Méthodologie d'extraction et d'analyse de réseaux de régulation de gènes : analyse de la réponse transcriptionnelle à l'irradiation chez S. cerevisiæ

Nizar Touleimat

Université d'Évry-Val d'Essonne
U.F.R. Sciences Fondamentales et Appliquées

THÈSE

présentée pour obtenir
le grade de Docteur en Sciences
de l'Université d'Évry-Val d'Essonne

Spécialité
bioinformatique

Mohamed Nizar Touleimat

Méthodologie d'extraction et d'analyse de réseaux de régulation de gènes

- Analyse de la réponse transcriptionnelle à l'irradiation chez *S. cerevisiæ* -

Soutenue le *26 novembre 2008* devant le jury composé de

Mme Florence d'Alché Buc	*Directeur de thèse*
Mme Marie Dutreix	*Directeur de thèse*
Mme Céline Rouveirol	*Rapporteur*
M. Frédéric Devaux	*Rapporteur*
M. Louis Wehenkel	*Examinateur*
M. Pierre Dupont	*Examinateur*
M. Nicolas Pollet	*Examinateur*

Thèse préparée au sein de l'équipe Apprentissage, Modélisation et Intégration de données pour les systèmes biologiques (AMIS-BIO) du laboratoire Informatique, Biologie Intégrative et Systèmes Complexes (IBISC) - FRE 2873 CNRS
et au sein de l'équipe Recombinaison et Instabilité des Génomes, du Département Recherche de Transfert de l'Institut Curie

Remerciements

Mes remerciements s'adressent en premier à mes directrices de thèse, Marie Dutreix et Florence d'Alché-Buc. Marie, dont l'engagement m'a permis de débuter ce projet et de reprendre goût à la science après une année sabatique. Ses qualités scientifiques m'ont permis de bénéficier d'un encadrement rigoureux, stimulant et productif. Florence, qui a pris le risque d'encadrer un "bio-bioinformaticien" et avec qui mon projet de thèse a rellement débuté. Ses qualités pédagogiques et scientifiques m'ont permis de rapidement évoluer dans un environnement de travail exceptionnel. Un grand merci à Florence pour ses qualités humaines qui ont fait de cette aventure scientifique une belle aventure humaine.

Je tiens à remercier Céline Rouveirol et Frédéric Devaux, qui m'ont fait l'honneur d'être les raporteurs de ma thèse. Leurs commentaires et corrections ont permis d'enrichir et de valoriser mon travail. Je remercie tous les membres du jury dont la présence lors de ma soutenance de thèse m'a honoré. Je les remercie également pour leurs questions et commentaires stimulants.

Je tiens à remercier Farida Zehraoui, Pierre Geurts et Maria Quanz pour leurs collaborations fructueuses. J'ai pris un réel plaisir à travailler avec eux, tant pour leurs talents scientifiques que pour leur grandeur d'âme.

Un grand merci à mon co-thésard, co-bureau, complice et ami, Cédric Auliac, pour son soutien, ses encouragements, son recul et son aide si précieuse dans la préparation de ma soutenance. Merci pour toutes les dicussions non scientifiques et en particulier pour l'échange de protocoles expérimentaux gastronomiques !

Je tiens à remercier mes autres camarades et collègues, Minh Quach, Cyril Combe et Nicolas Brunel pour leurs conseils et leur amitié. Merci à Nathalie, Christophe, Aurélie, Maryline et Céline qui, malgré mes sauts épisodiques à l'Institut Curie, m'ont toujours accueilli comme membre de l'équipe à part entière et comme ami. Un grand merci à tous les thésards d'IBISC pour l'ambiance chaleureuse et animée du laboratoire. Je serais toujours redevable à Laurent Poligny et à Benoît Calvez pour m'avoir si souvent dépanné et sauvé de catastrophes informatiques. Je tiens à remercier Mohamed Sghaier Ben Hammadi pour le travail remarquable qu'il a effectué sur l'outil *XRegPath*, lors de son stage dans notre équipe.

Merci à mes *potes* de la Cité Internationale Universitaire de Paris : Maïra, Prisca, Ghada, Manu, Nico, Ale, Ramo, Emiliano, Manuel, Aurélie, Hayet, Chloé et sa bande. Votre présence à mes côtés a été si précieuse... J'ai passé les plus beaux moments de ces années de thèse en votre compagnie !

Un grand merci à *mi hermaña* Maïra, mon *alter ego* durant ces années de thèse. J'ai modestement essayé de suivre ton exemple...

Merci à mes compagnons de toujours, mes frères et sœurs de cœur, Ruth, Adil et Marie

Lecorre.

Merci à ma si précieuse Anne. Tu m'as connu au moment le plus difficile de ma thèse, mais tu en as fais le commencement du plus doux de ma vie.

A mes parents, mon frère, ma sœur et à toute ma famille, de France et de Syrie, je dédie cette thèse. Merci pour vos encouragements, pour votre soutien et votre amour.

Résumé

Comprendre le fonctionnement d'un organisme vivant et pouvoir prédire son comportement en fonction des variations de son environnement sont des objectifs classiques en biologie et en médecine. Dans le cas du stress génotoxique provoqué par de fortes doses de radiations ionisantes, les cellules déclenchent des mécanismes moléculaires complexes pour tenter de réparer les lésions de l'ADN. La réponse cellulaire aux dommages de l'ADN est relativement bien étudiée mais de nombreuses observations montrent que la réponse à l'irradiation est beaucoup plus globale et implique l'expression de nombreux gènes. Dans le cadre de ce travail nous nous intéressons à la réponse transcriptionnelle de la levure *S. cerevisiae* à de fortes doses de radiations γ. Nous souhaitons identifier les différentes formes potentielles de cette réponse et reconstruire un réseau de régulation génique impliqué dans son contrôle. La problématique de ce travail réside d'une part dans l'exploitation des dynamiques d'expression des gènes, mesurées dans différentes conditions de perturbations génétiques, et d'autre part dans l'intégration d'informations biologiques systémiques, associées aux gènes, de natures hétérogènes. En nous appuyant sur cette problématique particulière, nous définissons une approche générale nommée *XRegPath*, qui exploite séquentiellement les données d'expression et les informations biologiques additionnelles. Notre approche est principalement constituée d'une étape automatisée de déduction logique de régulations à partir d'une stratégie de perturbations et de deux étapes d'induction qui permettent d'analyser la dynamique d'expression des gènes et d'extraire des liens de régulation potentiels à partir des données systémiques additionnelles. Cette approche permet d'extraire différentes bribes de réseaux de régulation et de les intégrer au sein d'un réseau global rendant compte de la dynamique d'une réponse cellulaire et des principaux mécanismes de régulation associés. L'application de notre méthodologie à l'analyse de la réponse transcriptionnelle à l'irradiation chez la levure a permis d'identifier une réponse complexe et a permis de proposer un modèle de régulation. Certaines relations au sein de ce modèle ont pu par la suite être validées expérimentalement.

Understanding how a living organism functions and being able to predict its behaviour according to different variations of its environment are classical objectives in biology and in medicine. In the case of the genotoxic stress induced by high doses of ionizing radiation, cells trigger complex molecular mechanisms to attempt to repair DNA damage. The cellular response to DNA damage is relatively well studied, however, many observations show that the irradiation response is more global and involves the expression of many genes. Within the framework of this study we are interested in the transcriptional response of the yeast *S. cerevisiae* to high doses of γ radiations. We propose to identify the different potential patterns of this response and to reconstruct a gene regulatory network involved

in its control. The first point of this work lies in the exploitation of the gene expression dynamics, measured in different conditions of genetic perturbations. The second point lies in the integration of systemic and heterogeneous biological informations that are associated to the genes. By addressing these particular issues we define a general approach called *XRegPath*, that sequentially exploits expression data and additional biological informations. Our approach is mainly composed of three steps : one step of automated logical deduction of regulations from a strategy of perturbations and two induction steps that allow the analysis of the gene expression dynamics and the extraction of potential regulation relations from additional systemic data. This approach allows to extract different pieces of regulation networks and to integrate them into a global network that is able to illustrate the dynamical property of a cellular response with its several associated regulatory mechanisms. The application of our methodology to the analysis of the yeast transcriptional response to irradiation allowed to identify a complex response and allowed to propose a regulation model of the irradiation response. Some of the regulations inside the model have been subsequently experimentally validated.

A mon grand-père, *Jiddo* Mahmoud.

Table des matières

Liste des abréviations :

ADNc : ADN complémentaire
ARNm : ARN messager
ARNr : ARN ribosomique
ARNt : ARN de transfert
ChIP : abrégé de *ChIP-on-chip* (*Chromatin ImmunoPrecipitation on Chip*)
EM : *Expectation Maximization* (algorithme)
FT : Facteur de Transcription (FTs au pluriel)
GO : *Gene Ontology*
IR : IRradiation
MMS : Methyl Methane-Sulfonate (composé chimique génotoxique)
ORFs : *Open Reading Frame* pour phase ouverte de lecture
RI : Radiations Ionisantes
RT-PCR : *Polymerase Chain Reaction after Reverse Transcription of RNA*
RX : Rayons X
SAGE : *Serial Analysis of Gene Expression*
SNP : *Single Nucleotide Polymorphism*

Introduction

Comprendre le fonctionnement d'un organisme vivant et pouvoir prédire son comportement en fonction des variations de son environnement sont des objectifs que l'on souhaite atteindre dans de nombreuses disciplines liées aux sciences du vivant (médecine, agronomie, biologie, etc.). Au siècle dernier, les études étaient basées sur l'observation des caractéristiques directement visibles de modèles expérimentaux (caractéristiques physiques, signes physiologiques ou pathologiques, survie ou mort, etc.). Récemment, les progrès de la physique et de la biologie moléculaire ont permis aux scientifiques d'explorer les différents niveaux d'organisation du vivant (des atomes aux molécules les plus complexes et des gènes aux organisations pluricellulaires). Cela a conduit les biologistes à développer une large palette d'outils leur permettant d'intervenir directement sur les gènes ou leurs produits, en empêchant par exemple l'expression d'un gène de façon transitoire ou permanente. La levure, organisme eucaryote unicellulaire, est l'exemple le plus marquant d'un modèle vivant transformé en véritable "laboratoire moléculaire".
Les biologistes ont pu ainsi observer de façon systématique l'effet de diverses mutations sur le comportement d'un organisme dans diverses conditions environnementales. C'est ce type de stratégies par perturbations génétiques qui a permis d'associer des gènes à des phénotypes particuliers dans différentes conditions environnementales.

Jusque dans les années 90, en raison des moyens technologiques dont ils disposaient alors, les biologistes moléculaires se sont concentrés sur l'analyse et l'étude du rôle d'un très petit nombre de gènes à la fois [1, 2], très souvent un seul gène. Pourtant on sait depuis, qu'une fonction biologique nécessite généralement la participation de plusieurs gènes organisés en réseaux. L'avènement de différentes techniques d'investigation à grande échelle, a permis d'observer des comportements cellulaires au niveau moléculaire. C'est grâce en particulier à la technologie des puces à ADN [3] qu'il est aujourd'hui possible d'obtenir des mesures quantitatives de l'expression de l'ensemble des gènes d'un organisme à un moment précis et dans une condition donnée. L'apport majeur des données transcriptomiques pour l'analyse des réseaux génétiques est lié à 3 caractéristiques principales :

- l'aspect quantitatif : apportant une grande précision d'observation du niveau de concentration de chaque ARNm, reflétant l'expression du gène correspondant corrigée par la stabilité de l'ARNm codé dans la population de cellules étudiées.
- leur production à grande échelle (celle du génome) : permettant l'observation de tous les gènes à la fois,
- la possibilité de produire des données dynamiques : plus proches des phénomènes biologiques et permettant d'identifier de potentielles relations (corrélations, causalité, etc.) entre les éléments observés.

La disponibilité de ces données, dites systémiques, a stimulé le développement de plusieurs

familles de méthodes d'inférence de réseaux de régulation allant de la fouille de données à la modélisation fine du comportement dynamique de petits réseaux d'interaction de gènes [4].

Dans le cadre de cette thèse, nous nous intéressons à la réponse transcriptionnelle de la levure *S. cerevisiae* à un stimulus génotoxique, de fortes doses de radiations γ. Nous avons pour objectif d'identifier les différentes formes potentielles de cette réponse et de reconstruire un réseau de régulation génique impliqué dans son contrôle. Nous disposons de trois sources d'informations biologiques différentes :

- une stratégie expérimentale de perturbations génétiques (7 souches de levures avec des combinaisons de caractéristiques génétiques différentes), permettant de tester différentes hypothèses de régulation,
- des cinétiques d'expression de gènes mesurées après irradiation pour approximativement tous les gènes de *S. cerevisiae* et dans toutes les conditions définies par la stratégie expérimentale,
- et des informations biologiques additionnelles, de sources et de natures différentes, extraites de la littérature ou des grandes bases de données biologiques.

Nous posons trois grands enjeux dans ce travail :

- exploiter la composante temporelle de ces données, et leur hétérogénéité, sans imposer un filtrage arbitraire des cinétiques d'expression de gènes
- tirer des règles de régulation générales à partir de l'analyse du comportement des gènes après irradiation dans toutes les conditions expérimentales à la fois,
- exploiter les différentes sources d'informations biologiques additionnelles pour permettre l'interprétation des profils de réponses identifiées et permettre aussi de compléter les relations de régulation déduites.

Nous sommes confrontés à deux grandes difficultés. Il s'agit d'une part de prendre en compte la structure particulière de chaque type d'information, que ce soit l'aspect dynamique de l'expression des gènes, l'aspect logique et combinatoire de la stratégie expérimentale ou l'aspect très hétérogène des données génomiques additionnelles. D'autre part, il s'agit d'exploiter l'ensemble de ces données au sein d'une seule méthodologie et d'obtenir en sortie un seul réseau de régulation.
Parmi les approches permettant l'inférence de réseaux de régulation de gènes à partir de données d'expression, les méthodes séquentielles, basées sur une réduction de la dimension des variables et sur l'intégration d'informations biologiques hétérogènes, nous semblaient les plus adaptées pour l'exploitation de nos données. Cependant, à notre connaissance, aucune de ces approches ne permet à la fois une analyse globale et dynamique de la réponse à un stimulus et la prise en compte d'une stratégie de perturbation pour formaliser et automatiser l'extraction de règles de régulation.

En nous appuyant sur cette problématique biologique particulière, nous avons défini une approche générale nommée *XRegPath*, qui exploite séquentiellement ces trois grands types d'informations biologiques en initiant le processus d'inférence de réseaux de gènes par l'analyse des données d'expression. Cette approche est constituée de 5 étapes hiérarchisées, dont une étape de déduction qui propose un cadre logique et automatisé pour la déduction de règles de régulation à partir d'une stratégie de perturbations et deux

étapes d'induction qui permettent pour l'une, d'identifier des ensembles de gènes cohérents du point de vue de leurs dynamiques d'expression et, pour l'autre, d'identifier des sous-ensembles de gènes cohérents du point de vue d'informations biologiques de natures diverses (annotations fonctionnelles, associations à des régulateurs transcriptionnels, etc.). Notre approche permet d'intégrer les différentes bribes de réseaux de régulation, extraites par les trois étapes d'inférences au sein d'un modèle global qui, sans prétendre à l'exhaustivité rend globalement compte de la dynamique d'une réponse cellulaire et des principaux mécanismes de régulation associés.

L'application de notre méthodologie à l'analyse de la réponse transcriptionnelle à l'irradiation chez la levure nous a permis d'identifier une réponse complexe, avec différentes caractéristiques fonctionnelles et différents mécanismes de régulation. Cette méthodologie nous a également permis de réunir toutes ces informations au sein d'un modèle de régulation global et dynamique, impliqué dans le contrôle de la réponse à l'irradiation. Certaines de nos observations ont pu être validées expérimentalement et confirment certaines relations au sein de notre modèle de réponse à l'irradiation.

Enfin, la méthodologie *XRegPath* a été implémentée sous la forme de deux logiciels complémentaires qui permettent la gestion de différents types d'informations biologiques et leur exploitation à travers les 5 étapes de notre méthodologie.

Organisation du document

Ce mémoire s'appuie sur des articles rédigés, soumis ou acceptés. Il s'organise en 4 chapitres :

- le chapitre 1 propose une courte introduction à la biologie moléculaire qui permet de définir les termes et notions de biologie utilisés dans ce manuscrit ainsi qu'une présentation des différents domaines de la bioinformatique et une introduction à la biologie des systèmes,
- le chapitre 2 décrit l'état de l'art dans le domaine de l'inférence de réseaux de régulation géniques en mettant l'accent sur l'exploitation des données du transcriptome,
- le chapitre 3 présente la problématique de ce travail, à la fois biologique et méthodologique,
- le chapitre 4 présente nos différents résultats sous la forme de deux articles : la conception et le développement de la méthodologie *XRegPath* et son application à l'analyse de la réponse à l'irradiation chez la levure avec une analyse détaillée des résultats obtenus.

Chapitre 1

De la biologie moléculaire à la biologie des systèmes

J'ouvre ce premier chapitre sur une introduction générale à la structure de la cellule et à sa physiologie. Cette introduction n'a pas prétention à être exhaustive ou à représenter l'étendue des connaissances actuelles mais plutôt à présenter simplement les termes et concepts en biologie qui seront utilisés dans la suite de ce manuscrit. Les termes les plus importants pour cette étude sont représentés en gras et les exemples cités concerneront le plus souvent la levure *S. cerevisiae* qui est le modèle vivant utilisé dans le cadre de nos travaux.

1.1 Du gène à la cellule

La **cellule** est la plus petite unité fonctionnelle d'un être vivant. Les organismes **unicellulaires** procaryotes (bactéries) ou **eucaryotes** (protistes), généralement microscopiques, sont composés d'une seule cellule, tandis que les organismes pluricellulaires (métazoaires) sont faits de nombreuses cellules réunies en ensembles spécialisés. Comparées aux cellules eucaryotes, les cellules procaryotes ont une structure très simple. Elles ne contiennent aucun organite et toutes les réactions biochimiques sont donc réalisées par des composants en solution dans le cytoplasme. De même, ne possédant pas de noyau, leur matériel génétique est libre dans le cytoplasme et forme un chromosome unique et circulaire. Chez les eucaryotes au contraire, le matériel génétique est composé de plusieurs chromosomes en bâtonnet et est enfermé dans un noyau délimité par une double membrane. Par ailleurs, les cellules eucaryotes contiennent de grandes surfaces membranaires permettant de compartimenter toutes les fonctions cellulaires au sein de structures spécialisées appelées organites.

1.1.1 Bases moléculaires de la cellule

Les organismes vivants sont majoritairement composés de molécules organiques même si de nombreux composés minéraux entrent aussi dans leur constitution et peuvent être indispensables à leurs fonctions physiologiques. Le fer par exemple est indispensable au fonctionnement de certaines enzymes comme la protéine Rnr2 chez la **levure** [5], im-

pliquée dans la biosynthèse des désoxyribonucléotides de l'**ADN**. Les 4 grandes familles de composés organiques sont les sucres (glucides), les protéines (protides), les graisses (lipides) et les **acides nucléiques** :

- Les glucides : ce sont des molécules très riches en énergie et sont principalement connues pour leur rôle de "carburant" de l'organisme. Les sucres peuvent aussi jouer un rôle structural (cellulose chez les plantes [6] et renforcement des membranes plasmiques chez toutes les espèces [7]), intervenir au niveau des fonctions immunitaires grâce à leurs propriétés antigéniques (groupes sanguins) [8] et participer à l'environnement local (les glucides sont des molécules très polaires).
- Les protides : appelés plus couramment **protéines** au niveau de la cellule. Elles sont constituées de polymères d'acides aminés et sont le produit de l'expression des **gènes** via la **transcription** et la **traduction**. On peut globalement attribuer 3 grandes fonctions aux **protéines** :

  1. Des fonctions structurales : les **protéines** interviennent au niveau des adhésions intercellulaires et dans l'adhésion de la cellule aux protéines de la matrice extra-cellulaire (voir revue dans [9]), mais elles interviennent surtout dans la constitution du cytosquelette [10] (filaments d'actine, microtubules, etc.) qui permet à la cellule de maintenir une forme caractéristique de son type et de son état physiologique [11, 12]. C'est ce cytosquelette qui confère aussi à la cellule ses propriétés dynamiques, que ce soit pour sa motilité ou pour le transport interne de petites molécules.

  2. Des fonctions enzymatiques au niveaux des réactions chimiques intervenant dans la quasi-totalité des voies métaboliques cellulaires et au niveau du transport de diverses molécules (protéines membranaires) [13].

  3. Au niveau de la signalisation cellulaire, que ce soit sous la forme d'hormones (par exemple les facteurs du ***mating-type*** (facteurs sexuels) chez la levure) [14, 15] ou en tant que récepteurs (fixation de virus [16], de toxines ou de substances chimiques [17]).

- Les lipides : comme les glucides ils peuvent représenter une réserve d'énergie pour l'organisme. Mais, leur rôle fondamental au niveau de la cellule consiste à séparer la matrice cellulaire du milieu extra-cellulaire par une bicouche lipidique fluide et dynamique [7].
- **Les acides nucléiques** : ce sont les macromolécules qui fournissent les informations nécessaires au développement et au maintien de la vie. Elles ont pour fonction la transmission du patrimoine génétique de génération en génération et le contrôle de la fabrication des **protéines** nécessaires à la vie. On distingue deux types d'acides nucléiques :
  - L'**ADN** (Acide DéoxyriboNucléique) qui est le support dépositaire de l'information génétique.
  - Les **ARN** (Acides RiboNucléiques) qui sont de plusieurs types : l'ARN messager ou **ARNm**, qui permet la conversion du code génétique des **gènes** en **protéines**, l'ARN de transfert ou **ARNt**, qui apporte les acides aminés nécessaires à une chaîne protéique en cours d'élaboration, l'ARN ribosomal ou **ARNr**, l'un des composants des **ribosomes**.

D'un point de vue chimique, les acides nucléiques sont des polymères géants obtenus par condensation de sous-unités appelées nucléotides. Un nucléotide est constitué d'une base azotée (Adénine (A), Guanine (G), Cytosine (C) ou Thymine (T). Pour les **ARN**, la Thymine est remplacée par l'Uracile (U)), d'un sucre (déoxyribose pour l'**ADN** ou ribose pour les **ARN**) et d'un groupe phosphate. C'est l'arrangement des nucléotides en une longue séquence qui encode l'information nécessaire à la construction des protéines.

1.1.2 Du génotype au phénotype

L'ensemble des allèles (différentes versions d'un même **gène** pour les organismes diploïdes) de l'ensemble des **gènes** d'un individu s'appelle le **génotype** et constitue le patrimoine génétique, héréditaire de tout individu. Les gènes, lorsqu'ils s'expriment, contrôlent la synthèse des **protéines** (voir figure 1.3) et, d'une façon plus générale contrôlent toutes les activités de la cellule. Le **phénotype** est l'ensemble des caractéristiques d'un organisme vivant ou d'un individu, que ce soit son apparence physique (couleur, taille, etc.), ou sa physiologie. Il est déterminé en partie par les gènes exprimés et en partie par l'environnement et le mode de vie. Dans un organisme pluricellulaire comme l'Homme, à l'exception de quelques cas particuliers, toutes les cellules possèdent le même génotype mais elles peuvent présenter des phénotypes différents qui dépendent de leur spécialisations. Pour des organismes plus simples comme la **levure**, une même cellule pourra présenter des phénotypes différents selon différents évènements physiologiques (cycle cellulaire, reproduction, etc.) et selon ses capacités d'adaptation aux variations de son environnement [12].

Le phénotype d'un individu ne se limite pas aux caractéristiques directement visibles à "l'oeil nu", il peut exister différents niveaux d'observation d'un phénotype, chacun rendant compte de caractéristiques différentes à des niveaux d'organisation différents. Nous pouvons commencer à observer un phénotype à partir du moment où les gènes d'un organisme (pluri- ou unicellulaire) sont exprimés. Actuellement, il est possible d'obtenir des mesures qualitatives et quantitatives pour l'ensemble des gènes exprimés, à un instant donné, au sein d'une population de cellules qui partagent le même génotype. Ce premier niveau d'observation est généralement obtenu avec la technologie des **puces à ADN** [3, 18] (voir section 1.3.1.2) et peut déjà permettre de différencier des génotypes différents ou alors de différencier, pour un même génotype, des états ou comportement cellulaires différents.
La relation plus ou moins directe entre génotype et phénotype est utilisée en biologie moléculaire pour identifier les gènes responsables des différentes caractéristiques d'un individu. L'approche classique des biologistes consiste à empêcher définitivement ou temporairement un ou plusieurs gènes de s'exprimer. Ils observent ensuite si ces perturbations ont un effet sur le phénotype et, en fonction de leurs observations ils déduisent si le ou les gènes perturbés sont responsables des variations phénotypiques observées [19]. C'est en utilisant des **stratégies de perturbations** systématiques que la plupart des gènes des organismes modèles ont pu être associés à des fonctions ou à des processus biologiques [20, 21, 22, 23, 24, 22].

1.1.3 Transcription et traduction de l'information génétique

Le "dogme central", introduit par Watson et Crick (les co-découvreurs de la structure de l'ADN) dans les années 50 [25, 26], stipule que dans tous les organismes vivants l'information soit transmise dans un seul sens : **ADN** → **ARNm** → **protéine** (voir figure 1.3). Le passage de l'information génétique de l'ADN à l'ARNm s'appelle la **transcription** et de l'ARNm à la protéine s'appelle la **traduction.**

Transcription des gènes en ARNm : il s'agit d'un phénomène biologique ubiquitaire qui consiste, au niveau de la cellule, à copier des régions dites codantes de l'ADN (gènes) en molécules d'ARNm. La transcription est réalisée par des complexes enzymatiques appelés des **ARN polymérases**. Chez les eucaryotes, trois sortes d'ARN polymérases sont responsables de la synthèse des 3 grands types d'ARN : la **polymérase I** pour la majorité des **ARNr**, la **polymérase II** pour les **ARNm** et la **polymérase III** pour les **ARNt** et certains **ARNr** [27, 28]. Pour transcrire un gène, une ARN polymérase va se fixer sur une séquence particulière, en amont du gène, appelée promoteur [29]. Ce promoteur diffère quelque peu selon les gènes, mais garde globalement des caractéristiques générales permettant de l'identifier. Lorsqu'elle se fixe sur le promoteur, l'ARN polymérase s'associe avec différentes protéines appelées **facteurs de transcription** (FT) pour former une particule d'initiation. Il existe des FT généraux et des FT spécifiques de certains gènes. La liaison entre l'ADN et l'ARN polymérase permet d'une part d'ouvrir la double hélice et d'autre part de catalyser l'insertion des ribonucléotides triphosphates pour former un brin d'ARN (voir figure 1.1). La synthèse de l'ARN est guidée par complémentarité d'appariement avec les bases azotées du brin transcrit de l'ADN. La zone de synthèse se déplace le long de l'ADN. Quand la transcription atteint la fin du gène, l'ARN est libéré et l'ADN se referme [30].

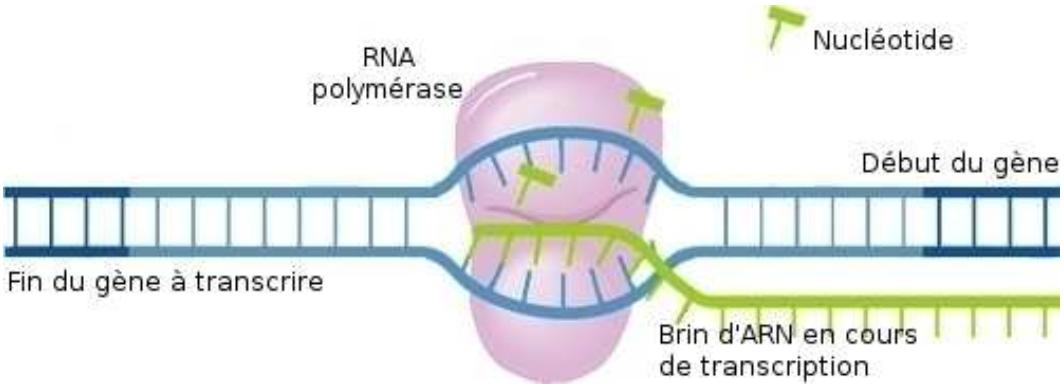


Figure 1.1 – Gène en cours de transcription.

Traduction des ARNm en protéines : le mécanisme de la traduction d'une séquence de codons (triplets de nucléotides) en une chaîne polypeptidique est complexe et implique un grand nombre d'étapes répétitives. La traduction se passe dans le cytoplasme, et plus particulièrement au niveau des **ribosomes** (voir figure 1.2). Les ribosomes sont constitués de deux sous-unités, une plus petite qui "lit" l'ARNm et une plus grosse qui se charge

de la synthèse de la protéine correspondante [31, 32]. Chaque sous-unité ribosomique est elle même constituée d'un complexe hétéromérique de protéines et d'ARNr qui portent l'activité catalytique de l'ensemble. L'information qui a été copiée à partir du gène est déterminée sur l'ARNm par les codons qui seront traduits selon le code génétique. On délimite le gène par un triplet initiateur qui code pour une méthionine (acide aminé qui commence le polypeptide) et par un triplet qui code pour un codon STOP (information qui provoque l'arrêt de la synthèse de la protéine). Le déroulement de la traduction est le suivant [33] : l'ARNm, tel un ruban défile dans le site du ribosome. Pendant ce temps, l'ARNt comportant l'anticodon complémentaire du codon d'ARNm en "cours de lecture" vient se placer en apportant l'acide aminé correspondant qui est relié au dernier acide aminé de la chaîne en cours d'élongation. Le ribosome va continuer à parcourir l'ARNm jusqu'à ce qu'il reconnaisse un codon STOP à l'aide de facteurs de terminaison, il se détache alors de l'ARNm et libère le peptide synthétisé. Un ARNm peut être traduit par plusieurs ribosomes à la fois. L'ensemble formé par un ARNm et les ribosomes qui le parcourent dessus s'appelle un polysome. Comme le phénomène de traduction a lieu plusieurs fois en même temps, il est fréquent de trouver des protéines identiques fabriquées à partir d'un même ARNm. On dit que la traduction est amplifiée.

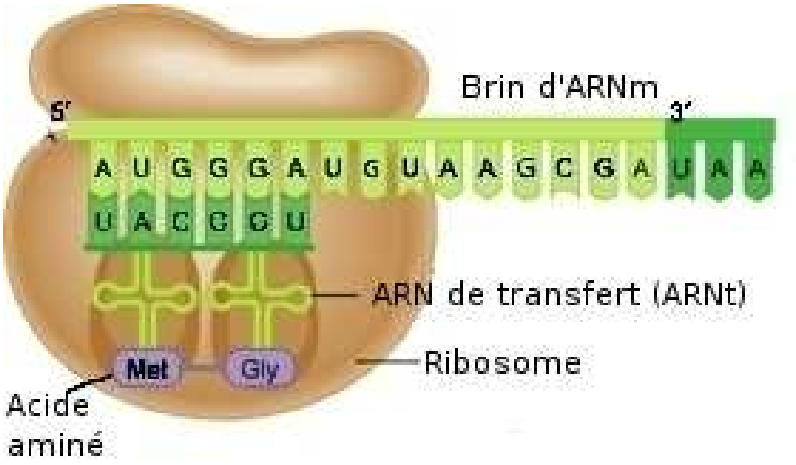


FIGURE 1.2 – Brin d'ARNm en cours de traduction.

Depuis le dogme introduit par Watson et Crick [25, 26], un certain nombre de phénomènes qui contredisent le dogme unidirectionnel, du gène vers la protéine, ont été décrits. Par exemple, un phénomène de transcription inverse a été mis en évidence chez les rétrovirus [34, 35] et les rétrotransposons [36, 37] qui transcrivent l'information génétique de l'ARN en une séquence d'ADN qui peut s'intégrer dans le génome de l'hôte. La découverte de ce phénomène et en particulier de l'enzyme qui en est responsable, la transcriptase inverse, a permis de développer la technique de ***RT-PCR*** (*Reverse Transcriptase Polymerase Chain Reaction*). Cette technique correspond à une transcription inverse (*RT*) des ARN en des fragments d'ADN simple brin suivie d'une étape d'amplifications itératives des fragments d'ADN. C'est le développement de la *RT-PCR* qui a permis de séquencer et de cloner les ARN ou d'en mesurer les quantités relatives à l'aide de **puces à ADN** même pour des transcrit très faiblement représentés.

1.1.4 Les différents niveaux de régulation : de l'expression des gènes à l'activité des protéines

Les gènes contiennent déjà toute l'information nécessaire à la vie de la cellule et c'est en grande partie à partir de cette information que le phénotype d'une cellule ou d'un individu s'exprime. Mais, dans un organisme pluricellulaire complexe, même si toutes les cellules de l'organisme contiennent le même patrimoine génétique, elles ne présentent pas le même phénotype que ce soit sur des critères physiques ou physiologiques. De plus, pour un organisme pluri- ou uni-cellulaire une même cellule pourra changer de forme ou connaître des perturbations de son fonctionnement en fonction des variations de son environnement. Ces variations de phénotypes sont les résultats de mécanismes de régulation au sein de la cellule opérant du niveau le plus en amont, l'expression des gènes, jusqu'au niveau le plus en aval, l'activité des protéines. La production d'une protéine active à partir d'un gène peut être régulée selon différents niveaux :

Régulation transcriptionnelle : le premier niveau de régulation de l'expression des gènes concerne l'accessibilité du matériel génétique. L'ADN peut exister sous une forme lâche appelée euchromatine, ou sous forme compactée, l'hétérochromatine. Sous cette dernière forme, les gènes ne sont pas accessibles aux **ARN polymérases** [38, 39]. Le deuxième niveau de régulation transcriptionnel fait appel à un mécanisme de contrôle beaucoup plus précis, basé sur des **séquences régulatrices**. En effet, les gènes des eucaryotes possèdent des séquences régulatrices, de 6 à 15 nucléotides, souvent présentes en amont du promoteur de ces gènes. Ces motifs d'ADN fixent de façon spécifique une ou plusieurs **protéines régulatrices** appelées **facteurs de transcription** (**FTs**) qui permettent d'activer ou d'inhiber l'expression du gène en aval [40, 41].

Régulation post-transcriptionnelle : chez la plupart des eucaryotes, les gènes existent sous la forme d'une succession de **parties codantes**, les **exons** et de **parties non codantes**, les **introns** (quasiment absents chez la levure) [42]. Lorsqu'un gène est transcrit en ARNm, celui-ci comporte la copie des parties exoniques et introniques du gène transcrit. Avant d'être traduit, l'ARNm devra être maturé par l'élimination de certaines parties introniques. Ce processus s'appelle l'**épissage** (voir figure 1.3), où des protéines peuvent se lier au transcrit primaire afin d'exciser des parties introniques [43]. En activant ou en inhibant la coupure de certains introns, ces protéines régulatrices peuvent influencer l'activité de la future protéine. La durée limitée de la vie de l'ARNm est également un facteur important dans la régulation de l'expression des gènes vers les protéines. La demi-vie d'un ARNm peut varier de quelques minutes, à quelques heures (majorité des ARNm). Il existe aussi des virus capables d'infecter des cellules par de courts fragments d'ARN viral double brin, contenant des séquences capables de s'hybrider avec certains ARNm de l'hôte, ce qui entraîne une destruction rapide de ces ARNm. Ce mécanisme, qui prive la cellule de certains de ses ARNm est connu sous le nom **d'ARN interférant** (**ARNi**). Il est actuellement instrumentalisé à grande échelle, par l'utilisation d'ARN double brin bien déterminés (sondes), pour perturber de façon transitoire ou définitive, la synthèse de protéines dont ont veut par exemple connaître la fonction [44].

Régulation de la traduction : la traduction d'un ARNm commence à partir du premier codon d'initiation rencontré. Cependant, un ribosome peut "sauter" un codon initiateur et modifier complètement la séquence en acides aminés de la protéine à traduire [45]. La détection du codon initiateur de la traduction par le ribosome dépend des codons voisins et de la présence de facteurs de traduction [46, 47]. La traduction d'un ARNm peut aussi être régulée par le propre substrat de la protéine à produire ; l'exemple le mieux étudié est celui du fer chez les mammifères. Le fer interagit avec des protéines qui se lient à des structures situées sur les ARNm codant pour la ferritine, la 5-aminolévulinate synthase et la transferrine [48]. Quand il y a peu de fer dans le cytoplasme, la traduction de ferritine et de 5-aminolévulinate synthase est diminuée alors que celle de la transferrine est augmentée.

Modifications post-traductionnelles et régulation de l'activité des protéines : pendant leur traduction (c'est-à-dire avant que la chaîne polypeptidique néosynthétisée ne soit relarguée par les ribosomes) ou quand elles sont entièrement synthétisées, certaines protéines subissent des modifications chimiques (voir revue dans [49]. Parmi celles-ci, on peut citer : les modifications impliquant l'addition d'un groupe fonctionnel (acétylation, alkylation, etc.), les modifications impliquant l'addition de groupes peptidiques ou de protéines (ubiquitination, sumoylation, etc.), les modifications changeant la nature chimique des acides aminés (citrullination, déamidation) et les modifications impliquant des changements structuraux (ponts di-sulfures et clivages protéiques). Ces modifications ont pour but de, réguler l'activité des protéines, de les "étiqueter" afin qu'elles soient reconnues par d'autres molécules ou par des systèmes de dégradation, les ancrer dans une membrane, les intégrer à une cascade de signalisation ou de les "adresser" à un compartiment cellulaire. La protéine ainsi modifiée adopte une structure et a des propriétés physico-chimiques très différentes de la molécule directement codée par le gène. Enfin, beaucoup de protéines ne sont fonctionnelles que sous forme d'hétéromères de peptides, résultant de la transcription de gènes différents.
C'est l'ensemble des interactions moléculaires et leurs dynamiques qui définissent le métabolisme d'une cellule. Les protéines sont capables d'interagir entre elles, de façon permanente, pour former des complexes dont l'activité est conditionnée par l'existence de différentes sous-unités protéiques, ou de façon transitoire pour activer l'une d'entre elles ou pour transmettre une information via une modification chimique. Leurs fonctions peuvent aussi dépendre d'interactions avec différents métabolites cellulaires. Ces interactions peuvent être transitoires, avec la transformation d'un substrat par une protéine, l'activation/inactivation d'une protéine, le transport d'un métabolite par une protéine d'un compartiment cellulaire à un autre (ou entre l'extérieur et l'intérieur de la cellule).

1.1.5 Adaptation et réponse cellulaire aux variations de l'environnement

Une cellule doit pouvoir s'adapter et faire face aux variations de son environnement. Généralement ces variations provoquent des stress physiques, chimiques ou des carences. Les réponses aux modifications de l'environnement sont relativement complexes et dépendent à la fois du type de stress et du type de cellule. Cependant le principe de ces

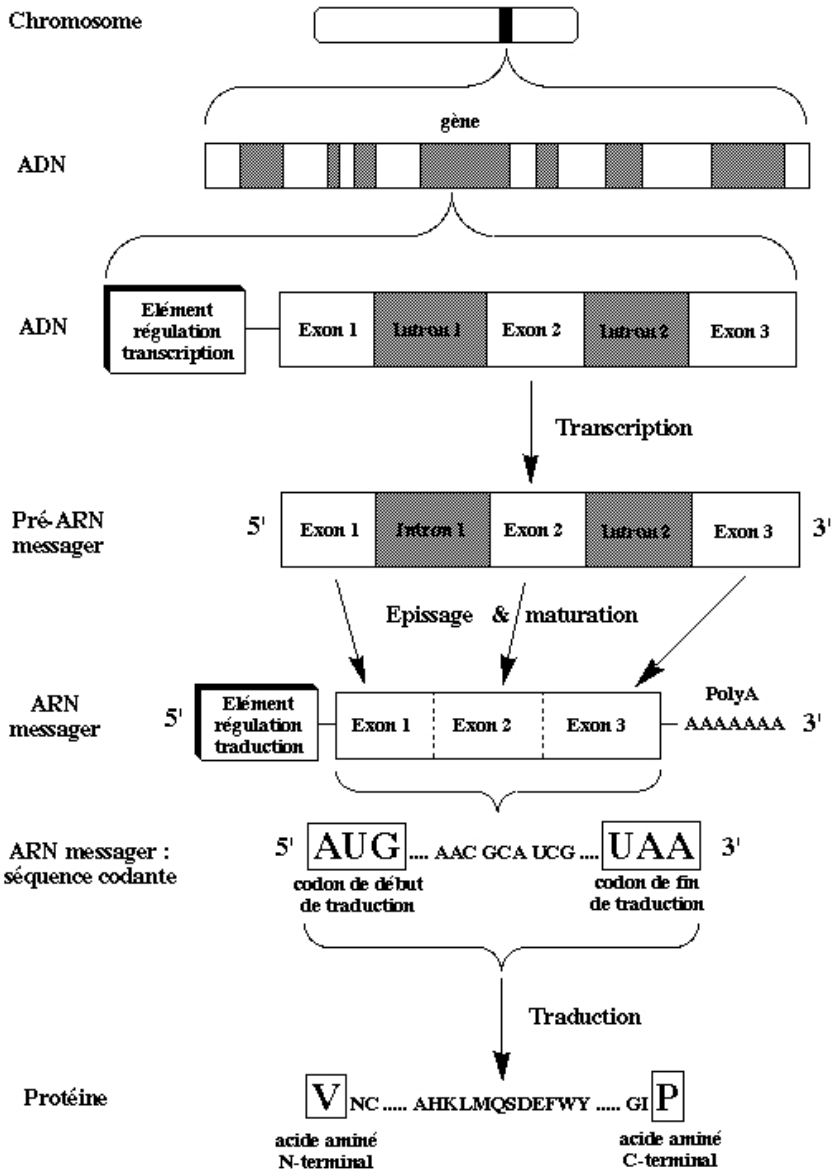


FIGURE 1.3 – Du gène à la protéine : structure, transmission et expression de l'information génétique (selon E. Jaspard (2004)).

réponses reste globalement le même avec des étapes de détection du stress, transduction de l'information et réponse cellulaire. La cellule peut détecter un changement de l'environnement directement, comme par exemple le gradient de concentration d'une molécule (chimiotaxie). Mais parfois, la cellule ne détecte que les conséquences des variations environnementales comme des dommages à l'ADN. Lors de mon travail de thèse, je me suis intéressé à la réponse de cellules de levure soumises aux effets de fortes doses de **radiations ionisantes**, connues pour leurs effets **génotoxiques**(induisant des lésions à l'ADN) [50, 51]. Dès que des dommages de l'ADN sont détectés, grâce à des complexes protéiques spécialisés, la cellule va arrêter la progression de son cycle cellulaire [52, 53, 54]. Cet arrêt du cycle est crucial car la cellule doit absolument réparer les lésions avant de

répliquer son génome et d'amorcer sa mitose sinon elle risque soit de propager des anomalies chromosomiques soit de rester bloquée en mitose sans pouvoir achever sa division. De nombreux complexes protéiques seront aussi recrutés pour réparer les lésions induites, avec des mécanismes spécifiques à chaque type de lésion [55]. Cette réponse cellulaire s'accompagne aussi d'une modulation de l'expression de nombreux gènes [56, 57, 58, 59, 60]. Nos résultats ont d'ailleurs révélé la mise en place d'un vaste programme de transcription qui réprimerait la transcription des gènes impliqués dans le métabolisme primaire de la cellule (métabolisme des sucres et des acides aminés) et favoriserait la transcription des gènes impliqués dans le renouvellement de composants cellulaires comme les ribosomes.

1.1.6 Choix d'un organisme modèle eucaryote : la levure *S. cerevisiae*

La **levure *Saccharomyces cerevisiae*** est un modèle **unicellulaire eucaryote** (voir figure 1.4) utilisé depuis de nombreuses années par les généticiens, les biologistes moléculaires et les microbiologistes. C'est aussi une espèce économiquement importante, puisqu'elle est utilisée par les industries agro-alimentaires. Le fait d'être un organisme unicellulaire simple, modèle pour la génétique classique, lui a permis aussi de devenir le premier organisme chez lequel on pu être posées les questions fonctionnelles de la biologie cellulaire. La levure a été le premier organisme eucaryote dans lequel il a été possible de réintroduire des gènes et il reste l'organisme chez qui il est le plus facile d'inactiver un gène par délétion ou de le remplacer par n'importe quelle construction d'ADN. Modèle de laboratoire, cellule simple pouvant se diviser à l'état **haploïde** (n chromosomes) et **diploïde** ($2n$ chromosomes), la levure a aussi la qualité d'avoir un génome compact quasiment exempt d'introns. C'est cette qualité qui en a fait le premier organisme modèle eucaryote à avoir son génome entièrement séquencé [61]. La connaissance du génome complet de cet organisme a ouvert de nouvelles dimensions à la recherche biologique et a permis la multiplication de banques de gènes, de mutants et de gènes fusionnés. Enfin, de très nombreuses protéines de *S. cerevisiae* ont des équivalents chez l'humain et, la plupart des maladies génétiques pour lesquels un gène humain a été impliqué ont un gène correspondant chez la levure. On peut donc souvent extrapoler à l'Homme les découvertes réalisées chez *S. cerevisiae*.
Le développement de techniques d'investigation moléculaires à grande échelle comme la technologie des **puces à ADN** [3] (voir section 1.3.1.2) a permis de passer du stade "un gène, une fonction" au stade "un génome, un réseau fonctionnel". C'est d'ailleurs chez la levure *S. cerevisiae* que la première puce à ADN contenant un génome complet a été construite [18]. Avec environ 6300 gènes (entre 20000 et 25000 gènes pour l'Homme), la levure représente un modèle eucaryote idéal pour ce type d'analyses systémiques. On considère aujourd'hui la levure comme un véritable "tube à essai cellulaire" et c'est ce qui a motivé le choix de cet organisme pour les travaux présentés dans ce document.
En un siècle environ, la biologie moléculaire est passée progressivement d'une science descriptive à une science analytique. Le développement de techniques d'investigation automatiques, précises et rapides a permis de générer d'énormes volumes de données et d'obtenir des descriptions précises des différents composants d'une cellule, des composés ioniques aux macromolécules. Le défi actuel consiste à identifier toutes les interactions possibles

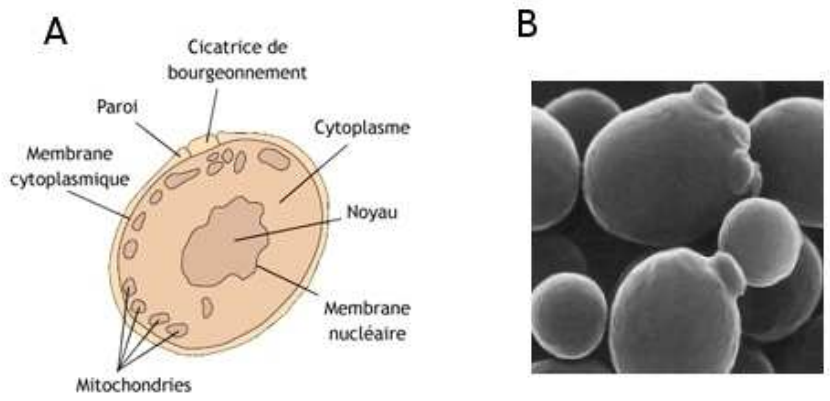


FIGURE 1.4 – Levure *S. cerevisiae*. A : représentation schématique d'une cellule de levure. B : image de levures obtenue par microscopique électronique à balayage.

entre les différents constituants de la cellule et surtout d'intégrer ces interactions au sein de modèles dynamiques pouvant décrire le comportement de la cellule dans différentes situations physiologiques ou pathologique, voire capables de simuler différents comportements cellulaires dans des conditions non testées expérimentalement. L'évolution de la biologie moléculaire a accompagné l'évolution de l'informatique en terme de capacité de stockage de l'information et de puissance de calcul. De nos jours, même avec la production d'énormes volumes de données biologiques, les principaux besoins ne résident plus dans la puissance des machines, mais plutôt dans le développement de méthodes de fouille de données et de modélisation capables d'analyser des données de plus en plus hétérogènes et complexes et d'en extraire des modèles dynamiques de plus en plus fins.

1.2 La bioinformatique

On regroupe généralement sous le terme de bioinformatique tous les outils et méthodes informatiques appliquées à la biologie. En pratique, ce que l'on entend par bioinformatique regroupe tous les moyens pour analyser, comparer, visualiser et distribuer les informations biologiques, ce qui nécessite à la fois des compétences en biologie ou médecine et dans au moins un des domaines suivants : informatique, mathématiques et statistiques. Bien que les champs d'application de la bioinformatique aillent de l'analyse du génome à la modélisation de l'évolution d'une population animale, cette discipline reste intimement liée à la biologie moléculaire et à la génomique. Ce sont en effet ces deux domaines de la biologie qui ont posé les premières problématiques nécessitant l'usage de l'informatique et qui ont aussi le plus œuvré au développement de la bioinformatique.

1.2.1 Brève histoire de la bioinformatique

Il n'est pas évident de définir avec précision quand a eu lieu le mariage entre biologie et informatique [62]. On estime que la bioinformatique est née au moment où Sanger a pour la première fois réussi à séquencer une protéine [63], en 1952. En effet, le séquençage des protéines s'est développé bien avant celui des séquences nucléiques. La méthode de séquençage de Sanger était relativement longue et fastidieuse mais elle a bénéficié de nombreuses améliorations jusqu'à devenir complètement automatique quelques années plus tard. Cette automatisation a généré un énorme accroissement de la quantité de données protéiques à analyser. Bien que le nombre de librairies protéiques augmentait rapidement, les ordinateurs de l'époque ne possédaient que peu de mémoire et peu de puissance de calcul. Rares étaient alors les scientifiques qui utilisaient un ordinateur. John Kendrew est le premier à avoir utilisé un ordinateur pour prédire en 1963 la structure 3-D de la myoglobine [64]. Le développement de langages de programmation adaptés aux applications scientifiques et faciles à apprendre a permis d'accélérer le développement de la bioinformatique. On peut citer en exemple Dayhoff, qui durant les années 60 s'était intéressée à l'évolution moléculaire et avait réussi à écrire des programmes en *FORTRAN*, qui permettaient de prédire des séquences d'acides aminés à partir de fragments peptidiques issus de l'hydrolyse des protéines.

Ce n'est que dans la deuxième moitié des années 70 que le séquençage des séquences nucléiques a été inventé. A l'époque, deux méthodes opposées se concurrençaient, celle de Gilbert, basée sur une dégradation chimique sélective, et celle de Sanger, basée sur une synthèse enzymatique sélective. C'est finalement la méthode de Sanger qui sera développée et automatisée. Lorsque les biologistes ont commencé à obtenir des séquences de gènes, des programmes similaires à ceux utilisés pour l'analyse de séquences protéiques ont commencé à être développés. Ces programmes d'analyse de séquences ont eu une influence cruciale sur le développement des librairies de séquences protéiques et génomiques. Dayhoff a par exemple conçu l'*Atlas of Protein Sequences*, une librairie qui projetait de contenir toutes les séquences connues de protéines. Cette librairie est devenue ce qui est aujourd'hui la base de données PIR (Protein Information Resource).
Les premiers programmes d'analyse de séquences ont été développés dans le cadre d'analyse phylogénétiques. On peut citer par exemple les travaux de Doolittle qui créa ses propres programmes pour rechercher des relations liées à l'évolution entre séquences [65, 66]. Les algorithmes de comparaison de séquences ont continué à se développer, et l'algorithme de recherche d'homologies à partir d'alignements de séquences conçu par Needleman and Wunsch [67] est devenu aujourd'hui un standard.
Une deuxième date importante dans l'histoire de la bioinformatique correspond à la création de la base de données *GenBank* au début des années 80. *GenBank* a pour mission principale le stockage d'informations sur les séquences d'ADN d'un grand nombre d'organismes. Puis, le développement de l'Internet a permis aux chercheurs du monde entier d'accéder librement à *GenBank*. Avec le développement des séquenceurs automatiques, les bases de données de séquences se sont développées et, à la fin des années 90, celles-ci doublaient de volume tous les 9 mois (voir figure 1.5). Avec le besoin de stocker l'information biologique, les algorithmes d'analyses de séquences se sont aussi développés pour permettre la recherche et la prédiction des régions codantes au sein de génomes complè-

tement séquencés.
Les deux derniers grands évènements marquant pour la bioinformatique correspondent à

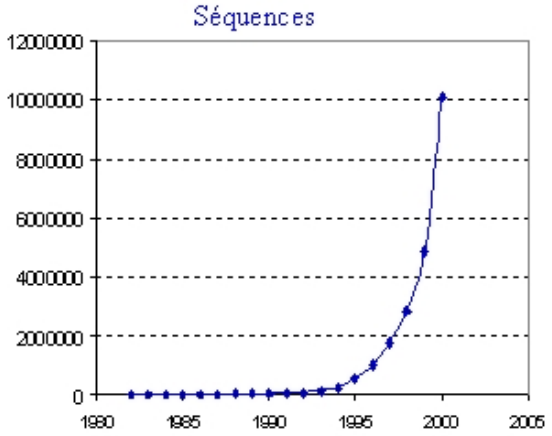


FIGURE 1.5 – Croissance de la banque *GenBank* en séquences nucléiques.

l'achèvement des programmes de séquençage des génomes de levure en 1996 (1er génome eucaryote séquencé) [61] et humain en février 2001 [68]. Les défis à relever consistaient alors à identifier tous les gènes et à leur associer des fonctions spécifiques (domaine de la génomique), à prédire la structure des protéines codées par ces gènes (domaine de la protéomique) et à comparer les rôles de certains gènes avec ceux d'autres espèces (en utilisant des outils du domaine de la transcriptomique comme les puces à ADN par exemple).
Les progrès et développements en bioinformatique ont été pendant longtemps principalement guidés par l'analyse des séquences protéiques et génomiques. Avec l'obtention de la séquence complète de nombreux génomes, la bioinformatique s'est diversifiée et spécialisée à chaque niveau d'organisation de la cellule et du vivant. Le développement de nouvelles techniques à haut débit comme les puces à ADN, la *Chromatin ImmunoPrecipitation* (*ChIP*) et les techniques d'identification de protéines partenaires ont orienté les chercheurs vers les domaines de la fouille de données et de l'apprentissage statistique. Et enfin, depuis quelques années, une nouvelles voie de recherche en bioinformatique, dite biologie systémique, propose une approches beaucoup plus intégrative et explicative que descriptive, où la cellule est vue comme un système complexe et dynamique et non comme la simple superposition de différents niveaux d'interactions statiques.

Les premiers résultats obtenus par des méthodes bioinformatiques avaient laissé les biologistes assez sceptiques à cause du manque de critères de validation et d'estimation de la fiabilité des informations produites. Mais, l'association des statistiques et la conception de protocoles expérimentaux *in silico* complets (hypothèses/contre hypothèses) ont permis à la bioinformatique de devenir une discipline scientifique à part entière, subdivisée en différentes spécialités. Durant les premières années de la bioinformatique, les chercheurs avaient du mal à se situer entre la biologie et l'informatique et trouvaient difficilement des journaux où publier leurs travaux. Les premiers articles de bioinformatique, essentiellement liés à l'analyse de séquences, ont commencé à paraître dans des journaux de biologie traditionnelle comme *Gene* et *the Journal of Molecular Biology*. En 1980, *Science* publiait

un des premiers états de l'art complet des méthodes bioinformatiques [69] d'analyse des séquences nucléiques. Aujourd'hui, de nombreux journaux de biologie "traditionnelle" publient régulièrement des travaux en bioinformatique et, les auteurs ont aussi le choix entre de nombreux journaux spécialisés dans la bioinformatique (par exemple, *Bioinformatics*, *BMC Bioinformatics*, *Journal of Computational Biology*, *Journal of Theoretical Biology*).

1.2.2 Les principaux domaines de la bioinformatique

1.2.2.1 Bases de données biologiques et outils de requête

Le développement des grandes banques de séquences généralistes telles que Genbank, l'EMBL ou DDBJ s'inscrit dans de grands projets internationaux. Celles-ci sont maintenant devenues indispensables à la communauté scientifique car elles donnent accès à des données et à des résultats essentiels dont certains ne sont plus reproduits dans la littérature scientifique. Devant la croissance quasi-exponentielle des données et l'hétérogénéité des séquences contenues dans les principales bases de séquences généralistes, d'autres bases spécialisées sont apparues.

Bases de données généralistes. Les grandes bases de données biologiques généralistes sont apparues dans les années 80 sous la tutelle de grands organismes publiques comme l'*EMBO* en Europe (*European Moleculary Biology Organisation*) qui a créé en 1981 la banque de séquences nucléiques *EMBL Data Library* [70, 71]. Du coté états-unien, soutenue par le *NIH* (*National Institute of Health*) la banque nucléique *GenBank* a été créée à Los Alamos [72, 73]. La collaboration entre ces deux banques a commencé relativement tôt. Elle s'est étendue en 1987 avec la participation de la *DDBJ*(*Dna Data Bank*) du Japon [74] pour donner naissance finalement en 1990 à un format unique dans la description des caractéristiques biologiques qui accompagnent les séquences dans les banques de données nucléiques [75].

Parallèlement, deux banques pour les protéines étaient créées : la première, la *Protein Identification Ressource* (*PIR-NBRF*) [76] est née sous l'influence du *National Biomedical Research Foundation* (*NBRF*) à Washington. La deuxième, *Swissprot* a été constituée à l'Université de Genève à partir de 1986 et regroupe entre autres des séquences annotées de la *PIR-NBRF* ainsi que des séquences codantes traduites de l'*EMBL* [77]. D'autres banques de séquences protéiques sont apparues ensuite comme *OWL*, banques composites de différentes bases protéiques ou *Genpept* et *TrEMBL* qui sont issues respectivement de parties codantes identifiées dans la base *GenBank* ou *EMBL*.
En 1990, une convention a été établie entre les trois banques nucléotidiques (*EMBL*, *Genbank* et *DDBJ*) pour permettre les échanges de séquences soumises à l'une ou à l'autre. Aujourd'hui, quelque soit la banque utilisée pour une requête, les résultats obtenus sont les mêmes et sont associés à des liens vers d'autres bases de données qu'elles soient nucléiques ou protéiques.

Bases de données spécialisées. Les bases de données spécialisées se sont constituées autour de thématiques biologiques ou caractéristiques biologiques précises comme le recen-

sement de familles de séquences impliquées dans les signaux de régulation, les promoteurs de gènes, les signatures peptidiques ou les gènes identiques issus d'espèces différentes. Elles peuvent aussi regrouper des classes spécifiques de séquences comme les vecteurs de clonage, les enzymes de restriction, ou toutes les séquences d'un même génome. Il existe actuellement plus de 250 bases de données d'intérêt biologique spécialisées outre les bases dont l'accès est limité (bases payantes ou propres à certains laboratoires).

Outils de requête. Les premières banques de données permettaient simplement aux utilisateurs de consulter leurs contenus, puis elles se sont structurées et développées pour offrir plus de fonctionnalités (recherche, saisie, modification des configurations du programme,...) et on parle maintenant plutôt de bases de données. L'exploitation des bases de données se fait via des logiciels dédiés à l'interrogation des bases de données biologiques (*ACNUC* ou *SRS*) qui sont programmés pour la manipulation des séquences biologiques et des informations qui leur sont associées. Les logiciels dédiés et en particulier *ACNUC* fonctionnent selon une construction de fichiers index représentant des critères de sélection (mot-clé, auteurs, espèces, revues, type de molécule...) et une organisation des fichiers permettant d'effectuer des liens entre critères ainsi qu'un langage de requête permettant l'utilisation d'expressions régulières.
Durant notre étude sur la levure, nous avons beaucoup utilisé la base de données *SGD* (*Saccharomyces Genome Database*, `http://www.yeastgenome.org/`), spécialisée dans l'annotation du génome de cette levure.

1.2.2.2 Analyse des séquences nucléiques et protéiques

L'analyse et la comparaison de séquences sont fondées sur la recherche d'un alignement optimal entre deux ou plusieurs séquences. Dans le cas de deux séquences, un alignement est un arrangement qui permet d'identifier les fragments où ces deux séquences sont similaires et ceux où elles diffèrent. Un alignement optimal sera évidement celui qui révélera le plus de similarités et le moins de différences. De façon générale, il existe 2 catégories de méthodes pour la comparaison de séquences :

- Méthodes d'alignement global : ces méthodes permettent l'alignement de séquences sur toutes leurs longueurs [67]. Les alignements globaux sont plus souvent utilisés quand les séquences mises en jeu sont similaires et de taille égale.
- Méthodes d'alignement local : ces méthodes permettent seulement des alignements sur de petites portions de séquences [78]. Les alignements locaux sont plus souvent utilisés quand deux séquences dissemblables sont soupçonnées de posséder des motifs semblables malgré l'environnement.

L'alignement multiple de séquences est une extension de l'alignement d'une paire de séquences à n séquences [79, 80]. Il permet d'aligner toutes les séquences d'un ensemble donné. L'alignement multiple est souvent utilisé pour l'identification de régions conservées à travers un ensemble de séquences soupçonnées d'avoir un ancêtre commun.

1.2.2.3 Phylogénie et évolution

L'évolution de la structure générale du génome conduit à des contraintes évolutives (composition en bases, vitesse d'évolution, par exemple) qui s'exercent simultanément sur

tous ou un grand nombre de gènes indépendamment de la fonction particulière de chaque gène. La phylogénie tente de reconstituer les filiations évolutives (arbres) aboutissant aux séquences étudiées. Elle permet, à partir de séquences alignées, la suggestion d'un arbre phylogénétique qui tente de reconstruire l'histoire des divergences successives durant l'évolution, entre les différentes séquences et leur ancêtre. Chaque nœud d'un arbre est une estimation de l'ancêtre des éléments inclus. Il faut toujours garder à l'esprit que l'on obtient toujours seulement une estimation de l'arbre. Cela revient à dire qu'en pratique les arbres sont imparfaits et que leur précision doit toujours être statistiquement établie. Le principe de base de toutes les méthodes consiste en :

1. Aligner proprement un ensemble de séquences (outil *CLUSTALW* [81] par exemple).
2. Appliquer des méthodes de génération d'arbres (méthode de parcimonie [82], méthode de vraisemblance [83, 84], méthode des distances [83, 84])
3. Évaluer statistiquement la robustesse des arbres (approches de type *bootstrap* [85, 86] ou *jacknife* [87]).
4. Représenter l'arbre consensus (outils *PHYLIP* par exemple *PHYLogeny Inference Package* `http://evolution.genetics.washington.edu/phylip.html`).

1.2.2.4 Structures des macromolécules biologiques

Le rôle principal de la bioinformatique structurale est de compléter des études expérimentales des structures par des études in silico. Que ce soit l'ADN, des protéines ou des complexes moléculaires homo- ou hétéromériques, la taille et la complexité des systèmes explorés constituent un véritable défi qui requiert des développements méthodologiques originaux [88, 89]. Ces développements vont consister en particulier à :

1. définir des "objets" simplifiés qui devront reproduire les propriétés physico-chimiques de groupements chimiques d'intérêt
2. prédire la déformabilité et la flexibilité des objets constituant les systèmes
3. prédire la stabilité des structures dans le temps
4. prédire l'interaction entre ces objets.

Selon l'échelle considérée, ces objets pourront donc correspondre à un atome ou à plusieurs dizaines voire centaine d'atomes. Les algorithmes de prédiction de structure et de modélisation dynamique de leurs comportement sont très coûteux en mémoire et en temps de calcul. Mais les performances en continuelle croissance des machines d'une part et des astuces méthodologiques d'autres part (prédiction de structures par homologie par exemple) ont permis de résoudre la structure de nombreuses molécules d'intérêt. On parle de structure 2D lorsque l'on s'intéresse à des motifs structuraux de base de type feuillets β ou hélices α, de structure 3D lorsque l'on s'intéresse à la structure en 3 dimensions d'une molécule et de structure 4D lorsqu'il s'agit d'assemblages de plusieurs polypetides en complexes fonctionnels (voir figure 1.6).

1.2.2.5 Analyse des données d'expression génique

Dans le prolongement du décryptage des génomes, le nouveau défi consistait à caractériser la fonction et l'activité de chaque gène identifié. Les biologistes se sont appuyés sur

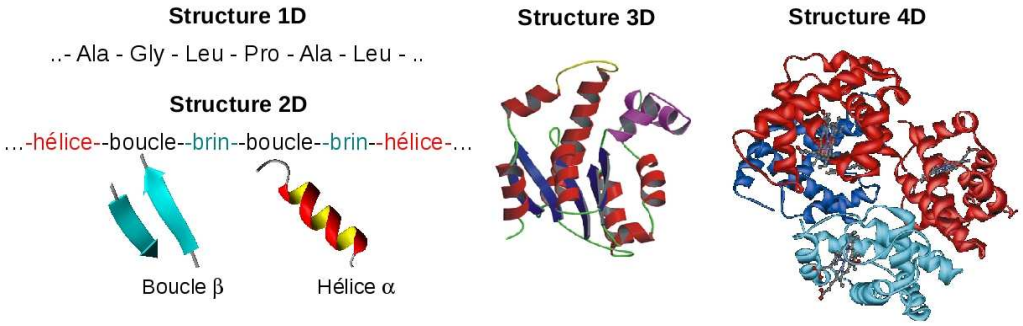


FIGURE 1.6 – Représentation des différents niveaux de structuration des protéines.

le développement et la mise en œuvre d'une variété de technologies sans cesse améliorées. Il s'agit notamment de technologies dites à haut débit telles que les puces à ADN [3] qui permettent d'étudier simultanément l'expression d'un grand nombre de gènes.
Afin d'éviter des redondances, les particularités de ce type de données, les outils d'analyse qui leurs sont dédiés et leurs différents champs d'application sont présentés dans la section 1.3.1.2, page 37.

1.2.3 Vers la biologie des systèmes

La recherche en biologie moléculaire et en génétique s'est pendant longtemps concentrée sur le rôle des gènes, des protéines ou de tout autres molécules, soit comme des entités isolées soit au sein de petits réseaux d'interaction [2, 1]. L'identification de toutes ces entités à l'échelle du génome reviendrait à lister tous les éléments d'un avion sans en comprendre la complexité et le fonctionnement [90]. Comprendre comment ces différents éléments sont assemblés pour former la structure d'un avion, reviendrait par analogie, au niveau de la cellule, à dessiner un diagramme complet d'un réseau de régulation de gènes et des différentes interactions biochimiques associées. De façon plus générale, cela revient à intégrer différents niveaux d'information pour comprendre comment fonctionne un système biologique : quelles sont ses différentes composantes, quelles sont les relations entre ces composants et comment se comporte l'ensemble, organisé en modules logiques. On appelle cette démarche **biologie des systèmes**. Plus qu'une discipline, la biologie des systèmes constitue une approche globale dans l'analyse du vivant.
Les origines de la biologie des systèmes remontent à la théorie des systèmes fondée sur l'hypothèse que tous les phénomènes peuvent être vus comme un réseau d'interactions entre différents éléments et que tous les systèmes peuvent être appréhendés par un même ensemble de méthodes [91, 92, 93, 94]. Les approches systémiques étaient déjà connues des biologistes, mais leur application à des systèmes biologiques complexes n'a été rendue possible que par le développement de moyens d'investigation à grande échelle pouvant produire des descriptions à différents niveaux d'organisation de la cellule et à l'échelle du génome.

Classiquement, adopter une approche systémique consiste à essayer de représenter une entité biologique physique et/ou logique sous la forme d'un modèle graphique ou ma-

thématique et de pouvoir prédire son fonctionnement dans des conditions non observées expérimentalement. Pour plus de clarté nous illustrerons la présentation de cette démarche par son application au niveau de la cellule. On peut considérer une cellule comme un système complexe et dynamique. La complexité est d'abord liée à la nature et au nombre de ses constituants (organites, protéines, métabolites, ADN, etc.). La complexité est ensuite liée à toutes les interactions possibles entre les éléments du système. Si le système possède plusieurs niveaux d'organisation (génome, protéome, etc) chaque niveau est à la fois organisé en réseau (interactions horizontales) et est en relation avec les niveaux supérieurs et inférieurs (interactions verticales). La cellule est un système dynamique car elle possède un cycle de vie (croissance, division, mort) et la capacité de s'adapter ou "répondre" à des variations de son environnement.

Pour modéliser la cellule et son fonctionnement, il est important de savoir à quel niveau de granularité on veut s'intéresser, si l'on souhaite représenter une cellule comme un ensemble d'organites spécialisés dans des fonctions vitales ou si l'on veut représenter tous les niveaux d'organisation de la cellule, du génome jusqu'à sa structure "visible" (phénotype) en passant par toutes les interactions moléculaires sub-cellulaires. Une approche typique de la biologie des systèmes, appliquée au niveau de la cellule, peut être formulée de la façon suivante :

1. Réunir toutes les informations disponibles sur les niveaux d'organisation que l'on estime important au fonctionnement du système que l'on souhaite modéliser. Et, à partir de ces informations, construire un premier modèle de son fonctionnement. Ce modèle peut être descriptif, graphique ou mathématique.
2. Définir tous les éléments du système à l'aide des outils expérimentaux disponibles, de l'identification de toutes les séquences des gènes impliqués jusqu'aux analyses du transcriptome, protéome, métabolome, interactome, etc.
3. Appliquer une stratégie de perturbations génétiques aux éléments centraux du système, généralement les gènes, (*knock-outs*, répression, etc.) dans un ou plusieurs environnements d'intérêt. Réunir des informations à partir de tous les niveaux possibles de régulation et de fonctionnement. A ce niveau, nous avons en notre possession des informations décrivant l'état statique du système. Pour analyser la dynamique du système (développement, cycle cellulaire ou phénomènes transitoires) il faut obtenir des informations cinétiques (cinétiques d'expression de gènes, cinétiques de concentrations de protéines, etc.).
4. Et enfin, intégrer toutes les informations réunies et extraites pour valider et/ou compléter le modèle de départ. Les éventuelles divergences ou zones d'ombres pourront susciter de nouvelles expériences et de nouvelles boucles *exploration/analyse* → *modélisation* → *simulation* → *expérimentation* → *exploration/analyse*. Ainsi, l'approche typique utilisée en biologie des système est à la fois itérative, intégrative et basée sur la vérification d'hypothèses.

L'intégration des informations peut initialement être "manuelle" et graphique. Cependant, on peut essayer de formaliser le modèle en des termes mathématiques et ultimement réussir à prédire le comportement du système dans des conditions qui n'ont pas encore été

évaluées expérimentalement. On peut aussi essayer de redéfinir le système afin de produire de nouvelle propriétés émergentes.

A mesure que l'on s'éloigne du génome pour s'intéresser à des niveaux d'organisation supérieurs, apparaît une hiérarchie d'informations hétérogènes : *ADN* → *ARN* → *protéine* → *modules fonctionnels (complexes protéiques intervenant dans un processus biologique particulier) et autres interactions moléculaires (protéines-protéines transitoires, protéines-ADN, etc.)* → *réseau de modules au sein d'une cellule* → *réseau de cellules formant un tissu ou un organe* → *individu (organisme pluricellulaire supérieur)* → *population d'individus* → *écosystème*. Le point important en biologie des systèmes est que pour l'analyse d'un niveau particulier d'organisation on aura besoin de recueillir les informations décrivant tous les niveaux d'organisation inférieurs et d'intégrer ces informations en un modèle représentant le plus fidèlement possible le niveau d'organisation étudié. Cela nécessitera alors de recueillir des données au niveau du génome, du transcriptome, de l'interactome (protéome et régulome) et du phénome (informations sur la croissance ou la viabilité de cellules en réponse à des perturbations génétiques et/ou environnementales). Il est important de toujours garder un lien entre les analyses phénotypiques et l'information génétique. C'est cette intégration transversale qui permet d'associer un réseau de régulation génique au fonctionnement d'un système biologique et qui permet de prédire le comportement du système, en fonction de perturbations génétiques simulées.

1.3 Données systémiques et réseaux biologiques

Différentes technologies peuvent être utilisées pour produire des données à l'échelle du génome ou à l'échelle d'un système biologique pour théoriquement tous les composants d'une cellule (ADN, protéines, métabolites etc.). Ces données permettent d'obtenir différents angles de vue et différents niveaux de granularité sur le fonctionnement interne de la cellule. Cependant, cette abondance d'informations présente aussi différents obstacles dont le principal correspond à l'extraction d'une information interprétable à partir du foisonnement de ces données. Néanmoins, le développement de méthodes de fouille de données et d'apprentissage automatique a permis aux biologistes d'exploiter et d'intégrer les différents types de données systémiques pour une meilleure compréhension de nombreux systèmes biologiques. Une autre difficulté consiste à intégrer en un même modèle plusieurs vues du système, à différents niveaux de granularité. En étudiant les relations et les interactions entre différentes parties du système biologique (organites, cellules, systèmes physiologiques, réseaux de gènes et de protéines), le biologiste pourra extraire un modèle de fonctionnement de la totalité du système.

1.3.1 Mesures systémiques des composants cellulaires

1.3.1.1 Génome

La génomique ou étude de toutes les séquences du génome et de l'information qui y est contenue est la mieux étudiée. Depuis 1995, environ 300 projets de séquençage de

génomes modèles parmi les trois grands règnes du vivant ont été achevés et depuis, des centaines d'autres ont été entrepris. La séquence d'ADN permet d'effectuer directement différents types d'analyses comme la recherche de motifs de régulation en amont des gènes, d'effectuer des analyses comparées entre espèces et d'améliorer notre compréhension des mécanismes d'évolution.

Au delà de la simple séquence, l'annotation des génomes permet de définir les portions d'ADN codant pour des protéines ou des ARNs fonctionnels et permet d'y associer des éléments de régulation. En plus de l'annotation des gènes, dernièrement de nombreux efforts ont été réalisés pour développer des outils permettant l'identification automatique de sites de fixation aux facteurs de transcription (FTs) et ainsi relier une information de structure (fragment de séquence) à une information liée à la régulation et à la fonction des gènes.

Toutes les informations et liens croisés concernant les génomes de la plupart des organismes modèles sont disponibles librement dans les bases de données décrites dans la section 1.2.2.1.

1.3.1.2 Transcriptome

L'étude du transcriptome correspond à l'analyse de l'expression des gènes dans différents contextes, physiologiques, pathologiques ou en réponse à des stimuli. Cette analyse passe par l'identification des gènes dont l'expression est modulée et par une mesure qualitative de la quantité relative d'ARNm transcrits à un instant donné. Ici, l'ARNm est simplement un intermédiaire sur la voie de production des protéines. L'analyse de la variation du contenu en ARNm donne une bonne indication sur l'état physiologique d'une cellule, voire d'un tissu ou d'un organisme et reste globalement corrélée à la variation du contenu cellulaire en protéines. L'analyse du transcriptome a pris une importance considérable ces dernières années et a permis une renaissance de la biologie des systèmes lorsque les premières données d'expression à grande échelle ont été disponibles.

Depuis la première démonstration de leur utilisation en 1995 [3], les biopuces (*microarrays*) ont révolutionné la biologie des systèmes par l'analyse du transcriptome en permettant l'analyse des processus biologiques à l'échelle du génome. Une biopuce, est constituée de fragments d'ADN ou d'oligonucléotides immobilisés sur un support solide selon une disposition ordonnée. Son fonctionnement repose sur le même principe que des technologies telles que le *Southern blot* ou le *Northern blot*, utilisées pour détecter et quantifier la présence d'une séquence nucléique spécifique au sein d'un échantillon biologique complexe, par hybridation à une sonde de séquence complémentaire portant un marquage radioactif. La confection des puces à ADN a permis d'étendre ce principe à la détection simultanée de milliers de séquences en parallèle. L'hybridation de la puce avec un échantillon biologique, marqué par un radio-élément ou par une molécule fluorescente, permet de détecter et de quantifier l'ensemble des cibles qu'il contient en une seule expérience. Il existe 3 grands types de puces :

- Les *macroarrays* ou filtres à haute densité : les dépôts (sondes) sont des clones d'ADNc ou des produits de PCR fixés à haute densité sur une membrane de ny-

lon. Le marquage est le plus souvent radioactif et le criblage est réalisé en excès de cible, on obtient ainsi une mesure de l'abondance relative de chacun des ARNm présent dans l'échantillon de départ. Actuellement, les *macroarrays* ont cédé la place aux puces (*microarrays*) permettant une meilleure mesure de l'expression d'un plus grand nombre de gènes en un minimum de temps.

- Les puces à ADN, classiquement à double fluorescence (voir figure 1.7) : elles permettent de fixer plusieurs milliers de sondes sur des lames de verre préalablement traitées chimiquement. Les sondes sont des fragments d'ADN simple brin de 200 à 2 000 paires de bases (pb) amplifiés par *PCR* ou de longs oligonucléotides (40-60 pb). Les cibles utilisées sont réalisées par transcription inverse, à partir d'ARN total ou messager utilisant deux fluorochromes différents (classiquement Cy3 et Cy5), ce qui permet d'hybrider simultanément deux cibles sur une même sonde. Les signaux d'hybridation sont analysés grâce à un lecteur capable de discriminer les deux fluorochromes et de générer deux images dont les niveaux de luminosité représentent l'intensité de la fluorescence lue. L'un des avantages de cette analyse comparée, repose sur le fait que le rapport Cy3/Cy5 n'est pas influencé par la qualité de la goutte déposée par le pipetteur robotisé. On calcule ensuite le ratio des intensités de fluorescence pour chacun des *spots*, ce qui permet de rechercher une expression différentielle des gènes dans les deux échantillons biologiques étudiés.

- Les puces à oligonucléotides : les puces les plus couramment utilisées sont les puces de la société *Affymetrix*. Dans ce cas, les sondes sont des oligonucléotides courts synthétisés *in situ* par une technique de photolithographie. Dans ce procédé, une lumière dirigée sur des sites spécifiques de la puce active la réaction d'oligo-synthèse. On peut synthétiser jusqu'à 300 000 oligonucléotides représentant 30 000 gènes sur une puce d'une surface d'environ 1 cm2. Une puce à ADN destinée à des études d'expression contient pour chaque gène un ensemble d'oligonucléotides mimant la séquence du gène, souvent choisis dans sa région 3', réduisant ainsi les risques d'hybridations croisées avec des séquences homologues de ce gène. Des oligonucléotides, dont la séquence varie pour une seule base, sont également ajoutés, ce qui permet de confirmer que le signal obtenu pour chacun des gènes est bien spécifique. On hybride une seule cible par puce et l'intensité de fluorescence mesurée par un scanner permet de mesurer l'abondance relative de chacun des ARNm présent dans l'échantillon biologique étudié.

Il existe d'autres façons de mesurer la quantité en ARNm ou les variations d'expression des gènes, mais que ce soit par *northern blots*, *RT-PCR*, *SAGE* ou tout autre technique déjà utilisée en biologie moléculaire, aucune ne permet de produire des mesures d'expression à l'échelle du génome de la qualité de celles réalisées par les puces à ADN.

En fonction de la nature de l'expérience et du type de question biologique de départ,

ont peut utiliser les informations obtenues à partir de puces à ADN pour différentes applications :

- analyse différentielle du transcriptome entre condition normale et pathologique [96, 97, 98, 99]

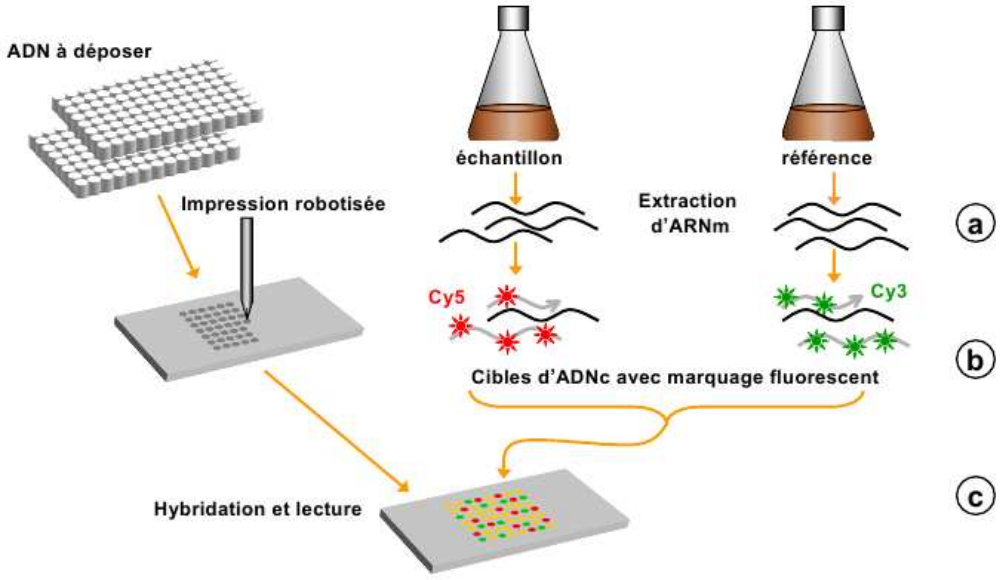


FIGURE 1.7 – Principe des puces à ADN de type "double fluorescence" (d'après [95]). (a) Extraction des ARNm des prélèvements de cultures cellulaires à analyser. (b) Transcription inverse et marquage des cibles avec des nucléotides fluorescents. On obtient pour chaque culture des cibles d'ADN complémentaires (ADNc) représentatives de l'ensemble des gènes exprimés. (c) Hybridation simultanée des échantillons marqués sur une même puce. Puis, mesure et analyse du signal différentiel pour chaque gène.

- identification de gènes marqueurs, spécifiques d'une maladie [100, 101]
- analyse de l'expression des gènes d'un système modèle [102, 103, 104]
- identification de motifs d'expression spécifiques de pathogènes [105, 106]
- analyse de la réponse transcriptionnelle à un stimulus, à un traitement médical ou l'identification d'un effet dose-dépendant [107, 108, 109]
- analyse toxicogénétiques où l'on va rechercher des motifs d'expressions dans un organisme modèle ou dans un tissu humain que l'on va utiliser comme prédicteurs préventifs de l'exposition à des environnements dangereux [110, 111, 112]
- du criblage médicamenteux [113, 114, 115]

En fonction de l'objectif, l'analyse des données issues des puces à ADN pourra être réalisées à différents niveaux de complexité : au niveau du simple gène, où l'on va rechercher quels sont les gènes qui sont différentiellement exprimés entre une situation contrôle et une ou plusieurs situations expérimentales [116], au niveau du groupe de gènes, où l'on va soit chercher à classer les gènes dans des classes connues (classification supervisée) [117], soit chercher à identifier de nouvelles classes de gènes au sein des données (classification non supervisée, voir section E.2) [118], et enfin, au niveau des systèmes ou du génome, où l'on va essayer d'identifier un réseau de régulation de gènes pouvant expliquer le contrôle des motifs d'expression observés (voir section 2).
Comme pour séquences protéiques et nucléiques, il existe de grandes bases de données d'expression publiques. Les trois principales bases de données issues de biopuces sont :

la *Stanford Microarray Database* http://genome-www5.stanford.edu/, *ArrayExpress* http://www.ebi.ac.uk/microarray-as/ae/ gérée par l'*EBI* et la *Gene Expression Omnibus* (*GEO*) gérée par le *National Institute for Biotechnology Information* (*NCBI*) http://www.ncbi.nlm.nih.gov/geo/.

1.3.1.3 Protéome

La protéomique a pour but d'identifier et de quantifier les niveaux cellulaires de toutes les protéines codées par le génome. Les analyses basées sur des données protéomiques sont historiquement postérieures aux analyses basées sur le transcriptome. C'est l'automatisation de méthodes de détection directes et indirectes comme l'électrophorèse sur gel 3D, la séparation de peptides par chromatographie, l'utilisation de protéines de fusion, les méthodes colorimétriques et la spectroscopie de masse qui permet de générer des informations protéomiques à l'échelle du génome.

Les protéomes de nombreuses structures cellulaires et organites comme le cytosquelette et la mitochondrie ont été produits [119]. Mais les efforts actuels visent à développer de nouvelles technologies qui permettraient de mieux caractériser les relations entre protéome et phénotype cellulaire et entre protéome et transcriptome pour différents processus cellulaires et dans différentes conditions expérimentales [120]. En particulier, une des voies développées consiste à améliorer la détection de protéines moins bien représentées dans les échantillons biologiques en ciblant la détection uniquement sur les peptides caractéristiques de chaque protéine ou de chaque isoforme. De nombreux laboratoires publics et compagnies privées développent actuellement des systèmes de puces à protéines [121, 122, 123]. Le fonctionnement de ces systèmes, qui permettraient idéalement d'identifier en parallèle plusieurs protéines au sein d'un mélange, est fondé sur la réaction antigène-anticorps entre les protéines à identifier et des anticorps spécifiques fixés à la surface de la puce.

L'intérêt des analyses protéomiques est qu'elles se basent sur le produit final de l'expression des gènes plutôt que sur un intermédiaire comme l'ARNm. Certaines techniques sont même capables de détecter des modifications post-traductionnelles des protéines [124, 125] (phosphorylations et glycosylations par exemple), des complexes protéiques [126, 127] et dans certains cas fournissent même des informations sur la localisation des protéines [128]. Ces informations ne pouvant pas être révélées par les techniques de mesures du transcriptome. Il faut cependant noter que pour l'instant ces techniques sont limitées par leur capacité de détection et ne donnent qu'une vue partielle du protéome. Les mesures du transcriptome et du protéome restent cependant complémentaires et de nombreuse analyses systémiques ont développé des modèles intégrant ces deux niveaux d'information.

Les données liées au protéome se présentent sous différentes formes. Pour accéder aux séquences protéiques de la plus part des organismes modèles on peut se référer aux bases de données généralistes présentées dans la section 1.2.2.1. Pour accéder aux familles et domaines protéiques on pourra se référer aux bases *InterPro* (http://www.ebi.ac.uk/interpro/), *Pfam* (http://www.sanger.ac.uk/Software/Pfam/) et *ProDom* (http:

//prodom.prabi.fr/prodom/current/html/home.php). Et enfin on trouvera un grand nombre de structures protéiques dans la bases *PDB* (*Protein Data Bank*, http://www.rcsb.org/pdb/home/home.do).

1.3.1.4 Métabolome

L'analyse du métabolome consiste à essayer d'identifier l'ensemble des petites molécules et métabolites (intermédiaires métaboliques, hormones et molécules signal et métabolites secondaires) qui peuvent être trouvées dans un échantillon biologique [129]. Une voie de recherche plus précise consiste à étudier la dynamique de la réponse métabolique aux stimuli environnementaux ou à des perturbations génétiques. Le métabolome peut être considéré comme le résultat de l'intégration au niveau de la cellule du transcriptome, du protéome et de l'interactome (voir 1.3.2) et ainsi ne fournit pas seulement une liste des métabolites cellulaires mais aussi un compte-rendu de l'état physiologique de la cellule [130, 131]. Comme pour les données de protéomique, les méthodes utilisées pour produire ces données sont en cours de développement et nécessitent encore des mises au points techniques. Ces méthodes consistent à analyser le contenu métabolique d'extraits cellulaires à l'aide de la spectroscopie de masse, de la spectroscopie par résonance magnétique nucléaire ou de la spectroscopie vibrationnelle.
Certaines études visent à cibler des familles de métabolites particulières ; on peut parler alors d'analyse du glycome [132, 133, 134] ou du lipidome [135, 136, 137, 138]. Ces informations peuvent avoir une importance capitale dans l'étude de maladies d'origine hormonale pour le lipidome et dans le cas du diabète pour le glycome.
Étant donné la grande diversité des molécules à détecter et les différentes échelles de dynamiques de concentration, les techniques doivent être capables de détecter de façon continue, plusieurs centaines d'espèces chimiques différentes [139]. Malgré ces difficultés et ces limitations, l'analyse du métabolome trouve des applications pour l'analyse de l'état cellulaire de différents systèmes biologiques [140, 141, 142] et dans des domaines comme l'ingénierie métabolique [143, 144, 145, 146], la pharmacologie [147, 148] et les études en nutrition humaine [149, 150].

La "méta-base" *BioCyc* (http://biocyc.org/) contient des informations sur le métabolome de la bactérie *E. coli* (*EcoCyc*) et sur un métabolome "multi-organismes" (*MetaCyc*) correspondant à une compilation de 1100 voies métaboliques obtenues à partir de 1500 organismes différents. BioCyc propose aussi des liens vers des bases de métabolomes de plus de 20 espèces animales et végétales différentes. On peut aussi accéder aux données produites grâce à différents projets de recherche comme le projet du métabolome humain (http://www.hmdb.ca/) et le projet du métabolome végétal (http://csbdb.mpimp-golm.mpg.de/).

1.3.1.5 Localisations sub-cellulaires des protéines ou *localizome*

Le *localizome* (terme introduit pour la première fois à ma connaissance par Kumar *et coll.* [151]) correspond à l'identification de la localisation sub-cellulaire de l'ensemble des protéines (voir figure 1.9). Ce type de données peut compléter nos informations sur la fonction des protéines et sur leurs éventuelles interactions [152]. La production de ces

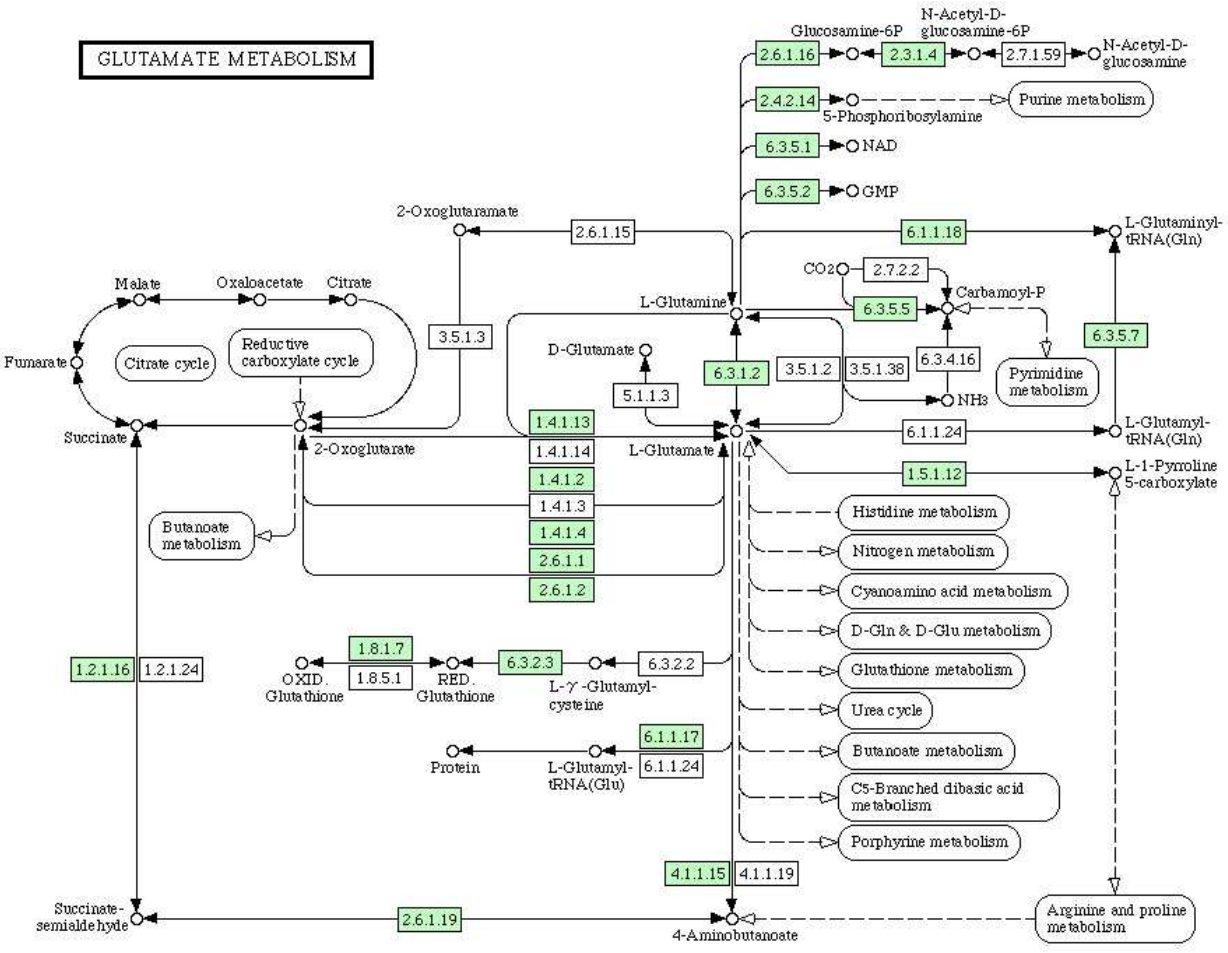


FIGURE 1.8 – Représentation du réseau métabolique du glutamate. En vert, les enzymes spécifiques de *S. cerevisiae* (source : *KEGG*).

données reste une des plus difficile parmi tous les types de donnés systémiques dans la mesure où chaque protéine doit être marquée ou détectée par un anticorps spécifique ou par une fusion avec une protéine fluorescente, visualisée en microscopie et attribuée, souvent manuellement, à une région sub-cellulaire ou à un organite particulier [153]. Malgrè ces difficultés, quelques *localizomes* complets ont été produits pour *S. cerevisiae* [151, 128] et *C. elegans* [154]. Il n'est pas étonnant que ces données aient été d'abord générées pour ces deux organismes modèles pour lesquels existe toute une palette de techniques en biologie moléculaire. Actuellement, des avancées techniques au niveau de la cytométrie par balayage laser et au niveau de "puces à tissus" devraient rendre obsolètes les techniques cyto- et histochimiques traditionnelles. Ces nouvelles méthodes à "haut débit" permettent de visualiser la présence de différents types de molécules dans différents échantillons cellulaires ou tissulaires.

Ces données de localisation ont déjà suscité le développement de méthodes bioinformatiques spécifiques. Par exemple, Guda et Subramaniam [155] ont analysé les occurrences

de certains domaines protéiques et la composition en acides aminés de protéines eucaryotes pour différentes localisation sub-cellulaires. A partir de ces informations ils sont en mesure de proposer d'assez bonnes prédictions (68-87% de vrais positifs pour une précision de 96-99%) pour la localisation de toute nouvelle protéine eucaryote dans un des 9 compartiments cellulaires qu'ils ont défini (cytoplasme, réticulum endoplasmique, protéines extra-cellulaires/sécrétées, appareil de Golgi, lysosomes, mitochondries, noyau, membrane plasmique et peroxisomes).

La principale source de donnée de *localizome* disponible en accès libre correspond au projet de localizome de *C. elegans* (`http://localizome.dfci.harvard.edu/`) et de la levure (`http://yeastgfp.ucsf.edu`). Pour les autres espèces il faudra se référer aux publications citées ci-dessus.

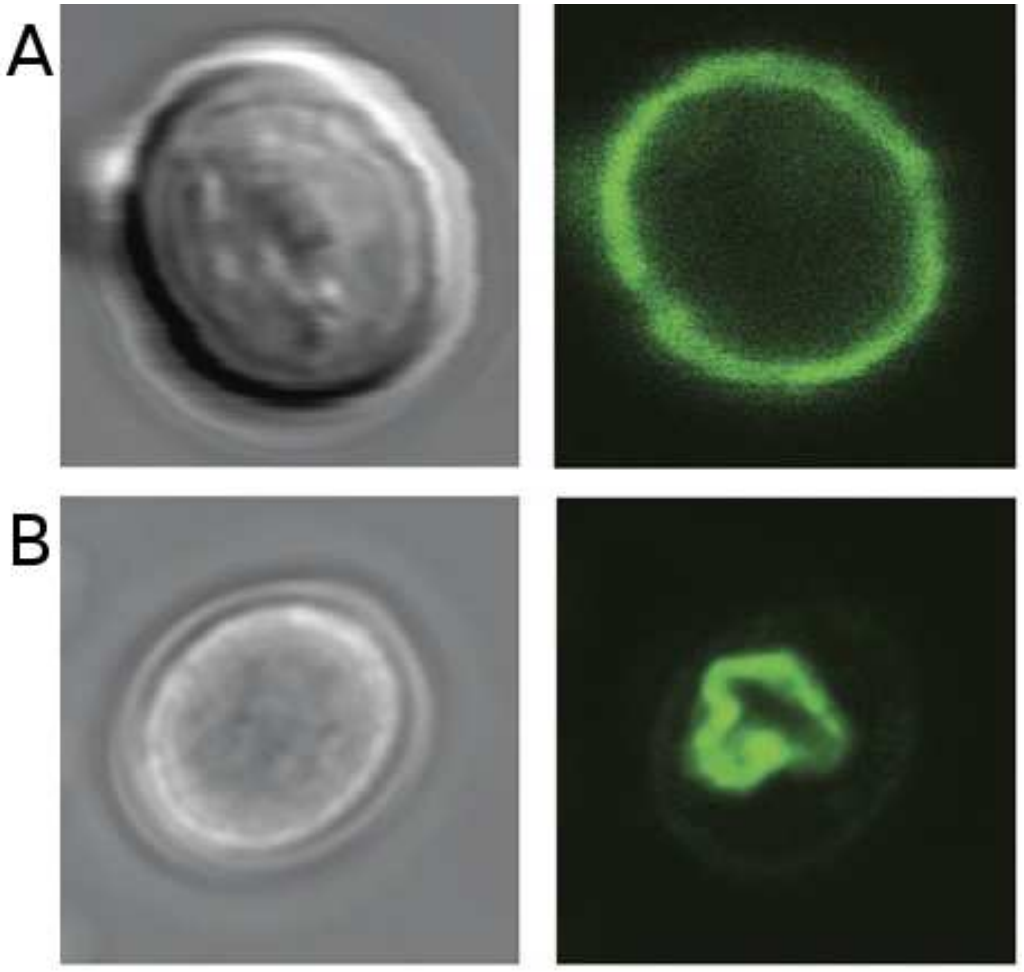


FIGURE 1.9 – Localisations sub-cellulaires de protéines de *S. cerevisiae* marquées par la GFP (d'après [156]) :(A) protéine de la membrane plasmique Ctr1p ; (B) protéine de la membrane golgienne Hut1p. Images obtenues par microscopie confocale.

