

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Applied Bioinformatics in *Saccharomyces cerevisiae*

Data storage, integration and analysis

Kwanjeera Wanichthanarak



CHALMERS

Department of Chemical and Biological Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2014

Applied Bioinformatics in *Saccharomyces cerevisiae*
Data storage, integration and analysis
KWANJEERA WANICHTHANARAK
ISBN 978-91-7597-032-5

© KWANJEERA WANICHTHANARAK, 2014.

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny series nr 3731
ISSN 0346-718X

Department of Chemical and Biological Engineering
Chalmers University of Technology
SE-41296 Gothenburg
Sweden
Telephone +46 (0)31 7721000

Cover: Yeast databases and data analysis
Back cover: Taken by K. Kusonmano

Printed by Chalmers Reproservice
Gothenburg, Sweden 2014

Applied Bioinformatics in *Saccharomyces cerevisiae*

Data storage, integration and analysis

Kwanjeera Wanichthanarak

Systems and Synthetic Biology, Department of Chemical and Biological Engineering,
Chalmers University of Technology

Abstract

The massive amount of biological data has had a significant effect on the field of bioinformatics. This growth of data has not only lead to the growing number of biological databases but has also imposed the needs for additional and more sophisticated computational techniques to proficiently manage, store and retrieve these data, as well as to competently help gaining biological insights and contribute to novel discoveries.

This thesis presents results from applying several bioinformatics approaches on yeast datasets. Three yeast databases were developed using different technologies. Each database emphasizes on a specific aspect. yApoptosis collects and structurally organizes vital information specifically for yeast cell death pathway, apoptosis. It includes predicted protein complexes and clustered motifs from the incorporation of apoptosis genes and interaction data. yStreX highlights exploitation of transcriptome data generated by studies of stress responses and ageing in yeast. It contains a compilation of results from gene expression analyses in different contexts making it an integrated resource to facilitate data query and data comparison between different experiments. A yeast data repository is a centralized database encompassing with multiple kinds of yeast data. The database is applied on a dedicated database system that was developed addressing data integration issue in managing heterogeneous datasets. Data analysis was performed in parallel using several methods and software packages such as Limma, Piano and metaMA. Particularly the gene expressions of chronologically ageing yeast were analyzed in the integrative fashion to gain a more thorough picture of the condition such as gene expression patterns, biological processes, transcriptional regulations, metabolic pathways and interactions of active components. This study demonstrates extensive applications of bioinformatics in the domains of data storage, data sharing, data integration and data analysis on various data from yeast *S.cerevisiae* in order to gain biological insights. Numerous methodologies and technologies were selectively applied in different contexts depended upon characteristics of the data and the goal of the specific biological question.

Keywords: Bioinformatics, database design, database system, gene expression analysis, integrative analysis

List of publications

The thesis is based on the work in the following publications:

- I. **Wanichthanarak K.**, Cvijovic M., Molt A. and Petranovic D. (2013). yApoptosis: yeast apoptosis database. Database (Oxford), 2013, bat068.
- II. **Wanichthanarak K.**, Nookeaw I. and Petranovic D. (2014). yStreX: Yeast Stress Expression Database. Database (Oxford). (Resubmitted after revision)
- III. Pornputtpong N.*, **Wanichthanarak K.***, Nilsson A., Nookeaw I. and Nielsen J. (2014). A dedicated database management system for handling multi-level data in systems biology. Source Code for Biology and Medicine. (Submitted)
- IV. **Wanichthanarak K.**, Wongtosrad N. and Petranovic D. (2014). Genome-wide expression analyses of the stationary phase model of ageing in yeast. Mechanisms of Ageing and Development (Submitted)

Additional publication not included in this thesis:

- V. Munoz A.J.*, **Wanichthanarak K.***, Meza E.* and Petranovic D. (2012). Systems biology of yeast cell death. FEMS Yeast Research, 12(2):249-65.

* Authors contributed equally

Contributions

- I. Designed and developed the database. Curated data and performed network analysis. Drafted and edited the manuscript.
- II. Designed and developed the database. Curated data and performed transcriptome and integrative analysis. Drafted and edited the manuscript.
- III. Participated in designing the database system and coding the parser class. Supervised the development of the database and data population, and implemented case studies. Took part in writing and editing the manuscript.
- IV. Performed transcriptome and integrative analysis. Supervised the work in time-series analysis. Drafted and edited the manuscript.

Additional publication not included in this thesis:

- V. Participated in writing and editing the manuscript.

Abbreviations

AIF	Apoptosis-inducing factor
AMID	Apoptosis-inducing factor-homologous mitochondrion-associated inducer of death
AP-MS	Affinity purification mass spectrometry
BioPAX	Biological Pathway Exchange
BP	Biological process
ChIP	Chromatin immunoprecipitation
CLS	Chronological lifespan
CVT	Cytoplasm to vacuole targeting
DEA	Differential expression analysis
DNA	Deoxyribonucleic acid
EF	Experimental factor
ESR	Environmental stress response
GEO	Gene Expression Omnibus
GO	Gene ontology
GSA	Gene set analysis
H ₂ O ₂	Hydrogen peroxide
JSON	JavaScript object notation
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAPK	Mitogen-activated protein kinase
NCBI	National Center for Biotechnology Information
PCD	Programmed cell death
PDI	Protein-DNA interaction
PDS	Post-diauxic shift
PKA	Protein kinase A
PPI	Protein-protein interaction
PSI-MI	Proteomics standards initiative molecular interaction
RLS	Replicative lifespan
RNA	Ribonucleic acid
ROS	Reactive oxygen species
<i>S.cerevisiae</i>	<i>Saccharomyces cerevisiae</i>
SBML	Systems biology markup language
SD	Synthetic complete medium
SQL	Standard query language
STRE	Stress response element
TCA	Tricarboxylic acid cycle
TF	Transcription factor
TOR	Target of rapamycin
Y2H	Yeast two-hybrid

List of figures and tables

Figure 1 Components and pathways of yeast apoptosis	5
Figure 2 Ageing paradigms in yeast	6
Figure 3 Stress-response and lifespan regulatory pathways	7
Figure 4 Bioinformatics and systems biology	10
Figure 5 Architecture of centralized and federated databases	12
Figure 6 Networks of apoptosis genes	19
Figure 7 Networks of yeast PCD	22
Figure 8 Overview of visitors to yApoptosis	22
Figure 9 Analysis workflow	23
Figure 10 Screenshot of navigations between different web pages	25
Figure 11 Part of the consensus heatmap of TF gene sets	26
Figure 12 Yeast data repository architecture	30
Figure 13 Resulting queries from a yeast data repository	31
Figure 14 Venn diagram of significant genes and enriched functional terms	33
Figure 15 Heatmap of enriched metabolic pathways from GSA	35
Figure 16 Consensus network of all identified active subnetworks	38
Table 1 List of software packages and resources for gene expression analysis	13
Table 2 Yeast apoptosis genes and human orthologues	21
Table 3 List of yeast data in the yeast data repository and resources	29
Table 4 Summary of biological physiologies and associated gene sets	37

Table of contents

1. Introduction	1
1.1 Thesis structure	2
2. Background	3
2.1 Yeast <i>Saccharomyces cerevisiae</i>	3
2.1.1 Yeast apoptosis	3
2.1.2 Yeast ageing	5
2.1.3 Yeast stress responses	8
2.2 Bioinformatics and systems biology	9
2.3 Biological database design and development	10
2.4 Gene expression analysis	12
2.5 Biological networks	14
3. Results and discussion	16
3.1 Design and development of yeast databases	16
3.1.1 Paper I: Yeast apoptosis database	16
3.1.2 Paper II: Yeast stress expression database	23
3.1.3 Paper III: Yeast data repository	28
3.2 Bioinformatics applications in transcriptome analysis	32
3.2.1 Paper IV: Genome-wide expression analyses of the stationary phase model of ageing in yeast	32
4. Conclusions and perspectives	39
Acknowledgements	41
References	42

Preface

This thesis is submitted for the partial fulfillment of the degree doctor of philosophy. It is based on work carried out between 2010 and 2014 at the Department of Chemical and Biological Engineering, Chalmers University of Technology, under the supervision of Assoc. Prof. Dina Petranovic. The research was funded by the Knut and Alice Wallenberg Foundation, the European Research Council and the Chalmers Foundation.

Kwanjeera Wanichthanarak

May 2014

1. Introduction

Saccharomyces cerevisiae or the baker's yeast is one of the famous model organisms and it has been extensively used in molecular biology, biotechnology and for study of processes related to human health and disease (Petranovic and Nielsen, 2008). The operating principles of fundamental cellular mechanisms, such as DNA replication, DNA recombination, cell division, protein homeostasis and vesicular trafficking, are well conserved among yeast and higher eukaryotes (Fields and Johnston, 2005; Winderickx, et al., 2008). Indeed, it was reported that yeast genes have mammalian homologues, for instance, RAS1 and RAS2 are homologues of the mammalian RAS proto-oncogenes (Botstein, et al., 1997). Conducting studies with yeast (either by classical complementation assays for human proteins that have a yeast homologue or by humanized yeast systems for human proteins that do not have a yeast counterpart) have contributed to reveal the functional roles or the biological consequences of mutations of human proteins (Botstein and Fink, 2011; Winderickx, et al., 2008). Nowadays yeast has relatively comprehensive sets of omics data such as genome, transcriptome, interactome and metabolome. These data facilitate research in the field of systems biology that observes more than individual genes and proteins but considers how these molecules interact and work together to establish the properties of living cells.

After the development of high-throughput DNA sequencing technologies, bioinformatics became an essential discipline to extract information in genome sequences (Barnes, 2007). The term informatics has been described by Altman (2012) that "it is the study of how to represent, store, search, retrieve and analyze information (Altman, 2012)". The biological data can be the information stored in the DNA sequences, RNA expression, three-dimensional protein structures, protein interactions, clinical data, and published literature. Bioinformatics thus involves beyond the alignment of DNA sequences. It is an area of science that uses various methodologies and computational technologies to structurally store biological information, to answer biological enquiries, to have new biological findings and to guide experimental design. Bioinformatics databases and software tools become indispensable parts of life sciences research these days.

The main objective of this thesis is to apply key domains of bioinformatics including database development and data analysis on data generated by studying yeast, particularly yeast cell death, ageing and stress pathways. The main intention of yApoptosis (Paper I) is to show how building a dedicated database which stores and collects specific molecular and cell biology information (focus on a specific pathway) and that it can easily be accessed, searched and shared, contributes to the development of this particular field of research. Given the growing amount of transcriptome data, and the facing challenge of efficiently utilizing such data, we developed the

yStreX database with integrated analyses (Paper II) where readily-processed data and analyses are collected and can be rapidly retrieved. This database was developed to facilitate exploitation of transcriptome data in the area of yeast stress responses.

To understand biological processes, systems biology research imposes integration of biological data from different levels. This issue puts challenges in data integration areas including establishing data standards and implementing computational tools and infrastructures. An effort is presented in this study by designing a database system and applying this system to build a yeast data repository containing multiple types of biological data (Paper III).

Data analysis was a key step performed throughout this PhD study for generating data to be stored in the databases and for studying a specific process (i.e. yeast ageing in Paper IV) using several software packages e.g. Limma package (Smyth, 2004) to identify differentially expressed genes compared to control. The integrated analysis was conducted by using biological networks such as metabolic pathways, transcriptional regulatory interactions and networks of biological processes as a scaffold for incorporating transcriptome data which results in a reduced dimension and a systemic view of the data to aid biological interpretation. Besides, integrating transcriptome and interactome data contributed to identification of functional modules which illustrates how molecules interact together under the ageing condition.

1.1 Thesis structure

The thesis contains two main parts. The first part is divided into four chapters containing extended descriptions of the work and related information. Chapter 1 gives the short introduction and the objective of the work in this thesis. Chapter 2 provides background of the topics related to the work in this thesis. In Chapter 3, the brief summaries of background and results from all papers are presented. This chapter is divided into two sections on database design and development, and data analysis. Chapter 4 concludes key aspects and perspectives of all the work. The second part is the collection of the research articles with the order following the sections in Chapter 3.

2. Background

2.1 Yeast *Saccharomyces cerevisiae*

The budding yeast *Saccharomyces cerevisiae* is a unicellular organism that has been recognized as a favorite model organism for eukaryotes. It was the first eukaryotic organism that had entire genome completely sequenced and yeast genome has proved to be a useful reference for the sequences of human and other higher eukaryotic genes (Goffeau, et al., 1996; Schneiter, 2004). The *Saccharomyces* Genome Database (SGD) is the main resource for the budding yeast (Cherry, et al., 2012). It contains piles of information about yeast genome, proteins and other related features.

Yeast has been used for both molecular research and industrial applications as there are several dominant features including ease of cultivation, well-studied organism, well-established molecular biology toolboxes and a bunch of molecular datasets generated from high-throughput technologies (Petranovic and Nielsen, 2008). Interestingly, about 40% of genes of human heritable disease have homologues in yeast (Oliver, 2002). This shows that yeast has a potential to be used in pharmaceutical and medical applications including as a platform for protein productions and as a toolbox for revealing biological insights of molecular mechanisms (Hou, et al., 2012; Munoz, et al., 2012). Examples of research on complex biological processes conducted in yeast include programmed cell death (PCD) and ageing (Carmona-Gutierrez, et al., 2010; Longo and Fabrizio, 2012). In addition, yeast genome-wide data serve as valuable resources for further research and applications in bioinformatics and systems biology.

2.1.1 Yeast apoptosis

Apoptosis is one form of PCD that was also reported in yeast. Yeast cells undergoing apoptosis show typical apoptotic features e.g. phosphatidylserine externalization, cytochrome c release, depolarization of mitochondrial membrane potential, chromatin condensation and DNA fragmentation (Madeo, et al., 2009).

To ensure proper development and maintain homeostasis, apoptosis in higher eukaryotes occurs under certain circumstances such as during cell differentiation, response to infection, removal of damaged cells, ageing and response to different stresses (Sharon, et al., 2009). In a unicellular organism like yeast, the purpose of apoptosis is comparable to metazoan as it arises during ageing, mating and stress-inducing to eliminate damaged or old cells, thereby promoting survival of major population (Carmona-Gutierrez, et al., 2010).

Generally speaking, yeast apoptosis pathways include caspase-dependent and caspase-independent pathway (Madeo, et al., 2009). The first pathway involves induction of yeast metacaspase by various stimuli such as H₂O₂, acetic acid, hyperosmotic stress, ageing and other agents. The gene YCA1 encodes a metacaspase which specifically catalyzes protein substrates and in turn triggers apoptosis (Mazzoni and Falcone, 2008). For caspase-independent process, the apoptosis-inducing factor 1 (AIF1) was one of implicated players (Madeo, et al., 2009). It translocates from mitochondria to the nucleus upon apoptotic stimulus and causes chromatin condensation and DNA fragmentation. The other players in caspase-independent pathways include NUC1, NMA111 and STE20. NUC1 is a yeast orthologue of mammalian endonuclease G (EndoG) and Nuc1 was found translocating to the nucleus from mitochondria upon apoptosis induction (Buttner, et al., 2007). Nma111 (nuclear mediator of apoptosis) localizes in the nucleus and Bir1, an inhibitor-of-apoptosis protein in yeast, is a substrate of its serine protease (Fahrenkrog, et al., 2004). Ste20 phosphorylates histone H2B at serine 10 causing chromatin condensation (Ahn, et al., 2005). It is also a part of pheromone-induced apoptosis resulting from an enhancement of mitochondrial respiration and cytochrome c release (Carmona-Gutierrez, et al., 2010; Pozniakovsky, et al., 2005).

Mitochondria are an organelle that has been considered to play an important role in ageing and apoptosis. For instance, release of cytochrome c from mitochondria after acetic acid treatment was found to induce apoptosis in yeast (Ludovico, et al., 2002), though, it is still skeptical how cytochrome c leads to caspase activation and apoptosis (Carmona-Gutierrez, et al., 2010). It is widely known that mitochondria are a main source of reactive oxygen species (ROS) in cells. These ROS arise as the products of cellular metabolism during aerobic respiration (Farrugia and Balzan, 2012). They have deleterious effects to a wide variety of molecules, such as nucleic acids, proteins and lipids. Changing mitochondrial morphology can also affect ROS production. It was demonstrated that inhibition of mitochondrial fission by deleting Dnm1 extends lifespan, and increases stress tolerance (Braun and Westermann, 2011; Cheng, et al., 2008; Perrone, et al., 2008). Mitochondrial fragmentation together with mitochondrial dysfunction and accumulation of ROS are general hallmarks of cell death (Braun and Westermann, 2011).

In brief, apoptosis can be induced by various stimuli both externally and internally. It involves several genes, proteins, cellular mechanisms and organelles such as mitochondria and nucleus as depicted in Figure 1. Besides, homologues of mammalian were found in yeast, such as a yeast caspase (Yca1), nuclear mediator of apoptosis (Nma111), endonuclease G homologue (Nuc1), apoptosis-inducing factor (Aif1) and yeast AMID (Ndi1) (Frohlich, et al., 2007; Madeo, et al.,

2009), which make yeast a promising research tool for elucidating cell death pathways in human and other higher eukaryotes.

