Exploration bioinformatique des relations entre mécanismes moléculaires et fonctions cellulaires

Claire Gaugain

Université Victor Segalen Bordeaux 2

Année 2007 Thèse n°1461

THESE

pour le

DOCTORAT DE L'UNIVERSITE BORDEAUX 2

Mention : Sciences Biologiques et Médicales

Option : Biologie-Santé

Présentée et soutenue publiquement

le 18 Décembre 2007

par **Claire GAUGAIN**

né(e) le 15 Novembre 1980 à Le Mans

Exploration bioinformatique des relations entre mécanismes moléculaires et fonctions cellulaires

Membres du jury

Mr. Alain BLANCHARD.....Président du Jury
Mme Marie-Dominique DEVIGNES.....Rapporteur de la Thèse
Mr. Jean-Loup RISLER.....Rapporteur de la Thèse
Mme Isabelle DUTOUR.....Examinateur
Mr. Jean-Marc SCHWARTZ.....Examinateur
Mr. Antoine DE DARUVAR.....Directeur de Thèse

Cette thèse a été préparée

au *Centre de Bioinformatique de Bordeaux*
Université Victor Ségalen Bordeaux 2
Plateforme Génomique Fonctionnelle
146 rue Léo Saignat
33076 Bordeaux Cedex

"Exploration bioinformatique des relations entre mécanismes moléculaires et fonctions cellulaires"

Résumé

L'intégration des données biologiques est un des principaux défis de la bioinformatique aujourd'hui. La mise à disposition de quantités importantes de données concernant tous les niveaux d'organisation de la cellule, nécessite la mise en place de stratégies d'intégration pour rassembler toutes ces données, et ainsi mieux comprendre le fonctionnement de la cellule. Nous nous sommes intéressés à l'exploitation du concept de voisinage pour représenter et intégrer des données biologiques. Dans un premier temps, notre travail met l'accent sur l'importance du choix de la représentation pour mener une intégration efficace. Notre étude sur la représentation du métabolisme a montré que les modes élémentaires sont une alternative pertinente à la représentation classique sous forme de voies métaboliques. De plus, les modes élémentaires nous ont permis de trouver des routes métaboliques utilisées par la cellule en réponse à divers stress. Nous avons également exploité le voisinage dans une perspective de génomique comparative. Nous avons cherché à déterminer si le voisinage d'expression peut être une signature pour les gènes, et s'il peut être utilisé pour caractériser des gènes en établissant des équivalences entre des génomes (orthologues ou gènes fonctionnellement similaires). Les résultats présentés confirment l'intérêt de l'exploration du voisinage, des gènes et de leur produit, pour intégrer des données hétérogènes. L'efficacité de cette exploration est fortement liée au choix de la représentation des connaissances.

Mots-clés: intégration de données; représentation des connaissances; transcriptomique; métabolisme; voisinage ; génomique comparative.

"Bioinformatic exploration of relationships between molecular mecanisms and cellular functions"

Abstract

Biological data integration is one of the major challenge in bioinformatics today. The availability of amounts of data concerning all the level of cell organisation, requires strategies of integration to bring together these data and thus better understand how the cell works. We have focused our work on the use of the concept of neighbourhood in order to represent and integrate data. First, our work emphasizes the importance of the choice of data representation for an efficient integration. Our study on metabolism representation shows that elementary modes are a relevant alternative to the classical representation of metabolism as metabolic pathways. Moreover, elementary modes have enabled us to find metabolic routes used by the cell in response to stressed. We have also used the neighbourhood in a new angle, the one of comparative genomics. We tested if expression neighbourhood of genes (set of genes with close expression profiles) can be a signature for genes, and if it can be used to define functional similarities between genes from different organisms. The work presented here, shows the interest of the exploration of gene and protein neighbourhood in order to integrate heterogeneous data. The efficiency of this exploration is highly related to the choice of knowledge representation.

Keywords: data integration; knowledge representation; transcriptomics; metabolism; neighbourhood; comparative genomics.

Remerciements

Mes remerciements s'adressent tout d'abord à mon directeur de thèse *Antoine de Daruvar* pour m'avoir accueillie dans son laboratoire, et pour son aide et son implication permanentes dans mes recherches malgré ses innombrables activités !

Un grand merci à mes collègues du Centre de Bioinformatique de Bordeaux pour leur soutien dans mes différents travaux: Nicolas Goffard et Alexis Groppi pour les travaux menés ensemble, pour nos discussions scientifiques et aussi celles plus futiles; Daniel Jacob, Roland Barriot et Aurélien Barré pour leur soutien technique tout au long de ma thèse; Monique Nazabal pour sa gentillesse et son soutien "administratif" sans lequel la paperasse n'aurait pas été si facile.
Merci aussi aux autres membres du CBiB avec lesquels j'ai créé des relations amicales et avec qui j'ai passé de bons moments au laboratoire et à l'extérieur: Alexandre, Nacer, Hélène, Cyril, Laurent; et tous les autres pour nos conversations de tous les jours et leurs conseils (Elisabeth, Pascal, Sandrine, Patricia), ainsi qu'à tous mes collègues du LaBRI.
Je tiens à remercier mes voisins de laboratoire, "la protéomique" ou "les humides", pour leur bonne humeur et leur sympathie. Je veux évidemment remercier plus particulièrement Delphine pour son amitié et pour être devenue rapidement *"ma copine de Bordeaux"*.

Je remercie Jean-Loup Risler et Marie-Dominique Devignes pour avoir accepté d'être les rapporteurs de ma thèse, et tous les autres membres du jury, Alain Blanchard, Isabelle Dutour et Jean-Marc Schwartz pour leur présence et leurs conseils à différents moments de ma thèse.

Merci aux amis des différentes associations de doctorants de France et de Navarre (Marc, Marion, Charles, Cathel, Sandrine, Aurélie, les Rémi, Claire, Christelle...), avec qui j'ai pu échanger, partager sur cette expérience qu'est la thèse et surtout, sur pleins d'autres sujets!

Merci à tous ceux qui me soutiennent depuis des années et qui ont toujours été présents,

Papa et Maman,
Mes sœurs,
Mes grands-parents et toute ma famille,
Mes amis du Mans et d'Angleterre,

A mon Loulou.

SOMMAIRE

LISTE DES FIGURES

LISTE DES TABLEAUX

GLOSSAIRE

Ce glossaire regroupe les termes importants et spécifiques utilisés dans ce manuscrit.

Entité biologique: type de molécules telles que l'ADN (les gènes), les ARNs, les protéines ou les métabolites.

Critère biologique: on utilise ce terme, dans cette thèse, pour désigner une propriété pouvant être associée aux entités biologiques. Cette propriété peut prendre un certain nombre de valeurs (numériques ou qualitatives) qui constituent une "annotation" pour les entités biologiques.

Groupe: ensemble d'entités biologiques (la plupart du temps des gènes ou des protéines) qui partagent une valeur similaire pour un critère biologique donné.

Collection: ensemble de groupes qui représente un critère biologique.

Voisinage: concept, introduit par Antoine Danchin en 1998, qui s'intéresse aux relations qui existent entre des entités biologiques plutôt qu'aux entités individuellement.

Clustering: ce terme réfère de façon générale à des méthodes basées sur une mesure de similarité entre entités biologiques, et qui permettent de regrouper, de classifier ces entités.

Mode élémentaire: ensemble minimum de réactions indispensables qui peut fonctionner à l'état stationnaire dans la cellule. Cet ensemble de réactions correspond à un chemin, ou à un des chemins possibles, pour aller d'un métabolite A à un métabolite B, en se basant sur la topologie du réseau étudié.

Route métabolique: ensemble de modes élémentaires connectés qui traversent plusieurs cartes/voies métaboliques.

ABREVIATIONS

Dans l'ordre alphabétique

BNN: 'Best Neighbours Nested'

BNS: 'Best Neighbours Single'

DTT: Dithiothreitol

EST: Expressed Sequence Tag

GO: Gene Ontology

KEGG: Kyoto Encyclopedia of Genes and Genomes

LAS: Sodium lauryl sulfate

MGI: Mouse Genome Informatics

MIAME: Minimum Information About Microarray Experiments

MIPS: Munich Information center for Protein Sequences

PCP: Pentachlorophenol

pI: point isoélectrique

SAGE: Serial Analysis of Gene Expression

SDS: Sodium n-dodecyl benzosulfonate

SRS: Sequence Retrieval System

TPN: Tetrachloro-isophthalonitrile

INTRODUCTION

1. Les données biologiques à l'échelle moléculaire et cellulaire: Génomique, Post-génomique et Bioinformatique

L'automatisation du séquençage au début des années 90 a considérablement augmenté la quantité de séquences et surtout de génomes entiers disponibles pour les chercheurs. Le séquençage du génome humain, et de nombreux autres organismes à un rythme de plus en plus rapide, a changé les perspectives de la recherche en biologie. Il y a encore 20 ou 30 ans, la démarche en biologie moléculaire ou médecine consistait à caractériser un gène impliqué dans une fonction biologique ou responsable d'une maladie. Aujourd'hui, ayant à notre disposition de nombreuses séquences, les chercheurs ont développé la démarche inverse: on dispose de l'ensemble des gènes de nombreux organismes et on s'intéresse à définir la fonction, le rôle de chacun de ces gènes. C'est ainsi qu'à la fin des années 80, on entre dans l'ère de la génomique, qui peut être décrite comme la discipline consistant à faire l'inventaire de l'ensemble des gènes d'un organisme afin d'en déterminer leurs fonctions. Dans ce contexte, le génome ne donne qu'une vision « figée » d'un organisme, et cela à deux points de vue :

- lorsqu'un gène est décrit comme une séquence nucléotidique, il apparaît comme un objet statique alors qu'il est dans une « dynamique » puisqu'il peut être différentiellement exprimé, régulé dans le temps ou dans différentes conditions ;

- lorsqu'un gène est étudié en dehors de son contexte, il apparaît comme un objet isolé alors qu'il agit en interaction avec d'autres *entités biologiques*[1] et participe à des réseaux complexes (interactions, régulations...).

Il faut être capable de connaître l'expression d'un gène et de comprendre les relations complexes qu'entretiennent les entités biologiques entre elles pour pouvoir découvrir la fonction des gènes et comprendre le fonctionnement de la cellule.

Dans ce but, et afin de compléter l'approche génomique, on entre alors dans l'ère de la post-génomique, démarche qui doit permettre d'avoir une vision plus dynamique de ce qui se passe dans la cellule. On peut définir la post-génomique ou génomique

[1] Les entités biologiques réfèrent aux molécules telles que l'ADN (les gènes), les ARNs, les protéines ou les métabolites.

fonctionnelle comme un domaine qui regroupe de nombreuses approches permettant l'étude des différentes entités biologiques en présence dans la cellule. Pour cela, il faut bien plus que la séquence nucléotidique d'un gène ou d'un génome pour connaître les différents mécanismes et interactions auxquels les entités biologiques participent. Ainsi, dans les années 90, de nouvelles technologies à grande échelle apparaissent. Elles permettent d'étudier des milliers de gènes ou de protéines simultanément et d'engranger toutes sortes d'informations sur ces entités et sur les mécanismes mis en place dans la cellule. Ces approches ont conduit à produire toujours plus de données, leur vitesse de développement pouvant se comparer à celle du séquençage. Ces technologies à grande échelle, largement exploitées aujourd'hui, sont nombreuses et correspondent à l'étude d'un ensemble particulier d'entités biologiques :

- le transcriptome correspond à l'ensemble des ARNm transcrits à un instant *t* dans une cellule. Ce sont des technologies telles que les microarrays (puces à ADN), la technique SAGE (Serial Analysis of Gene Expression) ou le séquençage d'ESTs (Expressed Sequence Tag) qui permettent de connaître le niveau d'expression de l'ensemble des gènes d'un organisme à un moment donné et dans une condition donnée (selon le tissu étudié, selon l'état de la cellule...) ;
- le protéome décrit l'ensemble des protéines effectivement traduites dans la cellule à un moment donné. On utilise l'électrophorèse bidimensionnelle sur gel, la spectrométrie de masse pour identifier et quantifier les protéines présentes dans un échantillon biologique et des approches telles que le système double-hybride pour déterminer celles qui sont en interaction ;
- le métabolome représente l'ensemble des métabolites, molécules intermédiaires ou produits du métabolisme, correspondant à l'empreinte spécifique des processus cellulaires qui ont eu lieu dans la cellule. La chromatographie couplée à la spectrométrie de masse et à la RMN (Résonance Magnétique Nucléaire) permet également d'identifier et quantifier tous les métabolites présents dans un contexte biologique défini.

Contrairement aux séquences qui donnent une image identique du génome à tout moment et dans chaque cellule d'un organisme, le transcriptome, le protéome et le métabolome varient dans le temps, et d'une cellule à une autre. Ils permettent d'avoir une vision plus dynamique de ce qui peut se passer dans une cellule.

Toutes ces approches et toutes les autres existantes (lipidomique, cytomique, etc) permettent de caractériser l'état moléculaire des cellules, à différents moments et dans

différentes conditions. Elles génèrent une grande quantité de données, appelées les données « omiques », qui nécessitent le développement d'outils informatiques appropriés, permettant le stockage et l'exploitation efficace de toutes ces informations. La bio-informatique est donc une discipline incontournable de la post-génomique et a subi une forte expansion ces dernières années. Elle est indispensable pour être capable d'extraire les informations biologiques pertinentes parmi la masse de données dorénavant disponibles. La bioinformatique telle que Minoru Kanehisa la définit en 2000 (*post genome informatics*) vise à une synthèse des connaissances biologiques depuis les informations génomiques jusqu'à la compréhension des principes de base de la vie, tout ceci avec l'aide de l'outil informatique (Kanehisa, 2000).

Cette multitude de données biologiques, issues des différentes approches de la génomique et de la post-génomique, est caractérisée par une forte hétérogénéité.

2. Hétérogénéité des données

Les nombreuses approches expérimentales développées ces dernières années pour caractériser les différents niveaux d'organisation de la cellule (génome, transcriptome, protéome, métabolome, etc) produisent des masses de données considérables et très hétérogènes. Cette hétérogénéité existe à différents niveaux :

- hétérogénéité dans le type de données ;
- hétérogénéité dans leur format et leur façon d'être stockées.

2.1 Les types de données

Il est difficile de discuter de l'ensemble des données existantes à ce jour et donc prétendre pouvoir le faire de façon exhaustive. Dans leur livre (Wooley and Lin, 2005), John Wooley et Herbert Lin ont divisé toutes ces données en une dizaine de catégories correspondant à différents types de données et illustrant parfaitement la variété et la complexité de celles-ci :

- les séquences nucléiques ou protéiques ;
- les informations géométriques qui concernent la structure tridimensionnelle des protéines et de l'ADN ;

- les motifs qui peuvent être exprimés comme des morceaux de séquences ou des expressions régulières que l'on peut trouver plus ou moins fréquemment dans les séquences et qui ont un intérêt biologique. On en trouve également dans les structures des protéines ou dans les réseaux métaboliques ;
- les données qui donnent des informations quantitatives sur les gènes (par exemple, leur niveau d'expression), les protéines et tout autre entité biologique ;
- la littérature scientifique qui contient des informations exploitées pour trouver des relations entre des entités, des mécanismes biologiques. Elle est également la base de l'annotation des entités biologiques ;
- les images qui sont issues d'observation au microscope, des dessins ou vidéos utilisés pour simplifier ou représenter des phénomènes ou systèmes complexes ;
- les graphes qui permettent de représenter les cartes génétiques, les voies métaboliques, les réseaux de régulation de gènes et autres relations entre des entités biologiques ;
- les modèles qui constituent une première intégration de toutes ces informations afin de simuler un mécanisme biologique ou le fonctionnement d'une cellule.

On peut noter parmi ces différents types de données que certains correspondent à des données brutes issues d'une expérience en laboratoire; d'autres résultent déjà de l'exploitation de ces données brutes pour élaborer des données plus « construites » telles que les dessins, les modèles ou les annotations.

Toutes ces données apportent des informations sur les différentes entités biologiques (niveau d'expression d'un gène, interaction entre protéines ou ADN-protéines, structure 3D d'une protéine, etc.). Si, au lieu de regarder ces différents types de données cités précédemment, on se place du point de vue des entités biologiques, on constate que ces données couvrent un ensemble de *critères biologiques* (expression, localisation, interaction, etc.). Pour un critère biologique donné, on a un ensemble de valeurs possibles. Une (ou plusieurs) de ces valeurs peut être associée à chaque entité biologique concernée par ce critère.

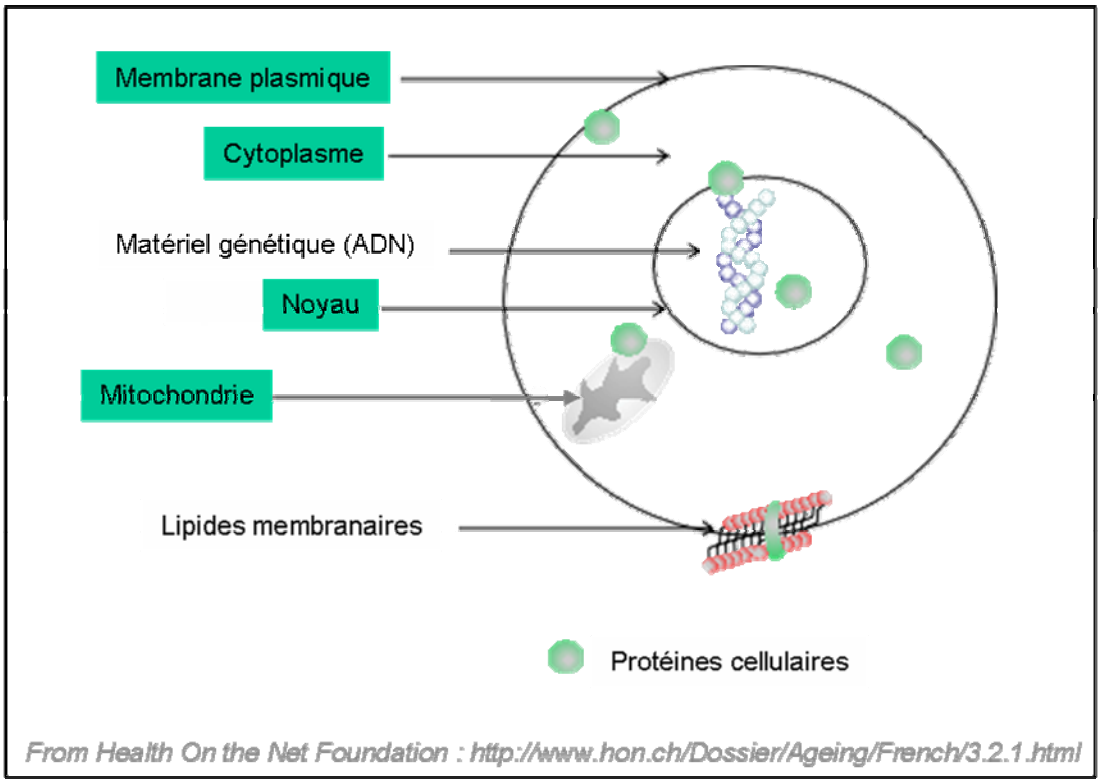
Les valeurs possibles d'un critère biologique, associées aux entités biologiques, peuvent être qualitatives ou numériques. On pourra alors parler de critères biologiques qualitatifs ou numériques.

✓ Les informations relatives à la fonction, à la structure d'une entité biologique et aux processus dans lesquels elle peut être impliquée, sont des informations qualitatives. Elles servent à annoter les entités biologiques, c'est-à-dire à clarifier

leur rôle dans la cellule. L'annotation consiste, par exemple, à associer une fonction biologique et/ou biochimique, une localisation cellulaire ou des domaines structuraux à ces entités biologiques. Ces informations servant à l'annotation correspondent à un ensemble fini de termes. Si on prend l'exemple d'un critère biologique tel que la localisation cellulaire des protéines (Figure 1), à chaque protéine (entité) on va pouvoir associer le nom du compartiment cellulaire (valeur) dans lequel elle a été détectée: la protéine P1 est située dans le noyau, la protéine P2 est dans le cytoplasme et la mitochondrie, etc. Selon la source de données considérée, l'annotation des entités peut varier plus ou moins. Par conséquent, chaque source pourrait constituer un critère biologique différent. Parmi les données qualitatives, on retrouve également les informations sur le métabolisme, sur les motifs, sur les interactions protéines-protéines, etc.

Figure 1: Les données qualitatives: exemple de la localisation cellulaire des protéines

Les protéines peuvent être caractérisées, entre autres, par leur localisation dans la cellule (en vert). La protéine peut donc appartenir à la membrane plasmique, au noyau, à la mitochondrie, etc. Cette localisation peut être encore plus précise si on regarde au niveau des compartiments sub-cellulaires (membrane externe de la mitochondrie, nucléole...). Tous ces compartiments ou sous-compartiments sont un ensemble fini de termes (valeurs qualitatives) qui peuvent être associés aux entités biologiques.



✓ Parmi les données numériques, on trouve par exemple, celles issues des puces à ADN qui sont des données quantitatives pouvant prendre un nombre infini de valeurs. A une entité biologique (un gène) est associé un ensemble de valeurs correspondant aux variations d'expression du gène dans différentes conditions biologiques (Figure 2a). On retrouve également des données numériques qui décrivent les propriétés physico-chimiques telles que le point isoélectrique, l'hydrophobicité d'une protéine, ainsi que des données ordinales qui rangent des éléments les uns par rapport aux autres telles que les coordonnées chromosomiques des gènes (Figure 2b).

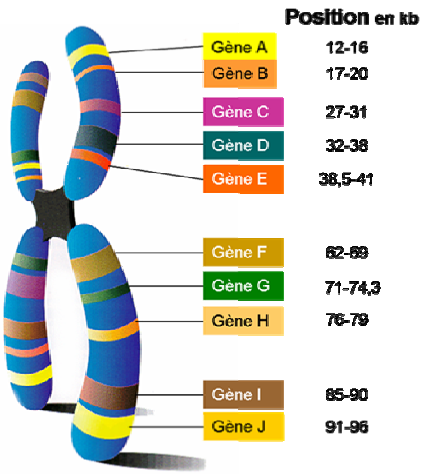
Figure 2: Les données numériques: les exemples des puces à ADN et de la localisation chromosomique

a) Des données de puces à ADN correspondent à un tableau de valeurs ; à chaque gène (première colonne) est associé un ensemble de valeurs numériques correspondant aux variations du niveau d'expression de ce gène dans différentes conditions. b) Données sur la localisation chromosomique des gènes : elles correspondent à des coordonnées sur le chromosome qui ordonnent les gènes comme c'est représenté sur le schéma (position en kb = kilobases).

a.

ORF	Heat 15min	Heat 30min	Heat 45min	Heat 60min	Heat 120min	Acid 10min
YAL001C	0.088379622	0.791333426	0.568478973	-0.490953941	-1.020532633	1.263826225
YAL002W	0.548436825	1.233939175	1.680114775	1.153805336	1.295233228	0.594451419
YAL003W	-0.145156949	-0.192716853	-0.224183728	-0.179805553	-0.364474552	0.005979426
YAL004W	1.671535504	2.3182941	0.782514068	0.174957079	1.919537972	1.171705008
YAL005C	1.382467175	1.936902742	1.587057832	1.398849929	1.793604423	0.982489832
YAL007C	0.692929864	1.265521383	1.601002512	0.712991404	0.922694274	-0.032678666
YAL008W	2.373100252	2.306001066	2.411358717	1.78608746	2.420233102	1.40747526
YAL009W	1.190451155	0.546845527	1.226091362	1.229997579	1.025695615	0.140092517
YAL010C	0.779490211	0.236416299	0.824283361	0.886747448	1.328489539	-0.239566125
YAL011W	0.375337375	0.817915567	0.022088759	0.354210334	0.186138299	0.679591947
YAL012W	-0.153688009	0.251663916	0.121878649	0.232095483	-0.315586568	-0.09047681
YAL013W	0.333454495	0.32331171	0.069452988	0.688141912	0.122297278	0.298917769
YAL014C	0.537561423	0.544639254	0.767199773	0.811884196	0.686842115	0.256040159
YAL015C	1.570452874	1.60964143	1.792342351	1.644453455	1.841510631	0.244588167
YAL016W	0.726801211	0.940680123	0.797077355	0.346517812	0.788395442	0.416297059
YAL017W	2.010430309	1.504066906	1.202957283	0.985654223	1.944215633	1.442380263
YAL018C	0	0.280956314	0	0	0.769771739	0
YAL019W	-0.259226661	1.051849283	1.116311725	0.79958278	0.797221853	-0.722162969

b.



De www.public.asu.edu/~acstone/humgen

L'intérêt de toutes ces informations (qu'elles soient qualitatives ou numériques) est qu'elles nous permettent d'identifier des relations possibles entre des entités biologiques. Quand il s'agit d'informations qualitatives, on peut facilement rassembler des entités annotées avec le même terme. En revanche, lorsqu'il s'agit d'informations numériques, il est nécessaire de mettre en place des méthodes de classification qui permettent de regrouper les entités qui ont des valeurs proches ou des comportements similaires (mesure de ressemblance/similarité entre les entités).

Ce problème de classification d'entités biologiques est l'un des points abordés dans cette thèse (voir Chapitre 2).

2.2 Origines et formats des données

Le séquençage d'un génome complet, quelle que soit sa taille, que ce soit le génome d'un eucaryote ou d'un procaryote, est présenté sous forme textuelle. De nombreuses données peuvent être stockées sous cette forme dans des fichiers plats (ou « à plat »). Ainsi, la majorité des données biologiques serait sous forme de fichiers plats (environ 80% au format texte, plus ou moins structuré). Ces mêmes données peuvent se retrouver dans des bases de données sous forme de fichiers indexés ou dans des bases de données relationnelles, ce qui facilite leur accès et leur exploitation.
Toutes ces données proviennent de sources différentes, de laboratoires disséminés dans le monde entier. Un certain nombre de bases de données généralistes (GenBank, UniProt) ou spécialisées (Comprehensive Yeast Genome Database, MGI Mouse Genome) ont été développées afin de stocker, gérer et avoir accès aux informations plus efficacement. Il est également nécessaire de créer des liens entre ces bases afin de relier les différents types d'informations qu'elles mettent à disposition sur les gènes et les protéines. Dans toutes ces bases de données, on trouve donc des références croisées; ce sont des liens reflétant une relation biologique : ils permettent de relier les informations présentes dans différentes bases et correspondant à un même objet biologique. Par exemple, un gène G, avec toutes les informations qui lui sont associées dans la base A, va pouvoir être relié à la protéine P_G pour laquelle il code, décrite dans la base B.

Des systèmes d'interrogations comme SRS (Etzold, et al., 1996) ou Entrez (Benson, et al., 1994), servent d'interface pour accéder à plusieurs bases de données simultanément. Ils permettent l'accès et l'interrogation de différentes bases de données grâce à des systèmes de requêtes. Ces systèmes aident à mettre en relation différents types de données provenant de différentes bases de données en exploitant les références croisées. Ces « méta-databases » correspondent à une couche « logicielle » supplémentaire regroupant un ensemble de bases de données.

Cette diversité dans la nature, l'origine et le format de ces données se retrouve aussi dans leur qualité. Les données issues des différentes approches et/ou techniques de génomique et post-génomique sont à utiliser avec précaution. En effet, il arrive que

certaines informations soient manquantes et/ou erronées selon les méthodes utilisées : faux-positifs et faux-négatifs. Par exemple, avec les méthodes d'identification d'interactions protéiques telles que le double hybride, on obtient des taux élevés de faux positifs (interactions détectées mais non avérées *in-vivo*) et faux-négatifs (interactions non détectées alors qu'elles existent) (Ito, et al., 2001). De même, pour les données de microarrays, on remarque qu'il peut y avoir des valeurs manquantes dans une expérience. Différentes méthodes ont été mises en place pour inférer ces valeurs manquantes (Johansson, 2006 ; Jornsten, 2007), et de nombreuses méthodes statistiques ont été développées pour normaliser, corriger les données brutes issues des technologies haut débit dans le but d'améliorer la qualité des données (Canales, et al., 2006; Kemmeren and Holstege, 2003; Klebanov and Yakovlev, 2007; Purvine, et al., 2004; Ramirez, et al., 2007).

Selon la technique employée et les méthodes de correction utilisées, des données de qualité différente sont obtenues. Ce point augmente encore l'hétérogénéité des données même si elles sont issues d'une même technologie.

3. Intégration des données

On a vu qu'il existait une grande diversité de données. Certaines sont de même nature mais sont générées par des techniques différentes (des données d'expression peuvent être produites par SAGE ou puces à ADN) et peuvent être comparées afin de ne garder que ce qui est retrouvé en commun entre elles, et ainsi améliorer la qualité des données.

D'autres sont issues d'approches différentes et indépendantes les unes des autres (données sur l'expression des gènes, sur les interactions entre protéines...). Ces données sont complémentaires, et leur confrontation est indispensable puisqu'elle doit permettre d'augmenter la fiabilité des résultats obtenus (Ge, et al., 2003) mais aussi de renforcer ou d'affiner la fonction, le rôle attribué aux différentes entités biologiques étudiées (Kemmeren and Holstege, 2003; Kemmeren, et al., 2002). En effet, les données résultant d'une seule approche (transcriptomique, protéomique ou autres) ne donnent qu'une idée limitée de la fonction des entités biologiques alors que l'intégration de données de plusieurs approches permet d'avoir des conclusions plus fiables (Kemmeren, et al., 2005).

3.1 Différents types d'intégration de données

L'intégration de données, quelle que soit leur nature, est un besoin résultant de l'explosion de leur quantité, de la diversité de leurs origines, et de la nécessité grandissante que nous avons à partager et à connecter ces données. Ce besoin d'intégration existe dans divers domaines tels que la finance, les assurances, le tourisme, l'environnement, les neurosciences, la génomique, etc. Il a même été récemment défini comme le *défi fondamental de la science du 21ème siècle* (Heidorn, et al., 2007). Les scientifiques ont pris conscience de l'importance de l'intégration de données à toutes les échelles de la biologie (de la biologie moléculaire aux écosystèmes).

Afin de pouvoir intégrer toutes ces données, il est nécessaire qu'on sache clairement ce que chacune représente, et qu'elles soient comparables. Pour cela, les données issues d'une même approche doivent être décrites sous forme standard. Les standards permettent d'unifier la façon de présenter les données afin de les comparer et de les échanger plus facilement. Plusieurs formats ont été spécifiés pour différents types de données biologiques :

- d'une part il y a des standards d'information minimum qui spécifient une manière de décrire une expérience avec un minimum requis d'informations tels que MIAME (Brazma, et al., 2001) qui signifie *Minimum Information About Microarray Experiments*, ou MIARE, *Minimum Information About an RNAi Experiment*. De nombreux autres ont été développés et seront développés car ces standards permettent d'unifier la description des données issues d'une même technologie et sont maintenant demandés par certains éditeurs pour publier des résultats s'appuyant sur des données « omiques ». Ceci doit permettre la reproductibilité des données en améliorant la qualité de la description de l'expérience, et une meilleure traçabilité des données. Ces points sont très importants du fait de la mise à disposition publique de ces données et par conséquent, de leur utilisation à long terme par des chercheurs ;
- d'autre part on a des ontologies ou vocabulaires contrôlés qui ont été créés afin d'unifier les termes permettant, par exemple, de décrire les gènes ou les protéines dans n'importe quel organisme vivant comme dans l'ontologie GO (Gene Ontology[2]) (Ashburner, et al., 2000), ou de décrire l'anatomie et la morphologie des plantes à fleurs comme dans l'ontologie PSO (Plant Structure Ontology) (Ilic, et al., 2007).

Ces représentations unifiées de l'information permettent de rapprocher des données de même nature, et doivent faciliter l'intégration de ces données .

[2] Gene Ontology [http://www.geneontology.org/]

L'intégration concerne des données de nature différente; c'est une stratégie, un processus qui doit permettre de rassembler ces données et en fournir une représentation unifiée. Elle peut donc être vue comme un outil pour extraire de nouvelles connaissances. Cette intégration peut être mise en œuvre de différentes façons et à différents niveaux.

✓ Une première forme d'intégration existe au niveau des bases de données. Des systèmes tels que SRS et Entrez, que l'on a déjà évoqués, permettent de mettre en relation différents types de données: ils exploitent les références croisées disponibles dans un ensemble de bases de données pour faire des liens entre des informations de nature et d'origine différentes. Par exemple, des références croisées permettent de faire le lien entre la fiche d'un gène dans GenBank et la fiche de la (ou des) protéine(s) pour laquelle il code dans UniProt. Ces liens sont utilisés pour aller chercher des informations à travers plusieurs bases de données à la fois, et permettent ainsi de rassembler, de recouper diverses informations sur les entités biologiques. Ces liens entre les bases de données sont pré-définis, et correspondent donc à des relations statiques entre des entités biologiques. Cette intégration se fait à l'échelle d'un gène ou d'une protéine.

A l'heure actuelle, les technologies (puces à ADN, double-hybride...) nous permettent d'étudier simultanément plusieurs gènes ou protéines. Il existe donc d'autres types d'intégration qui exploitent des ensembles de gènes ou protéines.

✓ Un autre type d'intégration s'intéresse à des ensembles d'entités biologiques. Celle-ci permet de croiser des données quelles que soient leur nature. Cette intégration peut passer par la représentation ou la visualisation de différents types de données simultanément afin de mettre en évidence les liens existants entre ces données. Par exemple, il existe des outils bioinformatiques qui permettent de colorer, dans une voie métabolique du KEGG, les enzymes correspondant à des gènes co-exprimés dans une condition particulière (Figure 3). Ainsi, d'éventuelles relations peuvent être mises en évidence grâce à l'intégration de différentes informations.

Figure 3: Exemple de visualisation de différents types de données

Cette figure illustre deux exemples d'intégration de données utilisant la visualisation. a) Le schéma représente un chromosome sur lequel on a ajouté des données d'expression issues de puces à ADN. Les gènes présents sur la puce à ADN sont marqués en jaune. Les gènes différentiellement exprimés dans cette expérience de puce à ADN sont notés avec une barre verticale, en dessous ou au-dessus du chromosome. Cette représentation permettrait d'observer une éventuelle correpondance entre les changements d'expression d'un ensemble de gènes et leur localisation sur le chromosome (Weniger, et al., 2007). b) Le schéma représente une carte (voie) métabolique du KEGG. Les enzymes colorées en rouge correspondent à des gènes retrouvés surexprimés dans une expérience de puce à ADN. Cette

représentation permet d'observer la relation entre un ensemble de gènes différentiellement exprimés et leur emplacement dans les voies métaboliques.

a.

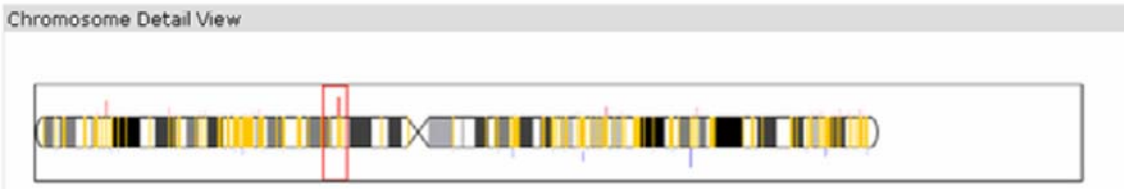
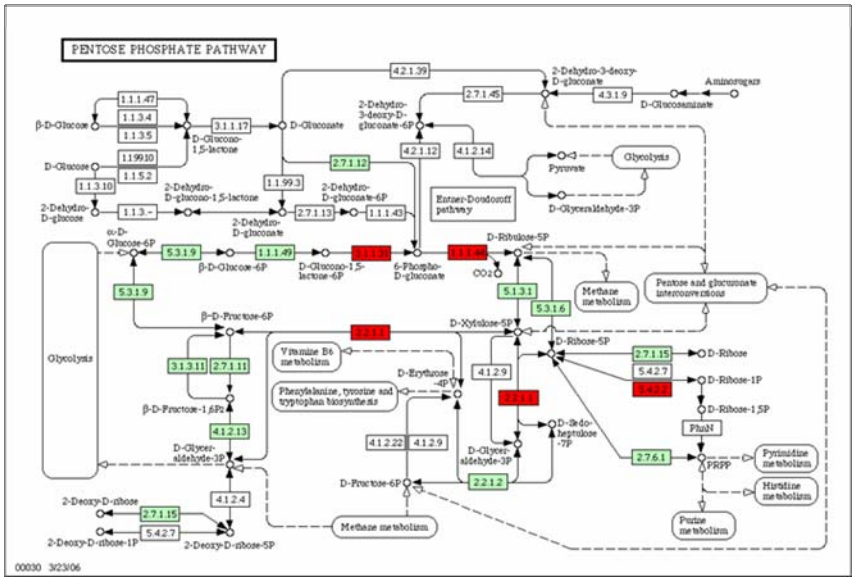


Illustration issue de Weniger, M., J. C. Engelmann, et al. (2007). BMC Bioinformatics 8(1): 179.

b.



Dans mon travail de thèse, je me suis intéressée à cette deuxième forme d'intégration, qui permet de mettre en relation des données hétérogènes. Cette intégration revient à celle employée en sociologie, c'est-à-dire le rapprochement d'une personne ou d'un groupe de personnes vers un autre groupe de personnes. C'est dans ce sens que j'ai travaillé sur l'intégration des données biologiques: comment rapprocher un type de données d'un autre type de données (c'est-à-dire deux critères biologiques) de manière appropriée, et de façon à ce que cela mette en évidence des contraintes ou mécanismes biologiques existant dans la cellule ?

L'intégration de données biologiques s'est développée avec la nécessité de rassembler et d'organiser la masse de données issue des approches de génomique et post-génomique afin de comprendre le fonctionnement de la cellule. La difficulté est d'être capable de croiser des sources de données hétérogènes et éparses pour mettre en évidence des relations entre ces informations. Les objectifs de cette intégration sont doubles :

- être capable de mettre en relation des données hétérogènes issues d'approches complémentaires afin d'augmenter la confiance et la pertinence

des résultats obtenus. Il s'agit, par exemple, d'être capable de croiser un groupe de gènes co-exprimés issus d'une expérience de puce à ADN étudiant une maladie, avec les voies métaboliques. Cette mise en relation doit permettre dans un premier temps de valider biologiquement le groupe de gènes co-régulés si on retrouve une correspondance avec une ou des voies métaboliques ;

- extraire des connaissances et affiner notre compréhension des mécanismes cellulaires. Si on reprend l'exemple du point précédent, une mise en relation efficace des gènes co-régulés et du métabolisme doit permettre dans un deuxième temps de mettre en évidence des processus métaboliques impliqués dans la maladie étudiée. Ces deux objectifs sont évidemment très liés.

De nombreux outils bioinformatiques s'appuyant sur des approches différentes, ont été développés pour répondre à ces objectifs.

3.2 Outils d'intégration de données

Un des défis de la bioinformatique est de proposer des outils et des méthodes permettant d'exploiter au mieux toutes les données biologiques disponibles. Il s'agit en particulier de permettre le rapprochement entre les informations moléculaires et les connaissances disponibles sur les fonctions cellulaires afin d'aider à l'interprétation des données massives issues des technologies à haut débit/à grande échelle. De nombreux outils ayant des approches différentes proposent des solutions pour croiser des données hétérogènes basées sur le second mode d'intégration cité au paragraphe précédent (partie 3.1). Il existe deux principales stratégies : celle qui se base sur une représentation visuelle des données, et celle qui se base sur des calculs statistiques.

✓ Certains outils se basent sur une représentation visuelle de différents types de données (voir Figure 3). Ils permettent de visualiser les liens entre des informations provenant de différentes sources en les superposant sur un même graphe ou schéma, ou en interrogeant des bases de données regroupant des réseaux tels que le métabolisme, les interactions, etc (Arakawa, et al., 2005; Baitaluk, et al., 2006; Mlecnik, et al., 2005). Certains proposent des méthodes de traitement des données "brutes" avant d'effectuer la superposition des données.

✓ Parmi les outils se basant sur des calculs statistiques, on observe deux types de méthodes. L'une permet de mettre en évidence un enrichissement fonctionnel dans un groupe de gènes/protéines, elle est appelée Gene Set Enrichment Analysis (Al-

Shahrour, et al., 2007; Shamir, et al., 2005; Subramanian, et al., 2005). Un test statistique est utilisé pour indiquer la fiabilité de cet enrichissement. Un autre type de méthode permet de mettre en relation des groupes d'entités biologiques ayant des propriétés particulières : par exemple, on peut mettre en évidence une relation entre un groupe de gènes co-exprimés et la localisation cellulaire des protéines qu'ils codent en utilisant un calcul statistique qui donne une valeur de confiance pour cette relation. Selon l'outil utilisé, les chercheurs ont la possibilité de combiner un plus ou moins grand nombre de types de données : un certain nombre de ces outils se limite à mettre en relation un groupe de gènes avec les termes GO (Khatri and Draghici, 2005; Wrobel, et al., 2005), d'autres se focalisent sur le rapprochement de données d'expression avec des données fonctionnelles (Montaner, et al., 2006), et d'autres permettent d'explorer les relations entre toutes sortes de données (Barriot, et al., 2004; Carmona-Saez, et al., 2007; Hu, et al., 2005; Robinson, et al., 2002; Shannon, et al., 2003; Zhang, et al., 2005).

Remarque: GEPAT est un outil qui propose les deux types d'approches (statistique et visuelle) pour analyser et interpréter des données de transcriptomique issues de puces à ADN. GEPAT (Weniger, et al., 2007) offre aussi la possibilité d'utiliser une approche statistique telle que l'analyse de l'enrichissement en termes de l'ontologie GO d'un groupe de gènes différentiellement exprimés, ou une approche visuelle en colorant ces gènes différentiellement exprimés dans des réseaux représentant les interactions protéiques et les voies métaboliques.

Comme on peut le voir à travers le nombre d'outils bioinformatiques développés ces dernières années, la question de l'intégration de données hétérogènes est d'un intérêt majeur pour beaucoup de chercheurs.

Mon projet de thèse a consisté à étudier, par une approche statistique, la représentation des données biologiques dans le but de rendre leur intégration plus efficace.

4. Problématique de la thèse et plan du manuscrit

Les approches de génomique et post-génomique qui se sont développées ces dernières années génèrent toujours plus de données. La quantité massive d'informations biologiques disponibles et leur nature hautement hétérogène, rend leur intégration nécessaire mais difficile, et fait l'objet de nombreux travaux en bioinformatique. Cette intégration est l'une des thématiques de recherche du Centre de Bioinformatique de Bordeaux et est au coeur de mon projet de thèse. Pour ce projet, j'ai étudié la levure *Saccharomyces cerevisiae* pour laquelle il existe un grand nombre de données.

Le premier chapitre de ce manuscrit décrit « le matériel » utilisé pour les différents travaux de cette thèse, c'est-à-dire les données, les outils et méthodes communs à tous les chapitres. Tout d'abord, nous abordons les principales caractéristiques de la levure *Saccharomyces cerevisiae* et les différentes données disponibles pour cet organisme modèle. Une deuxième partie est consacrée à la technologie des puces à ADN et à l'analyse des données transcriptomiques qu'elle génère. Une troisième partie présente la stratégie sur laquelle sont basés les différents travaux: représentation des données biologiques sous forme de groupes de gènes ou protéines, et comparaison des groupes pour mettre en correspondance les différentes données, et ainsi trouver d'éventuelles relations entre elles. Pour finir ce chapitre, nous présentons l'outil BlastSets, qui a été développé au Centre de Bioinformatique de Bordeaux pour l'intégration de données biologiques hétérogènes. Cet outil a été utilisé pour mener les recherches présentées dans ce manuscrit.

Le second chapitre présente le travail mené sur la représentation des données biologiques. Différentes méthodes pour représenter des données sous forme de groupes sont possibles. Nous avons développé une stratégie qui permet de déterminer, parmi différentes représentations, celle qui est la plus en accord avec les connaissances biologiques déjà établies. Cette stratégie a été appliquée à des données d'expression des gènes et aux points isoélectriques des protéines. Ce travail a montré l'importance du choix de la représentation lorsque l'on veut intégrer différents critères biologiques. Il a fait l'objet d'une publication, soumise récemment (voir Annexe 2 - Article n°1).

Dans le troisième chapitre, ce sont des travaux portant sur la représentation du métabolisme qui sont présentés. Ce travail a été effectué en collaboration avec une équipe de chercheurs du KEGG (Japon). Cette équipe a proposé une nouvelle façon

de décomposer le réseau métabolique qui s'appuie sur une méthode mathématique. Notre collaboration a consisté à valider la pertinence biologique de cette décomposition. Cela nous a conduit à proposer une représentation originale du métabolisme qui, combinée à des données d'expression, nous a permis de mettre en évidence de nouvelles routes à l'intérieur du réseau métabolique global. L'ensemble de ce travail a fait l'objet d'une publication donnée en annexe (Annexe 3 - Article n°2).

Dans le dernier chapitre, nous présentons une étude qui utilise la représentation sous forme de groupes des données d'expression dans un contexte de génomique comparative. Nous avons exploré la possibilité de caractériser un gène par son voisinage d'expression, afin de voir si cela peut permettre d'identifier des gènes fonctionnellement équivalents ("orthologues") dans différents organismes. Les résultats préliminaires présentés ici devraient déboucher sur la publication d'un troisième article.

CHAPITRE 1 : « MATERIEL »

1. La levure comme modèle

L'ensemble des travaux de cette thèse s'est focalisé sur l'organisme *Saccharomyces cerevisiae*. Cet organisme fait partie du règne des champignons (*Fungi*), du phylum des *Ascomycota*, et du sous-phylum des *Hemiascomycetes*. Cette levure est utilisée comme un organisme modèle en génétique : c'est un eucaryote simple dont le génome peut être facilement manipulé génétiquement et dont la croissance est rapide, ce qui en fait un organisme très approprié pour n'importe quelle étude biologique (Sherman, 2002). Cet organisme est un matériel facile à se procurer et peu coûteux à manipuler. De plus, il a grand intérêt en biotechnologie agroalimentaire, puisque cette levure est utilisée pour la fabrication de la bière, du vin, du pain.

Les souches de cette levure ont deux états stables haploïdes et diploïdes, chacun convenant plus particulièrement à tel ou tel type d'expériences (mutations récessives/test de complémentation). *Saccharomyces cerevisiae* est par ailleurs le premier eucaryote dont le génome fut entièrement séquencé (Goffeau, et al., 1996). Il est donc devenu un des organismes clés pour tout ce qui est recherche en génomique, il fait l'objet des premières études à grande échelle dans le domaine des puces à ADN, des interactions protéiques par analyse double-hybride, etc. L'utilisation massive de la levure comme un système modèle eucaryote vient aussi du fait que de nombreux gènes humains liés à des maladies ont un orthologue chez la levure, et qu'il y a une forte conservation des mécanismes métaboliques et de régulation entre ces deux organismes (Andrade, et al., 1998; Foury, 1997; Steinmetz, et al., 2002).
Le génome de *Saccharomyces cerevisiae* comporte 16 chromosomes nucléaires et un chromosome mitochondrial. Le génome est constitué de plus de 12 000 kilobases et 6609 gènes y ont été initialement identifiés. Aujourd'hui, après correction et amélioration des prédictions de gènes sur les différents chromosomes, le consortium Génolevures[3] répertorie 5673 gènes (Blandin, et al., 2000). Le consortium Génolevures pilote un projet de génomique comparative sur les levures hémiascomycètes, et met à disposition sur son site internet[4] les informations relatives aux diverses études menées sur ces levures. Tout le travail mené dans cette thèse sur la levure *Saccharomyces cerevisiae* a été effectué à

[3] Site web du consortium Génolevures [http://cbi.labri.fr/Genolevures]
[4] Génolevures [http://cbi.labri.fr/Genolevures/]

partir de la liste de gènes nettoyée et mise à jour par le consortium Génolevures. Parmi tous les gènes de *Saccharomyces cerevisiae*, environ 1000 n'ont toujours pas d'annotation fonctionnelle aujourd'hui malgré le nombre et la diversité des études effectuées sur cette levure (Pena-Castillo and Hughes, 2007).

Il existe de nombreuses bases de données spécialisées regroupant des informations sur *Saccharomyces* telles que :

- Saccharomyces Genome Database (SGD) qui fournit un accès aux séquences génomiques de la levure, à ses gènes et protéines, aux phénotypes des mutants et toute la littérature et les informations relatives à son génome (Cherry, et al., 1998);

- CYGD – The MIPS Comprehensive Yeast Genome Database qui regroupe des informations sur les fonctions cellulaires, sur les séquences, gènes et protéines de *Saccharomyces cerevisiae* ainsi que sur les interactions physiques et fonctionnelles entre les protéines et d'autres éléments génétiques (Guldener, et al., 2005);

- Yeast Microarray Global Viewer (Marc, et al., 2001), yMGV, qui regroupe des données de puces à ADN sur *Saccharomyces cerevisiae* et fournit une représentation graphique des variations d'expression d'un gène ou permet de récupérer des groupes de gènes qui ont des profils d'expression similaires. Depuis quelques années, ont été ajoutées des données d'expression d'une autre levure qui fait partie d'un sous-phylum différent (*Archiascomycètes*), la levure *Schizosaccharomyces pombe* (que nous avons également utilisée dans les travaux présentés au chapitre 4). L'intérêt de réunir ces deux levures se trouve en génomique comparative. Il s'agit de profiter de la masse de données disponible sur *Saccharomyces cerevisiae*, pour faire de l'inférence et aider les biologistes qui travaillent sur *Schizosaccharomyces pombe* (Lelandais, et al., 2004);

- SCPD – Promoter database of *Saccharomyces cerevisiae* (Zhu and Zhang, 1999) permet d'étudier de façon systématique et à l'échelle du génome entier, les promoteurs et les séquences régulatrices de la transcription des gènes de la levure.

Ces quelques exemples de bases de données ne sont pas exhaustifs; de nombreuses autres existent, regroupant plus particulièrement des informations sur les phénotypes associés à des changements/délétions génotypiques de la levure dans PROPHECY (Fernandez-Ricaud, et al., 2005), ou encore sur les différentes souches de levure (CBS Yeast Database). Au milieu de ces nombreuses bases de données, il est nécessaire de faire le tri, car certaines ne sont pas mises à jour régulièrement comme YIDB (Yeast Intron Database) ou font l'objet d'une exploitation commerciale et n'offrent qu'un accès restreint, comme Yeast Proteome Database (Csank, et al., 2002).

D'autre part, des bases de données généralistes contiennent également des informations sur la levure; on notera entre autre :

- Stanford Microarray Database (Ball, et al., 2005) stocke les données brutes et normalisées issues des puces à ADN de 50 organismes et permet également de récupérer, analyser et visualiser ces données;

- Genbank regroupe l'ensemble des séquences nucléiques disponibles publiquement avec leurs annotations pour plus de 240 000 organismes (Benson, et al., 2007);

- Protein Data Bank (Berman, et al., 2000) réunit les données tridimensionnelles (structure 3D) des macromolécules biologiques.

2. Puces à ADN et analyse du transcriptome

2.1 Introduction

La technologie des puces à ADN ou microarray est une des premières approches à haut débit à avoir été développée. Après s'être lancé dans le séquençage à grande échelle des génomes, c'est au tour des ARN messagers d'être passés au crible. Cette approche de transcriptomique permet de connaître à l'échelle de la cellule entière, et en une seule expérience, le niveau d'expression (quantité d'ARN messagers) de chacun des gènes de cette cellule. Le procédé se base sur la comparaison de deux états de la cellule: le niveau d'expression des gènes d'une cellule est comparé avec un niveau d'expression de référence, qui peut être celui d'une cellule dans un autre état ou dans son état de base (normal). Le niveau mesuré correspond donc à une expression différentielle. Une telle avancée technologique doit permettre aux chercheurs de comprendre les aspects fondamentaux de la croissance et du développement d'un organisme (quels gènes sont exprimés à un stade donné dans la cellule ?), de découvrir les gènes en cause dans les maladies humaines, etc.

La première publication (Schena, et al., 1995) présentant l'utilisation de puces à ADN a eu du mal à convaincre les scientifiques de l'époque. En effet, l'étude publiée concernait 45 gènes de la plante modèle *Arabidopsis thaliana*, la communauté scientifique était donc sceptique quant à l'utilisation à plus grande échelle de cette technique et à son application chez l'homme, surtout à des fins diagnostiques. De plus, au départ, il n'existait aucun outil commercialisé et le coût d'une telle

expérience était très élevé. On constate que 10 ans plus tard, la technologie des puces à ADN est devenue la pierre angulaire de la recherche dans plusieurs domaines (découverte de médicaments, recherche clinique, diagnostiques, etc.); l'article de Schena *et al.* a déjà été cité plus de 4700 fois.

Aujourd'hui la technologie s'est démocratisée et le coût d'une expérience sur puce à ADN a considérablement baissé. La plupart des chercheurs ont abandonné leurs équipements « faits maison » pour utiliser des puces à ADN et des équipements commerciaux proposés par diverses entreprises. Les chercheurs ont aussi la possibilité de demander des puces à ADN « personnalisées », pour lesquelles ils choisissent les sondes qui seront présentes sur la puce, ou des puces à ADN « pré-construites » telles que celles couvrant largement le génome de l'homme, de la souris ou de la levure. De nombreuses entreprises fournissent ce genre de puces à ADN : Affymetrix, Agilent, GE Healthcare, ArrayIt, Eppendorf et bien d'autres.

2.2 Principe de la technologie

Cette technologie repose sur l'hybridation entre deux séquences nucléotidiques, l'une étant la sonde l'autre la cible. Les sondes sont des séquences, simples brins, déposées ou synthétisées sur un support solide, l'ensemble constituant la puce à ADN. Ces séquences immobilisées sur leur support sont appelées *spot*. Chaque spot correspond à un grand nombre de copies d'une même séquence. Ces spots sont ordonnés sur la puce afin de pouvoir identifier la localisation de chaque séquence sur la puce: chaque sonde est clairement identifiée et une annotation lui est associée.
Les cibles sont des ARN messagers extraits d'une cellule, amplifiés, marqués puis hybridés sur les sondes.

Il existe deux principaux types de puces à ADN (le terme *puce à ADN* étant générique) et sont illustrés dans la Figure 4: les micro-arrays à ADN complémentaires (ADNc) et les puces à oligonucléotides. Les principales différences sont sur le nombre de sondes présentes par puce et leur préparation.

✓ Les micro-arrays à ADNc (Figure 4a) sont constitués d'une lame de verre ou d'une membrane de nylon sur laquelle sont déposées des centaines ou milliers de sondes correspondant à des séquences d'ADNc. D'un autre côté, différents échantillons d'ARN messagers sont extraits de cellules (échantillons correspondant à

2 conditions d'étude différentes, ou à un témoin et une condition d'étude). Ces ARN messagers sont utilisés comme matrices pour être rétro-transcrits en ADN complémentaires dans lesquels sont incorporées des molécules fluorescentes. Des fluorochromes différents (vert Cy3 pour la condition A, et rouge Cy5 pour la condition B) sont utilisés pour les deux conditions que l'on souhaite étudier. Les deux échantillons sont mélangés en proportions équivalentes puis déposés sur la puce à ADN. Alors a lieu l'hybridation compétitive: les séquences marquées des deux échantillons vont entrer en compétition pour s'apparier à leur sonde complémentaire. Une fois l'hybridation effectuée, la puce à ADN est scannée c'est-à-dire que chaque spot est excité par des lasers (un pour chaque fluorochrome) couplés à un microscope et à une caméra qui travaillent ensemble pour créer une image numérique de la puce à ADN qui est conservée dans un ordinateur. Après traitement de cette image, on obtient une retranscription de l'image de l'ensemble des spots sous forme de couleurs allant du vert au rouge en passant par le jaune :

- ceux apparaissant en vert correspondent aux gènes majoritairement exprimés dans la condition A;
- ceux apparaissant en rouge correspondent aux gènes qui sont exprimés majoritairement dans la condition B;
- ceux sans couleur (noir) correspondent à des gènes qui ne sont exprimés dans aucune condition;
- ceux ayant une couleur étant un mélange de vert et de rouge correspondent à des gènes exprimés de façon similaire dans les deux conditions.

