EVOLUTIONARY SYSTEMS BIOLOGY OF AMINO ACID BIOSYNTHETIC COST AND GENE IMPORTANCE IN SACCHAROMYCES CEREVISIAE

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Life Sciences

2010

By
Michael D. Barton

Contents

List of Tables

List of Figures

Nomenclature

$A_{nutrient}$	Amino acid absolute cost under a given nutrient limitation
ANOVA	Analysis of variance
C_x^y	Flux control coefficient of reaction x on reaction y
CAI	Codon adaptation index
COBRA	Flux balance analysis toolbox for MATLAB (see [1])
dN	Rate of non-synonymous evolutionary rate
dN/dS	Ratio of non-synonymous to synonymous evolutionary rate
dS	Rate of synonymous evolutionary rate
FBA	Flux Balance Analysis
gDW	Grams of Dry Weight biomass
hr	Hour
iJR904	*Escherichia coli* genome scale model (see [2])
InChI	IUPAC International Chemical Identifier
iND750	*Saccharomyces cerevisiae* genome scale model (see [3])
K_m	Substrate concentration where enzyme rate is at half maximum ($\frac{1}{2}V_{max}$)
MAD	Median absolute deviation
MCA	Metabolic Control Analysis
MIRIAM	Minimum Information Requested In Annotation of biochemical Models
mmol	Millimoles, quantity equal to one thousandth number of atoms in 0.012 grams of carbon^{12}
MOMA	Minimisation of Metabolic Adjustment
ORF	Open reading frame
$R_{nutrient}$	Amino acid relative cost under a given nutrient limitation
ROOM	Regulatory on/off minimisation
SBML	Systems Biology Markup Language
SGD	*Saccharomyces* Genome Database (see [4])
V_{max}	Maximum reactions per second per mole of enzyme

Abstract

Systems biology and the development of *in silico* models has allowed quantitative predictions about organism physiology. The aim of this PhD has been to use a model of *Saccharomyces cerevisiae* metabolism to make predictions about gene function and evolution. These predictions aim to link the selective pressures observed in the genome to the metabolic constraints that direct organismal evolution.

This PhD resulted in two systems biology derived measures related to *S. cerevisiae* physiology. The first examines the cost of synthesising an amino acid in terms of the constituent nutrients. This cost estimate provides a metric which may be used to determine if energy is a selective pressure on the use of amino acids in protein synthesis and evolution. The second measure estimates how small reductions in reaction activity encoded by proteins may effect organism fitness. The aim of this measure is to determine whether the importance of a reaction is reflected the expression and evolution of the corresponding gene.

Comparison of the derived amino acid costs with amino acid use and evolution across a range of *Saccharomyces* species revealed that amino acids are not maintained in genomes relative to cost but that interspecies substitution rates are strongly correlated with biosynthetic cost. Gene importance estimates were compared with *S. cerevisiae* gene dispensability, gene dosage dependence, interspecies evolutionary rate, and gene expression predicted through codon usage bias. The results of these analysis revealed limited explanatory power of this measure, but did indicate a weak relationship with gene evolutionary rate and gene expression.

This PhD has demonstrated that systems biology models can be used to predict evolutionary forces in yeast, and that further development and application of these models may yet yield greater understanding of evolution through modelling metabolism *in silico*.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

I would like to thank my supervisors Casey Bergman and Magnus Rattray for the scientific feedback and guidence which has allowed me to complete this PhD. I must especially thank Casey for accepting me as his PhD student halfway through my project, and for the financial support provided during the final stages.

I thank Steve Oliver for accepting me as his PhD student and giving the chance begin a career in scientific research. I also thank Steve for motivating my interest in systems biology. Balaźs Papp provided patient advice and feedback on performing flux balance analysis. Daniela Delneri, once my supervisor now advisor, provided professional, personal and culinary advice through-out the course of my PhD. Simon Whelan has given me much of his time to help me learn molecular evolution. Sam Griffiths-Jones provided constructive discussions throughout my PhD. Hans Westerhoff and Frank Bruggeman were kind enough to give me both theoretical and technical tuition in systems biology. Markus Herrgård gave helpful advice on performing flux balance analysis using the COBRA toolbox. Vangelis Simeonidis and Ralf Steuer also provided time to discuss flux balance analysis. Nick Gresham, Adam Huffman, and Victoria Kelly provided invaluable technical support during my research.

I thank NERC for providing PhD funding. Dennis Wall was generous to provide the *Saccharomyces* multiple sequence alignments used in this research. Richard Apodaca generously provided amino acid images in postscript format.

I thank the members of the Bergman lab for the pleasant working environment: Ian, Raquel, Max, Martin, Dave, Pedro, and Joachim. I thank my friends in the Manchester Faculty of Life Sciences for their invaluable emotional support during both good and bad times: Eamon, Zoher, Rob, James, Claire, Sarah, Farhana, Amanda, Ryan, Jamie, and Simon. Special thanks for Penny who has been graceful in accepting me in the most difficult periods.

Finally without the unwavering support of my family, Helen, Hazel, Mum and Dad I would, in the most literal sense, never have made it thus far.

1
Introduction

The aim of this PhD is to investigate if systems biology simulations can be used to accurately estimate selection pressures on organism growth in microbial eukaryotes. The work performed in this PhD uses mathematical models of metabolism to predict patterns of genome evolution and gene expression. In general this PhD employs system biology methods to test the accuracy at which the model predictions match observed experimental and genomic data. The brief overview in this chapter outlines the introductions, results and discussion of the chapters that follow.

Chapter 2 provides an introduction to systems biology, in particular focusing on the use and application of genome scale models of metabolism. These models, containing hundreds of reactions, can be used to predict organism growth and distribution of reaction flux through metabolic pathways.

Genome scale models are here used for the first time to predict the biosynthetic cost of each of the twenty amino acids. The aim of this work is to predict protein biosynthetic costs in terms of the sugars and other nutrients used to synthesise the encoding amino acids. By accurately estimating protein biosynthetic cost it may be possible to determine if biosynthetic cost is a selection pressure on gene expression and evolution.

The second half of Chapter 2 uses genome scale models to estimate the importance of a gene in terms of how much the encoded reaction activity affects growth. The aim of this analysis was to determine the metabolic selection pressures associated with maintaining a gene-encoded reaction activity. As with amino acid biosynthetic cost, accurate estimation of gene importance using systems biology may provide insight into gene expression and evolution.

Chapter 3 uses the estimated amino acid biosynthetic costs from Chapter 2 to determine the role of cost in gene expression and evolution. The cost minimisation hypothesis suggests that since amino acids have differing costs to produce it may be expected that energy can be saved in metabolism through the use of cheaper amino acids in expressed proteins. An energetic cost saving in protein may then correspond to a fitness advantage.

The role of amino acid biosynthetic cost is first examined in gene expression to determine if protein cost is reflected in the level of expression. Cost is simultaneously compared with other variables associated with translational optimisation that have previously been shown to be important in gene expression.

The second half of Chapter 3 compares amino acid sequence evolution with amino acid biosynthetic cost. The hypothesis for the analysis is that cheaper amino acids may be selected for in protein sequence evolution, again to provide a fitness advantage by minimising the use of energy. This analysis examines the genome of *Saccharomyces* species to determine if amino acid usage and substitution patterns are related to cost and therefore indicative of a cost selection trend in genome evolution.

Chapter 4 uses the gene cost estimates produced in Chapter 2 to examine if the systems biology predicted importance of a gene correlates with selection pressures on gene use and sequence evolution. Current *in silico* predictions of gene importance focus on whether the encoded reaction can be completely removed from metabolism without a fitness effect. The gene cost estimates produced in Chapter 2 however try to understand the importance of a gene based on how small changes in reaction activity, as opposed to complete reaction loss, affect organism fitness. The rationale behind this approach to mimic natural variation in reaction activity from mutations of small effect.

Gene importance estimates are compared with four variables representing gene use and sequence evolution and life cycle. The comparison with these variables aims to determine if systems biology can be used to predict gene importance, and if these estimates are correlated with the use and evolution of genes in *S. cerevisiae*. The four examine variables are: whether gene loss affects organism fitness, whether the loss of one copy of a gene affects organism fitness, the evolutionary rate of the gene, and the expression level of the gene.

Chapter 5 briefly summarises the main results of this thesis and includes a discussion of the some of the future directions this research could take.

2

Estimating amino acid biosynthetic cost and gene importance using flux balance analysis

Chapter summary

This chapter describes novel systems biology approaches to estimating amino acid biosynthetic cost and the functional importance of reactions encoded by single genes in *Saccharomyces cerevisiae*. The background section describes metabolic control analysis and how this is applied to model cellular metabolism. This is followed by a description of genome scale model construction and how flux balance analysis is used to simulate metabolism *in silico*. The results section first outlines a method to estimate the energetic cost of amino acid synthesis in glucose, ammonium, and sulphur limitation. The second half of the results section describes a method to determine the importance of single gene-encoded reactions in glucose and ammonium limited conditions with the aim of predicting the selection pressures on the encoding gene.

2.1 Introduction

2.1.1 Modelling metabolic systems

Early enzymology studies assumed the existence of rate limiting steps in biological pathways. The intuition was that the overall production rate of the end product of a pathway is constrained by the rate of the slowest reaction step. As the slowest reaction rate increases, the whole pathway rate increases in proportion until another step becomes limiting. Niederberger *et. al* [5] tested this assumption and instead found that individual up or down regulation of enzyme quantities at specific reaction steps had only marginal effect on tryptophan biosynthesis pathway in *Saccharomyces cerevisiae*. The rate was only accelerated by increasing the quantity of five related enzymes in tandem. This research demonstrated that control in a biological system is distributed over the system as a whole rather than concentrated at individual reactions.

The theory of metabolic control analysis (MCA) [6, 7] states that there is no rate limiting step in a pathway, but instead each reaction shares a measure of overall control. The control that the rate of reaction x has on reaction y is determined by applying a small change in the rate of x and measuring the degree of response in the rate of y. In metabolic control analysis this is described as the flux control of reaction x on reaction y. The unitless coefficient of flux control uses the symbol C_x^y and is defined as:

$$C_x^y = \frac{\delta y/y}{\delta x/x}\,. \tag{2.1}$$

The denominator ($\delta x/x$) represents the dimensionless effect of a small change in the reaction flux through x. The numerator ($\delta y/y$) represents the corresponding response in y resulting from the change in x. This analysis is typically performed in log space where the value of C_x^y represents the ratio of response in reaction y to changes in the rate of x. Larger values of C_x^y indicate the reaction x has a high degree of control on reaction y. Smaller values indicate reaction y experiences only a small change in flux in response to changes in x. Using this approach, a quantitative measure of control may be derived for each reaction in a biological system, whether *in vivo* or *in silico*.

Determining flux control coefficients for reactions in a metabolic network is important because it allows quantitative exploration of biological systems. For example, Rossell *et al.* [8] described the control of metabolism in terms of reactions whose rates are increased by changes in metabolite concentration, or whose activity is controlled through changes in enzyme quantity and therefore via gene expression [9]. MCA can also be applied to engineer an increase in the production of commercial biomolecules [5] and can also be used to identify possible drug targets through the combination of reaction knock-downs which have the greatest effect on the metabolic network [10, 11]. This approach can be taken further to develop antibiotic drugs for human pathogens such as trypanosome parasites, where likely drug targets are the reactions with a significant effect of the drug on metabolism in the pathogen but where the orthologous human reaction has a much decreased effect. This difference in metabolic control for the same reaction in both species translates into a lethal drug effect in the parasite with minimal impact on the host [12].

2.1.2 Genome scale models

Derivation of a flux control coefficient *in silico* requires the construction of a kinetic model of the system including enzyme kinetic parameters. In the case of reactions exhibiting simple Michaelis-Menten kinetics only substrate affinity (K_m) and maximum enzyme reaction rate (V_{max}) are necessary. More complex enzymatic kinetics include larger numbers parameters which requires further experimental effort to derive. Even with all experimentally derived kinetic parameters, a model must still be tested to determine if simulated behaviour matches expected *in vivo* behaviour [13]. The large experimental effort required to derive complete sets of enzyme kinetic parameters means kinetic models are relatively small compared with the anticipated size of total cellular metabolism [14].

In comparison with kinetic models, stoichiometric models do not require enzyme kinetic parameters, and instead only require a list of reactions and the metabolites used in each reaction. A stoichiometric model is represented by an $m \times n$ sized matrix S, which represents the reactions in metabolism and the metabolites consumed by each reaction. The size of m in the matrix is the total number of metabolites in the model, and n is the total number of reactions. The position S_{ij} in the matrix is the participation of metabolite i in reaction j. Positive values of S_{ij} indicate the metabolite is produced by the reaction, negative values mean the metabolite is consumed. A zero indicates the metabolite does not participate in the reaction.

As stoichiometric models require less data to create, it is easier to produce large metabolic representations using this modelling framework. Stoichiometric models comprising a significant proportion of an organism's metabolism have been created, and are referred to as genome scale models. The construction of genome scale metabolic models is reviewed by Feist *et al.* [15], and the process is outlined in brief here.

The construction of a genome scale model requires the production of the stoichiometric matrix detailing the participation of metabolites in each reaction. Enzyme databases such as BRENDA [16] and KEGG [17] provide information on conserved enzymes and pathways in metabolism while databases such as the *Saccharomyces* Genome Database (SGD) [4] allow the identification of reactions specific to yeast. The initial stages of model construction can be automated by cross referencing metabolic maps with the presence of an enzyme in a particular genome. Expert knowledge of an organism can identify the existence of specific metabolic pathway, particularly for organisms living in niche environments [18].

The first *Escherichia coli* stoichiometric model was developed almost 20 years ago [19]. A simplified model of acetate production was described by Majewski *et al.* in 1990, and the most recent 2007 genome scale model accounts for 1260 open reading frames [20]. The range of species for which a genome scale model exists continues to expand and includes attempts to reconstruct a human cell [21]. However, the field of genome scale model construction suffers from the lack of a single standard for construction and model format, making model interchange and comparison difficult. There are recent attempts to create a unified *S. cerevisiae* model [22] using standards such the systems biology markup language (SBML) [23], the minimum information requested in annotation of biochemical models (MIRIAM) [24], and the international chemical identifier codes (InChI) [25].

When constructing a genome scale model the goal is often the simulation of *in vivo* phenotypes, which requires an additional biomass producing reaction in the matrix S describing cellular growth, referred to as the biomass equation. In many cases this reaction consumes lipids, proteins, nucleic acids and other cofactors to produce a per hour (hr^{-1}) unit of new biomass. The required quantities of each molecule in this growth reaction are estimated from *in vivo* millimoles per gram of dry weight cellular biomass (mmol $^{-1}$ gDW $^{-1}$). Table 2.1 summarises the molecular requirement of the biomass equation in the *S. cerevisiae* iND750 model. These requirements include fats such as triglyceride and ergosterol, sugars such as glycogen and mannan, and small metabolites comprising sulphate, water, and ATP. The genome, transcriptome, and proteome content of the cell is included as the quantities of nucleotides and amino acids estimated from dry weight cellular composition.

The *in silico* simulation of a newly constructed stoichiometric model must be validated against the observed *in vivo* phenotypes, where *in silico* behaviour should reflect that of *in vivo* behaviour as closely as possible [26]. Differences between *in silico* predictions and *in vivo* results highlight gaps in the model or experimental data which can be updated and followed by repeated model validation or further experimental studies.

Stoichiometric model simulation using flux balance analysis

As described above a stoichiometric model is a matrix S describing the metabolites consumed and produced in each reaction. FBA uses linear programming to solve the equation $Sv = 0$ where v represents a vector of reaction rates, known as fluxes, for each reaction in S. Biologically this solution represents steady state in metabolism where, in theory, all mass produced and consumed by reactions is balanced. There are multiple solutions to this equation and so linear optimisation is used to determine the combination of fluxes that maximises or minimises a value in the vector v, referred to as the objective function. A common objective function used for genome scale models of unicellular organisms is growth rate as modelled by the biomass equation. Additional constraints can be placed on the flux balance solution to reflect *in vivo* reaction kinetics, such as upper and lower bounds for individual reactions. Constraining a reaction rate to be ≥ 0 restricts the solution space so the reaction may only proceed in the forward direction. Reaction constraints are also used to limit the flux of nutrients entering the cell and thereby simulate different environmental conditions.

The aim of FBA using a genome model is to simulate organism phenotypes related to metabolism. Several approaches have been developed for this purpose. Minimisation of metabolic adjustment (MOMA) [27, 28] is a technique that optimises the objective function while minimising the sum change in flux adjustment between two flux solutions. The aim of this method is to simulate perturbed metabolic phenotypes such a wild-type and gene knockout, where minimal flux changes more accurately reflect the *in vivo* behaviour of how metabolic adjustments are made. A related approach called regulatory on/off minimisation

Type	Molecule	Requirement (mmol^{-1} gDW^{-1})
Nucleic Acid	dAMP	0.00360
	dCMP	0.00240
	dGMP	0.00240
	dTMP	0.00360
	AMP	0.04600
	CMP	0.04470
	GMP	0.04600
	UMP	0.05990
Amino Acid	Alanine	0.45880
	Arginine	0.16070
	Asparagine	0.10170
	Aspartate	0.29750
	Cysteine	0.00660
	Glutamate	0.30180
	Glutamine	0.10540
	Glycine	0.29040
	Histidine	0.06630
	Isoleucine	0.19270
	Leucine	0.29640
	Lysine	0.28620
	Methionine	0.05070
	Phenylalanine	0.13390
	Proline	0.16470
	Serine	0.18540
	Threonine	0.19140
	Tryptophan	0.02840
	Tyrosine	0.10200
	Valine	0.26460
Carbohydrate	1 3 beta D Glucan	1.13480
	Mannan	0.80790
	Glycogen	0.51850
	Trehalose	0.02340
Lipid	Ergosterol	0.00070
	Triglyceride	0.00007
	Zymosterol	0.00150
Phospholipid	Phosphatidyl 1D Myo Inositol	0.00005
	Phosphatidate	0.00001
	Phosphatidylcholine	0.00006
	Phosphatidylethanolamine	0.00005
	Phosphatidylserine	0.00002
Small	ATP	59.27600
	H_2O	59.27600
	Sulphate	0.02000

Table 2.1: Biomass requirements of *S. cerevisiae* iND750 model. The requirement of each molecule is millimoles per gram of dry weight biomass. The estimation of biomass quantities for genome scale models is outlined by Feist *et al.* [15] the *S. cerevisiae* model construction is described by Duarte *et al.* [3].

(ROOM) minimises the number of reactions that switch between an active or inactive state between solutions [29]. This approach attempts to mimic realistic *in vivo* behaviour by favouring the use of already active reactions when adapting to different conditions, as opposed to activation or inactivation of multiple reactions.

Typically the maximal value of the objective function is optimised, however the FBA solution may be constrained by the objective function, where any changes in internal reaction flux results in a dramatic decrease in objective flux. Mahadevan and Schilling [30] discuss that metabolism may not be tightly optimised for a specific environment *in vivo*, but instead may retain a measure of flexibility in the metabolic network to adjust to changes in the environment. These authors showed that performing FBA to find a near optimal solution revealed greater flexibility in the metabolic network.

Predicting essential cellular components using genome scale models and flux balance analysis

The use of genome scale metabolic models is not limited to estimating the control of metabolic reactions on growth, but may also be used to estimate the importance of specific metabolites. Varma *et al.* [31] used FBA to determine the role of oxygen in *E. coli* energy metabolism by decreasing oxygen availability and identifying the corresponding change in the use of redox carriers. As oxygen availability decreased, the use of ethanol as the electron acceptor increased. The authors described this effect in terms of the ‘price’ of each molecule, where the price of oxygen use rises with decreasing availability while the price of ethanol use decreases. This highlights how systems biology models can be used to estimate the importance of metabolites to cell growth. Few subsequent studies have looked at the cost of genes or metabolites in a systems biology framework using a similar approach.

2.1.3 Estimating amino acid biosynthetic cost

Up to 80% of the cellular energy budget is predicted to be expended on protein production [32], and therefore cost of protein synthesis may be under selection pressure. To determine if differential amino acid usage in protein synthesis may effect organism fitness, a cost for each amino acid must be estimated.

Craig & Weber [33] estimated the biosynthetic cost of an amino acid as the sum of how many high energy phosphate bonds (e.g. ATP) and reducing molecules (e.g. NADPH) are diverted from glucose metabolism to produce each amino acid. Akashi & Gojobori [34] explored a similar approach for a range of different energy sources to estimate the role of amino acid cost in codon bias predicted *E. coli* and *Bacillus subtilis* gene expression. Heizer *et al.* [35] expanded this approach to estimate amino acid cost in four prokaryotic species including photoautotrophs. Wagner [32] considered biosynthetic amino acid cost in *S. cerevisiae* using a similar approach but considering both respiratory and fermentative energy metabolism. Seligmann [36] argued the cost of amino acid biosynthesis must take into account more than just the energy required for synthesis, and instead used molecular weight as a proxy for amino acid cost arguing that this may take into account investments made in synthesising more complex molecular structures. Using molecular weight also has the advantage of being independent of a particular species metabolism. Table 2.2 summarises these previously reported measures of amino acid biosynthetic cost.

The majority of amino acid cost estimates previously reported require manual examination of a metabolic network to determine the number of energy molecules expended in the synthesis. This is a time-consuming activity and susceptible to human error. No approaches to estimating amino acid cost have yet considered using genome scale models in an approach similar to Varma *et al.* [31], which would allow automated generation of cost in a systems biology framework.

2.1.4 Estimating gene importance

Protein biosynthesis may not be the only type of fitness cost where the another example is evolution of reaction encoding genes may effect organism fitness, depending on how important the reaction is. A common approach to estimating the importance of a reaction in a genome scale model is to remove the reaction completely, and compare the resulting solution with the original wild-type simulation. The *in silico* deletion of a reaction is equivalent to a knockout mutation *in vivo*, and may be considered to confer a fitness defect if there is a significant growth rate reduction [37]. The deletion of a reaction may also prevent model optimisation indicating the reaction is essential [38]. Even if a model solution can still be found, the reaction may still be considered essential if the predicted growth defect is large enough [39, 40].

	A&G Energy	C&W Energy	C&W Steps	Wager Fermentative	Wagner Respiratory	Molecular Weight
ala	11.7	12.5	1	2	14.5	89.1
arg	27.3	18.5	10	13	20.5	174.2
asn	14.7	4	1	6	18.5	132.1
asp	12.7	1	1	3	15.5	133.1
cys	24.7	24.5	9	13	26.5	121.2
gln	16.3	9.5	2	3	10.5	146.2
glu	15.3	8.5	1	2	9.5	147.1
gly	11.7	14.5	4	1	14.5	75.1
his	38.3	33	1	5	29	155.2
ile	32.3	20	11	14	38	131.2
leu	27.3	33	7	4	37	131.2
lys	30.3	18.5	10	12	36	146.2
met	34.3	18.5	9	24	36.5	149.2
phe	52	63	9	10	61	165.2
pro	20.3	12.5	4	7	14.5	115.1
ser	11.7	15	3	1	14.5	105.1
thr	18.7	6	6	9	21.5	119.1
trp	74.3	78.5	12	14	75.5	204.2
tyr	50	56.5	9	8	59	181.2
val	23.3	25	4	4	29	117.2

Table 2.2: Amino acid costs previously described in the literature. The Akashi & Gojobori [34], Craig & Weber energy [33], and Wagner [32] cost estimates are based on the curation of the number of high-energy molecules used during synthesis. The Craig & Weber 'steps' measure [33] is based on the number of biosynthetic steps between central metabolism and the produced amino acid. Molecular weight proposed by Seligmann [36] is in Daltons.

FBA of all possible gene deletions for *S.cerevisiae* has been calculated and compared with *in vivo* observations in several studies [26, 41]. Of the *in vivo* knockout phenotypes up to 85% were correctly predicted *in silico* depending on the simulated media. *In silico* knockout studies have been extended to detect metabolic redundancy by comparing all possible double reaction deletions and finding the pairs with no effect when deleted individually, but indicate a sick or lethal phenotype when deleted simultaneously [42, 43]. Reactions active in a specific range of conditions can also be determined through simulation of a wide range of environmental conditions, where a deletion may result in a fitness effect in only a small subset of biologically feasible conditions. This analysis indicated that many reactions are indeed necessary in only a limited set of anticipated environments [39].

2.1.5 Summary of results

The results in this chapter are presented in two sections. The first section describes an approach to estimating amino acid cost using the *S. cerevisiae* genome scale metabolic model similar to that described by Varma *et al.* [31] where each amino acid biosynthetic cost is estimated in the context of glucose, ammonium, and sulphate limitation. The estimated costs are validated by comparison with previous estimates of amino acid cost in the literature and show that this novel systems biology approach can be successfully used to estimate the biosynthetic cost of molecule in a variety of conditions. These results are use to study the impact of amino acid cost on gene expression and evolution in Chapter 3.

The second section presents a related approach to estimate the importance of a gene used the *S. cerevisiae* genome scale metabolic model. Gene importance is examined in the metabolic model by manipulating enzymatic reactions known to be encoded by the gene. Each reaction is examined using three measures: whether reaction flux may vary, the flux through the reaction, and how reductions in reaction flux affect the model objective function. This work represents a novel approach to estimate gene importance using genome scale models, and is used for analysing gene function and evolution in Chapter 4.

2.2 Materials and Methods

2.2.1 Genome scale metabolic models

The genome scale models used in this work were *S. cerevisiae* iND750 [3] and *E. coli* iJR904 [2]. FBA was performed using the COBRA toolbox [1] and the lpsolve optimisation library [44].

The COBRA toolbox for manipulating genome scale models and performing FBA defines a sparse matrix specifying which genes encode which reactions in the model. This representation defines if there is one or more genes known to encode an enzyme catalysing one of the reactions in the model's stoichiometric matrix. Gene-reaction associations can include multimeric enzymes where multiple gene products form an enzyme complex, or isoenzymes where multiple genes encode enzymes for the same reactions. Only a subset of the reactions in the *S. cerevisiae* metabolic model are associated with a single gene. This gene association matrix was parsed to determine the reactions catalysed by only a single gene, where the gene is not associated with any other reactions.

2.2.2 Flux balance analysis

Nutrient limitation was simulated by fixing the upper and lower bounds of the biomass production reaction to 0.3 hr^{-1} and optimising an objective function to minimise the entry of a specific nutrient into the cell. All other nutrient entry fluxes were set to have bounds of -10000 mmol^{-1} gDW^{-1} hr^{-1}, effectively non-limiting on simulated growth. This FBA configuration finds the minimum amount of a specific nutrient required to maintain cellular growth rate at 0.3 hr^{-1}. The growth rate 0.3 hr^{-1} was chosen as a realistic estimate of *S. cerevisiae* growth [26], though all FBA estimations may be expected to increase proportionally with any selected growth rate.

For the *S. cerevisiae* iND750 [3] genome scale model, there are reactions allowing entry of glucose, ammonium, sulphate, phosphate, oxygen and water into the cell. Each of these reactions represents the only external source for an essential nutrient. For example, when simulating the glucose limitation there are no other high energy sugars available. The FBA simulation does however allow a variety of molecules to exit the cell, these are metabolic by products such as carbon dioxide or ethanol.

2.2.3 Flux balance analysis of amino acid cost

Estimation of amino acid cost was performed by making a percentage change ($\pm 0.0002\%$) to the requirement of amino acid at position S_{ij} in the model stoichiometric matrix, where i corresponds to the amino acid at the jth position in the biomass reaction. When the model is optimised, each change in S_{ij} results in a corresponding change in the limiting nutrient uptake. Larger effects in nutrient uptake indicate a greater nutrient 'cost' for the amino acid. The slope between the change in amino acid requirement and the corresponding change in nutrient uptake flux was defined as the cost of the amino acid for that nutrient. This cost is a relative cost as it is derived from a percentage change in amino acid requirement rather than a fixed value change. Division of the relative cost by the original amino acid requirement results in an absolute estimate of amino acid cost equivalent to estimating the slope from a fixed value change in amino acid requirement.

Both relative and absolute amino costs were calculated for glucose, ammonium, and sulphate limitation. Each cost was then divided by the growth rate at which it was estimated (0.3 hr^{-1}) to give an estimate of cost independent of growth rate. The nomenclature used for relative and absolute costs defined in a given nutrient limitation is $R_{nutrient}$ and $A_{nutrient}$ respectively.

Units of absolute and relative measures of amino acid cost

The relative costs of amino acid synthesis are the value of m in the linear relationship $y = mx + c$ where x is the unitless percentage change in the amino acid requirement, and y is corresponding mmol^{-1} gDW^{-1} hr^{-1} response in nutrient intake flux. When divided by the hr^{-1} growth rate at which the cost was estimated, the units of the relative cost are mmol^{-1} gDW^{-1}. This represents the change in nutrient uptake flux given the fractional change in amino acid requirement in biomass.

The absolute cost is derived from the relative cost by division by the original mmol^{-1} gDW^{-1} requirement of the amino acid in the biomass equation. The absolute amino acid is therefore unitless and represents what a mmol^{-1} gDW^{-1} increase in amino acid requirement has on the mmol^{-1} gDW^{-1} uptake flux of the given nutrient.

2.2.4 Flux balance analysis of gene importance

Estimation of *in vivo* gene importance using optimal and suboptimal flux balance solutions

The gene importance based on reaction activity in the *S. cerevisiae* model for three nutrient limited conditions, glucose, ammonium, and sulphate, was estimated at a growth rate fixed to 0.3 hr^{-1} and the model optimised for the minimum intake of one of the three nutrients.

A suboptimal solution for the same nutrient limiting condition was found by a setting a lower boundary on the flux of nutrient into the cell at a 5% increase of the minimum required nutrient entry. For example if the minimum glucose flux into the cell at growth rate 0.3 hr $^{-1}$ is x, then the lower boundary of glucose into the cell was increased to $1.05x$, where the amount of glucose entering the cell may be more than this value but not less. This effectively relaxes the constraints on the use of the limiting nutrient, and allows a flux solution that uses more of the nutrient than is required [30].

This additional suboptimal FBA analysis was performed when estimating gene cost as the internal fluxes examined are dependent on the model solution. Therefore when perturbing a reaction, the original flux value is dependent on how the model was optimised. Performing this suboptimal FBA analysis may identify if model optimisation has any significant effect on the prediction of reaction flux, which in turn affects gene cost estimation. Only the optimal FBA solution was considered when estimating amino acid cost in the previous section. This is because amino acid cost was estimated from perturbing the amino acid requirements in biomass, which are model parameters rather than variables estimated from an optimised model solution.

Determination of reaction variability in glucose, ammonium, and sulphate limitation

The use of flux for single gene associated reactions was classified into four categories. The first two categories were based on whether the reaction flux was unused and exhibited zero flux ('zero flux'), or at the maximum allowable flux in the FBA solution of ±1000 mmol^{-1} gDW^{-1} hr^{-1} ('at maximum'). If the reaction did not fall into either of these two categories the reaction was classified on whether the flux could vary in the FBA solution given the cell growth rate and nutrient limitation. To do this for each reaction the model was re-optimised using MOMA to minimise the absolute flux through the reaction. If the reaction could not be minimised, this indicated the reaction was constrained in the solution space (‘constrained’), otherwise the reaction flux could be minimised and was classified as variable (‘variable’).

This resulted in one of four classifications for each single gene-associated reaction (‘at maximum’, ‘constrained’, ‘variable’, ‘zero flux’) in each of the three nutrient limitations for both optimal and suboptimal solution estimation.

Determination of reaction flux in glucose, ammonium, and sulphate limitation

The flux for each single gene associated reaction was recorded for all three nutrient limiting conditions for both optimal and suboptimal solutions. Reactions of near zero flux (± 0.001 mmol^{-1} gDW^{-1} hr^{-1}) in all three conditions were ignored.

The reaction flux of each of the single gene encoded reactions was classified into three categories. Classification was performed using one dimensional k-means clustering with three centres. The clustering was performed on the combined optimal and suboptimal non-zero variable reaction fluxes. The resulting clusters approximated reactions with zero flux, small flux, and large flux. Clustering was performed using the Hartigan and Wong algorithm [45] for k-means clustering.