Nowadays, research in yeast PCDs yeast still undergoes. It is not only about apoptosis but also covers other PCDs including autophagy and necrosis, and ageing processes. It remains several challenges to understand, for instance, how cells switch between those different PCD subroutines or how cells decide to live or die.

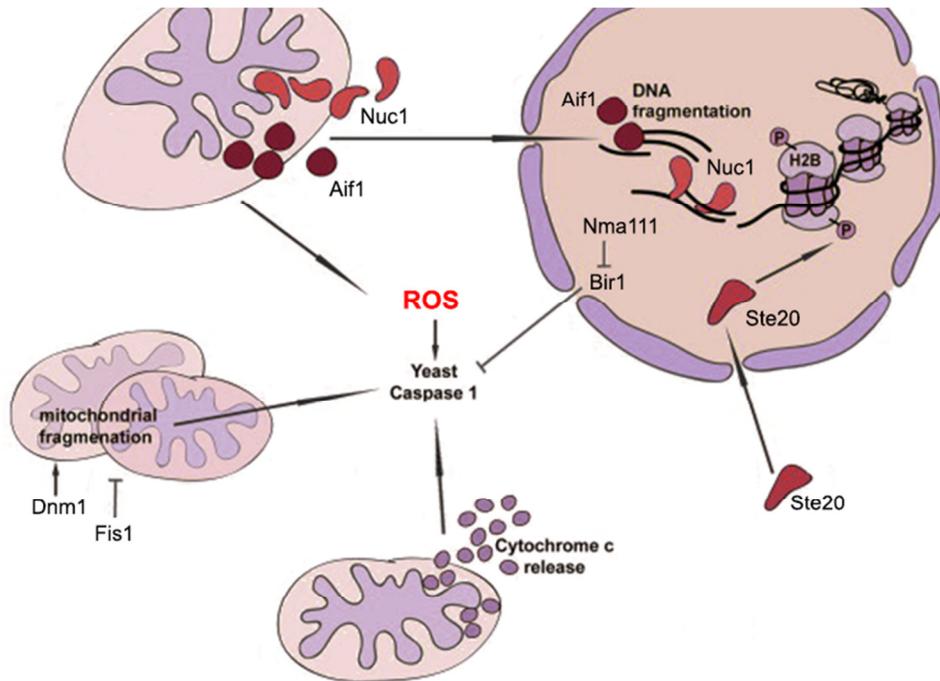


Figure 1 Components and pathways of yeast apoptosis. Upon apoptosis induction, it leads to activation of several apoptotic key players such as the yeast caspase Yca1. Apoptosis also engages other processes including mitochondrial fragmentation, cytochrome c release, DNA fragmentation and histone modification. Figure adapted from (Madeo, et al., 2009).

2.1.2 Yeast ageing

In general ageing is a process associated with progressive decline in the competence to compete against cellular stress and damage (Sharon, et al., 2009). It is also an endogenous stimulus of apoptosis. There are two ageing paradigms described in yeast: replicative lifespan (RLS) and chronological lifespan (CLS) (Figure 2). RLS is measured by the number of replications a mother cell produces before dying while CLS is the survival time of non-dividing populations in long-term cultivation (Longo and Fabrizio, 2012).

In both types, yeast cells die exhibiting markers of apoptosis such as accumulating oxygen radicals and caspase activation (Sharon, et al., 2009). Similar to cell death, yeast has been widely used as a model to study ageing processes. The yeast replicative ageing is a potential model for the ageing process of proliferating cells such as human stem cell, whereas chronological ageing serves as a model for the ageing of post-mitotic cell types such as brain and muscle (Rockefeller and Madeo, 2008). In this thesis, CLS is the main focus.

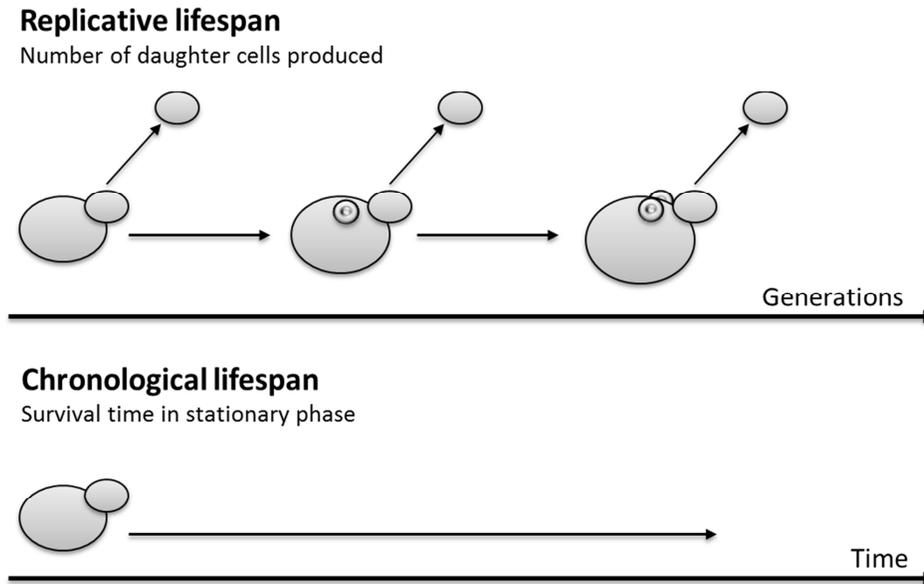


Figure 2 Ageing paradigms in yeast. Two ageing paradigms have been described in yeast including replicative (top panel) and chronological lifespan (bottom panel).

Unlike RLS, CLS is directly influenced by the availability of nutrients. Lacking of nutrients causes cells to enter diauxic shift and stationary phase showing several physiological changes such as low transcription rate, reduced metabolism, low protein synthesis, thick cell walls and non-budding. Yeast cell also accumulates storage molecules such as glycogen, triacylglycerol, polyphosphate and trehalose (Galdieri, et al., 2010). Signal transduction pathways that regulate longevity and stress responses were identified and were found evolutionally conserved. They are target of rapamycin (TOR), protein kinase A (PKA) and Snf1 pathway (Fabrizio and Longo, 2003; Galdieri, et al., 2010) (Figure 3).

These signaling pathways control transcriptions of downstream genes through different transcription factors (TF). In particular, during exponential phase, TOR and PKA pathway negatively regulate Rim15 protein kinase which in turn positively controls stress response TFs

such as Msn2, Msn4 and Gis1 (Wei, et al., 2008). In contrast, under calorie restriction genes containing the stress response element (STRE) (e.g. CTT1, DDR1, HSP12 and TPS2) or post-diauxic shift motif (PDS) (e.g. SSA3 and GRE1) in their promoters are transcriptionally induced by those TFs respectively as a result of deficiency in the signaling pathways (Orzechowski Westholm, et al., 2012; Pedruzzi, et al., 2000; Wei, et al., 2008). Snf1 is a protein kinase that inactive in the presence of glucose. Its target includes Adr1 (that activates transcription of genes implicated with utilization of non-fermentable carbon sources), and Mig1 (that represses transcription of genes in the presence of glucose) (Galdieri, et al., 2010). Under carbon stress, it was also found to regulate TFs Msn2 and Hsf1 (Hahn and Thiele, 2004).

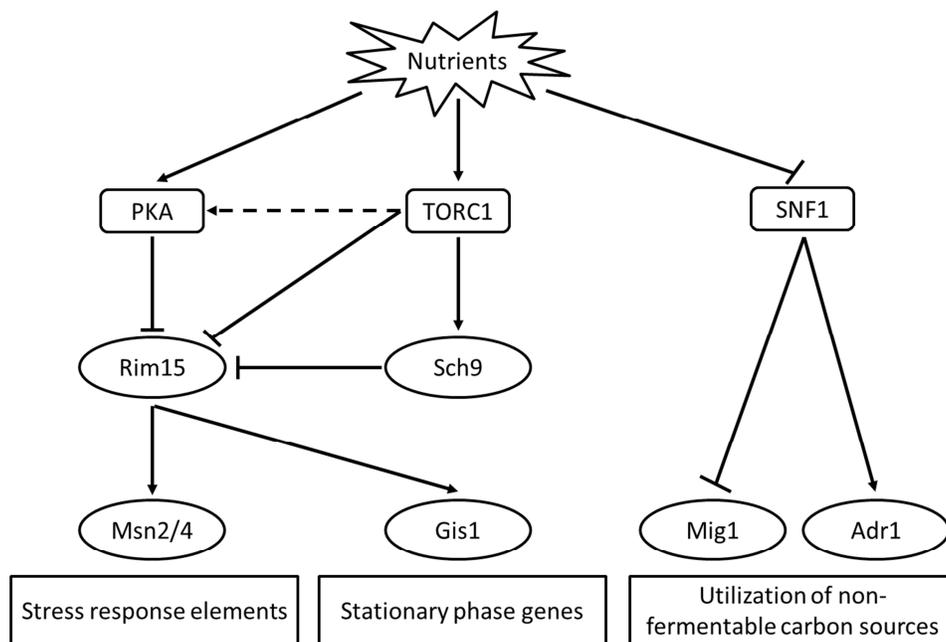


Figure 3 Stress-response and lifespan regulatory pathways. Stress responses and lifespan are regulated by signaling pathways (represented by rounded rectangular) including TOR, PKA and SNF1 pathway. Downstream components of the signal transduction pathways (represented by oval) further control genes involving in important processes such as stress responses, stationary growth and catabolism of non-fermentable carbon sources (represented by square). Figure adapted from (Galdieri, et al., 2010).

To summarize, longevity is determined by the capability to endure various stresses e.g. oxidative stress and genomic instability. It involves several lifespan regulatory pathways that control downstream players to exhibit, for instance, metabolic switches, anti-oxidant activities and other cellular protection processes to prolong the lifespan (Longo and Fabrizio, 2012).

2.1.3 Yeast stress responses

Changes in environmental systems can disturb internal homeostasis of cells which then might result in cellular malfunctions, no growth or death. Thus, living organisms have to adapt rapidly in order to survive under the new environment.

Yeast develops stress-response strategies by remodeling of its gene expression program to deal with fluctuations in, for example, pH, temperature, nutrient condition, osmotic pressure, oxygen level, drugs and toxic compounds (Gasch, et al., 2000). The reprogramming of gene expression can be captured by DNA microarrays as demonstrated in (Causton, et al., 2001; Gasch, et al., 2000; Knijnenburg, et al., 2009). These genomic studies allow us to gain insights into the regulation of responses to stressful conditions.

In most studies, yeast responds to environmental shifts through changes in genomic expression of thousands genes (Gasch and Werner-Washburne, 2002). Among those, around 900 genes changes commonly in diverse environments and they are denoted as the environmental stress response (ESR) genes. There are approximately 300 genes induced in several stresses and these genes are in carbohydrate metabolism (e.g. FBP26, TPS1,2,3 and GSY2), in protein folding (e.g. HSP26,42,78 and SSA3), in protein degradation (e.g. UBC5,8 and UBI4) and in oxidative stress response (e.g. CTT1 and SOD1) (Causton, et al., 2001; Gasch, et al., 2000; Gasch and Werner-Washburne, 2002). Most of these ESR genes contain STRE motif (AGGGG) in their promoters targeted by Msn2 and Msn4, which may be considered as general stress transcription factors. However, genes in the ESR are regulated by different TFs which are condition-specific, e.g. Yap1 and Hsf1 regulate ESR genes in response to oxidative stress or heat shock, respectively (Gasch, et al., 2000; Gasch and Werner-Washburne, 2002). Furthermore, around 600 common-repressed genes involve in RNA metabolism, ribosomal proteins and protein synthesis indicating that cells try to reserve energy during their adaptation to new conditions (Causton, et al., 2001; Gasch and Werner-Washburne, 2002).

The expression of ESR genes has been reported to be mediated by a number of signaling pathways which are activated by specific upstream signals, such as MAPK HOG pathway is active in response to osmotic stress. In contrast, pathways like PKA and TOR as shown in Figure 3 have been implicated to repress ESR genes. The capability of the cell to withstand and response appropriately is certainly vital and required to determine its cellular lifespan.

2.2 Bioinformatics and systems biology

Nowadays bioinformatics becomes a major part of most life sciences research (Marcus, 2008). Bioinformatics is a highly interdisciplinary field that drives knowledge discovery from biological data using computational-based analysis. It applies concepts and methods from many areas such as mathematics, statistics, genetics, computer science, physics, chemistry, medicine and biology to dig into information embedded in various biological data including data from high-throughput technologies, 3D protein structures, clinical data and scientific literature (Marcus, 2008).

The main aims of bioinformatics can be described in three areas: 1) to facilitate data management, access and sharing by structurally collecting biological data and existing information in the form of databases, such as GenBank DNA sequence database (Benson, et al., 2013), ArrayExpress functional genomics database (Rustici, et al., 2013) and BioGRID interaction database (Stark, et al., 2011), 2) to develop algorithms and tools for solving biological questions, for example a reporter feature algorithm for identifying enriched biological features of a gene list (Oliveira, et al., 2008), Cytoscape for visualization of interaction networks (Shannon, et al., 2003) and metaMA for meta-analysis (Marot, et al., 2009), and 3) to apply tools and methods for extracting useful knowledge from the data, for instance genome annotations, pathway reconstruction and genome-wide expression analysis (Luscombe, et al., 2001). In this thesis, my work focuses on two roles of bioinformatics including database development and data analysis.

Availabilities of high-throughput experiments these days have leaded us to consider cells as systems where systems properties ascend from the whole rather than individual parts (Palsson, 2006). Systems biology can be considered as an approach to understand biological systems that underlie with networks of interacting components (Munoz, et al., 2012). It also brings together biologists, mathematicians, computer scientists, engineers and physicists to explore complex biological systems.

The main objective of systems biology is to obtain a quantitative representation of the system of interest which can be in the form of a mathematical model. This model is used in behavior prediction under different conditions or it is used as a scaffold for integrative analyses (Kohl, et al., 2010; Munoz, et al., 2012). The model may then be improved in an iterative fashion. To achieve these, it involves the exploitation of biological datasets including those generated from high-throughput technologies (metabolomics, transcriptomics and interatomics) and the uses of computation approaches to integrate these products for the model reconstruction and for further analyses (Figure 4).

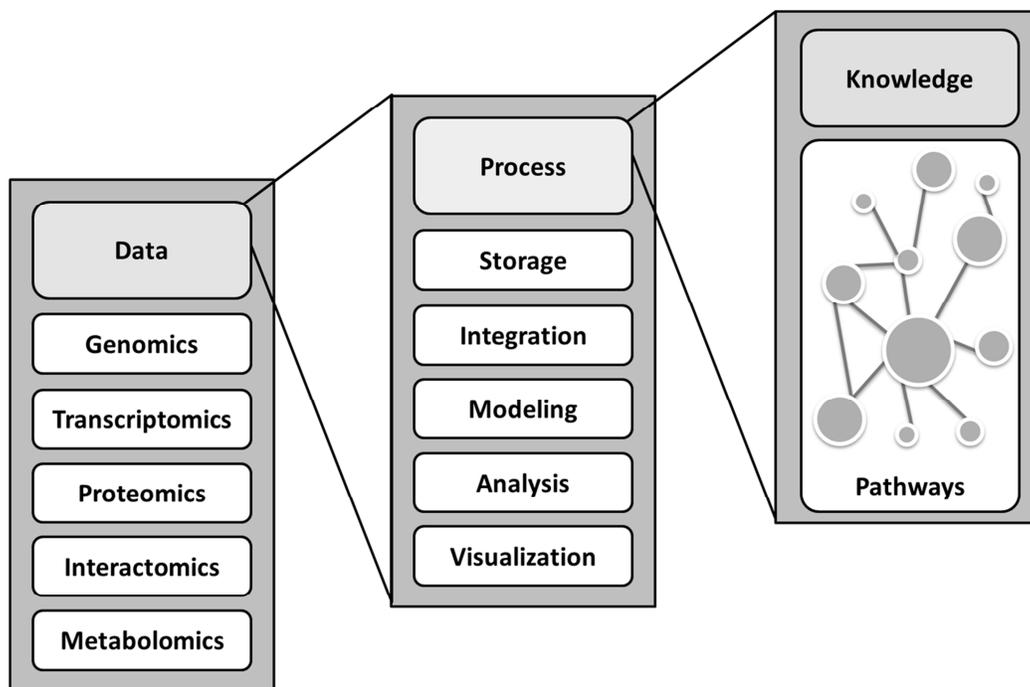


Figure 4 Bioinformatics and systems biology. Bioinformatics and systems biology are closely related with the ultimate goal to gain insights into biological systems. Systems biology studies biological systems as the integration of interacting components e.g. genes, proteins, metabolites and reactions. To integrate and analyze varied data from multiple sources, computational frameworks are required. This is where bioinformatics comes into play.

2.3 Biological database design and development

In the era of high-throughput technologies, we have seen an explosion of biological data in both the amount of data and types of data. Efficient computational infrastructure such as database system and web application for managing such data is unsurprisingly required (Berger, et al., 2013). There are a large number of public biological resources available these days and the development of such repositories continually grows.