Grâce à différents logiciels, on va pouvoir mesurer le rapport des intensités Cy3/Cy5 pour chaque spot, et donc déterminer l'expression relative de chaque ARNm dans les deux échantillons biologiques. On pourra ensuite procéder à diverses analyses en fonction de l'étude menée.

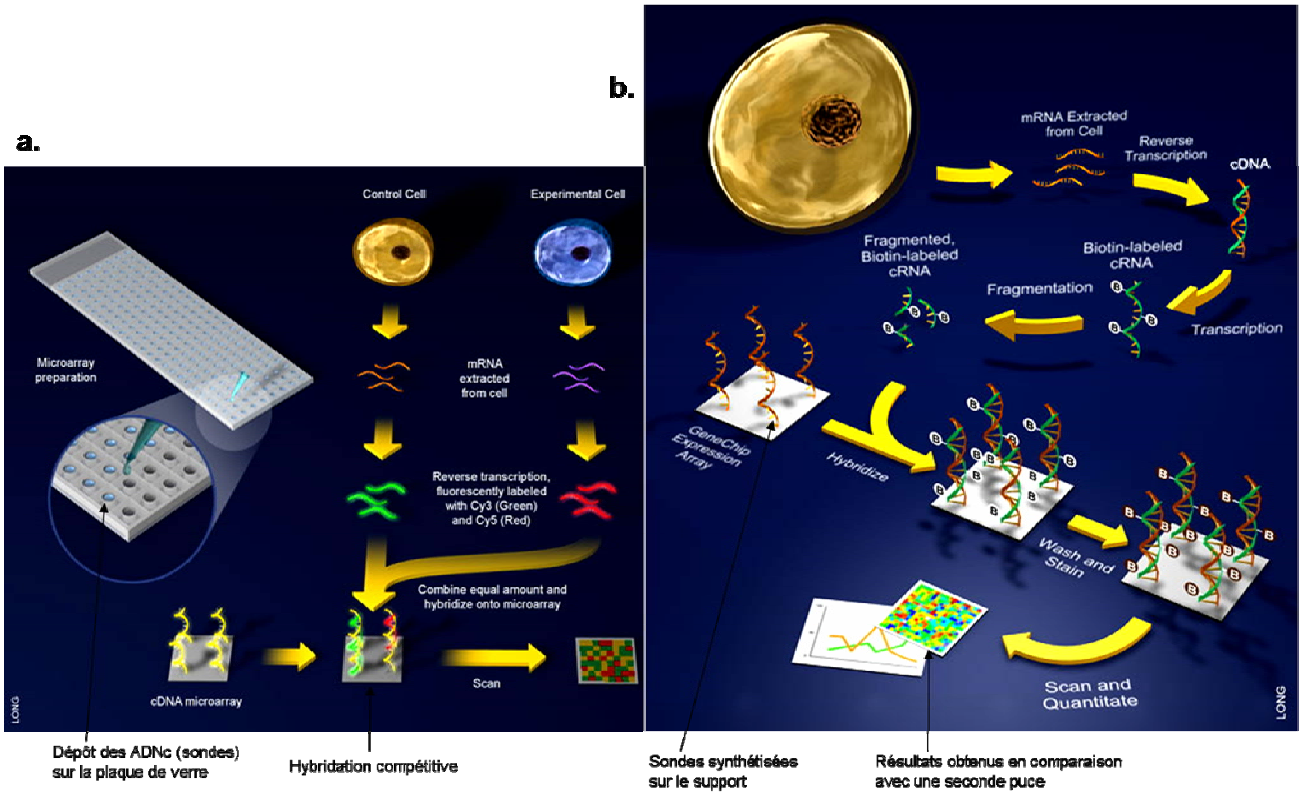
✓ Les puces à oligonucléotides, principalement développées par Affymetrix (Figure 4b), correspondent à un support sur lequel sont synthétisés des oligonucléotides d'environ 25 nucléotides. Ces oligonucléotides sont construits à partir du génome d'intérêt afin de correspondre à des séquences spécifiques de chacun des gènes de ce génome. Un génome entier peut être représenté sur ces puces (300000 oligonucléotides). D'un autre côté, les ARNm sont extraits d'une cellule, et contrairement aux puces à ADNc une seule condition est étudiée sur une puce. Les ARNm sont rétro-transcrits en ADNc et ces derniers sont transcrits *in vitro* en ARNc marqués à la biotine. Ces ARNc sont alors déposés sur la puce à oligonucléotides pour l'hybridation. La puce est lavée et colorée avec un fluorochrome qui se lie à la biotine.

Cette coloration permet de détecter les sondes sur lesquelles se sont hybridés les ARNc marqués. Comme pour les puces à ADNc, la puce à oligonucléotides est scannée et les signaux analysés grâce à des logiciels qui donnent le niveau d'expression des gènes présents sur la puce. Des puces étudiant des conditions différentes sont comparées entre elles pour pouvoir faire les études d'expression différentielle des gènes.

Figure 4: Principe des puces à ADN

Cette figure illustre les deux principales technologies de puces à ADN. a) Micro-array à ADNc. Des ADNc sont déposés sur une plaque ou une membrane et servent de sondes (jaune). Sur cette puce, deux échantillons, marqués de couleurs différentes, sont déposés et s'hybrident sur les sondes par hybridation compétitive. b) Puce à oligonucléotides. Ce schéma représente le principe des puces Affymetrix. Des oligonucléotides, correspondant à des gènes, sont synthétisés sur un support puis un échantillon est hybridé à ces sondes. Les puces sont comparées entre elles afin d'obtenir l'expression différentielle des gènes.

Images provenant de *http://plasticdog.cheme.columbia.edu/undergraduate_research/projects/sahil_mehta_project/work.htm*



Il est possible d'utiliser n'importe lequel des 2 marquages (biotine ou Cy3/Cy5) pour les micro-arrays à ADNc ainsi que pour les puces à oligonucléotides. Par exemple,

Agilent propose des puces à oligonucléotides pour lesquelles le marquage se fait avec les fluorochromes Cy3 et Cy5.

Cette technologie des puces à ADN permet de mettre en œuvre deux types d'approches. D'une part, on peut avoir une expérience qui compare simplement deux conditions (par exemple, cellule saine vs. cellule porteuse d'une maladie) et qui permet de mettre en évidence des groupes de gènes différentiellement exprimés (surexprimés ou sous-exprimés). Ces gènes sont un moyen de caractériser l'état des cellules, d'étudier ce qui est spécifique à leur état. Cette approche est utilisée pour la mise en évidence de marqueurs ou l'identification de gènes impliqués dans une maladie ou un processus particulier.
D'autre part, plusieurs expériences peuvent être rassemblées pour étudier plusieurs conditions à la fois (suivi de l'expression des gènes après plusieurs traitements et/ou au cours du temps). On obtient alors des profils d'expression pour chacun des gènes. A partir de ces données d'expression, on cherche à identifier des groupes de gènes qui ont des profils d'expression similaires afin de mettre en évidence des gènes qui semblent agir ensemble, ou dans un même processus. Pour cela, il est nécessaire d'utiliser des outils bioinformatiques de classification, ou de regroupement (*clustering*), comme Cluster (Eisen, et al., 1998). Cette approche est exploitée pour faire de la reconstruction de réseaux de régulation génique, ou pour identifier les processus biologiques mis en place par la cellule dans diverses conditions.

2.3 Les outils nécessaires à l'analyse

Comme il a été évoqué précédemment, le premier logiciel nécessaire pour traiter les résultats de puces à ADN est un logiciel de capture des couleurs et de traitement de l'image (exemple : ScanAlyze, ImaGene ou Spot). Ce genre de logiciel effectue trois tâches principales :

- le quadrillage consiste à localiser chaque spot sur la puce ;
- la segmentation consiste à différencier les pixels faisant partie du vrai signal et ceux faisant partie du bruit de fond ;
- l'extraction d'information consiste à mesurer l'intensité du spot et celle du bruit de fond.

Certains de ces logiciels vont plus loin en permettant la soustraction du bruit de fond et la normalisation des données.

Un grand nombre de méthodes et d'outils est maintenant disponible pour normaliser et analyser les données de puces à ADN. Selon la puce à ADN produite et la démarche sous-jacente du chercheur, des outils de plus en plus spécialisés ont été développés afin de répondre aux besoins, questions, problématiques de chaque laboratoire. Diverses packages de R/Bioconductor (Gentleman, et al., 2004) sont dédiés à l'analyse des puces à ADN : normalisation, calcul de log-ratio pour évaluer l'expression différentielle d'un gène, filtrage des spots d'intensité insuffisante, etc.
Plusieurs bases de données stockent les données de transcriptomique utilisant les puces à ADN. Par exemple, la base de données SMD met à disposition des outils et méthodes permettant à l'utilisateur qui veut récupérer des données de leur appliquer le traitement souhaité et de les obtenir sous le format désiré (mesures brutes ou normalisées, log-ratio ou autres). Il existe également des outils de stockage et de traitement des données tels que BASE (BioArray Software Environment). BASE est une plateforme qui contient une base de données permettant aux utilisateurs de stocker toutes les données produites à chaque étape de leur traitement. Des modules sont également disponibles pour le traitement, la visualisation et l'analyse des données. Les données peuvent être importées et exportées sous divers formats.

Une fois les données normalisées, traitées, des groupes de gènes (différentiellement exprimés, ou regroupés selon leur profil d'expression) peuvent être mis en évidence pour effectuer des analyses plus approfondies. Si on utilise une démarche manuelle, l'annotation de chacun des gènes du groupe identifié est examinée afin de trouver une voie métabolique impliquée dans la réponse aux conditions appliquées, ou un groupe de gènes dont l'expression est marqueur d'une pathologie, etc. Ce travail pouvant être très fastidieux de nombreux outils ont été développés pour mettre en relation un tel groupe de gènes avec des données fonctionnelles, permettant ainsi une analyse automatisée (Khatri and Draghici, 2005; Montaner, et al., 2006; Shamir, et al., 2005) (voir partie 3.2 dans l'Introduction).

3. Méthode de recherche de similarité entre des groupes

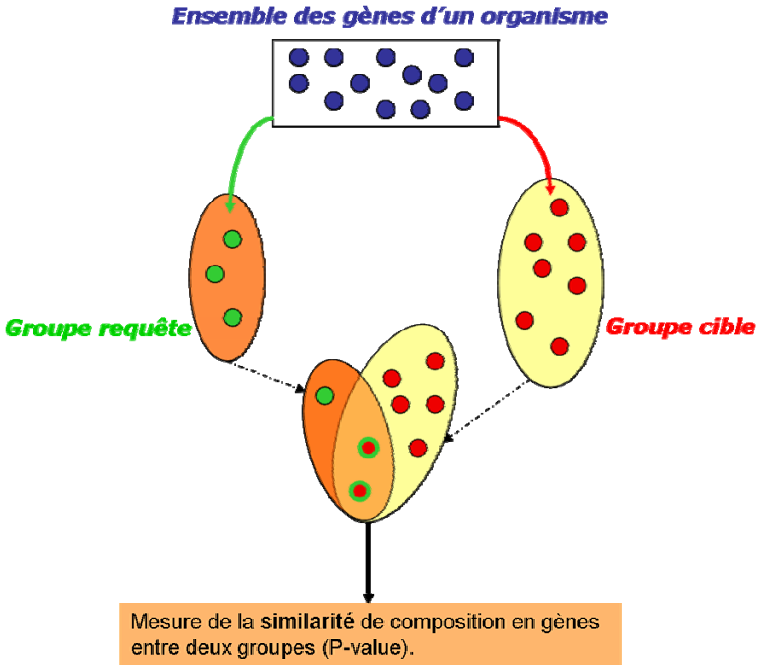
En introduction de cette thèse, l'intégration des données a été présentée comme un besoin pour analyser la masse de données très hétérogènes produites dans le domaine de la biologie. Une façon d'intégrer ces données si hétérogènes est de choisir une manière unique de les représenter quelle que soit leur nature. L'approche utilisée dans l'ensemble de cette thèse, et qui a déjà été exploitée (Barriot, et al., 2004), consiste à structurer les données sous forme de groupes afin d'être capable de les comparer.
Cette approche se base sur le concept de voisinage (Danchin, 1998). Un voisinage est un groupe correspondant à un ensemble d'entités biologiques (protéines ou gènes) ayant une valeur en commun pour un critère biologique donné : protéines appartenant à la même voie métabolique, gènes co-localisés sur un chromosome, protéines ayant des tailles similaires, etc. Comparer ces groupes permet de retrouver des relations entre des critères biologiques différents. On peut alors répondre à des questions biologiques telles que « Est-ce que les gènes que je viens d'isoler dans mon expérience sont déjà connus pour avoir une propriété en commun: la même localisation cellulaire, l'appartenance à un même complexe, ou la même annotation fonctionnelle ? ». On peut aller plus loin en recherchant des relations entre des critères biologiques, par exemple, « est-ce que toutes les protéines présentes dans un même compartiment cellulaire sont co-régulées, et cela, systématiquement pour chaque compartiment ? ».

Pour comparer des groupes, il est nécessaire d'avoir une mesure de similarité. Les groupes sont comparés sur la base de leur composition en gènes ou protéines (Figure 5): plus leur intersection sera grande plus ils seront similaires. Afin de pouvoir comparer aussi bien des groupes construits à partir de données sur les protéines que des groupes construits à partir de données sur les gènes, on utilise un identifiant unique pour chaque protéine et le gène codant cette protéine.
Pour mesurer la similarité entre deux groupes, nous utilisons la loi hypergéométrique qui permet de calculer la probabilité (P-value) d'avoir au moins le nombre observé de gènes en commun entre deux groupes qui peuvent différer en taille, et qui sont construits à partir d'un tirage dans une population de gènes. Cette population correspond à l'ensemble des gènes codant les protéines de l'organisme étudié (Figure 5). Cette P-value reflète la similarité entre deux groupes. Elle est considérée comme significative, c'est-à-dire que deux groupes sont significativement similaires, si elle est inférieure ou égale à un certain seuil. Ce seuil est choisi avant

d'effectuer la comparaison et correspond au niveau d'erreur accepté. Dans nos travaux, nous avons utilisé le niveau d'erreur $\alpha = 0.1$.

Figure 5 : Recherche de similarité entre deux groupes

Le groupe requête est comparé à un second groupe, le groupe cible. Ces deux groupes sont issus de l'ensemble des gènes d'un organisme : la similarité de composition entre ces deux groupes est mesurée en utilisant la loi hypergéométrique. On obtient une P-value qui est la probabilité d'avoir au moins le nombre observé de gènes ou protéines en commun entre les deux groupes.



Cette mise en correspondance de deux groupes, composés de gènes ayant un lien biologique (co-localisés, co-exprimés, etc.), peut être effectuée à plus grande échelle. On peut mettre en relation un groupe requête avec un ensemble de groupes cibles, correspondant à un critère biologique donné, pour chercher à annoter le groupe requête ou explorer la pertinence biologique des groupes cibles (en mettant en évidence une similarité entre groupe requête et groupe(s) cible(s), voir Chapitre 3). On peut également mettre en relation deux ensembles de groupes correspondant à deux critères biologiques différents afin de mettre en évidence des correspondances entre ces critères (voir Chapitres 2 et 4).

Pour faire ces mises en correspondance à plus grande échelle, on construit une *collection*[5] de groupes représentant chaque critère biologique que l'on veut étudier. Par exemple, si on s'intéresse à la localisation cellulaire des protéines d'un organisme, on peut construire une collection de groupes, qui représente le critère biologique « Localisation cellulaire ». Dans cette collection, chaque groupe est un ensemble de protéines annotées comme appartenant à un même compartiment cellulaire. Selon les critères, et donc selon les collections, les groupes sont ou non indépendants les uns des autres. Les liens entre ces groupes peuvent être illustrés par des graphes (Figure 6).

Figure 6 : Relations entre les groupes d'une collection

Exemple des relations hiérarchiques qui existent entre les différents compartiments cellulaires et sub-cellulaires. Les compartiments représentent des groupes d'une même collection « Localisation cellulaire ». Ce dessin illustre de deux manières différentes les relations existantes entre les compartiments.

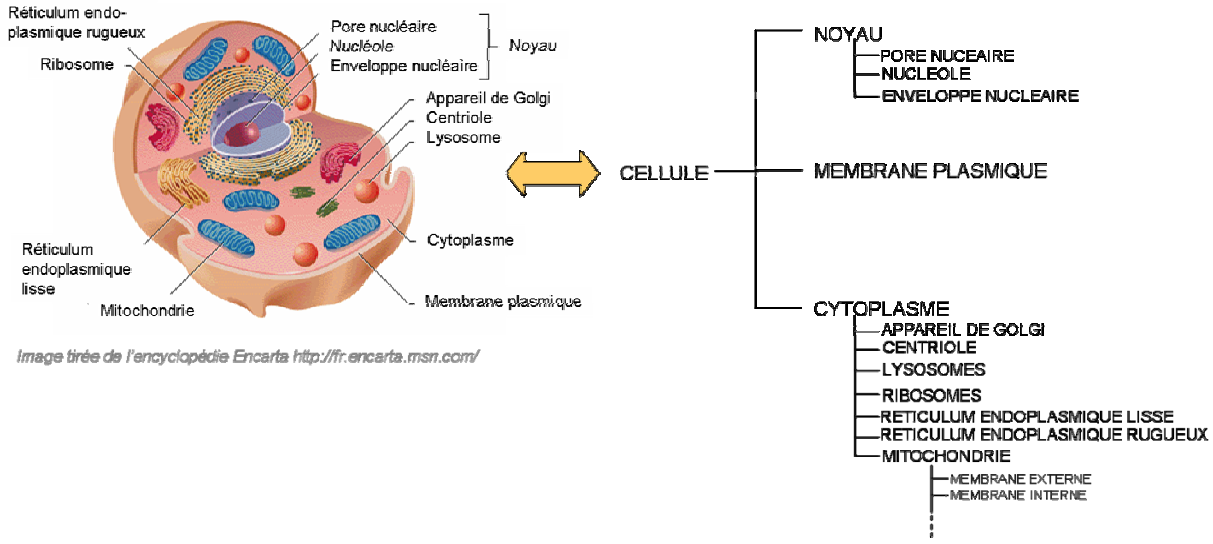


Image tirée de l'encyclopédie Encarta http://fr.encarta.msn.com/

Lorsqu'on souhaite comparer un groupe avec une collection de groupes représentant un critère biologique (par exemple, un groupe de gènes trouvés co-régulés dans une expérience de transcriptome comparé à la collection correspondant à l'ensemble des voies métaboliques) ou des critères entre eux (la collection des groupes de gènes trouvés co-régulés dans une expérience de transcriptome comparée à la collection correspondant aux voies métaboliques), de nombreuses comparaisons de groupes vont

[5] Tout au long du manuscrit, le terme *collection* est utilisé pour faire référence à l'ensemble des groupes (constitués de gènes ou protéines) représentant un critère biologique donné.

être effectuées. Pour corriger le biais que ces comparaisons multiples peuvent générer c'est-à-dire la probabilité de trouver des similarités par hasard, le seuil α doit être ajusté. Nous avons utilisé la correction de Bonferroni, souvent appliquée pour ce type d'ajustement (Castillo-Davis and Hartl, 2003; Robinson, et al., 2002; Wrobel, et al., 2005). Cette correction consiste à diviser le seuil de significativité α par le nombre de comparaisons effectuées, qui correspond au nombre de groupes comparés :

$$T = \alpha / n*m$$

T est le seuil ajusté ; n est le nombre de groupes soumis (1 ou une collection entière) ; m est le nombre de groupes appartenant à la collection sélectionnée qui va être comparée au(x) groupe(s) soumis.

La comparaison de groupes ou de collections de groupes est une approche générique qui peut aider à l'annotation de groupes, et permet d'étudier les relations pouvant exister entre des critère biologiques, à partir du moment où l'on peut les convertir sous forme de collections de groupes.

Pour différents critères biologiques tels que la localisation cellulaire des protéines, les interactions physiques entre protéines, la représentation sous forme de groupes est directe :

- l'ensemble des protéines présentes dans un même compartiment cellulaire correspond à un groupe, et ainsi chaque compartiment cellulaire devient un groupe, et l'ensemble de ces groupes forme une collection correspondant au critère biologique « Localisation cellulaire » ;
- de même, les protéines interagissant dans un même complexe forment un groupe, et l'ensemble de ces groupes forme une collection correspondant au critère biologique "Complexes Multi-protéiques".

La représentation d'un critère biologique sous forme de collections de groupes n'est pas toujours aussi intuitive. Pour certains critères biologiques, il est nécessaire d'appliquer un traitement sur les données pour générer des groupes de voisins; c'est particulièrement le cas pour des données numériques. Le choix d'une méthode de regroupement, et donc d'une représentation qui en découle pour un critère biologique donné, est la problématique principale abordée dans ces travaux de thèse.

Pour effectuer la manipulation et la comparaison (mesure de similarité) des groupes ou collections de groupes, nous avons utilisé l'outil BlastSets préalablement développé au Centre de Bioinformatique de Bordeaux par Roland Barriot (Barriot, 2005).

4. BlastSets : outil d'intégration de données hétérogènes

BlastSets[6] (Barriot, et al., 2004) est un outil d'intégration et de comparaison de données biologiques hétérogènes. Il a été créé pour pouvoir rassembler des données hétérogènes à l'échelle de génomes ou protéomes entiers afin de découvrir de nouvelles correspondances entre ces données. Cet outil permet de stocker et comparer des groupes en implémentant la méthode de recherche de similarité entre des groupes ou collections de groupes, présentée précédemment.

BlastSets est un outil ouvert puisqu'il peut intégrer n'importe quelles données biologiques définies sous un format standard: les données doivent être structurées sous forme de collections de groupes, décrites au format XML. Ce format permet d'insérer les collections dans la base de données de BlastSets (Figure 7). Une fois les collections stockées dans la base de données, on peut les comparer les unes aux autres, ou les comparer avec un groupe extérieur, en utilisant simplement l'interface web.

Figure 7: Architecture de l'outil BlastSets

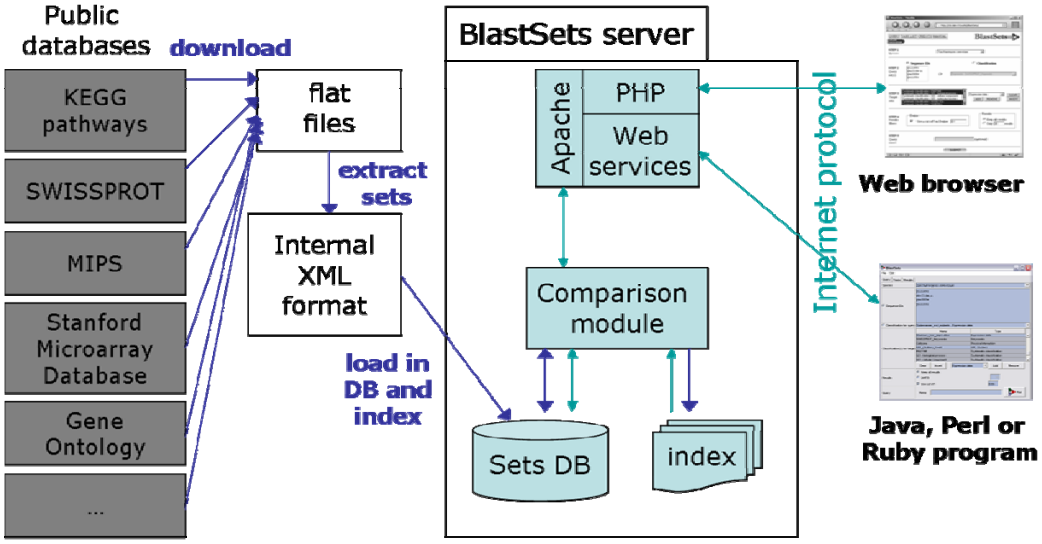


Schéma réalisé par Roland Barriot

[6] BlastSets [http://cbi.labri.fr/outils/BlastSets/]

Afin d'exploiter facilement n'importe quel type de données, il est nécessaire de faire le lien entre les gènes et leur produit. BlastSets utilise AliasServer (Iragne, et al., 2004), un outil qui permet d'avoir un identifiant unique pour chacune des protéines d'un organisme, et le même identifiant est affecté au gène codant cette protéine.

BlastSets est accessible sur internet à l'adresse http://cbi.labri.fr/outils/BlastSets/ et propose une interface simple pour comparer des groupes (Figure 8).

La première étape ('STEP 1') est de choisir l'organisme sur lequel on veut travailler grâce à un menu déroulant.

A la deuxième étape ('STEP 2'), BlastSets donne le choix entre lui donner un groupe de gènes ou protéines que l'on veut comparer aux groupes de sa base de données, ou choisir une collection ('Classification') de sa base que l'on comparera à une autre collection.

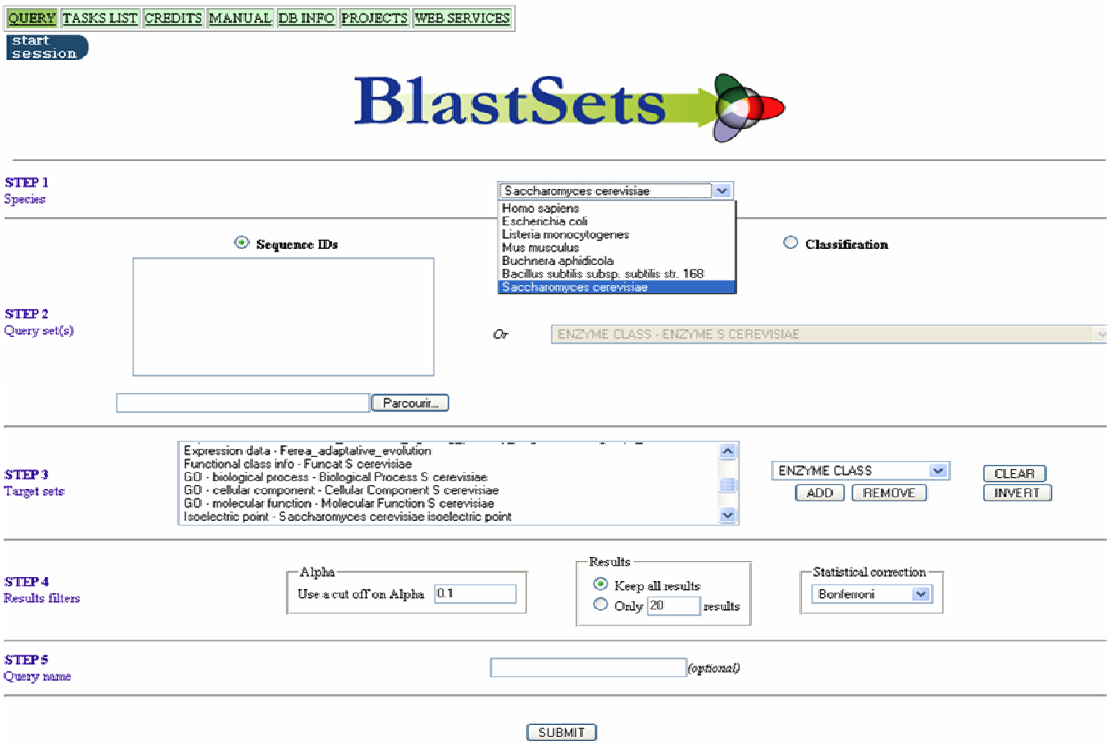
Ensuite ('STEP 3'), il faut sélectionner la ou les collections auxquelles on désire comparer notre groupe ou la collection choisie à l'étape précédente.

Enfin ('STEP 4'), l'utilisateur définit le seuil d'erreur accepté α et choisit une méthode de correction statistique (Bonferroni, par défaut).

Il ne reste plus qu'à donner un nom à la requête ('STEP 5') que l'on soumet au système (bouton 'SUBMIT').

BlastSets nous renvoie une liste de hits : ces hits sont des paires de groupes trouvés significativement similaires. Quelques informations supplémentaires sont données comme la taille de chacun des groupes du hit, la taille de l'intersection et la P-value associée à ce hit.

Figure 8 : Interface web de BlastSets



CHAPITRE 2 : Représentation des données biologiques en vue de leur intégration

En introduction, nous avons vu que de nombreux outils bioinformatiques ont été créés pour mettre en correspondance différents types de données biologiques provenant de sources variées. La plupart de ces outils sont spécialisés: un groupe de gènes ne pourra être mis en relation qu'avec un ou deux critères biologiques tels que les termes de l'ontologie GO ou les voies métaboliques du KEGG. Afin de pouvoir intégrer un large éventail de critères biologiques (localisation sub-cellulaire des protéines, interactions protéines-protéines, données d'expression des gènes, données sur les motifs protéiques, etc.), nous avons utilisé la méthode de recherche de similarités présentée dans le chapitre précédent. C'est une méthode très générale qui permet la mise en correspondance de nombreux critères biologiques de façon plus systématique, afin de retrouver des relations.
Afin d'être capable de mettre en relation ces différents critères, cette méthode s'appuie sur une représentation des données sous forme de groupes de gènes ou de protéines ayant une valeur similaire pour un critère biologique donné. Ainsi, n'importe quelle donnée biologique doit pouvoir être convertie sous forme de collection de groupes et ainsi être comparée à une autre collection (représentant un autre critère). Cette conversion est en fait une étape essentielle pour la comparaison de données, et elle est plus ou moins aisée selon le critère biologique.
Pour des critères biologiques tels que la localisation cellulaire des protéines, les interactions physiques entre protéines, le choix d'une représentation sous forme de groupes ne pose pas de problème : l'ensemble des protéines identifiées dans un même compartiment cellulaire correspond à un groupe ; l'ensemble des protéines interagissant dans un même complexe correspond à un groupe. En revanche pour d'autres critères biologiques, il est nécessaire d'appliquer une méthode de clustering, un traitement statistique et/ou bio-informatique sur les données pour être capable de les convertir en groupes: c'est plus particulièrement le cas pour des données quantitatives. Selon la méthode et les paramètres choisis pour créer des groupes, plusieurs représentations sous forme de collections de groupes peuvent être possibles pour un même critère biologique.

Ces méthodes de classification que l'on vient d'évoquer, correspondent à des algorithmes de regroupement en classes d'entités biologiques. Il en existe une large

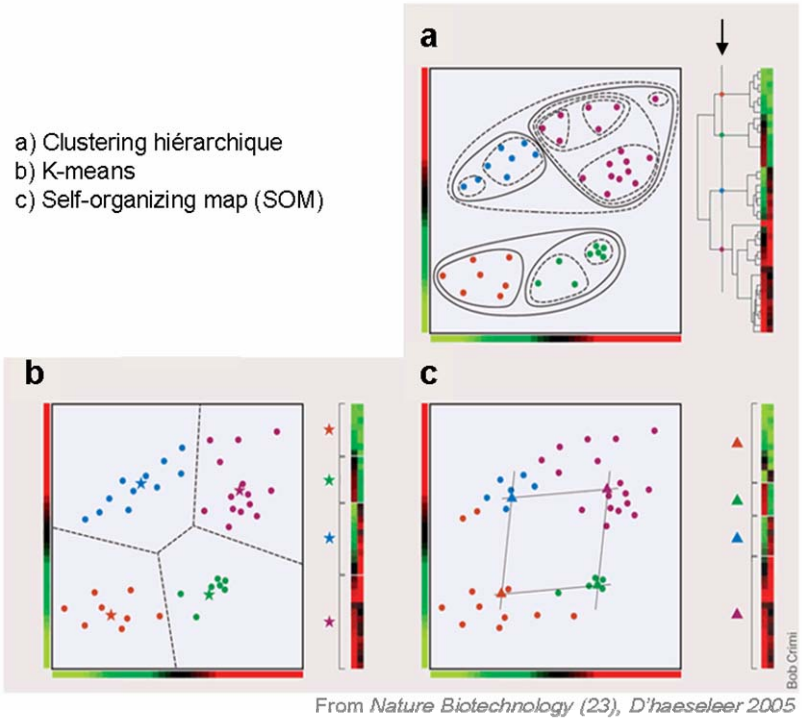
variété appartenant à deux grandes familles : les méthodes non-supervisées (clustering hiérarchique, k-means, self-organizing map, etc.) et supervisées (réseau de neurones, arbre de décision, etc.). Les méthodes supervisées nécessitent l'utilisation de connaissances à priori pour effectuer le clustering.
Les méthodes de clustering s'emploient pour toutes sortes de données. Sasson et ses collègues ont développé une technique de clustering ascendant (hiérarchique) pour classer les séquences protéiques selon leur similarité (Sasson, et al., 2002) et obtenir une granularité dans la taille des groupes pour retrouver des familles de protéines. Yoon et ses collègues ont utilisé une autre méthode, celle des k-means, pour classer des sites protéiques (sites actifs, sites de fixation, etc.) selon la similarité des environnements physico-chimiques des résidus (Yoon, et al., 2007) et ainsi prédire la fonction de ces sites.
Le clustering hiérarchique est l'une des méthodes les plus utilisées pour les données d'expression de gènes (D'Haeseleer, 2005). Les biologistes l'emploient pour identifier des groupes de gènes ayant des profils d'expression similaires et/ou des conditions générant des réponses trancriptionnelles similaires sur une puce à ADN. L'utilisation massive de cette méthode est en partie due à la mise à disposition rapide du logiciel Cluster (Eisen, et al., 1998) qui a rendu cette méthode accessible à tous. De plus, Cluster est facile d'utilisation et permet une visualisation graphique et colorée des résultats facilitant leur interprétation.
Il existe donc de nombreuses méthodes de clustering qui peuvent être utilisées sur diverses données biologiques. Le choix est souvent difficile parmi toutes ces méthodes. Elles ont donc été comparées, évaluées à plusieurs reprises afin d'aider et d'orienter les chercheurs dans leur choix (Datta and Datta, 2003; Datta and Datta, 2006a; Handl, et al., 2005; Yin, et al., 2006), mais chacune a ses avantages et ses inconvénients. La méthode la plus appropriée dépend des données utilisées et de la question à laquelle veut répondre le chercheur. Il faut noter que ces études comparatives sont faites majoritairement sur des données d'expression car le clustering est une étape importante dans l'analyse des profils d'expression de gènes.

Toutes ces approches de clustering permettent d'identifier des groupes de gènes ou protéines ayant une similarité, et ce, à partir de n'importe quelle donnée biologique. Si on revient à notre problème de conversion de données en collection de groupes pour mettre en oeuvre notre méthode de recherche de similarité, on peut utiliser toutes ces méthodes. Donc, pour chaque critère biologique, on peut générer autant de résultats de clustering différents, donc de représentations différentes qu'il y a de méthodes (Figure 9).

Le choix d'une méthode, et donc d'une représentation qui en découle, pour un critère biologique donné est la problématique abordée dans ce chapitre.

Figure 9: Représentations issues de différentes méthodes de clustering

Cette figure est extraite d'un article de D'Haeseleer (D'Haeseleer, 2005) et illustre 3 types de clustering sur des données de puces à ADN. Elle représente un exemple simple avec 40 gènes pour lesquels l'expression a été mesurée dans 2 conditions. a) Clustering hiérarchique : à droite, l'arbre binaire obtenu par clustering hiérarchique ; à gauche, le dessin représentant les groupes trouvés lorsque l'arbre binaire est coupé à un seuil indiqué par la flèche noire (4 groupes). b) k-means, avec k=4, partitionne l'espace en 4 sous-espaces en fonction des 4 centroïdes (étoiles). c) Self-organizing map (SOM) trouve 4 groupes organisés en grille.



Plus précisément, dans le travail présenté dans ce chapitre, il s'agit d'être capable d'exploiter des données quantitatives en mettant en place des méthodes de clustering qui permettent de construire des groupes de gènes ayant des valeurs proches pour un critère biologique donné. Ce travail est effectué en vue de l'intégration de ces données, c'est-à-dire que les méthodes utilisées doivent conduire à des représentations qui permettent de mettre en évidence des relations entre ces données et d'autres données biologiques.

C'est dans le but de trouver la représentation adéquate pour chaque critère biologique que j'ai mis en place une stratégie d'évaluation de différentes représentations. Cette stratégie

doit aider à choisir, parmi les différentes représentations, celle qui est le plus en accord avec les connaissances associées aux données et donc de faire ressortir des relations entre des critères biologiques.

Dans un premier temps, je décrirai de façon générale la stratégie mise en place pour comparer différentes représentations pour un critère biologique donné. Ensuite, j'illustrerai l'utilisation de cette stratégie avec d'une part les données de transcriptomique issues de puces à ADN, et d'autre part les données sur le point isoélectrique des protéines.

1. Approche générale : Stratégie d'évaluation de différentes représentations d'un critère biologique

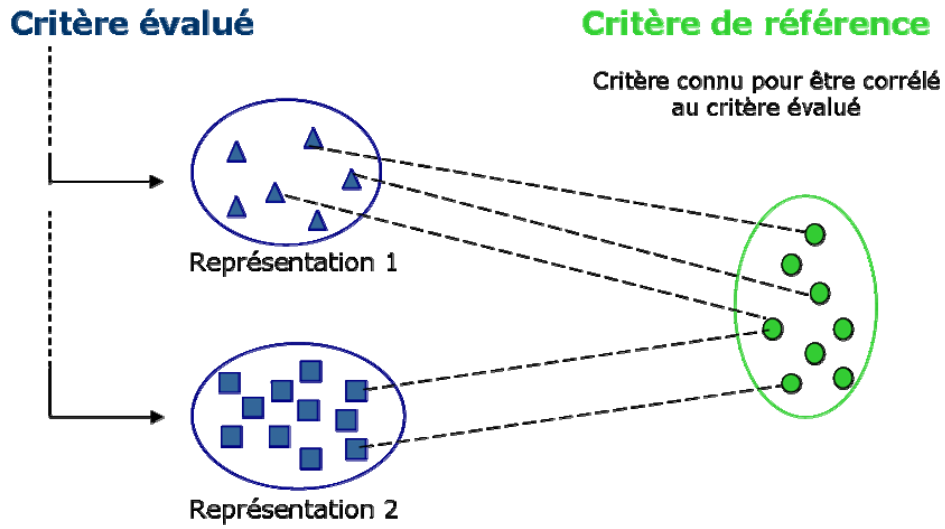
Afin que l'intégration de données hétérogènes soit performante et permette de révéler des relations entre des données, il y a deux étapes importantes : la façon dont on représente les données et la méthode de comparaison des données.
La représentation choisie pour tous les critères biologiques correspond à une collection de groupes. Pour certains critères, plusieurs représentations sous forme de groupes sont possibles. On veut donc pouvoir les comparer les unes aux autres afin d'être en mesure de choisir quelle représentation sera la plus appropriée dans une perspective d'intégration.

La stratégie que je propose repose sur la mise en relation de deux critères biologiques pour lesquels une correspondance a déjà été montrée (les opérons sont un exemple de correspondance entre les critères "localisation chromosomique" et "co-expression" des gènes). C'est pour l'un de ces critères (*critère évalué*) que l'on veut comparer différentes représentations, obtenues avec plusieurs méthodes de clustering ou de classification. L'objectif est de pouvoir évaluer la pertinence de ces représentations en utilisant un critère extérieur de référence. Ce *critère de référence* doit satisfaire deux caractéristiques essentielles: avoir une correspondance connue (publiée) avec le critère évalué, et sa représentation ne doit pas faire appel à une méthode ou à un traitement quelconque (représentation "naturelle" comme pour la localisation sub-cellulaire, par exemple).
La correspondance entre le critère de référence et le critère évalué étant connue, la représentation la plus pertinente doit être celle qui met en évidence de la façon la plus

significative cette correspondance, lorsqu'on compare chacune des représentations du critère évalué à celle du critère de référence. Cette stratégie est illustrée dans la Figure 10.

Figure 10: Schéma de la stratégie d'évaluation de différentes représentations

Chaque élément (triangles, carrés et ronds) représente un groupe appartenant à une collection symbolisée par une ellipse. Pour le critère biologique évalué, deux représentations (collections) sont possibles, les représentations 1 et 2. Chacune est comparée à la représentation (collection) du critère biologique de référence. La représentation montrant la plus forte correspondance avec le critère de référence sera considérée comme la plus appropriée. Sur cette figure, la représentation 1 a plus de groupes similaires avec la collection de référence, elle est donc la « meilleure représentation ».



Pour s'assurer que la correspondance que l'on retrouve, lorsqu'on compare deux critères, est significative, nous créons systématiquement pour l'un ou l'autre des critères (évalué ou de référence), une collection qui est un ensemble de groupes générés aléatoirement. Cette collection aléatoire est comparée à celle(s) du second critère et permet de vérifier que les résultats trouvés avec les collections réelles sont bien significatifs.

Cette stratégie est comparable à des approches d'évaluation de méthodes de clustering ou classification d'objets, mais nos objectifs sont différents.
Les approches d'évaluation de clustering cherchent à optimiser la méthode de regroupement afin de n'obtenir que des groupes pertinents. Elles se basent généralement sur des caractéristiques internes à la méthode de clustering telles que des indices de cohérence, de validité ou de stabilité des clusters obtenus.

Notre approche est différente car nous allons évaluer une représentation en utilisant un critère biologique extérieur, qui n'a aucun lien avec la méthode en elle-même. Ce critère externe qui sert de référence, est en fait un critère pour lequel on connaît une relation spécifique avec le critère biologique évalué. Ce qui est important dans notre approche, c'est de capturer les connaissances associées aux données pour optimiser la mise en relation avec le critère extérieur de référence, et non d'optimiser le clustering en soi.

Cependant, certaines approches d'évaluation de clustering reposent sur un principe similaire à notre stratégie. Le système implémenté par Bolshakova *et al.* (Bolshakova, et al., 2005a), nommé Machaon CVE[7], a été mis à jour afin d'utiliser l'ontologie GO comme source de connaissances sur les gènes, et ainsi, de calculer un indice de validité des groupes. Les groupes obtenus avec différentes méthodes de clustering peuvent être validés grâce à cet indice (Bolshakova, et al., 2005b). Par exemple, différentes partitions obtenues avec la méthode des k-means et avec plusieurs valeurs de k ont été évaluées. Celle qui optimise l'indice de validité est considérée comme la meilleure partition. Cette approche sert principalement à estimer le nombre optimal de cluster dans un jeu de données issu des puces à ADN ou autres (biomédicales, physiques...).

D'autres indices ont été proposés (Datta and Datta, 2006b) pour évaluer les algorithmes de clustering en utilisant également l'ontologie GO comme information de référence. Les indices calculés dans cette étude sont des indices de performance basés sur la correspondance des groupes obtenus et l'annotation des gènes dans l'ontologie GO (indices d'homogénéité et de cohérence biologique des clusters).

Les approches décrites ici s'appuient sur l'annotation fonctionnelle de l'ontologie GO pour évaluer les groupes obtenus avec les différentes méthodes. Elles reposent donc sur l'hypothèse forte que les gènes qui ont des profils d'expression similaires, donc qui appartiennent aux mêmes groupes, doivent avoir la même fonction ou des fonctions similaires, ce qui peut être discutable si ce n'est basé que sur une seule expérience. En effet, on sait que les résultats de puces à ADN contiennent du bruit, et que par conséquent, certains gènes retrouvés avec des profils d'expression proches peuvent être des faux positifs. Il faut donc être prudent face à ces résultats et vérifier que l'on n'a pas obtenu ce groupe par hasard.

[7] Machaon CVE [http://machaon.karanagai.com/]

2. Représentation des données de puces à ADN

Le clustering est la première étape de l'analyse de données de puces à ADN lorsqu'il s'agit d'expériences étudiant plusieurs conditions (voir Chapitre 1, partie 2.2). Ces expériences produisent pour chaque gène un profil d'expression. Il est nécessaire d'utiliser une méthode de clustering pour rassembler des gènes ayant des profils d'expression similaires dans un même groupe. Ainsi, on n'étudie plus un à un des milliers de gènes, mais des groupes des gènes ayant des profils d'expression communs.
Dans la littérature, on retrouve de nombreuses méthodes de clustering développées spécialement pour l'étude de ces données d'expression. Non seulement ces méthodes sont variées mais il existe également pour certaines d'entre elles, un ou plusieurs paramètres à configurer, initialiser. Tout ceci accroît la difficulté pour déterminer quel est le clustering adéquat qui va permettre d'extraire des groupes de gènes intéressants (qui ont un sens biologique) à partir des données brutes de puces à ADN.
Plusieurs articles ont montré que le clustering hiérarchique, pourtant largement utilisé dans l'analyse de données d'expression, n'est pas toujours le plus performant (D'Haeseleer, 2005; Datta and Datta, 2003; Huttenhower, et al., 2007). Nous avons également voulu l'évaluer mais dans le cadre de l'intégration de données.
Dans le but de confronter efficacement les données d'expression à d'autres données biologiques, nous avons évalué plusieurs représentations des données d'expression. Nous avons choisi de comparer la représentation, sous forme d'arbre binaire, issue du clustering hiérarchique à une autre représentation issue d'une nouvelle méthode que nous proposons pour identifier des groupes de gènes dont l'expression est co-régulée.

2.1 Données d'expression étudiées

Les données de puces à ADN utilisées pour cette étude ont été récupérées dans la base de données Stanford Microarray Database[8] (Gollub, et al., 2003). Je me suis servie des outils d'analyse disponibles sur ce site web pour récupérer les données sous la forme de taux de variation[9].

[8] Stanford Microarray Database [http://genome-www5.stanford.edu/]
[9] Stanford Microarray Database – *Data retrieval and Analysis* [http://smd.stanford.edu/cgi-bin/search/QuerySetup.pl]

Deux expériences de puces à ADN ont été sélectionnées pour ce travail sur la représentation des données d'expression :

- la première est celle de Spellman *et al.* (Spellman, et al., 1998) qui étudie les variations d'expression des gènes de la levure durant les différentes phases de la mitose[10]. Afin de mettre en évidence les gènes régulés par le cycle cellulaire, les cultures de cellules de levure ont été synchronisées par trois différentes méthodes et suivies au cours du temps, ce qui correspond à 77 conditions étudiées;

- la seconde est celle de Gasch *et al.* (Gasch, et al., 2000) qui s'intéresse aux modifications de l'expression des gènes de la levure en réponse à différents changements et stress environnementaux[11]. Dans les données brutes, il y a plus de 150 colonnes correspondant à une quinzaine de conditions testées à des temps différents ou avec diverses variations de température.

Nous avons choisi ces deux expériences pour cette étude car chacune d'elles considère des régulations génétiques dans des conditions biologiques bien différentes. Les résultats de ces deux études montrent que ce sont des groupes distincts de gènes co-régulés et divers processus biologiques qui sont mis en jeu dans ces expériences.

Les données brutes de ces expériences contiennent tous les gènes de la levure identifiés au moment où ont été effectuées ces expériences. Les gènes présents sur ces puces à ADN sont filtrés selon la liste des gènes fournie par le consortium Génolevures (Blandin, et al., 2000) (voir Chapitre 1, partie 1). Ce traitement permet de conserver uniquement les gènes des puces à ADN validés par Génolevures soit 5630 et 5649 gènes pour l'expérience de Spellman et de Gasch respectivement.

2.2 Description des deux représentations évaluées

Le clustering hiérarchique est l'une des méthodes les plus utilisées pour identifier des groupes de gènes co-exprimés dans une expérience de puce à ADN. Seulement, est-ce vraiment la représentation la plus pertinente lorsqu'on souhaite mettre en relation les données d'expression avec d'autres données biologiques ?

Pour répondre à cette question, nous avons utilisé la stratégie d'évaluation de plusieurs représentations en confrontant la représentation sous forme d'arbre binaire

[10] Informations supplémentaires sur l'expérience et les données de puce à ADN de Spellman [http://cellcycle-www.stanford.edu]

[11] Informations supplémentaires sur l'expérience et les données de puce à ADN de Gasch [http://genome-www.stanford.edu/yeast_stress/]

issu du Clustering Hiérarchique avec une nouvelle méthode que nous avons développée et que nous avons appelée la méthode des 'Best Neighbours'.

2.2.1 Clustering hiérarchique

La méthode hiérarchique agglomérative commence avec des groupes contenant un gène (autant de groupes que de gènes), et fusionne successivement les groupes les plus similaires jusqu'à n'obtenir qu'un seul « super-groupe » contenant tous les gènes. Le résultat produit un arbre binaire avec des groupes imbriqués de taille 1 à *N* (Figure 11). Un paramètre important à choisir est le type de distance ou de similarité que l'on va mesurer entre les gènes (Priness, et al., 2007). Il en existe plusieurs, et la sélection se fait en fonction des profils d'expression que l'on veut voir apparaître dans le même groupe (D'Haeseleer, 2005). Nous avons décidé d'utiliser le coefficient de corrélation de Pearson pour calculer la similarité entre les profils d'expression de chaque paire de gènes. Ce coefficient donne un score élevé à deux gènes dont les profils d'expression ont une forme globalement similaire, c'est-à-dire que l'expression de ces gènes varie dans le même sens en même temps ; il ne tient pas compte de l'intensité de l'expression[12] (Eisen, et al., 1998).

Pour chaque expérience de puce à ADN, les différentes étapes du clustering hiérarchique sont :

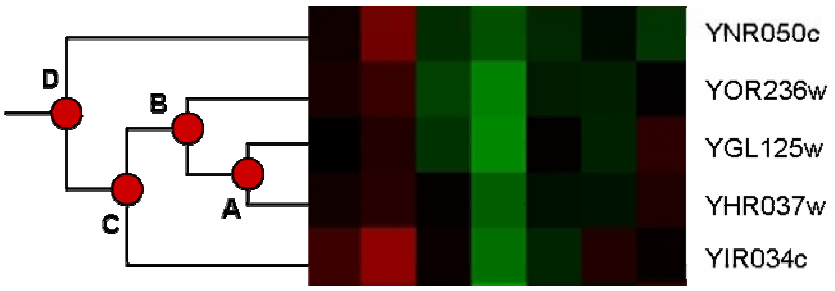
1) la construction de la matrice de similarité entre chaque paire de gènes de la puce (*N*x*N* valeurs) ; chaque gène, individuellement, devient un groupe;
2) la recherche des deux groupes dont la valeur de similarité est la plus élevée pour les fusionner en un nouveau groupe;
3) le calcul de la nouvelle matrice où on remplace les deux groupes fusionnés par la nouvelle valeur de similarité (calculée avec la méthode « average linkage » qui correspond à la moyenne des valeurs de similarité des deux groupes fusionnés);
4) l'itération des étapes 2) et 3) jusqu'à n'obtenir qu'un seul groupe de *N* gènes.

[12] Documentation du logiciel Cluster [http://rana.lbl.gov/manuals/ClusterTreeView.pdf]

Figure 11: Clustering hiérarchique

Figure montrant un exemple de résultat d'un clustering hiérarchique sur des données de puce à ADN. Chaque ligne représente un gène et chaque colonne correspond à une condition expérimentale. Cette méthode génère un arbre binaire dont chaque noeud (rond rouge) représente un groupe de gènes présentant un certain niveau de co-expression.

Composition des groupes : **A:** YGL125w, YHR037w; **B:** YGL125w, YHR037w, YOR236w; **C:** YGL125w, YHR037w, YOR236w, YIR034c; **D:** YGL125w, YHR037w, YOR236w, YIR034c, YNR050c.



Chaque nœud de l'arbre matérialisé par un rond rouge sur la Figure 11 devient un groupe. L'ensemble de ces groupes forme une collection représentant le critère biologique "expression des gènes issus d'une expérience de puce à ADN". Aucun seuil n'est appliqué pour stopper le clustering à un certain niveau de l'arbre ou à un certain nombre de groupes, donc tous les groupes du clustering hiérarchique sont présents dans la collection.

On obtient une collection par expérience; 'Gasch Hierarchical Clustering' et 'Spellman Hierarchical Clustering' contiennent respectivement 5648 et 5629 groupes.

2.2.2 Best Neighbours

Nous avons proposé une méthode alternative au Clustering Hiérarchique, c'est une méthode qui est simple et intuitive, que nous avons appelée la méthode des 'Best Neighbours' ou 'Meilleurs Voisins'. Elle permet de capturer un maximum d'informations en groupant les meilleurs voisins (gènes ayant les profils d'expression les plus proches) pour chacun des gènes présents sur la puce à ADN étudiée.

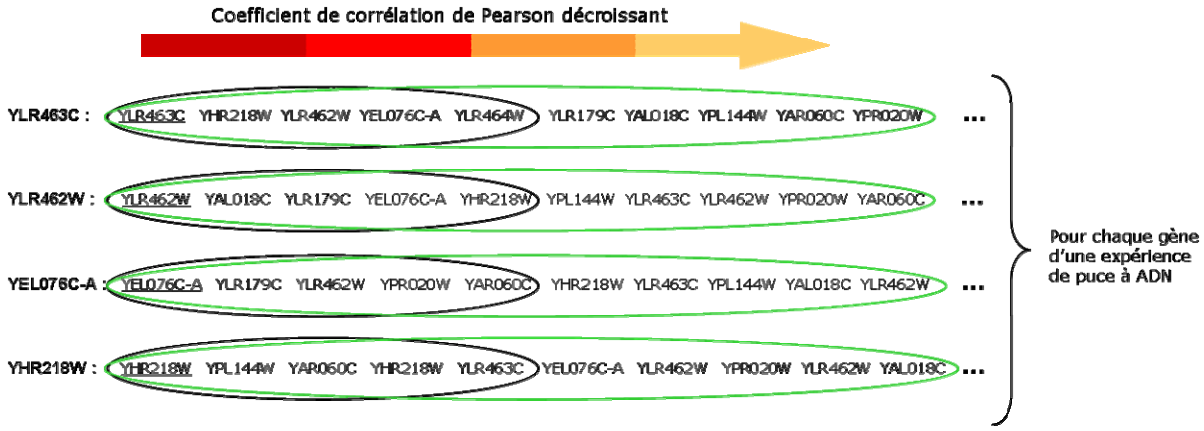
Cette méthode tout comme le clustering hiérarchique requiert une matrice de similarité de toutes les paires de gènes de la puce à ADN. Nous avons utilisé le même coefficient de corrélation que pour le Clustering Hiérarchique (Pearson) pour calculer la similarité entre les profils d'expression des gènes. Pour chaque gène, les autres

gènes présents sur la puce à ADN sont rangés dans l'ordre décroissant de leur coefficient de corrélation avec le gène de référence (Figure 12). Chaque gène, tour à tour, sert de gène de référence pour aller chercher les meilleurs voisins de celui-ci. Les meilleurs voisins sont les gènes ayant les profils d'expression les plus proches du gène de référence, c'est-à-dire les gènes qui ont un coefficient de corrélation élevé avec ce gène.