Determination of reaction flux control on glucose, ammonium, and sulphate uptake

The degree of control for each of the variable flux reactions was estimated for glucose, ammonium, and sulphate limiting conditions, in both the optimal and suboptimal solution using a similar approach to Varma *et al.* [31]. This approach aims to estimate the importance of a reaction through perturbations in reaction flux as opposed to complete reaction removal. This method attempts to be more biologically relevant than knockout simulations since small mutations in enzyme encoding genes are likely to be more common than complete gene loss.

To estimate the reaction control on nutrient uptake flux, the flux through each reaction was fixed over five points in the range of 100% - 95% of the original reaction flux in the initial solution. For each point change in reaction flux the model was re-optimised using MOMA, to identify the corresponding change in the nutrient uptake flux.

The slope of the change in the nutrient uptake flux (mmol^{-1} gDW^{-1} hr^{-1}) as a response to changes in reaction flux (mmol^{-1} gDW^{-1} hr^{-1}) was the unitless control, or fitness effect of small changes in reaction rate, for that reaction in that nutrient limitation. This was performed in both optimal and suboptimal FBA solutions.

2.3 Results

2.3.1 The *S. cerevisiae* iND750 genome scale model

Numbers of reactions and genes

Each of the 1266 reactions in the *S. cerevisiae* model may be associated with one or more genes. Each gene may also be associated with multiple reactions. Table 2.3 illustrates the number and types of associations between genes and reactions. Each reaction in the model can be categorised based on the type of associations with known *S. cerevisiae* genes.

The total number genes included in the model is 750. Of the 1266 reactions, 810 are associated with one or more genes. Reactions unassociated to any gene may represent gaps in metabolic knowledge, but the reaction must exist to allow flux balance analysis. An common example of this are membrane transporters moving a metabolite between compartments.

The set of gene-associated reactions includes those catalysed by single genes, paralogs and multimeric enzymes, where the genes may be associated with multiple other reactions. A subset of this are 579 reactions associated with only a single gene. This excludes paralogs and multimeric genes, but still includes enzymes associated with more than one reaction. A further subset of this category are the reactions associated to a single gene, where the gene is associated with no other reactions. These genes may be expected to have only a single point of effect in *S. cerevisiae* metabolism, and are therefore useful for analysis of the direct relationship between genes and their corresponding metabolic effect.

Description	Number
Genes	750
Reactions	1266
Reactions with at least one associated gene	810
Reactions with a single gene associated	579
One-to-one association between reaction and gene	262

Table 2.3: Comparison of numbers and types of associations between genes and reactions in the *S. cerevisiae* iND750 model.

2.3.2 A novel approach using genome scale models to estimating amino acid cost

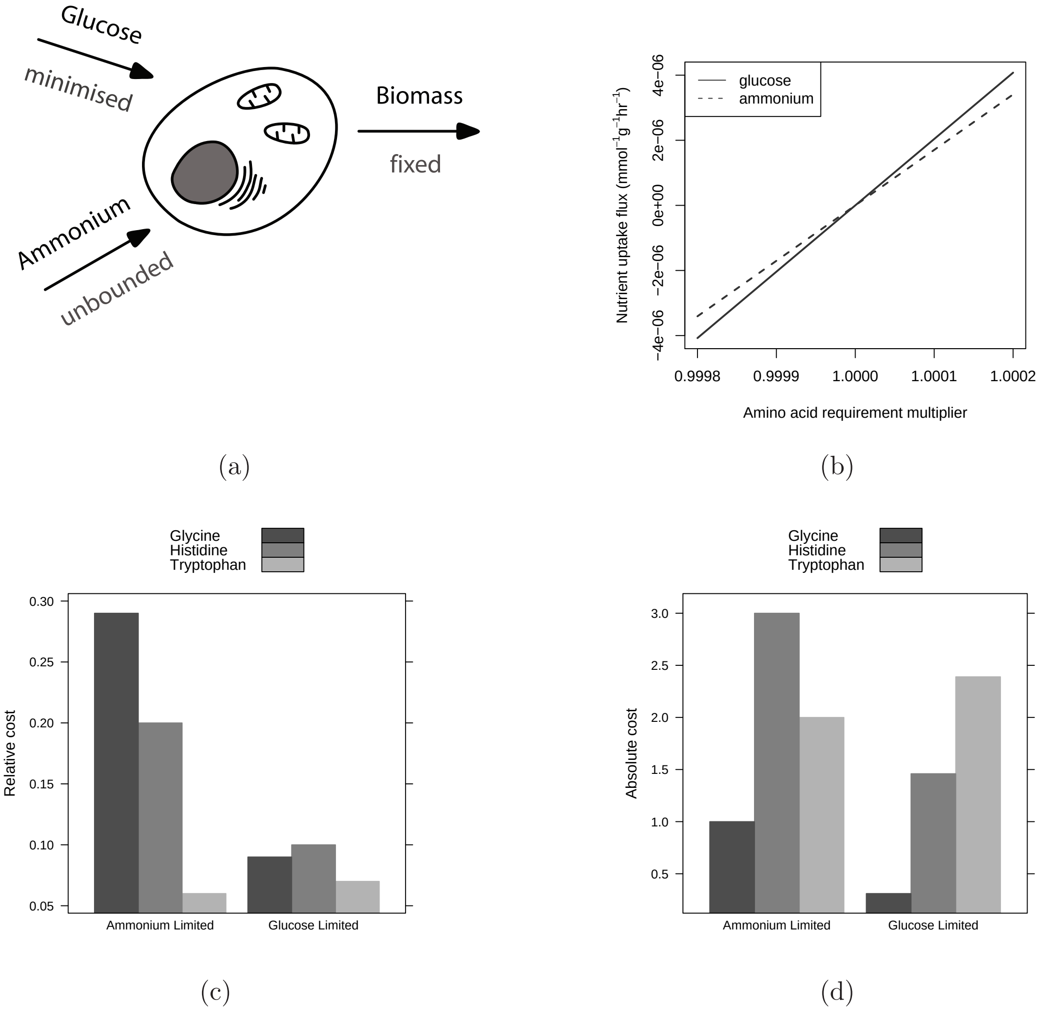
The introduction of this chapter described how metabolic control analysis can be used to estimate the importance of each reaction in a kinetic model. This section shows how similar principles can be used to estimate the sensitivity of *S. cerevisiae* growth to changes in amino acid requirements using a stoichiometric model. This sensitivity can be interpreted as the biosynthetic cost of the amino acid in the limiting nutrient.

Figure 2.1 shows a schematic version of the genome scale model method used to estimate the amino acid biosynthetic cost using FBA. Figure 2.1a illustrates the model configuration for the simulation of a glucose limited environment. Here the growth rate is fixed, and glucose uptake is minimised as the objective function. The entry of other nutrients, shown in the figure by ammonium, is unbounded and therefore ammonium is assumed to be non-limiting on growth.

In each nutrient-limiting environment, the biomass requirement of each amino acid is manipulated around the original value in the biomass equation in the range of ±0.0002%, and the model is re-optimised to determine the corresponding effect on the uptake flux of the limiting nutrient. Figure 2.1b provides examples of the slopes in changes in tryptophan requirement and the corresponding response in flux of either glucose and ammonium entry, depending on which nutrient entry is minimised.

Two types of amino acid costs were estimated by changing the biomass requirement of the model. The first is described as relative amino acid cost and is estimated from small percentage changes in biomass requirement. This relative estimate of amino cost can be rescaled to an absolute per molecule cost to reflect absolute changes in amino acid requirement.

One further consideration in estimating amino acid cost is the growth rate at which amino acid costs were estimated. For instance the absolute cost of tryptophan estimated at 0.2 hr^{-1} is twice the cost of tryptophan estimated at 0.1 hr^{-1}. To control for growth rate when estimating amino acid cost, the costs were estimated at a three levels of feasible yeast growth rates 0.1 hr^{-1}, 0.2 hr^{-1} and 0.3 hr^{-1}. Each amino acid cost is then rescaled by the growth rate to be independent of the growth rate at which it was estimated. The amino acid costs presented in this work are those estimated at the 0.3 hr^{-1} growth rate, though



(a)

(b)

(c)

(d)

Figure 2.1: Outline of the method used for the estimation of amino acid cost. *a*) is a schematic representation of simulating glucose limited conditions using a genome scale model and FBA. The growth rate of the model is fixed at a constant value and the objective function of the simulation minimises the cellular entry of a specific nutrient, which in the example is glucose. All other nutrient entering the model, illustrated here with ammonia, are unbounded. *b*) Relationship between altered tryptophan requirement and the response in either glucose and ammonium uptake. *c*) Estimated relative costs for tryptophan, glycine and histidine in ammonium and glucose limiting conditions. *d*) Absolute costs for the same amino acids in the same conditions.

growth rate had little effect of cost estimation with the largest variation being 0.000537^{-1} gDW^{-1} hr^{-1} for the phenylalanine $R_{glucose}$ cost, and 0.0145 for the tryptophan $A_{glucose}$ cost.

Figures 2.1c and 2.1d contrast the relative and absolute costs for tryptophan, histidine and glycine in ammonium and glucose limitation. The relative costs indicate that glycine is the most expensive amino acid, followed by histidine and tryptophan under ammonium limitation. This indicates that even though glycine is a small amino acid, the overall use in biomass makes glycine expensive in terms of synthesis from ammonium. The same relative costs in glucose limitation are comparably much less than the ammonium limited relative costs, which may be indicative of the greater expense of using nitrogen in *S. cerevisiae* biomass.

Comparing the absolute costs, tryptophan is a biosynthetically expensive amino acid in glucose limitation but is cheaper in nitrogen limitation. In contrast, histidine, which is a cheap amino acid in glucose limitation, is more expensive in ammonium limitation illustrating how per-molecule biosynthetic costs may vary depending on which nutrient is limiting. Glycine, a small amino acid with a single hydrogen side chain is cheap in both nutrient limitations. These differences across environments in biosynthetic cost underscore the importance of estimating cost using a flexible systems biology approach.

Estimated amino acid costs in glucose, ammonium, and sulphate limitation

Using the method described in the previous section, relative and absolute amino acid costs were estimated for all twenty amino acids in three simulated nutrient limiting conditions: glucose, ammonium, and sulphate. For clarity costs are described as either $R_{nutrient}$ or $A_{nutrient}$ for relative and absolute costs, respectively, in a given nutrient limiting condition. The estimated amino acid costs are shown in Table 2.4 and illustrated in Figure 2.2 together with amino acid costs previously described in the literature. The correlation between the costs estimated here and described in the literature is shown in Table 2.5.

The $A_{glucose}$ cost is highly correlated with several previous measures of amino acid cost. The energetic cost derived by Akashi and Gojobori [34] has the highest correlation coefficient (Spearman R = 0.94, $p < 0.0001$), but $A_{glucose}$ is also correlated (Spearman R values > 0.7, $p < 0.0001$) with Craig and Weber's energetic

	$A_{glucose}$	$A_{nitrogen}$	$A_{sulphur}$	$R_{glucose}$	$R_{nitrogen}$	$R_{sulphur}$
ala	0.50	1	0	0.23	0.46	0.00
arg	1.39	4	0	0.22	0.64	0.00
asn	0.79	2	0	0.08	0.20	0.00
asp	0.61	1	0	0.18	0.30	0.00
cys	0.75	1	1	0.00	0.01	0.01
gln	0.92	2	0	0.10	0.21	0.00
glu	0.86	1	0	0.26	0.30	0.00
gly	0.31	1	0	0.09	0.29	0.00
his	1.46	3	0	0.10	0.20	0.00
ile	1.21	1	0	0.23	0.19	0.00
leu	1.21	1	0	0.36	0.30	0.00
lys	1.31	2	0	0.38	0.57	0.00
met	1.25	1	1	0.06	0.05	0.05
phe	1.84	1	0	0.25	0.13	0.00
pro	0.99	1	0	0.16	0.16	0.00
ser	0.49	1	0	0.09	0.19	0.00
thr	0.69	1	0	0.13	0.19	0.00
trp	2.39	2	0	0.07	0.06	0.00
tyr	1.77	1	0	0.18	0.10	0.00
val	0.96	1	0	0.25	0.26	0.00

Table 2.4: *S. cerevisiae* FBA estimated absolute and relative amino acid costs. The units of the relative costs ($R_{nutrient}$) are millimoles of nutrient per gram dry weight biomass (mmol^{-1} gDW^{-1}). The absolute costs ($A_{nutrient}$) are unitless. All costs are estimated in the iND750 *S. cerevisiae* model using the COBRA toolbox. Estimates are rounded to two decimal places where necessary.

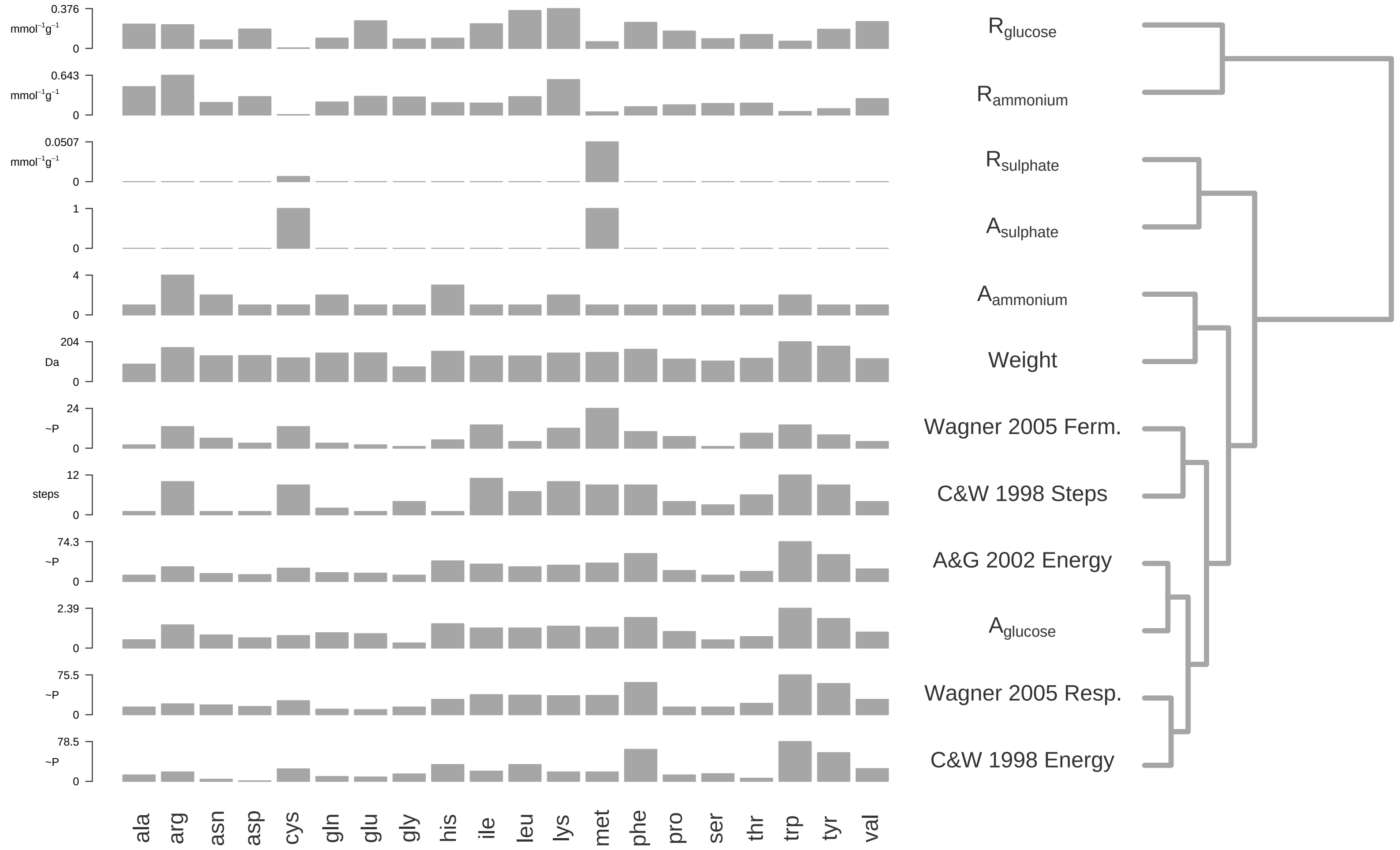


Figure 2.2: Amino acid cost estimates are shown as bar charts on the left hand side. Each axis shows the minimum and maximum value of each cost type. The correlations between different costs are illustrated as a dendrogram on the right hand side computed by complete agglomerative clustering using Spearman's rank correlation distance between costs. The cost values are shown in Table 2.2 and Table 2.4, the correlation between costs are given in Table 2.5

	A&G Energy	C&W Energy	C&W Steps	Wager Ferm.	Wagner Resp.	Weight	A_{glu}	R_{glu}	A_{amm}	R_{amm}	A_{sul}	R_{sul}
A&G Energy		0.81	0.70	0.77	0.87	0.75	0.94	0.06	0.51	-0.45	0.62	0.37
C&W Energy	0.00		0.62	0.45	0.81	0.44	0.73	0.07	0.34	-0.40	0.36	0.17
C&W Steps	0.00	0.00		0.80	0.72	0.37	0.59	0.03	0.34	-0.33	0.38	0.27
Wagner Ferm.	0.00	0.05	0.00		0.71	0.52	0.65	-0.16	0.52	-0.46	0.68	0.50
Wagner Resp.	0.00	0.00	0.00	0.00		0.52	0.75	0.08	0.31	-0.43	0.42	0.24
Weight	0.00	0.05	0.11	0.02	0.02		0.82	0.02	0.57	-0.16	0.46	0.20
A_{glu}	0.00	0.00	0.01	0.00	0.00	0.00		0.18	0.56	-0.25	0.54	0.26
R_{glu}	0.81	0.78	0.89	0.50	0.74	0.94	0.44		-0.25	0.63	-0.55	-0.30
A_{amm}	0.02	0.14	0.14	0.02	0.19	0.01	0.01	0.29		-0.11	0.55	0.09
R_{amm}	0.05	0.08	0.16	0.04	0.06	0.51	0.30	0.00	0.65		-0.71	-0.50
A_{sul}	0.00	0.12	0.10	0.00	0.06	0.04	0.01	0.01	0.01	0.00		0.76
R_{sul}	0.10	0.46	0.25	0.02	0.30	0.41	0.26	0.20	0.70	0.02	0.00	

Table 2.5: Spearman's rank correlation between amino acid costs estimated here and those previously reported in the literature described. Correlation coefficients are shown in upper right half, p values in the lower left. Where $p < 10^{-16}$ the value is approximated to 0. The abbreviations for each cost estimate are as follows: A&G Akashi and Gojobori [34], C&W Craig and Weber [33], Wagner fermentative and respiratory [32]. Molecular weight was used by Seligmann [36]

.

cost [33], Wagner's respiratory cost [32] and molecular weight [36]. The correlation with manually curated cost measures validates a systems biology approach using FBA to capture the energetic cost of synthesising amino acids.

The $A_{ammonium}$ costs show a weak non-significant correlation (Spearman R between 0.5 - 0.6, $p < 0.5$) with three previous measures of cost: Akashi & Gojobori, Wagner fermentative, and molecular weight. The $A_{ammonium}$ and $A_{sulphate}$ costs are proportional to the corresponding nitrogen and sulphur content of the amino acid. This suggests the biosynthetic costs in these two limitations can be measured as the requirement of the atoms that are limiting in the amino acid.

The $R_{glucose}$ and $R_{ammonium}$ costs show no correlation with previous measures of amino acid cost. The strongest (non-significant) correlation is with Wagner fermentative growth (Spearman R = -0.16, $p = 0.5$) but this is a negative non-significant correlation. The $R_{glucose}$ and $R_{ammonium}$ costs are however correlated with each other (Spearman R = 0.63, $p < 0.0001$).

Summarising both the relative and absolute estimates of amino acid cost, the absolute costs reflect the per molecule biosynthetic cost, while the relative costs reflect the absolute cost scaled by the quantity of the amino acid in biomass. For example, the $A_{sulphate}$ costs of methionine and cysteine are both 1, which reflects the single sulphur atom in each amino acid. The $R_{sulphate}$ cost for methionine is however much greater than that of cysteine, as methionine is used proportionally more than cysteine in the biomass equation.

Comparison of amino acid costs in *S. cerevisiae* and *E. coli*

The relative ease with which amino acid costs can be calculated for species where there is a genome scale model, compared with previous manual curation of amino acid cost, allows comparison of amino acid costs estimated in divergent species such as *S. cerevisiae* and *E. coli*. Figure 2.3 compares both $R_{glucose}$ and $A_{glucose}$ costs for *E. coli* and *S. cerevisiae* with two costs described in the literature: molecular weight [36] and Akashi and Gojobori's energetic cost [34].

Figure 2.3 shows that the absolute costs for both species show a small amount of variance in their correlation with the cost measures described in the literature. Both *E. coli* and *S. cerevisiae* absolute cost measures are well correlated with the Akashi and Gojobori's energetic cost [34]. Comparing variation between species, the *E. coli* $A_{glucose}$ cost is better correlated with the Akashi and Gojobori energetic cost than the *S. cerevisiae* equivalent (Spearman R = 0.99 vs. 0.94, $p < 0.0001$).

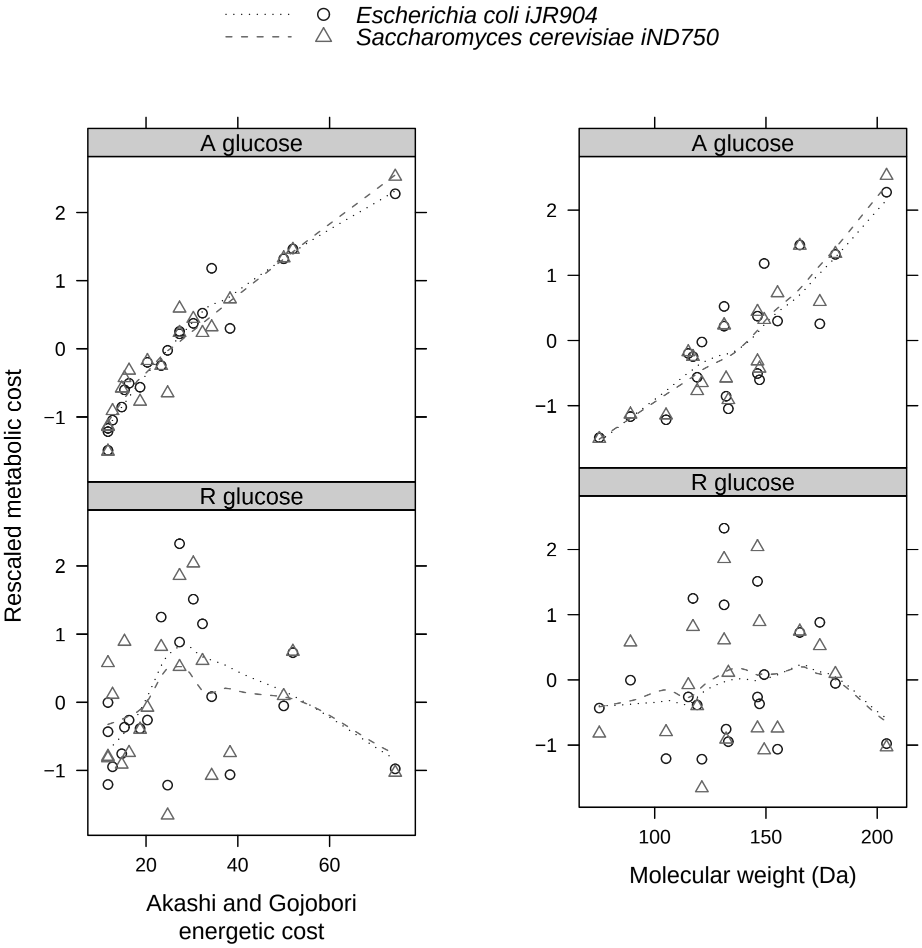


Figure 2.3: Comparison of estimated $A_{glucose}$ and $R_{glucose}$ costs Akashi and Gojobori's energetic cost [34] (left), and molecular weight [36] (right). On the y axis are the amino acid costs estimated using FBA. Both *S. cerevisiae* and *E. coli* measures are included to illustrate correlation of cost estimates between species and general application of the FBA approach to estimating amino acid cost. Estimated cost values have been rescaled around their mean value to allow comparisons across species. The trends in each plot are drawn using 'loess' smoothing.

This may be expected given the Akashi and Gojobori measures of cost were estimated for *E. coli*. In comparison molecular weight shows a linear relationship with the absolute measures of costs, but shows a greater degree of variation with the costs from both species. Overall the species difference between the two $A_{glucose}$ cost measures is marginal and they are highly correlated (Spearman R = 0.94, $p < 0.0001$) with each other indicating the glucose cost of synthesising amino acids does not vary between species.

The comparison of relative measures of cost illustrates the weak correlation with previous estimates of cost. For the both molecular weight and Akashi and Gojobori's energy cost estimates, neither *E. coli* or *S. cerevisiae* $R_{glucose}$ cost estimates show any correlation. Furthermore relative costs differ more between species (Spearman R = 0.74, $p < 0.0002$) than absolute costs. This indicates absolute costs are relatively invariant of the species in which they were estimated, while relative costs are species specific and vary as a function of biomass utilisation.

2.3.3 A novel approach to estimating gene importance

The previous section described how genome scale models can be used to estimate biological cost in amino acid synthesis using small scale perturbations to the stoichiometric model. This section extends this approach to estimate the importance of a gene to organism fitness via examination of the effects of perturbing the corresponding enzymatic reaction. Gene dispensability has already been explored extensively in the literature through deletion of reactions in genome scale models [37, 38, 39, 40]. Gene loss however may be expected to be relatively infrequent event in nature compared with small mutations in the gene sequence [37]. Because of this, gene importance to fitness is instead examined in terms of constraints on changes in reaction flux, estimated reaction flux, and how changes in reaction flux affect growth rate. The aim of each of these measures is to determine a single quantitative value for each gene which may then be used to understand the control and impact of the encoding gene on *S. cerevisiae* metabolism.

Effect of constraints on changes in reaction flux

The first measure of reaction importance considered is the constraint on changes in reaction flux in maintaining cellular growth. This approach considers whether

		Optimal		Suboptimal	
Type	Limitation	single gene	all reactions	single gene	all reactions
at maximum	glucose	3	18	3	16
	ammonium	4	25	4	20
	sulphate	4	25	4	24
constrained	glucose	60	180	58	181
	ammonium	58	173	59	174
	sulphate	58	173	59	176
variable	glucose	25	144	27	133
	ammonium	30	146	33	166
	sulphate	30	146	34	164
zero flux	glucose	174	924	174	936
	ammonium	170	922	166	906
	sulphate	170	922	165	902

Table 2.6: Types of reaction constraints in glucose, ammonium and sulphate limiting simulations. Reactions at maximum are at ±1000 mmol^{-1} gDW^{-1} hr^{-1} flux. Flux through constrained reactions cannot be decreased in the FBA solution whilst flux through variable reactions can. Zero flux reactions have no flux and are unused in the given condition. Fluxes for single gene encoded reaction are compared with all genes in the *iND750* model.

reaction flux can be varied while still allowing a valid FBA solution, focusing on whether the reaction is used and if the flux can be varied. This approach aims to identify genes under a large degree of selective constraint where the reaction flux cannot be varied, even if the reaction flux is small.

Each reaction was categorised based on its activity in different environmental conditions and whether the flux could be decreased whilst still allowing the model to be optimised. The aim of this approach was to determine if the flux for a reaction is used or maintained at a particular value across each of the three nutrient limited conditions. This approach may identify genes under an evolutionary pressure for one or more environmental conditions. Each reaction category determined is described as follows and the results are shown in Table 2.6 and Figure 2.4.

Reaction rate at maximum. The reaction is at the allowable limit (±1000 mmol^{-1} gDW^{-1} hr^{-1}) for flux in the FBA solution. These represent artificial boundaries placed on the model solution when performing linear optimisation. The numbers of these reactions was limited and no more than four were observed in any condition.

Constrained reaction rate. The flux through the reaction is constrained in the solution and cannot be reduced. The flux through this reaction is essential to producing biomass in the given nutrient limitation. The numbers of these reactions was relatively invariant between all FBA solutions considered ranging from 58 to 60.

Variable reaction rate. These reactions are used in the solution, for which the absolute reaction flux can be reduced toward zero. These reactions represent flexibility in the flux phenotypes of the metabolic network. The number of these reactions type varied between solutions ranging from 25 to 34.

Reaction unused. These reactions have zero flux and are unused in the FBA solution. This is the largest proportion of reactions across all simulations ranging from 165 to 174.

As Table 2.6 shows there is little difference in the number of genes assigned to different reaction categories between the optimal and suboptimal FBA simulations, with a trend for greater numbers of variable reactions in the suboptimal simulations compared with the optimal simulation. This might be expected given that the suboptimal solution is less constrained in solving the flux phenotype because of the 5% excess of nutrient entering the cell. Estimated flux constraints for all model reactions, not just single gene associated reactions, are included in the table and suggest that flux constraints for the single gene associated reaction subset show similar trends with all reactions in the model.

The overlap of reaction categories between in the optimal and suboptimal solution is illustrated in Figure 2.4. The reactions at maximum rate are not included given the small number of genes in this class across conditions. The optimal and suboptimal FBA solutions showed similar distributions of reaction constraints across environments. There were only small numbers of reactions categorised as constrained in only one of the three environments and almost no reactions constrained in two environments.

The numbers of reactions categorised as variable in only two environments was almost half the number categorised as variable in all three. The number of

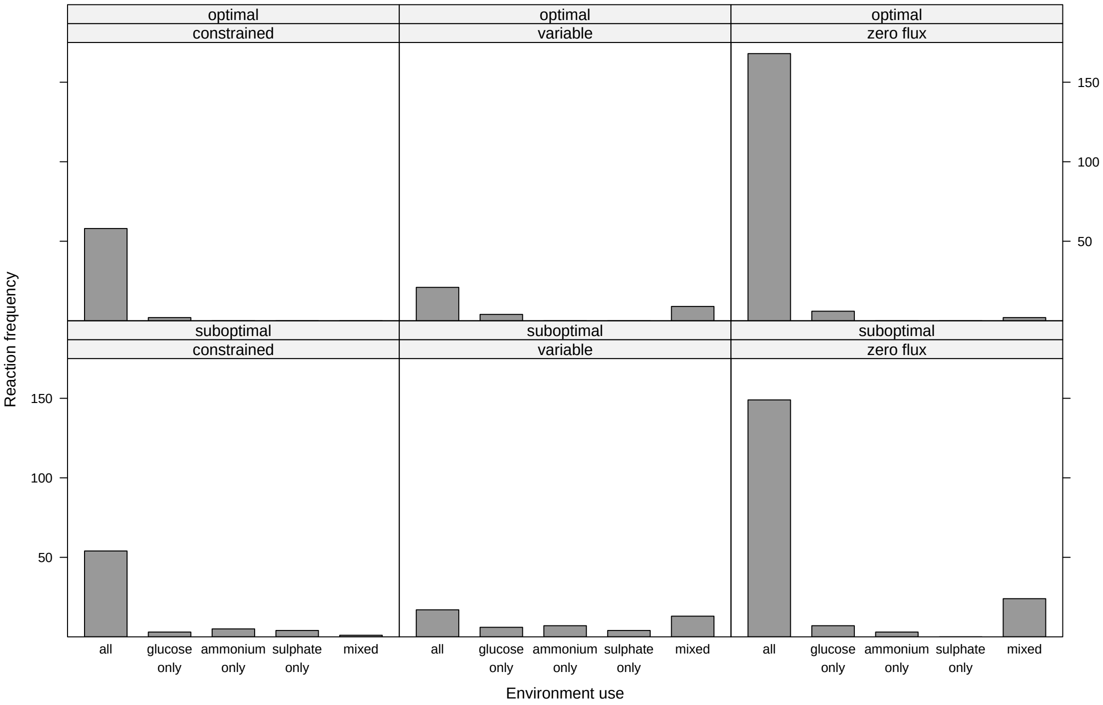


Figure 2.4: Overlap of single gene-associated reaction flux type between glucose, ammonium and sulphate limiting conditions for the optimal and suboptimal FBA solutions. The mixed category represents a reaction in two of the three nutrient limiting environments. Constrained reaction flux cannot be reduced whilst variable reactions can. Zero flux reactions are unused in the solution.

reactions that were observed as variable in only a single environment was much greater in the suboptimal solution than the optimal. In comparison to the other two reaction categories, the variable category showed the least consistency between environments. The variable category reactions however showed the lowest frequency of the three considered categories.

