UNIVERSALITY, SPECIFICITY AND REGULATION OF S. *cerevisiae* GROWTH RATE RESPONSE IN DIFFERENT CARBON SOURCES AND NUTRIENT LIMITATIONS

NIKOLAI GEORGIEV SLAVOV

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
MOLECULAR BIOLOGY

ADVISOR: DAVID BOTSTEIN

JUNE 2010

Abstract

Many studies have used chemostats and gene expression microarrays to characterize the growth rate response of the buddying yeast (*Saccharomyces cerevisiae*) growing on glucose carbon source (Hayes *et al*, 2002; Pir *et al*, 2006; Regenberg *et al*, 2006; Castrillo *et al*, 2007; Brauer *et al*, 2008). These studies demonstrated a common growth rate response (GRR) in continuous exponentially growing cultures, both aerobic and anaerobic and limited by different natural nutrients as well as by auxotrophic requirements. However, in all studies the carbon source was glucose, which is highly preferred by *S. cerevisiae* and special in many ways (Zaman *et al*, 2008). Thus, it is not clear how much of the identified GRR is specific to growth on glucose (Zaman *et al*, 2009; Futcher, 2006) as a sole carbon source and how much of the GRR is general to growth and independent of the carbon source. In fact, Zaman *et al* (2009) have suggested that much of the observed common growth rate response can be due to glucose. To explore whether the common growth rate response is still going to be present in cultures grown on non–fermentable carbon source, I grew *S. cerevisiae* continuous cultures on ethanol carbon source and measured physiological parameters, gene expression, and metabolites.

I found that the growth rate response of a large number of genes (about 1500) remains very similar on ethanol carbon source and I call this common growth rate response *universal growth rate response.* Genes with positive universal growth rate response include ribosomal and translation genes. Genes with negative universal growth rate response include autophagy, vacuolar and stress response genes. Remarkably, all genes having universal growth rate response are expressed periodically in the yeast metabolic cycle (*YMC*) (Slavov and Botstein, 2011; Slavov *et al*, 2011). Genes whose expression levels increase with growth rate are expressed in *YMC* phase with high oxygen consumption while genes whose expression levels decrease with growth rate are expressed in *YMC* phase with low oxygen consumption. To understand better the relationship between the

YMC and the growth rate response, I synchronized metabolically continuous cultures and quantified the relationship between the growth rate and the periods of *YMC* phases. The relative duration of the *YMC* phase with high oxygen consumption increases with growth rate, which can account quantitatively for the observed universal growth rate response. Furthermore, I measured a linear dependence between the periods of the *YMC* and the cell cycle, which suggests a switch from *YMC* to fermentation at growth rates too high for the *YMC* to ensure reductive period that is long enough for DNA replication.

In contrast to the universal growth rate response, the growth rate response of many other genes is carbon source and/or limitation specific. Some of the carbon source specific growth rate response genes are expected (such as the stronger induction of mitochondrial and ethanol utilization genes in ethanol carbon source compared to glucose) while other carbon source specific growth rate response genes are more surprising, such as genes related to generation of precursor metabolites and energy, microtubules and the cell–cycle. To characterize the underlying regulatory mechanisms behind the observed growth rate response, I identified transcription factors (TFs) likely to mediate the growth rate response and inferred their activities in different nutrients and growth rates using $\mathcal{RC}$*web* (Slavov, 2010). Based on the gene expression data, I inferred that some TFs have carbon source dependent activities (*GCN4, HAP4, FHL1, YAP5*) and even more TFs have growth rate dependent activities, including *RAP1, GAT3, CBF1, MET4, INO4, HAP4*. Interestingly, for most TFs the change in activity is not reflected in the level of the corresponding mRNA. In ethanol carbon source, I found very strong induction, positive growth rate response and differential usage of isoenzymes in pathways (such as gluconeogenesis, TCA, and ethanol utilization) whose metabolic fluxes are expected to increase with growth rate and to be higher in ethanol compared to glucose carbon source. These findings suggest that transcription likely plays a role in regulating those metabolic pathways, but not in regulating the activities of TFs.

To identify growth rate response differences between auxotrophs and prototrophs, I grew *his* and *lys* auxotrophs limited on their auxotrophic requirements at different growth rates (Slavov and Botstein, 2013). The gene expression data from these experiments indicate significantly weaker induction of autophagy genes in slowly growing auxotrophic cultures compared to prototrophic cultures growing at the same growth rate. From the growth rate experiments with *his* and *lys* auxotrophs as well as from batch experiments I discovered very wide distribution of cell sizes (3-5 fold difference in cell volumes) in cultures of auxotrophs starving for their auxotrophic requirement. Both the failure to induce autophagy and the poor control of cell–size are likely to contribute to the lower viability of starving auxotrophs. Based on analysis used by Slavov and Dawson (2009), I identified the genes whose combinatorial regulation is most different between auxotrophic and prototrophic cultures. These genes are likely to mediate glucose wasting by auxotrophs and I experimentally demonstrated that single deletions for of *SFP1, CCC1* or *CCP1* have highly significant effects on glucose wasting.

Acknowledgments

After my fascination with chemistry and intense immersion in experimental laboratory work in junior–high and high–school, I had largely neglected experimental work in favor of more theoretical work that I found to be much more intellectually demanding and exciting. David Botstein revived for me a quote that I remember vividly from the wall of my high–school physics classroom: *"Everything we know about nature begins with an experiment and ends with an experiment"*. The revival was not merely in the direction of my scientific work but also in the direction of my conceptual thinking and scientific maturity. In this context, the work described in this thesis depended crucially on David not only in the usual, trivial sense but in a much more fundamental and substantive sense (Slavov, 2012).

I would also like to thank my thesis committee, Ned Wingreen and Josh Rabinowitz, for their guidance and suggestions in completing this work. I am grateful to Sandy Silverman and Viktor Boer for teaching me how to use chemostats and assisting me with many questions and problems that I had along the way of performing the experiments in this thesis. They and everybody in the Botstein lab (Charles Lu, David Gresham, Greg Lang, Ryan Briehof, Allegra Petti, Scott McIsaac, Patrick Gibney and Mark Hickman) helped make the 3 years I worked here enjoyable. Thinking of my days in CIL, I cannot miss mentioning Amy Caudy who brightened the lab with her cheerful attitude and helped with alacrity.

I had a wonderful experience working with Kenneth A. Dawson in Ireland. He, more than anybody else, taught me what it means to understand deeply something simple and the power that such deep understanding can have even with far more complex systems. I am very grateful for his time and the many insightful discussions through which he broadened my education and stimulated my personal growth.

The work in this thesis would not have been possible without the help and encouragements of my teachers starting from primary school all the way to graduate school. In particular, I am forever indebted to Rumiana Stefanova and Galina Kaneva (my chemistry and math teachers in high school) who encouraged me to follow my passion for science and participate in international Olympiads. The scientific education that I gained during my participation in those competitions was fundamental to all work in this thesis and the memories from my success were my principle source of support and confidence in overcoming difficulties along the way of completing the work described here.

Dedicated to my parents and grandparents

Contents

List of Figures

List of Tables

Preface

Many studies have exploited chemostats and gene expression microarrays to characterize the growth rate response in the buddying yeast (*Saccharomyces cerevisiae*) growing on glucose carbon source (Hayes *et al*, 2002; Pir *et al*, 2006; Regenberg *et al*, 2006; Castrillo *et al*, 2007; Brauer *et al*, 2008). These studies demonstrated a common growth rate response (GRR) in continuous exponentially growing cultures, both aerobic and anaerobic and limited by different natural nutrients as well as by auxotrophic requirements. However, in all cases the carbon source was glucose, which is highly preferred by *S. cerevisiae* and special in many ways (Zaman *et al*, 2008). Thus, it is not clear how much of the identified GRR is specific to growth on glucose (Zaman *et al*, 2009; Futcher, 2006) as a sole carbon source and how much of the GRR is general to growth and independent of the carbon source. In fact, Zaman *et al* (2009) have suggested that much of the observed common growth rate response can be due to glucose. To explore whether the common growth rate response is sitll going to be present in a culture grown on non–fermentable carbon source, I grew *S. cerevisiae* continuous cultures on ethanol carbon source and measured physiological parameters, gene expression, and metabolites. The experiments, some of the challenges and the data are presented and discussed in chapter 1.

The statistical analysis of the data from these and the experiments by Brauer *et al* (2008) is the subject of chapter 2. I discovered that the growth rate response of a large number of genes (about 1500) is very similar across all limitations, ethanol and glucose

carbon source and I call this common growth rate response *universal growth rate response.* In contrast, the growth rate response of many other genes is carbon source and/or limitation specific. Some of the carbon source specific genes are expected (such as mitochondrial genes for ethanol carbon source) while other carbon source source genes are more surprising, such as genes related to generation of precursor metabolites and energy, microtubles and cell–cycle. Once the genes with common and differential growth rate responses are identified, one might ask what are the underlying regulatory mechanisms behind the observed growth rate response. The analysis from the fourth section of chapter 2 is focused on using 2 very different methods for inferring the transcriptional regulation underlying the growth rate response.

The comparison of the methods for identifying transcriptional regulators motivates the need for a network inference algorithm. Such algorithm, $\mathcal{RC}$*web*, is derived, tested and applied to the data in the fourth chapter. There I describe an approach to inferring regulators that avoids inaccurate assumptions and arbitrary thresholds while preserving high statistical power and inferring systematically not only the regulators, but also their combinatorial interactions and activities. The use of this approach requires solving an *NP–hard* problem for which the existing algorithms do not perform very well. I derive a radically different algorithm $\mathcal{RC}$*web* that significantly outperforms (by several fold) even the best competitors in accuracy and by orders of magnitude in computational efficiency (Slavov, 2010). Then I apply $\mathcal{RC}$*web* to my data and to growth rate response data on glucose carbon source from Brauer *et al* (2008).

Remarkably, all genes having universal growth rate response are expressed periodically in the yeast metabolic cycle (*YMC*) (Tu *et al*, 2005). Based on this observation, experimental data on the *YMC* at different growth rates and recent work demonstrating the cell autonomous nature of the *YMC* (Silverman *et al*, 2010), I developed a model that explains quantitatively the observed growth rate response as a superposition of expression

levels of genes expressed during different phases of the *YMC*. The model requires that the durations of *YMC* phases change with growth rate, which is confirmed experimentally. Furthermore, I develop a general non–linear approach to using FISH data for studying the connection between the cell–cycle, growth rate and the *YMC*.

The final chapter summarizes briefly differences in the growth rate response of prototrophic cultures and auxotrophic (for *his, lys, leu* and *ura*) cultures. I identify a significant difference in the extent of autophagy induction during slow growth which is one of the main functional groups of genes with universal growth rate response. Furthermore, I use analysis developed by Slavov and Dawson (2009) to identify the genes whose combinatorial regulation is most different between auxotrophic and prototrophic cultures. Based on this analysis, I identify a set of genes likely to mediate glucose wasting by auxotrophs and experimentally demonstrate that deletions of those genes have highly significant effects on glucose wasting.

Chapter 1

Studying Growth Rate Experimentally

1.1 History & Background

As a doctoral student in the $1930s$, Jacques Monod explored bacterial growth on different sugars and described two distinct growth phases separated by a switch which he denoted by the term *"diauxie"* meaning two growth phases. An important result of this work was the phenomenological description of the growth rate as a function of the nutrient concentration by a simple monotonically increasing function (rectangular hyperbola) known to this date as Monod's growth law (Monod, 1942, 1949). This result indicated quantitatively that nutrient concentrations regulate the rate of microbial growth.

After Jacques Monod, prominent scientists that furthered understanding on bacterial growth growth include Kjeldgaard and Maaløe. Maaløe and Kjeldgaard used the Monod's growth law to grow bacteria at different growth rates depending on the richness of the nutrient media and documented changes in cell–size, protein, DNA and RNA composition that depend only of the growth rate and not on the particular media composition used for achieving the growth rate (Kjeldgaard *et al*, 1958; Maaløe, 1979). Kjeldgaard *et al* (1958)

interpreted their results to mean that protein production (ribosomes and translation) is growth rate limiting.

More recent studies on growth rate have used the chemostat, a device designated by Novick and Szilard (1950) (a contemporary of Ole Maaløe) to use the nutrient dependence of growth rate (as discovered Jacques Monod) for growing exponential cultures at any desired growth rate lower than the fastest growth rate that the microorganism can sustain. Furthermore, modern studies have used DNA microarray technology developed in the mid 1990 by Schena *et al* (1995); Lashkari *et al* (1997) for monitoring genome–wide gene expression patterns associated with growth rate.

Modern studies using chemostats and gene expression microarrays to characterize the growth rate of the buddying yeast (*Saccharomyces cerevisiae*) growing on glucose carbon source include (Hayes *et al*, 2002; Pir *et al*, 2006; Regenberg *et al*, 2006; Castrillo *et al*, 2007; Brauer *et al*, 2008). These studies demonstrated a common growth rate response (GRR) in continuous exponentially growing cultures, both aerobic and anaerobic and limited by different natural nutrients as well as by auxotrophic requirements. However, in all cases the carbon source was glucose which is highly preferred by *S. cerevisiae* and special in many ways (Zaman *et al*, 2008). In fact, Zaman *et al* (2009) have suggested that much of the observed common growth rate response can be due to glucose. Thus, it is not clear how much of the identified GRR is specific to growth on glucose Zaman *et al* (2009); Futcher (2006) as a sole carbon source and how much of the GRR is general to growth and independent of the carbon source. Many other fundamental questions about the origin, regulation and significance of the growth rate response remain open. What are the nutrient specific differences in the GRR and how significant are they? What is the origin of the growth rate response and how is it mediated in terms of cell–signaling cascades? To explore these question I grew *S. cerevisiae* continuous cultures on ethanol carbon source and measured physiological parameters, gene expression, and metabolites.

The next sections of this chapter describe the apparatus used for controlling the growth rate (a chemostat) and the following sections detail the experiments, the materials and methods used in the experiments and briefly describe the data.

1.2 Chemostat

A population of microorganisms growing in the wild (or in a test tube) is likely to be composed of cells growing at different growth rates and changing their growth rate in time as nutrients are being depleted or other physical parameters (such as temperature) change in time. Studying the physiological responses to changes in growth rate is thus greatly confounded by many factors that might be unobserved and changing in time. The chemostat (Novick and Szilard, 1950) is a convenient experimental apparatus allowing one to avoid such complications. In its essential design, a chemostat is a fermenter for continuous cultures with a constant influx of fresh nutrient media and equal efflux of reaction media containing cells, residual nutrients and secreted metabolites, Fig.1.1.

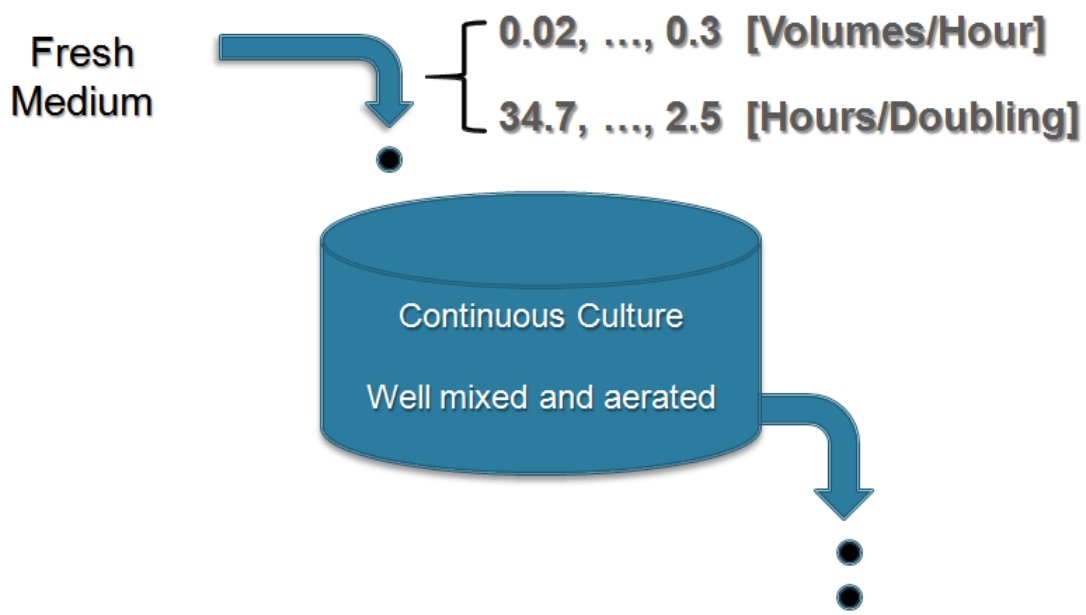

Figure 1.1: Schematic diagram of a chemostat (Novick and Szilard, 1950).

The dynamics of a chemostat cultures can be modeled by a very simple ordinary differential equation (1.1) with a single dependent variable N denoting the number of

cells and the time t being the independent variable.

$$\frac{dN}{dt} = \mu N - DN \tag{1.1}$$

Equation (1.1) describes the change in the number of cells N as a function of the growth rate of the population μ and the dilution data D. At steady–state $dN/dt = 0$ and thus the growth rate equals the dilution rate, $\mu = D$. This is an important result and equation (1.1) is the simplest way to derive it. However, equation (1.1) is rather limited in describing dynamics because the growth rate itself is a function of the growth conditions and the nutrient concentrations which are not included in equation (1.1). Equation (1.1) cannot even predict how the steady–state biomass density depends on the concentration of nutrients in the media.

These limitation can be overcome by only slightly more complicated model with two first–order ordinary differential equations (1.2–1.3) assuming that the growth rate dependence on the nutrient concentration can be described by a Monod equation (Monod, 1942, 1949).

$$\frac{dS}{dt} = D(S_o - S) - \frac{1}{\gamma} N \mu_{max} \frac{S}{S + K} \tag{1.2}$$

$$\frac{dN}{dt} = N \mu_{max} \frac{S}{S + K} - DN \tag{1.3}$$

The 2 dependent variables are the number of cells (N) and the amount of nutrient substrate (S) in the fermenter vessel. For simplicity and enhanced experimental control, it is convenient to provide all nutrients except for one in excess so that the growth of the culture depends only on the nutrient that is limited. Thus in Eq. (1.2–1.3), S is the amount of limiting nutrient. It is often more convenient to work with the concentration of the limiting nutrient and the biomass density (number of cells or cell volume per unit volume) rather than the absolute amounts. Making this change in Eq. (1.2–1.3) requires a simple scaler scaling by the volume and changes only the unites of N and S.

In the model above Eq. (1.2–1.3), the dynamics of N and S depend on four parameters:

- S_o – concentration of the limiting nutrient in the feed (fresh media)
- μ_{max} – The highest growth rate that the organisms can archive if all nutrients are abundant, e.g. above saturating nutrient concentrations.
- γ – A parameter quantifying the efficiency of converting the limiting nutrient into biomass. For example, grams of biomass generated from a gram of glucose.
- D – Dilution rate. The rate at which new media drips into the chemostat. By design, D is also the rate at which reaction media leaves the fermenter vessel.

At steady–state $dN/dt = 0$. Setting the derivative equal to zero in Eq. (1.3), results in:

$$\mu = \mu_{max} \frac{S}{S + K} = D \tag{1.4}$$

Thus, at steady–state the growth rate of the chemostat culture (μ) equals the dilution rate. Since the dilution rate can be set conveniently by the experimenter, this allows growing a culture at any desired growth rate that does not exceed μ_{max}. Throughout this thesis, I will used growth rate and dilution rate interchangeably for cultures at steady–state. Solving equations (1.2–1.3) for steady–state enables expressing the steady–state biomass as a function of the concentration of the limiting nutrient in the media which will be used in the following sections.

1.3 Experimental Design

One of the crucial factors in measuring the growth rate response accurately is using yeast strains and growth conditions that allow for a wide dynamical range of growth rates. Very slow growth rates are harder to establish and maintain accurately in a chemostat because small leaks of the pumps or obstructions of the tubing become significant relative

to the set dilution rate. Furthermore, the physiological state of the cells approaches asymptotically *G1/G0* arrest (quiescence) and measured changes become comparable to inevitable fluctuations associated with experimental measurements. As a result, at very slow growth rate the signal to noise ratio *(SNR)* decreases to rather unfavorable level. Thus to generate high quality data, it is highly desirable to extend the dynamical range of the growth rates by using strains and conditions allowing short doubling times. Initially, I started the experiments with *WT CEN.PK*. Its maximal growth rate on ethanol is rather slow and this not only fundamentally limited the dynamical range of my experiments but also resulted in rather low biomass densities even at $\mu = 0.08h^{-1}$. To overcome this problem, we looked for a strain that can grow faster on ethanol. Such a strain was isolated by Thomas Fox for its ability to grow well on ethanol (Fig.1.2) and cataloged as *DBY11369* in the strain collection of David Botstein. I used *DBY11369* for all growth rate

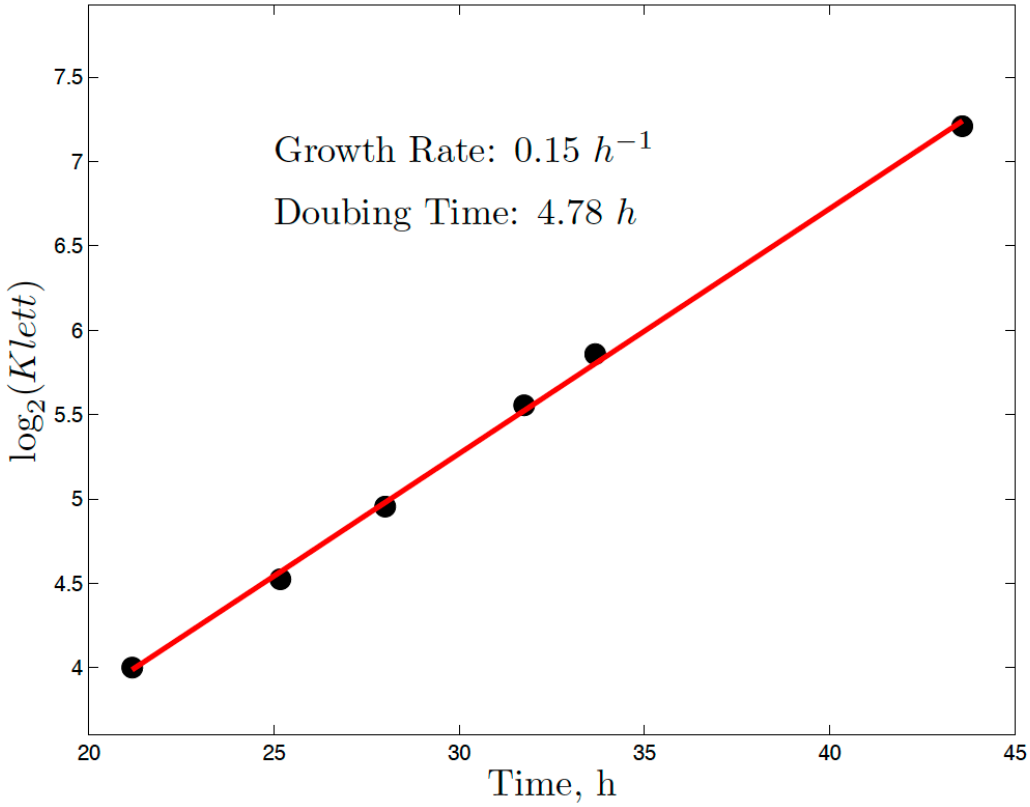

Figure 1.2: *DBY11369* was grown in $100mM$ ethanol media as a batch culture

experiments on ethanol carbon source described in this thesis. In excess of all nutrients (exponential phase in a batch culture) and at 30^oC, the highest growth rate I measured with *DBY11369* is $\mu = 0.15h^{-1}$ (Fig.1.2) which is higher than the growth rate I measured with any of the other strains I grew including *WT CEN.PK* and *WT S288C, HAP1*$^+$. In fact, after using *DBY11369* for the growth rate experiments on ethanol carbon source, I attempted growing *WT CEN.PK* in the conditions that worked great for *DBY11369*. Even at $\mu = 0.08h^{-1}$, the biomass density became so low that I had to lower the dilution rate to prevent the culture from washing out, once again demonstrating that my starting strain (*WT CEN.PK*) was not optimal for growth rate studies on ethanol carbon source. *S288C* with *HAP*1^+ has growth rate much closer to the growth rate of *DBY11369* and could have been a viable alternative.

A second important consideration in my experimental design was finding nutrient concentrations that are both limiting and result in optimal biomass densities at steady–state. As already mentioned, limiting the growth of the culture on a single nutrient simplifies the control of the experiment and the analysis of the data. In particular, solving Eq. (1.2–1.3) at steady–state ($dS/dt = dN/dt = 0$), indicates that the steady–state biomass depends linearly on S_0, the concentration of the limiting nutrient in the fresh media. To establish optimal concentrations for limiting nutrients, I grew *DBY11369* in batch in chemostat media with varying amounts of the limiting nutrient. Fig.1.3 shows the the final biomass (the biomass of the culture after the growth stops and the biomass density reaches an asymptotic value) as a function of the ethanol concentration: To quantify the biomass, I used OD_{600}, the absorbance (optical density) of the culture for light with wavelength $\lambda = 600nm$. Since I scanned a wide dynamical range of limiting nutrient concentrations, the biomass also varied widely, reaching high levels at which the optical density is not linearly proportional to the biomass density. To avoid such aberrations, I did serial dilutions ensuring that for all datapoints the measured optical

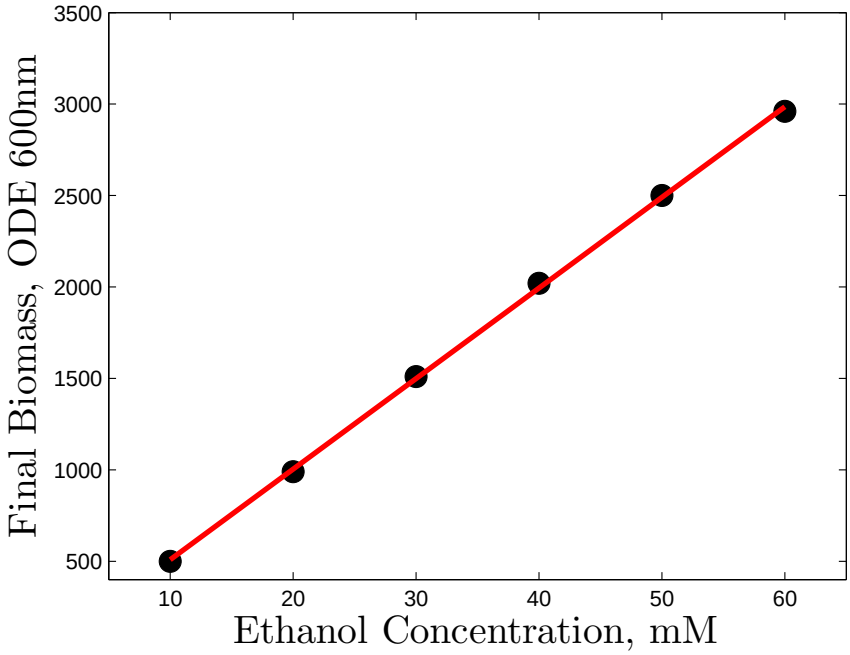

Figure 1.3: Dependence of the final biomass of *DBY11369* grown as a batch culture in chemostat media on the initial concentration of ethanol.

density is close to the optimum of the spectrophotometer. Fig.1.3 shows that the final biomass depends linearly on the concentration of ethanol over a wide dynamical range. For the chemostat experiments, I chose a concentration of $30mM$ ethanol since it gave optimal biomass and was well within the linear range. The fact that $30mM$ ethanol is limiting for growth in batch implies that it is likely to be liming for steady–state growth in chemostats but does not guarantee that. Therefore, I experimentally tested whether $30nm$ ethanol is indeed limiting at steady–state. I grew *DBY11369* at $\mu = 0.10h^{-1}$ to steady–state feeding from media containing $30mM$ ethanol and using a Coulter counter measured a cell density of 1.5×10^7 *cells/mL*. Then I switched to media with identical composition except for two times higher ethanol concentration ($60mM$). When the culture reached steady–state (again at $\mu = 0.10h^{-1}$), I measured culture density of 3×10^7 *cells/mL* indicating that $30mM$ ethanol is indeed limiting and in the linear response regime not only in batch but also at steady–state.

I used the same type of experiments to establish optimal limiting concentrations for PO_4^{3-} and NH_4^+, Fig.1.4. The optimal concentrations chosen for the chemostat experiments and verified to be limiting at steady–state are $[KH_2PO_4] = 20mg/L$ and $[(NH_4)_2SO_4] = 100mg/L$.

The determined amounts for the limiting nutrients were mixed with minimal defined (MD) media, metals, and vitamins. The only carbon source in all cases was ethanol $30mM$ for the ethanol limitation and $(100mM)$ for the nitrogen and phosphorus limitations. All media was added via sterile filtration to autoclaved chemostats as described previously (Saldanha *et al*, 2004; Brauer *et al*, 2005, 2008).

Chemostats were established in $500mL$ fermenter vessels (Sixfors; Infors AG, Bottmingen, Switzerland) containing $250mL$ of culture volume, stirred at $400rpm$, and aerated with humidified and filtered air. Chemostat cultures were inoculated, monitored and grown to steady state as described previously (Brauer *et al*, 2005). All cultures were monitored for changes in cell density and dissolved oxygen and grown until these values remained steady for at least $48h$.

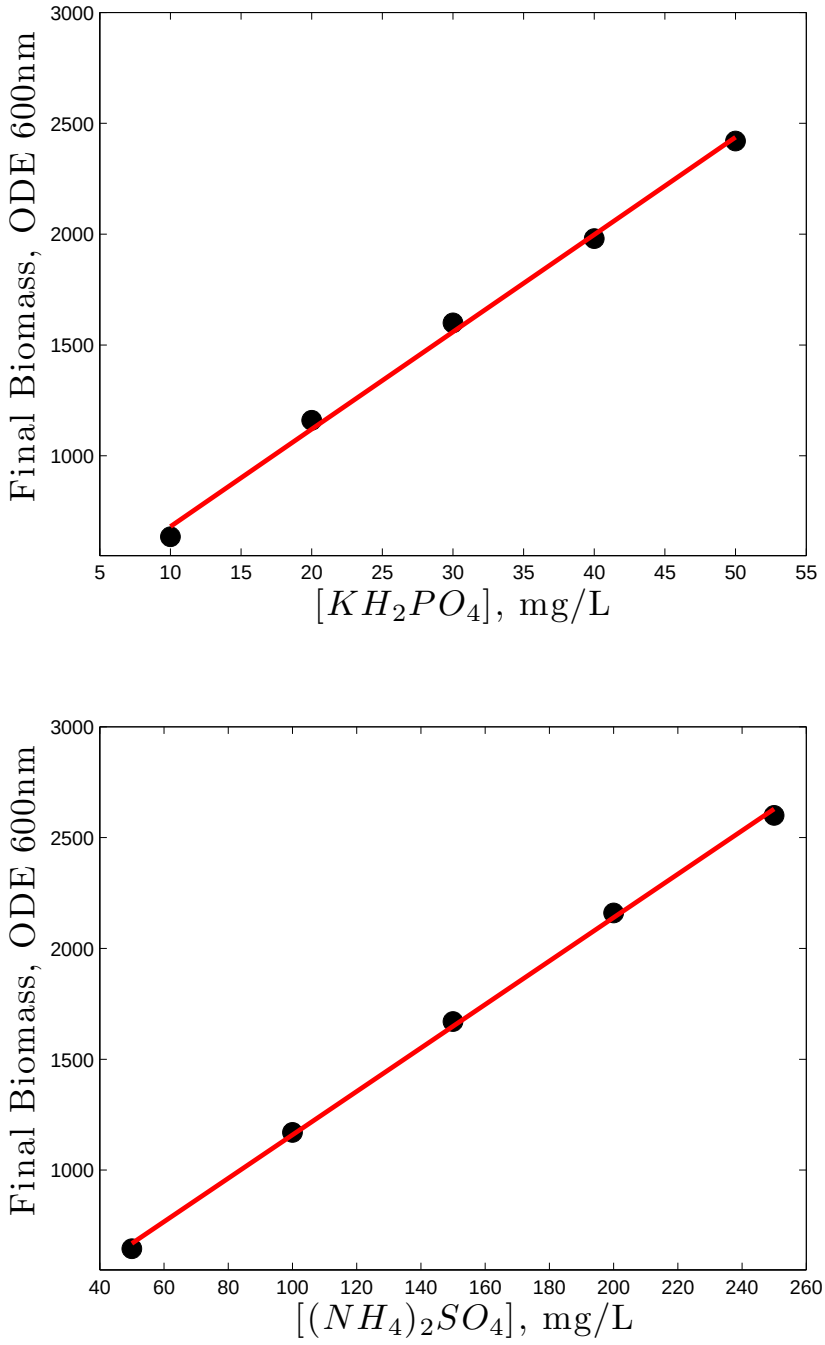

Figure 1.4: Dependence of the final biomass of *DBY11369* grown as a batch culture in chemostat media on the initial concentration of phosphate (PO_4^{3-}) and on the initial concentration of ammonium, (NH_4^+)

1.4 Physiological Growth Rate Response

Growing microbial cultures at steady–state in chemostats not only allows for precise control of many experimental parameters but also for precise measurements of physiological parameters that can be highly informative. This section is devoted to reporting the results of such measurements and discussing their biological significance.

1.4.1 Residual Ethanol

At steady–state, (Fig.1.5) the concentration of residual ethanol in the fermenter vessel for the culture limited on ethanol follows the trend expected from Eq.(1.2–1.3). While

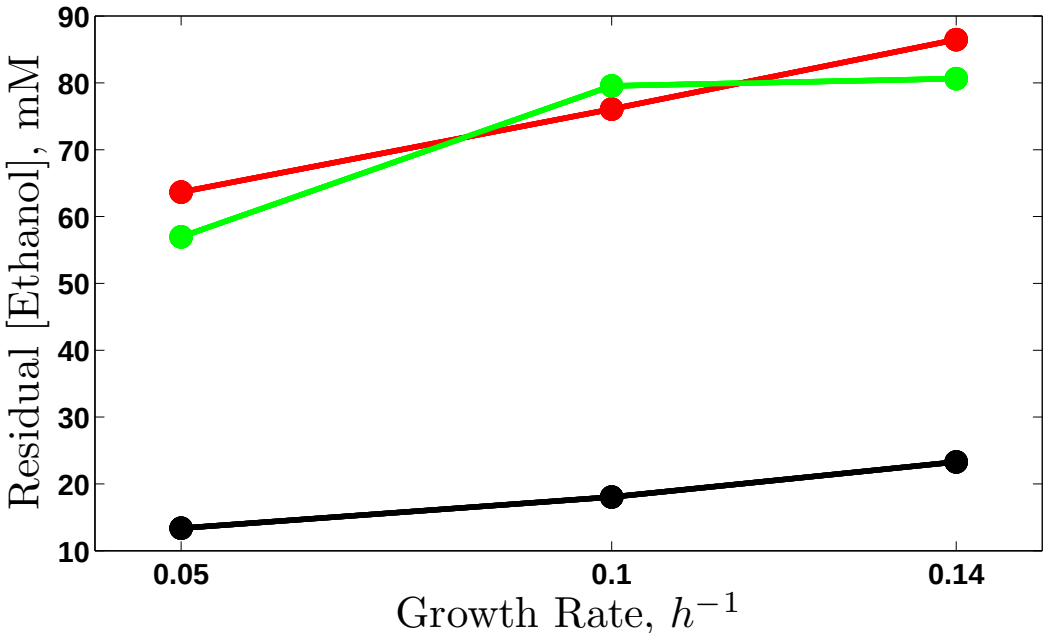

Figure 1.5: Residual ethanol concentration as a function of the growth rate across nutrient limitations of cultures growing on ethanol carbon source.

Eq.(1.2–1.3) are insufficient to describe the residual concentration of ethanol for the other limitations quantitatively, the qualitative trend can be predicated intuitively. As cells grow slower (low dilution rates), they spend more time in the reaction vessels and the

flux of media is slower. Both of these factors suggest that the concentration of residual ethanol should be inversely correlated to the growth rate of the cultures which is in fact what is experimentally observed, Fig.1.5. The specific consumption of ethanol (*ethanol consumed/steady–state biomass*) is lower for the ethanol limited culture suggesting that cultures not limited on ethanol might metabolize some fraction of the ethanol to acetate, possibly to generate reducing *NADPH* required for biosynthetic processes. This is consistent with acetate detected in the cultures limited on NH_4^+ and PO_4^{3-} but not in the cultures limited on ethanol. Problems with quantifying acetate limit the ability to make this argument more quantitative.

1.4.2 Bud Index

One one the advantages of budding yeast is that the cell morphology is indicative of the cell–cycle phase, (Hartwell, 1974; Hartwell *et al*, 1974). In particular, cells in *G0/G1* are not budded while cells in *S/G2/M* are budded. Buds, daughter cells, start emerging during *S* phase and grow until they separate as independent cells at the end of *M*. The fraction of budded cells (Fig.1.6) in steady–state cultures was quantified by counting a few hundred budded and non–budded cells in sonicated samples from steady–state chemostat cultures. The fraction of budded cells shows linear dependence (within experimental error) with

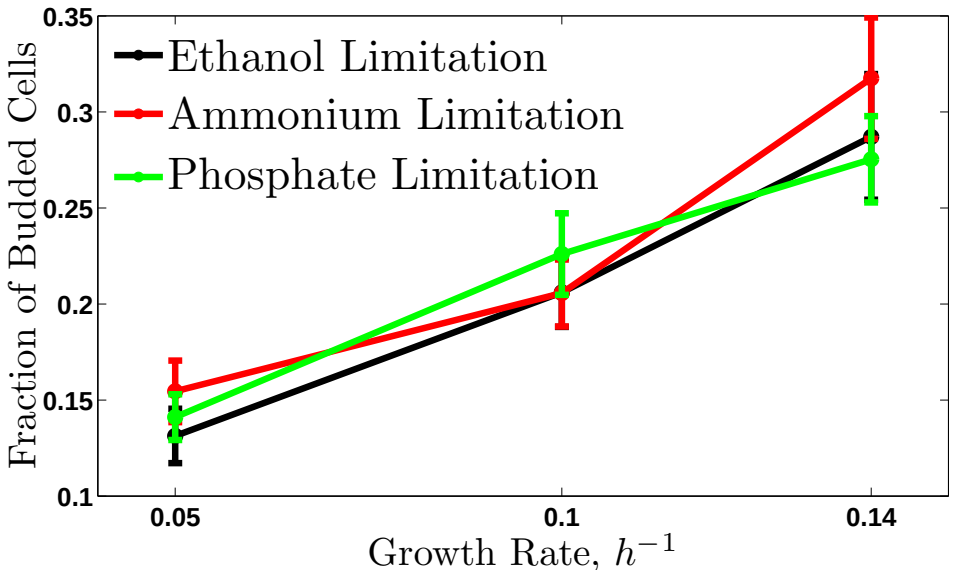

Figure 1.6: Fraction of budded cells as a function of the growth rate across nutrient limitation of cultures using ethanol carbon source.

respect to the growth rate. The budded fractions are not only qualitatively but also quantitatively the same as the ones observed on glucose carbon source for the corresponding growth rates (Brauer *et al*, 2008). These data indicate that the duration of *S/G2/M* phases of the cell–cycle does not change with growth rate and the duration of *G0/G1* is inversely proportional to the growth rate, (Hartwell, 1974; Brauer *et al*, 2008).

1.4.3 Biomass Density and Cell Size

The biomass density (Fig.1.7) also follows the trend expected from Eq.(1.2–1.3). The

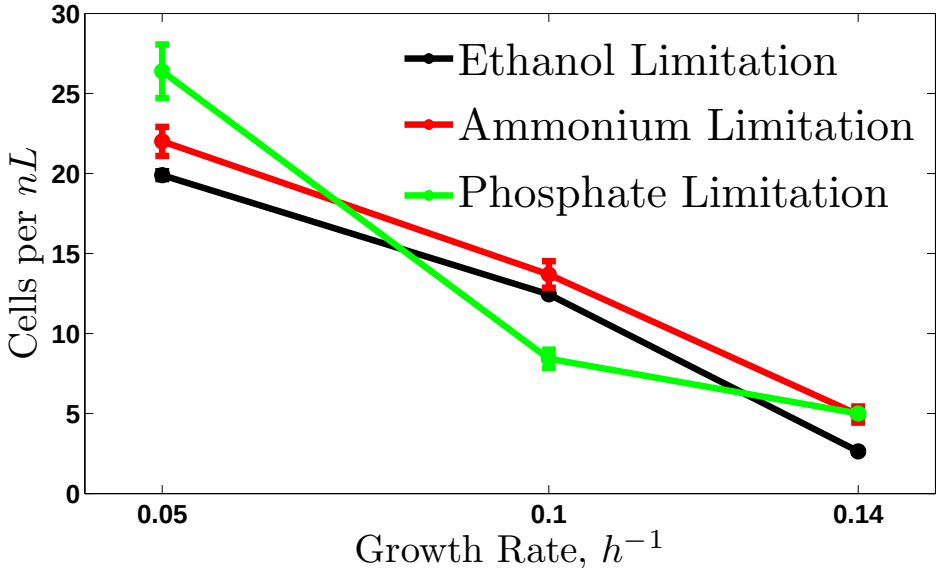

Figure 1.7: Cell density as a function of the growth rate across nutrient limitations of cultures using ethanol carbon source.

standard deviations (plotted as error bars) computed from 3 measurements spaced by 48 hours show that the biomass density of the cultures did not change significantly, and therefore is fully consistent with the cultures being at steady–state.