Figure 12: Principe de la méthode des 'Best Neighbours'

Pour chaque gène d'une expérience sur puces à ADN, chacun pris comme gène de référence (seulement 4 sont montrés ici), des groupes correspondant aux *N* meilleurs voisins (*best neighbours*) sont créés. Les meilleurs voisins sont les gènes ayant le coefficient de corrélation le plus élevé avec le gène de référence (gènes en début de ligne). Ce gène de référence est inclus dans tous les groupes quelle que soit leur taille (gène souligné).

Sur cette figure, deux exemples sont donnés: les gènes entourés en noir constituent des groupes de 5 meilleurs voisins; les gènes entourés en vert constituent des groupes de 10 meilleurs voisins. Il faut noter que les groupes de taille 10 incluent les groupes de taille 5.



Deux versions de cette méthode des 'Best Neighbours' ont été développées et ont été nommées 'Best Neighbours Single' (BNS) et 'Best Neighbours Nested' (BNN).

✓ La méthode des 'Best Neighbours Single' consiste, pour chaque gène d'une expérience de puces à ADN, à construire un groupe contenant un nombre fixe de meilleurs voisins. Comme nous n'avons aucun a priori sur le nombre de meilleurs voisins qui doit être inclu dans un groupe pour que l'on capture l'information pertinente, plusieurs tailles de groupes sont testées: 12 tailles différentes en commençant par les tailles 10, 20, 30, 40, 50, etc. jusqu'à 120. Nous avons donc créé 12 collections dont le nombre de groupes est égal au nombre de gènes dans

l'expérience mais dont la taille des groupes varie: la collection 'BNS 10' contient les groupes de taille 10 (groupes en bleu dans la figure 12), la collection 'BNS 20' contient les groupes de taille 20, et ce, jusqu'à la collection 'BNS 120'.

✓ La méthode des 'Best Neighbours Nested' consiste, pour chaque gène d'une expérience de puces à ADN, à construire une série de groupes contenant un nombre croissant de meilleurs voisins. Pour un gène de référence, on va avoir un ensemble de groupes imbriqués les uns dans les autres (sur le modèle des poupées russes). Pour cette méthode, 8 tailles maximales de groupes ont été testées:

- la collection 'BNN 50' contient tous les groupes de taille 5, 10, 15, 20, 25, 30, 35, 40, 45 et 50. Donc pour chaque gène, on va créer 10 groupes avec un nombre de meilleurs voisins croissants. Pour les expériences de Gasch et Spellman, cela correspond à des collections de 56490 et 56300 groupes, respectivement;
- la collection 'BNN 60' contient tous les groupes de la collection 'BNN 50' plus les groupes de taille 55 et 60. Donc pour chaque gène, on va créer 12 groupes avec un nombre de meilleurs voisins croissants. Pour les expériences de Gasch et Spellman, cela correspond à des collections de 67788 et 67560 groupes, respectivement;
- la collection 'BNN 70' contient tous les groupes de la collection 'BNN 60' plus les groupes de taille 65 et 70. Donc, pour chaque gène, on va créer 14 groupes avec un nombre de meilleurs voisins croissants. Pour les expériences de Gasch et Spellman, cela correspond à des collections de 79086 et 78820 groupes, respectivement.

La même démarche est utilisée pour construire les cinq collections restantes: 'BNN 80', 'BNN 90', 'BNN 100', 'BNN 110' et 'BNN 120'. Comme on peut le constater, contrairement à la méthode BNS, le nombre de groupes dans chacune de ces collections varie (Tableau 1). De plus, avec la méthode BNN, on crée beaucoup plus de groupes dans une même collection et de la redondance due au fait de l'enchevêtrement des groupes.

Tableau 1: Nombre de groupes créés pour chaque collection issue des données de puces à ADN

Pour chaque expérience de puces à ADN (deuxième et troisième colonne), ce tableau donne tout d'abord le nombre de gènes présents sur la puce et validés par Génolevures, puis le nombre de groupes générés pour chacune des méthodes.
BNS='Best Neighbours Single'; BNN= 'Best Neighbours Nested'.

Méthode de clustering	Spellman *et al.*	Gasch *et al.*
Nombre de gènes	5630	5649
Clustering Hiérarchique	5629	5648
BNS	5630	5649
BNN 50	56300	56490
BNN 60	67560	67788
BNN 70	78820	79086
BNN 80	90080	90384
BNN 90	101340	101682
BNN 100	112600	112980
BNN 110	123860	124278
BNN 120	135120	135576

2.3 Choix du critère biologique de référence

2.3.1 Définition

Les complexes multi-protéiques constituent un critère biologique approprié pour cette étude car il a été montré qu'il existe une correpondance entre les protéines appartenant à un complexe et l'expression des gènes codant ces protéines. De plus, leur représentation sous forme de groupes ne nécessite pas de traitement particulier.

Tout d'abord qu'entend-on par complexes multi-protéiques ? La plupart des protéines agissent en interaction avec d'autres protéines afin de remplir leur rôle dans la cellule. Un complexe protéique ou multi-protéique est un groupe de protéines qui fonctionne ensemble de façon permanente ou transitoire comme le protéasome et le ribosome par exemple (Jansen, et al., 2002). Il forme un complexe structural où toutes les protéines ne sont pas physiquement en interaction. Chaque protéine d'un complexe peut être définie comme une sous-unité de ce complexe.

Ces complexes correspondent à un niveau d'organisation des protéines qui est très important dans la cellule. Ils jouent un rôle essentiel dans de nombreux processus cellulaires. Leur analyse doit permettre de comprendre les mécanismes mis en jeu et l'organisation à différents niveaux de la cellule. Ils ont donc été largement étudiés et plusieurs méthodes de détection à large échelle de ces complexes ont été développées ces dernières années (purification+spectrométrie de masse, co-immunoprécipitation).
Dans ces études, la relation entre la co-expression de gènes et les complexes protéiques a été décrite à plusieurs reprises (Ge, et al., 2001; Jansen, et al., 2002; Simonis, et al., 2004). En effet, il a été démontré que les sous-unités d'un complexe tendent à être co-régulées, ou plus exactement les gènes codant ces sous-unités sont co-exprimés. Ce phénomène peut aisément s'expliquer de façon intuitive : un complexe multi-protéique est fonctionnel seulement si les différentes protéines qui le composent sont présentes simultanément dans la cellule, c'est-à-dire si les gènes codant ces protéines sont exprimés en même temps.

2.3.2 Source de données

Certains complexes ont été caractérisés individuellement par des méthodes biochimiques telle que la co-immunoprécipitation; d'autres ont été identifiés à grande échelle par des approches systématiques. La description de ces complexes est disponible dans des bases de données telles que MIPS-CYGD[13], YPD[14], SGD[15]. Les différentes études mettant en évidence une relation entre les complexes et la co-expression, ont exploité les complexes disponibles dans la base de données MIPS (Mewes, et al., 2004a) qui présente les complexes de façon simple et facilement exploitable. Nous avons donc travaillé avec les complexes décrits dans cette base de données afin de créer notre collection de référence.
Lorsque nous avons récupéré les données sur les complexes, il y en avait 1059 dans MIPS.

2.3.3 Représentation

Les données sur les complexes récupérées dans MIPS sont converties en une collection de groupes qui correspond au critère biologique de référence "Complexes Multi-protéiques".

[13] MIPS - CYGD [http://mips.gsf.de/genre/proj/yeast/]
[14] Yeast Proteome Databse [http://www.biobase-international.com/pages/index.php?id=ypd]
[15] Saccharomyces Genome Database [http://www.yeastgenome.org/]

Chaque groupe de cette collection correspond à un complexe c'est-à-dire à l'ensemble des protéines/sous-unités appartenant à ce complexe.

Une collection de complexes aléatoires est créée afin de s'assurer de la fiabilité de la relation mise en évidence avec la collection des complexes réels. Dans cette collection, la distribution du nombre de protéines dans les complexes est conservée. Chaque protéine qui fait partie d'un complexe réel est ré-attribuée aléatoirement à un groupe formant ainsi des complexes aléatoires.

2.4 Résultats et Discussion

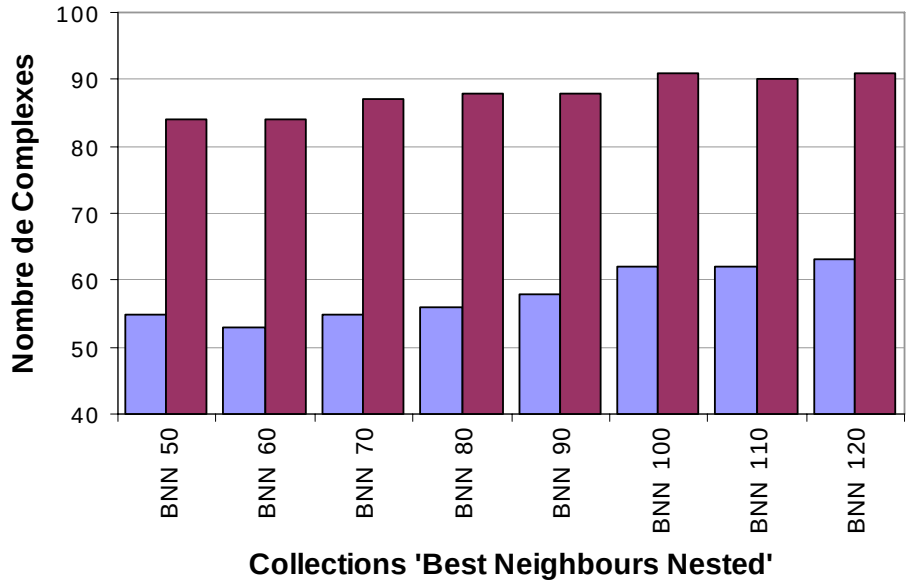
Nous avons utilisé la stratégie décrite au début de ce chapitre (partie 1) pour évaluer différentes représentations des données d'expression issues de puces à ADN. Cette stratégie consiste à comparer les différentes collections de groupes de gènes co-exprimés avec la collection de référence représentant les complexes multi-protéiques. Cette approche a été utilisée pour déterminer quelle représentation est la plus appropriée pour des données de puces à ADN dans le but de les mettre en relation avec d'autres données biologiques. Les différentes représentations de données d'expression ont été évaluées en examinant le nombre de complexes protéiques montrant une similarité significative avec au moins un groupe de chacune des collections. Ces complexes correspondent à ceux dont certaines sous-unités ont leurs gènes co-exprimés. La méthode la plus efficace est celle qui génère la collection qui permet de révéler au mieux la correspondance déjà connue entre les complexes et la co-expression.

La collection représentant les complexes est comparée (1) aux collections issues de la méthode BNN, (2) à la collection issue du Clustering Hiérarchique, (3) et aux collections issues de la méthode BNS, et cela, pour chacune des deux expériences de puces à ADN étudiée (Gasch et Spellman).

Dans un premier temps, nous avons analysé les résultats issus de la comparaison des complexes avec les groupes d'expression dérivant de la méthode BNN avec différentes tailles de groupes (Figure 13).

Figure 13: Résultats de la comparaison de la collection de complexes avec les collections 'Best Neighbours Nested'

Cet histogramme montre le nombre de complexes trouvés significativement similaires à au moins un groupe de gènes co-régulés appartenant à chacune des collections 'Best Neighbours Nested' (en bleu pour l'expérience de Spellman et en violet pour l'expérience de Gasch). Les noms des différentes collections indiqués sur l'axe des abscisses correspondent aux différentes tailles de groupes testées avec la méthode BNN.



Les résultats sont comparables pour les deux expériences (Gasch et Spellman): le nombre de complexes trouvés significativement similaires à un groupe d'expression issu de la méthode BNN est croissant lorsqu'on les compare à des groupes de tailles croissantes allant jusqu'à 100 (collection BNN 100). Avec les collections contenant des groupes de tailles supérieures à 100, on ne trouve plus de complexes supplémentaires significativement similaires à un groupe d'expression. Cette observation peut paraître surprenante car ces collections qui sont les plus larges contiennent les mêmes groupes que les plus petites collections (par exemple BNN 50, BNN 60) avec quelques groupes supplémentaires (voir 2.2.2). Ces collections les plus larges doivent donc refléter davantage les connaissances associées aux données d'expression. En regardant en détail les résultats, on se rend compte qu'on retrouve effectivement de nouveaux complexes qui sont similaires à des groupes d'expression supplémentaires des collections les plus larges. Mais, en même temps, on constate que certains complexes qui étaient trouvés similaires aux groupes des plus petites collections sont perdus avec les grandes collections. Cela s'explique par la correction statistique (Bonferroni) que nous avons utilisée qui permet d'éliminer les similarités qu'on pourrait trouver par chance à cause des comparaisons

multiples effectuées. Le nombre croissant de groupes dans les plus grandes collections (Tableau 1) conduit à une correction statistique plus stricte qui élimine certaines similarités qui étaient significatives avec de petites collections. Ceci est un premier inconvénient de l'utilisation de larges collections. Il en existe un deuxième, qui est le temps de calcul nécessaire pour effectuer toutes les comparaisons de groupes.
Parmi toutes ces collections, il apparaît que la collection BNN 100 est la meilleure pour mettre en évidence la correspondance existant entre les complexes et les groupes de gènes co-régulés, tout en gardant un temps de calcul raisonnable. C'est donc cette collection que nous avons retenue pour continuer l'étude, et plus particulièrement pour comparer les résultats du Clustering Hiérarchique.

Dans un deuxième temps, nous avons comparé la collection des complexes aux groupes issus du Clustering Hiérarchique. Le Tableau 2 présente le nombre de complexes qui ont été trouvés significativement similaires à au moins un groupe issu du Clustering Hiérarchique. Ces résultats sont comparés à ceux de BNN 100. On constate que pour les deux expériences de puces à ADN, la collection BNN 100 met en évidence plus de complexes similaires à au moins un de ses groupes d'expression que la collection issue du Clustering Hiérarchique. Une analyse plus approfondie de ces résultats montre que la plupart des complexes retrouvés similaires à un groupe issu du Clustering Hiérarchique sont ceux retrouvés avec BNN 100. De plus, parmi ces complexes, plusieurs ont déjà été présentés comme étant des régulons (Simonis, et al., 2004) validant la fiabilité de nos résultats, par exemple le complexe ribosomal cytoplasmique, l'ARN polymérase III, le complexe du nucléosome, le complexe du cytochrome bc1.

Tableau 2: Nombre de Complexes similaires aux groupes issus du Clustering Hiérarchique et de BNN 100

La première ligne de ce tableau donne le nombre de complexes ayant une similarité significative avec au moins un groupe d'expression (groupes issus du Clustering Hiérarchique ou groupes issus de BNN 100). La seconde ligne montre qu'aucun complexe généré aléatoirement n'est similaire à un groupe de gènes co-exprimés quelque soit la méthode utilisée.

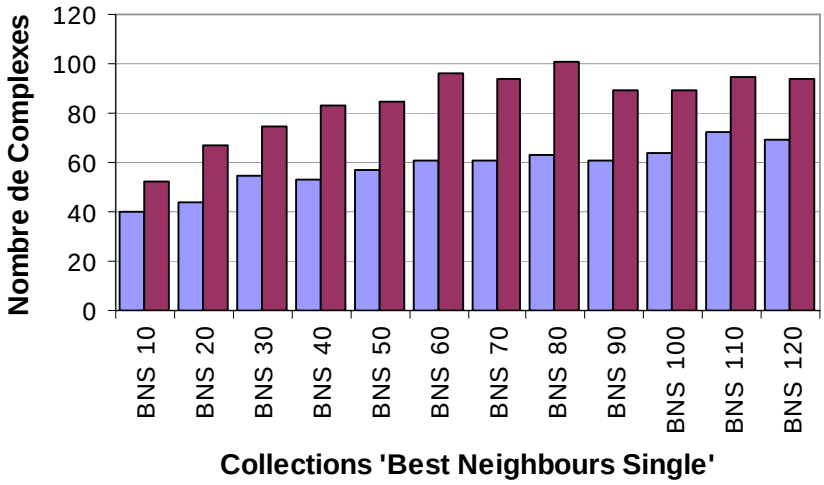
	Spellman experiment		**Gasch experiment**	
	Clustering Hiérarchique	**BNN 100**	**Clustering Hiérarchique**	**BNN 100**
MIPS Complexes (1059)	48	62	56	91
Complexes aléatoires (1059)	0	0	0	0

D'après ces premiers résultats, un nombre plus important de complexes sont trouvés corrélés avec un groupe d'expression lorsqu'on utilise une collection comportant plus de groupes: la collection BNN 100 contient 20 fois plus de groupes que la collection issue du Clustering Hiérarchique (Tableau 1). Cependant, la comparaison de la collection de complexes avec les collections issues du Clustering Hiérarchique, a l'avantage d'être moins lourde en calcul du fait du plus petit nombre de groupes dans ces collections.
Afin de vérifier si la meilleure efficacité de la collection BNN 100 n'est pas due à un effet de "taille de la collection", nous avons créé les collections 'Best Neighbours Single' (voir la partie 2.2.2 pour les détails). Ces collections contiennent le même nombre de groupes que celle du Clustering Hiérarchique (Tableau 1), et permettent comme les BNN d'évaluer différentes tailles de groupes de meilleurs voisins.

De nouveau, la collection des complexes est comparée aux groupes d'expression des différentes collections issues de la méthode 'Best Neighbours Single' (BNS). Les résultats montrent que si l'on utilise des groupes de taille supérieure à 30, on trouve plus de complexes significativement similaires à au moins un groupe d'expression issu de la méthode BNS qu'à un groupe issu du Clustering Hiérarchique (Figure 14). Par conséquent, même en utilisant des collections ayant le même nombre de groupes que le Clustering Hiérarchique, on trouve plus de complexes similaires à un groupe d'expression issu de la méthode 'Best Neighbours'. De plus, on constate que les résultats obtenus avec la collection BNS 100 sont similaires à ceux obtenus avec la collection BNN 100: 89 et 64 complexes avec BNS 100 pour les expériences de Gasch et Spellman respectivement, comparé à 91 et 62 pour BNN 100.

Figure 14: Résultats de la comparaison de la collection de complexes avec les collections 'Best Neighbours Single'

Cet histogramme montre le nombre de complexes trouvés significativement similaires à au moins un groupe de gènes co-régulés appartenant à chacune des collections 'Best Neighbours Single' (en bleu pour l'expérience de Spellman et en violet pour l'expérience de Gasch). Les noms des différentes collections indiqués sur l'axe des abscisses correspondent aux différentes tailles de groupes testées avec la méthode BNS.



Tous ces résultats montrent qu'une méthode simple comme celle des 'Best Neighbours Single' (un groupe de *N* gènes co-régulés pour chaque gène d'une expérience) pour représenter les données d'expression issues de puces à ADN, apparaît comme étant la plus adaptée lorsqu'on cherche à mettre en évidence la correspondance avec les complexes multi-protéiques. Le Clustering Hiérarchique bien que très utilisé pour les données d'expression ne semble pas être la représentation la plus efficace lorsque l'on veut mettre ces données en relation avec d'autres données biologiques.

Pour confirmer que les résultats obtenus ne sont pas dus au hasard, nous avons fait un contrôle qui consiste à comparer la collection de complexes protéiques aléatoires avec les différentes collections de données d'expression. Nous observons qu'aucun complexe généré aléatoirement n'a de similarité avec un groupe de gènes co-exprimés (Tableau 2), excepté pour les collections BNS 90 et 100 où l'on trouve 1 complexe aléatoire similaire à un groupe d'expression.
Ces résultats confirment que la relation mise en évidence entre les complexes et les groupes de gènes co-exprimés issus de la méthode des 'Best Neighbours' est fiable.

3. Représentation des données sur le point isoélectrique des protéines

Le point isoélectrique (pI) est une des propriétés physico-chimiques des protéines. Ces propriétés des protéines sont rarement exploitées car ce sont des données continues pour lesquelles il est difficile de définir une représentation. A ma connaissance, il n'a pas été mis en place de méthode permettant de représenter ces données sous forme de groupes. Nous avons donc imaginé plusieurs méthodes permettant de faire des groupes à partir de données continues telles que le pI, sachant que contrairement aux données d'expression où l'on a plusieurs valeurs pour un gène, pour le pI, on n'a qu'une seule valeur pour chaque protéine (c'est aussi vrai pour d'autres propriétés physico-chimiques telles la taille, le poids des protéines).
Le point isoélectrique d'une protéine correspond au pH pour lequel la charge électrique de la protéine est égale à zéro. En fonction du pH du compartiment cellulaire dans lequel se trouve la protéine, plusieurs de ses propriétés peuvent être modifiées (repliement, structure, solubilité, activité...) selon la valeur de son pI. Le point isoélectrique est une propriété physico-chimique très importante des protéines, qui peut être liée à leur fonction (Nandi, et al., 2005).

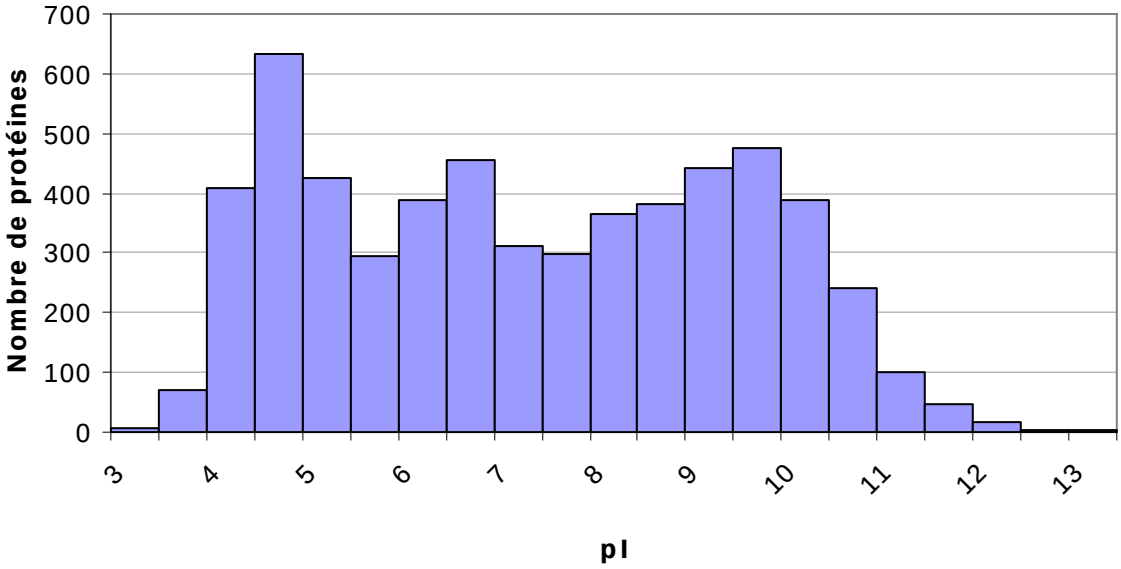
3.1 Données sur le point isoélectrique

Le pI théorique d'une protéine dénaturée (linéaire) peut facilement et précisément être calculé en utilisant les valeurs de pK des acides aminés portant une charge électrique (pH de demi dissociation des groupes ionisables) (Bjellqvist, et al., 1993). Plusieurs logiciels permettent de calculer les valeurs de pI des protéines à partir des séquences protéiques.
J'ai utilisé les séquences protéiques de *Saccharomyces cerevisiae* que nous avait fourni le Consortium Génolevures afin de calculer le point isoélectrique de chacune des protéines (5763 séquences protéiques). Ces calculs ont été effectués avec l'application « iep » du package EMBOSS[16] (Rice, et al., 2000).
La distribution des points isoélectriques du protéome de la levure est illustrée dans la Figure 15. Elle présente trois pics principaux qui ont déjà été mis en évidence (Schwartz, et al., 2001) et que l'on retrouve dans tous les protéomes, on parle de distribution multimodale.

[16] Informations sur EMBOSS [http://emboss.sourceforge.net/]

Figure 15: Distribution du nombre de protéines en fonction de leur pI

Histogramme de fréquence illustrant la distribution des valeurs des points isoélectriques du protéome de *Saccharomyces cerevisiae*. La distribution est montrée pour des gammes de pI de 0,5 et pour illustrer également la distribution dans les groupes de la collection 'pI_r0.5_flat'. On peut distinguer trois pics à 4,5 , 6,5 et 9,5.



3.2 Description des représentations évaluées

Les points isoélectriques des protéines correspondent à des nombres réels qui couvrent potentiellement une gamme continue de valeurs. Afin de construire des groupes de protéines dont les pIs sont proches, il est nécessaire de mettre en place des méthodes de regroupement. Ce sont ces méthodes que nous avons évaluées. Nous avons utilisé deux méthodes de regroupement:

- la première consiste à faire des groupes de taille fixe (*N* protéines ayant des valeurs de pIs adjacentes);
- la deuxième consiste à faire des groupes couvrant une gamme fixe de valeurs de pI (protéines dont les pIs vont de 0,5 à 1, de 1 à 1,5, etc).

Pour pouvoir créer ces groupes, les protéines de levure sont triées par valeurs de pI croissantes. Ces deux types de regroupement servent de base pour construire différentes collections ou représentations (4 collections au total dont l'organisation est expliquée au paragraphe suivant).

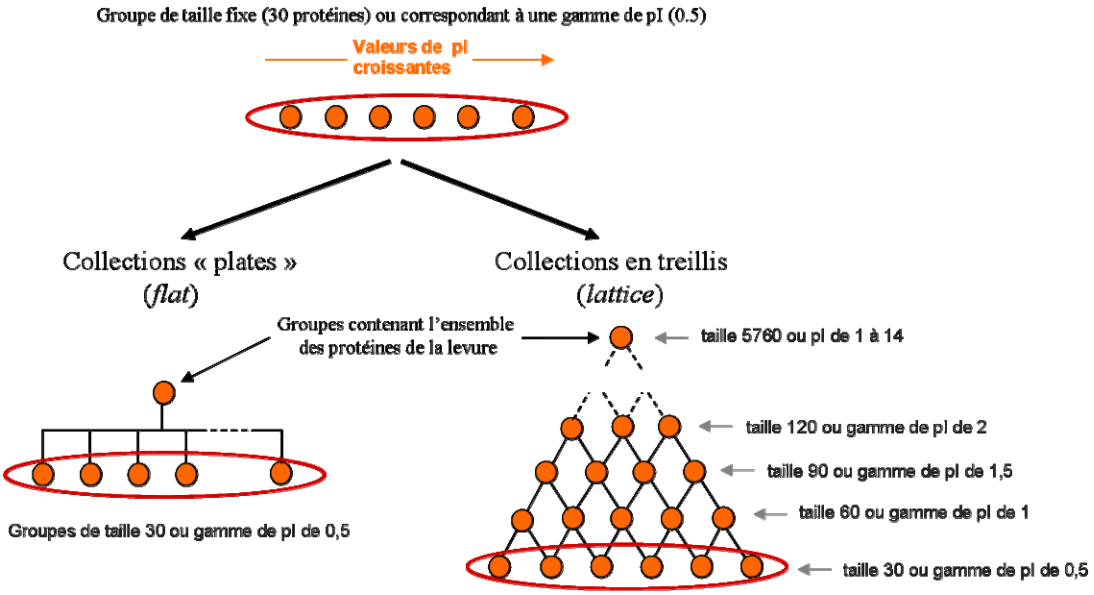
3.2.1 Représentations « plates »

Les collections construites en structure « plate » correspondent à des groupes indépendants les uns des autres (Figure 16). Pour construire ces collections, on utilise les deux types de regroupements définis ci-dessus :

- des groupes de taille fixe contenant 30 protéines ; la collection est notée 'pI_s30_flat';
- des groupes contenant les protéines appartenant à une gamme de pI de 0,5 ; la collection est notée 'pI_r0.5_flat'. Le nombre de protéines par groupe de gamme de pI de 0,5 est illustré dans la Figure 15.

Figure 16 : Représentations des points isoélectriques d'un protéome

Représentations des pIs: ce schéma montre les différentes façons de représenter des groupes de protéines ayant des points isoélectriques proches. En haut, les cercles oranges dans l'ovale rouge représentent des groupes de protéines avec des pIs voisins (cela peut aussi bien être des groupes de taille fixe que des groupes correspondant à une gamme de pI). Ces groupes peuvent être gardés tel quel (collections « plates » sur la gauche) ou ils peuvent être organisés en treillis (sur la droite).



3.2.2 Représentations en treillis

Les collections construites en treillis utilisent également les deux types de regroupement décrits précédemment, dont les groupes sont agrégés récursivement afin de créer un treillis (Figure 16). On obtient donc deux nouvelles collections qui ont été nommées 'pI_s30_lattice' et 'pI_r0.5_lattice'.

On obtient au final, 4 collections dont le nombre de groupes est donné dans le Tableau 3.

Tableau 3: Nombre de groupes créés pour chaque représentation des pI du protéome de levure

Ce tableau indique le nombre de groupes créés dans les 4 collections construites à partir des points isoélectriques.

	pI_s30_flat	pI_r0.5_flat	pI_s30_lattice	pI_r0.5_lattice
Nombre de groupes	195	27	18529	232

3.2.3 Collections aléatoires

A chaque protéine de levure, nous avons réassigné une des valeurs de pI calculées pour le protéome de la levure. Les protéines sont de nouveau ordonnées selon leur pI aléatoire. Pour chacune des 4 méthodes employées, nous avons construit une collection de pI en utilisant sur les pIs aléatoires. Les collections générées sont appelées 'random-pi-0.5_lattice', 'random-pi30_lattice', 'random-pi30_flat' et 'random-pi-0.5_flat'. Ces collections sont comparées à la collection de référence afin de s'assurer que les résultats obtenus avec les collections réelles sont fiables.

3.3 Choix du critère biologique de référence

3.3.1 Définition

Quelques études ont montré qu'il exite une relation entre la localisation sub-cellulaire des protéines et leur point isoélectrique. De plus, la localisation sub-cellulaire des protéines est un critère pour lequel la représentation ne demande pas de traitement des

données. La localisation sub-cellulaire des protéines remplit donc les caractéristiques pour servir de critère biologique de référence, et évaluer différentes représentations des points isoélectriques d'un protéome.

La localisation sub-cellulaire d'une protéine correspond au compartiment dans lequel la protéine a été identifiée dans la cellule. C'est une des principales caractéristiques d'une protéine, elle est importante pour comprendre le rôle de la protéine dans les processus cellulaires. Il existe une machinerie dans la cellule qui permet de trier et de transporter les protéines nouvellement synthétisées vers le compartiment auquel elles sont destinées: c'est le mécanisme d'adressage. Ceci est possible grâce à des peptides signaux se trouvant à l'extrémité des protéines, qui sont détachés de la protéine (protéines matures) lorsqu'elle arrive à sa « destination cellulaire ».

Avec le développement du séquençage de génome entier, on connaît la séquence de toutes les protéines d'un organisme mais pas toujours sa localisation dans la cellule. De nombreux logiciels de prédiction de la localisation cellulaire des protéines ont été développés. Ils se basent sur des informations relatives à la séquence des protéines telles que leur longueur, leur peptide signal, leurs domaines, leur point isoélectrique, etc (Drawid and Gerstein, 2000; Guda and Subramaniam, 2005; Horton, et al., 2007; Nakai and Horton, 1999). Le point isoélectrique est donc utilisé dans certains de ces outils de prédiction de localisation car il est une caractéristique importante des protéines (Drawid and Gerstein, 2000). De plus, une relation a été montrée entre le point isoélectrique d'une protéine et sa localisation sub-cellulaire. Schwartz et ses collègues ont montré que les trois pics observés dans la distribution des points isoélectriques (voir partie 3.1) correspondent à des compartiments cellulaires différents (le cytoplasme, le noyau et les membranes). Une autre étude menée par Wu et ses collègues (Wu, et al., 2006) nuance les conclusions de Schwartz en disant que ce n'est pas l'appartenance à un compartiment cellulaire qui explique la distribution trimodale des pI chez les eucaryotes, mais qu'effectivement il y a une relation entre point isoélectrique et compartiments cellulaires.

3.3.2 Source des données

Les données sur la localisation sub-cellulaire des protéines de levure sont disponibles dans plusieurs bases de données. Certains des outils bioinformatiques de prédiction cités précédemment mettent à disposition la localisation cellulaire prédite des protéines dans des bases de données (par exemple, Yeast Protein Localization

Server[17] développé par Drawid *et al.*). D'autres bases de données réunissent les données de leur laboratoire tel que YPL.db[18] (Habeler, et al., 2002), ou d'autres compilent les données extraites de la littérature et des données expérimentales comme MIPS SubCell (Mewes, et al., 2004b). C'est cette source de données, MIPS SubCell[19], que nous avons choisie puisqu'elle regroupe un grand nombre de données sur la localisation cellulaire des protéines de levure.
Pour créer la collection correspondant à la localisation cellulaire des protéines, nous avons utilisé les différents compartiments cellulaires décrits dans la base de données SubCell disponible en 2005 dans MIPS. Ces compartiments sont organisés hiérarchiquement les uns par rapport aux autres (voir Figure 6). Pour chaque compartiment cellulaire, la base de données fournit la liste des protéines détectées dans ce compartiment. La liste des 49 compartiments cellulaires de MIPS est complétée par un « compartiment » supplémentaire (Annexes 1 - Tableau 1.1): le complexe ribosomal cytoplasmique récupéré parmi les complexes protéiques cités dans le partie 2.3.2 de ce chapitre. Nous avons ajouté ce compartiment car il était absent de MIPS "*Protein Localization*" et on sait que les points isoélectriques des protéines appartenant au ribosome ont déjà été étudiés et ont montré une certaine conservation (Nandi, et al., 2005).

3.3.3 Représentation

Les données sur la localisation sub-cellulaire des protéines sont converties en une collection de groupes de protéines pour servir de critère biologique de référence.
Chaque compartiment cellulaire correspond à un groupe composé de protéines qui ont été détectées dans ce compartiment. L'ensemble de ces groupes forme la collection représentant le critère biologique "Localisation sub-cellulaire". La collection contient donc 50 compartiments sub-cellulaires organisés hiérarchiquement (voir Figure 6).

3.4 Résultats et Discussion

On observe une distribution trimodale des points isoélectriques du protéome de la levure (Figure 15). Cette observation a déjà été faite à plusieurs reprises : la distribution des pIs d'un protéome est multimodale. Il a même été montré en 2001 que, chez les eucaryotes, la distribution est trimodale alors qu'elle est bimodale chez

[17] Yeast Protein Localization Server [bioinfo.mbb.yale.edu/genome/localize/]
[18] Yeast Protein Localization database [http://ypl.uni-graz.at/pages/home.html]
[19] MIPS Protein Localization [http://mips.gsf.de/genre/proj/yeast/Search/Catalogs/catalog.jsp]

les procaryotes (Schwartz, et al., 2001). Ce qui a été réfuté plus tard en démontrant que la distribution des pIs des procaryotes pouvait également avoir ce troisième pic (Weiller, et al., 2004; Wu, et al., 2006). Plusieurs études ont tenté de trouver les raisons biologiques ou biochimiques de cette multimodalité: certains ont montré qu'elle était en relation avec la localisation cellulaire des protéines (Schwartz, et al., 2001), d'autres l'ont mise en relation avec la niche écologique des organismes (Knight, et al., 2004), d'autres encore ont observé cette même distribution dans des protéomes générés aléatoirement et concluent donc qu'elle est due à des propriétés (chimiques) des acides aminés (Weiller, et al., 2004; Wu, et al., 2006).
Le calcul des pIs des protéines de *Saccharomyces cerevisiae* nous a permis d'étudier leur distribution. Nous avons retrouvé les trois pics déjà observés, à environ 4,5 , 6,5 et 9,5 (Schwartz, et al., 2001).

La stratégie mise en place, décrite dans la partie 1 de ce chapitre, nous permet de comparer des collections de groupes de protéines représentant des critères biologiques différents. Cette stratégie est utilisée pour tester différentes représentations pour un critère donné en évaluant sa correspondance avec un critère de référence (critère pour lequel cette correspondance a déjà été mise en évidence; voir partie 3.3). Nous avons utilisé cette approche pour déterminer quelle représentation était la plus appropriée pour les pIs d'un protéome, dans le but de mettre en relation le pI des protéines avec d'autres critères biologiques. Afin d'évaluer les 4 collections représentant des protéines avec des pIs proches, nous avons utilisé comme référence la collection "Localisation sub-cellulaire", représentant la localisation cellulaire des protéines.
Les différentes représentations des pIs ont été évaluées en examinant deux caractéristiques : le nombre de compartiments sub-cellulaires trouvés significativement similaires à au moins un groupe de pIs proches, et la nature de ces compartiments. Ces compartiments correspondent à ceux dont une partie des protéines appartient à une gamme de pIs proches.

Les résultats, donnés dans le Tableau 4, montrent que chacune des 4 représentations permet d'identifier globalement les mêmes compartiments cellulaires correspondant à des groupes de protéines ayant des pIs proches. Pour chaque représentation de l'ensemble des pIs, le nombre de compartiments trouvés significativement similaires à au moins un groupe de pIs proches a été compté (Tableau 4) :

- seulement 3 compartiments avec 'pI_s30_flat';
- 8 compartiments avec 'pI_r0.5_flat' et 'pI_s30_lattice';
- 12 compartiments avec 'pI_r0.5_lattice'.

Tableau 4: Résultats de la comparaison de la collection des compartiments cellulaires avec les 4 collections de pIs

Ce tableau donne les compartiments cellulaires qui ont été trouvés significativement similaires à au moins un groupe de pIs d'une des 4 collections. Les croix indiquent avec quelle(s) collection(s) la correspondance, entre compartiment et pI, a été mise en évidence.

	pI_s30_flat	**pI_s30_lattice**	**pI_r0.5_flat**	**pI_r0.5_lattice**
Vacuole			x	x
Noyau		x	x	x
Matrice mitochondriale		x		x
Membrane interne de la mitochondrie	x	x	x	x
Mitochondrie		x	x	x
Extracellulaire		x	x	x
Protéines ribosomales cytoplasmiques	x	x	x	x
Cytoplasme		x	x	x
Paroi cellulaire	x	x	x	x
Nucléole				x
Cytosquelette				x
Membrane plasmique				x
TOTAL	3	8	8	12

C'est donc la méthode créant des groupes avec des gammes de pI de 0,5 organisés en treillis qui met en évidence la plus forte relation avec les compartiments cellulaires.
De plus, seulement cette représentation permet de mettre en évidence la relation entre un sous-compartiment du noyau (nucléole) et un groupe de pI, déjà décrite par Bickmore et ses collègues (Bickmore and Sutherland, 2002). Ils ont montré que de nombreuses protéines du nucléole ont un pI de 9-10, bien que les protéines de ce compartiment aient des pIs couvrant une large gamme de valeurs. Ceci confirme le résultat obtenu avec 'pI_r0.5_lattice' qui nous indique que le nucléole contient un enrichissement en protéines dont le pI est entre 9 et 10,5 (Tableau 5).

Avec la collection 'pI_r0.5_lattice', nous avons pu mettre en évidence des correspondances significatives, qui ont déjà été décrites dans la littérature, entre plusieurs compartiments cellulaires et des gammes de pI proches (Tableau 5).

Dans nos résultats, nous avons observé que les protéines ribosomales cytoplasmiques correspondent à des protéines dont le pI est entre 10,5 et 12,5. Nandi *et al.* (Nandi, et al., 2005) ont déjà montré que les pIs des protéines ribosomales sont très conservés et sont basiques (pI ~ 9-10). Dans notre étude, nous avons également pu retrouver certaines observations faites par Schwartz et ses collègues (Schwartz, et al., 2001):

- les protéines du cytoplasme ont des pIs plutôt acides, entre 4 et 7. Ceci a également été observé plus récemment par Kiraga *et al.* (Kiraga, et al., 2007) dans une étude plus générale basée sur 1784 protéomes.
- pour les protéines de la membrane plasmique et de la membrane interne de la mitochondrie nous avons pu observer des pIs plutôt basiques (environ 9) et cela a déjà été établi plusieurs fois (Kiraga, et al., 2007; Schwartz, et al., 2001). Cependant le groupe noté "*integral to membrane*" de SubCell, constitué de toutes les protéines membranaires non associées à une membrane en particulier, n'apparaît pas dans les résultats. Cela semble indiquer une disparité dans les pIs pour les protéines des différentes membranes de la cellule (cette disparité se voit déjà avec nos deux exemples de membranes: mitochondriale (9-10,5) et plasmique (7-9)).
- les protéines du noyau ont tendance à avoir des pIs acides: la gamme de pI que nous avons obtenue correspond à des pIs allant de 4 à 6,5. Cela a déjà été observé par Schwartz (Schwartz, et al., 2001) qui explique que le pic de pI à 6,5 dans la distribution des pIs d'un protéome eucaryote correspond aux protéines situées dans le noyau. Ce pic de pI à 6-6,5 pour les protéines nucléaires apparaît également dans l'étude faite par Kiraga et ses collègues (Kiraga, et al., 2007), mais ils ont montré qu'il y a un deuxième pic de pI à environ 9. En effet, comme on l'a vu précédemment avec le nucléole (pIs majoritairement entre 9 et 10,5), certains sous-compartiments du noyau rassemblent des protéines qui ont des pIs très différents de 6,5. On sait qu'il existe effectivement une hétérogénéité de pIs dans les différents sous-compartiments du noyau (Bickmore and Sutherland, 2002) et que la conclusion de Schwartz est à prendre avec précaution (elle a d'ailleurs été discutée par la suite). De plus, parmi les différents compartiments étudiés par Kiraga et ses collègues, plusieurs montrent des proportions importantes de protéines ayant un pI d'environ 6. Par conséquent, le pic de pI observé à 6,5 ne s'explique pas seulement avec une forte proportion de protéines du noyau mais par des protéines provenant de divers compartiments.

Kiraga et ses collègues ont analysé la distribution des pIs des protéines dans chaque compartiment cellulaire et ont montré que pour la plupart d'entre eux on retrouve un ou deux pics correspondant à un enrichissement de protéines dans des gammes de pIs particulières. Parmi ces compartiments, on en retrouve certains dans la liste de ceux que nous avons identifiés dans notre approche (Tableau 5) :

- la mitochondrie présente 2 pics aux environs des pIs 6 et 10, ce deuxième pic a une intensité plus important ; de notre côté, nous avons pu mettre en évidence un enrichissement significatif de protéines ayant des pIs entre 8,5 et 11,5 (deuxième pic de Kiraga) dans la mitochondrie, sa matrice et sa membrane interne;
- la vacuole a un pic un peu diffus autour de 5 ; notre approche a pu identifier de nombreuses protéines (111 sur les 277 de la vacuole) ayant des pIs entre 3,5 et 5,5;
- le cytosquelette montre un pic très dense à 5-5,5 et un autre à 9 ; notre étude a permis de déterminer un groupe de protéines (77 sur les 203 du cytosquelette) dont les pIs sont entre 4,5 et 6 correspondant au pic plus dense observé par Kiraga.

Tous ces résultats présentent des tendances qui convergent même s'ils ne sont pas complètement superposables. En effet, l'étude de Kiraga et ses collègues donne des résultats sur un ensemble de protéomes alors que notre travail ne porte que sur un seul organisme, la levure. Cependant, l'utilisation de la représentation la plus appropriée, 'pI_r0.5_lattice', a permis de retrouver des correspondances déjà connues entre pI et compartiments cellulaires, et d'en mettre en évidence de nouvelles, par exemple pour la paroi cellulaire et les protéines extracellulaires.

Tableau 5: Compartiments cellulaires dont les protéines appartiennent à une même gamme de pI

Ces compartiments ont été mis en évidence avec la représentation 'pI_r0.5_lattice'. Ce tableau montre les compartiments pour lesquels on trouve un enrichissement de protéines ayant des pI proches (les résultats donnés dans ce tableau ne montrent que la plus forte correspondance trouvée entre un compartiment et une gamme de pI).

Compartiments cellulaires	Nombre de protéines dans le compartiment	Gamme de pI	Nombre de protéines dans la gamme de pI	Intersection	P-value
Protéines ribosomales cytoplasmiques	119	10.5-12.5	411	89	10e-79
Cytoplasme	2798	4-7	2603	1514	2x10e-40
Mitochondrie	1019	8.5-11.5	2039	531	10e-33
membrane interne de la mitochondrie	139	9-10.5	1309	84	4x10e-22
matrice mitochondriale	69	10-11.5	742	27	3x10e-8
Paroi cellulaire	28	3.5-4.5	468	22	4x10e-15
Extracellulaire	51	3.5-5	1097	32	6x10e-12
Noyau	2114	4-6.5	2145	892	10e-9
nucleole	208	9-10.5	1309	81	6x10e-8
Vacuole	277	3.5-5.5	1530	111	3x10e-7
Membrane plasmique	183	7-9	1355	71	2x10e-6
Cytosquelette	203	4.5-6	1357	77	2x10e-6

Les correspondances observées individuellement entre un compartiment et une gamme de pI, et confirmées par diverses publications, permettent de penser qu'il existe une relation plus générale entre compartiments cellulaires et points isoélectriques. L'approche que nous avons développée permet de valider l'existence de cette relation, plus particulièrement avec la représentation 'pI_r0.5_lattice', en apportant une validation statistique. Cette représentation apparaît comme celle la plus appropriée pour représenter un ensemble de points isoélectriques d'un protéome, lorsque l'on souhaite confronter ces données avec la localisation cellulaire des protéines.

Afin de s'assurer que cette relation est bien fiable, nous avons comparé la collection "Localisation sub-cellulaire" avec chacune des collections aléatoires reprenant les 4 méthodes de représentations des points isoélectriques. Aucune similarité n'a été détectée entre un groupe représentant un compartiment cellulaire et un groupe aléatoire

de pIs quelle que soit la méthode utilisée. Ce résultat nous permet de confirmer que la relation mise en évidence entre localisation sub-cellulaire et les pIs des protéines d'un organisme, correspond à une réalité biologique.
Il est important de rappeler que l'étude a été effectuée avec des points isoélectriques théoriques. Ils ont été calculés à partir de séquences protéiques et des valeurs de pK des groupements ionisables. Il a été montré que ces calculs sont fiables (Bjellqvist, et al., 1993), et que le pI calculé d'une protéine repliée (ayant sa structure 3D) et le pI calculé d'une protéine non repliée (séquence linéaire) ont tendance à être similaires (Chan, et al., 2006). Par conséquent, les points isoélectriques calculés sont probablement peu différents des points isoélectriques réels (in-vivo) des protéines. Cette constatation nous permet d'avoir confiance dans nos résultats, même si l'on sait que les protéines in-vivo subissent d'autres modifications qui peuvent faire varier leur pI telles que les modifications post-traductionnelles.

4. Discussion

Dans ce premier travail, nous avons montré que le choix de la représentation des données biologiques est très important lorsque l'on souhaite mettre en relation des données, dans une perspective d'intégration et d'analyse. La représentation choisie peut avoir un impact fort sur l'efficacité de l'intégration. La stratégie que nous avons employée pour évaluer différentes représentations pour un critère biologique, et déterminer laquelle est la plus adaptée, peut être appliquée pour de nombreux critères biologiques. Avant de commencer une analyse de génomique fonctionnelle, il serait intéressant d'appliquer cette approche afin de s'assurer que la meilleure représentation (la plus appropriée) est utilisée et ainsi être capable de mettre en évidence des relations entre différents types d'informations biologiques, quelle que soit leur nature, et ce, de façon fiable.

Les résultats obtenus ont prouvé l'intérêt de la stratégie puisque, avec les données de puces à ADN, nous avons montré que selon la représentation choisie, on peut obtenir une nette différence dans l'intensité de la relation (33% et 60 % de complexes multi-protéiques supplémentaires identifiés dont les sous-unités sont co-régulées, pour les expériences de Spellman et Gasch respectivement). La meilleure représentation pour mettre en évidence cette relation est celle que nous avons proposée, 'Best Neighbours Single' avec des groupes de taille 100. Ce résultat est intéressant car cette

représentation rassemble simplement des gènes avec des profils d'expression similaires, alors que le Clustering Hiérarchique cherche à optimiser les gènes rassemblés dans un même groupe. Une méthode simple semble donc préférable à une méthode plus élaborée. L'efficacité de la méthode 'Best Neighbours Single' vient probablement du fait qu'un groupe de meilleurs voisins est créé pour chaque gène de l'expérience, ce qui n'est pas le cas du Clustering Hiérarchique. La méthode 'Best Neighbours Single' permet donc d'explorer les voisinages d'expression de chacun des gènes, tout en gardant un temps raisonnable de calcul.
De plus, en analysant plus précisément les résultats obtenus avec la méthode 'Best Neighbours Single', nous avons trouvé quelques complexes multi-protéiques qui n'avaient pas été mis en évidence avec le Clustering Hiérarchique, et qui pourtant ont été identifiés comme régulons (Simonis, et al., 2004): le complexe cdc28p et l'ARN polymérase II. Ceci constitue une preuve supplémentaire de la pertinence et de la fiabilité de notre méthode pour représenter les données d'expression.

Dans un deuxième temps, avec les points isoélectriques de levure, nous avons du mettre en place des méthodes nouvelles permettant de capturer les connaissances associées à ces données. La difficulté pour ces données vient du fait que ce sont des valeurs continues parmi lesquelles il n'est pas facile de définir des groupes. Notre approche nous a donné la possibilité de tester différentes méthodes et paramètres pour regrouper des protéines ayant des points isoélectriques proches. Ainsi, nous avons pu déterminer une représentation des points isoélectriques plus performante que les autres 'pI_r0.5_lattice'. Cette représentation a montré que la correspondance entre pI et localisation cellulaire des protéines est probablement plus générale que ce qui a été décrit jusqu'à présent.
A travers les résultats de cette étude sur les points isoélectriques, on se rend compte que la valeur du pI d'une protéine n'est pas suffisante pour pouvoir prédire la localisation cellulaire de celle-ci. Comme on a pu le voir (Tableau 5), plusieurs compartiments cellulaires contiennent des proportions importantes de protéines dans les mêmes gammes de pI (la vacuole et la paroi cellulaire: pIs entre 3 et 5; le nucléole et la mitochondrie: pIs entre 9 et 11). Néanmoins, cette propriété physico-chimique est exploitée par les logiciels de prédiction de localisation cellulaire, en combinaison avec d'autres propriétés des protéines. Drawid *et al.* (Drawid and Gerstein, 2000) ont montré que le point isoélectrique est un paramètre important pour la prédiction de la localisation des protéines (sixième paramètre le plus important après les peptides d'adressage).

Les méthodes considérées comme les meilleures (aussi bien pour les pIs que pour les données d'expression), nous permettent de capturer les connaissances associées aux données afin de rendre leur intégration plus performante. On remarque que ce n'est pas la collection (représentation) qui contient le plus grand nombre de groupes qui donne les meilleurs résultats; ceci est vrai aussi bien pour les données d'expression que pour les pIs. Ce paramètre « nombre de groupes dans une collection » est important car il est difficile de trouver quelle taille de collection est la plus appropriée: pour cela il faudrait savoir quelle fraction de groupes de la collection contient les connaissances pertinentes. Mais ceci n'est pas un paramètre que l'on peut prédire, et c'est notre approche, utilisant un critère de référence, qui nous a permis de sélectionner une collection plutôt qu'une autre.
Un autre paramètre important dans notre étude a été la correction statistique utilisée dans le cadre de comparaisons multiples. Nous avons choisi la correction de Bonferroni car elle est simple à calculer et elle s'est montrée efficace pour éliminer les faux positifs: aucune "fausse similarité" n'a été détectée avec des groupes aléatoires. Cette correction tient compte du nombre de groupes comparés, et donc dépend du nombre de groupes présents dans chacune des collections comparées. Elle rentre donc indirectement en jeu dans le choix de la collection la plus performante puisqu'elle va avoir tendance à pénaliser les grandes collections en diminuant le seuil de significativité. Une autre méthode de correction comme celle du False Discovery Rate (FDR), pourrait également être appliquée dans notre approche; ce point sera abordé dans la conclusion finale.

Le travail mené sur les données d'expression et les points isoélectriques peut être effectué pour n'importe quels critères biologiques dont la représentation sous forme de groupes nécessite le choix d'une méthode de clustering. C'est plus particulièrement le cas pour des données quantitatives, continues, telles que les propriétés physico-chimiques des protéines, la localisation chromosomique des gènes, etc. Ces données sont rarement exploitées dans les outils d'intégration de données; ceci est probablement du à la difficulté de les représenter.

5. Conclusion

Il y a de plus en plus de demandes en ce qui concerne l'intégration de données, tout particulièrement pour analyser les données issues de technologie à grande échelle. Cette intégration passe par une réflexion sur la façon de représenter toutes ces données très hétérogènes afin qu'on puisse les combiner. Les travaux présentés dans ce chapitre ont montré l'importance du choix de la représentation des données. A travers une approche permettant d'évaluer diverses représentations sous forme de groupes, nous avons mis en évidence que l'efficacité de l'intégration dépend de la représentation choisie, qui elle-même dépend de paramètres tels que le nombre de groupes dans la représentation et la correction statistique.
Notre approche offre un moyen de rationaliser le choix d'une représentation de données biologiques sous forme de collections de groupes. Cette approche peut être utilisée pour améliorer l'efficacité d'outils d'intégration de données de génomique fonctionnelle.

Annexe 2 - Article n°1: *"Clustering of genes and proteins, and optimising the integration of heterogeneous data"*.

CHAPITRE 3 : Représentation du métabolisme

Le métabolisme correspond à l'ensemble des réactions métaboliques qui peuvent avoir lieu dans la cellule. Toutes ces réactions forment un réseau complexe d'enzymes et de métabolites. Le réseau métabolique est donc une représentation du métabolisme sous forme de graphe dont les arêtes représentent les réactions (enzymes) et les nœuds, les métabolites. Aujourd'hui, la structure des réseaux métaboliques est assez bien connue pour plusieurs organismes modèles, et diverses données sur ces réseaux sont disponibles.

Le métabolisme intéresse beaucoup la communauté des biologistes car il définit des processus biologiques essentiels à la vie de la cellule. Il est donc important de pouvoir l'intégrer avec d'autres données biologiques pour comprendre le fonctionnement de la cellule. Par exemple, le métabolisme est largement utilisé par les biologistes pour analyser, interpréter les données d'expression issues des puces à ADN (Ghazalpour, et al., 2005; Goffard and Weiller, 2007; Li and Chan, 2004; Wei, et al., 2006; Yang, et al., 2005) afin de comprendre les processus mis en place par un organisme pour répondre à différents environnements/changements (stress, maladies, pollution, etc.). Cependant le réseau métabolique est volumineux et très connecté, ce qui le rend difficile à manipuler et à exploiter particulièrement dans un but intégratif.
Afin de faciliter leurs études, les chercheurs ont l'habitude d'exploiter le réseau métabolique sous forme d'un ensemble de voies métaboliques. Cette représentation est très utilisée et plusieurs sources mettent à disposition ce découpage en voies métaboliques pour divers organismes : KEGG[20] (Kanehisa, et al., 2006), Reactome[21] (Vastrik, et al., 2007), BioCyc[22] (Karp, et al., 2005), etc. Des outils bio-informatiques ont également été développés pour permettre la visualisation, grâce à un code couleur simple, des données biologiques dans le contexte des voies métaboliques:

- certains permettent de visualiser des gènes identifiés comme différentiellement exprimés dans une expérience ou les profils d'expression de gènes (Dahlquist, et al., 2002; Mlecnik, et al., 2005), on peut ainsi mettre en évidence les voies métaboliques impliquées dans les différentes conditions étudiées dans l'expérience;

[20] Voies métaboliques du KEGG [http://www.genome.ad.jp/kegg/pathway.html]
[21] Voies métaboliques de Reactome [http://www.reactome.org/]
[22] Base de données BioCyc [http://biocyc.org/]

- d'autres permettent de visualiser des données issues des différentes approches « omiques » (Arakawa, et al., 2005; Kanehisa, et al., 2006) afin d'aider à leur analyse (voir Figure 3b de l'Introduction, partie 3.1).