In the last category unused reactions tended to be consistently unused across all three environments. In the optimal solution the numbers of reactions unused in only one or two environments was very small, with the greatest number observed in glucose limitation. In the suboptimal solution the number of reactions unused across all three environments was again the majority, however there was a much greater frequency of reactions unused in two of the three examined environments.

Overall in each of the examined reaction categories, the majority of reactions are consistently the same category across all three environments. This suggests that constraints on reaction flux changes do not flux vary between environments.

Estimated reaction flux across environments

The second measure of gene importance considered is the flux of the reaction encoded by the gene. The hypothesis for this approach is that genes encoding reactions with a large flux value may be under greater selection pressure against mutations than genes catalysing reactions with relatively small flux activity. The assumption is that a mutation in a gene will cause a deleterious reduction in encoded reaction activity which will have a proportional effect on organism fitness.

The *in silico* flux for of the each single gene-associated reaction (described in Table 2.3) was calculated for glucose, ammonium and sulphate limiting conditions. The analysis was restricted to the subset of single gene-associated reactions where mutations in the sequence of these genes should have only a single corresponding phenotype in *S. cerevisiae* metabolism. Reaction flux was estimated for both the optimal and suboptimal FBA solution, to determine if constraints on finding an optimal FBA solution affect the predicted reaction flux.

For each optimal or suboptimal solution, reactions with a flux less than ±0.001 mmol^{-1} gDW^{-1} hr^{-1} in all three nutrient limiting conditions were interpreted as being zero and ignored, as were reactions with identical flux in all three conditions. This identified the set of reactions which exhibited variable fluxes between glucose, ammonium and sulphate limiting conditions. The resulting set of reactions is summarised in Table 2.7. The distribution of reaction fluxes for the three nutrient

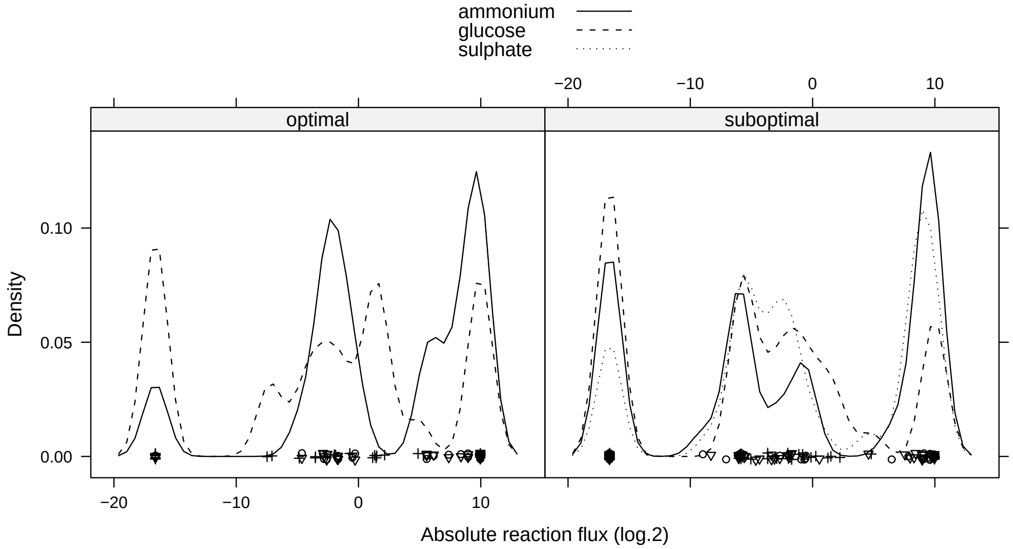


Figure 2.5: Reaction flux in glucose, ammonium and sulphate limiting conditions. The points correspond to the reactions with variable non-zero flux differences between conditions described in Table 2.7. Each line is the Gaussian kernel density for the observations using a bandwidth of 1. The x-axis is the absolute mmol^{-1} gDW^{-1} hr^{-1} reaction flux on a $\log_2$ scale. Zero flux reactions have a small arbitrary coefficient added and correspond to the leftmost peak in each plot.

limiting conditions in optimal and suboptimal FBA solutions is illustrated in Figure 2.5.

Of the genes that vary in the optimal solution, the ammonium and sulphate limiting environments have identical flux distributions (the solid line of the ammonium distribution covers dotted line of the sulphur distribution). In both figures the leftmost peak corresponds to unused reactions with zero flux where a small arbitrary coefficient was added to display these reactions in the figure. The

Reaction Type	Optimal	Suboptimal
Near zero flux in all conditions	176	171
Identical flux in all conditions	61	51
Variable non-zero flux differences between conditions	25	40

Table 2.7: Comparison of reaction flux in glucose, ammonium, and sulphate limiting conditions. The total number of reactions assessed was 262 and comprises all reactions in the *S. cerevisiae* genome scale model with single gene-association.

middle peak corresponds to reactions with a flux ∼1 mmol^{-1} gDW^{-1} hr^{-1} while the rightmost peak indicates reactions with a relatively larger flux. The shape of the ammonium and sulphate flux distributions is trimodal and skewed towards reactions with a high flux value. The ammonium and sulphate limiting condition have a larger number of single gene-encoded reactions whose flux is greater than 1 in optimal conditions. Glucose optimal flux is distributed towards reactions with a low flux or reactions which are unused.

The right plot of Figure 2.5 illustrates reaction flux in the suboptimal FBA simulation. All three distributions are again trimodal, but in this case the use of flux in ammonium and sulphate limitation are no longer identical. The distribution of flux in ammonium limitation is still skewed towards high flux reactions but exhibit larger numbers of unused reactions, compared with the optimal solution. This observation indicates an increased number of inactive reactions in the single gene reaction set suggesting activation in the set of reactions with multiple gene associations, or that the suboptimal solution requires a distribution of smaller flux values. Sulphate limitation maintains a large numbers of both low and high flux reactions with few unused reactions. Of all three nutrient limitations the glucose flux distribution remains relatively unchanged between optimal and suboptimal FBA solutions. This indicates the reactions considered in this analysis may be constrained to specific fluxes even in a suboptimal FBA solution. Comparing both optimal and suboptimal solutions, single gene associated reactions tend to have a higher flux in ammonium and sulphate limitation. This indicates that single gene reactions carry more flux and are more likely to be used in ammonium or sulphate limitation, when contrasted with glucose limitation. Furthermore the flux distribution in ammonium and sulphate limitation are identical indicating these reactions are used in the same way across the environmental conditions.

Figure 2.6 compares the reaction fluxes each of the three limiting environments for either optimal or suboptimal FBA solutions. In the optimal simulation, as expected given the above overlap in distribution, the ammonium and sulphate flux distributions are perfectly correlated. The glucose limited optimal flux simulation is correlated with these identical ammonium and sulphate flux distributions (Spearman R = 0.455, $p = 0.022$). In contrast the suboptimal simulations the ammonium and sulphate limiting conditions show limited correlation with each other (Spearman R = 0.38, $p = 0.014$), but a greater degree of correlation with the glucose limited flux distribution (ammonium Spearman R = 0.424, $p = 0.006$;

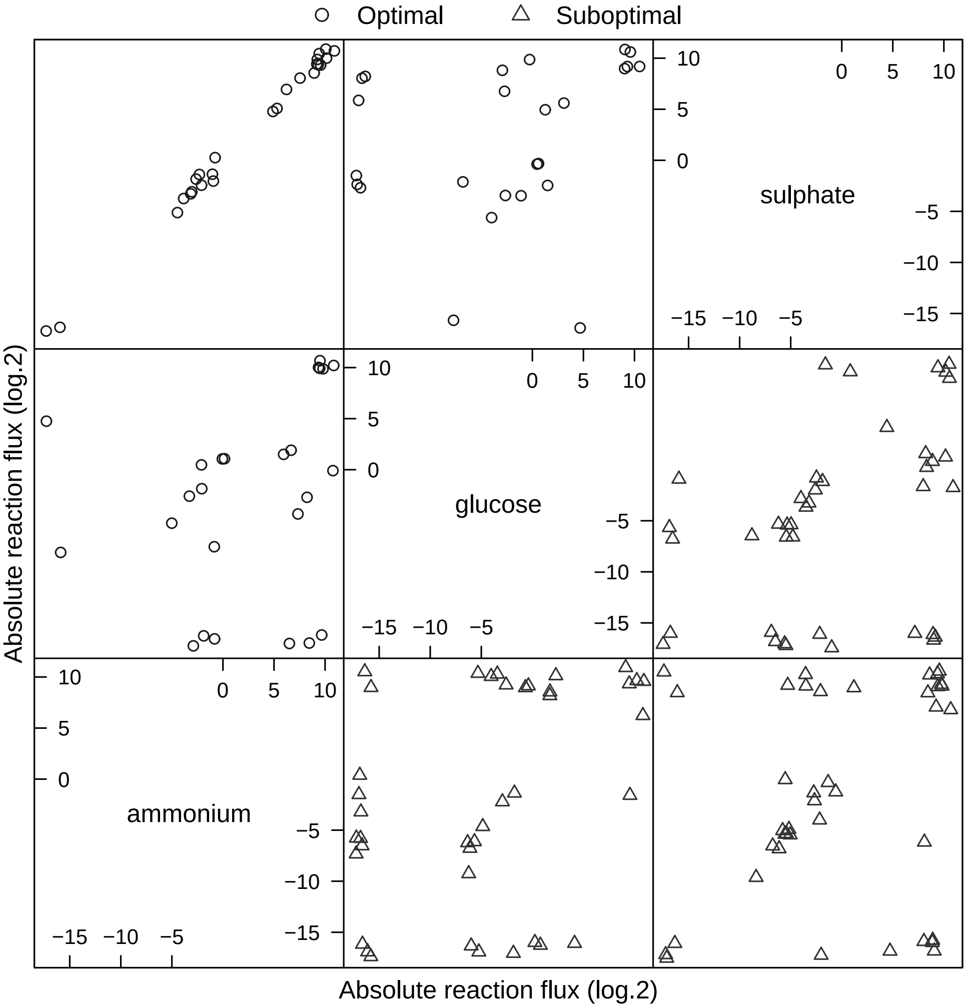


Figure 2.6: Flux of single gene encoded reaction with variable non-zero values between glucose, ammonium and sulphate limitation. The optimal FBA solution is shown by the upper left panels, the suboptimal solution in the lower right panels. The optimal solution leads to a lesser degree of correlation between glucose and the ammonium and sulphate flux distributions compared to the suboptimal solution. The ammonium and sulphate solutions are identical in the optimal solution. Each axis shows the absolute mmol^{-1} gDW^{-1} hr^{-1} reaction flux on a log$_2$ scale. A small amount of noise is added to each plot to illustrate overlapping points.

sulphate Spearman R = 0.549, $p = 0.0003$). Overall, the non-trivial variable flux reactions in each nutrient limitation show a moderate degree of correlation. Ammonium and sulphate limitation flux distributions are identical in the more constrained optimal FBA solution. This suggests that the FBA simulations for the single gene reaction set between each of the three considered environments show a relative degree of similarity in the use of flux.

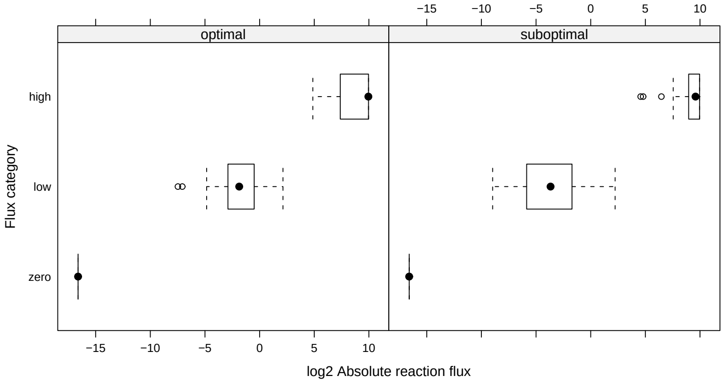
The two plots in Figure 2.5 show a trimodal distribution of reaction fluxes in each environment. The value at each peak may be approximated to reactions with either zero flux, with a flux close to 1, or a larger flux greater than 1. K-means clustering of the flux data was therefore used to classify the reactions into three categories based on flux, and is illustrated in Figure 2.7a. As expected the resulting k-means categories correspond to the peaks observed in the flux distribution plots. These categories are described as zero, low, and high flux reactions.

Figure 2.7b illustrates the use of these reaction flux categories between environments. In the optimal FBA solution a large proportion of high and low flux reactions show the same flux category across all three environments. The remainder of reactions in the high flux category are 'mixed' and show a high flux in two of the three on the environments but the not the third. The low flux category in the optimal solution show a number of reactions specific to just glucose limitation. There are no zero flux category reactions in all three environments as expected given the pre-filtering of the data. The majority of these zero flux reaction category appear only in glucose limitation.

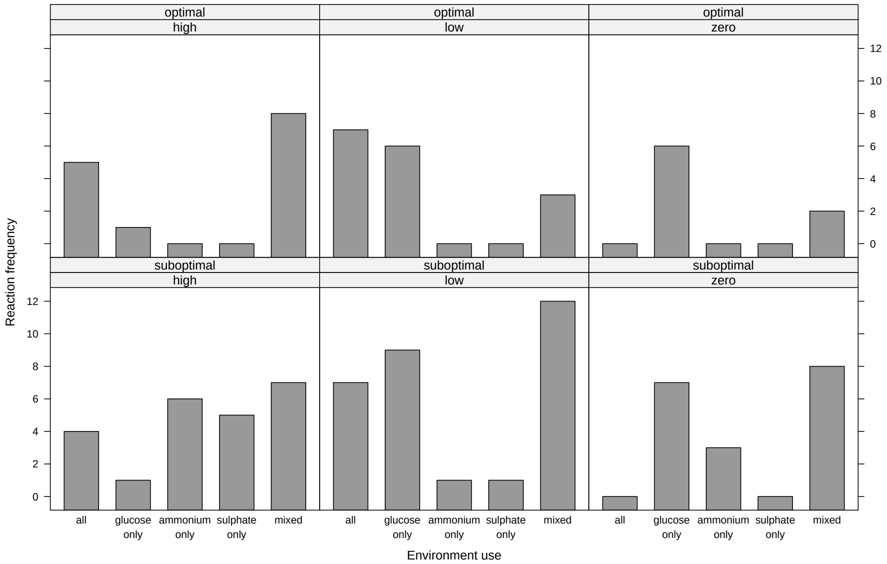
The flux categories estimated in the suboptimal FBA solution show a different distribution between environments. The high flux category reactions have a comparably greater frequency of reaction specific to either only ammonium or sulphate limitation. In comparison to the optimal FBA solution the low and zero flux reactions both show a greater frequency of reactions mixed between environments: appearing in two out of the three simulated environments. The zero flux category has three reactions unused specifically in ammonium limitation that were not observed for the suboptimal flux distribution.

Reaction flux control on nutrient uptake

In this section in a similar vein to MCA, perturbations are applied to the subset of variable flux single gene reactions identified above to estimate the effect of



(a)



(b)

Figure 2.7: Flux categories in glucose, ammonium, and sulphate limitation. a) Box and whisker plot of k-means determined qualitative categories for the non-zero flux variable reactions described in Table 2.7. b) Comparison of the qualitative reaction flux between the three nutrient limiting environments.

changes in reaction flux on nutrient uptake flux. The aim of this approach is to determine how sensitive each of the nutrient uptake fluxes are to changes in each of the examined reactions. In comparison to the previous sections where gene importance has been considered either in terms of reaction flux or constraints on reaction flux, this section attempts to derive a quantitative measure of the fitness effects of changes in reaction flux on organism growth. The set of gene encoding reactions with a greater fitness effects may be expected to be under greater selective pressure compared with those that show only small fitness effects.

In the above, a set of reactions was identified as being to able to vary in flux while still allowing a valid FBA solution. For each of these reactions, shown as their corresponding gene in Table 2.8, a series of small perturbations was used to reduce the absolute reaction flux in the FBA closer to 0, simulating a small deleterious mutation. The fitness cost of each reaction was determined as the absolute value of the slope between the change in reaction flux and the corresponding effect in the re-simulation of nutrient uptake flux. Figure 2.8 shows the distribution of these reaction fitness effects for all nutrient limitations in both the optimal and suboptimal FBA solutions. The median reactions sensitivities are shown in Table 2.8. In the optimal solution the reactions with the greatest flux control are found in glucose limitation with a peak furthermost to the right.

In the optimal solution, ammonium limitation exhibits one highly sensitive reaction while all other reactions appear to a negligible effect on uptake flux. Sulphate limitation shows a distribution of reaction control in between the distributions for glucose and ammonium. The median reaction coefficients for glucose, ammonium, and sulphate limitations are 10^{-9}, 10^{-12}, 10^{-7} respectively and indicate the reactions generally have little effect on nutrient uptake flux in the optimal solution.

The suboptimal distribution of reaction coefficients shows a lesser degree of control on uptake flux than was observed in the optimal solution. Glucose and sulphate limitation show a small number of sensitive reactions with a value near 1 while the remainder of the reactions exhibit a small amount of control on uptake flux in any nutrient condition.

Figure 2.9 compares the control coefficients for the reactions where a coefficient could be estimated in both the suboptimal and optimal FBA solutions. The glucose sensitivity estimates are not correlated (Spearman R = 0.366, p = 0.09)

Solution	Gene	Protein	System
both	YOL064C	Met22	Cysteine Metabolism
	YGR088W	Ctt1	Tyrosine, Tryptophan and Phenylalanine
	YDR300C	Pro1	Arginine and Proline Metabolism
	YKL120W	Oac1	Transport, Mitochondrial
	YDR050C	Tpi1	Glycolysis and Gluconeogenesis
	YDR111C	Alt1	Alanine and Aspartate Metabolism
	YMR083W	Adh3	Glycerolipid Metabolism
	YJL121C	Rpe1	Pentose Phosphate Pathway
	YOL126C	Mdh2	Oxidative Phosphorylation
	YER053C	Phc1	Transport, Mitochondrial
	YGR208W	Ser2	Glycine and Serine Metabolism
	YEL047C	Frds1	Transport, Mitochondrial
	YBL015W	Ach1	Pyruvate Metabolism
	YJR105W	Ado1	Nucleotide Salvage Pathway
	YER052C	Hom3	Alanine and Aspartate Metabolism
	YBR263W	Shm1	Glycine and Serine Metabolism
	YNL169C	Psd1	Phospholipid Biosynthesis
	YEL046C	Gly1	Threonine and Lysine Metabolism
	YCR012W	Pgk1	Glycolysis and Gluconeogenesis
	YIL155C	Gut2	Glycerolipid Metabolism
	YDL022W	Gpd1	Glycerolipid Metabolism
	YLR058C	Shm2	Glycine and Serine Metabolism
	YDL066W	Idp1	Citric Acid Cycle
	YKL085W	Mdh1	Oxidative Phosphorylation
	YJR077C	Mir1	Transport, Mitochondrial
	YBR166C	Tyr1	Tyrosine Tryptophan, and Phenylalanine
	YDR158W	Hom2	Alanine and Aspartate Metabolism
	YNL277W	Met2	Methionine Metabolism

Table 2.8: Single gene encoded reactions with variable reaction flux in either optimal, suboptimal or both FBA solutions.

Solution	Gene	Protein	System
optimal	YNL220W	Ade12	Purine and Pyrimidine Biosynthesis
	YLR174W	Idp2	Citric Acid Cycle
	YOR095C	Rki1	Pentose Phosphate Pathway
	YLR377C	Fbp1	Anaplerotic reactions
	YNL241C	Zwf1	Pentose Phosphate Pathway
	YDL215C	Gdh2	Glutamate metabolism
suboptimal	YOL140W	Arg8	Arginine and Proline Metabolism
	YLR142W	Put1	Arginine and Proline Metabolism
	YMR202W	Erg2	Sterol Biosynthesis
	YFR055W	Cys1	Methionine Metabolism
	YLL043W	Fps1	Transport, Extracellular
	YML008C	Erg6	Sterol Biosynthesis
	YPR035W	Gln1	Glutamine Metabolism
	YPL091W	Glr1	Other Amino Acid Metabolism
	YMR015C	Erg5	Sterol Biosynthesis
	YOR130C	Ort1	Transport, Mitochondrial
	YJR057W	Cdc8	Nucleotide Salvage Pathway
	YEL011W	Glc3	Alternate Carbon Metabolism
	YDL171C	Glt1	Glutamate metabolism
	YLR438W	Car2	Arginine and Proline Metabolism
	YAR035W	Yat1	Alanine and Aspartate Metabolism
	YKR080W	Mtd1	Folate Metabolism
	YLR056W	Erg3	Sterol Biosynthesis
	YGR012W	Csy1	Cysteine Metabolism
	YAL012W	Cys3	Methionine Metabolism

Continued

Limitation	Solution	Median ($\log_{10}$)
glucose	optimal	-9.34
	suboptimal	-10.35
ammonium	optimal	-11.53
	suboptimal	-8.11
sulphate	optimal	-7.73
	suboptimal	-7.10

Table 2.8: Median $\log_{10}$ reaction sensitivity on nutrient uptake in each nutrient limiting FBA simulation.

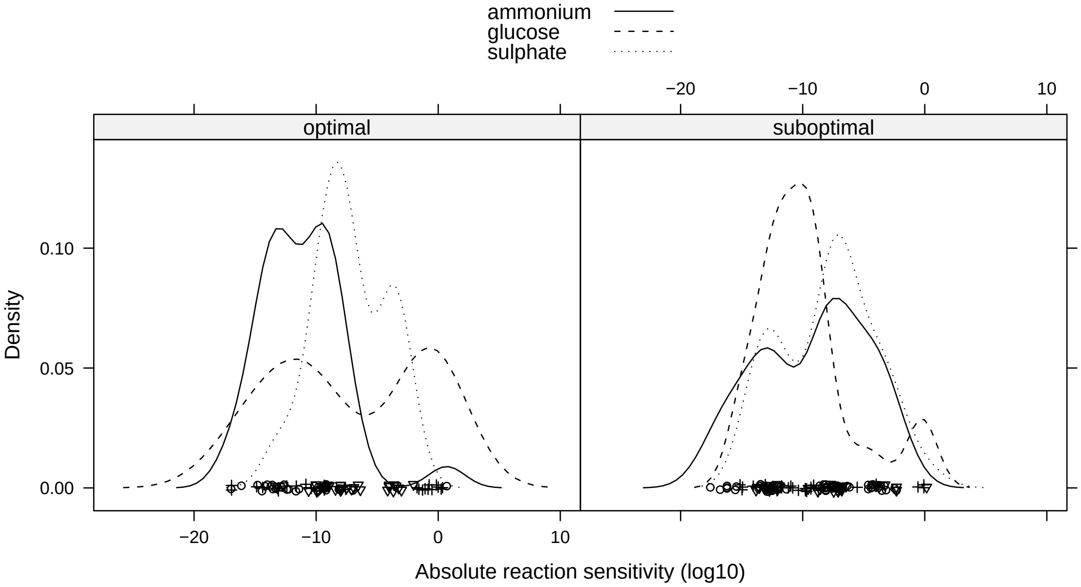


Figure 2.8: Reaction sensitivity in glucose, ammonium and sulphate limiting conditions. Each point is the $\log_{10}$ absolute value of the slope between the decrease in reaction flux, and the corresponding effect on nutrient uptake flux. The density line is the Gaussian kernel density for the observations using a rule-of-thumb[1] smoothing bandwidth.

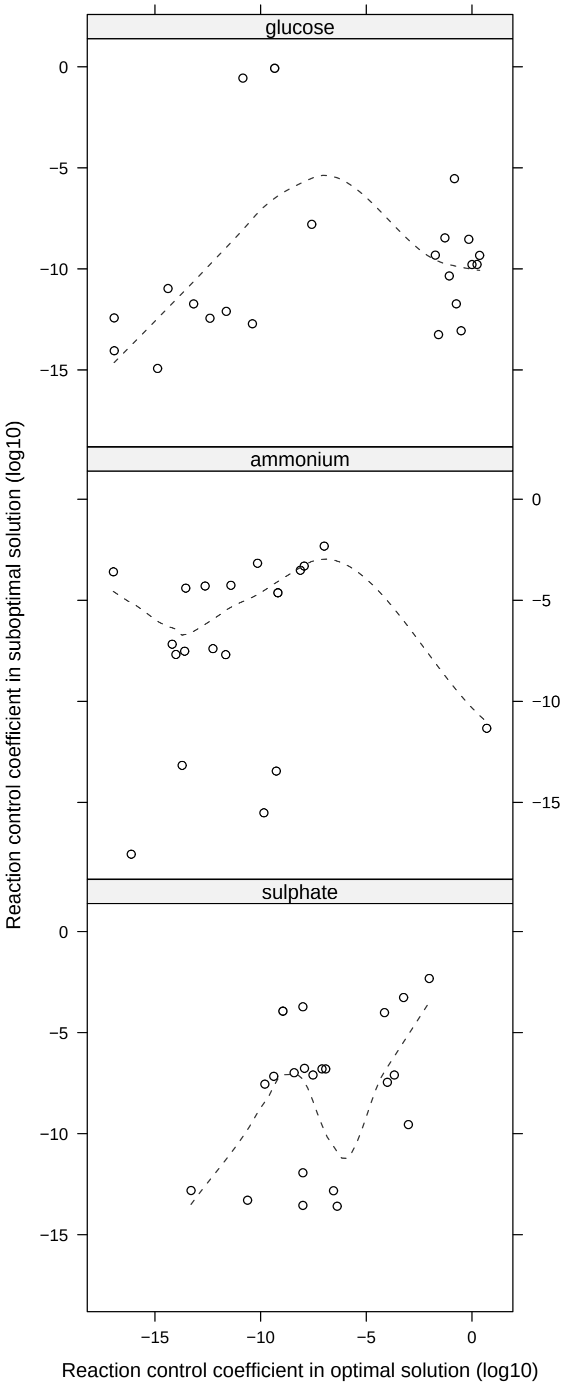


Figure 2.9: Comparison of reaction control coefficient between optimal and suboptimal solutions. Each point is the $\log_{10}$ absolute value of the slope between the decrease in reaction flux and corresponding effect on nutrient uptake flux. Loess smoothing is used to indicated trend.

and show a group of reactions which are highly sensitive in the optimal solution, but with limited sensitivity in the suboptimal solution. This indicates a set of reactions particularly constrained in the glucose limited optimal solution but much less so in the suboptimal solution. The ammonium (Spearman R = 0.291, p = 0.21) and sulphate (Spearman R = 0.242, p = 0.28) reaction sensitivities are non-linear between optimal and suboptimal conditions and show no significant correlation. These results highlight how constraints on the FBA solution can affect the estimated sensitivity of nutrient uptake to changes in individual reaction flux.

[1] Rule-of-thumb smoothing bandwidth is estimated from 0.9 times of either the standard deviation or interquartile range (which ever is smaller) divided by 1.34 times the sample size to the power of -0.2.

2.4 Discussion

2.4.1 Flux balance analysis predicts amino acid cost

A predicted 80% of cellular generated energy is expended in amino acid and protein synthesis [32]. This use of this energy in protein synthesis is expected to be under substantial cost minimisation selection pressure, particularly in organisms where the availabilty of energy sources is limited.

To the determine the effect of cost minimisation the majority of previous estimates of amino acid cost have been derived from manual curation of metabolic pathways. Using genome scale metabolic models allows the fast evaluation of amino acid cost for any species for which a genome scale model exists. This technique can be expanded to estimate the cost of any reaction or metabolite described in a genome scale model. This is substantial improvement over manual curation methods, both in speed and reproducibility. Furthermore, costs can be estimated under a variety of simulated environmental conditions.

The amino acid costs presented in this chapter were estimated for glucose ammonium, and sulphate limiting conditions (Table 2.4). The $A_{ammonium}$ and $A_{sulphate}$ amino acid costs are directly proportional to the nitrogen and sulphur content of the molecule. This indicates when either of these two nutrients are limiting that the biosynthetic cost of an amino acid is the nitrogen or sulphur atoms by the molecule. Glucose, however, is a source of both carbon atoms and energy and the $A_{glucose}$ cost may be expected to reflect a combination of both these factors. When compared with the literature reported cost measures the $A_{glucose}$ cost is most correlated with the Akashi and Gojobori [34] energetic cost, where small amino acids such as glycine and alanine are cheaper while larger more complex amino acids such as tryptophan and tyrosine are biosynthetically more expensive (see Table 2.4).

The relative amino acid cost estimates represent the absolute cost of an amino acid scaled by the requirement in *S. cerevisiae* biomass. The relative measures of amino acid cost computed here show little correlation with previous measures of amino acid cost. Furthermore, amino acids that are expensive based on their absolute cost are sometimes cheaper in terms of relative cost. For example, tryptophan has a high $A_{glucose}$ cost, but is instead much cheaper when the $R_{glucose}$ cost is considered. This reflects that expensive amino acids may be minimised in biomass composition. On the other hand lysine has one of the most expensive $R_{glucose}$ costs, but is relative inexpensive in $A_{glucose}$. Therefore a proportional increase in the requirement of lysine would have a greater effect on glucose uptake than of tryptophan, since lysine appears at greater quantities in *S. cerevisiae* biomass.

The absolute measures of amino acid cost are clearly useful as they show a strong correlation with previous estimates of amino acid cost as shown in Figure 2.2. In contrast the relative measures of amino acid cost show no correlation with previous cost estimates and their relevance to understanding the role of metabolic cost is ambiguous. The importance of a relative measure of amino acid cost may in understanding how the overall use of amino acids in genome has been shaped by cost selection pressures.

As an example the $R_{glucose}$ and $R_{ammonium}$ costs show a significant degree of correlation (Spearman R = 0.63), greater than that observed for the absolute estimates (Spearman R = 0.56). Assuming sugar-based energy limitation is the natural environment for *S. cerevisiae*, this leads to the possibility that optimisation of energy/carbon based costs in the proteome also results in optimisation of the nitrogen cost. If the optimisation of a protein sequence for carbon limitation also results in nitrogen limited optimisation this may provide a fitness advantage for maintaining protein production in both carbon or nitrogen limited environments.

The determination of a wide range of molecular costs will allow analysis of whether cost minimisation is a selection pressure for other aspects of the metabolome rather than just the subset of twenty amino acids. One application for this method would be to the estimate cost for the same metabolites across all existing species models. This would enable comparative metabolomic analysis, and potentially yield insights on the evolutionary selection pressures involved in maintaining metabolite levels. This *in silico* analysis, which was shown to work in *E. coli* as well as *S. cerevisiae*, would be simple and inexpensive to perform in comparison with *in vivo* metabolome analysis.

2.4.2 Predicting the importance of a gene

Interpreting gene importance using reaction phenotype

Gene importance through large scale gene knockout studies has been considered both *in vivo* and *in silico*. The aim of these studies is to determine the dispensability of a gene, and therefore estimate the importance of the gene to organism fitness. The hypothesis behind gene knockout studies is that a greater phenotypic effect indicates the greater the gene function to organism fitness. Nevertheless, it may be argued [46] that gene loss, and therefore loss of function, is a rarer event compared with gene mutation resulting in small changes in gene function.

The aim of this *in silico* study is to attempt to derive a measure of gene importance to organism fitness through methods other than complete removal from metabolism. The different estimates of gene importance considered were constraints on changes in reaction flux, reaction flux, and how reductions in reaction flux affect nutrient uptake. Each of these estimates was performed for glucose, ammonium, and sulphate limitation to relate fitness effects to specific environmental conditions. These fitness effects were estimated for both optimal and suboptimal FBA solutions to examine if a lesser level of constraint on model affects gene fitness estimates

Estimating the importance of a gene on an *in silico* phenotype requires determining the expected effect of each gene in a metabolic model. This analysis was restricted to the set of 262 genes in the *S. cerevisiae* model associated with only a single metabolic reaction. This excluded the set of 488 genes with multiple reaction associations. Alternative approaches have included multiple gene associated reactions in *in silico* gene knockout studies by treating reactions as still functional if there remains a paralog for the reaction, or have assumed the reaction is non-functional if the one unit of a multimeric enzyme is deleted [37]. In this analysis reactions are not being removed, but instead the size of the effect on nutrient uptake is estimated. Focusing on the set of genes with only one point of effect in metabolism allows greater confidence in assuming simulated gene function perturbations affect only a single reaction.

