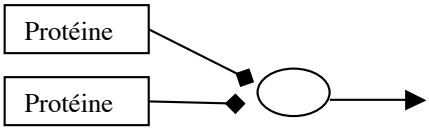
Données filtrées 'S1' publiques de Cellzome.

Synthetic Lethality [8]

Données produites par Mering et al 02



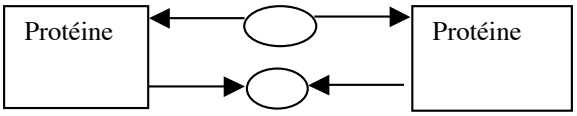
à partir de données de MIPS



Synexpression

- Affymetrix [8]

Données produites par Mering et al 02
à partir de données de microarray
obtenues avec dans plus de 300 conditions différentes ou au cours du cycle cellulaire Deux gènes sont alors dits synexprimés si ils ont un profil semblable d'expression au cours de ces expériences.



L'introduction dans la base est alors assez transparente comme indiqué sur la droite, la crédibilité de chacune de ces réactions est fonction des paramètres fournis par chacun des laboratoires, et de la technologie. Tous les champs prévus dans la base ne sont pas instanciés. Seuls ceux qui ont pu être remplis automatiquement à partir des données publiées sont effectivement instanciés. Cela est suffisant pour une première étude mais une ontologie sera nécessaire pour une étude plus approfondie.

La base de donnée contient avec toutes les données ci-dessus citées introduites :

Espèces :	7337	dont ~2000 complexes
Réactions :	42892	dont ~30000 de synexpression
Arêtes :	98784	

3ème partie : Analyse

Cette partie d'analyse s'appuie sur le *modèle formel,* et le *modèle de données* construits. Elle est encore très incomplète, et sera le sujet de toute mon attention dans le mois à venir.

III.1 Analyse préliminaire des données, recouvrement des données

Voici quelques données relatives aux données de la base servant aux calculs statistiques suivants.
Nombre de transformations par type :

Direct_transcriptional_regulation	938
Physical_interaction	8 531
Complexe	1 440
Coexpression	16 496
Phenotype_production	886

Nombre de couples de protéines participant à un complexe : 69 651
Nombre de couples de protéines participant à une interaction : 96 502
Nombre de protéines : 5655

Recouvrement

L'analyse est principalement orientée autour des gènes régulateurs [1]. Ces gènes participent pour leur très grande majorité à d'autres interactions, comme nous allons le voir.

Données sur lesquelles porte l'étude de recouvrement :

1 **Guelzim** : réactions de régulation transcriptionelle directe de Guelzim et al
2 **HGX** : Interactions double hybride d'Hybrigenics sur la levure.
3 **Curagen Uetz**: Interactions double hybride de Curagen
4 **Ito**: Interactions double hybride de Ito et al
5 **Ho**: données de complexes de Ho et al
6 **Mips complex**: Complexes de « plus basse hiérarchie » de Mips
7 **Cellzome**: Résultats bruts des purifications de Cellzome
Dont 40 petits complexes apparaissent en double car trouvés avec des 'baits' différents.
8 **Syn Lethality**: Interactions de Synthetic lethality rapportées par Mering et al [Nature May 2002]
9 **Synexpression**: Interactions de Synexpression calculées par Mering et al [Nature May 2002]

Ces ensembles sont également regroupés en 3 catégories plus larges :
10 **Y2H**: Ensemble des réactions des 3 types d'interactions double hybride.
11 **Compl** : Ensemble des réactions des 3 types de complexes
12 **Not Guelzim**: Ensemble des réactions de tout type sauf les réactions de Guelzim.

Afin de synthétiser et formaliser la présentation des analyses, quelques définitions sont nécessaires :

α représente une technologie ou un ensemble de technologies
Ici, α prendra une des 12 valeurs citées ci-dessus.

G : désigne l'ensemble des protéines de la levure
Pour cette analyse, G comprend exactement les ORFs présents dans la base de donnée. (C'est-à-dire participants au moins à une des 9 interactions précitées. $|G|=5655$. Ce choix est arbitraire mais la valeur de la constante $|G|$ n'a pas en soi un grande d'importance)

H : désigne un sous-ensemble de G.
Dans cette analyse, H prend successivement les valeurs suivantes :
- l'ensemble des gènes régulateurs ou régulés (suivant Guelzim et al)
- l'ensemble des gènes régulateurs suivant (suivant Guelzim et al)
- l'ensemble des gènes régulés suivant (suivant Guelzim et al)

L : représente l'ensemble des liens présents dans la base
(au sens du modèle abstrait de données, cf le rappel plus haut ; un lien est toujours relié à exactement une entité et une transformation)

L(H, α) : est le multiensemble des liens de type α reliés à un gène de H.
L(H, α)={ l | l ∈ L / type(l)= α & ∃ g ∈ H / entité(l)=g l }
Remarques :
- une interaction biologique entre deux entités de H apparaît deux fois dans l'ensemble, via les deux liens de part et d'autre de la transformation.
- L(G,*)=L

CG(H, α) : est l'ensemble des gènes de H possédant au moins un lien de type α.
CG(H, α)= {g ∈ H / ∃ l ∈ L , entité(l)=g & type(l)= α}
Si un gène appartient à CG(H, α), on dira qu'il est α-connecté.

Définition des ratios

Dans la suite, les ensembles et leur cardinalité seront confondus pour alléger les notations.
Sur ces cardinalité d'ensemble, 6 ratios sont définis :

L(G, α) / G
Nombre moyen de liens de type α par gène de la levure.

L(H, α) / H
Nombre de liens de type α par gène de H.