Cependant cette représentation du réseau métabolique en un ensemble de voies métaboliques correspond à un découpage arbitraire du réseau. En effet, les voies métaboliques ont été définies au fur et à mesure de leur découverte. Elles ne répondent donc pas à une méthode de décomposition précise mais plutôt à une décomposition intuitive du réseau (Gagneur, et al., 2003; Schilling, et al., 2000; Schuster, et al., 2000).

La communauté des mathématiciens s'est intéressée à ce problème de décomposition du réseau métabolique afin d'analyser sa structure et ses propriétés (robustesse, polyvalence, etc.). Ils ont proposé plusieurs méthodes basées sur la connectivité du réseau permettant ainsi d'analyser la structure du réseau métabolique global. Différents résultats ont montré qu'elle est modulaire (Ravasz, et al., 2002) et qu'elle peut être décomposée en sous-réseaux qui peuvent être organisés hiérarchiquement (Gagneur, et al., 2003; Ma, et al., 2004; Schilling, et al., 2000; Schuster, et al., 2000; Schuster, et al., 2002; Spirin, et al., 2006). Ces études s'appuient sur la topologie du réseau pour retrouver des modules à l'intérieur de celui-ci, mais il n'a pas été montré si ces modules correspondent ou non à des unités fonctionnelles dans la cellule.

Dans ce chapitre, nous nous sommes intéressés à la représentation du métabolisme de la levure dans le cadre d'une collaboration avec l'équipe de Minoru Kanehisa. Cette équipe fait partie du centre de bioinformatique du KEGG (Kyoto Encyclopedia of Genes and Genomes) à l'Université de Kyoto. Elle travaille sur une grande base de données de génomique incluant les cartes métaboliques pour divers organismes.
Suite au travail méthodologique présenté dans le chapitre précédent, j'ai eu l'opportunité de rencontrer Jean-Marc Schwartz, un des membres de l'équipe du KEGG. Son travail, qui se situe à l'interface entre les mathématiques et la biologie, porte sur l'analyse des réseaux métaboliques en utilisant le calcul de modes élémentaires (Schwartz and Kanehisa, 2005; Schwartz and Kanehisa, 2006). Un mode élémentaire est un ensemble minimum de réactions qui peut fonctionner à l'état stationnaire dans la cellule (Papin, et al., 2003; Schilling, et al., 2000; Schuster, et al., 2000). Afin d'approfondir son étude du métabolisme, cette équipe recherchait une approche qui lui permette de combiner les modes élémentaires avec des données

biologiques dans le but de classifier et d'annoter ces modes élémentaires. Il souhaitait plus particulièrement exploiter des données de puces à ADN.
Notre approche, basée sur la représentation de données sous forme de groupes, pouvait nous permettre de mettre en relation les modes élémentaires avec des données d'expression à l'échelle d'un génome entier. Ce travail devait nous permettre dans un premier temps de voir si l'approche que nous avions développée était applicable à un nouveau type de données, le métabolisme, et donc voir sa pertinence, sa fiabilité. Dans un deuxième temps, elle devait leur permettre de savoir si cette décomposition du réseau métabolique sous forme de modes élémentaires correspondait à une réalité biologique, et si cette représentation pouvait servir pour l'intégration du métabolisme avec diverses données « omiques ».

1. Représentation du réseau métabolique sous forme de modes élémentaires

1.1 Les modes élémentaires

1.1.1 Le calcul des modes élémentaires

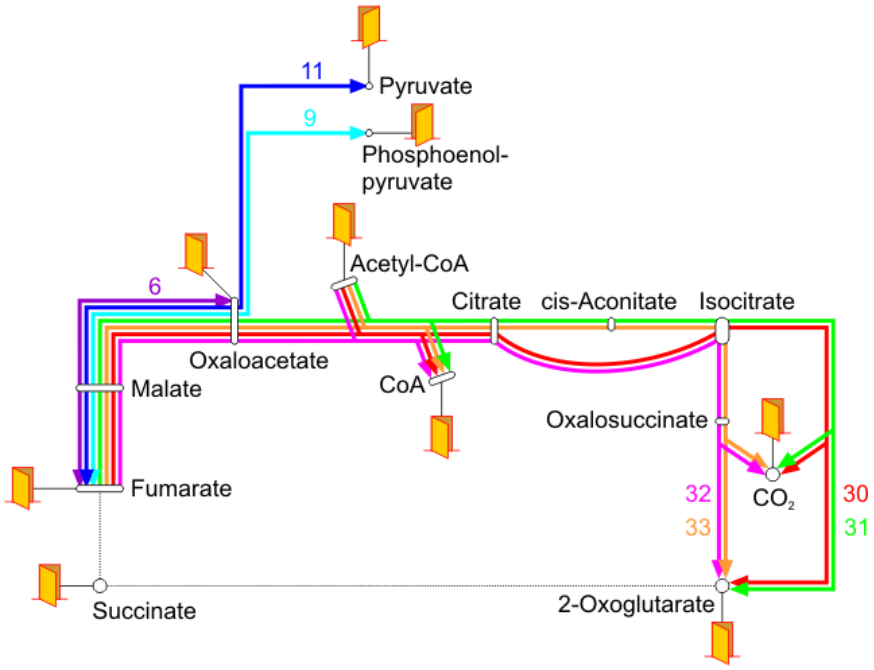
Parmi les méthodes proposées par les mathématiciens pour décomposer le réseau métabolique, il y a celles qui consistent à calculer les modes élémentaires ou *extreme pathways* (Papin, et al., 2003; Schilling, et al., 2000; Schuster, et al., 2000), qui sont des concepts très similaires (Klamt and Stelling, 2003). Pour calculer ces modes élémentaires, il faut commencer par identifier les métabolites externes qui sont ceux se trouvant aux extrémités du réseau. Plus précisément, ces métabolites participent à d'autres réactions qui ne font pas partie du réseau étudié. Par conséquent, ces métabolites n'ont pas une concentration stable dans le réseau puisqu'ils peuvent être consommés ou produits par d'autres réactions externes. Pour chaque paire de métabolites externes, on va calculer l'ensemble des chemins, réactions à suivre pour aller du premier métabolite au second, en se basant sur la topologie du réseau étudié. Chacun des chemins calculés est un chemin direct c'est-à-dire un ensemble minimal de réactions indispensables pour que le mode élémentaire opère à l'état d'équilibre.

Les modes élémentaires se sont montrés très utiles pour étudier plusieurs aspects du métabolisme. Ils permettent notamment d'organiser et d'analyser de façon systématique le réseau métabolique: en analysant l'impact possible d'enzymes déficientes (Schuster, et al., 2000), en étudiant l'importance et la pertinence des réactions ou la flexibilité et la robustesse des voies métaboliques (Gagneur and Klamt, 2004). Plus récemment, les modes élémentaires ont été utilisés pour décrire et comprendre les propriétés des réseaux de signalisation et de régulation de la transcription (Gianchandani, et al., 2006; Klamt, et al., 2006). Dans toutes ces applications, les modes élémentaires correspondent à des unités structurelles du réseau métabolique, et il n'a pas été montré que ces unités sont réellement fonctionnelles dans la cellule.

Le calcul des modes élémentaires à l'échelle de la cellule (du réseau entier), est freiné par le problème d'explosion combinatoire (Klamt and Stelling, 2002). Une approche alternative pour gérer ce problème, est de découper le réseau métabolique en sous-réseaux sur lesquels on va pouvoir calculer et interpréter les modes élémentaires. C'est cette approche que l'équipe du KEGG a proposée pour pouvoir calculer les modes élémentaires à l'échelle de la cellule.
Les voies métaboliques disponibles dans la base de données KEGG Pathway correspondent à des sous-réseaux du métabolisme représentés sous forme de cartes, et qui ont un sens biologique. Ces voies sont suffisamment petites pour qu'on puisse calculer les modes élémentaires pour chacune d'entre elles, et obtenir un nombre raisonnable de modes élémentaires à analyser. C'est pour ces raisons que les cartes métaboliques du KEGG ont été choisies comme sous-réseaux pour analyser le métabolisme de *Saccharomyces cerevisiae*. Ainsi, pour chaque carte métabolique, Jean-Marc Schwartz a calculé l'ensemble des modes élémentaires en adaptant l'algorithme classique développé par Stefan Schuster (Schuster, et al., 2000). Les détails sur l'algorithme utilisé sont disponibles dans l'article que nous avons publié (Schwartz, et al., 2007). Nous obtenons alors une représentation du métabolisme sous forme d'un ensemble de modes élémentaires.
Le calcul de modes élémentaires permet de calculer tous les chemins directs possibles entre des métabolites externes à l'intérieur de chaque voie métabolique. De ce fait, les modes élémentaires calculés dans une même voie peuvent avoir des réactions en commun. Certains modes élémentaires peuvent même être presque identiques (Figure 17).

Figure 17: Illustration de modes élémentaires calculés sur une carte métabolique

Ce schéma représente quelques-uns des modes élémentaires calculés entre le fumarate et le 2-oxoglutarate sur la carte métabolique du cycle du citrate. Chaque couleur correspond à un mode élémentaire différent auquel on a associé un chiffre (identifiant du mode élémentaire qui apparaît sur le dessin) ; les portes indiquent les métabolites servant d'entrée et de sortie (métabolites externes) pour le calcul des modes élémentaires. Cette figure illustre la nature combinatoire des modes élémentaires : plusieurs modes élémentaires sont presque identiques, et une même réaction peut appartenir à plusieurs modes élémentaires.



1.1.2 Représentation sous forme de groupes

Les voies métaboliques du KEGG

Tout d'abord, une façon simple et classique de représenter le réseau métabolique est celle correspondant à l'ensemble des voies métaboliques du KEGG. Chaque voie ou carte métabolique devient un groupe : chaque groupe est constitué de l'ensemble des enzymes impliquées dans une même voie métabolique. L'ensemble des groupes forme une collection nommée KEGG Pathways.

Les modes élémentaires

Le réseau métabolique est également défini par l'ensemble des modes élémentaires calculés sur chacune des voies métaboliques. A partir des modes élémentaires, une collection de groupes est créée : chaque groupe correspond à l'ensemble des enzymes impliquées dans un mode élémentaire. La collection contient donc autant de groupes

que de modes élémentaires calculés dans toutes les voies métaboliques du KEGG. Cette collection est nommée EM1.

Le réseau métabolique, ainsi converti en deux collections de groupes, peut être comparé aux données d'expression, en appliquant notre stratégie de comparaison de groupes (Chapitre 1, partie 3).

1.2 Méthode pour la validation biologique des modes élémentaires

1.2.1 Les données d'expression

Les organismes unicellulaires doivent être capables de s'adapter rapidement aux fluctuations qui peuvent avoir lieu dans leur environnement. Des expériences de puces à ADN sur la réponse aux stress permettent de mettre en évidence comment la levure adapte l'expression de ses gènes aux différents changements environnementaux qu'elle peut subir (Gasch and Werner-Washburne, 2002). Ces adaptations ont lieu à tous les niveaux de l'organisation cellulaire, et doivent donc également apparaître au niveau du métabolisme. Les gènes codant des enzymes impliquées dans une même voie métabolique ont plus de chances d'être co-régulés que des gènes pris au hasard (Yang, et al., 2004). A travers l'étude des gènes co-régulés lors de réponses aux stress, on s'attend à mettre en évidence des parties du réseau métabolique correspondant à la réponse métabolique de la cellule.

Nous avons choisi de travailler avec des expériences de puces à ADN étudiant la réponse aux stress et à différentes modifications environnementales chez la levure *Saccharomyces cerevisiae*.

Trois expériences ont été sélectionnées pour cette étude :

- celle de Causton *et al.* (Causton, et al., 2001) qui décrit la réponse transcriptionnelle à différents changements environnementaux chez la levure. Les données ont été récupérées sur le site du laboratoire de R.A. Young[23] ;

[23] Site web du laboratoire de R.A. Young [http://web.wi.mit.edu/young/environment/]

- celle de Gasch *et al.* (Gasch, et al., 2000) qui analyse l'expression des gènes de la levure durant l'adaptation aux stress. Les données ont été téléchargées à partir du site de Stanford MicroArray Database (SMD)[24] ;
- celle d'Iwahashi qui étudie la réponse transcriptionnelle de la levure à différents stress physiques et chimiques. Les données utilisées sont disponibles sur http://kasumi.nibh.jp/~iwahashi/.

Ces trois jeux de données nous ont permis d'étudier 35 conditions de stress au total. Certains de ces stress se rapportent à des changements environnementaux et des diminutions en substances nutritives, d'autres correspondent à des expositions à des composés toxiques tels que des pesticides ou fongicides.

Contrairement au travail présenté dans le deuxième chapitre, nous ne nous intéresserons pas aux profils d'expression des gènes dans chacune des trois expériences, mais à la variation d'expression de chacun des gènes pour chacune des conditions étudiées dans chacune des trois expériences. L'intensité de la variation d'un gène est donnée par le taux de variation de son expression. Par exemple, un taux de variation de 2 signifie que l'intensité de l'expression d'un gène a été multipliée par 2 (gène surexprimé); à l'inverse, un taux de variation de 0,5 signifie que son intensité d'expression a été divisée par 2 (gène sous-exprimé).
Pour chacun des trois jeux de données que nous utilisons, nous déterminons la distribution du logarithme des valeurs du taux de variation puis la déviation standard. La déviation standard de chaque expérience est multipliée par 2 afin d'obtenir une valeur seuil qui nous permet de sélectionner une proportion raisonnable de gènes surexprimés et sous-exprimés. Ce seuil correspond donc à la valeur du taux de variation de l'expression utilisée pour déterminer quels gènes sont considérés comme sur ou sous-exprimés dans chacune des conditions étudiées. Les gènes dont la valeur du taux de variation est supérieure au seuil sont considérés comme surexprimés, les gènes dont la valeur du taux de variation est inférieure au ratio 1 divisé par le seuil sont considérés comme sous-exprimés. Ainsi, pour chacune des 35 conditions étudiées, un groupe de gènes induits et un groupe de gènes réprimés sont définis.

1.2.2 Mise en relation du métabolisme et des données d'expression

On utilise les groupes de gènes induits et réprimés définis précédemment comme « groupes requêtes ». Chacun des groupes est mis en relation avec les deux collections

[24] Site web de SMD [http ://genome-www5.stanford.edu/]

métaboliques grâce à la méthode statistique de comparaison de groupes décrite dans le deuxième chapitre. Pour chaque groupe induit ou réprimé, on obtient une liste de voies métaboliques (collection KEGG Pathways) ou de modes élémentaires (collection EM1) trouvés significativement similaires au groupe requête. Cette liste représente des morceaux du réseau métabolique qui sont activés ou réprimés en réponse aux différents stress étudiés dans les expériences de puces à ADN.

Pour s'assurer que les résultats obtenus sont bien significatifs, nous avons créé des groupes de gènes aléatoires : pour plusieurs stress, les valeurs du taux de variation des gènes ont été permutées afin qu'une valeur aléatoire du taux de variation soit attribuée à chacun des gènes de l'expérience. On crée alors des groupes aléatoires de gènes surexprimés ou sous-exprimés. Ces groupes sont mis en relation avec chacune des collections sur le métabolisme afin de vérifier qu'on ne retrouve pas de correspondance entre ces groupes générés aléatoirement et des parties du métabolisme.

1.3 Résultats

1.3.1 Jeu de données d'expression analysé

Pour chacun des 35 stress étudiés, on crée deux groupes (induit et réprimé) : dans plusieurs conditions nous n'avons aucun gène considéré comme réprimé (Annexes 1 - Tableau 1.2). Au final, nous obtenons donc 58 groupes de gènes induits ou réprimés sur les 70 attendus. Parmi les 58 groupes, seulement 25 ont été trouvés significativement similaires à au moins un groupe d'une des collections métaboliques (Annexes 1 - Tableau 1.3). C'est avec ces 25 groupes de gènes induits ou réprimés, correspondant à 21 stress différents, que nous avons continué cette étude.

1.3.2 Validation biologique des modes élémentaires

Nous avons étudié la réponse aux stress chez la levure afin de savoir si les modes élémentaires sont des unités fonctionnelles du réseau métabolique et s'ils sont une représentation plus appropriée que les voies métaboliques globales pour mettre en évidence des réponses/adaptations métaboliques.

Dans un premier temps, nous avons pu mettre en évidence une correspondance entre les groupes de gènes induits ou réprimés et le métabolisme quelle que soit la représentation utilisée (EM1 ou KEGG Pathways). Cependant, nous avons constaté que la relation entre un de ces groupes d'expression et les modes élémentaires est, dans la plupart des cas, plus significative que celle avec les voies métaboliques. Ce résultat est illustré par quelques exemples dans le Tableau 6. Dans ce tableau, on peut voir que la valeur de la P-value entre un groupe de gènes induits ou réprimés est plus faible lorsque ce groupe est mis en relation avec les modes élémentaires; la relation est donc plus significative.

Tableau 6: Significativité des résultats obtenus avec les voies métaboliques du KEGG et les modes élémentaires (EM1)

Ce tableau montre des résultats obtenus avec BlastSets pour quelques stress. La deuxième colonne nous donne la voie métabolique du KEGG la plus similaire au groupe requête c'est-à-dire la voie métabolique avec la plus faible P-value (troisième colonne) dans la liste des résultats. Dans la quatrième colonne, on a le mode élémentaire le plus similaire au groupe requête c'est-à-dire celui avec la plus faible P-value (donnée en cinquième colonne).

La lettre entre parenthèses dans la première colonne correspond à l'expérience de puces à ADN dans laquelle le stress a été étudié : (I)=Iwahashi, (G)=Gasch, (C)=Causton.

Stress	Voie métabolique du KEGG la plus significative	*P*-value	Mode élémentaire (EM1) le plus significatif	*P*-value
Cendres industrielles (I), réprimé	sce00230 (Métabolisme des purines)	2.7e-8	sce00230.em279	1e-11
Pentanol (I), réprimé	sce00230 (Métabolisme des purines)	3.3e-6	sce00230.em341	1.8e-8
Tetrachloro-isophthalonitrile (I), réprimé	sce00230 (Métabolisme des purines)	2.5e-8	sce00230.em280	3.3e-10
Phase stationnaire (G), induit	sce00020 (Cycle du citrate)	3.4e-14	sce00020.em36	5.9e-16
Chaleur (C), induit	sce00500 (Métabolisme de l'amidon et du saccharose)	3.8e-4	sce00500.em13	4.2e-6

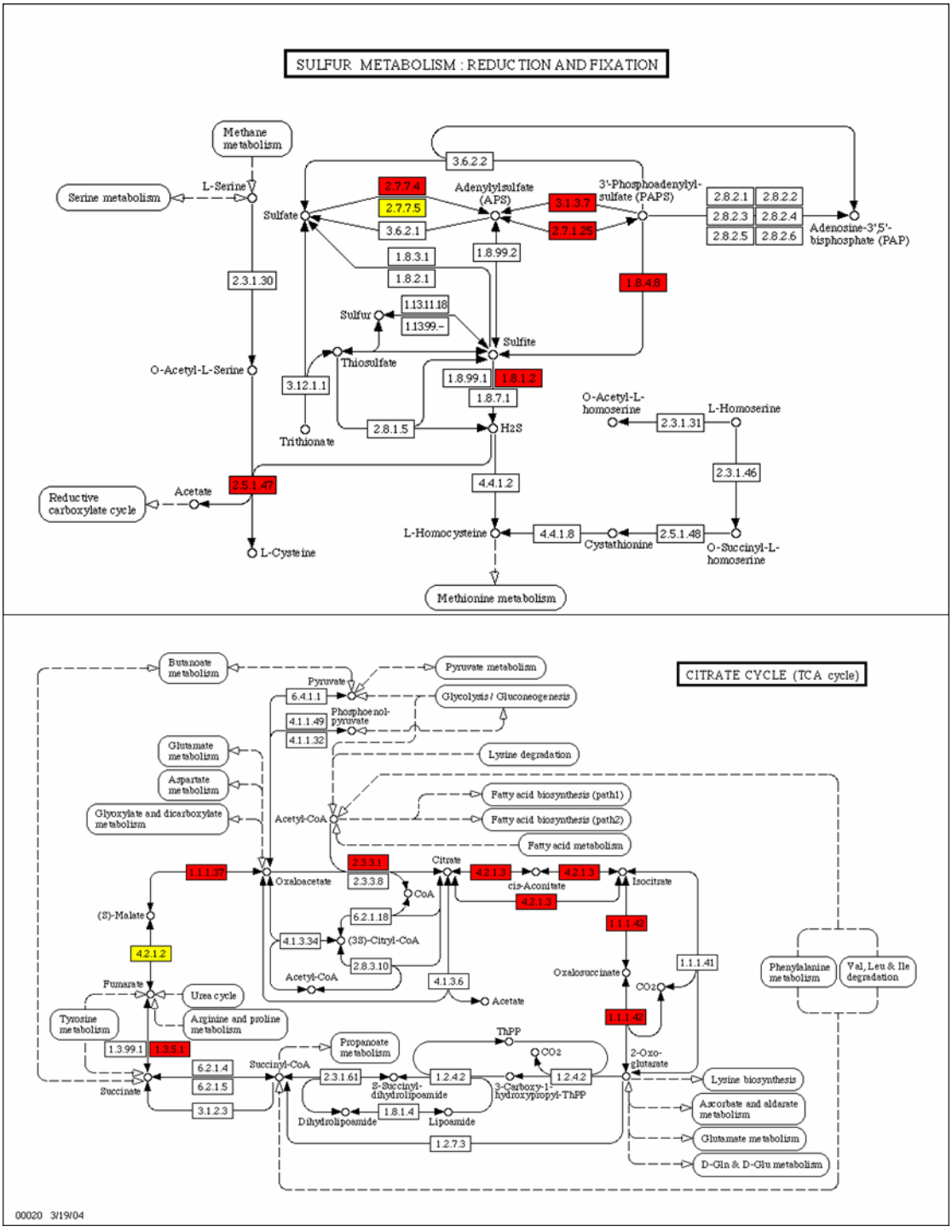
Par conséquent, les gènes trouvés surexprimés ou sous-exprimés lors d'une réponse à un stress ont une correspondance plus « forte »avec un mode élémentaire qu'avec une voie métabolique entière; ceci est illustré dans la Figure 18. Cela signifie que certains modes élémentaires capturent avec plus d'efficacité la réponse métabolique de la cellule lors d'un stress, et constituent effectivement des unités fonctionnelles du réseau métabolique.

Figure 18: Illustration de modes élémentaires induits en réponse à différents stress

Cette figure montre deux voies métaboliques du KEGG : le métabolisme du soufre et le cycle du citrate. Sur la première carte, représentant le métabolisme du soufre, les enzymes colorées en rouge correspondent à des enzymes codées par des gènes trouvés surexprimés lors de l'exposition au TPN (tetrachloro-isophthalonitrile, un fongicide). Ce chemin que dessinent les enzymes, correspond exactement à un des modes élémentaires calculés pour cette voie (mode élémentaire allant de l'acétate au sulfate), à l'exception de l'enzyme 2.7.7.5 qui a le même rôle que l'enzyme 2.7.7.4, qui elle est induite.

Sur la seconde carte, représentant le cycle du citrate, les enzymes colorées en rouge correspondent à des enzymes codées par des gènes trouvés surexprimés lors de la phase stationnaire chez la levure. L'ensemble de ces enzymes constitue un mode élémentaire allant du succinate au 2-oxoglutarate dans le cycle du citrate. Seule l'enzyme 4.2.1.2 manque (on a constaté qu'elle est également surexprimée mais qu'elle est juste en dessous du seuil que nous avons fixé pour considérer des gènes surexprimés ou non dans cette expérience).

Ces dessins illustrent le fait que les fonctions métaboliques mises en jeu lors de réponses aux stress chez la levure semblent correspondre à des modes élémentaires particuliers au sein des voies métaboliques entières.



Pour s'assurer de la significativité et de la fiabilité de ces premiers résultats, nous avons mis en relation des groupes de gènes induits ou réprimés générés aléatoirement avec les collections sur le métabolisme (voir partie 1.2.2). Aucune correspondance n'a été trouvée entre ces groupes et le métabolisme quelle que soit la façon dont il soit représenté. Ceci confirme la fiabilité des résultats mis en évidence précédemment.

Il faut noter que les chemins définis par les modes élémentaires à travers le réseau métabolique sont dépendants de la carte métabolique dans laquelle ils ont été calculés. Par conséquent, ils ne permettent pas de mettre en évidence un groupe d'enzymes régulés en réponse à un stress appartenant à plusieurs voies métaboliques. Nous nous sommes donc intéressés à ce problème qui fait l'objet de la partie suivante.

2. Combinaison des modes élémentaires

2.1 Méthode de combinaison des modes élémentaires

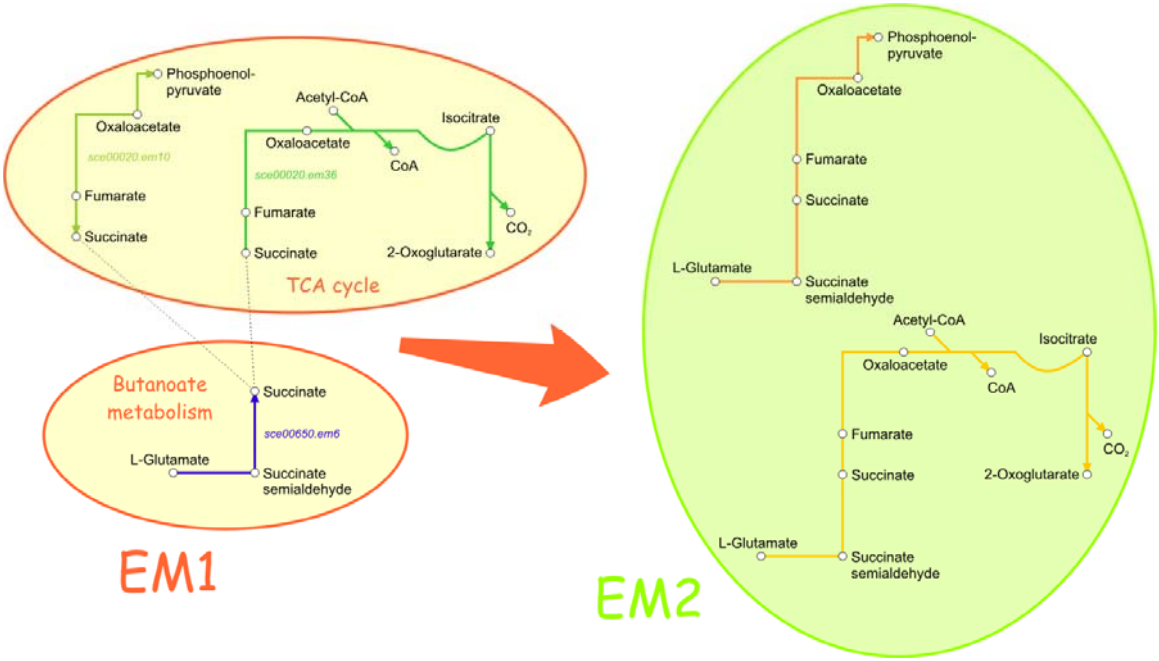
2.1.1 Assemblage de modes élémentaires par paires

La décomposition du métabolisme sous forme de cartes puis de modes élémentaires ne nous permet pas d'identifier des *routes métaboliques*[25] qui pourraient s'étendre sur plusieurs cartes. Afin de remédier à ce problème, nous avons décidé de combiner les modes élémentaires par paire. Nous avons alors construit une troisième collection : chaque groupe correspond à une paire de modes élémentaires qui appartiennent à deux voies métaboliques différentes, et qui sont connectés par un métabolite commun à une de leurs extrémités (Figure 19). Chaque groupe est alors constitué de l'ensemble des enzymes appartenant aux deux modes élémentaires connectés. Cette collection est appelée EM2.

[25] Une route métabolique est un ensemble de modes élémentaires traversant plusieurs voies/cartes métaboliques.

Figure 19: Construction de paires de modes élémentaires

Cette figure illustre comment sont formées les paires de modes élémentaires. A gauche (cercles oranges), sont représentés des modes élémentaires simples appartenant à deux voies métaboliques différentes (TCA cycle : deux modes élémentaires en vert ; Butanoate metabolism : un mode élémentaire en bleu). A droite (cercle vert), sont représentées les paires de modes élémentaires correspondant à la fusion des modes élémentaires de gauche. Ces modes élémentaires sont fusionnés car à leurs extrémités ils ont le même métabolite : le succinate. Cet assemblage est effectué de façon systématique pour toutes les paires de modes élémentaires ayant un métabolite externe commun et appartenant à deux voies métaboliques différentes.



2.1.2 Traitement des résultats

Nous utilisons les groupes de gènes induits et réprimés (partie 1.2.1), issus de chaque condition des trois expériences de puces à ADN, comme groupes requêtes. Chacun des groupes est mis en relation avec la collection de paires de modes élémentaires EM2. Pour chaque groupe induit ou réprimé, on obtient une liste de paires de modes élémentaires trouvés significativement similaires au groupe requête.

Afin d'exploiter davantage ces résultats, nous avons développé une application qui nous permet d'analyser les résultats obtenus avec les collections EM1 et EM2. Pour chacun des groupes requêtes, nous avons cherché à mettre en évidence des connexions entre les modes élémentaires (EM1) et les paires de modes élémentaires (EM2) trouvés similaires au groupe requête. Cette approche doit nous permettre d'identifier des routes métaboliques (ensemble de modes élémentaires connectés) traversant les

différentes voies métaboliques et correspondant à la réponse métabolique mise en place par la cellule pour répondre aux stress.
La démarche utilisée pour identifier ces routes est la suivante :

- la liste des modes élémentaires (simple, EM1, ou par paire, EM2) trouvés similaires au groupe requête est ordonnée par significativité décroissante (valeur croissante de P-value, voir Chapitre 1 partie 3) ;
- la paire ou le mode élémentaire ayant la plus forte similarité avec le groupe requête est récupéré(e), et appelé(e) *best hit* ;
- si le *best hit* est un mode élémentaire simple,
    - (1) on parcourt la liste ordonnée des hits jusqu'à trouver une paire de modes élémentaires contenant le *best hit*. Une fois que cette « meilleure paire » est identifiée,
    - (2) le reste de la liste est parcouru pour trouver la seconde meilleure paire de modes élémentaires connectée à la première meilleure paire (c'est-à-dire des paires ayant un mode élémentaire en commun avec cette meilleure paire). On forme ainsi une chaîne de paires de modes élémentaires,
    - (3) on continue à parcourir la liste pour trouver à nouveau une paire de modes élémentaires connectée à la chaîne que l'on a commencée à construire. Cette étape est répétée jusqu'à ce que toute la liste soit parcourue.
- sinon, si le *best hit* est une paire de modes élémentaires, on effectue seulement les étapes (2) et (3).

Ainsi, on obtient pour chaque groupe requête une chaîne de paires de modes élémentaires qui définit la colonne vertébrale de la réponse métabolique[26].

Pour s'assurer que les résultats obtenus avec EM2 sont bien significatifs, nous avons réutilisé nos groupes de gènes générés aléatoirement (partie 1.2.2) et nous les avons mis en relation avec la collection EM2.

[26] Remarque : parmi les paires de modes élémentaires qui sont ajoutés à cette chaîne, nous avons supprimé celles qui contenaient moins de trois enzymes pour être sûr qu'elles sont bien significatives, et pour éviter d'inclure des modes élémentaires trop courts qui ne seraient pas spécifiques d'une voie métabolique.

2.2 Résultats

2.2.1 Construction de routes métaboliques

Jusqu'à présent nous avons montré que :

- certains modes élémentaires constituent bien des unités fonctionnelles du métabolisme ;
- la représentation du réseau métabolique sous forme de modes élémentaires est très intéressante quand il s'agit de le combiner à d'autres données biologiques telles que l'expression des gènes ;
- les modes élémentaires permettent de mettre en évidence des chemins métaboliques correspondant à la réponse de la cellule lors de stress.

La collection EM2 est composée de paires de modes élémentaires connectés par un métabolite en commun à une de leurs extrémités. Ce métabolite joue le rôle de « pont » entre des voies métaboliques, nous permettant de définir des routes métaboliques plus étendues, traversant différentes cartes métaboliques.
Pour chaque stress, l'ensemble des paires de modes élémentaires trouvées significativement similaires au groupe requête et ayant les P-values les plus faibles, est utilisé pour construire une chaîne de modes élémentaires (voir partie 2.1.2). Cette approche nous permet de constituer des routes métaboliques induites ou réprimées à travers le réseau métabolique entier. Pour chaque stress, nous avons ainsi construit une route métabolique induite ou réprimée correspondant à la colonne vertébrale de la réponse au stress étudié ; elles sont illustrées dans la Figure 20.

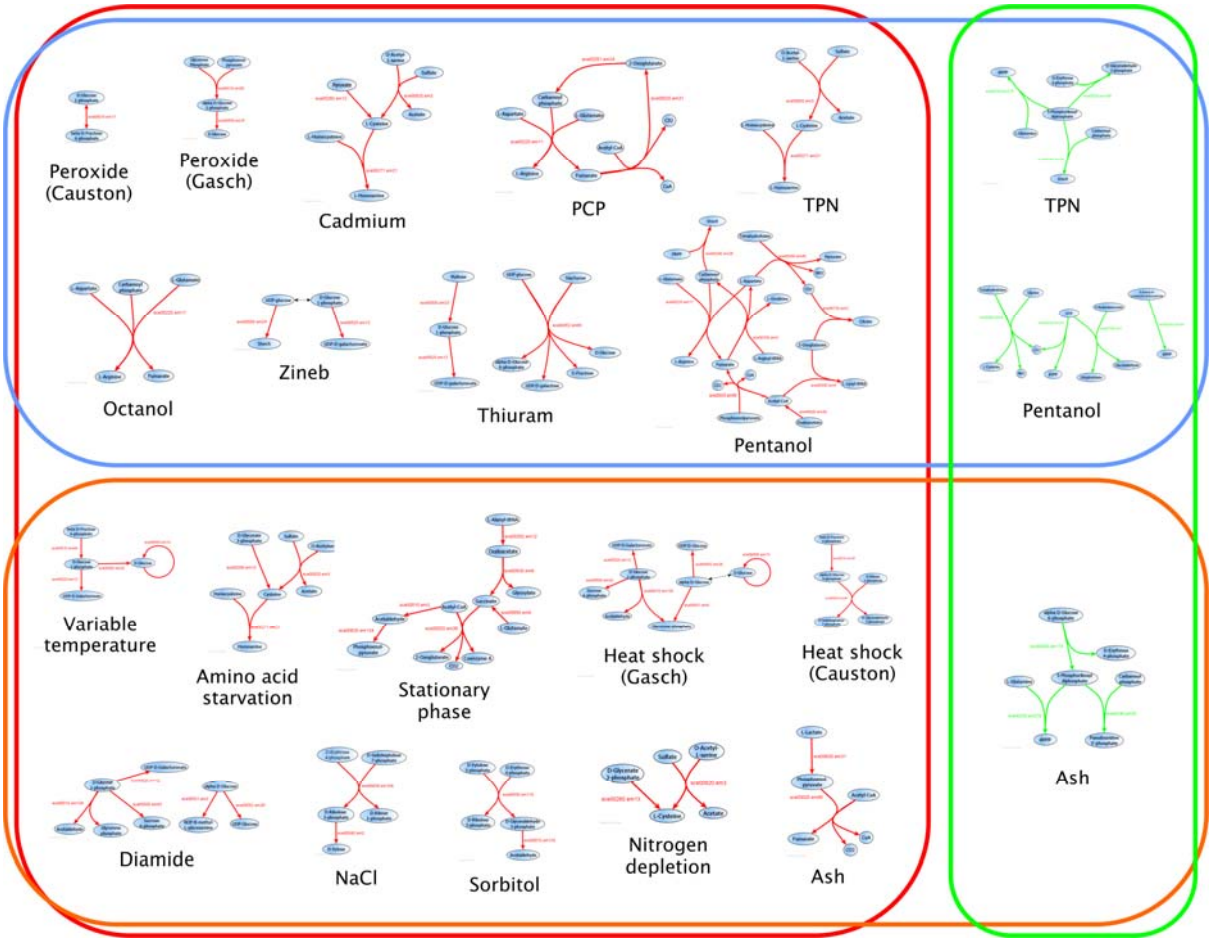
Figure 20: Routes métaboliques induites ou réprimées en réponse à différents stress

Cette figure illustre les routes métaboliques correspondant à la colonne vertébrale de la réponse métabolique à divers stress chez la levure, que l'on a pu identifier grâce à notre approche. Les routes incluses dans l'encadré rouge sont des routes induites, les routes dans l'encadré vert sont réprimées, celles dans l'encadré bleu sont des réponses à des conditions toxiques et celles dans l'encadré orange sont des réponses à des conditions non toxiques.

Les métabolites (ellipses bleues) correspondent à des métabolites d'entrée ou sortie de modes élémentaires dont certains permettent de faire des connexions entre modes élémentaires. Les modes élémentaires sont représentés par les flèches rouges ou vertes (modes élémentaires induits ou réprimés, respectivement).

TPN= Tetrachloro-isophthalonitrile; PCP= Pentachlorophenol

Variable température: variations de température; *Amino acid starvation*: carence en acides aminés; *Stationary phase*: phase stationnaire; *Heat shock*: chaleur; *Nitrogen depletion*: diminution en azote; *Ash*: exposition à des cendres industrielles.



La mise en relation des groupes de gènes aléatoirement générés avec les groupes de la collection EM2 ne donne aucun résultat significatif, confirmant la fiabilité des résultats exposés dans ce paragraphe.

2.2.2 Validation des routes métaboliques mises en évidence

Pour certains stress, les réponses métaboliques que l'on a observées (Figure 20) à partir des données d'expression confirment des observations faites précédemment à partir d'approches expérimentales. Ces réponses semblent donc correspondre à des fonctions de la cellule mises en place pour s'adapter à différentes conditions.

Vido *et al.* (Vido, et al., 2001) ont montré que l'exposition au cadmium augmente la synthèse de cystéine et peut-être de glutathione, ce qui est essentiel pour la détoxification cellulaire. La synthèse de ces deux composés est possible par l'activation de la voie des acides aminés soufrés. C'est ce que nous observons dans la route métabolique que nous avons mise en évidence : parmi les trois modes élémentaires induits lors de la réponse au cadmium, deux ont la cystéine comme produit final. Parmi ces deux modes élémentaires, un appartient au métabolisme de la cystéine, l'autre au métabolisme du soufre.
La carence en acides aminés dans le milieu est connue pour activer le facteur de transcription Gcn4p qui active la transcription des gènes impliqués dans des voies de biosynthèse des acides aminés, à l'exception de la voie de la cystéine bien que les gènes impliqués dans la biosynthèse des précurseurs de la cystéine soient eux aussi activés (Natarajan, et al., 2001). La réponse métabolique que nous observons lors d'une carence en acides aminés est constituée de plusieurs modes élémentaires provenant de voies de biosynthèse des acides aminés mais d'aucun de celle de la cystéine.
Les gènes surexprimés dans des cultures en phase stationnaire de levures sont associés à des fonctions mitochondriales telles que la respiration aérobie et le cycle du citrate (Martinez, et al., 2004). Par conséquent, la synthèse d'ATP est importante pour les levures en phase stationnaire. La chaîne de modes élémentaires que nous avons trouvée induite en phase stationnaire est constituée de modes élémentaires provenant de la respiration aérobie (glycolyse), du cycle du citrate, du métabolisme du pyruvate et de la phosphorylation oxydative.

Ces quelques exemples de résultats que nous avons obtenus, et confirmés par des publications précédentes, nous montrent que nous pouvons avoir confiance dans nos résultats et que l'approche développée est pertinente. On peut donc penser que les routes métaboliques identifiées (induites ou réprimés) par les autres stress reflètent elles aussi une réalité biologique.

Ce travail apporte de nouvelles perspectives pour étudier le métabolisme en proposant une nouvelle décomposition du réseau métabolique et en permettant de découvrir à l'intérieur de ce réseau des routes particulières correspondant à une fonction métabolique induite ou réprimée en réponse à telle ou telle condition.
Le découpage du réseau métabolique sous forme de paires de modes élémentaires est une représentation qui peut être utilisée pour rapprocher le métabolisme d'autres données biologiques dans un but d'intégration.

2.2.3 Caractérisation de l'activité des modes élémentaires

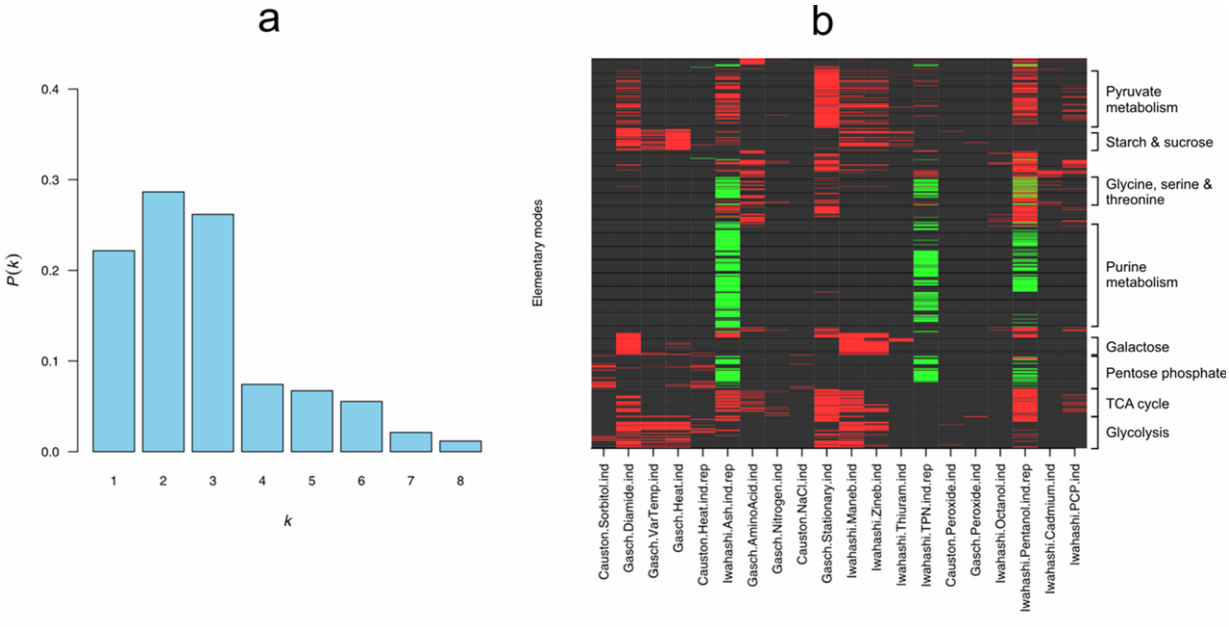
Des modes élémentaires spécialisés, d'autres «multi-tâches »
Dans le but d'avoir une vision globale de « l'activité » des modes élémentaires en réponse à divers stress, nous avons calculé la probabilité *P(k)* d'observer un mode élémentaire activé (qu'il soit induit ou réprimé) dans *k* conditions (Figure 21a). On observe deux situations bien distinctes : d'un côté, une proportion élevée de modes élémentaires activés en réponse à quelques stress particuliers (modes élémentaires spécialisés), et de l'autre une faible proportion de modes élémentaires activés en réponse à de nombreux stress (modes élémentaires multi-tâches). Environ 75% des modes élémentaires sont spécialisés, c'est-à-dire qu'ils ne sont activés en réponse qu'à un, deux ou trois stress différents, tandis que les 25% restants sont des modes élémentaires multi-tâches qui sont impliqués dans la réponse générale au stress chez la levure. Cette distribution ne correspond pas à une distribution aléatoire et révèle une organisation plus complexe des modes élémentaires impliqués dans la réponse au stress, en formant des modules à travers le réseau métabolique.

Nous avons également construit une carte représentant l'implication des modes élémentaires dans les différentes réponses aux stress. Nous avons choisi une représentation qui reprend celle des puces à ADN, où chaque ligne correspond à un mode élémentaire, et chaque colonne, à un stress. Les modes élémentaires induits sont en rouge et les réprimés en vert. Cette représentation basée sur l'activité transcriptionnelle des gènes permet d'avoir une vision globale de l'activité des modes élémentaires à l'échelle de la cellule dans différentes conditions de stress. Sur cette carte (Figure 21b), on constate que la plupart des modes élémentaires impliqués dans une ou plusieurs réponses aux stress sont soit induits soit réprimés; il est rare qu'un même mode élémentaire trouvé induit dans un stress soit trouvé réprimé dans un autre stress, et vice versa. De plus, on peut voir que les réponses réprimées ont des profils

similaires (le même ensemble de modes élémentaires est réprimé). Au contraire, les réponses induites ont des profils très divers et on retrouve peu de modes élémentaires induits à travers toutes les conditions, ce qui confirme la tendance observée dans la Figure 21a.

Figure 21: Activité des modes élémentaires

a) L'histogramme présente la probabilité de retrouver un mode élémentaire activé ou réprimé dans *k* conditions de stress. b) Représentation à l'échelle de la cellule de l'activité des modes élémentaires. Chaque ligne correspond à un mode élémentaire, et chaque colonne correspond à un stress. En rouge, on a les modes élémentaires induits, en vert les modes élémentaires réprimés.



Deux classes de réponses

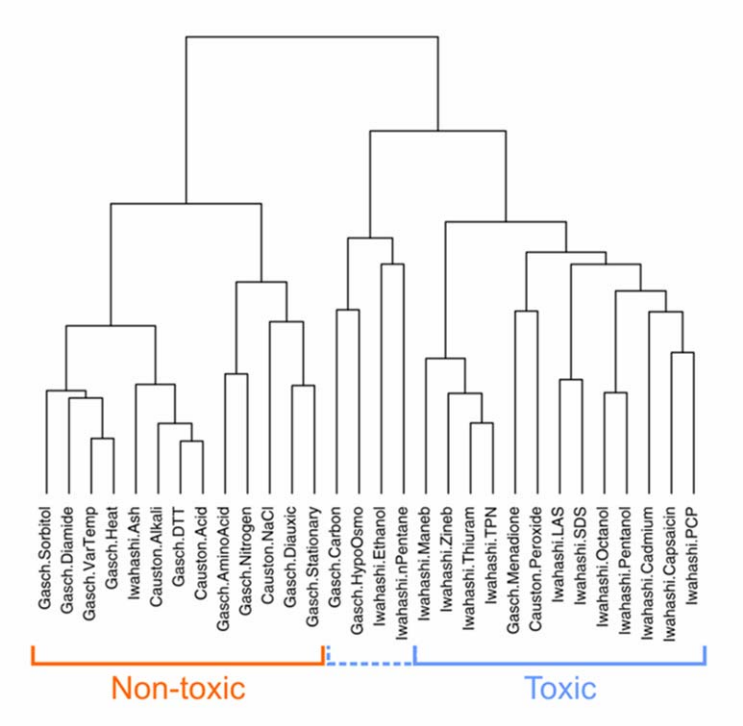
Un travail préliminaire effectué avant de mettre en œuvre l'étude avec les modes élémentaires, nous avait permis de mettre en évidence deux classes de stress. Ce travail avait été mené sur les données d'expression brutes de 30 stress (parmi les 35 stress utilisés dans notre étude sur les modes élémentaires, voir partie 1.3.1) afin de voir quels stress conduisaient à des réponses transcriptionnelles similaires (Schwartz, et al., non publié)[27]. Pour cela, la corrélation entre les taux de variation des gènes entre toutes les paires de stress a été calculée, générant une matrice. A partir de cette matrice, un clustering hiérarchique a été effectué afin de retrouver les stress les plus

[27] Résumé long du poster disponible à cette adresse [http://ismb2006.cbi.cnptia.embrapa.br/poster_abstract.php?id=J-19]

similaires (Figure 22). Ce clustering nous permet d'identifier deux classes de stress : toxique et non toxique.

Figure 22: Identification des deux classes de stress

Cet arbre présente le clustering des différents stress étudiés, basé sur les données brutes d'expression pour chacun des stress. Ce clustering a été effectué à partir de mesure de corrélation entre les taux de variation des gènes dans les différents stress.



La classe *toxique* correspond à des stress tels que l'exposition des cellules à des produits chimiques ou métaux toxiques. La classe *non toxique* inclut des stress tels que des variations de température, des chocs osmotiques ou des modifications dans les substances nutritives apportées aux cellules. La liste des stress contribuant à chacune des classes, et pour lesquels nous avons pu définir une réponse métabolique (19 stress), est donnée dans le Tableau 7.

Tableau 7: Répartition des différents stress dans les deux classes

Ce tableau présente les stress appartenant à chacune des deux classes de stress: toxique et non toxique. Les lettres entre parenthèse correspondent à l'expérience de puces à ADN dans laquelle le stress a été étudié : (I)=Iwahashi, (G)=Gasch, (C)=Causton.

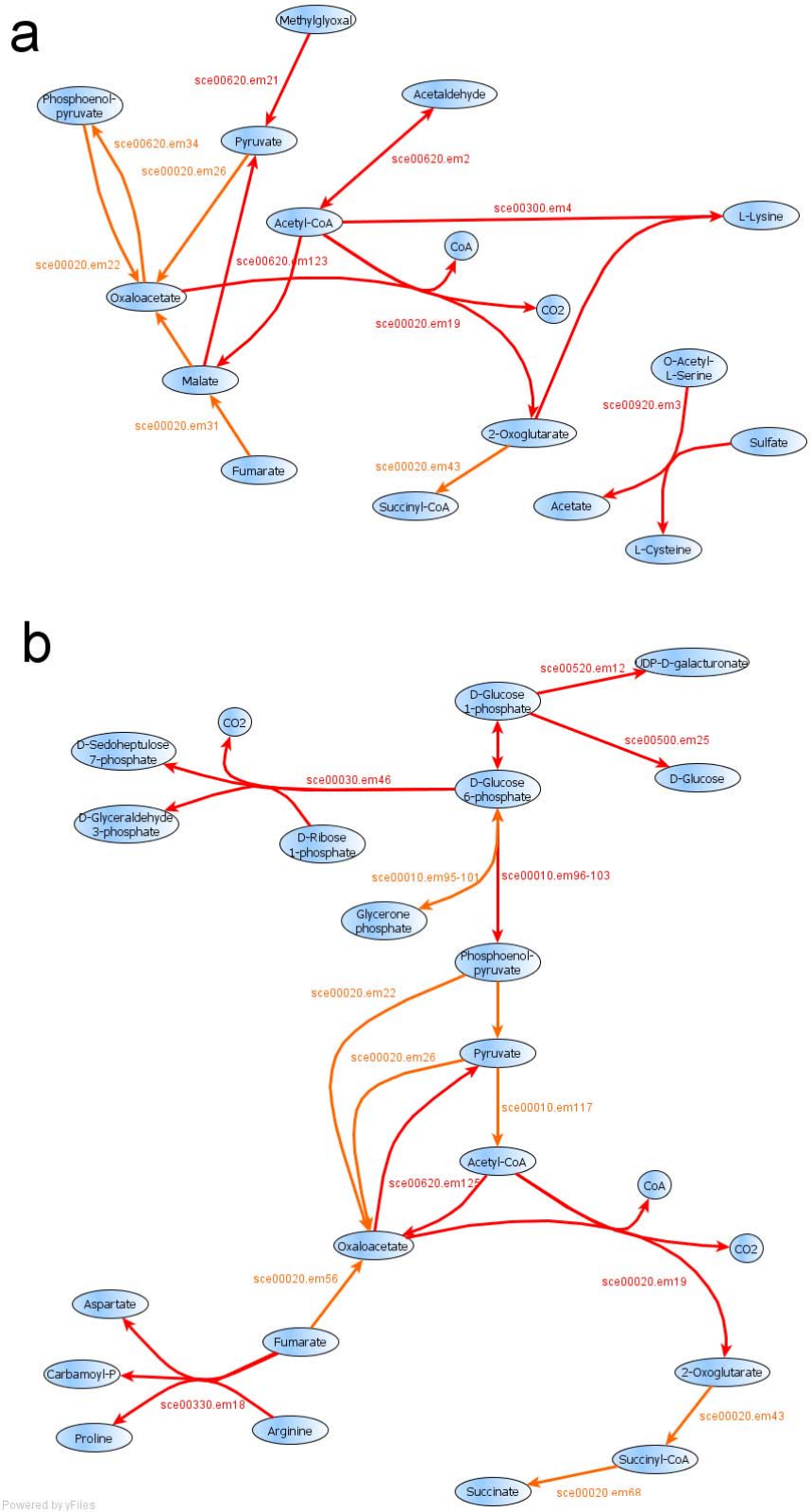
Variable température: variations de température; *Amino acid starvation*: carence en acides aminés; *Stationary phase*: phase stationnaire; *Heat shock*: chaleur; *Nitrogen depletion*: diminution en azote; *Ash*: exposition à des cendres industrielles.

Toxique	Non toxique
Peroxide (C)	Sorbitol (C)
Cadmium (I)	NaCl (C)
Maneb (I)	Acid (C)
Octanol (I)	Heat shock (G)
Pentachlorophenol (I)	Amino acid starvation (G)
Pentanol (I)	Diamide (G)
Thiuram (I)	Nitrogen depletion (G)
Tetrachloro-isophthalonitrile (I)	Stationary phase (G)
Zineb (I)	Variable temperature (G)
	Ash (I)

Pour chacune de ces classes, on a observé des modes élémentaires récurrents dans les réponses métaboliques. Ces similarités dans les réponses à l'intérieur d'une classe, nous ont permis de construire une route métabolique correspondant à la réponse globale de chacune des classes (Figure 23). Ces deux réponses métaboliques sont bien distinctes l'une de l'autre, seuls deux modes élémentaires du cycle de citrate sont impliqués dans les deux réponses.

Figure 23: Routes métaboliques globales correspondant aux deux classes de stress

a) Route métabolique correspondant à la réponse globale aux stress toxiques. b) Route métabolique correspondant à la réponse globale aux stress non toxiques.



3. Discussion

Dans un premier temps, mon travail a consisté à évaluer la pertinence de la décomposition du réseau métabolique sous forme de modes élémentaires proposée par Jean-Marc Schwartz, en utilisant l'approche basée sur la comparaison de groupes. L'utilisation des modes élémentaires offre l'avantage de tenir compte de la topologie du réseau lorsque l'on crée les groupes contrairement aux voies métaboliques utilisées telles quelles. En effet, dans une étude comme la nôtre, la proximité des enzymes appartenant à un mode élémentaire correspondant à des gènes retrouvés différentiellement exprimés, est une observation plus significative (qui rend les résultats plus fiables) que si ces enzymes sont éparpillées à l'intérieur d'une voie métabolique. Draghici *et al.* (Draghici, et al., 2007) propose de combiner plusieurs facteurs pour améliorer l'analyse du métabolisme: ils utilisent la topologie de la voie métabolique, la position dans la voie métabolique des gènes différentiellement exprimés et l'intensité du taux de variation de ces gènes. Ils montrent que leur approche est très prometteuse puisque leurs résultats sont plus significatifs et plus pertinents que ceux obtenus avec une approche classique telle que le Gene Set Enrichment Analysis.