Biological databases contain information from life sciences research ranging from raw high-throughput data to results from various analyses. Such databases can be classified in to different categories as listed in Nucleic Acids Research online Molecular Biology Database Collection (Fernandez-Suarez, et al., 2014), for instance nucleic acid sequence and structure databases (e.g. GenBank), protein sequence and structure databases (e.g. UniProt (Apweiler, et al., 2004)), metabolic and signaling pathways databases (e.g. KEGG, <http://www.genome.jp/kegg/>), organism-specific databases (e.g. SGD), and microarray and gene expression databases (e.g.

ArrayExpress). These databases are important resources for systems biology research to obtain biological insights. Proficient data management, storing and retrieving are important matters for progressing data integration and analyses (Berger, et al., 2013), however building a biological database is not a trivial task.

Biological data are complex (e.g. hierarchical structures of gene ontology), heterogeneous (i.e. composing of different types) and highly dynamic (e.g. data updates and new discoveries) (Ozsoyoglu, et al., 2006). These characteristics pose challenges in choosing existing technologies and designing database schema. In several biological databases, relational databases are used since a standard query language (SQL) is very well known and well standardized (Stein, 2013). The relational database, however, requires a predefined schema. Changing data contents might need revision of the database schema. Besides, complex and heterogeneous data result in a sophisticated schema which reduces query performance afterward. Recently NoSQL databases (e.g. document-oriented databases, graph databases and object-oriented databases) have emerged particularly for big data applications, heterogeneous data contents and data structures with complex relationships (Stein, 2013). Though such next generation databases are not much used in current biological databases (i.e. Cellular Phenotype Database (<http://www.ebi.ac.uk/fg/sym>) is only one example found), they become promising technologies for handling the massive amount of next generation sequencing data in this era.

Because high-throughput data are important for systems research to gain meaningful information. Different strategies, databases and software applications have been designed and developed toward facilitating integration of those data from various sources. Gene ontology (GO) (Ashburner, et al., 2000), Biological Pathway Exchange (BioPAX) (Demir, et al., 2010), Proteomics standards initiative molecular interaction XML format (PSI-MI) (Kerrien, et al., 2007) and Systems Biology Markup Language (SBML) (Hucka, et al., 2003) are examples of standards for representation and exchanges of biological information in which ontologies are for controlling a vocabulary of terms (Ashburner, et al., 2000) and the rest are for pathway data exchanges (Stromback and Lambrix, 2005).

The models used to integrate data in several databases and web applications include centralized and distributed model (Sreenivasaiah and Kim do, 2010). The former model has a unified schema and transfers data from diverse resources into one central repository, whereas the later has central interface to automate access across resources without data transfer. The centralized model is widely used in data warehouse (e.g. BioWarehouse (Lee, et al., 2006), cPath (Cerami, et al., 2006) and Pathway Commons (Cerami, et al., 2011)) because of its key advantages including performance and data consistency. However, its major problems are an issue on data update and

database expansion. Federated databases such as BioMart (Kasprzyk, 2011) apply the distributed model which resolve mentioned issues, instead they are limited by query speed over internet and facing interoperability of exchange protocol among data sources (Lee, et al., 2006; Sreenivasaiah and Kim do, 2010). Figure 5 illustrates the architecture of centralized and federated databases.

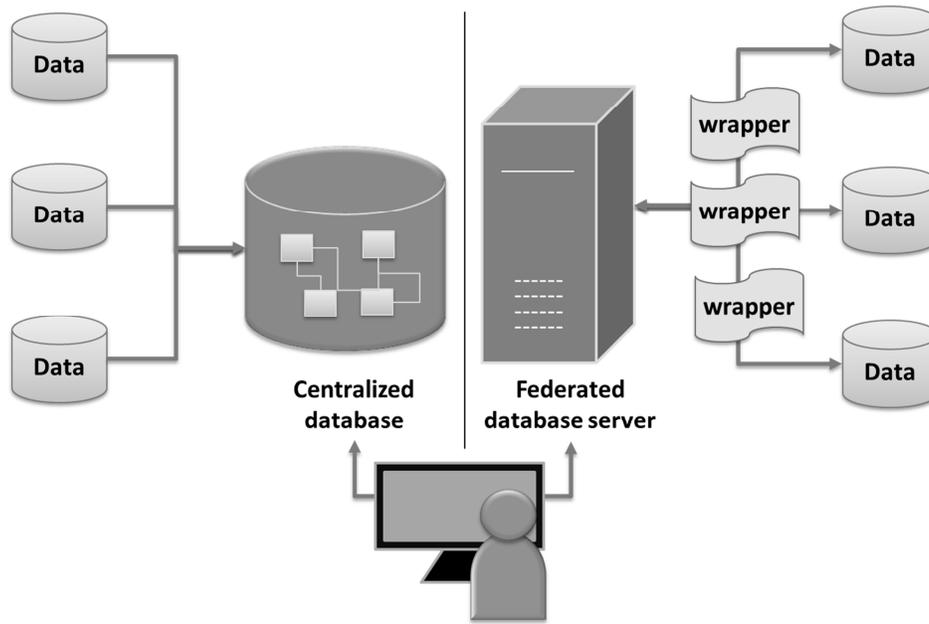


Figure 5 Architecture of centralized and federated databases. Figure on the left is the architecture of centralized system where data are transferred to store in one central repository. Figure on the right is the federated system that contains uniform interface for accessing data at remote sources.

2.4 Gene expression analysis

According to the central dogma of molecular biology, gene expression is the process when DNA sequence is transcribed into a gene product or RNA. Microarray and recently RNA sequencing (RNA-seq) are extensively used to measure gene expression levels which have been applied in several contexts including investigating gene functions, revealing regulatory patterns, studying co-expression and identifying putative markers.

Transcriptome data are accessible in several public databases such as ArrayExpress (Rustici, et al., 2013) and Gene Expression Omnibus (GEO) (Barrett, et al., 2013). Computational and statistical methods play an important role in processing and retrieving information in expression datasets. One of the most famous platforms is Bioconductor which contains R software packages for analysis of microarray and other high-throughput genomic data (Reimers and Carey, 2006). Table 1 lists resources and software packages used in this study for gene expression analysis.

Table 1 List of software packages and resources for gene expression analysis

Name	Description	Reference
Affy	R package for low-level analysis of Affymetrix GeneChip	(Gautier, et al., 2004)
PLIER	R package for normalizing Affymetrix probe-level expression data	(Hubbell, et al., 2005)
Limma	Linear model for differential expression analysis	(Smyth, 2005)
metaMA	R package for effect size and p-value combination method in meta-analysis	(Marot, et al., 2009)
Piano	R package for integrative analysis including various methods for gene set analysis	(Varemo, et al., 2013)
BioNet	R package for the integrative analysis of gene expression data in the context of biological networks to identify functional modules	(Beisser, et al., 2010)
GEO	Database of raw/processed functional genomic data including microarray and next-generation sequencing data	(Barrett, et al., 2013)
ArrayExpress	Functional genomic data repository including raw/processed microarray and next-generation sequencing data. Data are imported from GEO and from direct submission	(Rustici, et al., 2013)

Typical types of gene expression analyses include the identification of differentially expressed genes in conditions of interest comparing to reference condition using methods such as t-test statistic. Gene clustering is for pattern finding and genes grouped together are highly correlated in terms of expression profiles. Gene set enrichment analysis or gene set analysis (GSA) is an approach to discover significant biological themes or gene sets under specific states based on the incorporation of differential expression evidence to priori biological knowledge such as biological processes and metabolic pathways (Varemo, et al., 2013). This analysis helps the biological interpretation of a large gene list. There are a number of GSA methods available these days, for instance the mean of gene-level statistics (e.g. t-values) is simply set as a gene set statistic, and the reporter feature algorithm combines gene-level statistics (e.g. p-values) for a gene set (Oliveira, et al., 2008).

Nowadays there are a large number of gene expression data generated providing an important chance to increase statistical power, reliability and generalizability of analysis results. Meta-analysis is a statistical method which combines results from multiple-related studies to obtain homogeneous effects and heterogeneity across related studies (Ramasamy, et al., 2008). Several meta-analysis techniques have been implemented, for example vote counting, p-value

combination and effect size combination, the choice of which is subject to an objective of study. For the effect size combination approach which was used in this thesis, an effect size is a degree of the strength of an event or effect. In the meta-analysis, the effect size is calculated from each study and then combined to an overall size of the effect for each gene (Choi, et al., 2003; Feichtinger, et al., 2012). This method is preferential over p-value combination because it provides the magnitude of the effect.

Furthermore, availability of interactome data such as protein-protein interaction (PPI), genetic interaction and protein-DNA interaction (PDI) or transcription factor bindings allows us to perform analysis in the context of network-based analysis. Integration of transcriptome data to the network scaffold can contribute to module or subnetwork discovery. Here, gene expression results are the input for calculating the score of a module. With a search algorithm active modules with high score can be identified (Dittrich, et al., 2008). Such analysis aids our understanding of how molecules work together to drive cellular processes under a particular condition (Berger, et al., 2013).

2.5 Biological networks

Complex biological processes are executed from interactions and regulations of a number of molecules. Through high-throughput experiments, various large-scale datasets have been generated and they have been assembled into different biological networks. Biological networks compose of nodes which are biological components such as gene or protein, and edges that represent interactions among the nodes. Thus far, about five types of biological networks have been characterized: protein-protein interactions (PPIs), protein phosphorylation networks, transcription factor binding networks, metabolic pathways and genetic interactions (Zhu, et al., 2007). PPIs and transcription factor binding networks are briefly described below.

PPIs are usually represented as undirected graph. They can be identified by a variety of methods such as yeast two-hybrid (Y2H), affinity purification mass spectrometry (AP-MS) and X-ray crystallography (Koh, et al., 2012). Y2H is capable to identify a huge number of binary interactions while AP-MS is to screen interactions between several proteins. X-ray crystallography provides very detailed structure of chosen interactions. Several interaction databases are available to date. STRING database includes protein interactions from both experiments and computational predictions (Franceschini, et al., 2013). BioGRID database contains both protein and genetic interactions. It comprehensively curates interaction data generated from both low- and high-throughput experiments (Stark, et al., 2011).

Transcription factor-binding networks have been identified from direct experiments including combining chromatin immunoprecipitation with microarrays (ChIP-chip) and combining ChIP with DNA sequencing (ChIP-seq). They can also be discovered by predicting targets of a transcription factor (TF) based on binding site preferences and the sequence of promoter (Blais and Dynlacht, 2005). YEASTRACT database is an online resource that contains regulatory associations among TFs and their target genes in yeast (Teixeira, et al., 2013). This information was deduced from ChIP-chip assays and from genome-wide expression analysis where TFs were knocked out (Teixeira, et al., 2006).

Generally speaking, cellular networks are huge and complex. They can be decomposed into groups of interacting components or modules. Network modules are densely connected molecules that are formed to achieve specific functions (Kwoh and Ng, 2007; Zhu, et al., 2007). Modules can be concluded into two types: a protein complex and a functional module (Spirin and Mirny, 2003). A protein complex (e.g. complexes of TFs) comprises groups of proteins forming molecular machinery to perform a specific activity, whereas functional modules are groups of molecules that orchestrate a particular process. Functional modules can be identified by incorporating gene expression data to locate active subnetworks that have significant expression changes under specific conditions (Aittokallio and Schwikowski, 2006). The concept of network motifs is to explain patterns of subgraphs that occur recurrently in complex networks (Zhang, et al., 2005). Motifs have been investigated for better understanding of network architecture.

Though incompleteness and errors in interaction networks may limit and bias results from network analyses, the exploration of these biological networks still contributes to novel insights in understanding how cellular mechanisms are driven by underlining molecular components.

3. Results and discussion

In this chapter, I will summarize the publications that are the basis of this thesis. They can be divided into two main parts. The first part is about the design and implementation of yeast databases. The second part presents the application of bioinformatics approaches for the analysis of yeast ageing microarray data.

3.1 Design and development of yeast databases

Yeast is a model organism that has been used extensively to study essential cellular processes such as apoptosis, ageing and stress responses. As a result, a large amount of yeast-related data including comprehensive knowledge and genome data is available these days. Three yeast databases in Paper I-III were designed and developed by addressing different issues. Paper I focuses on collecting and structurally organizing related-information of a specific cell death pathway. Paper II facilitates exploitation of transcriptome data. Paper III concerns the integration of multi-level data.

3.1.1 Paper I: Yeast apoptosis database

Since apoptotic markers in yeast including DNA fragmentation, chromatin condensation and an exposure of phosphatidylserine at the cytoplasmic membrane were discovered by Madeo et al. (Madeo, et al., 1997), yeast has been used as a model organism for studying apoptosis subsequently. Apoptosis can be triggered by both endogenous triggers such as ageing and external stimuli including chemical and physical stress. The apoptotic core machinery and related pathways have been described and some have been found conserved in yeast such as apoptosis-inducing factor (AIF), endonuclease G and caspase pathway (Munoz, et al., 2012). Novel knowledge and important components are continually identified. Though, this information is dispersed in literature.

A yeast apoptosis database (yApoptosis) was implemented concerning a need of the research community for well organization of information. So that vital information can easily be shared, accessed and enhance research in yeast cell death areas. It also introduces a collaborative channel among research groups. The database was designed to collect information of curated apoptotic and related genes such as supported literature, human homologues, genome and functional information. In particular, a gene will be included into the database if it follows at least one of the following criteria:

- 1) It is annotated to the GO term ‘apoptotic process’ (GO: 0006915). There are 29 genes in the database that are assigned with this GO term.

- 2) It directly regulates the basic machinery of apoptosis. For example, PDS1 was included in the database because after apoptosis is induced, the caspase-like protease Esp1 is released from the anaphase inhibitor Pds1 (Yang, et al., 2008).
- 3) It belongs to another pathway that induces apoptosis downstream. For instance, we included RAS2 since it was reported that osmotin-induced cell death is regulated by PHO36 via a RAS2 signaling pathway (Narasimhan, et al., 2005).

Vital information for each gene such as detailed description, pathway information, protein sequence, GO annotations, links to original literature explaining the role of this gene in apoptosis and crucial external links are provided. These links to external resources facilitate exploration of other biological contexts of apoptosis genes. They include links to SGD for comprehensive biological information, UniProt for functional information of proteins, InterPro for information on protein families and domains (Hunter, et al., 2012), PSICQUIC View for molecular interactions (Aranda, et al., 2011), and Gene Expression Atlas for expression profiles in various conditions (Kapushesky, et al., 2010). This information is manually curated.

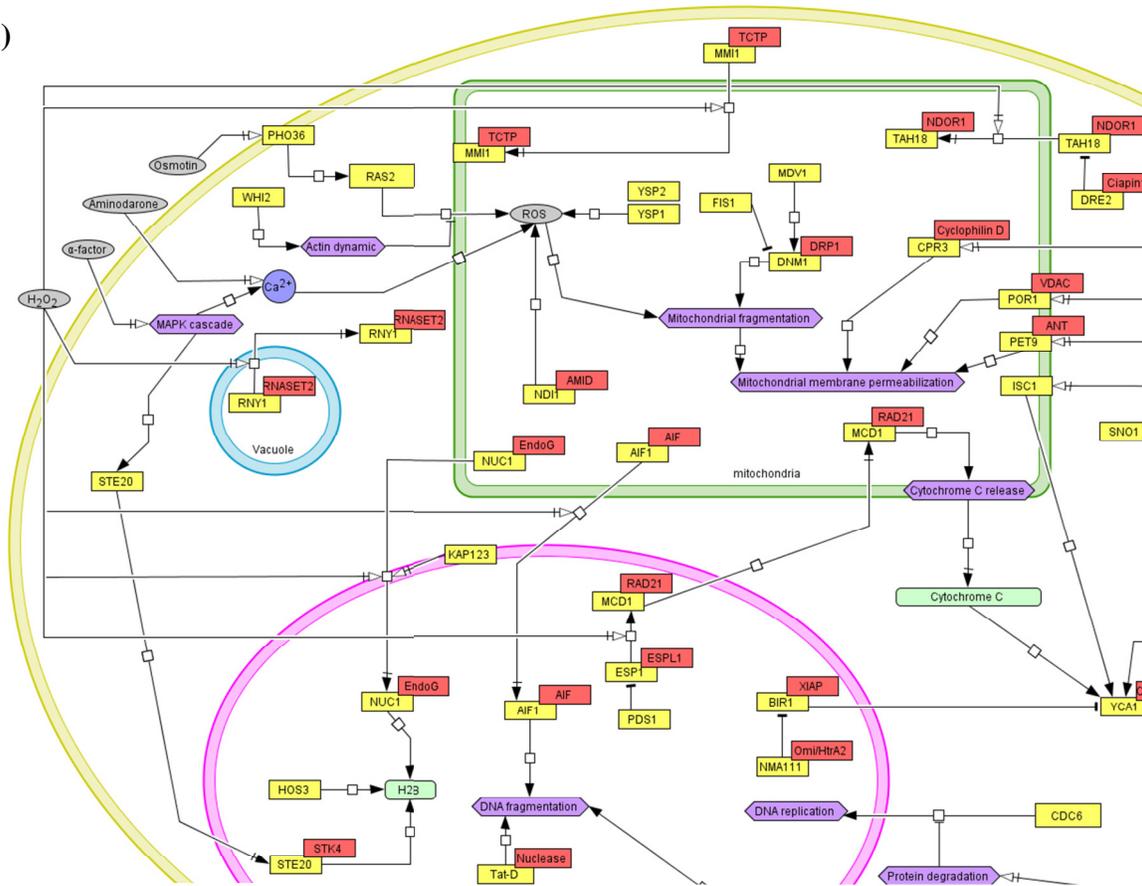
Based on extensive curation, the functional network of yeast apoptosis was drawn and also included to illustrate activities and relations between apoptotic-related components (e.g. triggers, genes and processes) in different locations (Figure 6A). Apart from cytoplasm, we found that the most number of apoptosis genes locate in mitochondria, nucleus and vacuole respectively. Mitochondria not only play a role in supplying cellular energy, it has also been accepted that mitochondria are important in execution of apoptosis both yeast and mammalian. Mitochondrial events triggering apoptosis in both mammals and yeast include mitochondrial fragmentation, collapse of membrane potential and release of AIF and cytochrome c, however the sequence of these incidences have not been confirmed (Eisenberg, et al., 2007). Components of mitochondrial death pathways in yeast (e.g. NDI1, MMI1 and DNM1) and their human orthologous (red boxes) are displayed in Figure 6A. In mammalian cells, activation of caspase by cytochrome c through caspase-9-Apaf1 pathway is well defined. Though yeast has a metacaspase Yca1 as an orthologue of mammalian caspases, yeast orthologue of Apaf1 has not been identified (Mazzoni and Falcone, 2008). An observation of how yeast caspase is activated would contribute to a complete picture of apoptosis process.