(L(H, α) / H) / (L(G, α) / G)
Surconnectivité (de type α) des gènes de H par rapport ceux de G.

CG(H, α) / H
Proportion des gènes α-connectés parmi les gènes de H.

L(H, α) / CG(H, α)
Nombre de connexions de type α pour les gènes de H rapporté au nombre de gènes ayant au moins une connexion de type α.
Autrement dit : le nombre moyen d'α-liens par gène α-connecté.

---

résultats :

G=	**5655**								
a	**L(G,a)**	**L(H,a)**	**CG(H,a)**	**L(H,a) / H**	**L(H,a) / G**	**CG(H,a) / H**	**L(G,a) / G**	**L(H,a) / CG(H,a)**	**G,L(H,a) / H,L(G,a)**
Genes Guelzim									
Guelzim	1876	1876	**480**	3,91	0,33	1,00	0,33	3,91	11,78
HGX	4442	248	105	0,52	0,04	0,22	0,79	2,36	0,66
Curagen Uetz	4534	485	172	1,01	0,09	0,36	0,80	2,82	1,26

Ito	8950	601	226	1,25	0,11	0,47	1,58	2,66	0,79
Ho	4945	559	150	1,16	0,10	0,31	0,87	3,73	1,33
Mips complex	1203	150	124	0,31	0,03	0,26	0,21	1,21	1,47
Cellzome	3872	297	110	0,62	0,05	0,23	0,68	2,70	0,90
Syn Lethality	1772	143	60	0,30	0,03	0,13	0,31	2,38	0,95
Synexpresssion	32992	2763	201	5,76	0,49	0,42	5,83	13,75	0,99
Y2H	17926	1334	342	2,78	0,24	0,71	3,17	3,90	0,88
Compl	10020	1006	254	2,10	0,18	0,53	1,77	3,96	1,18
not2 Guelzim	62710	5246	430	10,93	0,93	0,90	11,09	12,20	0,99

Régulateur									
Guelzim	1876	1009	**109**	9,26	0,18	1,00	0,33	9,26	27,90
HGX	4442	168	47	1,54	0,03	0,43	0,79	3,57	1,96
Curagen Uetz	4534	144	51	1,32	0,03	0,47	0,80	2,82	1,65
Ito	8950	180	47	1,65	0,03	0,43	1,58	3,83	1,04
Ho	4945	69	31	0,63	0,01	0,28	0,87	2,23	0,72
Mips complex	1203	56	38	0,51	0,01	0,35	0,21	1,47	2,42
Cellzome	3872	89	30	0,82	0,02	0,28	0,68	2,97	1,19
Syn Lethality	1772	38	16	0,35	0,01	0,15	0,31	2,38	1,11
Synexpresssion	32992	264	27	2,42	0,05	0,25	5,83	9,78	0,42
Y2H	17926	492	82	4,51	0,09	0,75	3,17	6,00	1,42
Compl	10020	214	59	1,96	0,04	0,54	1,77	3,63	1,11
not2 Guelzim	62710	1008	99	9,25	0,18	0,91	11,09	10,18	0,83

Régulé									
Guelzim	1876	1277	**405**	3,15	0,23	1,00	0,33	3,15	9,50
HGX	4442	168	75	0,41	0,03	0,19	0,79	2,24	0,53
Curagen Uetz	4534	397	139	0,98	0,07	0,34	0,80	2,86	1,22
Ito	8950	454	191	1,12	0,08	0,47	1,58	2,38	0,71
Ho	4945	519	131	1,28	0,09	0,32	0,87	3,96	1,47
Mips complex	1203	111	97	0,27	0,02	0,24	0,21	1,14	1,29
Cellzome	3872	235	87	0,58	0,04	0,21	0,68	2,70	0,85
Syn Lethality	1772	117	51	0,29	0,02	0,13	0,31	2,29	0,92
Synexpresssion	32992	2532	182	6,25	0,45	0,45	5,83	13,91	1,07
Y2H	17926	1019	287	2,52	0,18	0,71	3,17	3,55	0,79
Compl	10020	865	216	2,14	0,15	0,53	1,77	4,00	1,21
not2 Guelzim	62710	4533	363	11,19	0,80	0,90	11,09	12,49	1,01

Complex									
Guelzim	1876	1004	254	0,36	0,18	0,09	0,33	3,95	1,08
HGX	4442	2879	731	1,03	0,51	0,26	0,79	3,94	1,31
Curagen Uetz	4534	3287	1119	1,18	0,58	0,40	0,80	2,94	1,47
Ito	8950	4401	1391	1,57	0,78	0,50	1,58	3,16	0,99
Ho	4945	4945	1577	1,77	0,87	0,56	0,87	3,14	2,02
Mips complex	1203	1203	1047	0,43	0,21	0,37	0,21	1,15	2,02
Cellzome	3872	3872	1444	1,38	0,68	0,52	0,68	2,68	2,02
Syn Lethality	1772	1493	514	0,53	0,26	0,18	0,31	2,90	1,70
Synexpresssion	32992	13110	856	4,69	2,32	0,31	5,83	15,32	0,80
Y2H	17926	10567	2082	3,78	1,87	0,74	3,17	5,08	1,19
Compl	10020	10020	**2797**	3,58	1,77	1,00	1,77	3,58	2,02
not2 Guelzim	62710	35190	2797	12,58	6,22	1,00	11,09	12,58	1,13

Commentaires sur les résultats :

Les 3 premiers tableaux portent sur l'étude des gènes régulateurs et régulés. Le 4ème porte sur le recouvrement des données de complexe.

Les valeurs les plus remarquables de la sur connectivité (L(H, α) / H) / (L(G, α) / G) est celle des données de MIPS pour les gènes régulateurs qui sont 2,42 fois plus connectés parmi les gènes régulateurs que parmi l'ensemble des gènes.