The cell–size distributions in all limitations (Fig.1.8) show very good reproducibility as demonstrated by the close overlap of the red traces corresponding to individual measurements. Since in all cases the samples analyzed on the Coulter counter where diluted $1/1000$, the highest growth rates having the lowest biomass (Fig.1.7) exhibit higher noise levels. This problem can be easily overcome by using a lower dilution ratio for the highest growth rate cultures, such as $1/200$. Varying the dilution level to achieve optimal biomass densities was applied successfully in later experiments and these were the pioneering experiments that indicated the need to adjust the dilution ratio. The cell–size distributions in all limitations Fig.1.8 show a growth rate trend of increased variance with the growth

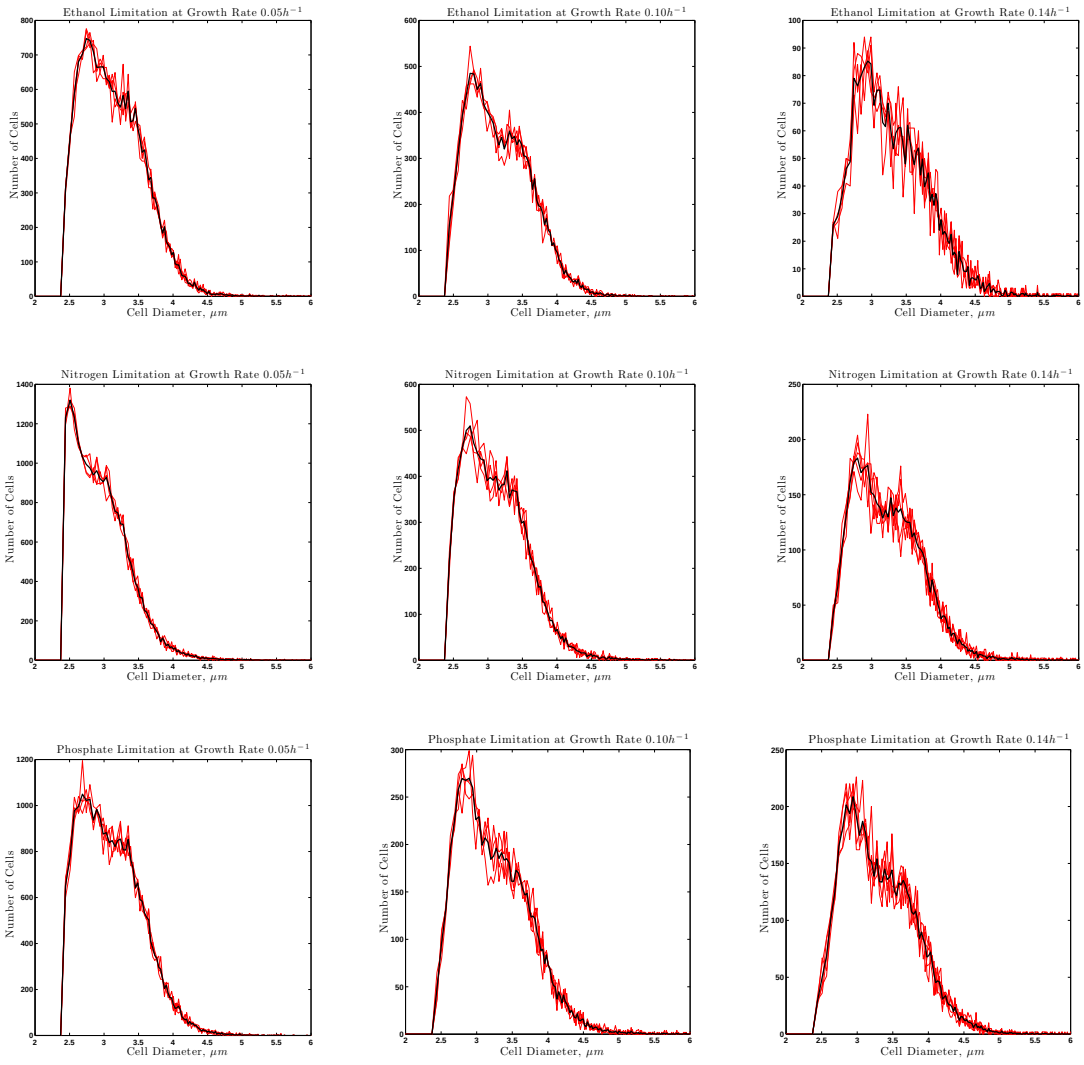

Figure 1.8: Distributions of cell sizes as a function of the growth rate across nutrient limitations of cultures using ethanol carbon source. The red traces are individual measurements and the black curves are the average of the individual measurements.

rate due to increasing fraction of cells with larger diameters, Fig.1.8. Furthermore, some distributions, such as the cell–sizes of the ethanol limitation at $\mu = 0.10h^{-1}$ appear to be bimodal. The most likely reason for this trend is the increased fraction of budded cells as indicated by Fig.1.6. Based on this argument, I developed and fit a model explaining the observed cell–size distributions as a mixture of the two populations (budded and non–budded cells) with different cell–size distributions, Fig.1.9. The simplest model that

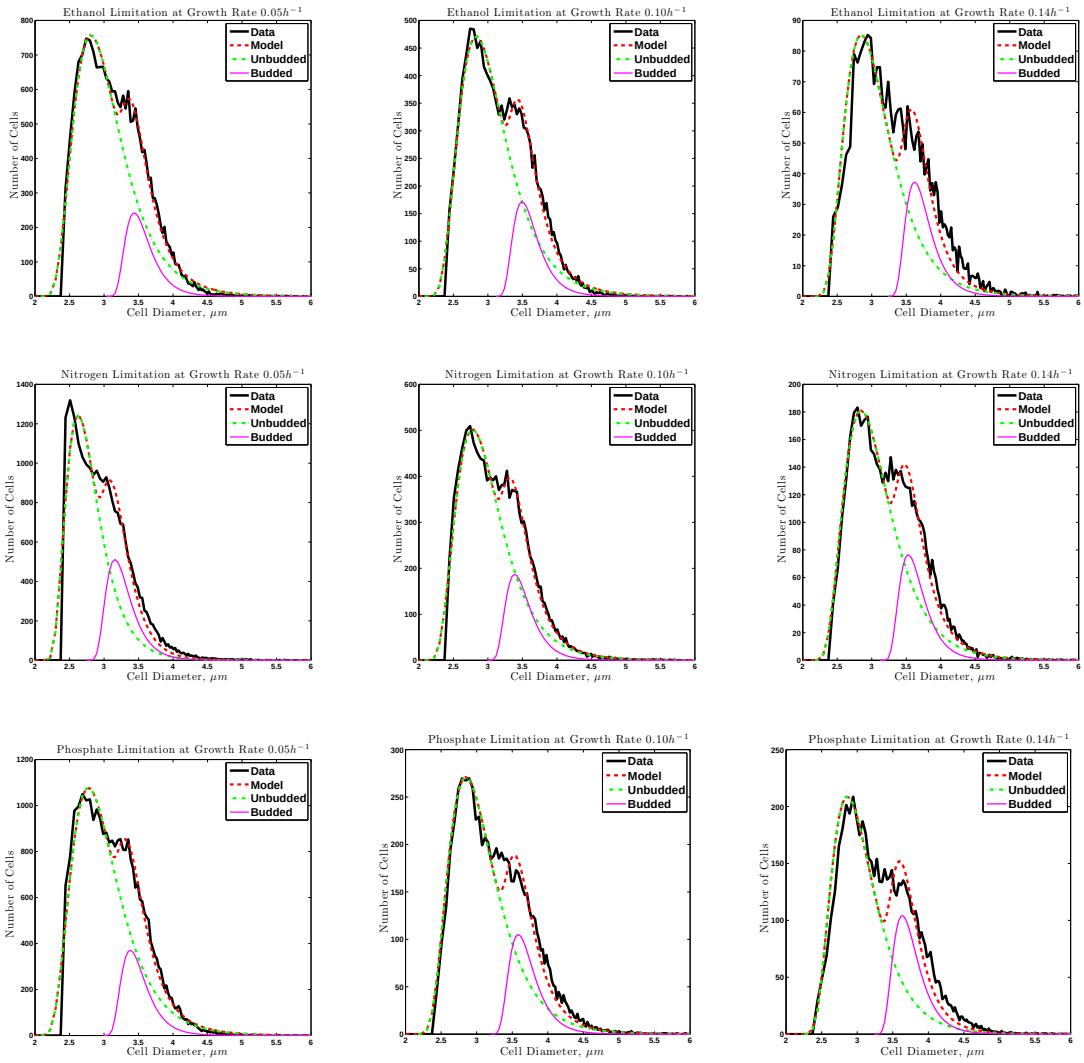

Figure 1.9: The empirical distributions (black) of cell sizes are decomposed into a mixture of two distributions, non-budded cells (green) and budded cells (magenta). The cell–size distribution predicated by the model (a superposition of the budded and non-budded distributions) is plotted as dotted red curves.

was first fit to the data used Gaussian distributions but it gave rather large systematic deviations from the data. The reason for those deviations is the significant asymmetry in the distributions, Fig.1.8. Thus generalized Gaussian distributions were used resulting in a much better fit to the data without significant systematic deviations, Fig.1.9. Why might the cell–size distributions follow generalized rather than simple Gaussian distributions? A very simple and likely explanation is the non–linearity in the interactions of

the factors that determine cell size in yeast. A Gaussian distribution is expected based on the Central Limit Theorem which is derived and proved based on the assumption of linear superposition of many independent factors. If the cumulative effect of the factors determining cell–size is not based on simple summation of their individual effects and/or those factors are not independent (both of which are very likely) the expected distribution is more likely to be generalized Gaussian as the data indicate, Fig.1.8.

The model fits depicted on Fig.1.9 can be used to infer the fraction of budded cells. This model predication is compared (Fig.1.10) to the experimentally measured budded fraction, Fig.1.6. While the model prediction and the measurement show strong correla-

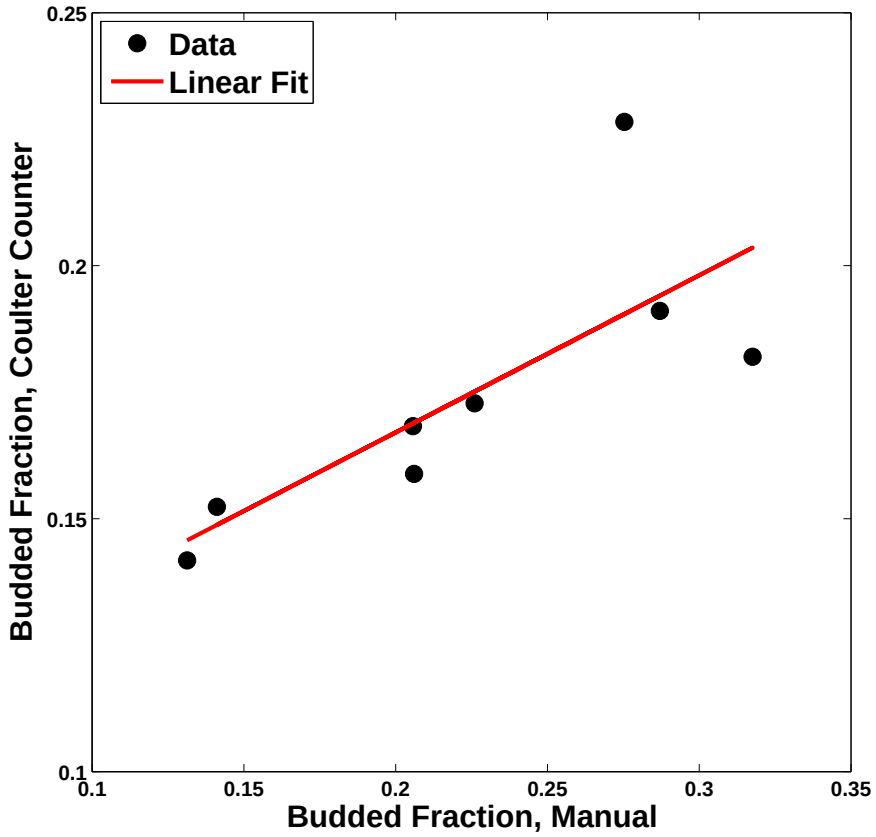

Figure 1.10: Bud Index Comparison. The fraction of budded cells predicted by the model of distribution sizes is plotted versus the fraction of budded cells counted experimentally, see Fig.1.6

tion, the deviation at the highest growth rates are significant which is most likely due to the noisier data because of the lowest biomass density at the highest growth rates, Fig.1.7. More significantly, the intercept of the linear fit does not go through $(0, 0)$ which most likely is caused by different shapes of the distributions of budded and non–budded cells. Indeed, adding a fudge parameter for different shapes moves the intercept toward zero and further improves the fit to the data.

1.5 Transcriptional Growth Rate Response

To measure RNA levels (Fig.1.11), I sampled *10-30ml* of steady–state culture from each limitation and growth rate and vacuum filtered the cells followed by immediate refrigeration in liquid nitrogen and then in a freezer at -80^oC. RNA for microarray analysis was extracted by the acid–phenol method. RNA was amplified and labeled using the Agilent low RNA input fluorescent linear amplification kit (P/N 5184-3523; Agilent Technologies, Palo Alto, CA). This method involves initial synthesis of cDNA by using a poly(T) primer attached to a T7 promoter. Labeled cRNA is subsequently synthesized using T7 RNA polymerase and either *Cy3* or *Cy5* UTP. Each *Cy5*-labeled experimental cRNA sample was mixed with the *Cy3*-labeled reference cRNA and hybridized for 17 h at 60^oC to custom Agilent Yeast oligo microarrays having 8 microarrays per glass slide. Microarrays were washed, scanned with an Agilent DNA microarray scanner (Agilent Technologies), and the resulting TIF files processed using Agilent Feature Extraction Software version 7.5. Resulting microarray intensity data were submitted to the PUMA Database for archiving. When data are clustered hierarchically, the similarity metric (non–centered correlations) is computed using all data shown in the plot unless otherwise specified. For implementing hierarchical clustering I used either the heuristic algorithm implemented by the PUMA Database `http://puma.princeton.edu/`or solved the combinatorial permutation problem of hierarchical to optimal solution using traveling salesman algorithm, see appendix 5.1.

Initially, the reference used for all experiments was from a glucose $(0.8g/L)$ limited culture growing at $\mu = 0.025h^{-1}$, Fig.1.11. To compare the growth rate data from ethanol carbon source to the growth rate data from glucose carbon source, I also hybridized 4 of my samples from ethanol carbon source with the reference used by Brauer *et al* (2008), glucose $(0.8g/L)$ limited culture growing at $\mu = 0.25h^{-1}$. From each sample,

I compute a gene specific offset indicating how much the expression level of each gene has to be shifted for renormalizing my expression data to the reference used by Brauer *et al* (2008), Fig.1.12. The correlation between offsets computed from the 4 samples is excellent, (Fig.1.12) indicating good reproducibility of the experiments. The median offset (from the four samples) was used for converting the expression data to the Brauer *et al* (2008) reference. The converted data is displayed together with the data from growth rate experiments on glucose carbon source, Fig.1.13:

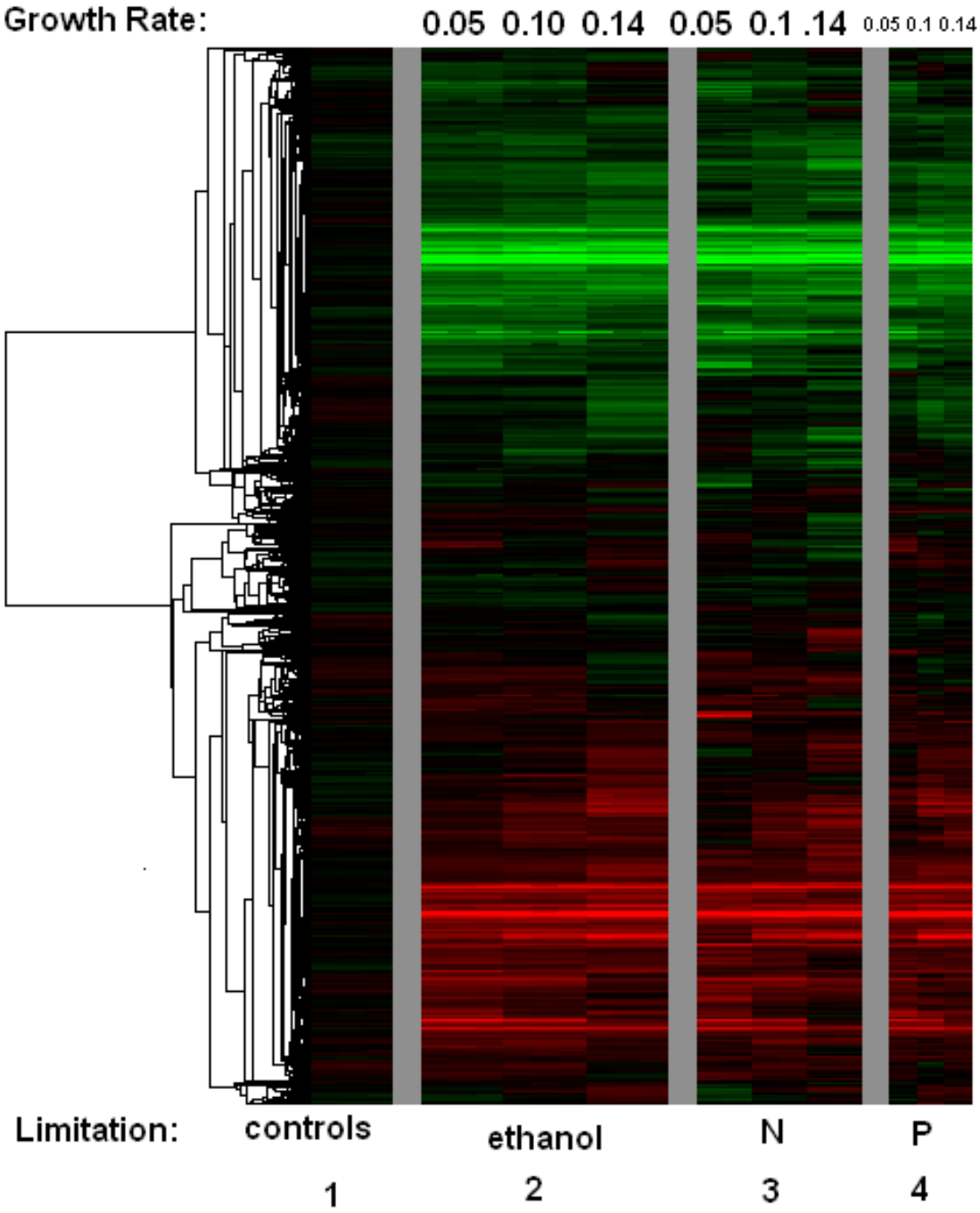

Figure 1.11: A clustergram of the gene expression data from continuous cultures on ethanol. The first column (1) corresponds to controls, second column (2) to ethanol limitation 3 replicas per growth rate ordered from slowest to fastest growth $\mu = \{0.05, 0.10, 0.14\}h^{-1}$, third column (3) to nitrogen limitation 2 replicas per growth rate ordered from slowest to fastest growth $\mu = \{0.05, 0.10, 0.14\}h^{-1}$, fourth column (4) to phosphorus limitation 1 replica per growth rate ordered from slowest to fastest growth $\mu = \{0.05, 0.10, 0.14\}h^{-1}$. The similarity metric (non–centered correlations) is computed using all data shown in the clustergram.

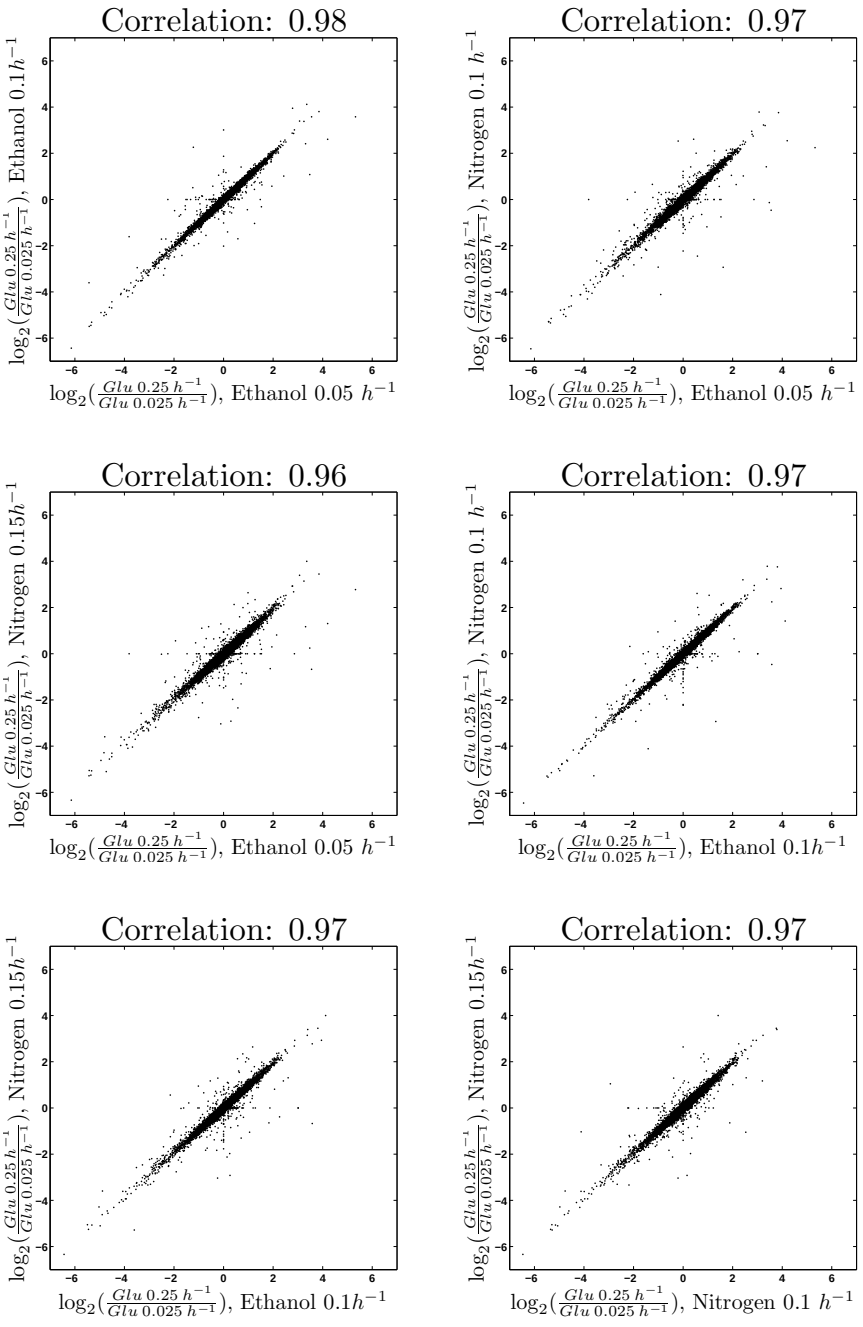

Figure 1.12: Correlation between offsets for reference switching computed from the 4 samples from ethanol carbon source.

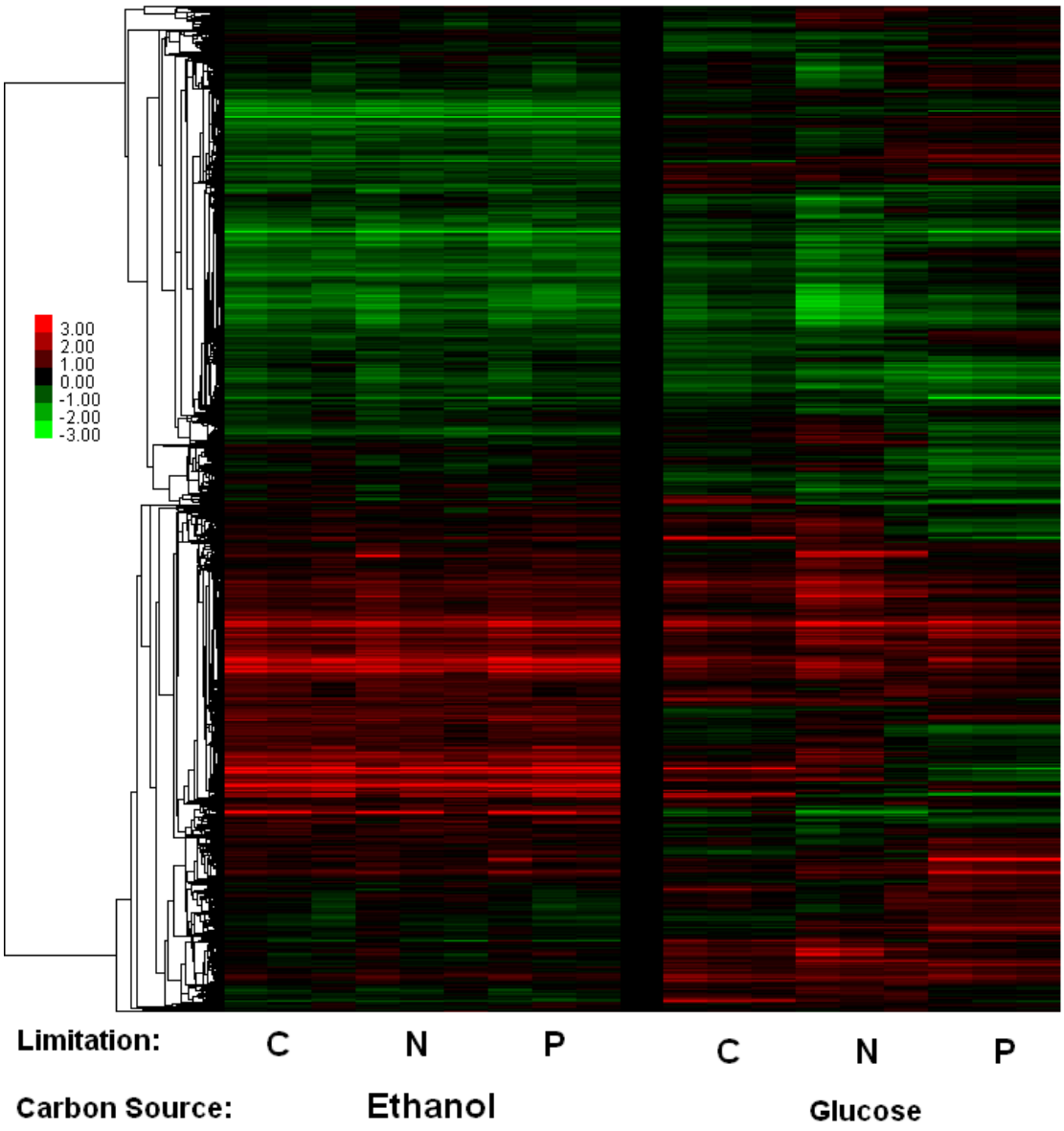

Figure 1.13: A clustergram of the gene expression data from continuous cultures on ethanol (left set of columns) and glucose (right set of columns). Each set of columns (carbon source) contain 3 limitations (carbon, nitrogen, phosphorus) and each limitation has 3 growth rates ordered from slowest to fastest growth $\mu = \{0.05, 0.10, 0.14\}h^{-1}$, for ethanol carbon source and $\mu = \{0.05, 0.10, 0.15\}h^{-1}$ for glucose carbon source. The similarity metric (non–centered correlations) is computed using all data shown in the clustergram.

1.6 Metabolic Growth Rate Response

I used the Vacuum Filtering Method (Boer *et al*, 2008). Between 13 and 30 milliliters of culture (depending on the biomass density) was rapidly sampled from the chemostat and vacuum filtered over a $0.45\mu m$ pore size, $25mm$ nylon filter (Millipore), and the filter was immediately quenched in $0.6ml$ of 20^oC extraction solvent (acetonitrile:methanol:water, $40:40:20$). After 15 min at 20^oC, I mixed the cell material with the extraction solvent, washed the filter with an additional $0.1ml$ of extraction solvent, centrifuged the resulting suspension at 4^oC, and set aside the supernatant. The pellet was extracted again with $0.1ml$ of extraction solvent for $15min$ at 4^oC, the suspension was again centrifuged, and the supernatants were pooled (total extraction volume, $0.8ml$).

1.6.1 C-MS/MS Analysis

Chris Crutchfield input the cell extracts that I prepared to liquid chromatography-electrospray ionization-triple quadrupole mass spectrometry in multiple reaction monitoring (MRM) mode. Positive ionization mode analysis was on a Quantum Ultra triple quadrupole mass spectrometer (Thermo Electron, San Jose, CA), coupled to hydrophilic interaction chromatography on an aminopropyl stationary phase. Negative ionization mode analysis was on a Finnigan TSQ Quantum DiscoveryMax triple quadrupole mass spectrometer (Thermo Electron) coupled to tributylamine ion-pairing reversed phase chromatography on a C18 stationary phase. Autosampler temperature was 4^oC and injection volume was $10\mu l$.

1.6.2 Normalization

To convert raw *LC-MS/MS* ion counts to relative cellular concentration data, ion counts were first normalized by the total cell volume extracted and the volume of the extraction solvent. Normalized ion counts were then converted to relative concentrations by dividing the value for the experimental samples by the corresponding value from the phosphate limited reference chemostat cultures $\mu = 0.05h^{-1}$ used by Boer *et al* (2010).

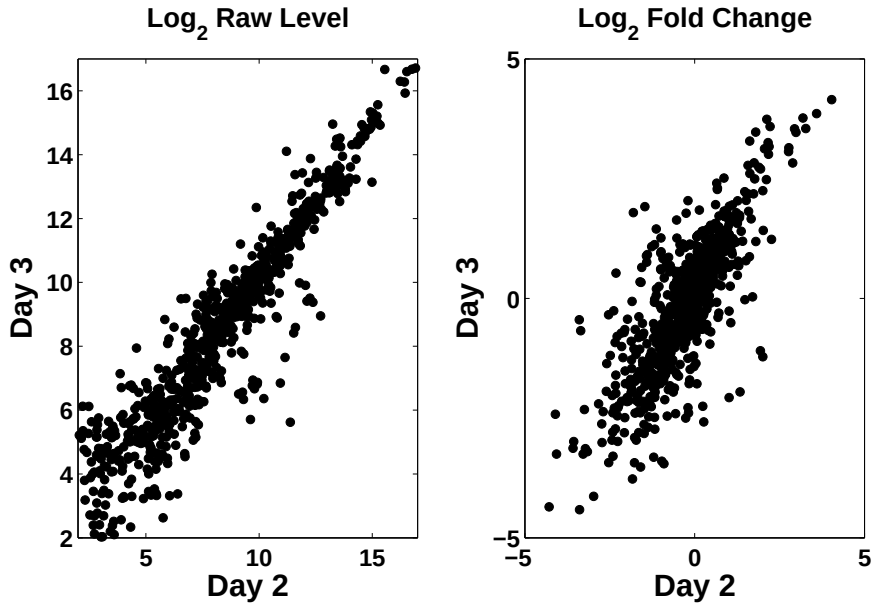

Figure 1.14: Reproducibility in the metabolite measurements

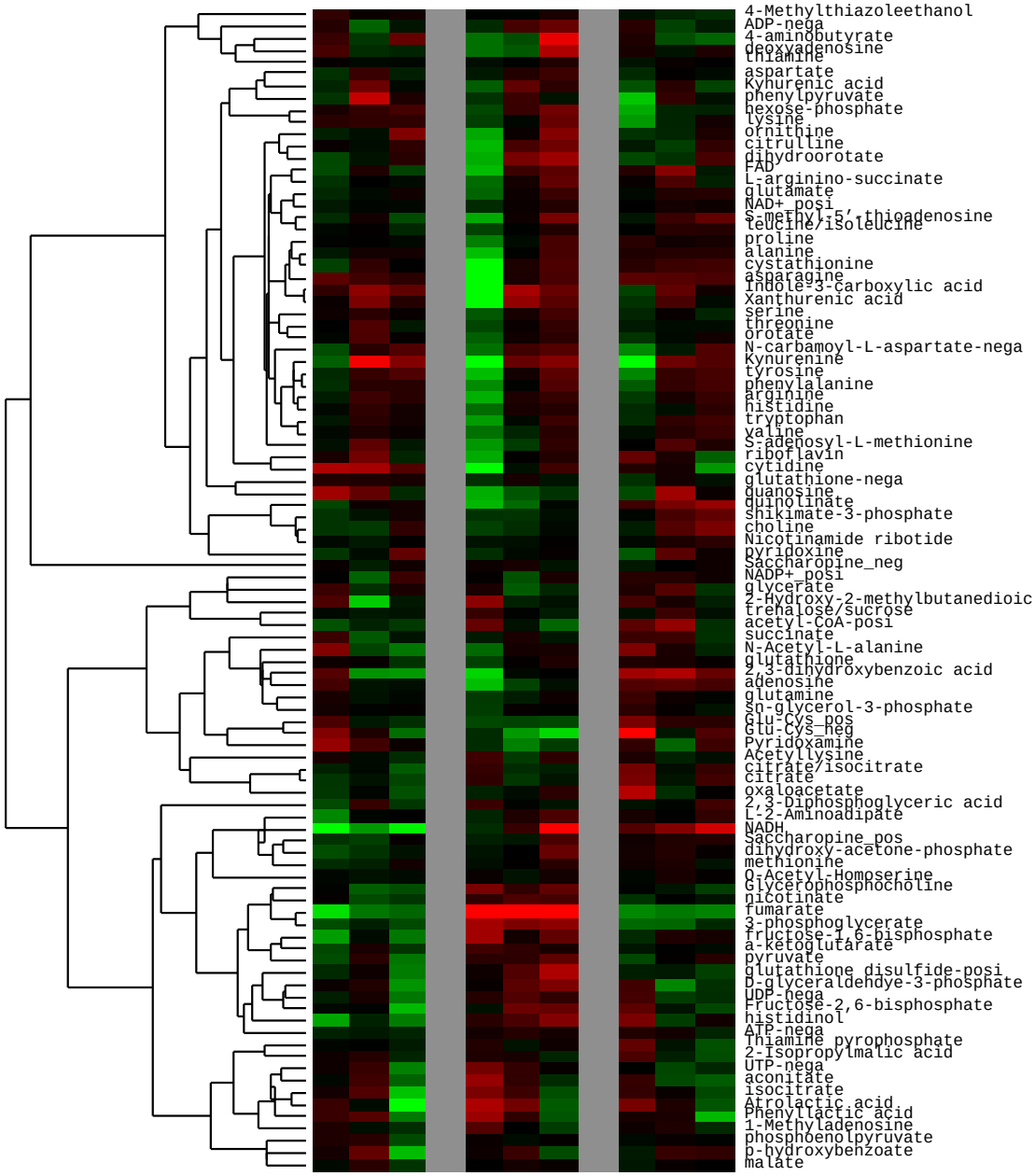

Figure 1.15: Metabolites Clustergram. Each column corresponds to a nutrient limitation. From left to right: ethanol, ammonium and phosphate. Each limitation has 3 growth rates ($\mu = \{0.05, 0.10, 0.14\}h^{-1}$) in increasing order. The similarity metric (non–centered correlations) is computed using all data shown in the clustergram.

Chapter 2

Analysis

2.1 Introduction

The most basic question to ask from the gene expression and the metabolite datasets is which genes and metabolites differ significantly (in a formal statistical sense) between different nutrient limitations and carbon sources. Such differences may be a different trend (increasing or decreasing levels of mRNAs/metabolites as a function of the growth rate), different mean levels or both. This kind of formal analysis aimed toward identifying a universal growth rate response will be the focus of the second section of this chapter. To facilitate understanding of the identified sets of genes, I use a standard and widely–used approach to finding gene ontology (GO) terms that are significantly overrepresented in the sets of growth rate response genes. To find overrepresented GO terms, I employ the CIL implementation of GO Term Finder (Boyle *et al*, 2004).

While the GO terms finder is a useful tool for identifying functional groups of genes, it has a number of drawbacks including hard thresholding, data discretization (binarization

and the associated loss of information) and decreased statistical power for GO terms with a small number of genes. These drawbacks are discussed briefly in the third section and were greatly mitigated by applying non–parametric statistical analysis for identifying sets of genes with statistically significant growth rate response specific to sets of nutrient limitations or to carbon sources. This analysis makes use of GO terms and prior knowledge of biological networks and pathways.

The fourth section is devoted to analyzing transcriptional changes in well–characterized metabolic and regulatory pathways. Our understanding of the many biochemical reactions involved in nutrient catabolism makes robust predictions about the reactions whose metabolic flux should change significantly and it is interesting to explore the isoenzymes whose levels change significantly and are thus likely to mediate the expected changes in metabolic fluxes.

Once the genes with common and differential growth rate responses are identified one might ask what are the underlying regulatory mechanisms behind the observed growth rate response. The analysis from the fourth section is limited only to well known metabolic and signaling pathways. Ideally, one would like to make this kind of analysis more quantitative and extend it to all signaling networks transducing the signal from sensing nutrient levels through biochemical pathways to systems–level physiological responses such as growth rate. The existing technologies and the lack of data about many of the intermediate levels fundamentally limit comprehensive inference. The data I have collected allow me to infer regulatory interactions primarily at the level of transcription and mRNA degradation. This is the subject of the fifth section. Different methods are applied and the results compared. The comparison and the emerging shortcomings motivate the need for a network inference algorithm. Such an algorithm, $\mathcal{RC}web$, is derived, tested and applied to the data in the next chapter.

2.2 Identifying Universal Growth Rate Response

2.2.1 Slopes/Exponents

A useful way to analyze the growth rate data is to fit a model that captures the growth rate response trends. Given the limited number of datapoints per gene, only simple models can be used without over–fitting the data. The explicit form of the model that should describe the data is hard to derive based on first principles. Among the simplest models is a linear model with two parameters per gene, one parameter for the mean nutrient specific level and one in which the mRNA concentration depends linearly on the growth rate, $mRNA_i = a_i\mu + b_i$. Based on sigmoidal (or as a sub case Michaelis Menten) signal transduction transfer functions and using some approximations, one may expect a power–law dependence (also needing 2 parameters per gene) between the growth rate (μ) and the change in mRNA levels, $mRNA_i = b_i\mu^{a_i}$. Brauer *et al* (2008) and Airoldi *et al* (2009) used a model with 2 parameters per gene with exponential dependence between the mRNA level and the growth rate, $mRNA_i = b_i e^{a_i\mu}$. In semi–log space, such exponential model "appears" linear $\log(mRNA_i) = \log(b_i) + a_i\mu$. Similarly, in log–log space the power–law model becomes linear, $\log(mRNA_i) = \log(b_i) + \log(\mu)a_i$.

An unbiased way to assess the performance of those models (having the same level of complexity) is to use a common statistical criterion R^2, which is the fraction of variance in the data explained by the model over the total variance in the data, Fig.2.1. Based on goodness of fit as quantified by R^2, the power–law model captures the largest fraction of the variance for all limitations (Fig.2.1A) and especially on some limitations, such as uracil Fig.2.1B. Since Brauer *et al* (2008) used the exponential model and my use of a model is purely phenomenological (to quantify trends in the transcriptional growth rate response), in the rest of this thesis I will use *only* the exponential model for consistency.

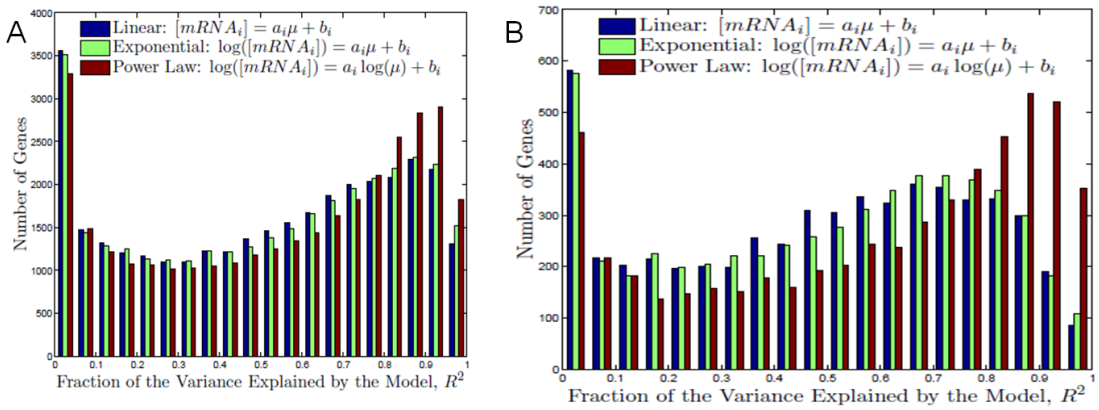

Figure 2.1: Explained fraction of the variance in the gene expression data on glucose carbon source by a linear, power–law and an exponential model. All 6 growth rates $\mu = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}h^{-1}$ are used by all models.

In semi–log space, the exponents in the model "appear" to be slopes, and thus I will use the term slopes in keeping with previous work (Brauer *et al*, 2008).

Computing Slopes and Quantifying Significance

All slopes throughout this thesis are computed based on regression models in the $\ell 2$ sense, minimizing the sum of squared residuals. The regression models use design matrices $\mathbf{X}$ in which each row corresponds to an experiment. For comparison to previous work (Brauer 2007), I start by fitting models that incorporate a nutrient specific constant reflecting the mean expression of each gene for each nutrient. In that case, the first column of $\mathbf{X}$ corresponds (contains) the growth rates, e.g. $X_{i1} = \mu_i$, where μ_i is the growth rate in the i^{th} experiment. The second column corresponds to the first limitation and has ones in the rows corresponding to experiments of the first limitation and zeros in all other rows. The third column corresponds to the second limitation and so on. Then I regress the matrix against the data $\mathbf{Y} \in \mathbb{R}^{M \times N}$, where rows correspond to M conditions and columns to N genes. Then the parameters (slopes and limitation specific constants) in matrix $\hat{\mathbf{C}}$ can be found as the maximum likelihood estimate assuming Gaussian noise in $\ell 2$ sense,

$\hat{\mathbf{C}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. The actual numerical implementation uses *QR* decomposition because of its better stability and numerical properties. As a measure of goodness of fit, I use the fraction of variance explained by the model which for the j^{th} gene is quantified by R_j^2:

$$R_j^2 = 1 - \frac{\sum_{i\in\alpha}(y_{ij} - f_{ij})^2}{\sum_{i\in\alpha}(y_{ij} - \bar{\mathbf{y}}_j)^2} \tag{2.1}$$

In (2.1), $\mathbf{y}_j$ is a vector of expression levels of the j^{th} gene (j^{th} column in $\mathbf{Y}$), $\bar{\mathbf{y}}_j$ is its mean expression level, i is index enumerating the set of conditions α used in the model and f_{ij} is the model predication for the i^{th} condition and j^{th} gene.

Based on analysis described in the appendix 5.2, the slopes computed from the 6 growth rates $\mu = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}h^{-1}$ on glucose carbon source are more similar to the slopes computed on ethanol carbon source. Therefore, all 6 growth will be used for computing the glucose slopes and analysis in this thesis.

2.2.2 Universal Growth Rate Response

To understand some aspects of the growth rate response, one has to look at growth rate response specific to nutrient limitations and to subgroups of limitations. Such specific growth rate responses are discussed in the following subsections. Here, I first consider the growth rate response that is common to all limitations on both glucose and ethanol carbon source. To identify such universal growth rate response, I fit a regression model (2.2.1) explaining the expression levels of each gene with a single gene specific slope. For comparison to previous work (Brauer *et al*, 2008), I also fit models that incorporate nutrient specific constants reflecting the mean expression of each gene for each nutrient. The goodness of fit is quantified by R^2, (2.1). Then, the data for each gene is permuted 10^6 times and the R^2 for each permutation is computed. Based on these computations,

the significance of the growth rate response for each gene is quantified by a *p value* which equals the fraction of R^2 values in the permuted data that are larger than the R^2 for the non–permuted data, Fig.2.2. It might be tempting to interpret the high R^2 and low *p*–

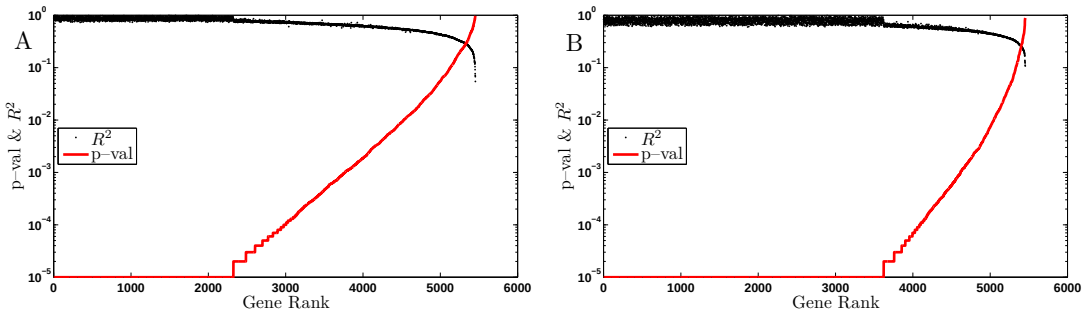

Figure 2.2: Rank ordered p values and the corresponding R^2 from a model accounting for both the nutrient mean effect (by a nutrient specific constant) and for the growth rate response by a gene specific slope, (Brauer *et al*, 2008). (A) Results for carbon, nitrogen and phosphate limitations on glucose and ethanol carbon source and (B) Results for all limitations on glucose and ethanol carbon source.

vals (Fig.2.2) as evidence for universal growth rate response shared by half the genome, Fig.2.2A or even $3/4$ of the genome, Fig.2.2B. Yet this conclusion overlooks the fact that the model includes *both* the growth rate *and* the effect of the nutrient limitation on the mean level of gene expression. Therefore, a gene whose mean expression level differs significantly among limitations and carbon sources will fit the model well (high R^2) even if its slopes on individual limitations differ slightly; As long as most of the variance in the expression of such a gene comes from different limitation–specific mean levels, its R^2 will always be high independent of its slope on different limitations. The significance of the fit for such a gene is further bolstered by the large amount of high quality data coming from many limitations, Fig.2.2B. Prime examples for genes in this group are genes involved in ethanol utilization and gluconeogensis as discussed in section 2.4. In particular, genes such as the cytoplasmic malate dehydrogenase *MDH2* that have large positive slopes in all ethanol carbon source limitations and large negative slopes in all glucose carbon source limitations fit the models with relatively large negative slope and high R^2. The reason for the high R^2 is that most of the variance in the data for *MDH2*

comes from the much higher level of expression of *MDH2* on ethanol carbon source and the model explains this variance well because of the nutrient specific constant. The net negative slope reflects the fact that there are more glucose carbon source conditions and a net negative slope minimizes the sum of squared residues. At a more conceptual level, the fact that adding more limitations in Fig.2.2B compared to Fig.2.2A increases the number of genes with very significant fits indicates that the model cannot be used for identifying a common growth rate response. The number of genes with a common growth rate response can only decrease (but not increase) by adding more conditions. The increase in the number of genes for which the model explains a significant fraction of the variance reflects the fact that the larger the set of profiled conditions, the more likely that a gene will be expressed differentially for at least one of the conditions.

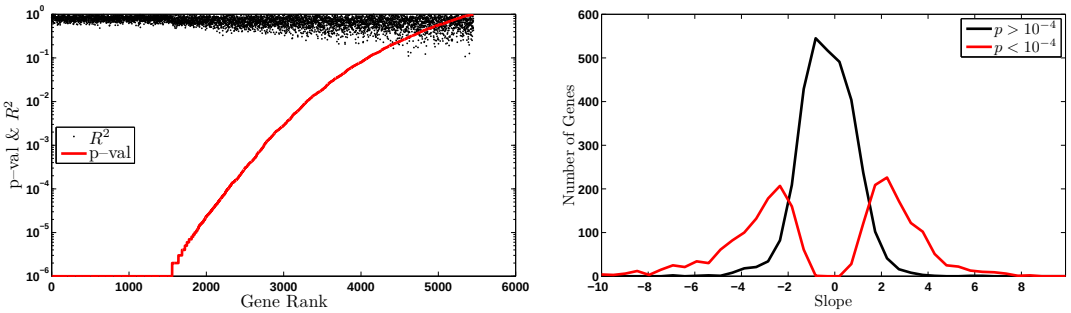

Figure 2.3: Rank ordered p values and the corresponding R^2 for all growth rate response gene expression data using a model accounting for the growth rate response *only* by a gene specific slope and a constant. The p values are computed from 10^6 bootstrap resamplings of the data. (A) Goodness of Fit (B) Distribution of Slopes.

Because of the nutrient mean effect, the results from the growth rate model used by Airoldi *et al* (2009) are hard to interpret as a common growth rate response especially in the case of conditions (such as different carbon sources) associated with substantial differences in condition specific mean expression levels. One approach to finding the genes with common growth rate response among the genes that fit the model is to use nested models with multiple slopes and assess the improvement in the goodness of fit. A much simpler approach, which I am using, is to eliminate the contribution coming

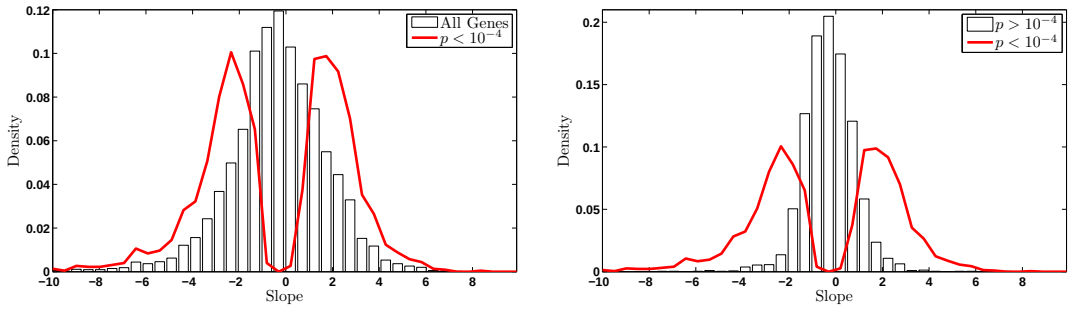

Figure 2.4: Distribution of slopes according to the significance of the explained variance by a model accounting for the growth rate response *only* by a gene specific slope and a constant and fit too all growth rate response data. (A) The bar graph corresponds to the distributions of slopes of all genes (B) The bar graph corresponds to the distributions of slopes of genes that do not fit the model with high significance

from explaining the nutrient dependent mean level of expression. This can be done by replacing R^2 with an F–statistic comparing the variance explained by two models: 1)A model accounting for both the growth rate and the nutrient mean effect and 2)A model accounting only for the nutrient mean effect. A yet simpler approach to eliminating the nutrient dependent mean level of expression is to normalize the data for each gene across each nutrient to mean (arithmetic average) zero (zero–centering) and then explain the observed variance in the normalized data by only two parameters: the growth rate μ and a gene specific constant. Because of its parsimonious and conceptual simplicity, I will use this approach. The results of fitting such a model to the data for all carbon sources and limitations (Fig.2.3) indicate that at a stringent p value cutoff of 10^{-6}, the model explains a significant fraction of the variance in the expression of about 1500 genes. In keeping to previous work, I also plot the distribution of slopes for all genes as a bar–plot histogram Fig.2.4A. Because of the large number of genes that fit the model well and the significant difference between their slopes and the slopes of the genes that do not fit the model well, the results are more clearly emphasized when the bar–plot distribution is limited whose fit to the model has high p values, Fig.2.4B.