1.3.2 Interactome

L'ensemble des interactions moléculaires au sein d'une cellule est appelé interactome. De ces interactions résulte l'ensemble des réactions formant le métabolisme d'une cellule. Les protéines sont capables d'influer sur l'ADN en tant que facteurs de transcription ou en modifiant la structure de la chromatine. Les protéines sont capables aussi d'interagir entre elles, de façon permanente ou transitoire. Leurs fonctions peuvent aussi être dépendantes d'interactions avec différents métabolites cellulaires, que ces métabolites participent directement à la fonction de la protéine ou interviennent comme substrat d'une réaction enzymatique. Les ARN interviennent aussi dans la régulation de la physiologie cellulaire en perturbant la traduction d'autres ARN. Les micro ARN (ARNmi) par exemple, sont de petits ARN simple-brin qui agissent comme répresseurs post-transcriptionnels en s'appariant à des ARN messagers. Les ARNmi ont été montrés comme étant impliqués dans un grand nombre de fonctions physiologiques essentielles telles la croissance cellulaire [157], l'apoptose [158], le métabolisme [159], etc.
Pour comprendre en détail tous les mécanismes moléculaires qui régissent la physiologie de la cellule il faudrait reconstruire tous les réseaux d'interaction possibles entre tous les composants cellulaires. Actuellement, les techniques d'investigation moléculaires ne nous permettent d'obtenir des données à grande échelle que pour les interactions entre protéines et entre protéines et ADN. Cependant, ces deux niveaux d'interaction suffisent à la construction des modèles cellulaires actuels dont la granularité reste au niveau des protéines.

1.3.2.1 Interactions protéines/ADN

Les données d'interactions entre protéines et ADN et en particulier entre facteurs de transcription et ADN définissent le réseau de régulation de gènes de la cellule encore appelé régulome. L'identification de la topologie et des interconnections au sein de ce réseau représente un des buts ultimes pour dans la compréhension des mécanismes de réponses cellulaires aux stimuli de l'environnement et des processus biologiques comme le développement et le cycle cellulaire.

La principale technique permettant de produire des données d'interaction entre FTs et ADN est la méthode dite *ChIP-on-chip* pour *Chromatin ImmunoPrecipitation on Chip* (voir figure 1.10) qui permet l'étude des protéines interagissant avec l'ADN. La technique de *ChIP on chip* combine la technique d'immunoprécipitation de la chromatine et les puces à ADN [160]. Elle permet, pour un facteur de transcription particulier, d'identifier l'ensemble des promoteurs où celui-ci se fixe dans les conditions testées. Le facteur de transcription étudié est covalement lié à l'ADN *in vivo*. Après purification, ses fragments d'ADN cibles sont marqués et hybridés sur des puces à ADN. Cette approche permet très rapidement d'obtenir un ensemble de régions promotrices partageant les mêmes sites de fixation pour un facteur de transcription. L'originalité et l'intérêt de cette méthode viennent du fait que l'extraction d'ADN se fait in vivo, ce qui permet d'avoir une idée plus réaliste des processus à l'œuvre dans les cellules lors de l'initiation de la transcription. L'utilisation de données ChIP peut être associée à celle de données transcriptomiques, ainsi on pourra vérifier pour chaque interaction FT/gène si la fixation du FT a vraiment un

effet sur l'expression du gène. Workman *et coll.* [161] utilisent ce type de stratégie pour analyser la réponse aux dommages à l'ADN et reconstruisent un réseau de régulation impliqué dans cette réponse.

On pourra avoir accès à des données d'interactions protéines/ADN de différents organismes dans les bases *BIND* (*Biomolecular Interaction Network Database*), *Transfac* (`http://www.biobase-international.com/`), *RegulonDB* (`http://regulondb.ccg.unam.mx/`) et *JASPAR* (`http://jaspar.genereg.net/`).

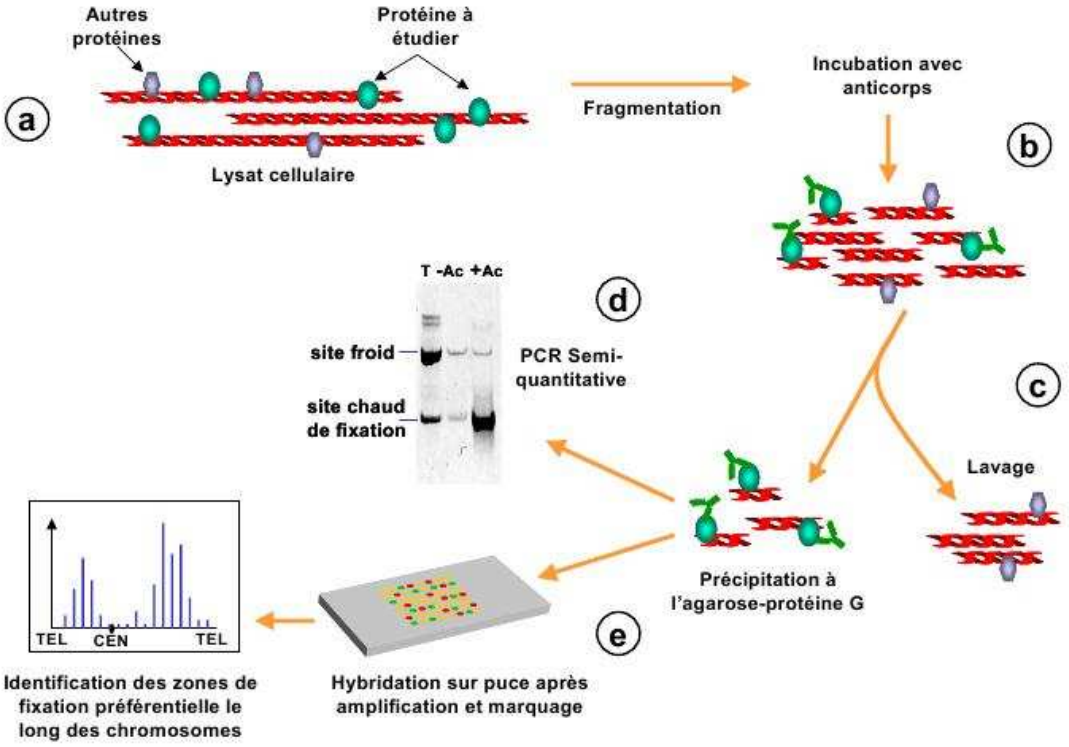


FIGURE 1.10 – Principe de la technologie des *ChIP-on-Chip* (d'après [95]). (a) Fixation covalente des protéines à l'ADN génomique par un traitement au formaldéhyde (*cross-linking*). (b) Fragmentation de l'ADN par des ultra-sons et incubation de l'extrait avec un anticorps spécifique de la protéine d'intérêt. (c) Purification et isolation de la protéine d'intérêt avec le fragment d'ADN associé par immunoprécipitation. (d) *PCR* semi-quantitative pour révéler les fragments d'ADN enrichis, correspondant potentiellement à des sites d'interaction. (e) Hybridation des fragments amplifiés et marqués sur une puce pour cartographier les sites de fixation sur les chromosomes et calculer la fréquence de fixation de la protéine sur le site correspondant.

1.3.2.2 Interactions protéines-protéines

De nombreuses protéines ne deviennent fonctionnelles qu'au sein de complexes homo- ou hétéromériques. L'activité des protéines peut être aussi dépendante de modifications post-traductionnelles nécéssitant une interaction physique transitoire entre deux protéines. L'identification de toutes les interactions protéines-protéines permet de reconstruire un

réseau cellulaire à la base de nombreux processus physiologiques comme les cascades de signalisation ou la constitution de complexes enzymatiques [162].

Récemment, plusieurs techniques à "haut débit" ont permis d'obtenir des cartes d'interactions protéines-protéines à l'échelle du génome. Parmi ces techniques on trouve la méthode par double-hybride, la purification par co-affinité couplée à la spectroscopie de masse et des outils de fouille de données [163] et d'apprentissage qui permettent de prédire des interactions protéines-protéines [164] en se basant sur les caractéristiques physico-chimiques communes de partenaires protéiques connus et sur tout autre type de relation connue entre deux protéines [165, 166].

De grands réseaux d'interactions protéines-protéines ont pu être extraits pour les bactéries *E. coli* [167]piloriextitH. pylori [168], pour *P. falciparum* [169], *S. cerevisiae* [170], *D. melanogaster* [171], *C. elegans* [172] et l'Homme [173]. D'autres études ne se sont pas "contentées" de produire une carte des interactions mais ont appliqué aussi des outils d'apprentissage ou de classification pour mieux comprendre la structuration globale de ces interactions [174, 126, 127]. Ils ont pu mettre en évidence la très forte connectivité et la structure modulaire du réseau d'interactions protéines-protéines chez la levure *S. cerevisiae*. Les modules fortement connectés ont pu aussi être associés de façon significative à des processus biologiques particuliers. Les données d'interaction protéines-protéines représentent également une source d'information complémentaire des réseaux de régulation de gènes en biologie des systèmes comme dans le travail d'Ideker et coll. [175] où ils ont réussi à intégrer 3 niveaux d'observation du comportement cellulaire (transcriptome, protéome et interactions physiques) afin d'extraire des régulations directes et indirectes au niveau transcriptionnel et afin d'identifier de potentielles régulations post-traductionnelles.

Enfin, l'identification des interactions protéines-protéines dans des organismes où le taux d'annotation des gènes reste encore faible, comme chez l'Homme, peut confirmer l'existence de protéines qui n'étaient que prédites, et attribuer un rôle potentiel aux protéines de fonction inconnue selon le principe de "coupable par association".

On peut accéder à différents ensembles de données d'interactions protéiques grâce à des banques spécialisées comme *BIND* (*Biomolecular Interaction Network Database*, `http://bind.ca`), *DIP* (*Database of Interacting Proteins*, `http://dip.doe-mbi.ucla.edu/`) et IntAct (`http://www.ebi.ac.uk/intact/site/index.jsf`).

1.3.3 Fonctions et états fonctionnels

1.3.3.1 Ontologies fonctionnelles

La biologie est un domaine qui manque encore de formalisme strict. Malgré les efforts du consortium HGNC (*HUGO Gene Nomenclature Committee*) pour standardiser la nomenclature des gènes, des améliorations étaient encore nécessaires pour définir les fonctions des gènes et de leurs produits [176]. Ceci a incité la communauté scientifique à développer des ontologies pour annoter les gènes et leurs produits. D'après le *Robert*, dictionnaire de la langue française,l'ontologie est la "partie de la métaphysique qui s'applique

à l'être en tant qu'être, indépendamment des ses déterminations particulières" [177]. De façon moins absconse, en informatique, une ontologie est l'ensemble structuré des termes et concepts fondant le sens d'un champ d'informations. L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que les relations entre ces concepts (définition *Wikipédia*).

Le projet Gene Ontology [176] (`http://www.geneontology.org/`) est un effort commun d'élaboration d'un langage dynamique et contrôlé pour l'annotation des gènes et de leurs produits. Les annotations GO constituent une banque de connaissances dans un vocabulaire commun à tous les domaines du vivant. Un gène est ainsi décrit selon trois ontologies différentes : processus biologique, composant cellulaire et fonction moléculaire. GO est actuellement très utilisé pour l'annotation fonctionnelle car de nombreux organismes ont leurs gènes classés par cette ontologie. Des annotations systématiques sont établies par les consortiums maintenant les séquences de différents organismes modèles et par la banque *GOA* de l'*European Bioinformatics Institute*. Les principaux organismes modèles et les consortiums associés :

- *Arabidopsis thaliana* (`http://www.arabidopsis.org/`)
- *Caenorhabditis elegans* (`http://www.wormbase.org/`)
- *Danio rerio* (`http://zfin.org/`)
- *Dictyostelium discoideum* (`http://dictybase.org/`)
- *Drosophila melanogaster* (`http://flybase.bio.indiana.edu/`)
- *Escherichia coli* (`http://www.ecolicommunity.org/`)
- *Gallus gallus* (`http://www.agbase.msstate.edu/`)
- *Homo sapiens* (`http://www.ebi.ac.uk/GOA/human_release.html`)
- *Mus musculus* (`http://www.informatics.jax.org/`)
- *Rattus norvegicus* (`http://rgd.mcw.edu/`)
- *Saccharomyces cerevisiae* (`http://www.yeastgenome.org/`)
- *Schizosaccharomyces pombe* (`http://www.genedb.org/genedb/pombe/`)

Chaque terme de GO correspond à un identifiant alphanumérique unique, un nom, des synonymes et une définition. Un terme n'est classé que dans une des trois ontologies de GO et chaque ontologie est structurée selon un graphe acyclique dirigé.
De nombreux outils ont été développés pour interroger GO comme *AmiGO* pour l'annotation et *EASE* pour l'identification des annotations statistiquement sur-représentées dans un groupe de séquences. Pour plus d'informations sur les outils utilisant GO, on peut se reporter au site du consortium GO, section *GO tools* (`www.geneontology.org`).

Pour leur part, les auteurs de la *Kyoto Encyclopedia of Genes and Genomes* (*KEGG*) [178] ont pour ambition d'accéder à un niveau supérieur d'information non plus en annotant la fonction isolée des gènes et de leurs produits, mais en les situant dans des voies métaboliques. Avec une architecture fondée sur des relations binaires, *KEGG* conçoit les voies métaboliques comme des réseaux de molécules en action et le génome comme un réseau unidimensionnel de gènes. Son but est alors de reconstruire ces réseaux métaboliques et de replacer chaque séquence biologique dans la ou les voies métaboliques dans lesquelles elle intervient. En 2005, *KEGG* gèrait les enzymes et composés impliqués dans 17682 voies métaboliques de diverses espèces et proposait également 255 voies métabo-

liques de référence construites à partir des éléments conservés entre les espèces. En tirant parti des relations phylogéniques, lorsqu'un élément d'une voie est inconnu dans une espèce, la comparaison avec le modèle de référence peut permettre d'émettre des hypothèses sur l'identité de cet élément. L'interrogation de *KEGG* à partir d'une liste de gènes permet de vérifier à quelles voies métaboliques ils appartiennent et de définir des "unités fonctionnelles" lorsque leurs protéines sont proches dans une voie métabolique.

Les annotations fonctionnelles sont souvent utilisées en biologie des systèmes pour soit valider la cohérence de classes de gènes construites à partir d'autres données systémiques (transcriptome, réseaux d'interactions, etc..) soit pour participer directement à la définition de modules fonctionnels (voir section 2.4.2.3).

1.3.3.2 Etats fonctionnels ou *phénome*

Les données du *phénome* font référence à la production à "haut débit" d'informations sur la croissance ou la viabilité de cellules en réponse à des perturbations génétiques et/ou environnementales. Après les puces à ADN et les puces à protéines, les puces à cellules ont réellement permis de passer du gène à l'échelle du système en permettant l'analyse en parallèle du phénotype de centaines de cellules sauvages ou mutées dans différentes conditions expérimentales [179, 180]. De la même façon les outils de *chemogenomics* permettent de tester en pharmacologie les effets phénotypiques de banques entières de molécules [181]. Et, dernièrement, l'utilisation d'ARN interférants a permis d'étudier, à l'échelle du génome, l'effet de l'inactivation transitoire de gènes cibles [44].
L'intégration des données du phénome à l'analyse de réseaux de régulation et d'interactions prédits à partir des autres types de données systémiques permet de valider les liens prédits et de vérifier l'importance de chaque élément du réseau pour le fonctionnement global du système [175].
A l'ère de la biologie des systèmes et de l'automatisation des techniques expérimentales,

les informations biologiques qui étaient auparavant produites à l'échelle de la variable unique (un gène, une protéine, une cellule) ont tendance à toutes devenir des données en "ome" (voir `http://omics.org/` pour une description de différents 'omes'). Nous avons vu au cours de ce chapitre les principales sources de données systémiques mais d'autres types de données [182] commencent à apparaître avec un nouveau lexique de termes en 'ome'. Ces différentes données 'omiques' permettent de définir de nouveaux domaines de connaissance :

- le *physiome*, regroupant toutes les informations pouvant aider à modéliser le fonctionnement d'un tissu ou d'un organe (voir le projet *Saphir*, soutenu par l'ANR, qui vise à développer un environnement pour une collection de modèles dédiés au transport des fluides : cœur, rein, poumon, muscles, hormones `http://physiome.ibisc.fr/saphir/`).
- l'*orféome*, projet de génomique fonctionnelle à haut débit visant à cloner systématiquement les *ORFs* (*Open Reading Frame* ou phase ouverte de lecture) d'un organisme pour mener des études fonctionnelles (voir le projet *ATOME* pour *A. thaliana*, `http://www.versailles.inra.fr/urgv/atome/`).

- le *foldome*, qui correspond à l'identification de toutes les formes de repliements de molécules biologiques comme les protéines.
- le *traductome*, qui décrit la population de protéines exprimée dans un organisme à un moment donné. A la différence du protéome, le traductome, comme le transcriptome, est dynamique.
- le *secretome*, sous-ensemble du protéome défini par les protéines qui seront exportées hors de la cellule.
- le *fonctiome*, à rapprocher des annotations fonctionnelles associées à chaque génome, décrit toutes les fonctions associées aux protéines codées par un génome.
- le "*unknowme*", terme un peu provocateur décrivant la partie du génome pour laquelle aucune annotation fonctionnelle n'a encore été associée.

On pourra ainsi créer de nouveaux 'omes' à mesure que l'on produira des données pouvant définir de nouveaux sous-domaines de connaissance liés à la cellule ou au vivant (*ribonome*, *mitochondriome*, *immunome*, etc.).

Dans une approche systémique, il faudrait idéalement obtenir des informations précises et quantitatives pour toutes les interactions importantes au sein d'un système. Actuellement, les méthodes de génération de données ne permettent pas d'obtenir ce niveau de granularité (impossibilités techniques ou coût en temps et en argent trop important). Une alternative possible consiste à inférer ou déduire des informations manquantes à partir des seules observations possibles ou disponibles. Dans l'exemple de la modélisation d'une dynamique cellulaire, on a généralement facilement accès à la dynamique d'expression des gènes (dynamique du transcriptome) mais on observe plus difficilement les relations de régulation entre gènes. La reconstruction de ce réseau de régulation génique est une étape déterminante pour le modèle final. On va alors essayer d'inférer ou de déduire ce réseau de régulation entre gènes en exploitant les données du transcriptome. Plus généralement, l'inférence de réseaux de régulation ou d'interaction constitue actuellement une des étapes les plus importantes en biologie des systèmes et a suscité de très nombreux développements mathématiques, statistiques, algorithmiques et surtout méthodologiques.

Chapitre 2

Inférence et modélisation de réseaux biologiques

2.1 Importance des données dynamiques

L'inférence de réseaux de régulation transcriptionnels se base essentiellement sur l'exploitation de données d'expression de gènes obtenues à l'aide de puces à ADN. Un grand bond méthodologique a été réalisé grâce à la génération de données cinétiques plutôt que de simples points de mesure indépendants. D'un point de vue biologique, il s'agit de capturer des phénomènes dynamiques, que ce soit une réponse cellulaire à des changements environnementaux ou bien l'analyse de phénomènes physiologiques comme le cycle cellulaire ou un processus de différentiation. Dans ces différents cas, un ensemble de cinétiques d'expression reflétera mieux l'hétérogénéité des réponses cellulaires et les mécanismes de régulation associés. D'un point de vue méthodologique, c'est dans la dynamique d'expression des gènes que l'on va rechercher des relations de causalité entre gènes ou entre gènes et conditions environnementales.

2.1.1 Importance de la stratégie de mesure du signal

La qualité des données d'expression influencera directement la qualité d'un réseau inféré. Cela ne concerne pas seulement les problèmes techniques inhérents à la technologie des puces à ADN et au traitement du signal mais concerne aussi les choix stratégiques et méthodologiques effectués lors de la conception d'un protocole expérimental. Lorsque l'on veut capturer la dynamique d'un phénomène physiologique, se pose la question de la fenêtre d'observation. Durant quel laps de temps doit on maintenir l'observation pour capturer toute la dynamique du phénomène ? Il ne suffit pas d'observer une modulation d'expression mais il faut aussi déterminer quand elle est apparue, si elle n'est que transitoire et à quel moment elle disparaîtra. Une cinétique d'expression correspond à un ensemble de mesures discrètes, le choix de l'intervalle entre ces mesures est critique, surtout en ce qui concerne des phénomènes cycliques tel le cycle cellulaire ou le rythme circadien. Un intervalle trop grand ou des temps de mesure mal placés risqueraient de faire passer un phénomène cyclique pour un phénomène transitoire ou bien ne rendraient compte d'aucune modulation d'expression.

Les méthodes d'analyse doivent aussi prendre en compte le fait que l'intervalle de temps entre deux mesures n'est pas forcément constant [183, 184]. Les mesures d'expression sont réalisées sur des populations de cellules, celles-ci peuvent être hétérogènes, à des stades de différenciation différents ou dans des phases différentes du cycle cellulaire (population asynchrone). Le signal mesuré correspond à une moyenne des signaux cellulaires individuels. L'analyse d'une réponse à un signal ou l'analyse d'un phénomène cyclique au sein d'une population asynchrone risque de correspondre pour un même gène à une moyenne de modulations non synchronisées où tous les signaux "s'écrasent" mutuellement. Pour certaines analyses, les cellules sont d'abord artificiellement synchronisées avant d'effectuer les mesures d'expression [185], pour d'autres, le stimulus étudié provoque "naturellement" une synchronisation des cellules et permet d'observer une réponse homogène pour chaque gène [60, 59].

2.1.2 Dynamiques cellulaires non transcriptionnelles

Si la grande majorité des études se base sur la dynamique du transcriptome, d'autres types de dynamiques cellulaires existent. Les modifications post-traductionnelles des protéines correspondent à un niveau de régulation majeur qui conditionne l'activité de nombreuses protéines et voies de signalisation. De récentes études basées sur l'analyse globale des sites de phosphorylation ont permis d'identifier et d'enrichir les composants du phosphoprotéome [186, 187, 188, 189]. La dynamique de chaque type de phosphorylation peut aujourd'hui être suivie et peut être prise en compte pour identifier les différents processus et voies cellulaires qui sont régulées par phosphorylation et ainsi enrichir un réseau de régulation plus global.

Un autre type de dynamique concerne cette fois les mouvements des protéines au sein d'une cellule. Il est clairement établis que l'activité de nombreuses protéines, notamment celle des protéines régulatrices, est conditionnée par leur localisation au sein de la cellule [190]. La régulation de certains évènements, comme la réponse rapide à un stress, est régulée par les déplacement de facteurs de transcription entre cytoplasme, génome mitochondrial, noyau et nucléole [191, 192, 193, 194]. Ce type de régulation est très rapide car il ne nécessite pas la synthèse de nouveaux facteurs de transcription, il utilise un "réservoir" de régulateurs déjà présents dans la cellule et qui sont rendus opérationnels par leur changement de localisation. L'absence de données cinétiques reflétant ces dynamiques spatiales rend l'intégration de ce type de régulation difficile.
Certaines stratégies expérimentales comme le marquage d'une protéine par la *GFP* (*Green Fluorescent Protein*) permettent de suivre les variations de l'intensité de fluorescence moyenne entre compartiments cellulaires [195, 196, 197, 198, 199, 153](voir [200] pour une présentation de différentes techniques d'imagerie cellulaire pouvant produire des données quantitatives). Même si ce type de technique reste difficile à mettre en œuvre à grande échelle, des réseaux représentant les dynamiques de localisation des protéines ont étés construit à l'aide de modèles basés sur des équations différentielles (voir [201, 202] et [203] pour une revue de ces modèles).

2.2 Inférence de réseaux : objectifs et formalismes

L'étude systémique d'une organisation biologique comme la cellule passe généralement par une première étape de modélisation. Cette modélisation consiste à abstraire le réseau réel sous une forme graphique ou mathématique facile à représenter et à manipuler. La façon la plus simple de représenter un système constitué d'éléments qui interagissent entre eux est la représentation en graphe, où des nœuds correspondent aux éléments du réseau et des arcs représentent les relations entre ces éléments. Un graphe est défini par un couple (V, E), avec V représentant l'ensemble des nœuds et E l'ensemble des arêtes. Un arc reliant deux nœuds peut être orienté, représenté par une flèche par exemple. Ce cas de figure peut illustrer une relation entre une source et une cible et est bien adapté pour représenter des transformations chimiques au sein de réseaux métaboliques ou des régulations au sein de réseaux de régulation de gènes. Les arcs orientés peuvent être associés à un signe (positif pour une activation ou négatif pour une inhibition) ou peuvent être pondérés par des valeurs quantifiant une probabilité ou une vitesse de réaction par exemple. Les graphes non orientés servent plutôt à représenter des interactions mutuelles comme les réseaux d'interactions protéiques, ou alors peuvent temporairement représenter des relations pour lesquelles sources et cibles n'ont pu être encore différenciées.

Différentes opérations peuvent être appliquées aux graphes afin d'en extraire des informations caractérisant un réseau biologique. La recherche de chemins entre deux composants du graphe permet d'identifier des régulations manquantes ou des redondances. Un cycle dans un graphe sera interprété comme une boucle de rétroaction, caractéristique de phénomènes biologiques comme l'homéostasie. On peut aussi rechercher l'existence de modules fonctionnels dans un graphe par des stratégies de coupes minimales d'arcs ou de recherche de cliques. L'analyse de la connectivité globale d'un graphe permet d'obtenir une mesure de la complexité d'un réseau. Et, enfin la comparaison de réseaux de régulation d'organismes différents permet d'analyser l'évolution et la conservation de voies de régulation [204].

Dans la suite de ce chapitre nous nous pencherons essentiellement sur l'inférence de réseaux de régulation à partir de données d'expression. D'une part, il s'agit du niveau de granularité le plus fin ou plutôt le plus en amont au niveau de la régulation des processus physiologiques cellulaires. D'autre part, il s'agit de la source de données systémique la plus abondante et correspondant à la plus grande diversité de conditions expérimentales (voir `http://ihome.cuhk.edu.hk/~b400559/arraysoft_public.html` pour une liste exhaustive de bases de données d'expression de gènes). Parmi tous les types de réseaux biologiques, les réseaux de régulation transcriptionnels sont ceux qui ont posé les problématiques méthodologiques les plus intéressantes et qui ont suscité de nombreux développements algorithmiques et mathématiques différents.

Nous pouvons classer les méthodes d'inférence de réseaux de régulation génétiques en deux grandes familles, différentes mais complémentaires. La première famille s'intéresse à la modélisation fine, statique ou dynamique, de relations de régulation au sein de petits réseaux de gènes. Ces méthodes postulent généralement une hypothèse globale sur le

réseau à inférer que ce soit au niveau de sa structure et/ou de sa dynamique et essayent d'estimer les paramètres associés au modèle. Ces méthodes s'appliquent généralement à de petits réseaux (<100 gènes) dont les nœuds sont bien identifiés. La deuxième famille de méthodes, beaucoup plus hétérogène que la première, correspond à des approches exploratoires, qui ne s'appuient pas sur un modèle global du réseau. On parle généralement d'extraction de réseaux, car ces méthodes vont plutôt chercher à identifier les nœuds et les relations de régulation les plus importants du réseau. On parlera plus souvent de méthodologies que de méthodes car ces approches peuvent mettre en œuvre des stratégies complexes exploitant diverses sources d'informations biologiques à partir de combinaisons d'outils informatiques et mathématiques différents. Dans les deux sections suivantes je présente différents exemples parmi les plus représentatifs des ces deux familles de méthodes d'inférence de réseaux de régulation de gènes. Je ne chercherai pas à être exhaustif au cours de cette présentation mais j'essayerai plutôt d'illustrer les différentes approches possibles avec leurs intérêts et leurs limites.

2.3 Méthodes d'inférence de paramètres ou de structures de modèles

Nous nous plaçons dans le cas où un nombre réduit de variables (gènes) a déjà été identifié soit par le biologiste soit par les méthodes décrites dans la section suivante (voir section 2.4, page 59). Les méthodes présentées ici vont chercher à définir ou à compléter un réseau de régulation impliqué dans le contrôle des variables déjà identifiées en exploitant des données du transcriptome. Étant donné une classe de modèles de réseaux de régulation, les méthodes d'apprentissage automatique vont permettre de définir une ou plusieurs solutions candidates (graphe d'interaction et paramètres du modèle) que le biologiste pourra ensuite tester en générant d'autres expériences pour vérifier la justesse d'une partie ou de l'ensemble du graphe inféré. Lorsque l'algorithme fournit le paramétrage complet d'un modèle, il devient possible d'utiliser ce modèle en simulation et donc à nouveau, on peut comparer données simulées et données expérimentales de test. Dans le cas des réseaux bayésiens dynamiques ou statiques, la plus grande difficulté au niveau de l'apprentissage résidera dans la recherche de la structure du réseau, ce problème étant reconnu comme *NP*-complet. L'inférence de réseaux, repose soit sur l'utilisation de modèles idéalisés comme les réseaux booléens soit sur des modèles continus, nécessaires pour relier le comportement dynamique de différents éléments à une topologie de réseau particulière.

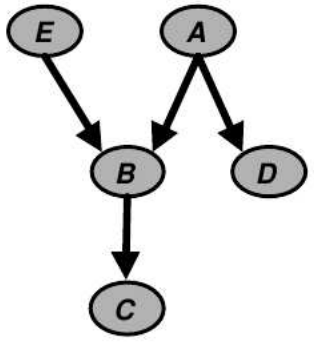
2.3.1 Modèles statiques : l'exemple des réseaux bayésiens

L'intérêt particulier des réseaux bayésiens est de tenir compte simultanément de connaissances a priori d'experts (dans le graphe) et de l'expérience contenue dans les données. De plus, leurs bases solides en statistiques leur permettent de bien appréhender les aspects stochastiques et bruités des mesures d'expression de gènes [205]. Un réseau bayésien correspond à la représentation d'une distribution de probabilité jointe sur un ensemble de variables aléatoires. Un réseau bayésien est constitué de deux composants (voir figure 2.1) :

- un graphe acyclique dont les sommets correspondent à des variables aléatoires et des arêtes indiquant les relations de dépendance conditionnelle,
- une famille de distributions conditionnelles pour chaque variable, étant donnés les parents de cette variable.

Ensemble, ces deux composants déterminent une unique distribution jointe. Lorsque l'on applique les réseaux bayésiens aux systèmes de régulation géniques, les sommets sont identifiés aux gènes et à leurs niveaux d'expression, les arêtes représentent les interactions entre gènes et les distributions conditionnelles décrivent ces interactions. Pour un ensemble de données d'expression de gènes, l'apprentissage d'un réseau bayésien consiste à inférer différents réseaux qui correspondent au mieux à l'ensemble des données analysées. Il a été montré par Ott *et coll.* [206] que ce problème, la recherche du réseau optimal, était *NP*-difficile. On doit alors faire face au dilemme suivant, soit restreindre les réseaux à un faible nombre de gènes soit inférer des réseaux sous-optimaux par des méthodes de recherche basées sur des heuristiques (voir Friedman *et coll.* [207]).

La principale limite des modèles de réseaux bayésiens réside dans l'impossibilité pour ces modèles de construire des réseaux cycliques alors que ce type de de régulations peut exister dans les réseaux de régulation réels. Cette limitation peut cependant être levée dans le cadre de l'inférence de réseaux de régulation de gènes à partir du formalisme des réseaux bayésiens dynamiques (voir section 2.3.2.2).



$P(E), P(A), P(B \mid E, A), P(C \mid B), P(D \mid A).$

$P(A, B, C, D, E) = P(A)\,P(B \mid A, E)\,P(C \mid B)\,P(D \mid A)\,P(E)$

$I(A; E), I(B; D \mid A, E), I(C; A, D, E \mid B), I(D; B, C, E \mid A), I(E; A, D).$

FIGURE 2.1 – Exemple d'un réseau bayésien simple constitué d'un graphe, de probabilités conditionnelles, de la probabilité de distribution jointe et des indépendances conditionnelles (selon Friedman *et coll.*, 2000 [207])

2.3.2 Modèles dynamiques

2.3.2.1 Modèles booléens

L'activité d'un gène peut être discrétisée sous une forme booléenne pour exprimer une forme active (1) ou inactive (0). L'activité d'un gène peut aussi être exprimée en fonction de l'état d'activation des autres gènes par l'intermédiaire de fonctions booléennes (voir figure 2.2). C'est en 1969 que Kauffman a démontré l'intérêt de ce formalisme pour la modélisation de réseaux de régulation [208], formalisme repris depuis dans de nombreux

travaux (voir [209] pour revue et [210, 211]).
Le formalisme des réseaux booléens a été un des premiers à être utilisé par les méthodes d'induction de modèles. Ici, le problème de l'identification d'un réseau de régulation génique à partir de données d'expression consiste à identifier le réseau booléen sous-jacent par interpolation. Un des premiers algorithmes d'identification à utiliser le formalisme booléen est l'algorithme REVEAL (REverse Engineering ALgorithm)[212]. Cet algorithme prend en entrée un ensemble de séries temporelles, recherche la connectivité entre les différents éléments du réseau, puis les fonctions booléennes qui déterminent l'évolution du réseau. Les réseaux booléens ont été appliqués à des modèles simples de réseaux de régulation [213] mais aussi à de larges réseaux par l'analyse de leur propriétés générales [214, 209, 210, 215].

Soit le vecteur x de taille n qui représente l'état du système dans un réseau de régulation booléen de n éléments. Chaque x_i peut prendre la valeur 1 ou 0, l'espace d'état du système correspond alors à 2^n états. L'état x_i d'un élément au temps $t+1$ est calculé grâce à une fonction booléenne ou règle b_i à partir de l'état de k éléments parmi n au temps t. La variable x_i est aussi appelée *sortie* de l'élément et les variables k sont appelées *entrées*. Pour k *entrées*, le nombre total de fonctions booléennes b_i reliant les *entrées* aux *sorties* est de 2^{2^k}. Pour $k = 2$ il y aura donc 16 fonctions possibles incluant le *nand*, *or* et *nor* (voir figure 2.2). En résumé, la dynamique d'un réseau booléen décrivant un système de régulation peut être écrite sous la forme

$$x_i(t+1) = b_i(x(t)), 1 \leq i \leq n, \tag{2.1}$$

où b_i relie k *entrées* à une valeur en *sortie*. La représentation du graphe des interactions

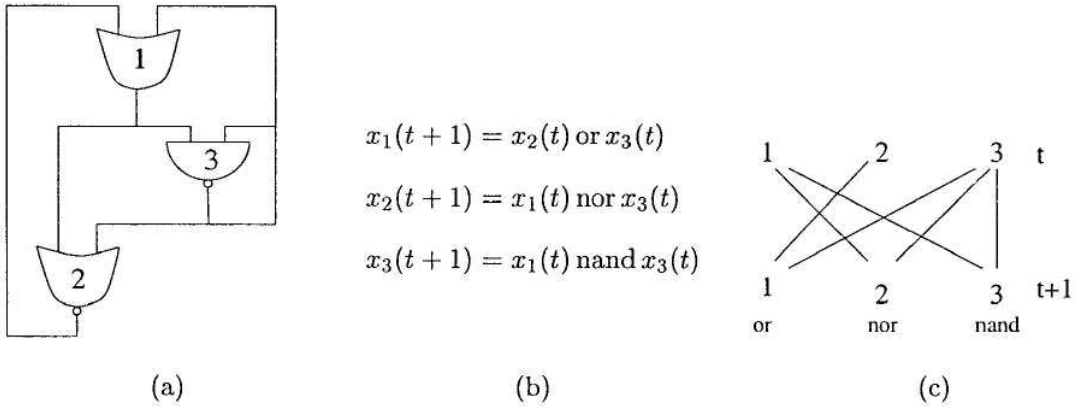


FIGURE 2.2 – (a) Exemple de réseau booléen avec (b) les équations correspondantes. Dans ce cas, on a $n = 3$ éléments dans le réseau et $k = 2$ entrées. (c) diagramme en circuit du réseau booléen. (Selon de Jong, 2002 [4].)

sous-jacent au réseau booléen (voir figure 2.2(c)) permet de prendre en compte les transitions d'état d'un temps t (première ligne) à un temps $t + 1$ (deuxième ligne) avec les fonctions booléenne associées (voir figure 2.2(b)). Les transitions d'état au sein d'un réseau booléen sont *déterministes* et *synchronisées*, ainsi, le calcul des transitions d'états de t à $t + 1$ des différentes entrées se fait de façon parallèle. Par exemple, un vecteur d'état 000 à t aura pour valeurs 011 à $t + 1$, dans l'exemple de la figure 2.2 cela correspond à

l'activation des gènes 2 et 3. On appelle *trajectoire* du système la séquence des différents états reliés par leurs fonctions de transition. Le nombre d'états possibles au sein d'une *trajectoire* est fini et tous les états initiaux peuvent potentiellement converger vers un état stable ou cyclique qui définissent à eux deux un *bassin d'attraction*.

Les réseaux booléens permettent une analyse efficace de réseaux de régulation de grande taille par la simplification de leurs structures et de leurs dynamiques. Cependant, la synchronisation des transitions d'état et la discrétisation des valeurs d'expression des gènes ne permet pas de refléter correctement certains types de comportements biologiques nécessitant par exemple de modéliser des chaînes d'activations ou d'inhibitions successives. Dans ces cas, le formalisme booléen ne sera pas approprié et des méthodes plus générales seront nécessaires.
Thomas *et coll.* [216, 217] ont proposé une généralisation des méthodes booléennes en donnant aux variables la possibilité de prendre des valeurs intermédiaires et aux transitions d'états de se dérouler de façon asynchrone. Ces modèles permettent de prendre en compte certaines propriétés dynamiques des réseaux et ont montré leur efficacité avec des petits réseaux de régulation comme le modèle d'infection d'*E. coli* par le phage λ [218, 219, 220], certaines phases du développement chez la drosophile [221, 222] et la morphogénèse chez *A. thaliana* [223]. Cependant, ces modèles ne sont pas applicables à la modélisation quantitative de réseaux de régulation à partir de grands volumes de données d'expression.

2.3.2.2 Réseaux bayésiens dynamiques

L'utilisation de réseaux bayésiens dynamiques a été proposée dans le cadre de l'extraction de réseaux de régulation géniques [224] pour permettre au formalisme des réseaux bayésiens statiques de prendre en compte la dimension temporelle des phénomènes transitoires comme les régulations transcriptionnelles. Cette approche dynamique a été mise en œuvre dans [225] et en particulier dans [226] où a été exploitée la possibilité d'introduire des connaissances *a priori* dans le modèle.

Gharamani [227] appelle *réseau bayésien dynamique* tout modèle graphique reflétant une évolution temporelle. Le plus simple des réseaux bayésiens est alors une chaîne de Markov : soit $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T$. Pour une hypothèse markovienne d'ordre 1, le présent ne dépend que du passé proche. Dans le cas continu, le modèle se décrit à l'aide de l'équation suivante :

$$x_{t+1} = f_\theta(x_t) + \epsilon_t \tag{2.2}$$

où f est une fonction linéaire ou non et ϵ_t est la réalisation d'un bruit gaussien. Dans cette famille de modèles la loi jointe peut être factorisée de la manière suivante :

$$P(\mathbf{x}_1, ..., \mathbf{x}_T) = P(\mathbf{x}_1) \prod_{t=1}^{T-1} p(\mathbf{x}_{t+1}|\mathbf{x}_t) \tag{2.3}$$

Le temps indique ici la causalité sur laquelle sera basée l'inférence. Il est aussi possible de déployer dans le temps un réseau bayésien statique, on emploie alors soit le terme de réseau

de croyance bayésien dynamique soit le terme générique de réseau bayésien dynamique. Il n'y a plus de contrainte sur l'acyclicité du graphe qui définit les indépendances conditionnelles à travers le temps. On s'intéresse à l'évolution des variables unidimensionnelles $x_i(t)$:

$$p(x_1(1), ..., x_n(T)) = \prod_{i=1}^{n} p(x_i(1)). \prod_{t=1}^{T-1} \prod_{i=1}^{n} p(x_i(t+1)|Pa_i(t)). \tag{2.4}$$

L'apprentissage d'un tel réseau ressemble à l'apprentissage d'un réseau bayésien statique : il se décompose en apprentissage de la structure et apprentissage des probabilités conditionnelles. Il est aussi possible de complexifier le modèle markovien si on suppose la présence d'un processus caché :

$$x_{t+1} = F_\theta(x_t) + \epsilon_t^h \tag{2.5}$$

$$y_t = H(x_t) + \epsilon_t^o \tag{2.6}$$

où (x_t) est le processus caché et $y(t)$ est le processus observé. On parle alors de modèle à espace d'états.
Dans les réseaux bayésiens dynamiques [228], la structure et les paramètres sont appris, ce qui les rend pertinents pour l'inférence des réseaux de régulation génétiques. Ils constituent une alternative intéressante aux réseaux booléens probabilistes. Les modèles à espace d'états sont très utilisés pour apprendre des réseaux de régulation en l'absence de mesures des concentrations de protéines (les variables cachées). Ils ont donné lieu à des développements au sein du laboratoire IBISC [226, 229, 230].

2.3.2.3 Systèmes d'équations différentielles

Les réseaux booléens sont particulièrement intéressents pour l'extraction des propriétés topologiques de réseaux de régulations mais restent assez rudimentaire en ce qui concerne la dynamique des réseaux. Les équations différentielles permettent une meilleure description de la dynamique des réseaux en modélisant de manière explicite les variations de concentration de molécules en fonction du temps [231, 232, 233, 234]. L'équation différentielle classique s'écrit sous la forme :

$$\frac{dx_i}{dt} = f_i(x), 1 \leq i \leq n, \tag{2.7}$$

où $x = [x_1, \ldots, x_n]' \geq 0$ est le vecteur de concentration de protéines, d'ARNm ou de petites molécules, et $f_i : \mathbb{R}^n \to \mathbb{R}$ est généralement une fonction non linéaire. Le taux de synthèse de i est vu comme dépendant de x, voire de x_i également. Des délais de type temps d'achèvement de la transcription, de la traduction et temps de diffusion vers le lieu d'activité de la protéine peuvent aussi être représentés dans ces équations [235, 236].
Ce type d'équation différentielles a permis par le passé de développer des méthodes mathématiques performantes pour la modélisation de systèmes de réactions biochimiques spécialement dans le contexte des processus métaboliques (voir [93, 94]). A partir de ces méthodes, des modèles cinétiques des processus de régulation génétiques peuvent être construits en spécifiant les fonctions f_i.

Les modèles basés sur des systèmes d'équations différentielles dépendent cependant de paramètres numériques difficiles à mesurer expérimentalement. Et un des problèmes posés concerne leur *stabilité*, c'est à dire si le comportement du système dépend des valeurs exactes des paramètres initiaux où si le comportement reste stable pour certaines plages de variations. Les systèmes instables semblent mieux correspondre aux phénomènes biologiques réels, d'un autre côté, la stabilité du système permet de s'affranchir des valeurs exactes de certains paramètres. Lorsque l'on construit un système d'équations différentielles décrivant le plus fidèlement possible le système à étudier il est nécessaire d'avoir des connaissances préliminaires sur le type et la dynamique du processus biologique que l'on souhaite modéliser ainsi qu'un nombre suffisant de données et des cinétique d'expression suffisamment longues pour inférer les paramètres inconnus du modèle. Un exemple de connaissances préalables concerne la topologie du réseau et les taux de dégradation et de production des protéines associées. Cependant, ces informations sont souvent inconnues et nécessitent l'utilisation de méthodes d'estimation des paramètres du réseau à partir des données d'expression.

2.4 Méthodes d'extraction ou de prédiction d'interactions

Contrairement aux méthodes d'estimation de paramètres ou de structures de modèles, les approches présentées dans cette section ne s'appuient pas *a priori* sur un modèle global du réseaux à inférer. Ces approches ne souffrent pas du problème de dimension des données et peuvent être utilisées de façon exploratoire. Cette très grande famille d'approches est par contre extrêmement hétérogène du point de vue des algorithmes d'apprentissage ou d'extraction de connaissances. On peut cependant discerner deux grandes sous-familles, des approches basées sur des algorithmes d'apprentissage et de prédiction de régulations et, des approches basées sur une réduction de dimension des variables à analyser et sur l'intégration de connaissances additionnelles pour extraire des relations de régulation potentielles.

2.4.1 Méthodes de prédiction d'interactions

Cette sous-famille, assez hétérogène est généralement caractérisée par deux étapes, une étape d'apprentissage et une étape de prédiction. La première étape consiste à apprendre des règles, une fonction ou un modèle qui permettent de prédire s'il existe une relation de régulation entre deux gènes.

Dans une étude récente, Hvidsten *et coll.* [237] utilisent la théorie des ensembles approchés [238] et le raisonnement booléen [239] comme cadre mathématique pour l'induction de règles à partir d'exemples. Ils utilisent une implémentation de ce cadre dans le système Rosetta [240] pour analyser chez la levure les associations entre motifs de fixation de FTs et des mesures d'expression de gènes obtenues à l'aide de puces à ADN. Ils obtiennent des règles de type *IF-THEN* associant des combinaisons de motifs de fixation à des gènes ayant des profils d'expression particuliers et proposent des hypothèses pouvant expliquer

la co-régulation de certains gènes. Par exemple, la règle : ***IF (RAP1 and SWI5 and MCM1) THEN (similar expression in cell cycle, sporulation, diauxic shift, heat and cold shock, and DNA-damaging agents)*** explique que si des gènes possèdent des motifs de fixation aux trois FTs $RAP1$, $SWI5$ et $MCM1$, ces gènes sont potentiellement co-régulés dans les conditions expérimentales impliquées par le *THEN*. Ce même formalisme a aussi été utilisé pour inférer des règles logiques dans le cas de la prédiction de fonctions [241, 242, 243] et dans l'analyse de voies métaboliques [244].
Fröhler et Kramer [245] utilisent le formalisme logique pour représenter différentes informations biologiques comme les niveaux d'expression des gènes, des motifs de séquences régulatrices, des ontologies fonctionnelles et les interactions protéines-protéines. Ils utilisent ensuite le cadre de la Pogrammation Logique Inductive (PLI) [246] et en particulier le système de Tilde [247] pour apprendre des arbres de décision logiques et inférer un modèle de régulation. Le cadre proposé par la PLI a été aussi utilisé pour l'apprentissage d'un réseau de régulation de gènes à partir de cinétiques d'expression dans [248] et d'un réseau d'interactions entre protéines à partir de données hétérogènes [249] (voir revue de Muggleton dans [250]).
Le formalisme logique a le double avantage de permettre l'intégration d'une grande variété de connaissances biologiques et d'être suffisamment intuitif pour être utilisé par des "non initiés". La difficulté principale concerne l'intégration des données quantitatives et dynamiques, discrétisées dans les travaux cités précédemment. Cependant, des voies de recherche sont actuellement explorées [251] pour prendre en compte l'aspect dynamique des séries temporelles en utilisant une notion d'intervalles de temps inspirée des travaux de Allen [252]. Cette discrétisation des cinétiques d'expression serait associée a des méthodes combinant logique du premier ordre et probabilités [253].

Un autre exemple de méthodologie basée sur l'apprentissage est illustré par l'approche proposée par Middendorf *et coll.* [254]. Dans ce travail, le but est d'apprendre une fonction permettant de prédire un réseau de régulation de gènes impliqué dans la réponse à différentes conditions expérimentales. Leur méthodologie prend en entrée des motifs de fixation à des régulateurs et les données d'expression d'un ensemble de régulateurs dans différentes conditions expérimentales. En sortie, on obtient une prédiction du niveau d'expression des gènes régulés (discrétisé en +1, 0 et -1). L'apprentissage se fait grâce à un algorithme de *boosting* avec une généralisation basée sur la marge d'arbres de décision, appelés arbres de décision alternants (*alternating décision trees*). Cette méthodologie originale présente deux grands intérêts, le premier correspond à la possibilité d'analyser l'influence de perturbations génétiques expérimentales de type *Knock Out* (*KO*) sur le comportement d'un réseau de régulation prédit. Le deuxième intérêt consiste à simuler des perturbations *in silico* et à prédire le comportement du système dans des conditions qui n'ont pas encore été testées expérimentalement. Cependant, cette méthodologie reste limitée à des organismes ou systèmes où les potentiels régulateurs sont relativement bien connus, car ils constituent une des données d'apprentissage en entrée. Cette méthodologie permet d'inférer des relations de régulation entre gènes mais elle ne permet pas d'identifier de nouveaux régulateurs. Enfin, cette approche permet d'analyser l'influence de perturbations génétiques mais celles-ci restent limitées aux perturbations de type *KO* ou aux perturbations touchant directement l'expression des gènes régulateurs sur lesquels se base

l'apprentissage. Des variations de ploïdie (nombre de copies de chromosome) par exemple ne pourront pas être prises en compte pour prédire le comportement du système.

Parmi les méthodes d'apprentissage les plus populaires figurent les méthodes utilisant les arbres de décision. Dans le cadre de l'apprentissage supervisé, le but est de trouver une fonction d'un certain nombre de variables d'entrée qui estiment au mieux une variable de sortie, cela à partir d'une base de données d'exemples de paires entrée-sortie. Dans le cas de l'inférence de réseaux biologiques les bases de données à traiter comportent un nombre très important de variables d'entrée. Les méthodes d'apprentissage telles que les méthodes d'ensembles d'arbres de décision sont capables de fournir des résultats acceptables (en terme de précision) dans ces conditions extrêmes. Soinov *et coll.* [255] reconstruisent un réseau de régulation de gènes en construisant des classifieurs à base d'arbres de décision qui permettent de prédire l'expression d'un gène à partir de l'expression des autres gènes. Cette approche permet d'identifier les gènes qui affectent l'expression d'un gène cible directement à partir du classifieur sans passer par une discrétisation arbitraire des niveaux d'expressions.

Récemment de nouvelles approches se sont intéressées à la prédiction d'interactions entre protéines à partir de données hétérogènes [165, 256, 166]. Le but là aussi consiste à construire un classifieur à partir d'une base d'exemples et à prédire l'existence ou non d'une interaction entre deux protéines pour lesquelles on ne dispose d'aucune donnée sur l'expression de leurs gènes respectifs mais pour lesquelles d'autres types d'informations biologiques existent. L'intérêt de ces méthodes réside dans l'utilisation de méthodes à noyaux (voir description des méthodes à noyaux dans en annexe, page 227) en sortie des classifieurs leur permettant de traiter des sorties complexes. Ces méthodes à noyaux sont plus gourmandes en temps de calcul mais elles sont généralement plus précises. Les méthodes de Yamanishi *et coll.* et de Kato et *coll.* encapsulent les données d'entrées dans une fonction noyau et proposent un modèle de type "boite noire" qui ne donne qu'un faible aperçu du problème d'apprentissage. Geurts *et coll.* proposent une nouvelle méthode basée sur la transformation en noyau de l'espace de sortie d'arbres de régression. Contrairement aux autres méthodes utilisant les noyaux, cette méthode utilise l'espace d'entrée original (non transformé en noyau) et ainsi hérite pleinement de la facilité d'interprétation et de la robustesse des méthodes à base d'arbres de décision classiques.

Bien que ces méthodes n'aient été appliquées qu'à la prédiction d'interactions protéiques, nous pourrions envisager leur application à la reconstruction de réseaux de régulation de gènes. Cependant, ces méthodes ne s'appliquent qu'à la complétion de graphes partiellement identifiés. Cela signifie qu'il faut qu'une partie du réseau, suffisamment informative, aie déjà été identifiée pour servir de base d'apprentissage. Ces méthodes interviendraient alors dans une deuxième phase de complétion après une première phase, exploratoire, d'extraction de relations de régulation.

2.4.2 Méthodes basées sur la réduction de dimension et l'intégration de données systémiques

Toutes les méthodes d'inférence de réseaux ou de prédiction d'interactions vues précédemment essayent de reconstruire les interactions potentielles entre tous les gènes d'un réseau. Cependant, bien souvent le grand nombre de données à analyser ne permet pas d'utiliser les méthodes d'inférence de paramètres ou de structure de modèles. De plus, le nombre de mesures par gène reste très inférieur au nombre de paramètres à estimer. Ce problème de sous-détermination est aggravé par le faible rapport signal/bruit des données d'expression et le caractère stochastique des systèmes biologiques. Ces derniers aspects posent aussi un problème de précision et de fiabilité des données d'apprentissage pour les méthodes basées sur la prédiction d'interactions entre gènes ou protéines. Ces contraintes ont amené certains chercheurs à reformuler le problème de la reconstruction d'un réseau de régulation de gènes en un problème de reconstruction d'un réseau de régulation d'ensembles de gènes. Nous définissons un ensemble comme un groupe de gènes co-exprimés dans une ou plusieurs conditions expérimentales et potentiellement sous le contrôle de mêmes mécanismes de régulation. Le but consiste à identifier des ensembles de gènes co-exprimés puis de rechercher un ou plusieurs régulateurs communs (généralement des FTs) à chaque ensemble.
L'hypothèse de l'existence d'ensembles de gènes co-régulés est étayée par la structure des réseaux de régulation géniques connus. D'une part, un faible nombre de FTs doit réguler l'expression d'un génome entier : la proportion de séquences codant pour des protéines régulatrices varie peu entre des espèces comme *A. thaliana* (5,9%), *D. melanogaster* (4,5%), *C. elegans* (3,5%) et *S. cerevisiae* (3,5%) [257]. Cela signifie que les gènes doivent obligatoirement se partager un petit nombre de régulateurs. D'autre part, il a été montré qu'à l'image de l'organisation des gènes en opérons (gènes partageant une seule séquence promotrice) chez les procaryotes il existe une organisation des gènes en régulons (ensemble de gènes possédant leurs propres promoteurs mais dépendant d'une même combinaison de facteurs de transcription) chez les eucaryotes. Nous pouvons citer par exemple l'opéron *ribi* qui correspond chez la levure à un ensemble de plus de 200 gènes qui interviennent dans la biogenèse des ribosomes et qui sont régulés par une combinaison de 3 FTs [258]. Ces co-régulations sont bien conservées au sein des eucaryotes et des procaryotes et correspondent souvent à des ensembles de gènes intervenant au niveau des mêmes processus biologiques [259, 260, 261, 262]. Stuart *et coll.* [263] ont démontré l'existence de paires de gènes restant co-exprimés à travers un ensemble de 3182 puces à ADN chez l'Homme, la drosophile, le ver et la levure. De plus, l'analyse de la structure du réseau de régulation transcriptionnel chez la levure a montré que ce réseau est organisé en modules logiques distincts dépendant chacun d'une même combinaison de FTs et implémentant souvent une fonction commune [264].
L'identification de groupes de gènes co-exprimés réduit considérablement la complexité du problème car un réseau de régulation à l'échelle du génome (quelques milliers de gènes) va se réduire à quelques centaines voire dizaines de nœuds au maximum avec aussi une réduction du nombre des interactions possibles entre nœuds. De plus, travailler avec des groupes de gènes réduit l'incertitude due à la qualité et à la fiabilité du signal et permet l'application de tests statistiques pour analyser le contenu en information biologique de

ces groupes.

Les méthodes de classification sont parmi les plus utilisées pour identifier des groupes de gènes co-exprimés à partir des données d'expression. Ces méthodes s'apparentent plus à de la fouille de données qu'à de l'apprentissage (voir un état de l'art des différentes méthodes de classification de gènes présenté en annexe, page 222). Cependant, la classification de données d'expression fait quand même l'hypothèse d'une structuration intrinsèque des données et va tenter de révéler cette structure. Par contre, il n'apparaîtra pas de notion de causalité entre les gènes. Alors que les méthodes d'inférence de réseaux de régulation présentées précédemment tentent d'inférer des liens de dépendance entre les variables, la classification va occulter le caractère singulier de chaque variable et va plutôt chercher à extraire une organisation du réseau sous la forme de sous-ensembles. L'apport de la classification au niveau de l'inférence de réseaux de régulation réside principalement dans la réduction du nombre de variables à modéliser. L'extraction de liens entre modules ou entre un module et un régulateur peut cependant être effectuée *a posteriori* à l'aide d'études complémentaires comme l'analyse des décalages de modulations d'expression [265, 266, 267], l'exploitation de stratégies de perturbations [268, 269, 270], la projection des gènes co-exprimés dans les voies métaboliques ou dans les processus biologiques [178, 271] et la recherche de promoteurs ou de motifs d'ADN communs en amont des séquences codantes des gènes co-exprimés [270, 185, 272].

2.4.2.1 Exploitation de la dynamique d'expression

Les approches qui permettent à la fois d'identifier des groupes de gènes co-exprimés en recherchant des similarités entre cinétiques et d'extraire des relations de régulation en exploitant les décalages de temps entre des modulations d'expression restent assez rares. Qian *et coll.* [273] utilisent un algorithme d'alignement local, inspiré de l'algorithme d'alignement de séquences de Smith-Wathermann [78], pour identifier différents types de corrélations entre gènes :

- des corrélations "simultanées" (profils synchrones et superposables) : ce type de corrélations peut être le reflet de gènes soumis à une même régulation transcriptionnelle.
- des corrélations "décalées" (mêmes profils mais avec un décalage dans le temps) : le gène dont la modulation d'expression est précoce peut être un potentiel régulateur positif du gène dont la modulation d'expression est similaire mais plus tardive.
- des corrélations "inversées" (les profils de deux gènes sont synchrones mais inversés selon une symétrie axiale) : ces deux gènes peuvent être sous la dépendance du même programme de régulation transcriptionnelle mais celui-ci a des effets inverses sur les deux gènes.
- des corrélations "décalées" et "inversées" : le gène dont la modulation d'expression est précoce peut être un potentiel régulateur négatif du gène dont la modulation d'expression est inversement similaire mais plus tardive.

La probabilité de chaque relation entre deux gènes est validée statistiquement en la comparant avec sa probabilité d'apparition de façon aléatoire. La classification des gènes à partir des mesures de corrélation permet l'identification du réseau final.

Selon le même principe, Chen *et coll.* [267] identifient des groupes de gènes co-exprimés à l'aide d'un algorithme de classification. Ils résument chaque groupe à un nœud dans un graphe et le représentent par un profil d'expression consensus (moyenne des cinétiques d'expression des gènes du groupe). Chaque profil consensus est ensuite "lissé" de façon à faciliter l'identification des pics de modulation. Ils définissent une fonction "objectif" qui leur permet de mesurer à quel point le pic de modulation d'un profil peut être responsable de l'activation ou de l'inhibition d'un autre profil en utilisant à la fois la forme et l'amplitude des modulations mais surtout les décalages dans leurs temps d'apparition. Le graphe candidat obtenu est ensuite "élagué" pour ne garder que les motifs de régulation les plus représentatifs.

L'inférence de relations de régulation entre deux groupes de quelques dizaines à quelques centaines de gènes co-exprimés peut cependant paraître en contradiction avec ce que l'on connaît des mécanismes de régulation transcriptionnels. En effet, pour tous les génomes, il n'existe qu'un faible nombre de gènes codant pour des FTs (voir section 2.4.2, page 62). Il faudra alors prendre en compte le fait que les groupes correspondant aux nœuds régulateurs ne doivent contenir que peu de gènes soit directement lors de l'identification des co-expressions soit *a posteriori* en recherchant au sein de "groupes régulateurs" un sous-groupe de FTs. Pour un organisme comme la levure, les différents gènes codant pour des FTs sont relativement bien connus et la liste de ces gènes peut être utilisée comme information a priori.

2.4.2.2 Exploitation de perturbations

L'inférence de liens de régulation à partir de la dynamique d'expression des gènes ne nous permet pas de différencier les régulations directes des régulations indirectes. La génomique fonctionnelle a l'habitude d'utiliser des stratégies de perturbations, où par exemple un gène sera muté et, à partir de l'observation de l'état de la cellule après mutation on pourra potentiellement inférer une fonction au gène. Appliqué au transcriptome, ce type d'approche peut nous permettre d'identifier des relations de régulation entre gènes ou groupes de gènes et parfois, de confirmer l'effet direct ou indirect d'une régulation transcriptionnelle. Dans l'exemple suivant, $X \rightarrow Y \rightarrow Z$, X régule l'expression de Z via Y. Si on perturbe Y, on ne verra un effet que sur Z alors que si c'est X qui est perturbé, on verra l'effet de cette perturbation sur Y et Z à la fois. L'application d'une stratégie de perturbations peut permettre par simple déduction de révéler la structure d'un réseau de régulation avec le signe et l'orientation des liens. Il est cependant rare que l'on teste toutes les perturbations possibles au sein d'un réseau, on incorpore généralement des connaissances biologiques préalables dans la conception de la stratégie de perturbations.

Holstege *et coll.* [268] ont, parmi les premiers, associé stratégie de perturbations et analyse du transcriptome. Dans leur étude ils observent l'effet sur le transcriptome de la levure de 12 mutations ciblées au niveau des différents composants de la machinerie de transcription. Leur stratégie d'analyse reste assez simple : ils identifient à l'aide d'un seuil d'expression les sous-ensemble du génome qui sont dépendants de chaque gène muté. Puis, à l'aide de différents diagrammes de Venn ils identifient les dépendances multiples en recherchant les intersections des différents groupes de gènes sensibles aux mêmes mu-

tations.
Hughes *et coll.* [269] appliquent la même stratégie, mais à une plus grande échelle. Ils analysent la réponse transcriptionnelle de la levure dans 300 conditions correspondant à 266 mutations par délétions et 34 traitements avec divers composés chimiques. L'application d'une stratégie de classification hiérarchique en 2 dimensions leur permet d'identifier des sous-groupes de gènes sensibles à différents sous-groupes de mutants ou de composés chimiques. Puis, l'association d'ontologies fonctionnelles leur permet d'identifier des gènes impliqués dans la régulation de processus biologiques précis. Ils identifient également des combinaisons de régulateurs impliqués dans la régulation de fonctions communes et dessinent des ébauches de voies de régulation. Les deux précédentes études étaient basées sur l'analyse de données statiques et relevaient seulement l'état "allumé" ou "éteint" d'un gène.

Laub *et coll.* [270] se basent par contre sur des cinétiques d'expression de gènes pour identifier des motifs d'expression dépendant du cycle cellulaire de la bactérie *Caulobacter crescentus*. Ils analysent aussi l'effet de la mutation d'un régulateur connu du cycle cellulaire sur le transcriptome de cette bactérie. A l'aide d'un algorithme de classification de type cartes auto-organisatrices de Kohonen (voir description de cette méthodes de classification en annexe, page 231) ils identifient différents motifs d'expression périodiques dépendant du cycle cellulaire et identifient parmi ces motifs ceux qui sont sous la dépendance du gène muté. Bien que n'utilisant qu'une seule mutation, cette étude est intéressante par l'exploitation simultanée de données dynamiques et d'une stratégie de perturbation qui permet d'une part d'identifier de façon précise des motifs de co-expression et d'autre part de confirmer ou d'infirmer la dépendance de chaque groupe co-exprimé à la voie de régulation représentée par le gène muté. Gasch *et coll.* [60] et Mercier *et coll.* [59] appliquent exactement la même stratégie d'analyse chez la levure avec une mutation pour la première étude et 3 mutations, des variations du *mating type* et de la ploïdie pour la deuxième. Cependant, dans ces deux études l'aspect temporel des données reste sous-exploité.

2.4.2.3 Intégration de données hétérogènes

Les données du transcriptome ne contiennent pas suffisamment d'informations pour reconstruire un réseau de régulation de gènes global et complet en utilisant une approche par classification seule. Un nombre croissant de méthodes d'inférence de réseaux de régulation exploitent également d'autres types de données systémiques. Le très grand nombre de gènes généralement analysé par expérience, permet d'obtenir de nombreux motifs de co-expression. L'intégration de données biologiques additionnelles et leur "fusion" aux données d'expression permet de contraindre et éventuellement de valider les nombreuses hypothèses qui peuvent expliquer la co-expression de gènes. Pour reconstruire et comprendre un réseau de régulation par exemple, on aura besoin de combiner des données biologiques hétérogènes comme les associations protéines-ADN, protéines-protéines, l'analyse de promoteurs, les ontologies fonctionnelles et les données d'expression.

Outils d'analyse "clé en main" Les outils, sous forme de logiciels, qui permettent l'intégration de données biologiques hétérogènes aux données d'expression, ont été prin-

cipalement conçu pour évaluer la cohérence de groupes de gènes co-exprimés du point de vue d'annotations fonctionnelles issues de bases comme *Kegg* [178] ou *GO* [176]. Des outils comme *Pathway Processor* [274], *Pathway Miner* [275] et *THEA* (*Tools for High-throughput Experiments Analysis*) [276] offrent diverses fonctionnalités permettant l'analyse de données d'expression :

- Annotation automatique de groupes de gènes, obtenus par classification par exemple, avec des informations biologiques issues de bases de données ontologiques.
- Projection et visualisation des données d'expression dans les voies métaboliques issues de *GO*, *Kegg*, *GenMAPP* ou *BioCarta*.
- Visualisation de groupes de gènes et systèmes de requêtes à partir des annotations.
- Évaluation, à l'aide de tests statistiques, des voies métaboliques ou processus biologiques les plus affectés par des changements transcriptionnels.

Ces outils restent très spécifiques de l'analyse de groupes de gènes à partir d'ontologies fonctionnelles mais ils peuvent être détournés en remplaçant les bases de données inclues par défaut dans ces outils par des bases de données "maison" ou construites à partir d'autres types d'information comme les associations gènes-régulateurs ou protéines-protéines. Dans ces cas, les fonctionnalités statistiques et graphiques des outils peuvent aussi s'appliquer et permettre l'intégration et l'analyse de différents types de données biologiques systémiques sous la forme d'associations gène-descripteur.
Ces outils restent cependant limités à des analyses par groupes de gènes et ne vont pas jusqu'à la reconstruction et l'analyse de réseaux. *Cytoscape* [277] par contre propose un cadre pour intégrer des réseaux d'interactions moléculaires aux données d'expression. Il est alors possible de visualiser et d'effectuer des requêtes sur un réseau, d'intégrer un réseau à des données d'expression, des ontologies fonctionnelles, des phénotypes et tout autre type d'information systémique. De plus, *Cytoscape* est relativement plastique et évolutif grâce à l'ajout régulier de nouveaux composants (*plug-ins*). Certains composants proposent des outils pour l'inférence de réseaux à partir de différentes sources d'informations :

- *AgilentLiteratureSearch* : créé un réseau d'interaction biologique (*CyNetwork*) à partir d'informations biologiques hétérogènes et d'une recherche dans la littérature scientifique (voir `http://www.agilent.com/labs/research/litsearch.html` et [278]).
- *Cytoprophet* : permet de prédire de potentielles nouvelles interactions protéiques en y associant un score (voir `http://cytoprophet.cse.nd.edu/` et [279]).
- *MetaNetter* : permet l'inférence de réseaux métaboliques en se basant sur des données métaboliques expérimentales (voir `http://compbio.dcs.gla.ac.uk/fabien/abinitio/abinitio.html` et [280]).
- *MONET* : permet d'inférer des réseaux d'interactions génétiques à partir de données d'expression en utilisant un réseau bayésien (voir `http://delsol.kaist.ac.kr/~monet/home/` et [281]).

Cytoscape paraît être un bon cadre pour l'analyse de données transcriptomiques et l'intégration de données systémiques hétérogènes, cependant, ce logiciel ne reste qu'une plateforme proposant différents outils d'analyse ou d'inférence sans liens méthodologiques entre eux ni stratégie d'analyse globale.
Enfin, le *Laboratory for BioInformatics and Functional Genomics* (`http://function.princeton.edu/software.html`) dirigé par O. G. Troyanskaya propose une palette d'ou-

tils pour l'analyse de données d'expression et pour l'intégration d'informations génomiques. Ces outils n'existent cependant que sous la forme d'applications indépendantes sans liens méthodologiques entre eux. De plus, ils fonctionnent selon le principe de "requêtes d'experts", c'est à dire que l'on doit avoir déja identifié un groupe de gènes d'intérêt avant de le soumettre aux outils d'analyse. Ces outils ne sont donc pas utilisables dans des approches globales et exploratoires.

Une analyse des méthodologies globales et intégratives, plutôt que des logiciels ou des outils "clé en main", m'a permis de regrouper ces méthodologies en deux grandes familles. La première famille d'approches correspond à des méthodologies qui essayent d'intégrer différents types de données biologiques de façon unifiée pour en extraire des réseaux de régulation. Dans la deuxième famille d'approches, les différentes sources d'information sont exploitées de façon séquentielle ou hiérarchisée avec au moins deux étapes d'analyse différentes. Dans les deux sections suivantes je présente différents exemples parmi les plus caractéristiques de ces deux familles d'approches.

Méthodologies basées sur une intégration unifiée d'informations biologiques hétérogènes L'approche la plus connue correspond aux classifications réalisées à partir de plusieurs dimensions à la fois, approches de type *biclustering* (voir description de ces méthodes en annexe, page 239). Dans [282] Liu *et coll.* introduisent l'information issue de la base *GO* dans la phase de classification et à l'aide d'un algorithme de *biclustering* essayent d'extraire des sous-ensembles de gènes co-exprimés reflétant des fonctions communes. Dans [283], Zaho et *coll.* proposent un nouvel algorithme pour identifier des groupes cohérents et superposés dans des données d'expression en 3 dimensions (gènes, temps et conditions). Cette approche est assez intéressante dans la mesure où elle peut permettre une analyse pertinente de données d'expression dans des conditions de perturbations. Cependant, elle reste limitée aux 3 dimensions relatives à la démarche expérimentale et les auteurs ne semblent pas prendre en compte de possibles extensions à d'autres dimensions (ontologies fonctionnelles par exemple).

Il existe aussi des approches intégratives beaucoup plus complexes comme celle de Tanay *et. coll* [284] qui proposent une modélisation intégrative de données génomiques sous la forme d'un graphe biparti pondéré. A l'aide d'un algorithme de *biclustering*, ils essayent d'identifier des modules de régulation et de reconstruire une organisation globale, hiérarchique et modulaire de la levure vue comme un système. En plus des données d'expression, des interactions protéiques et des associations gènes-TFs, ils intègrent des données phénotypiques de croissance. Selon le même principe que celui adopté par Tanay *et coll.* [284], Reiss *et coll.* [285] essayent de détecter des groupes de gènes potentiellement co-régulés avec une approche de type *biclustering*. Dans ce travail, ils exploitent 3 types de données génomiques : des données d'expression, des séquences d'ADN régulatrices, et des données d'association (profils phylogénétiques, informations métaboliques, annotations fonctionnelles, interactions protéiques, etc.). Chaque bi-classe (*bicluster*) est modélisée via un processus de chaîne de Markov dans lequel la bi-classe est optimisée de façon itérative. L'originalité de cette approche réside dans sa flexibilité, en effet, il est possible de donner plus de poids à un type de données lors de la construction itérative des

bi-classes. Par exemple, leur stratégie permet de prendre en compte la qualité de chaque type de données ou l'existence de connaissances *a priori*. Ainsi, il sera possible d'initialiser une bi-classe à l'aide de l'information sur l'expression des gènes seulement et de raffiner ou d'enrichir de façon itérative la bi-classe à l'aide des autres types d'information.

D'autres types d'algorithmes se basent sur le principe d'organiser (partitionner) les données selon plusieurs dimensions à la fois. Kasturi et Acharya [286] proposent d'identifier des groupes de gènes dont la co-expression serait corrélée à la présence de motifs d'ADN communs en amont des séquences codantes. L'originalité de leur algorithme, inspiré des cartes auto-organisatrices de Kohonen, réside dans le choix aléatoire de la catégorie de données à utiliser à chaque étape de la phase d'apprentissage itérative. Testé avec deux types d'informations biologiques seulement, cet algorithme pourrait être étendu à d'autres dimensions. De plus, cet algorithme bénéficie de la même flexibilité que l'algorithme de Reiss *et coll.* [285] dans la mesure où l'on peut favoriser l'apprentissage à partir d'une dimension au détriment d'une autre.

Enfin, Steinhauser *et coll.* [287] proposent d'identifier des modules transcriptionnels à partir de données d'expression en y fusionnant des informations sur l'organisation des gènes sur les chromosomes (distances inter-géniques et organisation en opérons). L'originalité de cette approche réside dans l'intégration des informations biologiques additionnelles en amont de la phase d'apprentissage ou de classification. En effet, chaque type d'information permet de calculer une similarité "locale" entre gènes, puis, ces similarités interviennent dans le calcul d'une similarité globale qui est utilisée par un algorithme de classification hiérarchique. Cette approche diffère des précédentes dans la mesure où la fonction de coût à optimiser lors de la classification correspond à une requête plus contraignante : elle impose la recherche de classes où les gènes sont similaires selon toutes les dimension. Dans les approches de type *biclustring*, la fonction de coût permet d'identifier des biclasses de gènes qui ne sont cohérents que selon un sous-ensemble de chaque dimension. Les approches de *biclustering* peuvent être vues comme une généralisation et une relaxation de cette dernière approche.

Méthodologies basées sur une intégration séquentielle et hiérarchisée d'informations biologiques hétérogènes Dans ce deuxième type d'approches, les natures spécifiques des données à intégrer sont prises en compte pour en extraire des informations biologiques spécifiques. Chaque type d'information ne fait pas toujours l'objet d'une démarche analytique propre mais on trouve au minimum 2 étapes d'analyse avec une séparation entre la donnée générée par une stratégie expérimentale et des données biologiques additionnelles utilisées pour valider ou enrichir les résultats obtenus à partir des premières données. Dans ce type de démarche les données de départ sont généralement les données d'expression. Il existe cependant des cas où la première étape de l'analyse exploite un autre type d'information que les données d'expression. Dans [288], Lars et Steen partent d'informations de séquences pour identifier des groupes de gènes ayant les mêmes motifs de régulation, puis, ils recherchent de potentielles co-expression au sein de ces groupes. De la même façon, Ben-Shaul *et coll.* [289] construisent à partir de *GO* des

groupes de gènes partageant les mêmes fonctions puis analysent les variations d'expression de ces groupes dans différentes conditions expérimentales. Je présenterais essentiellement les méthodologies qui se basent en premier sur l'analyse des données d'expression. D'une part cela correspond au cas général, d'autre part, rares sont les autres types d'information contenant une notion de causalité, que ce soit en réponse à des modifications de l'environnement cellulaire ou entre les gènes eux-mêmes.

Les démarches les plus simples, généralement en 2 étapes seulement, consistent à identifier des motifs de co-expression puis à les confronter à un autre type de données génomiques. Simonis *et coll.* [290] identifient des groupes de gènes co-exprimés puis recherchent dans ces groupes des motifs de régulation communs significativement enrichis. Bar-Joseph et al [291] adoptent une démarche différente de la précédente, leur point de départ consiste à exploiter des données ChIP pour identifier des groupes de gènes partageant les mêmes régulateurs, puis ils vérifient si les gènes d'un même groupe sont co-exprimés à l'aide de données transcriptomiques. Ils identifient ainsi des modules de gènes co-régulés et relient les différents modules pour construire un réseau de régulation global.

Les démarches les plus intéressantes restent cependant celles qui mettent en place une stratégie d'analyse complexe et qui procèdent en plus de 2 étapes. Parmi ces démarches nombreuses sont celles qui s'intéressent à la réponse transcriptionnelle à un stimulus ce qui justifie l'exploitation différenciée des données d'expression ou des données directement liées à la réponse au stimulus. Ces méthodologies parviennent à la fois à analyser la réponse transcriptionnelle d'un système et à inférer un réseau de régulation génique impliqué dans cette réponse. Parmi les approches les plus intéressantes, Haverty *et coll.* [292] ont développé une méthodologie, appelée *CARRIE*, pour l'inférence d'un réseau de régulation transcriptionnel, impliqué dans la réponse à un stimulus. La première étape consiste à identifier l'ensemble des gènes qui répondent au stimulus à l'aide d'un test statistique appliqué aux données d'expression. L'étape suivante consiste à identifier les FTs pour lesquels des motifs de fixation sont sur-représentés dans les régions promotrices des gènes co-exprimés. Dans ce travail, Haverty *et coll.* réussissent à intégrer une information fonctionnelle (expression des gènes) à une information structurale (motifs de régulation) afin d'identifier une réponse à un stimulus et afin de construire un réseau de régulation spécifiquement impliqué dans cette réponse. Ici, la méthodologie est initiée par l'identification des gènes directement impliqués dans la réponse au stimulus, ce qui signifie que la relation de causalité entre les cibles du réseau de régulation et le stimulus est clairement identifiée. Cependant, les variations d'expression entre gènes ne sont pas exploitées pour identifier différentes formes de réponses au stimulus. La réponse au stimulus est ainsi identifiée comme un seul grand ensemble de gènes dont l'expression est modulée.

Workman *et coll.* [161] se sont intéressés à l'analyse une réponse cellulaire (réponse aux dommages à l'ADN) et à l'extraction d'un réseau de régulation de gènes impliqué dans cette réponse. Leur méthodologie est constituée de 6 étapes successives et exploite 4 types d'informations biologiques. Ils commencent par sélectionner, à partir de la littérature et de l'annotation des gènes, les FTs impliqués de façon directe ou indirecte dans la réponse aux dommages à l'ADN. Puis, ils identifient les cibles de ces FTs en utilisant

une analyse différentielle de données ChIP en comparant les résultats obtenus dans une condition normale à ceux obtenus en condition génotoxique. Ensuite, à partir d'ensembles de gènes associés à un même TF, ils recherchent les motifs de régulation communs au sein de chaque ensemble et identifient des combinaisons de TFs qui agissent sur les mêmes gènes. Ils ajoutent à leur méthodologie une nouvelle étape expérimentale où ils valident l'effet de chaque régulateur sur l'expression des gènes à l'aide de mesures d'expression réalisées dans différentes conditions de perturbations (mutation du gène de chaque FT) et avec ou sans agent génotoxique. A l'aide d'une approche bayésienne, ils identifient ce qu'ils appellent des gènes "tamponnés", dont la réponse aux dommages est "tamponnée" par la délétion d'au moins un FT. Dans la dernière étape de leur méthodologie, ils comparent les interactions inférées à partir des données ChIP à celles inférées à partir des données d'expression. Cette comparaison leur permet d'identifier, à l'aide d'un modèle bayésien, des relations de régulation directes et des voies de régulation indirectes impliquant des interactions proteines-proteines et proteines-ADN plus complexes. L'originalité de cette approche est qu'elle débute par la construction du réseau impliqué dans une réponse cellulaire où l'identification des régulateurs potentiels permet de fixer une partie de la topologie du réseau. De plus, l'identification de ces régulateurs à l'aide d'une analyse différentielle entre condition normale et condition génotoxique permet d'avoir une relation de causalité directe entre les régulateurs du réseau et la réponse au stress génotoxique. Il est intéressent de noter que Workman *et coll.* utilisent une stratégie de perturbation clairement définie pour extraire des liens de régulation. Cependant, les effets de ces perturbations ne sont pas analysés au sein d'un cadre formel, par un processus de déduction automatique par exemple. De plus, les ontologies fonctionnelles (GO) ne sont utilisées que pour l'interprétation des régulateurs mais pas pour celle des gènes régulés et répondant au stress. Les annotation de GO auraient pu être utilisées par exemple pour analyser les groupes de gènes sous la dépendance d'une même combinaison de régulateurs.

Ideker *et coll.* [175] ont aussi proposé une méthodologie d'analyse qui se base sur l'exploitation d'une stratégie de perturbations génétiques pour la construction d'un réseau de régulation. Ils proposent de construire et d'affiner un modèle de régulation d'une voie cellulaire en exploitant de façon séquentielle 3 types d'information : les expressions de gènes mesurées dans différentes conditions de perturbations, des données protéomiques quantitatives et des interactions physiques connues (proteines-protéines et protéines-ADN). Leur approche consiste à définir une voie cellulaire d'intérêt, à identifier tous les gènes impliqués dans cette voie, puis à perturber chaque composant de cette voie (perturbations génétiques : délétions et sur-expressions et perturbations environnementales). Ils identifient et quantifient ensuite la réponse cellulaire à chaque perturbation (modulation du transcriptome et du protéome). Enfin, ils intègrent les réponses observées au niveau du transcriptome et du protéome et les interactions déjà connues au modèle spécifique de la voie cellulaire d'intérêt. Ce travail est remarquable à deux points de vue : premièrement, l'exploitation de la stratégie de perturbations permet d'affiner la topologie du modèle et d'identifier des groupes de gènes dont les expressions restent cohérentes à travers toutes les perturbations du système. Deuxièmement, un grand apport méthodologique est réalisé grâce à l'intégration de 3 niveaux d'observation du comportement cellulaire (transcriptome, protéome et interactions physiques) afin d'extraire des régulations directes et

indirectes au niveau transcriptionnel et afin d'identifier de potentielles régulations post-traductionnelles. Cependant, si cette approche est efficace pour compléter la topologie et la connectivité d'un modèle prédéfini, elle semble être plus difficile à mettre en œuvre pour une problématique plus exploratoire comme l'identification d'une réponse cellulaire globale.

Toutes les approches présentées dans cette section semblent particulièrement adaptées à à la caractérisation de réponses cellulaires à des variation de l'environnement ou à l'analyse de processus physiologiques précis. Elles permettent également d'inférer des réseaux de régulation impliqués dans les réponses cellulaires en exploitant à la fois des données expérimentales et des données biologiques externes de sources différentes. Cependant, aucune de ces méthodes ne prend en compte l'aspect dynamique des réponses cellulaires et négligent une grande partie de l'information qui pourrait être extraite de données cinétiques.

2.5 Méthodes "mixtes"

Par le qualificatif de "mixte" j'entends des méthodes qui vont associer une étape de réduction de dimension et/ou une étape d'intégration de données biologiques systémiques afin d'identifier une ébauche de structure de réseau et une étape d'estimation de paramètres ou de structures de modèles. Ces méthodes semblent assez prometteuses dans la mesure où elles essayent d'exploiter les avantages de plusieurs familles d'approches au sein d'une seule méthodologie. Leurs stratégies fonctionnent généralement par l'affinement progressif de la granularité d'un réseau de régulation et/ou la dimension des solutions possibles pour arriver à un réseau de relativement bien décrit.

Plutôt que de considérer chaque gène comme un nœud distinct et possédant ses propres paramètres au sein d'un modèle bayésien, Segal *et coll.* [293] regroupent d'abord les gènes qui partagent les mêmes paramètres et un même ensemble de régulateurs en modules. Cela réduit considérablement le nombre des paramètres du modèle à estimer et en même temps augmente le nombre de points de mesure disponibles pour l'estimation de chaque paramètre. Pour chaque module, l'effet de l'ensemble des régulateurs sur le profil d'expression de chaque gène du module est modélisé en tant que programme transcriptionnel en utilisant un arbre de régression. La procédure itérative utilise un algorithme *EM* pour la recherche du programme de régulation optimal pour chaque module et pour éventuellement réassigner chaque gène au module pour lequel le programme prédit au mieux son comportement. Bien que assez innovante, cette méthode ne permet pas l'exploitation d'autres sources d'informations que les données d'expression de gènes.

Myers *et coll.* [294] ont développé une méthodologie probabiliste pour l'identification de réseaux d'interactions biologiques en intégrant différents types de données systémiques. Les différents types d'informations sont intégrés grâce à un réseau bayésien qui pondère chaque information pouvant prouver une relation entre deux gènes. Ils obtiennent ainsi un graphe dont les nœuds sont des gènes et dont les arcs représentent les relations entre ces gènes pondérées par un niveau de confiance associé au type et à la qualité des don-

nées. L'exploitation de ce graphe est réalisée grâce à des requêtes correspondant à des ensembles de gènes (ou protéines). Pour chaque requête, un algorithme de prédiction va essayer de retrouver les relations qui lient ces gènes et va essayer de compléter le réseau extrait en identifiant de nouveaux composants (nœuds possédant le maximum de liens directs et indirects avec les nœuds de la requête). L'approche de Myers *et coll.* intègre dans un cadre probabiliste l'approche déjà développée sous la forme de l'outil en ligne *STRING* [163]. Cette méthodologie représente un outil de fouille de données assez intéressent dans la mesure où il permet d'extraire des relations entre gènes à partir de données de natures extrêmement diverses (données d'expression, interactions physiques entre protéines, interactions génétiques, analyse de séquences régulatrices, localisations chromosomiques des gènes et fouille de la littérature) et où ces relations sont associées à une mesure d'incertitude liée directement à la qualité des données. Cependant, cette méthodologie reste limitée à un système de requêtes (40 gènes maximum par requête) et ne permet pas d'inférer et d'analyser un réseau de régulation transcriptomique global. Cette méthodologie pourrait être par exemple complémentaire d'une étape de classification de gènes à partir de données d'expression, où l'on chercherait à identifier de potentielles corrélations entre co-expression et relations gène à gène d'autres natures. Même si elle est justifiée par un collège de biologistes, la pondération des liens du réseau reste arbitraire et subjective et doit être appréhendée avec précaution.

Enfin, un dernier exemple original d'une approche "mixte", a été récemment proposé par Elati et coll. [295] et appliqué à la reconstruction du réseau de régulation génétique de *S. cerevisiae*. Cette approche, en 3 étapes associe fouille de données et apprentissage/prédiction d'un modèle de régulation. La première originalité de ce travail réside dans l'appréhension du problème de l'inférence de réseaux de régulation génique "à rebours" de ce qui est généralement effectué par les autres méthodes. En effet, au lieu de chercher à obtenir directement les nœuds cibles ou les relations régulateur/cible d'un réseau de régulation, ils identifient d'abord tous les régulateurs potentiels du réseau. La deuxième originalité réside dans le modèle de régulation qu'ils adoptent, celui-ci représente ce qu'ils appellent un *programme de régulation* où un gène est potentiellement régulé par deux ensembles de co-régulateurs, l'un constitué de co-activateurs et l'autre de co-inhibiteurs. La première étape de leur méthodologie consiste à générer pour chaque gène un ensemble de groupes de co-régulateurs où chaque groupe correspond à des régulateurs fréquemment co-exprimés dans les données transcriptomiques. A cette étape, la dimension des régulateurs potentiels est déjà fixée par l'exploitation de connaissances biologiques *a priori* (une liste de 475 régulateurs tirés de la littérature). La deuxième étape de leur méthodologie consiste à calculer pour chaque gène cible un nombre limité de réseaux de régulation et à rechercher parmi cet ensemble, les groupes de co-activateurs et de co-inhibiteurs qui expliquent le mieux l'expression du gène cible. Et enfin, la troisième étape correspond à la sélection, grâce à une méthode de permutations, des réseaux de régulation significatifs d'un point de vue statistique parmi tous les réseaux inférés pour tous les gènes cibles. Elati *et coll.* appliquent leur méthode à deux jeux de données transcriptomiques, l'un statique [108] et l'autre dynamique [185]. Les réseaux inférés à partir de chaque jeux de données sont ensuite validés en les confrontant aux connaissances actuelles concernant la levure (littérature, régulations déjà connues, données ChIP, ontologies fonctionnelles, interactions

protéiques). Elati *et coll.* réussissent à résoudre diverses problématiques méthodologiques en une seule approche :

- Leur approche reste totalement exploratoire car elle prend en compte la totalité des données mesurées à l'aide de puces à ADN. L'espace des solutions à explorer est déja réduit par l'utilisation de connaissances biologiques *a priori* et par les contraintes liées à la structure du modèle de régulation à apprendre (recherche de co-régulateurs).
- Leur approche peut être appliqué à tout type d'organisme car elle n'exploite que les données du transcriptome généralement disponibles en grand nombre ou alors faciles à produire.
- Enfin, l'aspect dynamique des cinétiques d'expression est pris en compte grâce à l'adoption d'une discrétisation ternaire qui leur permet de représenter une cinétique d'expression par une suite de pentes : modulation positive (+1), modulation négative (-1), pas de modulation (0).

Chapitre 3

Problématique biologique et méthodologique : l'analyse de la réponse à l'irradiation

3.1 Réponse cellulaire à l'irradiation : état des connaissances

3.1.1 Effets biologiques des radiations ionisantes

Les radiations ionisantes vont potentiellement interagir avec toutes les macromolécules biologiques. Les radiations créent des réactions oxydatives qui modifient physiquement les macromolécules [296, 297] et entraînent l'hydrolyse des molécules d'eau, aboutissant à la formation de radicaux libres qui peuvent à leur tour altérer, dégrader ou lier les macromolécules cellulaires [298, 299].

Les radiations ionisantes peuvent causer des dommages au niveau des différentes membranes, plasmiques, nucléaires [300] et mitochondriales [301]. Un des principaux effets cellulaires des radiations ionisantes est l'endommagement de l'ADN [50, 51]. Les dommages à l'ADN, s'ils ne sont pas éliminés entraînent la mort de la cellule ou alors l'altération de son patrimoine génétique avec le risque de transmettre cette altération à sa descendance. Les radiations ionisantes peuvent causer différents types de lésions au niveau de l'ADN :

- Altérations au niveau des bases nucléiques.
- Cassures simple brin.
- Cassures double brin.
- Formation de pontages entre macromolécules.
- Sites de dommages multiples.

Dès qu'une cellule détecte des dommages à son ADN, elle arrête sa progression dans le cycle cellulaire. Cet arrêt du cycle est crucial car la cellule doit absolument réparer les lésions avant d'amorcer sa mitose sinon elle risque, soit de propager des anomalies chromosomiques, soit de rester bloquée en mitose sans pouvoir achever sa division. Différents cas de figures peuvent se présenter en fonction de la gravité des dommages à l'ADN :

- Si les lésions ne sont pas trop graves ni trop nombreuses la cellule va mettre en

place des mécanismes de réparation adaptés à chaque type de lésion. Si la cellule réussit à réparer fidèlement toutes les lésions, elle va pouvoir poursuivre son cycle normalement et entrer en mitose.

- Si jamais la cellule ne répare pas fidèlement une altération au niveau de bases nucléiques, cela peut générer une mutation au niveau de la partie codante ou régulatrice d'un gène. Cette mutation est alors transmise aux cellules filles après mitose. Si cette mutation n'est pas létale, elle peut avoir de grandes conséquences sur la physiologie de la cellule. Dans le cas d'eucaryotes supérieurs comme l'Homme nous pouvons différencier deux cas, soit la mutation touche une cellule germinale et alors elle risque d'être transmise à la descendance, soit cette mutation touche une cellule somatique et peut être la première cause d'un processus de cancérogénèse [302].
- Si les lésions à l'ADN sont trop importantes, la cellule ne pourra pas les réparer et mourra si la dose d'IR était très élevée.

3.1.2 Signalisation de la réponse à l'irradiation

En plus des effets des rayonnements ionisants, les cellules peuvent être exposées à différents types de stress cyto- et génotoxiques. La surveillance de l'intégrité du génome est un processus physiologique vital pour les cellules. Afin de maintenir l'intégrité de leurs génomes, les cellules ont donc mis en place, au cours de l'évolution, des mécanismes de surveillance et de réparation complexes. La réponse cellulaire à un stress génotoxique comme les radiations ionisantes sera initiée par la détection de dommages à l'ADN. Cette réponse cellulaire consistera essentiellement en l'induction de mécanismes de réparation complexes, adaptés à chaque type de lésion à l'ADN, afin de permettre une transmission fidèle du message génétique. Cette réponse sera aussi accompagnée d'un arrêt du cycle cellulaire, de la modulation de l'expression d'un grand nombre de gènes et parfois se terminera par l'entrée en apoptose et la mort de la cellule.

Les différentes formes de la réponse cellulaire aux dommages à l'ADN sont médiées par une cascade de protéines kinases qui semblent avoir été conservées à travers l'évolution pour tous les eucaryotes. Au sommet de cette cascade, on trouve une famille de phosphoinositol kinases qui incluent les protéines ATR et ATM chez les mammifères et Mec1p et Tel1p chez la levure. Ces complexes protéiques jouent un rôle central en tant que détecteurs des dommages à l'ADN [303, 304, 305] et en tant que régulateurs de la réparation ou de l'apoptose et de la survie cellulaire [55]. Ensuite, suivent deux classes de kinases dites de *checkpoint* : CHK1 et CHK2 chez les mammifères et Chk1p et Rad53p chez la levure. Chez la levure, Rad53p est suivie d'une kinase additionnelle, Dun1p. Cette dernière kinase est à la fois impliquée dans l'arrêt du cycle et dans la régulation transcriptionnelle de la réponse aux dommages à l'ADN [306, 307]. Chez les mammifères, CHK2 fait le lien entre les protéines de réparation et le contrôle du cycle cellulaire. L'activation de la voie ATM $\rightarrow$ CHK2 provoquera un arrêt du cycle en G2 et l'activation de la voie ATM $\rightarrow$ CHK2 $\rightarrow$ p53 provoquera un arrêt en G1 ou l'apoptose [308, 309]. Des mutations au niveau des voies ATM/Mec1p entraînent une hypersensibilité aux agents génotoxiques et des prédispositions aux cancers chez les mammifères. Chez la levure, les dommages à l'ADN activent des checkpoint au niveau de quatre positions du cycle cellulaire : la tran-

sition G1/S (checkpoint G1), pendant la phase S pour empêcher la réplication de l'ADN (checkpoint S), avant la mitose (S/M checkpoint) et G2/M (G2/M checkpoint). Chaque checkpoint activé doit être capable de reconnaître un type donné de lésion. Par exemple, les cassures double brin générées par les radiations ionisantes provoquent un arrêt G2/M avant l'entrée en mitose pour empêcher la perte de fragment de chromosomes [52, 53]. Les modifications de bases quand à elles inhibent le réplication et la progression dans la phase S [54]. Les systèmes de réparation sont directement intégrés au réseau cellulaire de régulation et de signalisation.

De plus en plus d'observations suggèrent un rôle important des différents organites cellulaires dans la détection du stress provoqué par les radiations ionisantes. Ces organites participeraient aussi à la détection des dommages, et à partir d'un certain seuil de dommages, déclencheraient des réponses locales et globales. Par exemple, des altérations au niveau du nucléole pourraient activer les voies mitochondriales qui ensuite activeraient p53 afin d'induire un arrêt du cycle ou l'apoptose. Ces échanges de signaux entre compartiments sub-cellulaires sont médiés par des translocations de protéines et sont aussi sujets à des régulations post-traductionnelles (principalement des phosphorylations) et transcriptionnelles [194]. Il reste cependant de nombreuses zones d'ombre dans la compréhension de la façon dont la réponse aux dommages à l'ADN est coordonnée. Et, concevoir l'IR comme une simple source de dommages à l'ADN est une vision certainement incomplète dans la mesure où des études récentes ont montré l'existence d'un effet indirect de l'IR sur les cellules (*bystander effect*) [310, 311] par l'intermédiaire de l'ionisation de l'eau, de probables composés dérivés encore à identifier [312, 313] et de jonctions intercellulaires [314, 315].

3.1.3 Vers une analyse globale et intégrative de la réponse à l'irradiation

La réponse cellulaire à l'IR est essentiellement décrite comme une réponse aux dommages à l'ADN, allant de la détection des lésions à la mise en place des mécanismes de réparation appropriés ou au déclenchement de l'apoptose dans le cas échéant. Chez l'Homme, la transduction du signal, de la détection des dommages jusqu'à l'activation des voies de réparation passe essentiellement par la cascade de kinases et par des modifications post-traductionnelles, de nombreux facteurs de transcription, qui ne sont pas tous identifiés, interviennent aussi en aval de la cascade de réactions de phosphorylation. Jusqu'à présent, les biologistes ont principalement cherché à disséquer les mécanismes de la réponse aux dommages à l'ADN, comprenant le contrôle du cycle cellulaire, la mise en place des mécanismes de réparation et l'activation des voies apoptotiques. Or, la réponse cellulaire l'IR semble être beaucoup plus globale et doit très certainement impliquer d'autres processus physiologiques. Différents niveaux de régulations peuvent potentiellement intervenir dans l'intégration de ces différentes réponses :

- au niveau transcriptomique, et en particulier chez la levure, de nombreux FTs peuvent être induits, inhibés, activés, délocalisés ou dégradés en réponses à différents stress.

- au niveau des modifications post-traductionnelles, de nombreuses voies de régulation par réactions de phosphorylation/déphosphorylation sont à l'œuvre.
- au niveau de la dynamique des protéines, on sait que suite à une IR, de nombreuses protéines se délocalisent d'un compartiments cellulaire à un autre. De plus, de nombreuses protéines acquièrent leur activité régulatrice ou enzymatique au sein de complexes protéiques le plus souvent hétéromériques.