Les données d'Hybrigénics recouvrent 4 fois plus les gènes régulateurs que les gènes régulés. (1,96 contre 0,53).

Les données de synexpression sont étrangement 0,53 fois sous connectés aux gènes formant des complexes, résultat d'autant plus remarquable que ce résultat porte sur les nombres les plus grand, et est donc le plus significatif. C'est-à-dire que 2 protéines qui s'expriment en même temps

Algorithmique (à revoir ou supprimer)

Les algorithmes à mettre en place doivent permettre de répondre à deux types de questions

- Rechercher un motif dans le graphe.
- Compter le nombre d'occurrence de chaque motif.

afin de compter systématiquement l'occurrence de chaque motif, et de comparer le résultat avec ceux trouvés dans un réseau généré aléatoirement.
Il serait intéressant de pouvoir chercher des motifs sans avoir à en préciser l'échelle. Il serait intéressant de travailler aussi bien avec de petits motifs élémentaires, que des plus grands constitués de 'boîtes noires'.
L'exploration du réseau se fait avec des méthodes classiques d'exploration de graphe quelque peu étendues.
Pour compter le nombre de motif, on parcourt une fois le graphe, en mettant dans une table de hachage les motifs nouveaux trouvés.

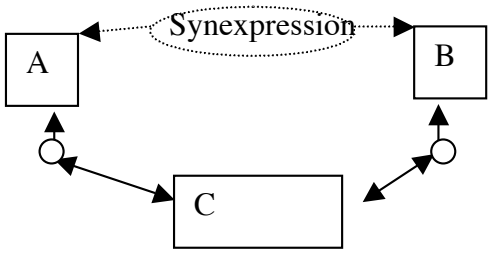
III.2 Exemples d'analyse des données

III.2.1 Etude sur les données de synexpression

Il est intéressant de confronter les données de synexpression aux autres données de la base qui sont complémentaires. En effet, les données de synexpression peuvent être considérées comme se situant à un niveau différent des données d'interactions physique car elles représentent les effets de l'action des circuits de régulations faisant intervenir des interactions directes physiques. Par conséquent il est intéressant à la lumière de ces effets mesurés en termes d'expression de rechercher un motif fondamental qui permet d'expliquer ou du moins d'émettre des hypothèses sur le circuit d'interactions régissant effectivement la synexpression.

Motif 1:

Description



- Ce motif représente de 2 gènes A et B corrégulés par un unique facteur de transcription C. De plus, A et B sont deux gènes synexprimés.

Intérêt du motif
- Le chemin réactionnel en noir est un bon candidat pour expliquer la donnée de synexpression en pointillé puisqu'il met bien en évidence la synchronisation des 2 gènes via C. L'expression de C contrôle en même temps celle de A et de B.

Résultats
- 77 instances de ce motif existent dans la base.

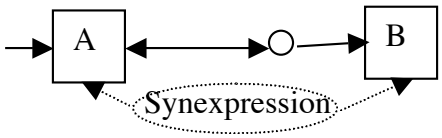
Interprétation

- Ce résultat montre qu'il est possible dans une certaine mesure d'expliquer le mécanisme sous-jacent d'une donnée de synexpression. D'autres motifs élémentaires permettent d'expliquer d'autres résultats de synexpression comme on va le voir plus bas.

Une fois le mécanisme validé, plusieurs traitements sont possibles. - On peut placer le synepression à un niveau hiérarchique supérieur afin de s'y référer directement sans plus avoir à analyser à chaque fois le rôle du motif sous-jacent. C'est-à-dire appliquer systématiquement le processus de renormalisation, en remplacant le motif global ci-dessus trouvé par motif de syn expression détaillé (en pointillé) au dessus du motif fondamental ce qui permet de travailler au choix avec l'une ou l'autre des représentations, ou encore les deux. - On peut accroître le score de crédibilité de la réaction afin de pouvoir formuler d'autres hypothèses par dessus avec un score de crédibilité important.

Motif 2 *:*

Description



- Ce motif représente une cascade de régulation : A est un gène régulé qui active l'expression de B. De plus A et B sont coexprimés.

Interprétation
- Le fait que lorsuque A soit exprimé, B le soit également peut expliquer la coexpression mesurée.

Résultats
- Ce motif est présent 7 fois dans la base.

Interprétation
- Le fait qu'on ait trouvé ces motifs, tout comme dans le cas du motif 1, montre que les données de synexpression recouvrent celles d'interactions physiques. En effet, à l'échelle de temps d'expérimentation sur l'expression de A et B, le retard qui existe entre l'expression de B et celle de A n'est pas visible.

Motif 3 :

Description

<table>
<tr><td>E1</td><td>L1</td><td>Coexpression</td><td>L2</td><td>E2</td></tr>
<tr><td>^</td><td colspan="3" rowspan="3">Occurence :200
Symétrie :2
Représentativité :100
UniqueE1E2E3E4 :49</td><td>^</td></tr>
<tr><td>Direct_transcriptional_ regulation</td><td>Direct_transcriptional_ regulation</td></tr>
<tr><td>L7</td><td>L4</td></tr>
<tr><td>E4</td><td>L6</td><td>Physical_interaction | Complex</td><td>L5</td><td>E3</td></tr>
</table>

- Le mécanisme 1 de coexpression montré ci dessus, peut être compliqué quelque peu en s'autorisant une interaction physique entre les 2 facteurs transcription.

Interêt
- Ce motif permet en effet d'expliquer la coexpression mesurée entre E1 et E2, sous hypothèse par exemple que E4 et E3 se dimérisent pour interagir avec l'ADN et ainsi réguler ensemble l'expression de E1 et de E2.