Au cours de ce travail sur la représentation du métabolisme, nous avons proposé une nouvelle décomposition du réseau métabolique sous forme de paires de modes élémentaires traversant plusieurs voies métaboliques et à l'échelle de la cellule entière. Les résultats obtenus, combinés à des données sur le transcriptome, nous ont permis de mettre en évidence des routes métaboliques à l'intérieur du réseau métabolique qui reflètent la réponse de la cellule lors de stress.

Malgré l'utilisation des voies métaboliques pour calculer les modes élémentaires, le nombre total de modes élémentaires calculés est relativement important, et l'on sait que certains sont très similaires. On devine alors que tous ne sont pas forcément fonctionnels dans la cellule, que seulement certains chemins sont préférentiellement utilisés. Le tableau 1.3 dans les Annexes présente le nombre de modes élémentaires issus de EM1 ou EM2 pour lesquels on a trouvé une similarité avec un ou des groupes de gènes surexprimés ou sous-exprimés lors d'un stress. Le tableau indique également le nombre de voies métaboliques auxquelles appartiennent ces modes élémentaires. On remarque que le nombre de voies métaboliques concernées est relativement faible par rapport au nombre de modes élémentaires identifiés, confirmant la redondance des modes élémentaires calculés dans une même voie métabolique. Il pourrait être intéressant d'ajuster notre approche en faisant un tri parmi tous ces modes

élémentaires : une méthode combinant une approche quantitative aux modes élémentaires telle que Jean-Marc Schwartz l'a proposée sur la glycolyse permettrait d'identifier les modes élémentaires les plus probablement utilisés par la cellule (Schwartz and Kanehisa, 2006). En utilisant les informations quantitatives sur la cinétique des enzymes et les concentrations des métabolites, et en les faisant varier, on peut sélectionner les modes élémentaires dont les flux sont affectés par ces variations, c'est-à-dire ceux qui sont actifs. Cependant cette approche nécessite de connaître les valeurs des flux dans toutes les réactions du métabolisme.

Nous avons choisi d'utiliser trois expériences de puces à ADN pour étudier un large panel de conditions de stress chez la levure. Ces expériences ont été menées dans des laboratoires différents et sont donc indépendantes les unes des autres. Néanmoins, on retrouve des conditions similaires étudiées dans les expériences de Causton et de Gasch mais on observe des divergences dans les réponses métaboliques obtenues (illustrées dans la Figure 20) pour ces conditions similaires (*peroxide et heat shock*). La question de la reproductibilité des expériences de puces à ADN est récurrente, et il a été montré que les expériences sont reproductibles et fiables si on prend soin du design expérimental et du traitement des données (Shi, et al., 2006) car la technique des puces à ADN et les méthodes d'analyse sont sensibles à de nombreux facteurs de variation. Les différences obtenues par notre approche dans les réponses métaboliques à des stress similaires peuvent s'expliquer par des différences dans la façon d'appliquer le stress. Elles peuvent également indiquer que la transcription de gènes impliqués dans ces réponses est très finement régulée et peut donc varier si le stress n'est pas appliqué de la même façon, avec la même intensité à la levure (Gasch, 2007).

Si on regarde les routes métaboliques que nous avons pu construire à partir de notre approche, on remarque que seulement 3 sur les 21 correspondent à des routes réprimées. En effet, seulement trois groupes de gènes réprimés lors de stress différents ont montré une relation significative avec des modes élémentaires. Cette observation peut s'expliquer du fait de la difficulté à étudier des groupes de gènes réprimés. Sur une puce à ADN, la détection de gènes faiblement exprimés est délicate et ces gènes ne seront donc pas toujours détectés. Pour cette raison, la fiabilité de la composition des groupes est remise en cause: ceci pourrait expliquer pourquoi on n'observe pas de relation significative avec le métabolisme.
Pour affiner notre approche et réduire le type de biais que nous venons de décrire, il faudrait modifier la méthode utilisée pour former les groupes de gènes sur ou sous-exprimés. En effet notre approche permet de classer les gènes en trois catégories

seulement : surexprimés, sous-exprimés et invariants. Il serait intéressant d'avoir une méthode qui permettrait de faire des classifications plus précises, plus fines des gènes afin d'éviter de définir des seuils sur les valeurs de taux de variation. Ainsi, on pourrait obtenir une description plus précise, plus subtile de la réponse transcriptionnelle et par conséquent métabolique.
D'autres facteurs de variations biologiques pourraient être utilisés dans notre approche pour essayer d'affiner notre vision des réponses métaboliques mises en place par la cellule dans diverses conditions. Nous n'avons tenu compte que du niveau de transcription des gènes; il serait intéressant de prendre en compte les autres niveaux de régulation ou modifications possibles des entités biologiques en présence dans la cellule. Nous pourrions également intégrer plus de deux types de données biologiques (protéomique, métabolomique) afin d'affiner les routes métaboliques que nous avons pu mettre en évidence.

4. Conclusion

Les données sur le métabolisme sont largement exploitées pour l'analyse de données issues de puces à ADN. En général, on utilise le découpage sous forme de voies métaboliques pour mettre en relation le métabolisme avec des groupes de gènes différentiellement exprimés. Ces voies métaboliques sont un découpage arbitraire du réseau métabolique global. L'équipe du KEGG a proposé de découper le réseau sous forme de modes élémentaires. Nous avons exploité notre approche basée sur la comparaison de groupes pour évaluer la pertinence et la réalité biologique des modes élémentaires. La comparaison du contenu en gènes des modes élémentaires avec des groupes de gènes différentiellement exprimés a effectivement permis de mettre en évidence la pertinence biologique des modes élémentaires. Nous avons donc pu montrer l'intérêt d'utiliser les modes élémentaires comme représentation du métabolisme dans une perspective d'intégration. Un avantage important de cette représentation du réseau sous forme de modes élémentaires est le fait de pouvoir les combiner. Leur combinaison nous a permis de reconstituer des routes qui reflètent des réponses métaboliques mises en place par la cellule lors de différents stress, et qui parcourent l'ensemble du réseau global à travers différentes voies.

Annexe 3 - Article n°2 : *"Observing metabolic functions at the genome scale", Genome Biology, 2007.*

CHAPITRE 4 : Caractérisation des gènes à travers leur voisinage d'expression

Avec le nombre toujours croissant de génomes entièrement séquencés, un des défis de la recherche en biologie est d'assigner une fonction à chacun des gènes identifiés. L'annotation d'un gène passe généralement par la comparaison de sa séquence avec celle de gènes dont la fonction est déjà connue chez d'autres organismes. Ce transfert d'annotation fonctionnelle d'un gène à un autre, d'un organisme à un autre, est une des principales applications de la génomique comparative; elle se base sur les liens d'orthologie. La qualité de la prédiction de l'orthologie est donc un point crucial pour l'inférence fonctionnelle.

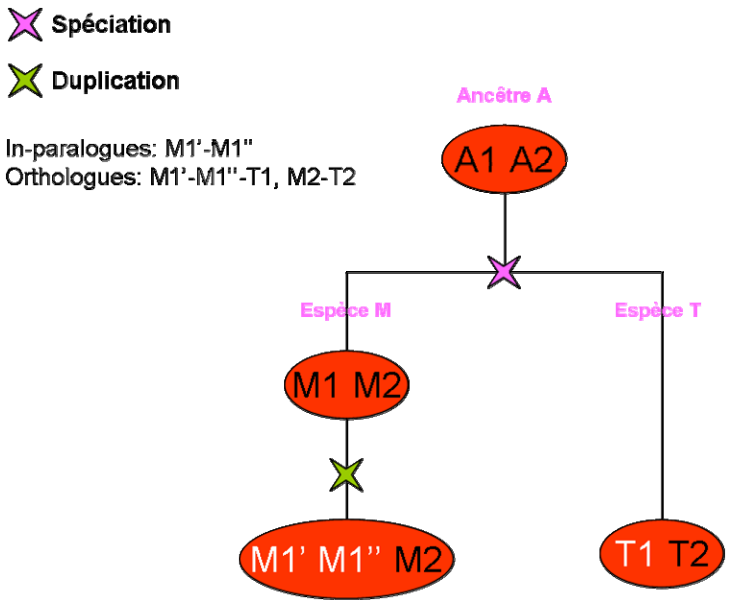
L'orthologie définit la relation existant entre des gènes ayant divergé suite à un évènement de spéciation. Des gènes sont dits orthologues s'ils appartiennent à des organismes différents, et dérivent d'un gène unique chez le dernier ancêtre commun à ces organismes. Par définition, l'orthologie ne décrit pas une relation fonctionnelle entre ces gènes; cependant on fait l'hypothèse que des gènes orthologues ont une fonction similaire, probablement celle du gène ancêtre. L'identité de fonction de gènes orthologues a pu être vérifiée dans de nombreux cas (Dolinski and Botstein, 2007; Hulsen, et al., 2006). C'est pour cette raison que la relation d'orthologie est largement exploitée pour faire de l'inférence fonctionnelle. La relation d'orthologie entre deux gènes est établie sur la base de la similarité qui existe entre leurs séquences du fait de leur origine commune. Cependant, cette relation d'orthologie n'est pas simple à mettre en évidence : au cours de l'évolution, les gènes peuvent subir des duplications et produire des gènes paralogues dont les fonctions sont rarement conservées (Koonin, 2005; Tirosh and Barkai, 2007). Par conséquent, le gène ancêtre de séquences homologues[28] peut subir une ou plusieurs duplications avant ou après spéciation, rendant plus difficile la détection des vrais orthologues. Dans l'exemple de la Figure 24, lors de la comparaison des séquences pour identifier les orthologues dans les espèces M et T, on va retrouver M2 comme le gène le plus similaire à T2, par contre on risque de retrouver M1' et M1" comme les gènes les plus similaires à T1. Il va alors être difficile de savoir si c'est M1' ou M1" qui a conservé la même fonction que T1. On définit alors M1' et M1" comme des gènes in-paralogues; les gènes M1', M1"

[28] Définition de gènes homologues: gènes partageant une origine commune. On distingue les gènes orthologues (issus d'une spéciation) et les gènes paralogues (issus d'une duplication).

et T1 forment un groupe d'orthologues (M1' et M1" sont appelés co-orthologues de T1). Il serait donc nécessaire d'avoir d'autres critères que celui de la similarité de séquences pour espérer distinguer les gènes qui ont réellement la même fonction. C'est ce que nous avons tenté de faire en mettant en place une stratégie basée sur les données de transcriptome issues des puces à ADN.

Figure 24: Schéma des relations d'orthologie et paralogie entre des gènes

Cette figure illustre un scénario possible engendrant des gènes orthologues et paralogues dans deux génomes. Soit deux gènes A1 et A2 appartenant à l'organisme A (ancêtre). Lors d'un évènement de spéciation, les gènes A1 et A2 sont transmis aux espèces descendantes M et T. L'espèce T conserve ces deux versions de A1 et A2, T1 et T2. Tandis que dans l'espèce M, le gène M1 subit une duplication qui donne deux gènes in-paralogues M1' et M1". Ce scénario produit deux groupes d'orthologues : une paire d'orthologues clairement identifiables M2-T2 qui a du conserver la même fonction, et un autre groupe M1'-M1"-T1 pour lequel les gènes ayant réellement la même fonction sont plus difficilement détectables.



Différentes sources, bases de données, mettent à disposition les groupes d'orthologues identifiés par des méthodes qui diffèrent mais toujours basées sur de l'alignement de séquences protéiques. OrthoMCL (Li, et al., 2003) et InParanoid (O'Brien, et al., 2005; Remm, et al., 2001) sont basés sur la même approche : ils utilisent BLAST pour comparer toutes les paires de protéines des génomes étudiés (plusieurs génomes pour OrthoMCL et seulement 2 pour InParanoid). Les séquences qui sont les meilleurs hits réciproques (plus forte similarité) et qui appartiennent à des génomes différents sont définies comme des orthologues. Les séquences dans un même organisme qui ont une

similarité plus élevée entre elles qu'avec celles d'organismes différents sont définies comme des in-paralogues et sont ajoutées aux groupes d'orthologues. Parmi les groupes d'orthologues définis avec cette méthode, certains se chevauchent. Afin d'obtenir des groupes d'orthologues bien distincts, OrthoMCL utilise une méthode de clustering, tandis que InParanoid définit plusieurs règles permettant de faire des groupes indépendants en les fusionnant, en les séparant ou en les éliminant. Une différence importante entre OrthoMCL et InParanoid est le nombre de génomes comparés: InParanoid compare les génomes par paire, alors qu'OrthoMCL peut en comparer plusieurs à la fois. Cependant une nouvelle extension de la méthode d'InParanoid appelée MultiParanoid[29] (Alexeyenko, et al., 2006) permet d'obtenir des clusters d'orthologues de plusieurs organismes à la fois en fusionnant les groupes d'orthologues d'InParanoid. Dans la base de données d'InParanoid, à l'intérieur de chaque groupe d'orthologues, la paire des deux principaux orthologues est différenciable des in-paralogues grâce à une valeur de confiance.
La construction des groupes d'orthologues dans COG (Tatusov, et al., 2003; Tatusov, et al., 2000) repose sur la comparaison par BLAST de toutes les paires de séquences protéiques de plusieurs génomes. Tout d'abord, les séquences d'un même génome qui sont trouvées plus similaires entre elles qu'avec n'importe quelle séquence d'un autre organisme sont rassemblées pour former des groupes d'in-paralogues. Ces groupes sont considérés comme une seule entité (séquence). On recherche alors les entités qui sont les meilleurs hits réciproques pour former des triangles de séquences orthologues. Les triangles ayant un côté en commun (deux séquences qui sont le meilleur hit l'une de l'autre) sont fusionnés pour former des groupes d'orthologues (COGs). Les groupes d'orthologues sont constitués de séquences appartenant à au moins trois organismes différents et qui sont les meilleurs hits réciproques d'au moins deux des génomes. Pour finir, une vérification manuelle de chaque COG permet d'éliminer les faux positifs, et diviser ceux contenant des protéines multi-domaines.

L'utilisation des puces à ADN, qui s'est beaucoup développée ces dernières années, est devenu un outil important dans la stratégie de recherche de fonction des gènes. En effet, on s'attend à ce que des gènes qui ont des fonctions similaires ou participent à un même processus biologique, aient plus de chances d'être co-exprimés que des gènes n'ayant aucun lien fonctionnel ; cette propriété est utilisée pour faire de l'inférence fonctionnelle. En revanche, l'inverse n'est pas toujours vrai: les gènes co-exprimés n'ont pas forcément un lien fonctionnel. La technologie des puces à ADN est une technique dont les résultats sont bruités : le signal détecté n'étant pas toujours fiable, il

[29] Base de données MultiParanoid [http://www.sbc.su.se/~andale/multiparanoid/html/index.html]

est possible que des faux positifs soient trouvés. Par conséquent, lorsqu'on utilise des données de transcriptome issues de puces à ADN pour retrouver la fonction de gènes, il est important d'être capable de distinguer parmi des gènes identifiés comme co-régulés ceux qui ont réellement un lien fonctionnel et ceux qui n'en ont pas.

Une première solution, qui a été largement explorée, consiste à étudier des paires de gènes co-exprimés retrouvés dans plusieurs organismes. Ces approches permettent de s'assurer de la validité du lien de co-expression identifié entre des gènes et de déduire un lien fonctionnel fiable entre ces gènes puisqu'il est conservé dans plusieurs organismes. Pellegrino *et al.* ont développé un outil, CLOE (Pellegrino, et al., 2004), permettant d'explorer les relations fonctionnelles entre des gènes et ainsi, d'identifier des partenaires potentiels pour un gène d'intérêt. Leur approche est basée sur l'étude de données d'expression de gènes orthologues au sein de deux organismes. Les gènes orthologues utilisés sont ceux définis dans la base de données Inparanoïd (Remm, et al., 2001). Pour un gène d'intérêt dans un organisme A, et son orthologue dans un second organisme B, une large gamme d'expériences de puces à ADN est choisie pour chacun des deux organismes. CLOE commence par calculer la corrélation entre les profils d'expression du gène d'intérêt et tous les autres gènes présents dans les expériences de l'organisme A (et fait de même pour le gène orthologue et tous les gènes de l'organisme B). Les gènes sont ensuite ordonnés selon leur corrélation avec le gène d'intérêt, ou son orthologue (de la corrélation la plus forte à la plus faible). On obtient alors deux listes ordonnées pour lesquelles un rang est attribué à chacun des gènes. Pour les paires d'orthologues retrouvées dans les deux listes, la moyenne des rangs est calculée. La liste des paires d'orthologues est alors réordonnée en fonction du rang moyen. Les gènes les plus corrélés de cette liste sont considérés comme des partenaires potentiels du gène d'intérêt. L'analyse des termes de l'ontologie GO surreprésentés dans la liste de ces partenaires est effectuée afin de mettre en évidence un processus biologique dans lequel ce gène d'intérêt pourrait être impliqué, ou même prédire ou confirmer une fonction pour ce gène.

Une autre approche développée par Stuart et ses collègues (Stuart, et al., 2003) se base sur la construction d'un réseau de co-expression conservé chez plusieurs organismes (4 dans cette étude). Ils commencent par créer des métagènes : un métagène se définit comme un groupe rassemblant les gènes orthologues dans les quatre organismes étudiés (les orthologues sont identifiés par similarité de séquences). A partir d'un ensemble d'expériences de puces à ADN, la corrélation entre les profils d'expression de toutes les paires de gènes d'un organisme est calculée. Pour chacun des gènes d'un métagène M1, les autres gènes du même organisme (qui eux-mêmes appartiennent à d'autres métagènes) sont ordonnés; un rang leur est attribué selon leur corrélation avec

le gène considéré, appartenant à M1. Afin de déterminer si le métagène M1 est co-exprimé avec un second métagène M2, les rangs de chacun des gènes de M2, attribués précédemment, sont récupérés, et la probabilité d'obtenir cette combinaison de rangs est calculée. Une valeur seuil est utilisée pour déterminer si la probabilité calculée entre deux métagènes est suffisamment significative pour conclure que ces deux métagènes sont co-exprimés. A partir de ces liens de co-expression, un réseau de métagènes est construit et étudié par visualisation en carte 3D sur laquelle est appliquée une méthode de clustering (K-means). Cette méthode permet de mettre en évidence des modules conservés dans plusieurs organismes (ensemble de métagènes interconnectés) ayant des fonctions similaires ou impliqués dans un même processus biologique.

Lee *et al.* ont développé le même genre d'approche en exploitant les données de puces à ADN d'un seul organisme (Lee, et al., 2004). La recherche de gènes co-exprimés est faite séparément dans chacune des expériences de puces à ADN humaine. Les liens de co-expression entre des gènes sont confirmés s'ils sont retrouvés dans au moins 3 expériences sur les 60 utilisées. Les liens ainsi confirmés sont utilisés pour construire un réseau de gènes co-exprimés. Ce réseau est étudié en appliquant deux types de clustering permettant de mettre en évidence des groupes de gènes très connectés (composants), qui sont ensuite associés à des termes de la GO. Cette approche leur a permis de montrer que les liens de co-expression sont reproductibles d'une expérience à une autre, et que ces liens entre les gènes sont corrélés avec leur(s) fonction(s).

Toutes ces approches reposent sur l'exploitation de données de transcriptome issues de puces à ADN, d'autres se basent sur les données d'ESTs (Expressed Sequence Tags, séquences courtes, marqueurs d'un gène exprimé). Par exemple, la base de données BodyMap-Xs (Ogasawara, et al., 2006) permet d'accéder aux données d'ESTs d'une quarantaine d'organismes. Les données d'ESTs exploitées dans cette base correspondent à l'abondance d'un EST dans un tissu ou un organe, donnant ainsi une estimation du niveau d'expression du gène correspondant (données extraites des bases de données EST de DDBJ[30], et UniGene[31]). De nombreuses recherches sont possibles pour étudier les profils d'expression des gènes dans différents tissus d'un même organisme, ou dans différents organismes (utilisation de InParanoïd (Remm, et al., 2001) pour les orthologues). Un des intérêts de cet outil est de chercher si des gènes homologues ont des profils d'expression similaires.

[30] Site internet de DNA Data Bank of Japan [http://www.ddbj.nig.ac.jp/Welcome-e.html]
[31] Site internet d'UniGene [http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene]

Les études présentées précédemment reposent sur l'analyse de la co-expression de paires de gènes pour trouver des gènes ayant un lien fonctionnel. Pour notre étude, nous avons utilisé le voisinage d'expression d'un gène (voir BNS, Chapitre 2), afin de savoir si cette caractéristique des gènes est susceptible d'être utilisée comme signature en génomique comparative. Le voisinage d'expression d'un gène correspond à l'ensemble des gènes ayant un profil d'expression similaire à ce gène (ensemble de gènes co-exprimés). Pour un gène étudié dans des conditions comparables mais dans des laboratoires différents ou avec des technologies différentes, on s'attend à observer une conservation du voisinage d'expression ; c'est ce que nous avons voulu mettre en évidence dans notre étude.

L'objectif de l'étude est donc de déterminer si le voisinage d'expression d'un gène d'un organisme donné est caractéristique c'est-à-dire s'il est conservé d'une expérience à une autre. Cela revient à généraliser ce que Lee a montré avec des paires de gènes co-exprimés (Lee, et al., 2004). Si c'est le cas, dans quelle mesure le voisinage d'expression peut-il servir à identifier des gènes orthologues, ou plus généralement à identifier des gènes ayant des fonctions similaires ? Comme nous l'avons évoqué précédemment, les relations entre les gènes d'organismes différents ne sont pas faciles à résoudre : les évènements de duplications successives, de transferts horizontaux compliquent la détection des orthologues, ou plus généralement celle de gènes ayant des fonctions similaires mais qui ne sont pas nécessairement des homologues (évolution convergente, déplacement d'orthologues).

Ainsi, le voisinage d'expression pourrait être utilisé comme un critère supplémentaire à la similarité de séquences pour l'identification des différents types de gènes homologues ou de gènes ayant des fonctions proches.

Nous avons utilisé la représentation du voisinage d'expression sous forme de groupes pour étudier la conservation de ce voisinage pour les gènes de *Saccharomyces cerevisiae*. Cette étude a été menée en deux étapes :

- elle a d'abord consisté à comparer les voisinages d'expression de chacun des gènes de *Saccharomyces cerevisiae* à partir de différentes expériences de puces à ADN étudiant des conditions comparables. Cette première étape doit nous permettre de valider la pertinence de l'approche; en effet, les voisinages comparés étant ceux d'un même gène mais dans des expériences indépendantes, on s'attend à retrouver des voisinages significativement similaires pour l'ensemble des gènes de la levure. Ce premier travail nous a donc servi de test, de « contrôle positif ». L'approche pourra ensuite être appliquée à des organismes différents ;

- dans un deuxième temps, la même approche a été utilisée pour comparer les voisinages d'expression de gènes orthologues chez *Saccharomyces cerevisiae* et *Schizosaccharomyces pombe*.

1. Méthodologie

1.1 Choix des données

1.1.1 Données de puces à ADN

Afin de pouvoir poursuivre l'étude jusqu'à la comparaison du voisinage d'expression de gènes orthologues, les expériences de transcriptome issues de puces à ADN que nous avons choisies devaient répondre à des critères particuliers. Il était nécessaire de trouver des données de puces à ADN comparables, par conséquent, les organismes choisis devaient avoir suffisamment de points communs. L'ensemble de la thèse portant sur l'organisme modèle *Saccharomyces cerevisiae*, nous avions besoin d'un organisme qui ait des similitudes avec cette levure. Nous avons donc cherché un second organisme dans la famille des levures, qui est très large phylogénétiquement, pour sélectionner une autre levure relativement distante de *Saccharomyces cerevisiae* telle que *Schizosaccharomyces pombe*.

Schizosaccharomyces pombe est une levure ascomycète tout comme *S. cerevisiae* mais elle ne fait pas partie de la même classe taxonomique (*Archiascomycètes*/*Hémiascomycètes*). Ces deux levures sont éloignées phylogénétiquement (Dujon, 2005) mais elles ont montré de fortes similitudes dans leur fonctionnement, par exemple lors de la réponse aux modifications du niveau de cuivre et de fer dans leur environnement (Rustici, et al., 2007), de la réponse à différents stress (Gasch, 2007), dans la régulation de leur cycle cellulaire (Rustici, et al., 2004; Spellman, et al., 1998).

Dans un premier temps, en tant que contrôle positif, nous avons utilisé deux expériences de puces à ADN concernant *Saccharomyces cerevisiae*, menées dans deux laboratoires différents mais étudiant le même type de stress :

- Gasch *et al.* (Gasch, et al., 2000) ont analysé l'expression des gènes de levure en réponse à différents stress au cours du temps, grâce à des puces à

ADN. Les stress appliqués vont du changement de température, d'osmolarité ou de la source en carbone jusqu'à l'exposition à différents produits (agent oxydant ou réducteur, agent mutagène). En réponse à ces différents stress, un changement rapide dans l'expression des gènes a été observé. Certains gènes répondent de la même manière quel que soit le stress, tandis que d'autres montrent une réponse spécifique à certaines conditions. Ainsi des processus biologiques mis en jeu par la cellule pour s'adapter aux différentes conditions peuvent être mis en évidence. Les données issues de cette expérience sont disponibles dans la base de données Stanford MicroArray Database[32];

- Causton *et al.* (Causton, et al., 2001) ont utilisé la technologie des puces à ADN pour étudier la cinétique de la réponse transcriptionnelle de la levure lors de changements environnementaux (modification de la source de carbone, chaleur, oxydation, changement d'osmolarité et de pH). Comme dans l'expérience précédente, ils ont pu mettre en évidence une réponse globale aux stress impliquant environ 10% des gènes de la levure, ainsi que des réponses plus spécifiques à chaque stress. Les données pour cette expérience sont accessibles sur le site web du laboratoire de Young[33].

Pour la seconde partie de ce travail, nous avons utilisé une troisième expérience de puces à ADN étudiant la réponse à différents stress chez *Schizosaccharomyces pombe* (Chen, et al., 2003). Dans cette expérience, plusieurs stress du même type que les stress appliqués à *Saccharomyces cerevisiae* (changement de température, d'osmolarité ou exposition à différents agents chimiques) ont été utilisés. L'expérience a permis d'identifier des gènes différentiellement exprimés en réponse à un stress spécifique, ainsi que des gènes dont l'expression varie en réponse à plusieurs stress. Ce deuxième groupe de gènes, qui fait partie de la réponse générale (quel que soit le type de stress) a été comparé à celui retrouvé par Gasch (Gasch, et al., 2000) et montre une intersection significative. De plus, les promoteurs des gènes répondant spécifiquement à un stress ont été étudiés afin de trouver des éléments régulateurs communs. Dans certains cas, des éléments déjà connus sont retrouvés et dans d'autres cas de nouveaux motifs potentiellement impliqués dans la régulation des gènes ont été mis en évidence.

[32] Informations supplémentaires sur l'expérience et les données de puce à ADN de Gasch. [http://genome-www.stanford.edu/yeast_stress/]
[33] Données issues de l'expérience de Causton [http://web.wi.mit.edu/young/environment/]

1.1.2 Données d'orthologie d'InParanoid

La base de données InParanoïd (O'Brien, et al., 2005) met à disposition les groupes d'orthologues pour plusieurs paires d'organismes eucaryotes dont *Saccharomyces cerevisiae* et *Schizosaccharomyces pombe*.

Les groupes d'orthologues ont été constitués à partir des résultats de l'alignement des séquences protéiques de *S. pombe* (P) et *S. cerevisiae* (C) obtenus avec BLASTp : comparaison des protéomes P contre C, C contre P, C contre C et P contre P. Dans le cas où des scores asymétriques seraient obtenus lors des comparaisons P-C et C-P, les scores des paires de séquences trouvées similaires (hits) sont moyennés. Afin de ne garder que les paires significatives, un filtre est appliqué sur le score ainsi que sur le pourcentage de la longueur totale de séquences alignées. Des gènes sont définis comme orthologues si leurs séquences sont les meilleurs hits réciproques et appartiennent à des organismes différents. Ces paires de séquences définissent les orthologues principaux pour un groupe d'orthologues donné, et à ceux-ci vont être ajoutés les in-paralogues de chaque organisme. Les in-paralogues sont des séquences appartenant au même organisme et qui présentent un score plus élevé entre elles qu'avec n'importe quelle séquence de l'autre organisme. Cette procédure est mise en œuvre pour traiter tous les hits. Si des groupes d'orthologues se chevauchent, des règles ont été établies pour permettent de séparer, éliminer ou fusionner ces groupes.

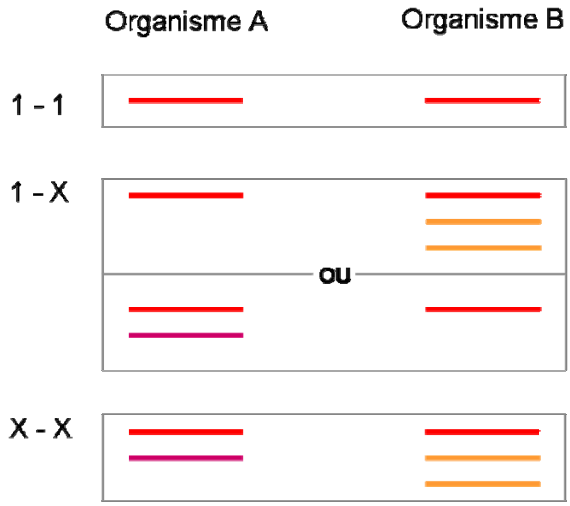
Pour *S. pombe* et *S. cerevisiae*, la base contient 2866 groupes d'orthologues. Ces groupes rassemblent 3099 gènes de *S. pombe* (sur ~ 4800) et 3321 gènes de *S. cerevisiae* (sur ~ 6000). Parmi ces 2866 groupes d'orthologues, 2406 sont des groupes d'orthologues simples, un gène chez un organisme a un seul orthologue chez le deuxième organisme (orthologues 1-1), et 460 sont des groupes d'orthologues multiples, un ou plusieurs gènes chez un organisme a un ou plusieurs orthologues chez le deuxième organisme (orthologues 1-X ou X-X) (Figure 25).

Figure 25: Différents types de groupes d'orthologues de la base InParanoid

Le schéma illustre les différents types de groupes d'orthologues que l'on peut retrouver dans la base de données InParanoid. Les groupes sont représentés par des encadrés gris et les gènes par des lignes horizontales.

Les groupes d'orthologues de type 1-1 correspondent à des paires de vrais orthologues : 1 seul gène de l'organisme A est associé à 1 seul gène de l'organisme B.

Les autres groupes correspondent à des groupes pour lesquels il est difficile de déterminer avec certitude les vrais orthologues. Les groupes de type 1-X correspondent à une paire principale d'orthologues (rouge) et, un ou des in-paralogues pour un des deux organismes (orange ou violet). Les groupes de type X-X correspondent à une paire principale d'orthologues (rouge) et, un ou des in-paralogues pour les deux organismes.



Dans le cas où l'on obtient des orthologues du type 1-X ou X-X, une valeur de confiance est calculée pour les in-paralogues afin d'évaluer leur similarité avec l'orthologue principal (une valeur de confiance de 100% correspond aux orthologues principaux).

1.2 La méthode

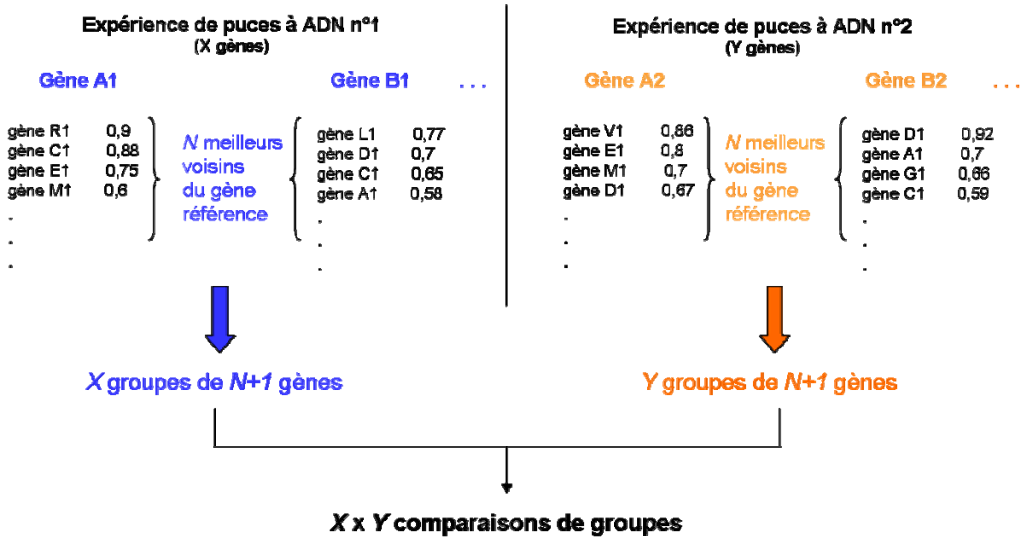
Pour déterminer si le voisinage d'expression d'un gène est caractéristique de celui-ci, il doit être retrouvé dans différentes expériences indépendantes. Pour cela, nous avons mis en place une méthode pour comparer les voisinages d'expression d'un gène issus de deux expériences de transcriptome indépendantes étudiant des conditions similaires. La méthode de constitution du voisinage d'expression d'un gène est basée sur le même principe que la méthode 'Best Neighbours Single' (Chapitre 2, partie 2.2.2).

L'approche comporte plusieurs étapes (Figure 26):

1) Les deux expériences à comparer sont sélectionnées, et seuls les gènes de la liste Génolevures (de *Saccharomyces cerevisiae* et de *Schizosaccharomyces pombe*) communs aux deux expériences et qui ont moins de 50% de valeurs manquantes dans leur profil d'expression sont conservés.
2) Pour chacune des deux expériences, les coefficients de corrélation de Pearson de toutes les paires de gènes sont calculés. Cette mesure nous donne la similarité entre les profils d'expression pour chaque paire de gènes d'une expérience.
3) Pour chaque gène, pris comme référence, on construit son voisinage d'expression c'est-à-dire qu'on définit un groupe composé de ses *N* meilleurs voisins, ce sont les gènes ayant les profils d'expression les plus proches (les coefficients de corrélation les plus élevés). Le groupe final est composé de *N+1* gènes : les *N* voisins + le gène de référence.
4) L'ensemble des groupes issu d'une expérience forme une collection de voisinage d'expression, comme décrit dans le chapitre 1 (partie 3).
5) Pour finir, la composition de chacun des groupes d'une expérience (première collection) est comparée à celle des groupes de la seconde expérience (seconde collection) afin de déterminer quels gènes ont un voisinage d'expression similaire.

Figure 26: Schéma de la méthode de comparaison de voisinage d'expression

Cette figure représente l'approche développée pour comparer les voisinages d'expression de 2 gènes (il peut s'agir de gènes d'une même espèce ou d'espèces différentes). Pour chaque expérience, on crée autant de groupes qu'il y a de gènes: *X* groupes dans l'expérience n°1 et *Y* groupes dans l'expérience n°2. Chaque groupe correspond aux *N* meilleurs voisins c'est-à-dire aux *N* gènes ayant les coefficients de corrélation (chiffres indiqués en face des gènes) les plus élevés avec le gène de référence (A1, A2, B1 et B2 sur la figure). Chaque groupe de la première expérience, en bleu, est comparé à chaque groupe de la seconde expérience, en orange. *X* x *Y* comparaisons de groupes sont alors effectuées.



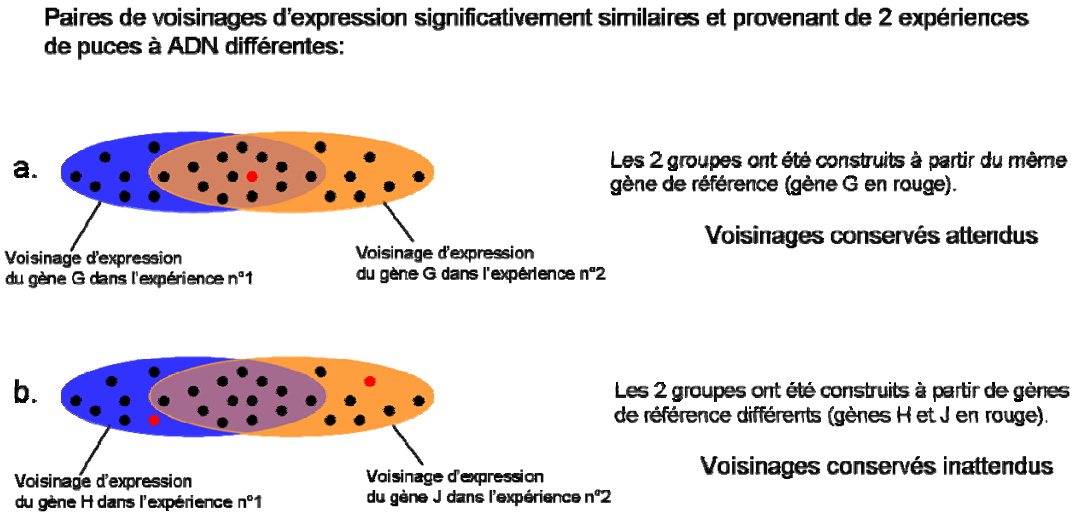
La comparaison des collections de groupes se fait grâce à notre approche de recherche de similarité entre des groupes (Chapitre 1, partie 3). Cette comparaison est effectuée avec BlastSets, et nous permet de mettre en évidence deux types de résultats parmi les groupes identifiés comme significativement similaires (voir Figure 27):

- des voisinages d'expression significativement similaires pour un même gène (gènes identiques ou faisant partie du même groupe d'orthologues) ; ces voisinages similaires sont appelés voisinages conservés attendus (Figure 27a). Ces cas correspondent à ce qu'on s'attend à observer, c'est-à-dire des voisinages conservés d'une expérience à une autre ou d'un organisme à un autre.

- des voisinages d'expression significativement similaires pour des gènes distincts (gènes différents ou ne faisant pas partie du même groupe d'orthologues) ; ces voisinages similaires sont appelés voisinages conservés inattendus puisqu'ils ne sont pas attendus mais correspondent peut-être aux situations originales que nous cherchons à mettre en évidence (Figure 27b).

Figure 27: Les différentes catégories de résultats obtenus

Cette figure représente les deux catégories de résultats qu'on a pu observer. (a) Les groupes trouvés similaires ont été créés à partir d'un même gène (gène G sur le dessin): voisinages conservés attendus. (b) Les groupes trouvés similaires ont été créés à partir de deux gènes différents (gènes H et J sur le dessin): voisinages conservés inattendus.



2. Résultats

2.1 Comparaison du voisinage d'expression des gènes de *Saccharomyces cerevisiae*

Comme premier test, nous avons construit des groupes de gènes contenant les 39 meilleurs voisins de chacun des gènes d'une expérience. Chacun des groupes créés contient 40 gènes (le gène de référence et son voisinage d'expression). L'ensemble forme une collection. Pour chacune des expériences de *Saccharomyces cerevisiae*, les collections sont nommées 'Gasch_40Neighbours' et 'Causton_40Neighbours'.

Nous avons comparé ces deux collections issues des deux expériences de puces à ADN concernant *Saccharomyces cerevisiae*. Chacune de ces collections contient 5595 groupes: plus de 31 millions de comparaisons sont donc effectuées. En effectuant une telle quantité de comparaisons, on retrouve seulement 982 voisinages significativement similaires (Tableau 8) c'est-à-dire 982 paires de voisinage provenant d'expériences différentes et qui ont au moins 8 gènes en commun. Parmi les voisinages retrouvés significativement similaires, 238 correspondent à des groupes construits à partir du même gène dans des expériences différentes, c'est-à-dire aux voisinages conservés attendus. Nous obtenons donc une couverture assez faible, puisqu'on s'attendait à ce que la majorité des 5595 gènes étudiés dans les deux expériences, aient des voisinages d'expression similaires. De plus, nous retrouvons une proportion relativement importante (744) de voisinages conservés inattendus, que nous considérons dans un premier temps comme faux positifs.

Tableau 8: Résultats de la comparaison des voisinages d'expression des gènes de *S. cerevisiae*

Ce tableau présente le nombre de groupes trouvés significativement similaires entre les deux collections 'Gasch_40Neighbours' et 'Causton_40Neighbours'. Le tableau indique le nombre de voisinages conservés attendus puis le nombre de voisinages conservés inattendus.

	Voisinages conservés attendus	Voisinages conservés inattendus
'Causton_40Neighbours' vs. 'Gasch_40Neighbours'	238	744

Il est nécessaire de trouver un moyen de différencier ces deux catégories de voisinages similaires. Nous avons alors étudié la constitution des voisinages. Les voisinages sont composés des 39 meilleurs voisins de chacun des gènes d'une expérience. Pour l'ensemble des voisinages trouvés conservés, la moyenne des coefficients de corrélation des 39 meilleurs voisins a été calculée. La distribution de cette moyenne a été analysée pour les deux catégories de voisinages conservés (Figure 28):

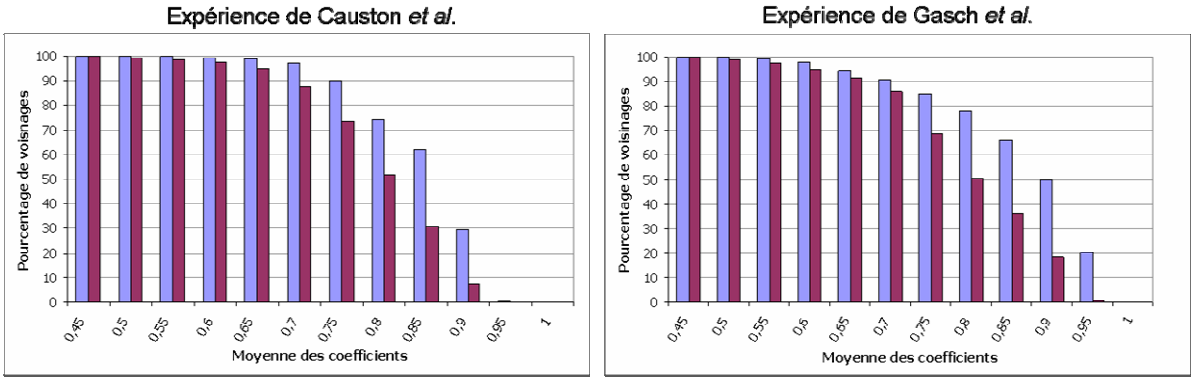
- pour les voisinages conservés attendus, on constate que 75 à 80% des 238 voisinages ont des coefficients de corrélation dont la moyenne est supérieure à 0,8 ;

- pour les voisinages conservés originaux, on observe que seulement 48 à 50% des voisinages ont des coefficients dont la moyenne est supérieure à 0,8.

Les voisinages d'expression conservés attendus sont donc plus homogènes et ont une corrélation plus élevée avec le gène de référence (95% des voisinages entre 0,7 et 1) que les voisinages inattendus (majorité des voisinages entre 0,6 et 0,95). Les groupes de voisinage d'expression conservés attendus correspondent donc à des groupes dont les gènes sont plus fortement corrélés et proches que les autres groupes conservés.

Figure 28: Distribution de la moyenne des coefficients de corrélation des voisinages

Les graphiques présentent la distribution de la moyenne des coefficients de corrélation pour les groupes de 39 meilleurs voisins pour chacune des deux expériences (Causton et Gasch). Chaque barre représente le pourcentage de groupes pour lesquels la moyenne des coefficients de corrélation des 39 meilleurs voisins est supérieure à la valeur donnée en abscisse. Les barres bleues correspondent aux voisinages d'expression conservés attendus ; les barres violettes correspondent aux voisinages d'expression conservés inattendus.



Au vue de ces résultats, et afin de construire des groupes plus pertinents pour vérifier si le voisinage d'expression d'un gène est spécifique de celui-ci, nous avons fait des

groupes des plus proches voisins en utilisant une valeur seuil sur le coefficient de corrélation : on définit les meilleurs voisins comme les gènes qui ont un coefficient de corrélation avec le gène de référence, qui est supérieur à un seuil choisi. Le seuil de 0,8 a été choisi afin de distinguer les 2 types de voisinage conservés.

Nous avons ainsi créé deux nouvelles collections correspondant à chaque expérience : pour construire le voisinage d'expression de chaque gène, au lieu de prendre les *N* gènes ayant les coefficients de corrélation les plus élevés (profils d'expression les plus proches) avec le gène de référence, nous avons pris les gènes dont les coefficients de corrélation avec le gène de référence est supérieur ou égal à une valeur seuil de 0,8. Pour certains gènes, aucun groupe ne sera formé car les coefficients de corrélation des autres gènes sont en dessous de 0,8. Le nombre de gènes dans les groupes va varier d'un voisinage à un autre en fonction du nombre de gènes ayant un coefficient de corrélation supérieur ou égal à 0,8. Par conséquent, dans ces collections, on aura moins de groupes, et des groupes de taille variable. Les deux nouvelles collections 'Causton_Neighbours0.8' et 'Gasch_Neighbours0.8' contiennent respectivement 4035 et 3012 groupes.
Nous avons comparé ces deux collections de la même façon que les collections précédentes ('Gasch_40Neighbours' et 'Causton_40Neighbours'). Ainsi, plus de 12 millions de comparaisons ont été effectuées. Parmi tous les voisinages comparés, on en retrouve 347 conservés attendus et seulement 363 conservés inattendus (Tableau 9). Si on se réfère aux résultats précédents (Tableau 8), on obtient plus de voisinages conservés attendus et moins de voisinages conservés inattendus. Ces résultats semblent plus cohérents : la méthode de construction des voisinages avec un seuil sur les coefficients de corrélation apparaît plus appropriée car elle nous permet de mettre en évidence davantage de voisinages conservés attendus.

Tableau 9: Résultats de la comparaison des voisinages d'expression des gènes de *S. cerevisiae*

Ce tableau présente le nombre de groupes trouvés significativement similaires entre les deux collections 'Gasch_Neighbours0.8' et 'Causton_Neighbours0.8'. Le tableau indique le nombre de voisinages conservés attendus, puis le nombre de voisinages conservés inattendus.

	Voisinages conservés attendus	**Voisinages conservés inattendus**
'Causton_Neighbours0.8' vs. 'Gasch_Neighbours0.8'	347	363

Le nombre de voisinages conservés attendus reste faible par rapport à ce que nous pourrions obtenir: seulement 347 sur les 3012 possibles soit environ 10% (3012 étant le nombre de groupes dans l'expérience de Gasch). Néanmoins, ces 347 voisinages conservés attendus correspondent à 50% des voisinages trouvés similaires; ce qui signifie, si on tient compte des 12 millions de comparaisons effectuées, que notre méthode permet de retrouver dans la moitié des cas, des voisinages similaires vrais c'est-à-dire qui sont avérés puisqu'ils caractérisent des gènes identiques (le même gène). Par conséquent, on peut dire que notre approche a une valeur prédictive plutôt bonne (prédiction correcte dans 1 cas sur 2).

Nous avons ensuite cherché la fonction des gènes de référence pour lesquels nous avons trouvé des voisinages similaires. Que ce soit pour les voisinages conservés attendus ou originaux, on retrouve des gènes dont les protéines appartiennent aux ribosomes, à l'organisation du cytoplasme et à la biosynthèse des ribosomes, au métabolisme des nucléotides. Ces fonctions ont déjà été mises en évidence comme celles correspondant à la réponse aux stress (Causton, et al., 2001; Gasch, 2007; Gasch and Werner-Washburne, 2002). Ces fonctions semblent spécifiques des expériences de puces à ADN choisies et ne permettent pas de faire de distinction entre les deux types de voisinages trouvés similaires. Ces observations nous laissent penser que les expériences utilisées pour cette approche doivent couvrir une large gamme de conditions afin de ne pas biaiser les résultats.

Dans cette première partie du travail, nous avons regardé si le voisinage d'expression d'un gène est une signature de ce gène. Pour cela, nous avons comparé les voisinages d'expression extraits de différentes expériences de transcriptome utilisant les puces à ADN. Nous avons constaté que pour certains gènes, le voisinage d'expression était effectivement conservé d'une expérience à une autre.

Afin de finaliser ce travail, nous allons approfondir l'étude des voisinages conservés inattendus. Afin de comprendre à quoi correspondent ces voisinages conservés inattendus, nous allons analyser en détail les paires de gènes concernés: regarder si leurs fonctions sont identiques (ou proches), ou si ce sont des gènes homologues. Ainsi nous pourrons savoir si les gènes pour lesquels notre approche a mis en évidence une similarité de voisinage d'expression, sont des gènes qui ont un lien fonctionnel et/ou d'homologie.

2.2 Comparaison du voisinage d'expression entre des gènes orthologues

Nous avons poursuivi notre étude en comparant le voisinage d'expression de gènes orthologues. Pour cela, nous avons créé des collections de groupes correspondant aux voisinages d'expression de deux organismes différents : *Saccharomyces cerevisiae* et *Schizosaccharomyces pombe*.

La construction des collections se fait comme c'est décrit dans la partie 1.2 de ce chapitre. Au lieu de prendre les *N* meilleurs voisins, on utilise le seuil de 0,8 sur les coefficients de corrélation pour créer les groupes représentant le voisinage d'expression puisque cette méthode s'est montrée plus efficace. Dans un premier temps, nous n'utilisons que les gènes orthologues de type 1-1 entre *Saccharomyces cerevisiae* et *Schizosaccharomyces pombe* décrits dans la base de données InParanoid afin de pouvoir effectuer la comparaison des groupes et pouvoir conclure clairement au sujet des observations faites. On obtient les collections suivantes 'Gasch_ortho_Neighbours0.8', 'Causton_ortho_Neighbours0.8' et 'Chen_ortho_Neighbours0.8'.

Nous avons comparé les voisinages d'expression issus des expériences de transcriptome de *S. cerevisiae* avec ceux de *S. pombe* :

- 'Gasch_ortho_Neighbours0.8' et 'Chen_ortho_Neighbours0.8': ces collections sont constituées de 1304 et 1303 groupes respectivement, et ont nécessité 1,7 millions de comparaisons;
- 'Causton_ortho_Neighbours0.8' et 'Chen_ortho_Neighbours0.8': ces collections sont constituées de 1695 et 1303 groupes respectivement, et ont nécessité 2,2 millions de comparaisons.

Nous avons également combiné les deux expériences de *S. cerevisiae* (Causton et Gasch) pour construire le voisinage d'expression des gènes communs aux deux expériences et orthologues 1-1 à *S. pombe*. La collection 'Gasch+Causton_ortho_Neighbours0.8' contient 727 groupes. Lorsque cette collection est comparée à 'Chen_ortho_Neighbours0.8', environ 1 million de comparaisons sont effectuées. Les résultats de ces comparaisons sont donnés dans le Tableau 10.

Tableau 10: Résultats de la comparaison des voisinages d'expression des gènes de *S. cerevisiae* et *S. pombe*

Ce tableau présente le nombre de groupes trouvés significativement similaires entre les collections : 'Gasch_ortho_Neighbours0.8' vs. 'Chen_ortho_Neighbours0.8' ; 'Causton_ortho_Neighbours0.8' vs. 'Chen_ortho_Neighbours0.8' ; et 'Gasch+Causton_ortho_Neighbours0.8' vs. 'Chen_ortho_Neighbours0.8'. Le tableau indique le nombre de voisinages conservés attendus, puis le nombre de voisinages conservés inattendus.

	Nombre de voisinages conservés attendus	Nombre de voisinages conservés inattendus
'Gasch_ortho_Neighbours0.8' vs. 'Chen_ortho_Neighbours0.8' (orthologues 1-1)	96	337
'Causton_ortho_Neighbours0.8' vs. 'Chen_ortho_Neighbours0.8' (orthologues 1-1)	75	228
'Gasch+Causton_ortho_Neighbours0.8' vs. 'Chen_ortho_Neighbours0.8' (orthologues 1-1)	74	176

La comparaison des voisinages d'expression issus des expériences de Gasch ou Causton (*S. cerevisiae*) avec ceux issus de l'expérience de Chen (*S. pombe*) donne des résultats comparables. L'expérience de Gasch permet cependant de mettre en évidence plus de gènes orthologues dont le voisinage d'expression est conservé que celle de Causton. En étudiant ces listes de gènes orthologues pour lesquels on a retrouvé un voisinage d'expression conservé, on constate qu'on retrouve, en majorité, les mêmes dans les deux expériences (environ 80%).

La combinaison des deux expériences de transcriptome de *S. cerevisiae* permet de mettre en évidence 74 voisinages conservés attendus et 176 voisinages conservés inattendus. Cette approche est intéressante car en combinant des expériences, on obtient une plus large proportion de gènes avec des voisinages conservés attendus par rapport aux inattendus. Cependant, pour le moment, nous ne sommes pas encore en mesure de conclure si les inattendus sont des faux positifs ou s'ils correspondent à des situations originales.

On constate que le nombre de voisinages conservés attendus (74) par rapport aux 727 paires d'orthologues correspond à la même proportion de voisinages conservés attendus trouvés dans l'étude précédente avec *S. cerevisiae* seulement, soit 10%. Donc, à nouveau, on trouve une couverture faible. Cependant ces 74 voisinages

conservés attendus constituent 30% de l'ensemble des voisinages trouvés similaires. Par conséquent, notre approche nous permet de façon sûre, de prédire dans 1 cas sur 3 une relation fonctionnelle entre des gènes d'organismes différents.

3. Conclusion et Discussion

Nous avons débuté cette étude dans la perspective de voir si le voisinage d'expression d'un gène est caractéristique de ce gène. Si c'est le cas, le voisinage d'expression pourrait alors servir comme critère supplémentaire à la similarité de séquences pour déterminer la fonction des gènes par inférence (transfert d'annotation). Pour répondre à cette question, nous avons fait un travail préliminaire nous permettant de savoir dans quels cas le voisinage d'expression des gènes est conservé, et est donc potentiellement une signature de ces gènes.

Le travail présenté dans ce chapitre est un travail exploratoire sur le voisinage d'expression. Ce travail est encore en cours, et demande donc à être approfondi.
Les premiers résultats obtenus montrent que l'approche que nous avons développée est intéressante puisqu'elle permet de vérifier que pour certains gènes au moins (identiques ou orthologues), le voisinage d'expression est caractéristique. Ces cas sont ceux que l'on attendait, ceux qui nous permettent d'évaluer la sensibilité de notre approche. Ces résultats préliminaires montrent l'existence d'une équivalence fonctionnelle pour certains gènes dont le voisinage d'expression est conservé.
Cependant, nous avons également observé des voisinages similaires pour des gènes distincts (inattendus), c'est-à-dire différents gènes d'un même organisme ou dans des organismes différents (gènes non orthologues). Ces gènes pourraient correspondre à des gènes qui ont une similarité fonctionnelle même s'ils n'ont pas d'homologie de séquences, c'est-à-dire aux cas difficiles à détecter du fait des multiples évènements pouvant avoir lieu au cours de l'évolution (duplications, transferts horizontaux, etc.). L'étude approfondie envisagée de ces paires de gènes inattendues, nous permettra de comprendre si la similarité de voisinage s'explique par une similitude fonctionnelle.