The relational database management system MySQL was applied and the friendly web interface was provided. We included typical functions to facilitate data query such as 'browse', 'quick search' and 'search' which allows specifying cellular location or process to constrain the query. The interactive representation of apoptosis networks was addressed to aid data visualization.

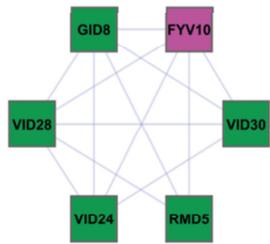
The apoptosis genes were further used in two analyses. The first analysis is predicting protein complexes from PPIs to find the complexes where apoptosis genes are a subunit using the algorithm ClusterONE (Nepusz, et al., 2012). The algorithm uses a greedy growth procedure to grow a cohesive group and calculates the cohesiveness score from edge weights to determine the possibility for a group of proteins to be a protein complex. The algorithm can also detect overlapping protein complexes based on the assumption that proteins might possess multiple functions. We found 11 apoptosis genes encoding protein subunits in 9 protein complexes. 5 complexes such as nuclear cohesin complex, small nucleolar ribonucleoprotein complex, decapping enzyme complex, chromatin silencing complex and GID complex were included in the MIPS benchmarks (Pu, et al., 2009). Apoptosis protein subunits of those complexes are Mcd1, Lsm1-4, Dcp1-2, Sir2 and Fyv10, respectively. Figure 6B shows an example of predicted complexes, GID complex, involving in proteasomal degradation of gluconeogenic enzymes: phosphoenolpyruvate carboxykinase and fructose-1,6-bisphosphatase (Santt, et al., 2008). Fyv10 is a protein subunit of the complex and it was also reported to have an anti-apoptotic role (Khoury, et al., 2008) suggesting that Fyv10 may possess dual roles.

The second analysis is to identify network modules in the integrated network of PPIs and transcription regulation interactions of apoptosis genes. We used the algorithm CyClus3D which use different types of network motifs to query the integrated network (Audenaert, et al., 2011). It allows us to find more realistic modules and investigate functional relationships between interaction types. Using this method we found 7 modules that have the same pattern of interactions. This simple pattern indicates transcriptional co-regulation of pairs of apoptosis proteins interacting together to perform their function (Figure 6C). For instance, it was reported that an oxidative stress-induced cell death in yeast is sensed by the Dre2-Tah18 complex (Vernis, et al., 2009). Higher doses of H₂O₂ destabilize its interaction causing delocalization of Tah18 to mitochondria and cell death afterward. This protein pair is regulated by Yap1, TF responds to oxidative stress. It can be seen from the network that different TFs control common pair of the proteins. This aspect supports the fact that cells have condition-specific TFs (such as Gcn4 acts in response to amino acid starvation) but they maintain common responses downstream (Gasch and Werner-Washburne, 2002).

A)



B)



C)

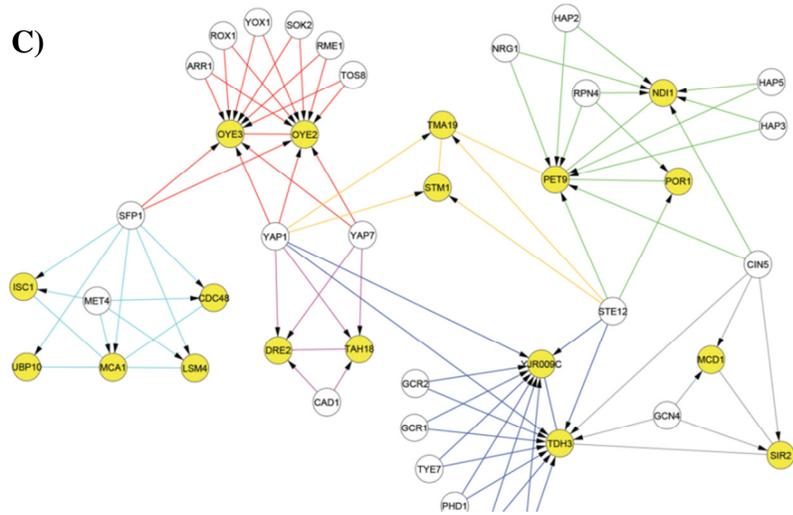


Figure 6 Networks of apoptosis genes. A) Functional network of yeast apoptosis. Human orthologues are in red boxes while yeast genes are in yellow boxes. Diagram legends can be found in detail from yApoptosis webpage (<http://ycelldeath.com/yapoptosis/index.php>). B) GID protein complex. An apoptosis-related gene, FYV10 is in purple box. C) Identified motif clusters. A directed edge represents transcriptional regulation while undirected edges are PPIs. Color of edges presents members of a cluster.

Programmed cell death (PCD) is considered as a complex process under tightly molecular controls. Based on morphological studies, there are three different phenotypes defined for PCD: apoptosis, necrosis and autophagy (Zappavigna, et al., 2013). As mentioned in the previous section, typical features of apoptosis include phosphatidylserine externalization, the condensation of chromosome and DNA fragmentation. Cell death by necrotic program shows important morphologies including increase in cell volume and plasma membrane rupture causing release of intracellular contents. Necrosis in yeast occurs after brutally exposing cells to chemical or physical substances. It also has been described to be the last fate of an apoptotic dead cell that has the loss of plasma membrane integrity due to the collapse of cellular system (Ludovico, et al., 2005). Besides necrosis is suggested to have correlations with ageing as markers of necrosis were found among dead cells in chronologically ageing colonies (Eisenberg, et al., 2010). Autophagy is a catabolic process that is conserved in eukaryotic cells. It involves the degradation of cytoplasmic components by lysosome in mammalian or vacuole in yeast to generate an amount of molecules to be recycled. Therefore, autophagy is induced in nutrient-limited conditions and it is crucial for regulation of organelle homeostasis such as degrading damaged mitochondria (Cebollero and Reggiori, 2009). In addition, an ageing process has been associated with autophagy in particular through TOR pathway (i.e. increasing autophagy results in lifespan extension.). It can be seen that autophagy itself serves a cytoprotective process (Madeo, et al., 2010; Rubinsztein, et al., 2011).

Research in yeast PCD is still ongoing. As mentioned in the previous paragraph, there are opening questions that need to be resolved. Studies on yeast necrosis are also at very beginning comparing to apoptosis and autophagy. Figure 7 illustrates molecular components in yeast PCD subroutines derived from published literature (Carmona-Gutierrez, et al., 2010; Cebollero and Reggiori, 2009; Eisenberg, et al., 2010). It can be seen that some components are shared among PCD types e.g. NUC1 and RAS2. Investigation of interplay among cell death machineries may contribute to understanding of how a cell decides its fate. Key players of mammalian PCD are presented in yeast genome supporting the use of yeast as a simple model for observing complex scenarios of mammalian PCD. The list of yeast orthologues of mammalian apoptosis components was retrieved from the database and summarized in Table 2.

In conclusion, yApoptosis is the first part of a yeast cell death database (yCellDeath) that has been established to gather information on various aspects of yeast PCD. yApoptosis represents a kind of specific repositories for sharing information and communicating between scientific communities. The total number of visitors after publishing the paper in October 2013 until May 2014 is 360 visitors. Among them, 45.6% are new visitors while 54.4% are returning visitors. An

overview of visitors is shown in Figure 8. The relational database was chosen by considering its ease of use and well standardization. We also performed network analyses using the collected genes as seed nodes to illustrate the applications of the database. The database structure of yApoptosis and data generation processes can be a blueprint for developing other PCD databases in the yCellDeath compendia.

Table 2 Yeast apoptosis genes and human orthologues

Yeast	Mammalian	Function in apoptosis
AIF1	AIF	Lead to chromatin condensation and DNA degradation (Wissing, et al., 2004)
AIM14	NADPH oxidases	Control non-mitochondrial ROS production (Rinnerthaler, et al., 2012)
BIR1	XIAP	Inhibitor of apoptosis protein (Walter, et al., 2006)
BXI1	Bax inhibitor-1 (BI-1)	Inhibitor of BAX protein (Cebulski, et al., 2011)
CDC48	VCP	Mutation causes ER stress (Carmona-Gutierrez, et al., 2010)
CDC6	CDC6	Require for DNA replication (Blanchard, et al., 2002)
CPR3	Cyclophilin D	Participate in refolding protein after import into mitochondria (Liang and Zhou, 2007)
DNM1	DRP1	Mitochondrial fragmentation (Fannjiang, et al., 2004)
DRE2	Ciapin1	Anti-apoptosis by binding to Tah18 (Vernis, et al., 2009)
ESP1	Separin	Caspase-like protease targets Mcd1 (Yang, et al., 2008)
MCD1	RAD21 (human cohesion)	Decrease mitochondrial membrane potential (Yang, et al., 2008)
NDI	AMID	Increase ROS production in mitochondria (Li, et al., 2006)
NMA111	Omi/HtrA2	Serine protease targets anti-apoptotic protein Bir1 (Walter, et al., 2006)
NUC1	Endonuclease G (EndoG)	Major mitochondrial nuclease (Buttner, et al., 2007)
PET9	VDAC	Mitochondrial permeabilization (Pereira, et al., 2007)
POR1	ANT	Mitochondrial permeabilization (Pereira, et al., 2007)
RNY1	RNASET2	Cleave tRNA (Thompson and Parker, 2009)
Tat-D	Nuclease	DNA degradation (Qiu, et al., 2005)
YCA1	Caspases	Caspase-like enzymatic activity (Mazzoni and Falcone, 2008)

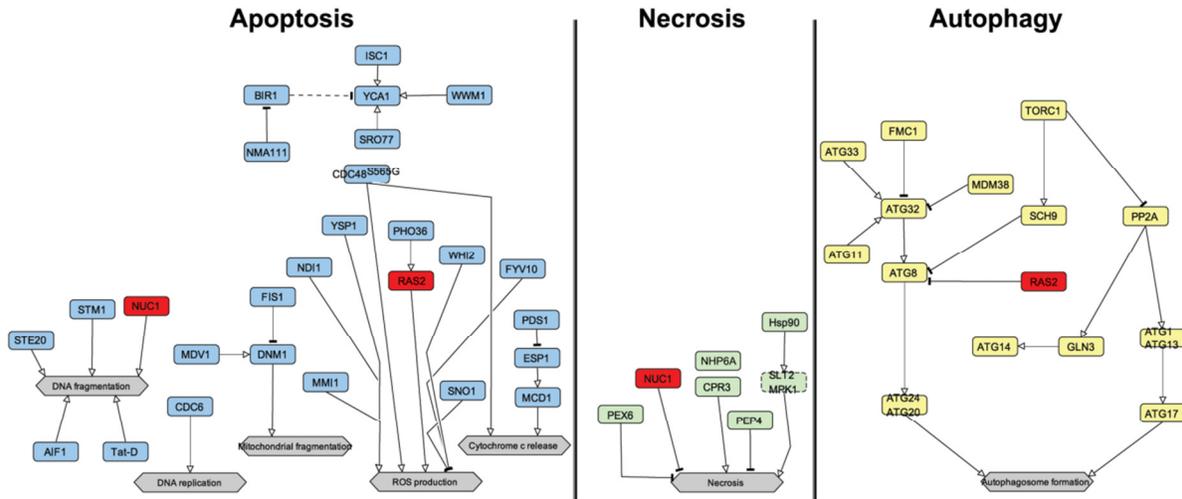


Figure 7 Networks of yeast PCD. The figure illustrates three modes of PCD: apoptosis (blue), necrosis (green), and autophagy (yellow). Different phenotypes represent in grey boxes. Nodes marked in red involve in more than one mode of PCD. Solid edges denote either triggering or inhibition. Dashes denote putative genes or interactions that have not been confirmed. Figure adapted from (Munoz, et al., 2012).

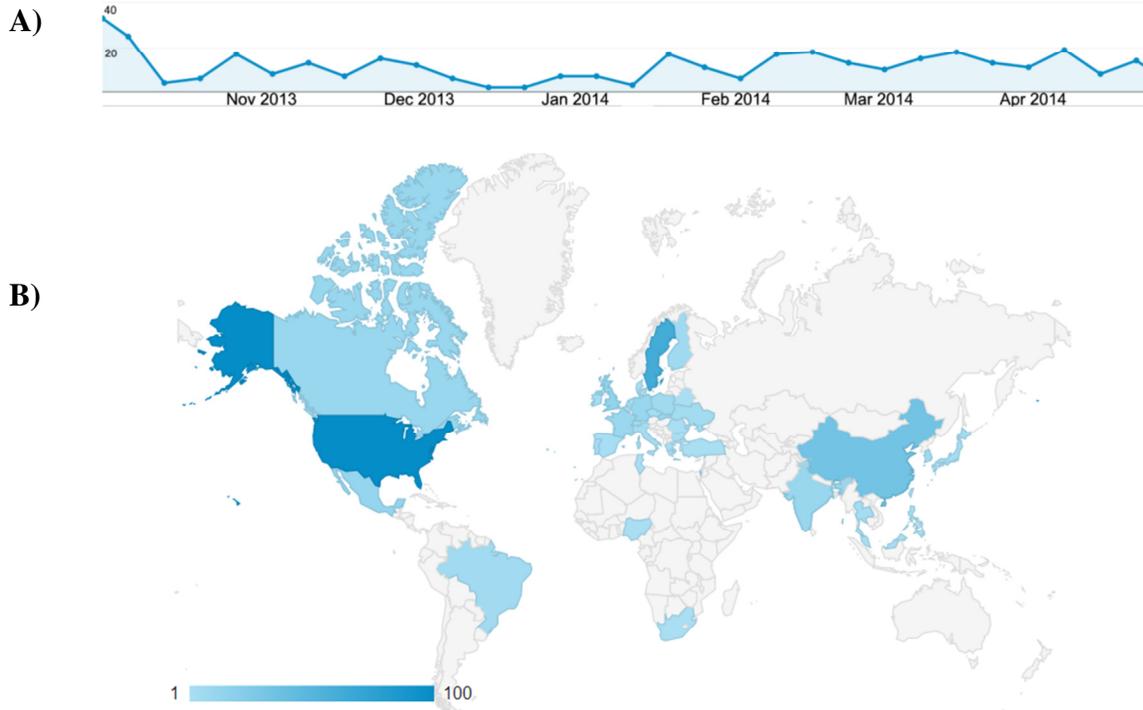


Figure 8 Overview of visitors to yApoptosis. A) The number of visitors in weeks. B) The number of visitors by countries. The color-scale is based on the number of visitors. Visiting statistics were generated from Google Analytics (<http://www.google.com/analytics/>) starting from October 2013 to May 2014.

3.1.2 Paper II: Yeast stress expression database

Similar to other organisms, yeast has to face several kinds of changes in its environments, for example fluctuations in pH, temperature, nutrient level, osmotic pressure, amount of oxygen, and the presence of various agents like toxic compounds and drugs. Appropriate responses to these variations are inevitably required for competitive fitness and cell survival. It was found that the stress-response schemes in yeast count significantly on genome-wide transcriptional changes (Causton, et al., 2001). As a result several studies used genome-wide expression experiments such as microarray to measure gene expression levels of yeast under different environments (Causton, et al., 2001; Gasch, et al., 2000). Nowadays, most gene expression data are deposited in public repositories such as ArrayExpress (Rustici, et al., 2013) and GEO (Barrett, et al., 2013). Even though these data are made available online, efficient utilization of the data from those databases is not straightforward.

We therefore developed a yeast stress expression database (yStreX) showing an effort to enhance exploitation of transcriptome data in the area of yeast stress responses. We adapted the concepts of Experimental factors (EFs) and EF values used in Gene expression atlas database (Kapushesky, et al., 2010) to describe experimental conditions in a united way allowing comparison between independent experimental setups and meta-analyses among related studies). Figure 9 shows the workflow of data preparation and analysis.