As expected, the distribution of R^2 values is shifted toward lower values compared to the more complex model (2.2) but the significance for a set of about 2000 genes remains

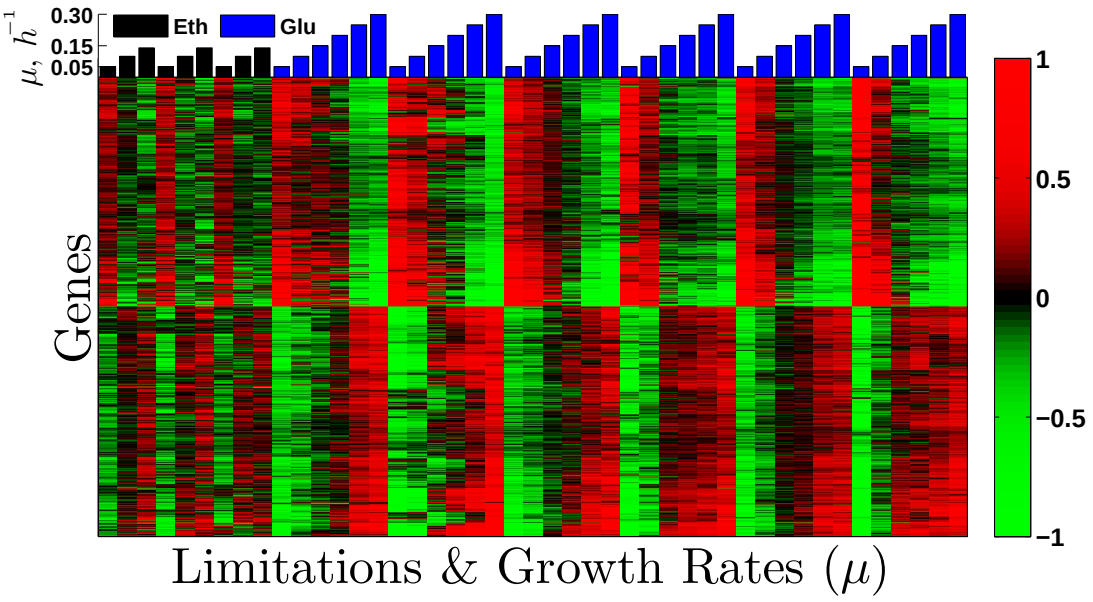

Figure 2.5: Zero–Centered Expression Levels of Genes with Universal*GRR*. Expression levels of the genes with best fits to the *GRR* model are normalized to mean zero for each limitation and clustered. The first 9 columns correspond to ethanol (black bars) carbon source and limitations on ethanol, nitrogen and phosphorus, 3 growth rates per limitation arranged in ascending order, $\mu = \{0.05, 0.10, 0.14\}h^{-1}$. The next columns (blue bars) correspond to glucose carbon source and limitations on glucose, nitrogen, phosphorus, sulfur, uracil and leucine, 6 growth rates per limitation arranged in ascending order $\mu = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}h^{-1}$.

very high, p value$< 10^{-5}$. A low p–value (and thus significant fit to the model) for a gene indicates that a significant fraction of the variance in the expression levels of that gene can be explained by a single growth rate slope; it does *not* mean that the gene has identical (or even statistically indistinguishable) slopes in all limitations and carbon sources. It does suggest, however, that the trends in the expression levels of genes with high R^2 are likely to be similar which can be seen to be the case indeed, Fig.2.5.

It is interesting to examine whether the genes used by Airoldi *et al* (2009) for growth rate predictions are among the genes with universal growth rate response, Fig.2.6. As expected and selected by Airoldi *et al* (2009), each of the genes used in the growth rate model has the same slopes across all limitations on glucose carbon source, Fig.2.6. On ethanol carbon source, many of those genes have expression profiles similar to those in

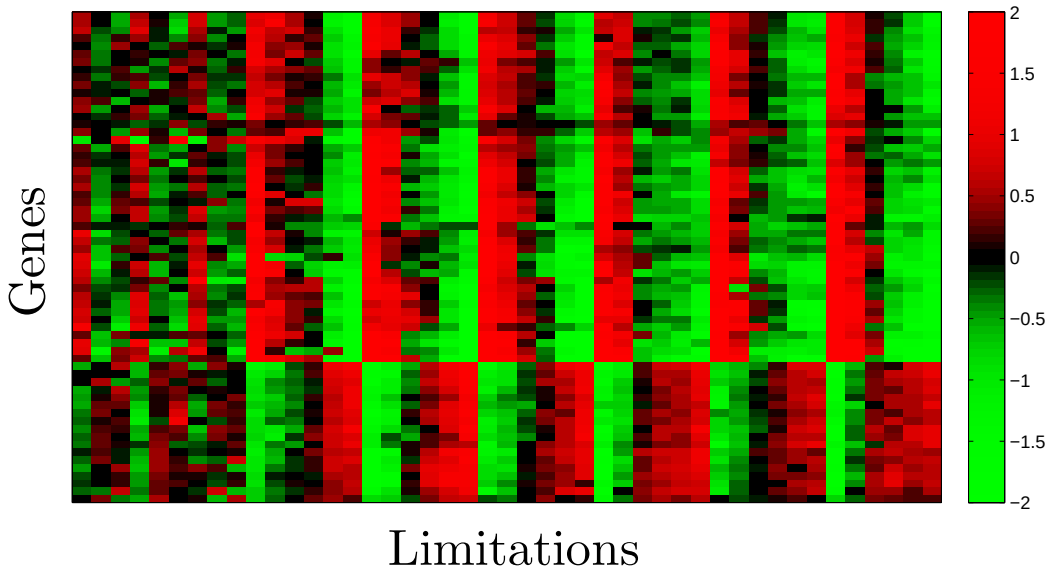

Figure 2.6: Genes Used in Predicting Growth Rate by Airoldi *et al* (2009). For notation see Fig.2.5

glucose and still correlate well to growth rate while the expression of other genes do not show good correlation to growth rate. Based on this result, any small set of genes from the cluster of universal growth rate response depicted on Fig.2.5 should perform better in estimating the growth rate of yeast cultures independent of the carbon source. Predicting growth rate by gene expression is easy and such a small set of genes might be chosen simply on the basis of large slopes and high R^2 as Airoldi *et al* (2009) did. Almost as easy and theoretically better justified is choosing such a gene set by sparse $\ell 1$–regularized logistic regression (Friedman *et al*, 2009; Liu *et al*, 2009), which is likely to result in a model with higher predictive accuracy.

Given the large number of growth rate response genes, one may expect the presence of a large growth rate component in the variance in gene expression levels. Such a growth rate response component is indeed present in the singular value decomposition (Golub and Kahan, 1965; Alter *et al*, 2000) of the gene expression data, Fig.2.7. The first left singular vector (accounting for 46% of the variance) is strongly correlated to

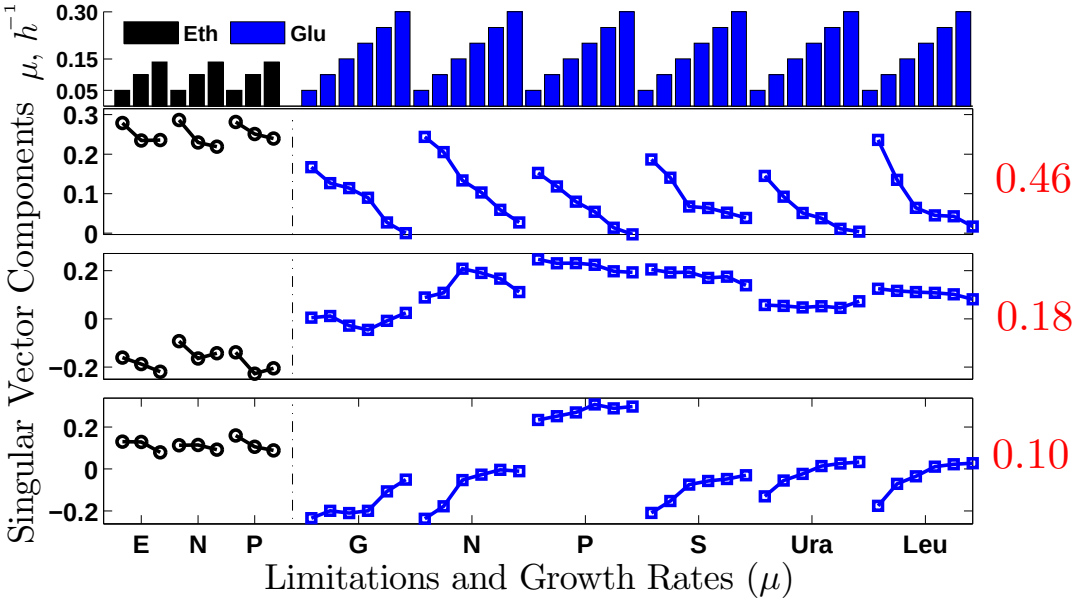

Figure 2.7: Singular value decomposition (SVD) of the gene expression data. Top panel corresponds to the first singular pair, middle to the second and bottom to the third. The fraction of variance explained by each singular pair is indicated by the numbers in red to the right of the corresponding panels.

the growth rate. Actually, it is anticorrelated on Fig.2.7 but in the context of singular value decomposition (SVD) correlation and anticorrelation to a factor are the same since flipping the signs of the corresponding singular vectors results in SVD that is just as valid and explains the variance in the data just as well. The strong correlation between growth rate and the first singular vector is not entirely surprising since the growth rate is the only systematic design variable common to all conditions and gene expression is strongly correlated. It is, however, reassuring and indicative of a very substantial growth rate response component common to nutrient limitations and carbon sources. The fact that the elements corresponding to ethanol carbon source are larger reflects the fact that the correlations between gene expression changes in ethanol carbon source conditions are stronger than the correlations between glucose carbon source conditions. The second singular vector is very hard to interpret. The third one correlates strongly to growth rate but in the opposite directions for glucose and ethanol carbon source, suggesting that this

Gene Ontology Term	Cluster freq.	Genome freq.	Corrected p–val	FDR
vacuolar protein catabolic process	84/1148	118/7274	9×10^{-39}	0.00%
stress response	269/1148	848/7274	1×10^{-33}	0.00%
autophagy	84/1148	118/7274	9×10^{-28}	0.00%
cell differentiation	79/1148	247/7274	1×10^{-7}	0.00%

Table 2.1: Overrepresented GO Terms for Genes with Negative Slopes. Full list can be found at: Negative Slope Genes

Gene Ontology Term	Cluster freq.	Genome freq.	Corrected p–val	FDR
ribosome biogenesis	174/1103	437/7274	1×10^{-33}	0.00%
cellular biosynthetic process	510/1103	2203/7274	8×10^{-31}	0.00%
regulation of translation	90/1103	190/7274	2×10^{-23}	0.00%
posttranscriptional regulation of gene expression	510/1103	2203/7274	9×10^{-21}	0.00%
translation	254/1103	962/7274	1×10^{-19}	0.00%
mitochondrial translation	57/1103	110/7274	2×10^{-16}	0.00%

Table 2.2: Overrepresented GO Terms for Genes with Positive Slopes. Full list can be found at: Positive Slope Genes

vector corresponds to genes having opposite growth rate responses in glucose and ethanol carbon source.

It is interesting to ask which are the biological processes overrepresented by the set of genes with universal growth rate response. The simplest approach to finding such over representation is to use the GO Term Finder, (Boyle *et al*, 2004). A very short summary of the most highly overrepresented GO terms is provided in Table.2.1 and Table.2.2. The full list together with the genes that belong to each GO term can be found at my growth rate response website: GRR website, and directly from GO Terms for Negative Slope Genes, and GO Terms for Positive Slope Genes The GO term trees for overrepresented GO terms can be found in the appendix both for genes with negative slopes Fig.5.5 and for genes with positive slopes Fig.5.6. The size and hierarchical structure of the tree

require significant scaling down of the tree to be able to fit it on a single page. Since the figure is vector graphics, however, it can be expanded to any size without any loss of resolution.

Another way to visualize the genes with growth rate response common to any pair of limitations and carbon sources is available at GRR website. Each set of genes defined on the basis of growth rate response is analyzed for significantly overrepresented functions and TFs likely to underly its regulation as described in section 2.5.

The large number of genes having universal growth rate response is a remarkable and important finding. It is fully consistent with the expectation that making protein (Table.2.2) is a major bottleneck for rapidly proliferating cells (Maaløe, 1979) and that slowly growing cells need to recycle proteins and organelles, Table.2.1. The universal growth rate response is also fully consistent with and supportive of the conjecture that an intrinsic yeast metabolic cycle (*YMC*) underlies a substantial part of the growth rate response, see Fig.4.6. Further evidence for and discussion of this conjecture can be found in section 4.

Nonetheless, a significant fraction of genes have nutrient and carbon source specific growth rate response. The identification and analysis of such genes is the subject of next section.

2.2.3 Metabolites

First, I fit an exponential (linear in semi–log space) model to the metabolite data computing a slope for each limitation. The rank ordered R^2 compared to a null model of randomized data for each metabolite indicates that there is clearly a growth rate trend in the data, Fig.2.8

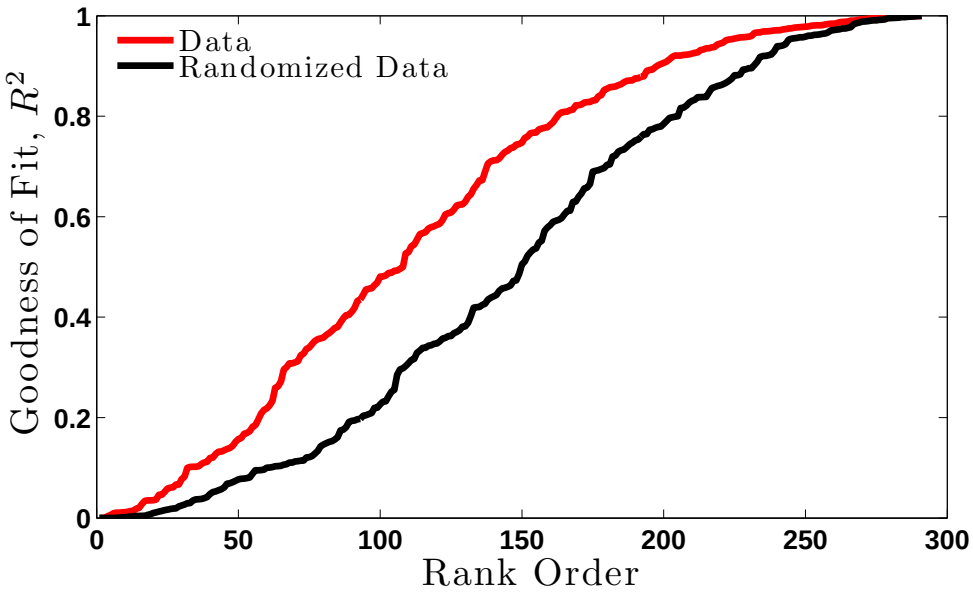

Figure 2.8: Goodness of fit of a linear model to the metabolite data.

The slopes computed for the phosphate and the ethanol limitations correlate remarbaly well Fig.2.9. Many of the metabolites with similar slopes in the ethanol and the phosphate limitations (Fig.2.9) are amino acids and have positive slopes, which is consistent with the strong up–regulation of genes catalyzing amino acid bio–synthetic reactions Fig.2.16. The metabolite slopes in the nitrogen limitation, however, are significantly different from those in the ethanol and phosphate limitations, Fig.2.10. The differences in the nitrogen limitation are reminiscent to differences observed by Boer *et al* (2008) on glucose carbon source. One of the clear distinctions of the metabolites in nitrogen limited cultures is the very low level of amino acids at the slowest growth rate at which the nitrogen shortage is most severe Fig.1.15.

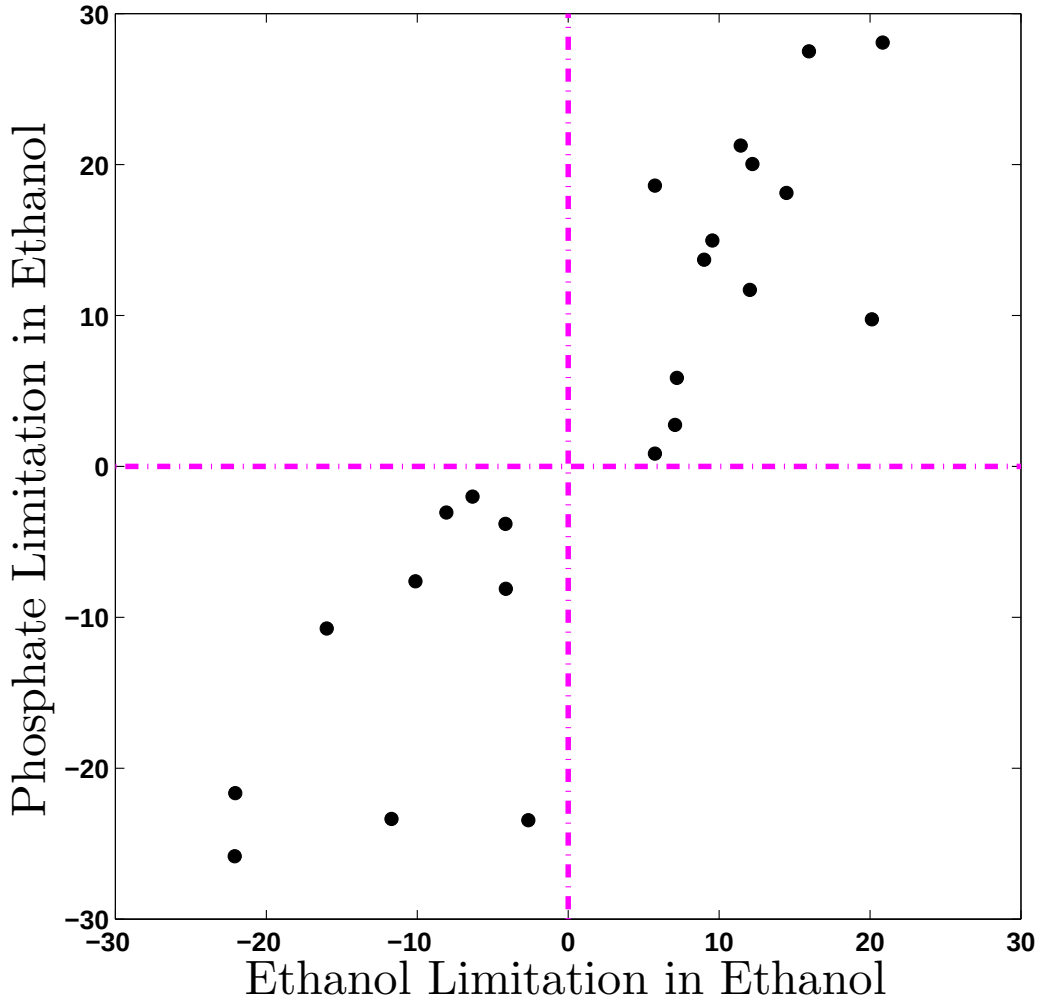

Figure 2.9: Correspondence of metabolite slopes for the ethanol and the phosphate limitations. Only metabolites with high R^2 ($R^2 > 0.85$) for *both* the phosphorus and the ethanol limitations are plotted.

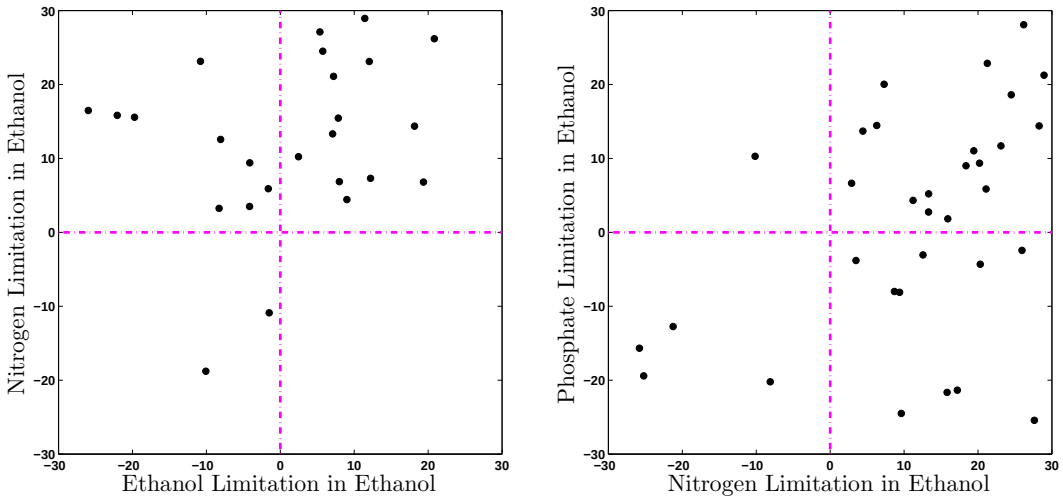

Figure 2.10: Correspondence of metabolite slopes between the nitrogen limitation and the ethanol & phosphate limitations. Only metabolites with high R^2 ($R^2 > 0.85$) for the limitations corresponding to the x and the y axes are plotted.

2.3 Processes and Networks with Differential Growth Rate Response

In addition to genes with a common growth rate response, there are sets of genes whose growth rate response is specific to nutrient limitations and carbon sources. To identify such sets of genes, I first compute a condition (limitation and carbon source) specific slopes for each gene. Next, I want to identify the biological processes overrepresented by those genes similarly to section 2.2. One possibility is to use the GO Term Finder (Boyle *et al*, 2004) again. Because of its drawbacks outlined in subsection 2.3.1, however, I will instead compare and quantify the differences between the distributions of nutrient specific slopes for predefined sets of genes (most of which correspond to GO terms) and the distribution of slopes (for the same nutrient) for all genes.

2.3.1 Methodology

The GO Term Finder has a number of drawbacks including:

1. The list of genes input to the GO Term Finder has to be defined based on thresholding a quantity of interest, such as the magnitude of growth rate response slopes or goodness of fit to a model, R^2. Often, the optimal position for such thresholding is hard to determine and justify.

2. Even if a good threshold is found and used, its application results in loss of quantitative information about the magnitudes of the quantity that is being thresholded. For example, when I thresholded genes based on R^2 and slope in section 2.2 and subsequently submitted a list to the GO Term Finder, no distinction was made between a gene with a slope 15 and a gene with a slope 3. Most quantitative information in the data is not used.

3. Some physiological processes, such as the growth rate response, have genome–wide consequences and affect a very large number of genes. Analyzing a long list of affected genes can greatly reduce the power to find statistically significant enrichment for biological functions performed by smaller sets of genes. For example, submitting a query list of 10^3 genes (such as genes with universally positive growth rate response, section 2.2) is associated with a relatively large probability that all 5 genes involved in a particular function will be in the list *even* when the list is sampled from the genome at random. Thus, such a test does not have the statistical power to identify a growth rate response specific to a biological processes performed by small sets of genes no matter how extreme and pronounced the growth rate response of those processes and genes is.

These drawbacks can be greatly mitigated by applying a simple non-parametric statistical analysis based on comparing the distribution of the quantity of interest (such as growth rate slope) for a predefined group of genes and the corresponding distribution for the whole genome. I will use this method for one dimensional comparison (only for the growth rate slopes) but it can be generalized easily to high dimensional comparisons including multiple characteristics of interest. As a simple summary statistic quantifying the magnitude of the growth rate response for each gene set, I will use the mean slope. I choose the mean slope over the median because I do not want to lose the effect of outliers, the genes whose growth rate response slopes are most extreme. Depending on the desired outcome, one might choose another summary statistic. To quantify the significance of the difference between the distribution for a particular gene set and the distribution for all genes, I use Wilcoxon rank–sum (Mann Whitney) test. It is a non-parametric test for comparing two distributions without making assumptions about their shapes. It gives the probability of obtaining greater observations in one population versus the other by

chance alone. The null hypothesis in the rank–sum test is that both samples have the same probability of exceeding each other.

2.3.2 Clustergram of Slopes

An intuitive way to display similarities and differences in growth rate response between different carbon sources and nutrient limitations is as a clustergram of slopes, Fig.2.11. For each set of growth rates on the same carbon source and nutrient limitation, I compute a slope for each gene. Those slopes are plotted as a heat map matrix plot color coded in red/white/blue spectrum to distinguish this type of derivative data from the gene expression data presented in red/black/green spectrum, Fig.2.11.

One obvious difference is that the slopes on ethanol carbon source tend to be larger in absolute value than the slopes on glucose carbon source as demonstrated by the more intense colors of the first set of three columns as compared to the second set of three columns. The same difference is quantified by the larger variance of the distribution of slopes in ethanol carbon source, Fig.5.2. A second global feature is that the slopes in carbon and phosphate limitations appear more similar to each other compared to the slopes in nitrogen limitation. This trend can be quantified by the Pearson correlations between the slopes computed by averaging across genes, Fig.5.3.

Another salient feature of the clustergram is that some genes have very similar slopes across all conditions (the universal growth rate response), some genes have very similar slopes for each carbon source but different slopes across carbon sources, and other genes have limitation specific slopes. To explore systematically the biological functions represented by those sets of genes and assess the statistical significance for each set of genes, in the next subsection I apply the analysis described in the Methodology subsection 2.3.1.

Since the absolute magnitude of the fold changes (third set of columns on Fig.2.11) are smaller than the absolute magnitude of the slopes, the clustering groups genes primarily by slopes but not by fold change. To detect structure in the fold changes, I cluster them alone, Fig.2.12. The clustergram of fold changes shows that substantial number of genes

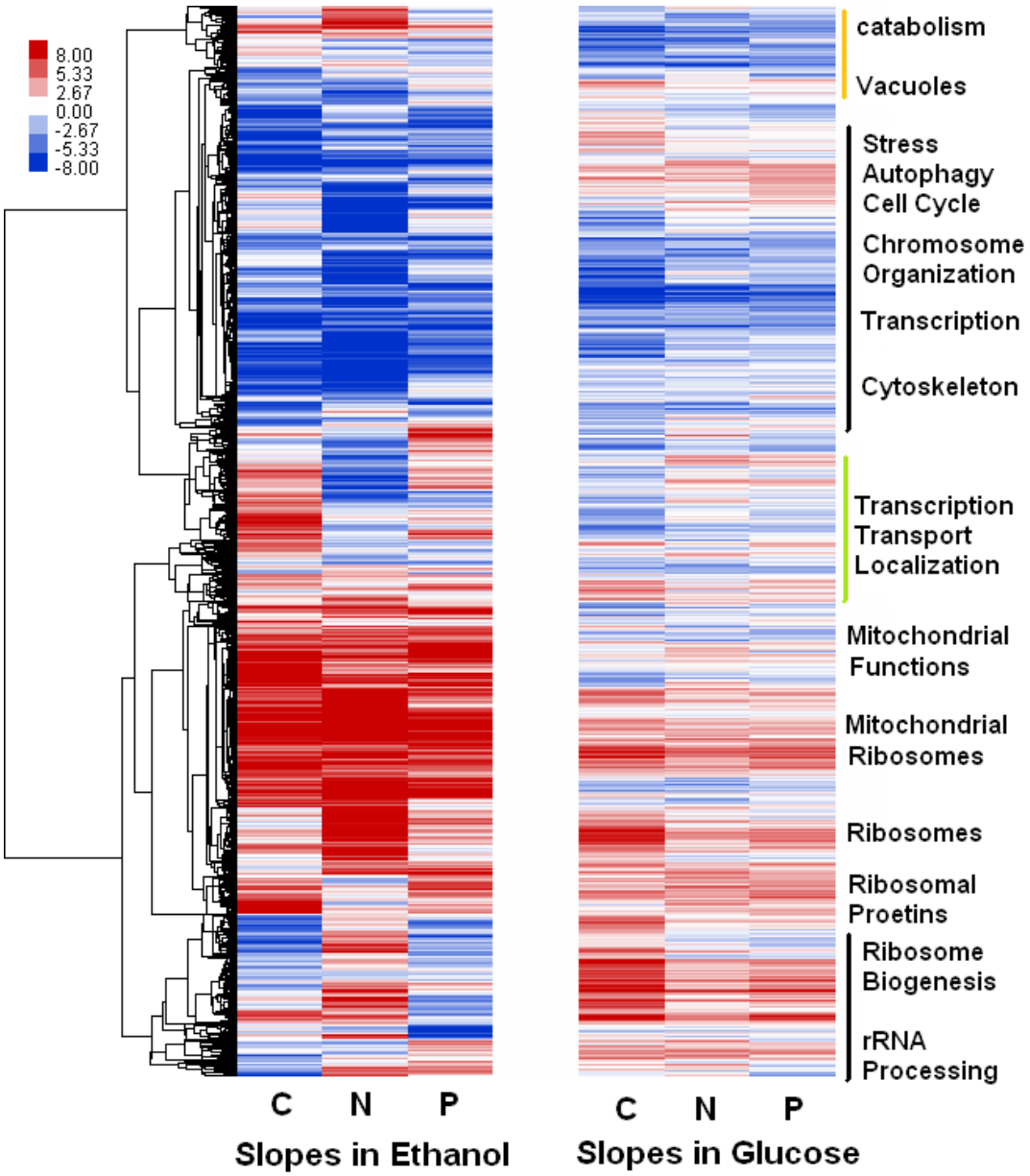

Figure 2.11: A clustergram of slopes and fold changes. The first set of 3 columns corresponds to slopes in cultures growing on ethanol carbon source and limited on carbon/ethanol (C), nitrogen (N) and phosphorus (P). The second set of 3 columns corresponds to slopes in cultures growing on glucose carbon source and limited on carbon/glucose (C), nitrogen (N) and phosphorus (P). The third set of 3 columns corresponds to fold change difference in expression levels between in cultures growing on ethanol carbon source and glucose carbon source. The columns again correspond to the same three limitations: carbon (C), nitrogen (N) and phosphorus (P).

are expressed at levels specific to the carbon source across all limitations. There are also some limitation specific differences in mean expression levels, Fig.2.12.

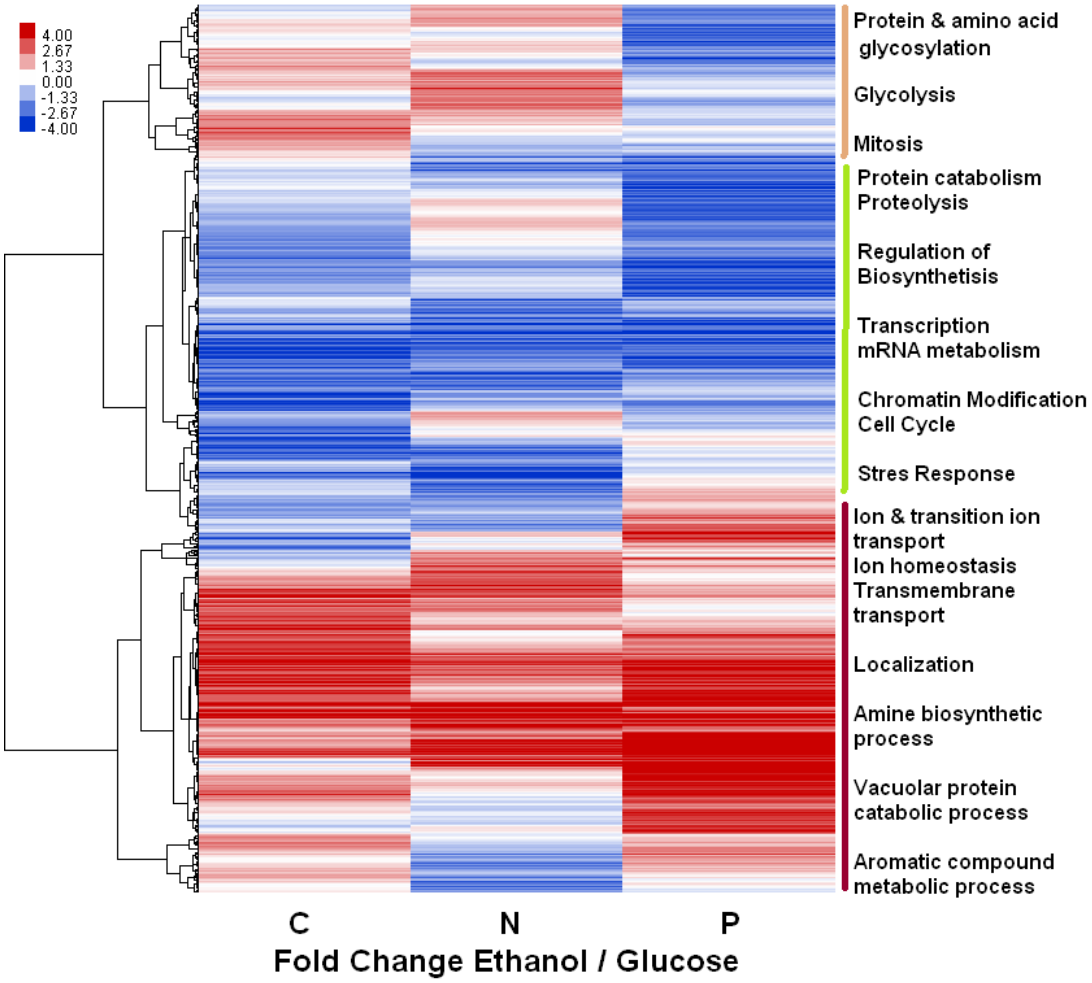

Figure 2.12: A clustergram of fold changes. The set of 3 columns corresponds to fold change between cultures growing on ethanol carbon source and culture grown on glucose carbon source and limited on carbon (C), nitrogen (N) and phosphorus (P). Each column corresponds to a limitation. The similarity metric (non–centered correlations) is computed using all data shown in the clustergram.

2.3.3 Clustergram of GO Terms

To summarize the mean slopes and fold changes between ethanol and glucose carbon source for genes from different functional groups, I plot a clustergram of GO terms, Fig.2.13. The clustergram is analogous to the one on Fig.2.11 except that the rows correspond to GO terms rather than individual genes.

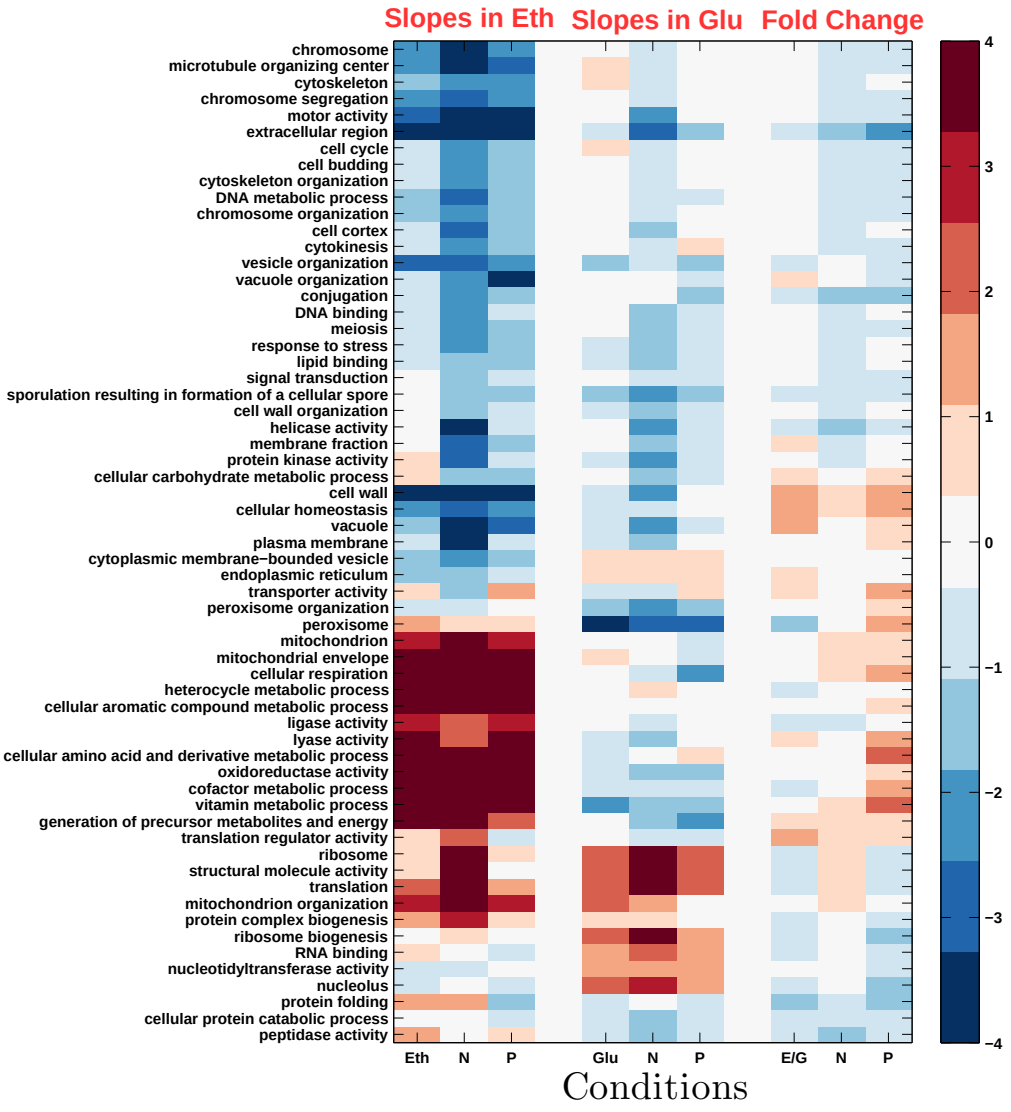

Figure 2.13: Slopes and Fold Change for GO Terms. The clustergram is analogous to the one on Fig.2.11 except that the rows correspond to GO terms rather than individual genes. The first set of 3 columns corresponds to slopes in cultures growing on ethanol carbon source and limited on carbon/ethanol (C), nitrogen (N) and phosphorus (P). The second set of 3 columns corresponds to slopes in cultures growing on glucose carbon source and limited on carbon/glucose (C), nitrogen (N) and phosphorus (P). The third set of 3 columns corresponds to fold change difference in expression levels between in cultures growing on ethanol carbon source and glucose carbon source. The columns again correspond to the same three limitations: carbon (C), nitrogen (N) and phosphorus (P). The similarity metric (non–centered correlations) is computed using all data shown in the clustergram.

2.3.4 Gene sets

The clustergram of GO terms provides a high–level comprehensive overview of the growth rate response of genes from various functional groups. To go beyond the high–level summary, I explore in more details the distributions of growth rate responses of the sets of genes defined based on the Gene Ontology and different expression levels in auxotrophic and prototrophic cultures. I start with functional groups whose growth rate response is expected and then more toward more unexpected results.

Gene set 1: **Mitochondrial envelope**

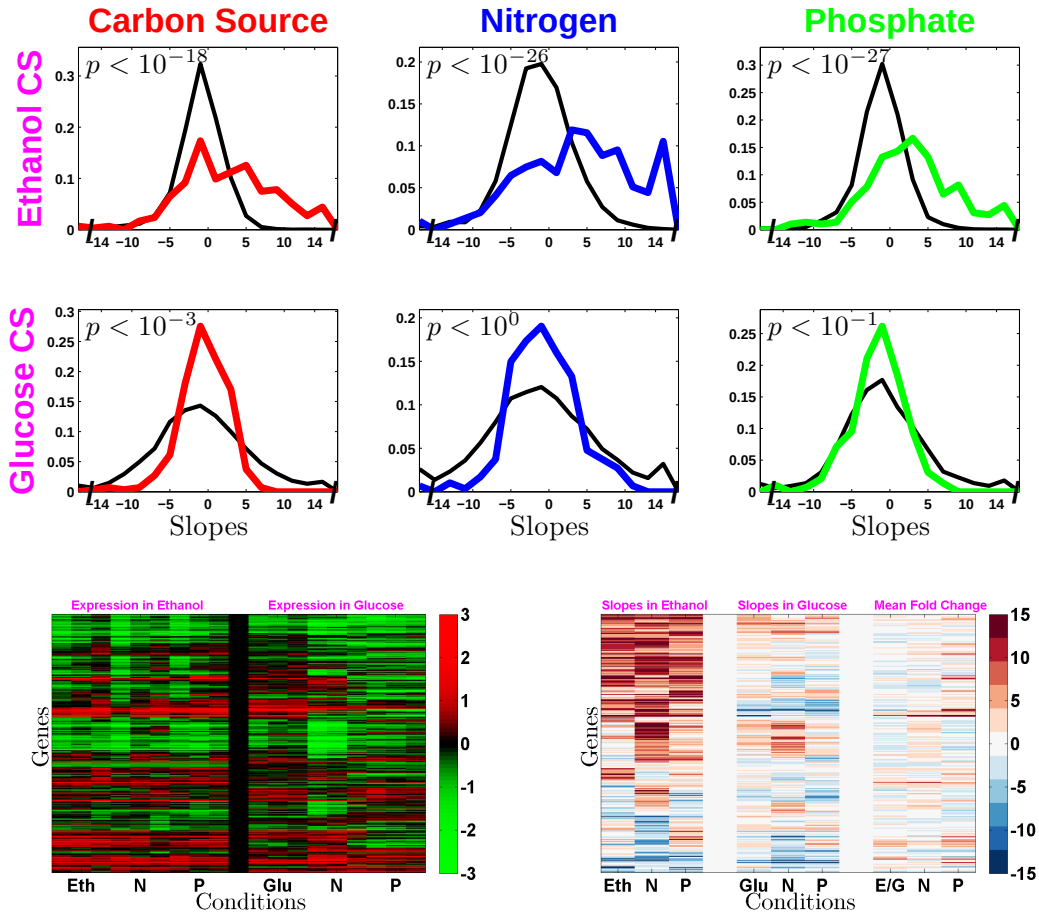

Figure 2.14: Mitochondrial envelope. Interactive Plots The top panel of distributions corresponds to ethanol carbon source and the bottom one to glucose carbon source. Each column corresponds to a limitation (carbon source, nitrogen and phosphate) as indicated on the top. The black distributions in each panel are for all genes in the genome and the colored distributions are for mitochondrial envelope genes only. The clustergrams display gene expression (left) and slopes & fold changes (right). The corresponding rows in the two clustergrams display data for the same genes and the clustering (permutation) is based on similarity metric (non–centered correlations) computed using *only* the slopes data (left panel).

Consistent with the requirement for increased aerobic–respiration at fast growth on ethanol carbon source, many mitochondrial genes have positive slopes. This trend is illustrated with the mitochondrial envelope genes, Fig.2.14. The genes with positive slopes are the same across all limitations on ethanol carbon source, first set of three columns in the clustergram of slopes. Interestingly a subset of the genes with positive

slopes in ethanol also have positive slopes in glucose carbon source. Once again the nitrogen limitations (middle columns in the first and second sets of columns of the slopes clustergram) are similar to each other and stand apart from the carbon and phosphate limitations.