Gene importance through constraints on reaction flux

The constraints on changes in reaction flux may give insight into potential variation in gene importance. Constraints on reaction flux were considered as four categories: 'at maximum', 'constrained', 'variable' and 'zero flux'. The reactions at maximum flux are at the boundaries imposed on the linear optimisation where all reactions are restricted to values within ±1000 mmol^{-1} gDW^{-1} hr^{-1}. The biological feasibility of whether a reaction may reach this level of flux is questionable. The number of observed reactions at this maximum level is low and discarding them does not significantly reduce the number of evaluated reactions.

The constrained reactions represent those whose flux is at what can be described as an edge in the FBA solution where reduction of the reaction flux prevents a viable linear optimisation solution. As the flux cannot be reduced, these reactions all represent inviable knockout mutants, since any *in silico* flux reduction produces a lethal phenotype and gene knockout represent reducing the reaction flux to 0. These reactions are likely to be functionally important *in vivo* as the nutrient uptake flux is most sensitive to changes in these reaction fluxes. The frequency of constrained reactions in the suboptimal FBA solution was also considerable, indicating constraint on reaction flux is present even in relaxed model optimisations, and therefore these results are not an artifact of the method of optimisation.

The variable reactions are those where the flux can be reduced and still produce a viable model solution. The variability observed in these reactions may be due to alternate pathways for the production of the same metabolite, or it may be that the flux through a different reaction can be increased to compensate the reduction in flux through the perturbed reaction. Variable reactions represent non-unique flux distributions where multiple different combinations of flux distributions can produce the same *in silico* phenotype. It is likely that effect of a reduction in flux for variable reactions would have a lesser phenotypic effect *in vivo* than that of a constrained reaction, as they are buffered by other parts of the metabolic network. Comparing the numbers of variable reactions in the suboptimal and optimal solutions, the suboptimal solution identifies more variable reactions, which may be expected as the solution space is less constrained.

Zero flux reactions are unused in the *in silico* flux phenotype and were the majority flux category observed in the single gene-associated reactions. These reactions could be removed from the model without any effect on producing a viable FBA solution. These reactions may represent environment-specific reactions, necessary in a different combination of nutrients not simulated in this analysis. The numbers of zero flux reaction is relatively unchanged between optimal and suboptimal FBA simulation. If these zero flux reactions are required in a specific environment then changes in solution constraints are unlikely to have any effect. The *in vivo* resulting effect of changes in these reactions is likely to be specific to only the environments where the reaction is required and are less likely to have general effects on fitness than constrained reactions.

Comparing the overlap of reaction categories between each of the three nutrient limited environments in Figure 2.4 highlights that the majority of reaction categories are the same across the three environments considered. One hypothesis may be that the core metabolic requirements of generating new biomass are independent of environment, placing consistent constraints on the same reactions. Alternatively changes in the types of flux categories may be observed for a wider range of simulated conditions as shown by Papp *et. al* [39] where half of the genes not essential in nutrient rich conditions can be predicted to be essential in at least one other type of condition.

Gene importance using variation in reaction flux

Of the single gene-associated reactions, only 10%-14% exhibited non-trivial difference in flux between glucose, ammonium and sulphate limiting environments. In the optimal solution, as Figure 2.5 shows, the reaction fluxes for ammonium and sulphate limitation were identical indicating the difference in flux distributions for these two conditions must be encoded by reactions not included in the set of single gene associated reactions. The glucose limited flux distribution for single gene-associated reactions was skewed towards lower values compared with ammonium and sulphate limitation flux distributions indicating that single encoded reactions, on average, carry less flux in glucose limitation.

When estimating reaction flux using suboptimal solutions the glucose limited flux distribution is similar compared with the distribution in the optimal solution. The ammonium and sulphate limiting flux distributions do however vary between optimal and suboptimal flux balance solutions. This indicates that relaxing solution constraint has little effect on single gene-associated reactions in glucose limitation, but does change the overall distribution of flux in sulphate and ammonium limitation. Therefore the relaxed constraints of the glucose limited suboptimal solution, compared with the optimal solution, affect the reactions outside of the single gene-associated reaction set considered here.

There were also fewer high flux reactions and more zero flux reactions in glucose limitation, supporting the idea that the distribution of flux in glucose limitation may be outside the set of single gene associated reactions and instead through reactions with multiple gene associations. This may indicate more enzymatic redundancy of reaction flux in glucose limitation compared with the other two environments. Further analysis of the gene count vs reaction flux in multiple different simulated environmental conditions may identify if this is the case and that *S. cerevisiae* has evolved to be more adapted for glucose limitation.

Comparing flux distributions between single gene associated reactions across nutrient limitation showed a moderate degree of correlation, and flux distributions in ammonium and sulphate limitation in optimal conditions are identical (Figure 2.6). The data compared in this analysis are however only the non-trivial flux differences between the three environments which excluded >85% of the reactions. As might be expected when all the single gene reaction are compared the flux distributions show a much higher degree of correlation (Spearman R > 0.74 between each environment $p < 0.025$, data not shown).

Breaking the reaction fluxes into qualitative categories, shown in Figure 2.7, allows a coarse grain comparison of the changes in reaction flux between the three nutrient limited conditions. Comparison of these flux categories shows differences in reaction use between environments, which indicates activity specific to a subset of nutrient conditions, particularly in the suboptimal FBA solution. Reactions active or inactive in a single environment may be expected to result in selection pressures on the encoding gene related to the nutrient conditions.

Gene cost through phenotypic effects on nutrient uptake

Given that only a small fraction of all single gene-associated reactions have variable flux, the number for which a sensitivity can be estimated is limited. Of the set of reactions examined few had large sensitivity values with the median reaction sensitivity approximately 10^{-8} (Table 2.8). This indicates that for the variable reactions for which sensitivity can be estimated, the majority of reactions produced only minor effects on nutrient uptake flux and therefore the phenotypic effect when reaction flux is reduced is small. Furthermore when estimating the slope between changes in reaction flux and nutrient uptake flux, the very small responsive changes in reaction flux may affect the accuracy at which the slope can be estimated.

The most sensitive nutrient uptake to changes in reaction flux was glucose

(Figure 2.9). This may be expected since glucose entry and metabolism is responsible for the production of the high energy molecules used in metabolism. The ammonium and sulphate uptakes might be expected to only be sensitive to changes in reactions that produce metabolites requiring nitrogen or sulphur, however further analysis would be required to confirm from this hypothesis.

Estimating gene importance using a genome scale model

When deriving gene cost estimates the number of analysed reactions is limited when considering only those with a 1:1 association with an annotated gene. The alternative, considering reactions with multiple gene associations, would produce a larger sample but the corresponding link between fitness effects of multiple genes encoding the same reaction would be difficult to predict.

The observation that most reactions have little effect on nutrient uptake flux could be attributed to small control coefficients, where changes in reaction flux have only tiny effects on metabolic activity. However there were a large number of highly constrained reactions, where any reduction in absolute reaction flux prevented the estimation of an FBA model solution. These reactions would therefore be assumed to have highly deleterious effects on growth, and this undermines the suggestion of a robust metabolic network. Furthermore, this work shows that the estimation of reaction sensitivity may instead be highly dependent on the solution space when optimising the model. This is apparent in the comparison of optimal and suboptimal reaction sensitivity for the same reactions, where there was little correlation (Figure 2.9).

No definitive fitness measure for gene importance or cost was found, but instead three possible values that may used to investigate the functional constraints on gene function and activity. This indicates that using systems biology to establish a link between the metabolism of an organism and its genome evolution, beyond simple gene knockout, will require further investigation into how gene importance or cost can be modelled. This may require further development of genome scale models and the methods used to simulate them, which will be necessary to understand and predict the metabolic factors involved in genome evolution using *in silico* methods.

3

Amino acid biosynthetic cost in gene expression and evolution

Summary

This chapter examines amino acid cost as a selective force in *Saccharomyces* gene expression and protein evolution. The motivation for this analysis is to determine if trends for the minimisation of biosynthetic cost are visible in the use and substitution of amino acids in protein sequences.

The first half of the chapter compares amino acid cost with gene expression at transcript and protein levels to show that cost minimisation is a much weaker selective force than optimisation for translation. Free amino acids in the metabolite pool are however maintained at levels relative to biosynthetic cost. These findings show cost minimisation is a weak force in shaping the gene expression in *S. cerevisiae*.

The second half of this chapter examines cost related changes in amino acid usage between related *Saccharomyces* species. Cost minimisation is examined in terms of the variation in amino acid usage between species, and relative substitution rates in orthologous genes. The results of this analysis show a strong trend for cost minimisation in protein sequence evolution.

3.1 Introduction

3.1.1 Amino acid biosynthetic cost as a protein selection pressure

The encoded amino acid composition of a genome varies across species, and is related to GC content and environment [47]. Understanding the selective forces acting on amino acid usage is an open question in genome biology [46]. One hypothesis to explain trends in amino acid usage is the cost selection hypothesis [34].

The premise of the cost selection hypothesis in protein evolution is that when two amino acids can perform a similar function at the same site in a protein sequence, selection favours a biosynthetically cheaper amino acid, as the difference in biosynthetic cost between the two amino acids can be diverted to other cellular processes. The strength of cost as a selective pressure is still unclear and is reviewed in the following section.

3.1.2 Biosynthetic cost and amino acid usage in the genome

In a few cases it has been shown that the genome composition of an organism may reflect trends for preferential amino acid usage. For example, cyanobacterium *Calothrix sp. 7601* encodes sulphur eradicated versions of proteins used in sulphur limiting conditions [48]. In *Escherichia coli* and *Saccharomyces cerevisiae*, proteins in pathways responsible for integrating carbon and sulphur from the environment are depleted for that nutrient [49]. The advantage of depleting a protein for a specific nutrient is that the protein may be more easily expressed in conditions limiting for the nutrient. These two studies indicate, that the use of amino acids with scarce atoms may be biosynthetically more costly when expressed in limiting conditions for that atom.

The trend for specific amino acid use was studied across 141 sequences comparing orthologous genes between species with differing sulphur content [50]. The sulphur content of each ortholog pair was observed to be correlated with the encoded proteome sulphur content of the host organism. This result suggests that changes in sulphur content are related to environment rather than function. In particular thermophiles which were observed to have a lower sulphur content than non-thermophiles.

Analysis of large numbers of complete genomes was also used to determine whether the use of biosynthetically more expensive amino acids is minimised [36]. In this study molecular weight was used as the measure of biosynthetic cost and across bacterial, archaeal and eukaryotic genomes the use of heavier amino acids (i.e. more costly) is minimised. Furthermore cost minimisation is expected to be a greater factor for free living organisms than for intracellular organisms, a trend corroborated by comparing free living *E. coli* with intracellular *Buchnera* species [51].

Comparison of paralogs between extremophile and non-extremophile related species, controlling for protein structure and function, highlights amino acid preferences dependent on the organism's lifestyle but not understood in terms of existing measures of cost [52]. One possible reason for this observation in hyperthermophiles may be different energetic constraints on metabolic reactions given the temperature [53]. However the use of specific amino acids for maintaining protein stability in high temperatures cannot be discounted either.

Sequencing species from related but non-identical environments can provide insight into how organisms adapt to change, in particular through reconstruction of the species' metabolic pathways. Metagenomic analysis of several oceanic environments identified variation related to the use of amino acid biosynthetic pathways. The variance was, however, unrelated to biosynthetic cost but instead the availability of the cofactors used in the synthesis of each amino acid [54]. This result hints at deeper complexity in understanding cost selection, where the metabolic use of cofactors are not readily apparent from the metabolic map of an organism but nonetheless still have a role in the evolution of protein sequence.

3.1.3 Biosynthetic cost and gene expression

The amino acids encoded in the genome are expressed in the proteome via the transcriptome. Cost trends should therefore be related to gene expression at the transcript and protein level. A large proportion of cellular energy is spent in protein synthesis [32]. The cost minimisation theory predicts that minimising the use of expensive amino acids provides a fitness advantage, and highly expressed genes should be under selective pressure to minimise this cost, where this pressure decreases with expression level.

The selection pressure for cost minimisation can be estimated by comparing the cost of amino acid synthesis encoded in highly expressed genes. Akashi and

Gojobori [34] compared protein per-residue biosynthetic cost with predicted expression level based on codon bias [55] and showed that protein cost does decrease with increasing expression level.

Akashi and Gojobori [34] performed their analysis in only two mesophiles, *E. coli* and *B. subtilis*. The importance of cost minimisation in gene expression may be related to organism lifestyle. Species with large differences in metabolism may display varying biosynthetic costs for each amino acid and therefore different trends in cost minimisation. Heizer *et al.* [35] examined cost related trends in four additional organisms including chemoheterotrophic, photoautotrophic, and thermophilic lifestyles. Across all species studied the trend between predicted gene expression and biosynthetic cost indicated a pressure to minimise expensive amino acids at high expression levels across a variety of environments.

Both of these studies examined the cost minimisation in gene expression using codon bias as a proxy for expression level. Validation of the extent of cost minimisation requires analysis of observed *in vivo* gene expression levels. Raiford *et al.* [56] examined the correlation of expensive amino acids with *S. cerevisiae* codon usage bias, transcript levels and protein levels. This analysis compared the aerobic and anaerobic costs for each amino acid with the usage in predicted or observed expression. The results of the study showed a trend for cost minimisation, but the trend varies according to amino acid class, and explains only a limited degree of variation in each of the codon usage, transcript and protein variables.

Further studies [57] compared the carbon content of up-regulated genes versus the rest of the proteome, and found no indication of cost minimisation in the up-regulated set. A yeast strain evolved in carbon limitation was found to use more carbon atoms per residue than the unadapted strain, in direct contrast to cost minimisation theory.

The current literature reviewed in this section describes a mixed view of the strength of cost minimisation in protein evolution. Analysis of genome data, comparing predicted expression level with predicted per residue biosynthetic cost indicates that cost minimisation may be a strong selective force. Examining actual expression data, cost minimisation appears to be a much weaker force. The two variables may not be directly comparable since expression data is subjective of the environment in which it was measured, while codon bias is result of long term evolutionary optimisation of the genome. Furthermore examination of the codon bias across the entire genome considers all possible expressed genes, while specific copies of a gene may be maintained for only a small subset of all possible environmental conditions [58].

3.1.4 Biosynthetic cost and protein sequence evolution

If there is a selective force for the optimisation of protein sequence for cheaper residues, then a pattern related to biosynthetic cost should also be detectable in the use and substitution of amino acids across species. The cost-selection hypothesis predicts that the use and mutation rates of expensive amino acids is much less than that of cheaper amino acids which are under less cost selection pressure. Whether such a pattern exists is unclear and remains to be investigated. Existing literature so far has mainly examined cost minimisation in either genome composition or gene expression.

Craig and Weber [33] found that replacement of amino acids in a small set of *E. coli* genes were cheaper than the overall mean or median amino acid cost. This result was however contradicted by a different set of *E. coli* proteins where the use of amino acids was the same as the overall amino acid cost. This result presents a mixed view of the importance of cost minimisation.

Metabolic cost was compared directly in a multivariate regression of possible *E. coli* and *B. subtilis* protein evolution determinants. Of all the protein characteristics included: expression level, functional category, essentiality, and biosynthetic cost - cost minimisation is a weak predictor of protein evolution [59].

Overall little work has been performed in analysing the role of metabolic cost in protein sequence evolution. Specific analysis of trends between amino acid substitution rates and biosynthetic cost will revel the strength of cost minimisation in shaping sequence evolution.

3.1.5 Summary of results

The first half of this chapter examines the cellular levels of transcripts, proteins and free amino acids *S. cerevisiae* to determine the extent of cost minimisation in shaping gene expression compared with other factors associated with gene expression. The second half examines cost minimisation in shaping the usage and relative substitution rates of amino acids in related *Saccharomyces* genomes.

The results in this chapter suggest cost minimisation is a weak selective force in shaping gene expression or amino acid usage in protein sequences. When cost is compared with the pattern of evolution in yeast proteins, biosynthetic cost related trends are apparent and indicate cost minimisation is indeed an important evolutionary force in protein sequence evolution.

3.2 Materials and Methods

3.2.1 Gene expression data

The transcript and metabolomic data used in this chapter are taken from Castrillo *et al.* [60]. The data was measured from *S. cerevisiae* continuously cultured in a chemostat. Each measurement was taken at one of three dilution rates 0.1 hr^{-1}, 0.2 hr^{-1} in 0.3 hr^{-1} in four different conditions, glucose, ammonium, sulphate and phosphate limitation. Transcript levels were estimated from four replicate microarray measurements of total RNA. Following measurement the RNA data was processed using RMA quantile normalisation [61]. The metabolite data was measured using Gas chromatography followed by time of flight mass spectrometry (GC/TOF-MS). The metabolite data was normalised using median absolute deviation (MAD). Missing metabolite data were inferred from replicates in the same conditions.

The protein data used were produced by Ghaemmaghami *et al.* [62]. These data were derived from tandem affinity purification of TAP-tagged *S.cerevisiae* ORFs. Each tagged ORF was measured using antibody quantification of the encoded tag. Absolute levels of protein per cell were estimated by comparing the *S.cerevisiae* gene measurements with a purified *E. coli* INFA-TAP construct scale.

S. cerevisiae codon adaptation index (CAI) data was estimated by Coghlan and Wolfe [63]. Genome encoded tRNA count data was produced by Akashi [64]. The $A_{glucose}$ and $R_{glucose}$ costs used were those described in the previous chapter. The other amino acid costs used are reported in the literature by Craig & Weber [33], Akashi & Gojobori [34], Wagner [32] and Seligmann [36]. Atomic content for each amino acid was taken according to universally available descriptions.

3.2.2 Gene expression multivariate regression

Each regression variable was transformed using the natural logarithm. The exception to this was the transcript data which were logged during the original processing. The sulphur content of individual amino acids contained observations with a value of zero, and therefore a small arbitrary value (0.0001) was added so the logarithm could be taken. All regression variables were scaled to have the same mean and variance. The aim of this data processing was to minimise heteroscedasticity or over-variation in model fitting. The metabolite data was mean averaged for each combination of experimental parameters to prevent pseudo-replication resulting from the inference of missing values.

The multivariate regression was performed using the R language for statistical computing [65]. For each type of data, transcript, protein or metabolite, the measured level in cell was the response variable in the regression equation. The explanatory variables in the regression equation were mean per residue carbon, nitrogen, and sulphur content, amino acid cost, and CAI. Environment and dilution rate were also included for the transcript and metabolite data.

For each of the transcript, protein, and metabolite data the complete model equation was fitted including all factors and multiple interactions between factors. The R function for linear modelling was (*lm*) was used as the regression function. Stepwise interaction and variable removable was then performed to test the removal of interactions and variables from the regression, to produce a reduced regression model. The next removed variable was identified from the greatest AIC difference. The reduced model was then compared with the previous complete model, and if the reduced model was considered more parsimonious based on the AIC difference, then the reduced model was then used as the new model. This process was repeated until no more interactions or variables could be removed without loss in explanatory power to produce the minimal regression model.

The importance of each explanatory variable in the regression was assessed by repeating the above process but with one of the explanatory variables removed from the initial model fitting. Starting with the minimum reduced regression model, one variable removed was then compared to the minimum reduced regression model with all variables included. The AIC difference between the two regression fits was used as a measure of importance for the removed variable. This process was repeated for each variable and each of the amino acid cost types as the biosynthetic cost variable.

3.2.3 Amino acid usage in *Saccharomyces* proteomes

The verified ORFs from seven *Saccharomyces* related species were downloaded from the *Saccharomyces* genome database. The species were *S. paradoxus*, *S. bayanus*, *S. castellii*, *S. cerevisiae*, *S. kluyveri*, *S. kudriavzevii*, and *S. mikatae*. The genomes of *S. paradoxus* and *S. bayanus* were sequenced by Kellis *et al.*

[66]. The *S. castellii*, *S. kluyveri*, *S. kudriavzevii*, and *S. mikatae* genomes were sequenced by Cliften *et al.* [67].

The percentage amino acid content of each species was calculated by summing the frequency of each amino acid in the total predicted proteome, then dividing by the total number of amino acids. Stop codons and ambiguous amino acid definitions were excluded, and only the twenty standard amino acids were considered. Amino acids from ORFs containing an internal stop codon were also excluded

The variation in amino acid usage between species was calculated using median absolute deviation. MAD is a robust measure of variance for non-normally distributed, or small variables. MAD is calculated as the median of the absolute deviations from the median average observation.

3.2.4 Amino acid cost in *Saccharomyces* protein evolution

The codon aligned gene sequences of *S. cerevisiae*, *S. mikatae*, *S. bayanus*, and *S. paradoxus* species were generated by Wall *et al.* [68]. Sequences containing less than four genes or where the alignment of the *S. cerevisiae* sequence did not match the SGD reference *S. cerevisiae* sequence were ignored.

The degree of conservation for each alignment column in each alignment sequence was classified based on the amino acids observed in the column using the BioRuby library [69]. Columns were classified as ‘identical’ if they contained the same amino acid for each of the four species. ‘Strongly conserved’ columns contained only amino acids with a Gonnet Pam250 matrix score greater than 0.5. ‘Weakly conserved’ columns contained amino acids with a score greater than 0.

The relative substitution rate of each alignment and each alignment column was estimated using the codeml tool, part of the PAML package [70]. The ancestral amino acid sequence was also determined for each alignment. The WAG amino acid rate matrix [71] was used for each estimation analysis.

The mean relative substitution rate was estimated for each ancestral amino acid across all the estimated ancestral protein sequences. Sites that contained a gap in the any of the descendant sequences were ignored. The frequency of gaps was calculated for each ancestral amino acid, where the number of instances of gaps in each descendant site was calculated given the ancestral amino acid for the position.

3.3 Results

3.3.1 Amino acid cost in gene expression

The role of amino acid cost in the evolution of gene expression was estimated using multivariate regression on transcriptomic, proteomic and metabolomic data. The aim of this analysis was to estimate the importance of amino acid biosynthetic cost in protein production in yeast cells.

The transcriptomic and metabolic data in this analysis were taken from a large scale systems biology analysis in *S. cerevisiae* [60]. This provided proteomic, transcriptomic and metabolomic data from four conditions: carbon, nitrogen, sulphur and phosphorus limitation. The proteomic data from this experiment was not used as relative change in protein level was estimated. Instead the proteomic data used in this analysis was taken the study by Ghaemmaghami *et al.* [62] which measured absolute levels of protein abundance.

Characteristics of gene expression

Multivariate regression was used to estimate the importance of amino acid cost in gene expression along with other factors associated with translation. The response variable in the regression model was either transcript, protein or free amino acid level. The explanatory variables were: mean per residue carbon, nitrogen and sulphur content of proteins, CAI of the encoding transcript, average tRNA number per codon in the transcript, and average amino acid cost. The aim of this analysis was to determine which of these factors are most relevant to explaining variation in gene expression.

Average carbon, nitrogen and sulphur content represents the total content for each atom in the protein divided by the length of the protein. In the free amino acid data the atomic content represents the content for each individual amino acid. These factors will identify any variation related to protein optimisation for nutrient limitation.

The CAI is the optimisation of the transcript for translation, where the use of certain codons is preferential for highly expressed transcripts [63, 72]. CAI is estimated from the usage of codons in each transcripts using highly expressed genes (i.e. ribosomal proteins) as a reference.

The average tRNA count per codon is the sum of the total number of genome encoded tRNAs per codon divided by the open reading frame length excluding stop codons. Transcripts with a greater than average tRNA count may indicate transcript optimisation for translation [64]. For the free amino acid data, the total number of tRNAs for each encoded amino acid was used.

The estimate used as the cost variable in the regression iterated over the range of costs described in the Chapter 2. This led to the regression process being repeated and the model fit being estimated for each amino acid cost type. The cost estimates used were as follows: Akashi & Gojobori energy [34], Craig & Weber energetic cost [33], Craig & Weber biosynthetic steps [33], Wagner fermentative and respiratory energetic cost [32], amino acid molecular weight (Da) as used by Seligmann [36], plus the $A_{glucose}$ and $R_{glucose}$ costs derived in the previous chapter. The aim of cycling the costs was to determine which estimate had the most explanatory power for explaining variation in gene expression data related to biosynthetic cost optimisation. When considering the transcript data, the atomic content and biosynthetic cost variables were that of the encoded protein. When considering the protein data, the CAI and tRNA count variables were that of the encoding transcript.

Multivariate regression

The importance of each variable in gene expression was examined in a multivariate regression model by removing a variable, refitting the model then comparing the reduced model with the model containing all the variables in the minimal model. Akaiki's information criterion (AIC) [73] was used to quantify the difference in explanatory power between the two models. The AIC is a log likelihood score for statistical models, which penalises models with more parameters. This allows models with differing numbers of parameters to be compared. For each regression, a complete model was fitted containing all interaction terms and stepwise automated variable removal was then used to reduce interactions and variables based on the AIC score. A negative AIC score indicates the reduced model is more parsimonious. A positive score indicates the larger model provides a better fit to the data. Table 3.1 shows the R^2 values for each type of expression data, for each cost type used in the multivariate regression. Figure 3.1 illustrates the effect of removal for each of the variables considered in the regression.

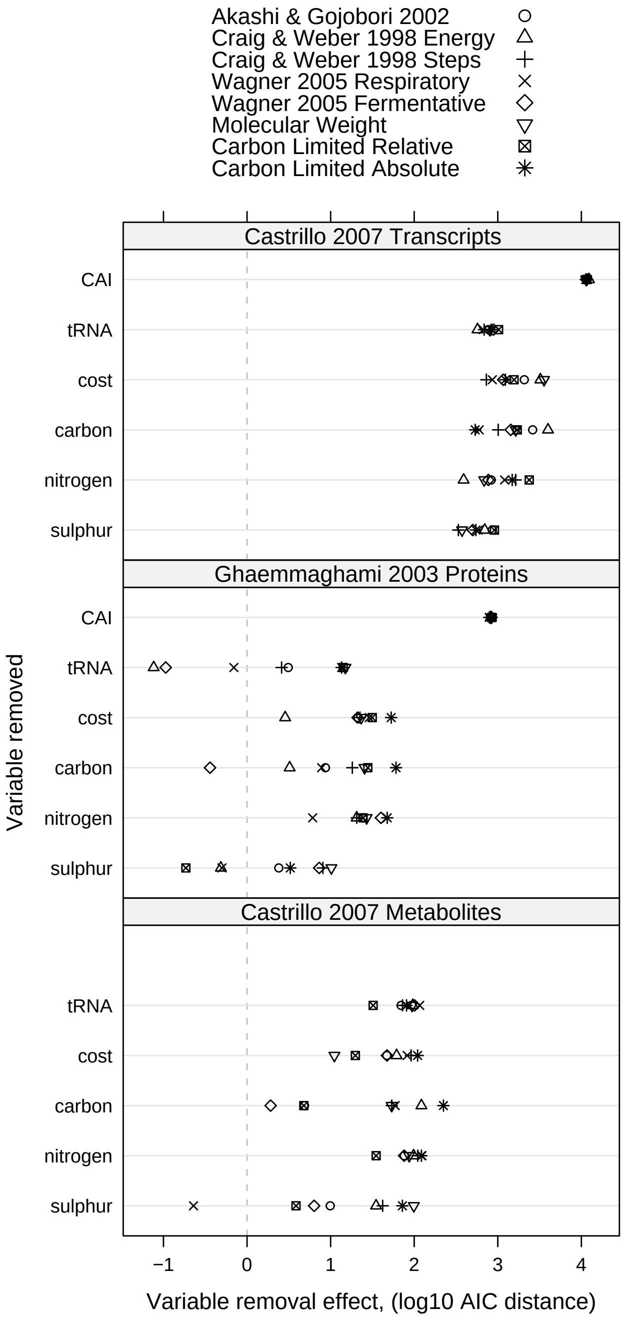


Figure 3.1: Variable importance in transcript, proteomic and metabolomic multivariate regression. The y-axis indicates the variable removed from the multivariate regression. The x-axis indicates the $\log_{10}$ AIC difference between the regression with the variable removed, and regression including all variables. A positive AIC difference indicates the model is a worse fit, a negative difference indicates the model is a more parsimonious fit. The legend indicates which amino acid cost type was used to calculate the per-residue biosynthetic cost. The transcript and metabolomic data were measured by Castrillo *et al.* [60]. The proteomic data were measured by Ghaemmaghami *et al.* [62].

Cost type	Transcripts	Proteins	Amino Acids
S. cerevisiae $A_{glucose}$	0.389	0.406	0.782
S. cerevisiae $R_{glucose}$	0.383	0.408	0.875
Akashi & Gojobori (2002)	0.398	0.405	0.805
Craig & Weber (1998) Energy	0.416	0.400	0.835
Craig & Weber (1998) Steps	0.375	0.404	0.866
Wagner (2005) Respiratory	0.382	0.405	0.822
Wagner (2005) Fermentative	0.377	0.406	0.851
Molecular Weight	0.422	0.405	0.767

Table 3.1: Variation explained by multivariate regression model fitting of transcriptomic, proteomic, and metabolomic variables. Each value in the table is the R^2 for the multivariate regression model using the indicated cost type, the carbon, nitrogen, sulphur content of the amino acid, average tRNA count, and CAI for the transcript and protein data. The transcript and metabolomic data were measured by Castrillo *et al.* [60]. The proteomic data was measured by Ghaemmaghami *et al.* [62].

Transcriptome and proteome analysis The regression analysis of the transcript data from Castrillo *et al.* [60] and the proteome data from Ghaemmaghami *et. al* [62] explained similar levels in variation. CAI was the most important variable in the regression of both these variables. Figure 3.1 shows that CAI was near half an order of magnitude greater in explaining transcript levels than other included factors. In the protein data CAI was a full order of magnitude greater in explaining protein levels. Together these results indicate that the optimisation of a transcript for translation through codon usage bias is a strong predictor of both transcript and protein levels - consistent with previous findings [72, 74].

Comparing other factors in the regression, amino acid cost, carbon and nitrogen content appear to show similar effects in explaining variation in both the transcript and protein data. The variables for tRNA count and sulphur content appear slightly less important. Overall the regression analysis indicates that a large proportion of the variation in the examined *S. cerevisiae* transcript and protein data is explained by codon usage bias. In contrast the variables of cost, carbon, nitrogen, sulphur content, or corresponding genome encoded tRNAs, appear less important. This is particularly apparent in the protein data analysis.

As the first two columns of Table 3.1 show, each of the different amino acid cost types had only marginal effects on the R^2 of the transcript and protein multivariate regression. The difference in R^2 due to the cost type used in the transcript analysis ranged from 37.5% for the regression using Craig & Weber's biosynthetic steps cost measure [33] to 42.2% for the regression using molecular weight [36]. In the analysis of the protein data, the difference in variation explained dependent on the cost measure was smaller ($\sim 0.8\%$), indicating all cost measures have approximately the same effect in explaining variation in protein levels.

Metabolome analysis In the regression of the metabolome data from Castrillo *et al.* [60] CAI is not applicable and therefore the variation is explained by tRNA count, biosynthetic cost, and carbon, nitrogen and sulphur content. The bottom plot of Figure 3.1 indicates that in the majority the regression models, each of the considered variables contributes equally to explaining variation in free amino acid levels. The exception to this is sulphur content which in half of the regression models explain less than the other factors. This contrasts with previous results for transcript and protein levels where a single variable explained the majority of variation.

Compared with the analysis of transcript and protein data, the multivariate regression of free amino acids explains a much larger degree of variation. The variation explained also differed by 10.8% between the best and worst regression model. The regression model that explained the least variance used molecular weight for amino acid cost, while the best fit used $R_{glucose}$ (explaining 87.5% of the variance). This result indicates that a large degree of the variation in amino acid levels can be explained based on their biosynthetic cost, atomic content, and the corresponding tRNAs encoded in the genome. Therefore cost minimisation may be a signficant factor in the maintanence and synthesis of amino acids in the metabolome.