Résultats
- 200 instances de ce motif existent dans la base de donnée.

Interprétation
- Comme l'a montré le motif 3 ci-dessus, là encore, on a un exemple où des données d'interaction peuvent expliquer des cas de coexpression, ce qui est tout à fait enthousiasmant. De plus il existe de nombreux motifs d'interactions fondamentaux entre ces couples puisque le motif est représenté 200 fois c'est-à-dire qu'il existe plusieurs chemins faisant intervenir les 4 partenaires du motif E1,E2,E3,E4.

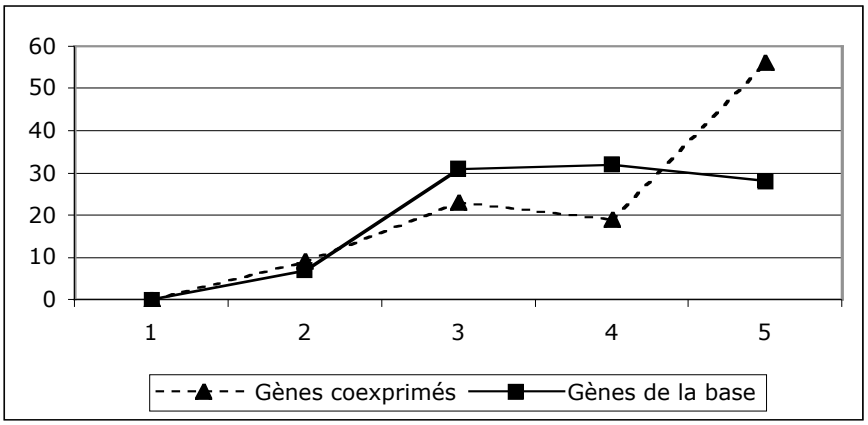
Ces premières analyses montre qu'il est possible de trouver des chemins faisant intervenir uniquement des interactions pour expliquer des mesures de coexpression.

Etude 4 : Distribution de la longueur minimal des chemins entre gènes coexprimés.
- Une recherche plus systématique sur ces données de synexpression consiste à regarder, pour tout couple de gènes coexprimés, la longueur du plus court chemin qui les relie dans la base au moyen uniquement d'interactions physiques protéine-protéine , protéine ADN, et de complexes. Un chemin minimal de longueur 1 signifie qu'il existe un chemin ne comportant qu'une seule interaction entre les deux gènes étudiés, comme par exemple le motif 2; de longueur 2 signifie que le chemin comporte 2 transformations comme par exemple le motif 1, et ainsi de suite.
- On trouve sur un échantillon de 3098 couples de gènes coexprimés pris au hasard :

	% de chemins minimaux de longueur
1 :	00
2 :	09
3 :	23
4 :	19
reste :	46

Ces résultats sont à comparer aux données globales, c'est-à-dire la distribution des plus courts chemins entre couples pris au hasard dans la base, pour un échantillon de1490 couples de protéines:

	% de chemins minimaux de longueur
1 :	00
2 :	07
3 :	31
4 :	32
reste:	28

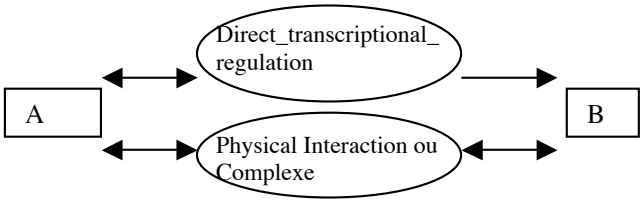


- Cela montre que les données de coexpression, c'est-à-dire que les gènes qui sont régulés ensembles sont paradoxalement plus éloignés que les autres gènes. La médiane du chemin le plus court entre gènes est autour de 3,5 alors qu'elle est de 4 pour des gènes coexprimés. Comme si la coexpression qui est une synchronisation est vraie qu'à longue portée, n'est pas le résultat de chemin d'interactions plus ou moins directes entre ces gènes. La synchronisation est ailleur…

III.2.2 Motif circulaires

Motif à 2 interactions

***Motif 5** :*
Description



- Cette boucle de rétroaction génétique fait intervenir une interaction protéine-protéine entre la protéine B le facteur de transcription A qui contrôle l'expression de B.

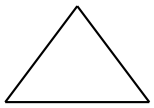
Interêt
- Les boucles de rétroaction sont importantes fonctionnellement mais rares dans le monde des régulations transcriptionelle direct. On s'attend donc lorsqu'on mélange les données génomiques et protéiques à trouver enfin un grand nombre de ces boucles. Cette boucle peut être considérée comme un boucle de rétroaction car A est un facteur de transcription qui permet l'expression de B. Et B se lie à A, ce qui peut représenter une réaction concurrente pour la ressource A, et ainsi inhiber l'expression de B.

Résultats
- Il existe 17 motifs où le facteur de transcription interagit physiquement avec la protéine qu'elle régule génétiquement. 7 ou l'interaction est de type Y2H (aucun HGX), et 15 ou l'interaction est de type formation de complexe.

Expression
5655 protéines existent dans la base.
63594 couples de protéines reliés par un transformation de type Complex ou Physical_interaction.
909 couples de gènes régulateur-régulé.
Nombre attendu du motif 5 : 2 * 909 * 63594 / 5655^2 ~ 5

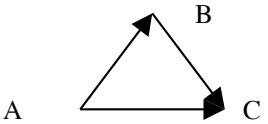
Interprétation
– C'est plus qu'attendu pour un réseau aléatoire, mais il faut un étude plus poussée sur les réseaux aléatoires et un calcul d'erreur approprié non réalisés à ce jour pour pouvoir véritablement conclure.