Gene set 2: Cellular respiration

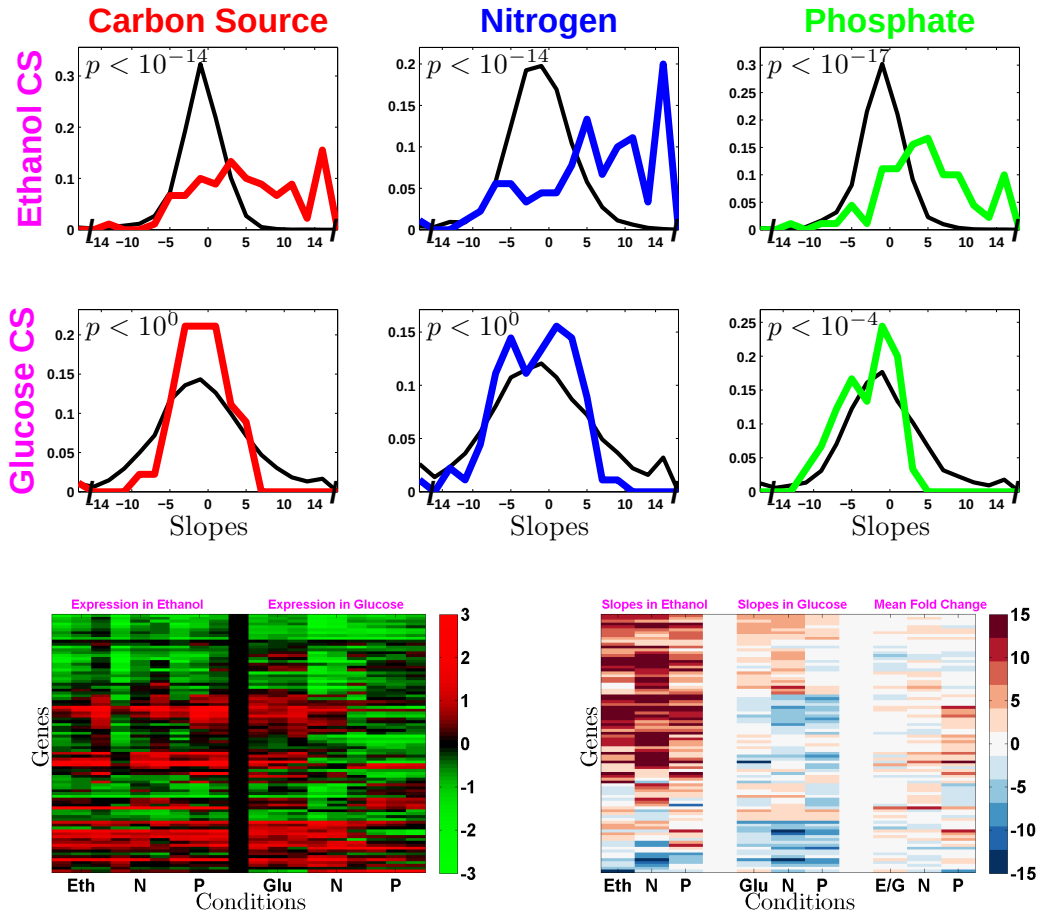

Figure 2.15: Cellular respiration Interactive Plots The top panel of distributions corresponds to ethanol carbon source and the bottom one to glucose carbon source. Each column corresponds to a limitation (carbon source, nitrogen and phosphate) as indicated on the top. The black distributions in each panel are for all genes in the genome and the colored distributions are for genes from the GO term generation of precursor metabolites and energy. The clustergrams display gene expression (left) and slopes & fold changes (right). The corresponding rows in the two clustergrams display data for the same genes and the clustering (permutation) is based on similarity metric (non–centered correlations) computed using *only* the slopes data (left panel).

Similar to gene set 6, growth rate related increase in the expression of cellular respiration genes is expected for cultures growing on ethanol carbon source.

Gene set 3: Generation of precursor metabolites and energy

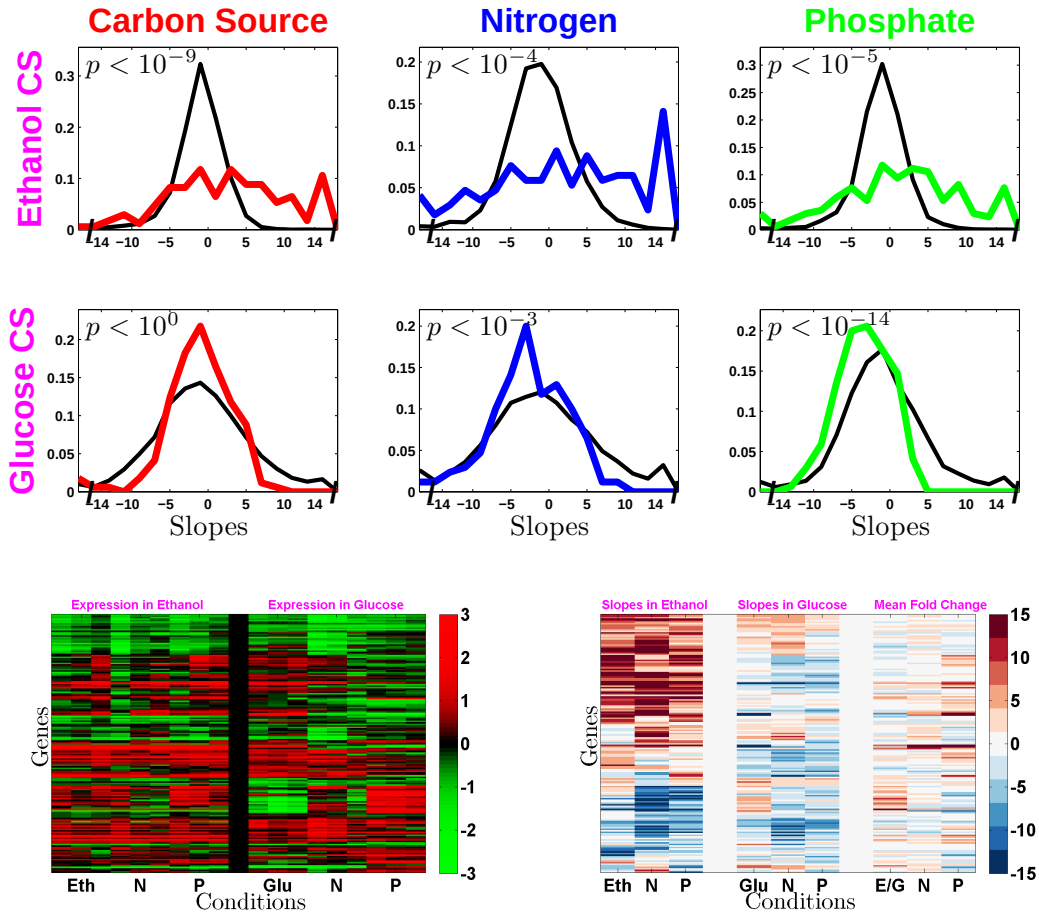

Figure 2.16: Generation of precursor metabolites and energy. See Interactive Plots. The top panel of distributions corresponds to ethanol carbon source and the bottom one to glucose carbon source. Each column corresponds to a limitation (carbon source, nitrogen and phosphate) as indicated on the top. The black distributions in each panel are for all genes in the genome and the colored distributions are for genes from the GO term generation of precursor metabolites and energy. The clustergrams display gene expression (left) and slopes & fold changes (right). The clustergrams display gene expression (left) and slopes & fold changes (right). The corresponding rows in the two clustergrams display data for the same genes and the clustering (permutation) is based on similarity metric (non–centered correlations) computed using *only* the slopes data (left panel).

Some of the genes in this set are mitochondrial such as the TCA cycle (Fig.2.31) which explains at least in part the significantly positive slopes of this set of genes. In fact given the central role of precursor metabolites and energy in growth the more surprising

funding is that this set of genes does not have significantly positive slopes on glucose carbon source.

Gene set 4: **Vacuole**

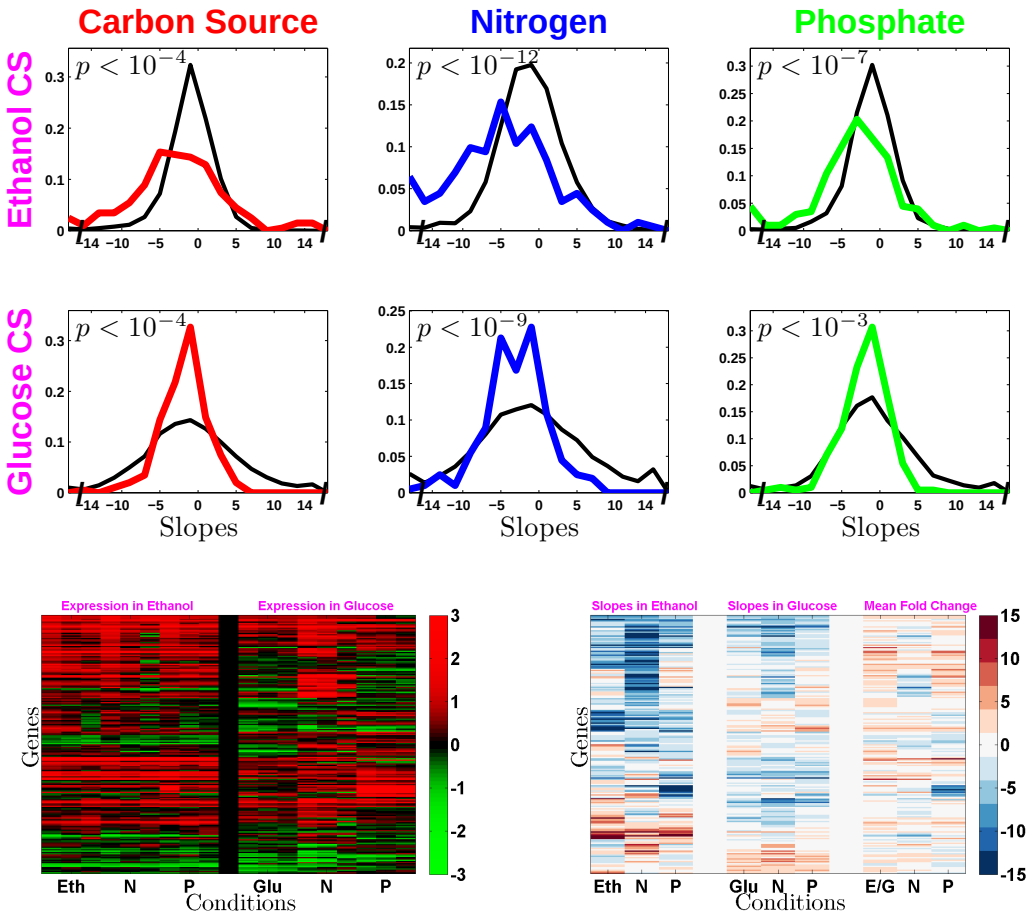

Figure 2.17: Vacuole. See Interactive Plots. The top panel of distributions corresponds to ethanol carbon source and the bottom one to glucose carbon source. The black distributions in each panel are for all genes in the genome and the colored distributions are for vacuolar genes only. Each column corresponds to a limitation (carbon source, nitrogen and phosphate) as indicated on the top. The clustergrams display gene expression (left) and slopes & fold changes (right). The clustergrams display gene expression (left) and slopes & fold changes (right). The corresponding rows in the two clustergrams display data for the same genes and the clustering (permutation) is based on similarity metric (non–centered correlations) computed using *only* the slopes data (left panel).

The vacuoles are part of the universal growth rate response, which most likely reflects the increased rate of recycling in slowly growing cells. The distribution of slopes for the vacuole genes is shifted significantly toward negative slopes (left) across all limitations on all carbon sources. This effect is particularly strong in the nitrogen limited cultures, middle column of distributions, Fig.2.17. Again the genes with negative slopes are the

same across limitations (as evident from the clustergram of slopes), especially the ones on ethanol carbon source.

Gene set 5: Peroxisome

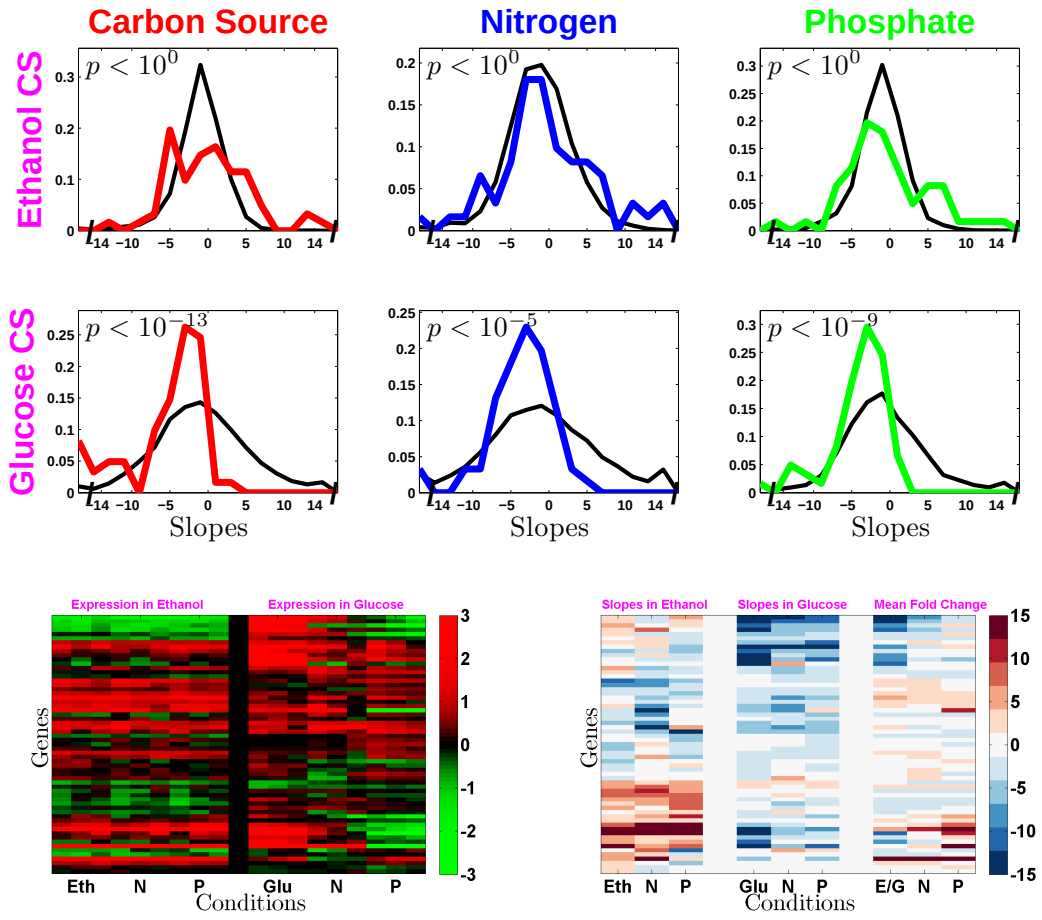

Figure 2.18: Peroxisome. Interactive Plots. The top panel of distributions corresponds to ethanol carbon source and the bottom one to glucose carbon source. Each column corresponds to a limitation (carbon source, nitrogen and phosphate) as indicated on the top. The black distributions in each panel are for all genes in the genome and the colored distributions are for peroxisomal genes only. The clustergrams display gene expression (left) and slopes & fold changes (right). The corresponding rows in the two clustergrams display data for the same genes and the clustering (permutation) is based on similarity metric (non–centered correlations) computed using *only* the slopes data (left panel).

As noted by Brauer *et al* (2008), peroxisomal genes are overrepresented among the genes with negative slopes as revealed here by the distribution of slopes for peroxisomal genes being shifted significantly to the left for all limitations on glucose carbon source. Interestingly, this is not the case for glucose carbon source, top row of distributions, Fig.2.18. In part, this difference can be explained by the fact that gluconeogenesis genes

(catalyzing reactions from the glyoxylate cycle) are localized in the peroxisomes and expressed at high levels and with positive slopes on ethanol carbon source.

Gene set 6: **Cofactor metabolic process**

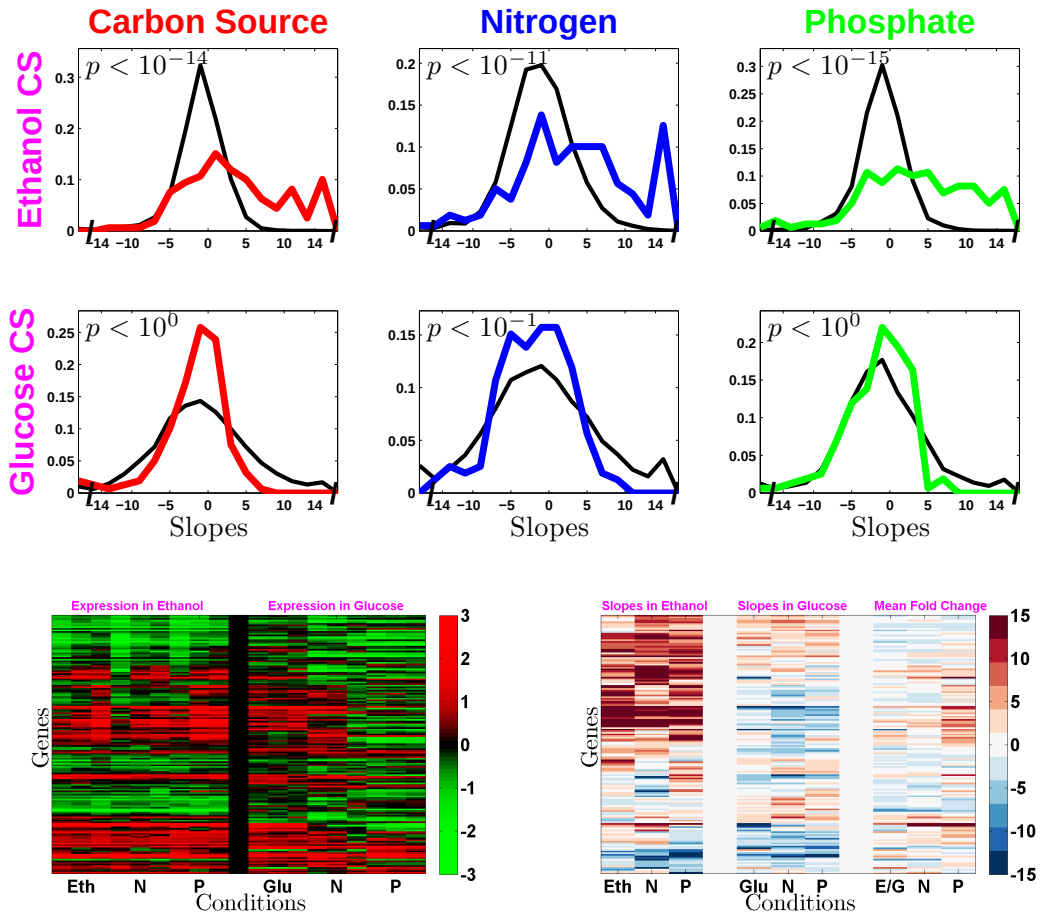

Figure 2.19: Cofactor metabolic process. Interactive Plots. The top panel of distributions corresponds to ethanol carbon source and the bottom one to glucose carbon source. Each column corresponds to a limitation (carbon source, nitrogen and phosphate) as indicated on the top. The black distributions in each panel are for all genes in the genome and the colored distributions are for cofactor metabolic process only. The clustergrams display gene expression (left) and slopes & fold changes (right). The clustergrams display gene expression (left) and slopes & fold changes (right). The corresponding rows in the two clustergrams display data for the same genes and the clustering (permutation) is based on similarity metric (non–centered correlations) computed using *only* the slopes data (left panel).

A diverse group of genes involved in the synthesis of cofactors or using cofactors have significantly positive slopes in ethanol carbon source across all limitations. A subset of those genes also have modestly positive slopes in glucose across all limitations but as a whole for glucose carbon source the slopes of this gene set is not significantly different from the slopes for all genes.

Gene set 7: Microtubule organizing center

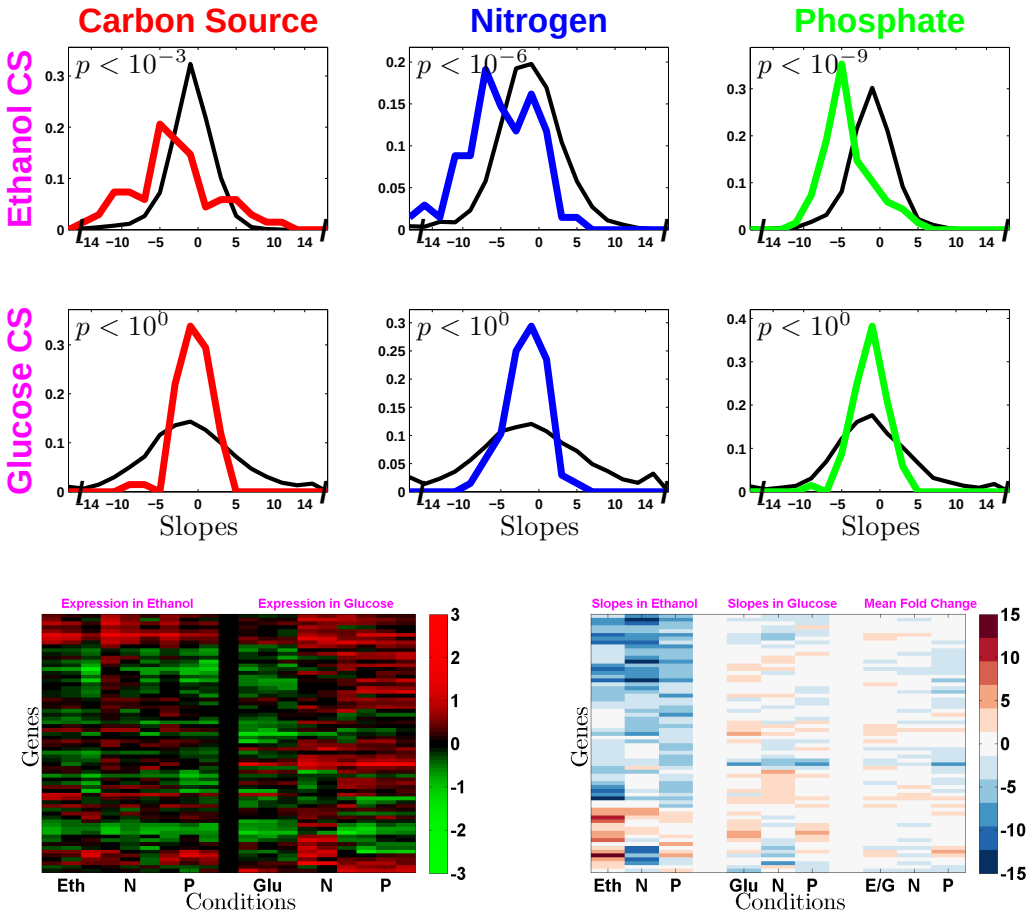

Figure 2.20: Microtubule organizing center. See Interactive Plots. The top panel of distributions corresponds to ethanol carbon source and the bottom one to glucose carbon source. Each column corresponds to a limitation (carbon source, nitrogen and phosphate) as indicated on the top. The black distributions in each panel are for all genes in the genome and the colored distributions are for microtubule organizing center genes only. The clustergrams display gene expression (left) and slopes & fold changes (right). The clustergrams display gene expression (left) and slopes & fold changes (right). The corresponding rows in the two clustergrams display data for the same genes and the clustering (permutation) is based on similarity metric (non–centered correlations) computed using *only* the slopes data (left panel).

As evident from *gene 7*, some genes with cell–cycle function have negative slopes. Looking at the slope distributions of cell–cycle related genes indicates that many other indeed gene sets for many functions related to mitosis and cell division have significantly negative slopes as exemplified here by the microtubule organizing center Fig.2.20. Such functions include DNA replication, chromosome segregation and cell–cycle regulated

genes. The clustergram of slopes (Fig.2.20) indicates that the genes with negative slopes are the same across all limitations on ethanol carbon source. This finding is surprising in the context of bud index linearly correlated to growth rate, subsection 1.4.2, Fig.1.6. The most probable explanation (consistent with all data) that I know of involves the YMC and is discussed in chapter 4.

Gene set 8: **Heterocycle metabolic process**

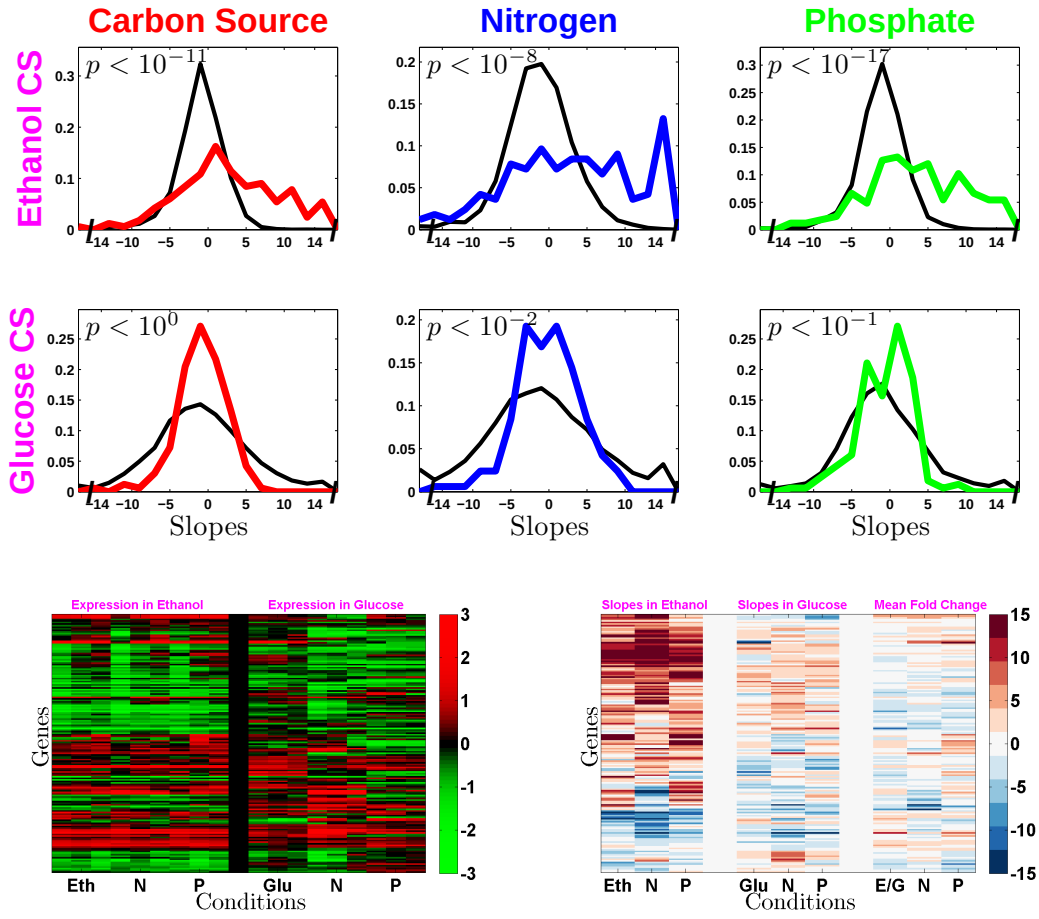

Figure 2.21: Heterocycle metabolic process. Interactive Plots. Notation is the same as in Fig.2.19 with colored distributions corresponding to the slopes of genes involved in heterocycle metabolic process.

This set includes gene participating in the biosynthesis of histidine, purines, pyrimidines, thiamine, and other heterocyclic compounds. Since all these compounds are required for biomass production and growth its not surprising to see the growth rate induction of the genes involved in their biosynthesis. The interesting observation is that they are not induced or induced to a much smaller extent in cultures growing on glucose carbon source. I may only speculate about possible reasons. Two prominent possibilities are:

1. In glucose carbon source, the abundance of metabolic intermediates is such that it does not require the same level of enzyme induction. Since I measured very few metabolites, this hypothesis is hard to evaluate on the basis of my metabolic data.
2. In glucose carbon source much of the regulation occurs at post-transcriptional level.

Gene set 9: **Vitamin metabolic process**

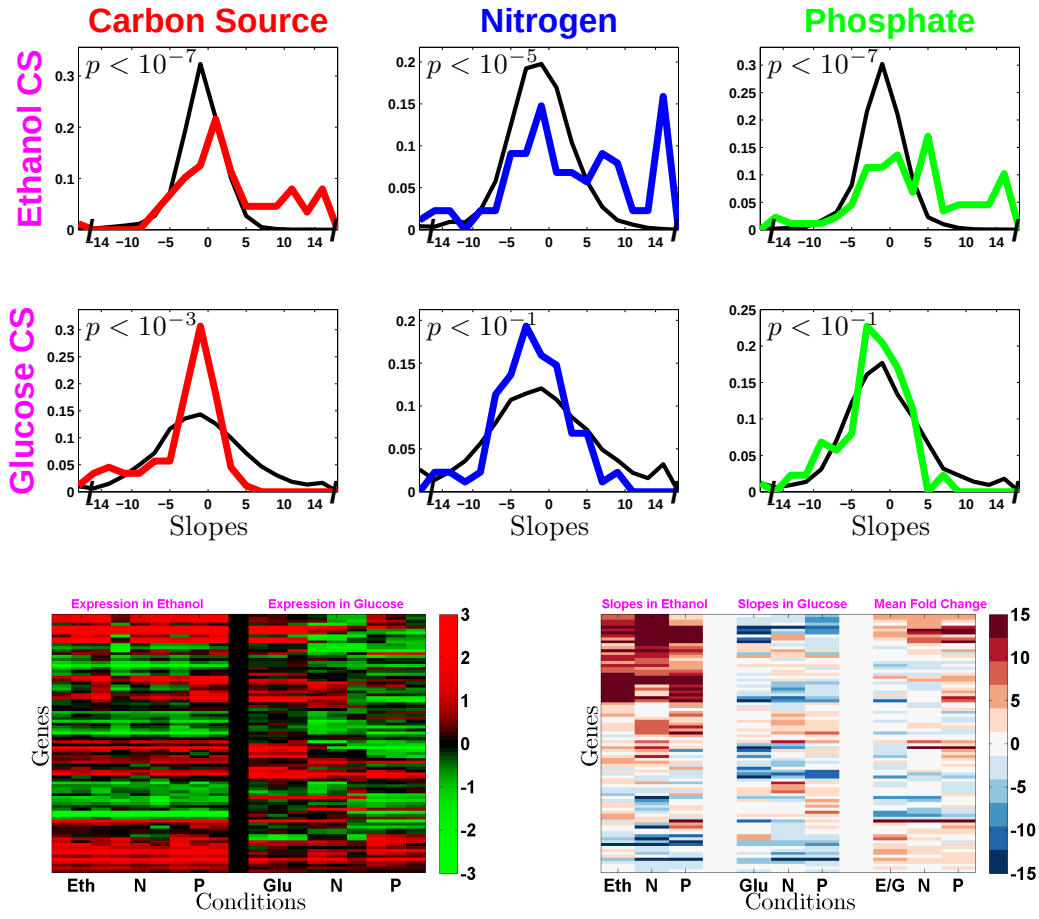

Figure 2.22: Vitamin metabolic process Interactive Plots Notation is the same as in Fig.2.19 with colored distributions corresponding to the slopes of genes involved in heterocycle metabolic process.

Many of the vitamin related genes in this group are cofactors for mitochondrial enzymes that show transcriptional induction themselves, such as the enzymes catalyzing the TCA cycle, see Fig.2.31. Thus up–regulation of gene from this set is another indicator of growth rate induced increase in the TCA cycle and related biochemical reactions involved in the production of energy and intermediate metabolites.

Gene set 10: Oxidoreductase activity

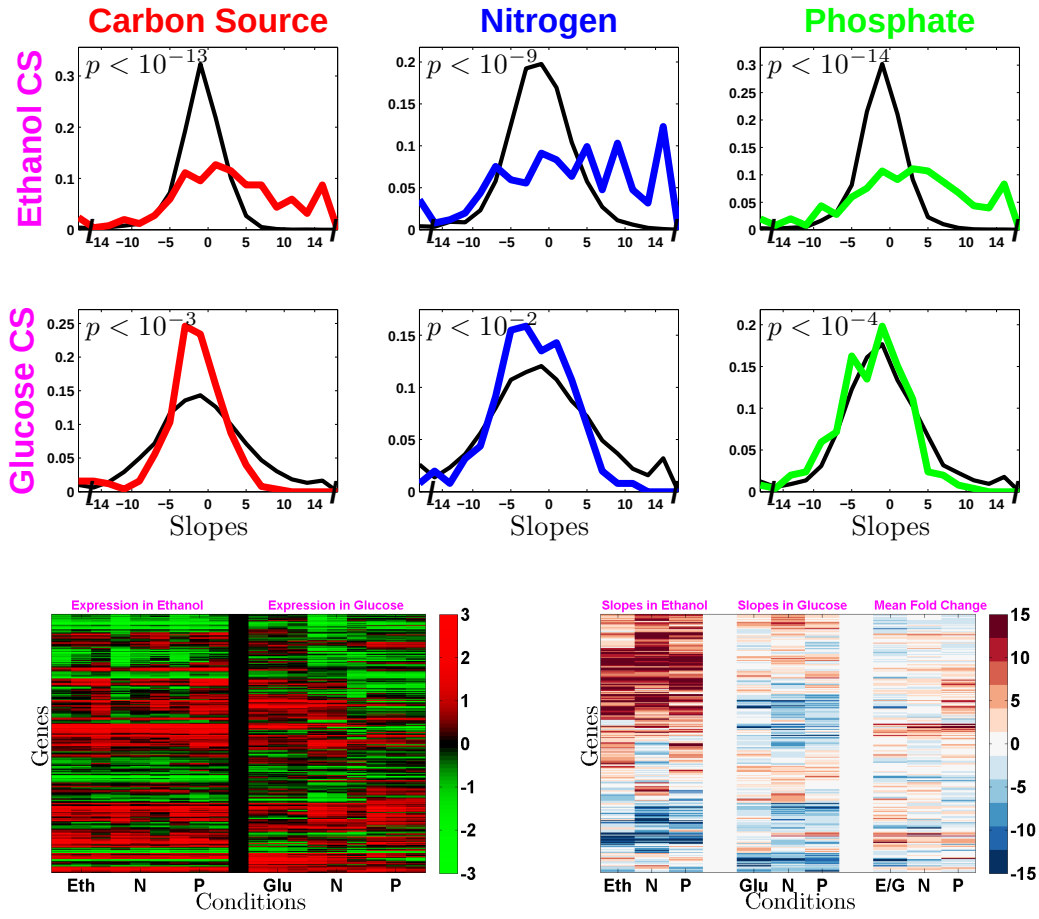

Figure 2.23: Oxidoreductase activity Interactive Plots Notation is the same as in Fig.2.19 with colored distributions corresponding to the slopes of genes involved in oxidoreductase activity.

Similar to some of the previous gene sets, many of the genes in this set that have positive slopes on ethanol carbon source are related to mitochondria, which reinforces the observation that most mitochondria related genes have positive growth rate response on ethanol carbon source.

Gene set 11: Auxotrophic starvation & cell–division

I first explore the growth rate response of a gene set (11) identified on the basis of different expression levels between prototrophs and auxotrophs growing on glucose carbon source, see Section **??**. The genes in set 1 have significantly ($p < 10^{-10}$) lower expression in auxotrophs compared to prototrophs, Fig.2.24. The difference in expression levels is

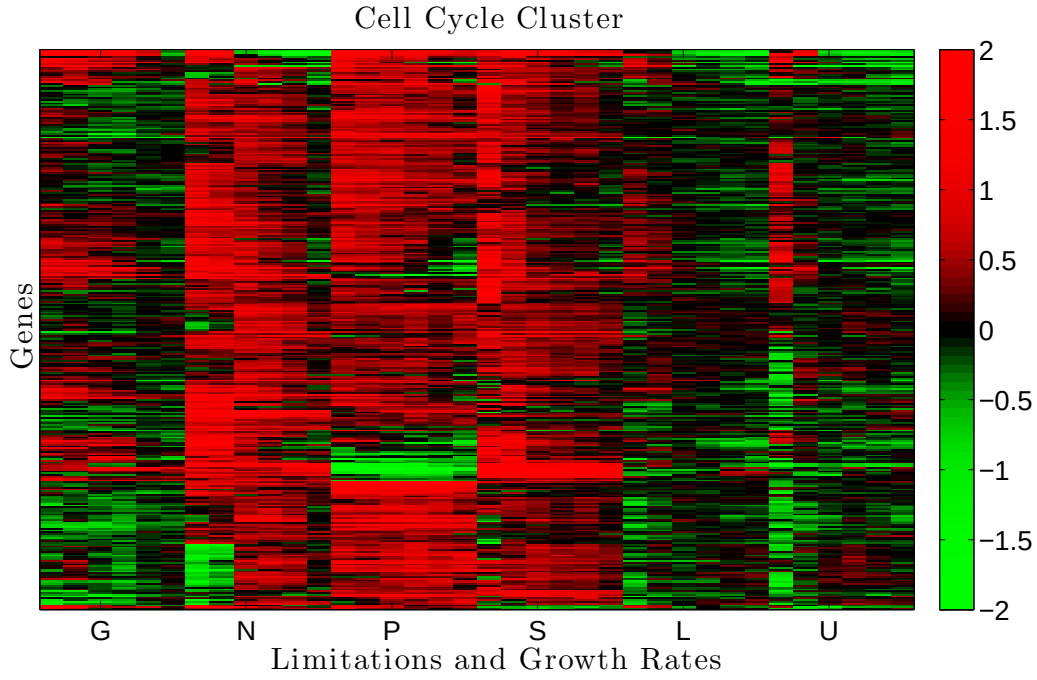

Figure 2.24: Gene expression profiles for genes in gene set 11. All cultures were grown on glucose carbon source and were limited on the nutrients indicated on the x–axis. Each nutrient limitation has 6 growth rate plotted in ascending order. Data from (Brauer *et al*, 2008).

also depicted in the form of distributions of fold changes and compared to the distribution of fold changes for ethanol carbon source, Fig.2.25. Based on analysis with the GO Term Finder, functions related to cell–division are strongly overrepresented ($p < 10^{-13}$) by genes in set 11. Such functions include chromosome organization, microtubule spindle, DNA replication, DNA repair, DNA packing, mitotic cycle, cell cycle, and M phase. In addition, genes in set 11 overlap significantly ($p < 10^{-8}$) with the targets of transcription

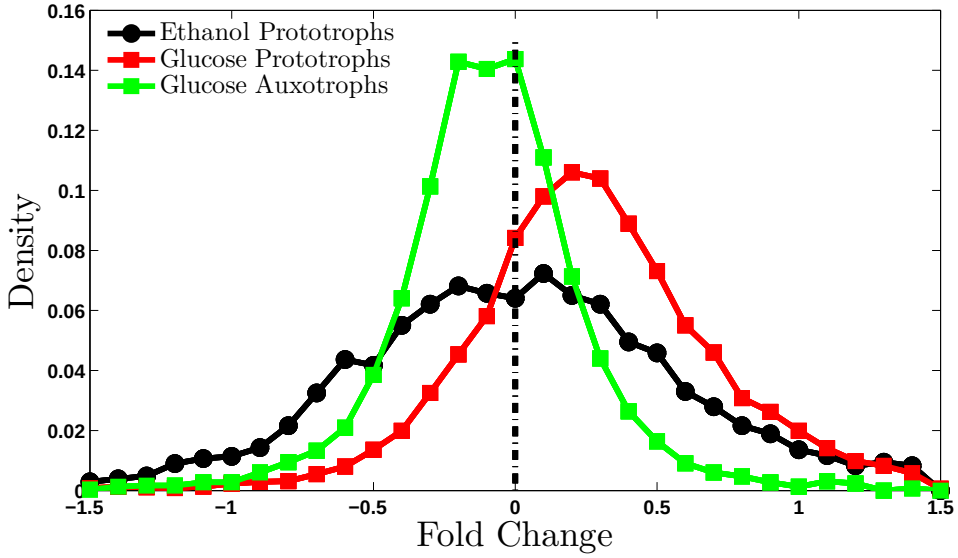

Figure 2.25: Distributions of fold changes for genes in gene set 11

factor *MBP1* identified by ChIP–chip studies (Harbison *et al*, 2004; MacIsaac *et al*, 2006), see section 2.5. Furthermore, set 11 genes have significantly negative slopes on ethanol carbon source across all limitations, Fig.2.26.

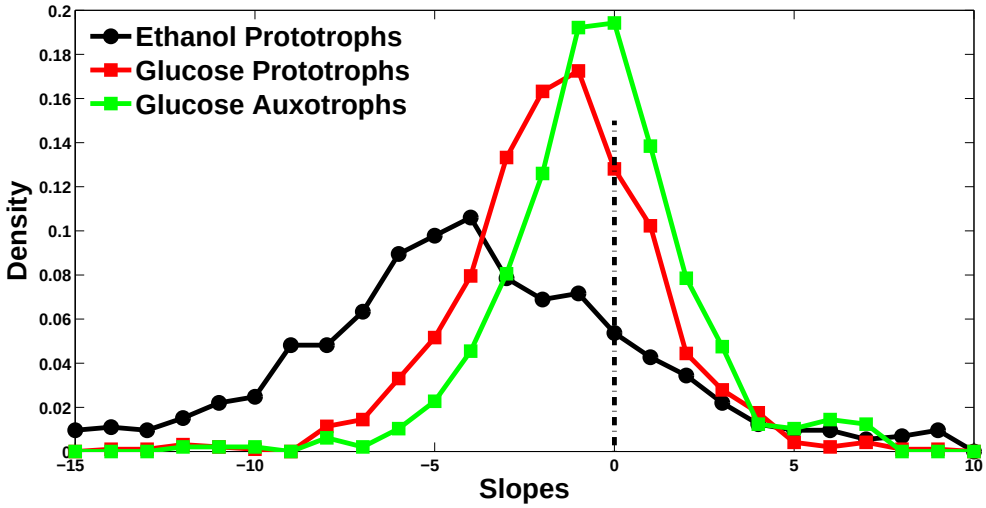

Figure 2.26: Distributions of slopes no different carbon sources and limitations for genes in gene set 11

2.4 *GRR* of Well Characterized Pathways

This section details transcriptional changes in mRNAs corresponding to enzymes catalyzing key biochemical reactions whose fluxes are expected to change significantly with growth rate, nutrient limitations and the carbon source.

2.4.1 Ethanol Utilization

The first reaction in ethanol utilization is its oxidation to acetaldehyde by alcohol dehydrogenases. Yeast has *5* isoenzymes catalyzing this reaction, *ADH1* to *ADH5*, Fig.2.27. The isoenzyme showing strongest induction (about 30 fold) on ethanol carbon source

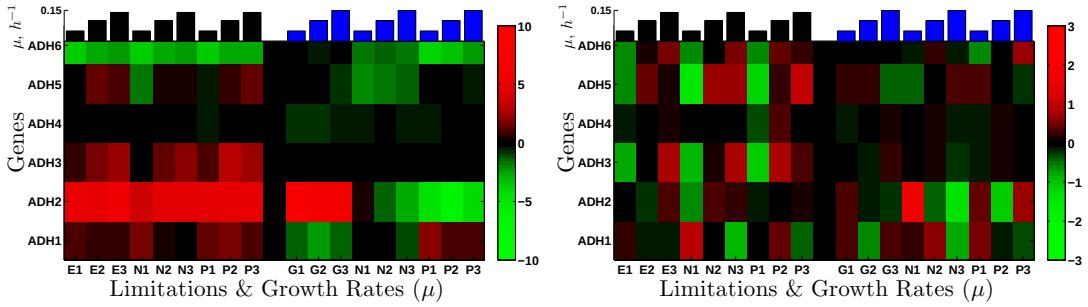

Figure 2.27: Expression levels of mRNAs coding for alcohol dehydrogenases in ethanol (left, black bars) and glucose (right, blue bars) carbon source. The growth rates are plotted as bars on the top and also denoted on the *x*–axis with number *1,2,3* corresponding to growth rates $0.05, 0.10, 0.14/0.15h^{-1}$. The letters correspond to limitations as follows: *E*– ethanol; *G*–glucose; *N*–nitrogen; *P*–phosphate; Left panel is the fold change gene expression data. The right panel is the data normalized to zero mean for each limitation to emphasize growth rate trends.

across all nutrient limitations is *ADH2* which is known to be the isoenzyme catalyzing the oxidation of ethanol. This indicates that at least some of the up regulation required for ethanol catabolism is accomplished by increasing the concentration of the *ADH2* mRNA which likely is reflected in making more enzyme as well. Interestingly, *ADH2* expression is induced equally strongly in the *Glu* limited cultures while repressed in P and N limited cultures using *Glu* as a carbon source. This expression pattern suggests

that *ADH2* expression is more likely repressed by glucose rather than induced by ethanol. *ADH3* is also induced in ethanol carbon source (about 2-4 fold) with positive slopes in all nutrient limitations, which likely reflects shuffling of *NADH* from the mitochondria to the cytoplasm that increases with the growth rate.

The second reaction in ethanol utilization is oxidation of acetaldehyde to acetate which is catalyzed by aldehyde dehydrogenase for which there are also *5* isoenzymes, *ALD2* to *ALD6*, Fig.2.28. The isoenzymes induced most strongly are *ALD2*, *ALD3* and *ALD6*. The

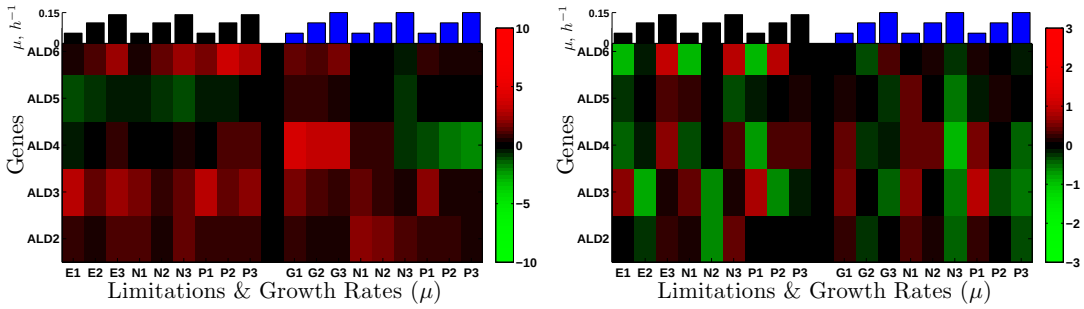

Figure 2.28: Aldehyde dehydrogenases. Notation as in Fig.2.27

first two are known to be repressed by glucose and induced by ethanol and stress and the latter is unique among the aldehyde dehydrogenases in using $NADP^+$ instead of NAD^+ as a cofactor. This cofactor specificity is crucially important in the context of growth as reduced $NADP^+$ (*NADPH*) is required in many biosythetic reactions but produced by only a few reactions. In growth on ethanol, the only other reaction generating *NADPH* reducing power is isocitrate dehydrogenase *ADP2* whose mRNA is also very strongly induced across all limitations on ethanol carbon source. Consistent with expectations for increasing demand for *NADPH* with increasing growth grate, *ALD6* (and *ADP2*) have positive slopes on ethanol carbon source, Fig.2.28.