3.3.2 Amino acid cost in protein evolution

Amino acid variance between *Saccharomyces* species

In the previous section biosynthetic cost was considered as a factor in gene expression, but cost minimisation may also be a selective pressure in sequence evolution, affecting the fixation and use of amino acids based on their cost. To determine the presence of such a selection pressure, the percentage amino acid usage across seven *Saccharomyces* species was calculated. Table 3.2 shows the total number of amino acids included for each species. Figure 3.2 compares the median percentage usage in the predicted proteome with three measures of amino acid cost:

Saccharomyces species	Amino acids
S. bayanus	2921220
S. castellii	2328652
S. cerevisiae	2916055
S. kluyveri	1248804
S. kudriavzevii	1546252
S. mikatae	1172495
S. paradoxus	2933095

Table 3.2: Total number of amino acids in each *Saccharomyces* species.

molecular weight, $A_{glucose}$ and $R_{glucose}$. These three estimates of amino acid cost were selected to provide a diverse set of cost estimates, where these three costs have a low Spearman's rank correlation coefficients with one another (see Figure 2.2 in Chapter 2). Molecular weight is also universal to across species while the FBA estimated costs are specific to *S. cerevisiae*.

Spearman's rank correlation was used to compared amino acid cost with median amino acid use across the seven *Saccharomyces* species. Spearman's rank correlation was used to allow for non-normal distributions in each pairwise comparison. The results suggests a weak and non-significant negative correlation between median percentage usage in the *Saccharomyces* genomes and either of the molecular weight or $A_{glucose}$ biosynthetic costs. The estimated correlation coefficients for both these measures indicates no strong selection pressure to maintain amino acids in the genome proportional to their cost of synthesis. This is also indicated by the large confidence intervals for the trend line in both plots.

The $R_{glucose}$ cost in contrast shows a positive correlation with median amino acid usage across the seven *Saccharomyces* species. This indicates genomic amino acid usage and the relative amino acid cost of synthesis in glucose are proportional, but in the opposite direction predicted by the cost minimisation hypothesis (see discussion).

The smallest and largest percentage amino acid usage across species illustrated in Figure 3.2 suggest a trend for decreasing variance in amino acid usage relative to increasing molecular weight or $A_{glucose}$ cost. This trend is further examined in Figure 3.3 which compares the median absolute deviation (MAD) in percentage amino acid usage with biosynthetic cost. The Spearman's rank correlation suggests a strong negative correlation between deviation in amino acid usage and either molecular weight or $A_{glucose}$. This is in contrast to the weak correlation

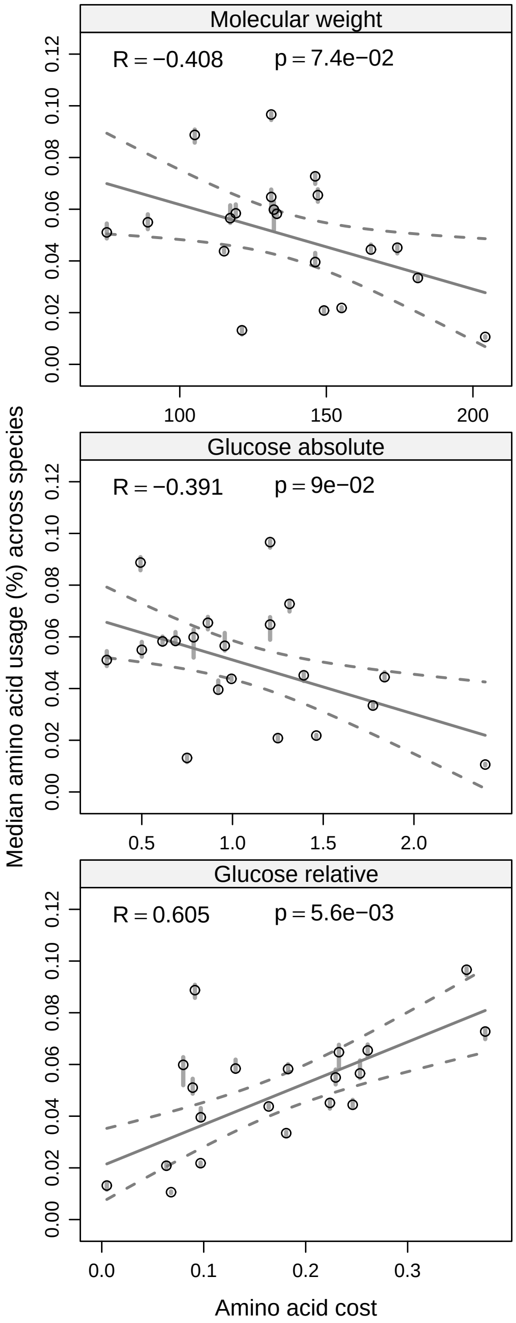


Figure 3.2: Comparison of percent amino acid usage with amino acid cost. Each point is the median percentage amino acid usage across 7 *Saccharomyces* species: *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii*, and *S. kluyveri*. Spearman's rank correlation between amino acid cost and median usage is indicated in each plot. A grey line at each point indicates the maximum and minimum usage across the seven species. Robust linear regression and 95% confidence intervals are used to indicate trend. In this and the following plots in this section each point represents one of the twenty amino acids.

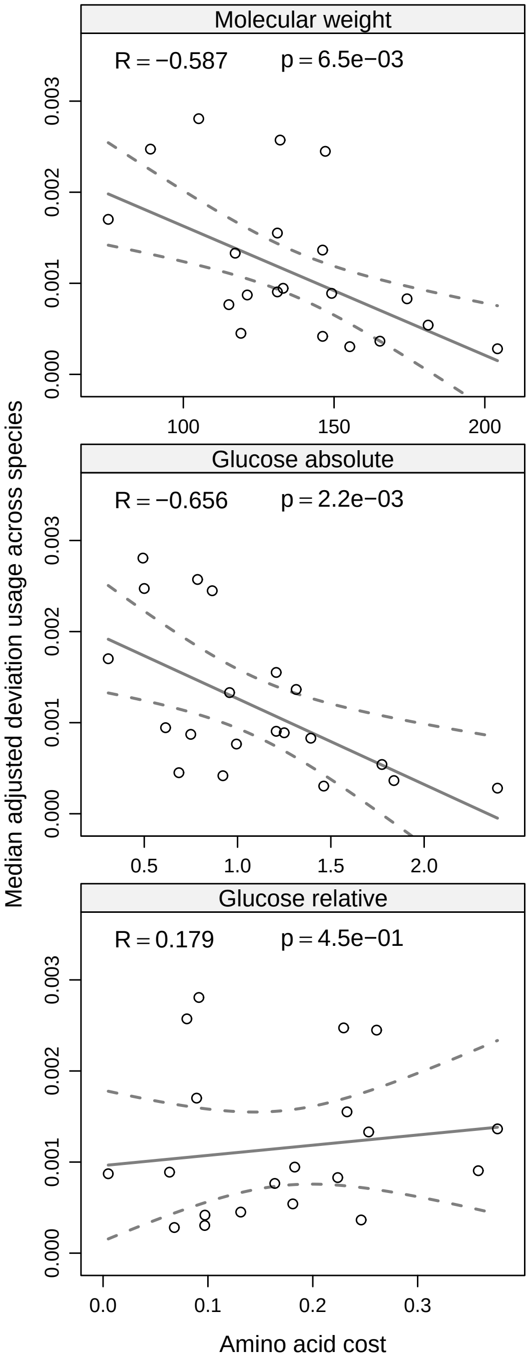


Figure 3.3: Comparison of median absolute deviation in percent amino acid usage with biosynthetic cost. The deviation is estimated from seven *Saccharomyces* species: *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii*, and *S. kluyveri*. Spearman's rank correlation is indicated in each plot. Robust linear regression and 95% confidence intervals are used to indicate trend.

between actual amino acid usage and these two measures of cost. This result suggest that the changes in amino acid usage between species may be related to the biosynthetic cost of the amino acid, and therefore indicative of a selection pressure in increased constraint on expensive amino acids but less so on inexpensive amino acids. This hypothesis is examined in the following section.

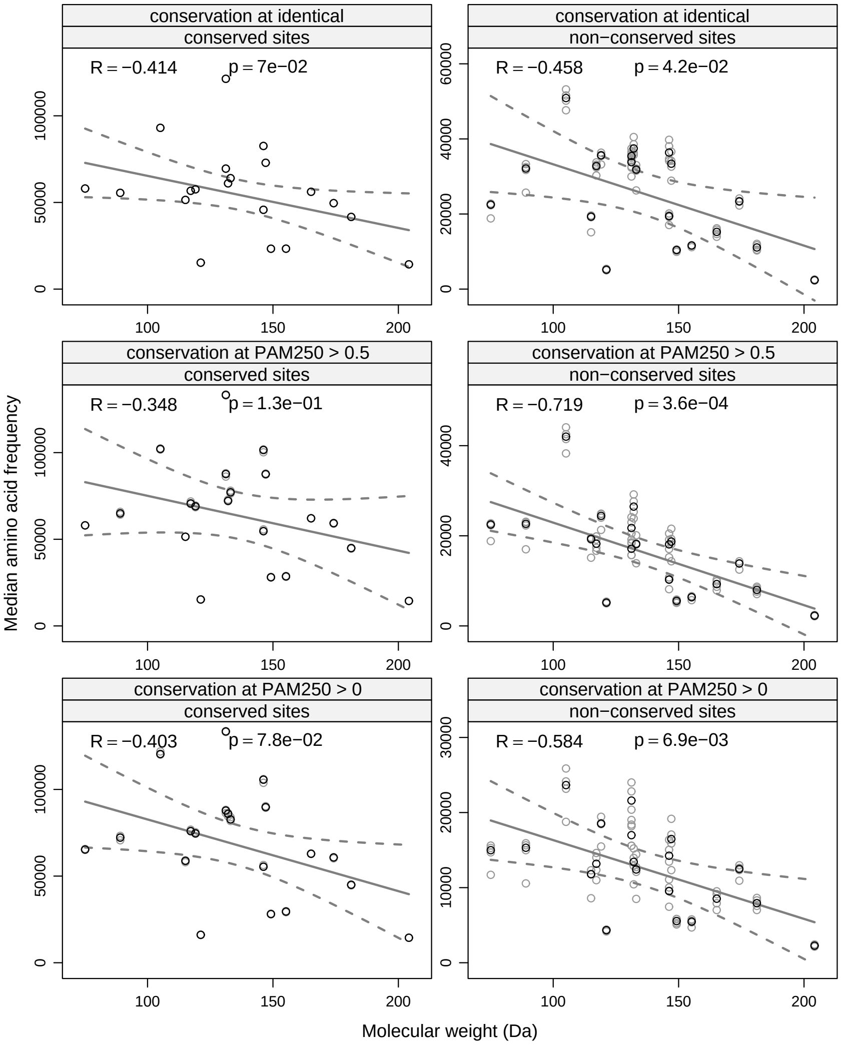
In contrast to the these two costs estimates, the $R_{glucose}$ cost shows no correlation with deviation in amino acid usage. This indicates that the variation in amino acid usage is not proportional to proteome usage scaled by cost of glucose synthesis.

Estimating cost minimisation in orthologs

The trend for cost minimisation was further investigated across four *Saccharomyces* species using 3334 of the codon-based protein sequence alignments generated by Wall *et. al* [68]. Each column in each alignment was classified as either conserved or non-conserved based on the estimated functional constraint on the four amino acids observed at that site. Three measures of conservation were considered: "identical" where the same amino acid appeared in at each position in the column, "strong conservation" where the observed amino acids had a high degree of functionally similarity based on PAM, or "weak conservation" where the observed amino acids had a lesser degree of functional similarity. For each definition of conservation, amino acids across the alignment were partitioned into one of two sets depending on whether it was observed in a conserved or non-conserved column. The expected outcome of this analysis was less conserved sites would evolve according to the cost minimisation.

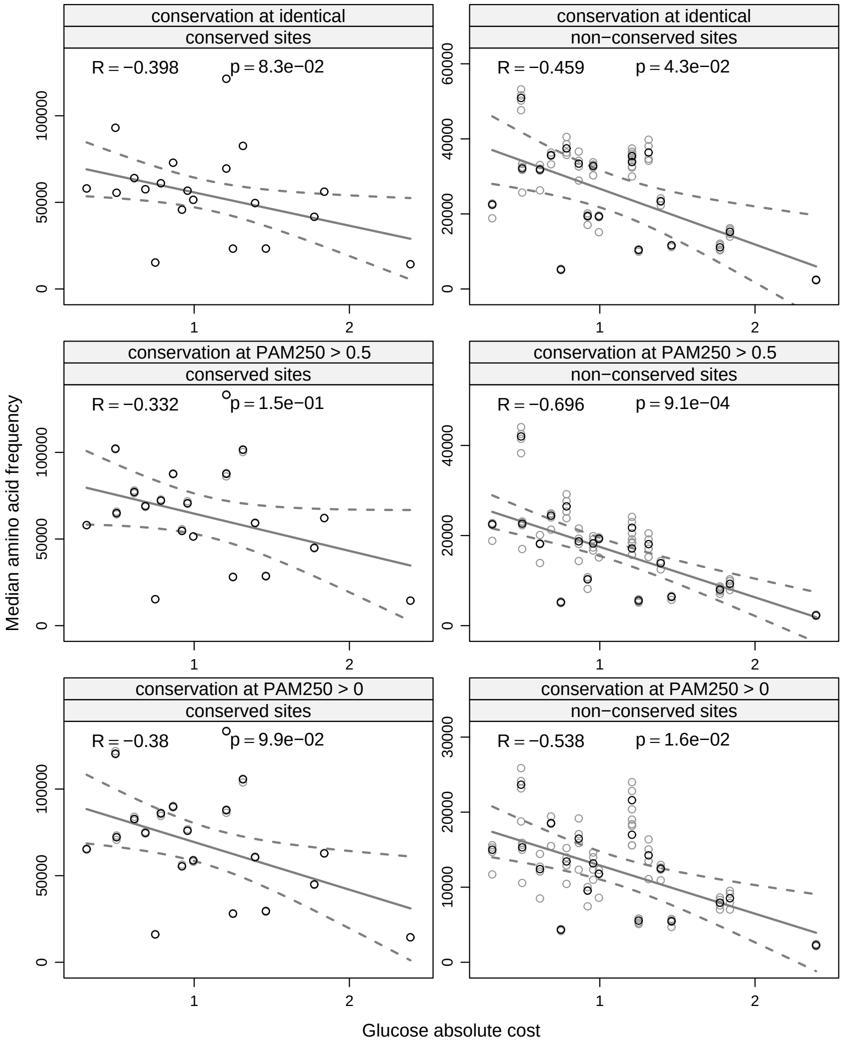
The frequency of each amino acid in each conserved or non-conserved set was compared with amino acid cost to determine if cost minimisation exhibits a trend as a function of the conservation of amino acids. Figure 3.4 illustrates this comparison for the same three amino acid cost estimates, with the median amino acid frequency across the four species. In each plot the median across species is indicate by the dark point, while the grey points represent observations in individual species.

Comparing molecular weight with amino acid conservation, there are small negative correlation coefficients for the conserved sites at all levels of conservation considered. Each Spearman Rank p-value is greater than $p = 0.05$, indicating the observed correlation is non-significant. In the non-conserved sites, in contrast,



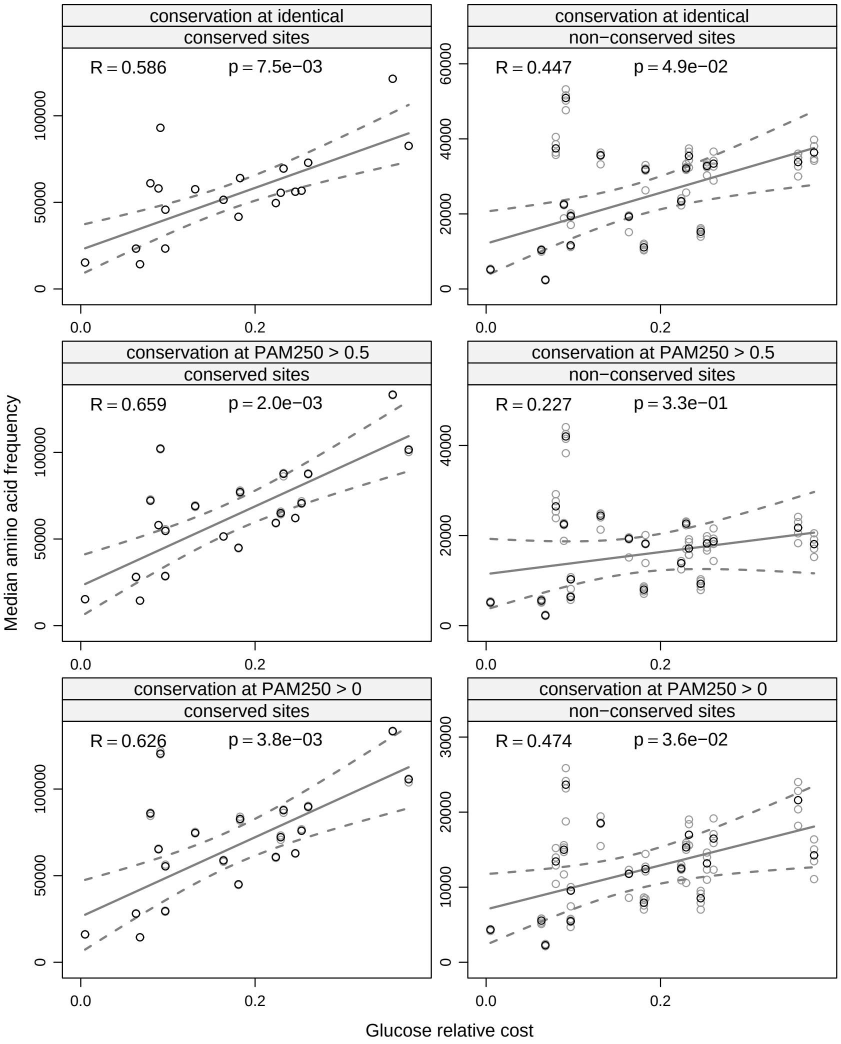
(a)

Figure 3.4: Comparison of median amino acid frequency with amino acid cost for conserved and non-conserved alignment columns. Figure (a) compares Molecular weight, (b) $A_{glucose}$, and (c) $R_{glucose}$. Conservation was determined from multiple sequence alignments of four *Saccharomyces* species: *S. cerevisiae*, *S. mikatae*, *S. bayanus*, and *S. paradoxus*. Each observation in each species is shown in grey, and the median usage across species in black. Amino acid frequency is partitioned into conserved and non-conserved sites based on the observed amino acids in each alignment column. The three definitions of conservation are: "identical" (where all amino acids in the column are the same), "strong conservation" (where the column amino acids have a Gonnet Pam250 score greater than 0.5), and "weak conservation" (where the column amino acids have a Gonnet Pam250 score greater than 0). Spearman's rank correlation and *p*-value is indicated for each plot. Robust linear regression with 95% confidence intervals are used to indicate trend.


conservation at identical
conserved sites
R = −0.398
p = 8.3e−02
conservation at identical
non−conserved sites
R = −0.459
p = 4.3e−02
conservation at PAM250 > 0.5
conserved sites
R = −0.332
p = 1.5e−01
conservation at PAM250 > 0.5
non−conserved sites
R = −0.696
p = 9.1e−04
conservation at PAM250 > 0
conserved sites
R = −0.38
p = 9.9e−02
conservation at PAM250 > 0
non−conserved sites
R = −0.538
p = 1.6e−02
Median amino acid frequency
Glucose absolute cost
0
50000
100000
20000
40000
60000
10000
30000
1
2

(b)

Continued.


conservation at identical
conserved sites
R = 0.586
p = 7.5e−03
conservation at identical
non−conserved sites
R = 0.447
p = 4.9e−02
conservation at PAM250 > 0.5
conserved sites
R = 0.659
p = 2.0e−03
conservation at PAM250 > 0.5
non−conserved sites
R = 0.227
p = 3.3e−01
conservation at PAM250 > 0
conserved sites
R = 0.626
p = 3.8e−03
conservation at PAM250 > 0
non−conserved sites
R = 0.474
p = 3.6e−02
Median amino acid frequency
Glucose relative cost
0
50000
100000
0
20000
40000
60000
10000
30000
0.0
0.2

(c)

Continued.

there is a significant negative correlation between median amino acid frequency and molecular weight. The largest Spearman's rank coefficient is observed in non-conserved columns when conservation is considered at PAM250 > 0.5.

The $A_{glucose}$ biosynthetic cost comparison with amino acid usage shows similar trends to that of molecular weight. Amino acid cost and amino acid frequency in the conserved sites shows no significant correlation for any level of conservation considered. As with molecular weight, the non-conserved sites show larger, significant correlation coefficients. The largest correlation observed in the non-conserved sites is for conservation at the PAM250 > 0 level, the middle of the three levels of conservation considered.

For both molecular weight and $A_{glucose}$ costs the observations for individual genomes (pale grey points) are close to the median (dark grey point) for each amino acid. This suggest the cost minimisation trend is consistent in each examined genome as well as the median average across genomes.

Examining the $R_{glucose}$ cost in Figure 3.4 there are significant positive Spearman's rank coefficients for conserved sites at all levels of conservation considered. In the non-conserved sites, there are two significant positive correlations at $p < 0.05$ when non-identical, or those that are non-conserved at the weakest level of conservation. This is against the expected cost minimisation trend. The sites considered non-conserved at the strong level of conservation, show no significant correlation between amino acid frequency and $R_{glucose}$ cost.

Biosynthetic cost trends in protein sequence relative substitution rates

Per site relative substitution rates were estimated for all columns in 3334 protein sequence alignments from four species in the *Saccharomyces* genus. This resulted in 1.66 million estimated relative substitution rates at each individual alignment column. In addition to the column substitution rate, the predicted ancestral amino acid at each sites was also determined. The aim of this analysis is to determine if amino acid biosynthetic cost is a trend in the substitution rates or deletion frequency of amino acids from protein sequences. Any trend observed will further indicate cost minimisation is a selective force in protein sequence evolution.

Figure 3.3 compares the mean substitution rate of each of the twenty ancestral amino acids in the alignment columns which did not contain any gaps. Each mean relative substitution rate is compared with three measures of amino acid cost: molecular weight, $R_{glucose}$, and $A_{glucose}$. The aim of this comparison is to determine if greater substitution rates are associated with cheaper amino acids, and if more expensive amino acids tend to be fixed in protein sequence.

This figure highlights a strong trend for cost minimisation where a lower relative substitution rate is associated with costly amino acids, in terms of molecular weight or $A_{glucose}$ biosynthetic cost. The $R_{glucose}$ cost however shows no correlation with mean amino acid relative substitution rate. This suggests cost minimisation is a significant selection pressure in the examined *Saccharomyces* species protein alignments for two out of the three considered cost types.

The previous figure compared the mean relative substitution rate for all alignment columns which did not contain any gaps, i.e. no insertions or deletions at any site. The sites which do contain gaps in the protein sequence alignment may also be indicative of a cost related trend in protein sequence evolution. Figure 3.4 compares the sites which contain a gap in one of the four aligned protein sequences. The frequency of sites containing a deletion given the ancestral amino acid, is compared with the ancestral amino acid cost. The aim of this analysis is determine if expensive amino acids are more likely to be fixed in a genome, while cheaper amino acids are more likely to be deleted.

This figure indicates no significant correlation between the frequency of deletions related to either the molecular weight or R$_{glucose}$ cost. There is however a large and significant Spearman's rank correlation with the absolute glucose measure of amino acid cost. This results suggests if there is a cost related trend for the deletion of amino acids, with the $A_{glucose}$ best describing the amino acid cost-retention selective force.

One caveat to this analysis is that the codeml tool in some cases will not correctly predict the ancestral sequence where the alignment column contains multiple gaps. This will lead to instances where the ancestral site is incorrectly identified as an amino acid where instead the amino acids in the child sequences are the result of an insertion. Future work will repeat this analysis using tools that may more accurately predict the ancestral site in sites with gaps.

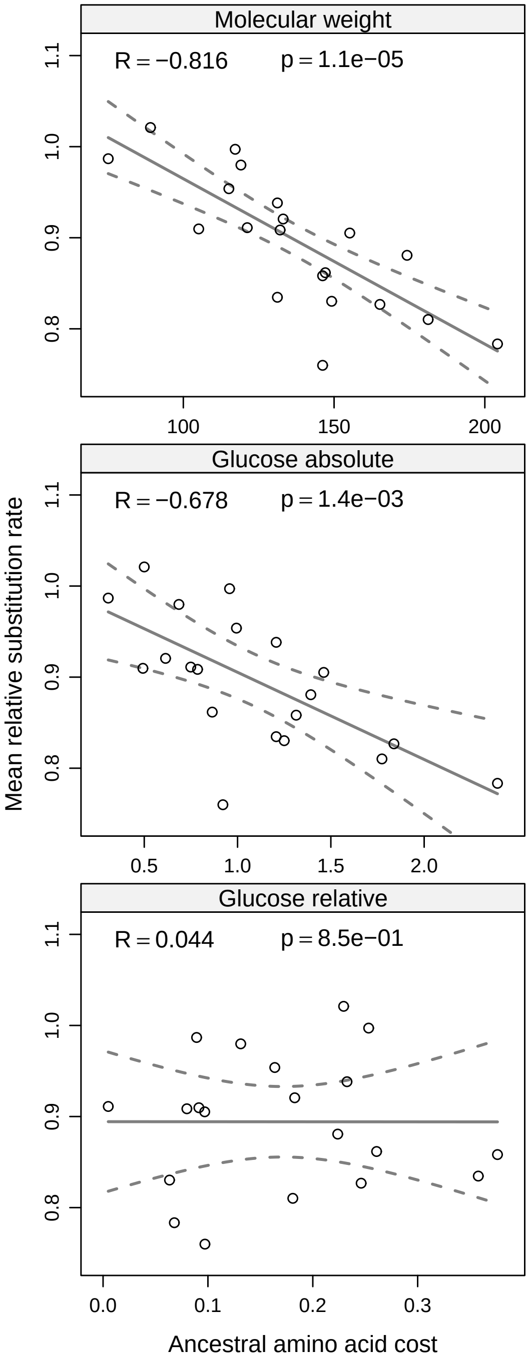


Figure 3.3: Comparison of amino acid mean relative substitution rate with three estimates of amino acid cost. The relative substitution rate is the mean of each alignment column. The amino acid cost is derived from the ancestral amino acid estimated at each position. Robust linear regression and 95% confidence intervals are used to indicated trend. Each plot indicates the Spearman's rank correlation.

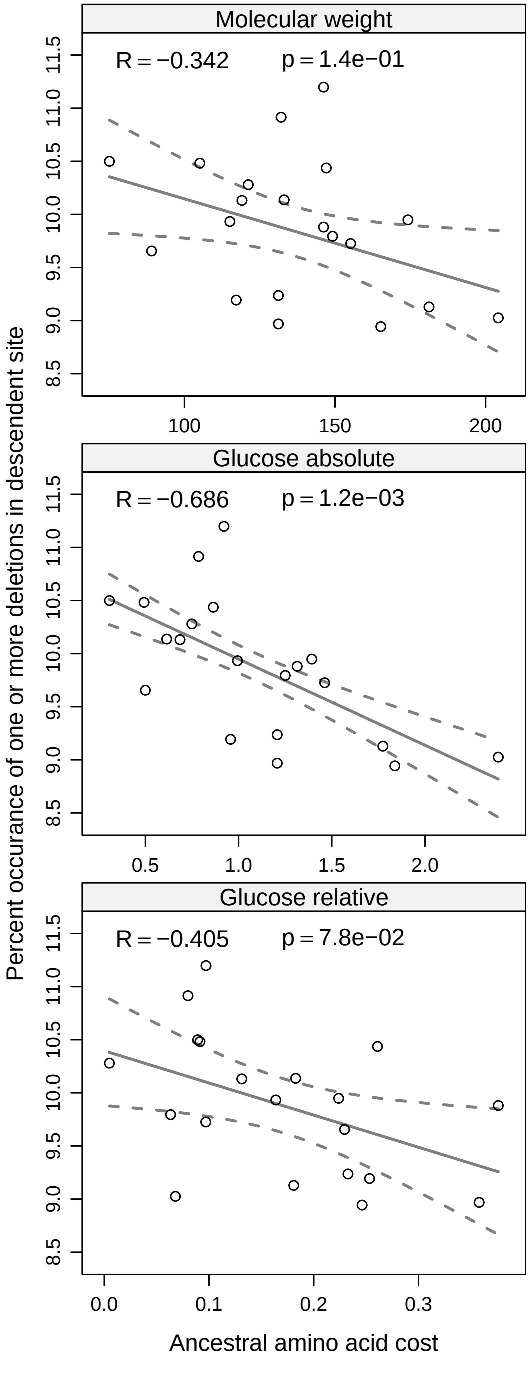


Figure 3.4: Comparison of amino acid frequency of deletion with three estimates of amino acid cost. Cost is that of the ancestral amino acid. Deletion frequency is estimated from number of times a deletion occurs at site in a descendant species given the amino acid at the ancestral site. Robust linear regression and 95% confidence intervals are used to indicated trend. Each plot indicates the Spearman's rank correlation.

3.4 Discussion

3.4.1 Amino acid biosynthetic cost is a small factor in yeast gene expression

Predictors of transcript and protein levels

The multivariate analysis of gene expression data suggests amino acid cost is a weak predictor of *S. cerevisiae* transcript and protein levels. Furthermore, regardless of which cost type was used in the regression, the optimisation of the encoding transcript for translation (predicted using CAI) was the most important variable. The correlation of CAI with transcript levels is not a new discovery, but this analysis compares CAI with other predictors of transcript and protein levels to show that CAI is a much greater predictor than the theorised role of amino acid cost [34], or atomic content [48, 49].

After CAI, the other factors considered in the regression all appear to play approximately a similar minor role in explaining variation in protein and transcript data. In general across each cost regression model of the transcript data, cost, carbon and nitrogen content appear more important than tRNA count or sulphur content. The regression models explains $\sim$2% more variation in the protein data, but the other factors were less significant than CAI indicating that the CAI has a large explanatory power for protein levels as well. In some cases the removal of tRNA or sulphur content, and even carbon content in one instance, produces a more parsimonious regression model of variance in protein levels.

Previous analysis has compared cost minimisation with predicted expression using CAI [34], however the emphasis of this analysis is compare many expression factors together against measured *in vivo* expression. As a result of this analysis CAI is shown to be the strongest predictor of both the transcript and protein expression data. The other considered factors, while showing some predictive power, are much less important than that of CAI and therefore translational optimisation.

Predictors of metabolite levels

Codon usage bias is not applicable to metabolite levels, and therefore all explanatory power in the regression model is based on amino acid cost, genome encoded tRNA, and carbon, nitrogen and sulphur content. The variation explained in the metabolite data was much greater than that of transcript or protein data, more than twice as much depending on which cost measure was used. This result indicates that the combination of these factors used in regressing free amino acid data are strong predictors of the variation. This result indicates, in *S. cerevisiae*, that amino acids are maintained in the cell at levels related to cost, atomic content and genome encoded tRNA levels. The caveat to this result, however, is the limited number of data points and possible pseudo-replication in metabolite measurements. Analysis of a larger sample will verify cost minimisation in free amino acid levels.

3.4.2 Amino acid biosynthetic cost is a selective pressure in protein sequence evolution

Variation in amino acid use between species

Analysis of the average usage of each amino acid indicates cost selection is a weak selection pressure in shaping overall protein sequence in *Saccharomyces* species. The large amount of unexplained variation in the relationship between cost and usage shows amino acid usage (seen in Figure 3.2) is not maintained in the genome proportional to molecular weight or $A_{glucose}$ biosynthetic cost. Amino acid usage may instead emerge as a result of specific structural and functional requirements. The limited relationship between these two costs and genomic usage may however still indicate a trend to minimise the use of expensive amino acids albeit after satisfying constraints on protein structure and function.