L'identification d'interactions protéine-gène ou protéine-protéine permettrait de construire à chacun des niveaux décrit ci-dessus des ébauches de réseaux de régulation dont l'intégration et l'association aux processus physiologiques permettrait d'obtenir un modèle de la réponse globale à l'IR et de sa régulation.

Plusieurs études globales et exploratoires ont entrepris d'analyser la réponse transcriptionnelle à de fortes doses d'IR chez la levure. Certaines études [56, 57] (*S. cerevisiae*) et [316] (*S. pombe*) utilisent un seul temps de mesure de l'expression des gènes après IR et identifient des réponses pour environ 200 gènes dans chaque condition expérimentale. D'autres études utilisent des mesures cinétiques de l'expression des gènes post-IR chez *S. cerevisiae*, ce qui leur permet d'identifier des réponses beaucoup plus globales (plus de 600 gènes avec 5 temps de mesures [58] et plus de 1300 gènes avec 7 temps [59] (voir notre article III en annexe) et 9 temps de mesures [60]). Cependant, dans ces derniers travaux, les données d'expression ne sont analysées que point par point ce qui ne permet d'identifier qu'un seul large groupe de gènes sensibles à l'IR (à diviser en gènes induits et en gènes réprimés) par condition expérimentale. Ces approches statiques ne permettent pas d'identifier différents groupes de gènes co-exprimés, ce qui refléterait potentiellement l'hétérogénéité des dynamiques de réponses. La plupart de ces travaux analysent l'effet de mutations sur les réponses à l'IR pour essayer d'identifier des voies de régulation. Les dépendances aux mutations sont identifiées par des intersections de groupes (diagrammes de Venn) ce qui limite le nombre des comparaisons et ne permet de sélectionner que de très petits groupes de gènes. Gasch *et coll.* quant à eux [60] appliquent une stratégie de classification qui permet de regrouper les gènes en fonction de leurs cinétiques dans toutes les conditions expérimentales analysées. Cette approche leur permet d'identifier plusieurs groupes de gènes caractérisant différentes réponses à l'IR (et au MMS). Cependant, leur approche reste statique dans la mesure où ils considèrent les différents points d'une cinétiques comme indépendants, de plus, les ensembles de gènes extraits ne sont pas analysés du point de vue de leurs dynamiques.
Tous les travaux cités ici restent cantonnés au niveau transcriptionnel et n'intègrent pas d'autres niveaux de réponses ou de régulation. Récemment, Molin *et coll.* [317] ont proposé une analyse de profils d'expression protéiques en réponse à l'IR chez *S. cerevisiae* et ont identifié l'induction spécifique de différentes protéines impliquées dans la réponse au stress oxydatif. Leur travail est particulièrement intéressant dans la mesure où il considère le produit final de l'expression des gènes mais l'approche utilisée ne leur permet d'identifier qu'une vingtaine de gènes et ne permet donc qu'une analyse partielle de la réponse à l'IR.

3.2 Données disponibles et problématique méthodologique

Nous voulons poursuivre nos précédents travaux concernant l'analyse de la réponse à l'IR [59] (voir article III en annexe) en formulant et en vérifiant de nouvelles hypothèses relatives à la régulation de cette réponse. Nous souhaitons aussi adopter une approche véritablement dynamique dans l'analyse des cinétiques d'expression des gènes et dans l'interprétation de la réponse à l'IR. Après l'identification de potentielles réponses à l'IR, nous voudrions les analyser du point de vue d'autres types de données systémiques (ontologies fonctionnelles, littérature) ou d'autres niveaux d'observation du comportement cellulaire (protéome, régulome). Enfin, nous souhaitons identifier les voies de régulation impliquées dans la réponse à l'IR et les réunir au sein d'un réseau plus global.
Nous formulons notre problématique de la façon suivante :

Étant donné :

1. différentes hypothèses de régulation à tester, traduites par une stratégie expérimentale de perturbations génétiques (7 souches de levures avec des combinaisons de caractéristiques génétiques différentes) (voir tableau 3.1),
2. des cinétiques d'expression mesurées après IR pour approximativement tous les gènes de *S. cerevisiae* et dans toutes les conditions définies par la stratégie expérimentale (voir tableau 3.1),
3. des informations biologiques additionnelles, de sources et de natures différentes, extraites de la littérature ou des grandes bases de données biologiques : ontologies fonctionnelles (*Kegg* et *GO*), données d'interactions protéines-protéines, données d'interactions protéines-ADN (données ChIP) et des informations sur la localisation chromosomique des gènes,

comment identifier l'hétérogénéité potentielle des réponses à l'IR et comment mettre en évidence des régulations impliquées dans le contrôle de ces réponses ?

Nous posons 4 grandes hypothèses de travail, dictées par des contraintes liées à la nature et à la structure de nos sources d'information.

1. Nous voulons privilégier les données expérimentales produites au sein du laboratoire et donc partir uniquement des cinétiques d'expression des gènes pour l'identification d'une réponse à l'irradiation. Ce parti pris nous permettra d'éviter de biaiser l'analyse exploratoire par des informations extérieures qui ne sont pas en rapport avec l'IR et qui sont souvent incomplètes.
2. Nous souhaitons effectuer une analyse globale de la réponse transcriptionnelle sans imposer un filtrage des cinétiques d'expression des gènes à l'aide d'un seuil arbitraire. Fixer un tel seuil risque d'éliminer de nombreux gènes sensibles à l'IR mais dont la modulation d'expression ne connaît que de faibles amplitudes. Ce peut être le cas des gènes de beaucoup de FTs qui n'ont besoin que de faibles variations d'expression pour moduler celle de leurs cibles.
3. Le traitement des cinétiques d'expression des gènes doit prendre en compte les dépendances entre les points d'une même cinétique. De plus, nous voulons que l'hé-

térogénéité des dynamiques de réponses à l'IR soit identifiée sur l'ensemble des conditions expérimentales à la fois. Ainsi nous devrions identifier les réponses dont la cohérence résiste aux perturbations génétiques.

4. Nous souhaitons prendre en compte la stratégie de perturbations génétiques pour analyser l'influence des variations génétiques sur la réponse à l'IR et dessiner l'ébauche d'un réseau de régulation.
5. Et enfin, nous voulons intégrer et exploiter les informations biologiques additionnelles pour interpréter et analyser la cohérence biologique des réponses et compléter un éventuel réseau de régulation impliqué dans le contrôle de la réponse à l'IR.

Nous nous retrouvons confrontés à un problème méthodologique lié d'une part à la diversité et à la structure des données en notre possession et certaines contraintes que nous imposons et qui rendent la résolution de notre problématique plus complexe.

Souches	LM18733	LM18734	LM18735	LM6053	FFD1	yIBPC205	LM79
Ploïdie	n	n	n	2n	n	n	n
Mating type	a	α	a(α)	a,α	a	α	α
Mutation	-	-	-	-	*rad*52	*sir*2	*ku*70
Tailles des cinétiques	7	7	7	7	7	7	5
Nb. de gènes analysés	5918	5249	5753	6120	6130	6190	6188

Table 3.1 – *Description des données d'expression et caractéristiques des fonds génétiques utilisés.*

3.3 Confrontation des méthodes existantes à notre problématique

Nous sommes confrontés à la fois à un problème d'analyse de données et à un problème d'inférence de réseaux. Notre problématique est exploratoire et devrait faire appel aux méthodes d'extraction ou de prédiction d'interactions plutôt qu'aux méthodes d'inférence de paramètres ou de structures de modèles (présentées dans la section 2.3, page 54). Les méthodes de prédiction d'interactions (présentées dans la section 2.4.1, page 59) visent à reconstruire un réseau et sont plus difficilement applicable à l'exploration des différentes formes de réponses à l'IR et à leur interprétation.

Les méthodes basées sur la réduction de dimension et l'intégration de données systémiques (présentées dans la section 2.4.2, page 62) semblent particulièrement intéressantes car elles permettent d'identifier des motifs d'expression qui peuvent refléter différentes réponses à un stimulus et permettent aussi d'exploiter des données hétérogènes pour extraire un réseau de régulation de gènes. Les méthodes et les outils d'analyse "clé en main"

(présentées dans les sections 2.4.2.1, 2.4.2.2 et 2.4.2.3 pages 63-65) qui exploitent au maximum 2 types de sources d'information (transcriptome/perturbations génétiques, transcriptome/ontologies fonctionnelles, etc.), ne peuvent exploiter simultanément les trois types d'informations dont nous disposons. De plus, les différents niveaux d'analyse nécessaires à la résolution de notre problématique étant connectés et interdépendants, ne peuvent être résolus par l'application parallèle de différentes méthodes indépendantes.

Parmi les méthodes intégratives plus complexes, les méthodologies basées sur une intégration unifiée d'informations biologiques hétérogènes [282, 283, 284, 285, 286, 287] permettraient d'intégrer toutes les données et informations biologiques en notre possession (voir section 2.4.2.3 page 67). Cependant, l'exploitation "uniforme" des différents types d'informations génomiques et en particulier la non différentiation des données d'expression des autres types de données risquerait de fortement biaiser notre analyse. En effet, les données d'expression sont souvent mesurées dans des conditions expérimentales particulières (analyse de la réponse à un stimulus externe, perturbations génétiques, etc.) et donc peuvent permettre d'extraire une relation de causalité entre le comportement des gènes et les conditions expérimentales dans lesquelles ils sont analysés. Le fait de fusionner l'information transcriptionnelle à d'autres types d'informations biologiques, n'ayant aucun rapport avec les conditions dans lesquelles ont été produites les données d'expression, reviendrait dans notre cas à noyer la relation de causalité qui existe entre l'expression des gènes et l'IR et présente le risque d'inférer des liens de régulation n'ayant rien à voir avec la réponse à l'IR. Ces méthodes semblent mieux correspondre à l'identification de la topologie d'un réseau général de régulation transcriptionnelle au sein d'un organisme qu'à l'analyse d'une réponse cellulaire à un stimulus et à l'identification du réseau de régulation impliqué.

Les méthodes "mixtes" proposent généralement des méthodologies complètes, parfaitement adaptées à la résolution d'une problématique biologique particulière. Cependant, même si elles adoptent aussi des étapes de réduction de dimension et/ou d'intégration de données biologiques hétérogènes, elles postulent quand même une hypothèse sur la structure du réseau à inférer, ce qui les rend difficilement généralisables à d'autres problématiques biologiques que celles qui ont inspiré leur développement. Parmi ces méthodologies, celle proposée par Elati et coll. [295] (voir sa description dans la section 2.5 page 71), permet la prise en compte de l'aspect dynamique des cinétiques d'expression. Cependant, elle va plutôt exploiter les différentes conditions où le transcriptome du système est mesuré pour en tirer les caractéristiques fréquentes et donc robustes du système (co-régulations fréquentes) plutôt que pour en déduire la sensibilité aux perturbations.

Enfin, les méthodologies d'analyse séquentielles (voir section 2.4.2.3 page 68) semblent les plus pertinentes et les plus adaptées à notre problématique. Certaines approches, comme celles de Haverty *et coll.* [292] et de Workman *et coll.* [161] appliquent leurs méthodologies à des problématiques proches de la notre et différencient explicitement les données qui reflètent la réponse à un stimulus (transcriptome ou interactions TFs-gènes) des autres sources de données biologiques. Cependant toutes ces approches restent statiques ou passent par une discrétisation des cinétiques d'expression des gènes. Certaines

de ces approches se basent également sur l'exploitation de perturbations génétiques mais aucune ne propose un cadre formel pour l'exploitation automatique d'une stratégie de perturbations d'un système biologique.

Chapitre 4

Inférence semi-automatique de voies de régulation à partir de données perturbées

4.1 Article I : *XRegPath* : méthodologie d'extraction semi-automatique de voies de régulation à partir de données perturbées

4.1.1 Conception de *XRegPath*

Nous sommes partis des hypothèses de travail présentées au chapitre 3 pour définir une méthodologie qui réponde à notre problématique. Notre approche nommée *XRegPath*, exploite séquentiellement les trois types d'informations à notre disposition en initiant le processus d'inférence par l'analyse des données d'expression. Nous distinguons 5 étapes successives (voir figure 4.1) :

Première étape : identification de groupes de gènes cohérents. Il s'agit d'une étape d'induction (recherche de lois générales et nouvelles à partir de l'observation de faits particuliers), elle consiste à réduire la dimension des variables (gènes) à analyser et à identifier des motifs d'expression caractéristiques de la dynamique de réponse à l'irradiation. Les outils de classification, et en particulier la classification spectrale à base de noyaux, trouvent naturellement ici un bon cadre d'application (voir l'état de l'art des outils de classification et des métriques en annexe, page 222). Dans cette étape, l'aspect dynamique de l'expression des gènes est pris en compte grâce au calcul d'une similarité entre gènes qui se base sur les accroissements et décroissements des profils d'expression et non sur les valeurs d'expression directement. Nous imposons cependant une contrainte liée à la stratégie expérimentale de perturbations génétiques. Nous cherchons à identifier des groupes de gènes co-exprimés dont les profils restent cohérents entre eux même si le système est perturbé. Cela est rendu possible par le développement d'une métrique qui permet de calculer une similarité globale entre gènes sur la base de leurs profils d'expression dans l'ensemble de nos conditions expérimentales.

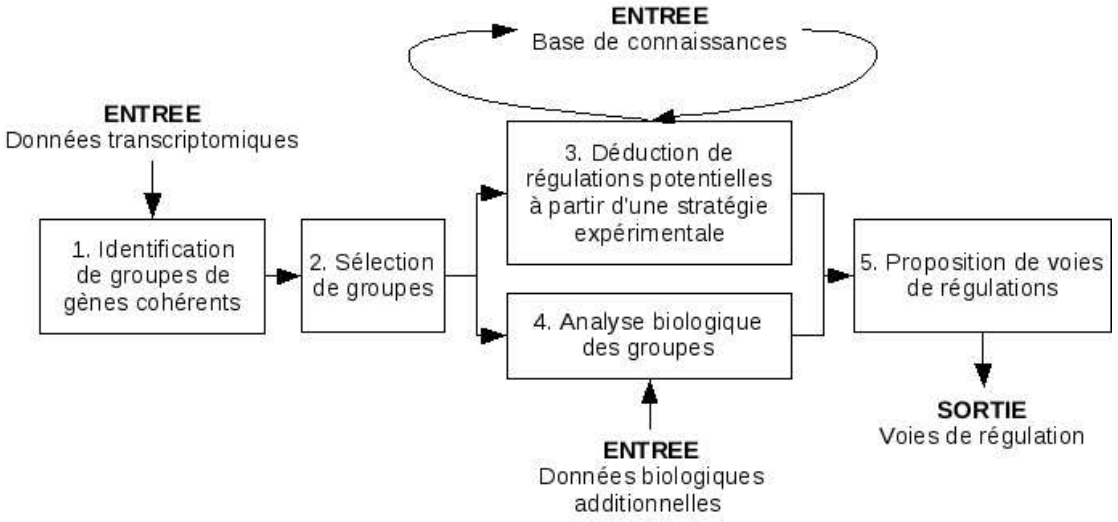


Figure 4.1 – Les différentes étapes de la méthodologie *XRegPath*.

Deuxième étape : sélection et annotation de groupes de gènes d'intérêt. La deuxième étape consiste à simplement sélectionner les groupes de gènes co-exprimés d'intérêt. Nous n'allons conserver pour la suite de l'analyse que les groupes dont le comportement moyen reflète une sensibilité à l'IR dans au moins une des conditions étudiées (ici les différentes souches de levures). Les profils d'expression moyens des groupes sélectionnés sont ensuite simplement annotés, dans les différentes conditions expérimentales, en tant que modulés ou non modulés.

Troisième étape : déduction de règles de régulations potentielles. La troisième étape est une étape de déduction, nous y analysons l'influence des différentes perturbations génétiques sur le profil moyen de chaque groupe de gènes co-exprimés pour en extraire des relations de régulation. Cette étape consiste simplement à déduire si une perturbation génétique a un effet ou non sur le comportement d'un gène ou d'un groupe de gènes. Le formalisme logique utilisé par Hvidsten *et al.* [237] semble particulièrement intéressant dans la mesure où il peut permettre de formaliser et d'automatiser cette étape de déduction. La logique de la stratégie de perturbation est d'abord formalisée en une base de connaissance codée en langage *Prolog*. Celle-ci contient des *faits* ou observations (annotations des réponses moyennes des groupes de gènes à l'IR et les caractéristiques génétiques des souches de levure) et des règles de régulation. Ces règles de régulation voient les différentes types de caractéristiques génétiques des souches de levure comme autant de variables pouvant prendre chacune un nombre fini de valeurs. Les règles peuvent être définies avant même le déroulement des expériences et permettent de décrire de façon formelle les différents effets (régulations simples, additives ou suppréssives) qu'une variable génétique ou une combinaison de ces variables peuvent avoir sur l'expression d'un gène ou d'un groupe de gènes. Un moteur d'inférence logique nous permet ensuite de dé-

duire quelles sont les règles qui sont vérifiées en fonctions des faits observés et d'extraire ainsi des règles de régulation entre les différentes valeurs des variables génétiques et les groupes de gènes co-exprimés. Le processus de déduction est séquentiel dans la mesure où les règles sont évaluées dans l'ordre de leur complexité croissante. Ainsi les règles les plus simples sont évaluées en 1er, le résultat de ces évaluations va ensuite enrichir la base de connaissances avec de nouveaux *faits* qui peuvent servir à leur tour à l'évaluation des règles plus complexes. Ce processus permet ainsi d'évaluer de façon itérative des règles très complexes que les *faits* observés à l'issue des expériences ne pouvaient suffire à vérifier.

Quatrième étape : analyse biologique de groupes de gènes co-exprimés. Dans la quatrième étape, étape d'induction indépendante de la précédente, l'intégration de données biologiques externes permet à la fois d'analyser la nature biologique de la réponse à l'irradiation (ontologies fonctionnelles, données d'interaction protéines-protéines) et d'identifier de nouvelles régulations potentielles (données d'interactions FTs-gènes, positions des gènes sur les chromosomes). Nous intégrons les différentes sources d'informations biologiques additionnelles sous la forme de fichiers d'associations gène-descripteur. Pour chaque groupe de gènes co-exprimés, chaque association entre un gène et un descripteur est évaluée statistiquement selon la loi hypergéométrique et n'est sélectionnée que si sa sur-représentation dans le groupe de gène est significative. A ce niveau de l'analyse, les différentes données biologiques additionnelles peuvent être intégrées de façon uniforme sans hiérarchisation particulière. Pour cela, les outils de type *biclustering* semblent assez intéressants car ils permettent d'analyser la cohérence biologique des groupes de gènes co-exprimés du point de vue des différents types d'informations biologiques additionnels en y recherchant l'existence de sous-ensembles de gènes associés à des sous-ensembles de descripteurs biologiques. Cette étape a été réalisée en collaboration avec Farida Zehraoui qui a implémenté un algorithme de *biclustering* spectral qui permet l'identification de bi-classes à partir d'une matrice binaire d'association gènes-descripteurs binaire. de cette étape A partir de ces sous-regroupements nous pouvons inférer de nouveaux liens de régulation entre des FTs et certains gènes co-exprimés. Cependant, ces liens restent hypothétiques car rien ne prouve leur implication dans la réponse à l'IR.

Cinquième étape : proposition de voies de régulation. Une fois toutes ces étapes réalisées, leur synthèse globale à la cinquième étape permet d'obtenir une vue intégrée de la réponse à l'IR et permet de construire un réseau de régulation impliqué dans le contrôle de cette réponse.

Dans cette approche, nous séparons clairement les étapes d'induction (identification de motifs de co-expression et exploitation d'informations biologiques additionnelles) de l'étape de déduction (exploitation de la stratégie de perturbations génétiques). La méthodologie *XRegPath* est totalement exploratoire et ne nécessite au départ aucune connaissance biologique pour l'identification de réponses transcriptionnelles à l'IR. Notre approche nous permet d'analyser la réponse transcriptionnelle à l'IR à l'échelle du génome sans sélection arbitraire des données d'expression et sans discrétisation des cinétiques. La

dynamique des cinétiques est à la fois prise en compte dans l'identification des réponses à l'IR est dans l'interprétation de ces réponses. Et enfin nous avons formalisé et automatisé la déduction de règles de régulations à partir des différents motifs de réponses à l'IR et des différentes caractéristiques des perturbations génétiques du système.
Le développement, l'implémentation et l'application de la méthodologie *XRegPath* sont présentés dans la section suivante, avec quelques résultats biologiques, sous la forme d'un article au format standard du journal *BMC Bioinformatics*. Les résultats complémentaires présenteront l'implémentation de *XRegPath* sous la forme de deux logiciels complémentaires. La totalité des résultats de l'analyse de la réponse à l'IR, leur interprétation biologique et la validation expérimentale d'une partie d'entre eux sont présentés sous la forme d'un deuxième article à la section 4.2 page 141.

4.1.2 *XRegPath* : semi-automated extraction of regulatory pathways using genetic perturbation data. N. Touleimat, F. Zehraoui , M. Dutreix and F. d'Alché-Buc

XRegPath: semi-automated extraction of regulatory pathways using genetic perturbation data

Nizar Touleimat*[1,2], Farida Zehraoui[1] , Marie Dutreix[2] and Florence d'Alché-Buc*[1]

[1]Informatique Biologie Intégrative et Systèmes Complexes (IBISC) 2873 FRE CNRS - Université d'Evry-Val d'Essonne, Tour Evry 2, 523 place des terrasses de l'Agora, 91000 Evry - FRANCE

[2]Recombination, Repair and Cancer, Translational department, Institut Curie, Centre Universitaire - Batiment 110-112, F-91405 Orsay cedex - FRANCE

Email: Nizar Touleimat*- nizar.touleimat@ibisc.fr; Farida Zehraoui - zehraoui@ibisc.fr; Marie Dutreix - marie.dutreix@curie.u-psud.fr; Florence d'Alché-Buc*- florence.dalche@ibisc.fr;

*Corresponding author

Abstract

Background: Inference of gene regulatory pathways from large scale gene expression data is still a bottleneck when no prior knowledge is available. In order to identify the transcriptional response of yeast to γ-radiation and its regulatory mechanisms, we consider the task of discovering gene regulatory pathways from gene expression kinetics measured accross several perturbations.

Results: Taking into account that a scientific discovery process is both a matter of induction and deduction, we conceived XRegPath, a general methodology based on automated deduction and statistical inference that helps the biologist to extract gene regulatory pathways involved in the cellular response of a given organism to some stress signal. XRegPath mines large datasets, extracts and filters information, deduces potential regulators, confronts different sources of data and finally gathers various pieces of evidence about regulatory processes.

Conclusions: We applied this methodology to the analysis of the yeast transcriptional response to γ-radiation. We extracted typical responses as co-expressed gene groups with typical behaviour. We showed the interest of this approach by focusing on two gene clusters, each of them gathering consistent gene expression profiles across multiple strains. Finally, we generated new hypotheses about yeast cellular response to radiation and its regulation that have been experimentally confirmed afterwards.

Background

The cellular response of an organism to a given input signal relies mainly on the dynamic modulation of a proteins pool. Currently, identifying and understanding such a response requires both to monitor gene expression modulation and to reveal the cascade of regulatory interactions involved in gene expression modulations through time in the cell. In this work, we consider the task of gene regulatory pathways extraction as a discovery process. Our goal is to provide the biologist with a tool that generates hypotheses that would be validated through new specific biological expriments. To achieve this goal, we observe that a classic way to unravel a gene regulatory network is to apply perturbations to the biological system and measure how it behaves when responding to the input signal. As far as a small number of genes are concerned, the biologist is able to deduce the influence of potential regulators. However when large scale data are available and perturbations are complex, e.g. not a single gene knock-out but a combination of various genetic perturbations, the biologist needs automated reasoning tools to propagate deductions without errors. For this purpose, we suggest to use first-order logic and Prolog, a language that allows to encode data and knowledge into facts and logical rules and among all, performs automatically proofs. Moreover, when faced to large scale gene expression kinetics without prior knowledge about interactions, the biologist also needs induction tools to extract patterns of co-expression and potential co-regulation. For this task, data-mining methods such as clustering can be of a great help, allowing to reduce dimension and thus providing inputs to the deduction process. When potentially co-regulated genes are identified, Prolog programs can be used to deduce potential complex regulations. In parallel, it also becomes possible to analyze potential co-regulees to the light of additional data such as functional information, association to transcription factors and position of genes on chromosomes. This is again a matter of statistical induction and biclustering tools seem very appropriate here. Finally, the results of the two approaches can be gathered into a set of hypotheses that highlight potential regulatory pathways involved in the response to the input signal. The originality of our approach lies both in the logical exploitation of perturbed data and in the sequential processing of data and knowledge.

Wagner in [1] first introduced the systematic exploitation of perturbation data for gene regulatory network reconstruction using the framework of graph theory. First order logic has already been used in other

context such as gene function prediction [2,3] and to a lesser extent for regulatory rules extraction from one kind of information in [4]. However, to our knowledge our work seems to be the first one that shows how automated logical reasoning can allow to exploit the logics underlying the perturbation experiments. Combining heterogeneous biological information with gene expression data in a single framework has already been done for instance in [5] and [6]. In contrast, our approach is entirely driven by expression data analysis without bias from additional information. This allows to clearly highlight patterns of co-expression, reduce dimension and ensure that the whole inference process is based on experimental observations and not on indirect related sources of information. Other approaches integrate heterogeneous genomic information in a sequential way in order to extract regulatory pathways [7–9]. However, none of them takes into account the dynamical nature of gene expression neither use a formalized deduction process to exploit genetic perturbation strategies.

In this paper, we illustrate our methodology through a study of the transcriptional responses of the yeast *S. cerevisiae* to irradiation. We show how it is possible to extract pieces of regulatory networks involved in the control of these responses. By means of this sequential strategy we characterized a set of gamma-rays yeast responses. Results confirm the relevance of our approach by retrieving a response of a large set of genes that has previously been identified in [10]. Moreover, we provide evidence for the hypothesis that this set of genes contains in fact two subsets with two distinct transcriptional regulatory mechanisms. The paper is organized as follows. In the results section, we first describe our methodology, called XRegPath. Then we detail our tests on the transcriptional response of yeast to irradiation. We close this section by a few words about the implementation. In the discussion section, we compare our methodology to several approaches that either integrate heterogeneous data in a unified framework or proceed to data-analysis through a sequential strategy. Finally, a conclusion and some details about the methods are provided.

Results

Methodology description

XRegPath is decomposed in five steps and exploits three kinds of information in an original way (Figure 1):

- kinetics of gene expression measured in various genetic systems: we assume that we have series of gene expression kinetics with potentially different sizes, each series corresponding to a genetic background. Each gene is then described by these kinetics of expression, one kinetic per genetic background. A genetic background corresponds to a specific genetic system. Two genetic backgrounds could differ in their "natural" genetic features (different alleles for a same gene, ploidy

variations...) or by artificial genetic perturbations (gene knock-out, artificial gene expression...). Note that for some cases the notion of different genetic backgrounds could be extended to differences between experimental conditions (differences between applyed stimuli, between growth conditions...).

- logical rules that describe the logic underlying the perturbations.
- additional biological data that come from various databases or ontologies.

All these information are not used at the same level in the methodology. We start from transcriptomic data obtained in the laboratory to identify groups of genes that remain co-expressed under all the perturbations (step 1). At this stage no prior knowledge is used. According to the step 1, we select clusters of genes whose profile are modulated at least in one strain (step 2). In step 3, we use automated reasoning to deduce potential regulations from the logic of perturbations among selected clusters of genes. In step 4, we analyze the clusters selected at step 2, by highlighting consistent subsets according to additional biological features. In step 5, we gather information inferred from the previous steps to generate consistent hypotheses about regulatory pathways involved in the response to stimulus.

Identification of coherent groups of genes

We define a stimulus response as a set of co-expressed genes whose expressions are significantly modulated by the stimulus and which are dependent of the same regulation mechanisms. We distinguish two notions: co-expression and potential co-regulation. Given an experimental condition, co-expressed genes present similar kinetics of expression. Co-regulated genes are genes that are under the control of the same regulatory mechanisms. We expect that these co-regulated genes show a coherent behaviour which means that they are co-expressed whatever the genetic background in which the gene expressions are measured. Thus we search for coherent groups of genes to get candidates for co-regulation. For this purpose, we use spectral clustering [11], a kernel-based clustering algorithm [12]. Spectral methods are based on normalized cuts of weighted graphs. These methods are attractive because they show much more stability that simple k-means.

To apply spectral clustering, one must define an appropriate similarity. In this paper, we have chosen a similarity which has also the properties of a kernel (it corresponds to some scalar product in some Hilbert space). There exists a large family of kernels already defined for various kinds of data. However temporal profiles measured for a given genetic background are very short so the various kernels that have been defined for time series are not appropriate. We thus proposed a kernel s that measures into what extent

the compared profiles share approximately the same slopes between time points. Given two genes G and G' and their respective temporal profiles, $\mathbf{x} = (x_1, ..., x_T)$ and $\mathbf{x'} = (x'_1, ..., x'_T)$, the proposed kernel s is defined as a Gaussian kernel based on the squared distance between the vectors $\mathbf{y}$ and $\mathbf{y'}$ of their respective instantaneous derivatives. We thus define directly:

$$s_{coexp}(G, G') = exp(-\gamma ||\mathbf{y} - \mathbf{y'}||^2) \tag{1}$$

where for any $\mathbf{x}$, $\mathbf{y} = (x(2) - x(1), ..., x(T) - x(T-1))$. This kernel takes into account the similarity of shapes (no matter of amplitude) by using derivative of gene expression kinetics and is translation-invariant. Using such a kernel as a similarity measure between genes for each genetic background, we define a kernel s_{coh} between two genes G and G' as the following normalized linear combination of the kernels computed for each genetic background:

$$s_{coh}(G, G') = \sum_{l=1}^{p} \tilde{\alpha}_l \times s^l_{coexp}(G, G') \tag{2}$$

where p is the number of genetic backgrounds, $\tilde{\alpha}_l = \frac{\alpha_l}{\sum_{l=1}^{p} \alpha_l}$ corresponds to the normalization of α_l, the length of the expression kinetics measured into the genetic background l and s^l_{copexp} is the co-expression kernel defined for genetic background l. The linear convex combination ensures that the combination has still the properties of a kernel and thus of a similarity. This kernel allows to measure the similarity between all the time-series.

A method of model selection is needed to choose the values of γ and the number of clusters K. The value of γ can be chosen for each of the genetic background using a visualization method based on histograms of kernel values (see Kernel parameter selection in Methods). Once γ is fixed, the number K of clusters can be determined by estimating the stability of the clustering algorithm, e.g. its robustness against subsampling of the data. We have developed a specific procedure of model selection inspired by the work of Ben-Hur *et al.* [13] to measure the similarity between the partitions obtained from different values of K and different subsamples (see clustering stability and partition size selection in Methods). Instead of computing pairwise similarities between partitions obtained from subsampling data for a fixed size of partition, we compute these similarities between clustering obtained with different sizes of partitions. This allowed us to monitor the data structure preservation from a partition of size K to a partition of size $K+1$ and to determine the optimal K value. Further details of this approach are given in the Methods section.

Cluster selection

Clustering provides a codebook of representative gene expression profiles (average profiles) reducing the dimension from thousands of genes to K clusters. The selection step consists in detecting the clusters whose average gene expression profile reveals expression variations after irradiation (IR). Flat profiles will not be considered for the next steps of the methodology. To ease this selection step, a modulation status is defined and associated to each of the genetic background kinetics inside the representative profile. In our study this annotation was performed by the biologist. We choose to annotate the cluster behaviour with only two values: *modulated* or *unmodulated*. The attribution of the modulation status to a representative profile depends on many criteria (listed in Cluster profiles annotation in the Methods section), some of which are based on the experimental design and others are based on the reliability of an observed kinetic profile. Once the clusters are annotated, the ones with at least one modulated profile are selected and provided as inputs in parallel steps 3 and 4. The discretization we used for cluster annotation is sufficient to extract the clusters of genes whose expression are modified by perturbations. However it is possible to provide a larger dictionary of values if required (e.g. that would take into account the temporal evolution of the profile: increase, decrease, etc).

Deduction of potential regulations using experimental design

In this step, we exploit the logic of the experimental design and analyze the sensitivity of the selected clusters to the genetic backgrounds and deduce the potential regulatory role of the genetic variations. Indeed, gene clusters whose expression level varies according to some genetic backgrounds can be identified as being potentially regulated by this genetic variation. Before experimentation, we can state logical rules that express what kind of conclusion can be implied by given observations. For instance a gene cluster C whose expression profile measured in an non gene G mutated strain differs from the profile measured in a strain with gene G mutated can be suspected to be regulated by gene G. Encoding into a logical program all these rules allows to mimic the kind of reasoning the biologists are familiar with. Inference engines can be used to process all the knowledge base and deduce automatically potential regulations. Moreover, the approach limits the error risk because one has only to pay attention to the rules definition. In order to implement this approach, Prolog, a language devoted to automated deduction in first order logic, has been used [14]. To detect clusters whose expression is affected by one or several factors, we build a knowledge basis (here a Prolog program) that includes:

- facts concerning the average gene expression profiles of clusters and information concerning genetic

backgrounds

- logical rules that state what implies the presence or absence of a potential regulation. Rules may involve one or more factors.

The facts describe the nature of the average profiles in each genetic background and the properties of each genetic background. In this study, the most important facts concern the description of expression profiles and genetic backgrounds.

- *background*(S, P, M, MS) which is true when the genetic background S corresponds to the three different genetic variables P (for ploidy), M (for mating type) and MS (for mutation status). The aim of this Prolog fact is to describe a genetic background in terms of its genetic variables.

- *expression*(C, S, E) which is true if a gene cluster C has expression E when measured in genetic background S.

We use Prolog rules (implication rules with a single conclusion) to describe what kind of observation can imply the presence or the absence of regulation. Assuming we test three genetic variables, we get different kinds of potential regulation: potential regulation by a single factor, by two factors or by three factors. Let us begin with single factor regulations. We first notice that when no appropriate data can be observed, conclusion about presence or absence of regulation cannot be drawn. For sake of clarity we define two predicates regulation and noregulation which are not exactly the contrary of each other. Predicate *regulation(C,*$X1$*,*$X2$*,X)* is true if we can observe that the gene expression of cluster C varies when genetic variable X differs according to the values $X1$ and $X2$ while all other characters are kept constant. In other words, predicate *regulation* is true when only the single variable X modulates the gene expression level of the cluster in response to stimulus. Predicate *noregulation(C,*$X1$*,*$X2$*,X)* is true if we can observe that the gene expression of cluster C remains the same when genetic variable X differs according to the values $X1$ and $X2$ while all other characters are kept constant. Given a cluster C and a genetic variable X, if both regulation and noregulation are false, then no conclusion can be derived about a potential presence or absence of regulation from X. Let us now detail the case of the regulation predicate: we need three Prolog clauses to encode the potential regulations according to the three kinds of genetic variables or backgrounds. For instance, here is the Prolog clause used for the ploidy character:

$regulation(C, P1, P2, ploidy)$:

$-background(S1, P1, M, MS), background(S2, P2, M, MS), expression(C, S1, E1), expression(C, S2, E2),$

$differs(E1, E2)$.

This Prolog clause can be read as: if we can observe a strain $S1$ characterized by $P1$, M, MS and a strain $S2$ characterized by a different ploidy value $P2$ and the same M and MS values and if the expressions of C in $S1$ and $S2$ differ, then there is a potential regulation of the ploidy variation $P1/P2$ on genes belonging to C. Similar rules can be defined for other genetic variables such as the mating type and the mutation status in the case of yeast (see Additional informations 5). When conclusions are drawn using Prolog, they are considered as new facts for the Prolog interpreter and as new hypotheses for the biologists. Our knowledge basis is first enriched with deductions involving a single type of genetic variations. In order to efficiently extract combined regulation hypotheses, we deduce presence or absence of cooperative regulations (rules that deal with more than one variable) only after the deduction process for single regulations. Using the same principles as previously, predicates *dependent2regulation* and *suppressive2regulation* are defined to be true if the co-occurrence of two variables is identified as responsible for either the stimulus response regulation or the absence of the response regulation observed with each variable alone. We have written rules that involve all possible combinations of variables modifications (see Additional informations 5 for the complete set of rules we applied in our deduction process). Once the knowledge base is built, Prolog can be used to automatically deduce new facts from experimental results. For instance, Prolog provides automatically responses to questions like: is cluster c potentially regulated by the variable X? In this case it automatically finds values of $X1$ and $X2$ for which the predicate *regulation(c,X1,X2,X)* is true.

Biological mining of clusters

Step 4 is dedicated to the biological analysis of the clusters selected from step 1 and step 2. Given a gene cluster, we would like to know if the genes share additional biological features such as Gene Ontology (GO) annotations [15], genes location on chromosomes, known physical interactions between proteins or relations to transcription factors (TFs). Note that this list is not restrictive and can be extended to any source of systemic biological data. This biological mining is performed by using biclustering for each genes cluster. Biclustering introduced in [16] for gene expression data measured in several conditions consists in searching for genes that have the same level of expression in the same subset of conditions. Biclusters in this case can overlap each other which may be difficult to interpret. Biclustering algorithms have also been developed and successfully applied to heterogeneous biological data for instance in [5,17]. In this step, we are interested in finding coherent groups of genes inside a cluster that share the same biological features. In order to extract such biclusters, we have chosen to use spectral biclustering as introduced in [18] that

produces a checkboard structure of the matrix to be analyzed. Like spectral clustering, spectral biclustering is also based on graph partitioning except that in this case, the graph is bipartite [19, 20]. A detailed reminder about this method can be found in the Method section. Two parameters define a spectral biclustering: K_1, the size of the partition among genes and K_2 the size of the partition among biological features. For sake of uniformity, all the results of biclustering can be written in terms of Prolog facts. In our methodology, the input of the biclustering algorithm is a matrix A whose rows correspond to genes and columns correspond to biological descriptors. However some biological descriptors may not be relevant to describe the genes of a given cluster. This is why we generate statistically-based rank scores that evaluate the over-representation of each candidate biological descriptor in the studied cluster compared to the whole set of genes analyzed. Rank scores are obtained using the hypergeometric distribution and a p-value is attributed to each gene descriptor. We select gene descriptors that have a p-value equal or lower than 0.05 as commonly used in literature [21]. Once a set of biological descriptors is selected, we build the corresponding Boolean features to describe the genes of a given cluster. As a result of biclustering, genes clusters partitioned into biclusters can be more finely described.

Proposal of regulatory pathways

From previous steps, we infer hypotheses about regulatory processes. All these hypotheses are gathered to propose a regulatory scheme for a cluster of interest. Note that biclusters inherit the regulatory mechanisms inferred for the clusters they belong to. In steps 3 and 4 predicates are not the same and do not deal with the same information content but they contribute to the enrichment of the regulatory scheme. In this study, we leave to the biologist the task to check consistency between the different gathered facts. Finally, steps 3 and 4 provide putative targets for new experiments to test new regulatory hypotheses.

Testing

The XRegPath approach was applied on yeast expression kinetic data in response to IR in 7 different yeast strains (4 datasets published in [10] and 3 new datasets produced for this work). The 7 yeast strains differ by predefined genetic variations (ploidy, mating type and 3 targeted mutations) (see Table 1).

Identification of coherent groups of genes

We applied the kernel based spectral clustering to a set of 4732 genes that have a complete expression kinetic in each one of the 7 yeast strains. We computed similarities between genes using the linear

combination of Gaussian kernels (see formula 2) with a γ value of 0.4 (see kernel parameter selection in the Methods section). Using the spectral *K-means* algorithm, we computed all partitions of size K varying from 2 to 36. We define an interval of partition sizes where data structure stay preserved (see Clustering stability and partition size selection in Method section). The choice of the optimal partition size for our data was done in this interval, jointly with the biologists as described in the Method section. For our data, the partition into 22 clusters gives the most satisfactory result representing all the responses to IR. Each cluster is represented by a combination of mean expression profiles (see figure 2 for clusters C2 and C15 and Additional information 4 for the other clusters) that will be further used for cluster annotation and selection.

Selection of clusters

All clusters have been annotated as $'U'$ for "unmodulated" or $'M'$ for "modulated" according to their profile in each strain (see Table 2). Thirteen clusters do not show any modulated profiles and were considered as insensitive to IR. Nine clusters show a modulated profile in at least one strain and are therefore considered as radiosensitive. For example, clusters C2 and C15 are respectively annotated as $MMUUMUU$ and $MMMUMUU$. Using three yeast strains (strains 18733, 18734 and 6053), Mercier *and al.* have already identified a set of genes induced by IR with some characteristic regulatory dependencies [10]. These genes are distributed for the most part into our clusters C2 and C15. Integrating the expression data produced in the 7 yeast strains in a same framework allowed us to split Mercier *and al.*'s radiosensitive genes into two different radiosensitive clusters. The mean expression profiles of these two clusters are similar through the set of yeast strains except in the strain 18735 where cluster C2 genes are insensitive to IR while cluster C15 gene expressions show a marked induction in response to IR. This result shows that our approach is able to integrate in a same framework expression data measured in a large number of different genetic backgrounds and that our method for the identification of coherent groups of genes is able to detect differences in the modulation of expression in a single genetic background and uses it to identify gene clusters with specific behaviours. In what follow, we illustrate the two last steps of our methodology with clusters C2 and C15.

Deduction of potential regulations using experimental design

We built a knowledge base as a Prolog program (see Additional informations 5) according to the methodology previously described. Then we asked the Prolog Interpreter to make deductions about the

genetic variables that would influence the response of clusters to IR. We applied this procedure for all the regulatory predicates for clusters C2 and C15 (see Table 4). Clusters C2 and C15 IR response annotations are similar except in strain 18735 (see table 2). From our Prolog program responses we deduced that cluster C2 is sensitive to mating type variation aα/α and cluster C15 is sensitive to ploidy variations $n/2n$. Moreover, the results indicate that in cluster 2, sir2 and ku70 are linked to mating type variation aα/α, and in cluster C15, sir2 and ku70 are linked to ploidy variation n/2n as shown by the co-occurrence of the two kinds of variables. In our study the facts were sufficient to test the influence of each genetic variable independently. However, the set of facts was insufficient to allow the testing of the influence of all the associations of variables for regulatory predicates that deal with more than one genetic variable. The experimental design does not generate a sufficient or an adequat set of facts to test all combination of variables associations. The deduction process can be used as a feedback to check the experimental design in order to highlight missing information. We could then propose new experiments designed to test regulatory predicates for a given association of genetic variables. For example, to test if there is a regulatory dependency between the mating type variation α/(a,α) and the rad52 mutation that affect cluster C15 expression (*dependent2regulation* or *nodependent2regulation*) we propose to the biologist to perform the same gene expression analysis in a yeast strain with the following genetic background: ploidy = n, mating type = (a,α) and a rad52 mutation. We are also able to highlight inconsistency in deduction results. If a genetic variation value or a combination of values appears to verify both a predicate and its negation that means that there is an inconsistency in either cluster annotation either expression results themselves. Beyond this feedback loop, our program could be used just after an experimental design and before any result generation to test if a set of genetic backgrounds is accurate enough (sufficient and not redundant) to test all the biologist's regulatory hypothesis. Thus, XRegPath appears more complete than a simple data mining tool: it helps the biologist for the design of experiments, it extracts knowledge from experimental results, it tests the consistency of the results and it allows a feedback to the experiments in order to complete or enrich a set of deduced results.

Biological mining of clusters

In parallel to automated deduction, clusters are analyzed for their content in biological information using five kinds of additional biological data: functional annotations, ChIP data, protein complexes information and two different annotations for gene chromosome positions (see Heterogeneous data integration in the Methods section). Note that gene positions on chromosomes information has been added to test the

hypothesis that response to IR might be controlled by telomere silencing. This exemplifies the fact that any source of information can be added at this step as far as it involves a large set of genes. Using the statistical test described in our methodology we selected the over-represented descriptors in each cluster in regard to the whole set of genes and look for biological relationships between these different biological information. We found that 125 genes among the 212 genes of cluster C2 could be described by at least one of the 14 selected biological descriptors: 5 chromosomal arms, subtelomeric position, 2 TFs, 3 biological pathways, 1 GO biological process and 2 protein-protein complexes. For cluster C15, 146 genes among the 198 genes could be described by at least one of the 74 selected descriptors: 1 chromosomal arm, 4 TFs, 7 biological pathways, 31 GO biological processes, 13 GO cellular components and 18 protein-protein complexes. The kind of biological information involved in the description of gene co-expression differs between the two clusters. Nearly half (42.85%) of cluster C2 descriptors belongs to chromosomal position categories (chromosome arms and subtelomeric position) while the major part (68.91%) of cluster C15 descriptors belongs to functional categories (cellular components, biological processes and biological pathways). Using the biclustering algorithm one can organize the genes within a cluster with respect to the biological descriptors. The biclustering patterns of cluster C2 and C15 are very different (see figures 3A and 3B): cluster C2 shows no clear and compact biclusters, genes are reorganized according to chromosome positions (chromosome arms and subtelomeric position). As we have already noticed after the statistical analysis, in this cluster, biological pathways and processes plays a very little role in pattern organization. In contrast, biclustering of cluster C15 shows a specialization of this cluster into two related processes: rRNA metabolism and ribosome biogenesis. Ribosomes are the workhorses of protein biosynthesis, the process of translating messenger RNAs (mRNAs) into protein. Eukaryotic ribosomes are composed of 80 ribosomal proteins and 4 rRNA species. These two kinds of ribosomes components have to be produced and processed before being assembled as preribosomes that are matured to be functional. We identify in cluster C15, five different but related biclusters all involved in several steps of the ribosome biogenesis. The functional analysis of the biclusters is confirmed by the association of two independent information, functional ontologies and protein complexes: in bicluster $b15_{01}$ (70 genes) we notice the presence of 7 interacting proteins organized into a complex (complex 11) and, in bicluster $b15_{02}$, 15 of the 19 genes products are shared by 2 protein complexes (complexes 7 and 9). We also notice the association of functional and regulatory information in 2 biclusters: bicluster $b15_{01}$ seems to be preferentially linked to the TF ABF1 and, most of the bicluster $b15_{03}$ genes are associated to the TFs FHL1, RAP1 and SFP1. In order to make the integration step easier, we wrote the biological properties highlighted by biclustering as

Prolog facts. For this purpose, we defined the predicate *potentialtarget(C,R)* that is true if the group of genes C is associated to regulatory descriptor R. That means that R could be an experimental target to study for a potential regulatory effect on C. The new facts inferred from biological mining for C15 and C2 are formalized as follows: $potentialtarget(b15_{01}, abf1)$, $potentialtarget(b15_{03}, fhl1)$, $potentialtarget(b15_{03}, rap1)$, $potentialtarget(b15_{03}, sfp1)$, $potentialtarget(c2, chrpos)$.
Where $b15_{01}$ and $b15_{03}$ are the biclusters 1 and 3 obtained in cluster C15 and $c2$ represents cluster C2. The regulatory descriptors $abf1$, $fhl1$, $rap1$ and $sfp1$ are the 4 TF's described above, and *chrpos* represents a gene regulation dependent of chromosome positions.

Proposal of regulatory pathways

Focusing on C2 and C15 clusters, we identified 2 different induced responses to IR : C15 genes are involved in the biogenesis of ribosomes and C2 genes do not seem to be involved in a particular biological process. We extracted for each cluster regulatory dependencies that enrich our knowledge basis and we proposed new regulatory hypotheses to be tested by further experiments. According to their regulatory schemes (see Figure 4), C2 and C15 responses to IR seem to be under the dependency of different regulatory mechanisms: C2 response regulation seems to depend upon gene chromosome location while C15 response seems to be regulated by specific TFs. However C2 and C15 responses to IR are both suppressed in *sir*2 and *ku*70 mutants (respectively LM79 and YiBPC205 strains), known to be perturbed in chromatin structure organization and in gene silencing at telomeres and some other loci. It is also well known that *Rap*1, one of the three TFs we proposed as potential regulators of C15 response to IR, binds telomere sequences and plays a role in telomeric position effect (silencing) and telomere structure. A loss of telomeric gene silencing had been shown to be concomitant with delocalization of Sir2, Ku70 and Rap1 proteins from telomeric DNA sequences following DNA damage induced by various agents [22, 23]. Mercier *et al.* [10] proposed that IR disturbs the silencing chromatin not only at telomeric regions, leading to transient expression of most of the genes under its control but also all over the chromosomes. We propose then a possible model where the responses regulations of the C2 and C15 are linked: chromatin structure modifications will have a direct effect on C2 genes transcription (that comprises a significant number of subtelomeric genes) and will have an indirect effect through the released Rap1 protein that could recruit Fhl1 and Sfp1 proteins and act as TFs to induce the transcription of C15 genes. It is already known that Sfp1 acts at ribosomal proteins promoter via Fhl1, and several experimental data indicate that Fhl1 may activate ribosomal proteins transcription by switching Rap1 between repression and activation modes [24].

Many biological questions arise from this hypothetical regulatory model: is the induction of C2 genes just a side effect of the chromatin structure modification induced by IR or is there a functional purpose to C2 genes activation? An induction of ribosome biogenesis while the cell cycle is blocked will result in proteins accumulation. Actually we do observe a significant increase of the size of irradiated cells. A possibility proposed by the biologists is that damaged ribosomes have to be renewed before cell growth restart.

Implementation

XRegPath has been implemented in Java and Matlab. The program dedicated to the analysis of potential regulators of gene clusters has been implemented in swi-Prolog.

Discussion

Results discussion

Parts of our data have already been used in a previous analysis [10] to characterize a specific cellular response to IR. Mercier *et. al* used the haploid strains 18733 and 18734, as well as the diploid strain 6053. First, they identified a set of genes that show a differential expression after IR in the three yeast strains. Those genes are defined as genes whose expression varied by twofold before and after IR in at least two independent measurements. When comparing the results in the different strains, they found that 471 genes displayed changes in the haploid strains only (HS-IR genes), 278 of these HS-IR genes were induced and 193 were repressed. The majority of the induced HS-IR genes belong to C2 and C15. We also retrieve a significant representation (p-values < 0.001) of the two kinds of HS-IR genes (induced and repressed) in C12, a cluster we annotated as not affected by IR. A precise analysis of the expression profiles contained in this cluster showed a wide heterogeneity of fluctuant profiles in the three strains studied in Mercier *et. al*. It seems that we identified in C12 a group of "false positives" genes that are very fluctuant and for whose an expression semantic is hard to attribute. Even if such genes show significant expression modulations intensities after IR, we cannot be confident with their uncontinuous fluctuations and cannot interpret the expression modulation as an IR response nor as experimental artefact. Therefore, we show that the XRegPath methodology is consistent with previous results obtained in Mercier *et. al* but it provides a more precise and reliable characterization of co-expressed genes using time-course profiles. Our results analysis reveals that one IR response corresponds to the induction of the transcription of genes involved in ribosome biogenesis. This result was quite surprising. Gasch *et. al* [25] analyzed the response kinetics of yeast to various perturbations including γ-rays in conditions close to those of ours experiences. They found

that, in one haploid yeast strain, genes involved in ribosomes biogenesis where globally repressed after IR. Our discovery contradicts Gasch *et. al*'s conclusions. To confirm our observations we used quantitative PCR technique to measure the transcriptional activity of several genes involved in the different steps of ribosome biogenesis in two different haploid yeast strains after IR. We found that the expression of all the tested genes was induced after IR (data not shown, manuscript in preparation). We also demonstrated that rRNA that were not represented on the microarrays are actually induced by IR. All these results confirm that ribosome biogenesis is induced after IR.

Comparison with other methods

There exist other global approaches that mine heterogeneous sources of information in order to extract regulatory networks. They can be classified into two families of approaches. In the first one, heterogeneous data are integrated in a unified framework in order to extract regulatory networks. The most common framework is based on biclustering strategies. In [5] Tanay *et. al* propose an integrative modelling of genomic data as a weighted bipartite graph and, using a biclustering approach, they infer regulatory modules and build a hierarchical and modular organization of the global Yeast system. In addition to gene expression data, protein interactions information and TFs binding, they include growth phenotype data in their study.

In [6], Reiss *et. al* propose to detect putative co-regulated genes groupings with a biclustering approach that exploits many kinds of genomic data: gene expression, upstream regulatory sequences and different kinds of association data. Each bicluster is modelled via a Markov chain process in which the bicluster is iteratively optimized. The originality of their approach is based on its flexibility, actually it is possible to stress one kind of data during the iterative building and refinement of the biclusters. Their strategy allows to take into account the quality of the data type or the existence of prior knowledge, and it could be possible for instance to seed a cluster only with expression data and then to iteratively refine or enrich it with the other sources of information.

The second family of approaches is characterized by a sequential processing of the data with at least two steps. In [7] Ideker *et. al* propose to build and refine a model of cellular pathway by using a sequential methodology that exploits three kinds of information: static gene expression measured in different perturbed conditions, quantitative proteomics data and known physical interactions (protein-protein and protein-DNA interactions). Their methodology is based on four successive steps: *a priori* definition of a specific pathway interaction model, selection and clustering of differentially expressed genes, integration of

mRNA variations, protein variations and physical interactions, and at least the generation of new hypotheses. This work is characterized by the following features: first, it exploits a strategy of perturbation in order to refine the topology of the model through gene expression data and to identify clusters of genes with coherent behaviours. Moreover, an important bringing-in of this work is the integration of 3 kinds of observations quantifying different levels of cellular responses (transcriptome, proteome and physical interactions) in order to extract direct and indirect regulations at the transcriptional level and to identify post-translational regulations.

In [8], Workman *et al.* use a six steps methodology that exploits four kinds of biological information in order to retrieve a network involved in the regulation of the yeast response to DNA damage. In this work it is interesting to notice the use of a clearly predefined genetic perturbations strategy in order to test regulatory hypothesis. The comparison of interactions extracted from ChIP data with interactions extracted from gene expressions produces a rich and fine regulatory network. Let us notice that in the works of Workman *et al.* and Ideker *et. al* simple perturbations (single gene knock-out) on the biological systems are systematically and logically used, but the deduction process is not formalized into first order logic. It is important to say that in our case we associate various perturbations in a same genetic background, it is worth to encode the logic of the perturbation strategy and the description of observations with first-order logic, enabling to automate the deduction process. This allows both to deal with a possibly high number of variables and to extract the influence of single or combined perturbations on the cell responses to the stimulus. Moreover, the objectives of Ideker *et. al* are quite different: their method is supposed to work on a reduced set of genes, focusing on a prior model of regulation while our goal is to start from the whole genome without prior knowledge except the genetic perturbation.

In [9] Haverty *et al.* consider various single perturbations including stimuli and develop a computational method, CARRIE, that identifies a specific cellular response for each perturbation and that infers the transcriptional regulatory networks involved in each case. In this work Haverty *et al.* achieve the integration of some functional information (gene expressions) with some structural information (DNA regulatory motifs) in order to extract transcriptional interactions involved in the regulation of the stimulus response. This methodology is initiated by the identification of the genes that are directly involved in the stimulus response from expression data. This means that the causal aspect of the response is clearly identified. However, the heterogeneity of the gene expression kinetics is not exploited to identify different cellular responses to the same stimulus. The response to the stimulus is identified as a unique large modulated gene set. The causal relation between the stimulus and the involvement of the TFs identified

through the DNA motif information cannot be proved with only structural information. These TFs have to be considered as potential regulators and regulatory hypothesis for further biological experiments. In [4] Hvidsten *et al.* use also gene expression data and DNA regulatory motifs to identify cellular responses and their potential regulators. They propose an original rule based approach to automatically extract IF-THEN rules of minimal binding-site combination shared by genes with a common expression profile. This can be compared to the way we apply biclustering on the co-expressed genes. Logical rules are here another mean to justify the co-expression. To our knowledge this work is the first that uses a rule-based approach to extract regulatory information that could explain genes coexpression. However the integration of structural information only leads to extract and highlight one kind of regulatory mechanism: some coregulated genes will not be taken into consideration if not associated with a rule. In this work they exploit GO information and ChIP data in order to evaluate the found genes modules but they do not use these informations as an input for the regulatory rules extraction step. Finally, to conclude, although these approaches also exploit various sources of biological information in a sequential way, our approach differs from these ones by taking into account the dynamical nature of gene expression and formalizing explicitely the underlying logic on which a strategy of perturbation is usually based.

Conclusions

XRegPath is a new and generic methodology that helps the biologist in the process of regulatory pathways inference from large scale data. Inductive data-mining such as spectral clustering and biclustering allows to extract patterns of information while automated reasoning exploits the logic of the experiments. Combining these two methods allows automating efficiently several steps in the pathways discovery process. By this way, XRegPath extends the classic approach led by the biologist that usually reasons about his/her experimental results to infer hypotheses to large scale data. Further work consists in the improvement of this approach by taking into account uncertainty at several levels: for instance in clustering, biclustering and at the level of profiles annotation. Managing uncertainty will allow to provide the biologist with probabilities associated to inferred facts. Finally, the discovered regulatory pathways can serve as prior knowledge to finer reverse-modeling approaches such as parameter estimation in Ordinary Differential Equations or dynamical Bayesian networks [26].

Methods

Gene expression kinetics features and genetic features of the yeast strains

The XRegPath approach was applied to the yeast data for 7 different yeast strains. An expression kinetic had been measured for each yeast gene during the period time of cell response to irradiation. Cells were irradiated at time 0 and samples were collected for microarray analysis at different times during the period of 5 hours which corresponds to the time after which the cell culture is again asynchronous. This has been done for 7 yeast strains that differ by predefined genetic variations (see table 1): ploidy (number of chromosomes copies, n for haploid or 2n for diploid), mating type (display simple sexual differentiation in yeast, can be either the a or α mating type in haploid cells and a/α in diploid cells. a/α mating type in haploid cells is an artificial construction) and 3 targeted gene mutations: rad52 (involved in DNA double-strand break repair), ku70 (involved in chromatin silencing, telomere maintenance and DNA double-strand break repair via nonhomologous end joining) and sir2 (involved in chromatin assembly or disassembly and DNA double-strand break repair via nonhomologous end joining). Normalized microarray data were generated as described in [10]. Raw data sets are available in MIAME format at http://microarrays.curie.fr/.

Hetereogeneous biological data integration

Data sources

We used five kinds of additional biological data sets to analyze the contents of the clusters:

- **functional annotations** that describe to which known biological pathways (KEGG) [27] or biological processes a gene belongs or in which cellular components the gene product is localized (GO) [15].
- **ChIP data**: a list of the associations between nearly all of the DNA-binding transcriptional regulators (203 TFs) and target genes promoters in the yeast has been provided by Harbison *et al.* in [28].
- **protein complexes information** : we used recent results presented in [29] that provides a list of association between 2703 proteins and 546 complexes.
- **gene positions on chromosomes** : two different annotations for gene chromosome positions were used in order to identify a potential correlation between genes positions on chromosomes and their co-expression: (1) we associate each gene to the chromosome arm to which it belongs (2) we

considered as subtelomeric the genes located at less than 20 Kbp from chromosome telomeres (a position susceptible to be controlled by chromatine silencing at telomeres).

Biological descriptors selection

The analysis methods have been adapted to the representation of the different data in their respective databank. For GO descriptors we used BiNGO [30] tool that is a plugin for Cytoscape [31]. Due to GO hierarchical tree structure, a gene may belong to multiple descriptors simultaneously. Thus, p-values where corrected [21] for each GO descriptor to account for over-representation that will invariably occur by chance when multiple tests are carried out. For the other biological descriptors we used the GeneMerge tool [32]. For all kinds of information we selected gene descriptors that have a p-value equal or lower than 0.05.

Kernel parameter selection

The value of the kernel parameter γ has been chosen for each of the genetic background using a visualization method based on histograms of kernel values. The expected ideal distribution of kernel values has to correspond to a compromise between the homogeneity of the values distribution in order to allow all the possible values and the minimization of high kernel values. We tested all values of γ between 0.1 and 2). The value of γ=0.4 gives the most satisfactory kernel matrix to represent similarities between genes (see Additional information 1).

Clustering stability and partition size selection

The method exploits measurements of the stability of clustering solutions obtained by perturbing the data set. We implemented a Ben-Hur *et al.* [13] like stability measure that is characterized by the distribution of pairwise similarities between clusterings obtained from randomly generated sub samples of the data and the clustering obtained from the whole set of the data. High pairwise similarities indicate a stable clustering pattern. A loss of stability at a given size K of partition will mean that the inner structuration of the data is no longer respected by the partition into K clusters. We measured, for 6 different sizes of partitions (K= 5, 10, 15, 20, 25 and 30) the mean pairwise similarity and its standard deviation between clusterings obtained from randomly generated sub samples of our data and the clustering obtained from the whole set of our data (see Additional information 2). We see that our algorithm is globally very stable, particularly for partitions of sizes superior to 10, where the mean similarity between random partitions and the reference partition is between 0.9 and 1. However it seems that for our data this criterion does not

show any loss of clustering stability and thus does not allow us to directly identify an optimal size of partition. We defined then a stability measure based on pairwise similarities between whole data partitions of different sizes directly inspired from the stability measure defined by Ben-Hur *et al.*

$$S(L_K, L_{K+1}) = 1 - \frac{||C_K - C_{K+1}||^2}{n^2}, \tag{3}$$

here L_K and L_{K+1} are respectively the partitions into K and $K+1$ subsets, C_K and C_{K+1} are the binary matrices representations (N by N matrices with $(C_K)_{i,j}$=1 if genes i and j are in the same cluster, 0 otherwise) of L_K and L_{K+1} and n^2 is the total number of pairwise similarities between genes.
We measured the pairwise similarities for all successive partitions couples between $K = 2$ and $K = 35$ (see Additional information 3). The curve representing the evolution of this stability measure shows a rapid gain of clustering stability between artitions of sizes $K = 2$ and $K = 5$, with a stabilization for the following partitions couples. The monitoring of this stability criterion allowed us to define an interval of partition sizes where data structure stay preserved (between $k = 18$ and $k = 35$). The choice of the optimal partition size for our data was done in this interval, jointly with the biologists by using the graphical representation of cluster mean expression profiles to compare partition of size $K + 1$ with the partition of size K. If no new profile combination is obtained by increasing the size from K to $K + 1$ we consider that K is the optimal number of clusters.

Cluster profiles annotation

We choose to annotate the cluster behaviour with only two values: *modulated* or *unmodulated*. The attribution of the modulation status to a representative profile depends on many criteria, some of which are based on the experimental design and others are based on the reliability of an observed kinetic profile (see table 2). The chosen criteria are:

- Irregular expression fluctuations are not considered as significative
- The absolute value of an expression modulation amplitude has to be high enough: more than 1.5 fold between the lowest and highest time points.
- The reliability of each time point and the experimental conditions in which the data have been generated are taken into account. For instance, if the first measure is a triplicate, it is considered as more reliable than a unique measure.

- Observation of the same kinetic profile accross different strains for a cluster reinforces the validity of such a profile.
- Standard deviations at each time point reflects the homogeneity degree of a given cluster (the smallest the standard deviation is the most significant the modulation is).

Spectral biclustering for biological mining of clusters

The clustering problem deals with gene similarities as reflected by their belonging to a set of biological descriptors. Biclustering ([12]) takes as an input the same matrix of biological descriptors, and tries to find significant sub-matrices in it, called biclusters. The biclustering approach has the advantage of detecting sets of genes which have similar subset of biodescriptors but share no common belonging under the other bio-descriptors. Spectral biclustering ([5],[24]) is based on bipartite graph partitioning. The partition is constructed by minimizing a normalized sum of edge weights between unmatched pairs of vertices of the bipartite graph. As a result, each cluster may involve only a subset of genes and a subset of bio-descriptors, which form a "checkerboard" structure.
For each cluster, we build a gene-descriptor binary matrix X with all kind of biological descriptors terms that have passed the statistical test. We have used the spectral biclustering algorithm described in ([10]).

1. Preprocessing: Compute the matrices: $M_1 = R^{-1}XC^{-1}X^T$ and $M_2 = C^{-1}X^TR^{-1}X$
   where:
   $R^{-1}X$ is the normalization of X by rows
   $C^{-1}X^T$ is the normalization of X by columns
2. Spectral mapping:
   compute u_1,u_2, ...,u_{k_1} the k_1 largest eigenvectors of $M1$ (eigen problem for "conditions" and v_1,v_2, ...,v_{k_2} the k_2 largest eigenvectors of $M2$ (eigen problem for "genes"). Form the matrices
   $U = [u_2u_3...u_{k_1}]$ and $V = [v_2v_3...v_{k_2}]$.
3. Post Processing - Grouping:
   cluster U and V rows by the k-means algorithm to obtain the conditions and the genes clustering and form biclusters.

Availability and requirements

- Project name: XRegPath.

- Project home page: http://touleimat.googlepages.com/ (temporarily).
- Operating system(s): Platform independent.
- Programming language: Matlab and Java.
- License: licence in process, free use for academics (code on demand).
- Any restrictions to use by non-academics: licence needed.

Authors contributions

MD formulated the biological question and provided experimental data. FAB defined the whole methodology with the help of NT and from discussions with MD. NT participated to the definition of the methodology, implemented most of it and analyzed all the biological results with MD. FZ applied and tested the biclustering approach to the yeast data. NT and FAB drafted the paper with the help of MD. All the authors contributed and approved the final version.

Acknowledgements

FAB would like to thank Pierre Tambourin for fruitful discussions about this work and Genopole for their financial support (ATIGE). NT has been funded both by Institut Curie (Orsay, France) and University of Evry. All the authors are grateful to Pierre Geurts, Louis Wehenkel and Aurélien Mazurie for helpful comments about the manuscript.

References

1. Wagner A: **How to reconstruct a large genetic network from n gene perturbations in fewer than n2 easy steps**. *Bioinformatics* 2001, **17**(12):1183–1197, [http://bioinformatics.oxfordjournals.org/cgi/content/abstract/17/12/1183].
2. Clare A, Karwath A, Ougham H, King RD: **Functional bioinformatics for Arabidopsis thaliana**. *Bioinformatics* 2006, **22**(9):1130–1136.
3. Hvidsten TR, Komorowski J, Sandvik AK, A L: **Predicting Gene Function From Gene Expressions And Ontologies**. In *PSB*, *Volume 06* 2001:299–310.
4. Hvidsten TR, Wilczynski B, Kryshtafovych A, Tiuryn J, Komorowski J, Fidelis K: **Discovering regulatory binding-site modules using rule-based learning.** *Genome Res.* 2005, **15**(6):856–866.
5. Tanay A, Sharan R, Kupiec M, Shamir R: **Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data.** *Proceedings of the National Academy of Sciences* 2004, **101**(9):2981–2986.
6. Reiss DJ, Baliga NS, Bonneau R: **Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks**. *BMC Bioinformatics* 2006, **7**(280).

7. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network**. *Science* 2001, **292**(5518):929–934.

8. Workman CT, Mak HC, McCuine S, Tagne JB, Agarwal M, Ozier O, Begley TJ, Samson LD, Ideker T: **A Systems Approach to Mapping DNA Damage Response Pathways**. *Science* 2006, **312**(5776):1054–1059.

9. Haverty PM, Hansen U, Weng Z: **Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification.** *Nucl. Acids Res.* 2004, **32**:179–188.

10. Mercier G, Berthault N, Touleimat N, Kepes F, Fourel G, Gilson E, Dutreix M: **A haploid-specific transcriptional response to irradiation in Saccharomyces cerevisiae**. *Nucl. Acids Res.* 2005, **33**(20):6635–6643.

11. Verma D, Meila M: **A comparison of spectral clustering algorithms**. In *Technical report uw-cse-03-05-0 and university of washington1* 2001.

12. Schölkopf B, Smola AJ: *Learning with Kernels*. MIT Press, Cambridge 2002.

13. Ben-Hur A, Elisseeff A, Guyon I: **A stability based method for discovering structure in clustered data.** In *PSB, Volume 7* 2002:6–17.

14. Wielemaker J, Hildebrand M, van Ossenbruggen J: **Using Prolog as the fundament for applications on the semantic web.** In *Proceedings of the 2nd Workshop on Applicatiions of Logic Programming and to the web, Semantic Web and Semantic Web Services*. Edited by SHeymans ERDP A Polleres, Gupta G 2007:84–98, [http://hcs.science.uva.nl/projects/SWI-Prolog/articles/mn9c.pdf].

15. Consortium TGO: **Gene Ontology: tool for the unification of biology**. *Nature Gen.* 2000, **25**:25–29.

16. Cheng Y, Church GMH: **Biclustering of expresssion data.** In *ISMB* 2000.

17. Madeira SC, Oliveira AL: **Biclustering Algorithms for Biological Data Analysis: A Survey**. In *IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume 01* 2004:24–45.

18. Kluger Y, Basri R, Chang JT, Gerstein M: **Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions.** *Genome Res.* 2003, **13**(4):703–716.

19. Dhillon IS: **Co-clustering documents and words using bipartite spectral graph partitioning.** In *Proc. ACM Int'l Conf Knowledge Disc. Data Mining* 2001.

20. Zha H, Ding C, Gu M, Simon HD: **Bipartite Graph Partitioning and Data Clustering.** In *ACM 10th Int'l Conf. Information and Knowledge Management*, Atlanta 2001:25–31.

21. Sokal RR, Rohlf FJ: *Biometry: The Principles and Practice of Statistics in Biological Research* 1995.

22. Mills KD, Sinclair DA, Guarente L: **MEC1-dependent redistribution of the Sir3 silencing protein from telomeres to DNA double-strand breaks**. *Cell* 1999, **97**:609–620.

23. Martin SG, Laroche T, Suka N, Grunstein M, M GS: **Relocalization of telomeric Ku and SIR proteins in response to DNA strand breaks in yeast.** *Cell* 1997, **97**:621–633.

24. Jorgensen P, Rupes I, Sharom JR, Schneper L, Broach JR, Tyers M: **A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size**. *Genes Dev.* 2004, **18**(20):2491–2505.

25. Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO: **Genomic Expression Responses to DNA-damaging Agents and the Regulatory Role of the Yeast ATR Homolog Mec1p**. *Mol. Biol. of the Cell* 2001, **12**:2987–3003.

26. Quach M, Brunel N, d'Alché Buc F: **Estimating parameters and hidden variables in non-linear state-space models based on ODEs for biological networks inference.** *Bioinformatics* 2007, (23):3209–3216.

27. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucl. Acids Res.* 2000, **28**:27–30.

28. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431**(7004):99–104.

29. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrìn-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MHY, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, OŚhea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF: **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.** *Nature* 2006, **440**(7084):637–643.

30. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks**. *Bioinformatics* 2005, **21**(16):3448–3449.

31. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks**. *Genome Res.* 2003, **13**(11):2498–2504.

32. Castillo-Davis CI, Hartl DL: **GeneMerge–post-genomic analysis and data mining and and hypothesis testing**. *Bioinformatics* 2003, **19**(7):891–892.

Figures

Figure 1 - The main steps of XRegPath.

XRegPath is decomposed in five steps and exploits three kinds of information in an original way: kinetics of gene expression measured in various genetic systems (differing by predefined genetic perturbations), logical rules that describe the logic underlying the perturbations and additional biological data that come from various databases or ontologies.

Figure 2 - Examples of multiclustering results: cluster behaviour representation into a set of genetic backgrounds.

We represented each cluster of coherent genes by the mean profile of its genes and by its standard deviation in each studied yeast strain. C2: cluster C2. C15: cluster C15. Expression intensity: log2 of normalized data intensity (see [10]). A to G : yeast strains in the study (A: 18733, B: 18734, C: 18735, D: 6053, E: FFD1, F: LM79, G: YiBPC205). Continuous lines: mean expression profiles of the gene cluster into the given genetic background. Dotted lines: mean expression profiles +/- the standard deviation into the given genetic background. NI: Non Irradiated.

Figure 3 - Biclustering results of clusters C2 and C15.

Each biclustering pattern is a graphical representation of the reorganized binary matrix associating a gene to a biological descriptor. A black square represents a '1' entry in the binary matrix and corresponds to a

significative association between a gene and a biological descriptor. Figure 3A: biclustering result of cluster C2; figure 3B: biclustering result of cluster C15.

Figure 4 - C2 and C15 IR response regulatory scheme.

Circles and rectangles: regulation factors deduced from the experimental design. Diamonds: potential experimental targets, deduced from ChIP data association.

Tables

Table 1 - Gene expression kinetics features and genetic features of the yeast strains under study

NI: Non Irradiated ; times after irradiation in hours ; NI - 5h : NI, 0.1h, 1h, 2h, 3h, 4h and 5h.

	Yeast strains						
	18733	18734	18735	6053	FFD1	LM79	YiBPC205
Ploidy	n	n	n	2n	n	n	n
Mating type	a	α	a,α	a,α	a	α	α
Gene mutation	none	none	none	none	rad52	ku70	sir2
Expression measures	NI - 5h	NI - 5h	NI - 5h	NI - 5h	NI - 5h	NI - 5h	NI, 0.1, 1, 3, 5h

Table 2 - Semantic representation of gene cluster behaviour in the different strains

M: modulated mean expression profile, U: unmodulated.

Yeast strain	18733	18734	18735	6053	FFD1	LM79	YiBPC205
c1 (138 genes)	U	U	U	U	U	U	U
c2 (212 genes)	M	M	U	U	M	U	U
c3 (284 genes)	U	U	U	U	U	U	U
c4 (219 genes)	U	U	U	U	U	U	U
c5 (231 genes)	U	U	U	U	U	U	U
c6 (239 genes)	U	U	U	U	U	U	U
c7 (263 genes)	U	U	U	U	U	U	U
c8 (25 genes)	M	M	M	U	M	M	M
c9 (210 genes)	M	U	U	U	U	U	U
c10 (220 genes)	U	U	U	U	U	M	M
c11 (320 genes)	U	U	U	U	U	U	U
c12 (172 genes)	U	U	U	U	U	U	U
c13 (101 genes)	M	M	M	M	M	M	M
c14 (259 genes)	U	U	U	U	U	U	U
c15 (198 genes)	M	M	M	U	M	U	U
c16 (216 genes)	U	U	U	U	U	U	U
c17 (247 genes)	U	U	U	U	U	U	U
c18 (304 genes)	U	U	U	M	U	U	U
c19 (277 genes)	U	U	U	U	U	U	U
c20 (226 genes)	U	U	U	U	U	U	U
c21 (252 genes)	U	U	U	U	U	M	U
c22 (119 genes)	M	M	M	U	M	U	U

Table 3 - Descriptions of predicates variables.

Variables	Description	Values
C	Set of genes (co-expressed gene clusters).	$c01$,...$c22$
S	Genetic system (yeast strain).	$s18733$, $s18734$, $s18735$, $s6053$, $sffd1$, $slm79$, $syibpc205$
E	Gene expression profile discretization	m (modulated) or u (unmodulated)
MS	Gene mutation status of S, takes into account single gene mutations.	$msrad52$ (gene rad52 mutated, $msku70$, $mssir2$, $mswt$ (wild type, no mutation)
M	Mating type status of S.	ma, $m\alpha$, $ma\alpha$
P	Ploidy status of S.	n, $2n$

Table 4 - Formalization and descriptions of regulatory predicates based on primary sources of data.

background(S,P,M,MS)	the genetic background (yeast strain) S has a ploidy status P, mating type status M and a mutation status MS
expression(C,S,E)	the expression of cluster C in the genetic background S is annotated as E
differexp(C,S1,S2)	A group of gene C has different expression behaviours between genetic background $S1$ and $S2$
regulation(C,X1,X2,X)	gene cluster C response to stimulus is regulated by the genetique variation $X1/X2$
noregulation(C,X1,X2,X)	negation of the previous predicate
dependent2regulation(C,X1,X2,X,Y1,Y2,Y)	gene cluster C response to stimulus is regulated by the cooccurence of two kinds of genetic variations $X1/X2$ and $Y1/Y2$.
nodependent2regulation(C,X1,X2,X,Y1,Y2,Y)	negation of the previous predicate.
suppressive2regulation(C,X1,X2,X,Y1,Y2,Y)	gene cluster C response to stimulus is regulated by 2 kinds of genetic variations $X1/X2$ and $Y1/Y2$ but the coocurence of these 2 genetic variations suppresses their single regulation effects on C.
nosuppressive2regulation(C,X1,X2,X,Y1,Y2,Y)	negation of the previous predicate.
dependent3regulation(C,X1,X2,X,Y1,Y2,Y,Z1,Z2,Z)	gene cluster C response to stimulus is regulated by the cooccurence of three kinds of genetic variations $X1/X2$, $Y1/Y2$ and $Z1/Z2$.
nodependent3regulation(C,X1,X2,X,Y1,Y2,Y,Z1,Z2,Z)	negation of the previous predicate.
suppressive3regulation(C,X1,X2,X,Y1,Y2,Y,Z1,Z2,Z)	gene cluster C response to stimulus is regulated by 3 kinds of genetic variations $X1/X2$, $Y1/Y2$ and $Z1/Z2$ but the coocurence of these 3 genetic variations suppresses their single regulation effects on C.
nosuppressive3regulation(C,X1,X2,X,Y1,Y2,Y,Z1,Z2,Z)	negation of the previous predicate.

Table 5 - Automated deduction results for clusters C2 and C15.

We use the theorems defined on the basis of the genetic perturbation strategy that aims to test the influence of genetic variables on the cell response to irradiation. For clusters C2 and C15 all the possible genetic variations combination have been tested for the regulatory rules that involve only one factor.

	Cluster 02	Cluster 15
regulation:		
. mating type	aα/α	
	aα/a	
. ploidy		n/2n
. mutation	none/ku70	ku70/none
	none/sir2	sir2/none
	rad52/ku70	rad52/ku70
	rad52/sir2	rad52/sir2
noregulation:		
. mating type	a/α	α/aα
		a/aα
		α/a
. ploidy	n/2n	
. mutation	none/rad52	none/rad52
	ku70/sir2	ku70/sir2
dependent2regulation:	no	no
nodependent2regulation:		
. mating type/mutation	(α/a)/(none/rad52)	(aα/a)/(none/rad52)
		(α/a)/(none/rad52)
suppressive2regulation:		
. mating type/mutation	(α/aα)/(none/ku70)	
	(α/aα)/(none/sir2)	
. ploidy/mutation		(n/2n)/(none/ku70)
		(n/2n)/(none/sir2)
nosuppressive2regulation:	no	no
dependent3regulation:	no	no
nodependent3regulation:	no	no
suppressive3regulation:	no	no
nosuppressive3regulation:	no	no

Additional File

Additional information 1 - Histograms of kernel values distributions according to all the studied gene couples.

EPS figure available on http://touleimat.googlepages.com/home/

Kernels presented here correspond to the normalizes linear combination of the 7 kernels computed for each genetic background. A: Gaussian kernel with parameter γ=0.3. B: Gaussian kernel with parameter γ=0.4 (best result). C: Gaussian kernel with parameter γ=0.5.

Additional information 2 - Stability measure based on comparison of partitions of random subsampled genes.

EPS figure available on http://touleimat.googlepages.com/home/

Circles: mean pairwise similarity measure between partitions obtained from randomly generated genes subsamples. Error bars: mean pairwise similarity measure +/- its standard deviation.

Additional information 3 - Stability measure based on comparison of partitions of increasing sizes.

EPS figure available on http://touleimat.googlepages.com/home/

The curve represent the evolution of the pairwise stability between two partitions of increasing sizes. This stability is defined in the Methods section: Clustering stability and partition size selection (see formula 3).

Additional information 4 - Mean expression profiles of all the co-expressed gene clusters.

EPS figures available on http://touleimat.googlepages.com/home/

We represented each cluster of coherent genes by the mean profile of its genes and by its standard deviation in each studied yeast strain. Common legend of figures 4a to 4v: A to G: yeast strains in the study (A: 18733, B: 18734, C: 18735, D: 6053, E: FFD1, F: LM79, G, YiBPC205) ; Expression intensity: log2 of normalized data intensity (see [10]) ; Continuous lines: mean expression profiles of the gene cluster in the given genetic background ; dotted lines: mean expression profiles +/- the standard deviation in the given genetic background ; NI: Non Irradiated.

Additional information 5 - Knowledge basis that reflects experimental design logic. Written in Prolog language.

The commented Prolog code of the knowledge basis is available on http://touleimat.googlepages.com/home/. Tables 3 and 4 describe the predicate variables and the regulatory predicates used in the knowledge basis.

Additional materials 1 - Lists of genes clusters

The 22 lists of genes corresponding to the 22 clusters are downloadable at: http://touleimat.googlepages.com/home/.

Additional materials 2 - Lists of classified genes and their cluster affectation

The list of the 4732 classified genes and their affectation to one of the 22 clusters is downloadable at: http://touleimat.googlepages.com/home/.

Additional materials 3 - Original reorganized binary matrices after biclustering

The original reorganized binary matrices with gene ID and biological descriptors, corresponding to clusters C2 and cluster C15, are available on http://touleimat.googlepages.com/home/.

XRegPath: semi-automated extraction of regulatory pathways using genetic perturbation data

Figures

Nizar Touleimat*[1,2], Farida Zehraoui[1] , Marie Dutreix[2] and Florence d'Alché-Buc*[1]

[1]Informatique Biologie Intégrative et Systèmes Complexes (IBISC) 2873 FRE CNRS - Université d'Evry-Val d'Essonne, Tour Evry 2, 523 place des terrasses de l'Agora, 91000 Evry - FRANCE

[2]Recombination, Repair and Cancer, Translational department, Institut Curie, Centre Universitaire - Batiment 110-112, F-91405 Orsay cedex - FRANCE

Email: Nizar Touleimat*- nizar.touleimat@ibisc.fr; Farida Zehraoui - zehraoui@ibisc.fr; Marie Dutreix - marie.dutreix@curie.u-psud.fr; Florence d'Alché-Buc*- florence.dalche@ibisc.fr;

*Corresponding author

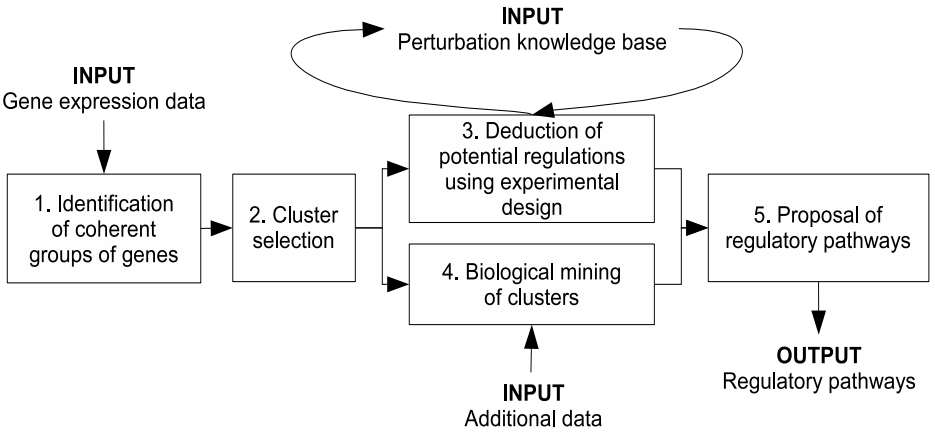


Figure 1: The main steps of XRegPath. XRegPath is decomposed in five steps and exploits three kinds of information in an original way: kinetics of gene expression measured in various genetic systems (differing by predefined genetic perturbations), logical rules that describe the logic underlying the perturbations and additional biological data that come from various databases or ontologies.

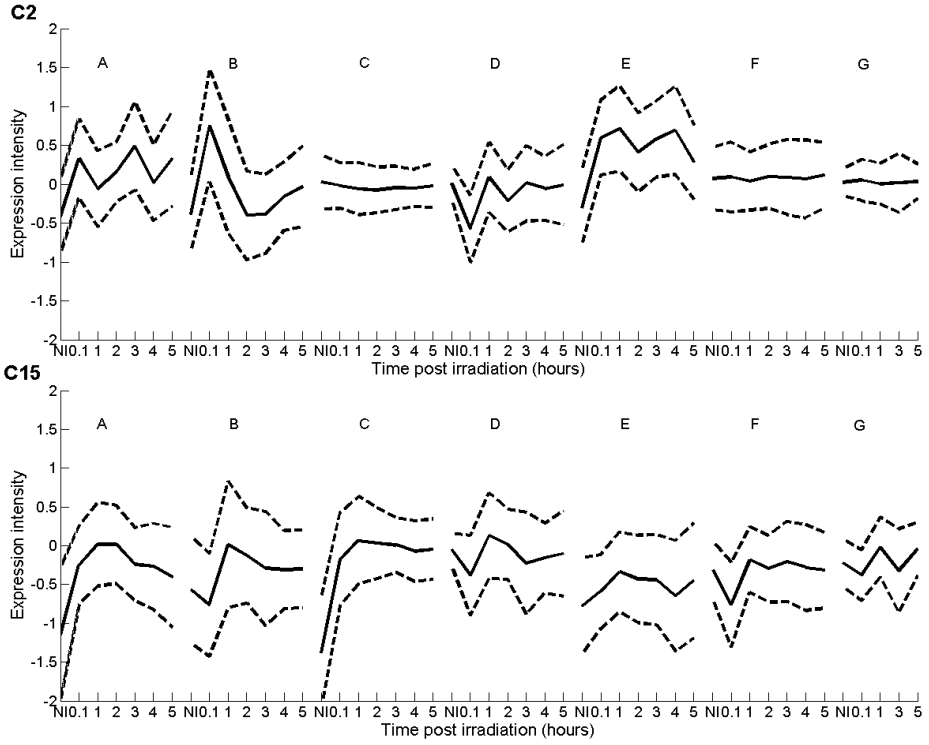
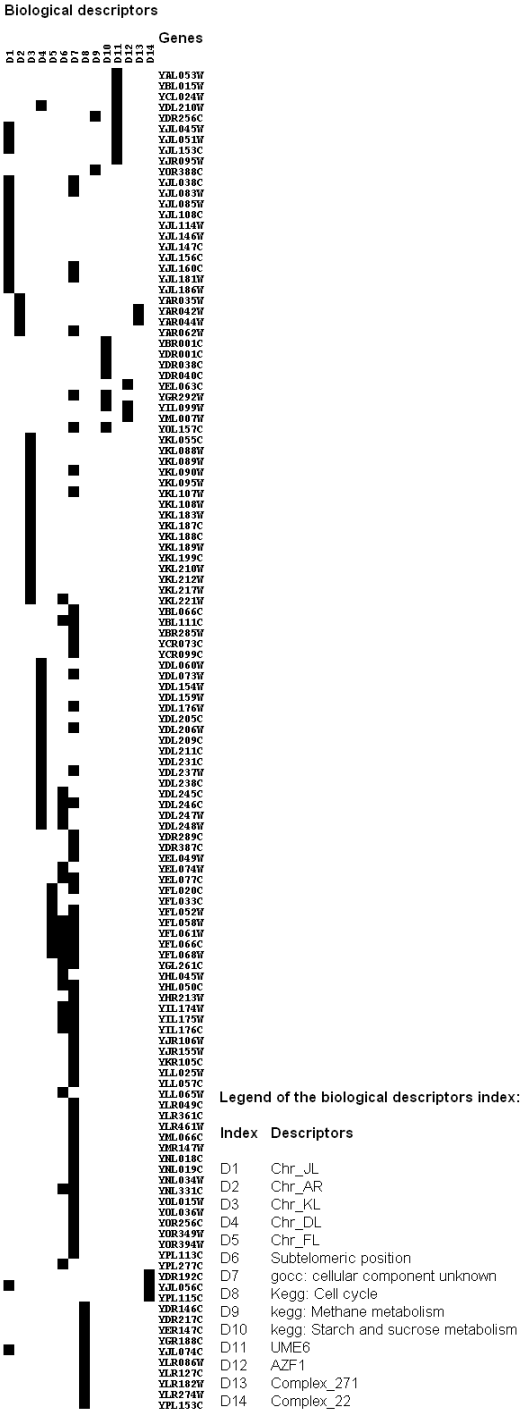


Figure 2: Examples of multiclustering results: cluster behaviour representation into a set of genetic backgrounds. We represented each cluster of coherent genes by the mean profile of its genes and by its standard deviation in each studied yeast strain. C2: cluster C2. C15: cluster C15. Expression intensity: log2 of normalized data intensity (see [10]). A to G : yeast strains in the study (A: 18733, B: 18734, C: 18735, D: 6053, E: FFD1, F: LM79, G: YiBPC205). Continuous lines: mean expression profiles of the gene cluster into the given genetic background. Dotted lines: mean expression profiles +/- the standard deviation into the given genetic background. NI: Non Irradiated.

A


Biological descriptors
Genes
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
D11
D12
D13
D14
YAL053W
YBL015W
YCL024W
YDL210W
YDR256C
YJL045W
YJL051W
YJL153C
YJR095W
YOR388C
YJL038C
YJL083W
YJL085W
YJL108C
YJL114W
YJL146W
YJL147C
YJL156C
YJL160C
YJL181W
YJL186W
YAR035W
YAR042W
YAR044W
YAR062W
YBR001C
YDR001C
YDR038C
YDR040C
YEL063C
YGR292W
YIL099W
YML007W
YOL157C
YKL055C
YKL088W
YKL089W
YKL090W
YKL095W
YKL107W
YKL108W
YKL183W
YKL187C
YKL188C
YKL189W
YKL199C
YKL210W
YKL212W
YKL217W
YKL221W
YBL066C
YBL111C
YBR285W
YCR073C
YCR099C
YDL060W
YDL073W
YDL154W
YDL159W
YDL176W
YDL205C
YDL206W
YDL209C
YDL211C
YDL231C
YDL237W
YDL238C
YDL245C
YDL246C
YDL247W
YDL248W
YDR289C
YDR387C
YEL049W
YEL074W
YEL077C
YFL020C
YFL033C
YFL052W
YFL058W
YFL061W
YFL066C
YFL068W
YGL261C
YHL045W
YHL050C
YHR213W
YIL174W
YIL175W
YIL176C
YJR106W
YJR155W
YKR105C
YLL025W
YLL057C
YLL065W
YLR049C
YLR361C
YLR461W
YML066C
YMR147W
YNL018C
YNL019C
YNL034W
YNL331C
YOL015W
YOL036W
YOR256C
YOR349W
YOR394W
YPL113C
YPL277C
YDR192C
YJL056C
YPL115C
YDR146C
YDR217C
YER147C
YGR188C
YJL074C
YLR086W
YLR127C
YLR182W
YLR274W
YPL153C
Legend of the biological descriptors index:
Index Descriptors
D1 Chr_JL
D2 Chr_AR
D3 Chr_KL
D4 Chr_DL
D5 Chr_FL
D6 Subtelomeric position
D7 gocc: cellular component unknown
D8 Kegg: Cell cycle
D9 kegg: Methane metabolism
D10 kegg: Starch and sucrose metabolism
D11 UME6
D12 AZF1
D13 Complex_271
D14 Complex_22

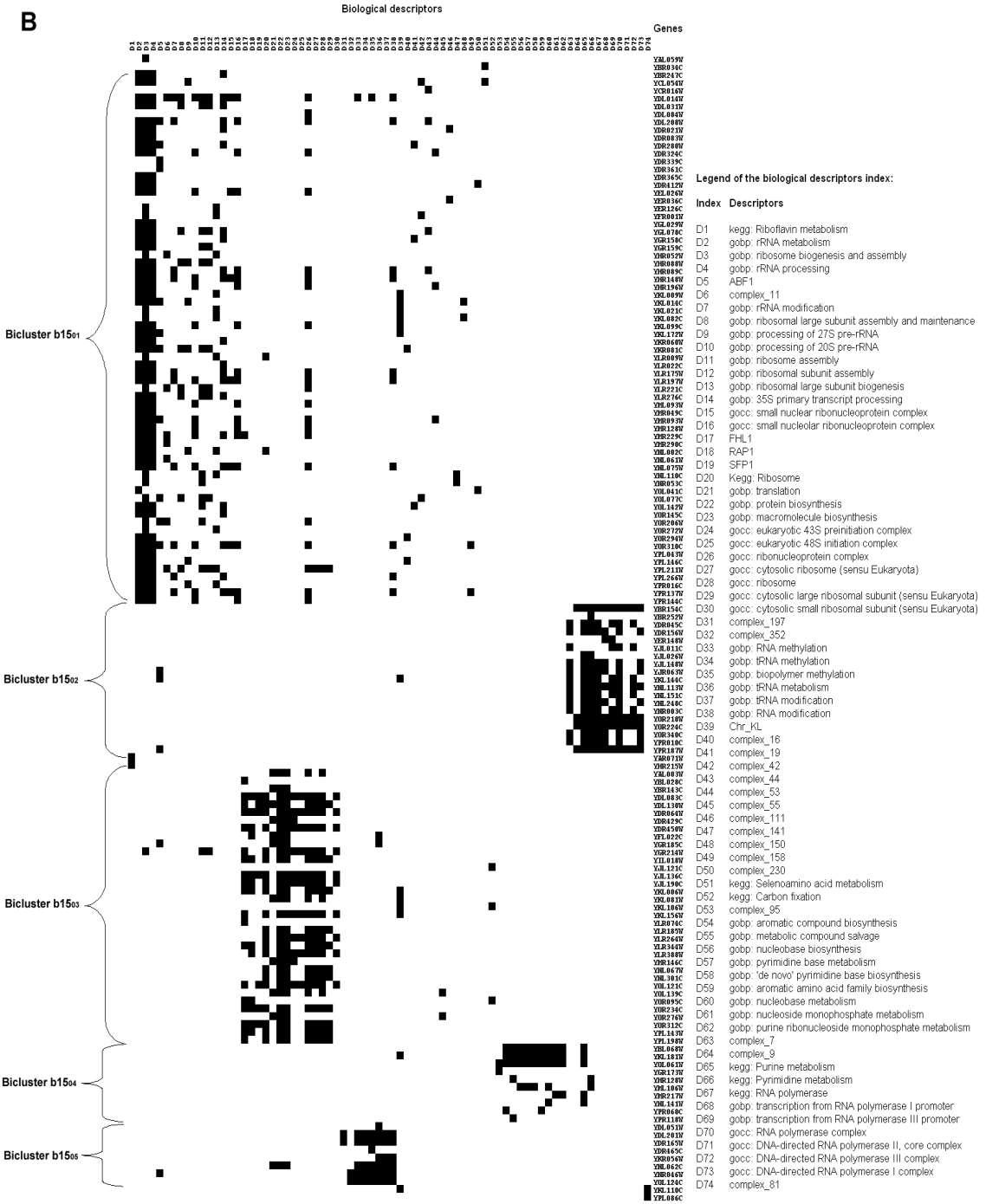


Figure 3: Biclustering results of clusters C2 and C15. Each biclustering pattern is a graphic representation of the reorganized binary matrix associating a gene to a biological descriptor. A black square represent a '1' entry in the binary matrix and corresponds to a significative association between a gene and a biological descriptor. Figure 3A: Biclustering results of cluster C2. Figure 3B: Biclustering results of cluster C15.

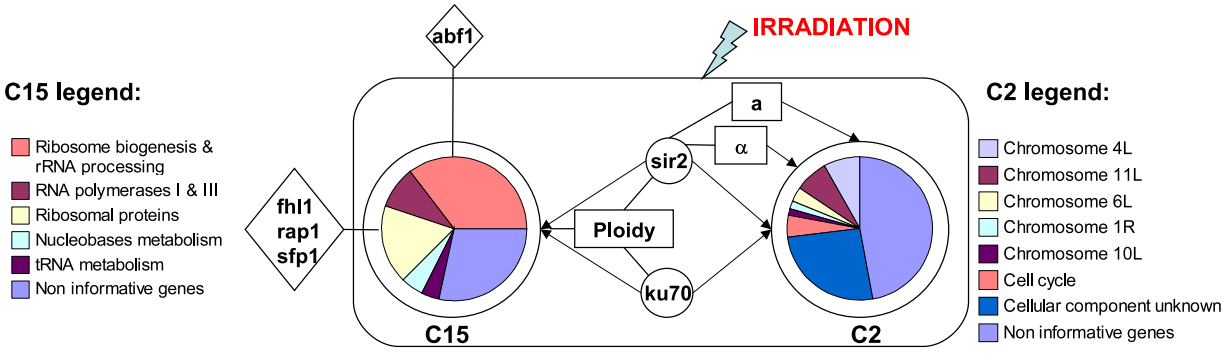


Figure 4: C2 and C15 IR response regulatory scheme. Circles and rectangles: regulation factors deduced from the experimental design. Diamonds: potential experimental targets, deduced from ChIP data association.

XRegPath: semi-automated extraction of regulatory pathways using genetic perturbation data
Additional file

Nizar Touleimat*[1,2], Farida Zehraoui[1] , Marie Dutreix[2] and Florence d'Alché-Buc*[1]

[1]Informatique Biologie Intégrative et Systèmes Complexes (IBISC) 2873 FRE CNRS - Université d'Evry-Val d'Essonne, Tour Evry 2, 523 place des terrasses de l'Agora, 91000 Evry - FRANCE

[2]Recombination, Repair and Cancer, Translational department, Institut Curie, Centre Universitaire - Batiment 110-112, F-91405 Orsay cedex - FRANCE

Email: Nizar Touleimat*- nizar.touleimat@ibisc.fr; Farida Zehraoui - zehraoui@ibisc.fr; Marie Dutreix - marie.dutreix@curie.u-psud.fr; Florence d'Alché-Buc*- florence.dalche@ibisc.fr;

*Corresponding author

Additional informations

Additional information 1 - Histograms of kernel values distributions according to all the studied gene couples.