De même, afin de savoir si on peut différencier parmi les in-paralogues ceux qui ont réellement conservé une fonction similaire, nous envisageons d'effectuer la comparaison du voisinage d'expression de tous les types d'orthologues (1-1, 1-X et X-X), et d'en faire une analyse détaillée. Dutilh *et al.* ont fait ce genre d'étude en

introduisant le concept de contexte d'expression (Dutilh, et al., 2006). Le contexte d'expression d'un gène est basé sur la co-expression avec un ensemble de gènes, et permet de comparer l'expression de gènes entre des organismes très distants en utilisant un large jeu d'expériences de transcriptome. Les contextes d'expression de gènes orthologues (1-1, 1-X et X-X) ont été comparés pour étudier leur conservation. Cette étude a montré que les gènes orthologues 1-1 présentent effectivement des contextes d'expression conservés. Cependant, quel que soit le type d'orthologues (1-1, 1-X ou X-X), il n'a pas été trouvé de corrélation entre la conservation du contexte d'expression et l'identité de séquences.

Cette étude de Dutilh et de ses collègues montre que l'utilisation d'un large jeu de données permet de retrouver une correspondance significative entre les contextes d'expression de gènes orthologues 1-1. Dans notre approche, nous n'avons pas retrouvé systématiquement cette relation entre les voisinages d'expression des orthologues 1-1. Pour étendre notre étude, il serait intéressant d'augmenter le nombre et la diversité des expériences de transcriptome dans le but d'améliorer nos résultats (couverture plus large). Le choix initial des expériences de transcriptome est en effet une étape cruciale afin de ne pas biaiser les observations finales.

Un autre choix important concerne les organismes comparés : il est nécessaire de sélectionner des organismes dont la distance phylogénétique est faible afin d'avoir suffisamment de points de comparaisons c'est-à-dire un nombre d'orthologues qui permettent la comparaison des voisinages de gènes et reflétant l'ensemble du génome.

CONCLUSION

Contexte

L'intégration de données n'est pas un concept clairement défini et peut donc faire référence à différentes notions telles que l'exploitation des liens entre les bases de données, la modélisation ou l'exploration du voisinage (voir Introduction). Tout au long du travail présenté dans ce manuscrit, nous nous sommes intéressés au voisinage.

Le concept de voisinage a été proposé par Antoine Danchin en 1998. Il propose de s'intéresser aux relations qui peuvent exister entre des entités biologiques (gènes, protéines) plutôt qu'à ces entités prises individuellement. Ce concept est apparu à une époque où de plus en plus de génomes entièrement séquencés étaient disponibles et où des technologies à grande échelle se développaient. Ces technologies permettent de connaître les changements d'états de l'ensemble des gènes ou protéines, à l'échelle de la cellule entière. Dans ce contexte, la mise en oeuvre du concept de voisinage est alors possible, et a été exploitée dans Indigo (Nitschke, et al., 1998). Cet outil permet une exploration visuelle de différents types de voisinage simultanément (co-localisation, similarité de fonctions, etc.). Par exemple, pour un gène donné, l'utilisateur peut voir dans une première fenêtre, les gènes voisins de ce gène sur le chromosome. Dans une seconde fenêtre, il peut faire apparaître les protéines impliquées dans la même voie métabolique que la protéine codée par ce même gène. L'utilisateur peut ainsi retrouver des similarités entre différents voisinages. Ces similarités, ou correspondances, entre des voisinages permettent de mettre en évidence d'éventuels mécanismes ou contraintes existant dans la cellule. Dans Indigo, seulement quelques types de voisinages sont exploités, et les groupes de voisins ne sont pas créés de manière systématique, mais plutôt selon les informations disponibles.

Afin de pouvoir exploiter ce concept de voisinage de façon méthodique, il était nécessaire de formaliser l'approche proposée dans Indigo. Le voisinage correspond à des groupes rassemblant des gènes ou des protéines qui partagent des propriétés : gènes ayant des profils d'expression similaires, protéines ayant des points isoélectriques proches, protéines voisines dans le réseau métabolique, etc. On peut donc exploiter le voisinage en convertissant les données sous forme de collections de groupes. Une fois ce formalisme défini, il est nécessaire de comparer les groupes entre

eux afin d'identifier des correspondances entre des informations différentes. Pour s'assurer de la fiabilité des similarités mises en évidence entre des groupes, un traitement statistique doit être appliqué. La loi hypergéométrique permet de mesurer la similarité existant entre deux groupes et de contrôler qu'elle n'est pas due au hasard. C'est dans cet esprit qu'a été développé l'outil BlastSets: il est constitué d'une base de données dédiée au stockage des collections de groupes représentant des critères biologiques, et d'un système de requêtes permettant de rechercher des similarités entre des groupes. Il résulte de ces comparaisons des listes de groupes significativement similaires, et qui traduisent des relations complexes pouvant exister dans la cellule.
L'ensemble des travaux présentés dans cette thèse repose sur l'exploitation du voisinage, représenté sous forme de collections de groupes qui sont comparés afin de mettre en évidence des correspondances.

Représentation des données et intégration

L'intégration telle que nous l'avons présentée précédemment passe par la conversion des données biologiques en collections de groupes. Pour certains critères biologiques, cette conversion n'est pas intuitive: il est nécessaire d'appliquer une méthode de clustering pour pouvoir créer une collection de groupes de voisins. Plusieurs méthodes existent, et pour chacune d'elles, divers paramètres peuvent être utilisés. Par conséquent, pour un même critère biologique, une multitude de collections peuvent être générées. Lorsqu'on souhaite intégrer des données, se pose donc le problème du choix de la méthode à utiliser pour obtenir une représentation (collection) pertinente d'un critère biologique. Cette problématique de la représentation a été abordée à travers deux volets au début de la thèse.
Dans un premier volet (Chapitre 2), nous nous sommes intéressés au choix de la représentation de données biologiques hétérogènes dans le but d'optimiser leur mise en relation, et ainsi mettre en évidence des correspondances entre elles. Nous avons proposé une stratégie qui permet de définir une représentation qui, pour un critère biologique donné, capture au mieux les connaissances associées à ce critère de façon à optimiser l'efficacité de l'intégration. Ainsi, en mettant en relation les données d'expression et les complexes multi-protéiques, nous avons montré qu'une représentation simple des données d'expression, basée sur la méthode des 'Best Neighbours' est plus efficace que la méthode classique du clustering hiérarchique. Nous avons également évalué plusieurs méthodes de regroupement des protéines ayant des points isoélectriques proches. Une des méthodes mise au point, nous a

permis d'étendre la liste connue des compartiments sub-cellulaires dans lesquels on retrouve des protéines avec des points isoélectriques voisins. Ces travaux confirment l'importance du choix de la représentation lorsqu'on souhaite intégrer des données.
Dans un second volet (Chapitre 3), nous avons étudié la représentation du métabolisme dans le cadre d'une collaboration avec une équipe du KEGG. Cette équipe nous a proposé une nouvelle façon de décomposer le réseau métabolique sous forme de modes élémentaires. Grâce à l'exploration du voisinage, nous avons validé la pertinence biologique de ce découpage, et montré qu'il pouvait être utilisé comme représentation du métabolisme. Cette représentation combinée avec des données d'expression nous a permis de mettre en évidence des fonctions métaboliques, impliquant plusieurs voies, utilisées par la cellule en réponse à différents stress.

Les principaux résultats de ces travaux ont contribué à:

- optimiser notre stratégie d'intégration en mettant au point des méthodes d'évaluation afin de sélectionner une représentation pertinente. Cette approche et nos résultats pourraient donc être utilisés pour améliorer la performance des outils d'intégration;
- proposer une représentation pour les points isoélectriques et mettre en évidence l'importance de l'exploitation des propriétés physico-chimiques des protéines;
- valider la pertinence biologique de la décomposition du métabolisme sous forme de modes élémentaires;
- proposer une nouvelle approche pour explorer le réseau métabolique dans son entier.

Les travaux présentés laissent entrevoir un certain nombre de perspectives.

✓ En ce qui concerne le Chapitre 2, nous allons nous intéresser plus précisément au choix de la méthode de correction statistique à appliquer lorsqu'on fait des comparaisons multiples. Rappelons que dans nos études, nous avons comparé des collections de groupes, ce qui implique que de nombreuses comparaisons ont été effectuées. De ce fait, le nombre de similarités trouvées par hasard augmente; il est donc nécessaire d'ajuster le seuil de significativité du test statistique. Il existe plusieurs méthodes qui permettent de corriger ce seuil. Nous avons choisi d'utiliser la correction de Bonferroni qui tient compte du nombre de comparaisons effectuées. Cette correction est simple et rapide à calculer quelle que soit la taille des collections comparées. De plus, en utilisant des collections de groupes aléatoires, nous avons démontré que cette correction ne donnait pas de faux positifs. Cependant, cette

correction implique une indépendance entre les groupes, et ce n'est pas le cas dans nos collections. Ceci rend cette correction probablement trop stringente et nous risquons donc de rejeter des similarités vraies. D'autres méthodes de correction peuvent être utilisées, notamment celle du FDR (False Discovery Rate). Cette méthode permet de contrôler la proportion attendue de faux positifs, et est moins stringente comparée à Bonferroni. La méthode FDR vient d'être implémentée dans le système BlastSets, et les comparaisons de collections seront refaites avec cette méthode de correction afin de voir si les résultats obtenus avec Bonferroni sont confirmés et peuvent être améliorés. De plus, le taux de faux positifs pourra être contrôlé grâce à nos collections de groupes aléatoires.

✓ Le travail présenté sur le métabolisme propose une représentation pertinente et originale pour explorer l'ensemble du métabolisme en exploitant les modes élémentaires. Nous avons validé cette approche en utilisant des données d'expression concernant différents stress. Nous souhaitons désormais appliquer notre approche plus largement en utilisant des jeux de données issus d'approches de transcriptomique étudiant d'autres conditions, et en élargissant l'étude à des jeux de données d'expression obtenus par des approches de protéomique.

Exploration du voisinage et génomique comparative

Un autre volet de cette thèse a consisté à exploiter le concept de voisinage d'expression dans une perspective de génomique comparative (Chapitre 4). Nous avons développé une approche pour vérifier si le voisinage d'expression d'un gène pouvait servir de signature. Une telle signature permettrait alors de mettre en correspondance des gènes équivalents sur le plan fonctionnel entre différents organismes. Les premiers résultats montrent que dans un certain nombre de cas, le voisinage permet effectivement de retrouver des paires d'orthologues (il faut souligner qu'aucune information sur la similarité de séquences n'a été utilisée). Cependant, la méthode n'a qu'une faible couverture (seule un petite fraction de toutes les paires d'orthologues est retrouvée). L'utilisation de jeux de données plus larges devrait permettre d'étendre cette couverture.

Pour élargir cette étude, nous souhaitons vérifier si d'autres types de voisinages (co-localisation chromosomique, points isoélectriques) que le voisinage d'expression, sont eux aussi des propriétés permettant de caractériser des gènes ou des protéines. Les analyses en cours confirment que le voisinage d'expression est une signature, mais sa

couverture est faible, donc seule, elle ne suffit pas. Nous souhaitons vérifier si la combinaison de propriétés conservées (conservation du voisinage d'expression, conservation du point isoélectrique ou d'autres propriétés physico-chimiques, etc.) pour des protéines fonctionnellement équivalentes, nous permettrait d'obtenir une meilleure couverture, et finalement d'améliorer les prédictions de fonctions par inférence.

Conclusion générale

L'ensemble de ces travaux a permis de valider la pertinence de la représentation des données sous forme de collection de groupes, et de valoriser la flexibilité, la souplesse, offerte par ce type de représentation. Ceci confirme l'intérêt de l'exploitation du concept de voisinage pour l'intégration de données, qui pourrait être exploité davantage qu'il ne l'est. En effet, peu d'outils d'intégration exploitent le voisinage de façon systématique, alors qu'il est généralisable à un grand nombre de données.

Il serait intéressant que les interfaces de requêtes des bases de données (telles que SRS ou Entrez) offrent la possibilité d'explorer le voisinage. En effet, ces interfaces ont l'avantage de regrouper différents types de données associés aux entités biologiques, et provenant de différentes sources. Elles permettraient donc l'exploration d'un grand nombre de voisinages (et même de combiner plusieurs sources pour un même voisinage) en proposant la recherche d'informations communes à un ensemble d'entités biologiques. Cela pourrait résoudre une des difficultés rencontrées dans ma thèse, et à laquelle on est fréquemment confronté: l'accès et la collecte de données.

BIBLIOGRAPHIE

Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J., Minguez, P., Montaner, D. and Dopazo, J. (2007). From genes to functional classes in the study of biological systems. *BMC Bioinformatics* **8,** 114.

Alexeyenko, A., Tamas, I., Liu, G. and Sonnhammer, E. L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22,** e9-15.

Andrade, M. A., Sander, C. and Valencia, A. (1998). Updated catalogue of homologues to human disease-related proteins in the yeast genome. *FEBS Lett* **426,** 7-16.

Arakawa, K., Kono, N., Yamada, Y., Mori, H. and Tomita, M. (2005). KEGG-based pathway visualization tool for complex omics data. *In Silico Biol* **5,** 419-23.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25,** 25-9.

Baitaluk, M., Sedova, M., Ray, A. and Gupta, A. (2006). BiologicalNetworks: visualization and analysis tool for systems biology. *Nucleic Acids Res* **34,** W466-71.

Ball, C. A., Awad, I. A., Demeter, J., Gollub, J., Hebert, J. M., Hernandez-Boussard, T., Jin, H., Matese, J. C., Nitzberg, M., Wymore, F., Zachariah, Z. K., Brown, P. O. and Sherlock, G. (2005). The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res* **33,** D580-2.

Barriot, R. (2005). Intégration des connaissances biologiques à l'échelle de la cellule, *Thèse*, LaBRI, Université Bordeaux 1.

Barriot, R., Poix, J., Groppi, A., Barre, A., Goffard, N., Sherman, D., Dutour, I. and de Daruvar, A. (2004). New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. *Nucleic Acids Res* **32,** 3581-9.

Benson, D. A., Boguski, M., Lipman, D. J. and Ostell, J. (1994). GenBank. *Nucleic Acids Res* **22,** 3441-4.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2007). GenBank. *Nucleic Acids Res* **35,** D21-5.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28,** 235-42.

Bickmore, W. A. and Sutherland, H. G. (2002). Addressing protein localization within the nucleus. *Embo J* **21,** 1248-54.

Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J. C., Frutiger, S. and Hochstrasser, D. (1993). The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **14,** 1023-31.

Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M., Bon, E., Casaregola, S., de Montigny, J., Gaillardin, C., Lepingle, A., Llorente, B.,

Malpertuy, A., Neuveglise, C., Ozier-Kalogeropoulos, O., Perrin, A., Potier, S., Souciet, J., Talla, E., Toffano-Nioche, C., Wesolowski-Louvel, M., Marck, C. and Dujon, B. (2000). Genomic exploration of the hemiascomycetous yeasts: 4. The genome of Saccharomyces cerevisiae revisited. *FEBS Lett* **487,** 31-6.

Bolshakova, N., Azuaje, F. and Cunningham, P. (2005a). An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics* **21,** 451-5.

Bolshakova, N., Azuaje, F. and Cunningham, P. (2005b). A knowledge-driven approach to cluster validity assessment. *Bioinformatics* **21,** 2546-7.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29,** 365-71.

Canales, R. D., Luo, Y., Willey, J. C., Austermiller, B., Barbacioru, C. C., Boysen, C., Hunkapiller, K., Jensen, R. V., Knight, C. R., Lee, K. Y., Ma, Y., Maqsodi, B., Papallo, A., Peters, E. H., Poulter, K., Ruppel, P. L., Samaha, R. R., Shi, L., Yang, W., Zhang, L. and Goodsaid, F. M. (2006). Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol* **24,** 1115-22.

Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J. M. and Pascual-Montano, A. (2007). GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol* **8,** R3.

Castillo-Davis, C. I. and Hartl, D. L. (2003). GeneMerge--post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics* **19,** 891-2.

Causton, H. C., Ren, B., Koh, S. S., Harbison, C. T., Kanin, E., Jennings, E. G., Lee, T. I., True, H. L., Lander, E. S. and Young, R. A. (2001). Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell* **12,** 323-37.

Chan, P., Lovric, J. and Warwicker, J. (2006). Subcellular pH and predicted pH-dependent features of proteins. *Proteomics* **6,** 3494-501.

Chen, D., Toone, W. M., Mata, J., Lyne, R., Burns, G., Kivinen, K., Brazma, A., Jones, N. and Bahler, J. (2003). Global transcriptional responses of fission yeast to environmental stress. *Mol Biol Cell* **14,** 214-29.

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. and Botstein, D. (1998). SGD: Saccharomyces Genome Database. *Nucleic Acids Res* **26,** 73-9.

Csank, C., Costanzo, M. C., Hirschman, J., Hodges, P., Kranz, J. E., Mangan, M., O'Neill, K., Robertson, L. S., Skrzypek, M. S., Brooks, J. and Garrels, J. I. (2002). Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD). *Methods Enzymol* **350,** 347-73.

D'Haeseleer, P. (2005). How does gene expression clustering work? *Nat Biotechnol* **23,** 1499-501.

Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C. and Conklin, B. R. (2002). GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* **31,** 19-20.

Danchin, A. (1998). *La Barque de Delphes: ce que révèle le texte des génomes*, Odile Jacob edn. Ed. Odile Jacob.

Datta, S. and Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* **19,** 459-66.

Datta, S. and Datta, S. (2006a). Evaluation of clustering algorithms for gene expression data. *BMC Bioinformatics* **7 Suppl 4,** S17.

Datta, S. and Datta, S. (2006b). Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics* **7,** 397.

Dolinski, K. and Botstein, D. (2007). Orthology and Functional Conservation in Eukaryotes. *Annu Rev Genet*.

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C. and Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome Res*.

Drawid, A. and Gerstein, M. (2000). A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J Mol Biol* **301,** 1059-75.

Dujon, B. (2005). Hemiascomycetous yeasts at the forefront of comparative genomics. *Curr Opin Genet Dev* **15,** 614-20.

Dutilh, B. E., Huynen, M. A. and Snel, B. (2006). A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation. *BMC Genomics* **7,** 10.

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 U S A* **95,** 14863-8.

Etzold, T., Ulyanov, A. and Argos, P. (1996). SRS: information retrieval system for molecular biology data banks. *Methods Enzymol* **266,** 114-28.

Fernandez-Ricaud, L., Warringer, J., Ericson, E., Pylvanainen, I., Kemp, G. J., Nerman, O. and Blomberg, A. (2005). PROPHECY--a database for high-resolution phenomics. *Nucleic Acids Res* **33,** D369-73.

Foury, F. (1997). Human genetic diseases: a cross-talk between man and yeast. *Gene* **195,** 1-10.

Gagneur, J., Jackson, D. B. and Casari, G. (2003). Hierarchical analysis of dependency in metabolic networks. *Bioinformatics* **19,** 1027-34.

Gagneur, J. and Klamt, S. (2004). Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics* **5,** 175.

Gasch, A. P. (2007). Comparative genomics of the environmental stress response in ascomycete fungi. *Yeast*.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. 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,** 4241-57.

Gasch, A. P. and Werner-Washburne, M. (2002). The genomics of yeast responses to environmental stress and starvation. *Funct Integr Genomics* **2,** 181-92.

Ge, H., Liu, Z., Church, G. M. and Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* **29,** 482-6.

Ge, H., Walhout, A. J. and Vidal, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* **19,** 551-60.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5,** R80.

Ghazalpour, A., Doss, S., Sheth, S. S., Ingram-Drake, L. A., Schadt, E. E., Lusis, A. J. and Drake, T. A. (2005). Genomic analysis of metabolic pathway gene expression in mice. *Genome Biol* **6,** R59.

Gianchandani, E. P., Papin, J. A., Price, N. D., Joyce, A. R. and Palsson, B. O. (2006). Matrix formalism to describe functional States of transcriptional regulatory systems. *PLoS Comput Biol* **2,** e101.

Goffard, N. and Weiller, G. (2007). PathExpress: a web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Res* **35,** W176-81.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996). Life with 6000 genes. *Science* **274,** 546, 563-7.

Gollub, J., Ball, C. A., Binkley, G., Demeter, J., Finkelstein, D. B., Hebert, J. M., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J. C., Schroeder, M., Brown, P. O., Botstein, D. and Sherlock, G. (2003). The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res* **31,** 94-6.

Guda, C. and Subramaniam, S. (2005). pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics* **21,** 3963-9.

Guldener, U., Munsterkotter, M., Kastenmuller, G., Strack, N., van Helden, J., Lemer, C., Richelles, J., Wodak, S. J., Garcia-Martinez, J., Perez-Ortin, J. E., Michael, H., Kaps, A., Talla, E., Dujon, B., Andre, B., Souciet, J. L., De Montigny, J., Bon, E., Gaillardin, C. and Mewes, H. W. (2005). CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res* **33,** D364-8.

Habeler, G., Natter, K., Thallinger, G. G., Crawford, M. E., Kohlwein, S. D. and Trajanoski, Z. (2002). YPL.db: the Yeast Protein Localization database. *Nucleic Acids Res* **30,** 80-3.

Handl, J., Knowles, J. and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics* **21,** 3201-12.

Heidorn, P. B., Palmer, C. L. and Wright, D. (2007). Biological information specialists for biological informatics. *J Biomed Discov Collab* **2,** 1.

Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J. and Nakai, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res* **35,** W585-7.

Hu, Z., Mellor, J., Wu, J., Yamada, T., Holloway, D. and Delisi, C. (2005). VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res* **33,** W352-7.

Hulsen, T., Huynen, M. A., de Vlieg, J. and Groenen, P. M. (2006). Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* **7,** R31.

Huttenhower, C., Flamholz, A. I., Landis, J. N., Sahi, S., Myers, C. L., Olszewski, K. L., Hibbs, M. A., Siemers, N. O., Troyanskaya, O. G. and Coller, H. A.

(2007). Nearest Neighbor Networks: clustering expression data based on gene neighborhoods. *BMC Bioinformatics* **8,** 250.

Ilic, K., Kellogg, E. A., Jaiswal, P., Zapata, F., Stevens, P. F., Vincent, L. P., Avraham, S., Reiser, L., Pujar, A., Sachs, M. M., Whitman, N. T., McCouch, S. R., Schaeffer, M. L., Ware, D. H., Stein, L. D. and Rhee, S. Y. (2007). The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. *Plant Physiol* **143,** 587-99.

Iragne, F., Barre, A., Goffard, N. and De Daruvar, A. (2004). AliasServer: a web server to handle multiple aliases used to refer to proteins. *Bioinformatics* **20,** 2331-2.

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

Jansen, R., Greenbaum, D. and Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Res* **12,** 37-46.

Kanehisa, M. (2000). *Post-genome informatics*. Oxford University Press.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34,** D354-7.

Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., Tsoka, S., Darzentas, N., Kunin, V. and Lopez-Bigas, N. (2005). Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* **33,** 6083-9.

Kemmeren, P. and Holstege, F. C. (2003). Integrating functional genomics data. *Biochem Soc Trans* **31,** 1484-7.

Kemmeren, P., Kockelkorn, T. T., Bijma, T., Donders, R. and Holstege, F. C. (2005). Predicting gene function through systematic analysis and quality assessment of high-throughput data. *Bioinformatics* **21,** 1644-52.

Kemmeren, P., van Berkum, N. L., Vilo, J., Bijma, T., Donders, R., Brazma, A. and Holstege, F. C. (2002). Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell* **9,** 1133-43.

Khatri, P. and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* **21,** 3587-95.

Kiraga, J., Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Biecek, P., Polak, N., Smolarczyk, K., Dudek, M. R. and Cebrat, S. (2007). The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. *BMC Genomics* **8,** 163.

Klamt, S., Saez-Rodriguez, J., Lindquist, J. A., Simeoni, L. and Gilles, E. D. (2006). A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics* **7,** 56.

Klamt, S. and Stelling, J. (2002). Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep* **29,** 233-6.

Klamt, S. and Stelling, J. (2003). Two approaches for metabolic pathway analysis? *Trends Biotechnol* **21,** 64-9.

Klebanov, L. and Yakovlev, A. (2007). How high is the level of technical noise in microarray data? *Biol Direct* **2,** 9.

Knight, C. G., Kassen, R., Hebestreit, H. and Rainey, P. B. (2004). Global analysis of predicted proteomes: functional adaptation of physical properties. *Proc Natl Acad Sci U S A* **101,** 8390-5.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39,** 309-38.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J. and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Res* **14,** 1085-94.

Lelandais, G., Le Crom, S., Devaux, F., Vialette, S., Church, G. M., Jacq, C. and Marc, P. (2004). yMGV: a cross-species expression data mining tool. *Nucleic Acids Res* **32,** D323-5.

Li, L., Stoeckert, C. J., Jr. and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13,** 2178-89.

Li, Z. and Chan, C. (2004). Integrating gene expression and metabolic profiles. *J Biol Chem* **279,** 27124-37.

Ma, H. W., Zhao, X. M., Yuan, Y. J. and Zeng, A. P. (2004). Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics* **20,** 1870-6.

Marc, P., Devaux, F. and Jacq, C. (2001). yMGV: a database for visualization and data mining of published genome-wide yeast expression data. *Nucleic Acids Res* **29,** E63-3.

Martinez, M. J., Roy, S., Archuletta, A. B., Wentzell, P. D., Anna-Arriola, S. S., Rodriguez, A. L., Aragon, A. D., Quinones, G. A., Allen, C. and Werner-Washburne, M. (2004). Genomic analysis of stationary-phase and exit in Saccharomyces cerevisiae: gene expression and identification of novel essential genes. *Mol Biol Cell* **15,** 5295-305.

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

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

Mlecnik, B., Scheideler, M., Hackl, H., Hartler, J., Sanchez-Cabo, F. and Trajanoski, Z. (2005). PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res* **33,** W633-7.

Montaner, D., Tarraga, J., Huerta-Cepas, J., Burguet, J., Vaquerizas, J. M., Conde, L., Minguez, P., Vera, J., Mukherjee, S., Valls, J., Pujana, M. A., Alloza, E., Herrero, J., Al-Shahrour, F. and Dopazo, J. (2006). Next station in microarray data analysis: GEPAS. *Nucleic Acids Res* **34,** W486-91.

Nakai, K. and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **24,** 34-6.

Nandi, S., Mehra, N., Lynn, A. M. and Bhattacharya, A. (2005). Comparison of theoretical proteomes: identification of COGs with conserved and variable pI within the multimodal pI distribution. *BMC Genomics* **6,** 116.

Natarajan, K., Meyer, M. R., Jackson, B. M., Slade, D., Roberts, C., Hinnebusch, A. G. and Marton, M. J. (2001). Transcriptional profiling shows that Gcn4p is a

master regulator of gene expression during amino acid starvation in yeast. *Mol Cell Biol* **21,** 4347-68.

Nitschke, P., Guerdoux-Jamet, P., Chiapello, H., Faroux, G., Henaut, C., Henaut, A. and Danchin, A. (1998). Indigo: a World-Wide-Web review of genomes and gene functions. *FEMS Microbiol Rev* **22,** 207-27.

O'Brien, K. P., Remm, M. and Sonnhammer, E. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33,** D476-80.

Ogasawara, O., Otsuji, M., Watanabe, K., Iizuka, T., Tamura, T., Hishiki, T., Kawamoto, S. and Okubo, K. (2006). BodyMap-Xs: anatomical breakdown of 17 million animal ESTs for cross-species comparison of gene expression. *Nucleic Acids Res* **34,** D628-31.

Papin, J. A., Price, N. D., Wiback, S. J., Fell, D. A. and Palsson, B. O. (2003). Metabolic pathways in the post-genome era. *Trends Biochem Sci* **28,** 250-8.

Pellegrino, M., Provero, P., Silengo, L. and Di Cunto, F. (2004). CLOE: identification of putative functional relationships among genes by comparison of expression profiles between two species. *BMC Bioinformatics* **5,** 179.

Pena-Castillo, L. and Hughes, T. R. (2007). Why are there still over 1000 uncharacterized yeast genes? *Genetics* **176,** 7-14.

Priness, I., Maimon, O. and Ben-Gal, I. (2007). Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics* **8,** 111.

Purvine, S., Kolker, N. and Kolker, E. (2004). Spectral quality assessment for high-throughput tandem mass spectrometry proteomics. *Omics* **8,** 255-65.

Ramirez, F., Schlicker, A., Assenov, Y., Lengauer, T. and Albrecht, M. (2007). Computational analysis of human protein interaction networks. *Proteomics* **7,** 2541-52.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297,** 1551-5.

Remm, M., Storm, C. E. and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314,** 1041-52.

Rice, P., Longden, I. and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16,** 276-7.

Robinson, M. D., Grigull, J., Mohammad, N. and Hughes, T. R. (2002). FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics* **3,** 35.

Rustici, G., Mata, J., Kivinen, K., Lio, P., Penkett, C. J., Burns, G., Hayles, J., Brazma, A., Nurse, P. and Bahler, J. (2004). Periodic gene expression program of the fission yeast cell cycle. *Nat Genet* **36,** 809-17.

Rustici, G., van Bakel, H., Lackner, D. H., Holstege, F. C., Wijmenga, C., Bahler, J. and Brazma, A. (2007). Global transcriptional responses of fission and budding yeast to changes in copper and iron levels: a comparative study. *Genome Biol* **8,** R73.

Sasson, O., Linial, N. and Linial, M. (2002). The metric space of proteins-comparative study of clustering algorithms. *Bioinformatics* **18 Suppl 1,** S14-21.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467-70.

Schilling, C. H., Letscher, D. and Palsson, B. O. (2000). Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol* **203,** 229-48.

Schuster, S., Fell, D. A. and Dandekar, T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol* **18,** 326-32.

Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I. and Dandekar, T. (2002). Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae. *Bioinformatics* **18,** 351-61.

Schwartz, J.-M., Gaugain, C., Nacher, J. C., de Daruvar, A. and Kanehisa, M. (non publié). Probabilistic comparison of gene sets for detecting key metabolic pathways in the stress response of yeast. *ISMB 2006 (14th International Conference on Intelligent Systems for Molecular Biology)* **Poster J-19**.

Schwartz, J. M., Gaugain, C., Nacher, J. C., de Daruvar, A. and Kanehisa, M. (2007). Observing metabolic functions at the genome scale. *Genome Biol* **8,** R123.

Schwartz, J. M. and Kanehisa, M. (2005). A quadratic programming approach for decomposing steady-state metabolic flux distributions onto elementary modes. *Bioinformatics* **21 Suppl 2,** ii204-ii205.

Schwartz, J. M. and Kanehisa, M. (2006). Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis. *BMC Bioinformatics* **7,** 186.

Schwartz, R., Ting, C. S. and King, J. (2001). Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome Res* **11,** 703-9.

Shamir, R., Maron-Katz, A., Tanay, A., Linhart, C., Steinfeld, I., Sharan, R., Shiloh, Y. and Elkon, R. (2005). EXPANDER--an integrative program suite for microarray data analysis. *BMC Bioinformatics* **6,** 232.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13,** 2498-504.

Sherman, F. (2002). Getting started with yeast. *Methods Enzymol* **350,** 3-41.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. M., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Schrf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P. K., Zhang, L., Slikker, W., Jr., Shi, L. and Reid, L. H. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24,** 1151-61.

Simonis, N., van Helden, J., Cohen, G. N. and Wodak, S. J. (2004). Transcriptional regulation of protein complexes in yeast. *Genome Biol* **5,** R33.

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,** 3273-97.

Spirin, V., Gelfand, M. S., Mironov, A. A. and Mirny, L. A. (2006). A metabolic network in the evolutionary context: multiscale structure and modularity. *Proc Natl Acad Sci U S A* **103,** 8774-9.

Steinmetz, L. M., Scharfe, C., Deutschbauer, A. M., Mokranjac, D., Herman, Z. S., Jones, T., Chu, A. M., Giaever, G., Prokisch, H., Oefner, P. J. and Davis, R. W. (2002). Systematic screen for human disease genes in yeast. *Nat Genet* **31,** 400-4.

Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302,** 249-55.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102,** 15545-50.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. and Natale, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4,** 41.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28,** 33-6.

Tirosh, I. and Barkai, N. (2007). Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol* **8,** R50.

Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E. and Stein, L. (2007). Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* **8,** R39.

Vido, K., Spector, D., Lagniel, G., Lopez, S., Toledano, M. B. and Labarre, J. (2001). A proteome analysis of the cadmium response in Saccharomyces cerevisiae. *J Biol Chem* **276,** 8469-74.

Wei, H., Persson, S., Mehta, T., Srinivasasainagendra, V., Chen, L., Page, G. P., Somerville, C. and Loraine, A. (2006). Transcriptional coordination of the metabolic network in Arabidopsis. *Plant Physiol* **142,** 762-74.

Weiller, G. F., Caraux, G. and Sylvester, N. (2004). The modal distribution of protein isoelectric points reflects amino acid properties rather than sequence evolution. *Proteomics* **4,** 943-9.

Weniger, M., Engelmann, J. C. and Schultz, J. (2007). Genome Expression Pathway Analysis Tool - Analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context. *BMC Bioinformatics* **8,** 179.

Wooley, J. C. and Lin, H. S. (2005). *Catalyzing Inquiry at the Interface of Computing and Biology*.

Wrobel, G., Chalmel, F. and Primig, M. (2005). goCluster integrates statistical analysis and functional interpretation of microarray expression data. *Bioinformatics* **21,** 3575-7.

Wu, S., Wan, P., Li, J., Li, D., Zhu, Y. and He, F. (2006). Multi-modality of pI distribution in whole proteome. *Proteomics* **6,** 449-55.

Yang, H. H., Hu, Y., Buetow, K. H. and Lee, M. P. (2004). A computational approach to measuring coherence of gene expression in pathways. *Genomics* **84,** 211-7.

Yang, Y., Engin, L., Wurtele, E. S., Cruz-Neira, C. and Dickerson, J. A. (2005). Integration of metabolic networks and gene expression in virtual reality. *Bioinformatics* **21,** 3645-50.

Yin, L., Huang, C. H. and Ni, J. (2006). Clustering of gene expression data: performance and similarity analysis. *BMC Bioinformatics* **7 Suppl 4,** S19.

Yoon, S., Ebert, J. C., Chung, E. Y., De Micheli, G. and Altman, R. B. (2007). Clustering protein environments for function prediction: finding PROSITE motifs in 3D. *BMC Bioinformatics* **8 Suppl 4,** S10.

Zhang, B., Kirov, S. and Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* **33,** W741-8.

Zhu, J. and Zhang, M. Q. (1999). SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* **15,** 607-11.

ANNEXES

1. Tableaux de données

Tableau 1.1: Tableau listant les compartiments cellulaires et sub-cellulaires de levure

Compartiments cellulaires	Compartiments sub-cellulaires	Nombre de gènes
• périphérie de la cellule		216
• paroi cellulaire		38
• membrane plasmique		183
• membranes / endomembranes (protéines non assignées à une membrane spécifique)		172
• cytoplasme	protéines du ribosome cytoplasmique (119)	2798
• cytosquelette	tubuline du cytosquelette (36); centrosomes (82); filament intermédiaire (3); cytosquelette d'actine (55)	203
• réticulum endoplasmique (RE)	lumière du RE (9); membrane du RE(129)	549
• appareil de golgi	membrane golgienne (59)	132
• vésicules de transport	vésicules de transport RE-golgi (14); vésicules de transport golgi-RE (60); vésicules de transport à l'intérieur de l'appareil de golgi (3); vésicules de transport golgi-membrane plasmique (4); vésicules de transport golgi-vacuole (55); vésicules de transport de l'endocytose (4); autres vésicules de transport (2)	138
• noyau	matrice nucléaire (37); nucléole (208); pores nucléaires (44); chromosome (43); enveloppe nucléaire(165)	2114
• mitochondrie	membrane interne de la mitochondrie (139); membrane externe de la mitochondrie(20); matrice mitochondriale (69); espace intermembranaire de la mitochondrie (13)	1019
• vacuole	membrane de la vacuole (97); lumière de la vacuole (6)	277
• peroxysome	membrane du peroxysome(20); matrice du peroxysome (19)	52
• endosome	endosome tardif (1)	57
• microsome		5
• particules lipidiques		26
• bud	Neck (112); bud tip (15)	148
• extracellular		51

Tableau 1.2: Tableau présentant le nombre de gènes dans chaque groupe induit ou réprimé

Les lettres entre parenthèse dans la première colonne correspondent à l'expérience de puces à ADN dans laquelle le stress a été étudié : (I)=Iwahashi, (G)=Gasch, (C)=Causton.

Condition de stress	Nombre de gènes sur-exprimés (dans BlastSets)	Nombre de gènes sous-exprimés (dans BlastSets)
Chaleur (C)	173	2
Acide (C)	32	6
Ammoniaque (C)	73	10
H_2O_2 (C)	99	35
NaCl (C)	193	113
Sorbitol (C)	136	8
Chaleur (G)	114	0
Carence en azote (G)	167	11
Phase stationnaire (G)	334	0
Choc hyperosmotique (G)	20	0
H_2O_2 (G)	60	0
Diauxic shift (G)	17	0
Menadione (G)	30	14
Dithiothreitol (G)	56	0
Choc hypoosmotique (G)	11	0
Diamide (G)	94	0
Variation de température (G)	91	0
Carence en acides aminés (G)	61	7
Sources alternatives de carbone (G)	0	0
Cadmium (I)	149	16
Cendres industrielles (I)	390	713
Sodium n-dodecyl benzosulfonate (I)	36	2
Sodium lauryl sulfate (I)	53	2
Capsaicin (I)	10	0
Thiuram (I)	273	166
Zineb (I)	62	17
Maneb (I)	21	4
Tetrachloro-isophthalonitrile (I)	347	518
Pentachlorophenol (I)	181	31
Trichlorophenol (I)	27	10
Ethanol (I)	210	30
Pentanol (I)	285	182
Irradiation (I)	6	6
Octanol (I)	119	55
Pentane (I)	28	40

Tableau 1.3: Tableau récapitulant le nombre de modes élémentaires simples ou par paire trouvés similaires à un groupe de gènes induits ou réprimés

Pour chacun des stress (première colonne), on a le nombre de modes élémentaires trouvés induit ou réprimés dans la collection EM1 et le nombre de voies métaboliques auxquelles ils appartiennent (seconde et troisième colonnes). De la même façon, dans les deux dernières colonnes, on a le nombre de modes élémentaires trouvés induits ou réprimés dans la collection EM2, et le nombre de voies métaboliques auxquelles ils appartiennent.

Les lettres entre parenthèse dans la première colonne correspondent à l'expérience de puces à ADN dans laquelle le stress a été étudié : (I)=Iwahashi, (G)=Gasch, (C)=Causton.

Condition de stress	Nombre de modes élémentaires induits ou réprimés (EM1)	Nombre de voies métaboliques du KEGG induites ou réprimées	Nombre de modes élémentaires induits ou réprimés (EM2)	Nombre de voies métaboliques du KEGG induites ou réprimées
Chaleur (C), induit	12	2	28	4
Chaleur (C), réprimé	2	2	2	2
NaCl (C), induit	5	1	4	2
H_2O_2 (C), induit	16	10	3	2
Sorbitol (C), induit	1	1	30	2
Acide (C), induit	6	1	0	0
Carence en acides aminés (G), induit	13	3	104	19
Diamide (G), induit	42	12	196	21
H_2O_2 (G), induit	6	2	3	2
Chaleur (G), induit	34	2	88	7
Carence en azote (G), induit	2	2	13	6
Phase stationnaire (G), induit	54	5	292	25
Variation de température (G), induit	20	3	57	7
Cendres industrielles (I), induit	24	11	153	19
Cendres industrielles (I), réprimé	200	2	284	8
Cadmium (I), induit	1	1	19	5
Maneb (I), induit	17	11	193	21
Octanol (I), induit	5	2	12	6
Pentachlorophenol (I), induit	7	5	56	12
Pentanol (I), induit	44	7	289	35
Pentanol (I), réprimé	184	2	166	7
Thiuram (I), induit	12	11	19	5
Tetrachloro-isophthalonitrile (I), induit	17	11	25	8
Tetrachloro-isophthalonitrile (I), réprimé	155	1	202	8
Zineb (I), induit	16	10	127	19

2. Article n°1

"Clustering of genes and proteins and optimising the integration of heterogeneous data"

Clustering of genes and proteins and optimising the integration of heterogeneous data

Claire Gaugain[1,2], Nicolas Goffard[3], Alexis Groppi[1], Aurélien Barré[1], Roland Barriot[4], Antoine de Daruvar[1,2§]

[1] Centre de Bioinformatique de Bordeaux (CBiB), Université V. Segalen Bordeaux 2, Bordeaux, France
[2] Laboratoire Bordelais de Recherche en Informatique (LaBRI), UMR CNRS 5800, Université Bordeaux 1, Talence, France
[3] Institut Louis Malardé, PO Box 30, 98713 Papeete, Tahiti, French Polynesia
[4] ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arengberg 10, 3001 Leuven, Belgium

[§] Corresponding author

Email addresses:
CG : claire.gaugain@etud.u-bordeaux2.fr
NG : ngoffard@ilm.pf
AG : alexis.groppi@u-bordeaux2.fr
AB : aurelien.barre@u-bordeaux2.fr
RB : roland.barriot@esat.kuleuven.be
AD : antoine.daruvar@u-bordeaux2.fr

Abstract

Background

Increasing amounts of heterogeneous biological data at cellular and molecular scales require data integration to determine relationships between molecular mechanisms and cellular functions. Through integration, gene and protein data can be brought together and correlations determined. A widely used data integration approach consists in (i) converting biological data into gene or protein sets, and (ii) searching for significant overlaps between these sets. While the approaches to search set overlap are well established, the researchers are left with the difficult task of converting their data into sets. Sometimes, the conversion is straightforward, e.g. a multiprotein complex defines a set, but this is generally not the case and the sets are constructed using a clustering method with specific parameters. Hence, methods are needed for choosing the right clustering algorithm amongst the numerous available.

Results

In this study, we used different clustering approaches and compared their data-integration efficiency. We present detailed results obtained on clustering methods for gene expression profiles ('expression sets') and isoelectric points of proteins ('pI sets'). We compared the expression sets to sets of genes involved in multiprotein complexes, and pI sets to the sets of proteins having the same subcellular localization.

Our results show that the clustering method strongly impacts integration efficiency. For the expression profile data, sets constructed using the 'best neighbours' perform better than hierarchical clustering. For isoelectric points, the most efficient method appears to be a recursive clustering of sets containing proteins that correspond to fixed pI ranges. This last method enables us to extend the list of cellular compartments that are known to be enriched with proteins having a pI within a defined range.

Conclusion

For data integration, the choice of a clustering method and its parameters has a strong impact on subsequent analyses. In general, the effectiveness of a clustering method in representing biological information as a collection of sets cannot be predicted because it results from a balance between the number of sets created and the stringency of the statistical correction required due to multiple comparisons. Our results and the proposed approach can be used to improve strategies and tools addressing functional genomics data integration.

BACKGROUND

The goal of functional genomics is to help understand how the cell factory works. For an increasing number of organisms, the knowledge of the complete genome sequence provides us with the predicted catalogue of pieces of the factory: the genes and their products. Some information, such as functional annotations or physico-chemical properties, is attached to each individual component. Complementary information is provided by a variety of recently developed experimental approaches that allow analyses at the cellular scale. These approaches highlight direct or indirect relationships between the components, e.g., the physical proximity of genes along the chromosomes, the co-regulated expression of genes, the subcellular co-localization of proteins, the physical interactions of proteins, etc. With all these approaches, the vast amount of highly heterogeneous data made available needs to be integrated so that the links between molecular mechanisms and cellular functions can be deciphered. This integration has become a major challenge for bioinformatics.

One important aim of functional genomics data integration, which we address in this paper, is revealing correlations between information attached to genes or proteins. Typically, when analysing results of expression profiling experiments, a biologist's goal is to determine if a set of co-expressed genes corresponds to available functional information, such as a metabolic pathway, by comparing the composition of gene sets. Specifically, in this example, the goal is to determine if a set composed of co-expressed genes is 'similar' to a set which contains the genes coding for enzymes involved in a given metabolic pathway.

Several recently developed tools use this approach to enable biologists to bring together heterogeneous biological data [1]. FunSpec [2] was among the first tools made available to the community. A webserver dedicated to the analysis of yeast sequences allows biologists to submit and compare their own sets of genes to pre-computed collections of sets corresponding to several biological criteria such as functional classes, cellular localization, physical interaction, etc. The data used were retrieved from various sources such as the MIPS [3], the GO databases [4] and various published datasets. Using similar principles, several tools, such as OntologyTraverser [5] and GObar [6], are dedicated to the identification of statistically over-represented terms from the Gene Ontology within a set of genes from various organisms. GENECODIS [7] also enables functional analyses of gene lists looking for co-occurrences of annotations. These terms or annotations are retrieved from several sources such as KEGG pathways, GO, InterPro motifs, etc. Other tools such as goCluster [8] are specifically designed to handle microarray data. goCluster offers various means of clustering microarray data; the clusters obtained are then compared to Gene Ontology terms. The strategy used by all these tools to identify correlations between biological information is very general and can be applied to any type of information that can be attached to a set of genes. This was used by BlastSets [9] (the tool used in this study) to enable not only the analysis of a set of genes, but also of any collection of gene sets corresponding to a biological criterion. Those collections are stored in a database and can be compared to each other in order to find correlations for an entire collection of gene sets.

For this type of approach, data integration efficiency can be defined as the capability to reveal true correlations between biological criteria. This efficiency relies mainly on two methodological steps: data representation and set comparison.

- Data representation. For each biological criterion, data is somehow converted in order to define a collection of gene sets. The conversion method depends on the type of information and should capture as much biological knowledge as possible. In some cases, the definition of sets is straightforward, i.e. each characterised protein complex defines a set composed of the genes that encode for the corresponding proteins. However, in other cases, sets are constructed using an arbitrary clustering method with

specific parameters, i.e. sets of physically neighbouring genes can be defined using a window of a certain size which is then slid along the chromosomes. In those cases, it should be noted that the number of sets generated to convert biological data will vary considerably depending on the method and the parameters chosen to build the sets.

- Set comparison. Once sets are defined, they are compared to each other to find similarities. The similarity between two sets is usually measured using the hypergeometric distribution in order to compute the probability of having at least the observed number of genes in common between two sets. This probability is denoted by the P-value and is considered significant if it is less than or equal to a certain threshold. However, two biological criteria are compared via multiple comparisons: the collection of sets corresponding to a biological criterion is compared to the collection of sets corresponding to the second biological criterion. Consequently the significance of the P-value has to be adjusted and several statistical corrections can be used.

These two steps—capturing biological knowledge into collections of gene sets and evaluating the significance of set similarity—are tightly dependent. Schematically, in order to increase the amount of biological knowledge captured, the number of sets is increased, e.g. several window sizes, rather than just one, are used in order to capture various scales of the physical neighbourhood of genes along the chromosome. Statistical correction will then take into account the increased number of comparisons in order to discard any potential false similarity. One issue that we address in this study is finding a balance between the number of sets defined for each biological criterion and the stringency of the statistical correction used.

In the work presented here, we set up a strategy to evaluate the effects of using different methods to represent biological criteria in the form of set collections. We used two examples exploiting correlations described in the literature:

- Expression data vs. Multiprotein complexes. Different studies have shown that the expression of the subunits of a protein complex tend to be co-regulated [10-12]. We use this correlation between the composition of complexes and the co-expression of the corresponding genes to evaluate the different clusterings of expression data;
- Isoelectric point vs. Subcellular localization. A correlation between the isoelectric point (pI) of proteins and their subcellular localization has been described in several publications [13, 14]. We use this correlation to assess different clustering approaches for representing the pI information as protein sets.

This work was carried out using the BlastSets system to upload and analyse yeast data taken from various public sources. Our approach and the results presented in this paper offer a way to rationalize the choice of clustering methods for the representation of biological criteria in the form of set collection in order to optimize the efficiency of functional genomics data integration.

RESULTS

Experimental setup

For a given biological criterion, various clustering methods can be used to create a collection of sets (see Methods for definitions). Each method results in different collections. In order to determine which of these collections best captures biological information for integration purposes, each assessed collection is compared with a unique reference collection (Figure 1). This reference collection corresponds to a biological criterion, which is known to be correlated with the assessed criterion. The most efficient clustering method is the one that generates the collection which best reveals the known correlation between the assessed and the reference criteria (collection 1 in the example in Figure 1).

Clustering genes according to expression profiles

Because multiprotein complexes are an important level of organization of proteins and play a key role in cellular processes, efforts have been made to identify protein complexes experimentally and predict their functions. Several studies have shown that, in a number of cases, the genes that encode for the subunits of a protein complex are co-regulated. This can be explained intuitively: a protein complex is functional only if all its subunits are present simultaneously in the cell.

We have assessed different methods that allow the clustering of expression data into collections of co-expressed genes sets. Those collections are compared to the reference collection of complex sets composed of the genes encoding the subunits of protein complexes. Our aim is to identify the clustering method of expression data that best reveals the known correlation between the physical interactions of proteins (complexes) and the co-regulation of the corresponding genes. The best method will then be the one which produces the collection that finds the highest number of complex sets significantly similar to at least one expression set.

The assessment was performed for two microarray experiments: the first studied cell cycle-regulated genes (Spellman experiment [15]) and the second studied genes involved in the adaptation to various stressful environmental conditions (Gasch experiment [16]).

We first compared the expression collections obtained using a ‘Best Neighbours Nested’ (BNN) approach (see Methods section for details) with different parameters. For both microarray experiments we built 8 collections corresponding to different set sizes ranging from 5 to 50, 5 to 60, up to 5 to 120. Each collection was used to detect similarities with sets from the complex collection. The results, shown in Figure 2, are similar for both microarray experiments: the number of complexes found to be similar to an expression set increases up to a maximal set size of 100 ('BNN 100' collection) and appears stable in larger collections.

For both experiments, more complex sets are found to be similar to an expression set from the 'BNN 100’ collection than from the Hierarchical Clustering collection (see Table 1). A detailed analysis of the results shows that most of the complexes that are found to be similar to an expression set from the Hierarchical Clustering are also found to be similar to an expression set from the 'BNN 100' clustering. Several of these complexes have already been shown to be regulons (i.e. cytoplasmic ribosomes, RNA polymerase III, nucleosomal protein complex, cytochrome bc1 complex, etc.) [12].

In these initial results, more complexes were found to correlate to expression sets when a larger collection was used: indeed, the 'BNN 100' collection contains approximately 20 times more sets than the Hierarchical Clustering collection (see Table 2). In order to verify that the superior efficiency of the 'BNN 100' collection was due to a "size collection" effect, we built the ‘Best Neighbours Single’ (BNS) collections (see Methods section for details). These BNS

collections contain the same number of sets as the Hierarchical Clustering collections. For both microarray experiments, we built 12 collections corresponding to different set sizes ranging from 10 to 120. Each collection was compared with the complex collection. The results (Figure 3) show that more complexes are found to be similar to an expression set from BNS collections, when using a set size of 30 or higher, than from the Hierarchical Clustering collections. Interestingly, the results obtained with small 'BNS 100' collections are similar to those obtained with the 'BNN 100' collections: 89 and 64 for the Gasch and Spellman experiments respectively with 'BNS 100' as compared to 91 and 62 with 'BNN 100'. These results show that a simple method for representing the expression data (one set of a fixed size for each gene) appears to be optimal for highlighting the correlation with multiprotein complexes keeping a small collection size and thus requiring a limited number of set comparisons.
Finally, sets containing randomly generated complexes were compared to expression sets of each method. We observed that only one randomly generated complex was found to be similar to an expression set thus proving that the similarities found between multiprotein complexes and expression sets were not observed by chance.

Clustering proteins according to Isoelectric points

The isoelectric point (pI) of a protein is the pH at which a protein carries no net electrical charge. The pI is an important physico-chemical property of proteins which could be related to their function [14, 17]. It is known that, depending on the pH, a protein can be more or less active or soluble. It is also known that pH varies from one cellular compartment to another.
The distribution of theoretical pI is multimodal: in eukaryotes three distinct peaks can be observed. Several studies have been conducted to find biological or chemical explanations for the multimodality of the distribution of pI [13, 14, 18]. For *Saccharomyces cerevisiae*, the peaks are around 5, 6.5 and 10 (Figure 4). Those peaks were described by Schwartz *et al.* [13] who correlated them with the localization of the proteins in the cell: proteins with a pI around 5.5 tend to be located in the cytoplasm, proteins with a pI around 6-7 correspond to nuclear proteins, and proteins with a pI around 9 often are membrane proteins. More generally, several authors agree that the subcellular localization of proteins is related to their pI value even if this relationship does not fully explain the multimodal distribution of pI [14]. Drawid *et al.* [19] have shown that isoelectric point data is among the most informative sequence attributes used by software for predicting protein subcellular localization.
We have assessed 4 methods for clustering proteins based on pI similarity. Each method results in a collection of pI sets. Each collection obtained was compared with the reference collection made of compartment sets (see Methods section for details). For each pI collection, the number of subcellular compartments found to be similar to at least one pI set was counted (Table 3):

- 3 compartments for "pI s30 flat"
- 8 compartments for "pI s30 lattice" and for "pI r0.5 flat"
- 12 compartments for "pI r0.5 lattice".

Moreover, only the last method ("pI r0.5 lattice") enables us to highlight subnuclear compartments (Table 4) as described by Bickmore *et al.* [20]. Bickmore shows that most of the proteins located in the nucleolus have a pI of 9-10 and this is the result we obtained by comparing proteins of the nucleolus and sets of proteins from "pI r0.5 lattice". Several of the compartments identified in our study (Table 4) have already been described in the literature as preferentially containing proteins from a certain pI range:

- ribosome with a highly conserved basic pI [17]
- cytoplasmic proteins with a pI around 5.5, membrane proteins with a pI around 9 and nuclear proteins with a pI of 6-7 [13].

A recent article presents a large study on the correlation between pI and other properties of proteomes [21], indicating that the relationship between pI values and some cellular compartments in eukaryotes is significant:

- cytosol, vacuoles and cytoskeleton have acidic proteome; we have also made these observations in our work (4-7, 3.5-5.5 and 4.5-6 respectively)
- mitochondrion has a basic proteome which we note in our results (8.5-11.5).

These results not only confirm known correlations between pI and protein subcellular localization, they also show that such correlation exists for compartments that were not previously described. These results also show that the most appropriate method for capturing pI similarity in order to highlight correlation with cellular localisation is by building sets of pI range of 0.5 organized in a lattice. As a control, we randomly re-assigned pI values to proteins in order to generate random pI collections. These random collections were compared to subcellular compartment sets and no similarity was observed, thus proving that the correlations between protein subcellular localization and protein pI were not observed by chance.

DISCUSSION

Our results show that the choice of clustering method for the representation of biological information in the form of a collection of sets, strongly impacts integration efficiency. It should be stressed that we have evaluated clustering methods to answer very specific questions: Which clustering can best represent expression profiles in the form of a collection of gene sets in order to find protein complexes whose components are coded by co-regulated genes? Which clustering can best represent pI similarity in the form of a collection of protein sets in order to find cellular compartments enriched with proteins of a certain range of pI? Clearly, using the most appropriate method strongly increases the capacity to integrate different biological information.