The third reaction in ethanol utilization is the transfer to *CoA* to acetate. This transfer might be catalyzed by acetyl-CoA synthetases, which uses *CoA* and *ATP* as substrates or by *CoA* transferase which transfers *CoA* from succinyl–CoA to acetate, *ACH1*. Yeast has

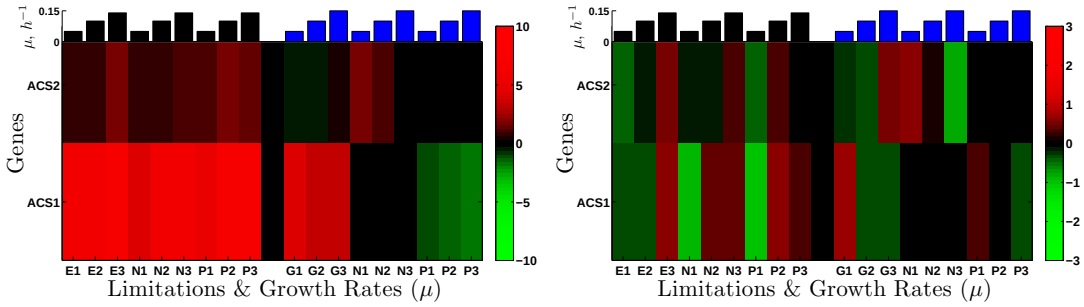

Figure 2.29: Acetyl-CoA synthetases. Notation as in Fig.2.27

two acetyl-CoA synthetases, *ACS1* and *ACS2*. *ACS1* has a positive slope and is induced very highly (up to 120 fold) across all limitations on ethanol carbon source and only in glucose limitation (to a slightly smaller extent) on glucose carbon source in a manner analogous to *ADH2*. *ACS2* shows only modest induction (up to 2-3 fold) with a slightly positive slope, Fig.2.29. The acetyl-CoA transferase *ACH1* is also induced highly, up to 30 fold.

The next step in ethanol catabolism is transporting acetyl–CoA across the mitochondrial and peroxisomal membranes. Major players in this process are the carnitine acetyl-CoA transferases. *CAT2*, *YAT1* and *YAT2*, Fig.2.30. All 3 genes show very strong up regulation (up to about 250 fold) across all limitations on ethanol carbon source and very large positive slopes (up to 40) except for *YAT1* which is induced highly even at slow growth.

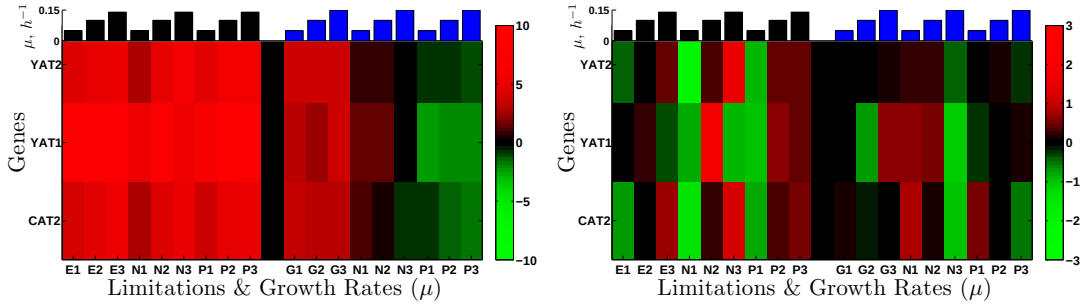

Figure 2.30: Acetyl and acyl transferases. Notation as in Fig.2.27

In summary, all biochemical reactions involved in the oxidation of ethanol to its inclusion in the *TCA* are catalyzed by enzymes whose corresponding mRNAs are up regulated on ethanol carbon source and most of them also have positive growth rate response (e.g positive slopes). This observation indicates that transcriptional regulation plays a role in in the utilization of ethanol and growth rate control.

2.4.2 Central Carbon Metabolism

Krebs cycle

Once in the mitochondria, acetyl–CoA is fed into the *TCA* and oxidized for energy or used for generating intermediates for anabolic processes. Since *TCA* is a major hub through which ethanol has to pass before it can be used for energy or building blocks, the flux through the *TCA* is expected to increase with growth rate on ethanol carbon source. Consistent with this expectation, the slopes of mRNAs corresponding to *TCA* enzymes are overwhelmingly positive, Fig.2.31 and 2.32.

Glyoxylate Cycle & Gluconeogensis

Yeast growing on ethanol must synthesize glucose (gluconeogensis) for the essential carbohydrates, glycosylated proteins and of course as an intermediate in the synthesis of pentose and deoxypentose for nucleic acids. The only known pathway for making glucose from acetyl–CoA is the glyoxylate cycle. The first biochemical reaction separating the glyoxylate cycle from *TCA* is isocitrate lyase (*ICL1* and *ICL2*) catalyzing the formation of succinate and glyoxylate from isocitrate. Consistent with expectation, *ICL1* is strongly induced on ethanol carbon source (Fig.2.33) while *ICL2* to a smaller extent. Remarkably, both isocitrate lyases are also induced in the glucose limitation suggesting that it is the

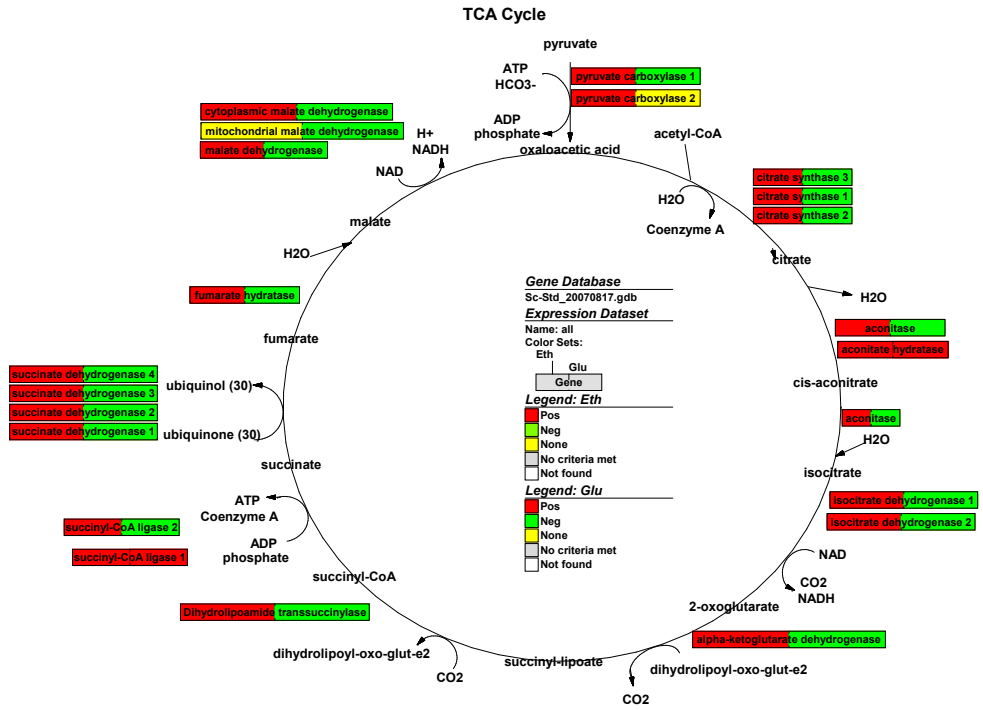

Figure 2.31: *TCA*. The slopes for each isoenzyme are colorcoded on the box enclosing its name. The first half of the box displays the slope on ethanol carbon source and the second half the slope on glucose carbon source.

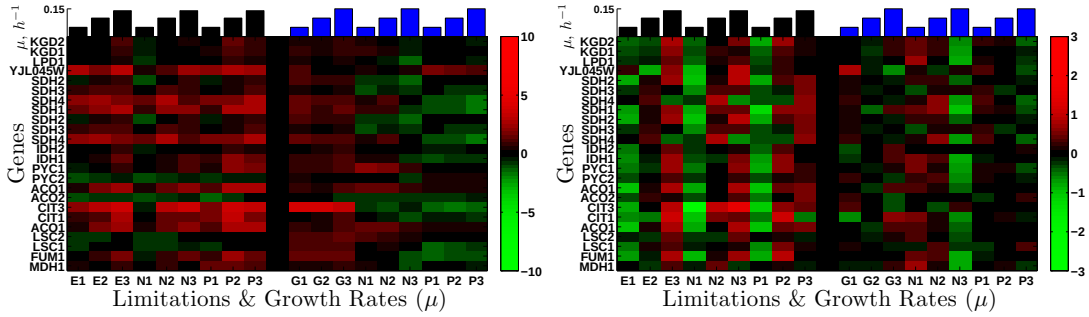

Figure 2.32: Expression of *TCA* mRNAs. Notation as in Fig.2.27

lack of glucose rather than the presence of ethanol that results in the increased expression levels. This mechanism of glucose repression rather than ethanol induction can be rationalized by realizing that a single repression mechanism can accomplish the regulation that otherwise might require many induction mechanism, e.g ethanol, glycerol, acetate

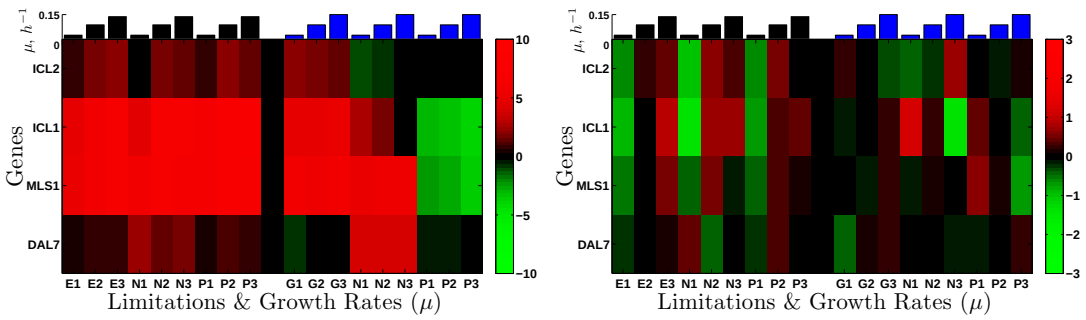

Figure 2.33: Glyoxylate Cycle. Notation as in Fig.2.27

and pyruvate induction mechanisms. The next pair of isoenzymes in the glyoxylate cycle (*MSL1*) and (*DAL7/MSL2*) catalyze the synthesis of malate from the acetyl-CoA and the glyoxylate produced in the first reaction. Those two malate synthases have elevated levels in ethanol carbon source, especially *MSL1* whose expression pattern is similar to *ICL1*. Both malate synthases and *ICL1* show nitrogen derepression in the ammonium limited cultures on glucose carbon source which most likely reflects the role of those enzymes in nitrogen salvage from purine catabolism. Some of growth rate slopes of mRNAs corresponding to the glyoxylate cycle enzymes are slightly positive (Fig.2.33) but they are relatively small compared to large differences in the expression of those mRNAs between ethanol and glucose carbon source. This observation is consistent with the possibility that regulation of the glyoxylate cycle is largely a function of glucose repression.

Discussion & Conclusion

Our understanding of the many biochemical reactions involved in nutrient catabolism makes robust predictions about the reactions whose metabolic flux should change significantly and it is interesting to explore the isoenzymes whose levels change significantly and are thus likely to mediate the expected changes in metabolic fluxes. The level at which those changes happen, however, is often not known. The flux of a reaction

can increase as a consequence of changing the concentrations of the reactants and the products and/or changing the activity of the enzyme catalyzing the reaction. Such change in enzyme activity may come from posttranslational modifications of the enzyme, changing its localization or changing its amount which can be mediated either by change in transcription or change in the concentration in the level of the corresponding mRNA. Therefore, even for the best studied biochemical reactions the changes in the mRNAs levels of the corresponding enzymes cannot be predicted without experimental data. Thus, I use mRNA data to characterize the level at which the flux changes and the isoenzymes that are likely to catalyze the reactions. Remarkably, I find that most expected changes in biochemical fluxes are reflected in changes of mRNA levels

2.5 Regulation of Growth Rate Response

In the preceding sections, I identified a set of about 1500 genes with universal growth rate response and many other sets of genes with differential carbon source and limitation dependent growth rate responses. The molecular interactions and regulatory mechanisms (Slavov, 2012) underlying the expression patterns of these genes, however, are not well characterized. Such characterization is the focus of this section. The data I have collected allows me to infer regulatory interactions primarily at the level of transcription and mRNA degradation. In this section I use methods that allow only the inference of TFs but in the next chapter I apply *RCweb* (Slavov, 2010) which allow me to infer mRNA degradation as well.

2.5.1 Overlap between Gene Sets and TF Targets

One of simplest approaches to identifying transcription factors (TF) that might underly the growth rate response is to compute the overlap between a set of genes defined to have a type of growth rate response (such as universally upregulated with growth rate) and the targets of a TF as identified independently from ChIP–chip (Harbison *et al*, 2004) and other experiments (MacIsaac *et al*, 2006). Given a TF with n targets, a set of m growth rate response genes selected out of N genes, the probability of observing an overlap of k genes by chance alone is given by the cumulative mass function of the hypergeometric distribution, Eq.2.2:

$$P(X \geq k) = \sum_{i=k}^{i=\infty} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} = 1 - \sum_{i=0}^{i=k-1} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \tag{2.2}$$

The probability for seeing by chance alone the observed overlap between a set of m genes and the targets of the i^{th} TF is the *p value* for the hypothesis that the i^{th} TF contributes to regulating the expression of the set of m genes. This approach has been used by

Fazio *et al* (2008) in the context of growth rate response. All results presented here are computed by a perl script, which I wrote and is available upon request.

The chief advantages of this approach are its simplicity and the avoidance of inaccurate assumptions typical to most other approaches. One of the most frequent assumption is approximating TF activity with the level of its corresponding mRNA (Segal *et al*, 2003) and I am going to use the results from the TF–targets overlap to assess the number of TF for which this assumption might be applicable. In particular, I will record whether a TF identified by the TF overlap method has an expression profile similar to its targets. I do this using a very lenient criterion: For each TF whose targets overlap significantly with a set of genes with a growth rate response, I record whether the mRNA corresponding to the TF belongs to that set of growth rate responsive genes or not, last columns of Tables 2.3 through 2.10.

The TF–targets overlap approach also has, however, a number of disadvantages including all problems discussed in the Methodology subsection 2.3.1. Furthermore, this approach is likely to have very large number of false negatives (it lacks power) because of the incomplete knowledge of the transcriptional network. This lack of power is most severe for the ethanol carbon source since the experimental conditions used in the ChIP–chip experiments do not include ethanol carbon source. Nonetheless, the rate of false positives is likely to be very low: A very low *p value* for the overlap of the targets of the i^{th} TF and the j^{th} gene set indicates high probability that the i^{th} TF regulates the j^{th} gene set. Other disadvantages include the inability to detect combinatorial interactions, make a quantitative model and derive experimentally testable predictions.

Universal growth rate response

Tables 2.3 and 2.4 list the results from applying the TF–targets overlap approach to the genes that fit the growth rate model from Section 2.2 best (highest R^2) and decrease in expression with growth rate, that is have negative growth rate slopes. I used the

TF Name	Targets	Set	Overlap	p val	Reg
HSF1	43	807	12	10^{-2}	Yes
STB4	10	807	5	10^{-3}	No
MSN2	89	807	29	10^{-6}	No
SUT1	68	807	17	10^{-3}	No
MSN4	73	807	21	10^{-4}	No
SKN7	147	807	35	10^{-4}	No

Table 2.3: TF regulating the negative growth rate response. The first column is the name of the TF (hyper–linked to *SGD*) The second column is the number of TF targets present in the MacIsaac *et al* (2006) results at $p \leq 0.001$ and conservation level 1. The third column is the number of input genes with universal positive growth rate response, which is dependent on a threshold. The fourth column is the number of genes common between the two sets (second and third columns). The forth column is the corresponding *p value* computed from Eq.2.2. The last column indicates whether the mRNA corresponding to the TF is present in the input set of genes that are regulated by the TF. The same designations and notations are used for all tables in this subsection.

TF Name	Targets	Set	Overlap	p val	Reg
HSF1	43	1127	17	10^{-3}	Yes
SIP4	10	1127	5	10^{-2}	No
STB4	10	1127	5	10^{-2}	No
MSN2	89	1127	32	10^{-4}	No
SUT1	68	1127	23	10^{-3}	No
MSN4	73	1127	27	10^{-4}	Yes
SKN7	147	1127	48	10^{-5}	No

Table 2.4: TF regulating the negative growth rate response. The columns and notation are the same as in table 2.3

TF targets published by (MacIsaac *et al*, 2006) at $p \leq 0.001$ and conservation level 1 for both tables and throughout this section. In table Fig.2.3 I used the top 807 genes with negative slopes while in table Fig.2.4, I used the top 1127 genes. For some genes the results do not change substantially, while for others they do change underscoring a problem discussed in the Methodology subsection 2.3.1. Based on these results the TFs

for which there is strong evidence for mediating the universal growth rate response are *MSN2/MSN4* and *SKN7* with weaker evidence for other TFs in the tables. For growth on glucose carbon source in both aerobic and anaerobic conditions, Fazio *et al* (2008) also found *SKN7* but not *MSN2/MSN4*. A possible reason why Fazio *et al* (2008) did not find *MSN2/MSN4* might be that those TFs are not involved in the growth rate response in anaerobic conditions. Given the strong overrepresentation of stress genes among the the genes with negative slopes, *MSN2/MSN4* are likely TFs to regulate the negative growth rate response and indeed they are also found by FIRE (Elemento *et al*, 2007). All TFs found by this analysis to be involved in the negative growth rate response are stress response related TF, Table 2.4.

The TFs likely to mediate the positive growth rate response (increasing mRNA abundance with growth rate) are listed in Tables 2.5 and 2.6 for two different levels of significance (R^2 thresholds). A prominent TF among those is *RAP1* which was also found on

TF Name	Targets	Set	Overlap	p val	Reg
BAS1	34	766	11	10^{-3}	No
ABF1	241	766	47	10^{-3}	No
SFP1	32	766	12	10^{-4}	No
GAT3	9	766	5	10^{-4}	No
FHL1	143	766	50	10^{-12}	No
RAP1	105	766	31	10^{-6}	No

Table 2.5: TF regulating the universal positive growth rate response. The columns and notation are the same as in table 2.3

TF Name	Targets	Set	Overlap	p val	Reg
BAS1	34	1089	14	10^{-3}	No
ABF1	241	1089	60	10^{-3}	No
SFP1	32	1089	14	10^{-4}	No
GAT3	9	1089	5	10^{-3}	No
FHL1	143	1089	58	10^{-10}	No
RAP1	105	1089	38	10^{-6}	No

Table 2.6: TF regulating the positive growth rate response. The columns and notation are the same as in table 2.3

glucose carbon source by FIRE (Airoldi *et al*, 2009; Elemento *et al*, 2007), by Regenberg *et al* (2006), and by Fazio *et al* (2008). *RAP1* is a DNA–binding protein involved in either activation or repression of transcription, depending on binding site context. It also binds telomere sequences and plays a role in telomeric position effect (silencing) and telomere structure. *FHL1* and *SFP1* are TFs related to ribosomal biogenesis and RNA processing and also identified by Fazio *et al* (2008) but not by FIRE (Airoldi *et al*, 2009; Elemento *et al*, 2007). Given the overrepresentation to ribosomal and translation related genes among the gene set with universal growth rate response, the involvement of *FHL1* and *SFP1* is not surprising. The other TFs in tables 2.5 and 2.6 are interesting but given the weak evidence for their involvement the results are rather inconclusive.

Specific Growth Rate Response

In addition to the universal growth rate response, some sets of genes have a growth rate response specific to some limitations and or to the carbon source. First consider the growth rate response specific to ethanol carbon source. There is a high positive correlation ($+0.75$) between the slopes in ethanol and phosphate limited cultures growing on ethanol carbon source for the genes whose expression levels fit the exponential (linear in semi–log space) model ($R^2 \geq 0.85$) in those conditions, Fig.2.34. An interactive plot, hyper–linked to the data can be found at my growth rate website. The genes with positive growth rate response at both limitations (and thus common to ethanol carbon source) are likely to be regulated by the TFs listed in Table 2.7. The involvement of the *HAP* transcription factors (*HAP1-5*) is consistent with the expected growth rate associated increase in respiration on ethanol carbon source.

There is also an ethanol carbon source specific negative growth rate response (third quadrant in Fig.2.34) and the TFs most likely to mediate it are in table 2.8.

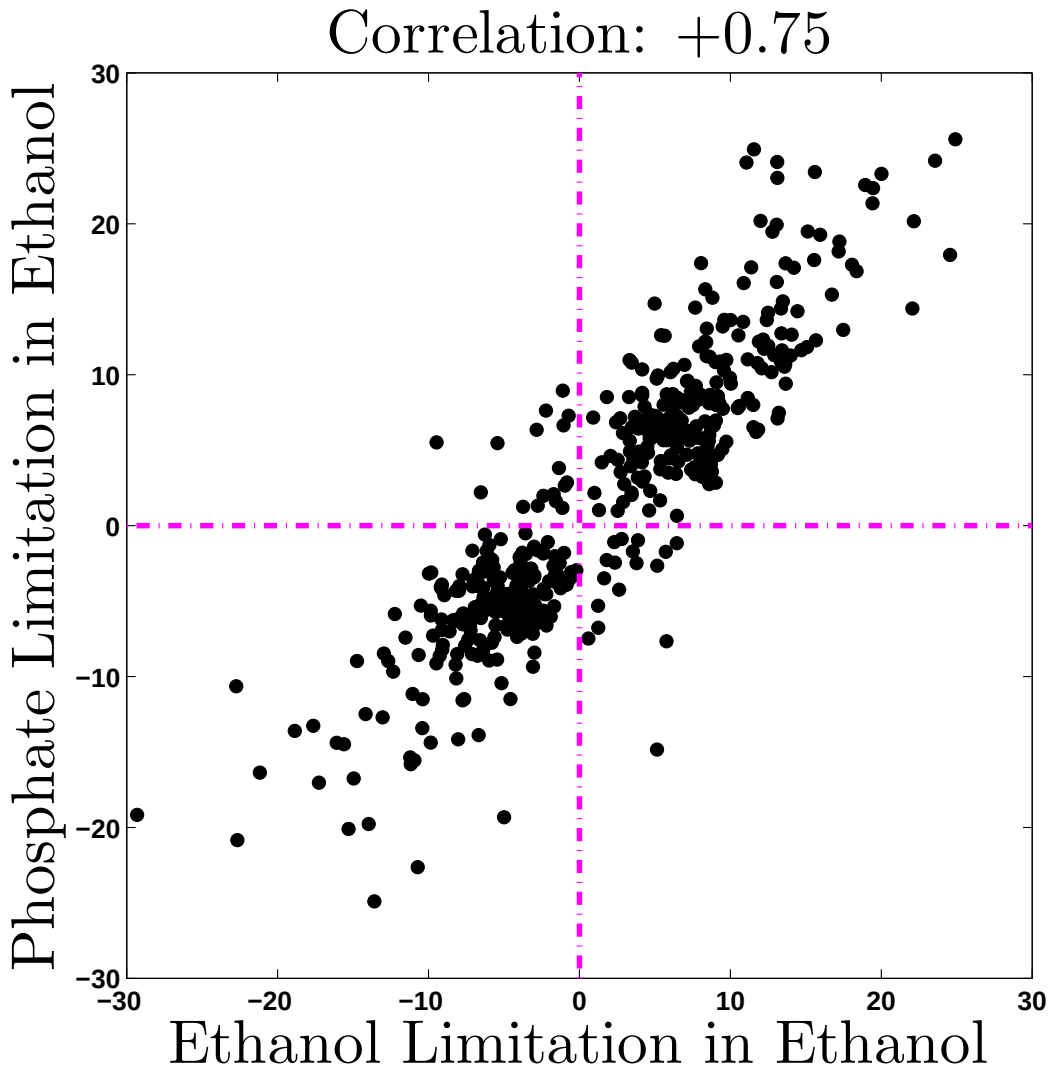

Figure 2.34: Correlation between slopes in ethanol and phosphate imitated cultures growing on ethanol carbon source.

The genes in the fourth quadrant of Fig.2.34 having positive slopes in the ethanol limitation but negative slopes in the phosphate limitation overlap significantly ($p < 10^{-4}$) with the targets of *PHO4* which reflects the effect of the phosphate limitation.

Next I use another pairwise comparison between phosphate limited cultures in ethanol carbon source and in glucose carbon source to identify TFs used differentially in those 2 carbon sources, Fig.2.35. The TF whose targets have more positive slopes in glucose

TF Name	Targets	Set	Overlap	p val	Reg
RTG3	48	914	17	10^{-4}	No
BAS1	34	914	20	10^{-9}	No
SIP4	10	914	5	10^{-3}	No
CBF1	211	914	50	10^{-3}	No
ABF1	241	914	55	10^{-3}	No
GLN3	67	914	18	10^{-2}	Yes
INO4	29	914	13	10^{-5}	No
HAP4	47	914	17	10^{-4}	No
GCN4	152	914	70	10^{-20}	No
LEU3	21	914	9	10^{-3}	No
HAP5	32	914	11	10^{-3}	No
HAP1	106	914	27	10^{-3}	No
HAP3	21	914	10	10^{-4}	No
HAP2	50	914	20	10^{-5}	No

Table 2.7: TF regulating the positive growth rate response in ethanol carbon source. The columns and notation are the same as in table 2.3

TF Name	Targets	Set	Overlap	p val	Reg
ZAP1	8	868	4	10^{-3}	Yes
MBP1	131	868	35	10^{-4}	Yes
RCS1	66	868	18	10^{-3}	No
GTS1	12	868	5	10^{-2}	Yes

Table 2.8: TF regulating the negative growth rate response in ethanol carbon source. The columns and notation are the same as in table 2.3

TF Name	Targets	Set	Overlap	p val	Reg
FKH1	98	230	11	10^{-3}	Yes
FKH2	100	230	11	10^{-3}	No
STB1	27	230	6	10^{-4}	No
MBP1	131	230	23	10^{-10}	No
SWI4	133	230	21	10^{-8}	No
SWI6	140	230	22	10^{-9}	No

Table 2.9: TF regulating the genes with different growth rate response between ethanol and glucose carbon source corresponding to quadrant 2 of Fig.2.35. The columns and notation are the same as in table 2.3

carbon source compared to ethanol carbon source are summarized in table 2.9. Remarkably all TFs in this group regulate cell–cycle genes. The *FKH1* and *FKH2* are involved in regulating *G2/M* genes. The protein products of *SWI4*, *SWI6* and *MBP1* form a complex that regulates the *G1/S* transition. *STB1* has a role in regulating MBF-specific

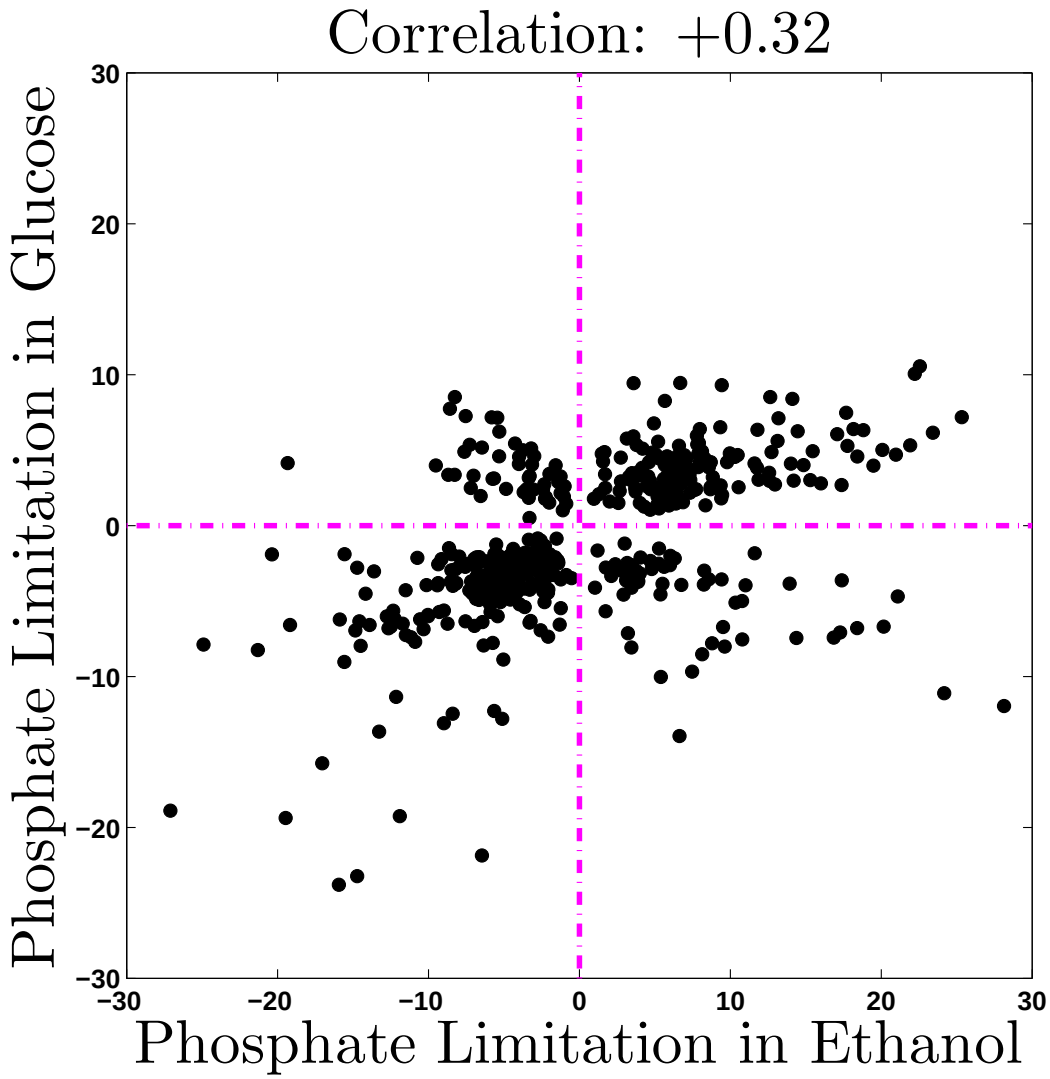

Figure 2.35: Correlation between slopes in phosphate imitated cultures growing on ethanol or glucose carbon source.

transcription at Start and expression is cell–cycle regulated. It is phosphorylated by *Cln-Cdc28p* kinases in vitro; unphosphorylated form binds Swi6p and binding is required for Stb1p function.

The TF whose targets have more positive slopes in ethanol carbon source are enumerated in table 2.10. *SIP4* most likely up–regulates the gluconeogensis genes required

for growth in ethanol but not for growth on glucose. The *HAP1-5* TFs are required for growth on ethanol to induce respiration.

TF Name	Targets	Set	Overlap	p val	Reg
RTG3	48	258	6	10^{-2}	Yes
SIP4	10	258	4	10^{-5}	No
STB4	10	258	4	10^{-5}	No
HAP4	47	258	9	10^{-5}	No
UME6	117	258	11	10^{-2}	No
HAP5	32	258	5	10^{-3}	No
HAP1	106	258	11	10^{-3}	Yes
HAP3	21	258	7	10^{-6}	No
HAP2	50	258	7	10^{-3}	No

Table 2.10: TF regulating the genes with different growth rate response between ethanol and glucose carbon source corresponding to quadrant 4 of Fig.2.35. The columns and notation are the same as in table 2.3

Complete and systematic identification of the genes, their GO terms and likely TF regulators for any pair of limitations and carbon sources can be found at my growth rate website.

Correlation between the Profiles of TFs and Their Targets

This overlap–based approach for identifying TFs is dependent of thresholds and lacks power to identify many of the regulators. For the TFs with very low *p values*, however, one may ask whether the mRNA corresponding to a TF is in the set of growth rate response genes used to identify the TF. In other words, does the expression profile of a TF correlate to the expression profiles of its targets ? For the majority of cases (final columns of the tables presented in this subsection) this is not the case. Thus a very important outcome from this analysis is that the level of a mRNA cannot be used as a surrogate for the activity of its corresponding TF. Many of the approaches (Segal *et al*, 2003) for inferring TF regulation make and heavily relay upon this assumption, thus I will not employ them for inferring regulatory interactions.

Conclusions

The overlap approach is very simple, intuitive and results in a high confidence predictions about the involvement of a few TF. However, the results are not quantitative, can be very sensitive to thresholds and strongly biased towards TFs with large number of targets that result in high enough statistical significance.

2.5.2 FIRE

FIRE (Elemento *et al*, 2007) is one of the few approaches for inferring transcriptional regulation that that does not "approximate" activity of TFs with their corresponding mRNAs. It is based on computing the mutual–information between DNA sequence motifs and the expression levels of mRNAs, the expression levels are represented as the presence or absence of mRNAs from a predetermined number of clusters.

To apply FIRE (Elemento *et al*, 2007), I clustered all growth rate data, both on ethanol and on glucose carbon source using k–means algorithm with 10 clusters and Euclidean distance as a similarity measure between expression profiles. The output of FIRE generated by the most current implementation *iGET* is depicted on Fig.2.36. The results from the application of FIRE contain only a few TF and consensus sequences. Some of the possible explanations for that include:

1. Loss of information due to representing large number of data points from the expression profile of the i^{th} gene with a single discrete number, the membership of the i^{th} gene to a cluster. This problem becomes increasingly significant with the number of transcriptionally profiled conditions, e.g the size of the dataset.

2. Data discretization (into clusters) whose optimal number is hard to determine.

3. Limited number of TF motifs

4. Estimation of mutual information depends critically on the accuracy of the estimated joint and marginal distributions which is generally rather hard and in the case of small samples impossible. This limits the application of FIRE only to genes from relatively large clusters.

While both the TF target overlap method and FIRE result in useful inferences, they have a number of drawbacks. To overcome these shortcomings, as well as to enable identifying combinatorial regulation (both by TFs and by mRNA degradation proteins) and build a quantitative model with predictive power, I developed a new inference algorithm, *$\mathcal{RC}$web* (Slavov, 2010). It is derived, tested and applied to the data in the next chapter.

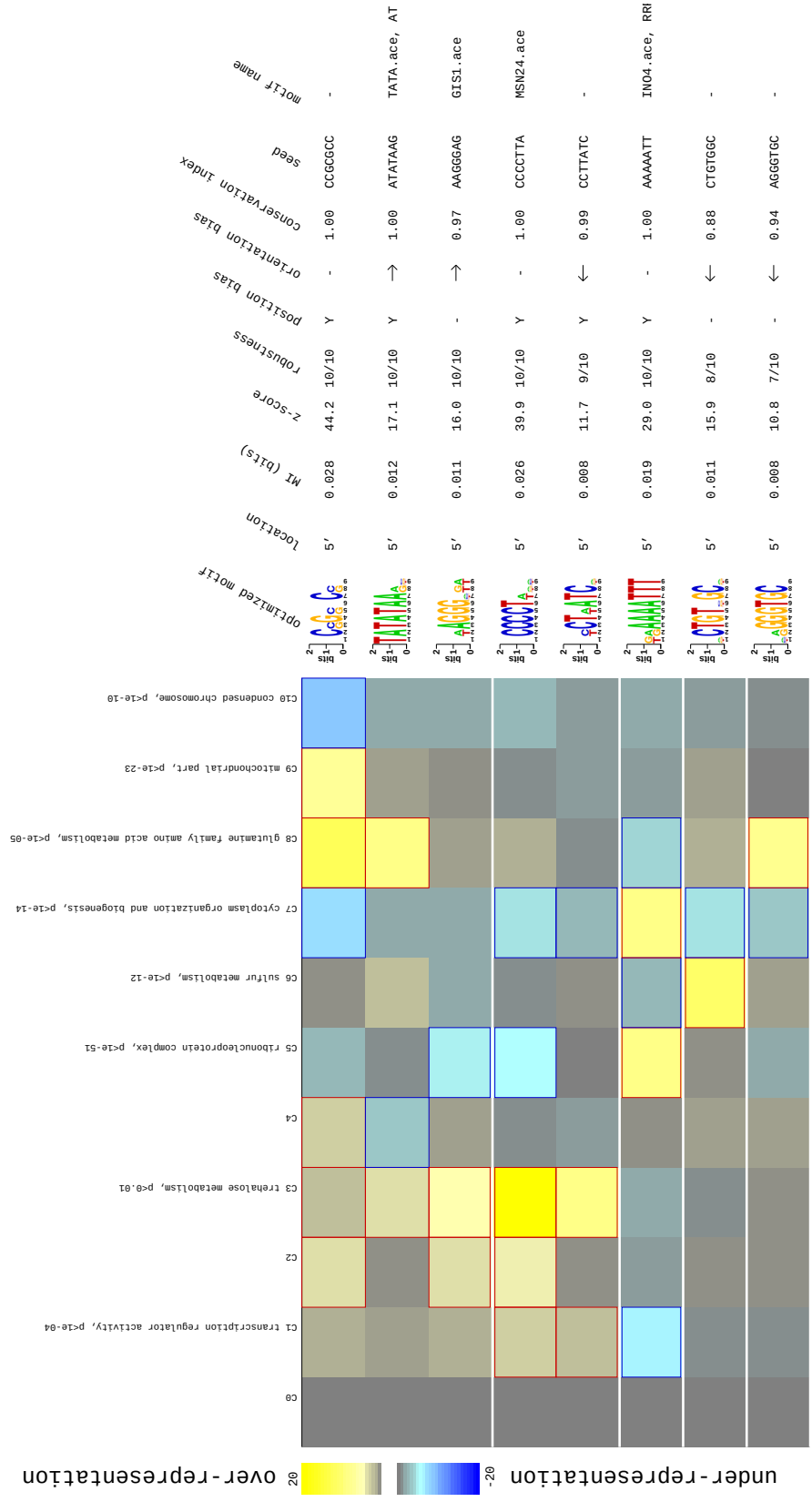

Figure 2.36: TFs identified by FIRE. FIRE was applied to the data clustered into 10 clusters,

Chapter 3

Regulation of the Growth Rate Response

3.1 Introduction

The use of networks has become common in biology as a framework within which to understand holistically complex biological systems and their emergent properties. Usually, the nodes (vertices) in biological networks are biomolecules and the links (edges) correspond to their interactions and biochemical reactions. Since our knowledge of the intermolecular interactions is incomplete, however, often the physical nature of the edges is not clearly defined, especially in the so called relational networks. Despite the strong interest in the problem and much research conducted on inferring intermolecular interactions from high-throughput data, there are still serious unresolved challenges. One of them is that intermolecular interactions and biochemical reactions are non-linear and their exact explicit forms are generally not known. A second major problem is that despite the significant advances in developing high-throughput experimental techniques for

simultaneous measurement of the levels of many biomolecules (such as mRNAs), many other molecules (such as metabolites, proteins and their posttranslational modifications) still cannot be measured as efficiently on the same scale. The high-throughput experimental detection of spatial localization of biomolecules and their modifications (such as phosphorylation and methylation) that influence crucially bimolecular interactions are even harder and still rather limited.

Despite these problems, researchers have attempted to infer network topologies, especially the topologies of transcriptional networks. These are bipartite networks (having two types of nodes) and consist of transcription factors (TFs) and the target genes regulated by the TFs. Most approaches to inferring transcriptional networks use the mRNA levels of TFs as predictor variables (surrogates for the unobserved concentrations of posttranslationally modified TFs) to explain the measured expression levels of mRNAs (Segal *et al*, 2003). Furthermore, the inference algorithms usually incorporate a model for computing conditional independence (using Bayesian networks or partial correlations that often rely on assuming linear dependence between the expression level of an mRNA and its regulator, the mRNA of the corresponding TF). The use of conditional independence aims to avoid indirect interactions and identify only the direct physical interactions in the inferred network.

In this work, I present a different perspective on the problem with emphasis on avoiding the aforementioned assumptions. Since there are multiple levels of regulation between a mRNA and its corresponding active TF (regulation of translation, posttranslational modifications and nuclear localization), $\mathcal{RC}web$ accepts that the active forms of TFs cannot be approximated with the levels of the corresponding mRNAs and treats TF activities as unobserved variables. Indeed, the correlation between the mRNA of the i^{th} TF and the mRNAs whose transcription is regulated by the i^{th} TF is rather poor for many TFs as demonstrated in section 2.5.1. Furthermore, $\mathcal{RC}web$ does not try to

approximate the functional dependence between mRNAs and their regulators with a particular functional form since the dependence can be very non-linear and highly co-operative (e.g. having irreducible many-body components). Rather, *RCweb* attempts to identify sets of genes regulated by the same regulators (which are treated as unobserved variables) without knowing the molecular identities of these regulators *a priori*, e.g. assign them to particular TFs or proteins degrading mRNA. Only later, based on further experimental or computational analysis the regulators can be identified.

3.2 Generative Model

As mentioned previously, *RCweb* does not need to assume explicit functional dependence between mRNAs and the levels of their regulators. The formal and general framework for treating non–linear functions is outlined in the Appendix 5.3. Here, I review briefly the application to a particular class of functions relevant to biological interactions and signal transduction.

Consider the i^{th} physiological condition in which the concentration of the j^{th} mRNA (G_{ij}) is determined by its Q_j regulators, $\vec{x} \equiv (x_1, ..., x_{Q_j}) \equiv \{x_k\},\ k \in \omega_j$, which are the active posttranslationally modified proteins localized to their organelles of activities, RNAs and small molecules (ligands) that control the production (transcription) and the degradation of the G_j. These include, transcription factors (TFs), enzymes modifying histones and nucleosomes, non-coding RNAs, proteins binding mRNAs and regulating their degradation. For simplicity and intuition building we first derive *RCweb* by assuming a very likely explicit form (1) for the rate of production and degradation of G_j that takes into account the active forms of TFs binding to the promoter of G_j, TF_k for $k \in \omega_j$, as well as τ_j quantifying the degradation rate of G_j. For derivation of *RCweb* for the general case $[G_j] = F_j(\vec{x})$ where F_j is any non-linear function of the regulators $\vec{x}$ see

5.3.

$$\frac{d[G_{ij}]}{dt} = \prod_{k \in \omega_j} V_{max_k} \frac{[TF_{ik}]^{n_k}}{[TF_{ik}]^{n_k} + K_k^{n_k}} - \frac{1}{\tau_j}[G_{ij}] \tag{3.1}$$

At steady-state $d[G_{ij}]/dt = 0$ and (1) simplifies to:

$$[G_{ij}] = \tau_i \prod_{k \in \omega_j} V_{max_k} \frac{[TF_{ik}]^{n_k}}{[TF_{ik}]^{n_k} + K_k^{n_k}} \tag{3.2}$$

If $[TF]_{ik}$ is approximated with the concentration of the corresponding mRNA and the non-linear term is approximated with a linear one (2) can be solved easily. These assumptions, however, are poorly justified and introduced only to ease the computation. To avoid them, we reframe the problem and treat the active forms of TF as unobserved variables:

$$\log([G_{ij}]) = \log(\tau_j) + \sum_{k \in \omega_j} \log(V_{max_k} \frac{[TF_{ik}]^{n_k}}{[TF_{ik}]^{n_k} + K_k^{n_k}}) \tag{3.3}$$

$$\underbrace{\log([G_{ij}])}_{y_{ij}} = \underbrace{\log(\tau_j) + \sum_{k \in \omega_j} \log(V_{max_k})}_{c_i^o} + \sum_{k \in \omega_j} \underbrace{\log(\frac{[TF_{ik}]^{n_k}}{[TF_{ik}]^{n_k} + K_k^{n_k}})}_{r_{ik}} \tag{3.4}$$

$$y_{ij} = c_i^o + \sum_{k \in \omega_j} r_{ik} \tag{3.5}$$

To infer the network, we need to find c_i^o (a gene specific constant), ω_j (the set of regulators that control the expression of the j^{th} gene) and r_{ik} (the numerical value of the k^{th} component function in the i^{th} conditions that corresponds to regulators whose levels and identities we do not know *a priori*) consistent with the measured levels of messenger RNA for the j^{th} gene for all physiological conditions, y_{ij}. If we assume 3.1 for the explicit form of the j^{th} expression function (F_j), the component function r_k is a non–linear function of a single TF and r_{ik} is its numerical value at the i^{th} physiological condition. In the general case, however, r_k can be any non-linear component function (that may or may not have a closed form) of one or many TFs that takes into account their

interactions (e.g. cooperative/synergistic effects), see 5.3. For a single mRNA (3.5) is an ill-posed problem without a unique solution. However, equation (3.5) can be written in a matrix form for all genes:

$$\mathbf{Y} = \mathbf{RC} \tag{3.6}$$

The solution of this mathematic problem (3.6) is the subject of the next section

3.3 Introduction to *RCweb*

Factor analysis (FA) decompositions are useful for explaining the variance of observed variables in terms of fewer unobserved variables that may capture systematic effects and allow for low dimensional representation of the data. Yet, the interpretation of latent variables inferred by FA is fraught with problems. In fact, interpretation is not always expected and intended since FA may not have an underlying generative model. A prime difficulty with interpretation arises from the fact that any rotation of the factors and their loadings by an orthogonal matrix results in a different FA decomposition that explains the variance in the observed variables just as well. Therefore, in the absence of additional information on the latent variables and their loading, FA cannot identify a unique decomposition, much less generative relationships between latent factors and observed variables.

A frequent choice for a constraint implemented by principle component analysis (PCA) and resulting in a unique solution is that the factors are the singular vectors (and thus orthogonal to each other) of the data matrix ordered in descending order of their corresponding singular values. Yet, this choice is often motivated by computational convenience rather than by knowledge about the system that generated the data. Another type of computationally convenient constraint applied to facilitate the interpretation of

FA results is sparsity, in the form of sparse Bayesian FA (West, 2002; Dueck *et al*, 2005; Carvalho and West, 2008), sparse PCA (d'Aspremont A, 2007; Sigg and Buhmann, 2008) and FA for gene regulatory networks (Srebroand and Jaakkola, 2001; Pe'er *et al*, 2002). However, papers that introduce and use sparse PCA do not consider a generative model but rather use sparsity as a convenient tool to produce interpretable factors that are linear combinations of just a few original variables. In sparse PCA, sparsity is a way to balance interpretability at the cost of slightly lower fraction of explained variance.