Comparison of the usage of amino acids between closely related species in the *Saccharomyces* genus (see Figure 3.3) shows, while cost minimisation is not apparent in amino acid content in the genome, the variability in usage does indicate a cost trend related to molecular weight or $A_{glucose}$ cost. Costly amino acids show a reduced interspecies variance in use compared with cheaper amino acids. This may indicate a selection pressure to maintain existing expensive amino acids for a specific functional role. Cheaper amino acids may vary more in genome usage as the cost selection pressure on their genomic use is proportionally less than expensive amino acids.

Cost minimisation trends in non-conserved sites

Amino acid frequency was compared in conserved and non-conserved sites across the alignments of four *Saccharomyces* species in Figure 3.4. The aim of this analysis was to determine if cost minimisation is a stronger trend outside positions with strong constraints for amino acid structure and function.

This comparison highlighted a strong trend for cost minimisation in the non-conserved sites when compared with the molecular weight or $A_{glucose}$ estimates of amino acid cost. The conserved sites however showed no strong relationship with amino acid cost. The greatest degree of cost-related correlation was observed at the non-conserved sites for the PAM250 > 0.5 level of conservation.

This suggests cost-minimisation may indeed be a significant selective force on variable sites. Given this analysis the most relevant definition of conservation may be the use of identical or highly functionally similar amino acids, and sites that are less conserved are where cost minimisation is a factor in protein sequence evolution. Further analysis of highly expressed proteins, or orthologs with differing tree lengths may shed further light on the extent of this trend.

Amino acid cost affects relative substitution rate

The relative substitution rates along with ancestral amino acid at each site in the same *Saccharomyces* protein sequence alignments were estimated for comparison with biosynthetic cost. The aim of this analysis was to determine cost related trends in the fixation of amino acids related to minimising biosynthetic cost in protein sequence.

Comparing mean substitution rates for the ancestral amino acid indicated a trend for the substitution of expensive amino acids inversely proportional to molecular weight or $A_{glucose}$ biosynthetic cost. This result further suggests cost minimisation is a selective force in the evolution of protein sequence. Biosynthetically expensive amino acids were observed to have a low substitution rate indicating they remain fixed in the protein sequence, which may be for a possible specific functional role. The reason biosynthetically cheaper amino acids may vary more is because the cost burden association with random insertion into the genome is much less. Alternatively, amino acid positions with little or no structural or functional constraint may evolve to a less costly state. The only constraint on cheaper amino acids in variable positions may be their hypothesised role in protein structure or function. If this hypothesis is indeed the case, this may provide further methods for understanding the cause and rate of amino acid substitutions and therefore genetic distance and divergence. On possible example is that expensive amino acid changes between orthologs may represent a greater genetic distance than the interchange of inexpensive amino acids.

Analysis of the frequency of deletions given the ancestral amino acid suggested that cost minimisation may also be a force in the rate at which amino acids are retained or removed from protein sequence. The results of this analysis were inconclusive for the $R_{glucose}$ and molecular weight costs, but there was a greater correlation coefficient for the $A_{glucose}$ biosynthetic cost. This showed that expensive amino acids are less likely to be removed from the protein sequence. This may be indicative of functional constraint to retain expensive amino acids, where the fixation of functionally constrained expensive amino acids is observed despite their negative effect on cost minimisation. In comparison there is less constraint against the deletion of biosynthetically cheaper amino acids. Assuming equilibrium between insertions and deletion this may also indicate cheaper amino acids are more likely to be inserted.

The relative glucose cost shows limited explanatory power

The $R_{glucose}$ cost showed a strong positive relationship with genomic amino acid usage, but showed no correlation with variation in amino acid use, amino acid conservation, evolutionary rates, or percentage deletion. This suggests the $\mathrm{R}_{glucose}$ cost shows no relationship with the substitution or fixation of amino acids in the *S. cerevisiae* genome. The analyses of these data suggest the relative cost measures may not be as meaningful as the absolute measures and that observed sequence evolution provides a useful filter to show which systems biology approach is better.

Future work to understand the role of metabolic cost

The previous chapter showed how estimation of amino acid cost may vary between species and environment. A similar trend may also be inferred for the importance of metabolic cost in sequence evolution in different species. The *Saccharomyces* genus examined have an extensive regulatory network and show distinct cellular responses to a variety of conditions [60]. Therefore in a fluctuating environment, metabolic cost may not be as strong a selection pressure as the ability to adapt to nutrient scarcity or excess. Future analysis could consider small oceanic species invariant to environmental conditions where metabolic cost may be a much stronger selection pressure given the genome is streamlined for a single environment [75]. Such species with compact genomes and limited number of paralogs may indicate that cost minimisation is inversely proportional to adaptation to multiple environments. Gene duplication may be a factor in the evolutionary adaptation to new environments [58], therefore the use of gene duplication events may be key to identifying the relationship between the importance of genomic cost minimisation versus the evolutionary pressure to encode responses to multiple environmental conditions.

4

Comparing predicted *in silico* gene importance with gene use and evolution

Summary

Current flux balance analysis approaches to understanding gene importance focuses on the effect of complete gene removal on organism fitness. This chapter instead uses the three novel estimates of gene importance produced in Chapter 2 to examine gene use and sequence evolution in *Saccharomyces cerevisiae*. Estimates of gene importance are compared with the *in vivo* effect of gene loss, the *in vivo* effect when one allele copy is removed, estimated gene evolutionary rate, and gene expression levels predicted from codon usage bias. The FBA estimates of gene importance explained only small levels of variation in gene evolutionary rate and codon usage predicted expression, however the significant correlates do suggest a marginal relationship. The results of this work represent an attempt to understand gene importance using more than just *in silico* dispensability, and future work may result in more accurate estimation of the importance of individual genes in metabolism and evolution.

4.1 Introduction

4.1.1 Genome annotation and functional genomics

The development of pyrosequencing technology [76], a cheaper and faster alternative to Sanger sequencing [77], allows the rapid sequencing or re-sequencing of a genome. The reduction in cost afforded by short read sequencing technologies will increase the number of available genome sequences on top of the large number of those already known.

The first-time sequencing of a genome requires annotation to determine the encoded content. Bioinformatics pipelines allow the identification of large kilobase length protein coding genes [78] to small non-coding transposable elements a few nucleotides in length [79]. Genome annotation is however imperfect, and some open reading frames may fail to be identified. An example of this is *S. cerevisiae*, a long studied model organism, the genome of which was first sequenced in 1996 [80], and later resequenced in 2009 [81]. Figure 4.1 shows that thirteen years after the original sequencing, 977 open reading frames are still uncharacterised, and 811 are classified as 'dubious'. This demonstrates that even in well studied organisms the function of a large proportion of open reading frames may remain unresolved. This serves to underline the point that the sequencing and annotation of genome is only the first step in understanding organism biology.

Identifying gene function

A common approach to studying gene function uses sequence alignment to identify similar sequences with a known function. This approach is less effective when closely related sequences are not available, or can result in incorrect annotation if the closest homolog has a divergent function [82]. Nevertheless homology searching is a fast and inexpensive first pass at genome functional annotation.

If the function of a gene remains unclear using *in silico* methods an *in vivo* approach to understand function is to remove a gene from the genome ('gene knockout') and measure the corresponding phenotypic effect [83]. In *S. cerevisiae* a library of homozygous gene deletion mutants has been used to screen for gene function in different conditions [84, 85]. A phenotypic effect resulting from knockout on a specific medium indicates the activity of the gene may be related to that medium. If the removal of the gene results in an inviable mutant, this indicates

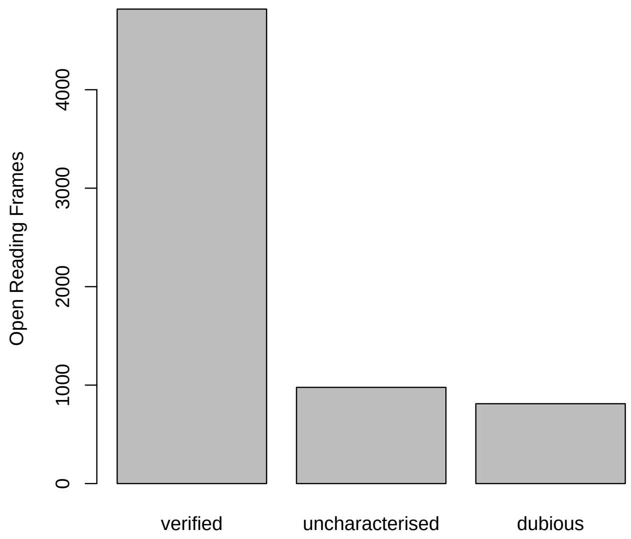


Figure 4.1: Summary of open reading frame classifications in the *Saccharomyces* Genome Database, July 2009.

the gene is also essential for growth and its role may be considered important to metabolism.

The function of essential genes in diploid organisms may also be examined through removing one of the two allelic copies and again measuring the corresponding fitness effect [86, 87]. Removing only a single gene copy also allows essential genes to be studied. This approach has been used to identify possible drug targets in *S. cerevisiae* where a large mutant phenotype defect on a specific drug may indicate the gene has a role in metabolising that drug [88]. These studies are useful to drug development as *S. cerevisiae* shares many homologs with the human genome [89, 90] and is cheaper to screen experimentally.

Such large-scale functional studies, while providing large quantities of data, must be taken in the context of the experimental conditions. For example two *in silico* gene knockout studies have shown that gene essentiality may be very specific to the environment in which the mutant is examined, as well as which other genes are present in the genome [37, 39]. Furthermore an *in vivo* study of gene knockouts indicated that mutants producing no discernible fitness phenotype may still have an effect on intracellular metabolite levels [91].

4.1.2 Understanding gene function *in silico*

A major disadvantage of using *in vivo* methods to study gene function is that the analyses are expensive and time consuming relative to *in silico* work. Flux balance analysis and genome scale models, described in Chapter 2, can however be used to predict these gene knockouts phenotypes computationally. Genome scale models can be useful to predict the existence of unidentified genes in a genome where the gene encoded reaction would be necessary for viable cell growth. This was the case for the initial *S. cerevisiae* construction which showed many membrane transporters have yet to be identified [41].

A genome scale model also allows gene essentiality predictions to be made for large numbers of conditions much more cheaply than testing *in vivo* [26]. Furthermore combinations of gene knockouts can be easily screened *in silico* to identify epistatic or antagonistic gene product interactions [42, 43]. The effort required to test combination knockouts *in vivo* may well take months or years, while *in silico* analyses may screen all combinations in days.

Flux balance analysis may be used to further characterise the role of a gene and the encoded reaction in metabolism beyond just screening essentiality. Genes with a high, but non lethal, impact on metabolism when deleted may identify critical metabolic subsystems [92]. The underlying factors causing gene dispensability can be examined using reaction activity between environments, and the number of encoding paralogs [39]. The identification of coupled reactions, whose flux is correlated, may also identify regions in metabolism that appear in the same operon or are co-expressed [93].

4.1.3 Genome scale models and evolution

Despite their utility in studying gene function using genome scale models for simulating *in silico* selection pressures may be of limited utility as gene essentiality is uncorrelated with evolutionary rate [39, 40]. In addition it may be hypothesised that the activity of a reaction is more likely to be altered through sequence substitutions than by complete gene loss from the genome.

Predicting selection pressures resulting from subtle functional affects such as changes in gene sequence is difficult using flux balance analysis. However explaining the action of evolution *in silico* is a worthy goal as for the same reason that accurate *in silico* gene deletion analyses is advantageous [94].

An obvious first step to modelling evolution using a genome scale model is to begin with genes encoding metabolic reactions. Modelling a small deleterious effect on a metabolic reaction is difficult, as the relationship between gene sequence and reaction activity is unclear. Estimating the deleterious effect of a gene mutation using flux balance analysis may require reducing the flux through the encoded reaction and determining the effect on growth rate. One technical problem is that the predicted reaction flux is the result of linear optimisation and may not necessarily be close to the actual reaction flux. Furthermore it is difficult to predict what effect a deleterious mutation will have on reaction activity where the effect may be expected to vary from reaction to reaction. Overall this makes predicting evolution using flux balance analysis a challenging problem.

In Chapter 2 estimates of gene importance were derived that complement those based on reaction deletion. The motivation for this analysis is to study gene evolution through simulating small mutations in gene sequence as opposed to complete gene loss. The three derived estimates of gene importance are: the category of constraint on changes in reaction flux, reaction flux, and the sensitivity of growth to changes in reaction flux. These measures were estimated for the subset of reactions in the model encoded by a single gene because the relationship between changes in gene sequence would be expected to only have a single point of effect in metabolism.

4.1.4 Summary of results

In this chapter the gene cost estimates produced in Chapter 2 are compared with four metrics of gene function and evolution. The aim of this analysis is to determine if these novel gene importance measures estimate the selection pressures on genes in *S. cerevisiae*. The variables examined are the *in vivo* homozygous mutant growth, *in vivo* mutant growth when one allele copy is removed (hemizygous), inter-species estimated gene evolutionary rate, and gene expression levels predicted using codon usage bias. The analysis indicates marginal correlation between each of the examined variables and suggests further development in *in silico* modelling is required to derive accurate estimates of gene function and evolution.

4.2 Materials and Methods

4.2.1 Gene dispensability

The gene dispensability estimates used in this analysis was downloaded from the *Saccharomyces* Genome Database [95]. The file downloaded (*phenotypes.tab*) was filtered for the homozygous gene deletion survey by Giaever *et al.* [84] where homozygous deletion viability was estimated on yeast, peptone, and dextrose (YPD) media. Of the 5,916 assessed genes 1,105 were identified as essential to growth.

4.2.2 Hemizygous mutant fitness

Hemizygous mutant fitness was estimated from a large scale *S. cerevisiae* deletion library [86]. Each strain represented a knockout of one allelic copy at a one gene loci. Each knockout was produced from homologous recombination using a null cassette with high sequence identity to the target region. The insertion cassette incorporated up and down tags unique to each mutant strain allowing estimation of mutant biomass from PCR amplification and hybridisation (see Giaever *et al.* for a detailed explanation of this method [88]).

The fitness of each mutant was estimated in four different media by Delneri *et al.* [87]. The media considered were glucose, ammonium and sulphate limitation, and grape juice. All strains were examined simultaneously using continuous culture. Culture biomass was sampled at 48 hour intervals and the strain fitness was estimated from regression of the log. ratios of tag levels over time. An increase in tag levels indicating a positive growth effect (haploproficient) and a decrease in tag levels indicating a negative growth effect (haploinsufficient). Each strain fitness is presented as the difference in mutant growth rate at a 0.1 hr^{-1} chemostat dilution rate.

The fitness values were normalised using median centreing. The fitness data was filtered using p values estimated from model-based resampling. False-discovery q values were also estimated to account for multiple testing (described in [96]).

4.2.3 Gene evolutionary rate

The evolutionary rate data used in this analysis were estimated by Wall *et al* [68]. The estimates were derived from multiple DNA sequence alignments of *S. bayanus*, *S. mikatae*, *S. bayanus*, and *S. cerevisiae*. The sequences were codon aligned using clustalw and the phylogenetic tree estimated using dnaml from the PAML package [70]. Gene dN/dS, dN, and dS rates were estimated using codeml with model parameter 0.

4.2.4 Codon adaptation index

The CAI for *S. cerevisiae* genes was estimated by Coghlan and Wolfe [63]. Each gene CAI is estimated as the bias in codon use relative to 24 highly expressed proteins, the values range from 0 (no bias) to 1 (maximum bias).

4.3 Results

4.3.1 Gene importance and dispensability

This section compares the requirement of a gene for growth (gene dispensability) with FBA estimated gene importance. The hypothesis for comparing these two variables is that essential genes may be expected to have a critical role in metabolism and therefore the reaction in the computational model should be correspondingly estimated as important. Comparison of *in silico* and *in vivo* gene dispensability has been performed previously [41] but no work has used the types of gene importance measures derived in Chapter 2. The *in vivo* gene dispensability measures were determined by Giaever *et al.* [84] through removing both copies of a gene and examining if the mutant is viable when grown on rich media.

Reaction constraint

Gene dispensability *in vivo* was compared with *in silico* estimated reaction constraint categories. The expected outcome is that reactions where the flux cannot be reduced are more likely to be encoded by essential genes, and that zero flux reactions are will be more likely to be encoded by dispensable genes. Table 4.1 compares the frequency of reaction constraint category with gene dispensability. Figure 4.2 illustrates the χ^2 residual deviance for each category.

In both the optimal and sub-optimal solutions the χ^2 residual deviance (shown in parentheses) indicates, in each environment, that constrained reactions are more likely to be essential than expected, and also less likely to be dispensable. In contrast the reverse is true for the variable reaction category where these reactions are more likely to be dispensable and less likely to be essential than would be expected from the χ^2 null distribution of no association. The zero flux category reactions however do not show any strong deviance between essential and dispensable.

Reaction flux

Gene dispensability was also compared with reaction flux to determine if high flux reactions are more likely to be essential. The expectation is that the removal of a high flux reaction will have a greater effect on metabolism and growth, particularly for the single gene associated reactions where there is no corresponding

solution	limitation	p	reaction	essential		dispensable	
optimal	glucose	0.024	constrained	19	(1.77)	36	(-0.97)
			variable	2	(-1.57)	23	(0.86)
			zero flux	36	(-0.41)	131	(0.22)
	ammonium	0.052	constrained	18	(1.55)	36	(-0.85)
			variable	3	(-1.43)	26	(0.79)
			zero flux	36	(-0.29)	127	(0.16)
	sulphur	0.052	constrained	18	(1.55)	36	(-0.85)
			variable	3	(-1.43)	26	(0.79)
			zero flux	36	(-0.29)	127	(0.16)
suboptimal	glucose	0.017	constrained	19	(1.85)	35	(-1.01)
			variable	2	(-1.63)	24	(0.89)
			zero flux	36	(-0.41)	131	(0.22)
	ammonium	0.018	constrained	18	(1.55)	37	(-0.84)
			variable	2	(-1.96)	30	(1.06)
			zero flux	36	(-0.03)	123	(0.02)
	sulphur	0.004	constrained	20	(2.11)	36	(-1.15)
			variable	2	(-2.01)	32	(1.09)
			zero flux	34	(0.33)	124	(0.18)

Table 4.1: Comparison of gene dispensability with reaction constraint category. Reaction constraints was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. Gene dispensability estimates were produced by Giaever *et al.* [84]. The Pearson χ^2 p-value is indicated for each environment. Each row shows the frequency of reaction type, and Pearson χ^2 residual deviance from the null distribution in parentheses. A positive value indicates a greater than expected frequency, and vice versa for negative values.

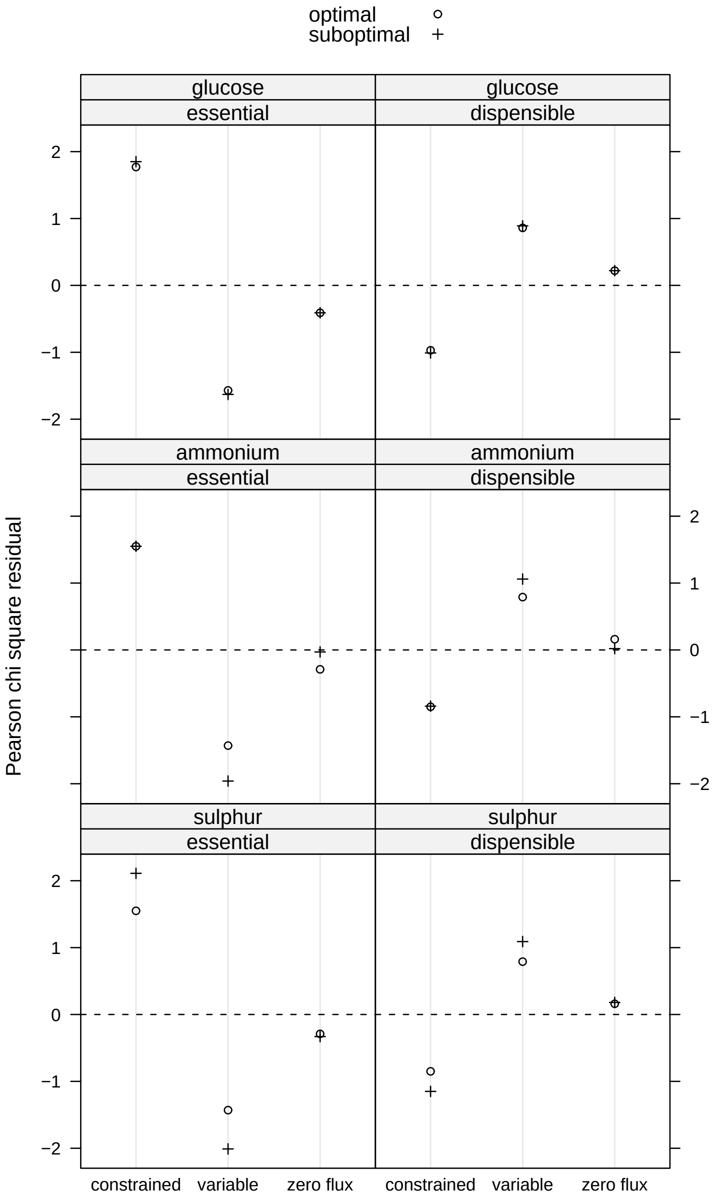


Figure 4.2: Pearson χ^2 deviation values for reaction constraint category for each environment in each solution. Each value represents the difference between the observed and expected value scaled by the expected value.

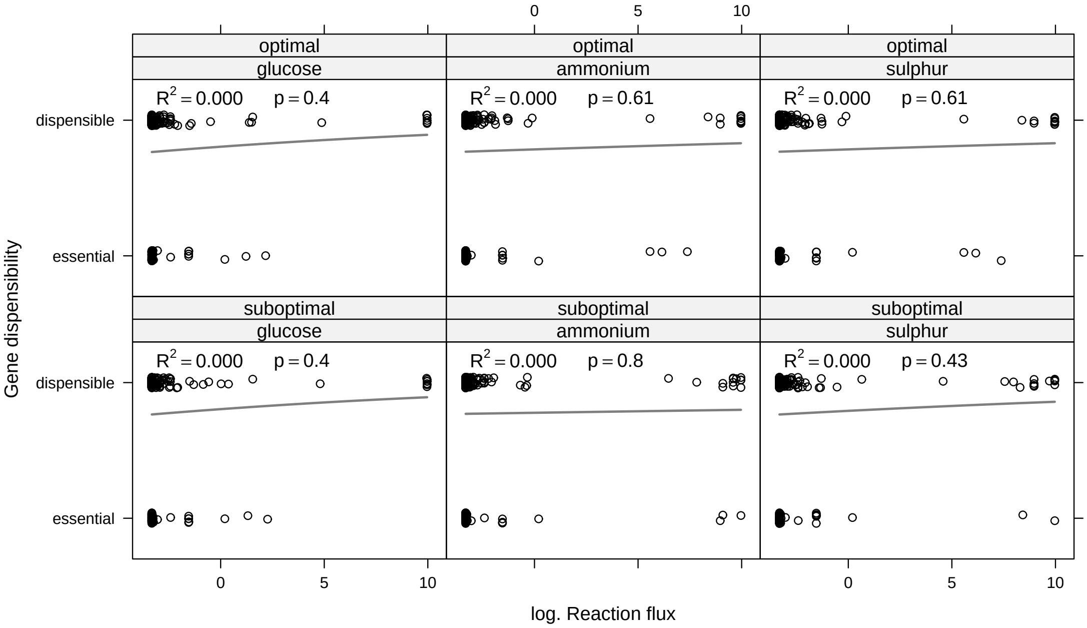


Figure 4.3: Comparison of reaction flux with gene dispensability. Reaction flux was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. Gene dispensability estimates were produced by Giaever *et al.* [84]. A generalised linear model with binomial errors was used to regress dispensability and reaction flux using a sigmoid relationship. The trend line indicates the model predicted values. Pseudo-R^2 values are estimated as the square of the Spearman's rank correlation between the observed data and regression predicted values.

paralog to buffer the reaction. Figure 4.3 compares gene dispensability with FBA estimated reaction flux in glucose, ammonium and sulphur limitation for both optimal and suboptimal FBA simulations.

The binomial regression demonstrates no relationship between reaction flux and gene dispensability. In each figure the estimated model R^2 is close to 0 and non-significant. The observed data points and trend line also show no bias towards high flux reactions being more likely to be essential. This suggests that gene dispensability is unrelated to predicted reaction flux.

Reaction sensitivity

In this section gene essentiality is compared with the sensitivity of the encoded reaction flux. The hypothesis is that essential genes will encode more sensitive

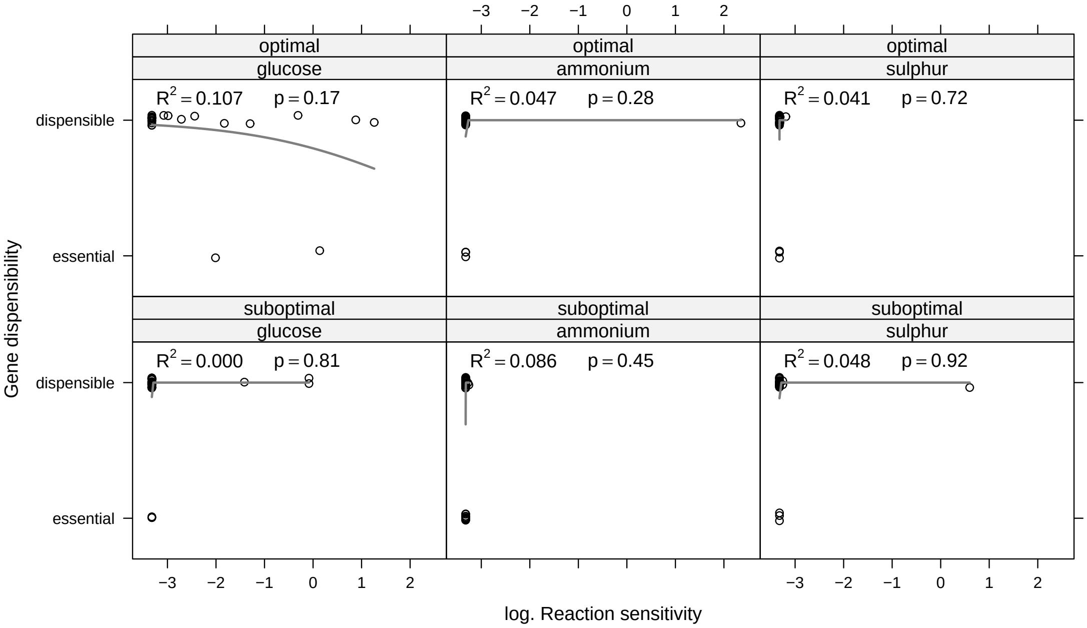


Figure 4.4: Comparison of reaction sensitivity with gene dispensability. Reaction sensitivity was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. Gene dispensability estimates were produced by Giaever *et al.* [84]. A generalised linear model with binomial errors was used to regress dispensability and reaction flux using a sigmoid relationship. The trend line indicates the model predicted values. Pseudo-R^2 values are estimated as the square of the Spearman's rank correlation between the observed data and regression predicted values.

reactions since the effect of a reduction in reaction flux is predicted to have a greater effect on predicted cell growth. Figure 4.4 compares reaction sensitivity with gene dispensability in glucose, ammonium and sulphur limitation for both optimal and suboptimal FBA simulations.

In each environment and FBA solution type there is no relationship between *in silico* reaction sensitivity and *in vivo* gene dispensability. The estimated R^2 value in each plot indicates reaction sensitivity is a poor predictor of gene dispensability, but the small number of data points is also a contributing factor in the lack of observable relationship.

4.3.2 Gene importance and hemizygous fitness

Delneri *et al.* [87] surveyed which genes in *S. cerevisiae* affect growth when one of the two diploid gene copies is removed. This hemizygous fitness effect was quantitatively estimated for each mutant strain. This section compares these *in vivo* estimates of gene effect on growth with the *in silico* FBA estimates developed in Chapter 2. The expectation is that genes predicted to be important *in silico* should correspond to genes shown to have a significant effect on growth *in vivo* when measured in a hemizygous state.

Reaction constraint

Hemizygous fitness is compared here with the reaction categories for the constraints on *in silico* reaction flux. The hypothesis for this analysis is that *in silico* predicted constrained reactions will correspond to the genes that induce a greater fitness defect when gene dosage is reduced by half *in vivo*. Figure 4.5 compares differential hemizygous mutant growth rate with the reaction constraint categories for glucose and ammonium limitation in both the optimal and suboptimal FBA solutions. Sulphur limitation was not included in this analysis as no hemizygous data was available for this environment.

The figure indicates no difference in the distributions of hemizygous growth effect for each of the reaction constraint categories. The hemizygous growth rates for each of the constrained, variable, and zero flux categories appears to show similar distributions which suggests the fitness effects are not different between categories. One factor analysis of variance (ANOVA) confirms this, where the R^2 for each ANOVA is close to 0 and all p values are close to 1. This result indicates *in vivo* differential hemizygous growth rate is not predicted by whether the reaction is constrained *in silico*.

Reaction flux

This analysis compares *in silico* estimated reaction flux with *in vivo* derived gene hemizygous growth effect. The hypothesis is that genes predicted to encode a high flux reaction will correspond to a growth defect when gene dosage is reduced by half. The reason for this assumption is that a high flux reaction will be important to growth and removing one gene copy may be expected to reduce the enzyme levels catalysing the reaction. Figure 4.6 compares the *in silico* reaction flux with

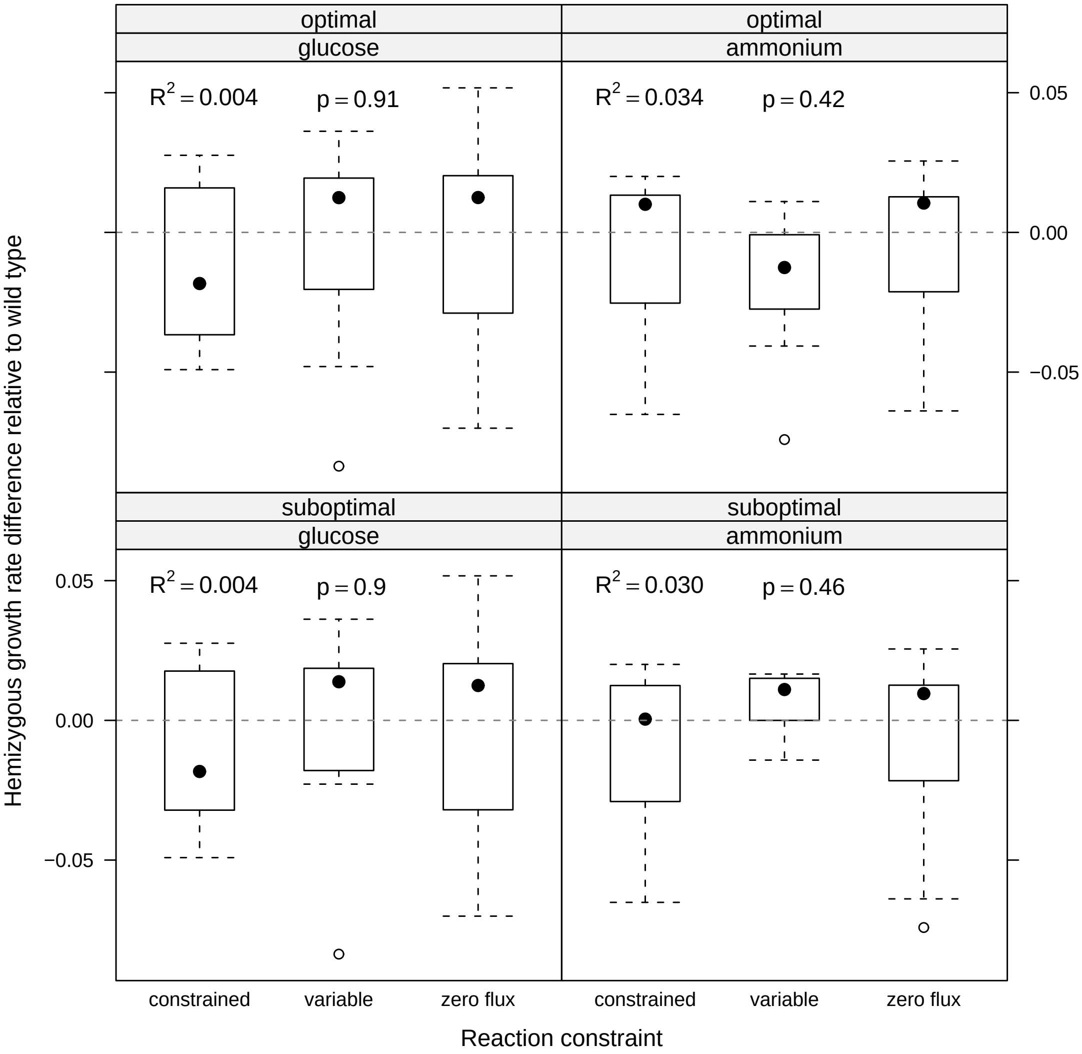


Figure 4.5: Comparison of hemizygous mutant fitness with estimated reaction constraint category. Reaction constraint categories were estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose or ammonium limitation for both optimal and suboptimal solutions. Hemizygous fitness was estimated in *S. cerevisiae* by Delneri *et al.* [87] and represents the difference in growth rate for each hemizygous knockout mutant from the chemostat dilution rate of 0.1 hr^{-1} in either ammonium or glucose limitation. One factor analysis of variance regression coefficients are indicated in each plot.

in vivo hemizygous growth effect in glucose and ammonium limitation using the optimal and suboptimal FBA solutions.