In the first part of the study, we initially compared collections of expression sets obtained using the 'Best Neighbours Nested' (BNN) clustering. We noticed that the number of complex sets found to be similar to an expression set (Figure 2) increases up to the collection 'BNN 100' and appears stable for larger collections ('BNN 110' and 'BNN 120'). These BNN collections are nested, thus a larger collection contains smaller ones plus additional sets with increasing sizes which should enable us to capture more information from expression data. Looking at the results in detail, we actually observe that, while enlarging the expression collection, new complex sets are indeed found to be similar to added expression sets. However, simultaneously, some similarities between complex sets and expression sets that were found significant with smaller collections are lost. This phenomenon is due to the statistical correction (Bonferroni) applied to discard similarities that might be obtained by chance due to multiple comparisons. The increase of the number of sets in a collection (Table 2) leads to a more stringent selection of similarities which might be penalizing. It should also be noted that using larger collections has another drawback, which can become a bottleneck: heavier computational requirements due to the greater number of set comparisons required.

The advantage of Hierarchical Clustering, which aims at capturing similarities of expression profiles, over our BNN approach, is the production of far fewer sets. However, when the results from Hierarchical Clustering and those from a BNN 100 collection were compared, we observed (Table 1) that, for both expression experiments, fewer complexes were found to be similar to expression sets with the Hierarchical Clustering.

To verify that this result was due to a "collection size effect", we constructed small collections based on a Best Neighbours approach with a fixed set size: each BNS collection contained the same number of sets as the Hierarchical Clustering collections (Table 2). BNS collections of set size 30 and larger are able to highlight similarities with significantly more complex sets than are Hierarchical Clustering collections. Thus, using a 'BNS 100' collection for the Gasch experiment, the number of multiprotein complexes that appear correlated with an expression set increases by 60% compared to Hierarchical Clustering.

It is noteworthy that a simple approach which roughly aggregates genes with similar expression profiles, rather than the Hierarchical Clustering which optimises gene aggregation, yields the best results. The strength of the BNS approach is probably the creation of an expression set of best neighbours for each gene, which is not the case for the Hierarchical Clustering.

A careful look at the results from the BNS clustering reveals different complexes that were not found with Hierarchical Clustering and that have been shown to be regulons [12]: this is the case of cdc28p and the RNA polymerase II complexes. This outcome indicates that results obtained with BNS clustering are at least as reliable as those obtained with Hierarchical Clustering.

These results suggest that the BNS is the most efficient way to cluster genes with close expression profiles in order to reveal correlation with multiprotein complexes. This method

enables a systematic exploration of the expression neighbourhood for each gene of a microarray experiment as well as a reasonable computing time.
In the second part of this work, we compared different methods for clustering yeast proteins showing a similar isoelectric point using correlation with the subcellular localization of proteins as a reference. The most efficient method appears to be a recursive clustering of sets based on a fixed size of pI ranges which allowed us to highlight correlations between pI ranges and the localization of proteins in a large number of cellular compartments. Interestingly, although some of the compartments that were found to correlate with a range of pI have already been described in the literature, others are new. Thus, the correlation between pI and subcellular localization is probably more general than originally thought and might even be stronger than what we observed since the pI used in our experiment was theoretical and thus probably differs from the real *in vivo* pI. Indeed, even if Chan *et al.* [22] have shown that the calculated pI of folded proteins and of unfolded proteins tend to be similar, it is known that proteins can undergo several post-translational modifications in the cell that can change their pI. It should be noted again that the clustering method ('pI r0.5 lattice'), which appears to be the most efficient for highlighting correlation, produces a reasonable number of sets (Table 5).
These results suggest that the best clustering methods are those which correspond to an optimal balance between the number of sets built for a criterion and the stringency of the necessary statistical correction due to multiple comparisons. The underlying key parameter is the signal/noise ratio within the set collection: what fraction of the collection has biological relevance? This parameter cannot be predicted.
An important parameter of set comparison approaches is the statistical correction for multiple comparisons. In our study, we used the Bonferroni correction which is simple to compute and which is available in most of the tools that perform sets comparisons [2, 23, 24]. This correction appeared efficient in filtering out false positives: none were found when randomised collections of sets were compared.
The work we presented here on expression data and pI can be carried out for any biological information where the construction of sets is not straightforward. This is the case for data which corresponds to continuous values and thus requires a clustering method for building the sets. Typical examples are genes/proteins physico-chemical attributes (such as pI, physical location on the chromosomes, etc.). This information is rarely taken into account by data integration tools, probably because of the difficulty in representing it. Indeed, if for example, genes are to be clustered to capture their physical proximity along the chromosome, there are no obvious criteria for deciding which cluster sizes are relevant and should be kept. This is especially true when the clusters are to be used for exploratory comparison with other criteria which could potentially involve any group of genes from a pair up to a complete chromosome. In those cases, empirical approaches, such as the one presented here, could be very helpful in optimizing the construction of sets.

CONCLUSIONS

For the purpose of data integration, the effectiveness of a clustering method in representing biological information in the form of a collection of sets cannot be predicted. A method's efficiency is not guaranteed by increasing the number of sets because the multiple comparisons that occur then necessitate statistical correction.
We have proposed an empirical approach for assessing the efficiency of different clustering methods for the purpose of data integration based on set comparison. Our approach and the results presented in this paper offer a means for rationalizing the choice of clustering methods

for representing biological data in the form of set collections. This method can be used to optimize the efficiency of strategies and tools for the integration of functional genomics data.

METHODS

Our study was performed on the *Saccharomyces cerevisiae* genome; the curated list of genes was provided by the Génolevures Consortium [25].

Definition

A set is composed of genes or proteins that have a similar value or attribute. Only the genes that are coding for proteins were considered, and a unique identifier was used to refer to a gene and its product, thus enabling the comparison of sets that contain either genes or proteins.

A collection of sets represents a given biological criterion (i.e. expression of genes measured in a given microarray experiment, all known multiprotein complexes, isoelectric point of proteins, subcellular location of proteins).

Several representations (collections) are possible for the same biological criterion.

Creation of set collections

Expression data representations

Expression data used in this work are raw data from microarray experiments that were retrieved from Stanford MicroArray Database [26, 27]. We used two microarray experiments:

- in the first, Spellman and his colleagues followed the expression of genes during the different steps of cell cycle in yeast [15],
- in the second, Gasch *et al.* compared the genetic adaptation of *Saccharomyces cerevisiae* in different stressful environmental conditions [16].

These two well studied experiments were chosen because they focus on gene regulation in various biological conditions and were shown to highlight the co-regulation of gene sets involved in different biological processes.

We used a collection of 1059 yeast multiprotein complexes retrieved at MIPS, Munich Information Center for Protein Sequences [3]. Some of these complexes are well described in scientific literature—they have been identified and annotated manually; others have been identified by several high-throughput experiments [28-30]. Each complex represents a set composed of genes encoding the subunits of this complex. All these sets ('complex sets') form the collection representing multiprotein complexes ('complex collection').

Different studies have shown that the expression of the subunits of a protein complex tend to be co-regulated [10-12]. We used this correlation to evaluate different representations of expression data. Multiprotein complexes from MIPS were used as the reference criterion. These sets were compared with different set collections derived from expression data. The best method for building a set collection from expression data highlights the highest number of reference complex sets that have a similarity with at least one set of co-regulated genes.

For each microarray experiment, we computed the Pearson correlation coefficient to measure the similarity between the expression profiles of all pairs of genes present on the microarray. Once this calculation was performed, two methods of clustering were applied to create a collection ('expression collection') of sets, grouping genes with similar expression profiles ('expression sets').

First, Hierarchical Clustering (Figure 5 A) was performed using Cluster software [31] and resulted in collections of 5629 and 5648 sets for the Spellman and Gasch experiments respectively. Each node of the resulting binary tree defines an expression set.

As an alternative, we proposed another method that we named the ‘Best Neighbours’. In this simple method, the genes showing the highest expression profile correlation (see Figure 5 B) are pulled together. Two versions of this method were used: BNS (Best Neighbours Single) and BNN (Best Neighbours Nested).

In the BNS approach, for each gene of an experiment, one set is built containing a fixed number of neighbours showing the highest expression profile correlation. We used 12 different set sizes, resulting in 12 collections of sets. Each collection contains a fixed number of sets equal to the number of genes in the experiment: the 'BNS 10' collection contains size 10 sets; the ‘BNS 20’ collection contains size 20 sets, and so on, up to 'BNS 120'.

In the BNN clustering, for each gene of an experiment, series of sets were built containing an increasing number of best neighbours. For this method, 8 settings (thresholds) were used, resulting in 8 collections of sets:

- The ‘BNN 50’ collection contains sets of size 5, 10, 15, 20, 25, and so on up to 50 (by steps of 5). The collection contains 56300 and 56490 sets for Spellman and Gasch experiments respectively (10 sets per gene).
- The ‘BNN 60’ collection contains all the sets from the ‘BNN 50’ collection plus sets of size 55 and 60. The collection contains 67560 and 67788 sets for Spellman and Gasch experiments respectively (12 sets per gene).

The six other collections were built in the same way up to a threshold of 120: ‘BNN 80’, ‘BNN 90’, ‘BNN 100’, ‘BNN 110’ and ‘BNN 120’.
The size of each expression collection is given in Table 2.

Protein isoelectric point representations

We used sequences of *Saccharomyces cerevisiae* proteins from the Génolevures Consortium to calculate the theoretical isoelectric point of each protein. We performed these calculations with the application “iep” from the EMBOSS package [32].

Data on the subcellular localization of *Saccharomyces cerevisiae* proteins were obtained from the MIPS website [33]. All these compartments were hierarchically organized. The list of 49 cellular and subcellular compartments described in the MIPS database was completed with another “compartment”: cytoplasmic ribosomal proteins retrieved from the MIPS multiprotein complexes database (cytoplasmic ribosomal large subunit and cytoplasmic ribosomal small subunit), see Table 6.

We used the correlation between the isoelectric point (pI) of proteins and their subcellular localization, which has been described in several publications [13, 14], to assess different methods for representing the pI. The pIs of proteins are real numbers which potentially cover a continuous range of values. In order to build sets ('pI sets') composed of genes that encode for proteins with close pI values, a clustering method must be chosen. Different methods for building these collections of sets ('pI collections') were tested using the subcellular localization of proteins as the reference criterion. The SubCell database from MIPS was used as source of information on the subcellular localization of yeast proteins. For each subcellular compartment, we created a set ('compartment set') composed of the genes encoding the proteins that are supposed to be localized there. A reference collection ('compartment collection') of 50 sets corresponding to those compartments was created (see Table 6). Using this setting we identified the most efficient method for representing protein pI as the one that

highlights the highest number of compartment sets that have a similarity with at least one pI set.

We assessed different methods to build sets corresponding to proteins with close pI values. We first created 2 pI collections of sets corresponding to proteins with adjacent pI values. One collection was made of sets containing a fixed number of genes; the other from sets containing genes that encode proteins that cover a fixed range of pI values. These collections were either used as is (flat collections) or aggregated recursively in order to build a lattice (lattice collections) as described in Figure 6.

To build these collections, yeast proteins were sorted by increasing values of pI. The corresponding distribution is shown in Figure 4. Based on this sorted list, we created the two set collections containing proteins with adjacent pI values:

- Fixed size sets: each set contains 30 genes ("pI s30")
- Fixed pI range sets: each set contains genes that correspond to proteins whose pI is within a range of 0.5 ("pI r0.5").

The resulting 4 collections were named: "pI s30 flat", "pI s30 lattice", "pI r0.5 flat" and "pI r0.5 lattice". The number of sets contained in each collection is shown in Table 5.

Comparison of biological criteria represented as set collections

To determine if a "correlation" exists between biological criteria, we captured the corresponding information into collections of gene sets. For a given criterion, each set of a collection is composed of genes which directly, or indirectly through their products, are similar (or neighbours according to the A. Danchin definition [34]). This unified representation of information enables the comparison of heterogeneous data. Indeed, if two sets were found to be significantly similar while corresponding to different criteria, those criteria could be considered "correlated". The similarity between 2 sets is simply based on the comparison of their gene composition. Assuming that the sets are made of genes retrieved from the complete list of genes of the organism, we used the hypergeometric distribution to compute the statistical significance of this similarity (P-value).

The comparison of 2 biological criteria involves a systematic pairwise comparison of all the sets from the corresponding set collections. Schematically, the number of sets, from a collection attached to a given criterion, that are found to be similar to sets from a collection of another criterion, reflects the level of correlation between the two criteria. When comparing 2 criteria, a statistical correction is necessary to discard similarities which might be due to multiple set comparisons. We used the Bonferroni correction which divides the significance threshold, denoted α, by the number of comparisons performed (the number of sets in the first collection (n) multiplied by the number of sets in the second collection (m)):

$$T = \alpha / n*m$$

T is the adjusted threshold. The similarity between two sets (P-value) is considered significant if P-value < T. In this study, we set α=0.1.

In order to ensure that the similarities found when comparing 2 biological criteria were significant, we systematically created a test collection by randomizing the composition of the sets corresponding to one of the criteria. The random collection was compared to the other collection with the expectation that no significant similarity would be detected.

We used the BlastSets system [35] to store and compare the different set collections.

Authors' contributions

AD and CG conceived of the experiment. CG performed most of the experiments with support from NG, AG, and AB. RB implemented the BlastSets software which was used to run most experiments. AD, CG and AG prepared the manuscript. All authors have approved the final manuscript.

Acknowledgements

CG has a PhD fellowship provided by the French Ministry of Education, Research and Technology. RB is a post-doctoral researcher with a grant from the research Council Katholieke Universiteit Leuven, Center of Excellence EF/05/007 SymBioSys. The BlastSets project is supported by funds allocated by ACI IMPBio from the French Ministry of Research. The computational resources were provided by the Centre de Bioinformatique de Bordeaux located at the Université Bordeaux 2 which is funded by the Région Aquitaine and the French Ministry of Education, Research and Technology. We thank Maxime Delorme for his help and his work on BlastSets.

References

1. Khatri P, Draghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems**. *Bioinformatics* 2005, **21**(18):3587-3595.
2. Robinson MD, Grigull J, Mohammad N, Hughes TR: **FunSpec: a web-based cluster interpreter for yeast**. *BMC Bioinformatics* 2002, **3**:35.
3. Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V *et al*: **MIPS: analysis and annotation of proteins from whole genomes**. *Nucleic Acids Res* 2004, **32**(Database issue):D41-44.
4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
5. Young A, Whitehouse N, Cho J, Shaw C: **OntologyTraverser: an R package for GO analysis**. *Bioinformatics* 2005, **21**(2):275-276.
6. Lee JS, Katari G, Sachidanandam R: **GObar: a gene ontology based analysis and visualization tool for gene sets**. *BMC Bioinformatics* 2005, **6**:189.
7. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A: **GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists**. *Genome Biol* 2007, **8**(1):R3.
8. Wrobel G, Chalmel F, Primig M: **goCluster integrates statistical analysis and functional interpretation of microarray expression data**. *Bioinformatics* 2005, **21**(17):3575-3577.
9. Barriot R, Poix J, Groppi A, Barre A, Goffard N, Sherman D, Dutour I, de Daruvar A: **New strategy for the representation and the integration of biomolecular knowledge at a cellular scale**. *Nucleic Acids Res* 2004, **32**(12):3581-3589.
10. Ge H, Liu Z, Church GM, Vidal M: **Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae**. *Nat Genet* 2001, **29**(4):482-486.
11. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein-protein interactions**. *Genome Res* 2002, **12**(1):37-46.

12. Simonis N, van Helden J, Cohen GN, Wodak SJ: **Transcriptional regulation of protein complexes in yeast**. *Genome Biol* 2004, **5**(5):R33.
13. Schwartz R, Ting CS, King J: **Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life**. *Genome Res* 2001, **11**(5):703-709.
14. Wu S, Wan P, Li J, Li D, Zhu Y, He F: **Multi-modality of pI distribution in whole proteome**. *Proteomics* 2006, **6**(2):449-455.
15. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, 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.
16. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes**. *Mol Biol Cell* 2000, **11**(12):4241-4257.
17. Nandi S, Mehra N, Lynn AM, Bhattacharya A: **Comparison of theoretical proteomes: identification of COGs with conserved and variable pI within the multimodal pI distribution**. *BMC Genomics* 2005, **6**:116.
18. Weiller GF, Caraux G, Sylvester N: **The modal distribution of protein isoelectric points reflects amino acid properties rather than sequence evolution**. *Proteomics* 2004, **4**(4):943-949.
19. Drawid A, Gerstein M: **A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome**. *J Mol Biol* 2000, **301**(4):1059-1075.
20. Bickmore WA, Sutherland HG: **Addressing protein localization within the nucleus**. *Embo J* 2002, **21**(6):1248-1254.
21. Kiraga J, Mackiewicz P, Mackiewicz D, Kowalczuk M, Biecek P, Polak N, Smolarczyk K, Dudek MR, Cebrat S: **The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms**. *BMC Genomics* 2007, **8**:163.
22. Chan P, Lovric J, Warwicker J: **Subcellular pH and predicted pH-dependent features of proteins**. *Proteomics* 2006, **6**(12):3494-3501.
23. Castillo-Davis CI, Hartl DL: **GeneMerge--post-genomic analysis, data mining, and hypothesis testing**. *Bioinformatics* 2003, **19**(7):891-892.
24. Kankainen M, Brader G, Toronen P, Palva ET, Holm L: **Identifying functional gene sets from hierarchically clustered expression data: map of abiotic stress regulated genes in Arabidopsis thaliana**. *Nucleic Acids Res* 2006, **34**(18):e124.
25. Blandin G, Durrens P, Tekaia F, Aigle M, Bolotin-Fukuhara M, Bon E, Casaregola S, de Montigny J, Gaillardin C, Lepingle A *et al*: **Genomic exploration of the hemiascomycetous yeasts: 4. The genome of Saccharomyces cerevisiae revisited**. *FEBS Lett* 2000, **487**(1):31-36.
26. Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, Hebert JM, Hernandez-Boussard T, Jin H, Kaloper M, Matese JC *et al*: **The Stanford Microarray Database: data access and quality assessment tools**. *Nucleic Acids Res* 2003, **31**(1):94-96.
27. **Stanford Microarray Database [http://genome-www5.stanford.edu/]**
28. Krogan NJ, Peng WT, Cagney G, Robinson MD, Haw R, Zhong G, Guo X, Zhang X, Canadien V, Richards DP *et al*: **High-definition macromolecular composition of yeast RNA-processing complexes**. *Mol Cell* 2004, **13**(2):225-239.
29. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K *et al*: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry**. *Nature* 2002, **415**(6868):180-183.

30. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM *et al*: **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature* 2002, **415**(6868):141-147.
31. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14863-14868.
32. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**(6):276-277.
33. **MIPS website [http://mips.gsf.de/genre/proj/yeast/]**
34. Danchin A: **La Barque de Delphes - Ce que révèle le texte des génomes**. Paris, France: Odile Jacob; 1998.
35. **BlastSets tool [http://cbi.labri.fr/outils/BlastSets/index.php]**

Figure legends

Figure 1 - Evaluation of different representations for a biological criterion: experimental setup

For the assessment of the representations of an assessed biological criterion, a reference criterion is used. The collection of sets created for each representation is compared to the collection of sets of the reference criterion. The dashed lines between sets of the assessed and the reference criterion correspond to a significant similarity. In this example, more sets from the reference criteria were found to be similar to the assessed criterion from collection 1 than from collection 2.

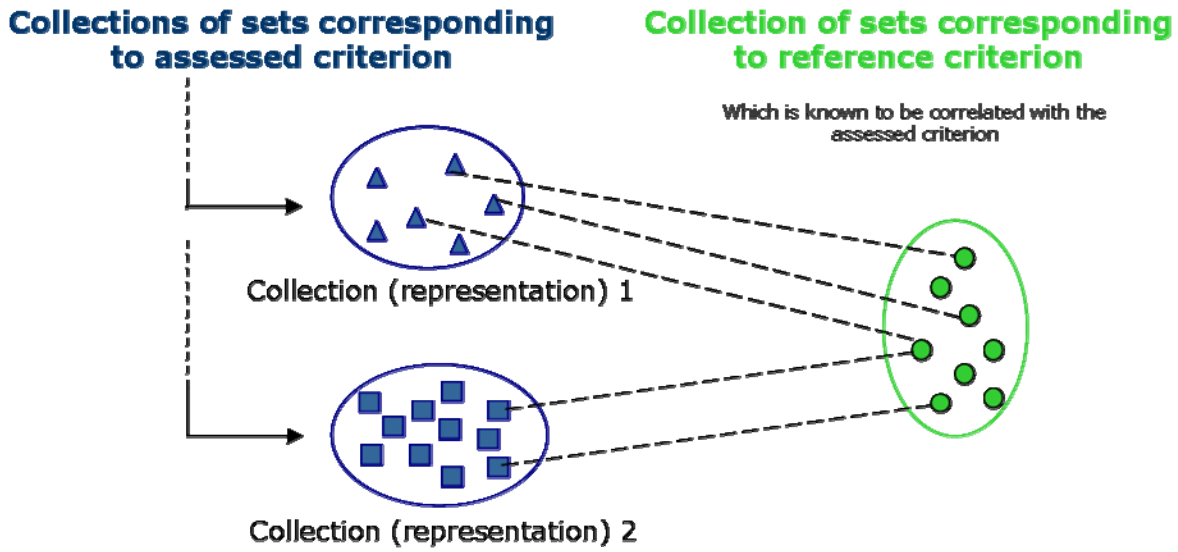


Figure 2 - Results of the comparison between Complexes and 'Best Neighbours nested' collections

This graph shows the number of complexes found to show a significant similarity with an expression set from a collection derived from the 'Best Neighbours Nested' clustering (blue for the Spellman experiment, and purple for the Gasch experiment). The names of the BNN collections (on the X axis) correspond to the range of set sizes present in each collection: from 5 to 50, 5 to 60, 70, 80, 90, 100, 110 and 120 respectively (see Methods section for details). Best results were obtained with a threshold of 100. This clustering (BNN 100) was retained for further comparison with hierarchical clustering.

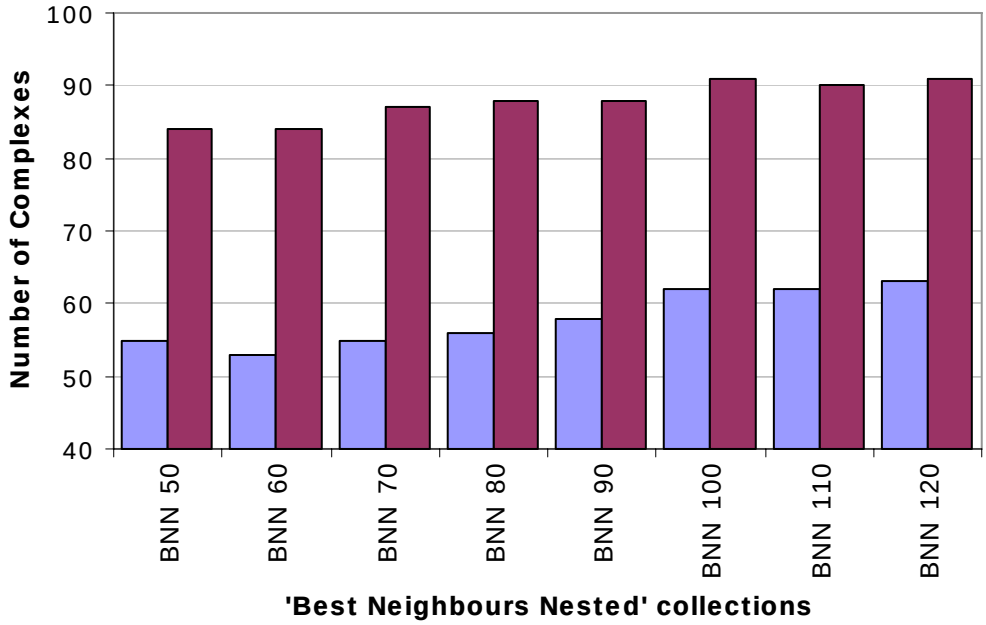


Figure 3 – Results of the comparison between Complexes and 'Best Neighbours Single' collections

This graph shows the number of complexes found to show a significant similarity with an expression set from a collection derived from the 'Best Neighbours Single' clustering (blue for the Spellman experiment, purple for the Gasch experiment). The names of the BNS collections (on the X axis) correspond to different set sizes: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110 and 120 respectively (see Methods section for details).

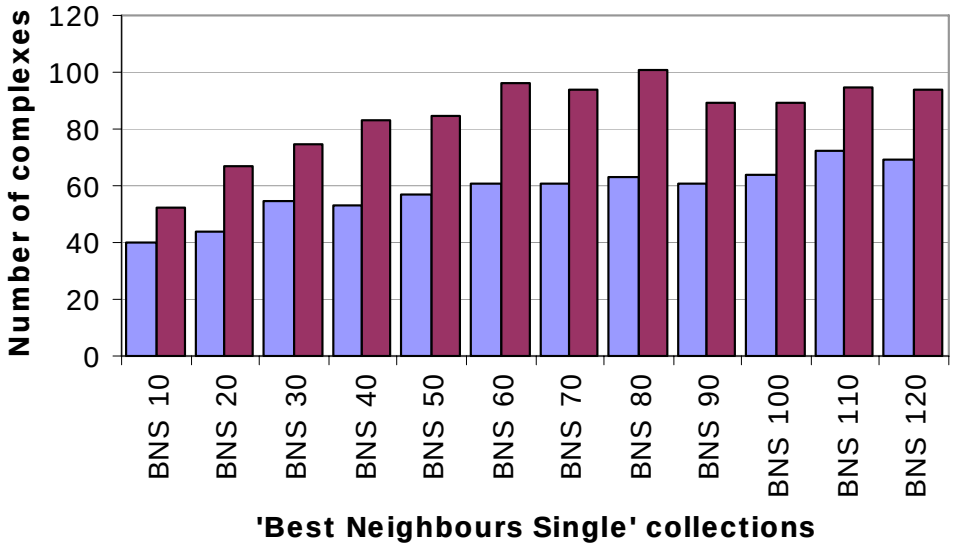


Figure 4 - Distribution of proteins isoelectric points (pI) in yeast proteome

This frequency histogram shows the pI values of the *Saccharomyces cereviasiae* proteome, calculated with the EMBOSS package 'iep'. Three peaks appear at pI ~ 4.8, 6.4 and 10 as has already been described for eukaryotes.

This distribution also corresponds to the size (number of proteins) of sets created in the flat and lattice collections based on sets covering a fixed range of pI (pI r0.5 flat and pI r0.5 lattice).

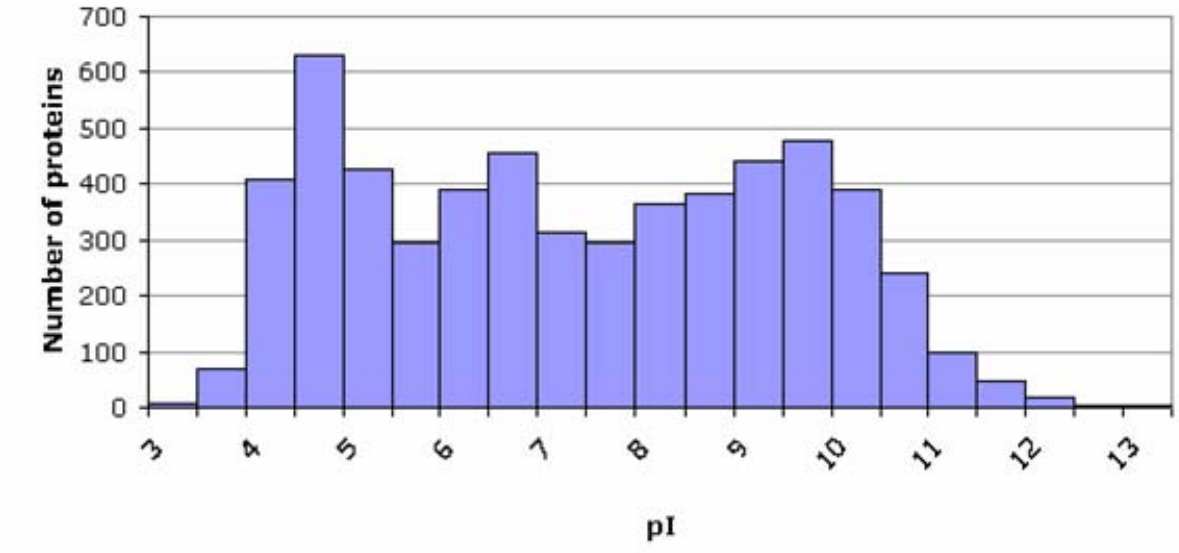


Figure 5 - Representations of expression data as set collections

A) Hierarchical clustering: example showing the principle of hierarchical clustering on microarray data. Each line represents a gene and each column corresponds to a condition. Hierarchical clustering results in a binary tree, each node (red circles) represents a set of potentially co-expressed genes (**A:** YGL125w, YHR037w; **B:** YGL125w, YHR037w, YOR236w; **C:** YGL125w, YHR037w, YOR236w, YIR034c; **D:** YGL125w, YHR037w, YOR236w, YIR034c, YNR050c).

B) 'Best Neighbours' clustering: for each gene, taken as a reference (only 4 are shown here), sets of genes of size *N* are built. These sets contain the best neighbours (genes with the highest correlation coefficients) for the reference gene which is itself included in all sets.

As examples, the genes encircled in black are the 5 best neighbours and in blue the 10 best neighbours.

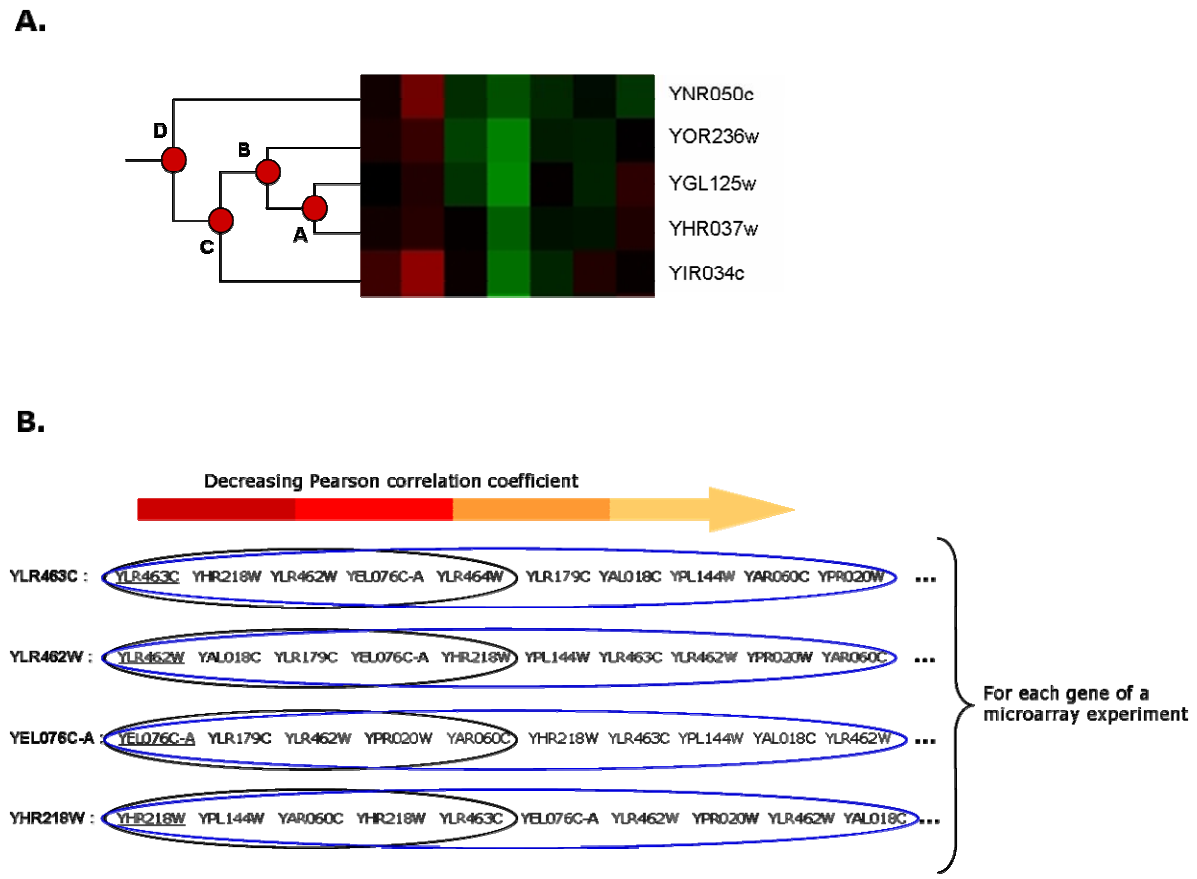
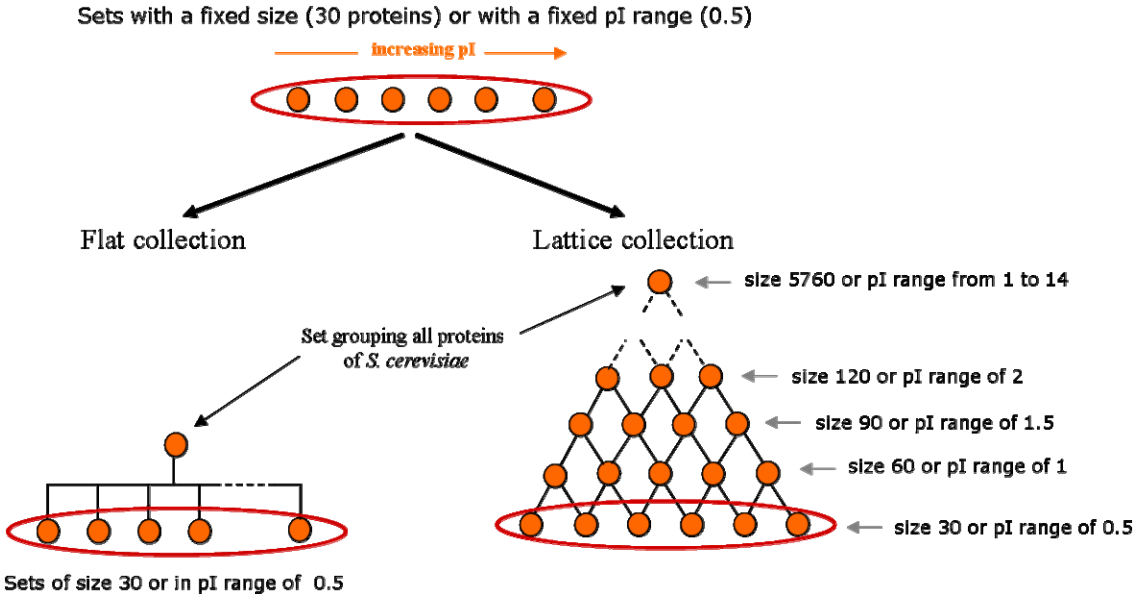


Figure 6 - Representations of protein Isoelectric points (pI) as set collections
This schema explains how proteins were clustered based on their isoelectric points. At the top, orange circles in the red oval represent sorted sets of proteins covering adjacent ranges of isoelectric points: those sets are defined either based on fixed size (30 proteins per set) or based on a fixed width range of pI (0.5 per set). Each of those initial collections of sets is either kept as such, leading to the "Flat collection" shown on the left side of the figure, or recursively aggregated to form a lattice as shown on the right side. As a result 4 collections of sets were built: "pI s30 flat" corresponds to a collection of adjacent sets of size 30, "pI s30 lattice" corresponds to the same collection recursively merged to form a lattice, "pI r0.5 flat" corresponds to a collection of adjacent sets in a range of pI of 0.5, and "pI r0.5 lattice" corresponds to the same collection recursively merged to form a lattice.



Tables

Table 1 - Complexes found to be similar to expression sets
This table shows the number of complexes, among the 1059 complexes described or randomised, having a similarity with an expression set derived either through Hierarchical Clustering or through 'Best Neighbours Nested' clustering (‘BNN 100’).

	Spellman experiment		Gasch experiment	
	Hierarchical clustering	**BNN 100**	**Hierarchical clustering**	**BNN 100**
MIPS complexes (1059)	48	62	56	91
Random complexes (1059)	0	0	0	0

Table 2 - Number of sets derived from expression data
For two microarray experiments (Spellman et al., 1998, and Gasch et al., 2000), the table shows the number of genes from the microarray that are also present in the curated list of genes from the Génolevures Consortium [25], and the sets built when using the ‘Best Neighbours nested’ clustering with different thresholds (see Methods section for details). Hierarchical Clustering produces 5629 and 5648 sets for Spellman and Gasch experiments respectively.

Methods of clustering	**Spellman *et al.***	**Gasch *et al.***
Number of genes	5630	5649
Hierarchical Clustering	5629	5648
BNS	5630	5649
BNN 50	56300	56490
BNN 60	67560	67788
BNN 70	78820	79086
BNN 80	90080	90384
BNN 90	101340	101682
BNN 100	112600	112980
BNN 110	123860	124278
BNN 120	135120	135576

Table 3 - Protein sets assigned to cellular compartments found similar to a set corresponding to a range of pI

This table shows the subcellular compartments (among the 50 that were used) for which the assigned set of proteins was found to have a significant similarity with a set derived from pI ranges (crosses). Different clustering for building sets of pI have been used: “pI s30 flat”, “pI s30 lattice”, “pI r0.5 flat” and “pI r0.5 lattice” (see Methods section for details). For each clustering, the total number of compartments found to be similar to a range of pI is given on the last line.

Cellular compartments	**pI clustering**			
	pI s30 flat	**pI s30 lattice**	**pI r0.5 flat**	**pI r0.5 lattice**
Vacuole			x	x
Nucleus		x	x	x
Mitochondrial matrix		x		x
Mitochondrial inner membrane	x	x	x	x
Mitochondria		x	x	x
Extracellular		x	x	x
Cytoplasmic ribosomal proteins	x	x	x	x
Cytoplasm		x	x	x
Cell wall	x	x	x	x
Nucleolus				x
Cytoskeleton				x
Plasma membrane				x
TOTAL	3	8	8	12

Table 4 - Details of similarities found between the composition of cellular compartments and sets built using the "pI r0.5 lattice" clustering

This table presents the results obtained when comparing the proteins of each subcellular compartment to proteins in the same range of pI in the "pI r0.5 lattice" collection. Only the most significant results for each compartment are shown.

Subcellular compartments	Number of proteins in the compartment	Range of pI	Number of proteins in the range of pI	Intersection	P-value
Cytoplasmic Ribosomal proteins	119	10.5-12.5	411	89	10e-79
Cytoplasm	2798	4-7	2603	1514	2x10e-40
Mitochondria	1019	8.5-11.5	2039	531	10e-33
- inner membrane	139	9-10.5	1309	84	4x10e-22
- matrix	69	10-11.5	742	27	3x10e-8
Cell wall	28	3.5-4.5	468	22	4x10e-15
Extracellular	51	3.5-5	1097	32	6x10e-12
Nucleus	2114	4-6.5	2145	892	10e-9
- nucleolus	208	9-10.5	1309	81	6x10e-8
Vacuole	277	3.5-5.5	1530	111	3x10e-7
Plasma Membrane	183	7-9	1355	71	2x10e-6
Cytoskeleton	203	4.5-6	1357	77	2x10e-6

Table 5 - Number of sets derived from pI data

The sequences of *Saccharomyces cerevisiae* proteins from Génolevures Consortium were used to compute the theoretical isoelectric point of each protein. Four clustering methods were then used, leading to four collections of sets of different sizes: "pI s30 flat" corresponds to a collection of adjacent sets of size 30, "pI s30 lattice" corresponds to the same collection recursively merged to form a lattice, "pI r0.5 flat" corresponds to a collection of adjacent sets in a range of pI of 0.5, and "pI r0.5 lattice" corresponds to the same collection recursively merged to form a lattice (see Figure 6 and Methods section for details).

	pI s30 flat	pI s30 lattice	pI r0.5 flat	pI r0.5 lattice
Number of sets	192	18528	21	231

Table 6 - Hierarchical organization of cellular compartments

The list of cellular compartments and sub-compartments is based on the MIPS SubCell database [3]. The number of distinct yeast proteins assigned to each compartment is given in parentheses.

Compartments	Sub compartments
cell periphery (216)	
cell wall (38)	
plasma membrane (183)	
integral membrane not assigned to a specific membrane (172)	
cytoplasm (2798)	cytoplasmic ribosomal proteins (119)
cytoskeleton (203)	tubulin cytoskeleton (36); spindle pole body (82); intermediate filament (3); actin cytoskeleton (55)
endoplasmic reticulum (549)	ER lumen (9); ER membrane (129)
golgi (132)	golgi membrane (59)
transport vesicles (138)	ER-golgi transport vesicles (14); golgi-ER transport vesicles (60); inter-golgi transport vesicles (3); golgi-plasma membrane transport vesicles (4); golgi-vacuole transport vesicles (55); endocytotic transport vesicles (4); other transport vesicles (2)
nucleus (2114)	nuclear matrix (37); nucleolus (208); nuclear pore (44); chromosome structure (43); nuclear envelope (165)
mitochondria (1019)	mitochondrial inner membrane (139); mitochondrial outer membrane (20); mitochondrial matrix (69); mitochondrial intermembrane space (13)
vacuole (277)	vacuolar membrane (97); vacuolar lumen (6)
peroxisome (52)	peroxisomal membrane (20); peroxisomal matrix (19)
endosome (57)	late endosome (1)
microsome (5)	
lipid particles (26)	
bud (148)	neck (112); bud tip (15)
extracellular (51)	

3. Article n°2

"Observing metabolic functions at the genome scale"

Research

Observing metabolic functions at the genome scale

Jean-Marc Schwartz[¤*†], Claire Gaugain[¤‡], Jose C Nacher[*§], Antoine de Daruvar[‡] and Minoru Kanehisa[*]

Addresses: [*]Bioinformatics Center, Kyoto University, Uji, Kyoto 611-0011, Japan. [†]Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK. [‡]Centre de Bioinformatique de Bordeaux, Université Bordeaux 2, 33076 Bordeaux, France. [§]Department of Complex Systems, Future University, Hakodate, Hokkaido 041-8655, Japan.

¤ These authors contributed equally to this work.

Correspondence: Jean-Marc Schwartz. Email: jean-marc.schwartz@manchester.ac.uk

Published: 26 June 2007

Genome ***Biology*** 2007, **8:**R123 (doi:10.1186/gb-2007-8-6-r123)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/6/R123

Received: 21 March 2007
Revised: 30 May 2007
Accepted: 26 June 2007

Abstract

Background: High-throughput techniques have multiplied the amount and the types of available biological data, and for the first time achieving a global comprehension of the physiology of biological cells has become an achievable goal. This aim requires the integration of large amounts of heterogeneous data at different scales. It is notably necessary to extend the traditional focus on genomic data towards a truly functional focus, where the activity of cells is described in terms of actual metabolic processes performing the functions necessary for cells to live.

Results: In this work, we present a new approach for metabolic analysis that allows us to observe the transcriptional activity of metabolic functions at the genome scale. These functions are described in terms of elementary modes, which can be computed in a genome-scale model thanks to a modular approach. We exemplify this new perspective by presenting a detailed analysis of the transcriptional metabolic response of yeast cells to stress. The integration of elementary mode analysis with gene expression data allows us to identify a number of functionally induced or repressed metabolic processes in different stress conditions. The assembly of these elementary modes leads to the identification of specific metabolic backbones.

Conclusion: This study opens a new framework for the cell-scale analysis of metabolism, where transcriptional activity can be analyzed in terms of whole processes instead of individual genes. We furthermore show that the set of active elementary modes exhibits a highly uneven organization, where most of them conduct specialized tasks while a smaller proportion performs multi-task functions and dominates the general stress response.

Background

The increasing availability of high-throughput data has allowed more and more analyses to be performed at the cell scale. After completion of genome sequencing for many species, the focus is shifting towards getting a global understanding of cell physiology. This task requires the integration of heterogeneous data at different scales, including genomic, transcriptomic, proteomic, and metabolomic data.

At the level of metabolism, good knowledge of the structure of metabolic networks has now been achieved for several species. A number of genome-wide models of metabolism have been reconstructed [1-4], but these structural models provide only a static representation of an organism's metabolism; the structure of a metabolic network is static for a given species, and only changes at a slow pace across species through evolution [5]. However, the usage of particular metabolic reactions by a given cell is highly dynamic. It changes very rapidly in time with modifications in the environment, in the cell cycle, or with stochastic fluctuations. Static representations, therefore, need to be extended toward truly dynamic descriptions.

Metabolic networks are also highly complex, formed by several hundreds of densely interconnected chemical reactions. To characterize such complex systems at the genome scale, it is necessary to identify smaller building blocks. Cellular networks have been shown to have a high degree of modularity, and are composed of groups of interacting elements and molecules that carry out specific biological functions [6]. In recent years, several methods have been proposed to decompose complex biological networks into subnetworks and to identify basic interaction modules [5,7-9]. Although relevant progress has been achieved in detecting motifs and modules in transcriptional regulatory and protein-protein interaction networks [10-16], the building blocks of metabolic pathways still remain largely undiscovered. Evidence for the existence of modularity in metabolic pathways was recently proposed by Ravasz *et al.* [17], who showed that the high clustering degree observed in metabolic networks may imply a hierarchical modularity, in which modules are made up of smaller and denser modules in a fractal manner.

A complementary approach is provided by the concept of an 'elementary mode'. Elementary modes, and the very similar concept of 'extreme pathways', are minimal sets of reactions that can operate in steady state in a metabolic network [18-20]. They have already proven useful for studying many aspects of metabolism, including the prediction of functional properties of metabolic pathways, the measurement of robustness and flexibility, inferring the viability of mutants, the assessment of gene regulatory features, and so on [21]. Recently, it has been shown that they could even provide a basis for describing and understanding the properties of signaling and transcriptional regulatory networks [22,23]. All these applications, however, consider elementary modes as purely 'structural units'. Although the biological significance of elementary modes has already been mentioned [24], the use of elementary modes as true elementary 'functional units' of cellular metabolism has not been attempted so far. A few studies [25,26] have combined metabolic and transcriptomic data in order to find out whether co-expressed genes are part of a given metabolic pathway, but most of these approaches used complete metabolic pathways as metabolic units.

Here, we address the problem of identifying metabolic units in a genome-scale model of the yeast *Saccharomyces cerevisiae* by relying on elementary modes. Our study is based on the integration of dynamic gene expression data in various stress conditions into a genome-scale model of metabolism, modularly structured in elementary modes. We used a bioinformatics tool called BlastSets [27] to combine these two types of data in order to answer the following question: do enzymes that are involved in the same elementary mode have their corresponding genes co-expressed in particular conditions? We were able to identify active elementary modes, that is, elementary modes whose enzymes are induced or repressed in response to different environmental stresses; these elementary modes can thus be seen as functional units of the metabolic stress response.

Results

Genome-wide computation of elementary modes

The computation of elementary modes in genome-wide models of metabolism is seriously hampered by the problem of combinatorial explosion. Even though the number of elementary modes is usually smaller in a real system than its theoretical limit and can be further reduced by taking into account various environmental or regulatory constraints, it is of no practical use to handle systems of thousands of elementary modes because such systems become impossible to interpret [28,29]. One possible approach to deal with this problem consists of decomposing a genome-scale metabolic network into smaller subunits. This kind of decomposition has already been proposed, but was based on network topology [30]; it consisted of finding the optimal decomposition that minimized the number of elementary modes. However, there is no guarantee that such subunits represent functionally coherent and biologically interpretable pathways.

We have developed an alternative approach for computing elementary modes at the genome scale. In the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, metabolic pathways are represented as a series of maps, where each map covers a precise biological function [31]. These maps are sufficiently small for the number of elementary modes inside each of them to remain in the hundreds (Table 1). Furthermore, because they have been manually drawn and annotated based on biological information, these units have a clear biological meaning and are easy to interpret. We thus considered each pathway map of the KEGG database as one subnetwork. We then computed the full set of elementary modes inside each of them using a classical algorithm [20] (Additional data file 1).

Because of their combinatorial nature, a number of different elementary modes usually share common reactions along their path. It often occurs that several elementary modes are almost identical except for a few branches at their extremities. Similarly, a given reaction can belong to a large number of

different elementary modes. Figure 1a illustrates this property by showing some of the elementary modes between fumarate and 2-oxoglutarate in the citrate cycle (note that only 7 elementary modes have been drawn out of 99 calculated for the entire citrate cycle map). This combinatorial property, which is a major problem in large networks, is, on the contrary, welcome in our study: as our aim is to search for the most active route in a system, it guarantees that the full set of topologically possible routes will be considered in the search.

The use of KEGG maps for defining subnetworks aims at having entities that are as much as possible biologically coherent. The start and end points of elementary modes are compounds located at the boundaries between subnetworks. One drawback of this approach is that active metabolic routes that are spread over different KEGG maps may not be easily identified. To overcome this problem, we constructed two different collections of elementary modes, EM1 and EM2. EM1 contains the full set of single elementary modes computed with each KEGG pathway map being used as a subnetwork; each elementary mode from EM1 is entirely included in a single pathway map. EM2 was formed by combining all pairs of elementary modes from EM1 that are connected through a common boundary compound; elementary modes from EM2 thus spread over two different pathway maps (Figure 1b). The use of EM2 reduces the dependence of results on subnetwork boundaries since active elementary modes spread over different KEGG maps can now be identified. More details are provided in the 'Genome-wide computation of elementary modes' section in Materials and methods, and the full description of single elementary modes is available in Additional data file 1.

Elementary modes represent true functional units of metabolism

Functional activity is more significant in elementary modes than in entire pathways

To elucidate whether elementary modes can be considered as true functional biological units, the stress response of yeast was investigated in a large number of different conditions. Towards this goal, we used microarray data obtained from several experimental analyses [32-34] (see the 'Expression data' section in Materials and methods) and a bioinformatics tool called BlastSets [27]. BlastSets enabled us to find similarities between the composition of two sets of genes or proteins derived from two different types of information (here, metabolic pathways and expression data). The elementary modes EM1 and EM2 were stored independently as two BlastSets collections. Entire KEGG pathways were also stored as a BlastSets collection, to find out whether stress responses involve entire pathways, as defined in KEGG, or only parts of these pathways, as represented by elementary modes. In many stress conditions, induced/repressed elementary modes were found with higher P values than whole pathways (Table 2).

The numbers of detected induced/repressed elementary modes for each stress condition are shown in Table 3, as well as the number of different KEGG pathways these elementary modes belong to. The numbers obtained with EM1 and EM2 are relatively well correlated but there is no absolute relationship between them; in most cases, the number of induced/repressed elementary modes is increased when compared to EM2, but a few of them show higher numbers with EM1. The same observation can be made about the number of KEGG pathways to which these elementary modes belong. In a majority of cases, elementary modes detected with EM1 are concentrated in a relatively small number of pathways, and EM2 increases this number by adding modes from adjacent pathways. But in a few cases, for example Thiuram, the number of pathways detected with EM2 is smaller than with EM1, indicating that these elementary modes tend to be isolated and poorly connected to adjacent pathways.

Examples of elementary modes induced in particular stress conditions are shown in Figure 2, including an induced elementary mode in the citrate cycle during stationary phase, and another induced one in sulfur metabolism in response to tetrachloro-isophthalonitrile exposure. The sets of induced enzymes detected by BlastSets are indeed highly connected. Fewer elementary modes could be identified from the sets of repressed enzymes and they are usually less connected, meaning that repressed enzymes are more dispersed in the mode. This fact has already been mentioned by Wei *et al.* [35] for the genetic model plant *Arabidopsis thaliana*, who observed that induced genes in the same metabolic pathway tend to be close and well connected to each other, while repressed genes are more distant.

Induced/repressed elementary modes are statistically significant

BlastSets applies a stringent threshold on *P* values (*P* value must be lower than 6.0×10^{-5} for EM1 and 3.4×10^{-6} for EM2; see 'Description of BlastSets' section in Materials and methods), which should already guarantee that identified elementary modes are statistically significant. Nevertheless, in order to further assess the reliability of our results, we created random gene expression values by random permutation of gene expression values in several stress responses. These random sets of induced/repressed genes were compared to elementary modes in BlastSets, in the same way as for stress-induced/repressed genes. No active elementary mode was identified using these random sets. The procedure was repeated for several conditions, always with the same result. This finding confirms that elementary modes found to be active in specific environmental stress conditions have a high statistical significance.

Pairing elementary modes to reconstruct induced/repressed routes

To identify complete metabolic routes that are spread over several KEGG pathway maps, we constructed the EM2 collection containing elementary modes grouped in pairs. Two elementary modes are grouped as a set in EM2 if they share a

Table 1

KEGG metabolic pathways for *Saccharomyces cerevisiae* and number of elementary modes for each

Pathway identifier	Pathway name	Number of computed elementary modes	in BlastSets
sce00010	Glycolysis/gluconeogenesis	163	112
sce00020	Citrate cycle (TCA cycle)	99	60
sce00030	Pentose phosphate pathway	206	203
sce00040	Pentose and glucuronate interconversions	4	2
sce00051	Fructose and mannose metabolism	12	11
sce00052	Galactose metabolism	81	63
sce00053	Ascorbate and aldarate metabolism	2	2
sce00061	Fatty acid biosynthesis	4	3
sce00071	Fatty acid metabolism	22	20
sce00072	Synthesis and degradation of ketone bodies	4	2
sce00100	Biosynthesis of steroids	6	5
sce00120	Bile acid biosynthesis	5	4
sce00130	Ubiquinone biosynthesis	4	1
sce00190	Oxidative phosphorylation	7	7
sce00220	Urea cycle and metabolism of amino groups	12	11
sce00230	Purine metabolism	350	346
sce00240	Pyrimidine metabolism	31	28
sce00251	Glutamate metabolism	40	38
sce00252	Alanine and aspartate metabolism	43	39
sce00260	Glycine, serine and threonine metabolism	102	94
sce00271	Methionine metabolism	26	25
sce00272	Cysteine metabolism	14	12
sce00280	Valine, leucine and isoleucine degradation	8	7
sce00290	Valine, leucine and isoleucine biosynthesis	12	11
sce00300	Lysine biosynthesis	5	4
sce00310	Lysine degradation	6	5
sce00330	Arginine and proline metabolism	29	24
sce00340	Histidine metabolism	5	4
sce00350	Tyrosine metabolism	11	8
sce00360	Phenylalanine metabolism	3	3
sce00361	gamma-Hexachlorocyclohexane degradation	6	1
sce00362	Benzoate degradation via hydroxylation	3	0
sce00380	Tryptophan metabolism	15	8
sce00400	Phenylalanine, tyrosine and tryptophan biosynthesis	38	30
sce00401	Novobiocin biosynthesis	6	2
sce00410	beta-Alanine metabolism	6	6
sce00430	Taurine and hypotaurine metabolism	2	1
sce00440	Aminophosphonate metabolism	5	3
sce00450	Selenoamino acid metabolism	6	5
sce00460	Cyanoamino acid metabolism	9	2

Table 1 *(Continued)*

KEGG metabolic pathways for *Saccharomyces cerevisiae* and number of elementary modes for each

sce00480	Glutathione metabolism	5	4
sce00500	Starch and sucrose metabolism	49	47
sce00520	Nucleotide sugars metabolism	15	11
sce00521	Streptomycin biosynthesis	2	1
sce00530	Aminosugars metabolism	13	13
sce00550	Peptidoglycan biosynthesis	3	0
sce00561	Glycerolipid metabolism	7	4
sce00562	Inositol phosphate metabolism	5	4
sce00563	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	3	0
sce00564	Glycerophospholipid metabolism	28	25
sce00590	Arachidonic acid metabolism	4	2
sce00600	Glycosphingolipid metabolism	7	5
sce00620	Pyruvate metabolism	139	132
sce00624	1- and 2-Methylnaphthalene degradation	7	3
sce00625	Tetrachloroethene degradation	4	1
sce00627	1,4-Dichlorobenzene degradation	9	0
sce00630	Glyoxylate and dicarboxylate metabolism	7	6
sce00632	Benzoate degradation via CoA ligation	7	2
sce00640	Propanoate metabolism	8	4
sce00650	Butanoate metabolism	9	7
sce00670	One carbon pool by folate	13	12
sce00680	Methane metabolism	5	3
sce00710	Carbon fixation	13	8
sce00720	Reductive carboxylate cycle (CO_2 fixation)	3	3
sce00730	Thiamine metabolism	2	0
sce00740	Riboflavin metabolism	3	2
sce00750	Vitamin B6 metabolism	4	2
sce00760	Nicotinate and nicotinamide metabolism	9	8
sce00770	Pantothenate and CoA biosynthesis	4	3
sce00780	Biotin metabolism	1	1
sce00790	Folate biosynthesis	17	6
sce00860	Porphyrin and chlorophyll metabolism	4	3
sce00900	Terpenoid biosynthesis	9	8
sce00903	Limonene and pinene degradation	9	2
sce00910	Nitrogen metabolism	17	15
sce00920	Sulfur metabolism	3	2
sce00960	Alkaloid biosynthesis II	3	3
sce00970	Aminoacyl-tRNA biosynthesis	20	15
sce00980	Metabolism of xenobiotics by cytochrome P450	2	2
sce04070	Phosphatidylinositol signaling system	4	4

The first and second columns give the identifier and the name of each KEGG metabolic pathway. For each of them, the number of elementary modes computed is indicated in the third column and the number of elementary modes entered in the BlastSets database in the fourth column. In most cases, there is a difference between these two numbers because BlastSets eliminates redundant elementary modes and the ones involving only one enzyme.