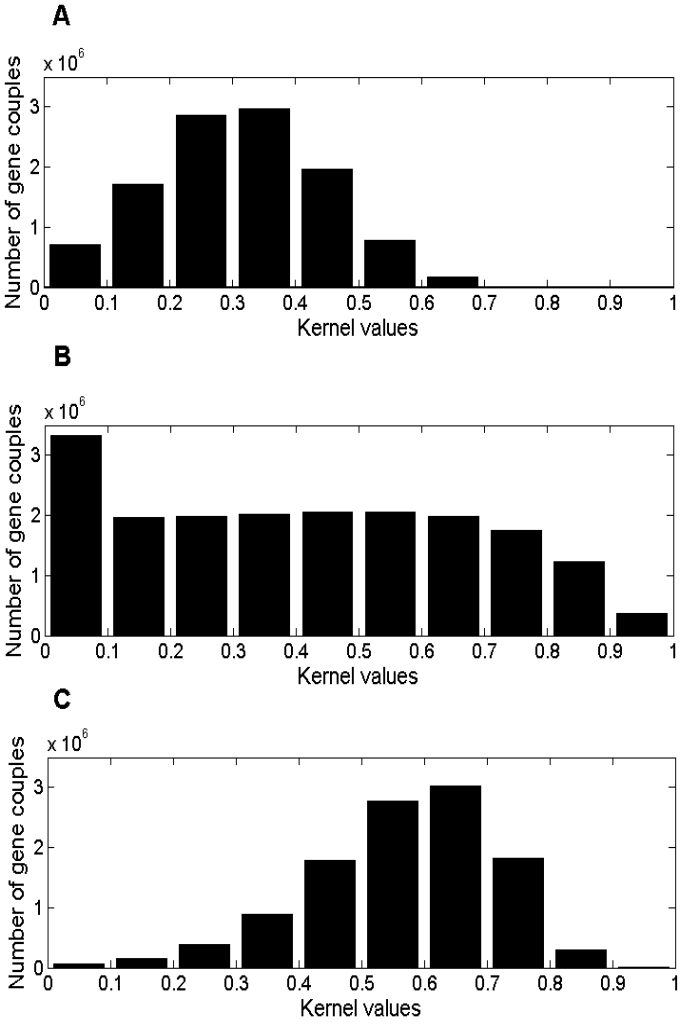


Figure 1: Histograms of kernel values distributions according to all the studied gene couples. Kernels presented here correspond to the normalizes linear combination of the 7 kernels computed for each genetic background. A: Gaussian kernel with parameter γ=0.3. B: Gaussian kernel with parameter γ=0.4 (best result). C: Gaussian kernel with parameter γ=0.5.

Additional information 2 - Stability measure based on comparison of partitions of random subsampled genes.

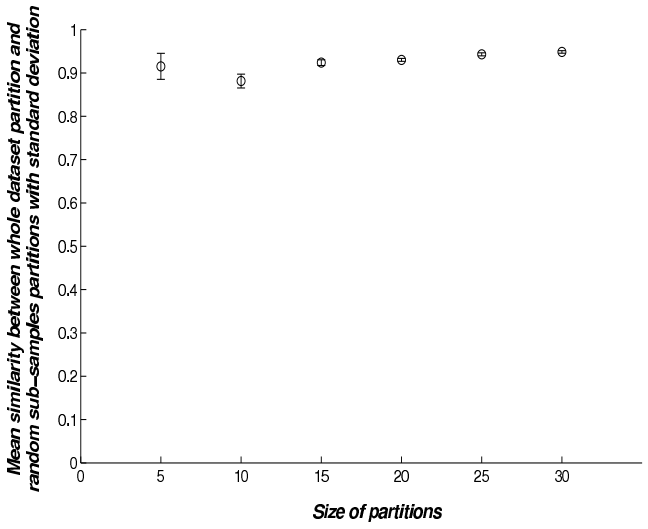


Figure 2: Stability measure based on comparison of partitions of random subsampled genes. Circles: mean pairwise similarity measure between partitions obtained from randomly generated genes subsamples. Error bars: mean pairwise similarity measure +/- its standard deviation.

Additional information 3 - Stability measure based on comparison of partitions of increasing sizes.

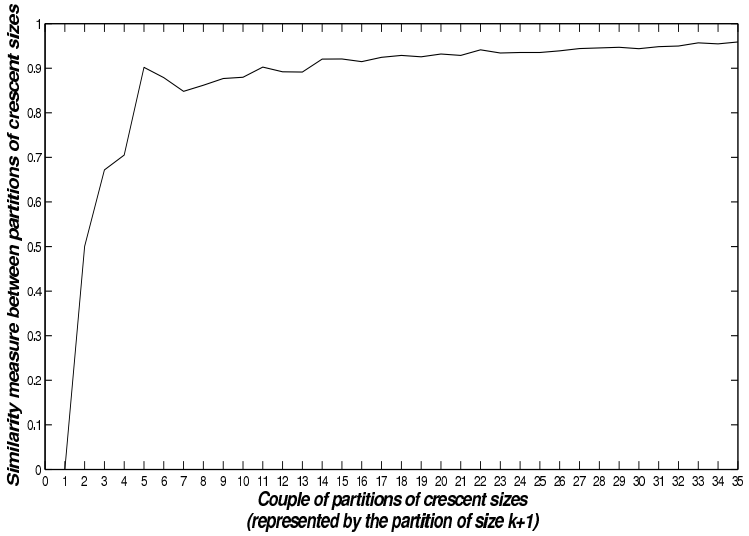
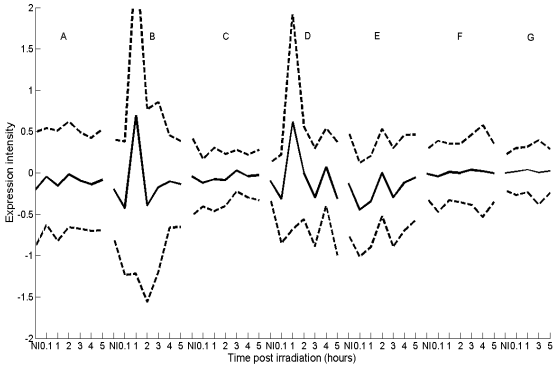
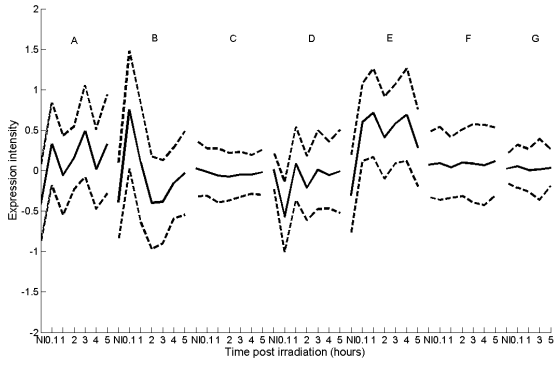


Figure 3: Stability measure based on comparison of partitions of increasing sizes. The curve represent the evolution of the pairwise stability between two partitions of increasing sizes. This stability is defined in the Methods section: Clustering stability and partition size selection (see formula 3).

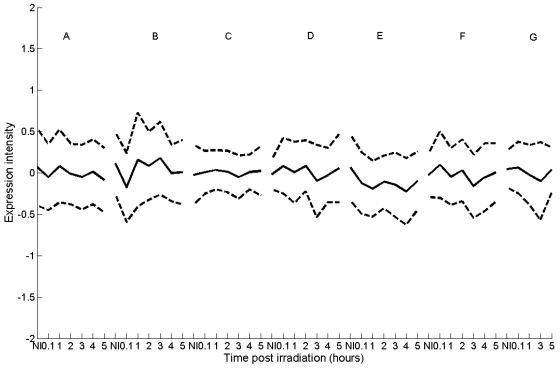
Additional information 4 - Mean expression profiles of all the co-expressed gene clusters.



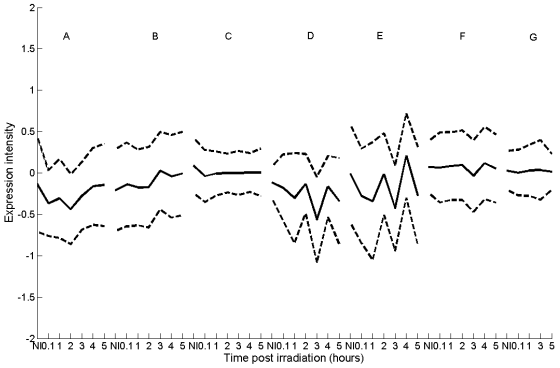
(a) Mean expression profiles of cluster C1 (138 genes)



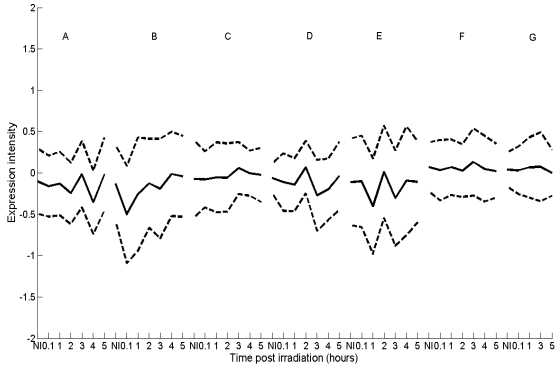
(b) Mean expression profiles of cluster C2 (212 genes).



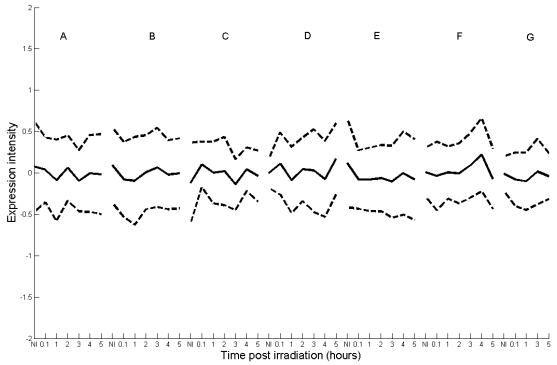
(c) Mean expression profiles of cluster C3 (284 genes).



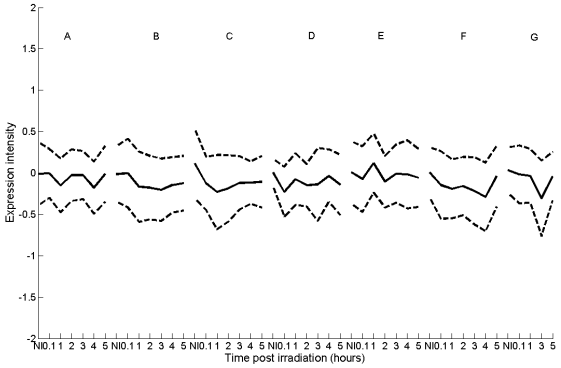
(d) Mean expression profiles of cluster C4 (219 genes).



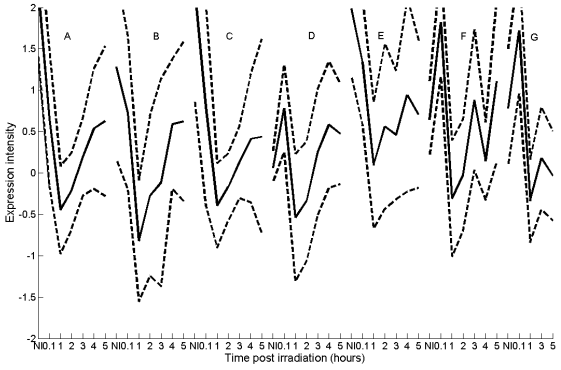
(e) Mean expression profiles of cluster C5 (231 genes).



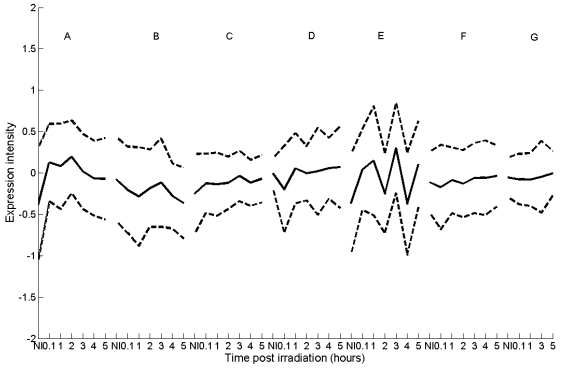
(f) Mean expression profiles of cluster C6 (239 genes).



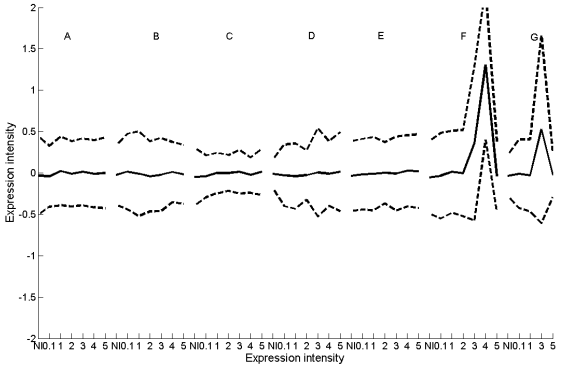
(g) Mean expression profiles of cluster C7 (263 genes).



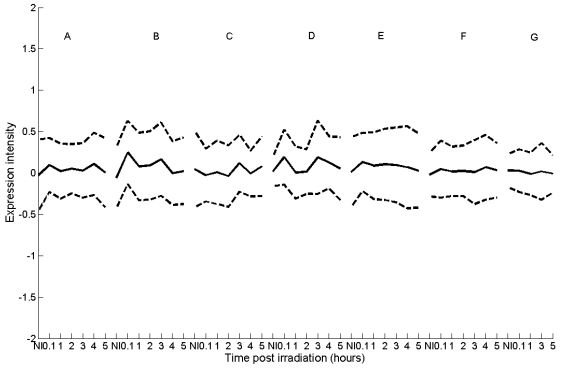
(h) Mean expression profiles of cluster C8 (25 genes).



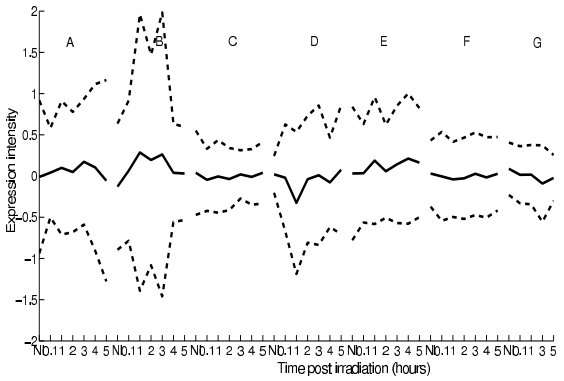
(i) Mean expression profiles of cluster C9 (210 genes).



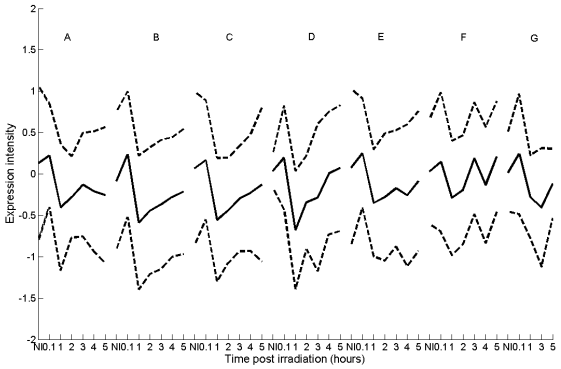
(j) Mean expression profiles of cluster C10 (220 genes).



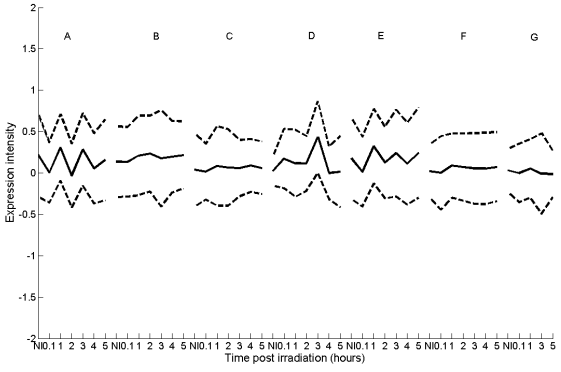
(k) Mean expression profiles of cluster C11 (320 genes).



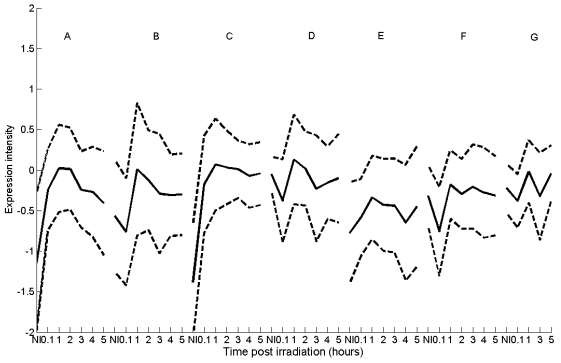
(l) Mean expression profiles of cluster C12 (172 genes).



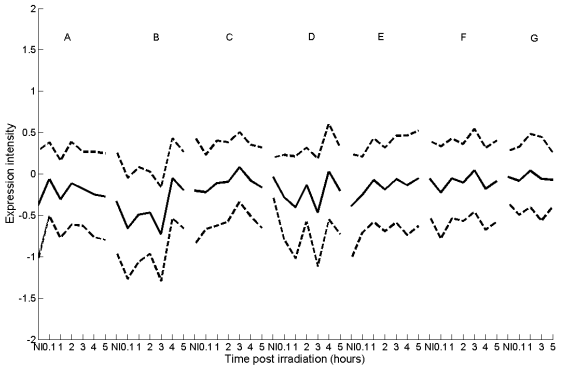
(m) Mean expression profiles of cluster C13 (101 genes).



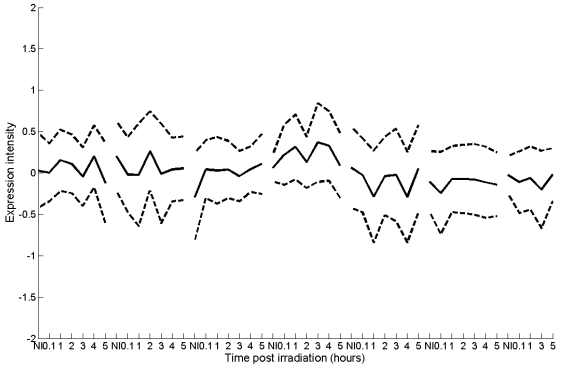
(n) Mean expression profiles of cluster C14 (259 genes).



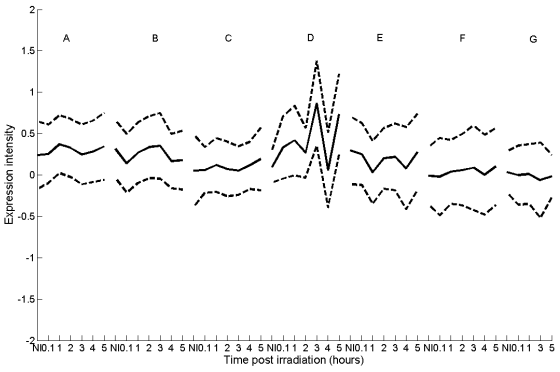
(o) Mean expression profiles of cluster C15 (198 genes).



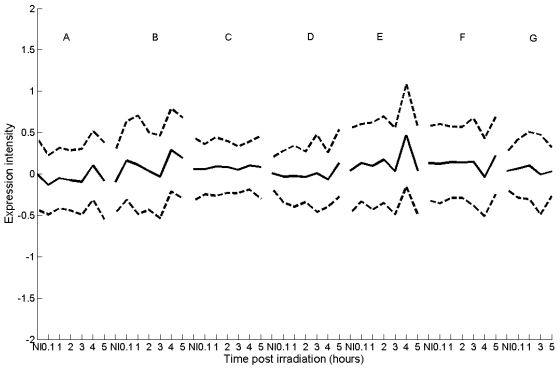
(p) Mean expression profiles of cluster C16 (216 genes).



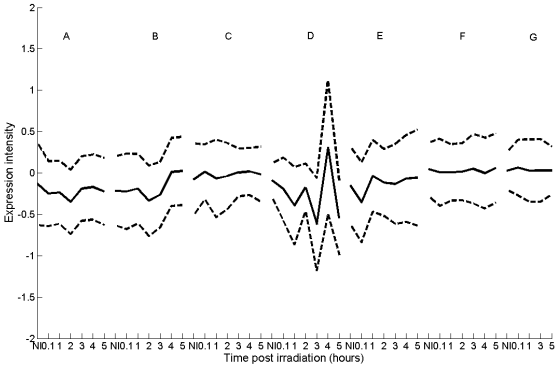
(q) Mean expression profiles of cluster C17 (247 genes).



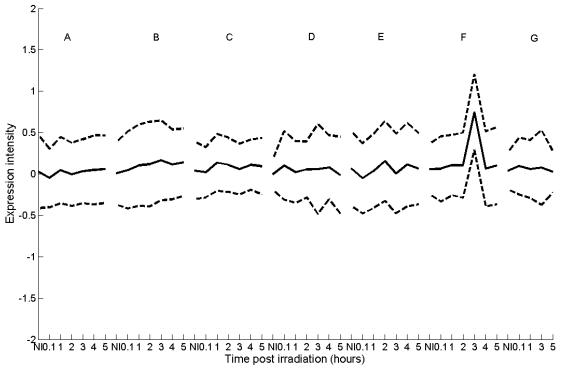
(r) Mean expression profiles of cluster C18 (304 genes).



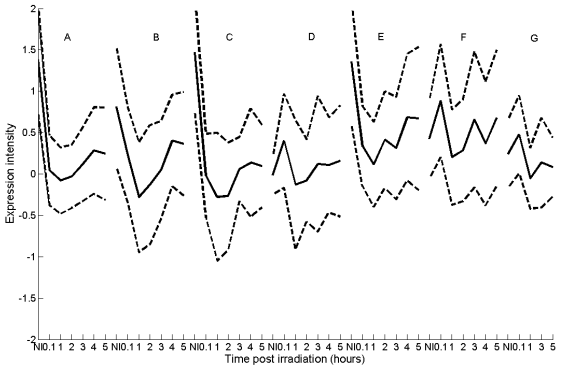
(s) Mean expression profiles of cluster C19 (277 genes).



(t) Mean expression profiles of cluster C20 (226 genes).



(u) Mean expression profiles of cluster C21 (252 genes).



(v) Mean expression profiles of cluster C22 (119 genes).

Figure 4: Mean expression profiles of all the co-expressed gene clusters. We represented each cluster of coherent genes by the mean profile of its genes and by its standard deviation in each studied yeast strain (see figures 8A to 8V). Common legend of figures 4a to 4v: A to G: yeast strains in the study (A: 18733, B: 18734, C: 18735, D: 6053, E: FFD1, F: LM79, G, YiBPC205) ; Expression intensity: log2 of normalized data intensity (see [10]) ; Continuous lines: mean expression profiles of the gene cluster in the given genetic background ; dotted lines: mean expression profiles +/- the standard deviation in the given genetic background ; NI: Non Irradiated.

Additional informations 5 - Knowledge basis that reflects experimental design logic. Written in Prolog language.

The commented Prolog code of the knowledge basis is available on http://touleimat.googlepages.com/home/. Tables 3 and 4 describe the predicate variables and the regulatory predicates used in the knowledge basis.

Additional materials

Additional materials 1 - Lists of genes clusters

The 22 lists of genes corresponding to the 22 clusters are downloadable at: http://touleimat.googlepages.com/home/.

Additional materials 2 - Lists of classified genes and their cluster affectation

The list of the 4732 classified genes and their affectation to one of the 22 clusters is downloadable at: http://touleimat.googlepages.com/home/.

Additional materials 3 - Original reorganized binary matrices after biclustering

The original reorganized binary matrices with gene ID and biological descriptors, corresponding to clusters C2 and cluster C15, are available on http://touleimat.googlepages.com/home/.

4.1.3 Analyses complémentaires

4.1.3.1 Application de *XRegPath* à des données publiées

Gasch et coll. [60] ont déjà analysé les effets de deux génotoxiques (MMS et radiations γ) sur le transcriptome de la levure *S. cerevisiae* (voir présentation de leurs résultats à la section 3.1.3, 77). Leurs données correspondent à des cinétiques d'expression de gènes mesurées en réponse aux deux génotoxiques. Ils analysent aussi l'effet d'une mutation (gène MEC1) sur la réponse des cellules au MMS et à l'IR. Leur données et leur problématique étant relativement similaires aux nôtres, nous avons appliqué notre méthodologie *XRegPath* sur un ensemble de leurs données d'expression. Nous étions essentiellement intéressés par les résultats que produiraient la 1ère étape de notre méthodologie, l'identification de groupes de gènes co-exprimés. Il s'agit de l'étape la plus importante de notre approche car c'est elle qui va identifier les composants essentiels du futur réseau de régulation, qui va en déterminer la granularité et qui permettra d'en interpréter le comportement. Nous souhaitions aussi vérifier si les résultats obtenus avec les données de Gasch *et coll.* pouvaient corroborer les résultats obtenus avec nos propres données et ainsi valider notre approche sur un autre jeu de données.

Nous avons sélectionné un ensemble de 3650 gènes ayant des cinétiques complètes dans les 6 conditions suivantes (voir tableau 4.1) : souche sauvage (wt) + IR, souche mutée (*mec1*) + IR, wt + MMS, *mec1* + MMS, wt + MOCK (témoin expérimental pour la condition irradié) et *mec1* + MOCK. Nous avons déterminé les paramètres du noyau calculant la similarité entre deux gènes et la taille de la partition optimale selon les procédures présentées dans notre article I (voir page 87). Les résultats qui nous ont permis de choisir ces différents paramètres sont présentés en annexe page 192. Nous avons déterminé que la taille de partition optimale K, permettant de représenter l'hétérogénéité des profils d'expression des gènes dans les 6 conditions, était $K = 13$.

Nous avons regroupés les 3650 gènes en 13 classes. Parmi ces classes, nous avons identifié 5 classes de gènes comme ayant un profil d'expression moyen modulé dans au moins une condition de stress (voir figure 4.2). Au total cela représente 929 gènes sensibles à au moins un des deux stimuli, soit environ 25% du total des gènes analysés.
Notre avons d'abord analysé les profils moyens de chaque classe pour en extraire d'éventuelles dépendances aux stimuli et/ou à la mutation :

- C1_gasch (voir figure 4.2 (a)) : les gènes de cette classe répondent positivement à l'IR mais perdent leurs cohérence lorsqu'ils sont exposés au MMS. Cette réponse semble donc spécifique de l'IR et est indépendante du gène MEC1.
- C3_gasch, C9_gasch, C10_gasch (voir figures 4.2 (c),(d) et (e)) : les expressions moyennes des gènes de ces trois classes sont réprimées par les deux types de stimuli. Mais alors que les profils de répression reviennent au niveau de base dans la condition d'IR, ils se maintiennent dans la condition MMS. De plus, ces classes ne sont sensibles au MMS uniquement dans la souche sauvage que les réponses à l'IR ne sont pas affectées par la mutation de MEC1. Ces classes représenteraient des gènes sensibles

Souches	Traitement	5 min	10 min	15 min	20 min	30 min	45 min	60 min	90 min	120 min
wt	MMS	x		x		x	x	x	x	x
mec1	MMS	x		x		x	x	x	x	x
wt	IR	x	x		x	x	x	x	x	x
mec1	IR	x	x		x	x	x	x	x	x
wt	MOCK	x				x		x	x	
mec1	MOCK	x				x		x		

Table 4.1 – *Description des données transcriptomiques de Gasch et coll. [60] utilisées par XRegPath. Sont décrits : souche mutée ou non pour le gène MEC1, le type de traitement appliqué aux souches de levures et les temps de mesures des expressions des gènes en minutes.*

aux deux types de stimuli, mais chaque stimulus déclencherait une voie de réponse différente avec l'implication du gène MEC1 dans la régulation de la réponse au MMS.

- C4_gasch (voir figure 4.2 (b)) : les gènes de cette classe répondent positivement aux deux stimuli mais avec des profils de réponse moyens différents. Les réponses aux deux stimuli seraient sous le contrôle de mécanismes de régulation différents mais tous deux indépendants du gène MEC1.

Une des réponses les plus marquantes, parmi les résultats que nous avions obtenus avec les données produites par l'équipe de Marie Dutreix à l'Institut Curie, révélait une induction de l'expression de gènes impliqués dans différentes étapes de la biogenèse des ribosomes (voir article II page 141). Nous avons recherché parmi les classes obtenues avec les données de Gasch *et coll.* si l'on retrouvait la même réponse ribosomale (classe C15_curie, 198 gènes). Les gènes de la classe C15_curie se retrouvent majoritairement dans 3 classes de gènes sensibles à l'IR (33 gènes dans C3_gasch, 22 gènes dans C9_gasch et 44 gènes dans C10_gasch). De plus l'analyse statistique de la représentativité des termes *GO* dans 3 classes C3_gasch, C9_gasch et C10_gasch indique exclusivement un enrichissement en termes en rapport avec le métabolisme des ribosomes. Cette observation peut expliquer la similarité des profils moyens de ces trois classes dans toutes les conditions, même si l'on note des différences de formes et d'amplitudes. Nous retrouvons ainsi une réponse ribosomale à l'IR à partir d'un autre jeu de données indépendant. Le sens de la modulation de l'expression moyenne des classes C3_gasch, C9_gasch et C10_gasch est cependant inversé par rapport à celui de la classe C15_curie et indique une répression des gènes impliqués dans la biogenèse des ribosomes après IR. Ces résultats semblent contradictoires mais cette différence dans le sens des réponses pourrait être dû à une différence entre les 2 protocoles d'irradiation : Gasch *et coll.* soumettent leurs cellules à l'IR durant 20 minutes alors que l'équipe de l'Institut Curie ne soumettent leurs cellules à l'IR que durant 20 secondes. Cette différence de temps d'irradiation pourrait être à l'origine de comportements différents des cellules de levures comme nous l'avons discuté dans l'article II (page 141).

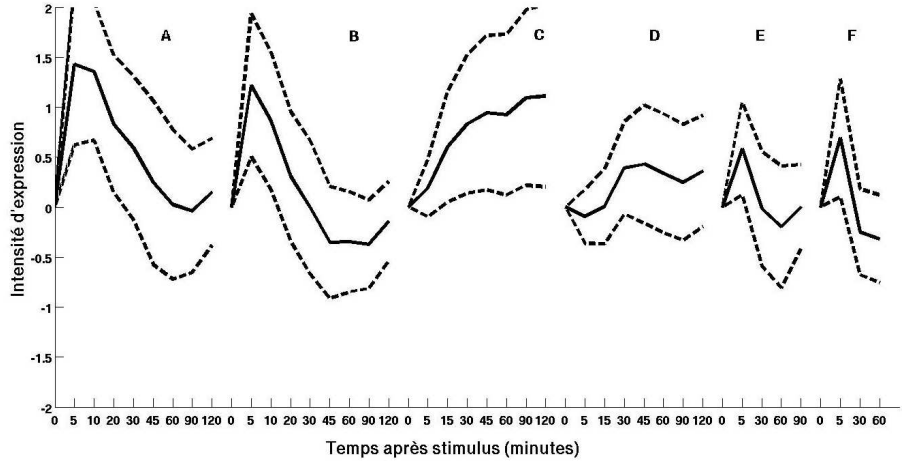
L'application de la première étape de notre méthodologie aux données de Gasch *et coll.* nous a permis de confirmer la stabilité et l'efficacité de notre algorithme de classification spectral à base de noyaux. Nous avons pu également retrouver une réponse à

l'IR impliquant le métabolisme des ribosomes même si le fait que les modulations soient de sens opposés prête à discussion. Enfin, nous avons pu démontrer l'efficacité de notre approche lorsque l'on combine des perturbations externes (IR et MMS) à des perturbation internes (ici, une seule mutation). En effet, nous avons réussi à identifier des influences spécifique à chaque type de perturbation sur le comportement des gènes. Notre approche permet donc d'extraire, en une seule étape de classification, des liens de régulation à partir de cinétiques d'expression de gènes de tailles différentes et mesurées dans des conditions expérimentales caractérisées par différentes combinaisons de perturbations.

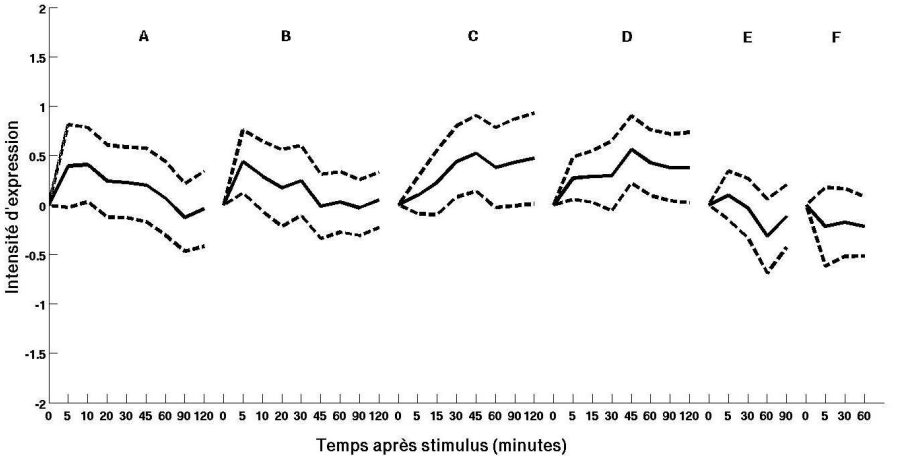
4.1.3.2 Prédiction d'interactions protéiques au sein de groupes de gènes co-exprimés

Certaines méthodes d'apprentissage proposent de prédire d'éventuelles interactions manquantes entre protéines au sein de réseaux supposés incomplets. P. Geurts et F. d'Alché-Buc, en collaboration avec L. Wehenkel, [166] ont récemment proposé une nouvelle approche appelée *Output Kernel Trees* (*OK3*), qui permet d'apprendre, à partir d'interactions existantes, un arbre de décision permettant de prédire, à partir de différents types de données génomiques, s'il existe ou non une interaction entre deux protéines. Nous apprenons un modèle à partir d'un ensemble d'interactions connues chez la levure et en utilisant comme données d'entrée des cinétiques d'expression, les localisations subcellulaires des protéines et des données phylogénétiques.
Cette méthode a été appliquée aux résultats obtenus avec l'approche *XRegPath* pour essayer de prédire au sein des groupes de gènes co-exprimés répondant à l'IR un réseau d'interaction entre protéines (voir article IV en annexe, page 203). Nous illustrons cette application avec un groupe de 198 gènes dont la co-expression est induite après IR. Ce groupe est particulièrement intéressant car environ 75% de ses gènes interviennent dans la biogenèse des ribosomes et, nous y avions identifié 5 sous-classes de gènes impliquées dans des processus de synthèse ou de régulation distincts. Les interactions protéines-protéines connues au sein de ce groupe ne concernaient que 60 gènes. Nous avons réussi à compléter ce réseau d'interaction pour arriver à connecter 130 gènes. En collaboration avec Marie Dutreix, nous avons analysé le réseau prédit et remarqué que les zones fortement connectées correspondaient aux 5 sous-classes de gènes identifiées par *XRegPath*. De plus, nous avons réussi à réintégrer au réseau d'interaction des gènes que nous n'avions pu prendre en compte dans notre analyse faute d'informations disponibles lors de l'intégration des données biologiques additionnelles à l'étape 4.

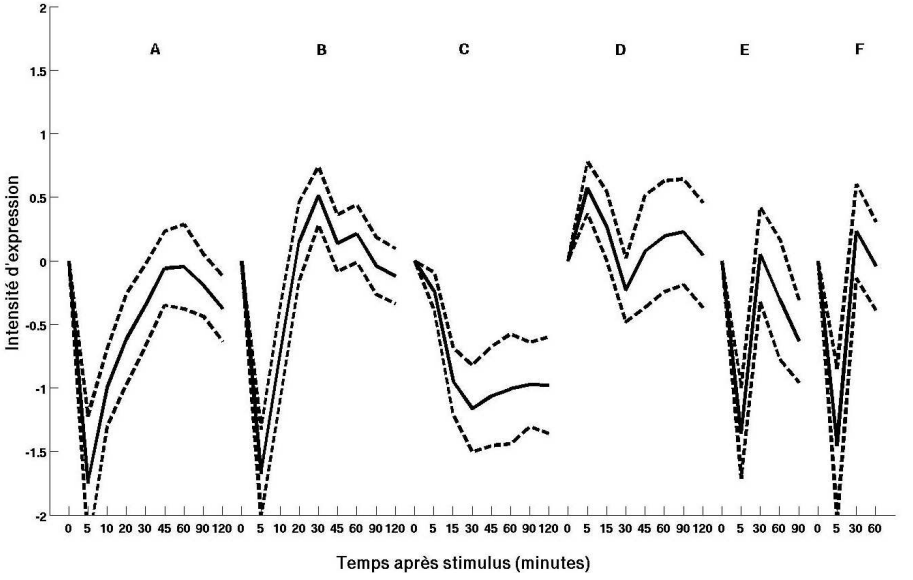
Dans notre cadre de recherche, l'approche *OK3* peut être utilisée en complément de notre approche *XRegPath* afin de réintégrer à l'analyse biologique des groupes de gènes co-exprimés des gènes que nous avions exclus faute d'informations suffisantes. L'approche *OK3* peut ainsi étendre notre méthodologie en proposant une méthode de validation, de complétion et d'interprétation de réponses à un stimulus indépendante de notre approche par biclustering.



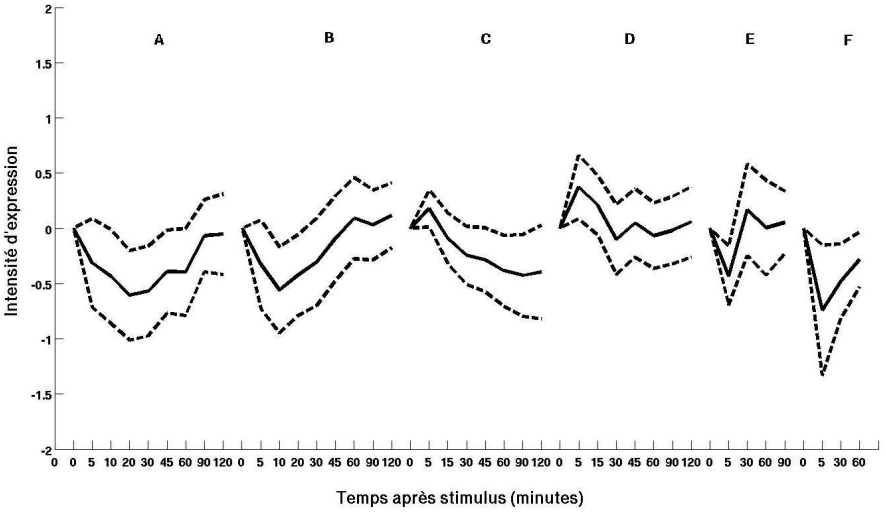
(a) Profils d'expression moyens du groupe C1_gasch (117 gènes).



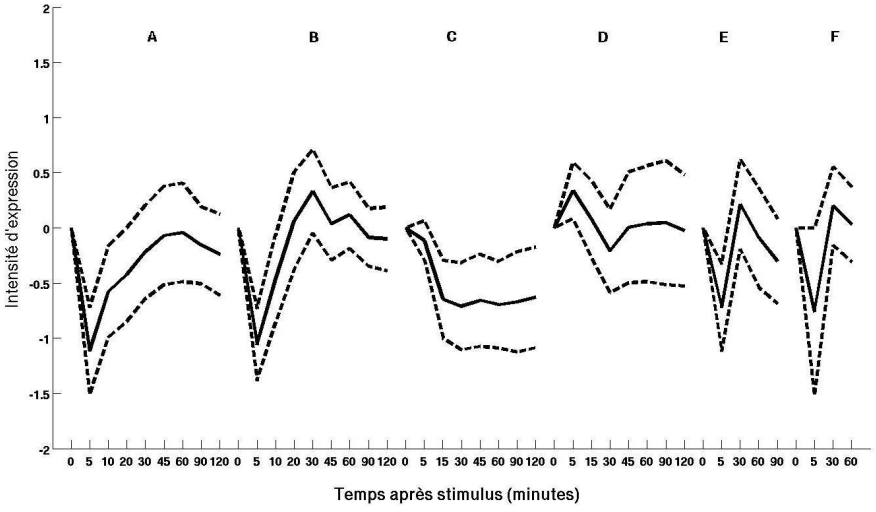
(b) Profils d'expression moyens du groupe C4_gasch (345 gènes).



(c) Profils d'expression moyens du groupe C3_gasch (75 gènes).



(d) Profils d'expression moyens du groupe C9_gasch (216 gènes).



(e) Profils d'expression moyens du groupe C10_gasch (176 gènes).

FIGURE 4.2 – Profils d'expressions moyens des groupes de gènes sensibles à au moins une condition de stress. Nous avons représenté chaque groupe par son profil moyen encadré par les écarts-types dans chaque condition expérimentale. A à F : conditions expérimentales de l'étude (A : wt + IR, B : *mec1* + IR, C : wt + MMS, D : mec1 + MMS , E : wt + MOCK , F : *mec1* + MOCK). Intensité d'expression : log2 des données d'expression (voir [60]. En ligne continue : profil d'expression moyen du groupe de gène dans la condition considérée. Ligne hachuré : profil moyen +/- l'écart-type.

4.1.4 Implémentation logicielle de la méthodologie *XRegPath*

Nous avons d'abord implémenté la méthodologie *XRegPath* sous la forme d'un ensemble de programmes codés dans des langages différents (*Matlab*, *JAVA*, *Prolog* et *LISP*). Ces programmes étaient utilisés de façon séquentielle avec les interfaces d'outils existants déjà (*Cytoscape* [318], *TreeView* [319] et *GeneMerge* [320]). Puis est né le besoin d'intégrer ces différents programmes et outils au sein d'une seule plate-forme d'analyse et d'inférence de manière à offrir un outil complet et facile d'utilisation. Cette intégration a posé différents problèmes, dont l'harmonisation et l'adaptation des formats des différents types de données utilisés par la méthodologie. Le développement de *XRegPath* s'est déroulé en deux étapes et a donné lieu à deux logiciels différents mais complémentaires. Le premier logiciel appelé du nom de la méthodologie, *XRegPath*, implémente les étapes 1, 2, 4 et 5 de la méthodologie (voir figure 4.1). Le deuxième logiciel appelé *XRegRules* implémente simplement l'étape 3, qui correspond à l'étape de déduction automatique de règles de régulation.

4.1.4.1 Logiciel *XRegPath*

Le cahier des charges de cet outil a été conçu en collaboration avec Florence d'Alché-Buc et Farida Zehraoui. Farida Zehraoui et moi même avons supervisé 11 étudiants en master professionnel en informatique pour le développement de cet outil. Cet outil offre les fonctionnalités suivantes (voir en annexe, page 216 pour les détails d'implémentation et les différents scénarios d'utilisation) :

- Gestion de l'information biologique : toutes les informations biologiques exploitées dans ce projet sont stockées dans une base de données relationnelles dont le 'gène' est l'entité principale. Notre outil propose un système d'interrogation par requête qui permet d'accéder aux différents types de données et qui permet de sélectionner des listes de gènes en fonction de leurs différents attributs biologiques.
- Outils de fouille de données : ce logiciel implémente les deux types d'algorithmes de classification (*clustering* spectral et *biclustering* spectral) utilisés par la méthodologie *XRegPath*. A ces outils sont évidement associés le calcul des similarités entre gènes et la sélection des descripteurs biologiques associés à des groupes de gènes co-exprimés basée sur un test statistique.
- Représentation de l'information : a chaque étape d'analyse, l'information inférée est visualisée par une représentation adaptée (figure ou tableau) et peut être sauvegardée sous la forme de fichier texte ou d'image.
- Interface d'utilisation : toutes les fonctionnalités de l'outil sont accessible à partir d'une seule interface graphique ergonomique qui permet soit de dérouler une analyse de façon séquentielle en suivant l'enchaînement des différentes étapes proposées par notre méthodologie, soit d'accéder directement à la fonctionnalité d'une étape particulière et de l'utiliser indépendamment des autres étapes.

4.1.4.2 Logiciel *XRegRules*

J'ai conçu l'outil *XRegRules* et ai encadré son développement par Mohamed Sghaier Ben Hammadi, étudiant en 1ère année de master informatique. Cet outil correspond à l'im-

plémentation de l'étape 3 de notre méthodologie *XRegPath* en un logiciel indépendant. Notre principal objectif était de fournir aux biologistes un outil qui les aide à formaliser et à écrire une stratégie expérimentale de type perturbation génétique en une base de connaissance *Prolog*, et qui leur permette de tester par eux-mêmes les règles de régulation de la base sans avoir à utiliser un environnement de programmation en "ligne de commande". Notre deuxième objectif consistait à développer cet outil en tant que module potentiellement intégrable au logiciel *XRegPath* qui ainsi pourrait proposer un déroulement complet de notre méthodologie.
L'outil *XRegRules* propose les fonctionnalités suivantes (voir en annexe, page 218 pour les détails d'implémentation et les différents scénarios d'utilisation) :

- Aide à l'écriture de bases de connaissances : l'outil permet le chargement de toute base de connaissances écrite en langage *Prolog*, il permet aussi l'écriture "en ligne" d'une nouvelle base de connaissance. Mais, la principale innovation de *XRegRules* consiste à aider l'utilisateur dans l'écrire d'une base de connaissance en permettant de générer de façon automatique une partie de cette base à partir d'observations expérimentales.
- Extraction de règles de régulation : à l'aide d'une interface de saisie l'utilisateur peut soit évaluer une règle particulière soit évaluer en une seule fois une liste de règle ou la base de connaissances dans sa totalité.
- Visualisation des résultats : à partir d'une seule interface graphique, on peut accéder de façon indépendante aux différentes fonctionnalités de l'outil. Les résultats sont actuellement représentés sous la forme de tableaux ou de listes qu'il est possible de sauvegarder sous la forme de fichiers texte.

Actuellement l'outil *XRegPath* est en phase de test, et n'est pas encore accessible hors du laboratoire. L'outil *XRegRules* par contre est disponible et peut être facilement chargé et utilisé.

4.2 Article II : Une ré-orchestration du programme du métabolisme primaire révélée par l'analyse dynamique de la réponse transcriptionnelle de la levure aux radiations ionisantes.

4.2.1 Présentation de l'article

Nous avons déjà publié une analyse de la réponse transcriptionnelle à l'IR chez la levure (voir article III en annexe, page 193) en exploitant des cinétiques d'expression de gènes mesurées dans trois souches de levure (une souches haploïde de *mating-type* a, une souches haploïde de *mating-type* α et une souche diploïde de mating type a/α). Nous avions identifié une réponse à l'irradiation au sein de chaque souche de levure en sélectionnant les gènes dont les cinétiques d'expression présentaient à deux temps d'analyse au moins une modulations d'expression deux fois plus grandes ou deux fois plus petites que le niveau d'expression avant IR. Puis, nous avions déterminé les sous-ensembles de gènes spécifiques ou communs aux 3 souches à l'aide d'un diagramme de Venn.

Nos observations nous ont permis d'identifier une réponses à l'irradiation spécifique des souches haploïdes et dont la régulation semblait impliquer les facteurs du *mating-type*. Nous avions aussi observé une sur-représentation de gènes répondant à l'IR localisés en positions subtélomériques, régions des chromosomes dont la transcription est connue pour être réprimée par une hypercompaction de la structure chromatinienne. Ce résultat nous faisait soupçonner un possible rôle de la chromatine dans la régulation de la réponse à l'IR.

Dans l'article précédent les influences des caractéristiques des différentes conditions expérimentales (mutations et variations du *mating-type* et de la ploïdie) ne pouvaient être mise en évidence que par des intersections de groupes, une méthode inefficace au delà de 3 conditions et difficile à interpréter. De plus, l'analyse statique basée sur un seuil arbitraire ne permet pas de prendre en compte l'aspect dynamique des cinétiques et néglige donc une grande partie de l'information produite par les puces à ADN. L'application de l'approche *XRegPath* peut être vue comme une évolution dynamique, automatique et enrichie de notre précédente approche. Dans l'article II présenté à cette section, nous avons repris l'ensemble des données précédentes et les avons enrichies avec de nouvelles séries de cinétiques d'expression puis, nous les avons exploitées par une méthode globale afin d'en raffiner l'analyse.

L'application de la méthodologie *XRegPath* à l'analyse globale de la dynamique de la réponse transcriptionnelle dans 7 souches de levure nous a d'abord permis d'identifier 8 réponses distinctes avec différentes caractéristiques fonctionnelles et différents régulateurs. Cette méthodologie nous a permis de construire un modèle de régulation global qui associe à la fois des modifications de la structure chromatinienne et l'influence de facteurs de transcription pour le contrôle de l'expression des gènes après IR. Notre analyse a révélé l'existence d'une modulation globale de l'expression du métabolisme primaire après irradiation chez la levure. Cette modulation apparaissant immédiatement après l'irradiation. Certaines de nos observations ont pu être validées expérimentalement et confirmer certaines relations au sein de notre modèle de réponse à l'irradiation.

4.2.2 Primary metabolism program re-orchestration revealed by dynamic analysis of yeast transcriptional response to ionizing radiation. Nizar Touleimat, Maria Quanz, Nathalie Berthault, Farida Zehraoui, Florence d'Alché-Buc and Marie Dutreix.

REORCHESTRATION OF THE PRIMARY METABOLISM PROGRAMME IN RESPONSE TO IRRADIATION REVEALED BY DYNAMIC ANALYSIS OF THE YEAST TRANSCRIPTIOME

Nizar Touleimat[1,2,4,5], Maria Quanz[1,2,3], Nathalie Berthault[1,2], Farida Zehraoui[4,5], Florence d'Alché-Buc[4,5] and Marie Dutreix[1,2, 3*].

[1]Institut Curie, Centre de Recherche, Orsay, F-91405, France; 2CNRS, UMR2027, Orsay, F-91405, France; [3]Institut Curie, Hôpital, Département de transfert, Orsay, F-91405; [4]Informatique, Biologie Intégrative et Systèmes Complexes, Université d'Evry-Val d'Essonne, Evry, F-91000; [5]CNRS, FRE3190, Evry, F-91000.

Keywords: response to irradiation, expression kinetics, ribosomes, chromatin

Running title: Irradiation-induced metabolism reorchestration

*** corresponding author**

Abstract

The deleterious genetic consequences of high doses of gamma radiation are well known, but the cellular response triggered by such damage remains little characterised or understood. We carried a global analysis of transcriptional dynamics in yeast, monitoring gene expression in different *S. cerevisiae* genetic backgrounds during the cellular response to irradiation. We used a new, generic method, XRegPath, based on inductive data mining to extract and mine patterns of gene expression and regulatory pathways in seven genetic backgrounds. We observed rapid and extensive modification of the transcriptional programme after irradiation, mostly involving genes related to primary metabolism. Carbohydrate and amino-acid metabolism genes were repressed, whereas the synthesis of all components of ribosome was rapidly induced by irradiation. These observations were confirmed *in vivo*, through monitoring of the derepression of the rRNA locus, and by estimating the ribosome content of cells with a GFP-tagged Rpl11b ribosome subunit. The increase in the number of ribosomes per cell was correlated with the increase in cell size following irradiation.

Introduction

Gamma radiation induces many types of damage to nucleic acids and other cell components. These damages trigger cell cycle arrest, DNA repair and, if unrepaired, cell death. All living species have evolved systems of fine regulation to control the chronological occurrence of these events. Complex surveillance mechanisms monitor genomic integrity during cell-cycle progression, and orchestrate the response to DNA damage. Several global analyses have shown that a plethora

of responses other than DNA damage repair and cell cycle control are induced by DNA damage (1) and are important for cell recovery. Indeed, it now seems inappropriate to consider irradiation as simply damaging DNA, as free radicals — a bystander product of water ionisation by radiation — are known to damage proteins, RNA, carbohydrates, lipids and many other cell components. Very little is known about the signalling and regulation of such damage. One of the major difficulties involved in studying the global response to irradiation is the very large number of genes displaying a change in transcription (about a third of the genome), making it very difficult to infer regulatory mechanisms from transcriptional data.

The signal inducing the response is probably triggered by damage to the chromosome or the disruption of replication in the presence of unrepaired damage. We therefore analysed the transcriptional response to irradiation in two mutant strains (*hdf1* and *rad52*) defective in the two main DNA double-strand break repair pathways (the Non-Homologous-End-Joining and the Homologous Recombination). We previously demonstrated the derepression of silenced subtelomeric genes by irradiation, in a study of mostly wild-type strains (2). Heterochromatic structure has been shown to be lost at telomeres in response to DNA damage (3,4). We investigated this phenomenon further by analysing transcriptional responses in a *sir3* mutant defective in chromatin remodelling in silent regions of the chromosome. The data obtained with these mutants were processed, together with other laboratory-generated data, through an original method (XRegPath) that clusters genes according to their expression kinetics in many independent experiments. Integration of all the information gathered in the various steps of our method led to the reconstruction, for each cluster, of the system regulating the response of *Saccharomyces cerevisiae* to irradiation (IR).

We identified eight different IR responses characterised by different functional or regulatory features. We constructed a global regulatory scheme combining changes to chromatin structure and transcription factor effects in the control of gene expression after irradiation. This analysis

highlights global changes in primary metabolism occurring immediately after irradiation. We present additional genetic and molecular findings supporting these results.

MATERIALS AND METHODS

Strains and biological experiments

The wild-type *Saccharomyces cerevisiae* strains used were the diploid FF 6053 (Mat*a*/α), and the haploids FF 18734 (Matα), FF 18733 (Mat*a*), FF 18735 (Mat*a*(α)). The FF 6053 diploid was obtained by mating the two haploids (FF 18734 and FF 18733), as described by Mercier *et al.* (2) The mutant strains used were FFD1 (Mat*a*, *rad52*), LM79 (Matα, *hdf1*) (5) and yIBPC205 (Matα, *sir2*). rRNA gene silencing was assessed in the Yg177 (Mat*a*) strain, which carries an m*URA3-HIS3* insertion downstream from the 5S rRNA gene. Yg177 cells with a repressed rRNA locus were selected based on their growth on two media: SC-rich medium lacking uracil and SC medium containing 5-fluoro-orotic acid (toxic to cells containing URA3, (6)). Repression was assessed as the ratio of colony forming units (cfu) on SC+5-FOA medium to that on SC-medium. Mutants (*ura3*) were excluded by replica-plating on SC–Ura medium. Fluorescent GFP-fusion (Mat*a*) strains 10D10 (*RRN3-GFP*), 32E9 (*RPL11B-GFP*) and 09C4 (*SFP1-GFP*) were obtained from the collection described by Huh *et al.* (7). We used DAPI staining, microscopy and FACS analysis, as previously described (8), to determine the duration of cell cycle arrest following irradiation.

Ionising irradiation conditions, time courses and the collection of microarray data

Overnight cultures in the exponential growth phase were centrifuged. The cell pellet was resuspended at a density of 10^9 cells/ml and irradiated (60 Gy/min with a ^{137}Cs source) at room temperature in rich medium to minimise temperature and osmotic variations during treatment. Irradiated cells were immediately resuspended in prewarmed rich medium, at the original density, for time-course experiments. Cells were irradiated with 200 Gy at time 0, and samples were

collected for microarray and cell cycle analysis at various times (0.1, 1, 2, 3, 4 and 5 h) after irradiation. Total RNA was extracted from frozen samples by the hot phenol method and Cy-5-labelled as described by Mercier (9). A fluorescent Cy-3-labelled control cDNA population was prepared from the same pool of total RNA extracted from five independent, exponentially growing cultures of the diploid strain (FF 6053). RNA was hybridised with microarrays, which were then scanned with a Genepix 4000B machine (Axon Instruments). Fluorescence intensities for all spots were normalised using the location and scale normalisation procedures described by Mercier *et al.* (2). Measurements on non-irradiated cells were carried out in triplicate.

Real-time PCR analysis

cDNA was synthesised from the extracted total RNA by incubation with SuperScript II (Invitrogen). Real-time PCR was carried out with an iCycler thermal cycle and the iQ SYBR Green Supermix (Biorad, Hercules, USA). The primers used for real-time PCR were designed using the *Saccharomyces* Genome Database (http://www.yeastgenome.org) and Oligo 6.4 software (Molecular Biology Insights, Cascade, USA). Oligonucleotides were synthesised by Eurogentec (Seraing, Belgium). The primers used to detect the expression of *RPC31*, *RPA49*, *RPB8*, *RPC40*, *NAP1* and *ACT1* are shown in Table S1 (supplementary material). The PCR profile consisted of an initial activation and denaturation step at 95°C for 1.5 minutes, followed by 60 cycles of 95°C for 15 s and 60°C for 1 minute. A dissociation curve was plotted to confirm that the signal was free of primer effects for data analysis. Each point was duplicated and a negative control, in which H_2O replaced cDNA, was used to check that the mixture was not contaminated. A standard curve was produced from cDNA samples from non-irradiated cells, pooled as a function of time point. Four different concentrations were used in the real-time PCR to generate a standard curve for each gene. The threshold cycle (Ct) value was determined with iCycler software. Gene copy number was calculated from the standard curve. A ratio was calculated by dividing the fold-change in

expression of the target genes (*RPC31*, *RPA49*, *RPB8*, *RPC40*, or *ACT1*) in the irradiated cells with respect to non-irradiated cells by the fold-change of *NAP1* expression in the irradiated cells with respect to the non-irradiated cells. *NAP1* was used as the reference gene because it displayed the lowest level of variation in all microarray data obtained in various conditions in the laboratory (our own unpublished results). Two independent extracts were analysed and each real-time PCR was carried out in duplicate. Results are expressed as means and standard deviations.

Gene clustering

We took into account the dynamic features of the data, by applying a clustering strategy based on a Gaussian similarity measure (the kernel method) of gene expression kinetics rather than simple Euclidean distances. This similarity measurement made it possible to compare kinetic curves on the basis of their shapes rather than their amplitudes. We calculated local similarity, within a given genetic background, between the kinetic curves for instantaneous derivatives. We then generated a linear combination of these local values to determine the global similarity between genes. We applied a spectral clustering algorithm (10) to the global similarity values, to group genes as a function of their transcriptional response in different genetic backgrounds. The optimal number of clusters was determined to be 22 — this number giving the most stable partition using a bootstrap variant based on that inspired from B Ben-Hur *et al.* (11). Adding more clusters did not affect the main partition, but did result in the generation of new clusters containing very small numbers of genes (less than ten).

Biclustering data sources

Biological descriptors were extracted from data banks. We used functional annotations describing metabolic pathways (KEGG (12)) or biological processes (Gene Ontology, (13)) (GO descriptors), chromatin immunoprecipitation (ChIP) data (14) giving the list of interactions between

transcriptional regulators (203 TFs) and gene promoters (TFs descriptors), the list of the 2703 proteins interacting in 546 complexes (15) (complex descriptors), and the chromosomal positions of the genes (*Saccharomyces* Genome database: ftp://ftp.yeastgenome.org/yeast/). Two different annotations for the chromosomal positions of genes were used: we associated each gene with the chromosome arm on which it is found and we annotated as "subtelomeric" genes located less than 20 kb from chromosome telomeres (genes in these regions are likely to be controlled by chromatin silencing at telomeres).

Biclustering method

With the aim of identifying significant biological functions or regulatory mechanisms common to a set of co-expressed genes, we performed statistical analyses of each cluster for the biological descriptors defined by the data source. We generated statistically based rank scores using the hypergeometric distribution to detect significant descriptor over-representation (defined by a P-value < 0.05) in each co-expressed gene group. We took into the account the hierarchical tree structure of GO, by correcting the P-value of GO descriptors as described by Sokal (16). The associations between selected descriptors and genes were represented as a binary matrix, with a '1' for association and '0' for non-association. The matrix was reorganised using a biclustering strategy (17) based on the density of '1' and '0' entries. A Treeview (18) representation of the reorganised matrix made possible the visual detection of sets of genes, called biclusters, associated with a similar subset of biological descriptors.

RESULTS

1 The principle of the XRegPath approach

We used an original strategy based on statistical and logical inference to identify sets of co-expressed genes displaying significant changes in expression in response to irradiation that are

dependent on the same regulation mechanisms. Genes were first clustered as a function of similarities in expression profile (see Materials and Methods). This method takes into account the similarities between genes within a given genetic background and adjusts the clustering to optimise global similarity. We then generated a mean kinetic profile for each cluster, to facilitate the analysis by attenuating the fluctuation of expression for individual genes. We used the differences between mean profiles in various yeast strains to deduce potential regulators according to the genetic properties of each genetic background. Biological information was then deduced by integrating various systemic data sets into co-expressed gene groups by a biclustering strategy (see Materials and Methods). Information as diverse as functional ontologies, interactions of transcription factors (TFs) with promoters, protein complex formation and the chromosomal location of genes was used for biclustering. Another Potential regulators of IR-sensitive clusters were identified by deduction based on the genetic properties of each strain and the mean expression profiles associated with clusters. These deductions may be made directly if only a few combinations of genetic variables are involved. However, the use of an automated tool becomes worthwhile if several genetic properties vary. This deduction process is automated in XRegPath, through the development of explicit logical rules and the use of a dedicated computational language such as Prolog. The complete methodology method allowed to get an integrated vision of the changes to the cell occurring in response to irradiation, deepening our understanding of the global regulation of this response) (Figure 6).

2. Identification of eight clusters of genes responding differently in the different strains

We analysed the IR response in each strain by monitoring global gene expression during the period of recovery from irradiation. The duration of this period was determined by flow cytometry analysis of the cell cycle arrest following irradiation with 200 Gy. In all strains analyzed, the irradiated cells were blocked in the G2/S phase for 1 to 2 hours after irradiation and cell division

did not resume until at least five hours later (illustrated in Figure 1, for the ORD18733 strain). Samples were taken throughout this period, at 0, 0.5, 1, 2, 3, 4 and 5 hours after irradiation, for RNA extraction and transcriptome analysis. We applied the XRegPath method to a selected set of 4732 genes for which complete expression kinetics data were available for all seven yeast strains tested. We obtained 22 clusters, each comprising 101 to 320 genes (with the exception for cluster C8, which comprised only 25 genes). Each cluster is represented as its mean profile (with standard deviation) in each yeast strain, for the kinetics of the genes it contains. An analysis of these mean expression profiles led to the identification of three types of behaviour. First, we identified 8 homogeneous clusters with mean expression profiles changing after irradiation in at least one yeast strain, comprising 1389 irradiation-modulated genes (Figure 2; Table 1). The second type of behaviour corresponded to 12 clusters with flat profiles in all the yeast strains, comprising 3033 genes considered insensitive to irradiation (data not shown). Two clusters (310 genes) displayed a third kind of behaviour, in which mean profiles were framed by large, irregular standard deviations, reflecting highly variable expression profiles, the fluctuations of which were interpreted as unreliable (data not shown). The number of irradiation-modulated genes was consistent with previous estimates obtained through threshold-based data selection (19,20). Our IR-modulated genes included 1144 that were globally induced and 245 that were globally repressed. These genes were clustered as a function of their expression kinetics and their characteristics in each genetic background were summarised (Table 1). A rapid overview showed an absence of behaviour inversion for a modulated cluster in different yeast strains after irradiation: none of the clusters was up-regulated by after irradiation in a one strain and down-regulated in another. This observation attests to the quality of our data and of our clustering approach. Current knowledge of the mechanisms regulating transcription suggests that it is unlikely that a given cluster will show opposite responses to irradiation in different genetic backgrounds.

Observation of the mean kinetics in each cluster revealed that most of the profiles were very simple, with a rapid change occurring after irradiation, before cell cycle arrest, and a slow return to basal levels coinciding with resumption of the cell cycle (Figure 1, 2). However, the two C15 and C22 clusters showed a very slow return to basal levels, with expression levels remaining strongly induced and repressed, respectively, five hours after irradiation. C10 and C18 differed from the other clusters in displaying maximal induction at late time points (4 h and 3 h, respectively) and being induced in only one (C18 in the diploid FF6053 strain) or two (C10 in the *sir2* and *hdf1* mutants) strains. The precise shape of the mean expression curve for a cluster was conserved in the various strains displaying expression changes, demonstrating the specificity of the profile for the group of genes observed. Interestingly, the C22 and C15 clusters seemed to have very similar profiles, in the various strains, despite displaying opposite patterns of regulation. These clusters displayed no response to IR in the *sir2* and *hdf1* mutants and a very weak response in the diploid (FF6053) strain. The *rad52* mutation did not affect the IR response of the various clusters.

The induction of cell cycle arrest by irradiation suggested that part of the response might be specific to the S/G2 phase. We investigated whether the initiation and resumption of the response phase were due to the IR or cell cycle arrest, by projecting the modulated genes, cluster by cluster, onto the raw data set of Spellman *et al.* (21). These authors performed a comprehensive series of experiments with the aim of identifying all protein-encoding transcripts in the genome of *S. cerevisiae* regulated by the cell cycle. They used DNA microarrays to analyse mRNA levels in cell cultures synchronised by three independent methods. None of our eight IR-modulated clusters displayed periodic expression when projected onto each of the data series of Spellman *et al.* (see supplementary figures S1). Thus, the correlation between IR responses and the cell cycle reflects the irradiation response rather than a regulatory dependence on the cell cycle.

3. Role of chromatin rearrangement

The C2, C10, C15 and C22 clusters, containing 54 % of the modulated genes, were found to be sensitive to *hdf1* and *sir2* mutations (Table 1). These results point reveal a new regulation pathway involving both the *sir2* and *hdf1* genes. However, both the proteins encoded by these genes seem to play an important role in regulating a large number of genes. These two proteins have been shown to be associated with silencing at subtelomeric positions (ref) and sir2 is involved in chromatin silencing at various chromosome loci. We therefore used XRegPath to search for a potential correlation between the location of genes and their co-expression within the same cluster. We found that cluster C2 was enriched in subtelomeric genes (21 subtelomeric genes, *P-value* = 2.41 x 10^{-11}). We also identified six clusters containing co-expressed genes for which locations on chromosome arms (1L, 2R, 4L, 6L, 7L, 7R, 8R, 9L, 10L, 10R, 11L) were overrepresented (Supplementary Figures S3). We investigated whether this overrepresentation on chromosome arms reflected the existence of physical gene clusters by applying a positional gene enrichment analysis method implemented in the PGE web tool (22), taking into account position on the chromosome in addition to chromosome number. Significant overrepresentation on chromosome arms was demonstrated for four clusters (Figure 2): C2 (one region of 12 genes on chromosome 4L); C18 (one region of 15 genes on chromosome 10R); C10 (one region of 14 genes on chromosome 12L); C15 (one region of 9 genes on chromosome 12R) (Supplementary Figures S4).

The main mechanism regulating transcription in yeast is the binding of one or several transcription factors to promoters. Analysis, through the integration of ChIP data (15), of the clusters identified showed these clusters to be generally highly sensitive to *hdf1* and *sir2* mutations (C2, C10, C15, C22) and associated with very few TFs. For example, cluster C2 was not found to be associated with any TFs. In the other clusters sensitive to *sir2* and *hdf1* mutations, most of the TFs controlling the IR-modulated genes are known to be directly or indirectly involved in chromatin structure modification (Figure 3; Supplementary Figures S3). For example, Sum1 (C10)

is known to be involved in chromatin silencing at telomeres and the mating-type cassette (23) and Sko1 (C22) are involved in heterochromatin establishment in the subtelomeric region of chromosomes (24) and gene derepression (25,26). Cluster C15 contains genes associated with Abf1 that directly mediate various chromatin-related events, such as gene silencing (27,28) and chromatin remodelling (29). It also contains genes associated with three TFs: Rap1, Fhl1 and Sfp1. The Rap1 factor binds telomere sequences and is involved in chromatin silencing at telomeres and in the maintenance of telomere structure. Fhl1 and Sfp1 are known to be recruited by Rap1 after its delocalisation from telomeres. The regulatory effects of Fhl1 and Sfp1 may therefore coincide with changes in chromatin structure at telomeres.

4. Chromatin alterations and the induction of RNA polymerases I and III allow rapid changes in ribosome biosynthesis and general metabolism

The global analysis of the IR response presented here, together with other published data, indicates that about 30% of yeast genes display changes in expression after irradiation. It is therefore extremely difficult to identify the biological processes affected by irradiation. The subdivision of these genes into subgroups of co-regulated genes provided a unique tool for addressing this question, making it possible to integrate many different sources of genomic information into the analysis of each cluster.

Biological mining of the C15 and C9 gene clusters highlighted showed that genes involved in ribosome synthesis after irradiation were induced in these clusters. Within C15, we identified five different, but related, gene biclusters (a bicluster is a subgroup of genes related to a set of descriptors overrepresented in the cluster), all involved in several steps of ribosome biogenesis (142 genes, 71.72% of C15) (Figures 2 and 3). The importance of ribosome biogenesis in this gene cluster was confirmed by the association of the C15 bicluster of ribosomal proteins with the four TF descriptors (Abf1, Fhl1, Rap1 and Sfp1) in the Chip data analysis performed by Krogan *et al.*

(15) (Figures 2-3). These TFs are involved in regulating the expression of genes involved in ribosome biogenesis (30). One C15 bicluster was found to contain almost all the subunits of RNA polymerases I (Pol I) and III (Pol III) (Table 2). Whereas C15 seems to involve all the genes involved in ribosome biogenesis, C9 displays specialisation in rRNA processing (17 genes). This specialisation is supported by two other C9 gene subgroups, one relating to general RNA metabolism (52 genes) and the other to RNA export from the nucleus and ribosomal protein import into the nucleus (12 genes) (Figure 3).

The IR response involves both the induction of ribosome biogenesis and the repression of carbohydrate metabolism (23 genes in C22; Figure 3). Consistent with the findings for GO descriptors, biclustering with TF descriptors indicated that 14 C22 gene promoters bound Hsf1, a TF known to regulate the transcription of genes involved in carbohydrate metabolism. C8 genes (9 genes) were also found to be associated with the repression of carbohydrate metabolism and to bind Hsf1 and Msn2. The C8 and C13 clusters include genes involved in purine (12 genes) and pyrimidine nucleotide metabolism (5 genes; Figure 4). A more precise analysis of these genes showed that most were related to amino-acid metabolism. C13 pyrimidine genes are involved in the histidine biosynthesis superpathway (*ADE2, ADE6, ADE8, ADE12, UEA4, ADK1* and *YNK1*), three C13 purine genes are involved in tryptophan metabolism (*BNA1, BNA3* and *BNA4*) and three C8 purine genes are involved in histidine biosynthesis and serine biosynthesis (*MTD1, ADE1* and *SHM2*). Consistent with these results, we found that 24 genes from the C13 cluster (25%) were involved in amino-acid metabolism, including precise functional subgroups relating to histidine metabolism (5 genes), serine family amino acid and glycine metabolism (3 genes), methionine and selenoamino acid and aspartate metabolism (4 genes). C13 genes were also found to be associated with the repression of oxidative phosphorylation (4 genes). Biclustering revealed several associations between C13 genes and Met2, Met4, Met31, Nrg1, Gln3, Cbf1, Gcn4. These TFs are known to be involved in amino-acid metabolism, purine biosynthesis and glucose metabolism. In

conclusion, the three clusters displaying repression in response to irradiation seem to be biologically related and indicate a decrease in general metabolism levels in irradiated cells.

Cluster C18 contained a large number of genes encoding proteins located in organelles and involved in the post-translational modification, trafficking and secretion of proteins: SNARE-interacting proteins (5 genes), endoplasmic reticulum proteins (27 genes), organelle endomembrane system proteins (14 genes) and mitochondrial components (42 genes) — mostly mitochondrial inner membrane proteins (13 genes) and mitochondrial ribosomal proteins (17 genes).

Induction of ribosome biogenesis

The most unexpected result of our study was the discovery that ribosome biogenesis was induced following irradiation. This observation seems to be incompatible with the arrest of cell division and the consequently low requirements for the biosynthesis of new cell components. We monitored the concentration of ribosomes in single cells, using a fusion of GFP with Rpl11b (Figure 5a) to confirm this finding (31,32). Rpl11b is one of the 45 ribosomal proteins forming the 60D ribosomal unit, together with 5S, 5.8S and 25S rRNA. It is an essential component of the ribosome (33). Flow cytometry analysis confirmed that the Rpl11b-GFP content of cells increases in the first three hours after irradiation. This increase in the number of ribosomes was associated with an increase in cell size (Figure 5c-e). These results are consistent with various studies highlighting the direct correlation between cell size, cell cycle and ribosome biosynthesis in bacteria and yeast ((34); for more recent reviews, see (35-37)). This correlation was also recently confirmed by a global analysis of mRNA levels in various growth conditions (38).

Subunits of the Pol I and Pol III RNA polymerases, responsible for the transcription of ribosomal 35S rDNA genes and control 5S rRNA synthesis, respectively, are overproduced following irradiation, as shown by C15 cluster profile (Table 2). We used quantitative PCR to confirm the

expression induction of *RPC31, RPA49, RPB8* and *RPC40* in irradiated FF1833 cells. The expression of all polymerase subunit genes increased after irradiation (Figure 5a). The RPA49 gene, encoding an RNA Pol I subunit, was the most responsive. Its expression increased by a factor of 10 during the first 10 minutes after irradiation. High Pol I concentrations in cells have been shown to lead to the derepression of Pol II-mediated transcription, which is restricted to genes encoding ribosomal proteins. This observation is consistent with our microarray and flow cytometry data indicating an increase in ribosomal protein synthesis after irradiation. Furthermore, Laferte *et al*. (39) showed that 5S rRNA, which is synthesised by Pol III, is also induced in cells overexpressing Pol I. The increase in Pol III concentration should induce an increase in 5S rRNA synthesis. We assessed this induction by evaluating derepression of the silenced rRNA locus, using a *URA3* reporter gene inserted downstream from the 5S rRNA locus in the rDNA array (Figure 5c). In exponentially growing cultures, 15% of the cells have a fully silenced *ura3* gene, as demonstrated by the ability of these cells to grow on selective medium containing 5-FOA. Immediately after irradiation, only 7 % of the cells retained this ability, indicating that half of the cells in which *URA3* was initially silenced lost this silencing after irradiation (Figure 5b). This derepression may result from higher levels of Pol III activity and/or the chromatin remodelling induced by irradiation. Our data confirm previous observations indicating tight regulation to maintain equimolar levels of ribosomal components.

Many different regulatory mechanisms (the most prominent of which is the *TOR1* signalling system) have been implicated in the regulation of ribosome biogenesis (for review, see (40,41)). The results of Jorgensen *et al*. (42,43) are of particular relevance: they found a connection between ribosome biosynthesis and cell cycle entry at START (44) involving the regulation of both processes by the products of *SFP1* and *SCH9*. Both Sch9, which is usually found mostly at the periphery of the vacuole, and Sfp1, which is usually confined to the nucleus, are dispersed

throughout the cell in response to nutritional or diauxic stress. These changes in localisation result in the repression of ribosomal protein gene transcription. We monitored the subcellular distribution of Sfp1, which has been identified as particularly responsive to stress (45), to determine whether the growth arrest in irradiated cells was responsible for inducing the delocalisation of Sfp1 responsible for the repression of ribosome biogenesis (Figure S4). We found that, by contrast to growth arrest in the stationary phase, the arrest of cell division after irradiation did not induce Sfp1 relocation. We also analyzed the distribution of Rrn3, an essential initiation factor for RNA polymerase I. Rrn3 relocates from the nucleolus to the cytoplasm in response to diverse stress signals (46) (Figure S5). Consistent with our findings for Sfp1, Rrn3 dispersed into the cytoplasm in the stationary phase but remained in the nucleolus at all times monitored after irradiation.

DISCUSSION

Analysis of the biological processes involved in our IR-sensitive clusters suggested a global change in the primary metabolism programme. We observed a rapid repression of carbohydrate metabolism (C22), followed by the repression of amino-acid metabolism (C8, C13). These observations are consistent with the arrest of cell division. In parallel, we observed an induction of the expression of genes encoding proteins involved in the metabolism or biogenesis of many cellular components, including cytoplasmic ribosomes (C15, C9), mitochondrial ribosomes, mitochondria and endoplasmic reticulum membranes (C18). The induction of new organelle synthesis in non-dividing cells suggests that renewal of the cell components damaged by IR may occur. However, it remains unclear how this switch in the transcription programme is regulated.

Several observations suggest that chromatin remodelling may partly account for the dynamics of IR responses. Indeed, three clusters (C2, C15 and C22) displayed a rapid response that was abolished in *sir3* and *hdf1* mutants. The *sir3 and hdf1* genes are involved in the silencing of subtelomeric genes g (2), which are overrepresented in C2. Sir3, together with Sir2 and Sir1, is involved in the silencing of various loci, including the mating type locus and the rRNA area on chromosome VII (ref). Our data clearly indicate that Hdf1 acts similarly to Sir2 in controlling the response to irradiation, but the role of this protein in chromatin remodelling is less clear. The Ku70 protein, encoded by *HDF1*, is released from telomeres after irradiation and relocated to double strand breaks (DSBs), in which it plays a non-essential role in NHEJ repair (47). The released Ku70, together with Ku80, may recruit the Remodel the Structure of Chromatin (RSC) complex, an abundant multisubunit protein complex involved in ATP-dependent chromatin remodelling in yeast (20). The chromatin remodelling activity of this complex depends on the physical interaction of RSC with the Ku80-Ku70 dimer (48). We found that four genes (RSC1, RSC2, RSC3 and RSC8) encoding components of the RSC complex were induced (C9) after irradiation. Once the chromatin remodelling-dependent IR response has been triggered, a more subtle response may occur through the recruitment of TFs sensitive to chromatin structure, such as Sko1 (involved in regulating C22 genes), which forms a complex with Tup1 and Ssn6 to repress salt stress defence genes and the oxidative stress response (49). Tup1 and Ssn6 are involved in establishing a repressive chromatin structure through interactions with histones H3 and H4 (50) and participate in heterochromatin formation in subtelomeric regions (24). Similarly, Rap1, which binds to telomere sequences and plays a role in telomeric silencing (51) through its recruitment of Fhl1 and Sfp1 (52,53), may be involved in controlling the transcription of genes of the C15 cluster.

The induction of ribosome biogenesis was confirmed by analyses of cluster C15, which contains a bicluster composed of genes encoding almost all the Pol I and Pol III subunits. Cross-talk between

the Pol I, Pol III and Pol II transcriptional machineries co-ordinates the synthesis of ribosome components. The high levels of RNA Pol I induced by irradiation may divert RNA Pol II away from its global transcription programme towards the preferential transcription of ribosomal proteins, as in *polI* overexpression mutants (39). This would result in the expression of genes involved in other aspects of global primary metabolism, such as carbohydrate metabolism and amino-acid metabolism, being decreased in favour of the transcription of ribosome-related genes. The hijacking of pol II for the transcription of ribosomal genes would decrease transcription at other loci, potentially leading to an apparent repression of general metabolism, as observed for the C8 and C22 clusters. We propose a model (Figure 6) involving the linked regulation of the various IR responses associated with chromatin alterations: changes in chromatin structure directly affect the transcription of C2 genes (including a significant number of subtelomeric genes) and have an indirect effect through the release of Rap1, which recruits Fhl1 and Sfp1 proteins to their binding sites in the promoters of C15 genes. The indirect effect would also induce C22 gene repression through the release of Tup1 and Ssn6 from telomeres and recruitment of the Sko1 repressor.

The induction of ribosome biogenesis after irradiation observed here conflicts with the observations of Gasch *et al.* (19), who described a repression of ribosomal protein production after irradiation. This repression of ribosome synthesis is surprising because cell size is directly correlated with ribosome biogenesis (42,54,55) and yeast cells (like almost all living cells) increase in size when they stop dividing after irradiation. We compared the increase in cell size with the increase in amount of an essential ribosomal protein (Rpl11b) in cells and showed that these two entities remained correlated in irradiated cells. Moreover, the Sfp1, Rrn3 and Sch9 TFs have been reported to relocate to the cytoplasm in response to genotoxic agents or oxidative stress. This change in the cellular distribution of TFs is observed whenever the transcription of ribosomal genes is repressed (42,45). However, we observed no change in the cellular distribution of Sfp1, Rrn3 or Sch9 after

irradiation. Instead, all our data suggest that ribosome synthesis increased after irradiation. By contrast, nutrient limitation and stress induce the repression of ribosome synthesis (42,45). Our results may differ from those of Gasch *et al*. because we used a much shorter irradiation time (200 seconds rather than 20 minutes) to limit cell storage in concentrated conditions. It is possible that the observed repression of ribosomes synthesis in Gasch's study resulted from starvation conditions during irradiation.

We detected an early response to irradiation but no late response corresponding to the reinitiation of cell division. Cell division is rapidly blocked in all cells of the population, whereas the reinitiation of cell division is highly heterogeneous in the population and depends on the efficiency of DNA damage repair in each cell. The identification of transient transcriptional changes occurring at the reinitiation of division is therefore likely to be hindered by the lack of synchrony.

In conclusion, using XRegPath to integrate various kinetic analyses, we were able to identify a general modification of metabolism in response to irradiation and to propose a regulatory mechanism for such a broad change. The XRegPath approach can be used with any kind of data and is an efficient tool for taking into account the expression profiles and properties of the studied strains.

Acknowledgements: This work was supported by Genopole®, the *Centre National de la Recherche* (CNRS), and the Institut Curie. Maria Quanz was supported by the German Academic Exchange Service (DAAD). The *Association Nationale pour la Recherche* Genomic Data to Graph Extraction project also partly supported the work of Nizar Touleimat.

References

1. Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.l.B., Storz, G., Botstein, D. and Brown, P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241-4257.
2. Mercier, G., Berthault, N., Touleimat, N., Kepes, F., Fourel, G., Gilson, E. and Dutreix, M. (2005) A haploid-specific transcriptional response to irradiation in *Saccharomyces cerevisiae*. *Nucl. Acids Res.*, **33**, 6635-6643.
3. Martin, S.G., Laroche, T., Suka, N., Grunstein, M. and M, G.S. (1999) Relocalization of telomeric Ku and SIR proteins in response to DNA strand breaks in yeast. *Cell*, **97**, 621-633.
4. Mills, K.D., Sinclair, D.A. and Guarente, L. (1999) MEC1-dependent redistribution of the Sir3 silencing protein from telomeres to DNA double-strand breaks. *Cell*, **97**, 609-620.
5. Maillet, L., Gaden, F., Brevet, V., Fourel, G., Martin, S.G., Dubrana, K., Gasser, S.M. and Gilson, E. (2001) Ku-deficient yeast strains exhibit alternative states of silencing competence. *EMBO J*, **2**, 203-210.
6. Boeke, J.D., LaCroute, F. and Fink, G.R. (1984) A positive selection for mutants lacking orotidine-5'-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance. *Mol Gen Genet*, **197**, 345-346.
7. Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S. and O'Shea, E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686-691.
8. Mercier, G., Denis, Y., Marc, P., Picard, L. and Dutreix, M. (2001) Transcriptional induction of repair genes during slowing of replication in irradiated *Saccharomyces cerevisiae*. *Mutation Research/DNA Repair*, **487**, 157-172.
9. Mercier, G., Berthault, N., Mary, J., Peyre, J.A.A., Comet, J.P., Cornuejols, A., Froidevaux, C. and Dutreix, M. (2004) Biological detection of low radiation doses by combining results of two microarray analysis methods. *Nucl. Acids Res.*, **32**, e12.
10. Verma, D. and Meila, M. (2001), *Technical report uw-cse-03-05-0 and University of Washington1*.
11. Ben-Hur, A., Elisseeff, A. and Guyon, I. (2002), *PSB*, Vol. 7, pp. 6-17.
12. Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.*, **28**, 27-30.
13. Consortium, T.G.O. (2000) Gene Ontology: tool for the unification of biology. *Nature Gen.*, **25**, 25-29.

14. Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99-104.

15. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P. *et al.* (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, **440**, 637-643.

16. Sokal, R.R. and Rohlf, F.J. (1995) *Biometry: The Principles and Practice of Statistics in Biological Research*.

17. Kluger, Y., Basri, R., Chang, J.T. and Gerstein, M. (2003) Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.*, **13**, 703-716.

18. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *PNAS*, **95**, 14863-14868.

19. Gasch, A.P., Huang, M., Metzner, S., Botstein, D., Elledge, S.J. and Brown, P.O. (2001) Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol. Biol. Cell*, **12**, 2987-3003.

20. Martens, J.A. and Winston, F. (2003) Recent advances in understanding chromatin remodeling by Swi/Snf complexes. *Curr. Op. in Gen. \& Dev.*, **13**, 136-142.

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

22. De Preter, K., Barriot, R., Speleman, F., Vandesompele, J. and Moreau, Y. (2008) Positional gene enrichment analysis of gene sets for high-resolution identification of overrepresented chromosomal regions. *Nucl. Acid Res.*, gkn114.

23. Laurenson, P. and Rine, J. (1991) SUM1-1: A Suppressor of silencing defects in *Saccharomyces cerevisiae*. *Genetics*, **129**, 685-696.

24. Robyr, D., Suka, Y., Xenarios, I., Kurdistani, S.K., Wang, A., Suka, N. and Grunstein, M. (2002) Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell*, **109**, 437-446.

25. Mennella, T.A., Klinkenberg, L.G. and Zitomer, R.S. (2003) Recruitment of Tup1-Ssn6 by yeast hypoxic genes and chromatin-independent exclusion of TATA binding protein. *Eukaryot Cell.*, **2**, 1288-1303.

26. Papamichos-Chronakis, M., Petrakis, T., Ktistaki, E., Topalidou, I. and Tzamarias, D. (2002) Cti6, a PHD domain protein, bridges the Cyc8-Tup1 corepressor and the SAGA coactivator to overcome repression at GAL1. *Mol. Cell*, **9**, 1297-1305.

27. Loo, S., Laurenson, P., Foss, M., Dillin, A. and Rine, J. (1995) Roles of ABF1, NPL3, and YCL54 in silencing in *Saccharomyces cerevisiae*. *Genetics*, **141**, 889-902.

28. Pryde, F.E. and Louis, E.J. (1999) Limitations of silencing at native yeast telomeres. *EMBO J.*, **18**, 2538-2550.

29. Lascaris, R.F., Groot, E., Hoen, P.B., Mager, W.H. and Planta, R.J. (2000) Different roles for abf1p and a T-rich promoter element in nucleosome organization of the yeast RPS28A gene. *Nucleic Acids Res.*, **28**, 1390-1396.

30. Jorgensen, P., Rupes, I., Sharom, J.R., Schneper, L., Broach, J.R. and Tyers, M. (2004) A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size. *Genes Dev*, **18**, 2491-2505.

31. Fatica, A., Oeffinger, M., Tollorvey, D. and Bozzoni, I. (2003) Cic1p/Nsa3p is required for synthesis and nuclear export of 60S ribosomal subunits. *RNA*, **9**, 1431-1436.

32. Stage-Zimmermann, T., Schmidt, U. and Silver, P.A. (2000) Factors affecting nuclear export of the 60S subunits *in vivo*. *Mol. Biol. Cell*, **11**, 3777-3789.

33. Rotenberg, M.O., Moritz, M. and Woolford, J.L., Jr. (1988) Depletion of *Saccharomyces cerevisiae* ribosomal protein L16 causes a decrease in 60S ribosomal subunits and formation of half-mer polyribosomes. *Genes Dev*, **2**, 160-172.

34. Maaloe, O. and Kjeldgaard, N.O. (1966) *Control of Macromolecular Synthesis*. WA Benjamin New York.

35. Nomura, M. (1999) Regulation of ribosome biosynthesis in *Escherichia coli* and *Saccharomyces cerevisiae*: diversity and common principles. *J Bacteriol.*, **181**, 6857-6864.

36. Warner, J.R. (1999) The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.*, **24**, 437-440.

37. Zhao, Y., Sohn, J.H. and Warner, J.R. (2003) Autoregulation in the biosynthesis of ribosomes. *Mol. Cell. Biol.*, **23**, 699-707.

38. Brauer, M.J., Huttenhower, C., Airoldi, E.M., Rosenstein, R., Matese, J.C., Gresham, D., Boer, V.M., Troyanskaya, O.G. and Botstein, D. (2008) Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. *Mol. Biol. Cell*, **19**, 352-367.

39. Laferte, A., Favry, E., Sentenac, A., Riva, M., Carles, C. and Chedin, S. (2006) The transcriptional activity of RNA polymerase I is a key determinant for the level of all ribosome components. *Genes Dev*, **20**, 2030-2040.

40. De Virgilio, C. and Loewith, R. (2006) The TOR signalling network from yeast to man. *Int. J. of Biochem. \& Cell Biol.*, **38**, 1476-1481.

41. De Virgilio, C. and Loewith, R. (2006) Cell growth control: little eukaryotes make big contributions. *Oncogene*, **25**, 6392-6415.

42. Jorgensen, P., Rupes, I., Sharom, J.R., Schneper, L., Broach, J.R. and Tyers, M. (2004) A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size. *Genes Dev.*, **18**, 2491-2505.

43. Jorgensen, P. and Tyers, M. (2004) How cells coordinate growth and division. *Current Biology*, **14**, R1014-R1027.

44. Hartwell, L.H. (1974) *Saccharomyces cerevisiae* cell cycle. *Bacteriol Rev.*, **38**, 164-198.

45. Marion, R.M., Regev, A., Segal, E., Y, B., Koller, D., Friedman, N. and O'Shea, E.K. (2004) Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *PNAS*, **101**, 14315-14322.

46. Grummt, I. and Pikaard, C.S. (2003) Epigenetic silencing of RNA polymerase I transcription. *Nat Rev Mol Cell Biol*, **4**, 641-649.

47. Barnes, G. and Rio, D. (1997) DNA double-strand-break sensitivity, DNA replication, and cell cycle arrest phenotypes of Ku-deficient *Saccharomyces cerevisiae*. *PNAS*, **94**, 867-872.

48. Shim, E.Y., Ma, J.L., Oum, J.H., Y, Y. and Lee, S.E. (2005) The yeast chromatin remodeler RSC complex facilitates end joining repair of DNA double-strand breaks. *Mol Cell Biol.*, **25**, 3934-3944.

49. Pascual-Ahuir, A., Posas, F., Serrano, R. and Proft, M. (2001) Multiple levels of control regulate the yeast cAMP-response element-binding protein repressor Sko1p in response to stress. *J Biol Chem*, **276**, 37373-37378.

50. Davie, J.K., Trumbly, R.J. and Dent, S.Y. (2002) Histone-dependent association of Tup1-Ssn6 with repressed genes *in vivo*. *Mol Cell Biol*, **22**, 693-rud703.

51. Shore, D. (1997) Telomere length regulation: getting the measure of chromosome ends. *Biol Chem*, **378**, 591-597.

52. Buchman, A.R., Lue, N.F. and Kornberg, R.D. (1988) Connections between transcriptional activators, silencers, and telomeres as revealed by functional analysis of a yeast DNA-binding protein. *Mol Cell Biol*, **8**, 5086-5099.

53. Vignais, M.L., Woudt, L.P., Wassenaar, G.M., Mager, W.H., Sentenac, A. and Planta, R.J. (1987) Specific binding of TUF factor to upstream activation sites of yeast ribosomal protein genes. *EMBO J.*, **6**, 1451-1457.

54. Cipollina, C., van den Brink, J., Daran-Lapujade, P., Pronk, J.T., Vai, M. and de Winde, J.H. (2008) Revisiting the role of yeast Sfp1 in ribosome biogenesis and cell size control: a chemostat study. *Microbiology*, **154**, 337-346.

55. Rudra, D. and Warner, J.R. (2004) What better measure than ribosome synthesis? *Genes Dev.*, **18**, 2431-2436.

56. Cramer, P. (2002) Multisubunit RNA polymerases. *Curr Opin Struct Biol*, **12**, 89-97.

Figure legends

Figure 1: Expression kinetics for Cluster 15 and relationship to the cell cycle. Cells (FF18733) were analysed before (0 h) and 0.1, 1, 2, 3, 4 and 5 h after 200 Gy irradiation for cell cycle (grey plots) and for global gene expression (black line: C15 mean expression profile; grey lines: standard deviations).

Figure 2. Mean kinetics of irradiation-sensitive clusters in the different yeast strains. Y-axis: log2 of normalised mean intensity value. X-axis: expression levels at seven time points (time 0, 0.1, 1, 2, 3, 4, 5 hours) in the yeast strains FF18733 (A), FF18734 (B), FF18735 (C), FF6053 (D), FFD1 (E), LM79 (F), yIBPC205 (G). Continuous lines: mean expression profiles. Dotted lines: standard deviation. NI: non-irradiated.

Figure 3: Analysis of the biological content of each IR-modulated cluster. Significant biological information gathered from biclustering is shown in various colours. TFs significantly associated with gene subsets are indicated.

Figure 4: Graphical representation of the biclustering of cluster C15. Five biclusters displaying specialisation in different steps of ribosome biogenesis. Y-axis: 146 C15 genes associated with at

least one enriched biological descriptor. X-axis: biological descriptors (see details in supplementary figure S2).

Figure 5. Increase in ribosome biogenesis and cell size after irradiation. a) Real-time PCR quantification of RNA polymerase genes expression: *RPC40* (circles), *RPB8* (triangles), *RPC31* (inverted triangles), *RPA49* (diamonds). *ACT1* (squares) was used as an irradiation-insensitive control gene. b) Quantification of the percentage of Yg179 cells displaying *URA3* silencing at the rDNA locus after irradiation. c-d) Flow cytometry quantification of ribosome content (Rpl11b-GFP fluorescence) and cell size (FSC-H) after irradiation. e) Images of Rpl11b-GFP-expressing fluorescent yeast cells before (upper panel) and 4 h after irradiation (lower panel). Scale bars:10 μm.

Figure 6: Model of regulation of the IR response. The area of the triangles is proportional to the number of genes. Arrows indicate positive controls, bars indicate negative controls. Red: induced clusters; green: repressed clusters.

Table 1: Cluster description

	Strain	FF18733	FF18734	FF18735	FF6053	FFD1	LM79	YiBPC205
genetic markers	Ploidy	n	n	n	2n	n	n	n
	Mating type	a	α	a(α)	a/α	a	α	α
	Mutations	-	-	-	-	rad52	ku70	sir2
Clusters (gene number)	C15 (198)	I	I	I	U	I	U	U
	C2 (212)	I	I	U	U	I	U	U
	C9 (210)	I	U	U	U	I	U	U
	C10 (220)	U	U	U	U	U	I	I
	C18 (304)	U	U	U	I	U	U	U
	C13 (101)	R	R	R	R	R	R	R
	C8 (25)	R	R	R	U	R	R	R
	C22 (119)	R	R	R	U	R	U	U

I: mean overall expression induced

R: mean overall expression repressed

U: unmodulated mean profile.

In brackets: number of genes in each cluster

Table 2: Induction of PolI and PolIII subunits by irradiation

Pol I	Pol II	Pol III	Class
A190	Rpb1	C160	Core
A135	Rpb2*	C128*	Core
AC40*	Rpb3	AC40*	Core
AC19	Rpb11	**AC19**	Core
Rpb6	**Rpb6**	**Rpb6**	Core/Common
Rpb5	**Rpb5**	**Rpb5**	Common
Rpb8	**Rpb8**	**Rpb8**	Common
Rpb10	**Rpb10**	**Rpb10**	Common
Rpb12	Rpb12	Rpb12	Common
A12.2	Rpb9	**C11**	
A14	Rpb4	**-**	
A43	Rpb7	**C25**	
A34.5	-	**C17**	
A49		**C31**	
		C34	

Bold: induced by irradiation (cluster 15)

*: not investigated.

Core: sequence partially homologous in all RNA polymerases;

common: shared by all eukaryotic RNA polymerases (56).