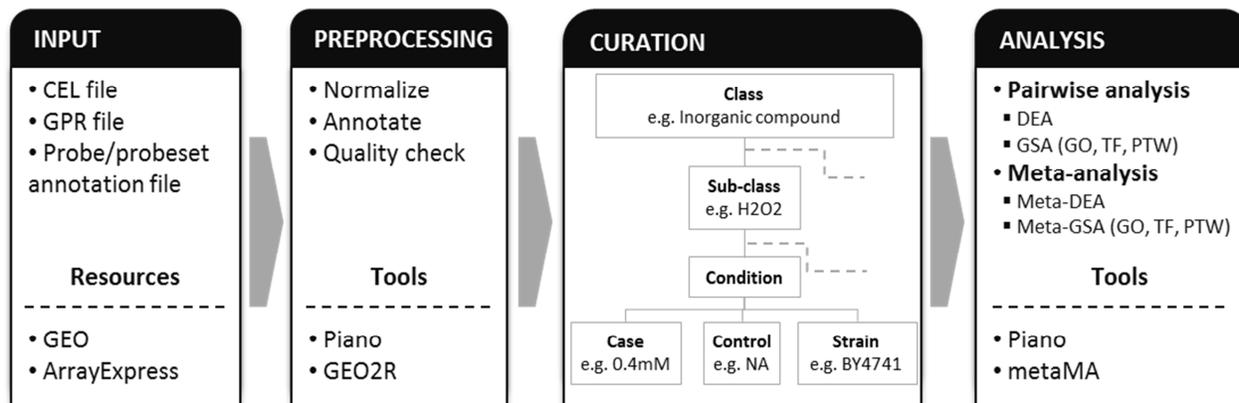


Figure 9 Analysis workflow. The diagram shows the workflow of data preparation and analyses. Microarray datasets were retrieved from GEO and ArrayExpress database and were preprocessed before curated into defined experimental classes and sub-classes. Statistical analyses were performed including DEA, GSA and meta-analysis. Data sources and tools used are also listed in the boxes.

Typical tasks of gene expression analyses including differential expression analysis (DEA), gene set analysis (GSA) and meta-analysis were performed on microarray datasets generated from different stress-induced conditions. R is the main platform for statistical computing in this study. Limma R package (Smyth, 2005) was used to calculate gene-level statistics (including fold-changes, t-values and p-values) for each gene on each experimental condition. For GSA, Reporter algorithm (Oliveira, et al., 2008) was used to compile significant gene sets on three classes of existing biological knowledge: GO, TF and pathway (PTW) in each experimental condition. Meta-analysis was performed on datasets from two or more independent, but relevant studies using a moderated effect size combination in metaMA R package (Marot, et al., 2009). The effect size was calculated for each gene in each study and then was combined before assessing for DE genes of a specific condition. The resulting test statistics, or we call meta-z-scores, were used in GSA for computing significant gene sets. Detailed descriptions about statistical analyses can be found in Paper II.

The document-oriented database MongoDB (Chodorow and Dirolf, 2010) was used to store dataset information and results from the analyses as JavaScript object notation (JSON) documents. The web interface with different functions to access and explore the analysis results including statistical values of genes and enriched biological features was provided. Mainly there are two approaches to query the data: 1) searching by a gene or a set of genes for conditions where the queried genes significantly expressed, and 2) querying by a condition to retrieve a set of differentially expressed genes from the meta-analysis of related studies.

Two examples were included to show applications of the database. The first example illustrated concurrent query of 51 apoptosis-related genes from yApoptosis database to find significant genes under oxidative stress induction by H₂O₂. It also showed database features and navigations between different pages such as Gene summary, Condition summary and GSA result (Figure 10). The second example showed the results from the meta-analysis, which combines gene expression data from independent, but related experiments. In this example, these experiments compared Rapamycin treated (TOR complex 1 (TORC1) inhibition) to untreated condition. It has been described TOR pathway coordinate signals with other signaling pathways including Sch9 and PKA pathway to regulate stress response, cell growth, ribosome biogenesis, and other cellular processes (Cheng, et al., 2007; Wei, et al., 2009). In accordance with the report in literature, common stress response TFs such as Gis1, Msn2 and Msn4 were found enriched in up-regulated genes. The consensus heatmap of TFs gene sets is shown in Figure 11.

Gene summary

YNR074C (AIF1)

Systematic name: YNR074C (GSD_300005257)

Standard name: AIF1

Description: Mitochondrial cell death effector that translocates to the nucleus in response to apoptotic stimuli, homolog of mammalian Apoptosis-Inducing Factor, putative oxidase

Gene ontology: [Click to show \(1\)](#)

Significant conditions: (Total found = 33, |log2FC| > 1, adj. p-val < 0.01)

Refine the resulting table by providing a new cut-off:
 Results are sorted descending by log2FC

log2FC > adj. p-val < **SUBMIT**

Class	Sub-class	Control	Case	Strain	Platform	LogFC	Pval	GSA
Inorganic compound	Hydrogen peroxide (H2O2)	Control	0.4mM	BY1471 (HMG mutant)	A-GEO0.5915_U1MAY5 Yeast Bk v1.0	4.545	0.000193300645	GO1 TPTW
Extra stress	Oxidative stress	Control	0.4mM	BY1471 (HMG mutant)	A-GEO0.5915_U1MAY5 Yeast Bk v1.0	4.545	0.000193300645	GO1 TPTW
Inorganic compound	Hydrogen peroxide (H2O2)	Control	0.4mM	NSY429	A-GEO0.5915_U1MAY5 Yeast Bk v1.0	3.101	0.000666232822	GO1 TPTW
Extra stress	Oxidative stress	Control	0.4mM	NSY429	A-GEO0.5915_U1MAY5 Yeast Bk v1.0	3.101	0.000666232822	GO1 TPTW
Inorganic compound	Hydrogen peroxide (H2O2)	Control	0.4mM	BY1414_Δel_p023	A-GEO0.5915_U1MAY5 Yeast Bk v1.0	3.007	0.0002765455613	GO1 TPTW
Extra stress	Oxidative stress	Control	0.4mM	BY1414_Δel_p023	A-GEO0.5915_U1MAY5 Yeast Bk v1.0	3.007	0.0002765455613	GO1 TPTW
Inorganic compound	Hydrogen peroxide (H2O2)	Control	0.4mM	BY1414_Δel_p023	A-GEO0.5915_U1MAY5 Yeast Bk v1.0	2.923	0.00017471918107	GO1 TPTW
Extra stress	Oxidative stress	Control	0.4mM	BY1414_Δel_p023	A-GEO0.5915_U1MAY5 Yeast Bk v1.0	2.923	0.00017471918107	GO1 TPTW

Click on a case to go to Condition summary page
 to go to GSA result page

GSA result

Condition details:

- Class: Inorganic compound
- Sub-class: Hydrogen peroxide (H2O2)
- Control: Control
- Case: 0.4mM

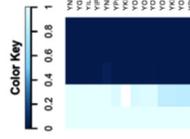
Specify cut off: 0.001

Downloaded data: **Download** **Save heatmap** **Save**

Number of TF gene sets: 64

Heatmap Data

A heatmap of selected gene set with color-scale based on p-values and gene expression directionality on each column is as following:
 Dir_Dn (distinct-directional, down-regulated), Mix_Dn (mixed-directional, down-regulated), Non_dir (non-directional), Mix_Up (mixed-directional, up-regulated).
 See Piano for more details.



Condition summary

Experimental class: Inorganic compound

Experimental sub-class: Hydrogen peroxide (H2O2)

Control: Control

Case: 0.4mM

Strain: BY1472

Raw data ID: E-GEO0-16346 (biological replicates)

Publication: PMID: 20109270

GSA: GO1 TPTW [Click on a type of gene set to go to GSA result page](#)

Differentially expressed genes: (Total found = 602, |log2FC| > 1, adj. p-val < 0.01)

Refine the resulting table by providing a new cut-off:
 Results are sorted descending by log2FC

log2FC > adj. p-val < **SUBMIT**

Gene	LogFC	Pval
SRX1	5.011	2.9E-12
YKL071W	5.789	3.4E-13
FIM2	4.63	1.2E-11
OYE3	5.037	1.2E-11

Click on a gene to go to Gene summary page

Figure 10 Screenshot of navigations between different web pages. Text in bold addresses an action that can be performed in each box. Gene summary page contains information about a gene and Condition summary page includes details about a selected condition. GSA result page contains both the heatmap and the data table of the selected gene set, TF gene sets in this example. The color intensity is based on p-values.

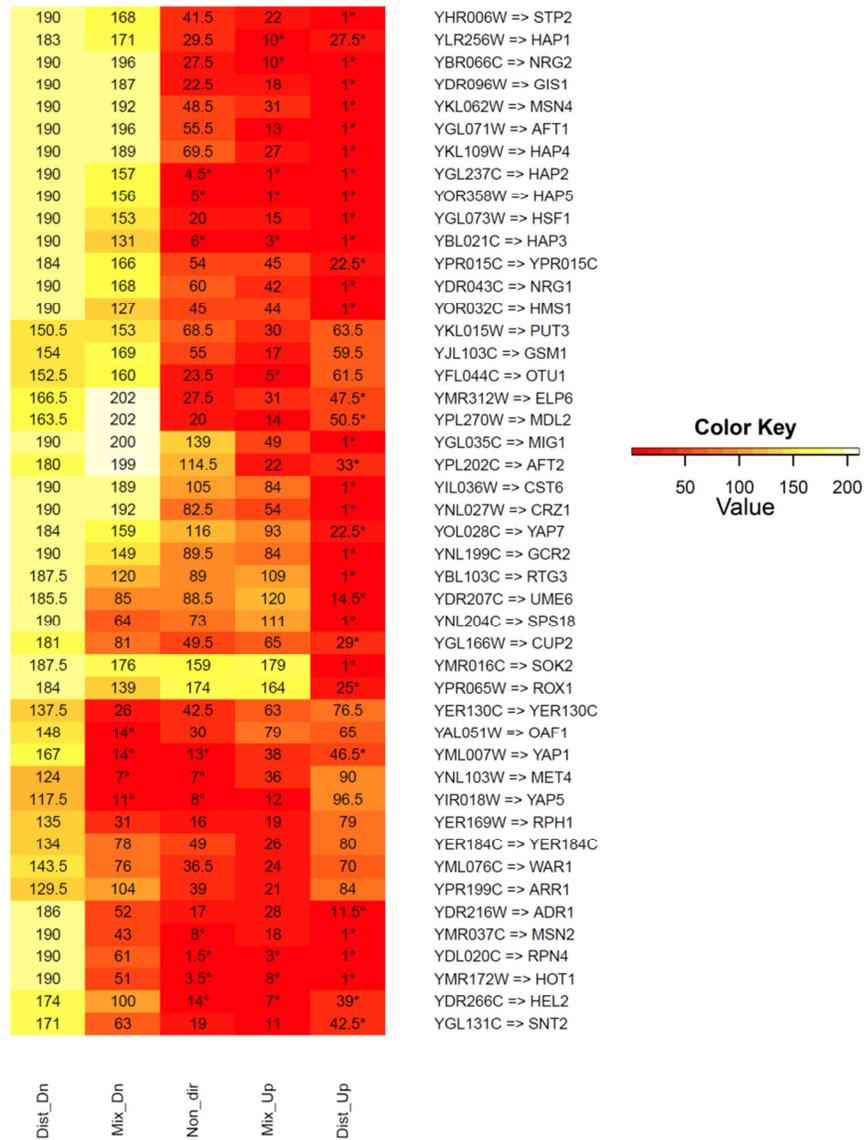


Figure 11 Part of the consensus heatmap of TF gene sets. The heatmap is a result from GSA on TF gene sets. The studies compared Rapamycin treated to untreated condition. The color-scale is based on the consensus score represented also by a number. Significant gene sets (p -value < 0.05) are marked with an asterisk. Each column indicates the gene expression directionality of the genes belonging to the TF gene sets.

In summary, yStreX is an online repository that highlights the needs to efficiently utilize and distribute gene expression data. It includes not only readily-analyzed gene expression but also enriched biological features of gene lists which would help scientists in the field to make sense of the data. A friendly web interface provided enables fast access and assists inference from these resulting data. Besides, creating the specific compendia of transcriptome data is enable us to perform meta-analyses regardless of different platforms and laboratories which can enhance statistical power, reliability and generalization of results.

The database aims to serve experimental molecular biologists who do not wish to re-analyze data. This aspect is usually a limiting factor for databases that just have a collection of raw data. Though such databases can be used in broader applications, they expect users to know in advance how to process the data and how to carry out data analysis. To compromise this issue, a link back to the original data was provided allowing users to access raw data and achieve additional work which might not be in the extent of the yStreX.

yStreX was developed on a document-oriented NoSQL database showing the use of a next generation database. The NoSQL database was chosen as it has appeared to be a preferred database choice for big data applications and a predefined schema is not needed in contrast to typical relational databases (Couchbase, 2013). The last aspect benefits for scaling out the database when the amount of data grows in the future.

We also demonstrated by two examples how the database can be exploited. In addition to these, pre-analyzed gene expression data can be broadly applied, for instance integrating to network scaffolds for identification of functional modules or using as additional information to improve PPI confidence (Bader, et al., 2004; Beisser, et al., 2010).

3.1.3 Paper III: Yeast data repository

The main aim of systems driven research such as systems biology is to understand how molecules, pathways and networks organize to drive complex biological systems (Sreenivasaiah and Kim do, 2010). This is usually done by integration of data from different levels including genome, transcriptome, proteome, metabolome, interactome and reactome to formulate a model that describes how the systems work (Ideker, et al., 2001). Advances in high-throughput technologies these days have resulted in explosive generation of multi-level of omics data which is, therefore, beneficial for driving systems biology research. However it is a challenging issue for data integration when facing with complex, heterogeneous, highly dynamic, incomplete and disassembling characteristics of those data.

With those in mind, a database system for handling multi-level data was developed. The database system aims to solve the vital issues in data management and to facilitate data integration, modeling and analysis within a single database. The database schema was adapted from ontology classes of BioPAX (Demir, et al., 2010) and implemented using an object-oriented concept. This concept represents a biological component as an object containing important attributes and a variety of relationships. It is applicable for heterogeneous and sophisticated biological data (Cooray, 2012; Okayama, et al., 1998).

A yeast data repository was then developed on top of the database system to represent a sole database environment or a centralized database underlining with multiple levels of yeast data (e.g. genome information, annotation data, interaction data and metabolic model) from different resources. List of the data is summarized in Table 3 and Figure 12 illustrates the database architecture.

We demonstrated usability and applications of the yeast data repository to achieve different tasks. A simple web interface was provided including a basic query interface which allows searching for different objects such as gene, protein and metabolite. Search results include not only the matched object but also other essential objects related to the resulting object. Figure 13A schematically presents results from querying a gene INO1. In general, it has been described that DNA makes RNA and RNA encodes protein. A protein then performs its function e.g. as an enzyme in a biochemical reaction. Based on yeast data in the database, related objects such as interacting partners of INO1-encoded protein, metabolites of the reaction catalyzed by INO1-encoded protein and TFs regulating expression of INO1 were reported as the part of the queried results. Furthermore, two research cases were conducted. First, the pheromone pathway segment was retrieved from PPI data using a protein receptor Ste3 as a starting protein and a TF Ste12 as an ending point. Reconstruction of PPI networks can contribute to identification of unknown

components and paths in signaling pathways. A vital issue of the PPI data is the dominance of false positives. Several efforts have been made to improve the confidence of the interactions including integration of GO annotations and gene expression data (Arga, et al., 2007). To simplify the case, specific GO terms were used as constraints to exclude proteins that are not relevant to the pheromone response pathway. The resulting pathway is displayed in Figure 13B. Secondly we identified downstream metabolic reactions that are regulated by TF targets of Snf1. This was done by first querying for TFs phosphorylated by Snf1 from kinase interaction data. Then from the transcriptional regulatory network, we retrieved gene targets of the phosphorylated TFs and we further identified metabolic reactions where these protein-encoding genes participate in (Figure 13C). These genes are involved in the glyoxylate cycle, amino acid biosynthesis, glycolysis / gluconeogenesis, acetate transport and oxidative phosphorylation.

In brief, we established the yeast data repository as an integrated platform of multi-level data. Different applications were demonstrated showing capabilities and effectiveness to perform specific tasks on a single database environment. The database was populated and managed using the implemented database system that was designed to handle complex data ensuring data integrity, consistency and reliability, and to facilitate data integration. Nonetheless, its major issues as a centralized database are data update and database expansion, its key advantages includes performance and data consistency. The data consistency is one of the key matters that needs to be addressed in data integration.