Motif 6
Description
- Motif triangulaire à 3 interactions



Résultats
- Il existe 1151 motifs circulaires à 3 interactions dont au moins une est une régulation transcriptionelle.

Description



- un motif en FeedForward est un motif introduit par Uri Alon et al. où un gène A régule un gène B et C, et B régule aussi C. Un motif FeedForward est dit cohérent si les deux chemins d'action de A sur C régulent C de la même manière, c'est-à-dire soit positivement soit négativement.

Résultats
- 141 motifs trouvés comportent 3 régulations transcriptionelles en Feedforward pour moitié cohérentes, et pour moitié incohérentes
- aucune boucle de sens unique (A régule B qui régule C qui régule A)

Interprétation
- Cela est en accord avec les résultats d'Uri Alon qui met l'accent sur le feedforward et l'absence de boucles.

***Motifs supplémentaires 7** :*

Il existe 465 motifs triangulaires à 2 interactions régulatrices transcriptionelles
De nombreux autres motifs en triangle ont également été compatiblisés à retrouver en annexe.

Motif à 4 interactions (carré)

Description générale des motifs à 4 en carré
Représentation des motifs à quatre :

<table>
<tr><td>E1</td><td>L1</td><td>T1</td><td>L2</td><td>E2</td></tr>
<tr><td>L8</td><td colspan="3" rowspan="3">Occurence
Symétrie
Représentativité
UniqueE1E2E3E4</td><td>L3</td></tr>
<tr><td>T4</td><td>T2</td></tr>
<tr><td>L7</td><td>L4</td></tr>
<tr><td>E4</td><td>L6</td><td>T3</td><td>L5</td><td>E3</td></tr>
</table>

E1,E2,E3,E4 : Entités
T1,T2,T3,T4 : Transformations
L1,L2,L3,L4,L5,L6,L7,L8 : Link

Avec la condition qu'aucune transformation ne soit un phénotype ou une coexpression.

Pour chaque motif, il est indiqué à l'intérieur de sa représentation le nombre de motifs trouvés. Mais il faut préciser cette notion :

Définitions :

- *L'occurrence*, est le nombre d'instances du motif qui a été trouvé dans la base, sachant que 2 motifs A et B sont distincts si $\exists$ i $\in$ |[1,4]| , A.Ti != B.Ti v A.Ei != B.Ei
autrement dit deux motifs sont distincts si ils ont une de leur entité ou transformation orientée distincte.

- Le nombre de *symétrie* est le nombre de motifs symétriques du motif présenté, c'est le nombre de motifs distincts obtenus par permutation ou rotation des entités et transformations, en vérifiant les même relations entre les protéines et les transformation.
Ce nombre n'est pas indiqué s'il n'existe pas.

- La *représentativité* est le rapport entre l' *Occurrence* et la *Symétrie*.
Par exemple, sur le motif 10, on a une symétrie d'ordre 2 en inversant E1 E2 avec E4 E3.

- *UniqueE1E2E3E4* est le nombre d'occurrences du motif réduit uniquement à l'ensemble de ses entités. C'est-à-dire que deux instances seront comptabilisées seulement si ils diffèrent par l'ensemble des protéines mis en jeu.

Motif 8 :
Description

<table>
<tr><td>E1</td><td>L1</td><td>Direct_transcriptional_
regulation</td><td>-></td><td>E2</td></tr>
<tr><td>L8</td><td colspan="3" rowspan="3">Occurence :62950
Symétrie :X
Représentativité :?
UniqueE1E2E3E4 :37043</td><td>L3</td></tr>
<tr><td>T4</td><td>T2</td></tr>
<tr><td>L7</td><td>L4</td></tr>
<tr><td>E4</td><td>L6</td><td>T3</td><td>L5</td><td>E3</td></tr>
</table>

Résultats
- Ils se comptent par dizaines de milliers.

Expression
5655 protéines existent dans la base.
64494 couples de protéines reliés par un transformation de type Complex,, Physical_interaction ou Direct_transcriptional_regulation
909 couples de gènes régulateur-régulé.
Nombre attendu du motif 8 : 32 * 909 * 64494^3 / 5655^4 ~ 7600 à comparer avec 37043.

Intérêt
- Dans la suite, on s'attache au décompte plus détaillé de tous ces motifs circulaires à 4 interactions comportant au moins une interaction de régulation transcriptionelle.

Etude préliminaire pour l'étude des motifs 10 et 11 :
Description
On désire étudier ici la probabilité d'interaction entre protéines régulée et facteur de transcription.

Motif servant au calcul de A :

E1	**L1**	Physical_interaction ou Complex	**L2**	**E2**
L8				**L3**
Direct_transcriptional_regulation				Direct_transcriptional_regulation
v				**v**
E4				E3

Motif servant au calcul de B :

E1	**L1**	Physical_interaction ou Complex	**L2**	**E2**
^				**^**
Direct_transcriptional_regulation				Direct_transcriptional_regulation
L7				**L4**
E4				E3

Motif servant au calcul de C *:*

E1	**L1**	Physical_interaction ou Complex	**L2**	**E2**
L8				**^**
Direct_transcriptional_regulation				Direct_transcriptional_regulation
v				**L4**
E4				E3

Résultats
Statistiques obtenues sur 10 000 couples de protéines pris au hasard :
- Probabilité d'interactions directe physique entre facteurs de transcription :
A=0.0112
- Probabilité d'interactions directe physique entre protéines régulées:
B=0.0071
- Probabilité d'interactions directe physique entre facteur de transcription et protéines régulées:
C=0.0041
2 facteurs de transcriptions ont 3 fois plus de chances d'interagir que deux protéines prises au hasard.