Figure 1

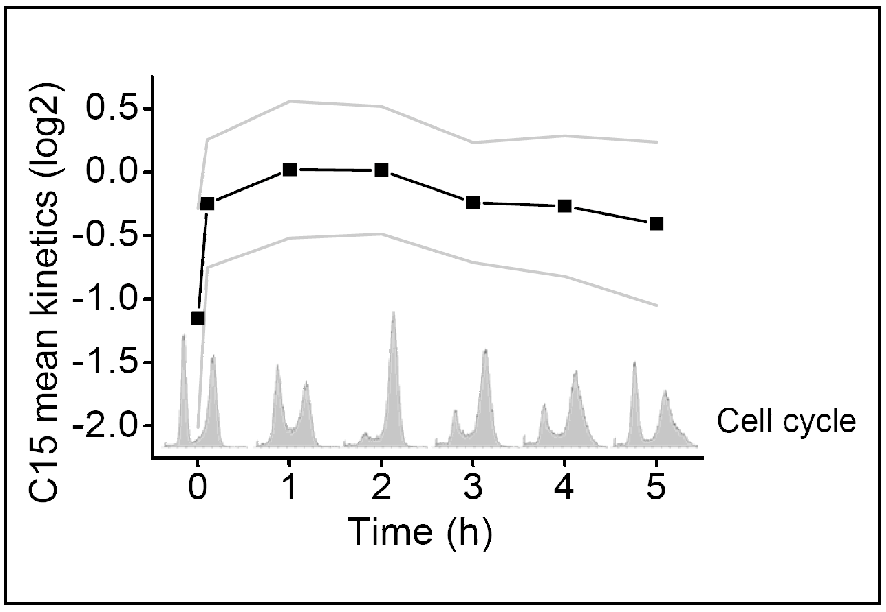

C15 mean kinetics (log2)
0.5
0.0
-0.5
-1.0
-1.5
-2.0
0
1
2
3
4
5
Time (h)
Cell cycle

Figure 2

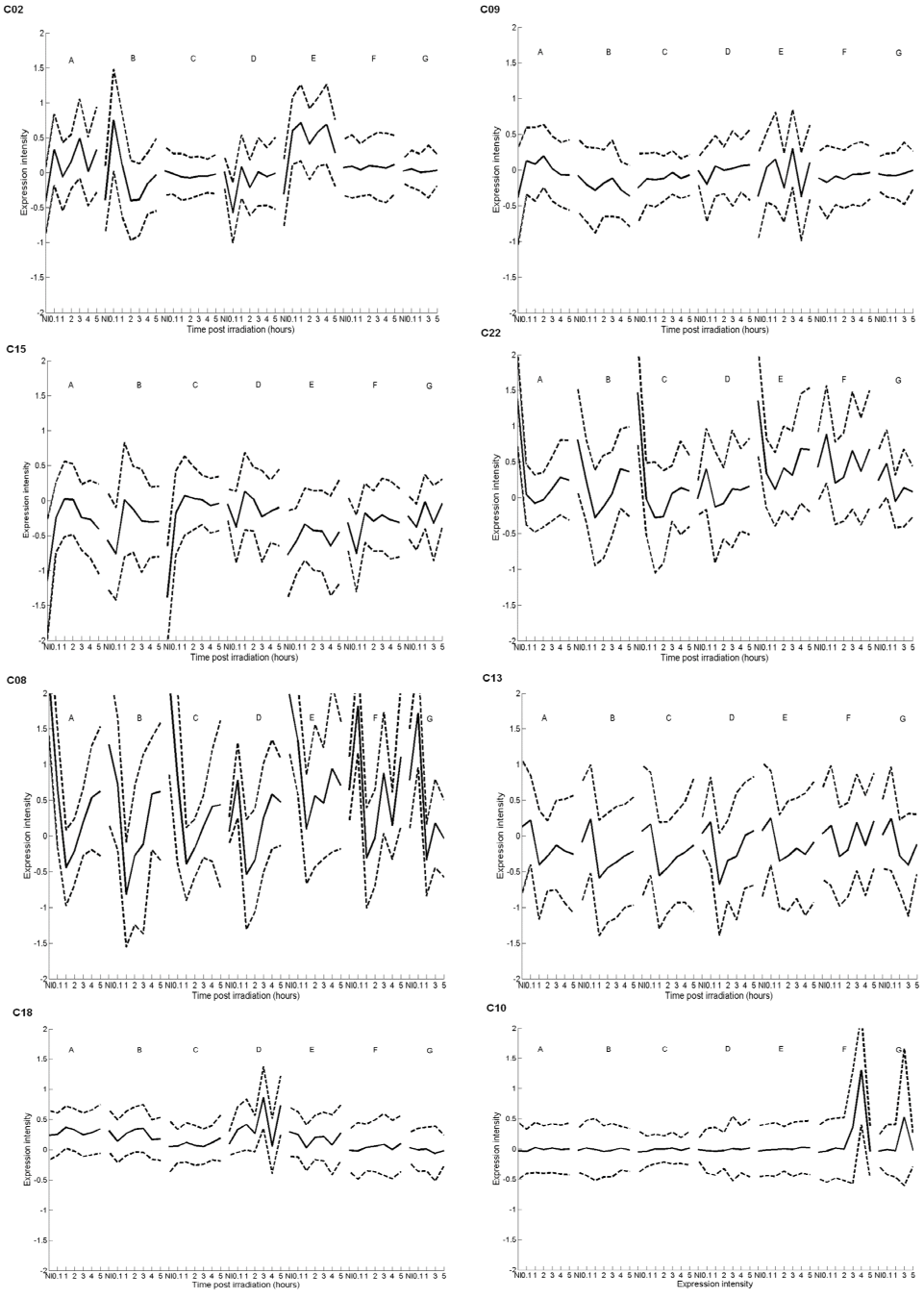

C02
C09
C15
C22
C08
C13
C18
C10
A
B
C
D
E
F
G
Expression intensity
Time post irradiation (hours)

Figure 3

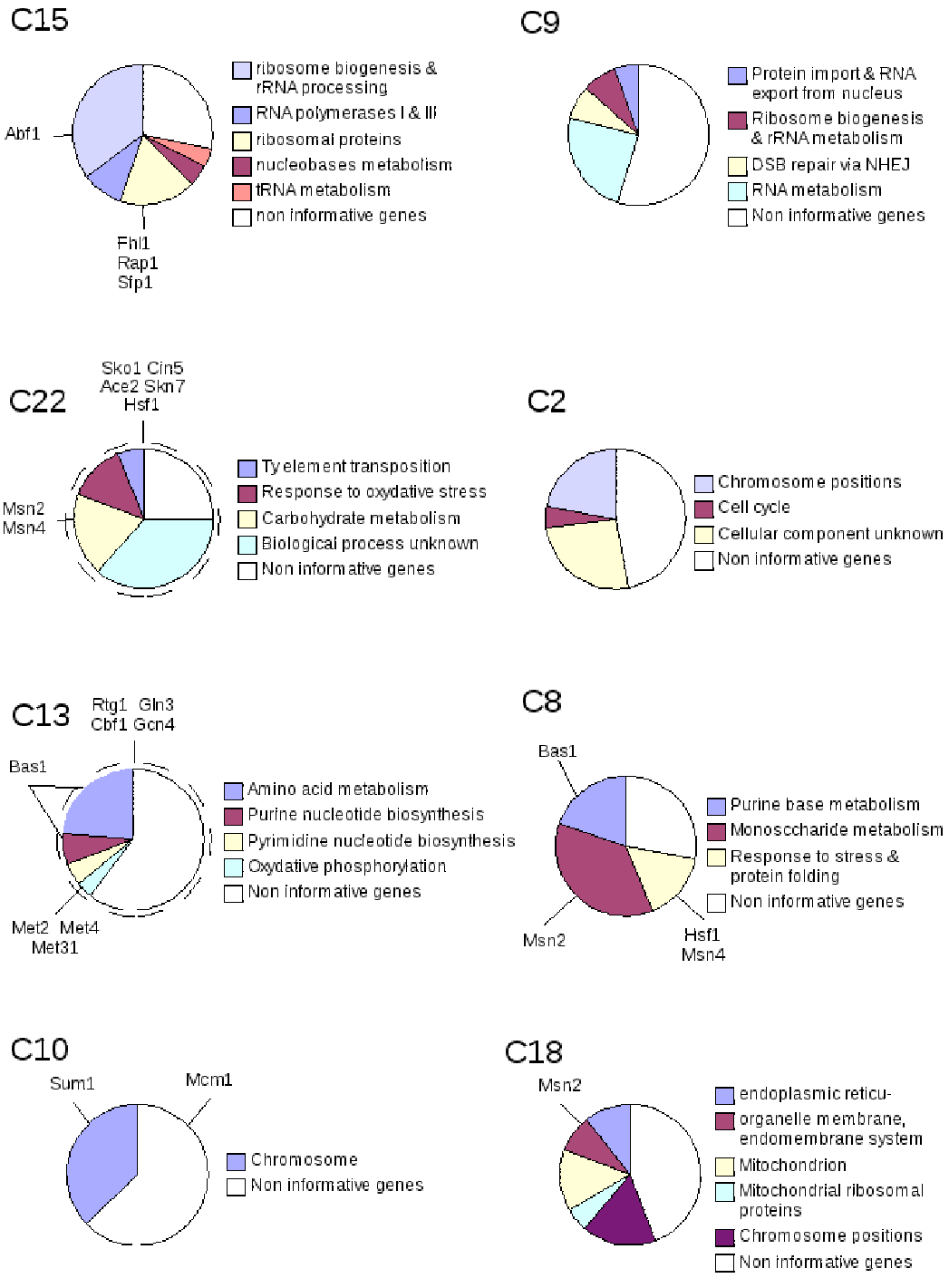

C15
Abf1
Fhl1
Rap1
Sfp1
ribosome biogenesis & rRNA processing
RNA polymerases I & III
ribosomal proteins
nucleobases metabolism
tRNA metabolism
non informative genes
C9
Protein import & RNA export from nucleus
Ribosome biogenesis & rRNA metabolism
DSB repair via NHEJ
RNA metabolism
Non informative genes
C22
Sko1 Cin5
Ace2 Skn7
Hsf1
Msn2
Msn4
Ty element transposition
Response to oxydative stress
Carbohydrate metabolism
Biological process unknown
Non informative genes
C2
Chromosome positions
Cell cycle
Cellular component unknown
Non informative genes
C13
Rtg1 Gln3
Cbf1 Gcn4
Bas1
Met2 Met4
Met31
Amino acid metabolism
Purine nucleotide biosynthesis
Pyrimidine nucleotide biosynthesis
Oxydative phosphorylation
Non informative genes
C8
Bas1
Msn2
Hsf1
Msn4
Purine base metabolism
Monosccharide metabolism
Response to stress & protein folding
Non informative genes
C10
Sum1
Mcm1
Chromosome
Non informative genes
C18
Msn2
endoplasmic reticu-
organelle membrane, endomembrane system
Mitochondrion
Mitochondrial ribosomal proteins
Chromosome positions
Non informative genes

Figure 4

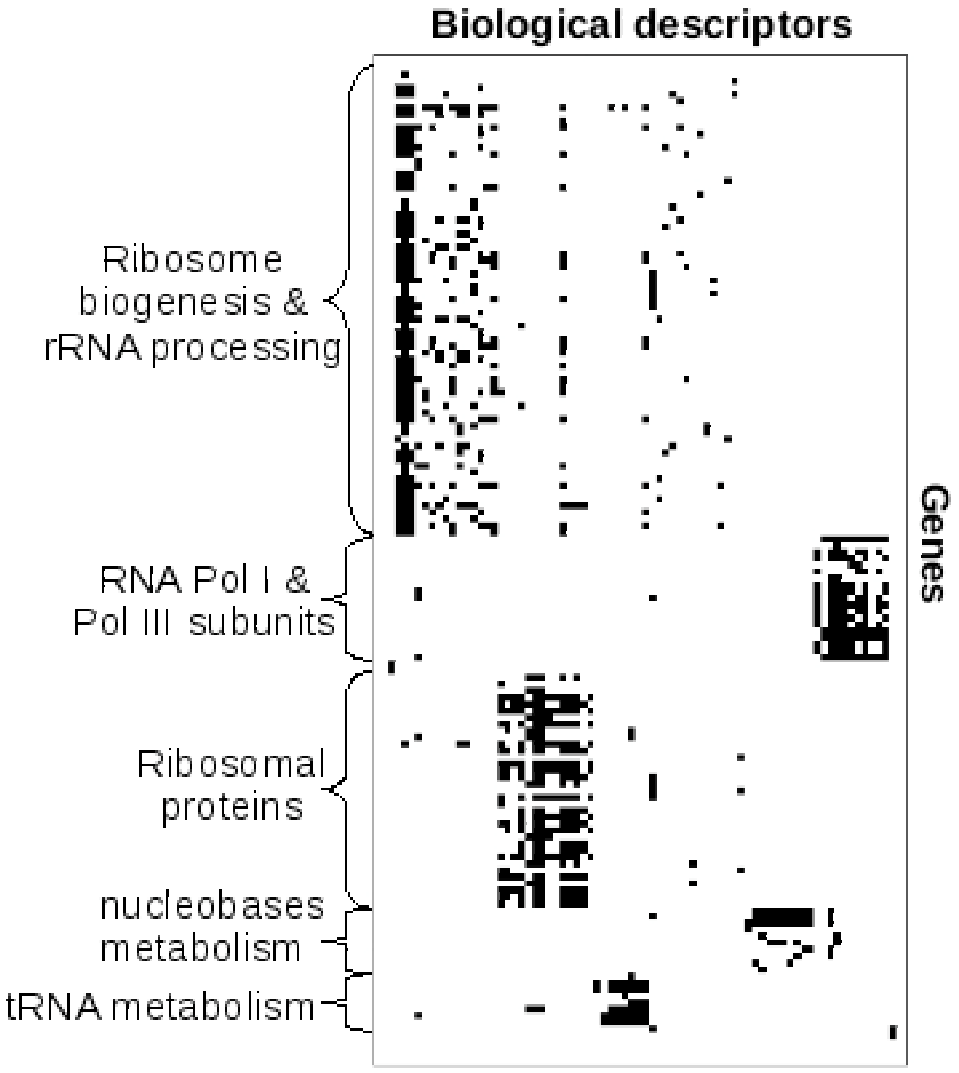

Biological descriptors
Genes
Ribosome biogenesis & rRNA processing
RNA Pol I & Pol III subunits
Ribosomal proteins
nucleobases metabolism
tRNA metabolism

Figure 5

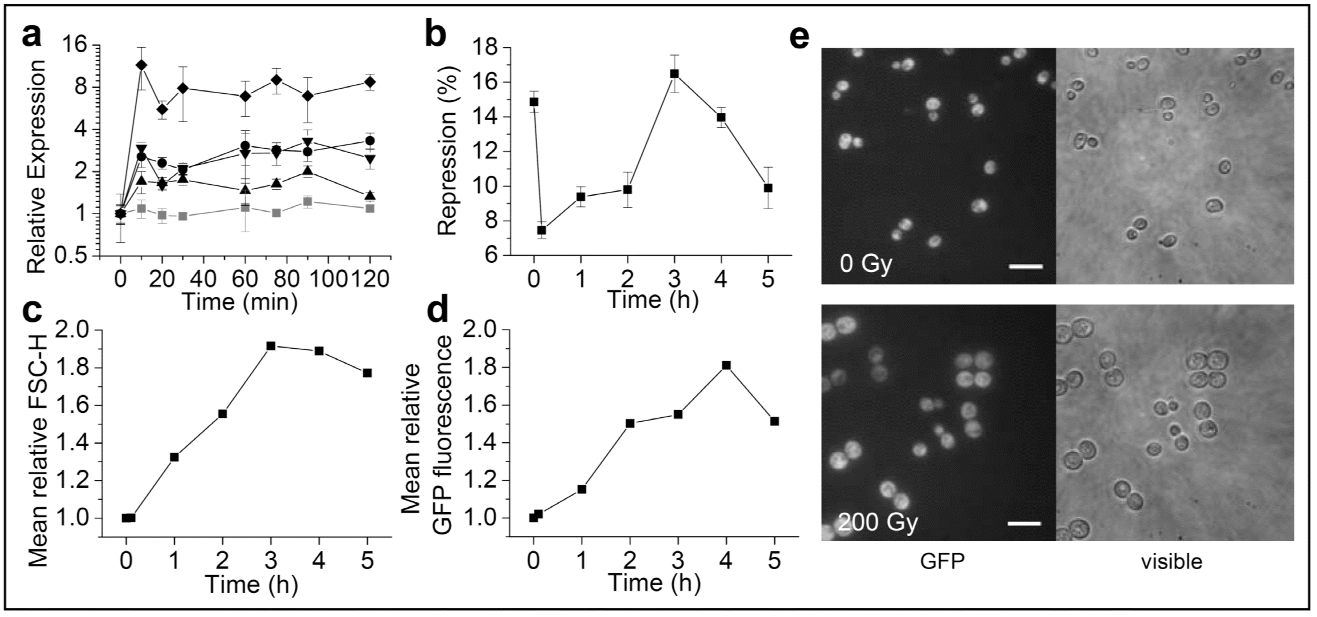

a
Relative Expression
16
8
4
2
1
0.5
0 20 40 60 80 100 120
Time (min)
b
Repression (%)
18
16
14
12
10
8
6
0 1 2 3 4 5
Time (h)
e
0 Gy
200 Gy
GFP
visible
c
Mean relative FSC-H
2.0
1.8
1.6
1.4
1.2
1.0
0 1 2 3 4 5
Time (h)
d
Mean relative GFP fluorescence
2.0
1.8
1.6
1.4
1.2
1.0
0 1 2 3 4 5
Time (h)

Figure 6

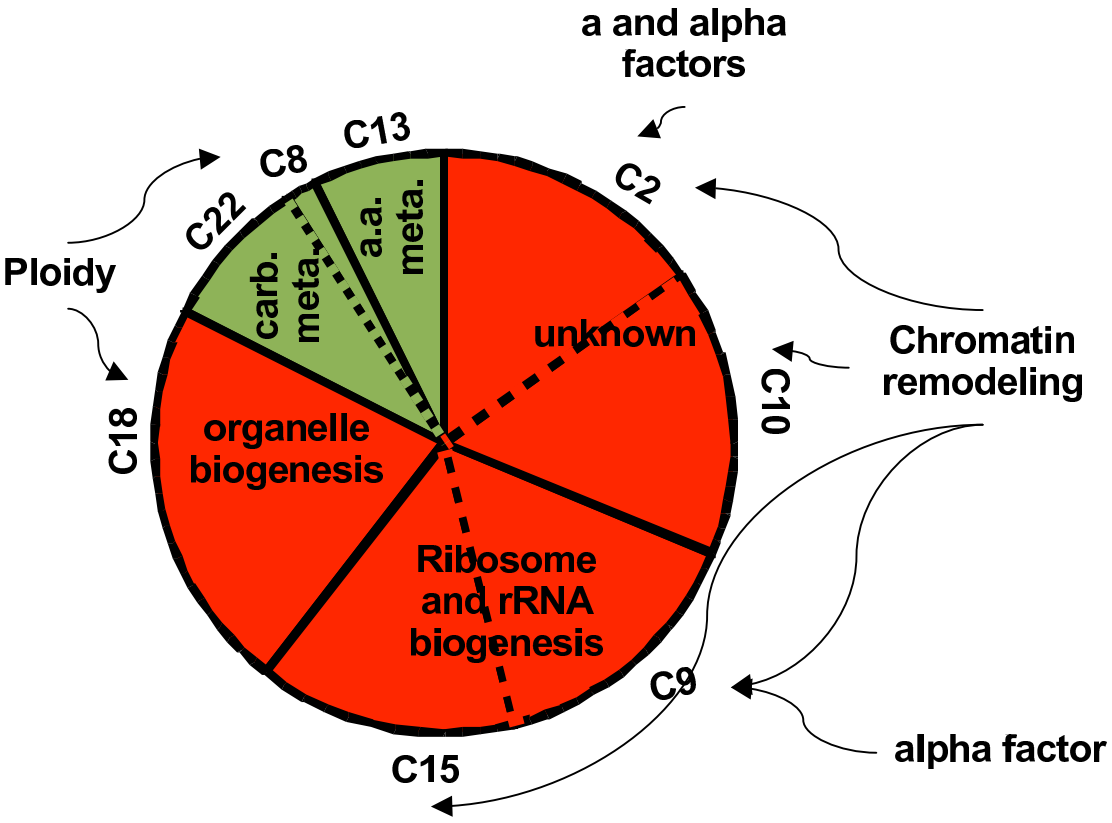

a and alpha factors
C13
C8
C22
C2
Ploidy
carb. meta.
a.a. meta.
unknown
Chromatin remodeling
C10
C18
organelle biogenesis
Ribosome and rRNA biogenesis
C9
alpha factor
C15

Figure S1 (Panel A and B): Projection of the genes of our eight IR modulated clusters onto t data of Spellman *et al*. (*Spellman et al., 1998*). A: projection onto alpha factor time course data, projection onto cdc15 time course data, C: projection onto cdc28 time course data and D: projecti onto elutriation time course data. Continuous lines: mean expression profiles of the gene cluster for t given data set. Dotted lines: mean expression profiles +/- the standard deviation for the given data se

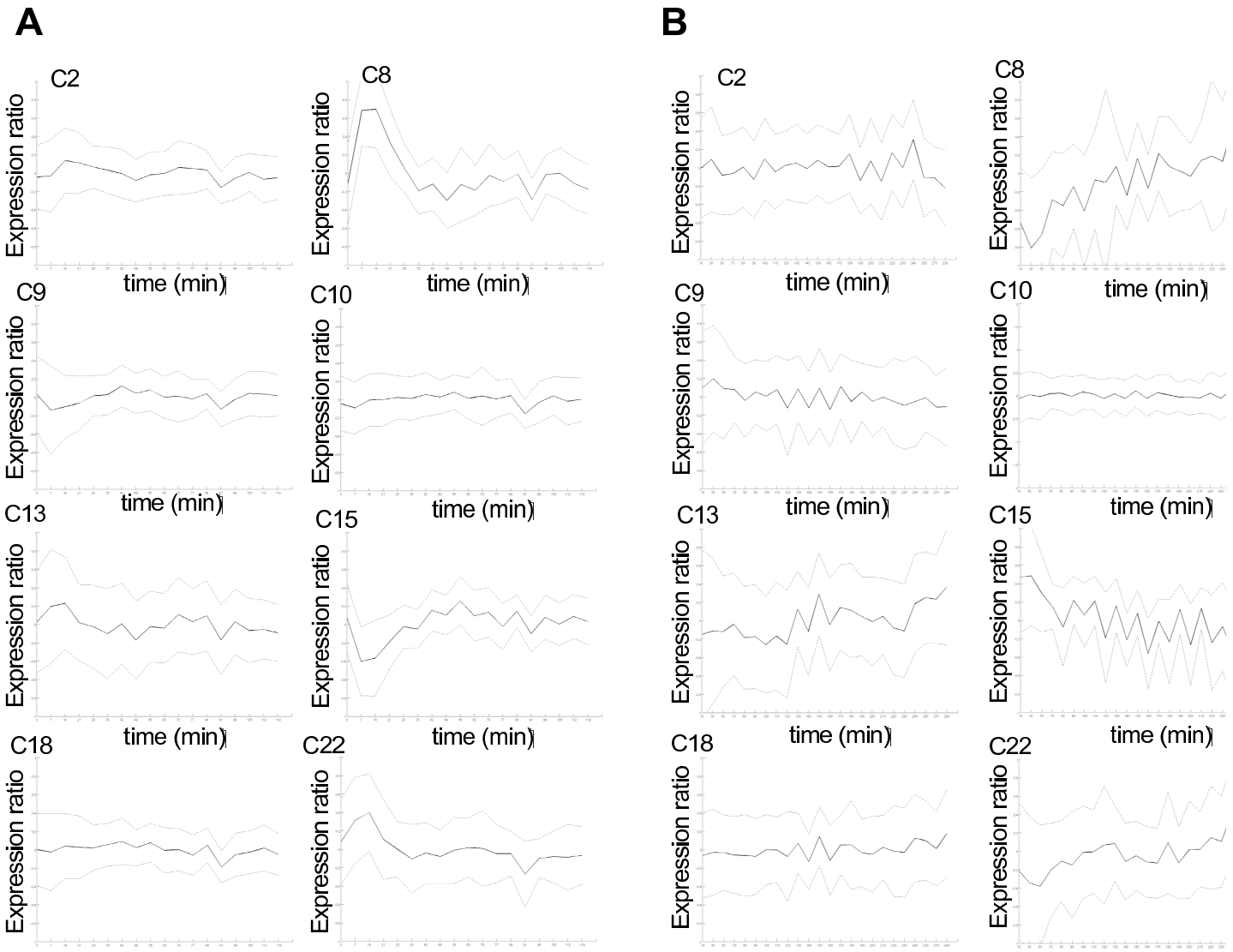


Figure S1 (Panel C and D

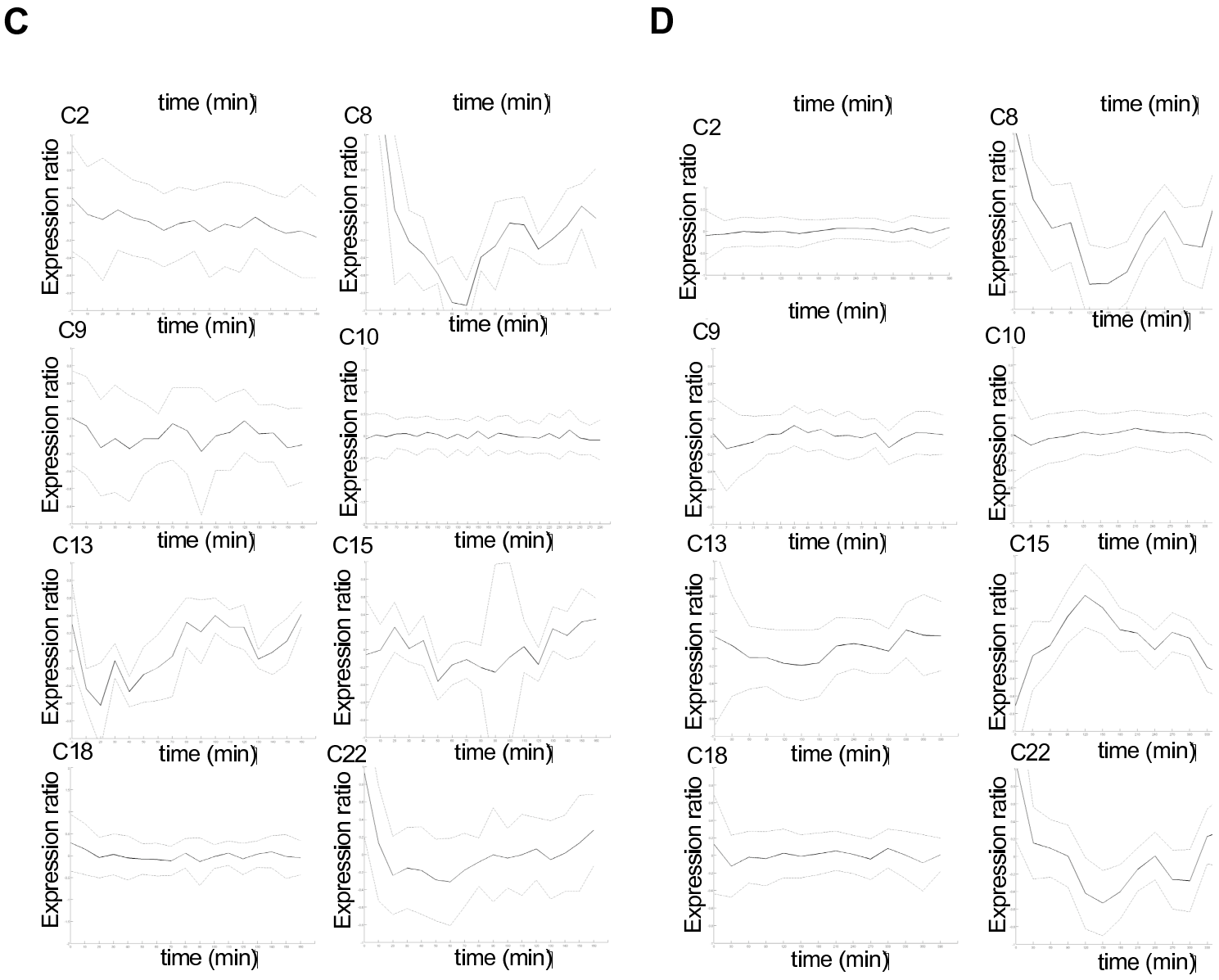
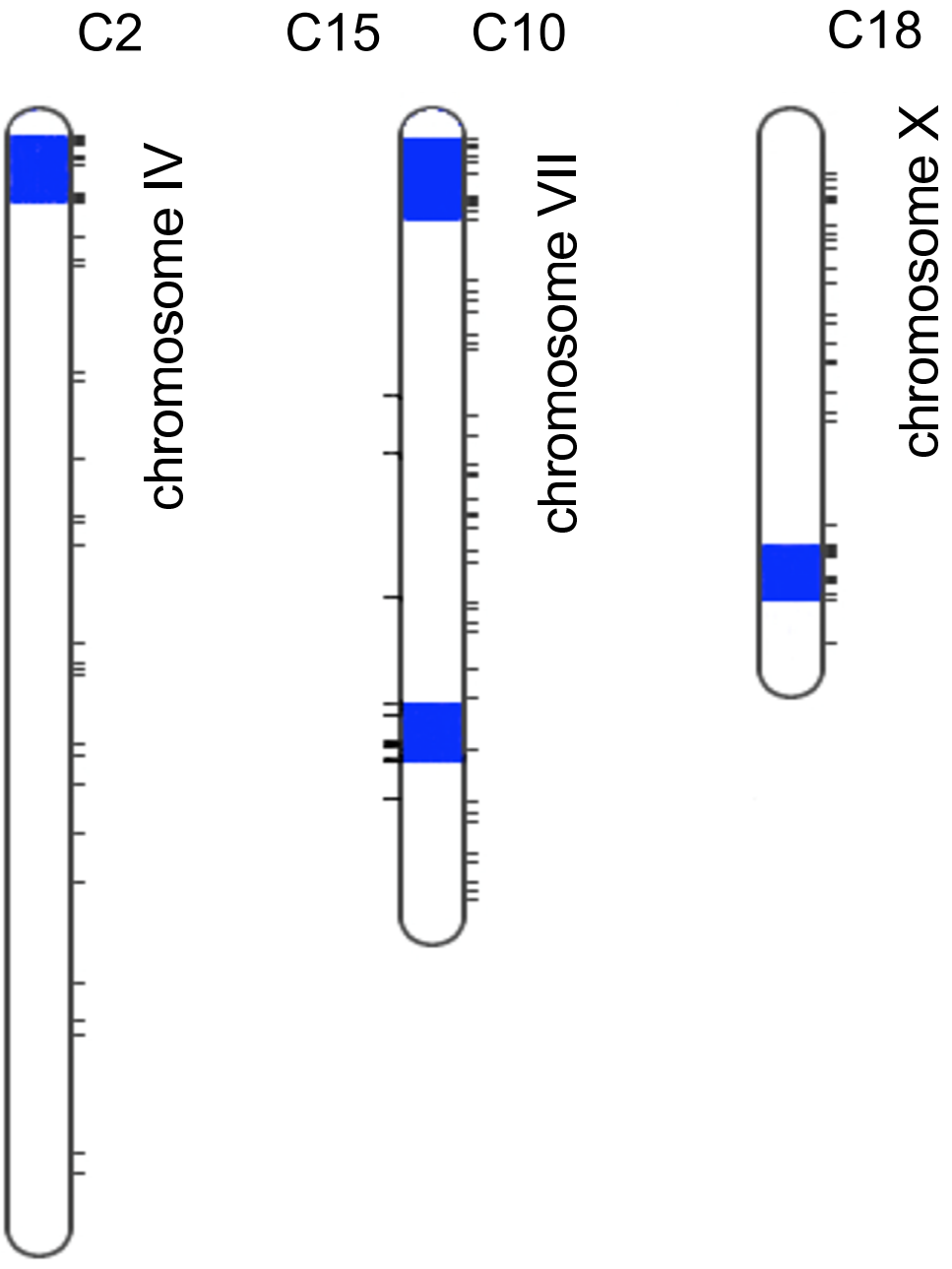


Figure S2: Chromosome regions significantly enriched in co-expressed genes. In blue: regior enriched in genes from the same cluster. C2 region on chromosome IV: 12 genes and *p-value* = 2.3 x 10^{-8}, C15 region on chromosome VII: 9 genes and *p-value* = 8.8 x 10^{-6}, C10 region on chromosome VII: 14 genes and *p-value* = 4.3 x 10^{-8}, C18 region on chromosome X: 15 genes and *p-value* = 1.66 10^{-8}.



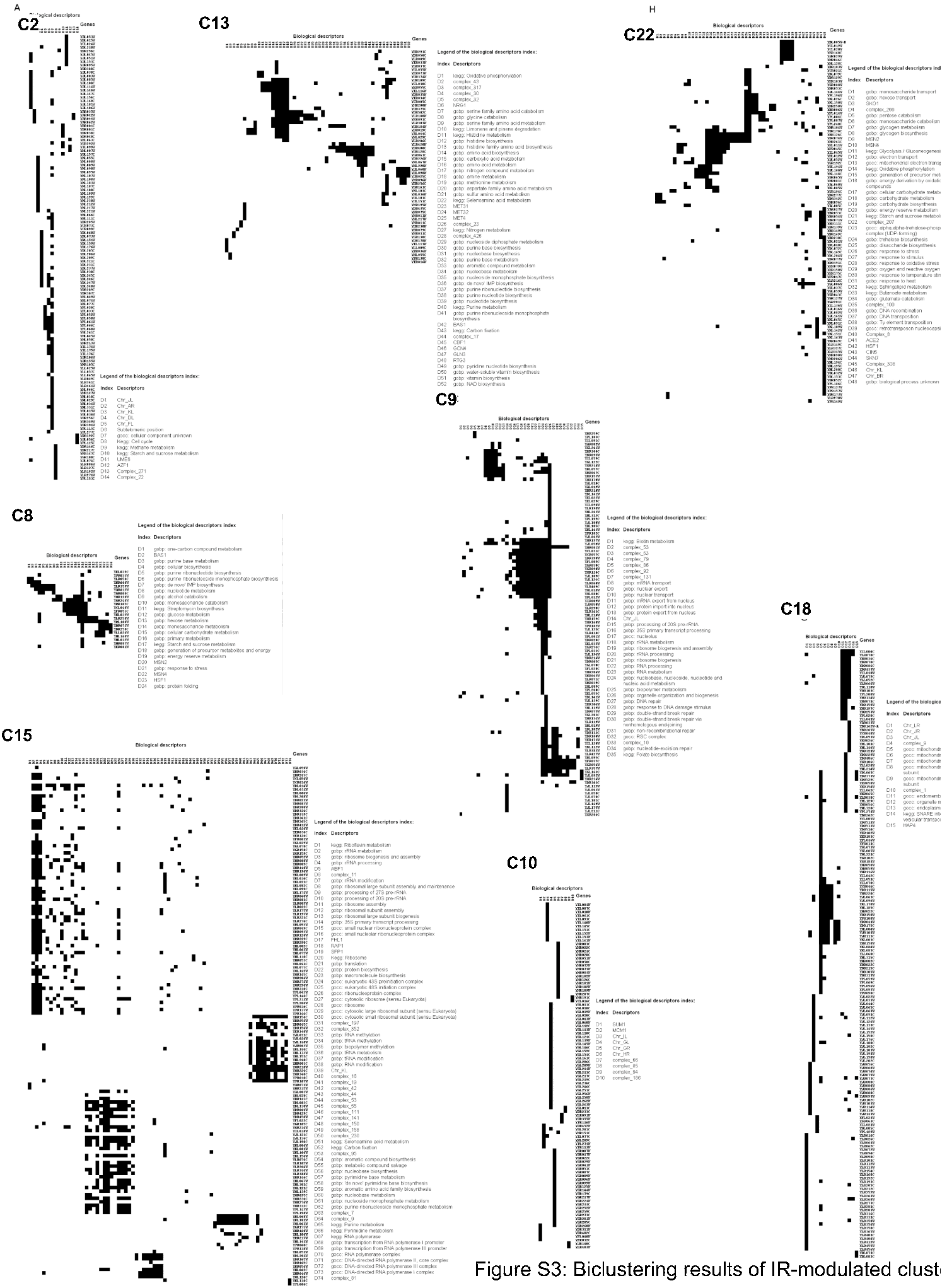


Figure S3: Biclustering results of IR-modulated clust

Figure S4 . Nuclear localisation of Sfp1 after irradiation. Yeast cells expressing an Sfp1-GFP fusion protein construct under control of the SFP1 promoter were analysed in the exponential growth phase before irradiation (NI), in stationary phase (Stat) and 5, 15, 30, 45, 60, 75, 90 minutes after irradiation (all images obtained were similar and only the 30-minute time point is presented) and DAPI staining was used to visualise chromatin. Scale bar: 5 µm. The cells were analysed by fluorescence and phase-contrast light microscopy (visible).

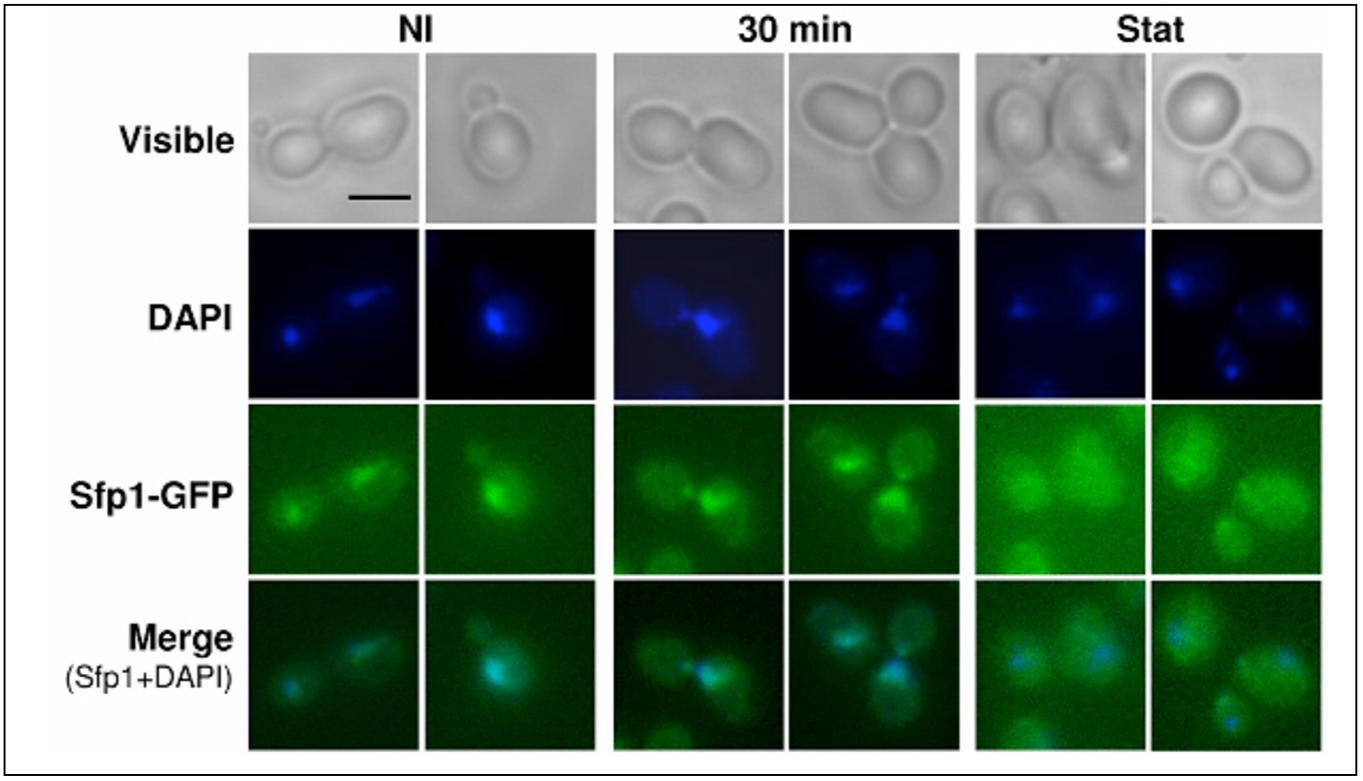


Figure S5 . Nucleolar localisation of Rrn3 after irradiation. Yeast cells expressing an Rrn3-GFP fusion protein construct under control of the *RRN3* promoter and carrying a plasmid expressing a Nop1- mCherry fusion protein construct under control of the *NOP1* promoter (for the visualisation of nucleoli) were analysed in the exponential growth phase before irradiation (NI), in stationary phase (Stat) and 5, 15, 30, 45, 60, 75 and 90 minutes after irradiation (all the images were similar and only the 30-minute time point is presented). Scale bar: 5 µm. The cells were analysed by fluorescence and phase-contrast light microscopy (visible).

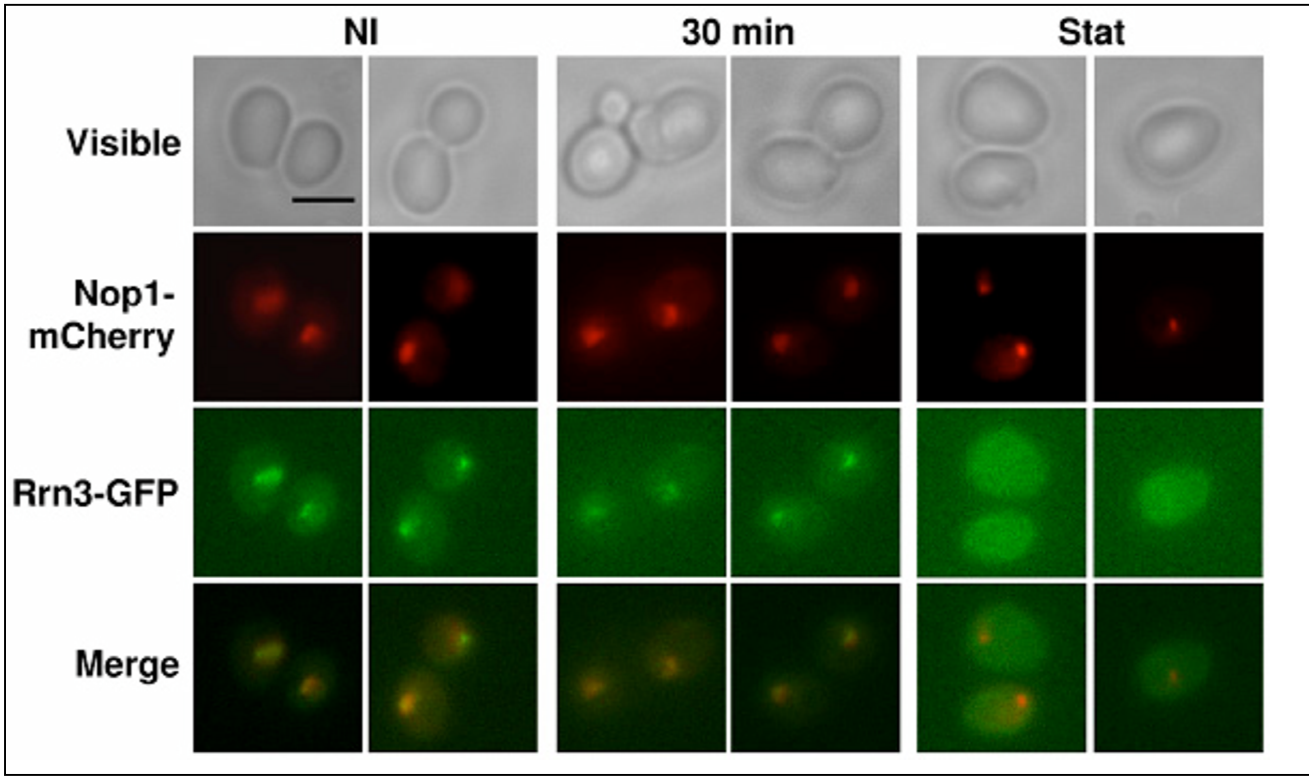


Table S1: Primer sequences for real-time PCR

Gene	Orientation	Sequence (5′ to 3′)
RPC31	forward	GGC AGC AAA AGG AAG TAA AA
	reverse	CAC CGT AAT CAT CAT CGT CA
RPA49	forward	TCG ATG CCA CTG ATG TAG AA
	reverse	GGA ATT GTT TTG GTA TGG GA
RPB8	forward	CAG GCT GGT GAC GAT CCC
	reverse	AAG AGG CAA CCG AAT GAG TA
RPC40	forward	GGA CGA TGA AAA GTT TAC GG
	reverse	TTC AAA TCA CGG GCG TAG
ACT1	forward	TGT TCC ATC CTT CTG TTT TG
	reverse	GCA TTC TTT CGG CAA TAC CT
NAP1	forward	TTC TTC CGC CAA CCC AT
	reverse	CCT TCT GCG TGG TCG TA

Conclusion

La principale originalité de notre méthodologie réside dans la juxtaposition de 5 étapes d'analyse qui permettent l'extraction de relations de régulation à partir de données d'expression cinétiques, de conditions de perturbations génétiques, et d'informations biologiques de sources différentes. Notre méthodologie emploie deux types d'inférences : une étape de déduction (exploitation logique de la stratégie de perturbations) et deux étapes d'induction (identification de groupes de gènes co-exprimés et identification de bi-classes au sein de ces groupes de gènes).

L'utilisation d'un cadre logique automatisé pour la déduction de règles de régulation est à notre connaissance relativement nouvelle pour l'interprétation et l'exploitation de stratégies de perturbations génétiques. Cette utilisation de la logique ne fait que formaliser et automatiser l'approche classique du biologiste, qui, à partir d'une stratégie de perturbations ciblées peut déduire à partir de ses observations quelles perturbations affectent le système. Ce processus simple devient plus complexe et source d'erreurs lorsque le nombre de perturbations augmente et que l'on combine différentes sortes de perturbations (mutations, ploïdie, double mutants, etc.). Dans ce cas, l'automatisation du raisonnement devient réellement intéressent. De plus, notre approche logique permet d'extraire, à partir de combinaisons de perturbations, des relations de co-régulation complexes, où plusieurs variations génétiques différentes doivent apparaître simultanément pour induire un gain (co-régulation additive) ou une perte (co-régulation suppressive) de sensibilité à un stimulus.

L'identification et l'analyse de la réponse transcriptionnelle à l'IR s'est basée exclusivement sur l'exploitation de la dynamique d'expression des gènes. C'est le traitement dynamique des cinétiques d'expression qui nous a permis de révéler l'hétérogénéité des formes que pouvait prendre la réponse à l'IR. L'identification de ces formes à l'aide d'une stratégie de classification s'est accompagné d'une analyse de la stabilité de notre algorithme de classification. Cette analyse nous a d'une part permis de prouver la stabilité et la robustesse de notre algorithme et nous a permis d'autre part d'aider au choix du nombre optimal de groupes de gènes co-exprimés permettant de représenter toute l'hétérogénéité des réponses des gènes.

Enfin, nous avons pu implémenter la totalité de notre approche dans deux logiciels complémentaires qui permettront à des biologistes d'exploiter leurs données expérimentales, d'intégrer des données externes et d'interpréter les résultats obtenus rapidement et simplement.

L'application de notre méthodologie s'inscrit dans le cercle vertueux qui caractérise la démarche de la biologie des systèmes : expérimentation → modélisation → simulation/extraction → expérimentation. Dans notre cas l'étape de modélisation a plutôt consisté à modéliser et à automatiser le raisonnement du biologiste. L'étape d'extraction nous a permis d'identifier de nouvelles hypothèses de régulation. Toutes ces hypothèses ont été minutieusement analysées et confrontées à la littérature. Nous avons pu ainsi proposer, pour la levure, un schéma de régulation global impliqué dans la réponse à l'IR. Différents mécanismes de régulation seraient impliqués dans la réponse à l'irradiation, faisant intervenir des régulations ciblées à l'aide de FTs mais aussi une régulation plus globale dépendant de la modulation de la structure chromatinienne. L'identification des différents acteurs de cette réponse à l'IR nous a permis de proposer l'hypothèse d'une ré-orchestration du programme de transcription, après IR, conduisant à la répression du métabolisme primaire de la levure et au renouvellement de ses différents composants cellulaires comme par exemple les ribosomes. Dans ce schéma global de réponse à l'IR, de nombreuses parties restent à vérifier mais, certains liens de régulation comme l'induction de la biogenèse des ribosomes ont pu être validés expérimentalement.

Perspectives

Perspectives à court terme

Chaque étape de notre méthodologie peut être améliorée ou étendue à des cas plus généraux.

Au niveau de la 1ère étape : à cette étape, nous avons focalisé notre attention sur les gènes dont les profils cinétiques restent cohérents quelle que soit la condition expérimentale dans laquelle ils sont observés. Cette sélection a été opérée grâce à une classification de type *multi-clustering* qui impose aux gènes de rester toujours proches les uns des autres. Il est cependant possible d'identifier des groupes de gènes qui ne sont co-exprimés que sur une partie des conditions et à certains intervalles de temps seulement. Les méthodes actuelles de *biclustering* peuvent être étendues à ce cas [321], où les données sont définies par différentes similarités selon trois dimensions : les gènes, les cinétiques d'expression et les conditions expérimentales (ici les perturbations). Cette approche peut être vue comme une relaxation de notre contrainte de "cohérence" de l'expression des gènes sur toutes les conditions.

Au niveau de la 2ème étape : notre méthodologie reste semi-automatique et nécessite une participation active de l'utilisateur au niveau de cette étape. L'utilisateur doit sélectionner les groupes de gènes co-exprimés qui présentent une modulation significative de leur expression moyenne dans au moins une condition expérimentale. Actuellement cette sélection est réalisée "manuellement" en fonction de l'interprétation qu'aura l'utilisateur des profils moyens des groupes de gènes. Pour automatiser cette étape de sélection, nous pourrions développer un test statistique de type *Student* multivarié afin de tester si la cinétique moyenne d'un groupe est significativement différente d'un profil type "non modulé". Ce type de test permet de prendre en compte à la fois l'amplitude des modulations et la variance au sein des groupes de gènes. Les résultats des tests nous permettraient également d'annoter automatiquement (modulation/non modulation) les profils moyens des groupes de gènes co-exprimés dans les différentes conditions expérimentales de l'étude.

Au niveau de la 3ème étape : nous intégrons au sein d'un même cadre différents types de données systémiques pour l'analyse de groupes de gènes co-exprimés. Cette intégration reste cependant limitée à des informations sous la forme d'associations gènes-descripteurs. Les données d'interactions protéines-protéines par exemple sont difficiles à intégrer sous leur forme originale et peuvent nécessiter un pré-traitement pour identifier de potentiels

complexes protéiques et produire ensuite des associations protéines-complexes. Ce type de pré-traitement est par exemple effectué par Krogan *et coll.* à l'aide d'outils d'apprentissage statistique (voir [126]). Il serait aussi possible d'utiliser un codage des différents types d'informations sous forme de plusieurs similarités et d'employer alors les approches de type *biclustering* présentés précédemment[321]. Nous pourrions imaginer une mesure de similarité entre deux gènes liée à la distance entre leur deux protéines au sein du réseau d'interaction. Cette mesure refléterait la probabilité que deux protéines interagissent au sein d'un même complexe.

Au niveau de la 4ème étape : ici, les perspectives que nous proposons correspondent essentiellement au développement de l'outil d'aide à la déduction automatique, XRegRules (voir section 4.1.4.2, page 140). Ces développements concernent la gestion des redondances dans les résultats des requêtes et la représentation des résultats (relations de régulation) sous la forme de graphes. Une nouvelle interface d'aide à la vérification de règles pourrait être développée, elle proposerait la liste des règles à vérifier et la liste des valeurs possibles pour chaque variable. Enfin, l'outil pourrait être étendu à un outil d'aide à la conception de stratégies expérimentales. En effet, en fonction d'une base de règles, de variables et de leurs différentes valeurs possibles il est possible de proposer automatiquement une base de *faits* (valeurs des perturbations) qui permette de vérifier toutes les règles, pour toutes les valeurs des variables.

Au niveau de la 5ème étape : cette étape correspond au recueil et à l'intégration de l'ensemble des informations inférées aux étapes précédentes pour construire un schéma global de la réponse à l'irradiation comprenant à la fois des informations fonctionnelles et un réseau de régulation de gènes. Actuellement ce schéma global est construit manuellement à partir de différents tableaux de résultats et de différentes solutions de règles logiques. Une première étape vers l'automatisation complète serait de coder tous les résultats sous la forme de prédicats logiques comme cela est déjà le cas pour les résultats de l'étape de déduction de règles de régulation. Il suffirait alors de traduire le fichier de résultats (arcs prédits) en un schéma de graphe. Il faudra cependant veiller à différencier par des prédicats différents les résultats liés directement à la réponse au stimulus de ceux issus de l'intégration de données biologiques de sources hétérogènes.

Perspectives à moyen terme

Prise en compte d'une notion d'incertitude. Une évolution majeure de notre approche consisterait à considérer toute la chaîne de traitement et y insérer un traitement des incertitudes pour associer à une prédiction d'arc un degré de confiance à chaque étape. Le premier niveau d'incertitude concernerait l'identification des réponses à l'irradiation par le partitionnement de l'ensemble des gènes en groupes de gènes co-exprimés. Une approche de type *soft-clustering* associée à un calcul de probabilité plutôt que de similarité permettrait d'associer un gène à différentes classes à la fois. Le deuxième niveau d'incertitude concernerait l'annotation des profils de co-expression moyens à l'aide d'un test

statistique. L'utilisateur saurait ainsi quel crédit donner aux différentes modulations d'expression observées. Ces deux probabilités pourraient ainsi être déroulées à travers l'étape de déduction logique et rester associées à chaque règle vérifiée (que ce soit au niveau du gène ou du groupe de gène). A la 4ème étape, l'intégration et l'exploitation des données biologiques hétérogènes pourraient être aussi associée à des probabilité. Plutôt que que de manipuler les associations gènes-descripteurs sous la forme d'une matrice binaire, chaque entrée de la matrice pourrait correspondre à une probabilité d'association. L'attribution de ces probabilité resterait spécifique de chaque type de données et dépendrait fortement de la qualité et de la fiabilité de la technique ayant produit les données. Chaque arc proposé dans le réseau final serait associé à une mesure d'incertitude qui permettrait à l'utilisateur d'orienter son analyse vers les parties du réseau qu'il estimera être les plus fiables. Pour permettre cette évolution de notre approche, il faudra cependant résoudre le problème de l'intégration séquentielle des différents types de mesures d'incertitudes.

Vers une modélisation fine et détaillée du comportement dynamique du réseau. Si l'on arrive à obtenir une description suffisamment fine des différents acteurs de la réponse à l'IR et que l'on arrive à isoler un petit réseau de régulation où les interactions potentielles ont pu être validées expérimentalement, ces informations peuvent servir de connaissances *a priori* pour des approches de modélisation fines comme l'estimation de paramètres dans le cadre des équations différentielles ordinaires ou les réseaux bayésiens dynamiques [230].

Généralisation de notre approche. Le stimulus utilisé dans notre étude, de fortes doses de radiations γ, pourrait être remplacé par tous types de stimulus ou variations environnementales pouvant influencer le comportement des cellules étudiées. En effet, l'application de notre méthodologie aux données transcriptomiques de Gasch *et coll.* [60] a montré que l'on pouvait combiner des conditions de perturbations internes du système (mutation) et des conditions environnementales différentes (exposition à 2 génotoxiques différents) et extraire des réponses spécifiques à chaque condition. Cette méthode n'est ni spécifique d'un organisme ni spécifique d'un type de stimulus. Notre analyse se base sur la dynamique du transcriptome, mais elle serait aussi transposable à d'autres types de données dynamiques comme par exemple des cinétiques du protéome ou des cinétiques de croissance cellulaires mesurées dans un grand nombre de conditions expérimentales différentes [322]. Dans ce dernier cas, la granularité de l'analyse n'atteindrait pas l'échelle du gène et ne nécessiterait pas la mise en œuvre de la 4ème étape de notre méthodologie. Le déroulement des autres étapes permettrait en revanche d'identifier des groupes d'expériences (traitements et perturbations) dans lesquels des populations de cellules ont des dynamiques de croissances similaires. L'étape de déduction logique permettrait quand à elle d'identifier automatiquement quels sont les traitements et les perturbations qui peuvent affecter la croissance des cellules. Ce type d'approche pourrait par exemple être appliqué dans le cadre de l'identification de gènes essentiels et la reconstruction de réseaux métaboliques.

Problème du passage à l'échelle des grand génomes. L'application de la méthodologie *XRegPath* à l'Homme par exemple pose un problème de passage à l'échelle. En effet, chez l'Homme, les dernières estimations font état d'environ 20000-25000 gènes codant pour des protéines. Toutes les données systémiques à l'échelle de ce génome deviennent difficiles à manipuler et à exploiter. Pour notre approche, l'étape critique concerne la classification des gènes à partir de leurs similarités respectives. Nous calculons et manipulons des matrices de similarité de taille $n \times n$, avec n le nombre de gènes. Chez la levure (6300 gènes environ) la taille de ces matrices ne pose pas de problèmes mais avec des matrices de l'ordre de 20000×20000 pour l'Homme il devient difficile de manipuler de si grands volumes de données tant du point de vue de la capacité mémoire que de celui du temps de calcul (problème de résolution des vecteurs propres lors de la classification spectrale). Une première solution consisterait à filtrer les données avant l'étape de classification. Pour nos données, des mesures de fluctuation d'expression (de 3 à 6 mesures par gène) avaient été réalisées pour tous les gènes avant irradiation. Nous pourrions mettre en place un test statistique qui permettrait de comparer les mesures d'expression des gènes post-IR aux mesures de fluctuations "normales" des gènes et éliminer les gènes dont les profils d'expression post-IR restent au niveau des fluctuations. Cette approche nécessite cependant de prendre en compte les mesures de fluctuations au moment de la conception du protocole expérimental et n'assure nullement que l'on pourra suffisamment réduire le nombre de gènes à classer pour pouvoir appliquer la méthodologie *XRegPath*. Une autre solution reviendrait à redéfinir la méthode de classification des gènes à partir de leurs profils d'expression en réduisant l'une des dimensions des matrices de similarité. Plutôt que de calculer des similarités entre couples de gènes nous pourrions les calculer entre profils d'expression de gènes et profils types de réponses. Ces profils types seraient définis à l'avance avec l'expertise du biologiste car ils dépendent de la nature des phénomènes analysés (réponse à un signal, analyse du cycle, etc...), du nombre de points de mesure, des intervalles et de la durée totale d'observation. Le travail ici consisterait d'abord à définir un bon codage pour représenter les profils des gènes et les profils des réponses "types" puis à utiliser une métrique appropriée au calcul des similarités.

Annexes

A Application de *XRegPath* à des données publiées : choix des paramètres de noyau, analyse de stabilité et choix de la taille de partition

A.1 Choix des paramètres de noyau

Selon la même procédure que celle présentée dans l'article I (voir page87), nous avons généré pour chaque condition d'analyse une série de matrices de similarités pour différentes valeurs du paramètre γ. Les valeurs optimales de γ pour les 6 conditions expérimentales choisies ont été déterminées à partir des histogrammes des distribution des similarités (voir la figure 4.3 comme exemple). Ces valeurs sont représentées dans le tableau 4.2.

Conditions	wt + MMS	*mec1* + MMS	wt + IR	*mec1* + IR	wt + MOCK	*mec1* + MOCK
γ optimal	0.35	0.3	0.22	0.22	0.3	0.4

TABLE 4.2 – *Valeurs des paramètres de noyau γ optimaux pour chaque condition expérimentale choisie d'après les travaux de Gasch et coll. [60].*

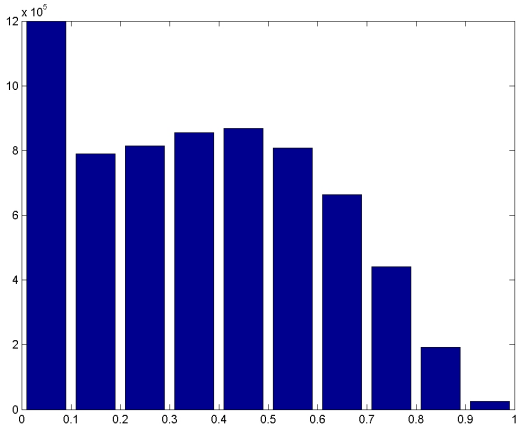


FIGURE 4.3 – Histogramme de la distribution des valeurs de noyau (similarités entre gènes) pour la condition *mec1* + IR et pour la valeur optimal $\gamma = 0.3$.

A.2 Analyse de stabilité et choix de la taille de partition

L'analyse de la stabilité des partitions nous a permis de définir un intervalle dans lequel choisir la taille de la partition optimale (voir figure 4.4). Ce choix d'une taille optimale $K = 13$ a été réalisé conjointement avec les biologistes selon les critères définis dans l'article I (page 87).

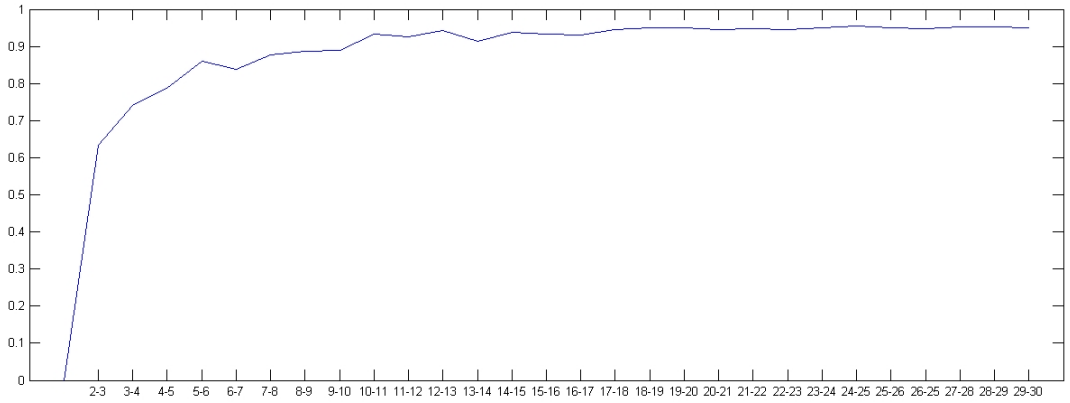


FIGURE 4.4 – Mesure de stabilité basée sur la comparaison de partitions de tailles croissantes. La courbe représente l'évolution des mesures de stabilité entre deux partitions de tailles croissantes. Cette mesure de stabilité est définie dans l'article I (page 87).

B Article III : A haploid-specific transcriptional response to irradiation in *Saccharomyces cerevisiae*

Published online November 30, 2005

Nucleic Acids Research, 2005, Vol. 33, No. 20 ***6635–6643***
doi:10.1093/nar/gki959

A haploid-specific transcriptional response to irradiation in *Saccharomyces cerevisiae*

G. Mercier, N. Berthault, N. Touleimat, F. Képès[1], G. Fourel[2], E. Gilson[2] and M. Dutreix*

CNRS-UMR 2027, Institut Curie, Bât. 110, Centre Universitaire, F-91405 Orsay, France, [1]Programme d'Épigénomique, Bât. G3, 93 rue Henri Rochefort, F- 91000 Evry, France and [2]Laboratoire de Biologie Moléculaire de la Cellule, l'Ecole Normale Supérieure de Lyon, CNRS-ENS UMR5161, 46 allée d'Italie, 69364 Lyon Cedex 07, France

Received August 31, 2005; Revised and Accepted October 26, 2005

ABSTRACT

Eukaryotic cells respond to DNA damage by arresting the cell cycle and modulating gene expression to ensure efficient DNA repair. We used global transcriptome analysis to investigate the role of ploidy and mating-type in inducing the response to damage in various *Saccharomyces cerevisiae* strains. We observed a response to DNA damage specific to haploid strains that seemed to be controlled by chromatin regulatory proteins. Consistent with these microarray data, we found that mating-type factors controlled the chromatin-dependent silencing of a reporter gene. Both these analyses demonstrate the existence of an irradiation-specific response in strains (haploid or diploid) with only one mating-type factor. This response depends on the activities of Hdf1 and Sir2. Overall, our results suggest the existence of a new regulation pathway dependent on mating-type factors, chromatin structure remodeling, Sir2 and Hdf1 and independent of Mec1 kinase.

INTRODUCTION

The cellular response to exogenous DNA damage involves a complex combination of cell cycle arrest, the modulation of gene expression and DNA damage repair, resulting in the survival or death of the cell. Diploid *Saccharomyces cerevisiae* strains (Mat*a*/Matα) are more resistant than haploid strains (Mat*a* or Matα) to gamma rays (1). The genetic basis of this difference remains poorly understood. Diploids and haploids differ in the expression of mating-type genes and the number of chromosome copies. Various cellular processes, including mating, meiosis and budding, are directly controlled by a/α mating-type, at the transcriptional level. Recent studies have also demonstrated the importance of mating-type status in the regulation of microtubule properties (2), the maintenance of cell wall integrity (3) and DNA repair by non-homologous end-joining (NHEJ) (4). Galitski *et al.* (5) investigated the contributions of mating-type and ploidy to gene expression in three isogenic sets of yeast strains differing only in terms of ploidy, which were subjected to whole-genome expression analysis. The results obtained confirmed the existence of both ploidy-dependent and mating-type-specific gene expression patterns under normal growth conditions. We used microarray analysis and gene reporter fusions to assess the contributions of ploidy and mating-type to the transcriptional response induced by irradiation.

Gamma irradiation generates various types of DNA damage, including double-strand breaks (DSBs). A single unrepaired DSB is deleterious for cells, as it may lead to genetic instability and the loss of chromosome fragments. Such damage may occur anywhere in the genome and may have a major effect on the general organization of chromosomes in the nucleus. The DSBs caused by ionizing radiation trigger G_2/M arrest before entry into mitosis, preventing the loss of chromosome fragments during division (6,7), whereas base modifications inhibiting DNA replication activate the S phase-progression checkpoint (8). Transduction of the resulting signals is thought to require the kinase cascade, which involves the activities of Mec1p, Rad53p, Chk1p and Dun1p [reviewed in Elledge *et al.* (9)] However, the transcription factors involved in the γ-induced response at the other end of the regulation cascade have not been identified.

One of the most important defense mechanisms against the lethal effects of DSB is the repair of broken DNA by homologous recombination (HR). The abolition of radiation resistance has been observed for a number of DNA repair mutants of the *RAD52* recombinational repair epistasis group (*RAD51*, *RAD52* and *RAD54*) (10), and for *RAD50*, *XRS2* and *MRE11*, affecting the resection of DSBs (11,12). The difference in radiosensitivity between haploids and diploids seems to result mostly from the lack of a template for HR during the G_1 and early S phases of the haploid cell cycle.

*To whom correspondence should be addressed. Tel: +33 1 69 86 71 86; Fax: +33 1 69 86 94 29; Email: marie.dutreix@curie.u-psud.fr

However, many studies have shown that genotype at the MAT locus also plays an important role in the response to irradiation, affecting DNA repair and the HR/NHEJ balance, (13–16). Diploid cells express the Mata1-Matα2 repressor, which turns off the transcription of a set of 'haploid-specific genes', including several components of the mating pheromone signaling pathway. NHEJ efficiency has been shown to be lower in diploid cells than in haploid cells (17,18). All the genes involved in controlling this balance have not been yet characterized. Recent studies have shown that *LIF1* and *LIF2* (*NEJ1*) are strongly regulated by mating-type, as the steady-state levels of these proteins are lower in diploid Mata/Matα strains than in haploid strains (4,19). However, mating-type heterozygosity is known to increase the frequency of HR (15,16) via Ku-dependent and -independent mechanisms (13). The available data therefore seem to indicate that both ploidy and mating-type locus affect the efficiency of DNA repair. The identification of all proteins induced by irradiation and subject to a/α regulation should further increase our understanding of the way in which the choice between repair pathways is controlled.

MATERIALS AND METHODS

Strains and culture conditions

The *S.cerevisiae* diploid FF 6053 (Mat*a*/α), and the haploids FF 18734 (Matα), FF 18733 (Mat*a*) and FF 18735 (Mat*a*/α) were used for transcriptional analysis. The FF 6053 diploid was obtained by mating two haploids (FF 18734 and FF 18733). FF 18735 was constructed by integration a plasmid (Yip5) encoding Matα into the FF 18733 haploid strain. The *hdf1* and *sir2* mutant strains are W303-1a (Matα) haploid derivatives. We analyzed URA3 gene silencing in haploid and diploid strains, using W303-1a derivatives containing a modified telomere VII-L, in which the ADH4 subtelomeric gene was replaced by the URA3 reporter gene and various portions of the X and Y′ element were inserted between the URA3 reporter gene and terminal telomeric DNA repeats (20). Yeast cells were grown exponentially in YPD medium at 30°C and oxygenated by shaking at 150 r.p.m. with a HT Infors AG shaker (Bottmingen, Switzerland).

Ionizing irradiation conditions and time-courses

Overnight exponential cultures were centrifuged, and the cell pellet was resuspended at a density of 10^9 cells/ml and irradiated (60 Gy/min and ^{137}Cs source) at room temperature in rich medium to minimize temperature and osmotic variations during treatment. Irradiated cells were plated directly on rich medium for survival analysis or immediately resuspended in rich medium at the original density for time-course experiments. Cells were irradiated at time 0, and samples were collected for microarray and cell cycle analysis at various times (0.1, 1, 2, 3, 4 and 5 h) after irradiation. Kinetic analysis was performed on strains exposed to a 200 Gy of ionizing radiation, which resulted in a cell survival rate of 25% for the two haploid strains (FF 18734 and FF 18733) and 75% for the diploid strain (FF 6053). We used DAPI staining, microscopy and FACS analysis, as described previously (21) to determine the duration of cell cycle arrest following irradiation. The transcriptional response was analyzed during this period.

Probe and microarray hybridization and data analysis

Total RNA was extracted from frozen samples by the hot phenol method. A fluorescently labeled first-strand cDNA was synthesized by RT, as described in Supplementary Data. For all microarray hybridizations, the fluorescent Cy-3-labeled cDNA control population was prepared from the same pool of total RNA extracted from five independent, exponentially growing cultures of the diploid strain (FF 6053). Hybridized microarrays were scanned with a Genepix 4000B machine (Axon Instruments). Fluorescence intensities for all spots were normalized using the location and scale normalization procedures described by Mercier *et al.* (22), details are provided in Supplementary Data.

As a unique pooled RNA sample (prepared from non-irradiated cultures of the diploid strain) was used as the reference in all experiments, we calculated a ratio by dividing the measured ratio for each irradiated haploid strain with the corresponding value for the same strain in the absence of irradiation. For this purpose, we prepared three independent normal growth cultures of each haploid strain for control experiments (0 Gy), and used the median ratio for these strains for the normalization of irradiation time-course data. Genes with expression levels differing between irradiated and non-irradiated samples by a factor of at least two for at least one time point filtering of the time-course experiment were identified as irradiation-regulated (IR) genes. Pairwise mean linkage clustering analysis was performed with Cluster (using uncentered Pearson correlation coefficients) and visualized using Treeview (23).

Measurement of telomere position effect (TPE) by analysis of reporter gene expression

TPE was assessed by analyzing variegated expression of the *URA3* gene. Cells with a repressed *URA3* gene were selected as colonies growing in the presence of 5-FOA (SC + 5-FOA), which is toxic to cells expressing a functional *URA3* gene product (24). We then distinguished ura- mutants and silenced cells by replica plating on medium lacking uracil (SC-URA). Cells growing on both SC + 5-FOA and SC-URA media were considered to have a repressed *URA3* gene. We compared the TPE in haploid and diploid strains in the absence of irradiation, using various *URA3* reporter gene constructs (detailed in Figure 4). Drop assays were performed with the URA construct, by spotting serial dilutions of three independent overnight cultures in SC liquid medium on to SC, SC-URA and SC + 5-FOA plates. The effect of irradiation dose was assessed by dilution assay for three exponential independent cultures of each strain, irradiated at different doses, serially diluted and plated on specific media to determine the percentage of cells with a repressed *URA3* gene. In parallel, we evaluated the survival rate of irradiated cells by calculating the ratio of viable cells in irradiated cultures to viable cells in non-irradiated cultures.

Online supplementary data

Details of probe, microarray hybridization protocols and data analysis are given provided in the Supplementary Materials and Methods. Haploid-specific (HS-IR) genes and their function are listed in Supplementary Table S1. A statistical

analysis of the effect of Gasch mutants on HS-IR gene expression is given in Supplementary Table S2. Supplementary Figure S1 shows Treeview expression analysis of the HS-IR genes in 300 mutants and Supplementary Figure S2 shows the chromosomal location of HS-IR genes. The raw data are available at the following URL http://microarrays.curie.fr/.

RESULTS

Differences in global responses to irradiation between haploids and diploids

We compared global gene expression responses to ionizing radiation between *S.cerevisiae* haploids and diploids by irradiating three isogenic strains: the Mat*a*/α diploid (FF 6053 strain), the Mat*a* haploid (FF 18733 strain) and the Matα haploid (FF 18734 strain). The patterns of gene expression induced by irradiation were deduced by comparing the patterns of expression of a given strain before and after irradiation. The three isogenic strains were exposed to a 200 Gray (Gy) dose of ionizing radiation. Cell survival rates were lower in both haploids (25%) than in the diploid (75%). The irradiated cells stopped dividing for about 4 h and then resumed mitosis in all three strains (data not shown). We studied the transcriptional response of irradiated cells during the full recovery period, by analyzing mRNA from samples taken immediately after irradiation and every hour for the next 5 h.

In all three strains, ionizing radiation led to significant changes in the gene expression program, with the relative abundance of about 1400 genes differing by a factor of two or more between irradiated and non-irradiated cultures. Most of these IR genes displayed rapid and strong changes in expression—a typical stress response feature that has already been reported after γ-irradiation and treatment with various other types of DNA-damaging agent (25–27). We compared the lists of IR genes for the three strains (Figure 1). Surprisingly, the two haploids had twice as many IR genes in common with each other (595) than in common with the diploid strain (291 and 265). Only 124 genes were found to be induced (or repressed) in all three strains following irradiation. Many of the genes known to be involved in inducible DNA damage repair, such as *RAD51*, *RAD54*, *RNR2*, *RNR4*, *HUG1* and *RFA2* (21,28–30), were induced in all three strains. We compared our data with published results obtained after MMS and ionizing radiation treatments in a Mat*a* strain (25). The eight genes displaying specific induction in response to DNA-damaging treatments (*RNR2*, *RNR4*, *RAD51*, *RAD54*, *PML2*, *YER004W* and *YBR070C*) were also found to be induced in our experimental conditions in the Mat*a* haploid strain. *DIN7* was the only gene of the 'DNA Damage Signature' set [see Gasch *et al.* (25)] unaffected by irradiation in our experiments.

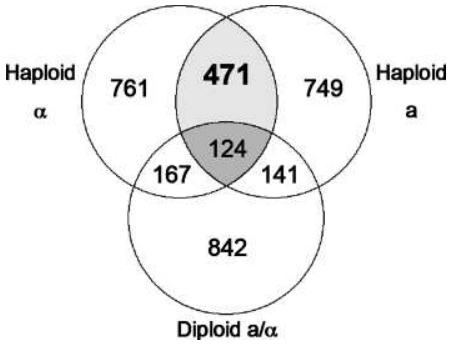


Figure 1. Venn Diagrams comparing the HS-IR genes modulated in the various strains. Circles indicate the number of genes showing changes in expression by a factor of at least two after irradiation in the haploids FF 18733 (Matα), FF 18734 (Matα) and the diploid FF 18735 (Mata/α). The intersections of the circles correspond to genes induced in at least two strains (i.e. the strains corresponding to the intersecting circles). Numbers indicate the number of genes in each group.

We focused on the 471 IR genes displaying changes in transcription after irradiation in both haploids, but not in the diploid strain. We found that 278 of these HS-IR genes were induced and 193 were repressed (Figure 1; Supplementary Tables S1).

Promoter analysis of HS-IR genes

We tried to identify transcription factors potentially involved in the regulation of HS-IR genes by analyzing the 800 bp sequence directly upstream from the coding region for consensus binding sites for known or unknown transcription factors. The Pbox, Qbox and PRE elements and sequences recognized by mating-type factor heterodimers were not significantly more frequent in the set of HS-IR genes than in the genome as a whole. This result suggests that the regulation of this response by mating-type factors is indirect. Only two sequences were found to occur at high frequency in the promoters of the 278 induced HS-IR genes: 114 genes (41%, versus 22% for the genome as a whole) contained the CTCATC sequence recognized by Rfa2. The binding of Rfa2p to upstream sequences has been shown to repress the expression of some repair genes and is decreased by UV irradiation or MMS treatment, thereby leading to the induction of these repair genes (31). We found that 81 of the 278 induced HS-IR genes possessed upstream regions containing the ATGAGC sequence, which has no known binding factor. The promoter regions of the 193 repressed HS-IR genes presented no overrepresentation of any specific sequence other than the frequent occurrence of G-rich sites.

Subtelomeric and even distributions of HS-IR genes on chromosomes

Visual inspection showed that an unexpectedly high proportion of induced HS-IR genes were located near chromosome ends ($\chi^2 = 114.6$, $P < 0.0001$; Supplementary Figure S1). Indeed, 18% of the 278 induced HS-IR genes were located within 20 kb of a telomere (subtelomeric). About one quarter (51/218) of the subtelomeric genes on our microarrays were induced by irradiation (Supplementary Table S1). The 51 subtelomeric HS-IR genes induced were evenly spread over 23 of the 32 chromosomes extremities, indicating that subtelomeric gene derepression is a general process affecting most chromosome ends (Supplementary Figure S1).

Genes controlled by the same sequence-specific transcription factor tend to be spaced at regular intervals along chromosome arms (32). We analyzed the distribution of HS-IR genes by calculating the distance between pairs of HS-IR genes from the same chromosome arm. In this analysis of coexpressed genes, eight yeast chromosome arms had too low

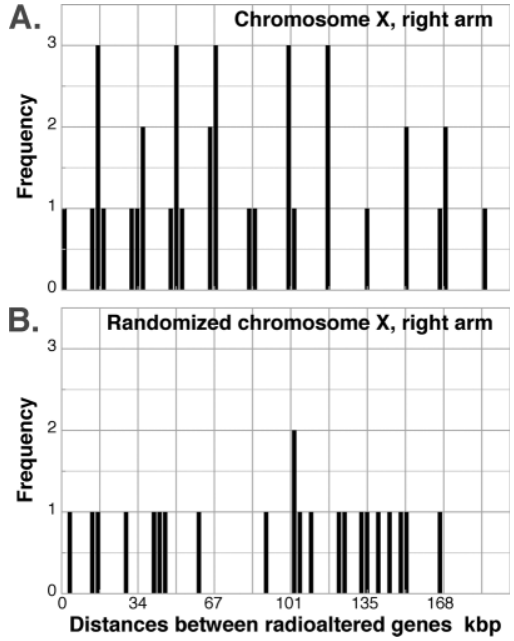


Figure 2. Distribution of distances separating radioaltered HS-IR genes along the right arm of yeast chromosome X. (**A**) Distances were measured between the starting points of the coding sequences of all gene pairs for the complete list of HS-IR genes. These distances are shown here on a bar graph, with a grid interval of 16 850 bp. Bar width (or 'bin size' for data discretization) is 2500 bp i.e. about the length of one yeast gene allowing fine distribution analysis. Varying bin size had no effect on the conclusions drawn. (**B**) As for (A), following the random attribution of gene positions. Gene content, chromosome length and target list are as in natural chromosome X. For each gene on this chromosome arm of length *L*, the randomization process replaces the start position with a random integer between 1 and *L*. Calculations used Microsoft® Excel VBA routines. The routines and data are available upon request.

a density of HS-IR genes for any firm conclusion to be drawn. Six of the remaining 24 arms displayed weak periodicity, and 18 displayed clear periodicity. For example, the HS-IR genes on the right arm of chromosome X tended to be regularly spaced, and were separated by 16 850 bp or by multiples of 16 850 bp (Figure 2A; grid step 16 850 bp). This even spacing is not consistent with the random attribution of gene positions (Figure 2B). Different periods were observed for different chromosome arms, as reported previously for coregulated genes (32). Thus, most HS-IR genes appear to be controlled by a few molecular factors involved in regular nuclear organization.

Chromatin modifying activities regulate HS-IR genes

We investigated whether all the HS-IR genes studied were regulated by the same pathway, by comparing the effect on their expression of deletions of genes encoding various regulatory proteins. We investigated microarray data for about 300 mutants (33). As expected for genes regulated by the same pathway, 134 of the 193 repressed HS-IR genes displayed similar sensitivities to a large set of mutations (Supplementary Figure S2-A), with basal expression levels decreasing for 32 mutants and increasing for 113 mutants. The number of HS-IR genes showing variation of expression was estimated and compared with the total number of genes showing variation of expression in each mutant. Only mutants giving a $P > 0.005$ in a hypergeometric test were considered (listed in Supplementary Table S2). Analysis of the molecular functions affected in mutants displaying specific HS-IR gene expression changes showed that most directly or indirectly involved chromatin remodeling and/or silencing. The largest changes in expression of repressed HS-IR genes were observed in mutants with impairments affecting chromatin (e.g. *sir4*, *rpd3*, *sin3*, *hat2*, *cyc8*, *hst3*, *ubp10* and *tup1*). The clustering of the induced HS-IR genes was consistent with a complex pattern of regulation, with very few common regulators (Supplementary Figure S2-B). However, most of the induced HS-IR genes also displayed significant changes in expression in mutants with impaired chromatin assembly and chromatin modifications (*sir2*, *sir3*, *hdf1*, *iswi1* and *isw2)* and DNA topology *(top1* and *top3*). Inactivation of the *TUP1* and *SSN6* genes encoding proteins acting as a transcription factor complex sensitive to chromatin structure (34) significantly increased the expression of repressed HS-IR genes, suggesting that irradiation may facilitate the recruitment of these repressors to the regulatory regions of HS-IR genes.

HS-IR gene expression and silencing are controlled by mating-type

HS-IR genes were identified as genes displaying changes in expression after irradiation in haploids but not in diploids. These genes seemed to be sensitive to chromatin regulation and some were subject to telomeric chromatin silencing. We investigated the contribution of mating-type status to control of the general transcriptional response to irradiation by analyzing the expression of HS-IR genes in a pseudo-diploid strain: a Mat*a* haploid strain, expressing the α factor. Control experiments involving microarray analysis confirmed that a combination of the a1 and α2 factors resulted in the repression of haploid-specific genes. As expected, the *STE2*, *STE6*, *MFA1*, *MFA2*, *AGA2*, *ASG7*, *Mfalpha1*, *Mfalpha2*, *STE3*, *FUS3* and *RME1* genes displayed similar levels of expression in the Mat*a*/α pseudo-diploid and in the diploid. Most of the IR genes shown to be induced after irradiation in haploids but not in diploids showed no induction in the pseudo-diploid strain (Figure 3). In contrast, 65% of the 124 genes induced (or repressed) in both haploids and diploids (see Figure 1), were also induced in the haploid expressing both mating-types (data not shown).

As the expression of HS-IR genes, including subtelomeric genes (Figure 3), seems to be controlled by mating-type factors, we investigated the effect of ploidy on silencing by means of reporter gene studies. Telomeric silencing at native ends has been reported to vary with gene location, depending on the combination of X and Y′ elements in yeast (35). We confirmed that this was the case in a TPE assay in strains carrying different subtelomeric sequences between the *URA3* gene and the TG_{1-3} repeat. However, for all constructs, the *URA3* gene was less strongly silenced in the diploid strain than in the haploid strain (Figure 4). Diploid-associated derepression was more pronounced in reporter constructs bearing the part of the X or Y′ element immediately adjacent to the telomere, suggesting that the natural subtelomeric sequences are involved in modulating TPE as a function of ploidy. Thus, TPE and microarray

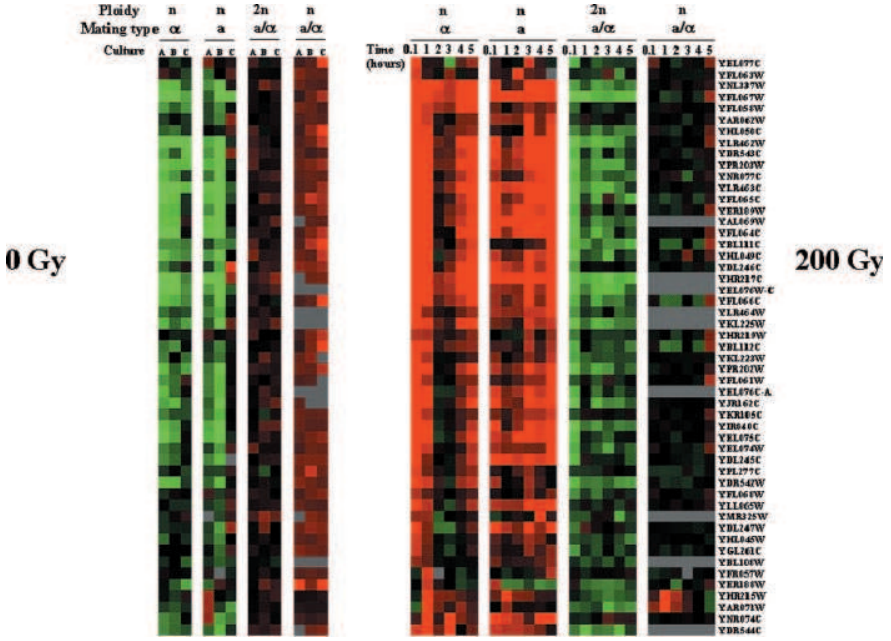


Figure 3. Analysis of the expression of subtelomeric HS-IR genes. Expression of the HS-IR genes in the four strains: the haploids (Mat α), (Mat a), [Mat a(α)] and the diploid (Mat a/α)—is shown with TreeView (23). Only data for telomeric genes are reported. The full analysis of the 471 HS-IR mutants is presented in Supplementary Figure S1. Panel A, non-irradiated cells: For each strain, RNA from three independent cultures was analyzed as described in materials and methods. Panel B, irradiated cells: RNA levels for the various genes were determined immediately after irradiation with 200 Gy (time 0.1) and after 1 h, 2 h, 3 h, 4 h and 5 h of incubation. Ratios were calculated with respect to the median of the three measures in non-irradiated cells.

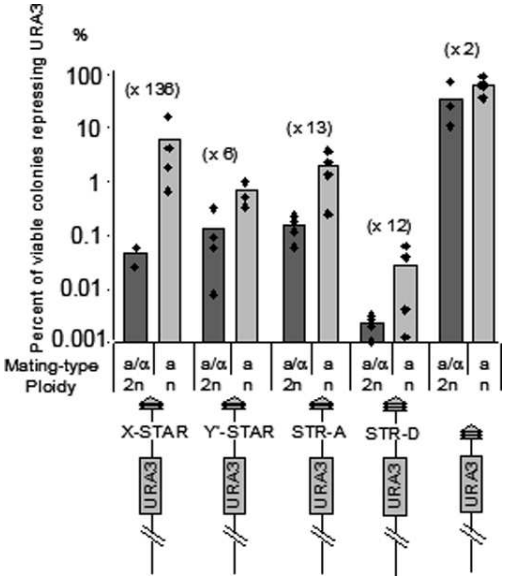


Figure 4. Measurement of TPE, using an artificial telomere-proximal gene, URA3, at tel VII-L. Various URA3 constructs were tested for TPE in haploid (gray rectangles) and heterozygous diploid (black rectangles) strains. Histogram bars represent the mean values obtained for a given strain. Each diamond indicates the proportion of colonies displaying URA3 gene repression (ratio of colonies growing in 5-FOA-SC versus SC).

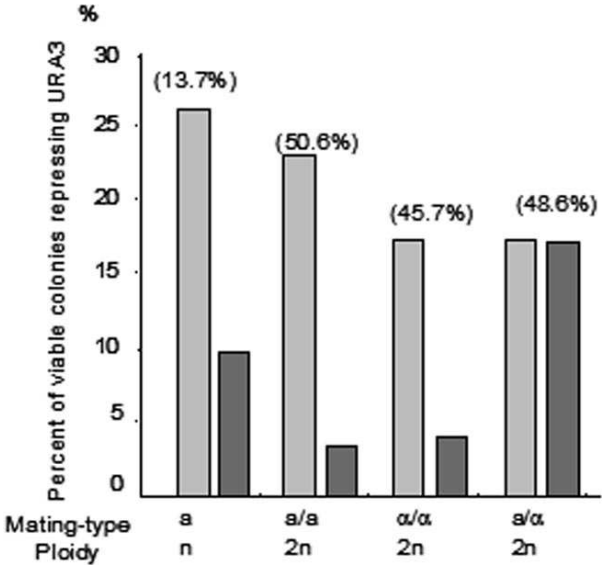


Figure 5. Effect of gamma rays on TPE in haploids and diploids with different mating-types. Four isogenic strains varying in ploidy and mating genotype, as indicated on the abscissa, were analyzed for TPE. The percentage of viable cells displaying URA3 repression in non-irradiated (dark gray) cultures and in cultures exposed to 200 Gy irradiation (light gray) was estimated by dilution assay on three independent cultures (diamonds). The cell survival rate is indicated in brackets.

analyses highlighted constitutive differences in silencing between diploid and haploid strains. As shown by microarray data (Figure 3), the silencing of subtelomeric genes measured by TPE (Figure 5) was not significantly affected by irradiation in diploids. The difference in silencing between diploids and haploids may be due to differences in the number of chromosomes or to differential regulation by mating-type factors, as shown for the HS-IR genes. We constructed diploid strains expressing only one mating-type to determine which of these two possibilities applied. The constitutive TPE in diploids

expressing only one mating-type was weaker than that in diploids, indicating that TPE may be sensitive to chromosome number. This would not be entirely surprising as the silencing protein Sir3 has already been shown to be present in limiting amounts in haploid cells (35,36). Interestingly, unlike (*a*/α) diploids, diploid strains expressing only the *a* or the α factor displayed complete derepression of the subtelomeric reporter gene after irradiation (Figure 5). This result is consistent with microarray data indicating that the response to irradiation, including the silencing switch-off, is inhibited by the expression of both mating-type factors in diploids.

The HS-IR response requires Ku70 and Sir2 but not Mec1

We further characterized the chromatin-dependent response to DNA damage by analyzing the response of *hdf1* and *sir2* mutant cells to 200 Gy of ionizing radiation over a 5 h period. These mutants displayed a much weaker overall transcriptional response than wild-type cells. Only 646 and 517 genes in the *hdf1* and *sir2* mutants, respectively, displayed radiomodulation, with only 185 genes being radiomodulated in both mutants. Only 6% of the genes insensitive to *hdf1* and *sir2* were HS-IR genes. Thus, the *hdf1* and *sir* mutations seem to block preferentially change in expression of the 471 HS-IR genes. We checked that the dependence of the gene responses to hdf1 and sir2 activity was specific to HS-IR genes, by carrying out the same analysis with the 124 genes induced in both diploids and haploids. This group was significantly less sensitive to the deletion of *hdf1* and *sir2* deletion, with only 19% (23/124) of the genes tested radiomodulated in the mutants ($\chi^2 = 14{,}2$, $P < 0.001$). These results suggest that the Sir2 and Ku70 proteins play specific roles in regulating the haploid-specific transcriptional response to irradiation.

We performed the same analysis on data published by Gasch *et al*. These authors studied the kinetics of gene expression after irradiation, for a wild-type strain and a *mec1* mutant. They found that 1369 genes were regulated by irradiation in the wild-type strain, versus only 962 in the *mec1* mutant. We found that 544 genes were radiomodulated in both strains, indicating that the expression of these genes was not entirely controlled by Mec1 kinase. We analyzed the effect of *mec1* mutation on the expression of HS-IR genes. Most (104) of the 169 HS-IR genes radiomodulated in the wild-type were also radiomodulated in the *mec1* mutant. This high proportion of Mec1-independent responses differed significantly from that reported by Gasch (544 of a total of 1369 radiomodulated genes were radiomodulated in both strains; $\chi^2 = 10.69$, $P < 0.005$). In contrast, only 22 of the 124 genes radiomodulated in our three wild-type strains were radiomodulated in the *mec1* mutant. This proportion does not differ significantly from that for the data published by Gasch ($\chi^2 = 0.9$, $P < 0.25$). Thus, *mec1* deletion has effects opposite to those of *hdf2* and *sir2* deletions, affecting HS-IR gene expression only weakly and having a strong effect on the haploid-independent response to irradiation. Global analysis of the response to irradiation in the *hdf2*, *sir2* and *mec1* mutants suggested that the response of HS-IR genes to irradiation depends on mating-type factors, Ku and Sir but not Mec1. This regulation pathway is different from the known Mec1-controlled pathway of radiomodulation in haploids and diploids and seems to be insensitive to mating-type factors, Ku and Sir proteins. Most of the genes responding to DNA damage (e.g. *RNR2*, *RNR4*, *RAD51*, *RAD54* and *RFA2*) studied by other laboratories are regulated by the Mec1 pathway.

DISCUSSION

We identified, by means of transcriptome analysis and reporter gene studies, a cellular response to irradiation dependent on chromatin structure and mating-type factors. Gene silencing has been shown to result from the inhibition of RNA pol II transcription activity by a specific compact chromatin structure. It requires the binding to histone tails of three unrelated proteins: Sir2, Sir3 and Sir4 initially recruited by Rap1. One possible mechanism accounting for the decrease in silencing after irradiation involves the repression by irradiation of genes encoding these proteins, resulting in a loosening of the compact structure of chromatin. The *RAP1* and *SIR3* telomeric structural genes displayed no change in expression following irradiation. The *SIR4* gene displayed a continuous increase in expression over the period analyzed. However *SIR4* overexpression was expected to increase silencing rather decrease it as observed, based on data for the overexpression of *SIR3* (35,36).

Many authors have highlighted the similarity between DSB and telomeric termini. Both bind proteins involved in the NHEJ repair pathway, such as the Ku heterodimer and the Mre11/Rad50/Xrs2 nuclease/helicase complex, which is thought to be involved in break religation. The Sir proteins, which are associated with transcriptional silencing at subtelomeric and mating-type loci, seem to be directly or indirectly involved in DNA damage repair. Sir4 interacts physically with Ku70, and mutations of the Sir complex result in deficiencies in the repair of linear plasmids (37–40). At both telomeres and DSBs, NHEJ and HR proteins compete in the maintenance of chromosome integrity. Ku proteins prevent HR at telomeres (41), whereas this process is the primary means of repair at DSBs (see for review van den Bosch, (42,43). The proteins binding around each site may be responsible for selecting the mechanism activated. DNA repair by HR is enhanced by mating-type heterozygosity. This increase in the rate of HR repair was not observed in an *a*/α diploid with only one mating-type, suggesting that the presence of a homologous chromosome is not sufficient to increase the rate of HR (13–15,18). Moreover, the DNA repair defect caused by the *rad51-K191R* mutant protein, which is responsible for a partial defect in ATP hydrolysis, is abolished in diploids and by mating-type heterozygosity in haploids (44). The effect of mating-type heterozygosity, which enhances repair, is not restricted to the HR pathway, because end-joining activity is also repressed in diploids (13,16,17). The abolition of radiation sensitivity in diploids has been observed for a number of DNA repair mutants (10), and in the case of the *rad18* and *rad55* diploids, is due to mating-type heterozygosity rather than ploidy (15,45). Vaillant *et al*. showed that NHEJ regulation involves the control of Lif2 protein production, which is repressed in diploids.

An analysis of published microarray data identified no known repair genes more strongly expressed in diploid strains

expressing the two mating-type alleles than in diploids expressing only one mating-type allele (46). Our data on the transcriptional response induced by irradiation showed that very few of the many genes displaying differential expression in haploids were directly involved in DNA damage repair (see Supplementary Tables S1 and S2). We performed an extensive ontology analysis on the microarray data, estimating statistical significance by means of a hypergeometric method. The processes retained were those corresponding to a number of genes significantly higher than would be expected on the basis of chance alone, considered the whole set of genes analyzed. This analysis revealed that induced HS-IR genes are involved primarily in ribosome biogenesis (23 genes, $\chi^2 = 15$), rRNA metabolism and RNA processing (28 genes, $\chi^2 = 6.5$), whereas repressed HS-IR genes are primarily involved in energy reserve metabolism (7 genes, $\chi^2 = 20.68$), mitochondrial electron transport (5 genes, $\chi^2 = 40.75$) and response to copper or desiccation (5 genes, $\chi^2 = 10,33$). Birell *et al.* (47) in their analysis of the transcriptional response to various DNA-damaging agents (including ionizing radiation), found no relationship between the genes required for survival following exposure to DNA-damaging agents and the genes displaying an increase in transcription after exposure. We identified 10 genes (*QCR9*, *QCR10*, *COX5B*, *QCR8*, *CYT1*, *GRX1*, *CTT1*, *SOD1*, *SOD2*, *CUP1-1* and *CUP1-2*) involved in oxidative phosphorylation, oxidative stress and Cu^{++} homeostasis that were induced after irradiation. The genes involved in these processes were recently shown to be induced by continuous exposure to low doses of ionizing radiation (22). Their contribution to the survival of irradiated cells remains unclear but it is possible that they act by regulating the free radical pool in the cell.