The algorithm described in this paper ($\mathcal{RC}web$) also uses a sparse prior, but $\mathcal{RC}web$ explicitly considers the problem from a generative perspective. $\mathcal{RC}web$ asserts that there is indeed a set of hidden variables that connect to and regulate the observed variables via a sparse network. Based on that model, I derive a network structure learning approach within explicit theoretical framework. This allows to propose an approach for sparse FA which is conceptually and computationally different from all existing approaches such as K-SVD (M. Aharon and and Bruckstein, 2005), sparse PCA and other LARS (Bradley *et al*, 2004) based methods (Banerjee *et al*, 2007). $\mathcal{RC}web$ is appropriate for analyzing data arising from any system in which the state of each observed variable is affected by a strict subset of the unobserved variables. To assign the inferred latent variables to physical factors, $\mathcal{RC}web$ needs either data from perturbation experiments or prior knowledge about the factors. This framework generalizes to non-linear interactions, which is discussed elsewhere. Furthermore, I analyze the scaling of the computational complexity of $\mathcal{RC}web$ with the number of observed and unobserved variables, as well as the parameter space where $\mathcal{RC}web$ can accurately infer network topologies and demonstrate its robustness to noise in the data.

3.4 Derivation

Consider a sparse bipartite graph $\mathcal{G} = (\mathcal{E}, \mathcal{N}, \mathcal{R})$ consisting of two sets of vertices $\mathcal{N}$ and $\mathcal{R}$ and the associated set of directed edges $\mathcal{E}$ connecting $\mathcal{R}$ to $\mathcal{N}$ vertices. Define a graphical model in which each vertex s corresponds to a random variable; N observed random variables indexed by $\mathcal{N}$ ($x_{\mathcal{N}} = \{x_s | s \in \mathcal{N}\}$) whose states are functions of P unobserved variables indexed by $\mathcal{R}$, $x_{\mathcal{R}} = \{x_s | s \in \mathcal{R}\}$. Since the states of $x_{\mathcal{N}}$ depend on (are regulated by) $x_{\mathcal{R}}$, I will also refer to $x_{\mathcal{R}}$ as regulators. The functional dependencies are denoted by a set of directed edges $\mathcal{E}$ so that each unobserved variable $x_i | i \in \mathcal{R}$ affects (and its vertex is thus connected to) a subset of observed variables $x_{\alpha_i} = \{x_s | s \in \alpha_i \subset \mathcal{N}\}$. Given a dataset $\mathbf{G} \in \mathbb{R}^{M \times N}$ of M configurations of the observed variables $x_{\mathcal{N}}$, *RCweb* aims to infer the edges $\mathcal{E}$ and the corresponding configurations of the unobserved variables, $x_{\mathcal{R}}$.

If the state of each observed variable is a linear superposition of a subset of unobserved variables, the data $\mathbf{G}$ can be modeled with a very simple *generative model* (3.7): The data is a product between $\mathbf{R} \in \mathbb{R}^{M \times P}$ (a matrix whose columns correspond to the unobserved variables and the rows correspond to the M measured configurations) and $\mathbf{C} \in \mathbb{R}^{P \times N}$, the weighted adjacency matrix of $\mathcal{G}$. The unexplained variance in the data $\mathbf{G}$ is captured by the residual $\boldsymbol{\Upsilon}$.

$$\mathbf{G} = \mathbf{RC} + \boldsymbol{\Upsilon} \tag{3.7}$$

This decomposition of $\mathbf{G}$ into a product of two matrices can be considered to be a type of factor analysis with $\mathbf{R}$ being the factors and $\mathbf{C}$ the loadings. Even when $P \ll M$ the decomposition of $\mathbf{G}$ does not have a unique solution since $\mathbf{RC} \equiv \mathbf{R}\mathcal{I}\mathbf{C} \equiv \mathbf{RQ}^T\mathbf{QC} \equiv \mathbf{R}^*\mathbf{C}^*$ for any orthonormal matrix $\mathbf{Q}$. Thus the identification of a unique decomposition corresponding to the structure of $\mathcal{G}$ requires additional criteria constraining the decompo-

sition. The assumption that $\mathcal{G}$ is sparse requires that $\mathbf{C}$ be sparse as well meaning that the state of the i^{th} observed vertex $x_i|i \in \mathcal{N}$ is affected by a strict subset of the unobserved variables $x_{\psi_i} = \{x_s|s \in \psi_i \subset \mathcal{R}\}$, the ones whose weights in the i^{th} column of $\mathbf{C}$ are not zeros, $\mathbf{C}_{i\psi_i} \neq 0$ and $\mathbf{C}_{i\bar{\psi}_i} = 0$ where $\bar{\psi}_i$ is the complement of ψ_i. Thus, to recover the structure of $\mathcal{G}$, *RCweb* seeks to find a decomposition of $\mathbf{G}$ in which $\mathbf{C}$ is sparse. The sparsity can be introduced as a regularization with a Lagrangian multiplier λ:

$$(\hat{\mathbf{C}}, \hat{\mathbf{R}}) = \arg\min_{\mathbf{R},\mathbf{C}} \|\mathbf{G} - \mathbf{RC}\|_F^2 + \lambda\|\mathbf{C}\|_0 \tag{3.8}$$

In the equations above and throughout the paper $\|\mathbf{C}\|_F^2 = \sum_{i,j} C_{ij}^2$ denotes entry-wise (Frobenius) norm, and the zero norm of a vector or matrix ($\|\mathbf{C}\|_0$) equals the number of non-zero elements in the array.

To infer the network topology *RCweb* aims to solve the optimization problem defined by (3.8). Since (3.8) is a NP-hard combinatorial problem, the solution can be simplified significantly by relaxing the $\ell 0$ norm to $\ell 1$ norm (Bradley *et al*, 2004). Then the approximated problem can be tackled with interior point methods (Banerjee *et al*, 2007). As an alternative approach to $\ell 1$ approximation, I propose a novel method based on introducing a degree of freedom in the singular-value decomposition (SVD) of $\mathbf{G}$ by inserting an invertible[1] matrix $\mathbf{B}$.

$$\mathbf{G} = \mathbf{USV}^T \equiv \underbrace{(\mathbf{US}(\mathbf{B}^T)^{-1})}_{\hat{\mathbf{R}}} \underbrace{(\mathbf{B}^T\mathbf{V}^T)}_{\hat{\mathbf{C}}} \tag{3.9}$$

The prior (constraint) that $\hat{\mathbf{C}}$ is sparse determines $\mathbf{B}$ that minimizes (3.8), and thus a unique decomposition. The goal of introducing $\mathbf{B}$ is to reduce the combinatorial problem to one that can be solved with convex minimization. When the factors underlying the observed variance are fewer than the observations in $\mathbf{G}$ there is no need to take the full SVD; if P factors are expected, only the first P largest singular vectors and values from

[1] $\mathbf{B}$ is always invertible by construction, see section 3.6.3 and equation (3.10)

the SVD of $\mathbf{G}$ are taken in that decomposition so that $\mathbf{USV}^T$ is the matrix with rank P that best approximates $\mathbf{G}$ in the sense of minimizing $\|\mathbf{G} - \mathbf{USV}^T\|_F^2$. Since conceivably sparse decompositions may use columns outside of the best $\ell 2$ approximation, *RCweb* considers taking the first P^* for $P^* > P$ singular pairs. Such expanded basis is more likely to support the optimal sparse solution and especially relevant for the case when P is not known. Such choice can be easily accommodated in light of the ability of *RCweb* to exclude unnecessary explanatory variables, see section 3.6.3.

Next *RCweb* computes $\mathbf{B}$ based on the requirement that $\mathbf{C}$ is sparse for the case $N > P$. To compute $\mathbf{B}$, one may set an optimization problem (3.10). Once $\mathbf{B}$ is inferred, $\hat{\mathbf{R}}$ and $\hat{\mathbf{C}}$ can be computed easily, $\hat{\mathbf{R}} = \mathbf{US}\hat{\mathbf{B}}^{-1}$ and $\hat{\mathbf{C}} = (\mathbf{V}\hat{\mathbf{B}})^T$.

$$\hat{\mathbf{B}} = \arg\min_{\mathbf{B}} \|\mathbf{VB}\|_0, \text{ so that } \det(\mathbf{B}) > 1 \tag{3.10}$$

The constraint on $\mathbf{B}$ in (3.10) is introduced to avoid trivial and degenerate solutions, such as $\mathbf{B}$ being rank deficient $\mathbf{B}$. Thus the introduction of $\mathbf{B}$ reduces (3.8) to a problem (3.10) that is still combinatorial and might also be approximated with a more tractable problem by relaxing the $\ell 0$ norm to $\ell 1$ norm and applying heuristics (Cetin *et al*, 2002; Candès *et al*, 2007) to enhance the solution. I propose a new approach, *RCweb*, outlined in the next section.

3.5 *RCweb*

Assume that a sparse $\mathbf{c}_i^T$ corresponding to an optimal $\hat{\mathbf{b}_i}$ (the i^{th} column of $\hat{\mathbf{B}}$) form $\mathbf{V}\hat{\mathbf{b}_i} = \mathbf{c}_i^T$ is known. Define the set of indices corresponding to non-zero elements in $\mathbf{c}_i^T$ with ω_- and the set of indices corresponding to zero elements in $\mathbf{c}_i^T$ with ω_0. Furthermore, define the matrix $\mathbf{V}_{\omega_0}$ to be the matrix containing only the rows of $\mathbf{V}$ whose indices are in ω_0. If ω_0 and thus $\mathbf{V}_{\omega_0}$ are known, one can easily compute $\hat{\mathbf{b}_\mathbf{i}}$ as the right singular vector

of $\mathbf{V}_{\omega_0}$ corresponding to the zero singular value. Since $\mathbf{c}_i^T$ and ω_0 are not known, *RCweb* approximates $\hat{\mathbf{b}}_i$ (the smallest[2] right singular vector of $\mathbf{V}_{\omega_0}$) with $\mathbf{v}_s$, the smallest right singular vector of $\mathbf{V}$. This approximation relies on assuming that a low rank perturbation in a matrix results in a small change in its smallest singular vectors (Benaych-Georges and Rao, 2009). Thus given that *RCweb* is looking for the sparsest solution and the set ω_- is small relative to N, the angle between the singular vectors of $\mathbf{V}_{\omega_0}$ and $\mathbf{V}$ is small as well. Therefore, $\mathbf{v}_s$ can serve as a reasonable first approximation of $\mathbf{b}_i$. Then *RCweb* systematically and iteratively uses and updates $\mathbf{v}_s$ by removing rows of $\mathbf{V}$ until $\mathbf{v}_s$ converges to $\mathbf{b}_i$ or equivalently $\mathbf{V}_{\omega_0}$ becomes singular for the largest set of ω_0 indices. When $\mathbf{V}_{\omega_0}$ becomes singular, all elements of $\mathbf{c}_i$ whose indices are in ω_0 become zero.

RCweb also has an intuitive geometrical interpretation. Consider the matrix $\mathbf{V}$ mapping the unit sphere in $\mathbb{R}^P$ (the sphere with unit radius from $\mathbb{R}^P$) to an ellipsoid in $\mathbb{R}^N$. The axes of the ellipsoid are the left singular vectors of $\mathbf{V}$. In this picture, starting with $\omega_- = \{\emptyset\}$ and $\omega_0 = \{1, \ldots, N\}$, solving (3.10) requires moving the fewest number of indices from ω_0 to ω_- so that $\mathbf{V}_{\omega_0}$ maps a vector from $\mathbb{R}^P$ to the origin. How to choose the indices to be moved? At each step *RCweb* chooses $i_{|max|}$, the index of the largest element (by absolute value) of the smallest axis of the ellipsoid which is the left singular vector of $\mathbf{V}$ with the smallest singular value. *RCweb* moves $i_{|max|}$ from ω_0 to ω_- effectively selecting the dimension whose projection is easiest to eliminate and removing its largest component, which minimizes as much as possible the projection in that dimension. *RCweb* keeps moving indices from ω_0 to ω_- using the same procedure until the smallest right singular vector of $\mathbf{V}_{\omega_0}$ converges to $\mathbf{b}_i$ and the smallest singular value of $\mathbf{V}_{\omega_0}$ approaches zero. *RCweb* is guaranteed to stop after at most $(N-P+1)$ steps since after removal of $(N-P+1)$ indices from ω_0, $\mathbf{V}_{\omega_0}$ will be at most rank $P-1$. If *RCweb* finds a sparse solution it will converge in fewer steps.

[2] By smallest singular vector I mean the singular vector corresponding to the to the smallest singular value

1. **Task:**

$$\hat{\mathbf{b}}_i = \min_{\mathbf{b_i^T b_i \geq 1}} \|\mathbf{V}\mathbf{b}_i\|_0$$

2. **Initialization:**

   - $\omega_- = \{\emptyset\}$ and $\omega_0 = \{1, 2, \ldots, N\}$
   - Set $\mathbf{K}_{\omega_0}^{-1} = (\mathbf{V}_{\omega_0}^T \mathbf{V}_{\omega_0})^{-1} = \mathcal{I} \in \mathbb{R}^{P \times P}$
   - Set $J = 1$;
   - $i_{|max|} = \arg\max \left(\sum_j |\mathbf{V}_{ij}| \right)$
     $\omega_- = \{i_{|max|}\}, \omega_0 = \{i | i \in \omega_0,\ i \neq i_{|max|}\}$
   - Update $\mathbf{K}_{\omega_0}^{-1} = RankUpdate(\mathbf{K}_{\omega_0}^{-1}, \mathbf{V}_{i_{|max|}})$

3. **Cycle:** $J = J + 1$ Repeat until convergence

   - Find the eigenvector $\mathbf{v}$ for $\mathbf{K}_{\omega_0}^{-1}$ with the largest eigenvalue λ_{max}
   - If $\lambda_{max}^{-1} \approx 0$ or $\mathbf{v}^J \rightarrow \mathbf{v}^{J-1}$, $\hat{\mathbf{b}}_i \equiv \mathbf{v}$; **STOP**
   - Compute the left singular vector of $\mathbf{V}_{\omega_0}$
     $\mathbf{u} = s^{-1} \mathbf{V}_{\omega_0} \mathbf{v}$
   - $i_{|max|} = \arg\max [(|u_1|, \ldots, |u_i|, \ldots, |u_N|)]$;
   - $\omega_- = \{\omega_-, i_{|max|}\}$
     $\omega_0 = \{i | i \in \omega_0,\ i \neq i_{|max|}\}$
   - Update $\mathbf{K}_{\omega_0}^{-1} = RankUpdate(\mathbf{K}_{\omega_0}^{-1}, \mathbf{V}_{i_{|max|}})$

The above algorithm can compute a single vector, $\hat{\mathbf{b}_\mathbf{i}}$, which is just one column of $\hat{\mathbf{B}}$. To find the other columns, *RCweb* applies the same approach to the modified (inflated) matrix, which for the i^{th} column of $\mathbf{B}$ is $\mathbf{V}^{(i)} = \mathbf{V}^{(i-1)} + \mathbf{V}\hat{\mathbf{b}_\mathbf{i}}\hat{\mathbf{b}_\mathbf{i}^\mathbf{T}}$ for $i = 2, \ldots, P$. Thus, after the inference of each column of $\mathbf{B}$ *RCweb* modifies $\mathbf{V}^{(i-1)}$ to $\mathbf{V}^{(i)}$ ($\mathbf{V}^{(1)} \equiv \mathbf{V}$) so

that the algorithm will not replicate its choice of ω_0. Note that in the i^{th} update of $\mathbf{V}^{(i-1)}$ the $\mathbf{V}\hat{\mathbf{b}}_\mathbf{i}\hat{\mathbf{b}}_\mathbf{i}^\mathbf{T}$ will modify only the ω_- rows in $\mathbf{V}^{(i-1)}$ since the rows of the $\mathbf{V}\hat{\mathbf{b}}_\mathbf{i}\hat{\mathbf{b}}_\mathbf{i}^\mathbf{T}$ whose indices are in ω_0 contain only zero elements. Applying *RCweb* to the inflated matrices avoids inferring multiple times the same $\hat{\mathbf{b}}_\mathbf{i}$, but a $\hat{\mathbf{b}}_\mathbf{i}$ inferred from the inflated matrix is generally going to differ at least slightly from the corresponding $\hat{\mathbf{b}}_\mathbf{i}$ that solves (3.10) for $\mathbf{V}$. To avoid that, *RCweb* uses the inflated matrices only for the first few iterations until the largest (by absolute magnitude) of the Pearson correlations between the smallest eigenvector of $\mathbf{V}_{\omega_0}$ from the current (i^{th}) iteration and the recovered columns of $\mathbf{B}$ is less than $1-\epsilon$ and monotonically decreasing; ϵ is chosen for numerical stability and also to reflect the similarity between the connectivity of $\mathcal{R}$ vertices that can be expected in the network whose topology is being recovered. A simpler alternative that works great in practice is to use the the inflated matrix for the first k iterations that are enough to find a new direction for $\hat{\mathbf{b}}_\mathbf{i}$ and then *RCweb* switches back to $\mathbf{V}$ so that the solution is optimal for $\mathbf{V}$. The switch requires k rank update of $\mathbf{K}_{\omega_0}^{-1} \equiv (\mathbf{V}_{\omega_0}^T\mathbf{V}_{\omega_0})^{-1}$ and thus choosing k small saves computations. Choosing k too small, however, may not be enough to guarantee that $\hat{\mathbf{b}}_\mathbf{i}$ will not recapitulate a solution that is already found. Usually $k = 10$ works great and can be easily increased if the new solution is very close to an old one.

There are a few notable elements that make *RCweb* efficient. First, *RCweb* does almost all computations in $\mathbb{R}^P$ and since $P \ll N$, $P < M$, that saves both memory and CPU time. Second, each step requires only a few matrix-vector multiplication for computing the eigenvectors (since the change from the previous step is generally very small) and $\mathbf{K}_{\omega_0}^{-1}$ is computed by a rank–one update which obviates matrix inversion.

The approach that *RCweb* takes in solving (3.10) does not impose specific restrictions on the distribution of the observed variables ($\mathbf{G}$), the noise in the data ($\mathbf{\Upsilon}$) or the latent variables $\hat{\mathbf{R}}$. However, the initial approximation of $\mathbf{b}_i$ with $\mathbf{v}_s$ can be poor for data arising from dense networks or special worst–case datasets. As demonstrated theoretically

(Benaych-Georges 2009) and tested numerically in the next section, $\mathcal{RC}web$ performs very well at least in the absence of worst–case scenario special structures in the data.

3.6 Validation

To evaluate the performance of $\mathcal{RC}web$, I first apply it to data from simulated random bipartite networks with two different topology types, (1) Erdös & Rènyi and (2) scale-free whose corresponding degree distributions are (1) Poisson and (2) power–law. The network topology is encoded in a weighted adjacency matrix $\mathbf{C}^{gold}$ and the values for the unobserved variables are drawn from a standard uniform distribution. The simulations result in data matrices $\mathbf{G} \in \mathbb{R}^{M \times N}$ containing M observations of all N unobserved variables. According to $\mathcal{RC}web$, the optimal sparse adjacency matrix ($\hat{\mathbf{C}}$) and the hidden variables ($\hat{\mathbf{R}}$) can be inferred by the decomposition, $\hat{\mathbf{G}} = \hat{\mathbf{R}}\hat{\mathbf{C}}$ so that $\hat{\mathbf{C}}$ is as sparse as possible while $\hat{\mathbf{G}}$ is as close as possible to $\mathbf{G}$.

In addition to $\mathcal{RC}web$, such decomposition can be computed by 3 classes of existing algorithms. For a comparison, I use the latest versions for which the authors report best performance: (*A*) *PSMF* for Bayesian matrix factorization as implemented by the author MatLab function PSMF1 (Dueck *et al*, 2005); (*B*) *BFRM 2* for Bayesian matrix factorization as implemented by the author compiled executable (Carvalho and West, 2008); (*C*) *emPCA* for maximum likelihood estimate (*MLE*) sparse PCA (Sigg and Buhmann, 2008); (*D*) *K-SVD* (M. Aharon and and Bruckstein, 2005). All algorithms are implemented using the code published by their authors, and with the default values of the parameters when parameters are required. The results are compared for various *M, N, P*, sparsity, and noise levels.

3.6.1 Limitations

Before comparing the results, consider some of the limitations common to all algorithms and the appropriate metrics for comparing the results. In the absence of any other information, the decomposition of $\mathbf{G}$ (no matter how accurate) cannot associate hidden variables (corresponding to columns of $\hat{\mathbf{R}}$) to physical factors. Furthermore, all methods can infer $\hat{\mathbf{C}}$ and $\hat{\mathbf{R}}$ only up to an arbitrary diagonal scaling or permutation matrix. First consider the scaling illustrated by the following transformation by a diagonal matrix $\mathbf{D}$, $\hat{\mathbf{G}} = \hat{\mathbf{R}}\hat{\mathbf{C}} = \hat{\mathbf{R}}\mathcal{I}\hat{\mathbf{C}} = \hat{\mathbf{R}}(\mathbf{D}\mathbf{D}^{-1})\hat{\mathbf{C}} = (\hat{\mathbf{R}}\mathbf{D})(\mathbf{D}^{-1}\hat{\mathbf{C}}) = \hat{\mathbf{R}}^{*}\hat{\mathbf{C}}^{*}$. Such transformation is going to rescale $\hat{\mathbf{R}}$ to $\hat{\mathbf{R}}^{*}$ and $\hat{\mathbf{C}}$ to $\hat{\mathbf{C}}^{*}$, which is just as sparse as $\hat{\mathbf{C}}$, $\|\hat{\mathbf{C}}^{*}\|_0 = \|\hat{\mathbf{C}}\|_0$. Since both decompositions explain the variance in $\mathbf{G}$ equally well *$\mathcal{RC}$web* (or any of the other method) cannot distinguish between them. Thus given $\hat{\mathbf{C}}$, there is a diagonal matrix $\hat{\mathbf{D}}$ that scales $\hat{\mathbf{C}}$ to $\mathbf{C}^{gold}$, the weighted adjacency matrix of $\mathcal{G}$.

The second limitation is that in the absence of addition information, *$\mathcal{RC}$web* can determine $\hat{\mathbf{C}}$ and $\mathbf{R}$ up to a permutation matrix. Consider comparing $\hat{\mathbf{C}}$ to the adjacency matrix used in the simulations, $\mathbf{C}^{gold}$. Since the identity of the inferred hidden variables is not known the rows of $\hat{\mathbf{C}}$ do not generally correspond to the rows of $\mathbf{C}^{gold}$; $\hat{\mathbf{C}}_i$ (the i^{th} row of $\hat{\mathbf{C}}$) is most likely to correspond to the $\mathbf{C}^{gold}$ row that is most correlated to $\hat{\mathbf{C}}_i$ and the Pearson correlation between the two rows quantifies the accuracy for the inference of $\hat{\mathbf{C}}_i$. To implement this idea, all rows of $\hat{\mathbf{C}}$ and $\mathbf{C}^{gold}$ are first normalized to mean zero and unit variance resulting in $\mathbf{C}_{nor}$ and $\mathbf{C}^{gold}_{nor}$. The correlation matrix then is, $\mathbf{\Sigma} = \mathbf{C}^{T}_{nor}\mathbf{C}^{gold}_{nor}$ and the most likely vertex (index of the unobserved variable) corresponding to $\hat{\mathbf{C}}_i$ is $k = \arg\max_j(|\Sigma_{i1}|, \ldots, |\Sigma_{ij}|, \ldots, |\Sigma_{iP}|)$, where $k \in \mathcal{R}$. The absolute value is required because the diagonal elements of $\hat{\mathbf{D}}$ can be negative. The accuracy is measured by the corresponding Pearson correlation, $\rho_i = |\Sigma_{ik}|$. An optimal solution of this matching problem can be found by using belief propagation algorithm for the simple case of a

bipartite graph even the *LP* relaxed version guarantees optimal solution (Sanghavi 2007). The overall accuracy is quantified by the mean correlation $\bar{\rho} = (1/p)\sum_{i=1}^{i=p}\rho_i$ where p is the number of inferred unobserved variables and can equal to P or not depending on whether the number of unobserved variables is known or not. In computing $\bar{\rho}$, each row of $\mathbf{C}^{gold}$ is allowed to correspond only to one row of $\hat{\mathbf{C}}$ and vice versa.

In addition to the two common limitations of permutation and scaling, some algorithms (d'Aspremont A, 2007) for sparse PCA require $M > N$ and since this is not the case in many real world problems and in some of the datasets simulated here, those methods are not tested. Instead I chose emPCA, which does not require $M > N$ and is the latest *MLE* algorithm for sparse PCA that according to its authors is more efficient than previous algorithms (Sigg and Buhmann, 2008).

3.6.2 Accuracy and Complexity Scaling

$\mathcal{RC}web$ has a natural way for identifying the mean degree[3]. However, some of the other algorithms require the mean degree for optimal performance. To avoid underestimating an algorithm simply because it recovers networks that are too sparse or not sparse enough, I assume that the mean degree is known and it is input to all algorithms. First all algorithms are tested on a very easy inference problem, Fig.1. Since PSMF and BFRM have lower accuracy and PSMF is significantly slower than the other algorithms, the rest of the results will focus on the *MLE* algorithms that also have better performance. PSMF gives less accurate results with power–law networks which can be understood in terms of the uniform prior used by PSMF. PSMF has the advantage over the *MLE* algorithms in inferring a probabilistic network structure rather than a single estimate. Special advantage of BFRM is the seamless inclusion of response variables and measured factors in the

[3]When $\mathcal{RC}web$ learns all edges, the smallest singular value of $\mathbf{V}_{\omega_0}$ approaches zero and its smallest singular vector converges to $\hat{\mathbf{b}}_\mathbf{i}$.

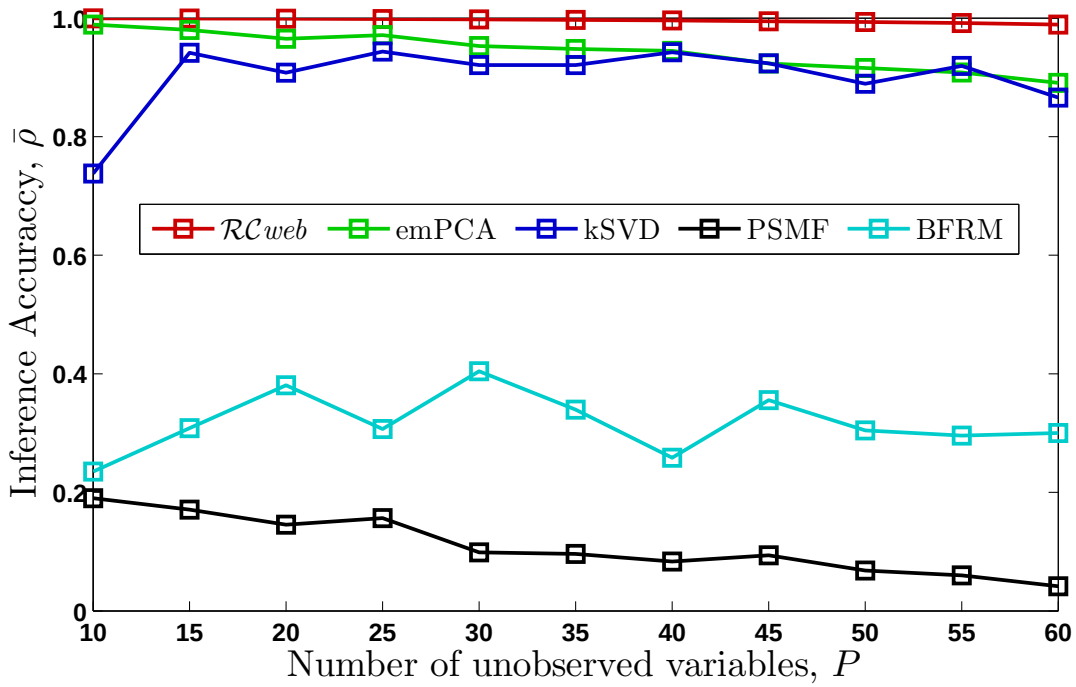

Figure 3.1: Accuracy of network recovery as a function of the number of unobserved variables, P. Number of observed variables $N = 500$; Number of observations, $M = 280$. All networks are with Poisson mean out–degree $0.10N = 50$, with 10 % noise in the observations.

inference. For the *MLE* algorithms, the accuracy of network inference increases with the ratio of observed to unobserved variables N/P (Fig.2) and with the number of observed configurations M, Fig.3. In contrast, as the noise in the data and the mean out–degree (mean number of edges from $\mathcal{R}$ to $\mathcal{N}$ vertices) are increased, the accuracy of the inference decreases. All algorithms perform better on Poisson networks and the lower level of noise in the data from power–law networks was chosen to partially compensate for that.

An important caveat when comparing the results for different algorithms is that K-SVD iteratively improves the accuracy of the solution, and thus the output is dependent on the maximum number of iterations allowed (I_{max}). For the results here, $I_{max} = 20$ and the accuracy of K-SVD may improve with higher number of iterations even though I did not observe significant improvement with $I_{max} = 100$. Even at 20 iterations K-SVD is significantly slower than $\mathcal{RC}web$ and emPCA, Fig.4. The scaling of the algorithms

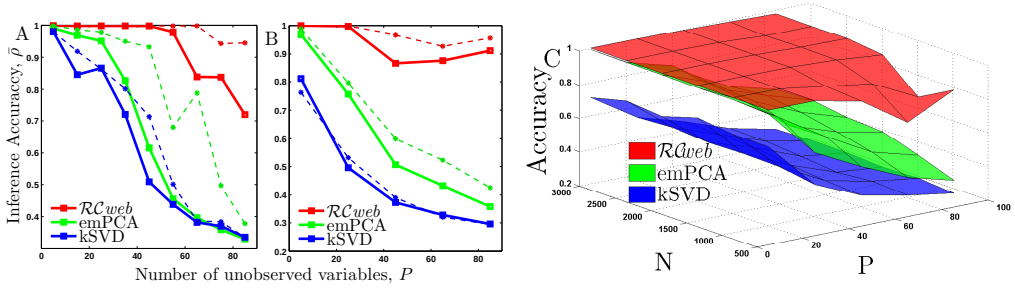

Figure 3.2: Accuracy of network recovery as a function of the number of unobserved variables, P. The thicker brighter lines with squares correspond to number of observed variables $N = 500$ and the dashed, thinner lines correspond to $N = 1000$. In all cases the number of observations is $M = 2000$. A) Poisson networks with mean out–degree $0.25N$ {125 and 250}, with 50 % noise in the observations and B) power–law networks (with mean out–degree $0.40N$ {200 and 400}, with 10 % noise in the observations. C) The same as (B) except for wider range of the observed variables.

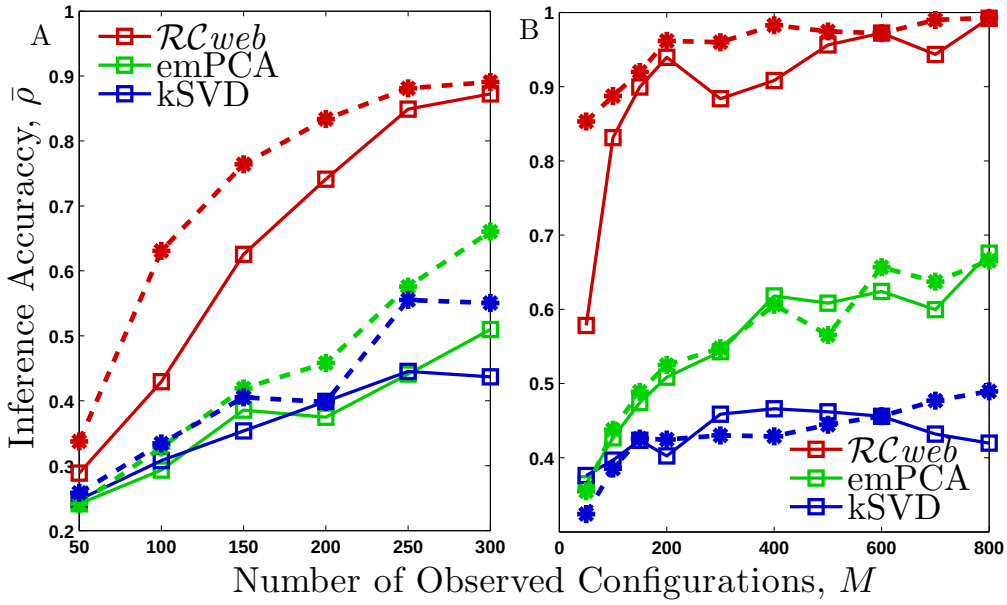

Figure 3.3: Accuracy of network recovery as a function of the number of observed configurations, M. Continuous lines & squares, $N = 500$; dashed lines & circles, $N = 1000$ A) Poisson networks with mean out–degree $0.20N$ {100 and 200}, with 50 % noise; B) power–law networks (with mean out–degree $0.40N$ {200 and 400}, with 10 % noise. In all cases $P = 30$.

with respect to a parameter was determined by holding all other parameters constant and regressing the log of the CPU time against the log of the variable parameter, Fig.4. The scaling with respect to some parameters is below the theoretical expectation since the

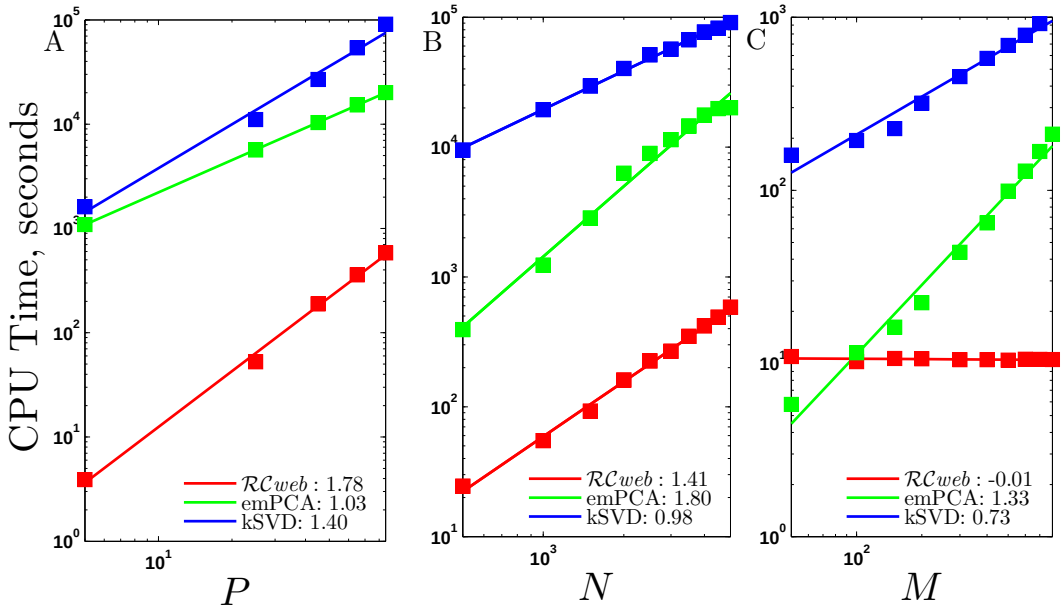

Figure 3.4: Computational efficiency as a function of P, M and N. The scaling exponents (slopes in log–log space) are reported in the legends. The networks are with power–law degree distribution and 60 % sparsity, with mean out–degree 0.40N

highest complexity steps may not be speed–limiting for the ranges of parameters used in the simulations.

3.6.3 Interpretability

In analyzing real data the number of unobserved variables (P) may not be known. I observe that if a simulated network having P regulators is inferred assuming $P^* > P$ regulators, the elements of $\mathbf{B}$ corresponding to the excessive regulators are very close to zero, $|B_{ij}| \leq 10^{-10}$ for $i > P$ or $j > P$. Thus, if the data truly originate from a sparse network $\mathcal{RC}web$ can discard unnecessary unobserved variables.

The results of $\mathcal{RC}web$ can be valuable even without identifying the physical factors corresponding to $x_{\mathcal{R}}$. Yet identifying this correspondence, and thus overcoming the limitations outlined in section 3.6.1 can be very desirable. One practically relevant

situation allowing in-depth interpretation of the results requires measuring (if only in a few configurations) the states of some of the variables that are generally unobserved ($x_{\mathcal{R}}$). This is relevant, for example, to situations in which measuring some variables is much more expensive than others (such as protein modifications versus messenger RNA concentrations) and some $x_{s\in\mathcal{R}}$ can be observed only once or a few times. Assume that the states of the k^{th} physical factor are measured (data in vector $\mathbf{u}_k$) in n_k number of configurations, whose indices are in the set ϕ_k. This information can be enough for determining the vertex $x_{s\in\mathcal{R}}$ corresponding to the k^{th} factor and the corresponding $\hat{\mathbf{D}}_{ss}$ as follows: 1) Compute the Pearson correlations $\vec{\rho}$ between $\mathbf{u}_k$ and the columns of $\hat{\mathbf{R}}_{\phi_k}$. Then, the vertex of the inferred network most likely to correspond to the k^{th} physical factor is $s = \arg\max_i(|\rho_1|, \ldots, |\rho_i|, \ldots, |\rho_P|)$. 2) $\hat{\mathbf{D}}_{ss} = (\hat{\mathbf{R}}_{\phi_k}^T \hat{\mathbf{R}}_{\phi_k})^{-1} \hat{\mathbf{R}}_{\phi_k}^T \mathbf{u}_k$. Similar to section (3.6.1), weighted matching algorithms (Sanghavi 2007) can be used for finding the optimal solution if there is data for multiple $x_{s\in\mathcal{R}}$.

Partial prior knowledge about the structure of $\mathcal{G}$ can also be used to enhance the interpretability of *RCweb* results. Assume, for example, that some of the nodes ($x_{s\in\alpha_k\subset\mathcal{N}}$) regulated by the k^{th} physical factor (which is a hidden variable in the inference) are known. Then the matching approach that was just outlined can be used with $\hat{\mathbf{C}}$ rather than $\hat{\mathbf{R}}$. If the weights are not known all non-zero elements of $\hat{\mathbf{C}}_{\alpha_k}$ can be set to one. The significance of the overlap (fraction of common edges) of the regulator most likely to correspond to the k^{th} physical and its known connectivity (coming from prior knowledge) can be quantified by a p-val (the probability of observing such overlap by chance alone) computed from the hyper–geometric distribution. This approach is exemplified with gene-expression data in the next section.

Conclusions

I introduce an approach ($\mathcal{RC}web$) for inferring latent (unobserved) factors explaining the behavior of observed variables. $\mathcal{RC}web$ aims at inferring a sparse bipartite graph in which vertices connect inferred latent factors (e.g. regulators of mRNA transcription and degradation) to observed variables (e.g. target mRNAs). The salient difference distinguishing $\mathcal{RC}web$ from prior related work is a new approach to attaining sparse solution that allows the natural inclusion of a generative model, relaxation of assumptions on distributions, and ultimately results in more accurate and computationally efficient inference compared to competing algorithms for sparse data decomposition.

3.7 Application to the Growth Rate Response Data

Simulated models have the advantage of having known topology, and thus providing excellent basis for rigorous evaluation. Yet the real opportunities to the application of *RCweb* are on real data where *RCweb* has the potential to transcend the descriptive macroscopic level of analysis (Slavov and Dawson, 2009). For real biological networks, however, such rigorous evaluation is *only* possible based on experimental testing. *RCweb* results in quantitative models that can be tested in a variety of different ways including predicted network structure, TF activities, dynamical responses of mRNA levels from perturbations and transcriptional effects of deleting and overexpressing genes. *RCweb* was conceived and designed to be used in conjunction with experimental testing of the results.

Since I have not had the opportunity to experimentally test the *RCweb* predictions, I will confine this section to a very short evaluation of the results based on the existing partial knowledge of transcriptional networks. In particular, I will use the approach outlined in section 3.6.3 for *partially* evaluating inference results from the growth rate data. I consider such evaluation preliminary and insufficient to demonstrate the power of *RCweb*.

I start by normalizing the growth rate response data for all 45 datasets on ethanol and glucose carbon source to z–scores so that the vector of expression levels for each gene has a mean zero and a unit variance. *RCweb* was initialized with $P^* = 24$ unobserved variables (corresponding to regulators of mRNA levels) and identified $P = 22$ co-regulated sets of genes in less than 3 seconds compared to several hours needed by FIRE. To evaluate whether some of those regulators correspond to known transcription factors, the inferred adjacency matrix $\hat{\mathbf{C}}$ is compared directly to the adjacency matrix identified by ChIP-chip experiments and published by MacIsaac *et al* (2006) using *weighted–matching*

that resulted in an optimal solution. From sets of genes inferred by *RCweb* to be co-regulated, 13 sets overlap significantly with sets of genes found to be regulated by TFs in ChIP-chip studies with the largest p value for those 13 sets being 2.3×10^{-8} and the smallest below the numerical precision of my computations. The inferred activities of those TF are shown in Fig.3.5:

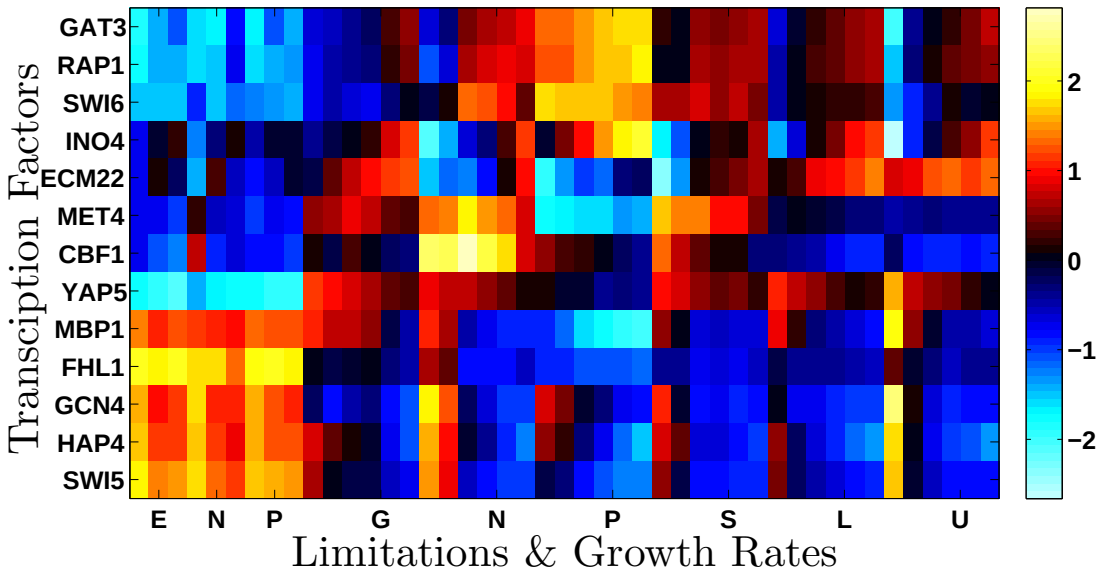

Figure 3.5: Activities of TFs Inferred by *RCweb*. Since the scale is arbitrary, all activities where normalized to z–scores. The first 9 columns correspond to 3 limitations (Ethanol (E), Nitrogen (N), Phosphor (P)) and each limitation has 3 growth rates ordered from slowest to fastest growth $\mu = \{0.05, 0.10, 0.14\}h^{-1}$, for ethanol carbon source and the remaining 36 correspond to 6 limitations (Glucose (G), Nitrogen (N), Phosphor (P), Sulfur (S), Leucine (L) and Uracil (U)) and each limitation has 6 growth rates ordered from slowest to fastest growth $\mu = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}h^{-1}$, for glucose carbon source. The results are subjects to the limitations discussed in section 3.6.1.

Many of the results on Fig.3.5 are consistent with biological expectations. For example, *HAP4* is much more active in ethanol carbon source than in glucose carbon source where it is active only in the slowest growth rates. Furthermore, *HAP4* has positive growth rate response in ethanol carbon source and negative growth rate response in glucose, again consistent with the different oxidative demands of cells growing on ethanol and glucose. Similar pattern of activities are inferred for *SWI5* and *GCN4*. Not

surprisingly, many of the inferred TF activities show strong growth rate dependence which is easiest to understand in the case of *RAP1* increasing in activity with growth rate for all limitations on glucose carbon source. The growth rate induction of the ribosomal genes in ethanol carbon source seems to be regulated primarily by *FHL1* whose activity is both higher and having positive growth rate slopes in ethanol carbon source.

The results from the $\mathcal{RC}$*web* inference are not limited to TF activity but also include identified potential regulators of mRNA degradation, new TF targets, patterns of combinatorial regulation and a model that can predict gene expression changes resulting from perturbations of regulators. I will not discuss those, however, before I am able to evaluate them rigorously.

Chapter 4

Growth Rate Response, *YMC* and Cell Cycle

4.1 Introduction

Brauer *et al* (2008) made the qualitative observation that genes expressed during different phases of the yeast metabolic cycle (*YMC*) (Klevecz *et al*, 2004; Tu *et al*, 2005) have growth rate response slopes (on glucose carbon source) that tend to be either positive or negative. In particular, genes expressed during the oxidative phases have positive slopes while genes expressed during reductive phases have negative slopes (Brauer *et al*, 2008). Much of the work outlined in this chapter is under development and the summary here is a succinct survey of recent progress rather than an exhaustive description of completed work.

The second section of this chapter focuses on demonstrating the existence of the *YMC* in single cells from non–synchronized populations and methods to analyze the data and maximize the extracted information while minimizing assumptions. In particular, I focus

on the mathematical problem of inferring high dimensional dynamical trajectories (the expected mRNA counts for all measured genes) from low–dimensional (just a few genes per experiment) time disordered observations (cells whose phases in the metabolic cycle are not known).