Table 3 List of yeast data in the yeast data repository and resources

Biological object	Physical object	Resource	Total
Chromosome	DNA	NCBI GenBank (Benson, et al., 2013)	17
Gene	DNA region	Ensembl (Flicek, et al., 2013)	7126
RNA Transcript	RNA	Ensembl (Flicek, et al., 2013)	7126
Protein	Protein	UniProt (Magrane and Consortium, 2011)	6617
Metabolite	Small molecule	iTO977 (Osterlund, et al., 2013)	484
Metabolic reaction	Biochemical reaction	iTO977 (Osterlund, et al., 2013)	717
Protein-protein interaction	Molecular interaction	BioGRID V3.2 (Stark, et al., 2011)	72453
TF-binding interaction	Control	YEASTRACT (Teixeira, et al., 2013) and (Harbison, et al., 2004)	48548
Kinase interaction	Control	(Breitkreutz, et al., 2010)	1333
Phosphorylase interaction	Control	(Breitkreutz, et al., 2010)	254

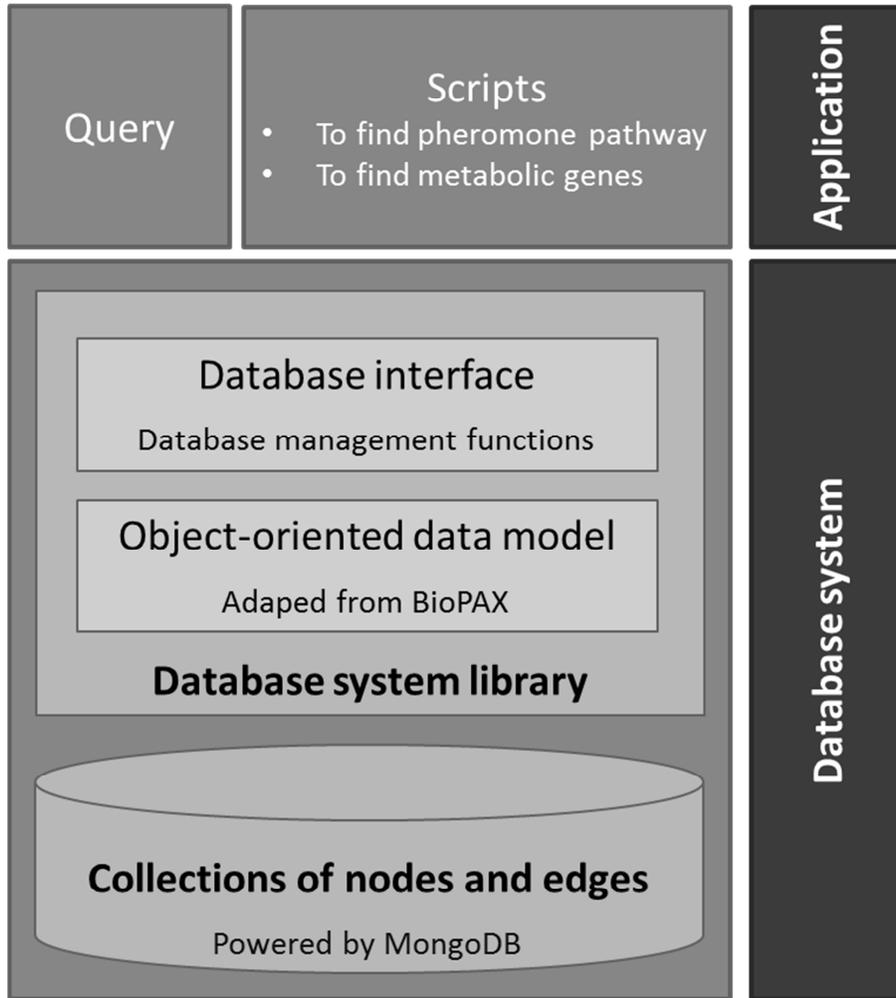


Figure 12 Yeast data repository architecture. The yeast data repository was implemented as the application on top of the database system. This database system composes of database management functions to manage operations between developers and the system, an object-oriented data model to represent biological objects, and a physical storage. The yeast data repository contains simple query interface and specific scripts to demonstrate two sample cases: querying the pheromone pathway in protein interaction networks and finding metabolic reactions regulated by Snf1 kinase.

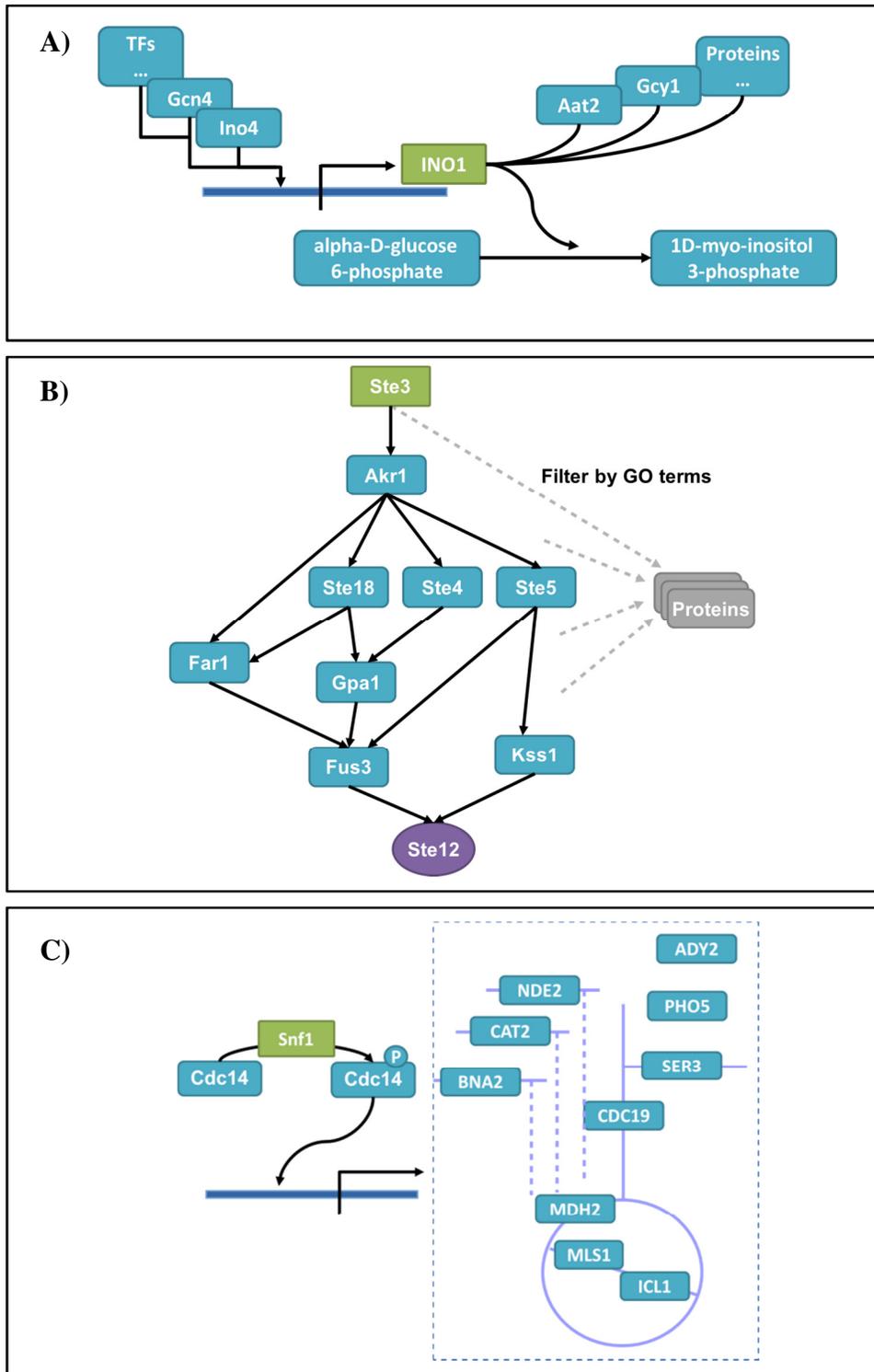


Figure 13 Resulting queries from the yeast data repository. A) Resulting objects from querying gene INO1 include TFs, interacting proteins and the list of metabolites. B) Pheromone pathway segment. C) Metabolic genes regulated by Cdc14, a TF target of Snf1. Green box is a queried term while blue rounded squares are the results of the query. A purple oval is the constraint of the query.

3.2 Bioinformatics applications in transcriptome analysis

Gene expression profiling is often used as a channel to study expression and regulatory patterns in different contexts. Computational approaches are inevitably required to make sense of such data. Paper IV outlined different bioinformatics methods used to analyze gene expression data of yeast aged chronologically.

3.2.1 Paper IV: Genome-wide expression analyses of the stationary phase model of ageing in yeast

Yeast has been used as a model to study two ageing processes in eukaryotes: replicative lifespan (RLS) and chronological lifespan (CLS). This is not only because it obtains easily manipulable genome, but conserved lifespan-regulatory components have been identified in yeast (Longo and Fabrizio, 2012). These include Sch9, Ras/PKA and TOR pathway. In this study, yeast CLS was the main focus. The yeast CLS has been considered to be an ageing model of post-mitotic cells of higher eukaryotes. To measure CLS, yeast is grown in synthetic complete medium (SD) without replenishing which leads to cell cycle arrest (Longo and Fabrizio, 2012). Under this condition, yeast cells are stressed from nutrient depletion and accumulation of toxic substances which promotes reprogramming of transcription machineries by several signal transduction pathways to maintain viability of major population (Galdieri, et al., 2010). DNA microarray has been extensively used to capture the transcriptional state of cells under different conditions. Gene expression assays have been included in ageing research to study transcription program under extreme calorie restriction that accounts for lifespan extension (Cheng, et al., 2007).

To explore changes in transcriptional regulations, biological processes and metabolisms of yeast in stationary phase, we grew *S. cerevisiae* wild type strain 113-7D in 2%-glucose medium (SD) and collected at log phase (Log), 2 (D2), 6 (D6) and 10 (D10) days of incubation, each in triplicate for microarray assay. Typical tasks of gene expression analysis performed in this study include: 1) identification of genes differentially expressed between aged yeast cells (D2, D6 or D10) against logarithmic growth cells (Log) with the linear model fitting procedure of the limma package (Smyth, 2005), 2) gene set analyses using a reporter features method in the Piano (Oliveira, et al., 2008; Varemo, et al., 2013) to identify enriched biological processes (BP), TFs and metabolic pathways of differentially expressed genes in each day relative to Log, and 3) integrated analysis to characterize active subnetworks of yeast at each day using BioNet R-package (Beisser, et al., 2010). Identification of genes significantly changed between two samples may supply putative markers and contribute to understanding of the molecular basis of varied conditions. However it usually comes with high dimension of the data or long list of genes. The gene set analysis helps reducing data dimensionality and biological interpretation of

the large gene lists. Existing biological information can be used to outline gene sets such as GO terms, TFs and pathways. Different types of gene sets provide information in different features. Assimilating information from these different gene sets provided a more comprehensive overview of the cellular changes under specified conditions. Integrating transcriptome data to network scaffold is to identify active modules or subnetworks revealing connected regions of the network that components significantly change in expression and enable us to identify key components based on network topology.

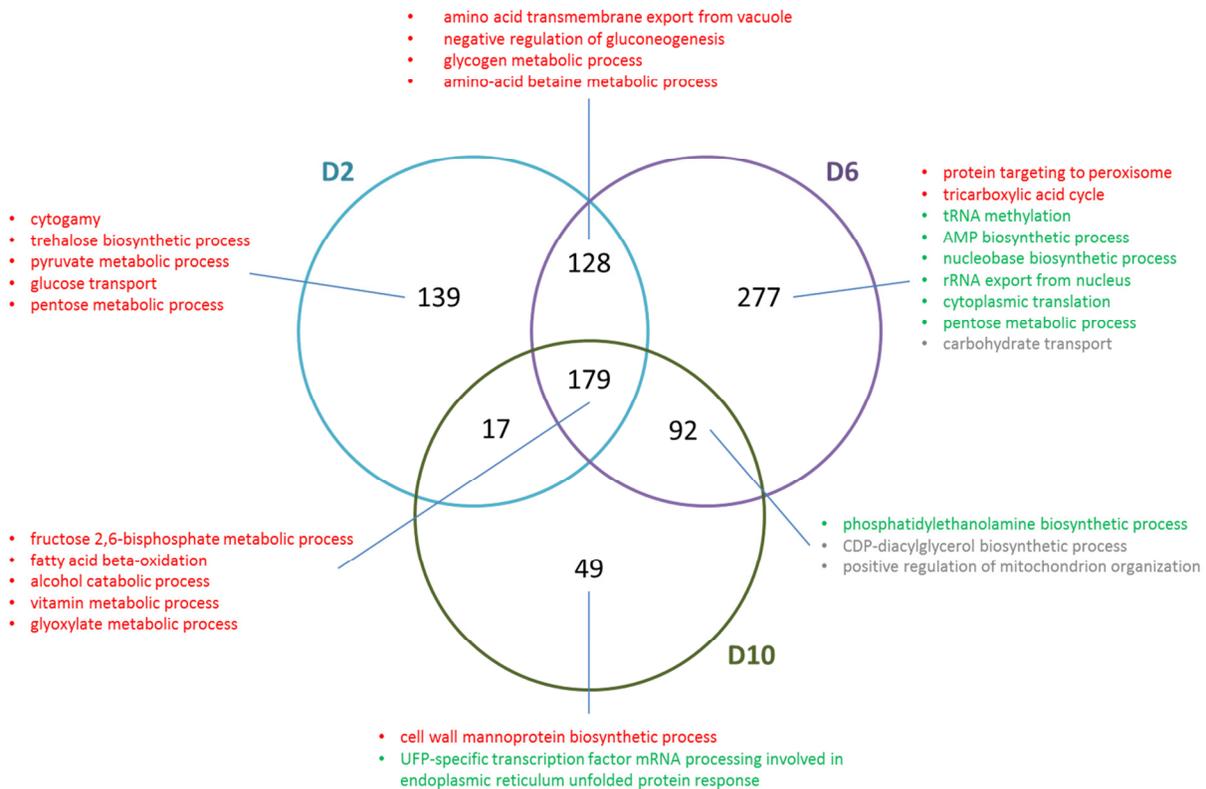


Figure 14 Venn diagram of significant genes and enriched functional terms. Significant genes were identified from comparing D2, D6 or D10 cells to Log phase cells. Functional terms enriched in the significant genes are listed. The red color is up-regulation while green indicates down-regulation and gray indicates equal number of up- and down-regulated genes associated to that functional term.

When comparing the gene expression of yeast cultured in stationary phase for several days to the log phase cells, we found in total 882 significant genes (cutoff adjusted p-value < 1e-09). The largest number of changes in gene expression was observed in D6 compared to Log, while the least was in D10. We then used the tool ClueGO (Bindea, et al., 2009) which is based on

hypergeometric test to acquire an overview of enriched biological processes among the list of significant genes compared to the Log. The results are shown in Figure 14. It can be seen that among up-regulated genes in D2, cytogamy, trehalose biosynthetic process and pyruvate metabolic process were enriched. Genes involving negative regulation of gluconeogenesis and glycogen metabolic process were up-regulated in both D2 and D6, and genes associated with tricarboxylic acid cycle (TCA) were up-regulated in D6. Though in D6, genes involved in bases biosynthetic processes, rRNA export from nucleus and cytoplasmic translation were mostly down-regulated, indicating that cell growth decreased. Most of genes participating the alcohol catabolic process, fatty acid beta-oxidation, and glyoxylate metabolic process were up-regulated in all cultures (D2, D6 and D10), compared to Log.

An important issue of ClueGO is that it requires a predefined list of genes which may cause information loss. To aid data interpretation and dig into what processes are affected during the aging in stationary phase, we therefore performed gene set analysis using the reporter feature method to investigate into three levels of biological information including biological processes (BP), TFs and metabolic pathways. The method incorporates gene-level statistics and considers directionality of gene expression changes (Varemo, et al., 2013).

We considered only gene sets of distinctly up- or down-regulated genes. As found in metabolic pathway-GSA, for stationary phase cells there was the expression of metabolic genes. These genes, which were all up-regulated, affected metabolic pathways such as TCA cycle, fatty acid degradation and glycerol metabolism (Figure 15). By integrating information from these different gene sets, we also summarized crucial physiologies from different types of gene sets in Table 4. For instance, stationary phase yeast was slowly growth or mostly no growth. This is inferred from the following gene sets: translation (BP), ribosome biogenesis (BP), nucleotide biosynthetic process (BP), Ifh1 (TF), tRNA synthesis (pathway), pyrimidine metabolism (pathway), and purine metabolism (pathway), which genes were all down-regulated.

Integrative analysis was performed to identify active subnetworks from an integrated network of PPIs and transcriptional regulatory network. Adjusted p-values at each day (D2, D6 and D10, each vs Log) were integrated to the network for calculation of high-scoring subnetworks based on the signal-noise decomposition and heuristic search (Dittrich, et al., 2008). The top 10 connected nodes (hubs) were retrieved using the 'degree sorted circle' layout in Cytoscape (Shannon, et al., 2003). A network of gene-metabolic pathways was then incorporated. The results showed that the sizes of active subnetworks are ranging from D10, D2 and D6 indicating highest expression activities on D6. This is also in accordance with our differential expression analysis results. From the active subnetworks identified on D2, D6 or D10, we found most of the

hubs were TFs for nutrient responses (e.g. Yap1, Xbp1 and Hsf1) and stress responses (e.g. Adr1, Gis1 and Pgm2). The active subnetworks comprised components of signaling pathways, for instance GPG1, GPA2, RAS1, TPK1, FUS1 and FUS3. The metabolic pathways that prominently change in all subnetworks are carbon metabolism (e.g. TCA cycle, fatty acid degradation, glycolysis/gluconeogenesis, oxidative phosphorylation, starch and sucrose metabolism, and trehalose metabolism) and amino acids metabolism. Most of genes for amino acids metabolism were down-regulated whereas most of genes for carbon metabolism were up-regulated. In addition, the genes involved in nucleotide synthesis were down-regulated in every subnetwork.