Our results demonstrate that chromatin structure is controlled by mating-type heterozygosity. Silencing at telomeric ends was weaker in diploids than in haploids, as shown by assessments of the expression of all genes located at subtelomeric positions and reporter gene fusions. In our reporter gene studies, we observed this ploidy dependence in different genetic backgrounds and various telomeric sequence combinations. The low level of silencing in diploids could be a consequence of saturating some silencing proteins by doubling the number of binding sites on chromosomes in diploids. However, we do not favor this hypothesis that does not explain the low level of silencing of most subtelomeric genes in the haploid strain expressing the two mating-types (Figure 3). In our search for *trans*-acting elements accounting for this difference, we found that several genes encoding proteins playing a direct or indirect role in chromatin structure displayed differences in expression in haploid and diploid cells. The *RSC6*, *ELP3* and *SMC3* genes, encoding proteins involved in chromatin remodeling, were constitutively and more strongly expressed in diploids. The *TEL2* and *EST2* genes, encoding proteins involved in telomere length regulation, were also more strongly expressed in basal conditions in diploids than in haploids. *TBF1* was also expressed more strongly in the diploid. Tbf1 displays insulating capacity, and binds to promoters and in subtelomeric anti-silencing regions (STARs) throughout the yeast genome (48,49). The interaction between Tbf1p and telomeres leads to a loss of silencing at these chromosomal loci (49). This observation is correlated with weaker telomeric silencing in the diploid strain than in haploid cells. The low level of gene silencing observed in diploids was also found in diploids expressing only one mating-type factor. However, unlike diploids heterozygous for mating-type, these strains displayed significant derepression after irradiation, like the haploid parent. Conversely, microarray analysis indicated that haploids expressing both mating-types showed no significant change in subtelomeric gene expression after irradiation (Figure 4). Thus, mating-type heterozygosity prevents chromosome remodeling after irradiation. Most studies on the expression of telomere-proximal genes have been carried out in a haploid background (50–52). Such a difference in the silencing status of subtelomeric regions between the haploid and diploid states has never before been described. Affecting silencing by inactivating Sir2 protein or by inducing chromatin remodeling by irradiation would have the direct consequence to suppress the HML and HMR loci repression leading to expression of both mating-type cassettes as in pseudo-diploids. Thus, based on our data and published results, it is difficult to determine the respective role of mating-type and silencing in the HS-IR response regulation. We propose that irradiation disturbs the silencing chromatin all over the chromosomes leading to transient expression of most of the genes under its control, including cassettes at HML and HMR loci. The expression of both mating-type in the irradiated haploid would decrease general silencing as observed for subtelomeric genes silencing in non-irradiated diploids and a(α) pseudo-diploid, and thereby delay silencing restoration and extend the HS-IR response. Actually whereas genes induced in haploids and diploids show a very rapid and transient induction of expression, most of the HS-IR genes remain overexpressed during all the cell division arrest.

The loss of TPE for artificial telomere-proximal genes has been shown to be concomitant with checkpoint-dependent delocalization of the heterochromatin structural proteins Sir1-4, Rap1 and Ku following DNA damage induced by various agents (EcoRI or HO endonuclease, MMS or bleomycin treatment) (53,39). Martin also showed, by chromatin immunoprecipitation (ChIP), that Sir3, Sir4 and Ku80 were redistributed from telomeric DNA to damaged sites. Thus, the local loss of telomeric proteins due to DNA damage may lead to the derepression of subtelomeric genes. The loss of heterochromatic structure at telomeres therefore appears to be a response to DNA damage. This response seems to be partially controlled by the kinases (Mec1 and Rad53) playing key roles in the checkpoint response to DNA-damaging treatments. For example, the redistribution of Sir3 after DNA-damaging treatment depends on Mec1, but not the Rad53 or Tel1 checkpoint proteins (39,53,54). However, Rad53 contributes to genome stability independently of Mec1, by preventing the damaging effects of excess histones, both during normal cell cycle progression and in response to DNA damage (55). Analysis of our data and those of Gasch *et al.* indicated that Sir2 and Hdf1 controlled the response to irradiation of HS-IR genes, whereas Mec1 did not. The mechanism underlying the dependence of this response on mating-type factors remains to be demonstrated.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank the 'Consensus' research group and F. Fabre for stimulating discussions, and the Genopole® and the DRIDF. We also thank S. Makhzami and M. Pierre for technical assistance. This work was supported by *l'Institut National de Recherche et de Sécurité* (convention no. 5011888), the CNRS, the Institut Curie, the *Association pour la Recherche sur le Cancer* (#5659) and the *Ligue Nationale Contre le Cancer*. Funding to pay the Open Access publication charges for this article was provided by Institut Curie.

Conflict of interest statement. None declared.

REFERENCES

1. Latarjet,R. and Ephrussi,B. (1949) Courbes de survie de levures haploïdes et diploïdes soumises aux rayons X. *C R Acad. Sci. Gen.*, **229**, 306–308.
2. Steinberg-Neifach,O. and Eshel,D. (2002) Heterozygosity in MAT locus affects stability and function of microtubules in yeast. *Biol. Cell*, **94**, 147–156.
3. Verna,J. and Ballester,R. (1999) A novel role for the mating-type (MAT) locus in the maintenance of cell wall integrity in *Saccharomyces cerevisiae*. *Mol. Gen. Genet.*, **261**, 681–689.
4. Frank-Vaillant,M. and Marcand,S. (2001) NHEJ regulation by mating-type is exercised through a novel protein, Lif2p, essential to the ligase IV pathway. *Genes Dev.*, **15**, 3005–3012.
5. Galitski,T., Saldanha,A.J., Styles,C.A., Lander,E.S. and Fink,G.R. (1999) Ploidy regulation of gene expression. *Science*, **285**, 251–254.
6. Weinert,T. (1998) DNA damage checkpoints update: getting molecular. *Curr. Opin. Genet. Dev.*, **8**, 185–193.
7. Weinert,T.A. and Hartwell,L.H. (1988) The RAD9 gene controls the cell cycle response to DNA damage in *Saccharomyces cerevisiae*. *Science*, **241**, 317–322.
8. Paulovich,A.G. and Hartwell,L.H. (1995) A checkpoint regulates the rate of progression through S phase in *S. cerevisiae* in response to DNA damage. *Cell*, **82**, 841–847.
9. Elledge,S.J., Winston,J. and Harper,J.W. (1996) A question of balance: the role of cyclin-kinase inhibitors in development and tumorigenesis. *Trends Cell Biol.*, **6**, 388–392.
10. Saeki,T., Machida,I. and Nakai,S. (1980) Genetic control of diploid recovery after gamma-irradiation in the yeast *Saccharomyces cerevisiae*. *Mutat. Res.*, **73**, 251–265.
11. Bressan,D.A., Baxter,B.K. and Petrini,J.H. (1999) The Mre11-Rad50-Xrs2 protein complex facilitates homologous recombination-based double-strand break repair in *Saccharomyces cerevisiae*. *Mol. Cell Biol.*, **19**, 7681–7687.
12. Belotserkovskii,B.P. and Zarling,D.A. (2002) Peptide nucleic acid (PNA) facilitates multistranded hybrid formation between linear double-stranded DNA targets and RecA protein-coated complementary single-stranded DNA probes. *Biochemistry*, **41**, 3686–3692.
13. Clikeman,J.A., Khalsa,G.J., Barton,S.L. and Nickoloff,J.A. (2001) Homologous recombinational repair of double-strand breaks in yeast is enhanced by MAT heterozygosity through yKU-dependent and -independent mechanisms. *Genetics*, **157**, 579–589.
14. Fasullo,M., Bennett,T. and Dave,P. (1999) Expression of *Saccharomyces cerevisiae* MATa and MAT alpha enhances the HO endonuclease-stimulation of chromosomal rearrangements directed by his3 recombinational substrates. *Mutat. Res.*, **433**, 33–44.
15. Heude,M. and Fabre,F. (1993) a/alpha-control of DNA repair in the yeast *Saccharomyces cerevisiae*: genetic and physiological aspects. *Genetics*, **133**, 489–498.
16. Morgan,E.A., Shah,N. and Symington,L.S. (2002) The requirement for ATP hydrolysis by *Saccharomyces cerevisiae* Rad51 is bypassed by mating-type heterozygosity or RAD54 in high copy. *Mol. Cell Biol.*, **22**, 6336–6343.
17. Astrom,S.U., Okamura,S.M. and Rine,J. (1999) Yeast cell-type regulation of DNA repair. *Nature*, **397**, 310.
18. Lee,S.E., Paques,F., Sylvan,J. and Haber,J.E. (1999) Role of yeast SIR genes and mating-type in directing DNA double-strand breaks to homologous and non-homologous repair paths. *Curr. Biol.*, **9**, 767–770.
19. Valencia,M., Bentele,M., Vaze,M.B., Herrmann,G., Kraus,E., Lee,S.E., Schar,P. and Haber,J.E. (2001) NEJ1 controls non-homologous end joining in *Saccharomyces cerevisiae*. *Nature*, **414**, 666–669.
20. Fourel,G., Revardel,E., Koering,C.E. and Gilson,E. (1999) Cohabitation of insulators and silencing elements in yeast subtelomeric regions. *EMBO J.*, **18**, 2522–2537.
21. Mercier,G., Denis,Y., Marc,P., Picard,L. and Dutreix,M. (2001) Transcriptional induction of repair genes during slowing of replication in irradiated *Saccharomyces cerevisiae*. *Mutat. Res.*, **487**, 157–172.
22. Mercier,G., Berthault,N., Mary,J., Peyre,J., Antoniadis,A., Comet,J.P., Cornuéjols,A., Froidevaux,C. and Dutreix,M. (2004) Biological detection of low radiation by combining results of two microarray analysis methods. *Nucleic Acids Res.*, **32**, e12.
23. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
24. Boeke,J.D., LaCroute,F. and Fink,G.R. (1984) A positive selection for mutants lacking orotidine-5′-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance. *Mol. Gen. Genet*, **197**, 345–346.
25. Gasch,A.P., Huang,M., Metzner,S., Botstein,D., Elledge,S.J. and Brown,P.O. (2001) Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol. Biol. Cell*, **12**, 2987–3003.
26. Jelinsky,S.A., Estep,P., Church,G.M. and Samson,L.D. (2000) Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes. *Mol. Cell Biol.*, **20**, 8157–8167.
27. Jelinsky,S.A. and Samson,L.D. (1999) Global response of *Saccharomyces cerevisiae* to an alkylating agent. *Proc. Natl Acad. Sci. USA*, **96**, 1486–1491.
28. Basrai,M.A., Velculescu,V.E., Kinzler,K.W. and Hieter,P. (1999) NORF5/HUG1 is a component of the MEC1-mediated checkpoint response to DNA damage and replication arrest in *Saccharomyces cerevisiae*. *Mol. Cell Biol.*, **19**, 7041–7049.
29. Elledge,S.J. and Davis,R.W. (1990) Two genes differentially regulated in the cell cycle and by DNA-damaging agents encode alternative regulatory subunits of ribonucleotide reductase. *Genes Dev.*, **4**, 740–751.
30. Kiser,G.L. and Weinert,T.A. (1996) Distinct roles of yeast MEC and RAD checkpoint genes in transcriptional induction after DNA damage and implications for function. *Mol. Biol. Cell*, **7**, 703–718.
31. Schramke,V., Neecke,H., Brevet,V., Corda,Y., Lucchini,G., Longhese,M.P., Gilson,E. and Geli,V. (2001) The set1 Delta mutation unveils a novel signaling pathway relayed by the Rad53-dependent hyperphosphorylation of replication protein A that leads to transcriptional activation of repair genes. *Genes Dev.*, **15**, 1845–1858.
32. Kepes,F. (2003) Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. *J. Mol. Biol.*, **329**, 859–865.
33. Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
34. Davie,J.K., Trumbly,R.J. and Dent,S.Y. (2002) Histone-dependent association of Tup1-Ssn6 with repressed genes *in vivo*. *Mol. Cell Biol.*, **22**, 693–703.
35. Pryde,F.E. and Louis,E.J. (1999) Limitations of silencing at native yeast telomeres. *EMBO J.*, **18**, 2538–2550.
36. Renauld,H., Aparicio,O.M., Zierath,P.D., Billington,B.L., Chhablani,S.K. and Gottschling,D.E. (1993) Silent domains are assembled continuously from the telomere and are defined by promoter distance and strength, and by SIR3 dosage. *Genes Dev.*, **7**, 1133–1145.
37. Boulton,S.J. and Jackson,S.P. (1998) Components of the Ku-dependent non-homologous end-joining pathway are involved in telomeric length maintenance and telomeric silencing. *EMBO J.*, **17**, 1819–1828.
38. Lieber,M.R., Grawunder,U., Wu,X. and Yaneva,M. (1997) Tying loose ends: roles of Ku and DNA-dependent protein kinase in the repair of double-strand breaks. *Curr. Opin. Genet. Dev.*, **7**, 99–104.
39. Mills,K.D., Sinclair,D.A. and Guarente,L. (1999) MEC1-dependent redistribution of the Sir3 silencing protein from telomeres to DNA double-strand breaks. *Cell*, **97**, 609–620.

40. Tsukamoto,Y., Kato,J. and Ikeda,H. (1997) Silencing factors participate in DNA repair and recombination in *Saccharomyces cerevisiae*. *Nature*, **388**, 900–903.
41. Polotnianka,R.M., Li,J. and Lustig,A.J. (1998) The yeast Ku heterodimer is essential for protection of the telomere against nucleolytic and recombinational activities. *Curr. Biol.*, **8**, 831–834.
42. Aylon,Y. and Kupiec,M. (2005) Cell cycle-dependent regulation of double-strand break repair: a role for the CDK. *Cell Cycle*, **4**.
43. van den Bosch,M., Lohman,P.H. and Pastink,A. (2002) DNA double-strand break repair by homologous recombination. *J. Biol. Chem.*, **383**, 873–892.
44. Morgan,E.A., Shah,N. and Symington,L.S. (2002) The requirement for ATP hydrolysis by Saccharomyces cerevisiae Rad51 is bypassed by mating-type heterozygosity or RAD54 in high copy. *Mol. Cell Biol.*, **22**, 6336–6343.
45. Cole,G.M., Schild,D., Lovett,S.T. and Mortimer,R.K. (1987) Regulation of RAD54- and RAD52-*lacZ* gene fusions in *Saccharomyces cerevisiae* in response to DNA damage. *Mol. Cell Biol.*, **7**, 1078–1084.
46. Galitski,T., Saldanha,A.J., Styles,C.A., Lander,E.S. and Fink,G.R. (1999) Ploidy regulation of gene expression [see comments]. *Science*, **285**, 251–254.
47. Birrell,G.W., Brown,J.A., Wu,H.I., Giaever,G., Chu,A.M., Davis,R.W. and Brown,J.M. (2002) Transcriptional response of *Saccharomyces cerevisiae* to DNA-damaging agents does not identify the genes that protect against these agents. *Proc. Natl Acad. Sci. USA*, **99**, 8778–8783.
48. Fourel,G., Boscheron,C., Revardel,E., Lebrun,E., Hu,Y.F., Simmen,K.C., Muller,K., Li,R., Mermod,N. and Gilson,E. (2001) An activation-independent role of transcription factors in insulator function. *EMBO Rep.*, **2**, 124–132.
49. Koering,C.E., Fourel,G., Binet-Brasselet,E., Laroche,T., Klein,F. and Gilson,E. (2000) Identification of high affinity Tbf1p-binding sites within the budding yeast genome. *Nucleic Acids Res.*, **28**, 2519–2526.
50. Fourel,G., Lebrun,E. and Gilson,E. (2002) Protosilencers as building blocks for heterochromatin. *Bioessays*, **24**, 828–835.
51. Holstege,F.C., Jennings,E.G., Wyrick,J.J., Lee,T.I., Hengartner,C.J., Green,M.R., Golub,T.R., Lander,E.S. and Young,R.A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.
52. Wyrick,J.J. and Young,R.A. (2002) Deciphering gene expression regulatory networks. *Current Curr. Opin. Genet. Dev.*, **12**, 130–136.
53. Martin,S.G., Laroche,T., Suka,N., Grunstein,M. and Gasser,S.M. (1999) Relocalization of telomeric Ku and SIR proteins in response to DNA strand breaks in yeast. *Cell*, **97**, 621–633.
54. McAinsh,A.D., Scott-Drew,S., Murray,J.A. and Jackson,S.P. (1999) DNA damage triggers disruption of telomeric silencing and Mec1p- dependent relocation of Sir3p. *Curr. Biol.*, **9**, 963–966.
55. Gunjan,A. and Verreault,A. (2003) A Rad53 kinase-dependent surveillance mechanism that regulates histone protein levels in *S. cerevisiae*. *Cell*, **115**, 537–549.

C Article IV : Inferring biological networks with Output Kernel Trees

BMC Bioinformatics

BioMed Central

Research

Inferring biological networks with output kernel trees

Pierre Geurts*[1,2], Nizar Touleimat[1,3], Marie Dutreix[3] and Florence d'Alché-Buc*[1]

Address: [1]IBISC FRE CNRS 2873 & Epigenomics project, GENOPOLE, 523, Place des Terrasses, 91 Evry, France, [2]Department of Electrical Engineering and Computer Science & GIGA, University of Liège, Institut Montefiore, Sart Tilman B28, 4000 Liège, Belgium and [3]UMR 2027 CNRS-IC, Institut Curie, Bâtiment 110, Centre Universitaire, 91405 Orsay, France

Email: Pierre Geurts* - p.geurts@ulg.ac.be; Nizar Touleimat - nizar.touleimat@ibisc.univ-evry.fr; Marie Dutreix - marie.dutreix@curie.u-psud.fr; Florence d'Alché-Buc* - florence.dalche@ibisc.univ-evry.fr

* Corresponding authors

from Probabilistic Modeling and Machine Learning in Structural and Systems Biology
Tuusula, Finland. 17–18 June 2006

Published: 3 May 2007

BMC Bioinformatics 2007, **8**(Suppl 2):S4 doi:10.1186/1471-2105-8-S2-S4

This article is available from: http://www.biomedcentral.com/1471-2105/8/S2/S4

Abstract

Background: Elucidating biological networks between proteins appears nowadays as one of the most important challenges in systems biology. Computational approaches to this problem are important to complement high-throughput technologies and to help biologists in designing new experiments. In this work, we focus on the completion of a biological network from various sources of experimental data.

Results: We propose a new machine learning approach for the supervised inference of biological networks, which is based on a kernelization of the output space of regression trees. It inherits several features of tree-based algorithms such as interpretability, robustness to irrelevant variables, and input scalability. We applied this method to the inference of two kinds of networks in the yeast *S. cerevisiae*: a protein-protein interaction network and an enzyme network. In both cases, we obtained results competitive with existing approaches. We also show that our method provides relevant insights on input data regarding their potential relationship with the existence of interactions. Furthermore, we confirm the biological validity of our predictions in the context of an analysis of gene expression data.

Conclusion: Output kernel tree based methods provide an efficient tool for the inference of biological networks from experimental data. Their simplicity and interpretability should make them of great value for biologists.

Background

The large spread-out of microarray data has recently renewed the interest for elucidating biological networks. Biological networks such as protein-protein interaction networks or metabolic networks are not real biological systems *per se* but are very convenient representations of the relations that underlie these complex systems. In this domain, the main challenge is to infer the structure of the

networks from all available data for a given organism. Both supervised and unsupervised methods have been proposed to address this problem. Unsupervised methods derive some interaction score for each protein pair on the basis of single or multiple sources of data (e.g., [1]). The great advantage of these methods lies in the fact that they do not require any prior knowledge about the network structure. However, they potentially perform poorly in comparison with supervised methods that incorporate more information. Among supervised methods, mainly two approaches have been adopted. Relational learning approaches exploit a sample of known interacting and non-interacting protein pairs to learn a classifier that can decide if a new pair of proteins is interacting or not from a set of features defined directly on pairs [2]. Other supervised approaches adopt a more global view of the problem, searching to complete the protein network from a known subnetwork. These algorithms use features of a single protein (or gene) to determine the position of this protein in the network [3-5]. The work presented in this paper falls into this latter family of methods. Existing supervised algorithms often embed the input data used to infer the network in a kernel and thus result in black-box models that do not provide much insight about the problem. In this paper, we propose a new method, called Output Kernel Trees, based on a kernelization of the output space of regression trees. Unlike existing kernel-based methods, it uses the original (non kernelized) input space and thus fully inherits the interpretability and robustness to irrelevant variables of standard tree-based methods. When applied to network inference, it provides useful information about the relationship between the input data and the existence of interactions.

The paper is structured as follows. We first introduce the general setting of supervised inference of biological networks and show how this problem can be addressed using Output Kernel Trees. Numerical experiments concern two kinds of networks in the yeast *S. cerevisiae*: a protein-protein interaction network and an enzyme network. We compare and discuss the role of various input features from expression data to phylogenetic profiles for the prediction of interactions. Our algorithm obtains results competitive with existing approaches and offers a way to rank the features according to their importance in the prediction. We also illustrate the biological validity of our predictions in the context of an analysis of gene expression data.

Methods

Supervised network inference

The problem of supervised network inference has been defined in [3,6] and subsequently considered in [5]. It may be formulated as follows.

Let $G = (V, E)$ be an undirected graph with vertices V and edges $E \subset V \times V$. $|V| = m$ is the number of nodes in the graph. We suppose that each vertex v_i, $i = 1...m$, can be described by some features in some input space $\mathcal{X}$, and we denote by $x(v_i) = x_i \in \mathcal{X}$ this information. Only the knowledge of a subgraph $G_n = (V_n, E_n)$ of G is available during the training phase: without loosing generality, we enumerate the nodes belonging to V_n as $v_1,....,v_n$ where n is the number of nodes in the subgraph denoted by $G_n = (V_n, E_n)$ with $V_n \subset V$ and $E_n = \{(v, v') \in E | v, v' \in V_n\}$. The goal of supervised graph inference is then to determine from the knowledge of G_n a function $e(x(v), x(v')): V \times V \rightarrow \{0,1\}$, ideally such that $e(x(v), x(v')) = 1 \Leftrightarrow (v, v') \in E$.

Following [3] and [5], our solution is based on a kernel embedding of the graph. A (positive semi-definite) kernel is defined as a function $k: V \times V \rightarrow \mathcal{R}$ which induces a feature map ϕ into a Hilbert space $\mathcal{H}$ such that $k(v, v') = \phi(v), \phi((v')$. To solve the problem of graph inference, we first define a kernel $k(v, v')$ such that adjacent vertices lead to high values of k and non-adjacent ones lead to smaller ones. The mapping of this kernel is thus such that $\phi(v)$ is close to $\phi(v')$ in $\mathcal{H}$ as soon as v and v' are connected. Then, the problem of graph inference may be solved as follows: from the $n \times n$ Gram matrix K with $K_{i,j} = k(v_i, v_j)$ and the input feature vectors x_i, find an approximation of the kernel values between pairs of new vertices described by their input values. A graph on unseen vertices is then obtained from the learnt kernel by connecting those vertices that correspond to a kernel prediction above some threshold.

A natural kernel between nodes of a graph is the diffusion kernel proposed in [7]. It defines the kernel value $k(v_i, v_j)$ between nodes v_i and v_j as the (i, j)-element of the matrix $K = exp(-\beta L)$, where $L = D - A$ is the Laplacian matrix of the graph, with D the diagonal matrix of node connectivities and A the adjacency matrix, and $\beta > 0$ is a user-defined parameter that controls the degree of diffusion. With respect to the adjacency matrix, the diffusion kernel defines a more global and smoother similarity measure between two nodes that takes into account all paths in the graph (even non direct) between these two nodes. When β increases, the kernel diffuses more deeply into the graph, making distant vertices in the graph closer in $\mathcal{H}$ with respect to directly adjacent vertices (see [7] for more details and several interpretations of the diffusion kernel).

Output Kernel Trees

Output Kernel Trees (OK3, [8]) are a kernelization of standard classification and regression trees [9] that can handle any output space over which a kernel may be defined. By extension, this method also allows to learn a kernel as a function of an input vector. We focus our presentation here on this particular feature of the method. The interested reader may refer to [8] for a more complete description.

Learning stage

Our algorithm follows the main steps of the CART algorithm [9]. Starting from a training set of vertices $\{v_1,...,v_n\}$ described by their input vectors $x_i = x(v_i)$, $i = 1,...,n$ and a Gram matrix K with $K_{i,j} = k(v_i, v_j)$, the idea of our method is to recursively split the training set with binary tests based on the input features. A test T is a boolean function of the input feature vector that usually involves only one feature at the same time: for a numerical variable, it compares its value to a threshold and for a categorical variable, it checks whether its value belongs to a subset of all possible values of the variable. Each split of a tree node aims at reducing as much as possible the variance of the output feature vector $\varphi(v)$ in the left and right subsets of graph vertices corresponding to the two issues of the test. (Note that to avoid confusion between nodes of the output graph and nodes of the tree model, we reserve the term "vertex" for the former, and "node" for the latter.) Given the definition of the output kernel, this amounts at dividing the set of vertices corresponding to that node into two subsets in which vertices are as much as possible connected between each other in the training graph.

More precisely, the score used to evaluate and select a test T given the local learning sample S at the current node is defined as follows:

$$\text{Score}(T,S) = \text{var}\{\phi(v) \mid S\} - \frac{N_l}{N}\text{var}\{\phi(v) \mid S_l\} - \frac{N_r}{N}\text{var}\{\phi(v) \mid S_r\}, \quad (1)$$

where N is the size of S, S_l and S_r are its left and right successors of size N_l and N_r respectively (corresponding to the test T being true or false respectively) and $\text{var}\{\varphi(v)|S\}$ is the empirical variance of the output feature vector in the subset S, computed using the kernel trick by:

$$\text{var}\{\phi(v) \mid S\} = \frac{1}{N}\sum_{i=1}^{N} \|\phi(v_i) - \frac{1}{N}\sum_{i=1}^{N}\phi(v_i)\|^2 = \frac{1}{N}\sum_{i=1}^{N} k(v_i,v_i) - \frac{1}{N^2}\sum_{i,j=1}^{N} k(v_i,v_j). \quad (2)$$

Like in the standard CART algorithm, an exhaustive search is carried out at each tree node to find the test that maximizes this score. The splitting of a node is stopped when the output feature vector is constant in S (ie. variance (2) is null) or some stopping criterion is met (e.g., the size of the local subsample is below some threshold).

By analogy with regression trees, this algorithm actually tries to find implicitly an approximation $\hat{\phi}(x(v))$ of the output feature vector $\varphi(v)$ corresponding to a vertex v from its input vector $x(v)$. The loss function that it minimizes (in average) over the learning sample is the square distance in $\mathcal{H}$, ie. $\|\hat{\phi}(x(v)) - \phi(v)\|^2$.

Prediction stage

Again, by analogy with regression trees, each leaf L of the tree is labeled with a prediction $\hat{\phi}_L$ in $\mathcal{H}$ computed as:

$$\hat{\phi}_L = \frac{1}{N_L}\sum_{i=1}^{N_L}\phi(v_i), \quad (3)$$

where N_L is the number of learning cases that reach this leaf. Our final goal however is to make predictions about the kernel value between two vertices v and v' described by their input vectors $x(v)$ and $x(v')$. Let us suppose that $x(v)$ (resp. $x(v')$) reaches leaf L_1 (resp. L_2) that contains vertices $\{v_1^1,...,v_{N_{L_1}}^1\}$ (resp. $\{v_1^2,...,v_{N_{L_2}}^2\}$). From (3), the kernel between v and v' is approximated by:

$$\hat{k}(v,v') = \langle\hat{\phi}_{L_1},\hat{\phi}_{L_2}\rangle = \frac{1}{N_{L_1}N_{L_2}}\sum_{i=1}^{N_{L_1}}\sum_{j=1}^{N_{L_2}}\langle\phi(v_i^1),\phi(v_j^2)\rangle = \frac{1}{N_{L_1}N_{L_2}}\sum_{i=1}^{N_{L_1}}\sum_{j=1}^{N_{L_2}}k(v_i^1,v_j^2), \quad (4)$$

which makes use of kernel values only. Then, this kernel can be thresholded to make a prediction about the existence of an edge between v and v'.

By construction, our method, called Output Kernel Trees (OK3), shares several features of standard tree-based methods. The most attractive ones being the interpretability of the model and the ability of the method to rank the features (see below).

Ensembles of output kernelized trees

While useful for interpretability reasons, single trees are usually not competitive with other methods in terms of accuracy, essentially because of their high variance. Thus, in the context of classification and regression problems, ensemble methods have been proposed to reduce variance and improve accuracy. In general, these methods grow an ensemble of diverse trees instead of a single one and then combine in some fashion the predictions of these trees to yield a final prediction. Among these methods, those which only rely on score computations to grow the ensemble of trees and which combine predictions by simply averaging them, can be directly extended to OK3. As a matter of fact, the prediction of an ensemble of trees in $\mathcal{H}$, which is an average of sums like (3), may be writ-

ten as a weighted sum of output feature space vectors from the learning sample, i.e. $\hat{\phi}_{ens}(x(v)) = \sum_{i=1}^{n} w_i(v)\phi(v_i)$. Then, kernel predictions are computed from the ensemble by $\hat{k}_{ens}(v,v') = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i(v)w_j(v')k(v_i,v_j)$. In our experiments, we grow ensembles of OK3 with the extra-trees method proposed in [10]. In this method, each tree of the ensemble is grown from the complete learning sample while randomizing the choice of the split at each node. We refer the interested reader to [10] for the exact description of this algorithm.

Attribute selection and ranking

An important feature of tree-based methods is that they can be exploited to rank attributes according to their importance for predicting the output value. The computation of this ranking is especially interesting with ensemble methods which are not interpretable by themselves. In the context of OK3, we propose to compute the importance of an attribute by computing for each split (in a tree, or in an ensemble of trees) where the attribute is used the total reduction of variance brought by the split, which is actually $N \times \text{Score}(S, T)$ (see Eqn. 1), and by summing these reductions. Thus, attributes that do not appear in a tree have an importance of zero, and those that are selected close to the root nodes of the trees typically receive high scores.

Results and discussion

Data

Biological networks

We carry out experiments on two kinds of protein networks in the yeast *S. cerevisiae*. The first one is a network of physical protein-protein interactions borrowed from [5] that consists of the high confidence interactions highlighted in [11]. It is composed of 2438 interactions that link 984 proteins. The second network is a network related to the metabolism of the yeast. Two proteins (enzymes) in this network are connected if they catalyze successive reactions in any metabolic pathway. It was obtained from the KEGG/PATHWAY database [12] by [4] and contains 668 proteins and 2782 edges. (Note that this network is slightly different from the one used in [3] and subsequently in [5].) 184 proteins are shared between the protein interaction network and the metabolic network.

As described in the Methods section, both networks were smoothed by a diffusion kernel. For comparison purpose with [5] and [4], the kernel matrix was normalized and the parameter β of the diffusion kernel was fixed to 3.0 for the protein-protein interaction network and to 1.0 for the metabolic network. We have nevertheless tried different values of $\beta \in [0.0, 3.0]$ but did not notice any important change in accuracy.

Input features

Different sources of data could be used for the inference of these biological networks. Experimental data obtained from various large scale methods are natural candidates but other kinds of data such as GO or KEGG annotations have also been used for this task [2]. In this paper, we used the same kinds of data as in [3] and [5].

Expression data (expr)

We considered two sets of gene expression data. The first dataset comes from the study in [13] and the second one comes from [14]. Both datasets contain small expression time series related to the cell-cycle in the yeast. Spellman et al's data gathers 77 time points and Eisen et al's data 80 time points. In our experiments, we use the original datasets accompanying the two publications, only filling missing values by the median of the corresponding column. Subsequently, we will refer to this data as "expr".

Phylogenetic profiles (phy)

The existence of orthologs of a given gene in a set of species is potentially an important source of information for the prediction of biological networks. In our experiments, we use the phylogenetic profiles gathered by [4]. They were obtained from the orthologous clusters in KEGG. Only fully sequenced genomes are taken into account. Each protein is described by a vector of 145 binary values, each one coding for the presence or the absence of an orthologous protein in a given organism.

Localization data (loc)

The localization of a protein in the cell is also potentially influencing its interactions with other proteins. The vector of features in this case consists of 23 binary values coding for the presence/absence of the protein in a given intracellular location. This data was obtained from the experiment in [15].

Yeast two hybrid network (y2h)

Such data is considered as a very noisy version of the true protein-protein interaction network and has been shown to contain many false positives. In our experiments, we use the networks obtained from the assays in [16] and [17].

Because of its pairwise nature, this kind of data can not be directly handled by tree-based methods that require that all proteins are described by an input feature vector. To still accommodate with it, we use the following procedure: following [3], we construct a graph with an edge between two proteins if these two proteins are connected in at least one of the two networks ([16] or [17]) and turn

this graph into a kernel matrix using a diffusion kernel with $\beta = 1.0$. This kernel is then transformed into a input feature vector for each protein by computing the first 50 directions with kernel PCA.

Results

For both networks, we use an ensemble of 100 output kernel trees grown with the extra-trees method with default parameters. To match the protocols used in [4] and [5], we evaluate the method by ten-fold cross-validation. On each run, we compute the diffusion kernel on 9 folds, apply OK3 and then compute from the resulting model all kernel predictions that involve at least one protein from the test fold. A network can then be reconstructed by connecting protein pairs with a kernel value above a threshold.

ROC analysis

We analyze ROC curves obtained by varying the threshold, the true positive rate being the proportion of existing edges correctly predicted and the false positive rate the proportion of non existing edges erroneously predicted. We (vertically) average the ROC curves obtained on the different folds and we also compute (average) areas under the ROC curves (AUC values).

We distinguish two types of edges for the computation of ROC curves: edges connecting an unseen protein (from the test fold, TF) to a seen protein (from the learning folds, LF) and edges connecting an unseen protein to another unseen protein (TF vs TF). We expect that the latter will be more difficult to predict than the former. Hence, we compute in each experiment three ROC curves and AUC values: the ROC curve computed on TF vs TF edges, on TF vs LF edges, and on both kinds of edges simultaneously. The TF vs. LF and TF vs. TF ROC curves with different sets of variable are given in Figure 1 for both networks. Average and standard errors of the AUC values are summarized in Table 1.

Overall, the results are quite good. They are better for the protein-protein interaction network than for the metabolic network. The way the method exploits each data source is very different in both networks. For the protein-protein interaction network, the most important source of information is the expression data followed by the y2h network, localization data, and phylogenetic profiles. For the prediction of the metabolic network, the most important source of information is the phylogenetic profiles followed by the expression data. Localization and y2h data are on the other hand not very useful on this latter database. On both networks, combining all data sources allows to improve the AUC values with respect to the use of each data source separately. As expected, TF vs. LF edges are easier to predict than TF vs. TF edges. The difference between the two kinds of edges is however less important on the protein-protein network than on the metabolic network.

This difference in AUC between the two networks probably reflects the biological significance of the input data. Actually, localization and y2h data directly reflect protein-protein interactions. In contrast, though interacting proteins belong per se to a same metabolic pathway, the inverse is not true. Indeed, non interacting proteins can participate to distant steps of a same pathway. In that case the localization and y2h network data would poorly contribute to prediction. Phylogenetic profiles are related to protein-protein interactions as well as pathway distribution since one expects all enzymes of the same pathway to be conserved or lost during evolution. The order of the different data set contributions to prediction nicely reflects all these biological constraints. Interestingly, expression data appear to be a good predictor for protein-protein interactions. This result could reflect the requirement that different partners of a protein complex should be co-expressed.

Comparison with full kernel-based methods

For comparison, the last column of Table 1 reports the results obtained in [5] for the protein-protein interaction network and in [4] for the metabolic network (when available). In both cases, the protocols are rigorously identical to ours, although the random folds of cross-validation are different. Both methods exploit a kernel on the inputs. [5] uses an algorithm based on expectation-maximization to learn simultaneously the missing kernel values and a weight for each different data source. [4] compares two approaches: kernel canonical correlation analysis and a distance metric learning method [6]. Several other approaches (such as a number of unsupervised methods) are also compared in these papers. We only report here their best results.

Looking at the AUC obtained when integrating all data sources (except y2h for the metabolic network that was not used in [4]), we get slightly worse results than the methods in [5] for the protein-protein interaction network and better results than the methods in [4] for the metabolic network. Note however that [5] reports an AUC of 0.858 for the prediction of TF vs. TF edges, which is slightly worse than our method (0.865). There are important differences with these methods in the exploitation of the individual data sources. On the protein-protein data, we are doing a much better use of the expression data and the y2h network while these methods are better in exploiting localization data and phylogenetic profiles. The results with y2h data is quite surprising since such kind of graph structured data seems at first more naturally handled by kernel-based methods. On the metabolic network however, we make a much better use of phylogenetic pro-

Table 1: AUC results.

Inputs	All	TF vs. LF	TF vs. TF	Kern. (All)
		Protein-protein interactions		
expr	0.851 ± 0.028	0.859 ± 0.027	0.819 ± 0.082	0.776
phy	0.693 ± 0.036	0.698 ± 0.035	0.617 ± 0.064	0.767
loc	0.725 ± 0.018	0.726 ± 0.017	0.710 ± 0.055	0.788
expr+phy+loc	0 887 ± 0 024	0 891 ± 0 023	0 845 ± 0 081	-
y2h	0.790 ± 0.023	0.795 ± 0.022	0.692 ± 0.068	0.612
expr+phy+loc+y2h	0.910 ± 0.019	0.914 ± 0.017	0.865 ± 0.057	0.939
		Metabolic network		
expr	0.714 ± 0.032	0.732 ± 0.035	0.619 ± 0.089	0.706
Phy	0.815 ± 0.033	0.819 ± 0.031	0.721 ± 0.086	0.747
loc	0.587 ± 0.022	0.587 ± 0.022	0.592 ± 0.042	0.577
expr+phy+loc	0.847 ± 0.025	0.853 ± 0.025	0.733 ± 0.057	0.804
y2h	0.639 ± 0.033	0.650 ± 0.034	0.490 ± 0.098	-
expr+phy+loc+y2h	0.844 ± 0.025	0.851 ± 0.026	0.721 ± 0.056	-

AUC results obtained with extra-trees and ten-fold cross-validation compared with full kernel-based methods. The best result in each row between tree-based and kernel-based methods (for all predictions) is underlined.

files than kernel-based methods and handle localization and expression data equivalently.

Kernel-based methods are usually not as efficient as tree-based ensemble methods to detect irrelevant inputs (although there exist techniques to incorporate specific attribute selection constraints into kernel-based methods). This may explain why they are not as good as our tree-based ensemble method on the expression data, which potentially contain irrelevant and noisy information. On the other hand, tests on phylogenetic variables in our trees are based on the presence or the absence of an ortholog protein in only one organism at a time. For the prediction of protein-protein interactions, the whole profile should be considered and hence these very local tests are somewhat inappropriate. For the prediction of the metabolic network, however, it is known that different organisms have developed different pathways.

Hence, the presence of an ortholog of the protein in a given organism is potentially informative. This may explain why our trees make a better use of phylogenetic profiles for the metabolic network than for the protein-protein network, while the opposite is true for kernel-based methods.

Interpretability

One of the main advantages of our tree-based approach is that it provides interpretable results. We illustrate this feature in this section.

Clustering

When used as single trees, output kernel trees provide a partition of the learning sample into clusters, one for each tree leaf, where proteins are as much as possible connected between each other. Each cluster is furthermore described by a rule based on the input variables. As an illustration, Figure 2 shows a tree that was obtained from the whole learning sample on the protein-protein interaction data, using phylogenetic profiles, expression, and localization data as inputs. The tree complexity was automatically adjusted by cost-complexity pruning with 10-fold cross-validation [9]. The left (resp. right) successor of each test node corresponds to the test at the node being true (resp. false). Each leaf is labeled with a pair (N, p), where N is the number of proteins in its cluster and p is the percentage of protein pairs that interact in the cluster. For comparison, the percentage of interactions in the whole learning sample is 0.5%. Of course, since the problem is quite difficult and noisy, several leaves do not correspond to significantly connected proteins. We projected the more significant leaves (arbitrarily defined as those that contain more than 5 proteins and 5% of connections) on the protein-protein interaction network (see Figure 3). As expected, these clusters correspond to highly connected regions in the graph. Looking at tree tests, we get furthermore a description of these clusters in terms of the input features. For example, the leaf L19 corresponds to those genes that satisfy two conditions on experiments CDC15 and CDC28 of Spellman et al's expression data. They are represented by red nodes in the graph of Figure 3. An analysis of the GO functions of these genes shows that most of them participate to ribosome biogenesis.

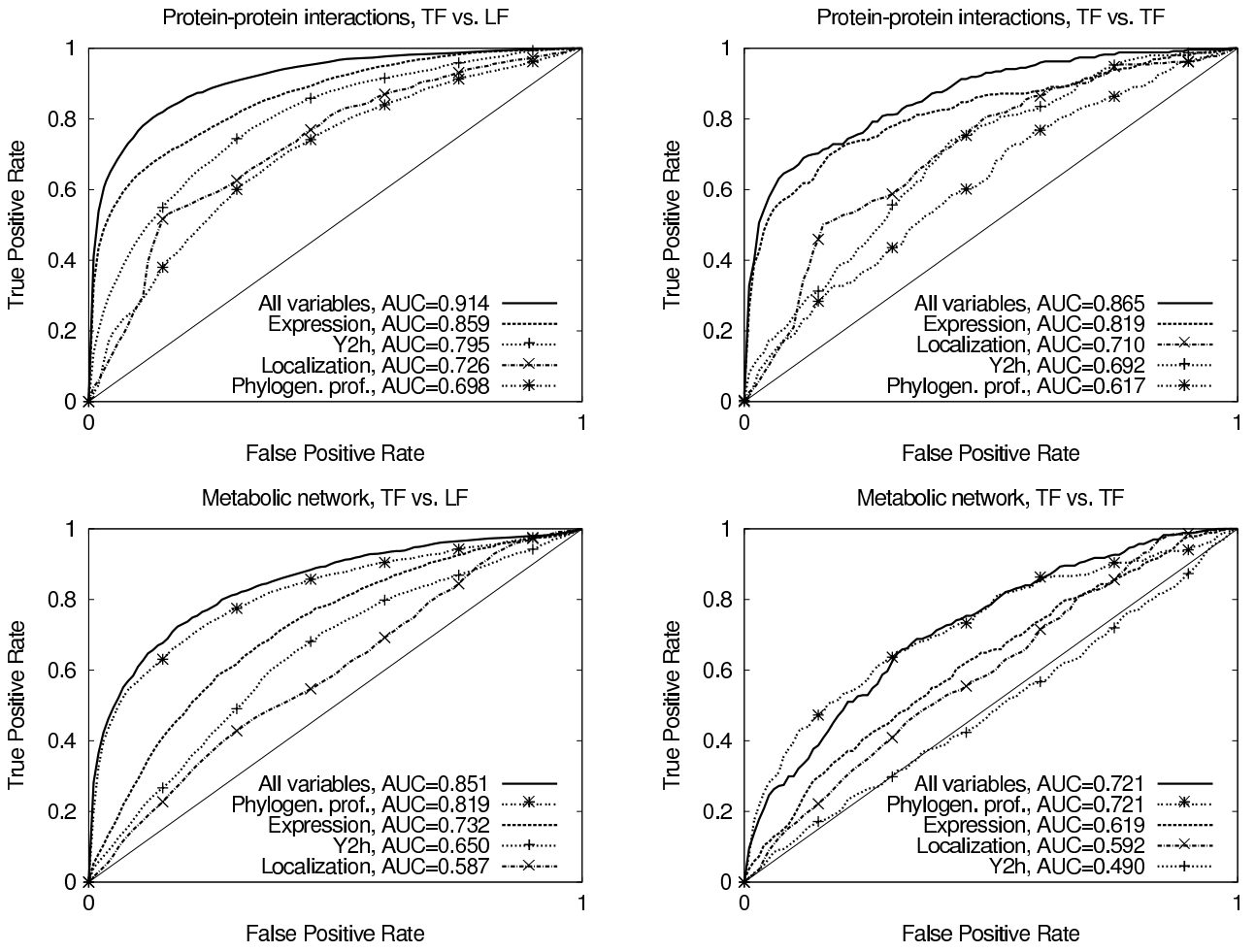


Figure 1
ROC curves. ROC curves for TF vs. LF edges (left) and TF vs. TF edges (right) with different sets of inputs, on the protein-protein interaction network (top) and the metabolic network (bottom).

Variable ranking

Table 2 shows the first 10 variables in the ranking obtained from the two datasets by ensembles of output kernel trees with expressions, phylogenetic profiles, and localization data. These rankings were obtained from ensembles of extra-trees with the importance measure developed in the Methods section. To further reduce the variance of these rankings, the importance of each feature is actually the average of the importances obtained over the 10 folds of the cross-validation.

Note that these rankings of individual features refine the ranking of the different data sources that was found in Table 1.

Biological validation

Previous experiments show the good general behavior of our algorithm on two benchmark problems. However, for this algorithm to be useful for biologists, it must be able to provide new and biologically sound predictions. To illustrate this capability, we run an additional experiment in the context of a bioinformatics analysis of a gene expression dataset. This transcriptome dataset, described in [18], includes gene expression kinetics of seven yeast strains submitted to a stress of radiation. A clustering analysis applied on these gene expression kinetics revealed several clusters of co-expressed genes, among which one cluster of 198 genes was deemed of particular interest for further analysis (see [19]). In this illustration, we focus on

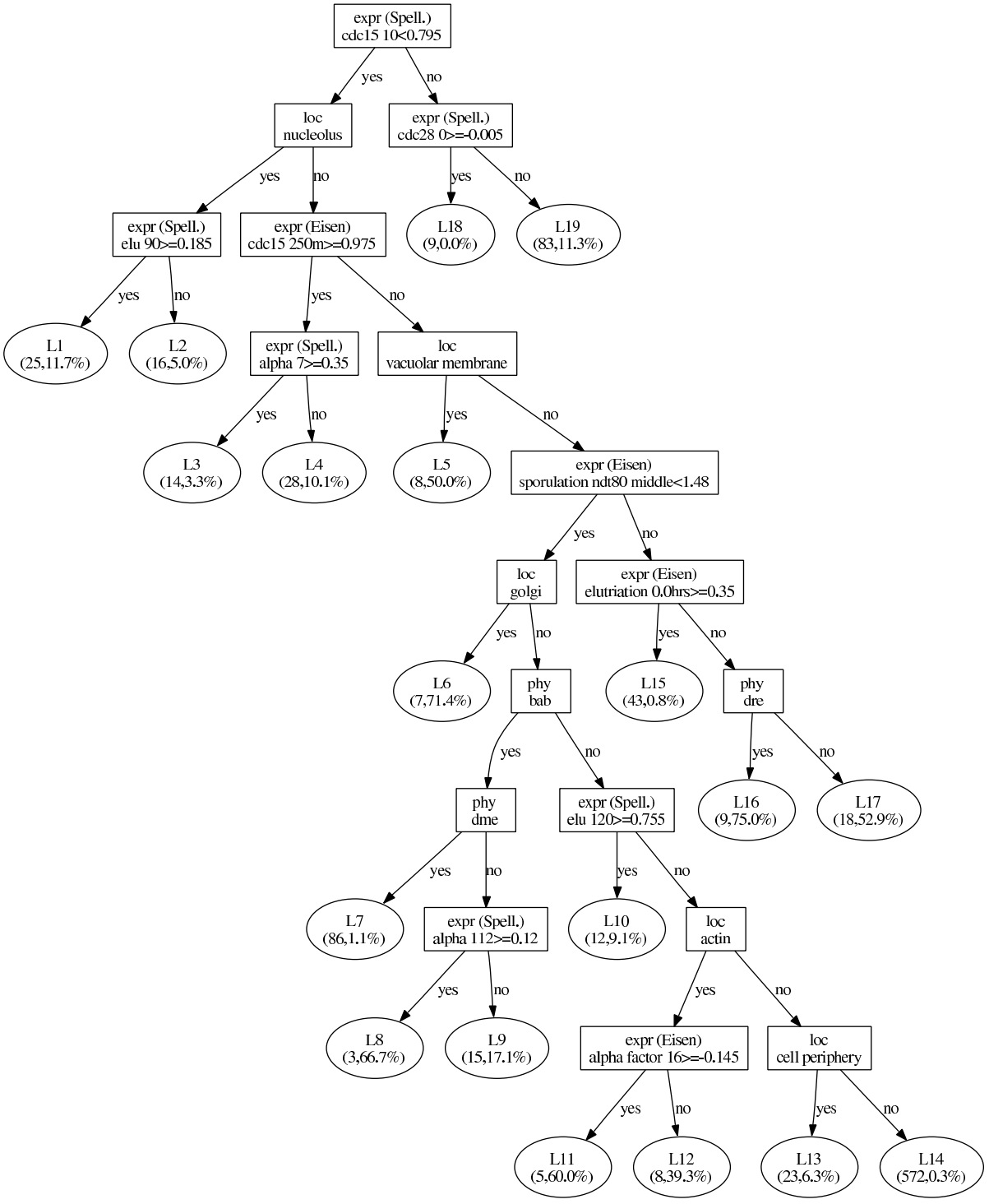


Figure 2

Decision tree. A decision tree obtained on the protein-protein interaction network using expression data, phylogenetic profiles and localization data as inputs. The tree size was determined by cost-complexity pruning with 10-fold cross-validation. The left (resp. right) edge from a test node corresponds to the test of the node being true (resp. false). Each leaf is labeled with a pair (N, p), where N is the number of proteins in its cluster and p is the percentage of protein pairs that interact in the cluster.

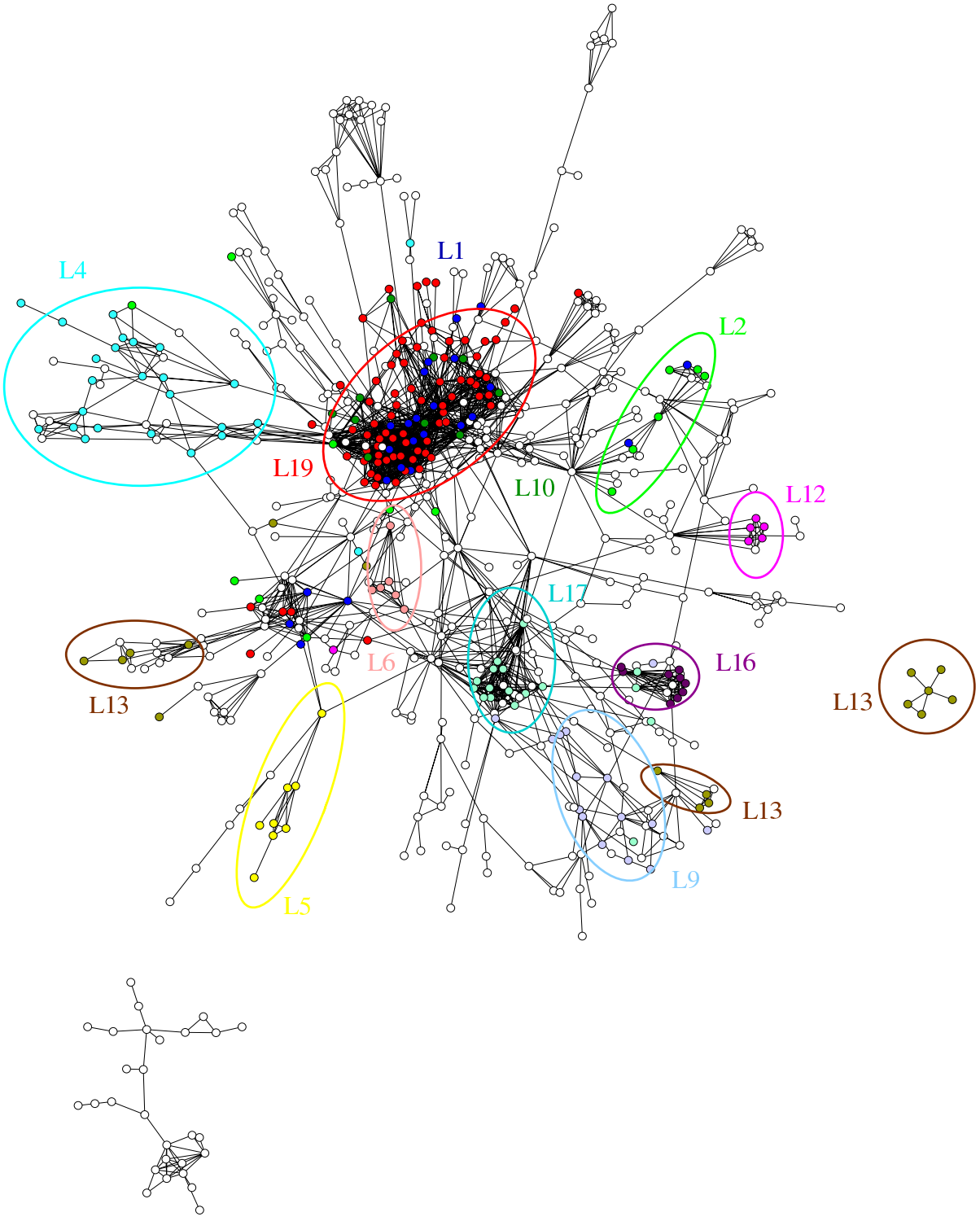


Figure 3
Graph clustering. The projection of the tree leaves in Figure 2 on the protein-protein interaction network. Only the leaves that contain more than 5 proteins and 5% of connections are represented.

Table 2: Variable ranking.

Protein-protein interactions			Metabolic network		
#	Att.	Imp	#	Att.	Imp
1	loc – nucleolus	0.021	1	phy – dre	0.011
2	expr (Spell.) – elu 120	0.013	2	phy – rno	0.009
3	loc – cytoplasm	0.012	3	expr (Eisen) – cdc15 120 m	0.008
4	expr (Eisen) – sporulation ndt80 early	0.012	4	phy – ecu	0.008
5	loc – nucleus	0.012	5	expr (Eisen) – cdc15 160 m	0.008
6	expr (Eisen) – sporulation 30 m	0.011	6	phy – pfa	0.007
7	expr (Eisen) – sporulation ndt80 middle	0.010	7	phy – mmu	0.007
8	expr (Spell.) – alpha 14	0.010	8	loc – cytoplasm	0.006
9	expr (Spell.) – elu 150	0.010	9	expr (Eisen) – cdc15 30 m	0.005
10	loc – mitochondrion	0.009	10	expr (Eisen) – elutriation 5.5 hrs	0.005

Variable rankings obtained with expressions, phylogenetic profiles, and localization data used as inputs to extra-trees.

the prediction of protein-protein interactions in this cluster. As a training set for our algorithm, we use a recent interactome dataset proposed in [20]. This high-quality data set was obtained by intersecting data generated by several different interaction detection methods. The resulting network, called "filtered yeast interactome" (FYI), contains 1,379 proteins and 2,493 interactions. Only 60 among the 198 proteins in the cluster of interest are present in the FYI dataset, leaving the connections between the remaining proteins unknown. We learned a model from the FYI data using as inputs expression, localization, and phylogenetic data (the y2h data was not considered as it was one of the sources exploited to make up the FYI dataset) and then used this model to complete the network of interactions for the 198 genes. This resulted in a network with 379 edges (using a kernel threshold of 0.85), among which only 35 edges were previously known. Figure 4 draws this network where nodes present in the training set are represented by blue diamond-shaped nodes and unseen nodes by red circle-shaped nodes. Only proteins that are connected to at least one other protein are represented (in total, 131 out of 198). The network with all protein names is available in a web appendix [21].

First, we note that this cluster of co-regulated genes is highly connected. Indeed, using the same kernel threshold, a set of 198 random genes would contain in average 10 times less edges than our clu networks.ster. The inferred network thus suggests that these proteins are likely to share some functions. It also clearly reveals several highly connected subclusters of nodes that could correspond to several functional modules. To check this hypothesis, we use the gene ontology to annotate the different subnetworks in Cytoscape [22]. Statistical significance of the annotation was checked with BiNGO [23].

Figure 4 shows these annotations. This analysis highlights four distinct but related biological processes, all involved in different steps of the production of ribosomes. Interestingly, rRNA polymerases subunits and ribosomal subunits have no direct connections between themselves but both are connected with proteins involved in rRNA processing and ribosome biogenesis, which translates some biological facts. We thus retrieve with our method biologically meaningful subnetworks. Note that these subclusters are also highlighted in [19] by exploiting other sources of information (a.o., functional annotations, protein complexes, regulation related descriptors).

A finer way to validate our method is to try to infer the functions of unannotated proteins by looking at functions of their direct neighbors in the inferred network. As an illustration of this possibility, we first focused on the highly connected subnetwork C1. This subcluster contains six proteins, YNL175C, YCR016W, YDR365C, YKR060W, YBL028C and YOR206W, that were not yet annotated in our version of the annotation (dating of June 2006). However, their positions in the inferred network suggest that they participate in ribosome biogenesis. For some of them, it is indeed possible to find some clues that they are related to this process. For example, YNL175C shares some sequence similarity with YOL041C, which is itself involved in ribosome biogenesis. As a strong evidence in favor of our prediction, we note that a recent computational analysis [24] based on gene expression data and sequence analysis has concluded that all these six proteins participate in ribosome biogenesis. As a matter of fact, the GO annotation of these genes has been introduced in the Saccharomyces Genome Database [25] in September 2006. Another interesting protein is YDR417C whose function is yet unknown but which lies in the middle of a subset of proteins that are all components of the ribosomal subunits (subcluster C2 in Figure 4). Actually, it turns out that this protein has a large overlap in terms of

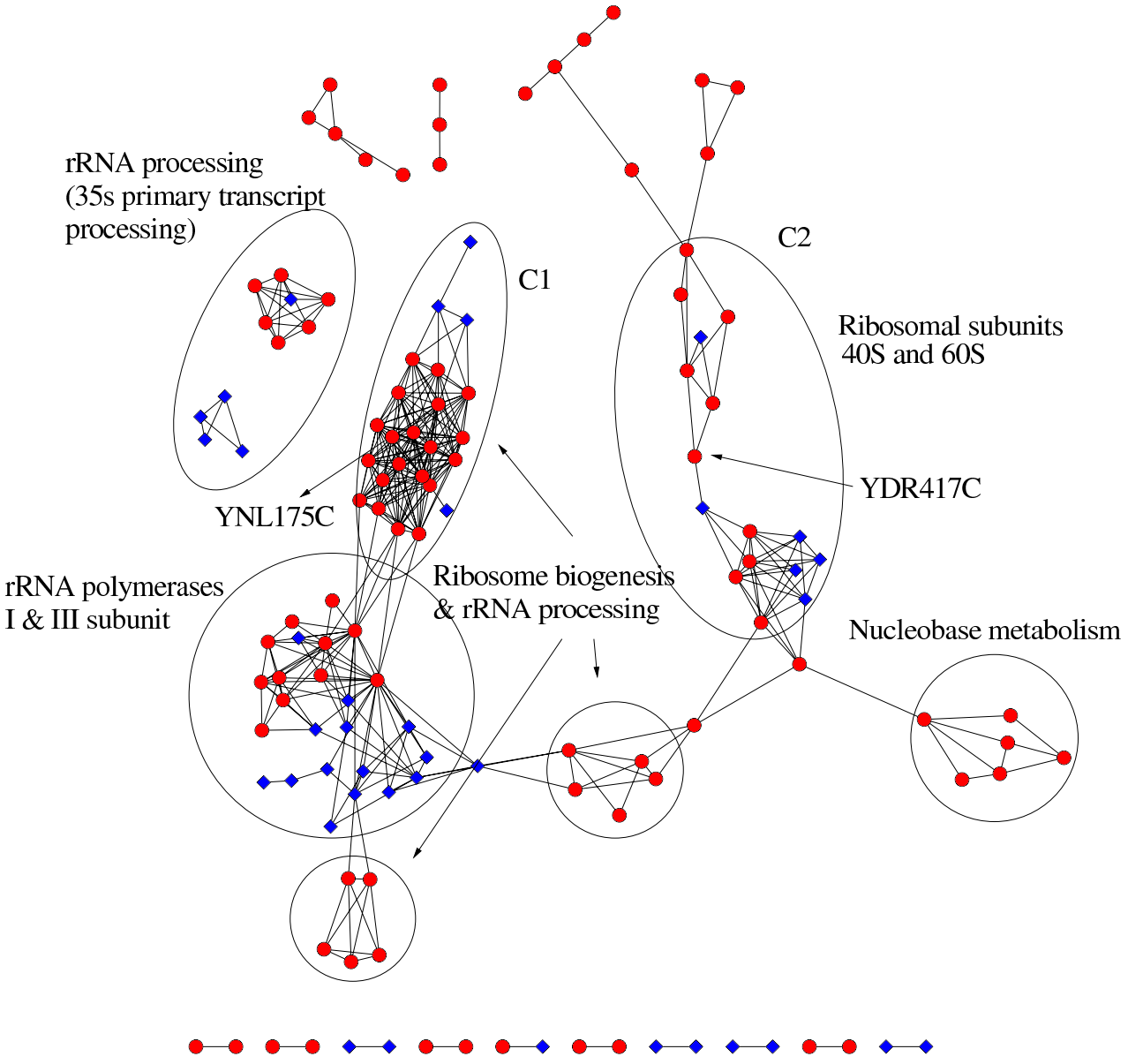


Figure 4
Cluster prediction. Predictions of protein-protein interactions in a cluster of 198 genes. Blue diamond-shaped nodes are proteins present in the training sample, red circle-shaped nodes were not seen by the learning algorithm. Annotation was found using BiNGO.

DNA sequence with another protein, YDR418W, which is a component of the large ribosomal subunit. Its position in the network may thus come from the fact that probes on microarray may not specifically distinguish between two messengers coded by the same chromosome sequence.

Conclusion

We proposed a new method for the supervised inference of biological networks. This method is based on a kernelization of the output space of tree-based methods. It yields competitive results with respect to full kernel-based methods on a protein-protein interaction network and on an enzyme network. In addition, it provides interpretable results in the form of a rule based clustering of the net-

work and a ranking of the variables according to their importance at predicting new edges. The ability to discover new information about unannotated proteins was further illustrated on a small-scale study. These results suggest that our tool could be helpful to point out proteins that are worth further experimental investigations.

Authors' contributions

PG developed the algorithm, carried out the experiments, and drafted the manuscript. FAB coordinated the collaboration, participated in the design of the algorithm, and helped in writing the manuscript. NT and MD helped in collecting and interpreting the data and did the biological validation of the results. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank the authors of [4] and [5] for providing their datasets. Pierre Geurts is a research associate of the FNRS (Belgium). This work has been done while he was a postdoc at IBISC laboratory (Evry, France) with support of the CNRS (France). Florence d'Alché-Buc's research has been funded by Genopole (France).

This article has been published as part of *BMC Bioinformatics* Volume 8, Supplement 2, 2007: Probabilistic Modeling and Machine Learning in Structural and Systems Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/8?issue=S2.

References

1. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms.** *Nucleic Acids Res* 2005, **33(Database issue):**D433-D437.
2. Ben-Hur A, Noble W: **Kernel methods for predicting protein-protein interactions.** *Bioinformatics* 2005, **21(Suppl 1):**i38-i46.
3. Yamanishi Y, Vert JP, Kanehisa M: **Protein network inference from multiple genomic data: a supervised approach.** *Bioinformatics* 2004, **20:**i363-i370.
4. Yamanishi Y, Vert JP, Kanehisa M: **Supervised enzyme network inference from the integration of genomic data and chemical information.** *Bioinformatics* 2005, **21:**i468-i477 [http://web.kuicr.kyoto-u.ac.jp/~yoshi/ismb05/].
5. Kato T, Tsuda K, Kiyoshi A: **Selective integration of multiple biological data for supervised network inference.** *Bioinformatics* 2005, **21(10**2488-2495 [http://www.cbrc.jp/~kato/faem/faem.html].
6. Vert JP, Yamanishi Y: **Supervised graph inference.** *Advances in Neural Information Processing Systems* 2004, **17:**1433-1440.
7. Kondor R, Lafferty J: **Diffusion kernels on graphs and other discrete input spaces.** *Proc of the 19th International Conference on Machine Learning* 2002:315-322.
8. Geurts P, Wehenkel L, d'Alché-Buc F: **Kernelizing the output of tree-based methods.** In *Proceedings of the 23rd International Conference on Machine Learning* Edited by: *Cohen W, Moore A. ACM*; 2006:345-352.
9. Breiman L, Friedman J, Olsen R, Stone C: *Classification and Regression Trees Wadsworth International*; 1984.
10. Geurts P, Ernst D, Wehenkel L: **Extremely randomized trees.** *Machine Learning* 2006, **36:**3-42.
11. von Mering C, Krause R, Snel B, Cornell M, Oliver S, S F, P B: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417(6887):**399-403.
12. Kaneshiha M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Res* 2004, **32(Database issue):**D277-D280.
13. Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, Eisen M, Brown P, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Mol Biol Cell* 1998, **9(12):**3273-3297.
14. Eisen M, Spellman P, Patrick O, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci* 1998, **95:**14863-14868.
15. Huh W, Falvo J, Gerke C, Carroll A, Howson R, Weissman J, O'Shea E: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425:**686-691.
16. Uetz P, Giot L, Cagney G, Mansfield T, Judson R, Knight J, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg J: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403:**623-627.
17. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y: **Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations of between the yeast proteins.** *Proc Natl Acad Sci* 2000, **97:**1143-1147.
18. Mercier G, Berthault N, Touleimat N, Kepes F, Fourel G, Gilson E, Dutreix M: **A haploid-specific transcriptional response to irradiation in Saccharomyces cerevisiae.** *Nucleic Acids Res* 2005, **33:**6635-6643.
19. Touleimat N, Zehraoui F, Dutreix M, d'Alché-Buc F: **Xpath: a semi-automated inference tool for regulatory pathways extraction from pertubed data.** . Submitted
20. Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, E Cusick M, Roth FP, Vidal M: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network.** *Nature* 2004, **430(6995):**88-93.
21. [http://www.ibisc.univ-evry.fr/Equipes/AMIS/papers/bmc-pmsb06/].
22. Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13(11**2498-2504 [http://cytoscape.org].
23. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks.** *Bioinformatics* 2005, **21:**3448-3449.
24. Wade C, Umbarger M, McAlear M: **The budding yeast rRNA and ribosome biosynthesis (RRB) regulon contains over 200 genes.** *Yeast* 2006, **23(4):**293-306.
25. **Saccharomyces Genome Database** [http://www.yeastgenome.org]

D Implémentation logicielle de la méthodologie *XRegPath*

D.1 Détails d'implémentation du logiciel *XRegPath* et scénarios d'utilisation

Gestion de l'information Le coeur de l'outil *XRegPath* est constitué d'une base de données relationnelle permettant de stocker et d'accéder à différents types d'informations biologiques associés aux gènes. L'entité 'gène' occupe une place centrale dans le schéma de la base, toutes les autres entités lui sont reliées plus ou moins directement (voir figure 4.5). Un gène est représenté par un identifiant unique (nomenclature officielle), l'identifiant de son *ORF* (*Open reading Frame*, ou phase ouverte de lecture), le nom de la protéine codée et ses différents alias et enfin sa position sur le génome (n° du chromosome et début et fin de la partie codante sur le chromosome). Sont associés au gène :

- des données d'expression, chaque donnée relie un gène à une expérience. Une donnée d'expression est représentée par deux vecteurs, l'un contenant les valeurs d'expression et l'autre contenant les temps de mesure.
- des groupes de gènes prédéfinis ou issus des étapes de classification ou de sélection de groupes via des requêtes sur les attributs ou tables associées aux gènes. Plusieurs groupes peuvent être associés à une même expérience.
- deux types d'ontologies fonctionnelles, *KEGG* et *GO*. La base de données *GO* est d'ailleurs entièrement chargée dans notre base et sa structure en arbre est conservée.
- des régulateurs (un gène peut être son propre régulateur).
- des complexes protéiques rendant comptes d'interactions protéines-protéines.
- tout autre type de descripteurs peut être associé à un gène via une table "exception".

La base peut être interrogée via des requêtes sur les attributs des gènes ou sur les informations qui leurs sont associées. On peut aussi composer des groupes de gènes à partir des données expérimentales à partir d'une requête qui sélectionne les gènes ayant des données complètes dans une série d'expériences (intersection d'expériences).
Cette base de données sert à la fois de source de données initiale pour la démarche de *XRegPath* mais peut aussi être enrichie avec les résultats obtenus par les différents outils de fouille de données proposés par cet outil. Cette base peut être facilement sauvegardée, chargée et actualisée (actualisation de la base *GO* par exemple).
XRegPath permet de sauvegarder sous la forme d'un fichier texte tous les résultats obtenus à chaque étape de l'analyse. On peut aussi reprendre une analyse à l'étape où on l'avait arrêtée en rechargeant le fichier sauvegardé. Les types de résultats que l'on peut sauvegarder sont : les groupes de gènes issus de requêtes sur la base de données ou issus de l'étape de classification, les similarités entre gènes calculées à partir de leurs données d'expression, les résultats des tests statistiques (calculs d'enrichissements et test selon la loi hypergéométrique) et les matrices d'associations entre gènes et descripteurs avant et après l'étape de *biclustering*. Il est aussi possible de sauvegarder les les représentations des cinétiques et des résultats de *biclustering* sous formes de figures.
Actuellement, les données disponibles par défaut dans la base sont celles qui correspondent à la levure *S. cerevisiae* et qui sont présentées dans notre article (voir section 4.1.2 page 87). Cette base peut être également chargée avec des données associées à tout type d'or-

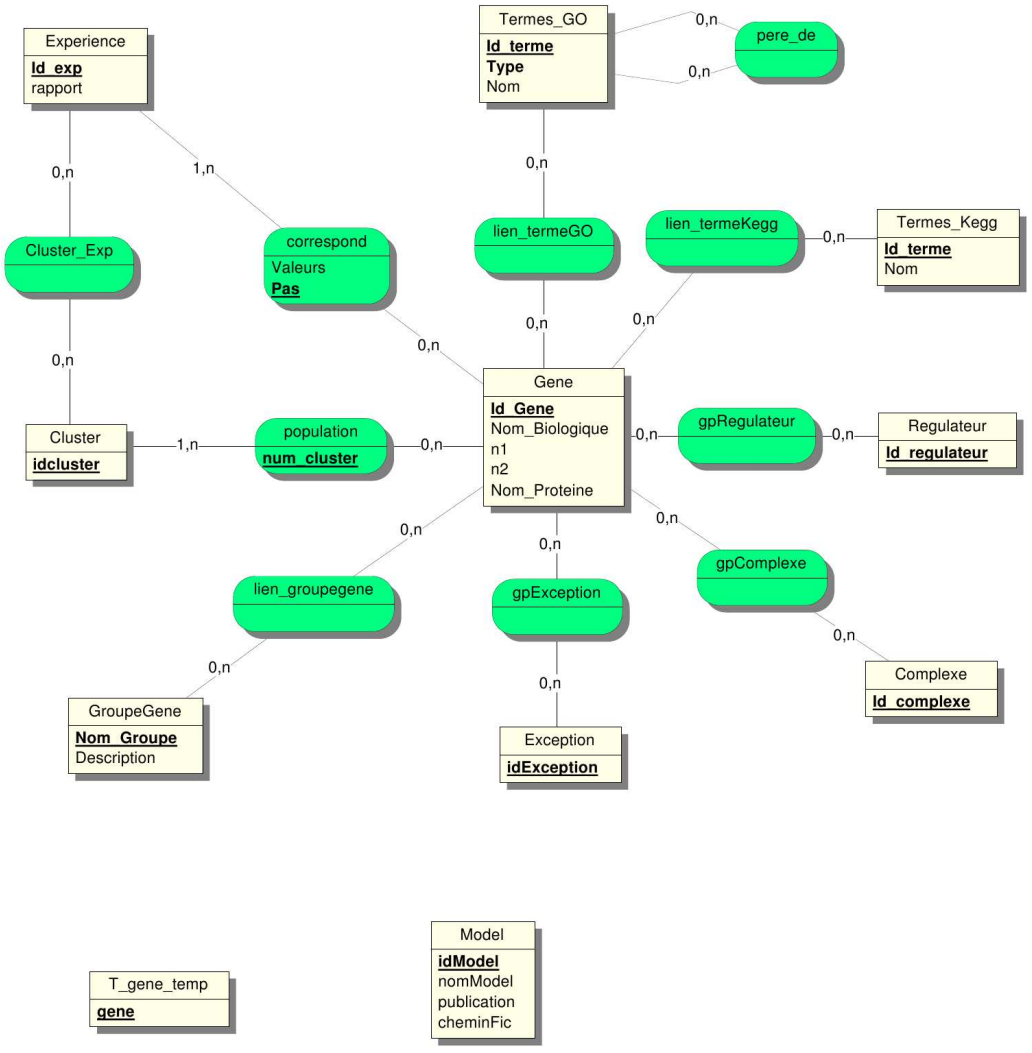


FIGURE 4.5 – Vue conceptuelle d'ensemble de la base de données d'XRegPath.

ganisme.

Outils de fouille de données L'outil *XRegPath* implémente les algorithmes de classification spectrale et de *biclustering* spectral présentés dans notre article (voir section 4.1.2 page 87). L'outil a été développé comme une succession de modules logiques de manière à laisser la possibilité de développer, d'enrichir ou de modifier chaque module indépendamment des autres.

Représentation de l'information A chaque étape de la méthodologie, une visualisation adaptée aux informations produites est proposée. Les résultats des requêtes sur la base de données sont représentés sous la forme de tableaux avec une liste de gènes associés aux attributs ayant permis leur sélection. Les résultats des tests statistiques (tests d'enrichissement de descripteurs associés aux gènes à partir de la loi hypergéométrique) sont représentés par des tableaux avec la liste des descripteurs, les valeurs statistiques associées

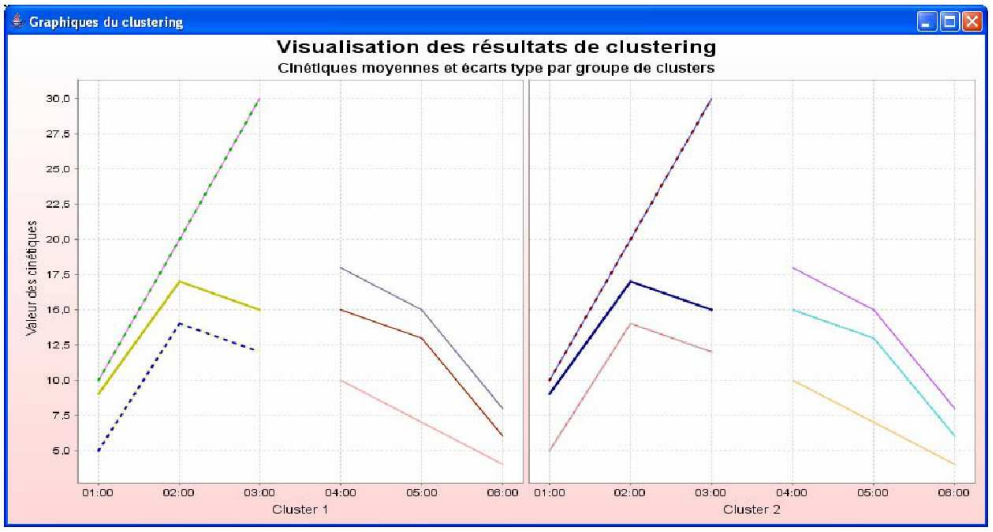


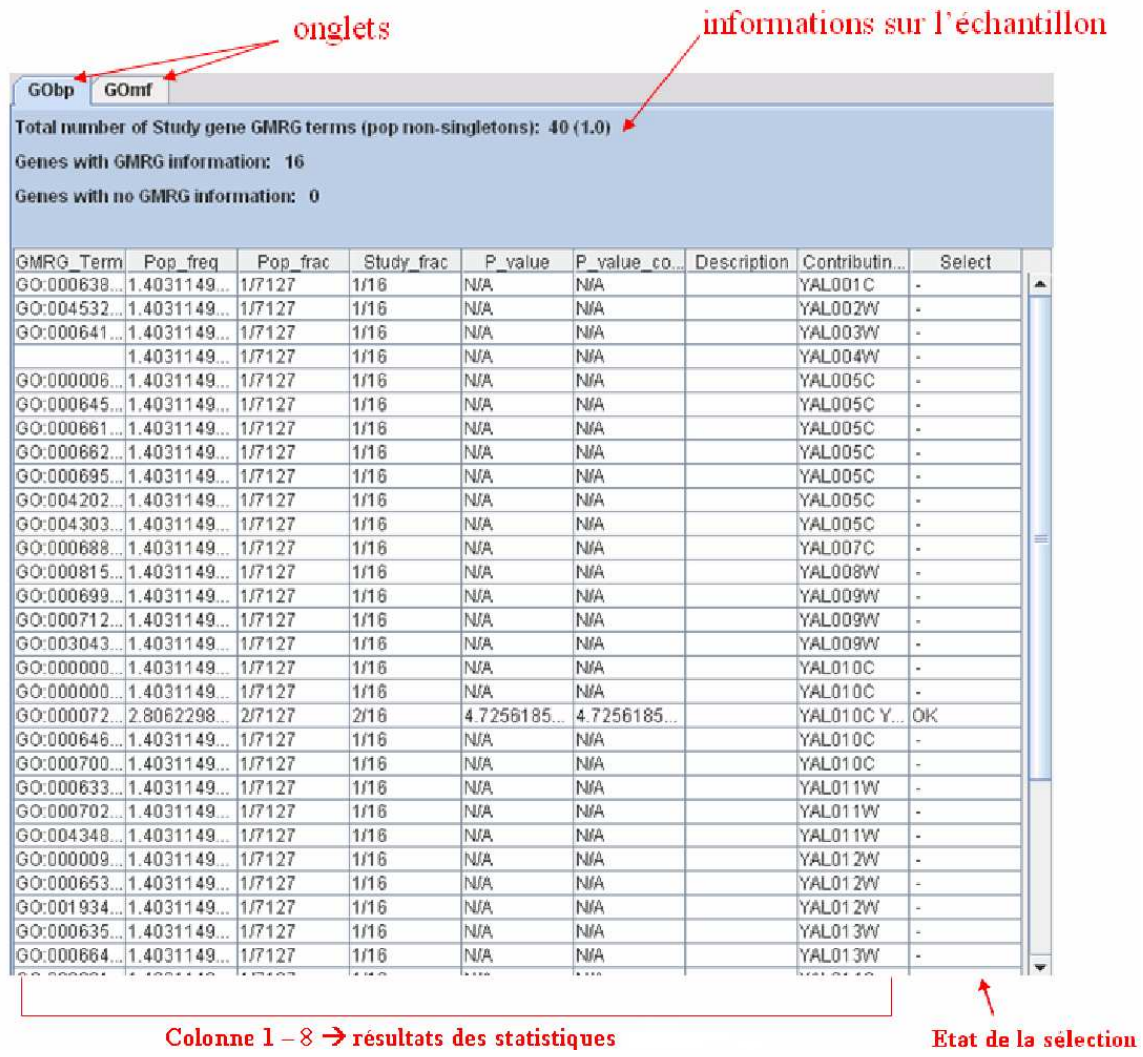
FIGURE 4.6 – Représentation de cinétiques d'expression moyennes, encadrées par les écarts types, par groupe de gènes et par expérience.

et la liste des gènes contribuant au calcul des statistiques (voir figure 4.7). Les groupes de gènes construits à partir de l'algorithme de classification spectral peuvent être visualisés sous la forme d'une liste de gènes ou sous la forme de la cinétique moyenne du groupe, encadrée par les écarts types, dans chaque expérience ayant contribué à leur classification (voir figure 4.6). Le résultat de l'étape de *biclustering* est visualisé sous la forme d'une matrice de couleur représentant les associations binaires entre gènes et descripteurs (voir figure 4.8). Cette représentation est inspirée de la visualisation des données d'expression de gène proposée par Eisen *et col.* dans le logiciel *Treeview* [319]. Enfin, les résultats du *biclustering* d'un groupe de gènes sont représentés sous la forme d'une répartition en secteurs des différents sous-groupes au sein du groupe de gènes (voir figure 4.9).

Interface graphique Les différentes fonctionnalités de l'outil *XRegPath* sont accessibles via une interface graphique ergonomique qui permet d'effectuer une analyse complète à partir d'une même fenêtre de commande. L'interface est réalisée de façon à proposer à chaque étape de l'analyse une série de paramètres, dont certaines valeurs sont parfois déjà définies par défaut. Par exemple la sélection des descripteurs de gènes à utiliser pour le *biclustering* est réalisée automatiquement à partir d'une *p-value* seuil fixée par défaut à 0.005. L'utilisateur a cependant la possibilité de changer la valeur de ce seuil ou alors de choisir un mode de sélection manuel et de sélectionner les descripteurs un par un (voir figure 4.7).

D.2 Détails d'implémentation du logiciel *XRegRules* et scénarios d'utilisation

Une grande partie du développement de cet outil a consisté à interfacer des requêtes codées en langage **Java** vers un moteur d'inférence utilisant le langage **Prolog.** Cet interfaçage a été réalisé via une bibliothèque développée en *Java*, *TuProlog* [323], qui code un



GMRG_Term	Pop_freq	Pop_frac	Study_frac	P_value	P_value_co...	Description	Contributin...	Select
GO:000638...	1.4031149...	1/7127	1/16	N/A	N/A		YAL001C	-
GO:004532...	1.4031149...	1/7127	1/16	N/A	N/A		YAL002W	-
GO:000641...	1.4031149...	1/7127	1/16	N/A	N/A		YAL003W	-
	1.4031149...	1/7127	1/16	N/A	N/A		YAL004W	-
GO:000006...	1.4031149...	1/7127	1/16	N/A	N/A		YAL005C	-
GO:000645...	1.4031149...	1/7127	1/16	N/A	N/A		YAL005C	-
GO:000661...	1.4031149...	1/7127	1/16	N/A	N/A		YAL005C	-
GO:000662...	1.4031149...	1/7127	1/16	N/A	N/A		YAL005C	-
GO:000695...	1.4031149...	1/7127	1/16	N/A	N/A		YAL005C	-
GO:004202...	1.4031149...	1/7127	1/16	N/A	N/A		YAL005C	-
GO:004303...	1.4031149...	1/7127	1/16	N/A	N/A		YAL005C	-
GO:000688...	1.4031149...	1/7127	1/16	N/A	N/A		YAL007C	-
GO:000815...	1.4031149...	1/7127	1/16	N/A	N/A		YAL008W	-
GO:000699...	1.4031149...	1/7127	1/16	N/A	N/A		YAL009W	-
GO:000712...	1.4031149...	1/7127	1/16	N/A	N/A		YAL009W	-
GO:003043...	1.4031149...	1/7127	1/16	N/A	N/A		YAL009W	-
GO:000000...	1.4031149...	1/7127	1/16	N/A	N/A		YAL010C	-
GO:000000...	1.4031149...	1/7127	1/16	N/A	N/A		YAL010C	-
GO:000072...	2.8062298...	2/7127	2/16	4.7256185...	4.7256185...		YAL010C Y...	OK
GO:000646...	1.4031149...	1/7127	1/16	N/A	N/A		YAL010C	-
GO:000700...	1.4031149...	1/7127	1/16	N/A	N/A		YAL010C	-
GO:000633...	1.4031149...	1/7127	1/16	N/A	N/A		YAL011W	-
GO:000702...	1.4031149...	1/7127	1/16	N/A	N/A		YAL011W	-
GO:004348...	1.4031149...	1/7127	1/16	N/A	N/A		YAL011W	-
GO:000009...	1.4031149...	1/7127	1/16	N/A	N/A		YAL012W	-
GO:000653...	1.4031149...	1/7127	1/16	N/A	N/A		YAL012W	-
GO:001934...	1.4031149...	1/7127	1/16	N/A	N/A		YAL012W	-
GO:000635...	1.4031149...	1/7127	1/16	N/A	N/A		YAL013W	-
GO:000664...	1.4031149...	1/7127	1/16	N/A	N/A		YAL013W	-

FIGURE 4.7 – Visualisation des résultats des tests d'enrichissement de descripteurs biologiques par groupe de gènes.

moteur d'inférence *Prolog* simple, facile à implémenter et à invoquer (voir figure 4.10).

Aide à l'écriture de bases de connaissances L'outil propose d'utiliser des bases de connaissances produites de trois façons différentes :

- Il est possible de charger une base déja existante, codée en *Prolog*, à partir d'un fichier texte.
- L'outil propose une interface de saisie qui permet d'écrire directement en *Prolog* la base à charger.
- L'outil propose une aide à l'écriture semi-automatique de prédicats de type *faits* à partir d'observations expérimentales. Cette fonctionnalité s'inscrit par exemple dans le cadre d'expériences qui observent le comportement de gènes ou de groupes de gènes dans différentes conditions de perturbations (combinaisons de variables). L'utilisateur peut soit charger des fichiers textes, contenant ses résultats formatés

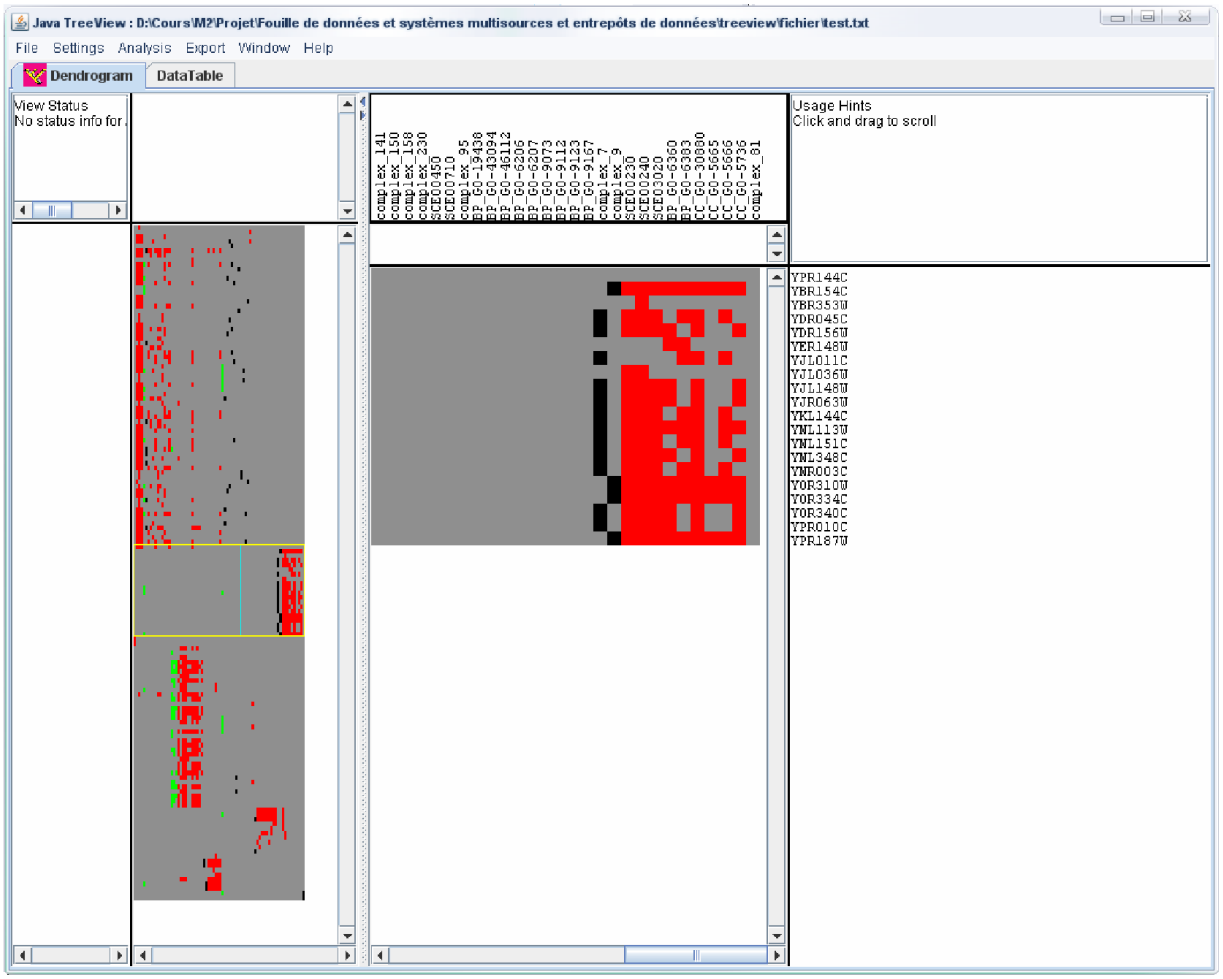


FIGURE 4.8 – Visualisation de la matrice d'association gènes/régulateurs, d'un groupe de gènes co-exprimés, réorganisée par l'algorithme de *biclustering* spectral. Chaque couleur indique un type de descripteur biologique (rouge : ontologies fonctionnelles, vert : facteurs de transcription, noir : complexes protéiques, gris : aucune association).

simplement en tableaux, soit utiliser l'interface graphique pour saisir "à la main" ses différentes observations. L'outil permet également de générer automatiquement des règles simples directement en lien avec les prédicats de type *faits* (voir figure 4.12). Les règles plus complexes devront être écrites par la suite de façon "manuelle" par l'utilisateur.

Extraction de règles de régulation L'interrogation d'une base de connaissances se fait simplement via une interface de saisie qui permet de composer une liste de "questions". L'outil va alors considérer chaque liste comme une requête à part entière et rendra la réponse à la requête sous la forme d'une liste de solutions (voir figure 4.11).

Interface graphique et visualisations des résultats L'interface graphique permet d'utiliser les différentes fonctionnalités de cet outil indépendamment les unes des autres,

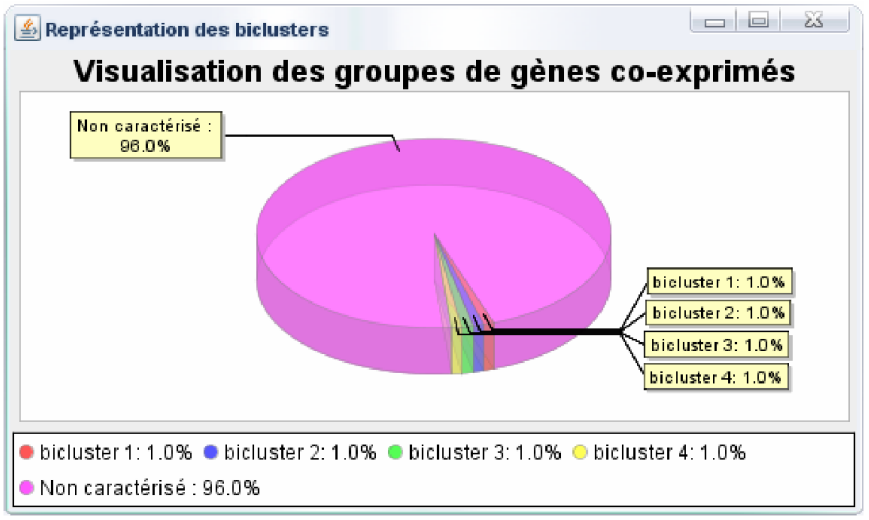


FIGURE 4.9 – Répartition en secteurs des différents sous-groupes gènes/descripteurs (*biclusters*) au sein d'un groupe de gènes co-exprimés.

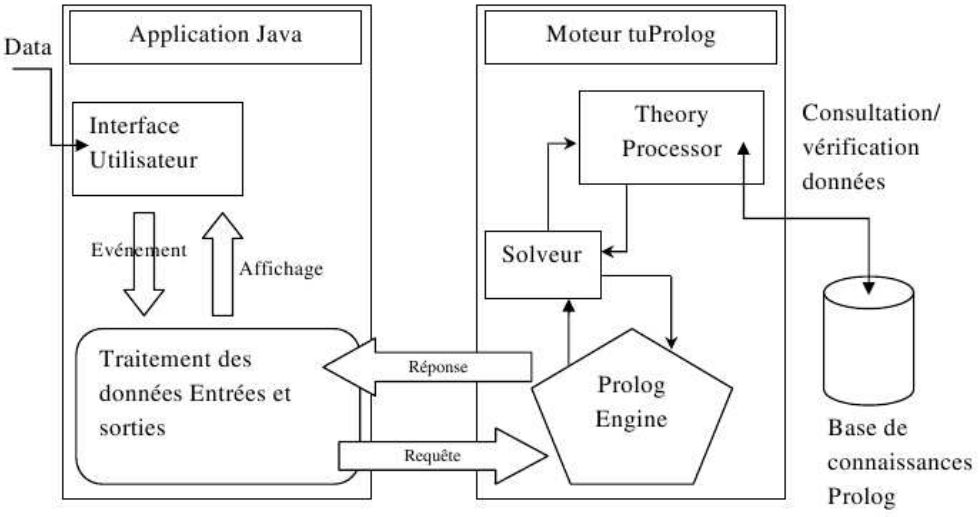


FIGURE 4.10 – Représentation conceptuelle de la structure logique de l'outil XRegRules

on peut par exemple dérouler un processus d'analyse de la création d'une nouvelle base de connaissances jusqu'à l'extraction de toutes les règles de régulation vérifiées. Mais, on peut aussi charger une base complète, l'enrichir, et sauvegarder les différents résultats temporaires dans des fichiers.
La visualisation des résultats après interrogation de la base de connaissances est beaucoup plus claire que ce que l'on obtiendrait en simple ligne de commande.

D.3 Développements futurs

Notre principal objectif sera d'intégrer le logiciel *XRegRules* au logiciel *XRegPath* afin d'obtenir un outil complet implémentant l'ensemble de notre méthodologie d'analyse

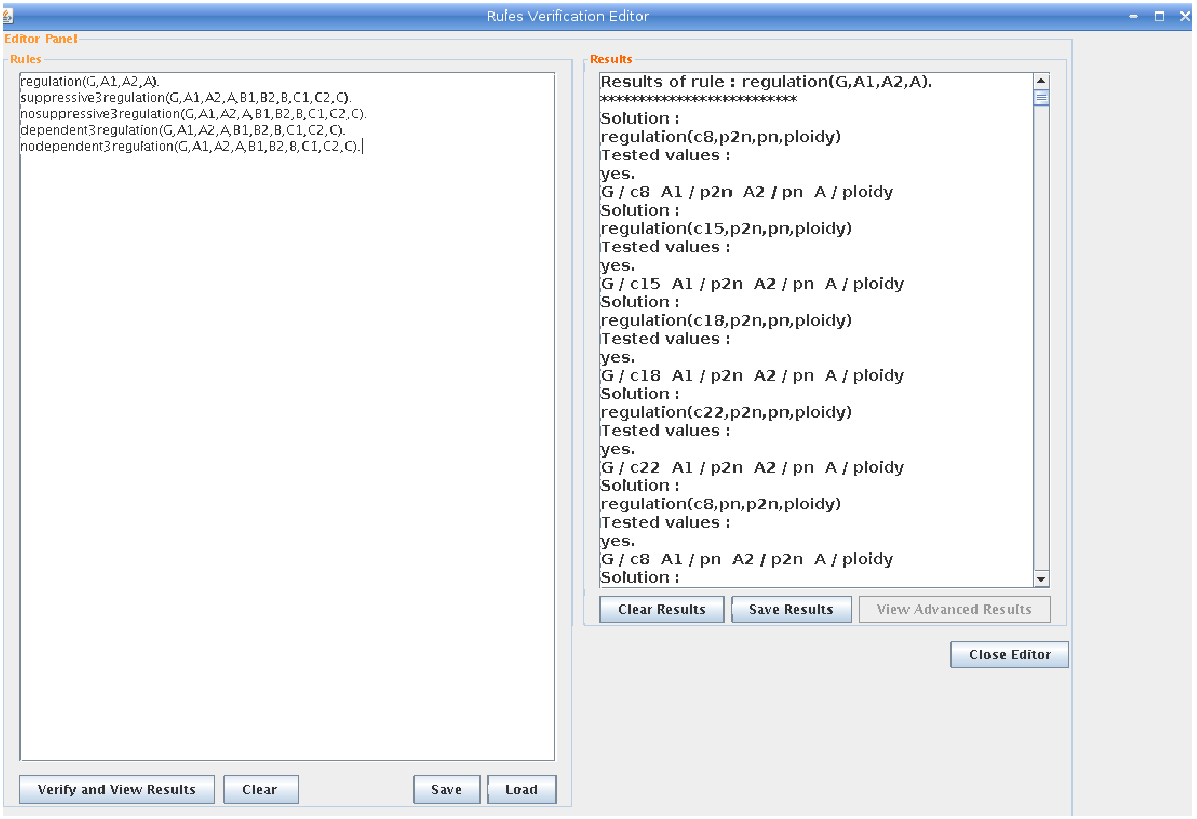


FIGURE 4.11 – Interface d'interrogation d'une base de connaissances dans XRegRules.

et d'extraction de réseaux de régulation. Nous souhaiterions également développer une fonction qui permette d'annoter automatiquement (présence de modulations et sens des modulations) les profils d'expression des gènes ou des groupes de gènes afin d'automatiser l'écriture des prédicats de type *faits* dans la base de connaissances.
Nous souhaiterions également rendre disponible cet outil complet sous deux formes :

- une application client/serveur avec une base de données systémiques concernant la levure mise à jour. Cette application faciliterait l'application de la méthodologie *XReglPath* aux données de la levure et ne nécessiterait pas d'installation complexe chez l'utilisateur.
- une application à charger et à installer par l'utilisateur qui rendrait possible la construction de bases de données pour d'autres organismes que la levure.