The third section of this chapter builds upon the qualitative observation by Brauer *et al* (2008) in developing a quantitative model of growth rate response based on the *YMC*. Furthermore, I show that model predictions are consistent with experimental measurements in synchronized cultures and discuss how the model can be used to propose and test the mechanistic connection between the growth rate response, the *YMC* and the cell cycle.

4.2 *YMC* in Single Cells

4.2.1 Correspondence between Correlations

The correspondence between gene–gene correlations in continuous yeast populations synchronized with respect to the *YMC* and in single cells from non–synchronized populations is evidence for the existence of the *YMC* even in non–synchronized cells (Silverman *et al*, 2010; Slavov *et al*, 2012, 2013). The image processing and statistical treatment which enabled this work is available (upon request) as a draft manuscript. Here I will present just the "tip of the iceberg" summary: A scatter plot depicting the correspondence of gene–gene correlations in synchronized populations and in single cells based on the data from the most recent versions of my image processing algorithms, Fig.4.1. Since the Poisson distribution is not symmetric (skewed to the right), uncorrelated Poisson noise can introduce systematic bias in the estimates of correlations. To minimize such a bias I used two approaches:

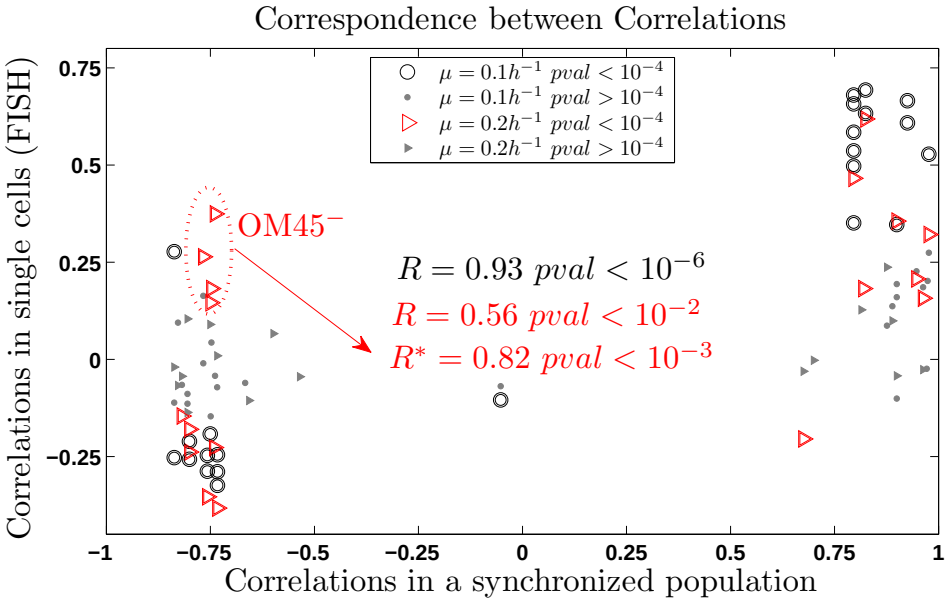

Figure 4.1: Correspondence between gene–gene correlations in continuous yeast populations synchronized with respect to the *YMC* (Tu *et al*, 2005) and in single cells from non–synchronized populations at growth rates $\mu = 0.10h^{-1}$ and $\mu = 0.20h^{-1}$. The significance of the correlations computed from the single cell data is reflected in the color and size of the markers. As an overall measure of the correspondence, I compute the correlation of correlations R using all significant ($p < 10^{-4}$) correlations and in the case of $\mu = 0.20h^{-1}$, I also compute R^* which excludes correlations including *OM45* since they are all obvious outliers.

1. The correlations in single cells were computed using only cells that have at least one mRNA.

2. The correlations in single cells were computed using only cells in *YMC* phases in which the mRNAs are expressed. To select such a subset of cells, I use a Poisson mixture model. In this model the observed cells may come from two Poisson distributions. To estimate the parameters of those distributions and assign cells to them, I use an expectation maximization (EM) approach described in the Appendix section 5.5.

On Fig.4.1, I display the results only from the first approach (excluding cells with no mRNAs) because they are very similar to the results from using the Poisson mixture model and are simpler to explain and understand.

4.3 *YMC*, Cell Cycle and Growth Rate Response

4.3.1 Non–synchronized Cultures

4.3.2 Introduction

Consider an non–synchronized population of yeast (or another microorganism) growing exponentially in a chemostat at a growth rate μ. After about 10 generations in the chemostat, measurements of biomass, bud index, and residual nutrients stop changing, which can be interpreted as a sign of steady–state. The individual cells, however, are not at steady–state. Based on the classical work by Hartwell (1974); Hartwell *et al* (1974) and more recent work by Silverman *et al* (2010), we know that individual cells go through cell cycles as well as yeast metabolic cycles (*YMC*) in the case of yeast. Therefore, the expression level of the i^{th} gene measured in a population $(\hat{G}_i)$ depends on the expression levels of the i^{th} gene during each phase of the cycles and on the fraction of cells in each phase of the cycles. Furthermore, if we assume that the population is non–synchronized with respect to the cycles, the number of cells in a phase is proportional to the duration of that phase. These dependencies can be used to derive a quantitative expression (4.1) for the level of the i^{th} gene measured in a non–synchronized population $(\hat{G}_i)$:

$$\hat{G}_i = \frac{1}{T}\int_{t=0}^{t=T} G(t)dt \approx \frac{1}{\sum_{j\in\omega} T_j}\sum_{j\in\omega} T_j G_{ij} \tag{4.1}$$

In the time continuous regime, $\hat{G}_i$ is the integral of expression levels during a cycle period divided by the duration of a cycle period. In the time discrete regime, $\hat{G}_i$ is the superposition (over the set of all phases ω) of expression levels during each phase (G_{ij}) weighted by the phase durations (T_j).

Given (4.1), one would expect that changes in the phase duration will affect the gene expression levels measured in non–synchronized populations. I will examine briefly the expectations for such changes starting with the cell cycle since 1) the budding index data from section 1.4.2 gives me direct measurement of the fraction of time cells spend in *G0/G1* verses *S/G2/M* phases, and thus I can compare the expectation based on (4.1) to the gene expression data; 2) The cell–cycle has been known for half a century, its better studied than the *YMC* and its existence has been verified by many researchers in many different systems.

4.3.3 Cell–Cycle

I am going to use the budding index as and indicator of the fraction of cell in different phases of the cell cycle. In particular, the fraction of budded cells equals the ratio of time cells spend in *S/G2/M* verses *G0/G1* phases. Based on that, I will estimate an expectation for the slopes of *S/G2/M* genes in the data by Brauer *et al* (2008) as follows:

1. At the slowest growth rate $\mu = 0.05h^{-1}$, the ratio of duration of *S/G2/M* verses *G0/G1* phases is:

   $$\mu = 0.05h^{-1} \mapsto \underbrace{(T_S + T_{G2} + T_M)}_{budded} \quad / \quad \underbrace{T_{G1}}_{non-budded} = 0.18$$

2. At the fastest growth rate $\mu = 0.30h^{-1}$, the ratio of duration of *S/G2/M* verses *G0/G1* phases is:

   $$\mu = 0.30h^{-1} \mapsto \underbrace{(T_S + T_{G2} + T_M)}_{budded} \quad / \quad \underbrace{T_{G1}}_{non-budded} = 0.80$$

3. If cell cycle genes specific to a phase of the cell–cycle are expressed only during that phase and their expression is regulated entirely by the cell–cycle the expected slopes for *S/G2/M* genes should be distributed around:

   $\mathbb{E}[Slope] = \log_2(0.8/0.18)/(0.3 - 0.05) = \mathbf{9}$

This expectation is inconsistent with the gene expression data in which cell cycle genes have slopes centered around zero (Brauer *et al*, 2008). This inconsistency indicates that the assumption at step (3) about cell–cycle genes being regulated only by the cell cycle is incorrect. In fact, FISH images with labeled cell–cycle genes demonstrate that indeed non–budded cells express cell–cycle genes. Furthermore, FISH data and cultures synchronized with respect to the *YMC* at different growth rates suggest that the *YMC* may counterbalance the effect of the cell–cycle. Before examining this possibility in more detail in the next subsection, consider the same type of slope estimate for cultures growing only on ethanol carbon source:

1. $\mu = 0.05h^{-1} \mapsto (T_S + T_{G2} + T_M)/T_{G1} = 0.15$
2. $\mu = 0.14h^{-1} \mapsto (T_S + T_{G2} + T_M)/T_{G1} = 0.28$
3. $\mathbb{E}[Slope]$ for ***S/G2/M*** $\log_2(0.28/0.15)/(0.14 - 0.05) = \mathbf{10}$

The slopes for cell-cycle genes computed from the gene expression data are negative indicating even larger differences between the expected and the measured trends in the growth rate response of cell–cycle genes.

4.3.4 *YMC*

There are many ways to study experimentally whether and how the *YMC* changes with growth rate, including the reconstruction of dynamical trajectories discussed in section **??**. I first consider the growth rate effect in a *YMC* synchronized population as measured directly by the level of dissolved oxygen, $d[O_2]$. The raw data for a glucose limited culture grown at $\mu = 0.10h^{-1}$ and $\mu = 0.05h^{-1}$ are shown on Fig.4.2. It is hard to estimate the change in shape of the oscillation and the duration of the different phases from Fig.4.2. To facilitate that, I scale the duration of one cycle for $\mu = 0.10h^{-1}$ so

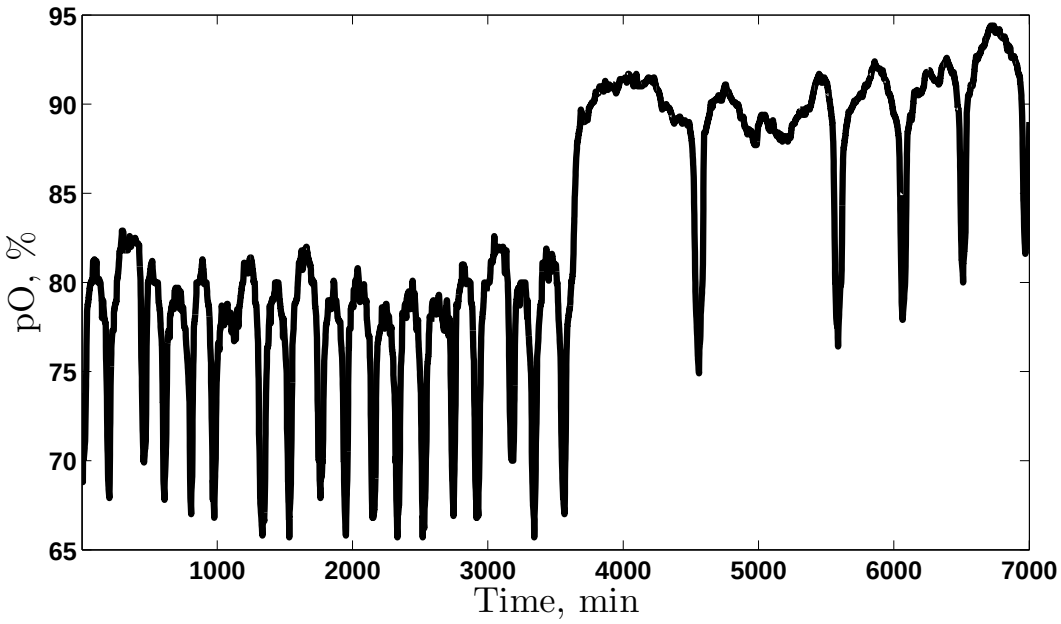

Figure 4.2: Changes in the *YMC* with growth rate. A culture synchronized at $\mu = 0.10h^{-1}$ was shifted to $\mu = 0.05h^{-1}$.

that it is equal to the duration of one cycle at $\mu = 0.05h^{-1}$ and plot 3 cycles for each growth rate one above the other, Fig.4.3. The traces in Fig.4.3 indicate that at the faster growth rate, $\mu = 0.10h^{-1}$, the relative duration of the oxidative phase is longer compared to the slower growth rate $\mu = 0.05h^{-1}$. To quantify this difference, I divide the first cycle into two parts at mid–height into an oxidative phase (high oxygen consumption) and at mid–height into a reductive (low oxygen consumption) phase. The ratio between the duration of the oxidative phase to the duration of the reductive phase is denoted by R. For $\mu = 0.10h^{-1}$ (doubling time $7h$) the ratio is denoted by R_{7h} and for $\mu = 0.05h^{-1}$ (doubling time $14h$) the ratio is denoted by R_{14h}. Remarkably, $R_{7h} = 2R_{14h}$ indicating that the duration of the oxidative phase is proportional to the growth rate and based on these data the coefficient of proportionality is 1.0.

Fig.4.4 shows data from additional experiments with metabolically synchronized cultures of a *WT S288C HAP*1^{+} strain supporting the conclusion that the relative duration

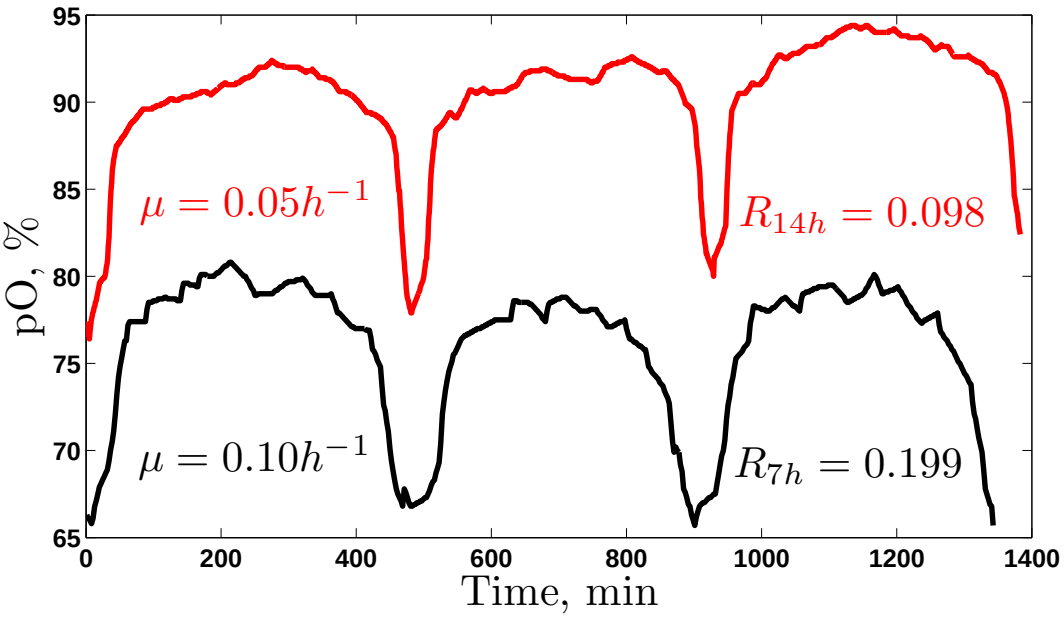

Figure 4.3: Changes in the *YMC* with growth rate in metabolically synchronized cultures of a haploid MATa *WT S288C HAP*1^{+} strain. Three cycles from the data on Fig.4.2 are shown for both $\mu = 0.10h^{-1}$ and $\mu = 0.05h^{-1}$. The period of the $\mu = 0.10h^{-1}$ cycle was scaled to be the same as the period of the $\mu = 0.05h^{-1}$. The ratio between the duration of the oxidative phase (high oxygen consumption) to the reductive (low oxygen consumption) phases is denoted by R. For $\mu = 0.10h^{-1}$ (doubling time $7h$) the ratio is denoted by R_{7h} and for $\mu = 0.05h^{-1}$ (doubling time $14h$) the ratio is denoted by R_{14h}.

of the high oxygen consuming *YMC* phase increases with the growth rate and expanding the dynamical ranges of growth rates.

Scaling the duration of the metabolic cycle is useful in emphasizing growth rate induced changes in the relative durations of its phases but it does not reveal another important parameter: the *YMC* period. Fig.4.5 shows how the period of the *YMC* changes with growth rate and the period of the cell–cycle:

Phenomenologically, the results on Fig.4.5A reconcile and explain the differences in *YMC* frequencies reported by Klevecz *et al* (2004) and by Tu *et al* (2005). More interestingly, the linear dependence between the periods of the *YMC* and the cell–cycle is reminiscent to the observation made first by Hartwell (1974); Hartwell *et al* (1974) that as cells grow slower (and thus increase the duration of their cell–cycle periods), the extra

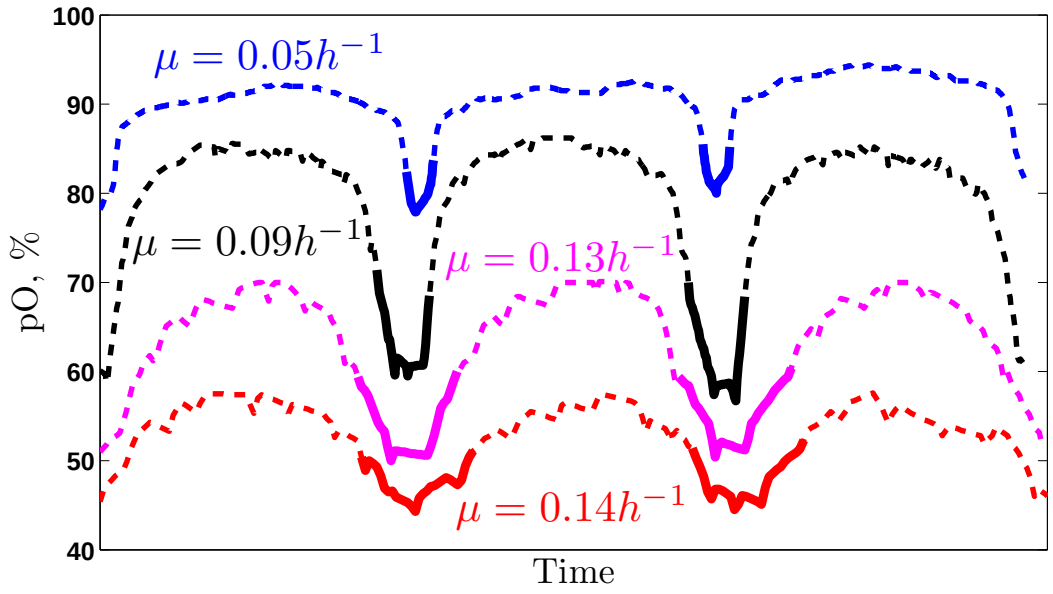

Figure 4.4: Changes in the *YMC* with growth rate. Three cycles from from metabolically synchronized cultures of the strain at different growth rates are scaled to have the same period for emphasizing the change in relative durations of the *YMC* phases. The *YMC* periods are shown in Fig.4.5 as a function of both the growth rate and the cell cycle period. For emphasizing the difference in relative durations, the oxidative (high oxygen consumption phase) is plotted in thicker solid trace.

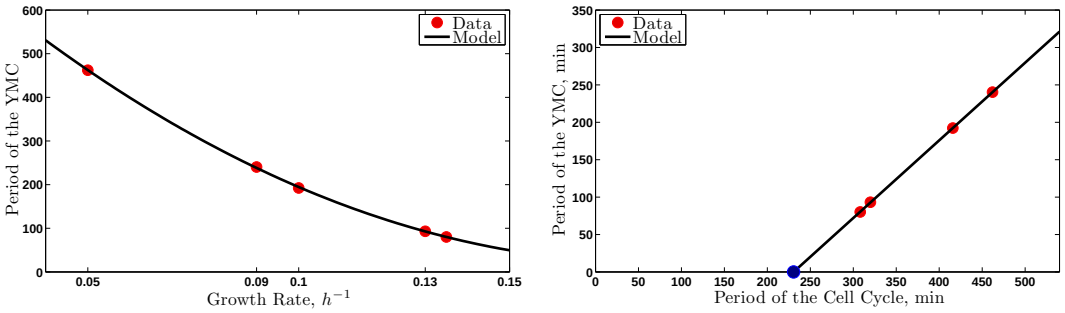

Figure 4.5: The period of the *YMC* as a function of growth rate μ (A) and the cell–cycle period (B) in metabolically synchronized cultures of a diploid *WT S288C HAP*1^{+}

time is spent in *G0/G1* stage. In the case of the *YMC*, the increase in the reductive (low–oxygen consumption) stage with the increase in the cell cycle period is likely to be due entirely to increase in the *YMC* phase with high expression of autophagy genes while the durations of the oxidative and the DNA replication phases of the *YMC* remain constant.

Given that the relative durations of the *YMC* change with growth rate, equation (4.1) predicts that the gene expression levels measured in non–synchronized populations should be growth rate dependent for genes expressed periodically during the *YMC*. Indeed, Fig.4.6 shows that all genes with universal growth rate response are expressed periodically during the *YMC*. Furthermore, Fig.4.6 indicates that genes expressed at higher

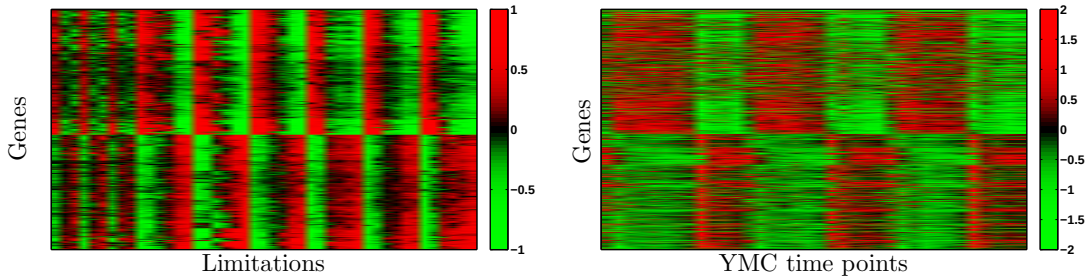

Figure 4.6: *GRR* Genes are Periodic in the *YMC*. The left panel depicts mean centered expression levels of *GRR* genes. Expression levels of the genes with best fits to my *GRR* model are normalized to mean zero for each limitation and clustered. The first 9 columns correspond to ethanol carbon source and limitations on ethanol, nitrogen and phosphor, 3 growth rates per limitation arranged in ascending order, $\mu = \{0.05, 0.10, 0.14\}h^{-1}$. The next columns correspond to glucose carbon source and limitations on glucose, nitrogen, phosphor, sulfur, uracil and leucine, 6 growth rates per limitation arranged in ascending order $\mu = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}h^{-1}$. The right panel depicts the expression levels (normalized to z–scores) of the same genes in the *YMC* (Tu *et al*, 2005). The genes in both panels are clustered based on their expression levels in the growth rate experiments and thus genes (rows) from the left panel correspond to the genes (rows) in the right panel. The over–represented GO terms for the genes with positive and negative slopes can be found in tables 2.2 and 2.1 respectively.

levels during the oxidative phases have positive slopes (expressed more highly at the higher growth rates) while genes expressed at higher levels during the reductive phases have negative slopes (expressed more highly at the slower growth rates) as expected from the data on Fig.4.3 and equation (4.1). There are no exceptions. All genes with universal growth rate response are periodic in the *YMC* and have the predicted growth rate response. Interestingly, the most overexpressed genes in cultures starving in glucose are all genes expressed during the reductive phase further suggesting a connection between the rate f growth (or lack of growth th this case) and the *YMC*.

I use equation (4.1) to build a simple model based on this remarkable correlation between the *YMC* and the universal growth rate response. The model uses the gene expression data measured by Tu *et al* (2005) in matrix $\mathbf{\Omega} \in \mathbb{R}^{1500\times 12}$ whose rows correspond to the 1500 genes with universal growth rate response and the columns correspond to the 12 time points in the *YMC* measured by Tu *et al* (2005). Each time point is the arithmetic average from the three cycles that Tu *et al* (2005) measured. I also use data from my ethanol carbon source experiments and from Brauer *et al* (2008) in a matrix $\mathbf{\Phi} \in \mathbb{R}^{1500\times 45}$ whose rows correspond to the 1500 genes with universal growth rate response and the columns correspond to the 45 growth rate conditions, 9 on ethanol carbon source and 36 on glucose carbon source. Applied to these data, the matrix form of equation (4.1) is simply:

$$\mathbf{\Omega C} = \mathbf{\Phi} \tag{4.2}$$

Here $\mathbf{C} \in \mathbb{R}^{12\times 45}$ is a matrix whose elements indicate the relative durations (fractions of the period of of the *YMC*) of the *YMC* sections corresponding to the 12 time points for all 45 cultures limited on different nutrients and growing at different growth rates. Assuming Gaussian distribution for the measurement errors, the maximum likelihood (*ML*) solution of this overdetermined problem (4.2) is:

$$\hat{\mathbf{C}} = (\mathbf{\Omega}^T\mathbf{\Omega})^{-1}\mathbf{\Omega}^T\mathbf{\Phi} \tag{4.3}$$

In the *ML* solution, some elements of $\hat{\mathbf{C}}$ are negative which has no physical meaning and may result from factors ranging from the trivial (high degree of co–linearity in the gene expression data) to more substantive reasons such as the growth rate response being a function not only of the *YMC* but also being regulated by other mechanisms. To avoid negative elements in $\hat{\mathbf{C}}$, I add the constraint $C_{ij} > 0$ and solve the resulting quadratic programming (*LP*) problem with a interior point method.

Similarly, one can build a model for cell–cycle genes and use the inference results to identify similarities and differences in the *YMC* in different nutrient limitations as well as the coupling between the *YMC* and the cell–cycle. The results can be compared to and evaluated with respect to the results from inferring dynamical trajectories from FISH data. Since the two approaches use different data and are orthogonal in many ways, they complement each other well. This is work in progress. I have both encouraging results and unresolved challenges.

4.3.5 *YMC* and Cell Cycle

The discrepancy between expected and observed slopes for the cell cycle genes can be resolved when the *YMC* is taken in into account. Indeed, Tu *et al* (2005) showed that cell–cycle genes are expressed during the reductive phases of the *YMC* whose relative duration is inversely proportional to the growth rate, Fig.4.3. Thus a correct expectation based on (4.1) should incorporate the *YMC* as well. Qualitatively, it is clear that the effects of the cell–cycle and the *YMC* are going to counterbalance each other since the durations of their phases with high expression of cell–cycle genes change in the opposite direction with growth rate. The exact quantitative resolution likely to come from applying *emDyn* and (4.1) is in progress.

4.3.6 Respiration in Cultures not Limited on Glucose

One of the interesting observation by Brauer *et al* (2008) is that cultures limited on natural nutrients do not ferment excess glucose. This observation might be related to the existence of an intrinsic *YMC* in non–synchronized cells limited on natural nutrients such as phosphate, (Silverman *et al*, 2010). Indeed, cells undergoing the *YMC* do not waste glucose and do not generate ethanol during most of phases of the *YMC* as shown

by Tu *et al* (2005) and confirmed by my measurements of residual glucose and ethanol in *YMC* synchronized cultures. A testable prediction of this supposition is that cultures not limited on any nutrients do not have an intrinsic *YMC*. To test this predictions, we use the method applied by Silverman *et al* (2010) to measure single cell gene–gene correlations in yeast cultures growing in excess nutrients in early phase batch cultures.

4.3.7 Metabolically Synchronized Cultures

There are many open questions regarding the exact mechanisms of *YMC* synchronization of yeast populations and the relationship between the *YMC* in such populations and the *YMC* in single cells (Silverman *et al*, 2010). Klevecz *et al* (2004) and Tu *et al* (2005) have demonstrated experimentally a relationship between the cell–cycle and the *YMC* but the exact nature and significance of the connection has not been established. The working hypothesis is that DNA replication is limited to the reductive phases of the *YMC* to limit DNA damage, (Tu *et al*, 2005; Klevecz *et al*, 2004). In this section, I present experimental data from *YMC* synchronized populations that are consistent with the possibility that the *YMC* observed in populations is emergent behavior from the coupling of the cell–cycle and the intrinsic *YMC*. Those two non–linear oscillators can be by a simple Duffing oscillator which exhibits the wide range of possible dynamical behaviors observed in *YMC* synchronized populations and some of which I outline below.

My attempts to metabolically synchronize chemostat cultures have significantly higher success rate with the *WT CEN.PK* strain compared to the haploid *WT S288C HAP*1^{+} strain. All results reported in this subsection are for the haploid *WT S288C HAP*1^{+} strain. I have been able to metabolically synchronize the haploid *S288C* only at growth rates $\mu = \{0.10, 0.05\}h^{-1}$ in which the period of the cell–cycle is equal to integer multiple of the period of the *YMC* observed in the population. Attempting to synchronize a population at a growth rate $\mu = 0.12h^{-1}$, which is slightly higher than the growth rate at which the *YMC* oscillations are stable ($\mu = 0.10h^{-1}$), results in initial synchrony that is gradually lost, Fig.4.7. In contrast, I have been able to metabolically synchronize the diploid *S288C* at faster growth rates. Even more interesting is the result when a population synchronized at $\mu = 0.10h^{-1}$ is shifted to a higher growth rate $\mu = 0.20h^{-1}$, Fig.4.8. The simple periodic waveform of the oxygen trace $p[O_2]$ becomes rather complex at $\mu = 0.20h^{-1}$

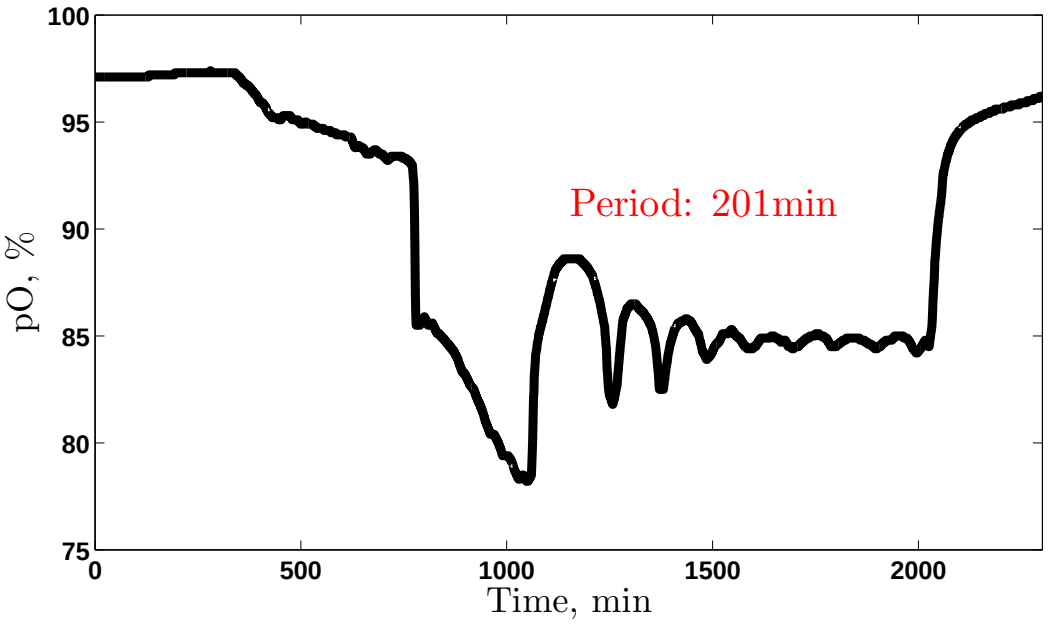

Figure 4.7: *YMC* at $\mu = 0.12h^{-1}$

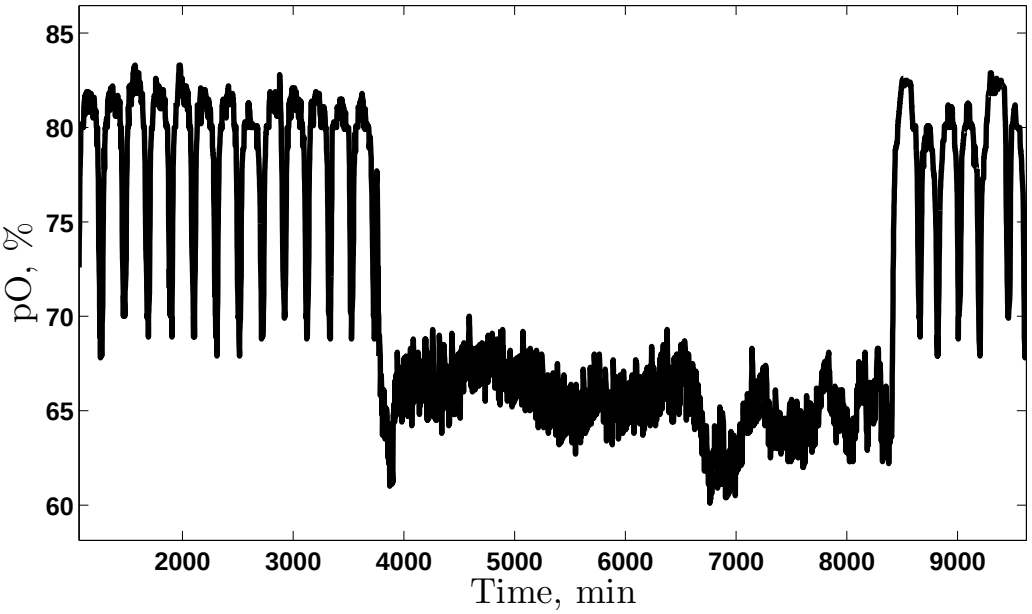

Figure 4.8: Growth Rate Transitions in the *YMC*. A population synchronized at $\mu = 0.10h^{-1}$ is shifted to a higher growth rate $\mu = 0.20h^{-1}$ and then back to $\mu = 0.10h^{-1}$

and recovers back to simple periodic waveform as soon as the growth rate is restored back to $\mu = 0.10h^{-1}$. To identify if the complex waveform at $\mu = 0.20h^{-1}$ has similar

frequencies to the frequencies found in the periodic waveform I perform discrete cosine transform of the time series data, Fig.4.9. While the general shapes of the harmonics are

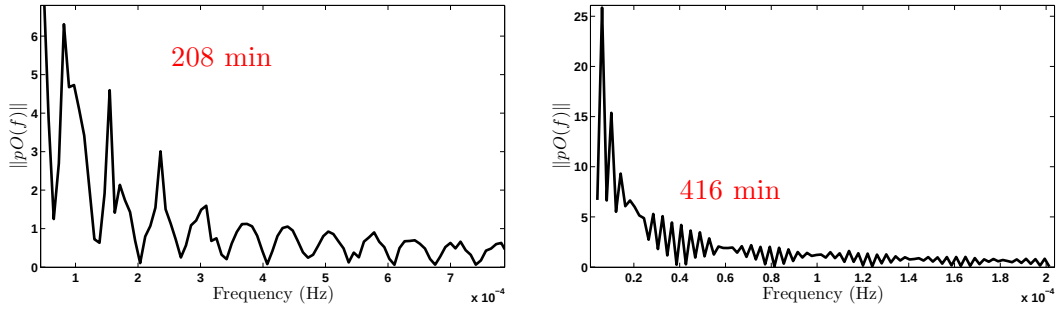

Figure 4.9: Power Spectra of the *YMC*. The left panel displays the power spectrum for $\mu = 0.10h^{-1}$ and the right panel displays the power spectrum for $\mu = 0.20h^{-1}$ for the oxygen traces shown in Fig.4.8

different for $\mu = 0.10h^{-1}$ and $\mu = 0.20h^{-1}$, they are related by an integer scaler of 2.

Interestingly, the power spectra of the complex waveforms on glucose and ethanol carbon source look very similar and differ only by a scaler multiple of the harmonic frequencies, Fig.4.10.

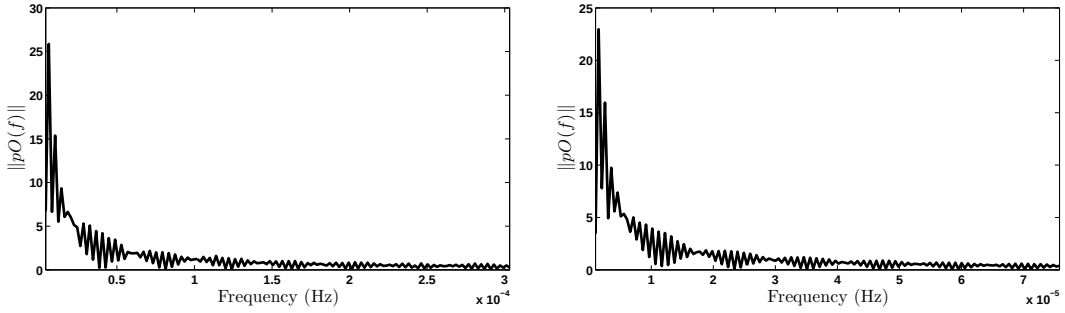

Figure 4.10: Power Spectra in Glucose and in Ethanol. The left panel displays the power spectrum for a complex waveform in glucose carbon source $\mu = 0.10h^{-1}$ and the right panel displays the power spectrum for a complex waveform in ethanol carbon source $\mu = 0.10h^{-1}$.

Key observations suggesting the emergent nature of the *YMC* in metabolically synchronized cultures include:

- Some of the differences between $d[O_2]$ waveforms in metabolically synchronized populations can be traced to a single parameter that has the expected effect. For

example, the amplitude of the oscillations depends strongly on the rate of air flow, Fig.5.7. Other differences, however, such as the dependence of the frequency and shape of the waveform on the nutrient composition of the media and biomass density of the population are harder to explain. A possible explanation is that changes in nutrients and biomass affect the strength of coupling between the *YMC* and the cell–cycle, which then affects the emergent synchrony of the population reflected in the measured $d[O_2]$.

- In some ranges of parameters, variation of biomass density and growth rate changes qualitatively the culture dynamics from regular periodic oscillations (limit ycle) through bursting to dynamics that appear chaotic.

- Spontaneous transitions between different $d[O_2]$ waveforms in metabolically synchronized *CEN.PK*.

- The marginal and joint distributions of mRNAs derived from FISH data from synchronized populations indicate that only a subset of the population is synchronized, Fig.5.8

Chapter 5

Appendix

5.1 Clustering using TSP

Despite the fact that clustering has been long–term preoccupation of the bioinformatics community, none of the numerous clustering algorithms for gene expression data that I am ware of guarantees finding the permutation of genes that minimizes the distance (computed by some metric of similarity) between all neighboring genes in a cluster. The fundamental problem is that finding such optimal permutation is an *NP*–hard problem and most of the heuristics used for gene-expression data (such as the many versions of hierarchical clustering) can result in rather suboptimal solutions. Those suboptimal solutions can contain many mistakes ranging from switching the order of large clusters to misplacing genes in clusters whose expression profiles are utterly different from the expression profiles of the genes.

In addition to the bioinformaticians, applied mathematicians have worked on a mathematical problem whose optimal solution is exactly the optimal permutation resulting in the best hierarchical clustering possible given the data and the chosen similarity metric.

This mathematical problem is known as the traveling salesman problem (*TSP*) and is one of the best studied problem in complexity theory. While no algorithm for solving the TSP can guarantee a polynomial time solution for all problems, there are algorithms (such as Concorde) that practically scale very well and can solve virtually any medium size (thousands and even tens of thousands of visited cities which for us are usually genes, metabolites or time points) problem to mathematically proven optimality within a few minutes of CPU time, (Hoos and Stutzle, 2009; Applegate *et al*, 2007b). Such algorithms have enabled, albeit more difficult, solving even large scale problems up to 85,900 cities Applegate *et al* (2007a) and much progress has been made in developing powerful algorithms during the last half century, Fig.5.1.

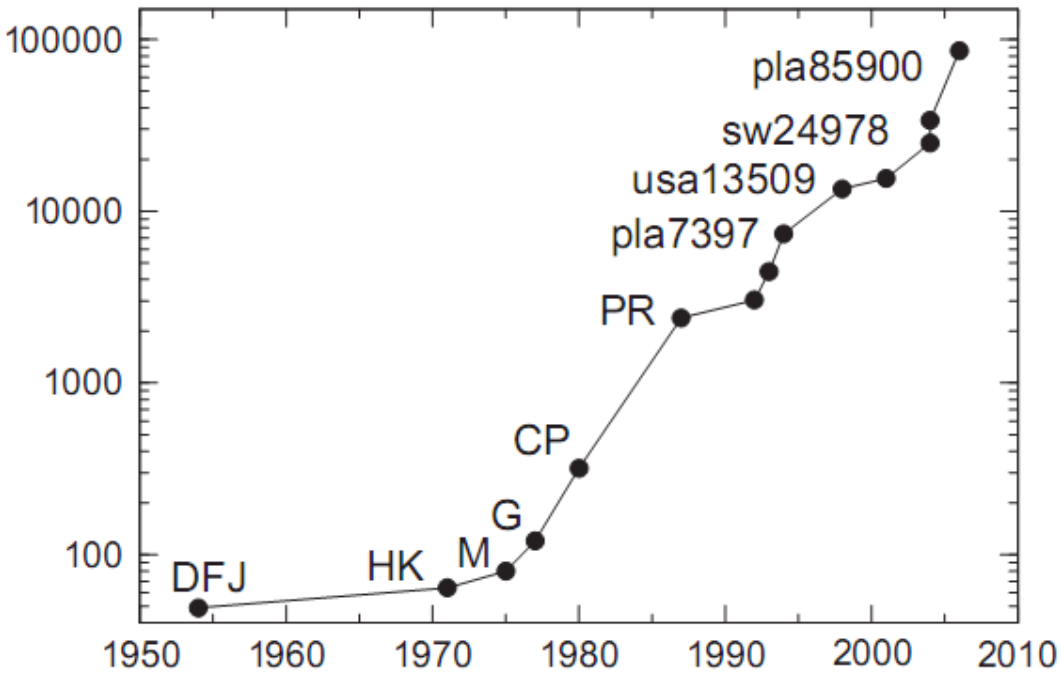

Figure 1.45 Further progress in the TSP, log scale.

Figure 5.1: TSP Progress, figure from Applegate *et al* (2007b)

Optimality proof

A complete description of the methods used for proving optimality of TSP solutions is beyond the scope of this thesis and can be found at (Applegate *et al*, 2007b). Below, I briefly outline one of the most successful and common approaches.

Once an optimal solution is found, its optimality can be proven even though the exhaustive enumeration of all possible solutions is impractical and will take many years for any super computer. The proof is often based on demonstrating equivalence between the hard combinatorial integer problem (*IP*) that we want to solve and its dial linear programming (*LP*) relaxed problem that is much easier to solve (Hoos and Stutzle, 2009; Applegate *et al*, 2007b). The TSP can be restated in graph theory terminology so that each city is a node (vertex) and each traveled distance is a link (edge) of a graph. Then finding the optimal path is a hard combinatorial optimization integer problem (*IP*) in which each edge either exists or not, and thus takes two possible discrete values, 0 or 1. However, the *IP* has a dual linear programming (*LP*) relaxation in which the vertex may belong to the interval $[01]$. In general, *LP* will not give an integer solution. When it does, however, then the *LP* solution is the optimal *IP* solution as the constraint set is strictly larger for *LP*, and includes all solutions for *IP*. Then the general procedure is to apply a family of tighter and tighter *LP* relaxations (e.g. by cutting plane methods) where non-integer solutions are iteratively eliminated until an integer solution is found. Once an integer solution is found, it is the optimal one. Alternatively the *LP* duality can be used to compute a lower bound and when the solution coincides with it, the solution is optimal.

5.2 Slopes/Exponents

Differences in slope distributions

Before considering the slopes of individual genes and sets of genes, I consider global characteristics (summary statistics) of the distributions of slopes for different limitations and discuss what might be the biological reasons for those characteristics. I start with the simplest summary statistics, the mean (first moment) and the variance (second centered moment). Then I compute the correlations between slopes (averaging across all genes) to quantify magnitude of global nutrient and carbon source effects. Furthermore, I address the question which is the relevant way of comparing slopes in glucose and ethanol carbon sources given the different ranges of growth rate possible in ethanol and in glucose carbon sources.

The mean slopes for all genes (Table 5.1) are very close to zero for all limitations. For both ethanol and glucose carbon source, the nitrogen limitation has the most negative means but in general the differences are very modest.

Carbon Source	Limitation		
	Carbon	Nitrogen	Phosphor
Ethanol	0.12	-0.12	0.12
Glucose	-0.10	-0.20	-0.19

Table 5.1: Arithmetic averages (means) for the slope across nutrient limitations and carbon sources

Much more interesting is the variation across limitations in the variance of the slope distributions, Fig.5.2. In all cases (as expected), using 3 growth rates to compute the slopes results in lower variances compared to using 6 growth rates. The interesting observation is that the nitrogen limitations have significantly higher variance for both ethanol and glucose no matter how many growth rates are used in computing the slopes. The

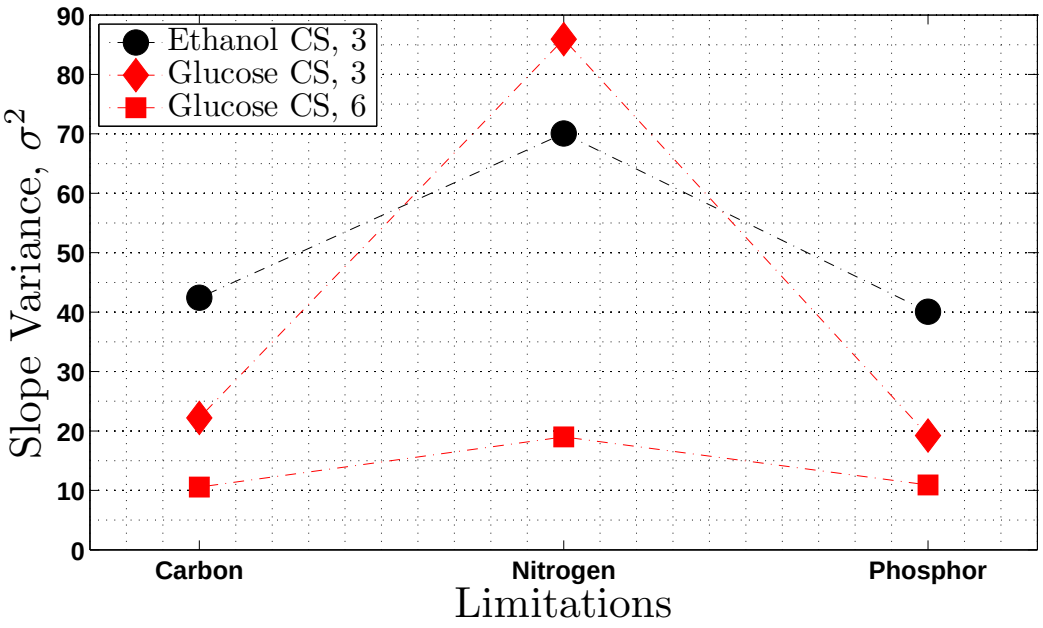

Figure 5.2: Slope variance for ethanol and glucose carbon source (CS). The numbers in the legend indicate the numbers of growth rates used in computing the slopes: $3 \mapsto \mu = \{0.05, 0.10, 0.14/0.15\}h^{-1}$, $6 \mapsto \mu = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}h^{-1}$

magnitude of the variance reflects (approximately) the magnitude of the response per unit change in growth rate. Thus, cultures limited on nitrogen have large growth rate response in mRNA levels compared to other limitations. This effect is particularly strong for the first 3 growth rates in glucose carbon source, Fig.5.2. Using the same logic and comparing the variance in slopes for the slowest 3 growth rates $\mu = \{0.05, 0.10, 0.14/0.15\}$ in ethanol and glucose, I conclude that the mRNA growth rate response is stronger in ethanol compared to glucose.