Interprétation
Ces résultats ne sont pas très étonnant car on sait d'après la littérature que des facteurs de transcription agissent souvent de pair pour se fixer sur l'ADN. Bien entendu ces résultats peuvent être également dûs à un biais introduits par l'incomplétude des données.

***Motif 10 et 11* :**
Description
Motif 10

E1	**L1**	Direct_transcriptional_ regulation	**->**	**E2**
L8	**Occurence :252 Symétrie :2 Représentativité :126 UniqueE1E2E3E4 :52**			L3
Physical_interaction | Complex				Physical_interaction | Complex
L7				**L4**
E4	**L6**	Direct_transcriptional_ regulation	**->**	E3

Motif 11 :

E1	**L1**	Direct_transcriptional_ regulation	**->**	**E2**
L8	**Occurence :44 Symétrie :2 Représentativité :22 UniqueE1E2E3E4 :15**			L3
Physical_interaction | Complex				Physical_interaction | Complex
L7				**L4**
E4	**<-**	Direct_transcriptional_ regulation	**L5**	E3

- Ces 2 motifs se ressemblent beaucoup, ils se distinguent par L5 et L6 uniquement. Ils mettent l'accent sur l'orientation de la propagation du signal dans l'organisme : le motif 11 permet une boucle de rétroaction, alors que le motif 10 seul ne le permet pas.

Résultats
- Le motif 10 est comptabilisé de 4 à 6 fois plus que le motif 11.

Interprétation
- Ces deux motifs devraient être trouvés en même nombre à priori. La seule différence qui existe entre ces motifs est l'orientation d'une régulation transcriptionelle par rapport à l'autre. Soit elles sont orientées dans le même sens (motif 10) soit en sens inverse, bouclant (motif 11). Or il s'avère que l'on trouve en bien plus grand nombre, le cas des sens identiques.
- Cela laisse supposer, que l'information est orientée avec un flux d'informations à sens majoritairement unique. On retrouve le résultat sur les motifs triangulaires, à savoir que les boucles rétroactives sont moins représentées. Cela ne s'explique cette fois, non pas par le fait que le traitement du signal laisse de meilleurs perspectives avec le motif 11 que le motif 10, mais est en grande parti du au fait que des facteurs de transcriptions interagissent plus probablement entre eux, comme vu lors de l'étude 9. En effet : on retrouve bien que le ratio A.B/C^2 ~= 5, correspond bien au ratio des occurrences du motifs 10 par rapport au 11. La surepresentativité du motif 10 contre le motif 11 ne nécessite aucune hypothèse supplémentaire spécifique au motif 10 et 11 que celles émises aux résultats préliminaires 9.

Motif 12 et 13
Description

Etudions à présent des motifs dont les motifs préalablement étudiés 10 et 11 sont des spécifications.
Motif 12 :

<table>
<tr><td>E1</td><td>L1</td><td>Direct_transcriptional_ regulation</td><td>-></td><td>E2</td></tr>
<tr><td>L8</td><td colspan="3" rowspan="3">Occurence :862
Symétrie :X
Représentativité :?
UniqueE1E2E3E4 :202</td><td>L3</td></tr>
<tr><td></td><td>T2</td></tr>
<tr><td>L7</td><td>L4</td></tr>
<tr><td>E4</td><td>L6</td><td>Direct_transcriptional_ regulation</td><td>-></td><td>E3</td></tr>
</table>

Motif 13 :

<table>
<tr><td>E1</td><td>L1</td><td>Direct_transcriptional_ regulation</td><td>-></td><td>E2</td></tr>
<tr><td>L8</td><td colspan="3" rowspan="3">Occurence :10744
Symétrie :X
Représentativité :?
UniqueE1E2E3E4 :2308</td><td>L3</td></tr>
<tr><td>T4</td><td>T2</td></tr>
<tr><td>L7</td><td>L4</td></tr>
<tr><td>E4</td><td><-</td><td>Direct_transcriptional_ regulation</td><td>L5</td><td>E3</td></tr>
</table>

- Ces deux motifs 12 et 13 sont identiques aux précédents, sauf qu'on la levé la contrainte sur T2 et T4 qui ne sont plus forcément des interactions physiques.

Résultats
- Cette fois les résultats de représentativités sont complètements inversés par rapport aux précédents : le motif en boucle 13 est bien plus représenté que le motif 12.

Interprétation
- Il est nécessaire de chercher quels types de transformation de T2 et T4 sont responsables de cette inversion.

Motifs 14
Description
- Etudions plus en détail la cause de l'inversion des résultats.

E1	**L1**	Direct_transcriptional_ regulation	**->**	**E2**
L8	**Occurence :10512 Symétrie :X Représentativité :? UniqueE1E2E3E4 :2230**			**L3**
Direct_transcriptional_ regulation				Direct_transcriptional_ regulation
L7				**L4**
E4	**L6**	Direct_transcriptional_ regulation	**L5**	E3

Résultats
- Le nombre d'occurrence de ce motif se compte en milliers

Interprétation
- Le nombre de motif explose lorsque T2 et T4 sont des Direct_transcriptional_
Regulation. La cause de l'inversion se trouve probablement là, mais il faut encore spécifier le motif.