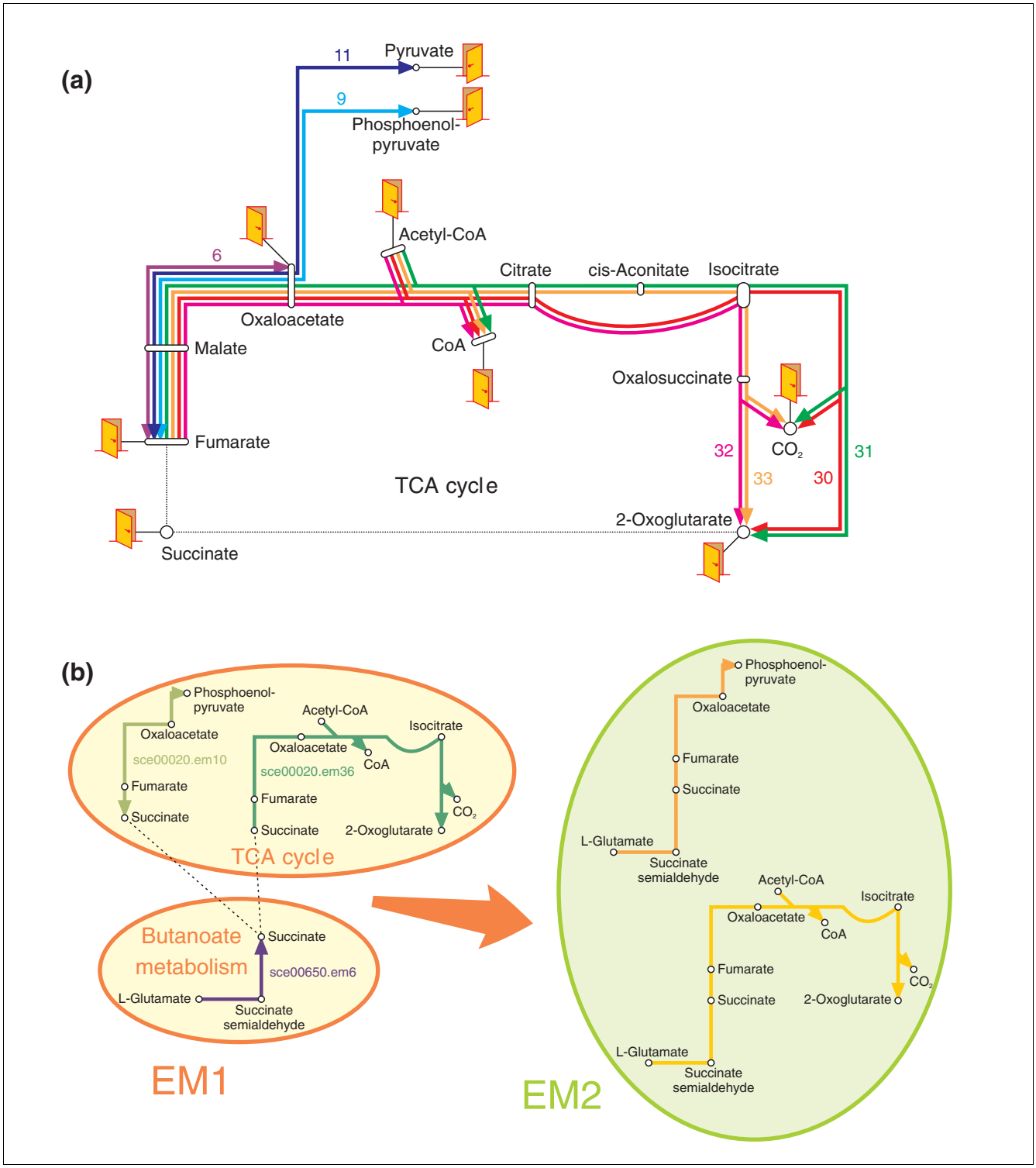


Figure 1
Construction of elementary mode collections. **(a)** This scheme represents some of the elementary modes calculated between fumarate and 2-oxoglutarate in the citrate cycle pathway. Each color corresponds to a different elementary mode; numbers indicate the identifiers of elementary modes as in Additional data file 1, and doors represent start and end compounds of elementary modes. This figure illustrates the combinatorial nature of elementary modes: several of them are almost identical except for one or two reactions, and a given reaction can belong to several elementary modes. **(b)** The composition of the EM1 collection (left) and how elementary modes were merged to build the EM2 collection (right). Three independent sets from EM1 can be merged into two sets in EM2 if they share a common boundary compound.

Table 2

First induced/repressed pathway and first induced/repressed elementary mode in particular stress conditions

Stress condition	First pathway	*P* value	First elementary mode (EM1)	*P* value
Ash [34], repressed	sce00230 (purine metabolism)	2.7e-8	sce00230.em279 (part of purine metabolism)	1e-11
Pentanol [34], repressed	sce00230 (purine metabolism)	3.3e-6	sce00230.em341 (part of purine metabolism)	1.8e-8
Tetrachloro-isophthalonitrile [34], repressed	sce00230 (purine metabolism)	2.5e-8	sce00230.em280 (part of purine metabolism)	3.3e-10
Stationary phase [33], induced	sce00020 (citrate cycle)	3.4e-14	sce00020.em36 (part of citrate cycle)	5.9e-16
Heat shock [32], induced	sce00500 (starch and sucrose metabolism)	3.8e-4	sce00500.em13 (part of starch and sucrose metabolism)	4.2e-6

Results given by BlastSets for particular conditions. The second column gives the most significant full KEGG pathway found to be induced/repressed (that is, the one with the lowest *P* value, given in the third column). The fourth column gives the most significant elementary mode from EM1 found to be induced/repressed. These results are sorted from the highest to the lowest difference between the two *P* values.

common boundary compound. These compounds act as bridges between individual pathway maps, enabling more extended induced/repressed routes to be identified by this approach.

In each stress situation, we could then infer a 'backbone' of induced/repressed metabolic routes. Backbones were constructed by selecting the pairs of elementary modes with the lowest *P* values and connecting them to each other, thanks to results from the EM2 collection (see 'Analysis of BlastSets results' section in Materials and methods). These backbones can be viewed as the main modules characterizing metabolic activity in terms of expression data in a given condition. They are provided for each individual condition in Additional data file 2.

Specialized and multitask elementary modes

To assess how the activity of elementary modes is distributed in response to a set of diverse environmental stresses, we computed the probability distribution $P(k)$ to find a given induced/repressed elementary mode in k stress conditions (Figure 3a). This distribution reveals a highly heterogeneous behavior: on one hand, a relatively low number of 'multitask' elementary modes are transcriptionally active in a large number of different conditions, while on the other hand, many 'specialized' elementary modes are active in a small number of conditions (less than three). About 77% of detected elementary modes appear to be conducting specialized tasks while the remaining 23% are involved in the more general stress response. This observed metabolic organization is far from a random distribution, where each induced/repressed elementary mode would have the same chance to be active in the vicinity of the average value. The deviation from a random distribution suggests that elementary modes involved in the stress response are governed by a more complex organization [36], that is, that they are organized into complex modules across the metabolic network.

Transcriptional activity of metabolic processes revealed by functional elementary modes

Map of elementary mode activities

It is possible to reveal the various patterns of stress responses by drawing the 'activity map' of elementary modes. In Figure 3b, each line represents an elementary mode and each column a stress condition; induced elementary modes are shown in red and repressed modes in green in this representation, which is deliberately chosen to look similar to a microarray. Indeed, in the same way a microarray represents a map of the transcriptional activity of individual genes, we are here able to construct a map of genome-scale elementary mode activities, revealing the transcriptional activity of entire metabolic processes. It is particularly clear on this map that most of the identified elementary modes are either only induced or only repressed. While the three repressed patterns are very similar, induced patterns are more diverse and very few elementary modes are induced over all conditions, confirming the trend revealed by the distribution in Figure 3a.

Two main classes of stress responses

Our approach is able to provide new insights about metabolic activity in terms of expression data in particular conditions. We analyzed the raw expression data obtained for each stress condition in order to see which stresses lead to similar responses; the clustering tree of stress conditions based on raw expression data is provided as Additional data file 3. Among the 31 different conditions we studied, 12 had a too weak transcriptional response for any induced or repressed elementary mode to be detected. We noticed that, among the remaining 19 conditions that produced a sufficiently strong response, stresses could be divided into two main classes, which we hence denote as 'toxic' and 'non-toxic'. The toxic stress class mostly includes exposure of cells to toxic chemicals and metals. The non-toxic class, on the contrary, mostly includes other types of stresses, such as temperature changes, osmotic shocks, nutrient starvation, and so on. The list of conditions assigned to each class is provided in Table 4.

The metabolic backbones inside each class show recurrent similarities, which allowed us to construct a common backbone for each class (Figure 4). The two classes show a clearly distinct global response and few elementary modes are induced in both backbones, with the exception of the citrate cycle and nucleotide sugar metabolism. In addition, we represented both classes by networks where each node corresponds to a metabolic pathway and each edge denotes that at least one pair of elementary modes spanning both pathways

Table 3

Number of induced/repressed elementary modes in each condition

Stress condition	Number of induced or repressed elementary modes (EM1)	Number of induced or repressed KEGG pathways (EM1)	Number of induced or repressed elementary modes (EM2)	Number of induced or repressed KEGG pathways (EM2)
Heat shock [32], induced	12	2	28	4
Heat shock [32], repressed	2	2	2	2
NaCl [32], induced	5	1	4	2
Peroxide [32], induced	16	10	3	2
Sorbitol [32], induced	1	1	30	2
Acid [32], induced	6	1	0	0
Amino acid starvation [33], induced	13	3	104	19
Diamide [33], induced	42	12	196	21
Peroxide [33], induced	6	2	3	2
Heat shock [33], induced	34	2	88	7
Nitrogen depletion [33], induced	2	2	13	6
Stationary phase [33], induced	54	5	292	25
Variable temperature [33], induced	20	3	57	7
Ash [34], induced	24	11	153	19
Ash [34], repressed	200	2	284	8
Cadmium [34], induced	1	1	19	5
Maneb [34], induced	17	11	193	21
Octanol [34], induced	5	2	12	6
Pentachlorophenol [34], induced	7	5	56	12
Pentanol [34], induced	44	7	289	35
Pentanol [34], repressed	184	2	166	7
Thiuram [34], induced	12	11	19	5
Tetrachloro-isophthalonitrile [34], induced	17	11	25	8
Tetrachloro-isophthalonitrile [34], repressed	155	1	202	8
Zineb [34], induced	16	10	127	19

This table shows the number of elementary modes found induced or repressed in each stress condition. These include all the results given by BlastSets independently of their *P* value. The numbers given in the fourth column are the numbers of individual elementary modes and not the numbers of pairs.

is present in a stress response (see 'Construction of toxic and non-toxic networks' section in Materials and methods). The toxic response network is shown in Figure 5a and exhibits two components. The inner component is composed of a group of strongly connected pathways centered on sulfur metabolism, pyruvate metabolism and lysine biosynthesis metabolism. These pathways thus have a strong tendency to be activated simultaneously. They constitute the core of the toxic stress response and cover most parts of the toxic backbone described previously. The external component, in contrast, is composed of a sparse network with thinner connections. In the non-toxic network this bi-component nature is less clear, but it is still possible to identify a more strongly connected central component containing starch and sucrose metabolism, the pentose phosphate pathway, glycolysis, and arginine and proline metabolism (Figure 5b).

Insights about specific stress conditions

In some cases, the observed transcriptional metabolic response confirms earlier findings. Vido et al. [37] reported that cadmium exposure increases the synthesis of cysteine and perhaps of glutathione, which is essential for cellular detoxification. The synthesis of these two compounds is possible through the activation of the sulfur amino acid pathway. We observe that, among the three elementary modes activated in response to cadmium exposure, two have cysteine as their final product, and among these two, one elementary mode is a part of cysteine metabolism and another is a part of sulfur metabolism. Cysteine is also one of the compounds produced in the general backbone of the response to toxic stresses (Figure 4a).

Amino acid starvation is known to activate the transcription factor Gcn4p, which induces genes involved in amino acid biosynthetic pathways, except the cysteine pathway [38], although the genes involved in the biosynthesis of cysteine precursors (homocysteine and serine) are induced. This is exactly what we observe in response to amino acid starvation: several elementary modes from amino acid biosynthetic pathways are activated but none from the cysteine pathway, even if some elementary modes from the cysteine pathway are linked to modes activated during amino acid starvation.

Genes induced in stationary-phase cultures of yeast are associated with mitochondrial functions, that is, aerobic respiration and the citrate cycle [39]. ATP synthesis is thus very

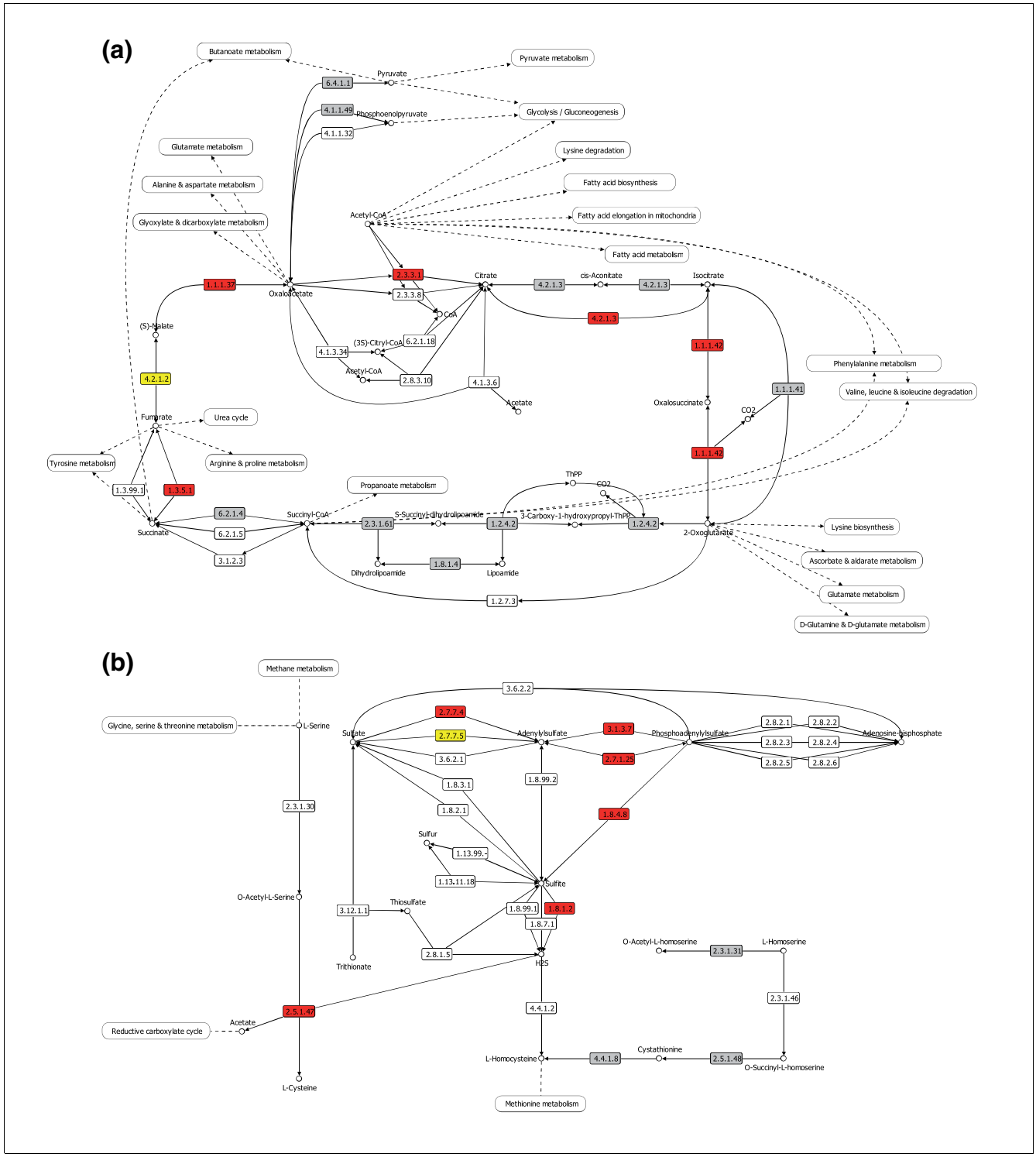


Figure 2
Examples of active elementary modes. **(a)** This figure shows the citrate cycle map from KEGG. Enzymes colored in red are coded by genes induced during the stationary phase. They correspond exactly to elementary mode number 36 of the citrate cycle, with the exception of one enzyme in yellow (4.2.1.2). **(b)** The sulfur metabolism map from KEGG. Enzymes colored in red are coded by genes found induced when yeast is exposed to tetrachloro-isophthalonitrile. These enzymes compose the entire elementary mode number 3 with the exception of two of them (in yellow): YGR012W is not induced but YLR303W is induced and fulfils the same function (EC 2.5.1.47); in the second case, two enzymes can fulfill the same function, so even if one is missing, the other completes the metabolic route (EC 2.7.7.5 and EC 2.7.7.4). Enzymes in grey are present in *S. cerevisiae* but do not belong to the elementary mode.

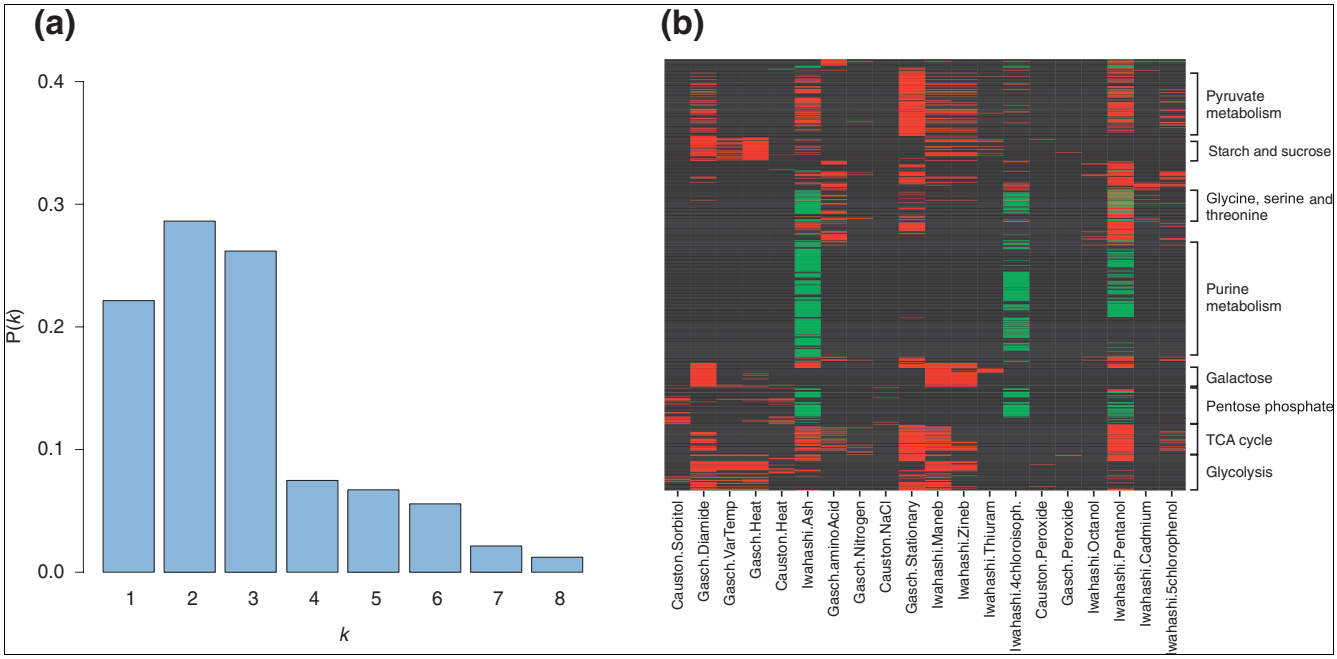


Figure 3
Transcriptional activity of elementary modes. **(a)** This histogram shows the probability of finding a given elementary mode induced/repressed in *k* stress conditions. **(b)** Map of genome-scale elementary mode activities. Each line of this figure corresponds to an elementary mode and each column to a stress condition. Repressed elementary modes are represented in green and induced modes in red.

important for yeast in the stationary phase. In our results, the elementary modes activated during the stationary phase are part of metabolic pathways linked to aerobic respiration, including glycolysis, the citrate cycle, pyruvate metabolism and oxidative phosphorylation.

Trehalose and glycerol are produced in large amounts by cells in stress situations [40]. Schade *et al.* [40] have shown that there is an overlap between the late cold response and the environmental stress response. This response corresponds to the production of glycerol and trehalose. This is what we observed in the general non-toxic backbone response (Figure 4b): glycerol is produced just a few reactions after glycerone

Table 4

Composition of toxic and non-toxic stress classes

Toxic class	Non-toxic class	Not assigned
Peroxide [32]	Sorbitol [32]	Alkali [33]
Cadmium [34]	NaCl [32]	Dithiothreitol [33]
Maneb [34]	Acid [32]	Diauxic shift [33]
Octanol [34]	Heat shock [32]	Alternative carbon [33]
Pentachlorophenol [34]	Amino acid starvation [33]	Hypo-osmotic [33]
Pentanol [34]	Diamide [33]	Menadione [34]
Thiuram [34]	Nitrogen depletion [33]	n-Pentane [34]
Tetrachloro-isophthalonitrile [34]	Stationary phase [33]	Ethanol [34]
Zineb [34]	Variable temperature [33]	Sodium n-dodecyl benzosulfonate [34]
	Ash [34]	Sodium lauryl sulfate [34]
		Capsaicin [34]
		Trichlorophenol [34]

Composition of the toxic and non-toxic stress classes, determined from the clustering tree of stress responses. The third column contains conditions whose response was too weak for any elementary mode to be identified by BlastSets.

phosphate, and trehalose is present one step before D-glucose in the starch and sucrose metabolism KEGG map (the only reason why it cannot appear as an end product in our study is that it is not a boundary compound in KEGG maps). These examples, confirming previously observed results, enable us to be confident in the identification of metabolic processes found to be induced/repressed in response to other stress conditions.

Discussion

There have been growing developments in recent years towards a more systems-level approach for understanding living organisms. On one side, microarray technologies have generalized the study of the transcriptome of biological cells in various conditions, and on the other side, numerous efforts have been undertaken to construct and describe the properties of metabolic networks at the genome scale. It is timely, therefore, to integrate both efforts and move towards a genome-scale analysis of cell metabolism.

At the same time, it is believed that a better understanding of the metabolome will be an important step towards improving the efficiency of the drug discovery process [41]. Instead of concentrating on the 'genomic universe', that is, the levels of gene regulation and transcription, our approach shifts the focus to the 'biochemical universe', that is, the small molecules or metabolites that actually perform biological functions and allow organisms to live and thrive. This shift is symbolized by the microarray-style representation of Figure 3b, which instead of showing the transcriptomic activities of individual genes, displays the transcriptomic activity of entire metabolic functions, represented by elementary modes, at the genome-scale. Although this is still a long way from an accurate and quantitative representation of the actual metabolic activity of a whole cell, which would require metabolic flux measurements, we believe that this shift opens a new perspective with a wide range of potential applications.

A major challenge addressed in this work consisted of embedding a suitable modularity into the highly complex and interconnected structure of metabolic networks. Our approach for computing elementary modes at the genome scale using KEGG pathway maps presents a number of advantages. These maps provide a decomposition of the metabolic network into well-defined subnetworks, which are biologically coherent and easy to interpret. Each map is sufficiently small for the number of elementary modes to remain in the hundreds, thus avoiding the necessity of having to cope with the problem of combinatorial explosion of elementary modes in large systems. Furthermore, these maps provide a manually curated representation of metabolic pathways where most secondary metabolites have been removed, thus avoiding the need to use complex procedures to identify principal metabolic routes and to eliminate invalid metabolic connections.

Microarray experiments are subject to a number of factors and we observed discrepancies in data obtained by different authors in similar conditions (peroxide treatment and heat shock experiments are available from both Gasch *et al.* [33] and Causton *et al.* [32]) The question of reproducibility of microarray experiments has been recurrent, but large-scale cross-platform experiments have shown that microarray data are indeed reliable and reproducible when adequate care is taken in experimental design and data treatment [42]. Differences may indicate that the transcription of genes involved in the metabolic response to stress is finely regulated and can fluctuate depending on a large number of factors.

Challenges also remain to obtain a more accurate description of the transcriptional activity of elementary modes in a cell. Our approach can be seen as 'discrete', since an elementary mode can only be assigned to three possible categories, that is, induced, repressed, or inactive. This division into three categories relies on a threshold on expression fold-change values, but enhanced statistical approaches could be researched to obtain a more subtle classification and avoid the need to set a threshold. Furthermore, with BlastSets, the localization of induced/repressed genes 'inside' an elementary mode is not taken into account, although this information could be relevant. For example, a repressed gene belonging to a group of genes coding for the same enzyme may have no influence on the activity of the elementary mode as a whole, while repression of a gene that is the only one to encode a particular enzyme would be important. Finally, the computation of elementary modes is based on a steady-state assumption and it remains to be seen to what extent these concepts can be extended to dynamic activity.

Materials and methods

Genome-wide computation of elementary modes

Combinatorial explosion prevents the computation of elementary modes in large networks. For example, in a network of about 110 reactions the number of elementary modes was shown to be higher than two million [21]. Furthermore, even if more efficient algorithms were found, it would not be particularly useful to compute all the elementary modes in a genome-wide model of metabolism because the resulting set would be extremely difficult to interpret. We therefore opted for an alternative modular approach for computing elementary modes at the genome scale. Pathway maps of the KEGG database constitute a good basis for this task, as each of them represents a coherent and well-defined biological function and is sufficiently small for the number of elementary modes to remain in the hundreds.

We used the KEGG XML files for *S. cerevisiae* as a source for the metabolic model [43]. These files have the advantage of having been manually curated and they contain the same information as the graphical maps displayed by the KEGG database. They thus have been cleaned from invalid meta-

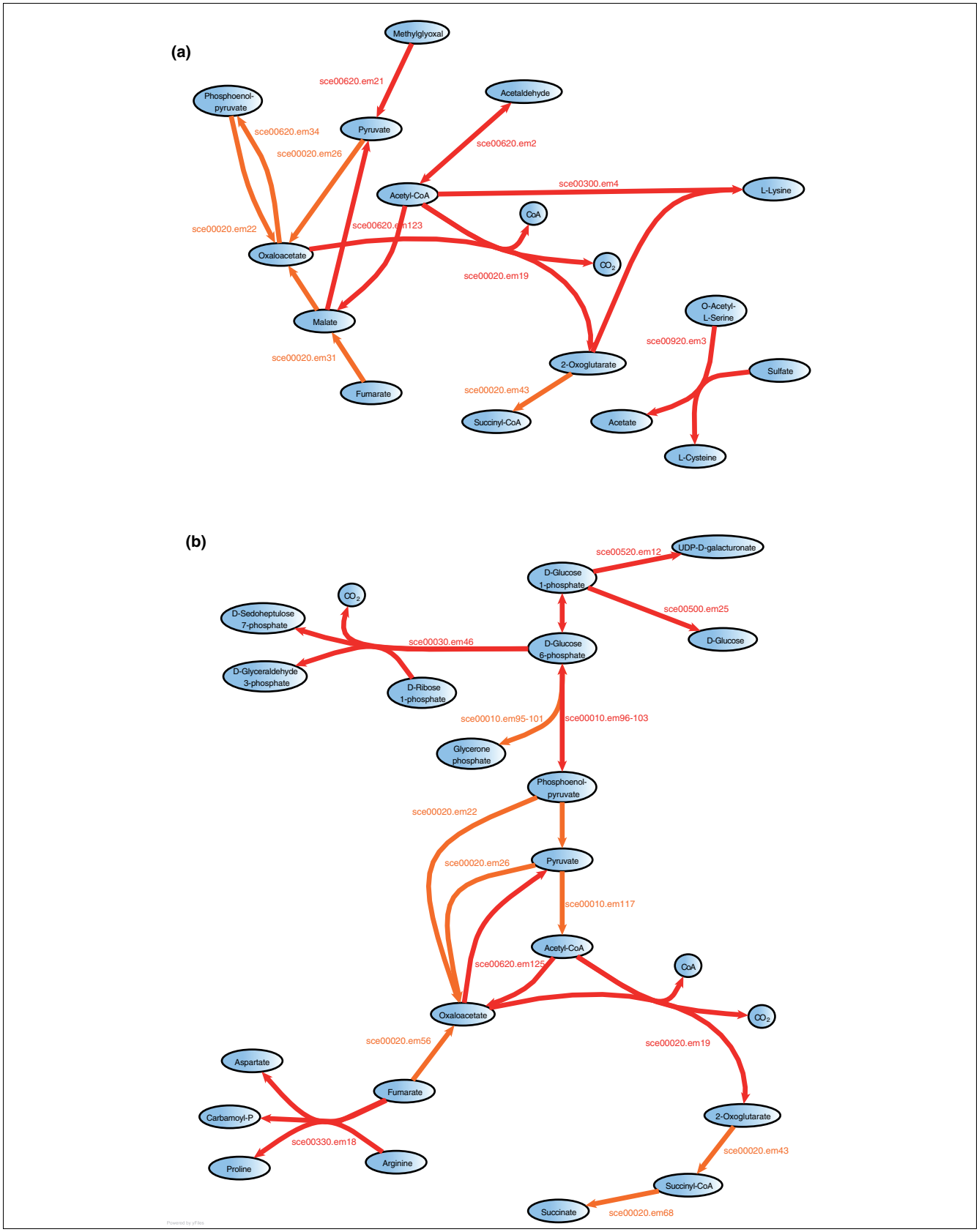


Figure 4 *(see legend on next page)*

Figure 4 *(see previous page)*
Backbones of metabolic stress response. **(a)** Toxic class. **(b)** Non-toxic class. These representations show all elementary modes induced in at least four different stress conditions. Main metabolic routes are drawn in red, and routes added by elementary modes that partly duplicate a main metabolic route but contain a separate short branch are drawn in orange.

bolic connections due to very common compounds (ADP, ATP, and so on), which otherwise create artificial links between metabolic compounds that do not correspond to biologically valid metabolic routes.

A stoichiometric matrix was constructed for each pathway based on its XML description. A point of major importance to the computation of elementary modes is the definition of 'external metabolites'. They act as start and end points of elementary modes, and in our hierarchical approach they additionally enable elementary modes from different pathway maps to be connected to each other. We adopted the following rules for defining external metabolites: one, a metabolite located at the interface between two or more pathway maps is considered external to all of them; two, a metabolite that can only be either produced or consumed is considered external; and three, unbalanced ubiquitous metabolites are considered external. Rule one creates the vast majority of entry/exit points to elementary modes and allows connections between pathway maps. Rule two prevents the existence of 'inactive' metabolic branches, that is, branches of a metabolic network that cannot participate in any elementary mode. This happens, for example, when a branch ends up in a dead end: as steady state conditions are assumed when elementary modes are computed, no flux can be present in a dead-end branch since this would lead to accumulation of the compound at its extremity. Rule three was introduced to prevent particular branches of the metabolic network from collapsing due to inappropriate balancing. For example, CO_2 appears on the map of the citrate cycle and must be considered external for the cycle to be able to operate, otherwise this route would contain a dead end and become inactive for the reason stated above. The complete list of metabolites covered by rule three comprises H_2O, O_2, P, CoA, CO_2, NH_3, UDP, H_2, and reduced and oxidized thioredoxin.

Once stoichiometric matrices had been constructed, elementary modes were computed using a classical algorithm [20]. The complete list of elementary modes for *S. cerevisiae* is provided as Additional data file 1, and was used to create the EM1 and EM2 BlastSets collections.

Expression data

Data sources

We have chosen experiments analyzing the gene expression responses of the yeast *S. cerevisiae* to various environmental stresses. Three sets of microarray experiments have been selected for our study. Causton *et al.* [32] described the transcriptional response to environmental changes using genome-wide expression experiments; data are available on the Young lab website [44]. Gasch *et al.* [33] analyzed gene expression of yeast cells during the adaptation to stressful environments in order to identify the main patterns of response in these different conditions; data were downloaded from the Stanford MicroArray Database website [45]. Iwahashi *et al.* [34] studied transcriptional responses of yeast to physical and chemical stresses using microarray; data are available from the Yeast Environmental Stress database [34].

These three datasets enabled us to study a total of 31 stress conditions. Some of these stresses involved environmental changes or nutrient depletion, while others involved exposure to toxic compounds such as pesticides or fungicides. The latter include: ash, which refers to exposure to burned ash from an industrial incinerator; maneb, which is a fungicide used in the control of several diseases of fruit, vegetable, field crops and ornamentals; pentachlorophenol (PCP), which is an effective fungicide, herbicide and algicide used as a wood preservative; tetrachloro-isophthalonitrile (TPN), which is a fungicide used to prevent biofouling on ships and in agriculture; thiuram, which is a compound used as fungicide to prevent crop damage and to protect harvested crops; and zineb, which was rated as a pesticide of low toxicity and may be a weak mutagen.

Data processing

We plotted the distributions of the natural logarithm of fold-change values for the Causton, Gasch and Iwahashi datasets. For each of the three sets of data, the standard deviation was determined. A threshold was defined by multiplying the standard deviation by a constant, and this threshold was used to determine which genes were considered as significantly induced or repressed in each condition. Genes whose fold change was higher than the threshold were considered induced; genes whose fold change was lower than 1 divided by the threshold were considered repressed. For each condition, a set of induced genes and a set of repressed genes were constructed. Table 5 indicates the number of genes present in each set.

Creation of random data sets

For three particular conditions (one from each dataset), we re-assigned gene expression values randomly to all genes of the experiment. We then processed these random expression data using the same procedures as described above. The resulting sets of random induced/repressed genes were compared to elementary modes using BlastSets in the same way as real expression data.

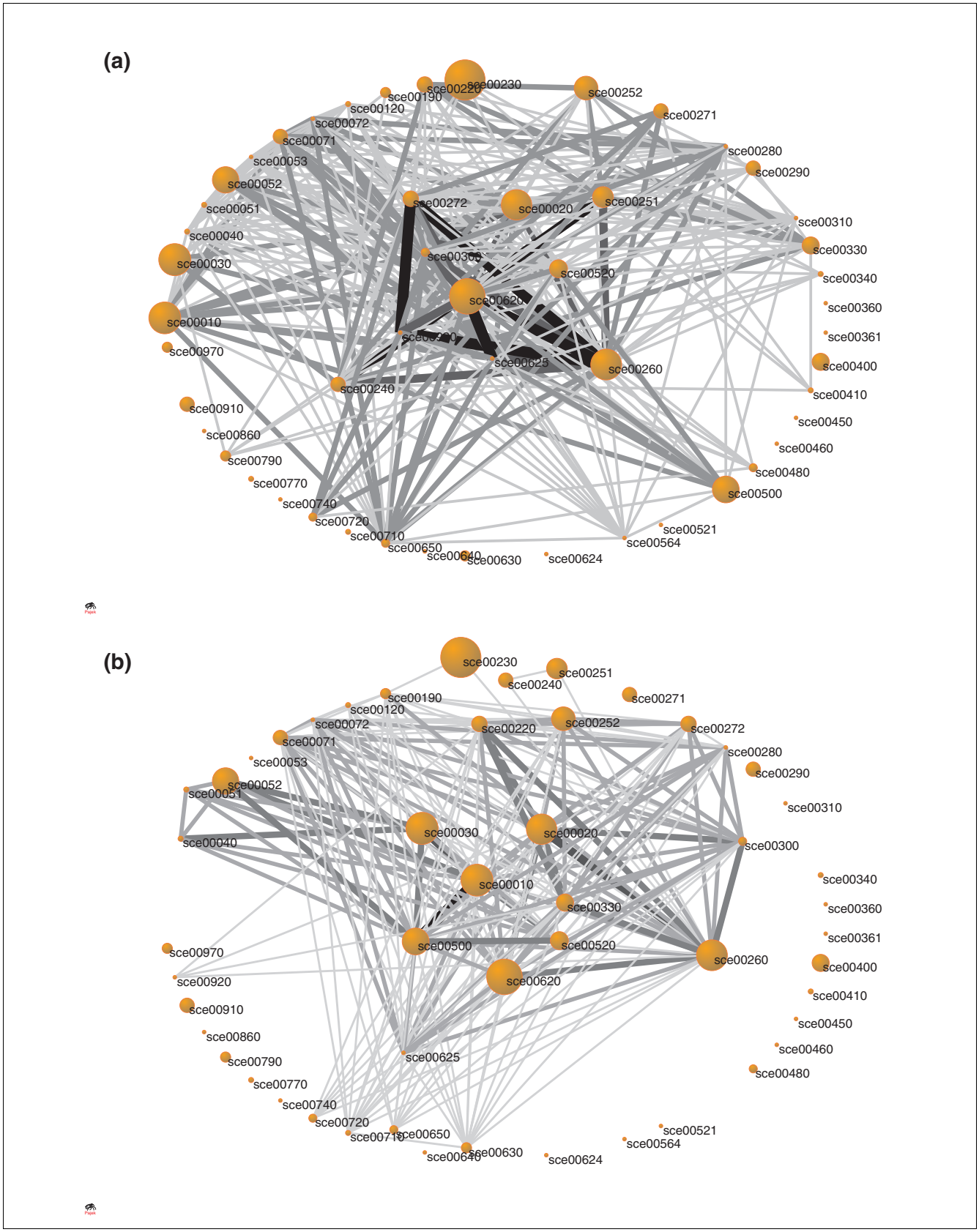


Figure 5 *(see legend on next page)*

Figure 5 *(see previous page)*
Interaction networks of metabolic pathways involved in the stress response according to the pairs of induced/repressed elementary modes spanning two pathways. **(a)** Toxic class. **(b)** Non-toxic class.

Data integration and analysis

Description of BlastSets

BlastSets is a bioinformatics tool that enables the integration of various biological data. This tool uses a standard representation for all types of data: data are structured in collections of sets of genes or proteins. Each collection corresponds to a biological source of information, and sets are composed of genes that share a similar property or value (close genes on a chromosome, co-expressed genes, proteins belonging to the same complex, proteins involved in the same metabolic pathway, and so on). The sets stored in the BlastSets database can

Table 5

Number of genes in each induced and repressed set

Stress condition	Number of genes in induced set (in BlastSets)	Number of genes in repressed set (in BlastSets)
Heat shock [32]	173	2
Acid [32]	32	6
Alkali [32]	73	10
Peroxide [32]	99	35
NaCl [32]	193	113
Sorbitol [32]	136	8
Heat shock [33]	114	0
Nitrogen depletion [33]	167	11
Stationary phase [33]	334	0
Hyperosmotic [33]	20	0
Peroxide [33]	60	0
Diauxic shift [33]	17	0
Menadione [33]	30	14
Dithiothreitol [33]	56	0
Hypoosmotic [33]	11	0
Diamide [33]	94	0
Variable temperature [33]	91	0
Amino acid starvation [33]	61	7
Alternative carbon [33]	0	0
Cadmium [34]	149	16
Ash [34]	390	713
Sodium n-dodecyl benzosulfonate [34]	36	2
Sodium lauryl sulfate [34]	53	2
Capsaicin [34]	10	0
Thiuram [34]	273	166
Zineb [34]	62	17
Maneb [34]	21	4
Tetrachloro-isophthalonitrile [34]	347	518
Pentachlorophenol [34]	181	31
Trichlorophenol [34]	27	10
Ethanol [34]	210	30
Pentanol [34]	285	182
Irradiation [34]	6	6
Octanol [34]	119	55
Pentane [34]	28	40

The number of genes identified by BlastSets in each stress condition from the three sets of microarray experiments.

be compared to each other or to submitted custom sets. To evaluate the similarity between two sets, their composition in terms of genes/proteins is compared, and the hypergeometric distribution is used to decide if the number of genes in common between the two compared sets is statistically significant (P value). As an example, one can check if the genes found co-expressed in an experiment correspond to a set containing proteins involved in the same pathway.

A *P* value is considered significant by BlastSets if it is less than or equal to a certain threshold. Multiple comparisons are performed as a set is compared to a collection of sets. The *P* value significance threshold is thus adjusted to the considered target sets using a Bonferroni correction. This takes into account the number of comparisons conducted, which depends on the number of sets in the collection. The resulting threshold is 6.0×10^{-5} when a set is compared to EM1 and 3.4×10^{-6} when compared to EM2. All hits with higher *P* values were automatically rejected. Additional details about BlastSets can be found in [27].

Integration of elementary modes in BlastSets

We used BlastSets to evaluate the biological relevance of elementary modes by comparing them to the sets of induced or repressed genes described above. We created two different collections of sets of elementary modes named KEGG_EM_1 (EM1) and KEGG_EM_2 (EM2) in BlastSets. EM1 is a collection of single elementary modes, that is, enzymes involved in a given elementary mode are gathered in a set labeled by the name of the mode. EM2 is a collection of pairs of elementary modes: all enzymes involved in two elementary modes that are connected through a common external link form one set in EM2. These two collections of sets of elementary modes are stored in the BlastSets database, and can be queried against user-submitted data via the BlastSets website [46].

Analysis of BlastSets results

Sets of induced and repressed genes in various stress conditions were compared to elementary modes using BlastSets, and lists of elementary modes found to be similar to the submitted sets of induced/repressed genes were obtained. A Perl script was developed to analyze these results and, thus, make it possible to reconstruct the chain of elementary modes that have been activated or repressed in response to each stress condition.

We retrieved the elementary modes (single or pair) that had the highest similarity with the set of induced/repressed genes, called the 'best hit'. First, if the best hit was a single elementary mode, we browsed subsequent hits until we found a pair of elementary modes containing this best hit. Second, once this 'best elementary mode pair' had been found, the rest of the list was browsed in order to find further pairs of elementary modes that were connected to the best pair, that is, pairs of elementary modes having one mode in common with the best elementary mode pair. Third, we could display a chain of pairs of elementary modes that defines the backbone of the metabolic response. If the best hit was a pair of elementary modes, only the second and third steps were performed. Among the elementary modes that could be added to the backbone, we removed all those that were composed of less than three enzymes to ensure that they were significant enough and to avoid the inclusion of short modes that are not specific to a single pathway.

Construction of toxic and non-toxic networks

Using the files containing BlastSets results with EM2, we constructed a matrix representing the usage of elementary modes in response to the different stresses, each row corresponding to an elementary mode and each column corresponding to a stress condition. In each element of the matrix, 1 was entered if the elementary mode was identified in response to the stress, 0 if it was not. A program was developed to compute, for each pair of pathways, the number of conditions where at least one pair of elementary modes spanning both pathways was found to be induced.

In Figure 5, each pathway was represented by a node whose radius was set proportional to the natural logarithm of the number of elementary modes contained in that pathway (for pathways with only one mode, the value was set to 0.5). This radius does not depend on the stress response and is only aimed at enhancing large pathways. Two pathways were connected by an edge if the number of induced pairs of elementary modes spanning both of them was non-zero. We weighted the connections by setting the thickness of edges proportional to the number of stress conditions in which such pairs were found. The weight thus does not depend on the number of active elementary modes in both pathways but on the number of conditions where both pathways contain simultaneously activated elementary modes. For a clearer representation, all weights were reduced by one unit, so that edges of weight 1 are not visible and the smallest visible edges are those of weight 2.

Additional data files

The following additional data are available with the online version of this paper. Additional data file is a list of elementary modes for *Saccharomyces cerevisiae*. Additional data file 2 is a figure showing induced and repressed metabolic backbones for all stress conditions. Additional data file 3 is a figure of a clustering tree of stress conditions.

Acknowledgements

JMS, JCN and MK were supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Society for the Promotion of Science, the Japan Science and Technology Corporation, and by Grant-in-Aid "Systems Genomics". CG has a PhD fellowship provided by the French Ministry of Education, Research and Technology. The BlastSets project is supported by funds allocated by ACI IMPBio from the French Ministry of Research. The computational resources were provided by the Centre de Bioinformatique de Bordeaux, Université Bordeaux 2, and by the

Bioinformatics Center, Institute for Chemical Research, Kyoto University. The Centre de Bioinformatique de Bordeaux is funded by the Région Aquitaine. We thank Professor Hitoshi Iwahashi for providing us his yeast microarray database of environmental stress response.

References

1. Duarte NC, Herrgard MJ, Palsson BO: **Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model.** *Genome Res* 2004, **14:**1298-1309.
2. Heinemann M, Kummel A, Ruinatscha R, Panke S: ***In silico* genome-scale reconstruction and validation of the *Staphylococcus aureus* metabolic network.** *Biotechnol Bioeng* 2005, **92:**850-864.
3. Oliveira AP, Nielsen J, Forster J: **Modeling *Lactococcus lactis* using a genome-scale flux model.** *BMC Microbiol* 2005, **5:**39.
4. Reed JL, Vo TD, Schilling CH, Palsson BO: **An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR).** *Genome Biol* 2003, **4:**R54.
5. Spirin V, Gelfand MS, Mironov AA, Mirny LA: **A metabolic network in the evolutionary context: multiscale structure and modularity.** *Proc Natl Acad Sci USA* 2006, **103:**8774-8779.
6. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402(Suppl):**C47-52.
7. Girvan M, Newman ME: **Community structure in social and biological networks.** *Proc Natl Acad Sci USA* 2002, **99:**7821-7826.
8. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, *et al.*: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network.** *Nature* 2004, **430:**88-93.
9. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs: simple building blocks of complex networks.** *Science* 2002, **298:**824-827.
10. Dobrin R, Beg QK, Barabasi AL, Oltvai ZN: **Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network.** *BMC Bioinformatics* 2004, **5:**10.
11. Ideker T, Ozier O, Schwikowski B, Siegel AF: **Discovering regulatory and signalling circuits in molecular interaction networks.** *Bioinformatics* 2002, **18(Suppl 1):**S233-240.
12. Isaacs FJ, Hasty J, Cantor CR, Collins JJ: **Prediction and measurement of an autoregulatory genetic module.** *Proc Natl Acad Sci USA* 2003, **100:**7714-7719.
13. Ishihara S, Fujimoto K, Shibata T: **Cross talking of network motifs in gene regulation that generates temporal pulses and spatial stripes.** *Genes Cells* 2005, **10:**1025-1038.
14. Mangan S, Alon U: **Structure and function of the feed-forward loop network motif.** *Proc Natl Acad Sci USA* 2003, **100:**11980-11985.
15. Maslov S, Sneppen K: **Detection of topological patterns in protein networks.** *Genet Eng NY* 2004, **26:**33-47.
16. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional regulation network of *Escherichia coli*.** *Nat Genet* 2002, **31:**64-68.
17. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297:**1551-1555.
18. Papin JA, Price ND, Wiback SJ, Fell DA, Palsson BO: **Metabolic pathways in the post-genome era.** *Trends Biochem Sci* 2003, **28:**250-258.
19. Schilling CH, Letscher D, Palsson BO: **Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective.** *J Theor Biol* 2000, **203:**229-248.
20. Schuster S, Fell DA, Dandekar T: **A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks.** *Nat Biotechnol* 2000, **18:**326-332.
21. Gagneur J, Klamt S: **Computation of elementary modes: a unifying framework and the new binary approach.** *BMC Bioinformatics* 2004, **5:**175.
22. Gianchandani EP, Papin JA, Price ND, Joyce AR, Palsson BO: **Matrix formalism to describe functional states of transcriptional regulatory systems.** *PLoS Comput Biol* 2006, **2:**e101.
23. Klamt S, Saez-Rodriguez J, Lindquist JA, Simeoni L, Gilles ED: **A methodology for the structural and functional analysis of signaling and regulatory networks.** *BMC Bioinformatics* 2006, **7:**56.
24. Peres S, Beurton-Aimar M, Mazat JP: **Pathway classification of TCA cycle.** *Syst Biol* 2006, **153:**369-371.
25. Hanisch D, Zien A, Zimmer R, Lengauer T: **Co-clustering of biological networks and gene expression data.** *Bioinformatics* 2002, **18(Suppl 1):**S145-154.
26. Yang HH, Hu Y, Buetow KH, Lee MP: **A computational approach to measuring coherence of gene expression in pathways.** *Genomics* 2004, **84:**211-217.
27. Barriot R, Poix J, Groppi A, Barre A, Goffard N, Sherman D, Dutour I, de Daruvar A: **New strategy for the representation and the integration of biomolecular knowledge at a cellular scale.** *Nucleic Acids Res* 2004, **32:**3581-3589.
28. Covert MW, Palsson BO: **Constraints-based models: regulation of gene expression reduces the steady-state solution space.** *J Theor Biol* 2003, **221:**309-325.
29. Klamt S, Stelling J: **Combinatorial complexity of pathway analysis in metabolic networks.** *Mol Biol Rep* 2002, **29:**233-236.
30. Schuster S, Pfeiffer T, Moldenhauer F, Koch I, Dandekar T: **Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae*.** *Bioinformatics* 2002, **18:**351-361.
31. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006:D354-357.
32. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA: **Remodeling of yeast genome expression in response to environmental changes.** *Mol Biol Cell* 2001, **12:**323-337.
33. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11:**4241-4257.
34. **Environmental Stress Database** [http://kasumi.nibh.jp/~iwahashi/]
35. Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page GP, Somerville C, Loraine A: **Transcriptional coordination of the metabolic network in Arabidopsis.** *Plant Physiol* 2006, **142:**762-774.
36. Barabási AL: *Linked: The New Science of Networks* Cambridge, MA: Perseus Publishing; 2002.
37. Vido K, Spector D, Lagniel G, Lopez S, Toledano MB, Labarre J: **A proteome analysis of the cadmium response in *Saccharomyces cerevisiae*.** *J Biol Chem* 2001, **276:**8469-8474.
38. Natarajan K, Meyer MR, Jackson BM, Slade D, Roberts C, Hinnebusch AG, Marton MJ: **Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast.** *Mol Cell Biol* 2001, **21:**4347-4368.
39. Martinez MJ, Roy S, Archuletta AB, Wentzell PD, Anna-Arriola SS, Rodriguez AL, Aragon AD, Quinones GA, Allen C, Werner-Washburne M: **Genomic analysis of stationary-phase and exit in *Saccharomyces cerevisiae*: gene expression and identification of novel essential genes.** *Mol Biol Cell* 2004, **15:**5295-5305.
40. Schade B, Jansen G, Whiteway M, Entian KD, Thomas DY: **Cold adaptation in budding yeast.** *Mol Biol Cell* 2004, **15:**5492-5502.
41. Kell DB: **Systems biology, metabolic modelling and metabolomics in drug discovery and development.** *Drug Discov* 2006, **11:**1085-1092.
42. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, *et al.*: **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.** *Nat Biotechnol* 2006, **24:**1151-1161.
43. **KEGG Database** [http://www.genome.jp/kegg/xml/sce/index.html]
44. **Young Lab Website** [http://web.wi.mit.edu/young/environment/]
45. **Stanford MicroArray Database** [http://genome-www5.stanford.edu/]
46. **BlastSets** [http://cbi.labri.fr/outils/BlastSets]

"Exploration bioinformatique des relations entre mécanismes moléculaires et fonctions cellulaires"

Résumé

L'intégration des données biologiques est un des principaux défis de la bioinformatique aujourd'hui. La mise à disposition de quantités importantes de données concernant tous les niveaux d'organisation de la cellule, nécessite la mise en place de stratégies d'intégration pour rassembler toutes ces données, et ainsi mieux comprendre le fonctionnement de la cellule. Nous nous sommes intéressés à l'exploitation du concept de voisinage pour représenter et intégrer des données biologiques. Dans un premier temps, notre travail met l'accent sur l'importance du choix de la représentation pour mener une intégration efficace. Notre étude sur la représentation du métabolisme a montré que les modes élémentaires sont une alternative pertinente à la représentation classique sous forme de voies métaboliques. De plus, les modes élémentaires nous ont permis de trouver des routes métaboliques utilisées par la cellule en réponse à divers stress. Nous avons également exploité le voisinage dans une perspective de génomique comparative. Nous avons cherché à déterminer si le voisinage d'expression peut être une signature pour les gènes, et s'il peut être utilisé pour caractériser des gènes en établissant des équivalences entre des génomes (orthologues ou gènes fonctionnellement similaires). Les résultats présentés confirment l'intérêt de l'exploration du voisinage, des gènes et de leur produit, pour intégrer des données hétérogènes. L'efficacité de cette exploration est fortement liée au choix de la représentation des connaissances.

Mots-clés: intégration de données; représentation des connaissances; transcriptomique; métabolisme; voisinage ; génomique comparative.

Abstract

Biological data integration is one of the major challenge in bioinformatics today. The availability of amounts of data concerning all the level of cell organisation, requires strategies of integration to bring together these data and thus better understand how the cell works. We have focused our work on the use of the concept of neighbourhood in order to represent and integrate data. First, our work emphasizes the importance of the choice of data representation for an efficient integration. Our study on metabolism representation shows that elementary modes are a relevant alternative to the classical representation of metabolism as metabolic pathways. Moreover, elementary modes have enabled us to find metabolic routes used by the cell in response to stressed. We have also used the neighbourhood in a new angle, the one of comparative genomics. We tested if expression neighbourhood of genes (set of genes with close expression profiles) can be a signature for genes, and if it can be used to define functional similarities between genes from different organisms. The work presented here, shows the interest of the exploration of gene and protein neighbourhood in order to integrate heterogeneous data. The efficiency of this exploration is highly related to the choice of knowledge representation.

Keywords: data integration; knowledge representation; transcriptomics; metabolism; neighbourhood; comparative genomics.