The figure demonstrates no relationship between the two variables. The Spearman's rank analysis further confirms this where all R estimates are close to 0 and p-values near 1. The plot also illustrates the skew in the data where the majority of the *in silico* estimates are small, while the *in vivo* effects are more uniformly distributed in the range of ± 0.05 hr^{-1}. The large *in silico* reaction fluxes predicted also do not correspond to the expected hemizygous growth defect. In the glucose limited estimates the large reaction flux values even correspond to a positive effect on growth. This result is the opposite of the expected and indicates *in silico* estimated reaction flux is a poor predictor of *in vivo* hemizygous fitness or that experimental noise in the derivation of the data may also be a factor.

Reaction sensitivity

Finally *in vivo* hemizygous growth effects were compared with *in silico* estimated reaction sensitivity. The expected outcome is that the growth sensitivity to reaction flux reduction will correspond to the *in vivo* effect of reducing gene dosage. The *in silico* measure estimates gene importance by reducing the encoded reaction flux and comparably the *in vivo* methods estimate gene importance by reducing gene copy number and measuring the effect on growth. Figure 4.7 compares the hemizygous growth rate with *in silico* estimated reaction sensitivity in glucose and ammonium limitation for the optimal and suboptimal solutions.

This figure highlights no clear trend between the two variables. As with reaction flux in the previous section the majority of sensitivities are very small with only a few larger observations. Spearman's rank correlation confirms the lack of significant relationship in three out of four of the plots: both types of ammonium limitation and the glucose limited optimal solution. The suboptimal glucose limited Spearman's rank however is borderline significant but has a negative coefficient indicating anti-correlation between the two variables - the opposite of the hypothesised trend.

4.3.3 Gene importance and evolutionary rate

This section explores examines whether genes encoding important metabolic reactions in *S. cerevisiae* are constrained by greater selection pressure. The theory

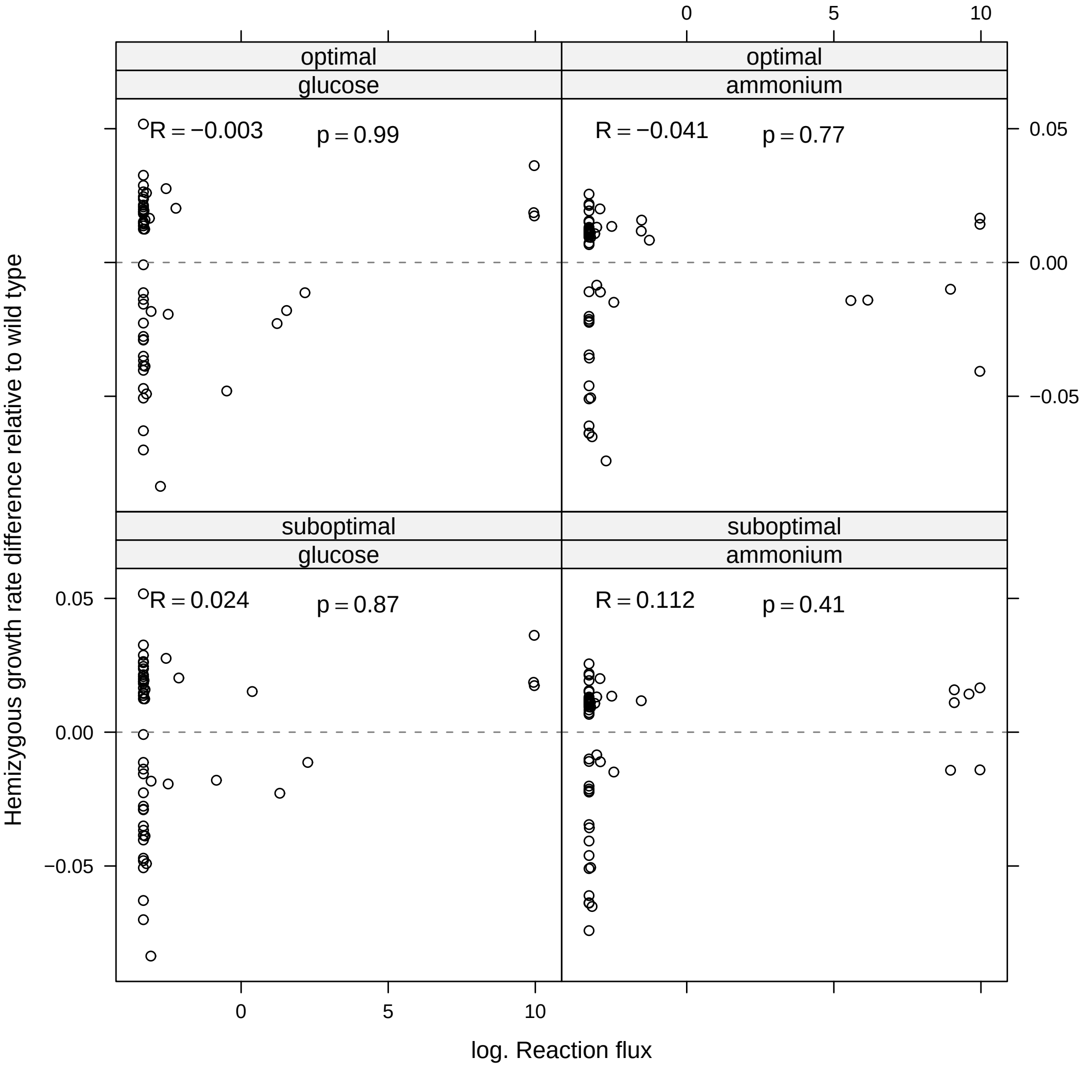


Figure 4.6: Comparison of hemizygous mutant fitness with estimated reaction flux. Reaction flux was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose or ammonium limitation. Gene flux was estimated for both optimal and suboptimal FBA solutions. Hemizygous fitness was estimated in *S. cerevisiae* by Delneri *et al.* [87] and represents the difference in growth rate for each hemizygous knockout mutant from the chemostat dilution rate of 0.1 hr^{-1} in either ammonium or glucose limitation. Spearman's rank correlation is indicated in each plot.

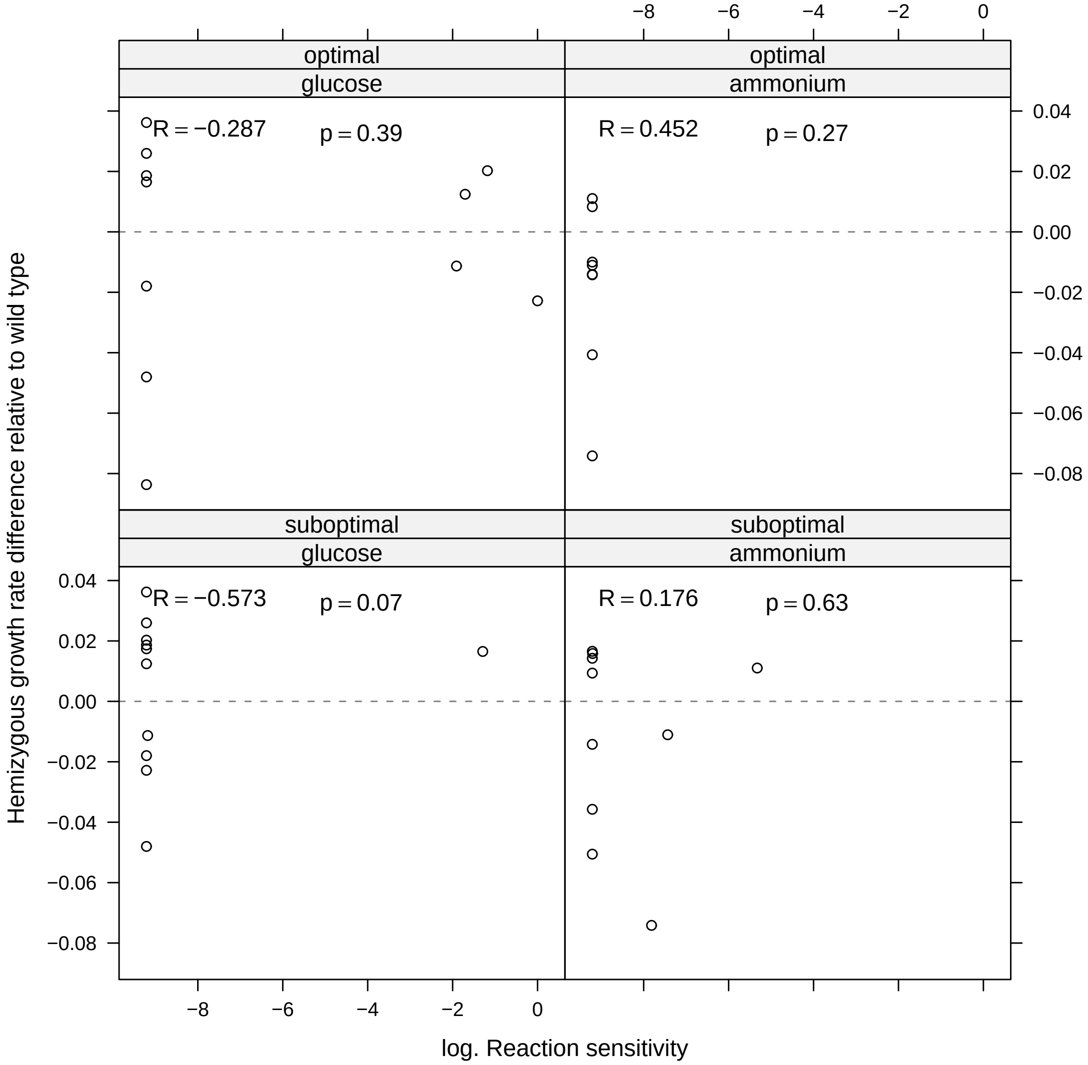


Figure 4.7: Comparison of hemizygous mutant fitness with estimated reaction sensitivity. Reaction sensitivity was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose or ammonium limitation for both optimal and suboptimal solutions. Hemizygous fitness was estimated in *S. cerevisiae* by Delneri *et al.* [87] and represents the difference in growth rate for each hemizygous knockout mutant from the chemostat dilution rate of 0.1 hr^{-1} in either ammonium or glucose limitation. Spearman's rank correlation is indicated in each plot.

for this analysis is that a deleterious mutation in an enzyme will have a proportionally greater fitness effect the more important the catalysed reaction is to growth.

The analysis here compares FBA estimated gene importance with three different measures of evolutionary rate: dN/dS, dN, and dS. The dN/dS ratio estimates the total evolutionary rate of the gene comparing the rate of non-synonymous sequence changes scaled by the number of synonymous changes and also controls for genomic variations in mutation rate affecting dN. The dN measure estimates the number of non-synonymous sequence substitutions - changes that may be expected to alter protein structure and function. The dS measure estimates the number of synonymous sequence changes which may be proportional to mutation rate as these changes do not alter the protein sequence and can therefore be considered nearly neutral. Each of these variables allows comparison and insight into different aspects of genomic evolution. All evolutionary rate measures used here were estimated by Wall *et al.* [68].

Reaction constraint

This section compares whether a reaction flux constrained or not with the evolutionary rate of the encoding gene. The expected outcome of this analysis was that constrained reactions have a lower evolutionary rate as deleterious mutations in these reactions have a greater fitness effects.

Estimated reaction constraint categories are compared with dN/dS gene evolutionary rate in Figure 4.8. ANOVA regression in each environment indicates reaction flux constraint explains a small but significant amount ($R^2 \sim 0.1$) of variation in gene evolutionary rate. In each of the environments considered the zero flux reactions have a larger distribution of evolutionary rates and a higher median. In the optimal FBA solution, the variable flux reactions appear to have a lower median evolutionary rate than the constrained reactions, but there is still a large overlap between the distributions.

Figure 4.9 compares reaction constraint category with non-synonymous evolutionary rate across the three examined environments and two considered FBA solutions. The distribution of dN values is similar to that of dN/dS with zero flux reaction having a larger distribution than constrained or variable reactions. ANOVA again indicates that reaction constraint only explains a small marginal amount of variation in dN rates.

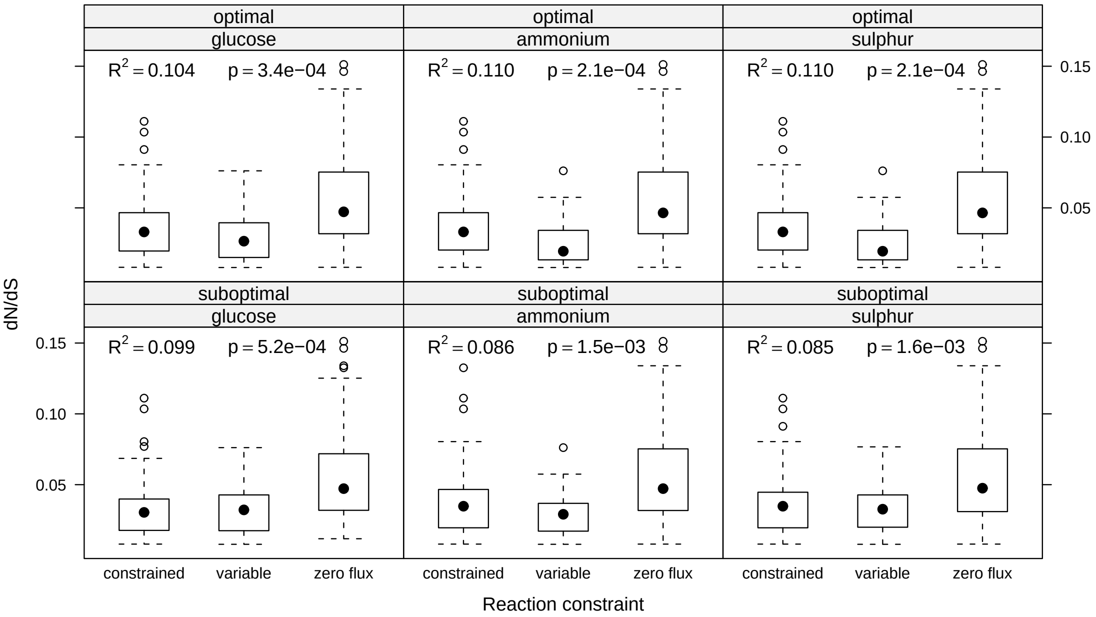


Figure 4.8: Comparison of gene evolutionary rate and predicted reaction constraint categories. Reaction constraint was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. *Saccharomyces* gene evolutionary rates (dN/dS) were estimated by Wall *et al.* [68]. One factor ANOVA regression coefficients are indicated in each plot.

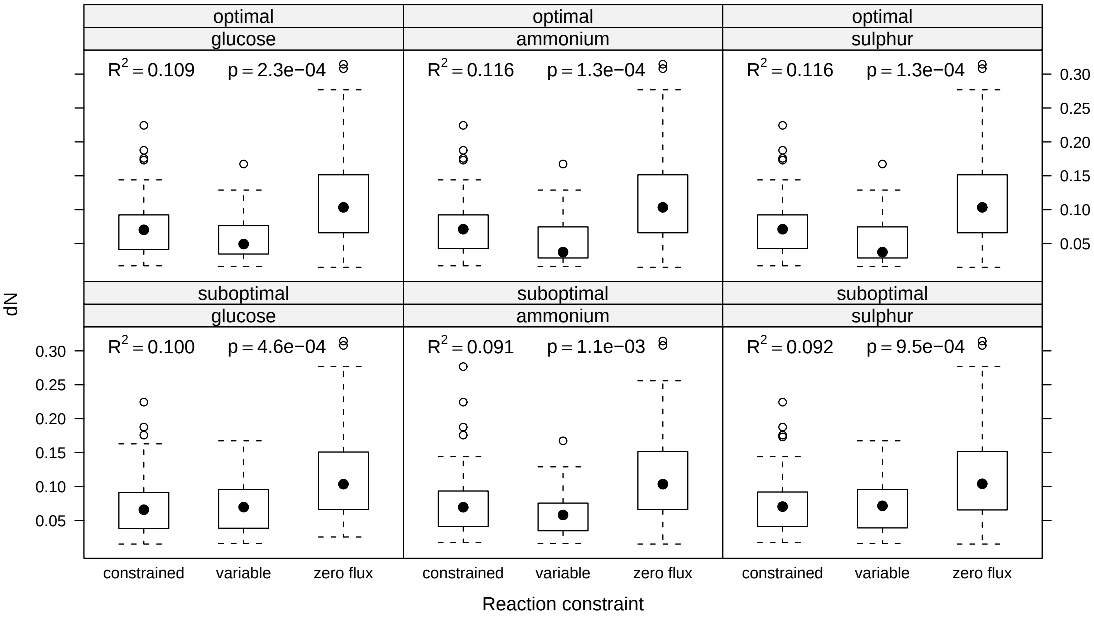


Figure 4.9: Comparison of gene non-synonymous (dN) evolutionary rate and predicted reaction constraint categories. Reaction constraint was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. *Saccharomyces* gene non-synonymous evolutionary rates were estimated by Wall *et al.* [68]. One factor ANOVA regression coefficients are indicated in each plot.

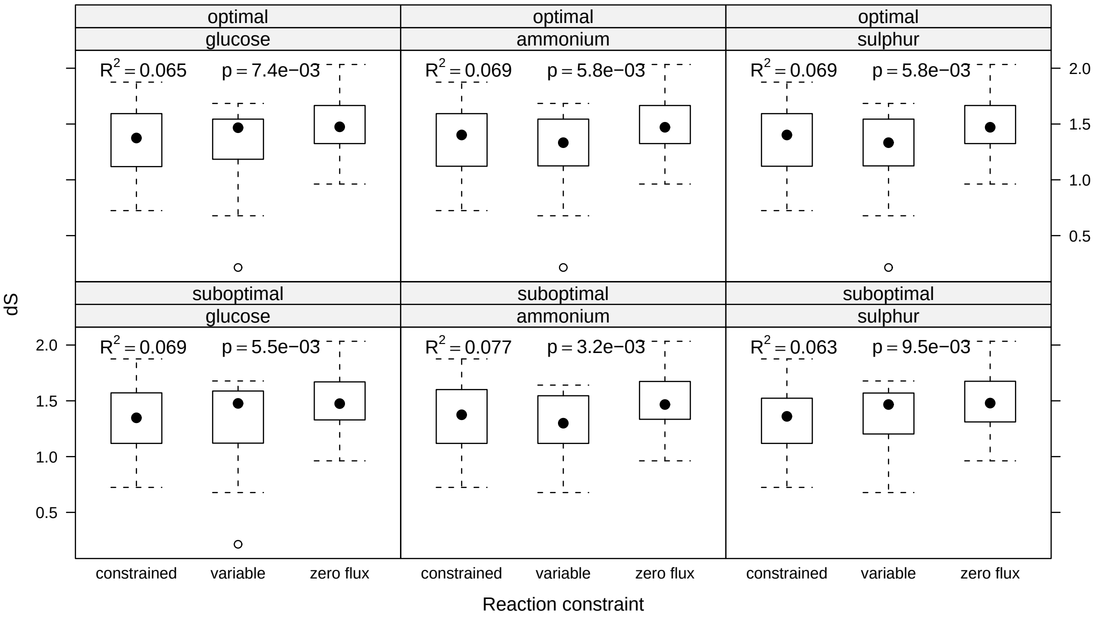


Figure 4.10: Comparison of gene synonymous (dS) evolutionary rate and predicted reaction constraint categories. Reaction constraint was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. *Saccharomyces* gene synonymous evolutionary rates were estimated by Wall *et al.* [68]. One factor ANOVA regression coefficients are indicated in each plot.

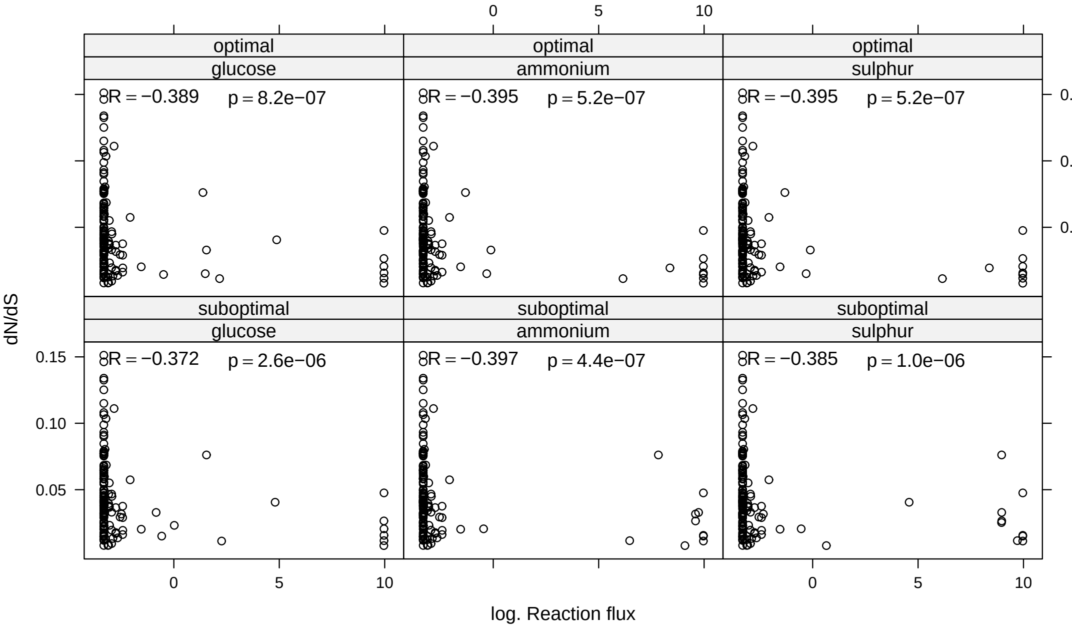


Figure 4.11: Comparison of gene evolutionary rate (dN/dS) and predicted reaction flux. Reaction flux was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. *Saccharomyces* gene evolutionary rates were estimated by Wall *et al.* [68]. Spearman's rank correlation coefficients are indicated in each plot.

Figure 4.10 compares reaction constraint with synonymous evolutionary rate across between three environments and two solutions. The distribution of dS values is less distinct between each reaction category compared with the previous two analyses. A lesser degree of explanatory power is also apparent in the smaller ANOVA R^2 value and the relationship appears to show a similar but weaker trend to that observed for dN. This indicates reaction constraint categories predict gene dS evolutionary rate less so than that of dN evolutionary rate.

Reaction flux

The three estimates of gene evolutionary rate were compared with predicted reaction flux for all single reaction associated genes. The expectation was that high flux encoding genes will have a lower evolutionary rate for dN and dN/dS. The reason being that deleterious mutations may be expected to have a greater fitness effect in enzymes catalysing high flux reactions relative to small flux reactions.

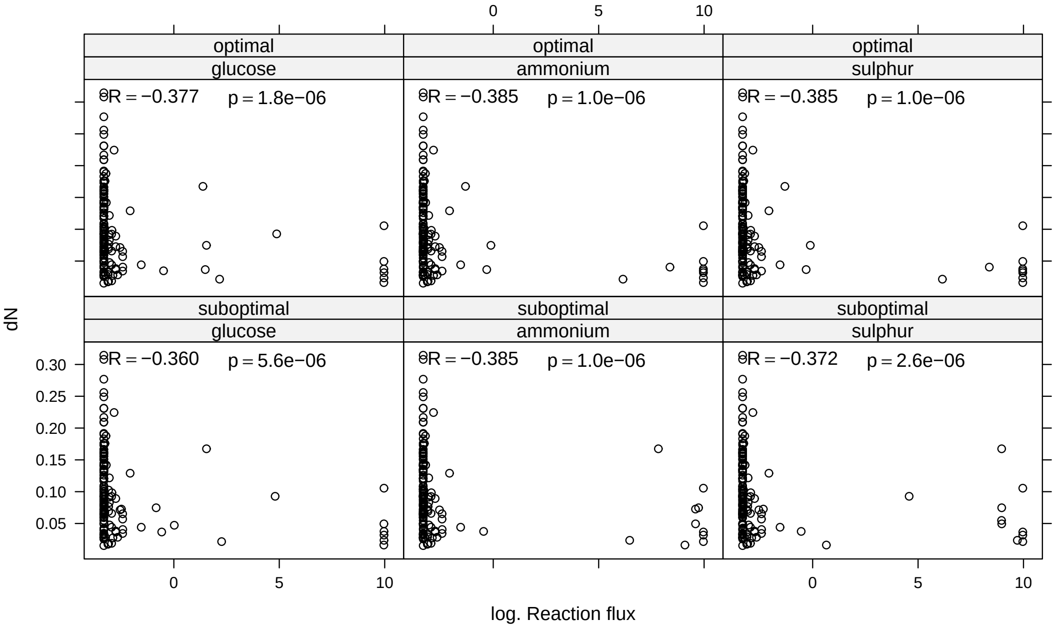


Figure 4.12: Comparison of gene non-synonymous (dN) evolutionary rate and predicted reaction flux. Reaction flux was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. *Saccharomyces* gene non-synonymous evolutionary rates were estimated by Wall *et al.* [68]. Spearman's rank correlation coefficients are indicated in each plot.

Figure 4.11 compares dN/dS evolutionary rate with predicted reaction flux. In each of the compared environments the Spearman's rank correlation indicates a small but significant negative correlation between reaction flux and dN/dS ratio. In the optimal solution, the largest negative coefficient was between the predicted ammonium and sulphur limited reaction fluxes. In the suboptimal solutions the largest Spearman coefficient was in ammonium limitation. This indicates a marginal relationship between the *in silico* predicted reaction flux and gene evolutionary rate.

Figure 4.12 compares non-synonymous evolutionary rate with predicted reaction flux. Across all environments and solutions dN also shows a weak negative correlation with reaction flux. Each of the correlation coefficients are similar to that observed to the dN/dS ration, indicating that the two measures share a similar small relationship with reaction flux.

Figure 4.13 compares synonymous evolutionary rate with predicted reaction

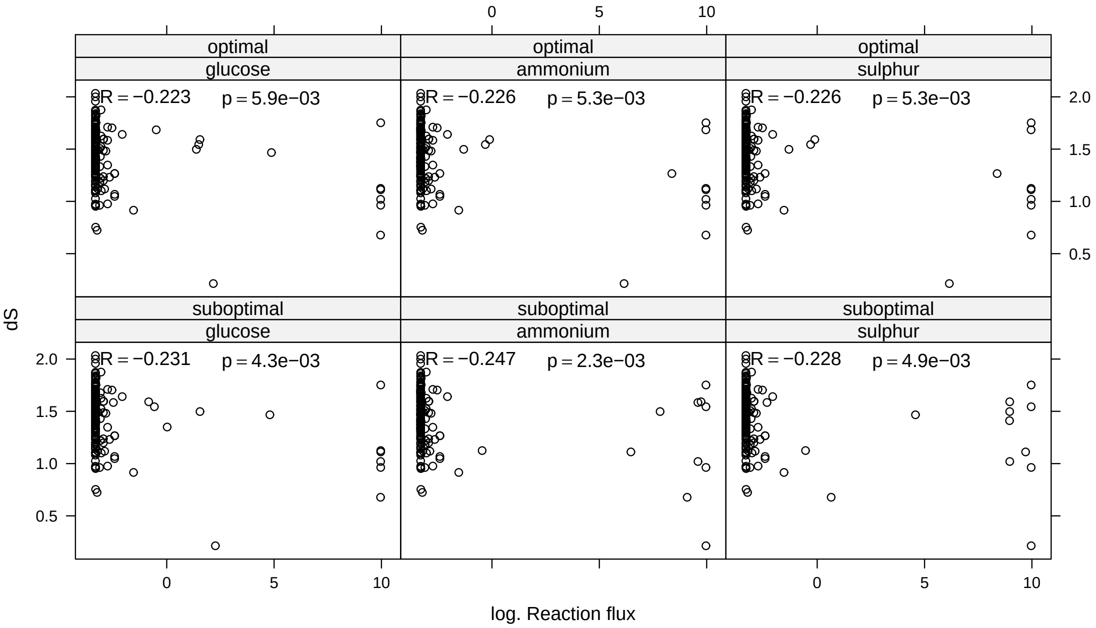


Figure 4.13: Comparison of gene synonymous (dS) evolutionary rate and predicted reaction flux. Reaction flux was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. *Saccharomyces* gene synonymous evolutionary rates were estimated by Wall *et al.* [68]. Spearman's rank correlation coefficients are indicated in each plot.

flux. In each of the considered environments and solutions the correlation coefficient is negative and small. The relationship between reaction flux and synonymous sequence changes is much weaker than that observed for the non-synonymous changes.

In each of the figures it is apparent that the data is skewed towards very small flux values. To test whether the observed correlations are the result of only a difference between zero flux and non-zero flux reactions the variables were re-analysed using Spearman's rank but excluding all zero flux reactions. In the dN/dS and dS comparisons the Spearman R was reduced by values ranging from 0.148 to 0.039 and all but two of the correlations were still significant at $p < 0$. In the dS data the R values dropped by half and were all non-significant. This suggests the estimated dS correlations are influenced by the number of zero-flux reactions.

Overall this result suggests a weak but significant trend for genes encoding high flux reactions to have a slower evolutionary rate. This is observed in dN/dS and dN, but less so for dS. This is consistent with the expected selective pressures on non-synonymous protein sequence changes.

Reaction sensitivity

Finally growth sensitivity to changes in reaction flux was compared with evolutionary rate. The expectation is that more sensitive reactions have a lower evolutionary rate, because these genes are under greater selection pressure against deleterious mutation to maintain fitness.

Figures 4.14, 4.15, and 4.16 compare reaction sensitivity with dN/dS, dN, and dS respectively. Across all environments and solutions there is no significant correlation between evolutionary rate and reaction sensitivity. Spearman's rank analysis further confirms this with small coefficients are small and non-significant p-values. This suggests that reaction sensitivity is a poor predictor of gene evolutionary rate. As in the two previous sections where reaction sensitivity has been compared, the lack of data points effects the degree at which the relationship between the two variables can be accurately estimated.