Motif 15 et 16
Description
Motif 15 :

E1	**L1**	Direct_transcriptional_ regulation	**->**	**E2**
L8	**Occurence :240 Symétrie :X Représentativité :? UniqueE1E2E3E4 :67**			**L3**
Direct_transcriptional_ regulation				Direct_transcriptional_ regulation
L7				**L4**
E4	**L6**	Direct_transcriptional_ regulation	**->**	E3

Motif 16 :

E1	**L1**	Direct_transcriptional_ regulation	**->**	**E2**
L8	**Occurence :10272 Symétrie :X Représentativité :? UniqueE1E2E3E4 :2203**			**L3**
Direct_transcriptional_ regulation				Direct_transcriptional_ regulation
L7				**L4**
E4	**<-**	Direct_transcriptional_ regulation	**L5**	E3

- Ces motifs 15 et 16 sont à comparer aux 10 et 11 et sont des spécifications de 12 et 13.

Résultats
- On retrouve les proportions ‘inverses’ accentuées obtenues avec les motifs 12 et 13.

Interprétation
- Ce sont bien les transformations de régulations transcriptionelles de T2 et T4 qui ont majoritairement modifié les résultats lors du passage des couples 10,11 aux 12,13.

Motif 17
Descrition
- Ce motif est une spécification du motif 15

E1	**L1**	Direct_transcriptional_ regulation	**->**	**E2**
L8	**Occurence :52 Symétrie :1 Représentativité :52 UniqueE1E2E3E4 :40**			**L3**
Direct_transcriptional_ regulation				Direct_transcriptional_ regulation
V				**v**
E4	**<-**	Direct_transcriptional_ regulation	**L5**	E3

- On retrouve ici le type de motif en FeedForward introduit avec les motifs triangulaires (Motif 6)

Motif 18
Description

E1	**L1**	Direct_transcriptional_ regulation	**->**	**E2**
^	**Occurence :0 Symétrie : Représentativité : UniqueE1E2E3E4 :0**			**L3**
Direct_transcriptional_ regulation				Direct_transcriptional_ regulation
L7				**v**
E4	**<-**	Direct_transcriptional_ regulation	**L5**	E3

Résultats
- Aucune occurrence de ce motif n’est présente.

Interprétation
- Ce nombre de 0 peut être comparé au nombre de motif 17 bien plus représenté. Le FeedForward est plus représenté que le motif circulaire.

Motif 19
Description

E1	**L1**	Direct_transcriptional_ regulation	**->**	**E2**
L8	**Occurence : 10168 Symétrie :2 Représentativité :5084 UniqueE1E2E3E4 : 2181**			**^**
Direct_transcriptional_ regulation				Direct_transcriptional_ regulation
v				**L4**
E4	**<-**	Direct_transcriptional_ regulation	**L5**	E3

- Ce motif représente on double contrôle

Résultats

- c'est le motif le plus représenté parmi les motifs 17,18,19.

Interprétation

- Peut être que ce motif à des propriétés intéressantes pour la cellule, notamment de robustesse. Ces résultats peuvent être dû à la grande symétrie du motif.

Motif 20

Description

E1	**L1**	Direct_transcriptional_ regulation	**->**	**E2**
L8	**Occurence :58 Symétrie :2 Représentativité :29 UniqueE1E2E3E4 :29**			**L3**
Physical_interaction | Complex				Direct_transcriptional_ regulation
L7				**v**
E4	**L6**	Direct_transcriptional_ regulation	**->**	E3

- Ce motif est une spécification du motif 12 et du motif 21

Interprétation

Une des interprétations de ce motif est un motif en FeeForward.

Motif 21

Description

E1	**L1**	Direct_transcriptional_ regulation	**->**	**E2**
L8	**Occurence :10479 Symétrie :X Représentativité :? UniqueE1E2E3E4 :2297**			**L3**
T4				Direct_transcriptional_ regulation
L7				**v**
E4	**<-**	Direct_transcriptional_ regulation	**L5**	E3

Ce motif, est un des autres nombreux motifs carrés qui figurent en annexe.

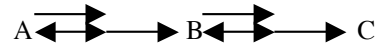
III.3 Perspectives

Sur les motifs

Il peut être intéressant de rechercher des motifs définis à priori par leur fonction. Ces motifs présentés ont des propriétés dynamiques particulièrement attrayantes qui ont par ailleurs été montrées in-vivo.
La modélisation des motifs ci-dessous peut se faire de manière discrète ou continue. Pur cela il faut se référer à la partie I de ce document. L'interprétation du traitement de l'information est légèrement différente si l'on considère des niveaux d'expression binaires ou continus, c'est pourquoi les propriétés mises en valeur ne doivent pas dépendre du modèle dynamique que l'on place, mais doivent être intrinsèques au modèle statique, tel qu'on le connaît.
L'interprétation des motifs est donnée avec un vocabulaire emprunté au monde discret ou continu.

Chaine Retardateur par cascade

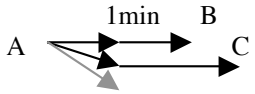
C(t)=B(t+1)+A(t+2)



Ce motif représente une cascade directe d'interactions de gènes. A contrôle l'expression de B, qui contrôle l'expression de C. Ainsi les 3 gènes sont exprimés séquentiellement.

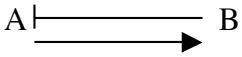
Chaine Retardateur par affinité

C(t)=B(t+1)+A(t+2)



Avec l'hypothèse que les vitesses de réactions sont différentes pour chacun des produits (B,C..) par exemple parce leur ARN messager est de longueur différente, ce petit réseau permet là encore d'exprimer séquentiellement plusieurs gènes.

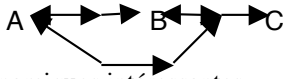
Boucle négative retardée

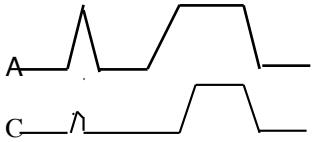


A = B cos (t/Δt) Homéostasie
Toute boucle avec un nombre d'inhibition impaire se comporte comme un oscillateur qui peut faire office d'horloge, ou comme un asservissement stable de l'expression de A.