E Méthodes de classification automatiques : application à la classification de gènes

L'analyse de grands volumes de données d'expression est une problématique relativement récente en comparaison de problématiques similaires, nécéssitant des méthodes de reconnaissance de formes, bien établies dans d'autres domaines d'application. Parmi ces méthodes, la classification permet d'une part de réduire la dimension des objets à

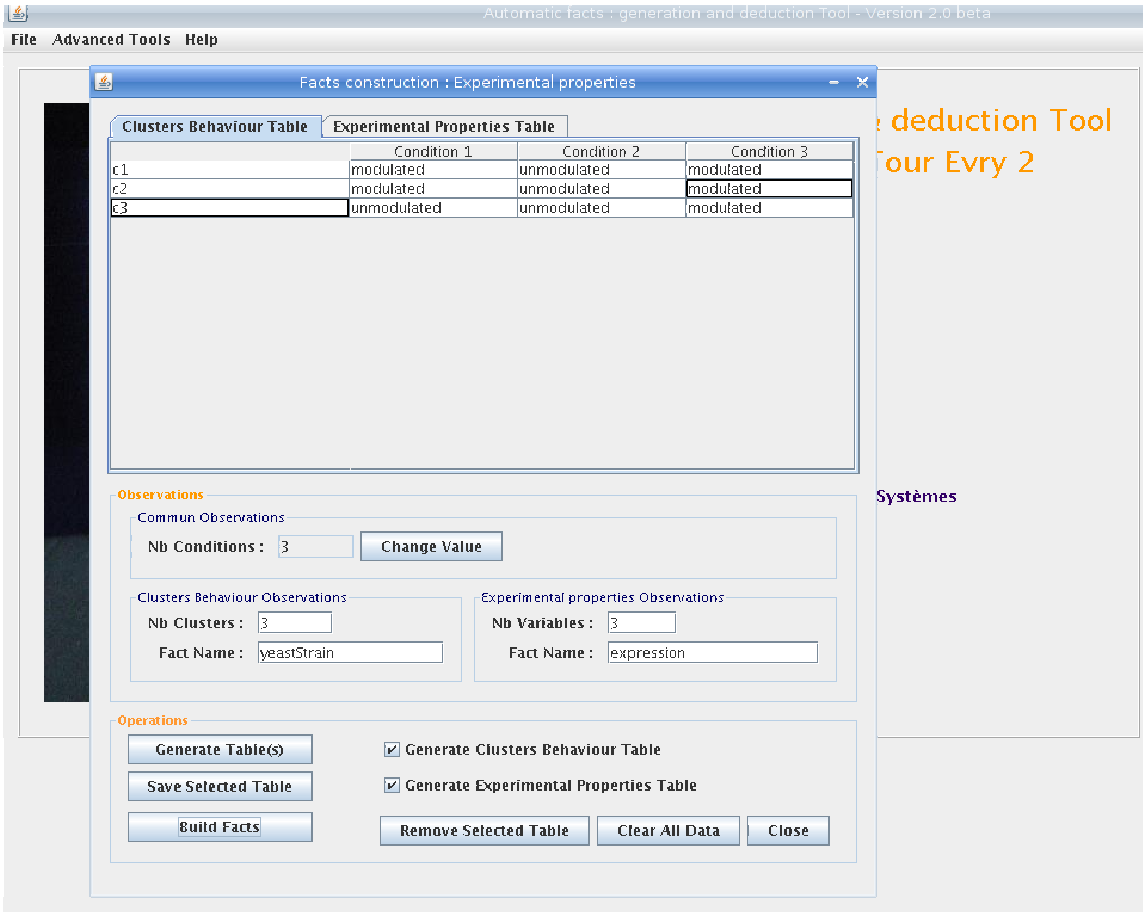


FIGURE 4.12 Interface d'aide à l'écriture automatique de prédicats de type *faits*.

manipuler et permet d'autre part de caractériser des formes et extraire des paramètres caractéristiques de ces objets. Nous nous cantonnerons à la présentation des méthodes de classification non supervisées, les méthodes supervisées dépassant le cadre de cette thèse. Dans le cas de la classification non supervisée, nous disposons de données non étiquetées et nous souhaitons découvrir des classes parmi ces données, c'est-à-dire des groupes contenant des données qui se ressemblent. Étant donné un ensemble de profils d'expression de gènes, le but est de construire des groupes de gènes ayant des motifs d'expression similaires.

Pour définir un problème de classification, il faut d'abord définir une notion de dissimilarité (ou de similarité) qui exprime à quel point les données sont dissemblables. Ensuite, le problème d'apprentissage des classes se traduit classiquement par un problème de maximisation de la compacité des classes et de la séparabilité de celles-ci. La compacité des classes est mesurée par l'inertie intra-classe W et la séparabilité des classes par l'inertie inter-classe B avec :

$$W = \frac{1}{2} \sum_{k=1}^{K} \sum_{x_i \in C_k, x_j \in C_k} d(x_i, x_j) \tag{4.1}$$

et

$$B = \frac{1}{2} \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{x_i \in C_k x_j \in C_{k'}, C_k \neq C_{k'}} d(x_i, x_j) \tag{4.2}$$

avec x_i et x_j couple de données parmi l'ensemble des données à classer, d une fonction mesurant une dissimilarité et K le nombre total de classes. On peut exprimer l'inertie totale T d'un ensemble de données partitionnées en K classes comme la somme $T = W + B$ des deux termes précédents. L'inertie totale étant une caractéristique des données et ne dépendant en rien des paramètres de la classification, on montre que minimiser l'inertie intra-classe revient à maximiser la séparabilité des classes. On ne considère donc qu'un seul de ces deux critères.

E.1 Dissimilarités et similarités entre données.

Définir une dissimilarité consiste à définir une mesure de la dissemblance ou, de façon équivalente la ressemblance (similarité) entre deux individus puisque l'on peut passer simplement de l'un à l'autre par une transformation linéaire.
L'objet mathématique le plus général traduisant une dissimilarité, est une fonction d, de l'ensemble des paires (x_i, x_j) d'individus dans l'ensemble des réels tel que :

- $d(x_i, x_j) \geq 0$,
- d symétrique : $d(x_i, x_j) = d(x_j, x_i)$ et,
- $d(x_i, x_i) = 0$ pour tout x_i.

Cette définition est peu contraignante et couvre un grand nombre de mesures possibles. En revanche, elle n'ouvre que peu de propriétés mathématiques et il est nécessaire d'ajouter d'autres contraintes pour acquérir certaines propriétés intéressantes.

E.1.1 Mesures de distances

Les dissimilarités les plus usuelles sont des distances. Une distance, est une dissimilarité particulière obtenue en ajoutant les conditions :

- $d(x_i, x_j) = 0 \Rightarrow x_i = x_j$ et,
- $d(x_i, x_j) \leq d(x_i, x_l) + d(x_j, x_l)$, l'inégalité triangulaire entre trois individus.

La manière la plus simple de définir une distance lorsque les données sont représentées sous forme vectorielle consiste à utiliser une p-norme. Parmi les distances, la famille la plus importante est constituée des distances dites de Minkowski d'ordre p ($p \geq 1$). Étant donné deux vecteurs $X_i = (x_{il})_{l=1}^{M}$ et $X_j = (x_{jl})_{l=1}^{M}$, de dimension M, la distance s'écrit alors :

$$d(x_i, x_j) = (\sum_{l=1}^{M} \mid x_{il} - x_{jl} \mid^p)^{1/p} \tag{4.3}$$

En utilisant différentes valeurs de p, on obtient les différentes distances à base de p-normes. Mais les seuls cas utilisés en pratique correspondent aux valeurs 1 et 2 de p.

Quand $p = 1$, l'indice est connu sous le nom de distance de Manhattan ou *city-block*,

$$d_{x_i,x_j} = \sum_{l=1}^{M} \mid x_{il} - x_{jl} \mid . \tag{4.4}$$

Pour $p = 2$, on retrouve l'habituelle distance euclidienne,

$$d_{x_i,x_j} = (\sum_{l=1}^{M} \mid x_{il} - x_{jl} \mid^2)^{1/2}. \tag{4.5}$$

Ces mesures sont bien adaptée pour la recherche de correspondances exactes entre deux vecteurs. La distance euclidienne permet d'identifier des structures sphériques dans les données et est efficace lorsque celles-ci sont bien compactes et isolées. Ces mesures restent cependant assez sensibles aux points distants. La distance de Mahalanobis est une version plus sophistiquée de ces mesures. Cette distance prend aussi en compte d'éventuelles corrélations entre cinétiques et est moins sensible aux différences d'amplitude.
Bien que très largement utilisée, la distance euclidienne n'est pas adéquate pour les séries temporelles. Cette mesure de distance est très sensible aux changement d'échelle et aux décalages. De plus, la forme, l'ordre des mesures, la longueur des séries temporelles et la taille des intervalles entre les mesures ne sont pas considérés dans la mesure où la distance est calculée à partir des sommes des distances à chaque temps de mesure.

E.1.2 Mesures de corrélation

Les mesures basées sur des corrélations cherchent à mesurer à quel point les coordonnées des vecteurs considérées varient de manière liée. Elles mesurent le degré de similarité entre les formes (modulations) des profils d'expression sans l'influence des différences d'amplitude. Les coefficients de corrélation incluent les métriques paramétriques (coefficient *Pearson*, *cosinus*) et non paramétriques (rang de *Spearman* et τ de *Kendall*).
Le coefficient de corrélation de Pearson entre deux vecteurs de données X_i et X_j de taille M est défini par :

$$R_{ij} = \frac{\sum_{l=1}^{M}(x_{il} - \bar{x_i})(x_{jl} - \bar{x_j})}{\sqrt{\sum_{l=1}^{M}(x_{il} - \bar{x_i})^2}\sqrt{\sum_{l=1}^{M}(x_{jl} - \bar{x_j})^2}} \tag{4.6}$$

Le calcul du coefficient de corrélation de Pearson exige que les données soient normalement distribuées. Lorsque la distribution ne suit pas une loi normale, le coefficient de corrélation de Pearson peut conduire à des résultats erronés. Une façon de remédier à cette situation consiste à utliser la corrélation de Spearman basée sur les rangs. Ce coefficient de corrélation ne diffère de celui de Pearson que par la conversion des valeurs en rangs avant le calcul du coefficient. L'équation 4.6 pourra alors être simplifiée en remplaçant les valeurs observées par les numéros des rangs.
En considérant les séries temporelles comme des vecteurs, les coefficients de corrélation mesurent le cosinus de l'angle entre les vecteurs d'expression moins leurs moyennes. Ainsi, des vecteurs avec un coefficient de corrélation égal à 0, seront dits orthogonaux (non corrélés), avec un avec un coefficient de corrélation égal à 1, seront dits parallèles dans la

même direction (corrélés positivement), et avec un coefficient de corrélation égal à -1, ils seront dits parallèles mais de directions opposées (corrélés négativement). Des vecteurs corrélés positivement correspondront à des cinétiques d'expression similaires.

Comme pour la distance euclidienne, les coefficients de corrélation vus ici postulent que les différents points d'une cinétique d'expression sont indépendant et distribués de façon identique. Ces deux types de mesures de dissimilarités restent invariant par rapport à l'ordre des observations : si l'ordre des valeurs d'une paire de cinétiques est permuté, leurs corrélations ou distances euclidiennes ne varieront pas. De plus, l'utilisation de coefficients de corrélation devient difficile lorsque les cinétiques d'expression de gènes sont courtes, ce qui est très souvent le cas.

Le choix d'une dissimilarité permettant d'évaluer à quel point deux données sont dissemblables (ou inversement, le choix d'une similarité) sous-tendent la définition des algorithmes de classification. Ce choix est influencé par le type des données et le contexte de la tâche à effectuer. Dans le cas des profils cinétiques de gènes, le choix d'une dissimilarité peut aussi correspondre à une requête biologique spécifique afin de classer les cinétiques selon des critères définis à l'avance. Dans ce cas, la dissimilarité devra refléter des propriétés mathématiques particulières et être adaptée au type de l'algorithme de classification choisi. Dans la suite de cette section, je présenterai quelques exemples de dissimilarités ou de similarités ayant des propriétés et usages différents.

E.1.3 Utilisation de fonctions *splines*

Lorsque les données sont fonctionnelles, on peut chercher à mesurer la distance entre les fonctions. Notons que c'est le cas des cinétiques d'expressions de gènes. Lorsque les cinétiques sont suffisamment longues, il est possible alors de les modéliser par des fonctions polynomiales appelées *splines* [324, 325, 326]. A partir de chaque profil mesuré, on réalise une régression non paramétrique en cherchant la décomposition dans une base de splines qui se rapproche le plus des données. Chaque profil d'expression peut être modélisé par une fonction *spline* estimée à partir des données observées. Chaque point de la cinétique intervient dans le lissage global du profil d'expression. Pour établir une distance entre deux profils, il suffit d'utiliser la distance euclidienne entre les vecteurs de coefficients obtenus par régression pour chacun des profils.
Lors de l'étape de classification, l'algorithme choisi n'utilise pas directement sur les profils d'expression mais leurs représentations lissées et effectuera les comparaisons sur les coefficients des fonctions polynomiales. Ce type de représentation permet de travailler sur des profils lissés et modélisés et donc d'être moins sensible à des points de cinétiques aberrants. De plus, cela permet de prendre en compte l'aspect dynamique et structuré des cinétiques. Notons cependant que, dans le cas général, les petites tailles des profils d'expression mesurés ne permettent que très rarement l'application de cette méthode.

E.1.4 Similarités et fonctions noyaux

Propriétés générales des similarités

Une similarité entre deux données mesure à quel point deux données se ressemblent. Elle possède les mêmes propriétés de symétrie qu'une dissimilarité mais avec aussi les propriétés suivantes :

- elle est définie positive : $\forall(x_i, x_j), s(x_i, x_j) \geq 0$,
- elle est bornée par la similarité d'un élément avec lui-même : $s(x_i, x_j) \leq s(x_i, x_i)$ et $\forall(x_i, x_j), s(x_i, x_j) = s(x_i, x_i) \Leftrightarrow x_i = x_j$.

La similarité est souvent normalisée en mettant sa valeur maximale à 1, créant ainsi la possibilité d'une interprétation probabilistique de la similarité.

Les fonctions noyaux

Les noyaux positifs semi-définis (appelés improprement noyaux) forment une classe de similarité intéressante. Les noyaux permettent d'étendre des algorithmes d'apprentissage linéaires classiques à des données qui ne sont pas décrites sous forme vectorielle, mais qui sont représentées par des objets plus complexes, tels que des séquences, des arbres ou des graphes.

Le principe de base des méthodes à noyau consiste à effectuer un changement de représentation des données : on considère que l'on applique aux données une transformation non linéaire $\phi : \chi \rightarrow F$. L'espace F est un espace de dimension élevée, voire infinie, qui doit mieux représenter les données. On impose comme seule contrainte à F d'être un espace de Hilbert, c'est-à-dire de disposer d'un produit scalaire.
Le deuxième principe de base des méthodes à noyau consiste à ne pas travailler directement sur ces transformées ϕ(x) qui pourraient donner lieu à des calculs coûteux étant donnée la dimension de l'espace F. Dans ce but, les fonctions de coût ou les algorithmes sont reformulés de telle sorte que les données n'interviennent pas individuellement, mais uniquement par le biais de produits scalaires. L'astuce du noyau consiste alors à introduire une fonction noyau k : $\chi \times \chi \rightarrow \mathbb{R}$ telle que :

$$\forall x_i, x_j \, \langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j) \tag{4.7}$$

Ainsi le produit scalaire s'écrit comme une fonction des données sous leur représentation initiale : la fonction noyau permet de calculer les produits scalaires dans l'espace des caractéristiques sans utiliser explicitement les transformées ϕ(x), et dépend uniquement des représentations des données dans l'espace initial. On peut donc travailler dans l'espace F sans même connaître sa dimension, et sans connaître la fonction de transformation ϕ. F sert ainsi uniquement sur le plan théorique, mais les données ne sont jamais manipulées dans cet espace.

Pour les données vectorielles, l'avantage est que l'on peut chercher des frontières de décision complexes, en conservant un formalisme linéaire : l'algorithme fournit un résultat linéaire dans l'espace F, mais non linéaire dans l'espace des données χ. De plus, l'astuce du noyau permet d'appliquer les algorithmes classiques à des données non vectorielles, et en particulier des données structurées telles que des séries temporelles.

Exemple de noyaux pour des données vectorielles

La première fonction noyau que l'on peut utiliser est le produit scalaire dans l'espace initial ou son équivalent normalisé :

$$k(x_i, x_j) = \langle x_i, x_j \rangle \qquad ou \qquad k(x_i, x_j) = \frac{\langle x_i, x_j \rangle}{\sqrt{\langle x_i, x_i \rangle \langle x_j, x_j \rangle}} \tag{4.8}$$

Les noyaux polynomiaux sont définis par :

$$k(x_i, x_j) = \left(\frac{\langle x_i, x_j \rangle}{a} + l \right)^b \tag{4.9}$$

où a,b et l sont les paramètres du noyau.
Enfin, les noyaux gaussiens sont définis quant à eux en fonction d'un paramètre σ comme suit :

$$k(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2} \right) \tag{4.10}$$

On se contentera ici d'une interprétation intuitive de ce noyau : la valeur du noyau pour x_i et x_j évolue inversement par rapport à leur distance euclidienne. L'exponentielle écrase la valeur du noyau pour les données très distantes. Le paramètre σ permet de contrôler la forme de la gaussienne.

Il existe de nombreuses méthodes pour construire des noyaux car cette famille de fonctions, semi-définies positives, est close par différentes opérations (voir [327] pour plus de détails). Ainsi, à part les noyaux de base pour lesquels la propriété de "semi-définie positivité" a été définie, tels les noyaux présentés ci-dessus, il est possible de construire d'autres noyaux. Dans les méthodes de classification automatique, la famille des noyaux représente donc une source possible de définition de similarités, même si toutes les propriétés de ces noyaux ne sont pas exploitées.

Fonctions noyaux pour l'analyse des cinétiques d'expression de gènes

Les noyaux gaussiens et polynomiaux correspondent aux fonctions noyaux les plus utilisées. Cependant, ces noyaux, dont l'application est classique pour les données vectorielles, ne peuvent capturer directement les dépendances locales existant entre les différents points d'une série temporelle.
Lors de l'analyse de la réponse transcriptionnelle à l'IR chez la levure, nous avons utilisé le noyau gaussien pour calculer les similarités, entre cinétiques d'expression, utilisées pour la classification des gènes en groupes co-exprimés (voir article I, page 87). Afin d'introduire la dimension temporelle dans le calcul des similarités, nous avons remplacé chaque cinétique par le vecteur de ses dérivées instantanées. Nous avons ainsi introduit la relation de dépendance qui lie les points consécutifs d'une série temporelle dans le calcul des similarités entre cinétiques d'expression. De cette façon également, les similarités calculées seront basées sur les formes des cinétiques plutôt que sur leurs amplitudes.
Une autre famille de similarités (ou distances), bien connue dans les domaines de la reconnaissance de voix et d'écriture, permet de projeter une séquence sur une autre à l'aide

d'un ensemble d'opérations élémentaires comme des substitutions et des répétitions de motifs. En associant à chacune de ces opérations un score, ces similarités utilisent la programmation dynamique pour calculer une séquence optimale d'opérations avec un score élevé. Différents travaux ont proposé de dériver des fonctions noyaux à partir de ces scores dont ceux utilisés dans l'algorithme d'alignement temporel dynamique ou *Dynamic-Time-Wraping* (*DTW*) [328, 329], celui utilisé dans l'algorithme de Smith et Waterman [330] et l'*edit-distance* [331], qui permet le calcul d'une distance entre deux chaînes de caractères. En s'inspirant de ces travaux, Cuturi et Vert [332] ont proposé un noyau qui permet de calculer une similarité basée sur l'alignement de deux séquences. Contrairement aux méthodes dont il est inspiré, leur noyau ne se base pas uniquement sur un alignement optimal mais sur un ensemble d'alignements proches. Ainsi, d'après leur noyau, deux séquences seront similaires si elles partagent un large ensemble de bons alignements plutôt qu'un seul alignement optimal. Cuturi et Vert proposent d'appliquer ce noyau à l'analyse de séries temporelles mais ne le testent que sur un ensemble de données vocales en choisissant des séries de données de dimension 13. Cette famille de noyau permet d'incorporer une plus grande richesse d'information que les noyaux de type gaussien ou polynomial, cependant, à cause des différentes opérations de substitution ou de répétition, leur application reste limitée à de longues séquences.

E.2 Algorithmes de classification non supervisée.

Un bon algorithme pour la classification de données d'expression doit pouvoir traiter avec un grand volume de données bruitées, doit être insensible à l'ordre des données en entrée, doit pouvoir réaliser la classification avec des complexités mémoire et temps raisonnables, ne doit pas nécessiter trop de paramètres à évaluer, doit pouvoir traiter des données structurées et complexes et enfin doit produire un résultat interprétable du point de vue de la biologie.

Les méthodes de classification regroupent deux familles de méthodes : la première famille, celle de la quantification vectorielle, se fonde sur une représentation vectorielle des données et la recherche d'un codage des données à l'aide d'un dictionnaire fini de vecteurs, représentant les différentes classes. Selon les méthodes ces vecteurs sont soit des prototypes, soit des centre de gravité. L'inertie intra-classe est remplacée par un critère de distorsion. Nous présentons succinctement deux méthodes de ce type : les k-moyennes et les cartes auto-organisatrices. La deuxième famille de méthodes est basée sur l'hypothèse que nous disposons d'une matrice de dissimilarité ou de similarité entre les données (les données ne sont pas représentées par des vecteurs). Les méthodes de cette deuxième famille cherchent à regrouper des données qui se ressemblent en minimisant avec plus ou moins d'efficacité un critère d'inertie. A l'issue de ces méthodes, on ne dispose pas d'un représentant des classes. Parmi cette famille de méthodes nous présentons la classification hiérarchique et la classification basée sur la théorie des graphes dont la classification spectrale.

E.2.1 Classification par quantification vectorielle

L'algorithme des *K-moyennes*

L'algorithme des *K-moyennes* [333, 334] est l'algorithme de partitionnement type basé sur le principe de la quantification vectorielle. Cet algorithme cherche à minimiser un critère de distorsion. Soit un ensemble de M données, une méthode de partition va diviser les données en K sous-ensembles où chaque sous-ensemble représente une classe et $K < M$. L'algorithme des *K-moyenne* est déroulé de façon à obtenir une partition des données qui optimise la fonction objectif suivante :

$$E = \sum_{i=1}^{K} \sum_{x_j \in C_i} \mid x_j - m_i \mid^2 . \tag{4.11}$$

Dans cette équation, x_j est une donnée à classer, m_i est le centre de gravité (centroïde) de la classe C_i, et K est le paramètre d'entrée qui fixe la taille de la partition.
Le déroulement des *K-moyennes* et présenté dans l'algorithme 1. L'algorithme des *K-*

Algorithm 1 Algorithme des *K-moyennes*

Entres: l'ensemble des M données, nombre de classes K
Sorties: une partition $P = (C_1, \ldots, C_i, \ldots, C_k)$

1. Initialisation :
   initialiser aléatoirement une partition $P = (C_1, \ldots, C_i, \ldots, C_k)$ et attribuer chaque objet de M à une classe de P ;
   générer une organisation aléatoire O des objects dans M ;
   pour $i = 1$ à k faire
   attribuer le prototype G_i au centre de gravité des objets de C_i ;
   fin ;
2. Étape d'allocation et de représentation :
   test $= 0$;
   pour $j = 1$ à n faire
   trouver la classe C_m de x_j ;
   trouver la classe C_l telle que $C_l = min_{i=1,\ldots,k} E(x_j, G_i)$;
   si $m \neq l$ alors
   test $= 1$;
   $C_l = C_l \cup x_j$ et $C_m = C_m - x_j$;
   recalculer les prototypes G_m et G_l ;
   fin ;
   fin ;
3. Condition d'arrêt :
   si test $= 0$ arrêt, sinon aller à (b) ;

moyennes est relativement simple et intuitif et est courament utilisé pour l'analyse de données transcriptomiques [272, 335, 336, 337, 338]. En l'absence de problèmes numériques, l'algorithme converge rapidement (de l'ordre de la dizaine d'itérations) vers une

solution stable. Cependant, cet algorithme nécessite de fixer deux paramètres : la taille K de la partition et les centres initiaux des classes. L'initialisation aléatoire des centres des classes est la stratégie la plus courante mais elle reste sensible à des problèmes de minima locaux. Dans ce cas la vitesse de convergence et la stabilité de l'algorithme peuvent varier en fonction de l'initialisation. Le choix de la taille K est réalisé *a posteriori* (voir section E.4, page 245 pour une revue des différentes méthodes d'estimation de la taille des partitions). Le choix de ces deux paramètres peut être facilité par des connaissances *a priori* sur la structure des données.

Cartes auto-organisatrices de Kohonen

Les cartes auto-organisatrices de Kohonen ou *Self-Organizing Maps* (*SOM*), ont été développées par Kohonen en 1984 sur la base d'un réseau de neurones à une couche [339]. Les données sont présentées en entrée et les neurones en sortie sont organisés selon une simple structure de voisinage, généralement une grille en deux dimensions (voir figure 4.13). Chaque neurone (chaque case de la grille) est associé à un vecteur de référence et chaque donnée est projetée sur la grille en étant associée au neurone qui possède le vecteur référence le plus proche. Durant le déroulement de l'algorithme, chaque donnée sera utilisée comme donnée d'apprentissage et influencera les mouvements des vecteurs référence vers les zones les plus dense de l'espace d'entrée. Ainsi, les vecteurs référence sont appris pour correspondre au mieux à la distribution des points d'entrée. Lorsque la phase d'apprentissage est achevée les classes sont identifiées par l'attribution de chaque donnée d'entrée à un neurone de sortie auquel un poids est associé.
Supposons que le vecteur référence d'un neurone N à l'itération i soit noté comme $f_i(N)$. L'initialisation de f_0 est aléatoire. Durant les étapes d'actualisation successives, un point x est sélectionné et le neurone N_x le plus proche de x est identifié. Le neurone est alors ajusté en le modifiant pour le rapprocher de x selon l'équation suivante :

$$f_{i+1}(N) = f_i(N) + \tau(d(N, N_x), i)(x - f_i(x)). \tag{4.12}$$

avec d la distance euclidienne au carrée norme L2. Le paramètre τ détermine le taux d'apprentissage et décroît lorsque le neurone N s'éloigne de N_x et lorsque le nombre d'itérations i augmente. Classiquement τ est choisi comme une fonction gaussienne.
On peut résumer l'algorithme *SOM* selon l'algorithme 2.

Algorithm 2 Algorithme de Kohonen

Entres: l'ensemble des M données, nombre de classes K, paramètre τ
Sorties: une partition $P = (C_1, \ldots, C_i, \ldots, C_k)$ sous la forme d'une grille de voisinage

1. initialisation aléatoire des poids des arcs entre neurones et vecteurs référence ;
2. sélection aléatoire d'un objet x parmi M ;
3. trouver le neurone $N_{\tau m}$ tel que $N_{\tau m} = min_{i=1,\ldots,k;j=1,\ldots,o} d(x, N_{i,j})$;
4. Actualiser les poids du neurone $N_{\tau m}$ et son voisinage pour les rapprocher de x selon la fonction f (voir équation 4.12 et τ)
5. reprendre à l'étape (b) jusqu'à ce que le critère d'arrêt soit atteint ;

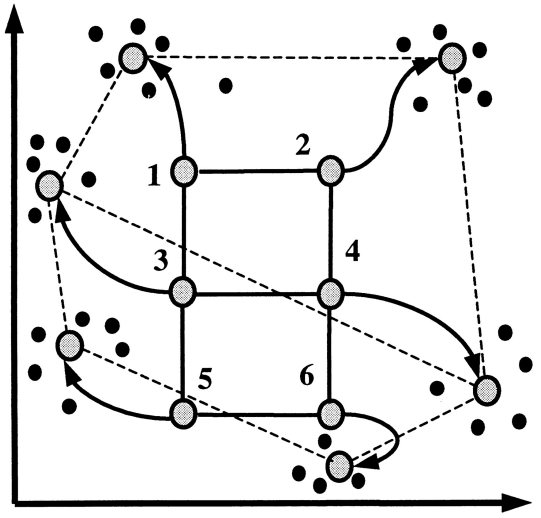


FIGURE 4.13 – Principe des cartes de Kohonen (selon [117]). L'organisation initiale des neurones dans une grille rectangulaire 3×2 est représentée par les traits pleins. Les trajectoires hypothétiques des neurones lorsqu'ils migrent, durant les itérations successives de l'algorithme SOM, pour se rapprocher des données sont représentées par des flèches. Les données sont représentées par des points noirs et les 6 neurones de la carte par des cercles grisés.

Les *SOM* produisent à la fois une partition des données et une visualisation cartographiée (généralement en deux dimensions mais des versions à trois dimensions ou sphériques existent aussi) de leur organisation. Cependant la procédure d'actualisation lors de la phase d'apprentissage reste similaire à celles d'autres algorithmes classiques comme les K-moyennes. La différence fondamentale reste que durant la phase d'apprentissage, les *SOM* utilisent chaque point pour actualiser tous les vecteurs référents (ceux du voisinage essentiellement) et pas seulement celui du neurone le plus proche. Les *SOM* ne résolvent pas les limitations du *K-moyennes* concernant le choix de la taille de la partition et la fixation des autres paramètres initiaux (initialisation des vecteurs, taux d'apprentissage, nombre d'itérations, taille et forme de la grille). Les ancêtres des cartes auto-organisatrices, les algorithmes comme *k-moyennes*, réalisent la discrétisation de l'espace d'entrée en modifiant à chaque cycle d'adaptation qu'un seul vecteur référent. Leur processus d'apprentissage est donc très long. L'algorithme de Kohonen profite des relations de voisinage dans la grille pour réaliser une discrétisation dans un temps très court. Il est essentiel d'observer que l'algorithme présente des opérations très simples donc qu'il est très léger du point de vue du coût de calculs.
Outre la rapidité d'exécution et sa relative stabilité, l'algorithme des SOM offre à l'analyse des cinétiques d'expression de gènes la possibilité d'interpréter facilement le résultat d'une classification grâce à la représentation d'une partition sous la forme d'une carte de voisinage [117, 340, 341]. Le voisinage dans les cartes de Kohonen est malheureusement fixe, et une liaison entre neurones ne peut être cassée même pour mieux représenter des données discontinues. Les *Growing Cell Structure*, ou *Growing Neural Gas* [342] peuvent représenter une solution à ce problème. Des neurones et les liaisons entre neurones peuvent

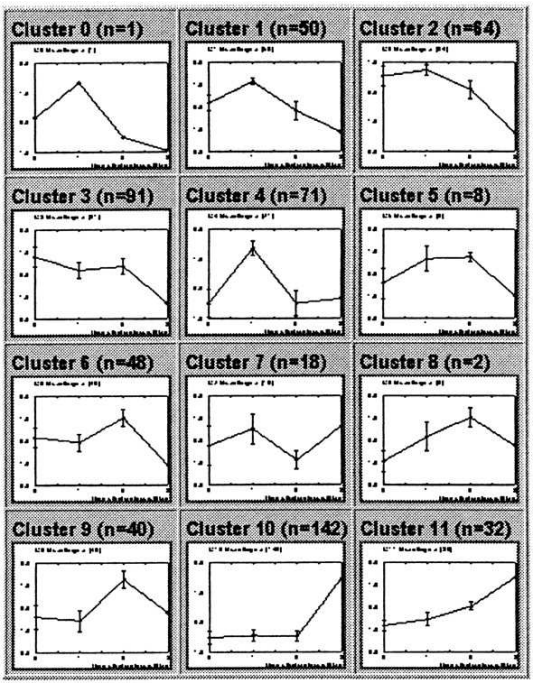


FIGURE 4.14 – Exemple de résultat de classification de cinétiques d'expression de gènes à l'aide de SOM (selon [117]). Chaque classe est représentée par le profil moyen de ses gènes avec les écarts types à chaque point.

y être supprimées ou ajoutées quand le besoin s'en fait sentir.

E.2.2 Méthodes de classification basée sur une matrice de dissimilarité : l'exemple de la classification hiérarchique.

Contrairement aux algorithmes des *K-moyennes* et des *SOM*, la classification hiérarchique ne nécessite pas que l'on spécifie le nombre de classes à rechercher. Par contre, on doit définir une mesure de dissimilarité entre des groupes d'observations, basée sur les paires de dissimilarités à travers les observations des deux groupes. Ces méthodes produisent une représentation hiérarchique dans laquelle les classes, à chaque niveau de la hiérarchie, sont crées par fusion des classes immédiatement inférieures. Au niveau le plus bas, chaque classe ne contient qu'une seule observation. Au niveau le plus élevé, il n'y a qu'une seule classe regroupant toutes les données.

Parmi les algorithmes hiérarchiques, on différencie les algorithmes agglomératifs (ascendants) des algorithmes divisifs (descendants). Les algorithmes agglomératifs initient le processus de classification en considérant chaque objet comme une classe et, à chaque itération fusionnent les paires de classes les plus proches en une nouvelle classe jusqu'à ce que tous les objets de l'ensemble à classer soient réunis en une seule classe. La paire d'observations choisie pour la fusion, correspond aux deux groupes ayant les plus petites dissimilarités inter-classes. A l'inverse, les algorithmes divisifs ou descendants débutent avec une seule classe contenant tous les objets et, à chaque itération vont diviser une classe en deux jusqu'à ce que l'on obtienne autant de classes que l'on a d'objets. L'al-

gorithme choisira la division qui produira deux nouveaux groupes avec la plus grande dissimilarité intra-classe.
Chaque niveau de la hiérarchie représente une partition particulière des données en classes disjointes. La hiérarchie complète représente une séquence ordonnée de ces partitions. Ce sera à l'utilisateur de définir le niveau qui correspondra à la partition "naturelle" de ses données. L'approche par *Gap statistic*, présentée à la section E.4 (245), peut aider au choix du niveau de partition optimal.

Pour les algorithmes agglomératifs, différentes types de mesures de dissimilarité existent et spécifient différentes stratégies de fusion de classes. Soit G et H deux classes, la dissimilarité $d(G, H)$ entre ces classes est calculée à partir de l'ensemble des dissimilarités entre paires d'observations $d(x_i, x_j)$ où x_i est dans G et x_j est dans H.
La classification agglomérative par lien simple (LS) va choisir la paire d'observations qui minimise la dissimilarité inter-classes :

$$d_{LS}(G, H) = min_{x_i \in G, x_j \in H} d(x_i, x_j). \tag{4.13}$$

La classification agglomérative par lien complet (LC) va choisir la paire d'observations qui maximise la dissimilarité inter-classes :

$$d_{LC}(G, H) = max_{x_i \in G, x_j \in H} d(x_i, x_j). \tag{4.14}$$

La classification agglomérative par lien moyen (LM) va utiliser la dissimilarité moyenne entre les classes :

$$d_{LM}(G, H) = \frac{1}{N_G N_H} \sum_{x_i \in G} \sum_{x_j \in H} d(x_i, x_j). \tag{4.15}$$

où N_G et N_H correspondent aux nombres d'observations dans chaque classe.
Si on défini le diamètre D_G d'un groupe d'observations comme la valeur de dissimilarité la plus élevée au sein de ses membres

$$D_G = max_{x_i \in G, x_j \in H} d(x_i, x_j), \tag{4.16}$$

la classification agglomérative par lien simple aura tendance à construire des groupes avec des diamètres larges. Le lien complet, représentera l'extrême opposé, avec des groupes compacts de faibles diamètres. Le lien moyen correspond à un compromis entre ces deux extrêmes, dans la mesure où il va produire des classes relativement compactes et distantes les unes des autres.

Les algorithmes de classification divisive débutent en prenant l'ensemble des données comme une classe unique, puis, ils divisent de façon récursive une des classes existantes en deux classes filles à chaque itération, de façon descendante. Le choix d'un algorithme divisif pourra être guidé par la volonté d'obtenir un petit nombre de classes. Les divisions successives pourront être obtenues en appliquant de façon récursive une méthode combinatoire comme les *k-moyennes*, avec $K = 2$, pour effectuer une séparation à chaque itération. Cependant, ce type d'approche dépendra de l'initialisation des *K-moyennes* au

début de chaque itération. De plus, ce type d'approche ne produira pas nécessairement une séquence de divisions possédant la propriété de monotonicité requise pour la représentation en dendogramme. Une solution pour éviter ces problèmes a été proposée par Kaufman et Rousseeuw [343] et consiste à diviser, à chaque niveau, la classe dont le diamètre est le grand. Une alternative serait de choisir la classe dont la dissimilarité moyenne, parmi ses membre,s est la plus grande

$$\bar{d}_G = \frac{1}{N_G} \sum_{x_i \in G} \sum_{x_j \in G} d(x_i, x_j), \tag{4.17}$$

Les divisions récursives continueront jusqu'à ce que, soit, toutes les classes deviennent des *singletons*, soit, tous les membres de chaque classe soient identiques entre eux ($d(x_i, x_j) = 0$).

Comme les approches heuristiques, les algorithmes de classification hiérarchique souffrent de leur manque de stabilité. En effet, les objets sont regroupés selon des décisions locales sans garantie d'une optimisation globale. Ce problème est exacerbé par la nature déterministe de la classification hiérarchique où, lorsqu'une itération est achevée, il n'y a pas de retour en arrière ou de correction possible. Enfin, la représentation des objets en arbres phylogénétiques semble mieux convenir à des organisations reflétant de vrais liens hiérarchiques comme l'évolution des espèces qu'à des similarités de comportement de gènes. Cependant, cette classe de méthodes, par sa simplicité, a été parmi les premières à être appliquées à la classification de cinétiques d'expression de gènes [185, 319, 344, 345, 108, 346] (voir figure 4.15). De plus, la structure hiérarchique des résultats produits permet de facilement visualiser et interpréter la représentation hiérarchique de tous les niveaux de partition possibles.

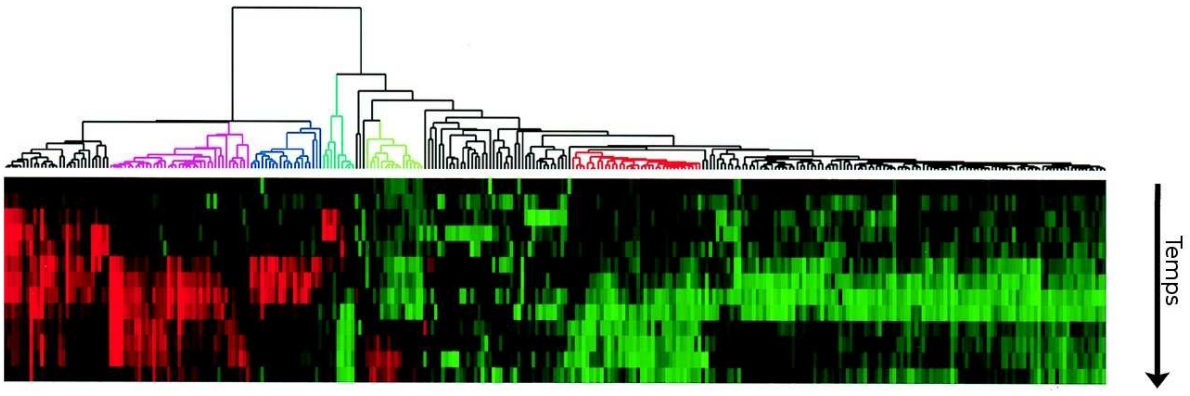


FIGURE 4.15 – Exemple de résultat de classification hiérarchique de cinétiques d'expression de gènes (selon [319]). Ici l'algorithme utilisé correspond à une classification hiérarchique agglomérative par lien moyen. Associé au dendogramme, la représentation par code couleur des intensités d'expression relatives (rouge : modulation positive de l'expression, vert : modulation négative de l'expression).

E.2.3 Théorie des graphes et classification spectrale

Classification basée sur la théorie des graphes

Les approches basées sur la théories des graphes connaissent depuis quelques temps un fort engouement pour l'analyse de grands jeux de données complexes. Ces approches posent toutes le problème de la classification comme un problème de partition au sein d'un graphe. Le graphe à partitionner est un graphe complet non dirigé dont les nœuds représentent les objets à classer et dont les arêtes sont pondérées par les valeurs contenues dans une matrice de similarité. Le problème consiste alors à partitionner le graphe $G = (V, E)$, où V représente l'ensemble des objets ($\mid V \mid = n$) et E a les mêmes valeurs que la matrice de similarité A.

Hartuv *et coll.* [347] proposent l'algorithme *HCS* qui cherche à identifier des composants fortement connectés, en tant que classe, dans un graphe de proximité. L'algorithme *HCS* basique prend en entrée la graphe de proximité G. Il calcule d'abord la coupe minimale de G, qui produit le graphe déconnecté avec le minimum d'arêtes enlevées. Si la coupe rejoint un certain critère (le nombre d'arêtes enlevées par la coupe est supérieur à la moitié du nombre de sommets), G est retourné en tant que classe. Sinon, il est séparé en ses deux plus larges parties connectées, selon cette coupe minimale. Ensuite, chaque partie suit le même processus de façon récursive. L'algorithme s'arrête lorsque chaque donnée est soit attribuée à une classe soit devient un *singleton*.

L'algorithme *CLICK* (CLuster Identification via Connectivity Kernals) [348] est basé sur le même principe que celui de l'algorithme *HCS*. Cependant, il suppose que : après normalisation, les valeurs des similarités entre paires d'éléments (qu'ils soient ou non dans la même classe) sont distribués normalement. Avec ce postulat, le poids w_{ij} d'une arête (ij) est défini de façon à refléter la probabilité que i et j soient dans la même classe. *CLICK* étend également l'algorithme *HCS* par une *étape d'adoption*, qui prends en charge les singletons restant et actualise les classes courantes et par une *étape de fusion* qui fusionne de façon itérative deux classes avec une similarité dépassant un seuil prédéfini.

L'algorithme *CAST* (*Cluster Affinity Search Technique*) [349] suppose que le graphe est constitué de *cliques* (une union disjointe de sous-graphes complets), $H = (U, E)$, qui représenteraient un ensemble idéal de données d'expression. Dans ce cas, l'ensemble des données à classer est vu comme une "contamination" du graphe idéal H par du bruit aléatoire. Dans un graphe constitué de *cliques*, chaque clique représente une classe. Pour chaque paire de gènes dans G, le modèle suppose qu'une arête/non-arête a été assignée par erreur avec une probabilité α. On suppose que la vraie partition de G est celle qui requière le moins de modifications d'arêtes pour générer H.
CAST utilise en entrée une matrice de similarités S, $n \times n$, réelle et symétrique et un seuil d'affinité t, où $0 \leq t \leq l$. L'algorithme construit les classes une par une. La classe en construction est notée C_{open}. Chaque élément x a une valeur d'affinité envers C_{open} définie

en tant que

$$a(x) = \sum_{y \in C_{open}} S(x, y). \tag{4.18}$$

Un élément a une forte affinité si il satisfait $a(x) \geq t \mid C_{open} \mid$. Sinon, x a une faible affinité. L'algorithme *CAST* alterne entre ajout d'éléments de fortes affinité avec la classe courante et suppression d'éléments de faibles affinités. Lorsque le processus se stabilise, C_{open} est considéré comme une classe complète et le processus continue avec une nouvelle classe, jusqu'à ce que tous les éléments soient assignés à une classe.
L'avantage principal de *CAST* est que le nombre des classes est déterminé par l'algorithme sans avoir besoin de connaissances *a priori* sur leur structure. Cet algorithme est basé sur un modèle stochastique pour la formation des classes et pour les erreurs de données, il peut donc être analysé de façon probabiliste. L'algorithme emploie aussi une stratégie incrémentale : il identifie d'abord un sous-ensemble des éléments, de grande qualité, comme noyau de la classe, puis, il génère et ajuste la classe complète avec des méthodes moins complexes. Cependant, il est important de noter que le seuil d'affinité et le taux d'erreur influencent indirectement la structure de la classe. De plus, il n'existe pas de preuves pour la complexité temporelle ni pour la convergence de l'heuristique. Le résultat final reste dépendant de l'ordre des données en entrée car lorsqu'une partition est initialisée, un nœud v est attribué à la classe pour laquelle il a la plus grande affinité.

Classification spectrale

Les méthodes de classification spectrale connaissent depuis quelques années un engouement et un développement spectaculaires. Ces méthodes commencent déjà à concurrencer les méthodes de classification classiques. Les méthodes spectrales sont particulièrement attractives car elle sont faciles à implémenter et sont relativement rapides (pour la classification de jeux de données de tailles supérieures à plusieurs milliers). Ces méthodes ne souffrent pas non plus de problèmes de minima locaux (cela reste cependant dépendant du type d'algorithme choisi). Les algorithmes de classification spectrale prennent en entrée des matrices de similarité, leur utilisation semble donc bien appropriée à la classification de gènes sur la base de leurs similarités de profils d'expression.

La classification spectral est né à travers le partitionnement de graphes en 1972 [350]. En effet, la classification peut être vu comme une partition d'un graphe de similarités, graphe composé d'un ensemble de noeuds représentant les données et d'un ensemble d'arêtes pondérées représentant la similarité entre les sommets. Une bonne classification aura par conséquent pour but de maximiser les poids des connexions intra-groupes et de minimiser les poids des connexions inter-groupes. Une coupe est définie comme ceci pour une bipartition en 2 groupes A et B :

$$Cut(A, B) = \sum_{i \in A, j \in B} W_{ij} \tag{4.19}$$

W_{ij} étant le poids de l'arête entre le sommet i dans A et le sommet j dans B.

Le critère de coupe d'un graphe sera par conséquent de minimiser la coupe vu précédemment :

$$\min_{A,B} Cut(A,B) \tag{4.20}$$

Mais dans certains cas, la coupe n'est pas du tout optimale car ce type de coupe considère uniquement les connexions externes des *clusters* et ne considère pas les densités internes du *cluster*. C'est pourquoi il existe différents critères de coupe et notamment celui de Shi et Malik [351] utilisant la coupe normalisée, qui considère la connexion entre les groupes relativement à la densité de chaque groupe et qui normalise l'association entre les groupes par le volume $vol(A)$ qui correspond au poids total des arêtes provenant du groupe A.

$$\min NCut(A,B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)} \tag{4.21}$$

Mais un problème se pose. Il s'avère que le calcul de la coupe de graphe normalisée est *NP*-difficile et seule une relaxation permet de résoudre le problème. L'approche possible afin de résoudre le problème est basée sur la théorie spectrale des graphes. Elle représente le graphe de similarité sous forme de matrice et applique des méthodes d'algèbre linéaire telles que le calcul de vecteurs propres ordonnés en utilisant les valeurs propres correspondantes sachant que les vecteurs et valeurs propres fournissent une information globale sur la structure de la matrice.

Construction d'une matrice d'adjacence A
La matrice d'adjacence du graphe est une matrice $n \times n$ ayant pour valeur :

$$A = [W_{ij}] \tag{4.22}$$

où comme auparavant W_{ij} est le poids de l'arête entre le sommet i et le sommet j. A est donc une matrice symétrique ayant des vecteurs propres réels et orthogonaux.

Matrice de degrés D
Cette matrice permet de normaliser la matrice d'adjacence. C'est une matrice diagonale de taille $n \times n$ et de valeur :

$$D(i,i) = \sum_j W_{ij} \tag{4.23}$$

La matrice "laplacien"
Le laplacien est le principal composant du clustering spectral. Nous allons donc voir les deux différents laplaciens. Le premier laplacien possible est le laplacien non normalisé. Il s'oppose au laplacien normalisé. Le non normalisé se défini comme ayant la matrice $L = D - A$ alors que la matrice du laplacien normalisé est définie comme ceci : $L = D^{-1/2}(D - A)D^{-1/2}$.

Une étude concernant ces deux laplaciens a été réalisée par Von Luxburg et Bousquet [352] afin de comparer ces deux méthodes différentes de clustering spectral. Ils ont

notamment analysé la convergence de ces méthodes. Ils en arrivent à la conclusion que le clustering spectral normalisé (utilisant le laplacien normalisé) converge sous certaines restrictions, contrairement au clustering spectral non normalisé qui converge uniquement avec de lourdes suppositions, qui pour la plupart sont difficilement satisfaites.
Un algorithme de multi-partitionnement de graphe par clustering spectral suit les étapes générales suivantes :

1. Construction du laplacien A' à partir de la matrice d'adjacence
2. Trouver les vecteurs propres et les valeurs propres de A'
3. Construire un espace à partir de vecteurs propres correspondant aux k premières valeurs propres
4. Utiliser l'algorithme *k-moyennes* afin de réduire l'espace en k classes

Quelques exemples d'algorithmes de classification spectrale Il existe plusieurs algorithmes possibles, implémentant la classification spectrale. Voici un aperçu des algorithmes les plus connus : *Shi and Malik* [351] et *Ng, Jordan and Weiss* [353]

Algorithm 3 Algorithme de Classification Spectrale : Shi and Malik [351]

Entres: La matrice noyau S $\in \mathbb{R}^{n \times n}$, k : nombre de classes choisies
Sorties: Classes $C_1, \ldots, C_k$

- Calculer le laplacien non-normalisé L
- Calculer les k premiers vecteurs propres $v_2, \ldots, v_{k+1}$ du problème de vecteur propre $\mathrm{L}v = \lambda \mathrm{D}v$
- Soit $V \in \mathbb{R}^{n \times k}$ la matrice contenant les vecteurs $v_2, \ldots, v_{k+1}$ en colonnes. V_i étant la i-ème colonne de V.
- Classer les points V_i avec l'algorithme des *K-means* dans les classes $C_1, \ldots, C_k$

Le succès des méthodes de classification spectrale est essentiellement basé sur le fait que l'on ne présuppose pas la forme des classes à obtenir. Contrairement à l'algorithme *K-means* classique où les classes identifiées forment des ensembles convexes, la classification spectrale peut résoudre des problèmes plus généraux comme l'identification de spirales entrelacées. De plus, l'implémentation de ces méthodes est relativement facile et leur application à de très grands jeux de données se révèle très efficace lorsque la classification est effectuée à partir d'un graphe de similarités épars.

Dans le cadre de cette thèse, nous avons choisi d'implémenter l'algorithme de classification spectrale de Ng *et coll.* [353], version normalisée et possédant de meilleurs propriétés de convergence que l'algorithme proposé par Shi *et coll.* [351].

E.3 Classifications de type *biclustering.*

La notion de *biclustering* a été introduite dans les années 70 par Hartigan [354]. Cheng et Church [355] ont été les premiers à appliquer cette stratégie à des données d'expression de gènes. Ils ont défini un bloc (*bicluster)* comme une sous-matrice uniforme (avec une faible somme des résidus quadratiques) et utilisent une approche gloutonne pour définir des blocs. En effet, l'information contenue dans un jeu de données d'expression, représenté

Algorithm 4 Algorithme de Clustering Spectral : Ng , Jordan and Weiss [353]

Entres: La matrice noyau S $\in \mathbb{R}^{n \times n}$, k : nombre de classes choisies
Sorties: Classes $C_1, \ldots, C_k$

- Calculer le laplacien normalisé
- Calculer les k premiers vecteurs propres $v_2, \ldots, v_{k+1}$ du laplacien L.
- Soit $V \in \mathbb{R}^{n \times k}$ la matrice contenant les vecteurs $v_2, \ldots, v_{k+1}$ en colonnes.
- Former la matrice Y à partir de V en normalisant chaque ligne de V pour avoir :

$$Y_{ij} = V_{ij}/(\sum_j X_{ij}^2)^{1/2}$$

- Classer chaque colonne Y_i avec l'algorithme des *K-moyennes* dans les classes $C_1, \ldots, C_k$

sous la forme d'une matrice *gènes* $\times$ *condition*, peut être étudiée dans les deux dimensions. On peut rechercher des motifs de co-expression au sein des gènes en fonction des conditions ou à l'inverse classer les conditions en fonction des gènes. Dans la suite de cette section nous présenterons quelques exemples d'algorithmes de *biclustering*. Ces algorithmes se basent sur l'idée qu'il existe des sous-ensembles de gènes qui ne sont co-régulés que dans des sous-ensembles de conditions. De la même façon il existerait des processus biologiques, illustrés par certaines conditions expérimentales, qui ne seraient associées qu'à certains sous-ensembles de gènes. Un gène dans ce cas pourrait soit faire partie de plusieurs classes différentes au sein d'une même partition soit ne faire partie d'aucune classe (voir [356] pour revue).

E.3.1 Les différents types de *biclusters*

On peut identifier 3 classes majeures :

- les *biclusters* à valeurs constantes,
- les *biclusters* à valeurs constantes sur les lignes et les colonnes,
- les *biclusters* à valeurs cohérentes.

Ces trois classes analysent directement les valeurs numériques de la matrice de données et essaient de trouver des sous-ensembles de lignes et de colonnes avec un "voisinage" similaire. Pour le cas de données d'expression de gènes, prenons l'exemple des biclusters à valeurs constantes, ceux-ci révèlent un sous-ensemble de gènes avec des expressions similaires pour un sous-ensemble de conditions.

Les biclusters à valeurs constantes :

Étant donnée la matrice $A = (X, Y)$, avec X comme ensemble de lignes et Y comme ensemble de colonnes, un *bicluster* est la sous-matrice (I, J) où I est un sous-ensemble de lignes et J un sous-ensemble de colonnes et a_{ij} est la valeur de la matrice A correspondant à la relation entre la ligne i et la colonne j. Un *bicluster* constant parfait est une sous matrice $A(I, J)$ où toutes les valeurs dans le bicluster sont égales pour tout $i \in I$ et tout $j \in J : a_{ij} = \mu$.

1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0

FIGURE 4.16 – *Biclusters* à valeurs constantes.

Les biclusters à valeurs constantes sur les lignes et les colonnes :

Un *bicluster* parfait à lignes constantes est une sous matrice $A(I, J)$ où toutes les valeurs dans le *bicluster* peuvent être obtenues par :

$$a_{ij} = \mu + \alpha_i$$
$$a_{ij} = \mu \times \alpha_i$$

Un *bicluster* parfait à colonnes constantes est défini quasiment de la même manière, à part que le coefficient α s'applique sur les colonnes de cette manière :

$$a_{ij} = \mu + \alpha_j$$
$$a_{ij} = \mu \times \alpha_j$$

1.0	1.0	1.0	1.0
2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0

1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0

FIGURE 4.17 – **Biclusters** à lignes et à colonnes constantes.

Les biclusters à valeurs cohérentes :

Pour qu'un *bicluster* à valeurs cohérentes soit parfait il faut qu'il soit défini comme ceci :

$$a_{ij} = \mu + \alpha_i + \mu_j$$
$$a_{ij} = \mu \times \alpha_i \times \mu_j$$

E.3.2 Structures des *biclusters*

Ces *biclusters* peuvent avoir différentes structures dans la matrice de similarité que l'on souhaite classer. Sachant que des *biclusters* peuvent se superposer, nous pouvons caractériser deux cas représentant plusieurs structures de *biclusters*. Le premier cas qui est le plus simple, est la situation où il n'y a qu'un unique *bicluster* dans la matrice (voir figure

1.0	2.0	5.0	0.0
2.0	3.0	6.0	1.0
4.0	5.0	8.0	3.0
5.0	6.0	9.0	4.0

1.0	2.0	0.5	1.5
2.0	4.0	1.0	3.0
4.0	8.0	2.0	6.0
3.0	6.0	1.5	4.5

FIGURE 4.18 – *Biclusters* à valeurs cohérentes

4.19(a)). Ensuite nous avons le cas où la matrice de données contient plusieurs *biclusters*. Pour ce second cas, différentes structures peuvent être obtenues (voir figure 4.19(b) jusqu'à figure 4.19(i)). En effet, nous pouvons tout d'abord avoir plusieurs *biclusters* où il n'y a aucune superposition, notamment la figure 4.19(c) où la structure *checkerboard*, ou encore structure en damier, est celle obtenue en effectuant une classification sur les lignes puis sur les colonnes. Puis nous pouvons avoir plusieurs *biclusters* mais avec superposition. Cette superposition peut être réalisée avec plus ou moins de structuration ou de hiérarchie entre les *biclusters*.

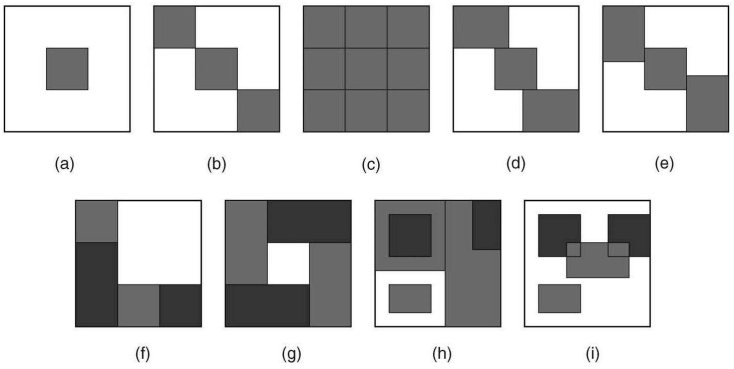


FIGURE 4.19 – Les différentes structures des *biclusters*
(a) *single bicluster*, (b) *exclusive row and column biclusters*, (c) *checkerboard structure*, (d) *exclusive rows biclusters*, (e) *exclusive columns biclusters*, (f) *nonoverlapping biclusters with tree structure*, (g) *nonoverlapping nonexclusive biclusters*, (h) *overlapping biclusters with hierarchical structure*, et (i) *arbitrarily positioned overlapping biclusters*.

E.3.3 Algorithme de Cheng et Church

Cheng et Church [355] ont développé une mesure de similarité appelée *somme des résidus quadratiques* pour mesurer la cohérence des gènes et des conditions au sein d'un bloc appelé *bicluster*. Un faible *résidu* sera un bon critère pour identifier un bloc. Soit un ensemble I de gènes et un ensemble J de conditions, le résidu quadratique moyen d'une sous-matrice est défini par :

$$H(I,J) = \frac{1}{\mid I \mid\mid J \mid} \sum_{i \in I, j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \tag{4.24}$$

où

$$a_{iJ} = \frac{1}{\mid J \mid} \sum_{j \in J} a_{ij}, a_{Ij} = \frac{1}{\mid I \mid} \sum_{i \in I} a_{ij} \tag{4.25}$$

et

$$a_{IJ} = \frac{1}{\mid I \mid\mid J \mid} \sum_{i \in I, j \in J} a_{ij} = \frac{1}{\mid I \mid} \sum_{i \in I} a_{iJ} = \frac{1}{\mid J \mid} \sum_{j \in J} a_{Ij} \tag{4.26}$$

sont les moyennes des lignes et de colonnes et la moyenne dans la sous-matrice A_{IJ}. Une sous-matrice A_{IJ} sera appelée δ-*bicluster* si $H(I, J) \leq \delta$ pour un seuil $\delta \geq 0$. Le score le bas $H(I, J) = 0$ indique que les expressions des gènes fluctuent de façon cohérente. Cela inclue les *biclusters constants* qui ne présentent aucune fluctuation. La variance des lignes peut être utilisée comme score additionnel pour rejeter les *biclusters constants* :

$$V_{IJ} = \frac{1}{\mid J \mid} \sum_{j \in J} a_{ij} - a_{Ij}^2. \tag{4.27}$$

La méthode "brute force" consiste à identifier un *bicluster* en calculant le score H pour toutes les possibles additions/délétions de lignes ou de colonnes et de choisir l'opération qui minimise H. Si aucune opération ne minimise H, ou si $H \leq \delta$, l'algorithme retourne le *bicluster*. Cheng et Church ont appliqué un algorithme *glouton*, avec une complexité inférieure à $O(mn)$, où n et m correspondent aux tailles des lignes et des colonnes de la matrice d'expression.

E.3.4 Classification par blocs

Tibshirani *et coll.* [357] ont ajouté une procédure *d'élagage* à l'algorithme de partitionnement par blocs proposé par Hartigan [354]. Ils ont aussi conçu une méthode, appelée *Gap Statistic*[358], qui permet de sélectionner le nombre optimal de blocs à partir de différentes permutations. Une représentation hiérarchique des classes peut être obtenue en limitant les coupes à des intersections entre blocs. La procédure débute avec toutes les données réunies en un seul bloc, puis l'algorithme choisi à chaque pas la colonne ou la ligne qui sépare un bloc existant en deux parties et qui minimise la variance *intra* bloc. L'algorithme continue le partitionnement des blocs jusqu'à ce qu'un grand nombre de blocs soit obtenu, puis, applique la méthode de *Gap Statistic* pour trouver le nombre optimal de blocs.

E.3.5 Modèle *plaid*

Lazzaroni *et coll.* [359] ont proposé une méthode basée sur la modélisation des données d'expression. L'idée consiste à considérer les données d'expression comme une somme de plusieurs termes appelés *couches* où chaque couche indique la présence d'un processus biologique particulier pour un sous-ensemble unique de gènes associé à un sous-ensemble unique de conditions. Chaque couche correspondra à un bloc. Cet algorithme est basé sur l'apprentissage d'un modèle pour les données d'expression. Les paramètres du modèle sont estimés à l'aide d'un algorithme *EM* dont le but est d'optimiser une fonction objectif en

ne parcourant qu'une direction à la fois jusqu'à ce qu'un minimum local soit atteint. Ce modèle permet à un gène d'appartenir à plusieurs blocs à la fois ou de n'être dans aucun bloc.

E.3.6 Algorithme *CTWC*

L'algorithme *CTWC* (*Coupled Two-Way Clustering*) [360] cherche à identifier des couples de petits ensembles de *caractéristiques* (F_i) et *d'objets* (O_j) où F_i et O_j peuvent être soit des gènes soit des conditions tel que lorsque l'on utilise seulement les *caractéristiques* dans F_i pour classer les *objets* O_j correspondants, on fasse émerger des partitions stables. CTWC fonctionne selon une heuristique et évite l'énumération *gloutonne* de toutes les combinaisons possibles : seuls les sous-ensembles de gènes ou de partitions reconnus comme stables sont retenus pour les itérations suivantes.

Parmi les travaux les plus récents, on peut noter les travaux de Ben-Dor et *coll.* [361] qui définissent un bloc comme un groupe de gènes dont les niveaux d'expression induisent un ordre linéaire à travers un ensemble de conditions. Segal *et coll.* [362] décrivent des modèles probabilistes afin d'étudier des relations entre expression, motifs de régulation et annotations des gènes.

E.3.7 Algorithme *SAMBA*

L'algorithme *SAMBA* recherche des blocs (*biclusters*) *gènes* $\times$ *conditions* en utilisant le *modèle d'évolution cohérente* [363]. Dans ce modèle les valeurs exactes des expressions des gènes ne sont pas directement prises en compte, mais, une classe est évaluée pour voir si elle présente un motif d'expression cohérent. Dans le modèle le plus simple, chaque mesure d'expression peut correspondre à un des trois états suivants : induit, réprimé et non modulé. Les seuils entre états sont cruciaux et on peut complexifier le modèle en introduisant des états intermédiaires comme par exemple "légèrement" induit et "fortement" induit.

Dans le modèle utilisé par *SAMBA*, la matrice des expressions est représentée par un graphe bipartite, $G = (U, V, E)$, où U est l'ensemble des nœuds correspondants aux conditions, $U \cap V = \varnothing$ et une arête (u, v) n'existe entre $v \in V$ et $u \in U$ si et seulement si il existe une modulation significative du niveau d'expression du gène v dans la condition u. Un des atouts de *SAMBA* réside dans la façon d'attribuer un score aux blocs (*biclusters*). Ce score correspond à la significativité statistique du bloc, où un poids est attribué à une arête (u, v) selon la vraisemblance d'obtenir ce poids par hasard, ($log\frac{P_c}{P_{(u,v)}} > 0$ pour les arêtes et $log\frac{1-P_c}{1-P_{(u,v)}} < 0$ pour les non-arêtes). La probabilité $P_{u,v}$ correspond à la proportion de graphes bipartites aléatoires, qui contiennent une arête (u, v), de degré identique à G. P_c est une probabilité constante $> max_{(u,v)\in U\times V} P_{u,v}$. L'attribution de ces poids aux arêtes et aux non-arêtes dans le graphe, la significativité statistique d'un sous-graphe H et les K plus grands (plus grands poids) sous-graphes pour chaque nœud

dans G peuvent être calculés et trouvés. Les auteurs proposent deux façon de calculer les poids des sous-graphes. La plus simple consiste à chercher les blocs qui représentent des modulations d'expression sans considération pour les signes de ces modulations. La deuxième façon consiste à cibler les *bi-cliques* régulières, c'est à dire les conditions dans lesquelles les modulations d'expression sont de même signe.

E.4 Estimation de la taille d'une partition.

La sélection automatique de la taille d'une partition dans une méthode de classification est à ce jour un problème non résolu. Il existe de nombreux critères ainsi que des algorithmes d'évaluation de ces critères qui peuvent aider à définir le choix du nombre de classes, toutefois l'aide de l'expert est en général requise pour prendre une décision. Une méthode pourra être plus efficace qu'une autre pour un jeu de données particulier mais échouera des données différentes. La diversité des problématiques liées à la l'analyse des données transcriptomiques, la diversité des données (nombres, conditions expérimentales, organismes), la diversité des techniques de classification mises en œuvre et la diversité des méthodes d'estimation du nombre optimal de classe empêche d'avoir une unique référence. Les quelques travaux comparatifs existants sont loin d'être exhaustifs et nécessiteraient des protocoles expérimentaux qui mettent en place toutes les combinaisons possibles de techniques et de jeux données. Il ne faut cependant pas désespérer, l'estimation d'un nombre de classes satisfaisant, si non optimal, reste possible tant que l'on est capable de justifier un choix méthodologique et d'interpréter un résultat tant du point de vue statistique que du point de vue biologique.
L'application d'un algorithme de classification à un jeu de données d'expression produira toujours une partition de ces données, que celle-ci reflète ou non une structure "naturelle". Dans le cas de la classification hiérarchique cela ne pose pas vraiment de problème (sauf si on veut choisir un niveau en particulier) car on obtient tous les niveaux d'organisation des données. Cependant, si le but est d'extraire l'organisation cachée de ces données, une partition artificielle ne sera pas satisfaisante et il faudra valider la pertinence et la stabilité d'une partition.
Classiquement, la validation d'une procédure de classification correspond à vérifier qu'une méthode a retrouvé la vraie structure contenue dans un jeu de données. On distingue dans ce cas deux types d'approches aux applications différentes. Dans le premier cas, on parle d'indices externes, utilisés plutôt dans le cadre de la classification supervisée. Ces indices ont pour but d'évaluer si une structure prédite correspond à une vraie structure connue. Dans le cadre de cette thèse je ne présenterais que le deuxième type d'approche, comprenant d'une part ce que l'on appelle les indices internes (basés sur l'analyse de la dispersion des classes), mais aussi la modélisation statistique des données, l'analyse de la stabilité des partitions, et une dernière classe, plus spécifique des données d'expression de gènes, où l'on va mesurer le contenu en information biologique des classes. Dans ces approches, l'approche est exploratoire et il n'existe pas d'information a priori sur la structure des données ou sur le nombre des classes.

E.4.1 Mesures de dispersion des classes

Ces indices se basent tous sur la comparaison des inerties ou dispersions intra-classes et inter-classes moyennes. On parle d'indices internes car ils sont calculés à partir des mêmes observations qui ont servi à la classification. On ne dispose pas en général d'un autre échantillon de données. Les méthodes de classification essayent de maximiser les distances entre classes, l'utilisation du test classique de comparaison des variances (*F-test*) ne peut dans ce cas être valide pour tester des différences entre classes. Fridlyand et Dudois ont présenté une excellente revue de ces méthodes [364] et ont en particulier reformulé certains indices sous forme d'hypothèses statistiques.
Supposons que le nombre maximal de classes dans un jeu de données soit fixé à M, $2 \leq M \leq n$. Une façon d'estimer le nombre de classes K est de rechercher le $\hat{K}$, $1 < \hat{K} \leq M$, qui fournit la statistique la plus significative contre l'hypothèse nulle H_0 ($K = 1$) selon laquelle il n'existe pas de classes au sein des données. Les deux hypothèses nulles courament utilisées sont l'hypothèse d'unimodalité et l'hypothèse d'uniformité.
Sous l'hypothèse d'unimodalité, les données sont vues comme étant un échantillon aléatoire issu d'une distribution normale multivariée. Ce modèle donne une grande probabilité de rejeter l'hypothèse nulle $K = 1$ si les données sont échantillonnées à partir d'une distribution dont le coefficient d'aplatissement est plus bas que celui de la distribution normale. L'hypothèse d'uniformité affirme que les données sont échantillonnée à partir d'une distribution uniforme dans un espace à p dimensions. Les méthodes basées sur l'hypothèse d'uniformité sont assez conservatives, c'est à dire qu'elles rejettent rarement l'hypothèse nulle H_0, lorsque les données sont échantillonnées à partir d'une distribution unimodale comme la distribution normale.
De nombreuses méthodes ont été proposées afin de tester l'hypothèse nulle $K = 1$ et estimer le nombre de classes contenues dans des données. La majorité de ces approches n'essayent pas de tester de façon formelle l'hypothèse nulle, $K = 1$, mais s'intéressent plutôt à la structure de la partition pour laquelle une mesure statistique sera optimale. Pour ces deux types d'hypothèses, les preuves permettant de rejeter l'hypothèse nulle H_0 peuvent être résumées de façon formelle en utilisant les différents indices internes décrit ci-dessous.
Pour une partition d'un ensemble de données en $1 \leq k \leq M$ classes, on défini B_k et W_k comme les matrices de tailles $p \times p$ des sommes des carrés inter- et intra-classes. B_1 restant non défini.

1. Kaufman et Rousseeuw [343] ont proposé une statistique dite *silhouette*. Pour une observation i, $a(i)$ est sa distance moyenne aux autres points de la même classe et $b(i)$ sa distance moyenne aux points contenus dans la classe la plus proche (celle qui minimise $b(i)$). La statistique *silhouette* est alors définie par :

$$s_i = \frac{b(i) - a(i)}{max(a(i), b(i))}, \tag{4.28}$$

   Un point sera considéré comme bien classé si s_i est grand. Les auteurs proposent de choisir le nombre de classes optimal $\hat{k}$ comme la valeur qui maximise le s_i moyen sur l'ensemble de données. Notons que s_i est non défini pour $k = 1$ classe.

2. Calinski et Harabasz [365] définissent pour chaque nombre de classe $k \geq 2$ un index :

$$ch_k = \frac{trB_k/(k-1)}{trW_k/(n-k)}, \tag{4.29}$$

où tr représente la trace d'une matrice, c'est à dire la somme des valeurs en diagonale. Le nombre de classes estimés est alors $argmax_{k\geq 2} ch_k$.

3. Krzanowski et Lai [366] définissent, pour chaque nombre de classes $k \geq 2$ les indices :

$$diff_k = (k-1)^{2/p} trW_{k-1} - k^{2/p} trW_k, \tag{4.30}$$

et

$$kl_k = \mid diff_k \mid / \mid diff_{k+1} \mid . \tag{4.31}$$

Le nombre estimé de classe est alors $argmax_{k\geq 2} kl_k$.

4. Hartigan [367] défini, pour chaque nombre de classes $k > 1$, un index :

$$hart_k = (\frac{trW_k}{trW_k})(n-k-1) \tag{4.32}$$

Le nombre de classes estimé correspond au plus petit $k \geq 1$ tel que $hart_k \leq 10$

5. Tibshirani et coll. [358] ont récemment proposé une méthode qui permet de comparer un indice interne comme la somme des carrés au sein d'une classe à son espérance sous une hypothèse nulle de référence. Pour chaque nombre de classes $k \geq 1$, on calcule la somme des carrés, trW_k, au sein d'une classe. On génère B (ici $B = 10$) ensembles de données de référence sous l'hypothèse nulle et on applique l'algorithme de classification pour chacun d'entre eux en calculant les sommes des carrés $trW_k^1, \ldots, trW_k^B$. On calcule l'indice *gap statistic* tel que :

$$gap_k = \frac{1}{b} \sum_b \log trW_k^b - \log trW_k \tag{4.33}$$

et, l'écart type standard sd_k de $logtrW_k^b$, $1 \leq b \leq B$. Soit $\tilde{sd}_k = sdk\sqrt{1+1/B}$. Le nombre estimé de classes est le plus petit $k \geq 1$ tel que $gap_k \geq gap_{k*} - \tilde{sd}_{k*}$, où $k* = argmax_{k\geq 1} gap_k$

Parmi tous ces indices il est important de noter que seuls les indices *hart* et *gap* permettent d'estimer s'il n'existe qu'une seule classe, $\hat{K} = 1$, au sein des données. Ce sont donc les seuls indices internes qui permettent de confirmer l'absence de structure dans les données.

Milligan et Cooper [368] ont réalisé une analyse complète de trente méthodes différentes pour l'estimation du nombre de classes. Parmi les méthodes les plus performantes l'index de Calinski et Harabsz venait en tête. Hastie et coll. [358] ont comparé l'efficacité des différents indices internes présentés dans cette section au *gap statistic*. Ils ont testé différents scénarios en jouant avec la dimension des données, le nombre de classes, la distribution des données et le nombre de simulations. Leurs résultats ont montré que la méthode d'Hartigan était la moins efficace alors que la méthode *gap statistic* arrivait en tête de la majorité des tests sauf dans les cas où les données n'était pas très bien séparés et pour le cas d'une forme oblongue des données en particulier.

E.4.2 Mesure de stabilité des partitions

La stabilité d'une partition à l'égard de perturbations comme l'échantillonnage ou l'ajout de bruit dans les données peut être considérée comme une propriété importante d'une méthode de classification. Ce type de propriété paraît particulièrement intéressent à étudier dans le cas des données d'expression, intrinsèquement bruitées. L'idée d'utiliser la stabilité pour évaluer un résultat de classification a déjà été utilisé [369, 370] et en particulier dans le contexte de la classification hiérarchique [371, 372]. Deux méthodes récentes utilisent la stabilité des partitions pour évaluer un résultat de classification. La première méthode, *Clest* [373], propose d'appliquer les techniques d'échantillonage pour estimer le nombre de classes contenues dans un jeu de données en observant la reproductibilité des affectations aux classes. Ils proposent d'estimer le nombre de classes K en effectuant des divisions aléatoires et successives des données de départ en deux ensembles disjoints, un ensemble d'apprentissage L^b et T^b. Pour chaque itération et pour chaque nombre de classe k, une partition $P(., L^b)$ de l'ensemble d'apprentissage L^b est réalisée et un prédicteur $C(., L^b)$ est construit en utilisant les étiquettes de classe à partir de la partition. Le prédicteur $C(., L^b)$ est ensuite appliqué à l'ensemble de test T^b et les étiquettes prédites sont comparée à celles produites par l'application de l'algorithme de classification à l'ensemble de test en utilisant une mesure de similarité. Le nombre de classes est estimé en comparant les similarités observées pour chaque K à leurs valeurs attendues sous une distribution nulle avec $K = 1$. Le nombre estimé de classes est défini comme étant le $\hat{K}$ correspondant à la plus grande statistique contre H_0 de $K = 1$.
La deuxième méthode, développée par Ben-Hur *et coll.* [374], échantillonne un espace de partition pour chaque choix K, et utilise une métrique basée sur la similarité entre partitions pour générer une distribution de valeurs de stabilité. Ben-Hur *et coll.* commencent par décrire une mesure de similarité introduite par Fowlkes et Mallows [372]. Premièrement, pour une partition d'un ensemble de données X, une matrice C est construite où $C_{ij} = 1$ si et seulement si x_i et x_j appartiennent à la même classe et $i \neq j$. Sinon, $C_{ij} = 0$. On peut alors définir une mesure de similarité entre deux partitions du même jeu de données en utilisant le produit scalaire de ces matrices C. Formellement ce produit scalaire peut être exprimé comme :

$$(C^1, C^2) = \sum_{i,j} C^1_{ij}, C^2_{ij} \tag{4.34}$$

En normalisant ce produit scalaire, on peut obtenir une corrélation ou un cosinus en tant que mesure de similarité définie par :

$$cor(C^1, C^2) = \frac{(C^1, C^2)}{\sqrt{(C^1, C^1)(C^2, C^2)}} \tag{4.35}$$

Cette métrique permet de comparer deux partitions de même taille K produites par le même algorithme de classification. De façon intuitive, nous pouvons avoir une idée de la stabilité d'un niveau de partitionnement de taille K. Pour plus de significativité, ces mesures sont réalisées sur un grand nombre de partitions réalisées à partir d'échantillons différents. Pour déterminer le nombre optimal de classes, deux sous-échantillons différents des données sont d'abord générés puis partitionnés en K classes. L'intersection de ces

partitions est alors utilisée pour le calcul de la similarité. Cela est répété pour un nombre d'itérations prédéfini et permet de générer une distribution des similarités pour chaque valeur de K comprise entre 2 et un maximum fixé. L'analyse des différentes distributions de similarité permet de choisir le nombre de classes K optimal. En pratique, Ben-Hur *et coll.* tracent les différentes distributions cumulées des valeurs de similarité et recherchent la présence d'une "transition de phase" claire indiquant une transition d'une partition stable à une partition moins stable (voir figure 4.20).

Fridlyand et Dudoit [364] et Ben-Hur *et coll.* [374] ont appliqué leur méthode à des

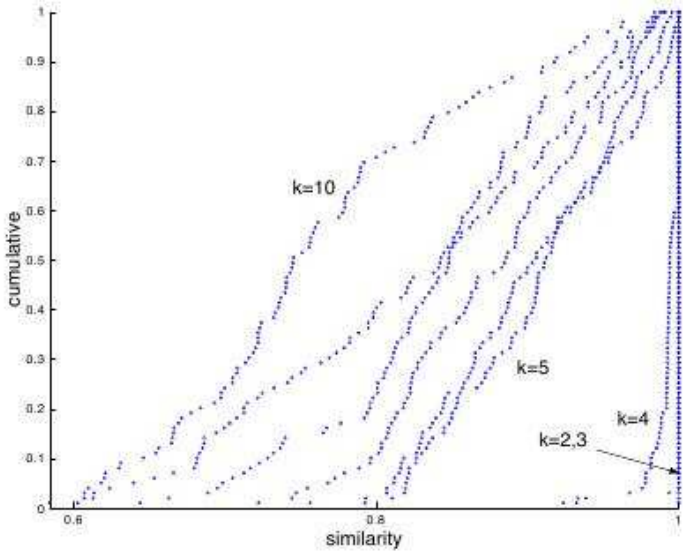


FIGURE 4.20 – Exemple de distributions de similarités cumulées entre partitions pour différents nombres de classes K (selon Ben-Hur *et coll.*, 2002 [374]). Dans cet exemple le passage de $K = 4$ à $K = 5$ met en évidence une "transition de phase" et une perte de stabilité de la partition, le K optimal dans ce cas est $K = 4$

données simulées et à des données d'expression expérimentales. Dans ces deux études, les résultats des méthodes présentées ont été comparés à ceux obtenus avec d'autres méthodes d'évaluation du nombre de classes. Dans chacune de ces deux études, les méthodes basées sur la stabilité des partitions arrivaient en tête, seul *gap statistic* produisait des résultats aussi satisfaisants. Contrairement au *gap statistic*, les deux méthodes présentées ici ne sont pas biaisées vers la recherche de classes compactes. De plus l'analyse de la stabilité des partition reste indépendante du type des données à classer et de l'algorithme de classification utilisé. Cependant, la significativité de ces méthodes reste liée au nombre de simulations effectuées pour chaque taille de partition K. A chaque taille de partition et pour chaque sous-échantillon des données on devra appliquer l'algorithme de classification. Dans ce cas la complexité temporelle sera fonction de la rapidité de l'algorithme de classification et de la taille des données à analyser (nombre de gènes) pour les algorithmes manipulant des matrices de similarités (méthodes spectrales par exemple).

E.4.3 Analyse de la pertinence biologique des partitions

Dans le cas de l'analyse des profils d'expression de gènes, la classification de gènes sur la base de leurs profils d'expression a pour but l'extraction de potentielles relations de régulation. La co-régulation de gènes est souvent corrélée à l'appartenance à un même groupe fonctionnel. De nombreuses études ont cherché à utiliser cette relation pour estimer la qualité d'une partition et un nombre optimal de classes. Ces différentes méthodes fonctionnent toutes sur le même principe : elles prennent comme point de départ une ou plusieurs partitions, et pour chacune des classes, vont essayer de mesurer la cohérence du groupe de gènes par rapport à un type d'information. En d'autre termes, elles vont mesurer le contenu en information biologique de ces classes et la significativité statistique de cette information. De façon intuitive, si un grand nombre de gènes d'une classe est associé de façon inattendue à une information ou une annotation biologique 'F', alors les gènes de cette classe peuvent être vraisemblablement décrit par cette annotation 'F'. Supposons qu'un nombre total G de gènes dans un génome ai été analysé à l'aide de puces à ADN. Parmi ces G gènes, m sont connus pour être associés à une annotation 'F'. Au sein d'une classe de gènes de taille D, h gènes sont associés à 'F'. Sous l'hypothèse nulle selon laquelle les gènes annotés par 'F' sont distribués de façon aléatoire dans les classes, h suit une distribution hypergéométrique. La P-valeur (probabilité d'observer h ou plus de gènes annotés dans la classe) est calculée selon :

$$P[X \geq h] = 1 - \sum_{i=0}^{h-1} \binom{D}{i} \binom{G-D}{m-i} / \binom{G}{M} \tag{4.36}$$

Les p-valeurs associées aux annotations fonctionnelles trouvées au sein d'une classe sont ensuite sélectionnées selon un seuil (inférieures au seuil) et combinées pour définir un score par classe. De la même façon les différents scores par classe sont combinés et normalisés pour définir à leur tour un score pour une partition de taille K. C'est ce score, associé à une partition qui permet de comparer le contenu en informations pertinentes de différentes partitions obtenues à partir d'algorithmes différents et pour des tailles K différentes.
Les différents travaux se basant sur ce type de stratégies utilisent quelques variantes pour l'évaluation de la pertinence d'une partition du point de vue d'informations biologiques. Gibbons et Roth [375], par exemple, utilisent un critère d'entropie pour attribuer un score à une partition.
La principale source d'information utilisée par ces méthodes [375, 376] repose sur les annotations fonctionnelles proposées par Gene Ontology (GO) [176]. Cependant, Raychaudhuri et Altman [377] proposent en plus des annotations GO un croisement des résultats avec le calcul d'un critère basé sur l'information contenue dans la littérature (associations gènes-publications). Jakt *et coll.* [378] enfin, proposent de valider la pertinence d'une partition à partir de la recherche de motifs communs de fixation à des régulateurs au sein de groupes de gènes co-exprimés. Ce dernier type d'information est particulièrement intéressent dans la mesure où il fait le lien avec l'hypothèse de co-régulation qui guide la plupart des stratégies de classification de gènes à partir de leurs motifs d'expression.
Ces méthodes paraissent séduisantes car parfaitement adaptées au type de données analysé. Cependant plusieurs objections peuvent être formulées. Tout d'abord, l'efficacité de ces méthodes dépend directement de la quantité d'information biologique disponible.

Il n'est pas étonnant de remarquer que ces méthodes ne sont pas ou peu appliquées à l'homme à cause du faible taux de gènes annotés. Ensuite, ces méthodes biaisent énormément les critères utilisés pour la validation d'une partition ou pour la détermination d'un nombre optimal de classes. D'une part, se focaliser sur des informations du type des ontologies fonctionnelles ou de motifs nucléiques communs ne signifie pas que d'autres types de relations entre gènes ne soient pas pertinents. Et, dans ce cas il y a le risque d'éliminer des classes pertinentes par rapport à d'autres types de relations (complexes protéiques, régulations indirectes, positions chromosomique, etc.). Enfin, dans une approche exploratoire, l'utilisation de ce type de méthodes est plus pertinente post-classification pour aider à l'interprétation des classes plutôt que pour guider leur obtention. De plus, l'analyse biologique des classes n'en sera que plus significative si on sait qu'elles ont été obtenues de façon indépendante de toute connaissance biologique *a priori*.

F Annexe VI : Les rayonnements ionisants et le vivant

F.1 Historique et aspects sociologiques

Il y a plus de cent ans, en Allemagne, Conrad Röentgen découvrait une source "inconnue" de radiation en plaçant la main de sa femme entre un tube cathodique et un écran fluorescent. Les rayons électromagnétiques traversèrent la main en révélant les phalanges et la bague de mariée. Ne sachant comment baptiser ces rayons invisibles et pénétrants, Röntgen les nomma "rayons X", du nom de l'inconnu algébrique habituel. Röntgen reçu pour sa découverte le premier Prix Nobel de physique en 1901. Fasciné par les travaux de Röntgen, le physicien français Antoine Henri Becquerel étudia les différences et similarités entre rayons X, fluorescence et phosphorescence et découvrait que dans tous ces cas l'énergie provenait d'une source émanante et pénétrait à travers la matière. Plus tard, Marie et Pierre Curie découvraient que ce phénomène correspondait à une propriété atomique de la matière et le nommait "radio-activité". Leurs travaux sur les éléments radioactifs leurs permirent de découvrir le polonium et le radium. Résultats qui leur permettront de recevoir le Prix Nobel de physique en 1903. Ernest Rutherford permettra plus tard une meilleur compréhension de la structure atomique et nucléaire de la matière. Les effets des radiations ionisantes (RI) sur le vivant restèrent cependant peu connus et il faudra plus de vingt années avant que le généticien Herman Müller ne démontre que les radiations ionisantes ne font pas que traverser la matière mais qu'elles peuvent aussi générer des mutations dans les organismes vivants.
Si les effets des RI sont aujourd'hui bien étudiés et leur utilisation bien maîtrisée, leur impact au niveau sociétal a engendré des comportements allant de l'inconscience absurde à la psychose irraisonnée. Pendant une bonne partie du XXe siècle, de nombreux médecins pensaient que le radium était totalement sûr. Dès 1908, le radium fut utilisé pour soigner de nombreuses affections, notamment cutanées. Dans les années 1930, des préparations à base de radium furent vantées pour guérir de presque toutes les maladies imaginables, de l'arthrite au cancer et de l'hypertension artérielle à la cécité. La défiance des population vis-à-vis de la radioactivité n'est réellement née qu'avec le développement de l'armement nucléaire et avec le choc des deux bombes A lâchées sur Hiroshima et Nagazaki en août

1945. La Seconde Guerre mondiale était terminée... mais à partir de ce moment le monde entrait dans la crainte d'une apocalypse nucléaire entretenue par la course à l'armement nucléaire des deux blocs est et ouest, de 1947 à la seconde moitié des années 1980. Le deuxième choc qui marqua fortement les esprits survenait le 26 avril 1986 ; le réacteur numéro 4 de la centrale nucléaire de Tchernobyl explosait et provoquait l'accident nucléaire le plus grave jamais survenu. Depuis ces évènements, l'inconscient collectif reste fortement défiant vis-à-vis de ce que l'on appelle le "nucléaire" et l'on fantasme beaucoup sur leurs effets. De plus, en France, quelques accidents récents de radiothérapie dus à des sur-doses d'irradiation (IR) n'ont fait qu'augmenter la défiance de la population à l'égard de tout type de radiation. Or de nos jours, tout le monde est exposé à des radiations qu'elles soient naturelles, dues aux activités humaines ou à but médical. Des sources aussi naturelles que la roche qui sert à construire nos maisons (granit radioactif en Bretagne ou dans le Massif Central) sont sources de radiations. Pour certains sociologues, la peur provoquée par les rayonnements radioactifs s'explique par l'ignorance et "l'invisibilité des radiations", le fait que l'homme ne soit pas "équipé" pour percevoir l'importance de la radioactivité. Sensible aux brûlures, suffoquant sous l'action d'un gaz polluant et fermant les yeux face à une intense source lumineuse, il subit, sans en avoir conscience, le passage des rayons ionisants.

F.2 Les différents types de radiations ionisantes

Si nous parlons généralement de radiations ionisantes, ce terme regroupe un ensemble de rayonnements physiques de différentes natures caractérisés par une fréquence (une énergie) leur permettant d'ioniser la matière. Il en existe différents types, aux sources, aux applications et aux conséquences différentes :

- Rayonnement alpha (α) : il s'agit de particules semblables à des noyaux d'hélium (2 protons + 2 neutrons). Ce rayonnement est très peu pénétrant, il ne traverse pas une feuille de papier. On utilise par exemple le rayonnement α dans les détecteurs de fumée des systèmes d'alarme incendie automatiques.
- Rayonnement bêta (β) : il s'agit d'électrons se déplaçant à haute vitesse. Ce rayonnement est arrêté par une plaque de plexiglas de 1 cm d'épaisseur. Dans les fabriques de papier, on mesure souvent l'épaisseur des feuilles à l'aide de sources de rayonnement β.
- Rayonnement neutronique : il s'agit de neutrons émis lors de la fission de noyaux lourds ou obtenus en bombardant certains noyaux avec des particules α (réaction α-n). L'application la plus courante des sources de neutrons est la mesure de l'humidité des sols.
- Rayonnement X : il s'agit d'ondes électromagnétiques, comme la lumière (UV), mais possédant une très grande énergie. Les radiations sont émises par l'enveloppe électronique de l'atome. Le rayonnement X est très pénétrant. L'application la plus connue des rayons X est la radiographie dans le diagnostic médical.
- Rayonnement gamma (γ) : il s'agit du même type de radiation que le rayonnement X, cependant ils sont émis par le noyau de l'atome et non par son enveloppe électronique. Le rayonnement γ est en général très pénétrant. A titre d'exemple d'utilisation du rayonnement γ, citons la radiographie de soudures et la radiothérapie médicale.

F.3 Effets des radiations sur le vivant

F.3.1 Effets physiques des radiations ionisantes

Lorsqu'un faisceau de RI pénètre dans les tissus, une partie du rayonnement est absorbée, une autre est déviée de sa trajectoire (diffusion) et la troisième partie est transmise sans interaction. La dose absorbée représente la quantité d'énergie absorbée par unité de matière. La dose absorbée est différente de l'énergie émise. Elle est exprimée en Gray (Gy), où 1 Gy correspond à un transfert à la matière d'une énergie de 1 joule par kilogramme. Par exemple, 70 Gy représentent la dose prescrite habituellement en radiothérapie dans le traitement des cancers ORL. Le phénomène de diffusion explique quand à lui pourquoi les régions situées hors du faisceau d'IR peuvent malgré tout recevoir une certaine dose de radiations. Les RI interagissent de manières différentes avec les tissus en fonction de leurs natures :

- Interactions des radiations électromagnétiques (rayons X et γ) avec les tissus : les photons incidents transfèrent leur énergie aux molécules du milieu traversé par différents mécanismes fondamentaux d'interaction. Ceci conduit à des ionisations ou à des excitations électroniques, puis à l'émission de photons secondaires d'énergie atténuée lors du retour des molécules à l'état stable. Ces photons secondaires seront par la suite, eux-mêmes à l'origine d'excitations et d'ionisations.
- Interactions des neutrons avec les tissus : les neutrons traversent les couches d'électrons des atomes en interagissent avec leurs noyaux. Ces interactions produisent des particules chargées secondaires : des protons rapides, des particules et des fragments nucléaires lourds responsables des ionisations et excitations du milieu. En fonction de leur énergie, les neutrons peuvent entrer en collision avec les noyaux avec lesquels ils interagissent et être ralentis ou ils peuvent être capturés par ces noyaux et provoquer des réactions nucléaires.
- Interaction des ions lourds avec les tissus : les particules lourdes et électriquement chargées, telles que les ions lourds, interagissent avec le milieu traversé par le biais de leur charge électrique. Ces particules cèdent immédiatement de l'énergie à un grand nombre d'électrons du milieu, dès qu'elles pénètrent dans ce milieu. Elles sont les seules à perdre leur énergie progressivement en arrachant des électrons aux atomes le long de leur parcours. Au début de leur trajet, elles se déplacent vite et le temps d'interaction avec la matière est faible : les ionisations sont peu nombreuses. Puis à chaque ionisation, elles perdent progressivement de la vitesse et le dépôt d'énergie par unité de longueur augmente jusqu'à l'arrêt de la particule. Ainsi, la plus grande quantité d'énergie est déposée en fin de parcours.

Tous les types d'IR aboutissent ainsi à une perte d'énergie le long d'un trajet. On définit ainsi un transfert linéique d'énergie qui est la quantité d'énergie perdue par unité de longueur. Plus la perte d'énergie est importante, moins la distance traversée par la particule est grande, et plus la zone traversée va subir d'ionisations. Ainsi, une particule légère très énergétique peut traverser le tissu sans provoquer de collision, et donc sans provoquer d'effet biologique.

F.3.2 Effets physico-chimique des radiations ionisantes

Ces effets durent quelques secondes à quelques minutes. Les molécules ionisées et excitées lors de la phase physique réagissent entre elles et avec les molécules voisines. On distingue l'effet direct dont la cible est représentée par les macromolécules cellulaires, notamment l'ADN, et l 'effet indirect qui est la conséquence de l'attaque des macromolécules par les radicaux libres issus de la radiolyse de l'eau et qui survient de façon prépondérante, dans 70 à 80% des cas. En présence d'oxygène, on assiste à la création de radicaux à fort pouvoir oxydant qui interagissent pour aboutir à la formation d'eau oxygénée, particulièrement oxydante. Une IR pratiquée en présence d'oxygène va générer la formation d'un plus grand nombre de molécules d'eau oxygénée qu'en conditions hypoxiques ; ceci permet de comprendre "l'effet oxygène" qui traduit une plus grande radio-sensibilité cellulaire en présence d'oxygène par rapport à des conditions hypoxiques.

F.3.3 Effets biologiques des radiations ionisantes

Il y a cinquante ans, juste après les effets dévastateurs des bombes atomiques au Japon, l'étude des effets des RI sur le vivant devenait un important sujet de recherche. De premières études posaient déja la question du risque de cancers après exposition à de fortes doses de radiations [379, 380]. Puis, l'utilisation de nucléosides radioactifs (thymidine tritiée) avec autoradiographies a permis d'observer les premiers effets au niveau cellulaire, notamment sur la prolifération in vivo [381]. Depuis les années 60, de très grands efforts ont été réalisés pour étudier et comprendre les conséquences biologiques des RI. Les RI vont potentiellement interagir avec toutes les macromolécules biologiques. Les effets ionisants des radiations créent des réactions oxydatives qui modifient physiquement les macromolécules, changeant ainsi leurs conformations voire leurs fonctions [297, 296]. De même, l'hydrolyse des molécules d'eau introduit une seconde source de stress oxydatif sous la forme de radicaux libres qui peuvent à leur tour altérer, dégrader ou lier les macromolécules cellulaires [298, 299].

Effets au niveau des membranes : les RI peuvent causer des dommages au niveau des différentes membranes, plasmiques, nucléaire [300] et mitochondriales [301, 382]. Ces dommages correspondent à des modifications de structures entraînant souvent des modifications fonctionnelles. Les deux cibles attaquées sont les protéines membranaires et les phospholipides constitutifs des membranes. Les protéines membranaires sont particulièrement sensibles aux attaques de radicaux libres. L'attaque de protéines constituants les canaux ioniques, les jonctions cellulaires ou impliquées dans les voies de signalisation entraîne des dysfonctionnements ou des inactivations perturbant gravement la physiologie cellulaire. Les attaques radicalaires affectent aussi les acides gras polyinsaturés des phospholipides en provoquant des peroxydations lipidiques à l'origine d'une diminution de la fluidité membranaire.

Effet génotoxique : un des principaux effets cellulaires des radiations correspond à l'endommagement de l'ADN [50, 51]. Ces dommages à l'ADN peuvent persister si ils ne sont

pas réparés et peuvent altérer irrémédiablement le message génétique voire transmettre cette altération à la descendance. Les radiations ionisantes peuvent causer différents types de lésions au niveau de l'ADN :

- Altérations au niveau des bases nucléiques : l'attaque de radicaux libres tout comme l'action directe des radiations peuvent endommager les bases nucléiques. Les bases peuvent être hydroxylées et changer de conformation. Les radiation peuvent aussi induire des délétions de bases nucléiques. Ces lésions sont généralement rapidement réparées.
- Cassures simple brin : l'IR dans ce cas provoque une rupture des liaisons phosphodiester au niveau d'un seul brin d'ADN.
- Cassures double brin : dans ce cas, soit un évènement ionisant affecte les deux brins d'ADN simultanément soit il s'agit de deux cassures simple brins indépendantes mais proches. Généralement, on considère qu'il s'agit d'une cassure double brin si les deux cassures simple brin ne sont pas éloignées de plus de 10 bases. Ces cassures sont les plus délétères pour la cellule et les plus difficiles à réparer.
- Formation de pontages entre macromolécules : l'IR peut provoquer des pontages intra- ou inter-brins sur l'ADN par fusion entre les deux bouts de brins d'ADN libres (suite à une cassure double brin ou à deux cassures simple brin). Ces constructions artificielles donnent lieu à des aberrations chromosomiques difficiles à réparer. Les radiations peuvent aussi provoquer des associations entre protéines et brins d'ADN [383]. L'ajout de protéines à l'ADN modifie sa structure et perturbe les fonctions qui y sont associées.
- Sites de dommages multiples : correspond à l'occurrence d'un grand nombre de lésions simples (combinaison des lésions vues précédemment) proches sur l'ADN [384, 385]. Ces associations de lésions sont extrêmement difficiles à réparer et sont le plus souvent létales pour la cellule.