Even more interesting and informative is the Pearson correlation between slopes (again computed by averaging across all genes), Fig.5.3. First, consider the slopes computed from all 6 growth rates ($\mu = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}h^{-1}$) in glucose and the 3 growth rates $\mu = \{0.05, 0.10, 0.14\}h^{-1}$ in ethanol, Panel (A). Two obvious trends from panel (A) are that all correlations are positive (indicating some universal growth rate response) and that carbon source has a major effect of the growth rate response. Another

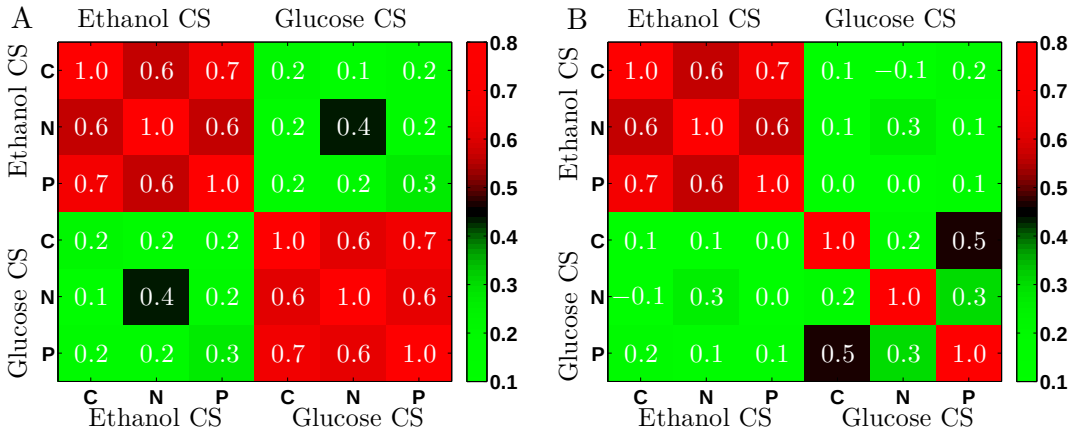

Figure 5.3: Pearson correlations between slopes for ethanol and glucose carbon source (CS). The data for ethanol is the same in both panels. For glucose, in panel (A) the slopes are computed from all 6 growth rates $\mu = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}h^{-1}$ while in panel (B) only the slower 3 growth rates were used $\mu = \{0.05, 0.10, 0.15\}h^{-1}$

salient observation is that the nitrogen limitation onces again stands out as having the most similar growth rate response. The limitations on phosphor are also more similar to each other indicating a slight PO_4^{3-} growth rate effect. If only the lowest 3 growth rates ($\mu = \{0.05, 0.10, 0.15\}h^{-1}$) on glucose (panel B) are used in computing the slopes, similarity between growth rate response is significantly lower as indicated by less positive correlations, Fig.5.3B.

Comparing the Euclidean distances between the slopes vectors (Slope in Carbon, Slope in Nitrogen, Slope in Phosphor) of each gene is another way to further verify and reinforce the conclusion that slopes in ethanol are more similar to slopes glucose when on all 6 growth rates, Fig.5.4.

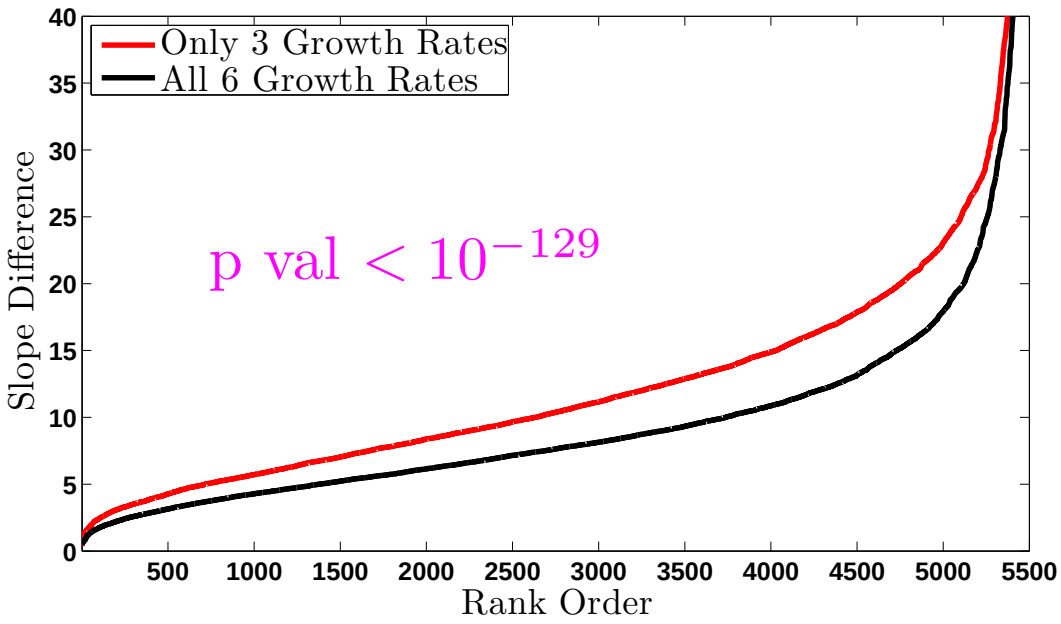

Figure 5.4: Euclidean distances between the slopes vectors for ethanol and glucose carbon source (CS). The data for ethanol is the same in both panels. For glucose, in panel (A) the slopes are computed from all 6 growth rates $\mu = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}h^{-1}$ while in panel (B) only the slower 3 growth rates were used $\mu = \{0.05, 0.10, 0.15\}h^{-1}$

5.3 Representation of non–linear functions for $\mathcal{RC}$*web*

The concentration of the j^{th} mRNA (G_j) is a function (F_j) of its Q_j regulators, $\vec{x} \equiv (x_1, ..., x_{Q_j}) \equiv \{x_k\},\ k \in q_j$, which are the active post-translationally modified proteins, RNAs and small molecules (ligands) that control the production (transcription) and the degradation of the G_j. The j^{th} expression function (F_j) has an exact expansion in terms of hierarchal component functions (r) and (c), the corresponding weighted superposition coefficients:

$$\begin{aligned}[G_j] = F_j(\vec{x}) \quad \equiv \quad & \sum_{m \in q_j} c_m r_m(x_m) + \\ & \sum_{m,n \in q_j} c_{mn} r_{mn}(x_m, x_n) + \ldots + \\ & c r_{1,\ldots,Q_j}(x_1, ..., x_{Q_j}) \end{aligned} \tag{5.1}$$

The component functions may not have closed forms and we do not know their explicit forms. In this most general case, we only assume that the expression of genes that have common regulators interacting with each other in the same manner can be expanded in terms of the same or very similar component functions. Then, we can treat the component functions as hidden (unobserved) variables and infer their numerical values form the data along with the most parsimonious combination (superposition) of such functions that can explain the measured (observed) expression of all genes.

Mathematically, we formulate a model (analogously to equations 4-6) so that the concentration of the j^{th} mRNA across all i conditions (G_{ij}) is a superposition of only a few component functions $r_k,\ k \in \omega_j$; here ω_j is a small subset of the Ω, the set of all P significant component functions, $r_1, r_2, \ldots, r_P$, and $\{r_p\},\ p \in \Omega$ from the expansions of the expression functions of all genes. We define matrix $\mathbf{R} \in \mathbb{R}^{M \times P}$ so that its columns are the full set of significant component functions and the i^{th} row

contains their corresponding numerical values for the i^{th} physiological conditions. The superposition of r_k corresponding to the expression function of the j^{th} gene (F_j) is encoded in a coupling vector $c_j \in \mathbb{R}^{P\times 1}$. Each element in c_j represents the coupling of the j^{th} gene to the corresponding component function. The coupling vectors form the columns of the adjacency matrix $\mathbf{C} \in \mathbb{R}^{P\times N}$ containing the couplings of all N genes to the component functions.

5.4 Supplementary Figures

5.4.1 GO term trees for the genes with universal GRR

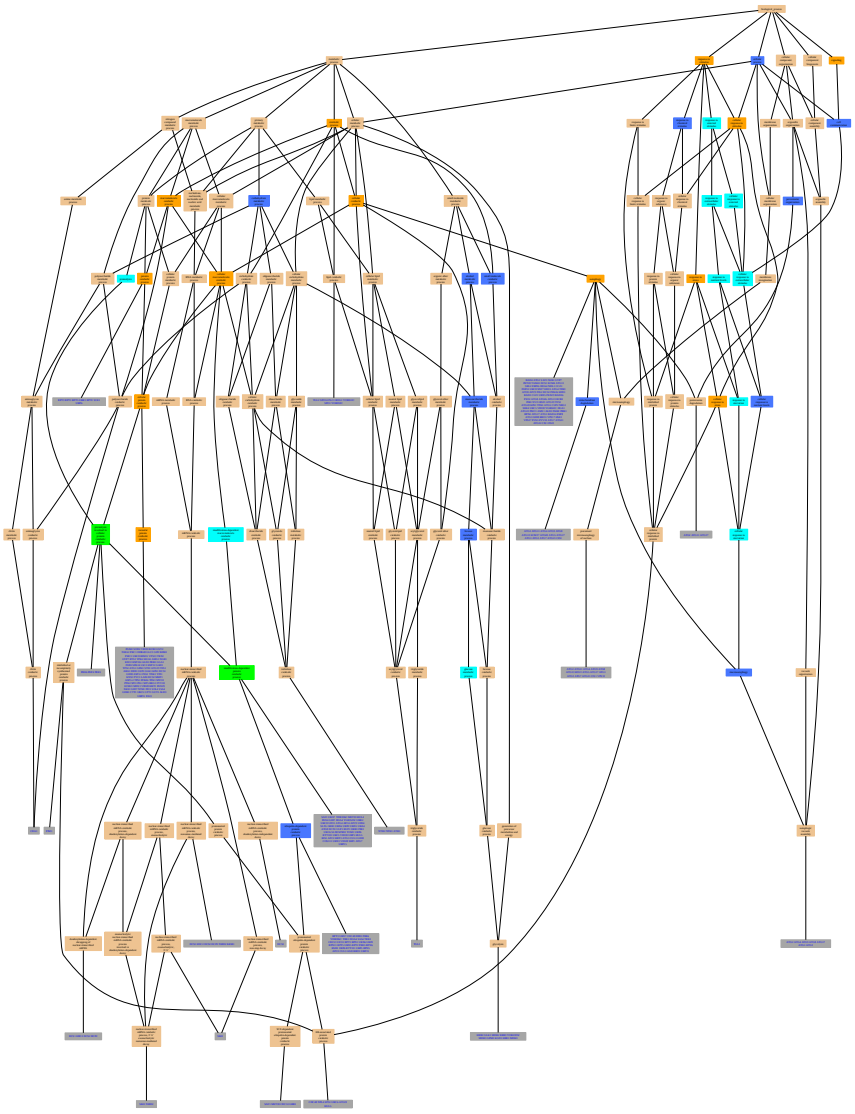

Figure 5.5: GO Term Tree for Genes with Negative Slopes. The color of each box reflects the the probability for seeing the observed overrepresentation by chance alone, as follows: Orange $\leq 10^{-10}$, Yellow $\{10^{-10}10^{-8}\}$, Green $\{10^{-8}10^{-6}\}$, Cyan $\{10^{-6}10^{-4}\}$, Blue $\{10^{-4}10^{-2}\}$ and Beige $> 10^{-1}$.

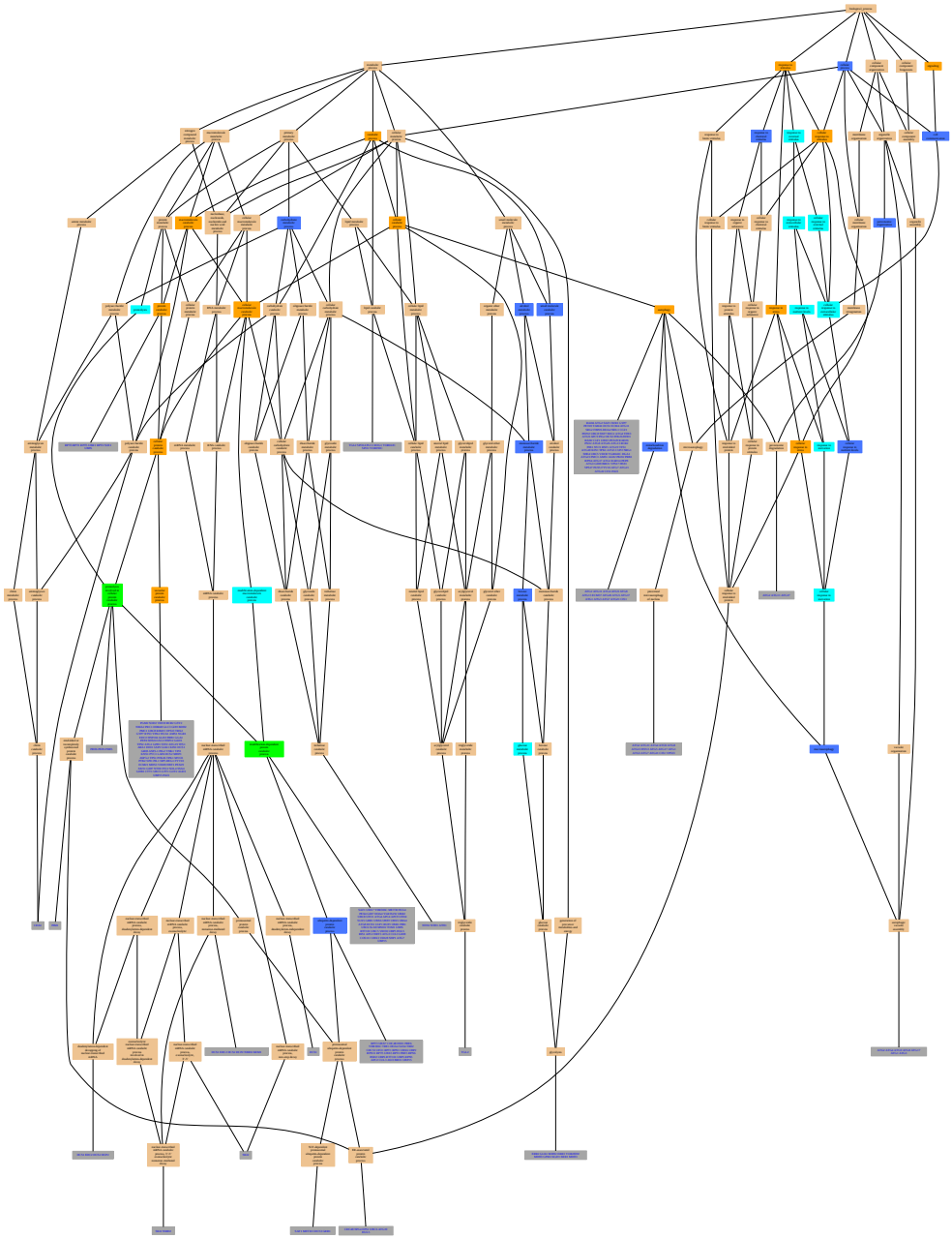

Figure 5.6: GO Term Tree for Genes with Negative Slopes. Notation the same as in Fig.5.5

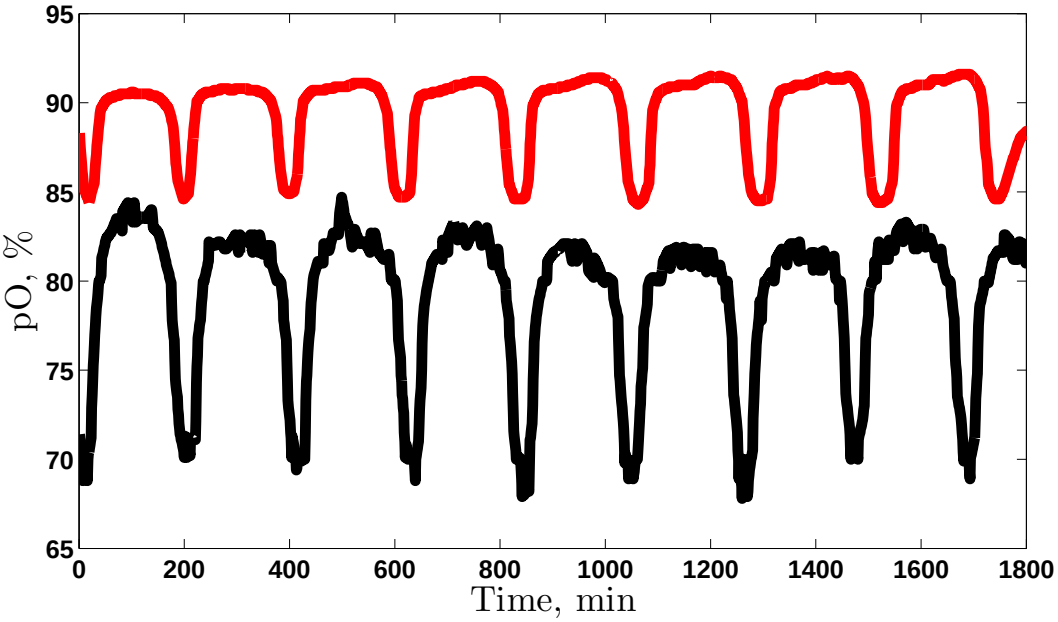

Figure 5.7: Effect of Air Flow on *YMC* Synchronization. Two culture of *S288C* started from the same colony, feeding from the same glucose limited media and growing at a dilution rate $\mu = 0.10h^{-1}$. The only difference is the rate of air flow.

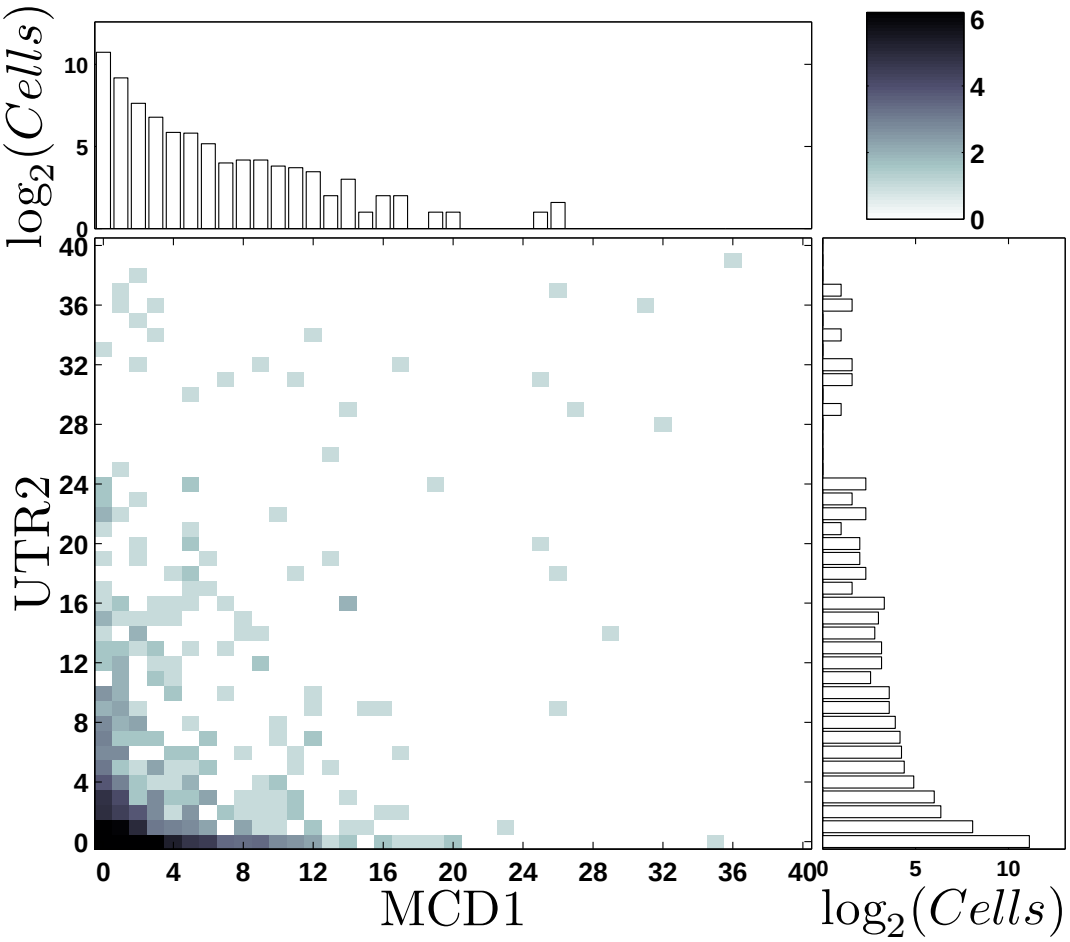

Figure 5.8: Marginal and Join Distributions of mRNAs in *YMC* Synchronized Cultures

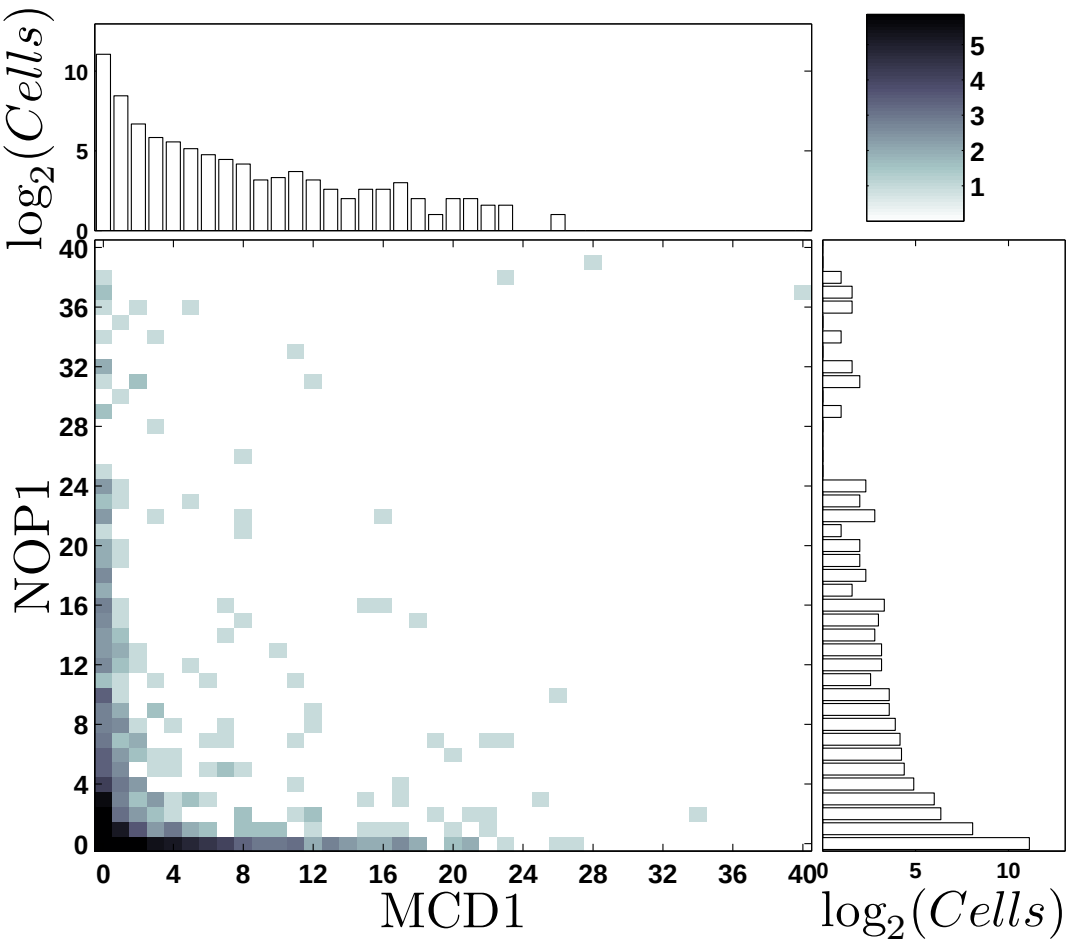

Figure 5.9: Marginal and Join Distributions of mRNAs in *YMC* Synchronized Cultures

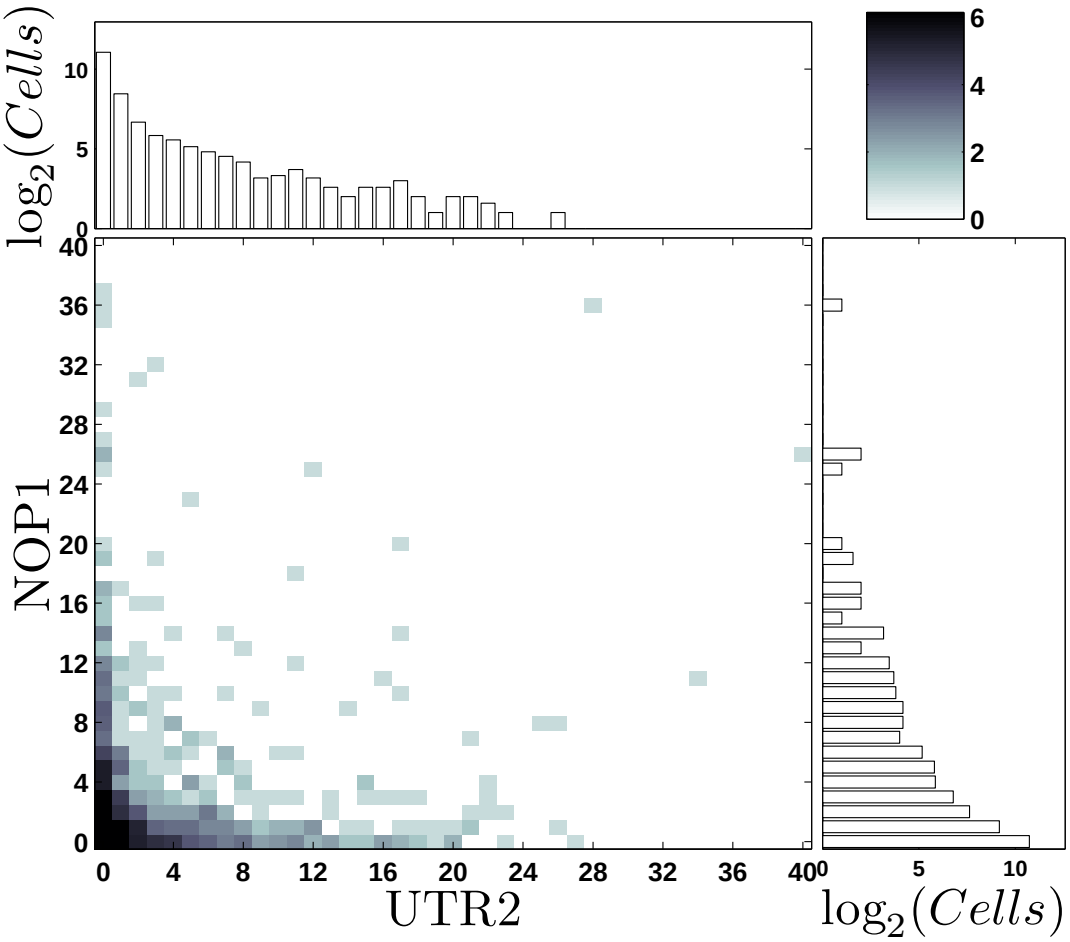

Figure 5.10: Marginal and Join Distributions of mRNAs in *YMC* Synchronized Cultures

5.5 Computing Correlations in FISH data

Pearson correlations between mRNAs computed by averaging across all observed cells is one type of summary statistic describing the data. Such correlations, however, do not capture all information from the joint distribution. The discrete nature of the count data and the presence of multiple sub-distributions can further limit and distort the information captured by the Pearson correlations.

Of particular concern is systematic positive bias that may be introduced by uncorrelated Poisson noise. Indeed such Poisson noise is very likely to exist in the mRNA counts for every pair of genes expressed at very low levels (or not expressed at all) during some phases of the *YMC*. During such *YMC* phases, the expected number of mRNAs for those genes may be zero but the observed number of mRNAs will not be necessarily zero for all cells; the stochastic nature of transcription and mRNA degradation implies that the observed number of mRNAs will follow a Poisson distribution with very low expectation. Since the Poisson distribution is asymmetric, such uncorrelated Poisson noise with low expectations may introduce rather strong positive bias in the estimates of the gene-gene correlations. Indeed, we observe such bias in simulated models, see supp info. That is also the reason why the ranges (min and max) of the Pearson correlations are so skewed. Since the noise is not correlated (and does not need to be correlated to introduce the bias) permutation are not likely to affect it.

To minimize systematic bias from Poisson noise (which unlike Gaussian noise cannot be eliminated simply by averaging many observations), we aim to compute the gene-gene correlations using only cells in which the genes are expressed systematically. To separate such cells from cell in which mRNAs are observed because of stochastic events, we separate all cells between two bivariate Poisson distributions. This separation is accomplished in a principled and systematic way using a mixture model (see supp info) and then the

correlation is computed using only cells from the systematic Poisson distribution having higher expectations for both genes.

To compute a gene-gene correlation using only cells from the *YMC* phases when the two genes are highly expressed (the systematic component) we applied an expectation-maximization (EM) algorithm for a mixture model. In the model, each cell may belong to one of two Poisson bivariate distributions:

1. A bivariate Poisson distribution (P_1) with covariance zero and low expectations (λ_x, λ_y) for both genes. Since the covariance is zero, the join probability mass function is given by the product of two univariate Poisson distributions:

$$P_1(X = x, Y = y|\lambda_x, \lambda_y) = P(X = x|\lambda_x)P(Y = y|\lambda_y)$$

   where $P(X = x|\lambda_x) = e^{-\lambda_x}\lambda_x^x/x!$

2. A bivariate Poisson distribution (P_2) with zero covariance which corresponds to the product of two univariate Poisson distributions.

$$P_2(X = x, Y = y|\lambda_1, \lambda_2) = P(X = x|\lambda_1)P(Y = y|\lambda_2)$$

   .

   The physical interpretation of this model is as follows. The counts mRNA can come from two univariate Poisson distributions: 1) A distribution corresponding to noise during *YMC* phases when the mRNA is not expressed but may be present because of leaky promoter and/or incomplete degradation, and 2) A distribution corresponding to signal during *YMC* phases when the mRNA is abundant. The conditional probability for each cell to be in a *YMC* phase when one of the measured mRNAs is expressed (abundant) or neither of the measured mRNAs is expresses is given by the corresponding bivariate Poisson distribution (P_1 and P_2) which for this

particular case of zero covariance are simply the product of the marginal univariate Poisson distributions.

(a) I also examined the more general case when the covariance θ of P_2 is not zero. In this case the joint mass function is given by (5.2):

$$P(X = x, Y = y|\lambda_1, \lambda_2, \theta) = \\ = e^{-(\lambda_1+\lambda_2+\theta)} \frac{\lambda_1^x}{x!} \frac{\lambda_2^y}{y!} \sum_{i=0}^{i=\min(x,y)} \binom{x}{i} \binom{y}{i} i! \left(\frac{\theta}{\lambda_1\lambda_2}\right)^i \tag{5.2}$$

Since this approach resulted in very similar results to the particular subcase when the covariance is zero in the paper we report only the results of the simple case when the joint mass function is given by $P_2(X = x, Y = y|\lambda_1, \lambda_2) = P(X = x|\lambda_1)P(Y = y|\lambda_2)$. The joint mass function P_2 is defined only for non-negative values of λ_{12}. Beyond this domain the definition has no physical meaning and the mathematical application of the function may result in negative functional values. For empirical estimate of λ_{12} that are beyond the defined domain, P_2 is computed as the product of the two univariate Poisson distributions with expectations λ_1 and λ_2. This approximation corresponds to λ_{12} being zero and will be reflected in the conditional probabilities for cells to belong to P_2. These approximated conditional probabilities may affect the distribution of cells that are close to the boundary of the two distributions (where separation is always problematic) but are unlikely to change the assignment of cells having high copy number of any of the two genes of or of both genes.

Detailed computational application:

1. Initialization

(a) The expectations and covariance for both distributions (P_1 and P_2) are initialized:

i. $\lambda_x = \frac{1}{10} mean$(observed counts of gene x)

ii. $\lambda_y = \frac{1}{10} mean$(observed counts of gene y)

iii. $\lambda_1 = 2mean$(observed counts of gene x)

iv. $\lambda_2 = 2mean$(observed counts of gene y)

(b) For each observed cell the algorithm computes the conditional probabilities for the cell to belong to either P_1 or P_2 using the initialized parameters and the probability mass functions.

(c) Each cell is assigned to the more likely distribution, which is the distribution for which the corresponding condition probability is greater.

2. Cycle

(a) The expectations of P_1 and P_2 are updated with their empirical maximum likelihood (ML) estimates given the cells in each distribution.

(b) The conditional probabilities for each cell (to belong to P_1 or P_2) are computed using the updated distribution parameters (expectations)

(c) Cells are redistributed between P_1 and P_2 based on the new conditional probabilities

(d) The algorithm iterates steps (a-c) until the expectations converge to values that do not change. In particular, the algorithm stops when the difference between corresponding expectations from successive iterations is less than 10^{-10}.

3. The gene-gene correlation then is estimated as the Maximum Likelihood Estimate (*MLE*) of the normalized covariance of P_1. This MLE corresponds to the Pearson

correlation computed by averaging only across the cells that the *EM* algorithm partitioned to belong to distribution P_2.

Constraints: The expectations for distribution $P_1(\lambda_x, \lambda_y)$ are not allowed to be greater than 1 or the one half the mean number of mRNAs from the corresponding gene:

$\lambda_x < \min(1, \textit{mean(}\text{observed counts of gene } x)/2)$

$\lambda_y < \min(1, \textit{mean(}\text{observed counts of gene } y)/2)$

These constraints ensures that only cell coming from *YMC* phases with low probability of expression for both genes are assigned to the non-systematic component and no valuable data are thrown away. Significance: The significance of the computed correlations is assessed by bootstrapping. The observed numbers of mRNAs in cells portioned to P_2 are permuted 10^6 times and the Pearson correlation for each permutation is computed. The reported p-value is the empirical probability that the randomly permuted mRNA counts are correlated more strongly than the counts observed in the data. In particular, for positive correlations the p-value is the fraction of correlations in the permuted data that are larger than the correlation in the non-permuted data. For negative correlations the p-value is the fraction of correlations in permuted data that are smaller than the correlation in the non-permuted data.

Intervals: The correlation interval defines the most negative and most positive correlations than can be observed with the empirical mRNA counts observed in the cells that the EM algorithm assigns to P_2. The strongest negative correlation is computed as the correlation between the counts for the two mRNAs sorted in opposite directions. The strongest positive correlation is computed as the correlation between the counts for the two mRNAs sorted in the same directions.

Bibliography

Airoldi EM, Huttenhower C, Gresham D, Lu C, Caudy AA, Dunham MJ, Broach JR, Botstein D, Troyanskaya OG (2009) Predicting Cellular Growth from Gene Expression Signatures. *PLoS Comput Biol* **5**: e1000257

Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 10101–10106, PMID: 10963673

Applegate D, Bixby R, Chvatal V, Espinoza D, Goycoolea M, Helsgaun K (2007a) Certification of an optimal TSP tour through 85,900 cities. *Operations Research Letters* **37**: 11–15

Applegate DL, Bixby RE, Chvátal V, Cook WJ (2007b) The Traveling Salesman Problem: A Computational Study. *Princeton University Press*

Banerjee O, Ghaoui L, d'Aspremont. A (2007) Model selection through sparse maximum likelihood estimation. *JMLR* **9**: 485–516

Benaych-Georges F, Rao R (2009) EThe eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *arXiv09102120v1*

Boer VM, Amini S, Botstein D (2008) Influence of genotype and nutrition on survival and metabolism of starving yeast. *Proceedings of the National Academy of Sciences* **105**: 6930–6935

Boer VM, Crutchfield CA, Bradley PH, Botstein D, Rabinowitz JD (2010) Growth-limiting Intracellular Metabolites in Yeast Growing under Diverse Nutrient Limitations. *Mol Biol Cell* **21**: 198–211

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G (2004) GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics Oxford England* **20**: 3710–3715, PMID: 15297299

Bradley E, Trevor H, Iain J, Tibshirani R (2004) Least Angle Regression. *Annals of Statistics* **32**: 407–499

Brauer MJ, Huttenhower C, Airoldi EM, Rosenstein R, Matese JC, Gresham D, Boer VM, Troyanskaya OG, Botstein D (2008) Coordination of Growth Rate, Cell Cycle, Stress Response, and Metabolic Activity in Yeast. *Mol Biol Cell* **19**: 352–367

Brauer MJ, Saldanha AJ, Dolinski K, Botstein D (2005) Homeostatic Adjustment and Metabolic Remodeling in Glucose-limited Yeast Cultures. *Molecular Biology of the Cell* **16**: 2503–2517, PMID: 15758028 PMCID: 1087253

Candès E, Wakin M, Boyd S (2007) Enhancing sparsity by reweighted $\ell 1$ minimization. *J Fourier Anal Appl* **14**: 877–905

Carvalho C, West M (2008) High-dimensional sparse factor modelling: Applications in gene expression genomics. *JASA* **103**: 1438–1456

Castrillo J, Zeef L, Hoyle D, Zhang N, Hayes A, Gardner D, Cornell M, Petty J, Hakes L, Wardleworth L, Rash B, Brown M, Dunn W, Broadhurst D, O'Donoghue K, Hester

S, Dunkley T, Hart S, Swainston N, Li P, *et al* (2007) Growth control of the eukaryote cell: a systems biology study in yeast. *Journal of Biology* **6**: 4

Cetin M, Malioutov D, Willsky A (2002) A Variational Technique for Source Localization based on a Sparse Signal Reconstruction Perspective. *ICASSP Orlando Florida* **3**: 2965–2968

d'Aspremont A Bach F. GL (2007) Optimal Solutions for Sparse Principal Component Analysis. *JMLR* **9**: 1269–1294

Dueck D, Morris Q, Frey B (2005) Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics* **21**: 144–151

Elemento O, Slonim N, Tavazoie S (2007) A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. *Molecular Cell* **28**: 337–350

Fazio A, Jewett M, Daran-Lapujade P, Mustacchi R, Usaite R, Pronk J, Workman C, Nielsen J (2008) Transcription factor control of growth rate dependent genes in Saccharomyces cerevisiae: A three factor design. *BMC Genomics* **9**: 341

Friedman J, Hastie T, Tibshirani R (2009) Regularization paths for generalized linear models via coordinate descent

Futcher B (2006) Metabolic cycle, cell cycle, and the finishing kick to Start. *Genome Biology* **7**: 107

Golub G, Kahan W (1965) Calculating the Singular Values and Pseudo-Inverse of a Matrix. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis* **2**: 205–224, ArticleType: primary_article / Full publication date: 1965 / Copyright Âl' 1965 Society for Industrial and Applied Mathematics

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

Hartwell LH (1974) Saccharomyces cerevisiae cell cycle. *Microbiol Mol Biol Rev* **38**: 164–198

Hartwell LH, Culotti J, Pringle JR, Reid BJ (1974) Genetic control of the cell division cycle in yeast. *Science* **183**: 46Ű51

Hayes A, Zhang N, Wu J, Butler PR, Hauser NC, Hoheisel JD, Lim FL, Sharrocks AD, Oliver SG (2002) Hybridization array technology coupled with chemostat culture: Tools to interrogate gene expression in Saccharomyces cerevisiae. *Methods* **26**: 281–290

Hoos H, Stutzle T (2009) On the Empirical Scaling of Run-time for Finding Optimal Solutions to the Traveling Salesman Problem

Kjeldgaard NO, Maaløe O, Schaechter M (1958) The transition between different physiological states during balanced growth of Salmonella typhimurium. *Journal of General Microbiology* **19**: 607–616, PMID: 13611203

Klevecz RR, Bolen J, Forrest G, Murray DB (2004) A genomewide oscillation in transcription gates DNA replication and cell cycle. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 1200–1205

Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America* **94**: 13057–13062

Liu J, Chen J, Ye J (2009) Large-scale sparse logistic regression : 547–556

M. Aharon and ME, Bruckstein A (2005) K-SVD and its non-negative variant for dictionary design. *Wavelets XI* **5914**: 591411–591424

Maaløe O (1979) *The Regulation of the protein-synthesizing machinery ribosomes, tRNA, factors, and so on*. In Biological Regulation and Development, R.F.Goldberger, ed. NewYork: PlenumPress

MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7**: 113–113, PMID: 16522208 PMCID: 1435934

Monod J (1942) Recherches sur la croissance des cultures bacteriennes. *ParisHerman*

Monod J (1949) The Growth of Bacterial Cultures. *Annual Review of Microbiology* **3**: 371–394

Novick A, Szilard L (1950) Description of the Chemostat. *Science* **112**: 715–716

Pe'er D, Regev A, Tanay A (2002) Minreg: Inferring an active regulator set. *Bioinformatics* **18**: 258–267

Pir P, Kirdar B, Hayes A, Onsan ZI, Ulgen K, Oliver S (2006) Integrative investigation of metabolic and transcriptomic data. *BMC Bioinformatics* **7**: 203

Regenberg B, Grotkjaer T, Winther O, Fausboll A, Akesson M, Bro C, Hansen L, Brunak S, Nielsen J (2006) Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in Saccharomyces cerevisiae. *Genome Biology* **7**: R107

Saldanha AJ, Brauer MJ, Botstein D (2004) Nutritional Homeostasis in Batch and Steady-State Culture of Yeast. *Mol Biol Cell* **15**: 4089–4104

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **270**: 467–470

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34**: 166–176

Sigg C, Buhmann J (2008) Expectation-Maximization for Sparse and Non-Negative PCA. *ICML Helsinki Finland* **9**

Silverman SJ, Slavov N, Petti AA, Parsons L, Briehof R, Thiberge SY, Zenklusen D, Gandhi SJ, Larson DR, Singer RH, Botstein D (2010) Metabolic cycling in single yeast cells from unsynchronized steady-state populations limited on glucose or phosphate. *Proceedings of the National Academy of Sciences*

Slavov N (2010) Inference of Sparse Networks with Unobserved Variables. Application to Gene Regulatory Networks. *JMLR WCP* **9**: 757–764

Slavov N (2012) THE MISSION OF MIT. *The MIT Tech*

Slavov N, Airoldi EM, van Oudenaarden A, Botstein D (2012) A conserved cell growth cycle can account for the environmental stress responses of divergent eukaryotes. *Molecular Biology of the Cell* **23**: 1986–1997

Slavov N, Botstein D (2011) Coupling among growth rate response, metabolic cycle, and cell division cycle in yeast. *Mol Biol Cell* **22**: 1997–2009

Slavov N, Botstein D (2013) Decoupling Nutrient Signaling from Growth Rate Causes Aerobic Glycolysis and Deregulation of Cell-Size and Gene Expression. *Molecular Biology of the Cell*

Slavov N, Botstein D, Caudy A (2013) Coordination of Gene Regulation, Metabolism and Cell Division in Yeast Cells from Metabolically Synchronized and Asynchronous Cultures. *In Preparation*

Slavov N, Dawson KA (2009) Correlation signature of the macroscopic states of the gene regulatory network in cancer. *Proceedings of the National Academy of Sciences* **106**: 4079–4084

Slavov N, Macinskas J, Caudy A, Botstein D (2011) Metabolic cycling without cell division cycling in respiring yeast. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 19090–19095, PMID: 22065748

Srebroand N, Jaakkola T (2001) Sparse Matrix Factorization for Analyzing Gene Expression Patterns. *NIPS* **9**

Tu BP, Kudlicki A, Rowicka M, McKnight SL (2005) Logic of the Yeast Metabolic Cycle: Temporal Compartmentalization of Cellular Processes. *Science* **310**: 1152–1158

West M (2002) Bayesian factor regression models in the large p, small n paradigm. *Bayesian Stat* **7**: 723–732

Zaman S, Lippman SI, Schneper L, Slonim N, Broach JR (2009) Glucose regulates transcription in yeast through a network of signaling pathways. *Molecular Systems Biology* **5**: 245–245, PMID: 19225458 PMCID: 2657534

Zaman S, Lippman SI, Zhao X, Broach JR (2008) How Saccharomyces responds to nutrients. *Annual Review of Genetics* **42**: 27–81, PMID: 18303986