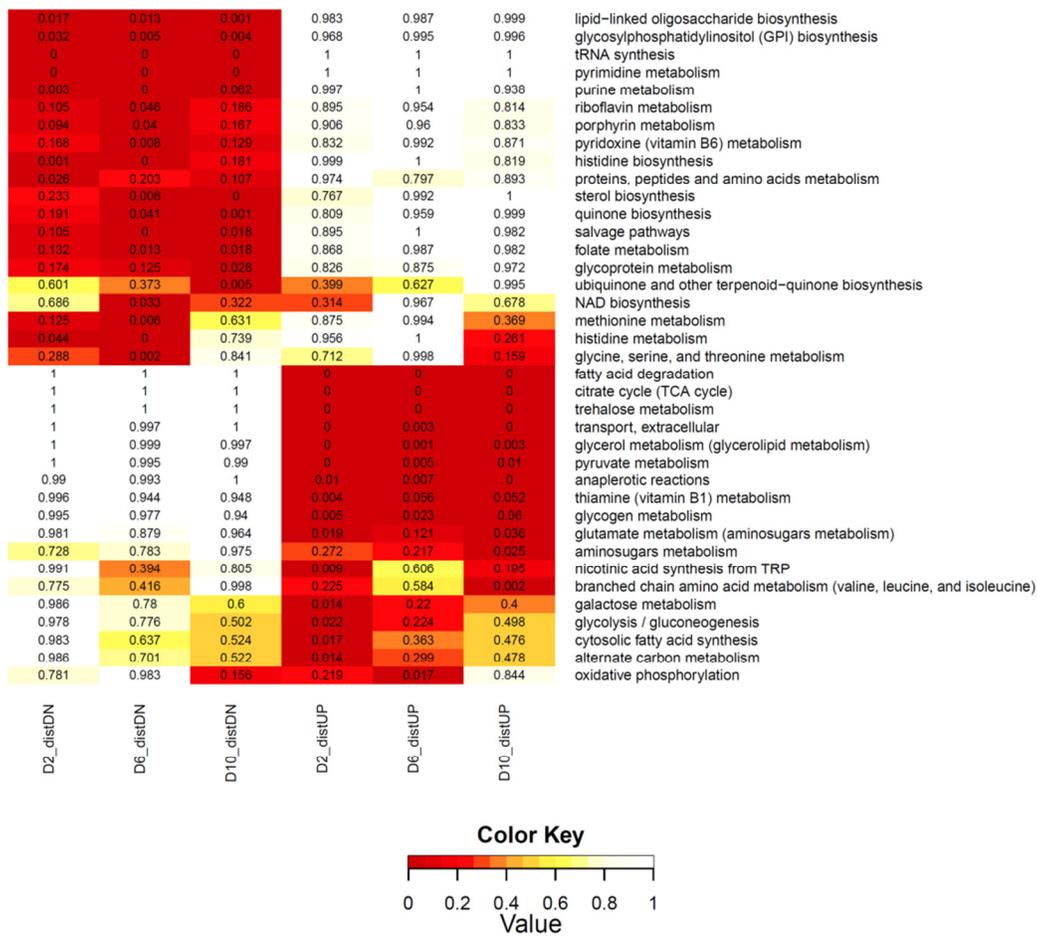


Figure 15 Heatmap of enriched metabolic pathways from GSA. P-value is also written as a number and determines the color intensity. On each column, gene expression directionality is as following: D2_distDN (down-regulated genes in D2), D6_distDN (down-regulated genes in D6), D10_distDN (down-regulated genes in D10), D2_distUP (up-regulated genes in D2), D6_distUP (up-regulated genes in D6) and D10_distUP (up-regulated genes in D10).

Mainly, signals are passed through components of signal transduction pathways (GPA2 and GPG1) to downstream TFs (Gis1, Yap1, Adr1 and Xbp1). The TFs transcriptionally regulate the expression of several stress-respond genes (SNO4, SSE2 and HSPs) and metabolic genes (POX1, FOX2 and GSC2) as shown in Figure 16. TFs found in all subnetworks include Ste12, Xbp1 and Tos8 implying putative TFs regulating genes during stationary aging in yeast. Ste12 is the TF part of the mitogen-activated protein kinase (MAPK) pathway. It has also been reported that the filamentous growth MAPK pathway are activated under nutrient limitation (Cullen and Sprague, 2012). Xbp1 is a TF found induced in quiescent cells after glucose depletion (Miles, et al., 2013). We found that Xbp1 regulates genes involving β -oxidation indicating its role in controlling metabolic adaptation. Tos8 is a TF controlling cell cycle regulation (Horak, et al., 2002) and we found it was regulated by both Ste12 and Xbp1.

To summarize, in this study we applied different analysis methods on gene expression profiles of yeast at different days during starvation to obtain information in various viewpoints such as gene expression activities, transcriptional regulations, biological processes, affected metabolisms and structure of underlying components. The signaling pathways including Ras/PKA, TOR and Snf1, pathways were explained to responsible for sensing the level of nutrients (Galdieri, et al., 2010; Ring, et al., 2012). They together regulate genes involving in stress and metabolic responses through several TFs (e.g. Msn2, Msn4, Gis1 and Hsf1) which cause, for instance, aerobic utilization of fatty acid, ethanol and glycerol. Our results are in accordance with previous reports. Furthermore, Ste12 and Xbp1 were found as highly connected nodes in all subnetworks implicating that they are important transcription factors in the stationary phase model of ageing yeast. We also show in this work that incorporating information from several methods and from various data would provide a more thorough basis of the condition of interest.

Table 4 Summary of biological physiologies and associated gene sets

Description	Biological process	Transcription factor	Metabolic pathway
aerobic respiration	glyoxylate cycle; mitochondrial electron transport, succinate to ubiquinone; tricarboxylic acid cycle	Hap4	anaplerotic reactions; citrate cycle (TCA cycle); glutamate metabolism (aminosugars metabolism); pyruvate metabolism
cell growth	nucleotide biosynthetic process; translation; ribosome biogenesis	Ifh1	purine metabolism; pyrimidine metabolism; tRNA synthesis
cell wall biogenesis	fungal-type cell wall organization	Rlm1; Swi4	
mitochondrial dysfunction	autophagy; CVT pathway; mitochondrion degradation	Rtg1; Rtg2; Rtg3	
protein folding	protein folding in endoplasmic reticulum; protein targeting to ER	Hac1	
Starvation	acetate biosynthetic process; autophagy; cellular aldehyde metabolic process; CVT pathway; ethanol catabolic process; fatty acid beta-oxidation; glycogen biosynthetic process; glycerol metabolic process; negative regulation of gluconeogenesis	Adr1; Cat8; Gis1; Gln3; Mig1; Oaf1; Pip2; Tec1; Xbp1	fatty acid degradation; glycerol metabolism (glycerolipid metabolism); glycogen metabolism; trehalose metabolism
stress response	response to heat; response to stress; trehalose biosynthetic process	Cin5; Gis1; Haa1; Hot1; Hsf1; Msn2; Msn4; Pdr1; Skn7; Sko1; Sok2; Xbp1; Yap6	trehalose metabolism
synthesis of cell membrane components	ergosterol biosynthetic process; phosphatidylethanolamine biosynthetic process	Ino2	glycosylphosphatidylinositol (GPI) biosynthesis; sterol biosynthesis



Figure 16 Consensus network of all identified active subnetworks. Red nodes are up-regulated and green nodes are down-regulated related to Log. Color intensity follows the fold change at D10. Significantly expressed nodes are with a cyan border. Gray nodes represent metabolic pathways and nodes with white label are regulated by both Ste12 and Xbp1. The pink edge denotes the binding of a TF to its target gene. A green edge is a protein-protein interaction and a dash blue line links a metabolic gene to the metabolic pathway.

4. Conclusions and perspectives

The ultimate goal of bioinformatics is to facilitate data management and analysis to obtain biological insights from the biological data (Moore, 2007). In this thesis, the key roles of bioinformatics in database design, database development and data analysis were addressed. In particular the main works were conducted for *S. cerevisiae* since a comprehensive set of yeast genome data is greatly available.

Different yeast databases were designed and developed for specific purposes. Diverse technologies ranging from a typical-relational database to NoSQL databases were applied to these databases based on data characteristics and the objectives of the databases.

MySQL relational database was applied in yApoptosis since it has a well-standardized query language and relatively easy to use which is suitable for databases without complicated and heterogeneous data. In addition, the defined database schema of yApoptosis can further be applied to implement databases of other PCDs such as autophagy and necrosis. On the other hand, NoSQL databases have arisen as next generation databases that allow the development of databases for complex and large datasets (e.g. sequencing data). yStreX was thus managed under MongoDB, a document-oriented database, because its schemaless feature allows conveniently scaling out the database when more gene expression data are included in the future. It not only represents a next generation database in life science research, but it also contains pre-analyzed gene expression data that can extensively be applied. Moreover, the development of a database system was performed to manage and integrate multi-level data from different sources into a single database contributing to extensive applications. This system applied an object-oriented concept to represent biological components and it was then used to build a yeast data repository representing a centralized database or a single access point to answer various questions.

Because biological data are mounting in size and complexity, database design and development are an active area of bioinformatics. It is important to both understand the data to be stored and purposes of the database so that programming logic and biological focus can be harmonized to produce successful databases.

In terms of data analysis, we applied various bioinformatics approaches to analyze data in different contexts, for instance gene expression analyses of transcriptome data from stress-induced conditions and from chronologically ageing yeast to observe changes in transcriptional regulations, biological processes and metabolic pathways in particular conditions, and the network-based analysis for module discoveries. Biological information obtained ranges from the list of key genes, to networks of interacting components and to significant pathways depended

upon level of integrated datasets. It can be seen from the Paper IV that incorporating more kind of information provides a more thorough picture of the condition of study.

In conclusion, diverse genome-scale data have been generated shifting life science research towards integrative methodologies to examine cellular processes on a systems level over the past several years. The constant improvement and development of effective algorithms, databases and technologies for large-scale and multi-dimensional data is to reduce the gap between data generation and computing capability. It is still challenging and bioinformatics is a research area that immensely ongoing with the ideal goal to fully understand biological systems.

Acknowledgements

First of all I would like to thank my supervisor Dina Petranovic for all supports and valuable guidance. I would like to thank my co-supervisor Jens Nielsen for suggestion and comment. I would also like to thank another co-supervisor Intawat Nookeaw for ceaseless contribution, inspiration and helpful discussion about bioinformatics issues. There were several times that I felt subdued about my work but all of you were always there for open discussion and answering questions. I am very grateful.

Apart from my supervisors, several people were involved in my work. Marija and Andrea had initiated the idea of having the apoptosis database and contributed to this work until it published, Thanks a lot. Natapol and Avlant were important teammates, Thanks so much for contributing such brilliant work. Thanks go to Arastoo and Nutvadee for contributing to a valuable part of this work. Thanks also go to Emma, Suwanee, Sakda and Ximena for helping me in the lab.

Thanks people in apoptosis group, Ana, Eugenio, Xin, Magnus, Arastoo, Jin, Laleh, Raya, Kumar, Celine and Matthew for having nice discussion, being great friends and having nice dinners together. Thanks computational guys, Rasmus, Luis, Tobias, Natapol, Amir x 2, Leif, Antonio M., Fredrik, Shaq, Francesco, Manuel, Subazini, Pouyan, Kaisa, Saeed and Adil for pleasant discussion about modeling and bioinformatics, and for creating joyful working environment.

I would be in a lot of troubles without all of you, Erica, Helena and Martina. Many thanks for helping me with practical and administrative things.

Thanks the Thai community, P'In, P'A⁺, P'Bon, P'Ple, Wanwipa, Balloon and Mote for having so much fun together, joining adventure trips, cooking nice Thai food and being both friends and a big family for me. A lot of thanks to Christoph, Bouke, Nina, Martin, all members of sysbio group and people I have not mentioned. It was such a great and unforgettable time for me to know all of you, work with you, chat with you and laugh with you. Many thanks go to my good old friends, Kate & Tuk. You guys always stay beside me in whatever situations I have.

I would like to give the biggest thanks and hugs to my beloved family, papa, mama, bros and grandma for your true love and endless support in every step of my life. Living far away was not easy for us. Many thanks for your waiting and your belief in me.

And lastly, thanks those who passed by for making me stronger.

References

- Ahn, S.H., *et al.* (2005) Sterile 20 kinase phosphorylates histone H2B at serine 10 during hydrogen peroxide-induced apoptosis in *S. cerevisiae*, *Cell*, **120**, 25-36.
- Aittokallio, T. and Schwikowski, B. (2006) Graph-based methods for analysing networks in cell biology, *Brief Bioinform*, **7**, 243-255.
- Altman, R.B. (2012) Introduction to translational bioinformatics collection, *PLoS Comput Biol*, **8**, e1002796.
- Apweiler, R., *et al.* (2004) UniProt: the Universal Protein knowledgebase, *Nucleic acids research*, **32**, D115-119.
- Aranda, B., *et al.* (2011) PSICQUIC and PSIScore: accessing and scoring molecular interactions, *Nature methods*, **8**, 528-529.
- Arga, K.Y., *et al.* (2007) Understanding signaling in yeast: Insights from network analysis, *Biotechnol Bioeng*, **97**, 1246-1258.
- Ashburner, M., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.
- Audenaert, P., *et al.* (2011) CyClus3D: a Cytoscape plugin for clustering network motifs in integrated networks, *Bioinformatics*, **27**, 1587-1588.
- Bader, J.S., *et al.* (2004) Gaining confidence in high-throughput protein interaction networks, *Nat Biotechnol*, **22**, 78-85.
- Barnes, M.R. (2007) *Bioinformatics for Geneticists: A bioinformatics primer for the analysis of genetic data*. Wiley, Chichester.
- Barrett, T., *et al.* (2013) NCBI GEO: archive for functional genomics data sets--update, *Nucleic Acids Res*, **41**, D991-995.
- Beisser, D., *et al.* (2010) BioNet: an R-Package for the functional analysis of biological networks, *Bioinformatics*, **26**, 1129-1130.
- Benson, D.A., *et al.* (2013) GenBank, *Nucleic acids research*, **41**, D36-42.
- Berger, B., Peng, J. and Singh, M. (2013) Computational solutions for omics data, *Nat Rev Genet*, **14**, 333-346.
- Bindea, G., *et al.* (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks, *Bioinformatics*, **25**, 1091-1093.
- Blais, A. and Dynlacht, B.D. (2005) Constructing transcriptional regulatory networks, *Genes Dev*, **19**, 1499-1511.

Blanchard, F., *et al.* (2002) Targeted destruction of DNA replication protein Cdc6 by cell death pathways in mammals and yeast, *Mol Biol Cell*, **13**, 1536-1549.

Botstein, D., Chervitz, S.A. and Cherry, J.M. (1997) Genetics - Yeast as a model organism, *Science*, **277**, 1259-1260.

Botstein, D. and Fink, G.R. (2011) Yeast: An Experimental Organism for 21st Century Biology, *Genetics*, **189**, 695-704.

Braun, R.J. and Westermann, B. (2011) Mitochondrial dynamics in yeast cell death and aging, *Biochem Soc Trans*, **39**, 1520-1526.

Breitkreutz, A., *et al.* (2010) A global protein kinase and phosphatase interaction network in yeast, *Science*, **328**, 1043-1046.

Buttner, S., *et al.* (2007) Endonuclease G regulates budding yeast life and death, *Mol Cell*, **25**, 233-246.

Carmona-Gutierrez, D., *et al.* (2010) Apoptosis in yeast: triggers, pathways, subroutines, *Cell death and differentiation*, **17**, 763-773.

Causton, H.C., *et al.* (2001) Remodeling of yeast genome expression in response to environmental changes, *Mol Biol Cell*, **12**, 323-337.

Cebollero, E. and Reggiori, F. (2009) Regulation of autophagy in yeast *Saccharomyces cerevisiae*, *Biochimica et biophysica acta*, **1793**, 1413-1421.

Cebulski, J., *et al.* (2011) Yeast Bax inhibitor, Bxi1p, is an ER-localized protein that links the unfolded protein response and programmed cell death in *Saccharomyces cerevisiae*, *PLoS One*, **6**, e20882.

Cerami, E.G., *et al.* (2006) cPath: open source software for collecting, storing, and querying biological pathways, *BMC Bioinformatics*, **7**, 497.

Cerami, E.G., *et al.* (2011) Pathway Commons, a web resource for biological pathway data, *Nucleic acids research*, **39**, D685-690.

Cheng, C., *et al.* (2007) Inference of transcription modification in long-live yeast strains from their expression profiles, *BMC Genomics*, **8**, 219.