FeedForward Anti-rebond

C(n+1)=A(n) AND A(n-1)

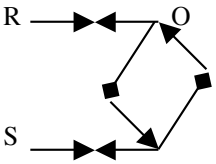




Ce motif qui fait intervenir 3 gènes a des propriétés dynamiques intéressantes.
A contrôle l'expression de B. Et C a besoin de B et A pour être exprimé, par exemple parce que A et B se dimérisent pour devenir facteur de transcription de C.
Ce motif permet de s'assurer que le signal A est stable avant d'exprimer C. Ce qui permet à l'organisme d'être peu sensible au bruit. Ce motif a été décrit par Alon [2] car il est sur représenté dans E.coli.
D'autres variantes similaires peuvent produire d'autres fonctions tout aussi intéressantes.

Bascule - Bit mémoire

S :Set R :Reset
O :Output



Il s'agit ici d'une bascule faisant intervenir 2 gènes. Plusieurs variantes possibles, La version présentée est celle qui a été étudiée par Gardner et Collins [18]

Sur la recherche

La base de données peut encore être enrichie par d'autres réactions, notamment des voies de transduction. D'autres données de transcriptome pourront également être introduites dans l'avenir notamment des données temporelles. Le modèle et la base mis en place, l'exploitation commence. Une étude théorique sur la sur-représentativité des motifs présents comparé à ceux que l'on peut trouver dans un graphe produit aléatoirement est à mener.

Cette analyse de motifs du réseau de régulation génétique étendu chez la levure devrait permettre la mise en évidence systématique des mécanismes de régulation utilisés par la levure, et devrait notamment faire apparaître un plus grand nombre de boucles de rétroactions que celles trouvées uniquement par l'étude du réseau de gènes régulateurs. En effet, les interactions du protéome sont plus rapides que l'expression de gènes ce qui est précieux pour permettre un bon asservissement de leur expression. De telles méthodes d'analyse par modules doivent jouer un rôle important à l'avenir et paraissent dès à présent inévitables pour une compréhension plus profonde de la physiologie d'organismes pris dans leur ensemble.

Conclusion

Ce rapport décrit un modèle et ses outils, nécessaires à l'exploration du réseau d'interactions de la levure en modules fonctionnels. De plus ce modèle permet de travailler sans échelle d'interaction prédéfinie.
La base sur laquelle s'appuie la recherche est constituée.
Les résultats des premières analyses sont intéressants et prometteurs. Des données de différentes technologies et de différentes échelles se recouvrent. Des motifs fonctionnels apparaissent. D'autres points exposés originaux demandent à être détaillés.

[1] N. Guelzim, S. Bottani, P. Bourgine & F. Képès
Topological and causal structure of the yeast transcriptional regulatory network
Nature Genetics 31: 60-63 (2002)

[2] S. S. Shen-Orr, R. Milo, S. Mangan & U. Alon,
Network motifs in the transcriptional regulation network of Escherichia coli.
Nature Genetics, Published online: 22 April 2002, DOI:10.1038/ng881.

[3] T. Ideker, O. Ozier, B. Schwikowski, A. F. Siegel,
Discovering regulatory and signaling circuits in molecular interaction networks, to be presented at ISMB 2002, to appear as Bioinformatics special issue

[4] P. Uetz, T. Ideker, B. Schwikowski,
Computational analysis, visualization and integration of protein-protein interactions. In: Golemis,E. (ed.) The Study of Protein-Protein Interactions - An Advanced Manual. Cold Spring Harbor Laboratory Press, in press (2001)

[5] Kurt W. Kohn.
Molecular interaction map of the mammalian cell cycle control and DNA Repair Systems, Mol Biol Cell. 10:2703-2734, (1999)

[6] R. Maimon, S Browning, GeneNetworkScience,
Diagrammatic Notation and Computational Structure of Gene Networks, presented at ICSB 2001

[7] Uetz,P. et al.
A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403: 623-627 (2000)

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

[9] Ho Y, Gruhler A, Heilbut A, et al.: *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature* 2002, 415: 180-183.

[10] Gavin A-C, Bösche M, Krause R, *et al.*: *Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature* 2002, 415:141-147.

[11] Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J. *From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli.* Bioessays. 1998May;20(5):433-40.

[12] Hybrigenics, données non public

[13] Ito T, et al. *A comprehensive two-hybrid analysis to explore the yeast protein interactome.*
Proc Natl Acad Sci U S A. 2001 Apr 10;98(8):4569-74.

[14] Matsuno, H., Doi, A., Nagasaki, M., and Miyano, S.,
Hybrid Petri net representation of gene regulatory network. Pacific Symposium on Biocomputing 2000, 338ú349, 2000

[15] Mips
[16] Yeast protein complex database , Cellzome, données filtrées S1

[17] De Jong H,
Modeling and Simulation of Genetic Regulatory Systems : A Litterature Review , Journal of Computational Biology, v9n1 2002

[18] S.Gardner T, C R.Cantor, J J.Collins
Construction of a genetic toggle switch in Escherichia coli, Nature 403 2000

[19] R Küffner, R Zimmer,T Lengauer,
Pathway Analysis in Metabolic Databases via Differential Metabolic Display (DMD), vol 16 no 9 Bioinformatics 2000

[20] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998b).
Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell, 9:3273-3297.

[21] H. Jeong, B. Tomber, R. Albert, Z.N. Oltvai, and A.-L. Barabasi,
The large-scale organization of metabolic networks , Nature 407 651 (2000)