F.3.4 Conséquences des lésions à l'ADN

Dès qu'une cellule va détecter des dommages à son ADN elle va arrêter sa progression dans le cycle cellulaire. Cet arrêt du cycle est crucial car la cellule doit absolument réparer les lésions avant d'amorcer sa mitose, sinon elle risque soit de propager des anomalies chromosomiques soit de rester bloquée en mitose sans pouvoir achever sa division. Différents cas de figures peuvent se présenter en fonction de la gravité des dommages à l'ADN :

- Si les lésions ne sont pas trop graves ni trop nombreuses, la cellule va mettre en place des mécanismes de réparation adaptés à chaque type de lésion. Si la cellule réussi à réparer fidèlement toutes les lésions, elle va pouvoir poursuivre son cycle normalement et entrer en mitose.
- Si jamais la cellule ne répare pas fidèlement une altération au niveau de bases nucléiques, cela peut générer une mutation au niveau de la partie codante ou régulatrice d'un gène. Cette mutation est alors transmise aux cellules filles après mitose. Si cette mutation n'est pas létale elle peut avoir de grandes conséquences sur la physiologie de la cellule. Dans le cas d'eucaryotes supérieurs comme l'Homme nous pouvons différencier deux cas, soit la mutation touche une cellule germinale et alors elle risque d'être transmise à la descendance. Soit cette mutation touche une cellule somatique et peut être la première cause d'un processus de cancérogénèse [302].

- Si les lésions à l'ADN sont trop importantes, la cellule ne pourra pas les réparer et mourra si la dose d'IR était très élevée.

F.3.5 Cancérogénèse et cancers radio-induits

Mécanismes généraux de la cancérogénèse. La cancérogénèse ou oncogenèse est un phénomène multi-étapes qui s'étend sur plusieurs années. Il correspond à l'accumulation d'évènements génétiques et épigénétiques dans une seule cellule jusqu'à sa conversion en cellule maligne. Chaque nouvel évènement apporte à la cellule qui en est le siège un avantage sélectif en terme de prolifération et/ou de survie par rapport à son environnement. Les évènements génétiques correspondent à une instabilité du matériel génétique souvent due à des agressions physiques ou chimiques de l'environnement (principalement cancérigènes chimique et radiations UV ou ionisantes). Néanmoins, un taux très faibles d'altérations du génome, dont certaines potentiellement oncogènes, est possible en l'absence de l'intervention de facteurs environnementaux qui augmentent la fréquence des lésions. L'évènement épigénétique le mieux caractérisé correspond à des mécanismes de méthylation/déméthylation de dinucléotides CpG. L'hyperméthylation de certaines séquences CpG serait associée à des évènements oncogéniques [386] et serait responsable de la perte d'expression de gènes suppresseurs de tumeurs [387]. L'oncogenèse correspond à une succession d'évènements de nature différente. Un seul évènement génétique affectant un gènes proto-oncogène ne va pas obligatoirement enclencher un processus oncogène. Différents facteurs peuvent limiter ou supprimer cette évolution pathologique. Tout d'abord, l'expression d'une forme mutée d'une protéine cellulaire risque d'alerter le système immunitaire et de conduire à l'élimination de la cellule exprimant cette protéine. D'autre part, la position d'une cellule par rapport à son programme de différenciation influence grandement la progression tumorale. Plus une cellule est en amont de son programme de différenciation plus un évènement génétique sera susceptible d'interrompre la différenciation et d'entraîner la cellule vers un processus d'oncogenèse. L'environnement direct de la cellule (matrice extracellulaire, environnement cellulaire) influence aussi grandement sa capacité de transformation par échange d'informations (facteurs de croissance, jonctions cellulaires) [388]. Par conséquent, seules certaines combinaisons d'évènements génétiques, construites dans un certain ordre, réussiront à détourner les différents obstacles à la transformation d'une cellule.

Différents modèles ont été proposés pour expliquer la progression tumorale, certains se basent sur deux étapes et d'autres en proposent plus de trois mais tous s'accordent sur le fait qu'un seul évènement génétique ou épigénétique ne suffit pas. La probabilité pour qu'une deuxième mutation apparaisse au sein d'une même cellule et entraîne cette dernière vers la voie oncogenique initiée par une première mutation est très faible. Or comment expliquer le taux de mutation très élevé dans une cellule ayant déja subi un premier évènement génétique ? Loab [389] a proposé qu'au cours de la progression tumorale s'installe une instabilité génétique qui augmente très fortement le taux de mutations. De plus, la découverte dans de nombreux cancers que des gènes de réparation des mésappariements étaient atteints a permis de construire un modèle expliquant l'apparition d'une instabilité génétique [390]. Ainsi, la probabilité qu'apparaisse un premier événement génétique est faible mais si cet évènement touche un gène proto-oncogène le génome devient moins

stable et la probabilité qu'un deuxième évènement apparaisse et se maintienne augmente avec l'instauration de cette instabilité. De plus, la prolifération cellulaire va grandement participer à l'instauration d'une instabilité génétique en augmentant la probabilité qu'une lésion dans l'ADN se maintienne et donne lieu à une mutation ou une aberration chromosomique.

Les RI ne vont pas provoquer une cancérogénèse par leurs seuls effets génotoxiques mais à cause aussi de leur pouvoir cytotoxique qui, en provoquant la mort de nombreuses cellules va stimuler la prolifération cellulaires au sein des tissus touchés.

Cancers Radio-induits. Il n'a pas été démontré de spécificité particulière ou de signature génétique marquante en ce qui concerne les cancers radio-induits à part l'existence d'un nombre élevé de multi-délétions. Si les causes de départ (altérations du génome) sont spécifiques des radiations ionisantes, les étapes intermédiaires (accumulation de mutations et évènements épigénétiques), suivent des voies communes à tous les processus d'oncogenèse. Les faits initiaux les mieux documentés correspondent à une action cancérigène par mutations géniques et anomalies chromosomiques augmentant l'instabilité génétique des cellules exposées et de leur descendance. Les évènements initiaux les plus connus correspondent globalement à trois types de phénomènes :

- la recombinaison de fragments chromosomiques pouvant provoquer l'activation de proto-oncogènes [391],
- la présence d'un grand nombre de délétions de plus d'une paire de bases [392],
- une grande fréquence de ruptures double brin qui n'apparaîtraient spontanément que très rarement. Ce type de lésion peut d'ailleurs permettre de quantifier la dose reçue par une cellule, sachant qu'un gray créé environ 30 cassures double brin par cellule [393]. Ces cassures double brin sont extrêmement difficiles à réparer car impliquant des mécanismes de réparation très complexes et dont les tentatives conduisent à des réparations fautives ou des recombinaisons non homologues.

En fonction de leur position dans le cycle cellulaire, les cellules n'auront pas la même sensibilité aux radiations. Les cellules en mitoses sont les plus sensibles à l'IR ce qui rend les tissus jeunes et à renouvellement rapide particulièrement fragiles et les plus susceptibles de développer des cancers. Dans ces tissus, les radiations créent un brusque accroissement du taux de mutations et stimulent un vieillissement accéléré et proportionnel à la dose reçue en augmentant la probabilité de survenue de mutations additionnelles en un temps plus court que le vieillissement naturel.

F.4 Ambivalence des effets des radiations ionisantes : effets délétères et applications thérapeutiques

F.4.1 Radiothérapie

Nous avons vu que les effets des radiations ionisantes sur les tissus vivants sont délétères à fortes doses et, on le sais maintenant, ne sont pas neutres à faibles doses. Or, il est un domaine où ce sont justement ces effets délétères que l'on va chercher à provoquer. En radiothérapie, on va sciemment chercher à altérer l'ADN de cellules cancéreuses afin

de provoquer leur mort. Très tôt, quelques années après la découverte des rayons X et du radium, les lésions cutanées provoquées par les radiations ont suggéré leur utilisation en biologie et en médecine. Dès 1896 les rayons X ont connus leurs premières applications en radiologie et en 1903, les rayonnements γ produits par le radium étaient utilisés pour le traitement de cancers. C'est dans les années 50 que la radiothérapie moderne a vraiment débuté avec l'utilisation de rayonnements à haute énergie ce qui a permis d'accroître le rendement en profondeur dans les tissus. Aujourd'hui, la radiothérapie externe est incontournable en cancérologie puisqu'elle est programmée dans deux tiers des schémas thérapeutiques, soit seule, soit associée à la chirurgie et/ou à la chimiothérapie. Contrairement à la chimiothérapie, la radiothérapie est un traitement localisé qui agit directement sur la zone du cancer. Il existe deux types de radiothérapies :

- La curiethérapie : la source radioactive, solide ou liquide, est placée à l'intérieur de l'organisme.
- La radiothérapie externe : la source d'IR est placée à l'extérieur du malade (appareils à rayons X, source de cobalt, accélérateurs).

On peut distinguer ensuite trois types d'applications thérapeutiques à la radiothérapie :

- la radiothérapie curative : le but est de stériliser définitivement toutes les cellules cancéreuses contenues dans le volume irradié afin d'éliminer définitivement une tumeur.
- la radiothérapie palliative : permet de freiner l'évolution de cancers évolués localement ou métastatiques dont on sait que l'on ne pourra pas les guérir.
- la radiothérapie symptomatique : utilisée pour soulager un symptôme majeur comme dans le cas de compressions ou de syndromes hémorragiques.

Le but de la radiothérapie est de détruire les cellules malignes tout en respectant les tissus environnants. L'IR provoquera l'arrêt de la progression du cycle et la mort des cellules dans le cas échéant. L'efficacité des radiations ionisantes repose essentiellement sur la capacité à induire l'apoptose des cellules cancéreuses. Les protocoles de radiothérapie sont définis principalement en fonction du type de tumeur, de sa localisation, de sa taille, de son extension et de son grade. Cependant, la dose totale ne suffit pas pour définir un traitement par IR, il faut également prendre en compte la dose par fraction, le nombre total de fractions (ou de séances) et le nombre de fractions par jour ou par semaine. Une radiothérapie classique délivre la dose totale par fractions de 2 Gy, une fraction par jour, 5 jours par semaine. Ce fractionnement de dose permet d'obtenir un meilleur ratio efficacité anti-tumorale/tolérance des tissus sains. Cette technique est fondée sur l'étalement et le fractionnement de la dose d'IR pour permettre entre chaque séance d'IR aux tissus sains traversés par le faisceau d'IR de se régénérer plus rapidement que la tumeur. En effet, il est établi, depuis très longtemps une relation de proportionnalité entre la radio-sensibilité des cellules cancéreuses et la vitesse de prolifération plus importante que les cellules saines et c'est principalement sur cette propriété que sont basés les protocoles de radiothérapie. Cependant, il existe différents niveaux de radio-sensibilité d'une lignée tumorale à l'autre, conditionnant l'efficacité de la radiothérapie. De plus, les radiations peuvent aussi affecter les cellules saines avoisinantes avec un risque important pour des cellules particulièrement radio-sensibles comme les cellules souches hématopoïétiques et les cellules souches de la lignée germinale. Et même si les différents protocoles de radiothérapie sont maintenant bien encadrés, on ne peut exclure les risques de cancers secondaires radio-induits suite à

une IR thérapeutique.

F.4.2 Phénomènes d'échappements à la radiothérapie

Certains cancers ont développé des mécanismes de résistance aux processus d'apoptose, soit par l'inactivation de certains gènes activant l'apoptose, soit par la sur-expression de protéines anti-apoptotiques [394, 395]. Ces cancers devenus radio-résistants, sont par la suite beaucoup plus difficiles à traiter car ils nécessitent de plus fortes doses d'IR, pouvant être nuisibles aussi pour les tissus sains en périphérie. L'acquisition de cette résistance est due principalement à l'instabilité génétique des cellules cancéreuses qui crée une hétérogénéité génétique au sein d'une tumeur. L'IR de cette tumeur provoquera la mort par apoptose de la majorité des génotypes cancéreux. Cependant, si une sous-population de cellules malignes avait déja acquis une forme de résistance elle sera la seule à survivre et à recoloniser l'espace laissé vacant. Ainsi, si la radiothérapie laisse s'échapper une sous-population de cellules tumorales elle risque de provoquer une sélection clonale de cellules malignes radio-résistantes.
Si les effets biologiques des RI sont très étudiés, peu d'informations sont disponibles quant aux mécanismes mis en oeuvre par la cellule en réponse aux radiations, en particulier en ce qui concerne l'induction de l'apoptose. Or, la compréhension des mécanismes apoptotiques induits par les différents types de radiations est essentielle pour l'amélioration des traitements existants et pour l'élaboration de nouvelles stratégies thérapeutiques visant à détruire les cellules malignes radio-résistantes.

F.5 Réponse cellulaire à l'irradiation

En plus des effets des rayonnements ionisants, les cellules peuvent être exposées à différents types de stress cyto- et génotoxiques. La surveillance de l'intégrité du génome est un processus physiologique vital pour les cellules. Si des constituants cellulaires autres que l'ADN sont endommagés, la cellule aura toujours la possibilité de les renouveler par la synthèse d'enzymes ou d'autres types de protéines or, si l'ADN, le support de l'information primaire est lui même touché, la cellule ne pourra plus rétablir une fonction perdue ou altérée et au mieux mourra ou au pire risquera de transmettre un message génétique erroné à sa descendance.
Afin de maintenir l'intégrité de leurs génomes les cellules ont donc mis en place, au cours de l'évolution, des mécanismes de surveillance et de réparation complexes. La réponse cellulaire à un stress génotoxique comme les RI sera initiée par la détection des dommages de l'ADN. Cette réponse cellulaire consistera essentiellement en l'induction de mécanismes de réparation complexes, adaptés à chaque type de lésion à l'ADN, afin de permettre une transmission fidèle du message génétique. Cette réponse sera aussi accompagnée d'un arrêt du cycle cellulaire, de la modulation de l'expression d'un grand nombre de gènes et parfois se terminera par l'entrée en apoptose et la mort de la cellule.

F.5.1 Détection des dommages radio-induits et signalisation

Les différentes formes de la réponse cellulaire aux dommages à l'ADN sont médiées par une cascade de protéines kinases qui semblent avoir été conservées à travers l'évolution

pour tous les eucaryotes. Au sommet de cette cascade on trouve une famille de phosphoinositol kinases qui incluent les protéines ATR et ATM chez les mammifères et Mec1p et Tel1p chez la levure. Ces complexes protéiques jouent un rôle central en tant que détecteurs des dommages à l'ADN [303, 304, 305] et en tant que régulateurs de la réparation ou de l'apoptose et de la survie cellulaire [55]. Ensuite, suivent deux classes de kinases dites de *checkpoint* : CHK1 et CHK2 chez les mammifères et Chk1p et Rad53p chez la levure. Chez la levure, Rad53p est suivie d'une kinase additionnelle, Dun1p. Cette dernière kinase est à la fois impliquée dans l'arrêt du cycle et dans la régulation transcriptionnelle de la réponse aux dommages à l'ADN [306, 307]. Chez les mammifères, CHK2 fait le lien entre les protéines de réparation et le contrôle du cycle cellulaire. Elle phosphoryle la protéine p53. Elle-même est la cible des kinases ATM et ATR. La phosphorylation de p53 en G1 permet sa dissociation de son inhibiteur mdm2 et ainsi de jouer son rôle d'activateur transcriptionnel sur les inhibiteurs du cycle et/ou de provoquer l'entrée de la cellule en apoptose. L'activation de la voie ATM → CHK2 provoquera un arrêt du cycle en G2 et l'activation de la voie ATM → CHK2 → p53 provoquera un arrêt en G1 ou l'apoptose [308, 309].
Des mutations au niveau des voies ATM/Mec1p entraînent une hypersensibilité aux agents génotoxiques et des prédispositions aux cancers dans les organismes supérieurs. Chez la levure ces mutations entraînent aussi une hypersensibilité aux génotoxiques et empêchent l'arrêt du cycle et la modulation de l'expression de gènes [396, 397]. Chez la levure les dommages à l'ADN activent des *checkpoints* au niveau de quatre positions du cycle cellulaire : la transition G1/S (*checkpoint* G1), pendant la phase S pour empêcher la réplication de l'ADN (*checkpoint* S), avant la mitose (S/M *checkpoint*) et G2/M (G2/M *checkpoint*). Chaque *checkpoint* activé doit être capable de reconnaître un type donné de lésion. Par exemple, les cassures double brin générées par les RI provoquent un arrêt G2/M avant l'entrée en mitose pour empêcher la perte de fragments de chromosomes [52, 53]. Les modifications de bases quand à elles inhibent la réplication et la progression dans la phase S [54]. Les systèmes de réparation sont directement intégrés au réseau cellulaire de régulation et de signalisation. Et, bien que différents groupes de protéines semblent être impliqués dans la détection des dommages, à différentes phases du cycle cellulaire, la transduction de tous ces signaux passe par la cascade de kinases et par les réactions de phosphorylations en chaîne.

F.5.2 Implication des organites et compartiments cellulaires dans la réponse aux dommages

Implication de la mitochondrie. La mitochondrie intervient dans la détoxication cellulaire suite à la création de composés radicalaires et de peroxydes par les RI. Ces composés peuvent aussi endommager la membrane mitochondriale, dans ce cas, des oxydations en chaîne peuvent se produire au niveau des mitochondries. Cela perturbe la perméabilité membranaire et stimule le relargage de Ca^{2+} [398] qui poursuit la cascade des dommages au niveau de la cellule [399]. Les dommages à l'ADN (créés directement par les RI ou issus des évènement oxydatifs initiés au niveau de la mitochondrie) peuvent déclencher des signaux pro-apoptotiques. La plupart de ces signaux vont converger vers la mitochondrie [400] comme par exemple p53 qui va se lier à Bcl2 et Bad [401]. Ces liaisons vont jouer sur la perméabilité de la membrane externe de la mitochondrie et relarguer d'autres fac-

teurs pro-apoptotiques [402]. La mitochondrie semble être ainsi un composant important dans la transduction du signal depuis la détection de dommages jusqu'à l'activation de molécules et de voies de signalisation médiant le programme de mort cellulaire.

Implication du nucléole. Le nucléole est un sous-compartiment nucléaire qui sert de site de synthèse pour les ARNr et de site d'assemblage des ribososmes. L'effet génotoxique des RI affecte la structure et le bon fonctionnement du nucléole. Des travaux ont suggéré que le nucléole pourrait agir comme un senseur et un site de stockage de protéines signal dans la transduction de la réponse aux dommages [403, 404]. De nombreuses protéines nucléolaires, dont la topoisomérase I et la nucléoline, sont en effet relâchées dans le nucléoplasme en réponse à des stimuli génotoxiques. Pour certaines protéines comme la nucléoline, et la Topo I [405], leur sortie du nucléole est dépendante de p53. La dé-localisation de la nucléoline hors du nucléole provoque sa liaison avec la protéine de réplication A et le blocage de l'initiation de la réplication de l'ADN, peut-être pour faciliter la réparation de l'ADN [405].

Implication des centrosomes. Les centrosomes sont les organites cellulaires qui initient la constitution du réseau de microtubules du cytosquelette à l'interphase et du fuseau mitotic. De nombreuses protéines cellulaires impliquées dans la régulation du cycle cellulaire ont été retrouvées au niveau des centrosomes, laissant suggérer un rôle des centrosomes dans la réponse aux dommages à l'ADN. Les centrosomes participeraient à la régulation de la réponse aux dommages à l'ADN via les protéines p53, BRCA1/BARD1, les kinases Chk1, Chk2 et d'autres régulateurs mitotiques [406, 407]. Les centrosomes sont inactivés lorsque le cycle cellulaire ne passe pas le *checkpoint* qui permet le passage à la mitose [408]. la kinase Chk2 va se localiser au niveau du nucléosome et inhibe la mitose et la formation du fuseau en réponse à des dommages à l'ADN. Cependant il n'a été identifié que relativement peu de protéines faisant la navette vers ou à partir du nucléole en réponse aux dommages à l'ADN. Par exemple, la protéine XRCC1, impliquée dans la réparation des cassures de l'ADN simple brin, est libérée du centrosome et interagit avec d'autres protéines impliquées dans la réparation du même type de lésions.

Implication du complexe golgien. L'appareil de Golgi est un complexe membranaire impliqué dans le stockage des protéines, leur transport et leur sécrétion. De nombreuses protéines pro-apoptotiques et anti-apoptotiques ont été retrouvées enrichies dans la membrane golgienne comme la caspase-2 qui relie le Golgi à la voie apoptotique [409]. La plupart des protéines jouant un rôle dans la réponse aux dommages à l'ADN et identifiées dans le Golgi tendent à se dé-localiser du Golgi vers les autres organites comme la mitochondrie ou le noyau.

Implication des membranes et du métabolisme des lipides. Il a été démontré que les rayonnements ultra violets (UV) activaient des récepteurs membranaires induisant la mort cellulaire, comme le récepteur CD95. La membrane cellulaire pourrait aussi agir comme un senseur de l'exposition à des stimuli cyto-toxiques. La membrane activerait les voies apoptotiques sans l'intervention d'un ligand sur ses récepteurs. Cette action impliquerait l'activation du facteur de transcription NFkB sans dépendre de la détection de

dommages à l'ADN [410]. Les lipides pourraient aussi directement, à travers le métabolisme des sphingolipides, avoir un rôle dans la régulation de la réponse cellulaire au stress cytotoxique. Il a été montré que de nombreux métabolites de type sphingolipides contribuaient à l'arrêt du cycle et à la régulation de la croissance cellulaire chez les mammifères et chez la levure. De plus, les céramides seraient aussi impliqués dans l'activation des voies apoptotiques en réponse aux stress radio-induits (voir état de l'art dans [411]).

F.5.3 Effet non ciblé de l'irradiation (*bystander effect*)

De nombreuses études ont montré que au sein d'une population, des cellules irradiées pouvaient transmettre des signaux de dommages à des cellules adjacentes non irradiées. Cette communication entre cellule pouvait induire chez les cellules non irradiées la même instabilité génétique que celle induite directement par l'IR. Ce phénomène a été décrit en tant que *bystander effect*, que l'on peut traduire par effet indirect [310]. Little [311] a compilé et analysé les différents et récents travaux ayant étudié ce phénomène et a proposé deux possibles mécanismes d'action pour expliquer cet effet indirect : le premier correspond à un contact direct, cellule à cellule, médié par une communication intercellulaire utilisant des jonctions de type gap. En effet, il a récemment été démontré que les jonctions gap était impliquées dans la transmission du signal de dommages radio-induits [412, 413] d'une cellule irradiée à une cellule non irradiée. De plus, des études ont montré que l'expression de la Connexin 43 est induite après irradiation [314, 315]. Le second mécanisme utilise quand à lui des facteurs solubles sécrétés dans le milieu extracellulaire par les cellules irradiées. Une étude récente a montré qu'un signal était transmis entre cellules irradiées et cellules non irradiées à travers des cultures ensemencées dans un même milieu mais à bonne distance les unes des autres [312]. Cette étude était particulièrement intéressante dans la mesure où l'effet indirect observé était associé à un stress oxydatif impliquant la production d'oxide nitrique. L'oxide nitrique est une molécule utilisée comme signal dans de nombreux processus comme les réactions d'inflammation par exemple.
De ces deux mécanismes de transmission du signal de dommages, la transmission via des jonctions gap semble être plus performante qu'une transmission via le milieu extracellulaire [313]. Cependant, aucune étude n'a encore pu identifier la nature exacte du signal transmis dans les deux types de mécanismes.

De plus en plus d'observations suggèrent un rôle important des différents organites cellulaires dans la détection du stress provoqué par les radiations ionisantes. Ces organites participeraient aussi à la détection des dommages, et à partir d'un certain seuil de dommages, déclencheraient des réponses locales et globales. Ces signaux pourraient être coordonnés par l'intermédiaire de la voie p53. Par exemple, des altérations au niveau du nucléole pourraient activer les voies mitochondriales qui ensuite activeraient p53 afin d'induire un arrêt du cycle ou l'apoptose. Ces échanges de signaux entre compartiments sub-cellulaires sont médiés par des translocations de protéines et sont aussi sujets à des régulations post-traductionnelles (principalement des phosphorylations) et transcriptionnelles (voir état de l'art dans [194]). Il reste cependant de nombreuses zones d'ombres dans la compréhension de la façon dont la réponse aux dommages à l'ADN est coordonnée. De

plus, il existe relativement peu d'informations sur la régulation de l'expression des gènes dépendants de Mec1p-Rad53p-Dun1p en réponse aux dommages à l'ADN. Chez l'homme, la transduction du signal, de la détection des dommages jusqu'à l'activation des voies de réparation passe essentiellement par la cascade de kinases et par des modifications post-traductionnelles mais implique aussi de nombreux facteurs de transcription vont intervenir en aval de la cascade de réactions de phosphorylation. Mais, la plupart de ces FT restent encore à identifier.

Bibliographie

[1] S. Oehler, E. R. Eismann, H. Krämer, and B. Müller-Hill. The three operators of the lac operon cooperate in repression. *EMBO J.*, 9(4) :973–979., 1990.

[2] G. K. Ackers, A. D. Johnson, and M. A. Shea. Quantitative model for gene regulation by λ phage repressor. *PNAS USA*, 79 :1129–1133, 1981.

[3] Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235) :467–470, 1995.

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

[5] P. Reichard. From rna to dna, why so many ribonucleotide reductases ? *Science*, 260 :1773–1777, 1993.

[6] A. Frey-Wyssling. *The Plant Cell Wall*. Gebruder Borntraeger, 1976.

[7] M. S. Bretscher. Membrane structure : Some general principles. *Science*, 181 :622–629, 1973.

[8] Reid E. and C. Christine Lomas-Francis. *The Blood Group Antigen Factsbook : Factsbook*. Elsevier Academic Press, 2003.

[9] R. O. Hynes. Cell adhesion : old and new questions. *Trends in Genetics.*, 15(12) :M33–M37, 1999.

[10] W. Birchmeier. Cytoskeleton structure and function. *Trends Biochem. Sci.*, 9 :192–195, 1984.

[11] D. Ingber, D. Prusty, Z. Sun, H. Betensky, and N. Wang. Cell shape, cytoskeletal mechanics, and cell cycle control in angiogenesis. *Journal of Biomechanics*, 28(12) :1471–1484, 1995.

[12] S. J. Kron and N. A. R. Gow. Budding yeast morphogenesis : signalling, cytoskeleton and cell cycle. *Current Opinion in Cell Biology.*, 7(6) :845–855, 1995.

[13] A. B. Pardee. Membrane transport proteins. *Science*, 162 :632–637, 1968.

[14] Kurjan J. The pheromone response pathway in saccharomyces cerevisiae. *Annual review of genetics*, 27 :147–179, 1993.

[15] M. Bölker and R. Kahmann. Sexual pheromones and mating responses in fungi. *Plant Cell*, 5 :1461–1469, 1993.

[16] L. Pedersen, M. Van Zeijl, S. V. Johann, and B. O'Hara. Fungal phosphate transporter serves as a receptor backbone for gibbon ape leukemia virus. *Journal of Virology*, 71(10) :7619–7622, 1997.

[17] K. Gulshan and W. S. Moye-Rowley. Multidrug resistance in fungi. *Eukaryotic Cell*, 6(11) :1933–1942, 2007.

[18] D. A. Lashkari and *al.* Yeast microarrays for genome wide parallel genetic and gene expression analysis. *PNAS USA*, 94 :13057–13062, 1997.

[19] A. Jones, M. Shyamsundar, M. Thomas, J. Maynard, S. Idziaszczyk, S. Tomkins, J. Sampson, and J. Cheadle. Comprehensive mutation analysis of tsc1 and tsc2-and phenotypic correlations in 150 families with tuberous sclerosis . *The American Journal of Human Genetics.*, 64(5) :1305–1315, 1999.

[20] N. Burns and *al.* Large-scale analysis of gene expression, protein localization, and gene disruption in saccharomyces cerevisiae. *Genes Dev.*, 8 :1087–1105, 1994.

[21] Winzeler E. A. *et al.* Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis. *Science*, 285(5429) :901–906, 1999.

[22] W. *et al.* Wurst. A large-scale gene-trap screen for insertional mutations in developmentally regulated genes in mice. *Genetics*, 139 :889–899, 1995.

[23] A. C. *et al.* Spradling. The berkeley drosophila genome project gene disruption project : Single p-element insertions mutating 25% of vital drosophila genes. *Genetics*, 153 :135–177, 1999.

[24] V. P. Sundaresan and *al.* Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes Dev.*, 9 :1797–1810, 1995.

[25] J. D. Watson and F. H. C. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171 :737, 1953.

[26] J. D. Watson and F. H. C. Crick. Genetical implication of the structure of desoxiribonucleic acid. *Nature*, 171 :964, 1953.

[27] A. Sentenac and B. Hall. *The molecular biology of the yeast Saccharomyces cerevisiae. Metabolism and gene expression.*, chapter Yeast nuclear RNA polymerases and their role in transcription, pages 561–606. Cold Spring Harbor Laboratory, N. Y., 1982.

[28] C. Mosrin and P. Thuriaux. The genetics of rna polymerases in yeasts. *Current Genetics*, 17 :367–373, 1990.

[29] J. Fickett and A. G. Hatzigeorgiou. Eukaryotic promoter recognition. *Genome Res.*, 7 :861–878, 1997.

[30] S. Hahn. Structure and mechanism of the rna polymerase ii transcription machinery. *Nature Structural & Molecular Biology*, 11 :394–403, 2004.

[31] R. A. Cox. Structure and function of prokaryotic and eukaryotic ribosomes. *Prog Biophys Mol Biol.*, 32(3) :193–231, 1977.

[32] I. G. Wool. The structure and function of eukaryotic ribosomes. *Annual Review of Biochemistry*, 48 :719–754, 1979.

[33] V. Ramakrishnan. Ribosome structure and the mechanism of translation. *Cell*, 108(4) :557–572, 2002.

[34] H. Temim and S. Mizutani. Viral rna-dependent dna polymerase : Rna-dependent dna polymerase in virions of rous sarcoma virus. *Nature*, 226 :1211–1213, 1970.

[35] D. Baltimore. Rna-dependent dna polymerase in virions of rna tumour viruse. *Nature.*, 226(5252) :1209–11, 1970.

[36] K. *et al.* Saigo. Identification of the coding sequence for a reverse transcriptase-like enzyme in a transposable genetic element in drosophila melanogaster. *Nature*, 312 :659–661, 1984.

[37] S. M. Mount and G. M. Rubin. Complete nucleotide sequence of the drosophila transposable element copia : homology between copia and retroviral proteins. *Mol Cell Biol.*, 5(7) :1630–1638, 1985.

[38] J. T. Kadonaga. Eukaryotic transcription : an interlaced network of transcription factors and chromatin-modifying machines. *Cell.*, 92(3) :307–13, 1998.

[39] G. Narlikar, H. Fan, and R. Kingston. Cooperation between complexes that regulate chromatin structure and transcription. *Cell*, 108(4) :475–487, 2002.

[40] D. S. Latchman. Eukaryotic transcription factors. *Biochem J.*, 270(2) :281–289, 1990.

[41] C. Lin. Combinatorial gene regulation by eukaryotic transcription factors. *Current Opinion in Structural Biology*, 9(1) :48–55, 1999.

[42] M Deutsch and M. Long. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Research*, 27(15) :3219–3228, 1999.

[43] S. M. Berget. Exon recognition in vertebrate splicing. *J. Biol. Chem*, 270(6) :2411–2414, 1995.

[44] D. M. Dykxhoorn and J. Lieberman. The silent revolution : Rna interference as basic biology, research tool, and therapeutic. *Annu. Rev. Med.*, 56 :401–423, 2005.

[45] M. B. Mathews. Lost in translation. *Trends in Biochemical Sciences*, 27(5) :267–269, 2002.

[46] W. C. Merrick. Mechanism and regulation of eukaryotic protein synthesis. *Microbiol Mol Biol Rev.*, 56(2) :291–315, 1992.

[47] M. Kozak. Mechanism of mrna recognition by eukaryotic ribosomes during initiation of protein synthesis. *Curr Top Microbiol Immunol.*, 93 :81–123, 1981.

[48] T. C. Cox, M. J. Bawden, A. Martin, and B. K. May. Human erythroid 5-aminolevulinate synthase : promoter analysis and identification of an iron-responsive element in the mrna. *EMBO J.*, 10(7) :1891–1902., 1991.

[49] M. Mann and J. Jensen. Proteomic analysis of post-translational modifications. *Nature Biotechnoloy*, 21 :251–261, 2003.

[50] J. F. Ward. Nature of lesions formed by ionizing radiation. *DNA Damage and Repair*, 1998.

[51] G. A. Nelson. Fundamental space radiobiology. *Gravitational and Space Biology Bulletin.*, 16(2), 2003.

[52] T. A. Weinert and L. H. Hartwell. The rad9 gene controls the cell cycle response to dna damage in *Saccharomyces cerevisiae*. *Science*, 241 :317–322, 1988.

[53] T. Weinert and L. Hartwell. Control of g2 delay by the rad9 gene of saccharomyces cerevisiae. *J. Cell Sci.*, Suppl. 12 :145–148, 1989.

[54] A. G. Paulovich and L. H. Hartwell. A checkpoint regulates the rate of progression through s phase in *S. cerevisiae* in response to dna damage. *Cell*, 82 :841–847, 1995.

[55] R. K. Schmidt-Ullrich, P. Dent, S. Grant, R. B. Mikkelsen, and K. Valerie. Signal transduction and cellular radiation responses. *Radiat. Res*, 153 :245–57, 2000.

[56] S. A. Jelinsky, P. Estep, G. M. Church, and L. D. Samson. Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells : Rpn4 links base excision repair with proteasomes. *Molecular and Cellular Biology*, 20(21) :8157–8167, 2000.

[57] V. De Sanctis, C. Bertozzi, G. Costanzo, E. Di Mauro, and R. Negri. Cell cycle arrest determines the intensity of the global transcriptional response of *Saccharomyces cerevisiae* to ionizing radiation. *Radiation Research*, 156(4) :379–387, 2001.

[58] G. W. Birrell, J. A. Brown, H. I. Wu, G. Giaever, A. M. Chu, R. W. Davis, and J. M. Brown. Transcriptional response of *Saccharomyces cerevisiae* to dna-damaging agents does not identify the genes that protect against these agents. *PNAS USA*, 99(13) :8778–8783, 2002.

[59] G. Mercier, N. Berthault, N. Touleimat, F. Kepes, G. Fourel, E. Gilson, and M. Dutreix. A haploid-specific transcriptional response to irradiation in saccharomyces cerevisiae. *Nucl. Acids Res.*, 33(20) :6635–6643, 2005.

[60] A. P. Gasch, M. Huang, S. Metzner, D. Botstein, S. J. Elledge, and P. O. Brown. Genomic expression responses to dna-damaging agents and the regulatory role of the yeast atr homolog mec1p. *Mol. Biol. Cell*, 12 :2987–3003, 2001.

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

[62] J. B. Hagen. The origins of bioinformatics. *Nature*, 1 :231–236, 2000.

[63] F. Sanger and E. O. Thompson. The amino-acid sequence in the glycyl chain of insulin. *Biochem. J.*, 52, 1952.

[64] J. C. Kendrew. Myoglobin and the structure of proteins : Crystallographic analysis and data-processing techniques reveal the molecular architecture. *Science*, 139(3561) :1259–1266, 1963.

[65] R. F. Doolittle, S. J. Singer, and H. Metzger. Evolution of immunoglobulin polypeptyde chains : carboxy-terminal of an igm heavy chain. *Science*, 154 :1561–1562, 1966.

[66] R. F. Doolittle. The evolution of vertebrate fibrinogen. *Fed. Proc.*, 35 :2145–2149., 1976.

[67] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.*, 48(3) :443–53, 1970.

[68] J. C. Venter and *al.* The sequence of the human genome. *Science*, 291(5507) :1304–1351, 2001.

[69] T. R. Gingeras and R. J. Roberts. Steps toward computer analysis of nucleotide sequences. *Science*, 209 :1322–1328, 1980.

[70] David B. Emmert, Peter J. Stoehr, Guenter Stoesser, and Graham N. Cameron. The european bioinformatics institute (ebi) databases. *Nucl. Acids Res.*, 22(17) :3445–3449, 1994.

[71] Guy *et al.* Cochrane. Priorities for nucleotide trace, sequence and annotation data capture at the ensembl trace archive and the embl nucleotide sequence database. *Nucl. Acids Res.*, page gkm1018, 2007.

[72] Howard S. Bilofsky, Christian Burks, James W. Fickett, Walter B. Goad, Frances I. Lewitter, Wayne P. Rindone, C.David Swindell, and Chang-Shung Tung. The genbank genetic sequence databank. *Nucl. Acids Res.*, 14(1) :1–4, 1986.

[73] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, Barbara A. Rapp, and David L. Wheeler. Genbank. *Nucl. Acids Res.*, 28(1) :15–18, 2000.

[74] Y. Tateno and T. Gojobori. Dna data bank of japan in the age of information biology. *Nucl. Acids Res.*, 25(1) :14–17, 1997.

[75] G. Stoesser, M. A. Tuli, R. Lopez, and P. Sterk. The embl nucleotide sequence database. *Nucl. Acids Res.*, 27(1) :18–24, 1999.

[76] David G. George, Winona C. Barker, and Lois T. Hunt. The protein identification resource (pir). *Nucl. Acids Res.*, 14(1) :11–15, 1986.

[77] Amos Bairoch and Brigitte Boeckmann. The swiss-prot protein sequence data bank, recent developments. *Nucl. Acids Res.*, 21(13) :3093–3096, 1993.

[78] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147 :195–197, 1981.

[79] D. G. Higgins and P. M. Sharp. Clustal : a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1) :237–44, 1988.

[80] K. Karplus, C. Barrett, and R. Hughey. Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14(10) :846–856, 1998.

[81] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res.*, 31 :3497–500, 2003.

[82] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.*, 27 :401–410, 1978.

[83] J. Kim and T. Warnow. Tutorial on phylogenetic tree estimation. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB99)*, 1999.

[84] W.-H. Li. *Molecular Evolution*. Sinauer Associates, Sunderland, M.A., 1997.

[85] J. Felsenstein. Confidence limits on phylogenies : An approach using the bootstrap. *Evolution*, 39 :783–791, 1985.

[86] D. M. Hillis and J. J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42(2) :182–192, 1993.

[87] B. Efron. Bootstrap methods : Another look at the jackknife. *Ann. Stat.*, 7 :1–26, 1979.

[88] A.R. Leach. *Molecular Modelling : Principles and Applications.* Pearson Education EMA, 2001.

[89] D. Frenkel and B. Smit. *Understanding Molecular Simulation : From Algorithms to Applications.* Academic Press, 2002.

[90] H. Kitano. Systems biology : A brief overview. *Science*, 295 :1662–1664, 2002.

[91] L von Bertalanffy. *General System Theory : Foundations, Development, Applications.* New York : George Braziller, 1968.

[92] G. Weinberg. *An Introduction to General Systems Thinking.* New York : Wiley-Interscience, 1975.

[93] R. Heinrich and S. Schuster. *The Regulation of Cellular Systems.* New York : Chapman & Hall, 1996.

[94] E.O. Voit. *Computational Analysis of Biochemical Systems.* Cambridge, UK : Cambridge University Press, 2000.

[95] W. Lin. *Applications de la technologie des Puces à ADN à l'étude de la différenciation méiotique et des mécanismes de recombinaison chez la levure Saccharomyces cerevisiae.* PhD thesis, UMR144 CNRS - Institut Curie, Section de Recherche, avril 2004.

[96] C. Coulouarn and *al.* Altered gene expression in acute systemic inflammation detected by complete coverage of the human liver transcriptome. *Hepatology*, 39(2) :353–364, 2004.

[97] G Petrovics and al. Frequent overexpression of ets-related gene-1 (erg1) in prostate cancer transcriptome. *Oncogene*, 24 :3847–3852, 2005.

[98] C. M. Costello, N. Mah, R. Hï¿½sler, Rosenstiel. P., G. H. Waetzig, and *et al.* Dissection of the inflammatory bowel disease transcriptome using genome-wide cdna microarrays. *PLoS Medicine*, 2(8) :e199, 2005.

[99] Marcia Saban, Helen Hellmich, Mary Turner, Ngoc-Bich Nguyen, Rajanikanth Vadigepalli, David Dyer, Robert Hurst, Michael Centola, and Ricardo Saban. The inflammatory and normal transcriptome of mouse bladder detrusor and mucosa. *BMC Physiology*, 6(1) :1, 2006.

[100] Matthew A. Ginos, Grier P. Page, Bryan S. Michalowicz, Ketan J. Patel, Sonja E. Volker, Stefan E. Pambuccian, Frank G. Ondrey, George L. Adams, and Patrick M. Gaffney. Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck. *Cancer Res*, 64(1) :55–63, 2004.

[101] M. Kittleson and J. Hare. Molecular signature analysis : Using the myocardial transcriptome as a biomarker in cardiovascular disease. *Trends in Cardiovascular Medicine*, 15(4) :130–138, 2005.

[102] V. E. Velculescu and *al.* Characterization of the yeast transcriptome. *Cell*, 88(2) :243–251, 1997.

[103] CS Richmond, JD Glasner, R Mau, H Jin, and FR Blattner. Genome-wide expression profiling in escherichia coli k-12. *Nucl. Acids Res.*, 27(19) :3821–3835, 1999.

[104] H. Caron, B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M.-C. Hermus, R. van Asperen, K. Boon, P. A. Voute, S. Heisterkamp, A. van Kampen, and R. Versteeg. The human transcriptome map : Clustering of highly expressed genes in chromosomal domains. *Science*, 291 :1289–1292, 2001.

[105] S. H. I. Kappe and *al.* Exploring the transcriptome of the malaria sporozoite stage. *PNAS USA*, 98(17) :9895–9900, 2001.

[106] E. Milohanic and *et al.* Transcriptome analysis of listeria monocytogenes identifies three groups of genes differently regulated by prfa. *Molecular Microbiology*, 47(6) :1613–1625, 2003.

[107] S. J. Jelinsky and L. D. Samson. Global response of *Saccharomyces cerevisiae* to an alkylating agent. *PNAS USA*, 96(4) :1486–1491, 1999.

[108] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M.l. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11(12) :4241–4257, 2000.

[109] J. E. Staunton, D. K. Donna K. Slonim, H. A. Coller, P. Tamayo, M. J. Angelo, J. Park, U. Scherf, J. K. Lee, W. O. Reinhold, J. N. Weinstein, J. P. Mesirov, E. S. Lander, and T. R. Golub. Chemosensitivity prediction by transcriptional profiling. *PNAS USA*, 98(19) :10787–10792, 2001.

[110] M. West and *et al.* Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS USA*, 98(20) :11462–11467, 2001.

[111] Wilbert HM Heijne, Anne S Kienhuis, Ben van Ommen, Rob H Stierum, and John P Groten. Systems toxicology : applications of toxicogenomics, transcriptomics, proteomics and metabolomics in toxicology. *Expert Review of Proteomics*, 2(5) :767–780, 2005.

[112] S. Ramaswamy and *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS USA*, 98(26) :15149–15154, 2001.

[113] C. Debouck and B. Metcalf. The impact of genomics on drug discovery. *Annual Review of Pharmacology and Toxicology*, 40 :193–208, 2000.

[114] E. Sausville and S Holbeck. Transcription profiling of gene expression in drug discovery and development : the nci experience. *European Journal of Cancer*, 40(17) :2544 – 2549, 2004.

[115] C. Freiberg, H. Brï¿½tz-Oesterhelt, and H. Labischinski. The impact of transcriptome and proteome analyses on antibiotic drug discovery. *Current Opinion in Microbiology*, 7(5) :451–459, 2004.

[116] G. Mercier, N. Berthault, J. Mary, A. Peyre, J. Antoniadis, J.-P. Comet, A. Cornuejols, C. Froidevaux, and M. Dutreix. Biological detection of low radiation doses by combining results of two microarray analysis methods. *Nucl. Acids Res.*, 32(1) :e12, 2004.

[117] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. D. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps : Methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A.*, 96(6) :2907–2912, 1999.

[118] A. P. Gasch and M. B. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3(11) :research0059.1–0059.22, 2002.

[119] J. R. Yates, A. Gilchrist, K. E. Howell, and J. J. Bergeron. Proteomics of organelles and large cellular structures. *Nature Rev. Mol. Cell Biol.*, 6 :702–714, 2005.

[120] B. Kuster, M. Schirle, P. Mallick, and R. Aebersold. Scoring proteomes with proteotypic peptide probes. *Nature Rev. Mol. Cell Biol.*, 6 :577–583, 2005.

[121] S. Sukhanov and P. Delafontaine. Protein chip-based microarray profiling of oxidized low density lipoprotein-treated cells. *Proteomics*, 5(5) :1274–1280, 2005.

[122] L. A. Kung and M. Snyder. Proteome chips for whole organism assays. *Nature Reviews*, 7 :617–622, 2006.

[123] C.-S. Chen, E. Korobkova, H. Chen, J. Zhu, X. Jian, S.-C. Tao, C. He, and H. Zhu. A proteome chip approach reveals new dna damage recognition activities in *Escherichia coli*. *Nature Methods*, 5 :69–74, 2008.

[124] A. Garcï¿½a, Y. A. Senis, R. Antrobus, C. E. Hughes, R. A. Dwek, S. P. Watson, and N. Zitzmann. A global proteomics approach identifies novel phosphorylated signaling proteins in gpvi-activated platelets : involvement of g6f, a novel platelet grb2-binding membrane adapter. *Proteomics*, 6(19) :5332–5343, 2006.

[125] M. D. Hoffmana, M. J. Sniatynskia, and J. Kast. Current approaches for global post-translational modification discovery and mass spectrometric analysis. *Analytica Chimica Acta*, page in press, 2008.

[126] N. J. *et al.* Krogan. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084) :637–643, 2006.

[127] A.-C. *et al.* Gavin. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440 :631–636, 2006.

[128] W.K. Huh, J.V. Falvo, L.C. Gerke, A.S. Carroll, R.W. Howson, J.S. Weissman, and E.K. O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425 :686–691, 2003.

[129] S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz. Systematic functional analysis of the yeast genome. *Trends Biotechnol.*, 16(10) :373–378, 1998.

[130] Anthony P. Burgard, Evgeni V. Nikolaev, Christophe H. Schilling, and Costas D. Maranas. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res.*, 14(2) :301–312, 2004.

[131] Markus Krummenacker, Suzanne Paley, Lukas Mueller, Thomas Yan, and Peter D. Karp. Querying and computing with biocyc databases. *Bioinformatics*, 21(16) :3454–3455, 2005.

[132] J. Hirabayashi, Y. Arata, and K. Kasai. Glycome project : concept, strategy and preliminary application to *Caenorhabditis elegans*. *Proteomics*, 1(2) :295–303, 2001.

[133] H. H. Freeze. Genetic defects in the human glycome. *Nat Rev Genet.*, 7(7) :537–51, 2006.

[134] Elena M. Comelli, Steven R. Head, Tim Gilmartin, Thomas Whisenant, Stuart M. Haslam, Simon J. North, Nyet-Kui Wong, Takashi Kudo, Hisashi Narimatsu, Jeffrey D. Esko, Kurt Drickamer, Anne Dell, and James C. Paulson. A focused microarray approach to functional glycomics : transcriptional regulation of the glycome. *Glycobiology*, 16(2) :117–131, 2006.

[135] X. Han and R. W. Gross. Global analyses of cellular lipidomes directly from crude extracts of biological samples by esi/ms : a bridge to lipidomics. *J. Lipid Res.*, 2003.

[136] D. F. Darvas, A. Guttman, and G. Dormï¿½n. *Chemical genomics*, chapter Defining the lipidome : a new target for therapeutics, page 215. CRC Press, 2004.

[137] P. Haimi, A. Uphoff, M. Hermansson, and P. Somerharju. Software tools for analysis of mass spectrometric lipidome data. *Anal. Chem.*, 2006.

[138] Laxman Yetukuri, Mikko Katajamaa, Gema Medina-Gomez, Tuulikki Seppanen-Laakso, Antonio Vidal-Puig, and Matej Oresic. Bioinformatics strategies for lipidomics analysis : characterization of obesity related hepatic steatosis. *BMC Systems Biology*, 1(1) :12, 2007.

[139] S. Schuster, D. A. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 18 :326–332, 2000.

[140] Fatma Kaplan, Joachim Kopka, Dale W. Haskell, Wei Zhao, K. Cameron Schiller, Nicole Gatzke, Dong Yul Sung, and Charles L. Guy. Exploring the temperature-stress metabolome of arabidopsis. *Plant Physiol.*, 136(4) :4159–4168, 2004.

[141] J. Munger, S. U. Bajad, H. A. Coller, T. Shenk, and J. D. Rabinowitz. Dynamics of the cellular metabolome during human cytomegalovirus infection. *PLoS Pathog*, 2(12) :e132, dć. 2006.

[142] Kay L. Ward, Ivan Tkac, Yuezhou Jing, Barbara Felt, John Beard, James Connor, Timothy Schallert, Michael K. Georgieff, and Raghavendra Rao. Gestational and lactational iron deficiency alters the developing striatal metabolome and associated behaviors in young rats. *J. Nutr.*, 137(4) :1043–1049, 2007.

[143] Lee J. Sweetlove, Robert L. Last, and Alisdair R. Fernie. Predictive metabolic engineering : A goal for systems biology. *Plant Physiol.*, 132(2) :420–425, 2003.

[144] L. Wu, W. A. van Winden, W. M. van Gulik, and J. J. Heijnen. Application of metabolome data in functional genomics : A conceptual strategy. *Metabolic Engineering*, 7(4) :302–310, 2005.

[145] Jorn Smedsgaard and Jens Nielsen. Metabolite profiling of fungi and yeast : from phenotype to metabolome by ms and informatics. *J. Exp. Bot.*, 56(410) :273–286, 2005.

[146] J. Schaub, C. Schiesling, M. Reuss, and M. Dauner. Integrated sampling procedure for metabolome analysis. *Biotechnology Progress*, 22(5) :1434–1442, 2006.

[147] H. C. Keun. Metabonomic modeling of drug toxicity. *Pharmacology & Therapeutics*, 109(1-2) :92–106, 2006.

[148] R. Kaddurah-Daouk, B. S. Kristal, and R. M. Weinshilboum. Metabolomics : a global biochemical approach to drug response and disease. *Annu. Rev. Pharmacol. Toxicol.*, 48 :653–83, 2008.

[149] J. B. German, M. A. Roberts, L. Fay, and S. M. Watkins. Metabolomics and individual metabolic assessment : The next great challenge for nutrition. *J. Nutr.*, 132(9) :2486–2487, 2002.

[150] M. J. Gibney, M. Walsh, L. Brennan, H. M. Roche, B. German, and B. van Ommen. Metabolomis in human nutrition : opportunities and challenges. *AM. J. Clin. Nutr.*, 82 :497–503, 2005.

[151] A. Kumar, S. Agarwal, J. A. Heyman, S. Matson, M. Heidtman, S. Piccirillo, L. Umansky, A. Drawid, R. Jansen, Y. Liu, K. H. Cheung, P. Miller, M. Gerstein, G. S. Roeder, and M. Snyder. Subcellular localization of the yeast proteome. *Genes Dev.*, 16 :707–719., 2002.

[152] Jeremy Simpson and Rainer Pepperkok. Localizing the proteome. *Genome Biology*, 4(12) :240, 2003.

[153] S. E. Mango. A green light to expression in time and space. *Nature Biotechnology*, 25 :645–646, 2007.

[154] Denis Dupuy, Qian-Ru Li, Bart Deplancke, Mike Boxem, Tong Hao, Philippe Lamesch, Reynaldo Sequerra, Stephanie Bosak, Lynn Doucette-Stamm, Ian A. Hope, David E. Hill, Albertha J.M. Walhout, and Marc Vidal. A first version of the caenorhabditis elegans promoterome. *Genome Res.*, 14(10b) :2169–2175, 2004.

[155] C. Guda and S. Subramaniam. Target : a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics*, 21(21) :3963–3969, 2005.

[156] D. Drew, S. Newstead, Y. Sonoda, H. Kim, G. von Heijne, and S Iwata. Gfp-based optimization scheme for the overexpression and purification of eukaryotic membrane proteins in *Saccharomyces cerevisiae*. *Nature Protocols*, 3 :784–798, 2008.

[157] V. Ambros, B. Bartel, D. P. Bartel, C. B. Burge, J. C. Carrington, X. Chen, G. Dreyfuss, S. R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun, and T. Tuschl. A uniform system for microrna annotation. *RNA*, 9(3) :277–9, 2003.

[158] D. Baulcombe. An rna microcosm. *Science*, 297 :2002–2003, 2002.

[159] R. C. Lee, R. L. Feinbaum, and V. Ambros. The *C. elegans* heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, 75 :843–854, 1993.

[160] A. Kirmizis and P J. Farnham. Genomic approaches that aid in the identification of transcription factor target genes. *Exp Biol Med*, 229 :705–21, 2004.

[161] C. T. Workman, H. C. Mak, S. McCuine, J.-B. Tagne, M. Agarwal, O. Ozier, T. J. Begley, L. D. Samson, and T. Ideker. A systems approach to mapping dna damage response pathways. *Science*, 312(5776) :1054–1059, 2006.

[162] M. Cusick, N. Klitgord, M. Vidal, and D. E. Hill. Interactome : gateway into systems biology. *Hum. Mol. Genet.*, 14 :R171–R181, 2005.

[163] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel. String : a database of predicted functional associations between proteins. *Nucl. Acids Res.*, 31(1) :258–261, 2003.

[164] A. Ben-Hur and W. S. Noble. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21(Suppl. 1) :i38–i46, 2005.

[165] Y Yamanishi, JP Vert, and M Kanehisa. Protein network inference from multiple genomic data : a supervised approach. *Bioinformatics*, 20 :i363–i370, 2004.

[166] Pierre Geurts, Nizar Touleimat, Marie Dutreix, and Florence d'Alché Buc. Inferring biological networks with output kernel trees. *BMC Bioinformatics*, 8(Suppl 2) :S4, 2007.

[167] G. Butland, J. M. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili. Interaction network containing conserved and essential protein complexes in escherichia coli. *Nature*, 433 :531–537, 2005.

[168] J. C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schechter, Y. Chemama, A. Labigne, and P. Legrain. The protein-protein interaction map of helicobacter pylori. *Nature*, 409 :211–215, 2001.

[169] D. J. LaCount, M. Vignali, R. Chettier, A. Phansalkar, R. Bell, J. R. Hesselberth, L. W. Schoenfeld, I. Ota, S. Sahasrabudhe, C. Kurschner, S. Fields, and R. E. Hughes. A protein interaction network of the malaria parasite plasmodium falciparum. *Nature*, 438 :103–107, 2005.

[170] T. Ito, K. Ota, H. Kubota, Y. Yamaguchi, T. Chiba, K. Sakuraba, and M. Yoshida. Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol. Cell Proteomics*, 1 :561–566, 2002.

[171] Etienne *et al.* Formstecher. Protein interaction mapping : A drosophila case study. *Genome Res.*, 15(3) :376–384, 2005.

[172] Siming *et al.* Li. A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657) :540–543, 2004.

[173] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksï¿½z, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. A human protein-protein interaction network : A resource for annotating the proteome. *Cell*, 122(6) :957–968, 2005.

[174] D. Scholtens, M. Vidal, and R. Gentleman. Local modeling of global interactome networks. *Bioinformatics*, 21 :3548–3557, 2005.

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

[176] The Gene Ontology Consortium. Gene ontology : tool for the unification of biology. *Nature Genetics*, 25 :25–29, 2000.

[177] J. Rey-Debove and A. Rey, editors. *Le Nouveau Petit Robert*. Dictionnaires Le Robert, Paris, Edition 2002.

[178] M Kanehisa and S Goto. Kegg : Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.*, 28(1) :27–30, 2000.

[179] B. R. Bochner. New technologies to assess genotype-phenotype relationships. *Nature Rev. Genet.*, 4 :309–314, 2003.

[180] C. Conrad, H. Erfle, P. Warnat, N. Daigle, T. Lorch, J. Ellenberg, R. Pepperkok, and R. Eils. Automatic identification of subcellular phenotypes on human cell arrays. *Genome Res.*, 14(6) :1130–1136, 2004.

[181] M. Bredel and E. Jacoby. Chemogenomics : an emerging strategy for rapid target and drug discovery. *Nature Rev. Genet.*, 5 :262–275, 2004.

[182] Dov Greenbaum, Nicholas M. Luscombe, Ronald Jansen, Jiang Qian, and Mark Gerstein. Interrelating different types of genomic data, from proteome to secretome : 'oming in on function. *Genome Res.*, 11(9) :1463–1468, 2001.

[183] Michael Shapira, Eran Segal, and David Botstein. Disruption of yeast forkhead-associated cell cycle transcription by oxidative stress. *Mol. Biol. Cell*, 15(12) :5659–5669, 2004.

[184] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F.H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksï¿½z, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. Ancestral antibiotic resistance in mycobacterium tuberculosis. *Cell.*, 122(6) :957–68, 2005.

[185] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 9(12) :3273–3297, 1998.

[186] J. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen, and M. Mann. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 127(3) :635–648, 2006.

[187] J. Ptaceka and M. Snyder. Charging it up : global analysis of protein phosphorylation. *Trends in Genetics*, 22(10) :545–554, 2006.

[188] M. Mann, S.-E. Ong, M. Grønborg, H. Steen, O. N. Jensen, and A. Pandey. Analysis of protein phosphorylation using mass spectrometry : deciphering the phosphoproteome. *Trends in Biotechnology*, 20(6) :261–268, 2002.

[189] R. Linding, J. Jensen, G. Ostheimer, M. van Vugt, C. Jørgensen, I. Miron, F. Diella, K. Colwill, L. Taylor, and K. Elder. Systematic discovery of in vivo phosphorylation networks. *Cell*, 7(29) :1415–1426, 2007.

[190] S. A. Gorski, M. Dundr, and T. Misteli. The road much traveled : traffick in the cell nucleus. *Current Opinion in Cell Biology*, 18(3) :284–290, 2006.

[191] CH Chen, D. P. von Kessler, W. Park, Y. Wang, B. Ma, and P. A. Beachy. Nuclear trafficking of cubitus interruptus in the transcriptional regulation of hedgehog target gene expression. *Cell*, 98(3) :305–16, 1999.

[192] A. G. Porter. Protein translocation in apoptosis. *Trends in Cell Biology*, 9(10) :394–401, 1999.

[193] T. Beck and M. N. Hall. The tor signalling pathway controls nuclear localization of nutrient-regulated transcription factors. *Nature*, 402 :689–692, 1999.

[194] V. Tembe and B. R. Henderson. Protein trafficking in response to dna damage. *Cellular Signalling*, 19 :1113–1120, 2007.

[195] Jennifer Lippincott-Schwartz and George H. Patterson. Development and use of fluorescent protein markers in living cells. *Science*, 300(5616) :87–91, 2003.

[196] Miroslav Dundr, Urs Hoffmann-Rohrer, Qiyue Hu, Ingrid Grummt, Lawrence I. Rothblum, Robert D. Phair, and Tom Misteli. A kinetic framework for a mammalian rna polymerase in vivo. *Science*, 298(5598) :1623–1626, 2002.

[197] Koret Hirschberg, Chad M. Miller, Jan Ellenberg, John F. Presley, Eric D. Siggia, Robert D. Phair, and Jennifer Lippincott-Schwartz. Kinetic analysis of secretory protein traffic and characterization of golgi to plasma membrane transport intermediates in living cells. *J. Cell Biol.*, 143(6) :1485–1503, 1998.

[198] U. Haupts, S. Maiti, P. Schwille, and W. W. Webb. Dynamics of fluorescence fluctuations in green fluorescent protein observed by fluorescence correlation spectroscopy. *PNAS USA*, 95(23) :13573–13578, 1998.

[199] G H Patterson, S M Knobel, W D Sharif, S R Kain, and D W Piston. Use of the green fluorescent protein and its mutants in quantitative fluorescence microscopy. *Biophys. J.*, 73(5) :2782–2790, 1997.

[200] R. Eils and C. Athale. Computational imaging in cell biology. *J Cell Biol.*, 161(3) :477–481, 2003.

[201] Chaitanya A. Athale, Morten O. Christensen, Roland Eils, Fritz Boege, and Christian Mielke. Inferring a system model of subcellular topoisomerase iiβ localization dynamics. *OMICS : A Journal of Integrative Biology*, 8(2) :167–175, 2004.

[202] Robert D. Phair, Paola Scaffidi, Cem Elbi, Jaromira Vecerova, Anup Dey, Keiko Ozato, David T. Brown, Gordon Hager, Michael Bustin, and Tom Misteli. Global nature of dynamic protein-chromatin interactions in vivo : Three-dimensional genome scanning and dynamic interaction networks of chromatin proteins. *Mol. Cell. Biol.*, 24(14) :6393–6402, 2004.

[203] R. D. Phair and T. Misteli. Kinetic modelling approaches to in vivo imaging. *Nat Rev Mol Cell Biol*, 2(12) :898–907, 2001.

[204] Gaelle Lelandais, Pierre Vincens, Anne Badel-Chagnon, Stephane Vialette, Claude Jacq, and Serge Hazout. Comparing gene expression networks in a multi-dimensional space to extract similarities and differences between organisms. *Bioinformatics*, 22(11) :1359–1366, 2006.

[205] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science.*, 303 :799–805, 2004.

[206] S. Ott, S. Imoto, and S. Miyano. Finding optimal models for small gene networks. *Pac Symp Biocomput.*, pages 557–67, 2004.

[207] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *J. Comput. Biol.*, 7(3-4) :601–20, 2000.

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

[209] S. A. Kauffman. *The Origins of Order : Self-Organization and Selection in Evolution,*. Oxford University Press, New York, 1993.

[210] R. Somogyi and C.A. Sniegoski. Modeling the complexity of genetic networks : Understanding multigenic and pleiotropic regulation. *Complexity*, 1(6) :45–63, 1996.

[211] S. Huang. Gene expression profiling, genetic networks, and cellular states : An integrating concept for tumorigenesis and drug discovery. *J. Mol. Med.*, 77 :469–480, 1999.

[212] S. Liang, S. Fuhrman, and R. Somogyi. Reveal : A general reverse engineering algorithm for inference of genetic network architectures. volume 3, pages 18–29, Singapore, 1998. World Scientific Publishing.

[213] S.A. Kauffman. The large-scale structure and dynamics of gene control circuits : An ensemble approach. *J. Theor. Biol.*, 44 :167–190, 1974.

[214] S.A. Kauffman. Antichaos and adaptation. *Sci. Am.*, 265 :78–84, 1991.

[215] G. Weisbuch. Networks of automata and biological organization. *J. Theor. Biol.*, 121 :255–267, 1986.

[216] R. Thomas. Regulatory networks seen as asynchronous automata : A logical description. *J. Theor. Biol.*, 153 :1–23, 1991.

[217] D. Thieffry and R. Thomas. Qualitative analysis of gene networks. *Pac Symp Biocomput*, 3 :77–88, 1998.

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

[219] R. Thomas and R. d'Ari. *Biological Feedback*. CRC Press, Boca Raton, FL., 1990.

[220] R. Thomas, A.-M. Gathoye, and L. Lambert. A complex control circuit : Regulation of immunity in temperate bacteriophages. *Eur. J. Biochem.*, 71 :211–227, 1976.

[221] L. Sánchez and D. Thieffry. A logical analysis of the drosophila gap genes. *J. Theor. Biol.*, 211 :115–141., 2001.

[222] L. Sánchez, J. van Helden, and D. Thieffry. Establishment of the dorso-ventral pattern during embryonic development of drosophila melanogaster : A logical analysis. *J. Theor. Biol.*, 189 :377–389, 1997.

[223] L. Mendoza, D. Thieffry, and E.R. Alvarez-Buylla. Genetic control of ower morphogenesis in arabidopsis thaliana : A logical analysis. *Bioinformatics*, 15(7-8) :593–606, 1999.

[224] K. Murphy and S. Mian. Modelling gene expression data using dynamic bayesian networks. Technical report, University of California, Berkeley, 1999.

[225] S. Kim, S. Imoto, and S. Miyano. Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. In *Proc. of CMSB*, pages 104–113, 2003.

[226] B.-E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d'Alché Buc. Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19(suppl2) :138–148, 2003.

[227] Z. Ghahramani. Learning dynamic bayesian networks. In *Summer School on Neural Networks.*, pages 168–197, 1997.

[228] M. De Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano. Inferring gene regulatory networks from time-ordered gene expression data of bacillus subtilis using differential equations. In *Proceedings of the Pacific Symposium on Biocomputing.*, pages 17–28, 2003.

[229] F. d'Alché Buc, P.-J Lahaye, B.-E. Perrin, L. Ralaivola, T. Vujasinovic, A. Mazurie, and S. Bottani. *Chapter in Bioinformatics Using Computational Intelligence Paradigms, Series : Studies in Fuzziness and Soft Computing, Vol.*, volume 176, chapter Dynamic model of gene regulatory networks based on inertia principle., pages 93–117. Springer, 2005.

[230] M. Quach, N. Brunel, and F. d'Alché Buc. Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference. *Bioinformatics*, 23(23) :3209–3216, 2007.

[231] T. Chen, H. L. He, and G. M. Church. Modeling gene expression with differential equations. *Pac Symp Biocomput*, 4 :29–40, 1999.

[232] P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Linear modeling of mrna expression levels during cns development and injury. *Pac Symp Biocomput*, pages 41–52, 1999.

[233] Y. Maki, D. Tominaga, M. Okamoto, S. Watanabe, and Y. Eguchi. Development of a system for the inference of large scale genetic networks. *Pac Symp Biocomput*, pages 446–458., 2001.

[234] M. Wahde and J. Hertz. Modeling genetic regulatory dynamics in neural development. *J. Comput. Biol.*, 8(4) :429–442, 2001.

[235] N MacDonald. *Biological Delay Systems : Linear Stability Theory.* Cambridge University Press, Cambridge., 1989.

[236] P. Smolen, D. A. Baxter, and J. H. Byrne. Modeling transcriptional control in gene networks : Methods, recent results, and future directions. *Bull. Math. Biol.*, 62 :247–292, 2000.

[237] T. R. Hvidsten, B. Wilczynski, A. Kryshtafovych, J. Tiuryn, J. Komorowski, and K. Fidelis. Discovering regulatory binding-site modules using rule-based learning. *Genome Res.*, 15(6) :856–866, 2005.

[238] Z. Pawlak. *Theory and decision library.*, system theory, knowledge engineering, and problem solving. Rough sets : Theoretical aspects of reasoning about data., page 229. D. Kluwer, Dordrecht, Boston., 1991.

[239] F. M. Brown. *Boolean reasoning : The logic of Boolean equations.* Kluwer Academic Publishers, Boston, 1990.

[240] J. Komorowski, A. Øhrn, and A. Skowron. *Handbook of data mining and knowledge discovery.*, chapter The ROSETTA rough set software system., pages 554–559. Oxford University Press, Oxford, 2002.

[241] T. R. Hvidsten, J. Komorowski, A. K. Sandvik, and Laegreid A. Predicting gene function from gene expressions and ontologies. In *PSB*, volume 06, pages 299–310, 2001.

[242] A. Clare, A. Karwath, H. Ougham, and R. D. King. Functional bioinformatics for arabidopsis thaliana. *Bioinformatics*, 22(9) :1130–1136, 2006.

[243] D. L. Brutlag, A. R. Galper, and D. H. Millis. Knowledge-based simulation of dna metabolism : Prediction of enzyme action. In *CABIOS*, volume 7, pages 9–19, 1991.

[244] A. R. Galper, D. L. Brutlag, and D. H. Millis. *Intelligence and Molecular Biology.*, chapter Knowledge-based simulation of DNA metabolism : Prediction of action and envisionment of pathways., pages 365–395. AAAI Press/MIT Press, Menlo Park, CA/Cambridge, MA., artificial edition, 1993.

[245] S. Fr¨ohler and S. Kramer. Inductive logic programming for gene regulation prediction. *Machine Learning*, 70 :225–240, 2008.

[246] S. Muggleton and L. De Raedt. Inductive logic programming : Theory and methods. *J Logic Prog*, 19(20) :629–679, 1994.

[247] H. Blockeel and L. D. Raedt. Top-down induction of first-order logical decision trees. *Artificial Intelligence*, 101(1-2) :285–297, 1998.

[248] I. M. Ong, S. E. Topper, D. Page, and V. S. Costa. Inferring regulatory networks from time series expression data and relational data via inductive logic programming. In S. Muggleton, R. Otero, and A. Tamaddoni-Nezhad, editors, *ILP 2006*, pages 366–378, Berlin Heidelberg, 2007. Springer-Verlag.

[249] T. N. Tran, K. Satou, and Y. B. Ho. Using inductive logic programming for predicting protein-protein interactions from multiple genomic data. In A. *et al.* Jorge, editor, *PKDD 2005*, pages 321–330, Berlin Heidelberg, 2005. Springer-Verlag.

[250] S. H. Muggleton. Machine learning for systems biology. In S. Kramer and B. Pfahringer, editors, *ILP 2005*, pages 416–423, Berlin Heidelberg, 2005. Springer-Verlag.

[251] C. Combe, V. Schächter, S. Matwin, and F. d'Alché Buc. Relational learning of biological networks. In *First FEBS Advanced Course on Systems Biology*, Gosau, Austria., 2005.

[252] James F. Allen and George Ferguson. Actions and events in interval temporal logic. Technical Report 521, University of Rochester, Computer Science Department, July 1994.

[253] L. De Raedt and K. Kersting. *Lecture Notes in Computer Science*, volume 3244/2004, chapter Probabilistic Inductive Logic Programming, pages 19–36. Springer Berlin, Heidelberg, 2004.

[254] M. Middendorf, A. Kundaje, C. Wiggins, Y. Freund, and C. Leslie. Predicting genetic regulatory repsponse using classification. *Bioinformatics*, 20(Suppl. 1) :i232–i240, 2004.

[255] Lev Soinov, Maria Krestyaninova, and Alvis Brazma. Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biology*, 4(1) :R6, 2003.

[256] T. Kato, K. Tsuda, and K. Asai. Selective integration of multiple biological data for supervised network inference. *Bioinformatics.*, 21(10) :2488–95, 2005.

[257] J. L. Riechmann, J. Heard, G. Martin, L. Reuber, Z. C. Jiang, J. Keddie, L. Adam, O. Pineda, O. J. Ratcliffe, R. R. Samaha, R. Creelman, M. Pilgrim, P. Broun,

J. Z. Zhang, D. Ghandehari, B. K. Sherman, and G. L. Yu. Arabidopsis transcription factors : Genome-wide comparative analysis among eukaryotes. *Science*, 290(5499) :2105–2110, 2000.

[258] P. Jorgensen, I. Rupes, J. R. Sharom, L. Schneper, J. R. Broach, and M. Tyers. A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size. *Genes Dev.*, 18(20) :2491–2505, 2004.

[259] Berend Snel, Peer Borkb, and Martijn A. Huynen. Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends in Biotechnology*, 20(10) :410, 2002.

[260] S. Bergmann, J. Ihmels, and N. Barkai. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biology*, 2(1), 2003.

[261] Vera van Noort, Berend Snel, and Martijn A. Huynen. Predicting gene function by conserved co-expression. *Trends in Genetics*, 19(5) :238–242, 2003.

[262] Berend Snel, Vera van Noort, and Martijn A. Huynen. Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. *Nucl. Acids Res.*, 32(16) :4725–4731, 2004.

[263] Joshua M. Stuart, Eran Segal, Daphne Koller, and Stuart K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643) :249–255, 2003.

[264] Nabil Guelzim, Samuele Bottani, Paul Bourgine, and François Képès. Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, 31 :60–63, 2002.

[265] A. Arkin and J. Ross. Statistical construction of chemical reaction mechanisms from measured time-series. *J. Phys. Chem.*, 99 :970–979, 1995.

[266] A. Arkin, P. Shen, and J. Ross. A test case of correlation metric construction of a reaction pathway from measurements. *Science*, 277 :1275–1279, 1997.

[267] T. Chen, V. Filkov, and S. S. Skiena. Identifying gene regulatory networks from experimental data. In ACM Press, editor, *In Proc. 3rd Ann. Int. Conf. Comp. Mol. Biol. (RECOMB'99)*, pages 94–103, New York, 1999.

[268] F. C. Holstege, E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. A. Young. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell.*, 95(5) :717–28, 1998.

[269] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, , and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1) :109–126., 2000.

[270] M. T. Laub, H. H. McAdams, T. Feldblyum, C. M. Fraser, and L. Shapiro. Global analysis of the genetic network controlling a bacterial cell cycle. *Science*, 290 :2144–2148, 2000.

[271] A. Zien, R. Kï¿½ffner, R. Zimmer, and T. Lengauer. Analysis of gene expression data with pathway scores. In AAAI Press, editor, *Proc. 8th Int. Conf. Intell. Syst. Mol. Biol. (ISMB 2000)*, pages 407–417, Menlo Park, 2000.

[272] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, , and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22 :281–285, 1999.

[273] J. Qian, M. Dolled-Filhart, J. Lin, H. Yu, and M. Gerstein. Beyond synexpression relationships : local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J. Mol. Biol.*, 314 :1053–1066, 2001.

[274] P. Grosu, J. P. Townsend, D. L. Hartl, and D. Cavalieri. Pathway processor : A tool for integrating whole-genome expression results into metabolic networks. *Genome Res.*, 12 :1121–1126, 2002.

[275] R. Pandey, R. K. Guru, and D. W. Mount. Pathway miner : extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics*, 20 :2156–2158, 2004.

[276] C. Pasquier, K. Girardot, F. Jevardat de Fombelle, and R. Christen. Thea : ontology-driven analysis of microarray data. *Bioinformatics*, 20 :2636–2643, 2004.

[277] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape : A software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11) :2498–2504, 2003.

[278] P. Vailaya, A.and Bluvas, R. Kincaid, A. Kuchinsky, M. Creech, and A. Adler. An architecture for biological information extraction and representation. *Bioinformatics*, 21(4) :430–438, 2005.

[279] C. Huang, F. Morcos, S. P. Kanaan, S. Wuchty, D. Z. Chen, and J. A. Izaguirre. Predicting protein-protein interactions from protein domains using a set cover approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(1) :78–87, 2007.

[280] R. Breitling, S. Ritchie, D. Goodenowe, M. L. Stewart, and M. P. Barrett. Ab initio prediction of metabolic networks using fourier transform mass spectrometry data. *Metabolomics*, 2(3) :155–164, 2006.

[281] P. H. Lee and L. Doheon. Modularized learning of genetic interaction networks from biological annotations and mrna expression data. *Bioinformatics*, 21(11) :2739–2747, 2005.

[282] J. Liu, W. Wang, and J. Yang. Gene ontology friendly biclustering of expression profiles. In *In Computational Systems Bioinformatics*, pages 436–447, 2004.

[283] Lizhuang Zhao and Mohammed J. Zaki. Tricluster : an effective algorithm for mining coherent clusters in 3d microarray data. In *SIGMOD '05 : Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 694–705, New York, NY, USA, 2005. ACM.

[284] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *PNAS USA*, 101(9) :2981–2986, 2004.

[285] D. J. Reiss, N. S. Baliga, and R. Bonneau. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, 7(280), 2006.

[286] J. Kasturi and R. Acharya. Clustering of diverse genomic data using information fusion. *Bioinformatics*, 21(4) :423–429, 2004.

[287] D. Steinhauser, B. H. Junker, A. Luedemann, J. Selbig, and J. Kopka. Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics*, 20(12) :1928–1939, 2004.

[288] J. J. Lars and K. Steen. Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics*, 16(4) :326–333, 2000.

[289] Y. Ben-Shaul, H. Bergman, and H. Soreq. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics*, 21(7) :1129–1137, 2005.

[290] N. Simonis, S. J. Wodak, G. N. Cohen, and J. van Helden. Combining pattern discovery and discriminant analysis to predict gene co-regulation. *Bioinformatics*, 20(15) :2370–2379, 2004.

[291] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol.*, 21(11) :1337–42, 2003.

[292] P. M. Haverty, U. Hansen, and Z. Weng. Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucl. Acids Res.*, 32(1) :179–188, 2004.

[293] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks : identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34 :166–176, 2003.

[294] Chad Myers, Drew Robson, Adam Wible, Matthew Hibbs, Camelia Chiriac, Chandra Theesfeld, Kara Dolinski, and Olga Troyanskaya. Discovery of biological networks from diverse functional genomic data. *Genome Biology*, 6(13) :R114, 2005.

[295] Mohamed Elati, Pierre Neuvial, Monique Bolotin-Fukuhara, Emmanuel Barillot, Francois Radvanyi, and Celine Rouveirol. Licorn : learning cooperative regulation networks from gene expression data. *Bioinformatics*, 23(18) :2407–2414, 2007.

[296] B. E. Lehnert. Exposure to low-level chemicals and ionizing radiation : reactive oxygen species and cellular pathways. *Human & Experimental Toxicology*, 21(2) :65–69, 2002.

[297] D. R. Spitz, E. I. Azzam, J. J. Li, and D. Gius. *Metabolic oxidation/reduction reactions and cellular responses to ionizing radiation : A unifying concept in stress response biology.* 2004.

[298] A. Martinez and R. Kolter. Protection of dna during oxidative stress by the nonspecific dna- binding protein dps. *J. Bacteriol.*, 179(16) :5188–5194, 1997.

[299] M. V. Prasad, J. M. Dermott, L. E. Heasley, G. L. Johnson, and N. Dhanasekaran. Activation of jun kinase/stress-activated protein kinase by gtpase-deficient mutants of g alpha 12 and g alpha 13. *J. Biol. Chem.*, 270(31) :18655–9, 1995.

[300] M. Benderitter. Structural radio-induced membrane modifications : A potential bio-indicator of ionising radiation exposure. *Int. J. of Rad. Biol.*, 75 :1043, 1999.

[301] A. Haimovitz-Friedman, C. C. Kan, D. Ehleiter, M. Persaud, R. S. McLoughlin, Z. Fuks, and R. N. Kolesnick. Ionizing radiation acts on cellular membranes to generate ceramide and initiate apoptosis. *Journal of Experimental Medicine*, 180 :525–535, 1994.

[302] J. B. Little. Radiation carcinogenesis. *Carcinogenesis*, 21(3) :397–404, 2000.

[303] T.A. Weinert, G.L. Kiser, and L.H. Hartwell. Mitotic check-point genes in budding yeast and the dependence of mitosis on dna replication and repair. *Genes Dev.*, 8 :652–665, 1994.

[304] K. Savitsky, S. Sfez, D. A. Tagle, Y. Ziv, A. Sartiel, F. S. Collins, Y. Shiloh, and G. Rotman. The complete sequence of the coding region of the atm gene reveals similarity to cell cycle regulators in different species. *Hum. Mol. Genet.*, 4 :2025–2032, 1995.

[305] K. A. Cimprich, T. B. Shin, C. T. Keith, and S. L. Schreiber. cdna cloning and gene mapping of a candidate human cell cycle checkpoint protein. *PNAS USA*, 93 :2850–2855, 1996.

[306] Z. Zhou and S.J. Elledge. Dun encodes a protein kinase that controls the dna damage response in yeast. *Cell*, 75 :1119–1127, 1993.

[307] D. Pati, C. Keller, M. Groudine, and S.E. Plon. Reconstitution of a mec1-independent checkpoint in yeast by expression of a novel human fork head cdna. *Mol. Cell. Biol.*, 17 :3037–3046, 1997.

[308] T. Caspari. How to activate p53. *Curr. Biol.*, 10 :R315–7, 2000.

[309] B. B. S. Zhou, P. Chaturvedi, K. Spring, S. P. Scott, R. A. Johanson, and R et al. Mishra. Caffeine abolishes the mammalian g2/m dna damage checkpoint by inhibiting ataxia-telangectasia-mutated kinase activity. *J. Biol. Chem.*, 275 :10342–8, 2000.

[310] W. F. Morgan. Non-targeted and delayed effects of exposure to ionizing radiation : Radiation-induced genomic instability and bystander effects in vitro. *Radiat. Res.*, 159 :567–580, 2003.

[311] J. B. Little. Cellular radiation effects and the bystander response. *Mutation Research*, 597 :113–118, 2006.

[312] C. Shao, M. Folkard, B. D. Michael, and K. M. Prise. Targeted cytoplasmic irradiation induces bystander responses. *PNAS USA*, 101 :13495–13500, 2004.

[313] S. A. Mitchell, G. Randers-Pehrson, D. J. Brenner, and E. J. Hall. The bystander response in C3H 10T1/2 cells : The influence of cell-to-cell contact. *Radiat. Res.*, 161 :397–401, 2004.

[314] E. I. Azzam, S. M. de Toledo, and J. B. Little. Expression of connexin43 is highly sensitive to ionizing radiation and other environmental stresses. *Cancer Res.*, 63 :7128–7135, 2003.

[315] D. Glover, J. B. Little, M. F. Lavin, and N. Gueven. Low dose ionizing radiation-induced activation of connexin 43 expression. *Int. J. Radiat. Biol.*, 79 :955–964, 2003.

[316] A. Watson, J. Mata, J. Bï¿½hler, A. Carr, and T. Humphrey. Global gene expression responses of fission yeast to ionizing radiation. *Mol. Biol. Cell*, 15(2) :851–860, 2004.

[317] M. Molin, J. P. Renault, G. Lagniel, S. Pin, M. Toledano, and J. Labarre. Ionizing radiation induces a yap1-dependent peroxide stress response in yeast. *Free Radic. Biol Med.*, 43(1) :136–44, 2007.

[318] S. Maere, K. Heymans, and M. Kuiper. Bingo : a cytoscape plugin to assess over-representation of gene ontology categories in biological networks. *Bioinformatics*, 21(16) :3448–3449, 2005.

[319] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS USA*, 95 :14863–14868, 1998.

[320] C. I. Castillo-Davis and D. L. Hartl. Genemerge : post-genomic analysis, data mining and hypothesis testing. *Bioinformatics*, 19(7) :891–892, 2003.

[321] F. Zehraoui and F. d'Alché Buc. Multi-spectral biclustering for data described by multiple similarities. In *International Workshop on Machine Learning in Systems Biology*, Brussels, 2008.

[322] W. Lee, R. P. St.Onge, M. Proctor, P. Flaherty, M. I. Jordan, A. P. Arkin, R. W. Davis, C. Nislow, and G. Giaever. Genome-wide requirements for resistance to functionally distinct dna-damaging agents. *PLoS Genet*, 1(2) :e24, aoÃ»t 2005.

[323] E. Denti, A. Omicini, and A. Ricci. tuprolog : A light-weight prolog for internet applications and infrastructures. In *Practical Aspects of Declarative Languages, 3rd International Symposium (PADL 2001)*, Las Vegas, NV, USA., March 2001. Springer-Verlag.

[324] M. J. L. de Hoon, S. Imoto, and S. Miyano. Statistical analysis of a small set of time-ordered gene expression data using linear splines. *Bioinformatics*, 18(11) :1477–1485, 2002.

[325] Z. Bar-Joseph, G. Gerber, T. Jaakola, D. Gifford, and I. Simon. Continuous representations of time series gene expression data. *J. Comput. Biol.*, 3-4 :341–356, 2003.

[326] Y. Luan and H. Li. Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics*, 19(4) :474–482, 2003.

[327] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University, England, 2004.

[328] H. Shimodaira, K.-I. Noma, M. Nakai, and S. Sagayama. Dynamic time-alignement kernel in support vector machine. In *Advances in NIPS*, volume 14. MIT Press., 2002.

[329] C. Bahlman, B. Haasdonk, and H. Burkhardt. *Frontiers in Handwriting Recognition*, chapter On-line handwriting recognition with support vector machines : A kernel approach., pages 49–54. 2002.

[330] J.-P. Vert, S. Hiroto, and A. Tatsuya. *Kernel Methods in Computational Biology.*, chapter Local alignment kernels for protein sequences. MIT Press., 2004.

[331] C. Cortes, P. Haffner, and M. Mohri. Rational kernels : Theory and algorithms. *JMLR*, 5 :1035–1062, 2004.

[332] M. Cuturi, J.-P. Vert, O. Birkenes, and Matsui T. A kernel for time series based on global alignments. *arXiv.org*, arXiv :cs/0610033v1, 2006.

[333] H. Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci.*, IV :801–804, 1956.

[334] S. P. Lloyd. Last square quantization in pcm. *IEEE Trans. Inform. Theory*, 28 :129–137, 1982.

[335] P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference : from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8) :707–726, 2000.

[336] Richard Clarkson, Matthew Wayland, Jennifer Lee, Tom Freeman, and Christine Watson. Gene expression profiling of mammary gland development reveals putative roles for death receptors and immune mediators in post-lactational regression. *Breast Cancer Res*, 6(2) :R92–R109, 2004. See related Research article : http ://breast-cancer-research.com/content/6/2/R75 and related Commentary : http ://breast-cancer-research.com/content/6/2/89.

[337] Thomas Eulgem, Victor J. Weigman, Hur-Song Chang, John M. McDowell, Eric B. Holub, Jane Glazebrook, Tong Zhu, and Jeffery L. Dangl. Gene Expression Signatures from Three Genetically Separable Resistance Gene Signaling Pathways for Downy Mildew Resistance. *Plant Physiol.*, 135(2) :1129–1144, 2004.

[338] F.-X. Wu, W. J. Zhang, and A. J. Kusalik. A genetic k-means clustering algorithm applied to gene expression data. *Lecture in Artificial Intelligence.*, 2671 :520–526., 2003.

[339] T. Kohonen. *Self-Organization and Associative Memory.* Spring-Verlag, Berlin, 1984.

[340] J. *et al.* Nikkila. Analysis and visualization of gene expression data using self-organizing maps. *Neural Networks*, 15 :953–966, 2002.

[341] H. Resson, D. Wang, and P. Natarajan. Clustering gene expression data using adaptive double self-organizing map. *Physiol. Genomics*, 14 :35–46, 2003.

[342] B. Fritzke. Fast learning with incremental rbf networks. *Neural Processing Letters*, 1(1) :1–5, 1994.

[343] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data : An Introduction to Cluster Analysis.* Wiley, New York, 1990.

[344] *et al.* Wen, X. Large-scale temporal gene expression mapping of central nervous system development. *PNAS USA*, 95 :34–339., 1998.

[345] V.R. *et al.* Iyer. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283 :83–87, 1999.

[346] J. *et al.* Qin. Kernel hierarchical gene clustering from microarray expression data. *Bioinformatics*, 19 :2097–2104, 2003.

[347] E. Hartuv, A. O. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrach, and R. Shamir. An algorithm for clustering cdna fingerprints. *Genomics*, 66(3) :249–256, 2000.

[348] R. Sharan and R. Shamir. Click : a clustering algorithm with applications to gene expression analysis. In *ISMB'00*, volume 8, pages 307–316, 2000.

[349] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *J. Comput. Biol.*, 6(3-4) :281–297, 1999.

[350] W.E. Donath and A.J. Hoffman. Algorithms for partitioning graphs and computer logic based on eigenvectors of connection matrices. *IBM Technical Disclosure Bulletin*, 15(3) :938–944, 1972.

[351] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 22(8) :888–905, 2000.

[352] U. von Luxburg, O. Bousquet, and M. Belkin. Limits of spectral clustering. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS) 17*, pages 857 – 864. MIT Press, Cambridge, MA, 2005.

[353] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering : Analysis and an algorithm. In *NIPS*, 2001.

[354] J. Hartigan. Direct clustering of a data matrix. *J. Amer. Statis. Assoc.*, 6 :123–129, 1972.

[355] Y. Cheng and G. M. H. Church. Biclustering of expresssion data. In *ISMB*, 2000.

[356] Amela Prelic, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Buhlmann, Wilhelm Gruissem, Lars Hennig, Lothar Thiele, and Eckart Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9) :1122–1129, 2006.

[357] R. Tibshirani, T. Hastie, M. Eisen, D. Ross, D. Botstein, and P. Brown. Clustering methods for the analysis of dna microarray data. Technical report, Stanford University, October 1999.

[358] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *J. Royal Statis Soc*, 2001.

[359] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica Sinica*, 12 :61–86, 2002.

[360] G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. *PNAS USA*, 97 :12079–12084, 2000.

[361] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini. Discovering local structure in gene expression data : The order-preserving submatrix problem. In ACM Press., editor, *Proceedings RECOMB'02*, pages 49–57, 2002.

[362] E. Segal, B. Taskar, A. P. Gasch, N. Friedman, and D. Koller. Rich probabilistic models for gene expression. *Bioinformatics*, 17 :S243–S252, 2001.

[363] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(S1) :S136–S144, 2002.

[364] J. Fridlyand and S. Dudoit. Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method. Technical Report 600, University of California, Berkeley, 2001.

[365] R. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3 :1–27, 1974.

[366] W. Krzanowski and Y. Lai. A criterion for determining the number of groups in a datast using sum of squares clustering. *Biometrics*, 44 :23–34, 1985.

[367] J. A. Hartigan. Statistical theory in clustering. *Journal of Classification*, 2 :63–76, 1985.

[368] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50 :159–179, 1985.

[369] E. Levine and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13 :2573–2593, 2001.

[370] M. Bittner and *al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(3), 2000.

[371] S. P. Smith and R. Dubes. Stability of a hierarchical clustering. *Pattern Recognition*, 12 :177–187, 1980.

[372] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383) :553–584, 1983.

[373] S. Dudoit and J. Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.*, 3(7), 2002.

[374] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *PSB*, volume 7, pages 6–17, 2002.

[375] F. D. Gibbons and F. P. Roth. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.*, 12 :1574–1581, 2002.

[376] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G. C. Tseng. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22 :2405–2412, 2006.

[377] S. Raychaudhuri and R. B. Altman. A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics*, 19(3) :396–401, 2003.

[378] L. M. Jakt, L. Cao, K. S. E. Cheah, and D. K. Smith. Assessing clusters and motifs from gene expression data. *Genome Res.*, 11 :112–123, 2001.

[379] C. J. Shellaberger, E. P. Cronkite, V. P. Bond, and S. W. Lippincott. The occurrence of mammary tumors in the rat after sublethal whole-body irradiation. *Radiat. Res.*, 6 :501–512, 1957.

[380] E. P. Cronkite, W. Moloney, and V. P. Bond. Radiation leukemogenesis : an analysis of the problem. *Am J Med*, 28 :673–682, 1960.

[381] W. L. Hughes, V. P. Bond, and G.et al. Brecher. Cellular proliferation in the mouse as revealed by autoradiography with tritiated thymidine. *PNAS USA*, 44 :476–483, 1958.

[382] P. Costantini, B. V. Chernyak, V. Petronilli, and P. Bernardi. Modulation of the mitochondrial permeability transition pore by pyridine nucleotides and dithiol oxidation at two separate sites. *J. Biol. Chem.*, 271(12) :6746–6751, 1996.

[383] T. J. Jenner, S. M. Cunniffe, D. L. Stevens, and P. O'Neill. Induction of dna-protein crosslinks in chinese hamster v79-4 cells exposed to high- and low-linear energy transfer radiation. *Radiat Res.*, 150(5) :593–9, 1998.

[384] D. T. Goodhead. Initial events in the cellular effects of ionizing radiations : clustered damage in dna. *Int J Radiat Biol.*, 65(1) :7–17, 1994.

[385] B. M. Sutherland, P. V. Bennett, E. Weinert, O. Sidorkina, and J. Laval. Frequencies and relative levels of clustered damages in dna exposed to gamma rays in radioquenching vs. nonradioquenching conditions. *Environmental and Molecular Mutagenesis*, 38(2-3) :159–165, 2001.

[386] A. M. De Marzo, V. L. Marchi, E. S. Yang, R. Veeraswamy, X. Lin, and W. G. Nelson. Abnormal regulation of dna methyltransferase expression during colorectal carcinogenesis. *Cancer Research*, 59 :3855–3860, 1999.

[387] P. A. Jones and P. W. Laird. Cancer epigenetics comes of age. *Nature Genetics*, 21 :163–167, 1999.

[388] H. Yamasaki, Y. Omori, M. L. Zaidan-Dagli, N. Mironov, M. Mesnil, and V. Krutovskikh. Genetic and epigenetic changes of intercellular communication genes during multistage carcinogenesis. *Cancer Detect Prev.*, 23(4) :273–9, 1999.

[389] L. A. Loeb. Mutator phenotype may be required for multistage carcinogenesis. *Cancer Res.*, 51(12) :3075–3079, 1991.

[390] K. W. Kinzler and B. Vogelstein. Lessons from hereditary colorectal cancer. *Cell*, 87 :159–170, 1996.

[391] S. Klugbauer, E. Lengfelder, E. P. Demidchik, and H. M. Rabes. High prevalence of ret rearangement in thyroid tumors of children from belarus after the chernobyl reactor accident. *Oncogene*, 11(12) :2459–2467, 1995.

[392] S. Chevillard and M. N. Guilly. Caractérisation cytogénétique et moléculaire des tumeurs radio-induites : Les effets des rayonnements ionisants. *Revue générale nucléaire.*, (3) :28–32, 2004.

[393] M. M. Vilenchik and A. G. Knudson. Endogenous dna double-strand breaks : Production, fidelity of repair, and induction of cancer. *PNAS USA*, 100(22) :12871–12876, 2003.

[394] G. I. Evan and K. H. Vousden. Proliferation, cell cycle and apoptosis in cancer. *Nature*, 411 :342, 2001.

[395] J. C. Reed. Bcl-2 family proteins. *Oncogene*, 17(25) :3225–3236, 1998.

[396] B. A. Desany, A. A. Alcasabas, J. B. Bachant, , and S. J. Elledge. Recovery from dna replicational stress is the essential function of the s-phase checkpoint pathway. *Genes Dev.*, 12 :2956–2970., 1998.

[397] V. I. Bashkirov, J. S. King, E. V. Bashkirova, J. Schmuckli-Maurer, and W. D. Heyer. Dna repair protein rad55 is a terminal substrate of the dna damage checkpoints. *Mol. Cell. Biol.*, 20 :4393–4404, 2000.

[398] C. Richter, V. Gogvadze, R. Laffranchi, M. Schlapbach Sweitzer, M. Suter, P. Walter, and M. Yaffee. Oxidants in mitochondria : from physiology to diseases. *Biochimica et Biophysica Acta*, 1271 :67–74, 1995.

[399] T. E. Gunter, K. K. Gunter, S.-S. Sheu, and C.E. Gavin. Mitochondrial calcium transport : physiological and pathological relevance. *Am. J. Physiol*, 267 :C313–C339, 1994.

[400] X. Saelens, N. Festjens, L. Vande Walle, M. van Gurp, G. van Loo, and P. Vandenabeele. Toxic proteins released from mitochondria in cell death. *Oncogene*, 23 :2861, 2004.

[401] P. Jiang, W. Du, K. Heese, and M. Wu. The bad guy cooperates with good cop p53 : Bad is transcriptionally up-regulated by p53 and forms a bad/p53 complex at the mitochondria to induce apoptosis. *Mol. Cell. Biol.*, 26 :9071, 2006.

[402] T. J. Preston and G. Singh. *Interorganellar Signaling in Age-Related Disease*. Elsevier Science, 2001.

[403] O. V. Zatsepina, L. N. Voronkova, V. N. Sakharov, and Y. S. Chentsov. Ultrastructural changes in nucleoli and fibrillar centers under the effect of local ultraviolet microbeam irradiation of interphase culture cells. *Exp. Cell Res.*, 181 :94, 1989.

[404] C. P. Rubbi and J. Milner. Disruption of the nucleolus mediates stabilization of p53 in response to dna damage and other stresses. *EMBO J.*, 22 :6068, 2003.

[405] Y. Daniely, D. D. Dimitrova, and J. A. Borowiec. Stress-dependent nucleolin mobilization mediated by p53-nucleolin complex formation. *Mol. Cell. Biol.*, 22 :6014–6022, 2002.

[406] L. Fletcher and R. J. Muschel. The centrosome and the dna damage induced checkpoint. *Cancer Lett.*, 243 :1–8, 2006.

[407] H. Loffler, J. Lukas, J. Bartek, and A. Kramer. Structure meets function : centrosomes, genome maintenance and the dna damage response. *Exp. Cell Res.*, 312(14) :2633–40, 2006.

[408] O. C. Sibon, A. Kelkar, W. Lemstra, and W. E. Theurkauf. Dna-replication/dna-damage-dependent centrosome inactivation in drosophila embryos. *Nature Cell Biol.*, 2 :90–95, 2000.

[409] M. Mancini, C. E. Nahamer, S. Roy, D. W. Nicholson, N. A. Thornberry, L. Casciola-Rosen, and Rosen A. Caspase-2 is localized at the golgi complex and cleaves golgin-160 during apoptosis. *J. Cell Biol.*, 149 :603, 2000.

[410] D. Kulms, B. Poppelmann, D. Yarosh, T. A. Luger, J. Krutmann, and T. Schwarz. Nuclear and cell membrane effects contribute independently to the induction of apoptosis in human cells exposed to uvb radiation. *PNAS USA*, 96 :7974–7979, 1999.

[411] K. E. Rizzieri and Y. A.. Hannun. Sphingolipid metabolism, apoptosis and resistance to cytotoxic agents : can we interfere ? *Drug Resist. Updat.*, 6 :359–76, 1998.

[412] C. Shao, Y. Furusawa, M. Aoki, and K. Ando. Role of gap junctional intercellular communication in radiation-induced bystander effects in human fibroblasts. *Radiat. Res.*, 160 :318–323, 2003.

[413] G. O. Edwards, S. W. Botchway, G. Hirst, C. W. Wharton, J. K. Chipman, and R. A. Meldrum. Gap junction communication dynamics and bystander effects from ultrasoft x-rays. *Br. J. Cancer*, 90 :1450–1456, 2004.