Cheng, W.C., Leach, K.M. and Hardwick, J.M. (2008) Mitochondrial death pathways in yeast and mammalian cells, *Bba-Mol Cell Res*, **1783**, 1272-1279.

Cherry, J.M., *et al.* (2012) *Saccharomyces* Genome Database: the genomics resource of budding yeast, *Nucleic acids research*, **40**, D700-705.

Chodorow, K. and Dirolf, M. (2010) *MongoDB: The Definitive Guide*. O'Reilly, Sebastopol.

Choi, J.K., *et al.* (2003) Combining multiple microarray studies and modeling interstudy variation, *Bioinformatics*, **19 Suppl 1**, i84-90.

- Cooray, M.P.N.S. (2012) Molecular biological databases: evolutionary history, data modeling, implementation and ethical background, *Sri Lanka Journal of Bio-Medical Informatics*, **3**, 11.
- Couchbase (2013) Making the Shift from Relational to NoSQL. Couchbase, pp. 5.
- Cullen, P.J. and Sprague, G.F., Jr. (2012) The regulation of filamentous growth in yeast, *Genetics*, **190**, 23-49.
- Demir, E., *et al.* (2010) The BioPAX community standard for pathway data sharing, *Nat Biotechnol*, **28**, 935-942.
- Dittrich, M.T., *et al.* (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach, *Bioinformatics*, **24**, i223-231.
- Eisenberg, T., *et al.* (2007) The mitochondrial pathway in yeast apoptosis, *Apoptosis*, **12**, 1011-1023.
- Eisenberg, T., *et al.* (2010) Necrosis in yeast, *Apoptosis*, **15**, 257-268.
- Fabrizio, P. and Longo, V.D. (2003) The chronological life span of *Saccharomyces cerevisiae*, *Aging Cell*, **2**, 73-81.
- Fahrenkrog, B., Sauder, U. and Aebi, U. (2004) The *S-cerevisiae* HtrA-like protein Nma111p is a nuclear serine protease that mediates yeast apoptosis, *J Cell Sci*, **117**, 115-126.
- Fannjiang, Y., *et al.* (2004) Mitochondrial fission proteins regulate programmed cell death in yeast, *Genes & development*, **18**, 2785-2797.
- Farrugia, G. and Balzan, R. (2012) Oxidative stress and programmed cell death in yeast, *Front Oncol*, **2**, 64.
- Feichtinger, J., *et al.* (2012) Microarray Meta-Analysis: From Data to Expression to Biological Relationships. In Trajanoski, Z. (ed), *Computational Medicine*. Springer, pp. 59-77.
- Fernandez-Suarez, X.M., Rigden, D.J. and Galperin, M.Y. (2014) The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection, *Nucleic acids research*, **42**, D1-6.
- Fields, S. and Johnston, M. (2005) Cell biology. Whither model organism research?, *Science*, **307**, 1885-1886.
- Flicek, P., *et al.* (2013) Ensembl 2013, *Nucleic acids research*, **41**, D48-55.
- Franceschini, A., *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration, *Nucleic acids research*, **41**, D808-815.
- Frohlich, K.U., Fussi, H. and Ruckenstein, C. (2007) Yeast apoptosis - From genes to pathways, *Semin Cancer Biol*, **17**, 112-121.
- Galdieri, L., *et al.* (2010) Transcriptional regulation in yeast during diauxic shift and stationary phase, *OMICS*, **14**, 629-638.

Gasch, A.P., *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes, *Mol Biol Cell*, **11**, 4241-4257.

Gasch, A.P. and Werner-Washburne, M. (2002) The genomics of yeast responses to environmental stress and starvation, *Funct Integr Genomics*, **2**, 181-192.

Gautier, L., *et al.* (2004) affy--analysis of Affymetrix GeneChip data at the probe level, *Bioinformatics*, **20**, 307-315.

Goffeau, A., *et al.* (1996) Life with 6000 genes, *Science*, **274**, 546, 563-547.

Hahn, J.S. and Thiele, D.J. (2004) Activation of the *Saccharomyces cerevisiae* heat shock transcription factor under glucose starvation conditions by Snf1 protein kinase, *Journal of Biological Chemistry*, **279**, 5169-5176.

Harbison, C.T., *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome, *Nature*, **431**, 99-104.

Horak, C.E., *et al.* (2002) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*, *Genes Dev*, **16**, 3017-3033.

Hou, J., *et al.* (2012) Metabolic engineering of recombinant protein secretion by *Saccharomyces cerevisiae*, *FEMS Yeast Res*, **12**, 491-510.

Hubbell, E., Liu, W.M. and Mei, R. (2005) Guide to Probe Logarithmic Intensity Error (PLIER) Estimation. http://media.affymetrix.com/support/technical/technotes/plier_technote.pdf.

Hucka, M., *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models, *Bioinformatics*, **19**, 524-531.

Hunter, S., *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database, *Nucleic acids research*, **40**, D306-312.

Ideker, T., Galitski, T. and Hood, L. (2001) A new approach to decoding life: systems biology, *Annu Rev Genomics Hum Genet*, **2**, 343-372.

Kapushesky, M., *et al.* (2010) Gene expression atlas at the European bioinformatics institute, *Nucleic Acids Res*, **38**, D690-698.

Kasprzyk, A. (2011) BioMart: driving a paradigm change in biological data management, *Database : the journal of biological databases and curation*, **2011**, bar049.

Kerrien, S., *et al.* (2007) Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions, *BMC Biol*, **5**, 44.

Khoury, C.M., *et al.* (2008) A TSC2-like motif defines a novel antiapoptotic protein family, *FEMS Yeast Res*, **8**, 540-563.

- Knijnenburg, T.A., *et al.* (2009) Combinatorial effects of environmental parameters on transcriptional regulation in *Saccharomyces cerevisiae*: a quantitative analysis of a compendium of chemostat-based transcriptome data, *BMC Genomics*, **10**, 53.
- Koh, G.C., *et al.* (2012) Analyzing protein-protein interaction networks, *J Proteome Res*, **11**, 2014-2031.
- Kohl, P., *et al.* (2010) Systems biology: an approach, *Clin Pharmacol Ther*, **88**, 25-33.
- Kwoh, C.K. and Ng, P.Y. (2007) Network analysis approach for biology, *Cell Mol Life Sci*, **64**, 1739-1751.
- Lee, T.J., *et al.* (2006) BioWarehouse: a bioinformatics database warehouse toolkit, *BMC Bioinformatics*, **7**, 170.
- Li, W., *et al.* (2006) Yeast AMID homologue Ndi1p displays respiration-restricted apoptotic activity and is involved in chronological aging, *Mol Biol Cell*, **17**, 1802-1811.
- Liang, Q. and Zhou, B. (2007) Copper and manganese induce yeast apoptosis via different pathways, *Mol Biol Cell*, **18**, 4741-4749.
- Longo, V.D. and Fabrizio, P. (2012) Chronological Aging in *Saccharomyces cerevisiae*, *Subcell Biochem*, **57**, 101-121.
- Ludovico, P., Madeo, F. and Silva, M. (2005) Yeast programmed cell death: an intricate puzzle, *IUBMB Life*, **57**, 129-135.
- Ludovico, P., *et al.* (2002) Cytochrome c release and mitochondria involvement in programmed cell death induced by acetic acid in *Saccharomyces cerevisiae*, *Mol Biol Cell*, **13**, 2598-2606.
- Luscombe, N.M., Greenbaum, D. and Gerstein, M. (2001) What is bioinformatics? A proposed definition and overview of the field, *Method Inform Med*, **40**, 346-358.
- Madeo, F., *et al.* (2009) Caspase-dependent and caspase-independent cell death pathways in yeast, *Biochem Bioph Res Co*, **382**, 227-231.
- Madeo, F., Frohlich, E. and Frohlich, K.U. (1997) A yeast mutant showing diagnostic markers of early and late apoptosis, *J Cell Biol*, **139**, 729-734.
- Madeo, F., Tavernarakis, N. and Kroemer, G. (2010) Can autophagy promote longevity?, *Nat Cell Biol*, **12**, 842-846.
- Magrane, M. and Consortium, U. (2011) UniProt Knowledgebase: a hub of integrated protein data, *Database (Oxford)*, **2011**, bar009.
- Marcus, F.B. (2008) *Bioinformatics and systems biology: collaborative research and resources*. Springer, Berlin, London.
- Marot, G., *et al.* (2009) Moderated effect size and P-value combinations for microarray meta-analyses, *Bioinformatics*, **25**, 2692-2699.

Mazzoni, C. and Falcone, C. (2008) Caspase-dependent apoptosis in yeast, *Biochimica et biophysica acta*, **1783**, 1320-1327.

Mazzoni, C. and Falcone, C. (2008) Caspase-dependent apoptosis in yeast, *Biochim Biophys Acta*, **1783**, 1320-1327.

Miles, S., *et al.* (2013) Xbp1 directs global repression of budding yeast transcription during the transition to quiescence and is important for the longevity and reversibility of the quiescent state, *PLoS Genet*, **9**, e1003854.

Moore, J.H. (2007) Bioinformatics, *J Cell Physiol*, **213**, 365-369.

Munoz, A.J., *et al.* (2012) Systems biology of yeast cell death, *FEMS Yeast Res*, **12**, 249-265.

Narasimhan, M.L., *et al.* (2005) Osmotin is a homolog of mammalian adiponectin and controls apoptosis in yeast through a homolog of mammalian adiponectin receptor (vol 17, pg 171, 2005), *Mol Cell*, **17**, 611-611.

Nepusz, T., Yu, H. and Paccanaro, A. (2012) Detecting overlapping protein complexes in protein-protein interaction networks, *Nat Methods*, **9**, 471-472.

Okayama, T., *et al.* (1998) Formal design and implementation of an improved DDBJ DNA database with a new schema and object-oriented library, *Bioinformatics*, **14**, 472-478.

Oliveira, A.P., Patil, K.R. and Nielsen, J. (2008) Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks, *BMC Syst Biol*, **2**, 17.

Oliver, S.G. (2002) Functional genomics: lessons from yeast, *Philos Trans R Soc Lond B Biol Sci*, **357**, 17-23.

Orzechowski Westholm, J., *et al.* (2012) Gis1 and Rph1 regulate glycerol and acetate metabolism in glucose depleted yeast cells, *PLoS One*, **7**, e31577.

Osterlund, T., *et al.* (2013) Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling, *BMC Syst Biol*, **7**, 36.

Ozsoyoglu, Z.M., Ozsoyoglu, G. and Nadeau, J. (2006) Genomic pathways database and biological data management, *Anim Genet*, **37 Suppl 1**, 41-47.

Palsson, B. (2006) *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press.

Pedruzzi, I., *et al.* (2000) *Saccharomyces cerevisiae* Ras/cAMP pathway controls post-diauxic shift element-dependent transcription through the zinc finger protein Gis1, *Embo J*, **19**, 2569-2579.

Pereira, C., *et al.* (2007) ADP/ATP carrier is required for mitochondrial permeabilization and cytochrome c release in yeast apoptosis, *Mol Microbiol*, **66**, 571-582.

- Perrone, G.G., Tan, S.X. and Dawes, I.W. (2008) Reactive oxygen species and yeast apoptosis, *Bba-Mol Cell Res*, **1783**, 1354-1368.
- Petranovic, D. and Nielsen, J. (2008) Can yeast systems biology contribute to the understanding of human disease?, *Trends Biotechnol*, **26**, 584-590.
- Pozniakovsky, A.I., *et al.* (2005) Role of mitochondria in the pheromone- and amiodarone-induced programmed death of yeast, *Journal of Cell Biology*, **168**, 257-269.
- Pu, S., *et al.* (2009) Up-to-date catalogues of yeast protein complexes, *Nucleic acids research*, **37**, 825-831.
- Qiu, J.Z., Yoon, J.H. and Shen, B.H. (2005) Search for apoptotic nucleases in yeast - Role of Tat-D nuclease in apoptotic DNA degradation, *J Biol Chem*, **280**, 15370-15379.
- Ramasamy, A., *et al.* (2008) Key issues in conducting a meta-analysis of gene expression microarray datasets, *PLoS Med*, **5**, e184.
- Reimers, M. and Carey, V.J. (2006) Bioconductor: an open source framework for bioinformatics and computational biology, *Methods Enzymol*, **411**, 119-134.
- Ring, J., *et al.* (2012) The metabolism beyond programmed cell death in yeast, *Exp Cell Res*, **318**, 1193-1200.
- Rinnerthaler, M., *et al.* (2012) Yno1p/Aim14p, a NADPH-oxidase ortholog, controls extramitochondrial reactive oxygen species generation, apoptosis, and actin cable formation in yeast, *Proc Natl Acad Sci U S A*, **109**, 8658-8663.
- Rockenfeller, P. and Madeo, F. (2008) Apoptotic death of ageing yeast, *Exp Gerontol*, **43**, 876-881.
- Rubinsztein, D.C., Marino, G. and Kroemer, G. (2011) Autophagy and Aging, *Cell*, **146**, 682-695.
- Rustici, G., *et al.* (2013) ArrayExpress update--trends in database growth and links to data analysis tools, *Nucleic Acids Res*, **41**, D987-990.
- Santt, O., *et al.* (2008) The yeast GID complex, a novel ubiquitin ligase (E3) involved in the regulation of carbohydrate metabolism, *Mol Biol Cell*, **19**, 3323-3333.
- Schneiter, R. (2004) Genetics, Molecular and Cell Biology of Yeast. University of Fribourg, Switzerland.
- Shannon, P., *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res*, **13**, 2498-2504.
- Sharon, A., *et al.* (2009) Fungal apoptosis: function, genes and gene function, *Fems Microbiol Rev*, **33**, 833-854.

Smyth, G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments, *Stat Appl Genet Mol Biol*, **3**, Article3.

Smyth, G.K. (2005) Limma: linear models for microarray data. In Gentleman, R.C., V.; Dudoit, S.; Irizarry, R.; Huber, W. (ed), *In: Bioinformatics and Computational Biology Solutions using R and Bioconductor*, R. Springer, New York, pp. 397-420.

Spirin, V. and Mirny, L.A. (2003) Protein complexes and functional modules in molecular networks, *Proc Natl Acad Sci U S A*, **100**, 12123-12128.

Sreenivasaiah, P.K. and Kim do, H. (2010) Current trends and new challenges of databases and web applications for systems driven biological research, *Front Physiol*, **1**, 147.

Stark, C., *et al.* (2011) The BioGRID Interaction Database: 2011 update, *Nucleic Acids Res*, **39**, D698-704.

Stein, L. (2013) Creating databases for biological information: an introduction, *Curr Protoc Bioinformatics*, **Chapter 9**, Unit9 1.

Stromback, L. and Lambrix, P. (2005) Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX, *Bioinformatics*, **21**, 4401-4407.

Teixeira, M.C., *et al.* (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*, *Nucleic acids research*, **34**, D446-451.

Teixeira, M.C., *et al.* (2013) The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in *Saccharomyces cerevisiae*, *Nucleic acids research*.

Thompson, D.M. and Parker, R. (2009) The RNase Rny1p cleaves tRNAs and promotes cell death during oxidative stress in *Saccharomyces cerevisiae*, *Journal of Cell Biology*, **185**, 43-50.

Varemo, L., Nielsen, J. and Nookaew, I. (2013) Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods, *Nucleic Acids Res*, **41**, 4378-4391.

Vernis, L., *et al.* (2009) A newly identified essential complex, Dre2-Tah18, controls mitochondria integrity and cell death after oxidative stress in yeast, *PLoS One*, **4**, e4376.

Walter, D., *et al.* (2006) The inhibitor-of-apoptosis protein Bir1p protects against apoptosis in *S. cerevisiae* and is a substrate for the yeast homologue of Omi/HtrA2, *J Cell Sci*, **119**, 1843-1851.

Wei, M., *et al.* (2008) Life span extension by calorie restriction depends on Rim15 and transcription factors downstream of Ras/PKA, Tor, and Sch9, *PLoS Genet*, **4**, e13.

Wei, M., *et al.* (2009) Tor1/Sch9-regulated carbon source substitution is as effective as calorie restriction in life span extension, *PLoS Genet*, **5**, e1000467.

- Winderickx, J., *et al.* (2008) Protein folding diseases and neurodegeneration: lessons learned from yeast, *Biochimica et biophysica acta*, **1783**, 1381-1395.
- Wissing, S., *et al.* (2004) An AIF orthologue regulates apoptosis in yeast, *The Journal of cell biology*, **166**, 969-974.
- Yang, H., Ren, Q. and Zhang, Z. (2008) Cleavage of Mcd1 by caspase-like protease Esp1 promotes apoptosis in budding yeast, *Mol Biol Cell*, **19**, 2127-2134.
- Zappavigna, S., *et al.* (2013) Autophagic cell death: A new frontier in cancer research, *Advances in Bioscience and Biotechnology*, **4**, 13.
- Zhang, L.V., *et al.* (2005) Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network, *J Biol*, **4**, 6.
- Zhu, X., Gerstein, M. and Snyder, M. (2007) Getting connected: analysis and principles of biological networks, *Genes Dev*, **21**, 1010-1024.