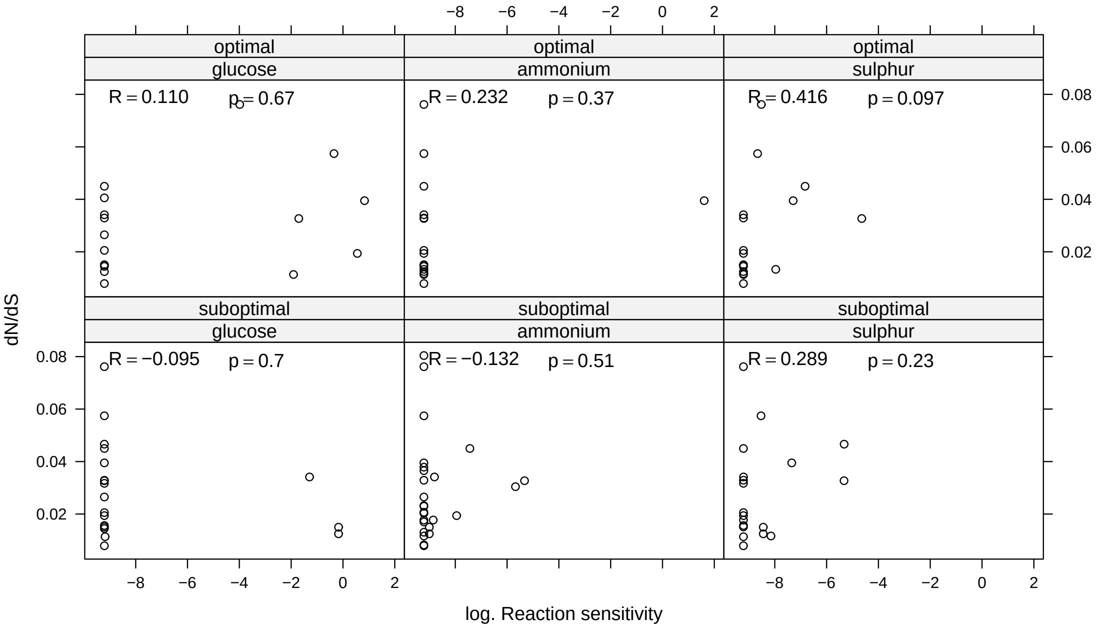


Figure 4.14: Comparison of gene evolutionary rate (dN/dS) and predicted growth sensitivity to reaction flux changes. Reaction sensitivity was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. *Saccharomyces* gene evolutionary rates were estimated by Wall *et al.* [68]. Spearman's rank correlation coefficients are indicated in each plot.

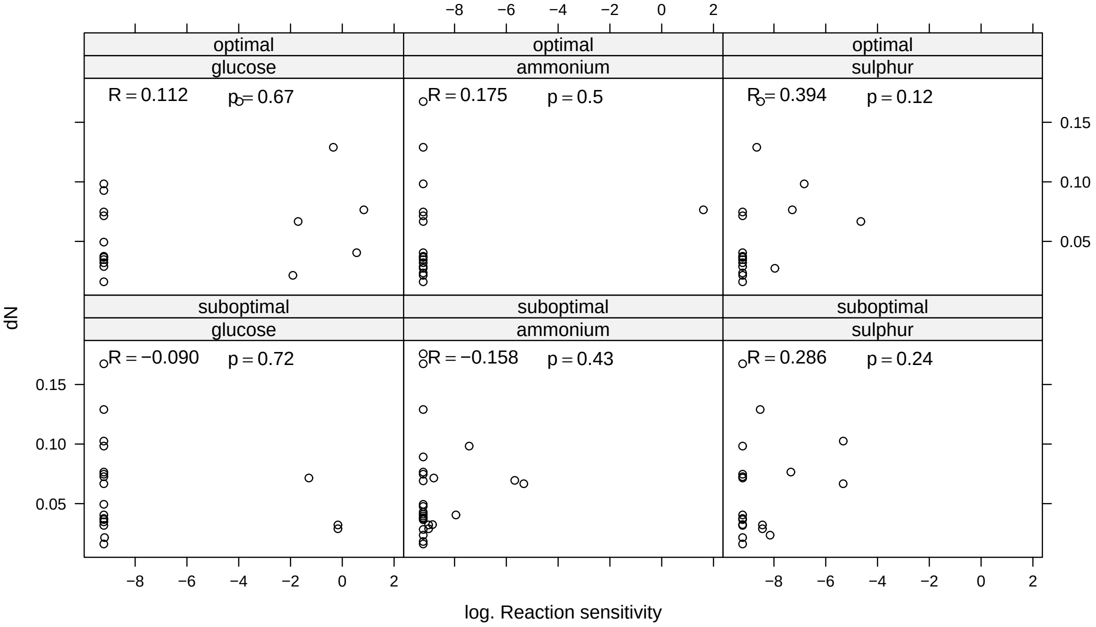


Figure 4.15: Comparison of gene non-synonymous (dN) evolutionary rate and predicted growth sensitivity to reaction flux changes. Reaction sensitivity was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. *Saccharomyces* gene non-synonymous evolutionary rates were estimated by Wall *et al.* [68]. Spearman's rank correlation coefficients are indicated in each plot.

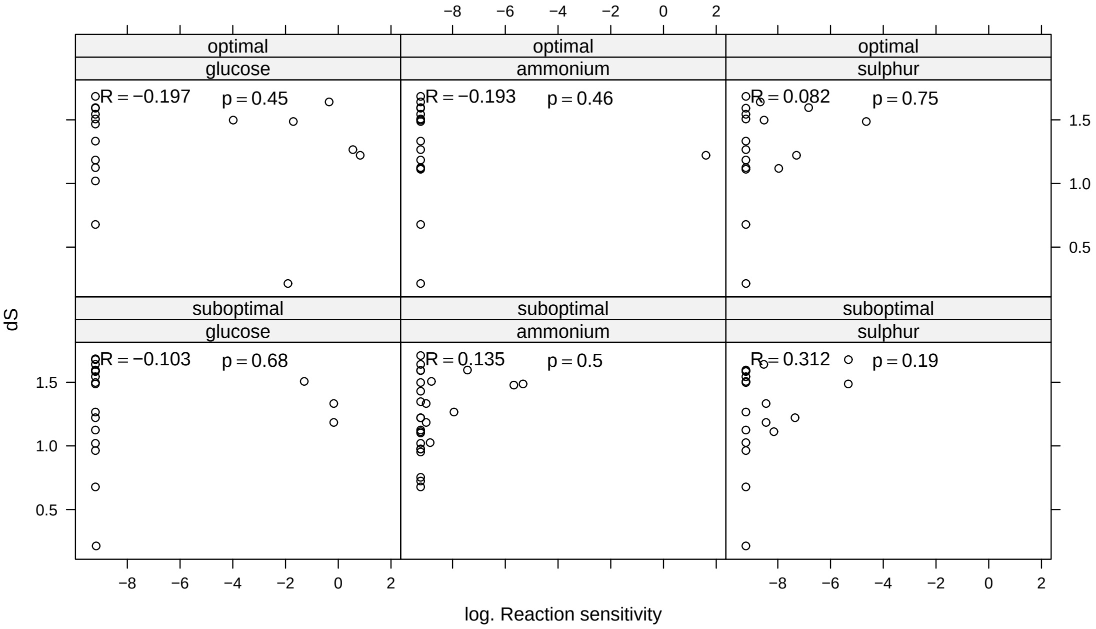


Figure 4.16: Comparison of gene synonymous (dS) evolutionary rate and predicted growth sensitivity to reaction flux changes. Reaction sensitivity was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. *Saccharomyces* gene synonymous evolutionary rates were estimated by Wall *et al.* [68]. Spearman's rank correlation coefficients are indicated in each plot.

4.3.4 Gene importance and codon usage bias

This section compares *in silico* estimates of gene importance with predicted gene expression levels estimated using codon usage bias. As shown in Chapter 3, CAI explains a substantial amount of transcript and protein level variation, and the aim of this section is determine if the estimates of gene importance are correlated with predicted gene expression. Furthermore some trends observed for dS may be interpreted in the context of codon bias and gene expression. The expectation for this analysis is that metabolically important genes may be expressed at greater levels to maintain reaction activity and therefore organism fitness. Codon usage is used over *in vivo* measured transcript or protein expression levels as codon bias reflects of long-term evolutionary pressures, while expression data may be influenced by experimental parameters. The CAI estimated by Coghlan and Wolfe [63] is used as the measure of codon usage bias in *S. cerevisiae* genes.

Reaction constraint

Figure 4.17 compares reaction constraint categories in each of the three environments and two solutions with CAI. In each simulated environment the constrained and variable reactions have a higher CAI than zero flux reactions suggesting the enzyme catalysing these reactions may indeed be more highly expressed. ANOVA regression coefficients in each plot however indicate gene constraint category is a weak predictor of CAI with $R^2 \sim 0.13$.

Reaction flux

Figure 4.18 compares CAI with estimated reaction flux across all environments and both solutions. The Spearman's rank correlation indicated in each plot shows a significant positive correlation in the range of 0.36 to 0.4. As with the previous section this correlation with reaction flux was re-analysed excluding all zero flux reactions. The resulting Spearman R values ranged from 0.26 - 0.392 and all but one correlation was still significant. This indicates a relationship between reaction flux and the optimisation of gene codons for expression.

Reaction sensitivity

Figure 4.19 compares CAI with gene sensitivity across all environments and FBA solutions. Across all environments and solutions there is a poor non-significant

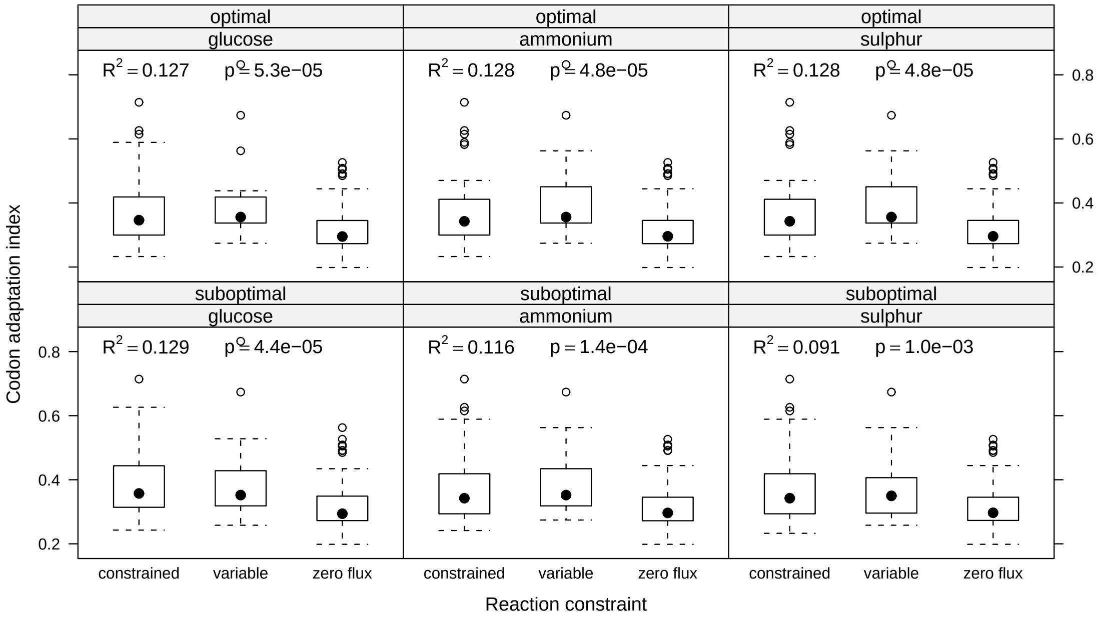


Figure 4.17: Comparison of CAI with predicted reaction flux constraint category. Reaction constraint was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. *Saccharomyces cerevisiae* gene codon adaptation indices were estimated by Coghlan and Wolfe [63]. One factor ANOVA regression coefficients are indicated in each plot.

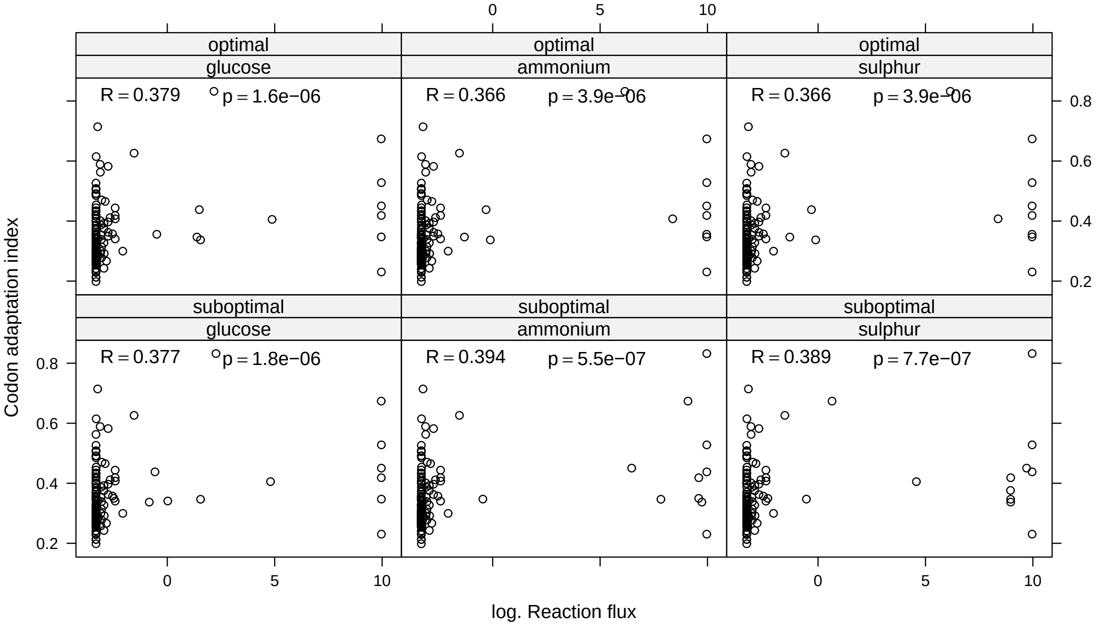


Figure 4.18: Comparison of CAI with predicted reaction flux. Reaction flux was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. *Saccharomyces cerevisiae* gene codon adaptation indices were estimated by Coghlan and Wolfe [63]. Spearman's rank correlation coefficients are indicated in each plot.

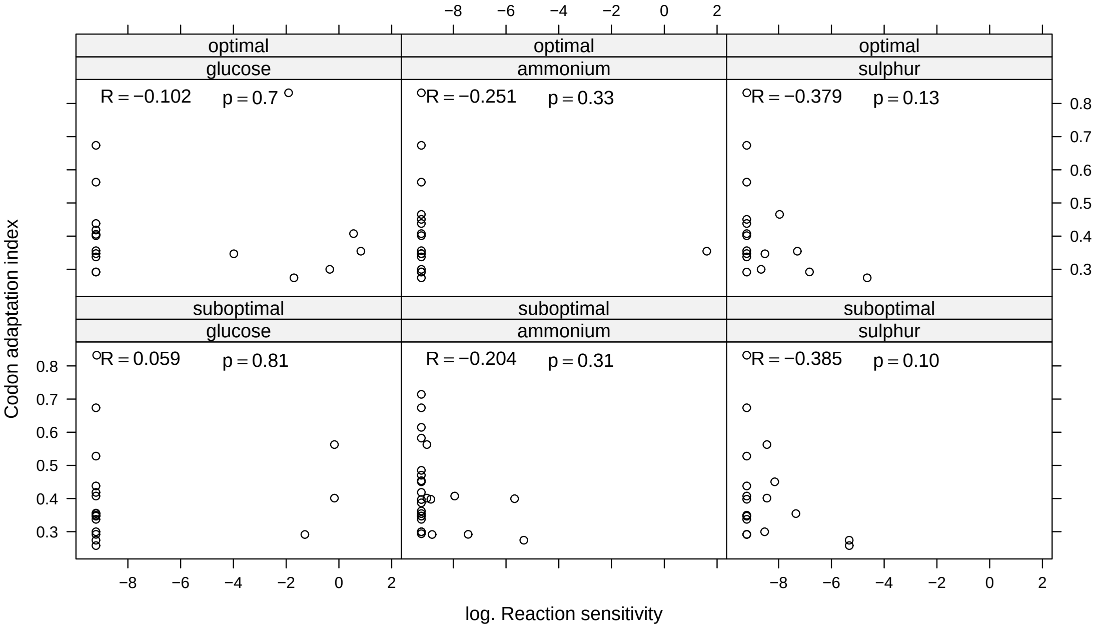


Figure 4.19: Comparison of CAI with predicted reaction flux sensitivity. Reaction flux was estimated using FBA of the *S. cerevisiae iND750* stoichiometric model in simulated glucose, ammonium and sulphur limitation for both optimal and suboptimal solutions. *Saccharomyces cerevisiae* gene codon adaptation indices were estimated by Coghlan and Wolfe [63]. Spearman's rank correlation coefficients are indicated in each plot.

correlation between gene sensitivity and CAI. This may result from a lack of data, as well as no real relationship between the two variables.

4.4 Discussion

The aim of this chapter was to determine if estimates of gene importance produced using a genome scale model could predict *in vivo* gene use and evolution in *S. cerevisiae*. Previous *in silico* estimates of gene importance have focused on complete gene knockout, however the estimates studied here have attempted to characterise the importance of gene based on the activity of the encoded reaction. The analyses presented here have examined the relevance of these measures by comparing them with gene dispensability, hemizygous growth fitness, evolutionary rate, and expression levels predicted through codon usage bias.

4.4.1 Poor predictors of gene dispensability

Gene dispensability was compared with reaction importance where the hypothesis was that highly constrained reactions, reactions with a large flux, or a large sensitivity would correspond to *in vivo* essential genes. However, when compared with gene dispensability the reaction flux and reaction sensitivity estimates were poor predictors. In each case ANOVA R^2 values were close to zero and all analyses were non-significant. This indicates these two measures are poor predictors of gene dispensability.

When comparing reaction constraint categories there was evidence that constrained reactions were more likely to be essential than would be expected. This is illustrated by the greater χ^2 residual deviance in Figure 4.2. In contrast however zero flux reactions were not more likely to be non-essential, as might be expected. This indicates that, overall, reaction constraint categories may be poor predictors of gene dispensability.

Previous estimates of gene dispensability using FBA have correctly predicted gene dispensability in $\sim$90% of 600 *in vivo* measured knockouts [26, 41]. There are two possible reasons why the analyses here show a much lower correspondence. First the *in vivo* mutants were tested on rich media while the *in silico* reaction importance measures here were estimated for glucose, ammonium or sulphur limitation - environments related to specific nutrient use. Second, data compared in this analysis were the one-to-one gene associated reactions while previous analyses have considered all types of gene associations. Therefore it may be possible if this analysis were repeated for all model reactions on a rich media there would be a greater correspondence between the two variables.

4.4.2 No relationship with hemizygous mutant growth effect

The *in silico* estimates of gene importance from Chapter 2 were compared with *in vivo* hemizygous growth difference with the expectation of a positive relationship. The result of the analyses was that each of the three *in silico* measures of gene importance showed little value as a predictor of hemizygous growth effect (Figures 4.5, 4.6, and 4.6). In all cases neither reaction flux or reaction sensitivity showed any correlation with *in vivo* fitness nor did gene constraint category show any difference in distribution of hemizygous growth rate. These results demonstrate that the *in vivo* hemizygous growth effects observed in the Delneri *et al.* [87] data cannot be predicted using the FBA estimates of gene importance developed here.

One reason for the lack of relationship may be that the two variables are comparable on a gene-by-gene basis, but examine different levels of the relationship between the genomic encoding and downstream effects on metabolism. The *in vivo* analysis removes one copy of a gene and determines the effect on organism growth while the *in silico* analysis examined the effect of reaction flux and changes in reaction flux on organism growth. Comparing these two variables assumes a linearity between a reduction in gene dosage and a corresponding effect on the encoded reaction. The relationship between gene copy and reaction effect may however be non-linear, since transcript levels only explain 40% - 70% of the variance in protein levels [60] and reaction activity is a non-linear function of both substrate and enzyme concentration [8].

4.4.3 Small association with gene evolutionary rate

The FBA-estimated measures of gene importance were also compared with three estimates of gene evolutionary rate: dN/dS, dN, dS. The aim of analysis was to determine if observed rates of evolution could be predicted using an *in silico* model. Gene expression levels have been shown to correlate with evolutionary rate [97], and if evolutionary rate is also correlated with reaction flux. This would represent an advance in understanding the selection pressures involved in protein sequence evolution.

The results of this analysis showed a limited correlation with reaction constraint categories (Figures 4.8, 4.9 and 4.10) and reaction flux (Figures 4.11, 4.12

and 4.13) but no correlation with reaction sensitivity (Figures 4.14, 4.15 and 4.16). The small correlations observed in the reaction constraint and flux data may indicate a marginal relationship between reaction activity and the corresponding evolutionary rate of a gene.

The larger of the observed correlations were for non-synonymous changes, with synonymous evolutionary rate showing reduced correlation. The difference in correlations with gene importance between the synonymous and non-synonymous evolutionary rates might be expected as synonymous changes should have little effect on protein structure and function and therefore reduced effects on fitness. Non-synonymous changes, however, would be expected to affect protein structure and activity, which will have a greater fitness effect the more important the encoded reactions.

It is worth noting again that only single gene associated reactions were considered, which may bias the evolutionary rates compared. If a reaction is catalysed by multiple paralogs it could be theorised that there is less selection pressure on each of the encoding paralogs given the enzymatic redundancy. Therefore, the weak trend between reaction flux and evolutionary rate may be conditional on that only single gene associated reactions were considered. Further analysis could consider multiple and single gene associated reactions in two separate analyses to see if there is a difference in the relationship between gene importance and evolutionary rate for genes with or without duplicates.

The evolutionary rates used in this analysis were estimated from four *Saccharomyces* species. Therefore observed evolutionary rate are mutations that have become fixed between species as the result of long periods of divergence and selective pressure. The resequencing of the *S. cerevisiae* genome allows the estimation of intraspecific variation evolution and therefore the detection of more recent mutations that may not yet have been purged by selection. A stronger relationship between evolutionary rate and gene importance may therefore be detected in intraspecific variation because of non-synonymous mutations that would never be observed in between-species evolutionary analysis.

4.4.4 Weak relationship with codon usage bias

Predicted gene expression measured using CAI was compared with the three estimates of gene importance. As mentioned in the results section, CAI was chosen for the analysis as expression data is subject to experimental parameters, while CAI is invariant and the result of long term evolutionary pressures across multiple environments.

The hypothesis for this analysis was a positive trend between CAI and gene importance, where the enzymes catalysing important reactions will be expressed at higher levels to maintain critical reaction activity. The results of this analysis showed reaction constraint categories (Figure 4.17) and reaction flux (Figure 4.18) explained similar levels of variance in gene CAI (Spearman $R = 0.36$; ANOVA $R^2 = 0.13$). This suggests a weak relationship between the optimisation of the gene for expression the flux maintained at the reaction. There was, however, no correlation between gene sensitivity and CAI, which may be due to the small number of data points for this analysis.

Codon usage bias and synonymous evolutionary rate

Codon usage bias reflects the use of optimal codons for faster translation where the use of specific codons allows for faster protein synthesis. Synonymous evolutionary rate estimates nucleotide changes resulting in the same encoded amino acid. Synonymous nucleotide changes may be used as a proxy for mutation rate because a synonymous codon change will not effect protein sequence and therefore should not be affected by selection. However, since codon bias is shown to have a strong trend with gene expression [60, 63] it might be anticipated that synonymous mutations may not be neutral in evolution. In the examined data a weak trend was observed for CAI with reaction category or reaction flux, as was also a weaker trend for dS evolutionary rate. This may indicate a possible trend between these two variables and further investigation may reveal the degree to which synonymous mutations are under selection to maintain codon optimisation.

4.4.5 Using FBA estimated measures of gene importance

The analyses presented in this chapter show that the gene importance estimates developed in this thesis explain at best only a limited amount of the variance in the four examined measures. A possible reason for the disparity is that the quantitative gene importance estimates are highly skewed where there are a few large values, and many small values. In particular reaction flux estimates at ±1000 mmol^{-1} gDW^{-1} hr^{-1} are constrained to this value by the boundaries on the FBA solution. Without these constraints on minimum and maximum flux it is conceivable that these large fluxes would be even higher. Since some flux values are unrealistically large this may indicate a divergence in the *in silico* predictions compared with what may be observed *in vivo*.

A second technical issue is that different approaches to FBA model optimisation, whether using an optimal or suboptimal solution, can generate large differences in gene importance estimation, as indicated by differences between the optimal and suboptimal solutions. This is both apparent in the gene constraint categories for the number of reactions whose flux could not be reduced, and also in the gene sensitivity estimates where certain reactions are highly sensitive and small flux reductions result in large fitness effects. *In vivo* it may be expected that reactions have a degree of robustness and that perturbations in reaction flux would not have such drastic fitness effects as would be suggested by different approaches to model optimisation.

Given these technical points, this leads to the question as to whether the small correlation observed in the analyses is due to an actual lack of biological relationship, or because the methods used are not yet sufficiently accurate to estimate gene importance. The approaches presented in this thesis do, however, represent an attempt to move genome scale model approaches beyond simulating simple gene activity knockouts. The aim of this investigation was to predict *in silico* phenotypes in other ways than just gene loss, which may be expected to be a rarer event in evolution when compared to small sequence mutations. Hopefully increasing development of genome scale models and simulation using FBA or related approaches will extend the initial approaches outlined here and become more accurate and useful in the future.

5 Discussion

5.1 Summary of findings

The aim of this PhD has been to use systems biology methods to understand and predict gene expression and evolution in *Saccharomyces cerevisiae*. As the field of systems biology progresses and becomes more able to simulate *in vivo* behaviour these *in silico* predictions will help understand how organism physiology and metabolism shapes the evolution and regulation of the genome.

The systems biology model used in this PhD is a genome scale stoichiometric model of *S. cerevisiae* metabolism comprising 750 reactions. The advantages of a large scale model is that metabolism can be simulated as a whole, including many of the metabolites required for cell growth, and can be used to understand selective pressures at an organism level.

The first application of the *S. cerevisiae* model was to predict amino acid biosynthetic cost, where previous estimations have relied on manual curation of metabolic maps. A systems biology approach allows predictions to be made rapidly and reproducibly and allows the analysis to be performed in any simulated environment. The absolute glucose cost of amino acid synthesis was shown to be correlated with previous measures of amino acid cost which validates the systems biology approach.

The second application of the *S. cerevisiae* model was to predict gene importance in relation to growth. Previously applications of genome scale models have predicted gene importance as a function of whether the gene is essential for growth (i.e. gene knockout). In this PhD gene importance was however examined using alternative approaches including whether reaction flux is constrained at a minimum value, reaction flux, and the effect of changes in reaction flux on nutrient uptake. This work represents a novel approach to examine reactions in genome scale models and allows insight into smaller scale genetic perturbations than full gene knockouts. During the estimation of gene importance, the flux balance analysis optimisation process was shown to have a significant effect on estimated gene importance as indicated by the variability in measures produced in optimal and suboptimal solutions.

Chapters 3 and 4 examined how both these systems biology estimates may be used to understand gene use and evolution in *S. cerevisiae*. The role of amino acid biosynthetic cost in gene expression was examined and shown to have limited effect. Instead translational optimisation was the single greatest factor in predicting transcript and protein levels. This suggests that fast and accurate gene expression may be a greater selective pressure than minimising biosynthetic cost. Amino acid cost was also compared with observed amino acid usage, conservation, substitution, and deletion rates in multiple *Saccharomyces* species. This yielded the most significant result of this thesis where amino acid cost appears to correlated with amino acid fixation. The result of this analysis suggest that less constrained protein residues evolve according the biosynthetic cost of the amino acid.

The systems biology estimated measures of gene importance were compared with four characteristics of *S. cerevisiae* gene use and evolution. The reaction constraint category and reaction flux value showed small and weak correlations with gene dispensability, non-synonymous evolution and codon usage bias. The gene sensitivity estimates showed no correlation with any of the compared variables. The yeast hemizygous fitness data also showed no relationship with any of the gene importance estimates. These limited results may indicate that flux balance analysis may not be suitable for predicting quantitative reaction activity and therefore the selective constraints on the encoding gene. Further developments in genome scale models and flux balance analysis may yet however make these types of estimates possible in the future.

5.2 Preliminary future analysis

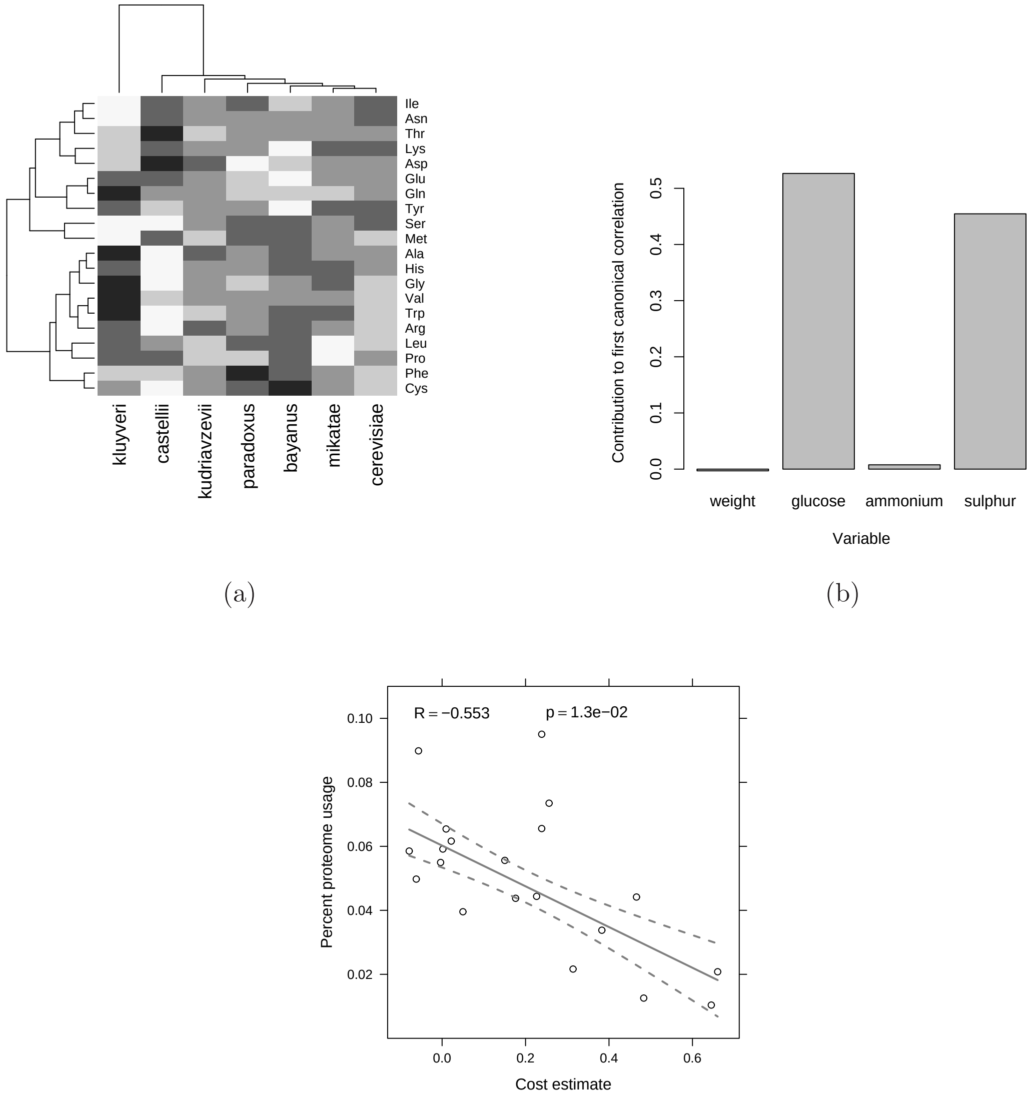
This final section explores a potential avenue of research following from the work described in the previous chapters. In previous chapters amino acid cost has been estimated according to the requirement of three nutrients: glucose, ammonium and sulphur. The results in Chapter 3 showed the usefulness of these cost estimates where the absolute glucose cost is proportional to amino acid substitution rates. In contrast however it may be argued that the biosynthesis of amino acid is a combination of all the required atoms and energy rather than individual nutrients. For example synthesising a cysteine residue will require energy to catalyse enzymatic steps as well as the required sulphur and nitrogen atoms. To explore this possibility, a preliminary analysis was performed to determine if a combination of estimates including absolute glucose, ammonium and sulphur costs, and molecular weight explains more variation that using single estimates alone.

The relative usage of each amino acid across each of the seven *Saccharomyces* species is shown as a heatmap in Figure 5.1a where each row represents the usage of the amino acid and each column represents a species. To determine the extent at which all four estimates of amino acid cost together explain variation in amino acid usage, this usage matrix was compared with a matrix of the four cost measures using canonical correlation.

Canonical correlation determines the correlation in unobserved variables between two matrices, where unobserved variables are derived from the weighted combination of the existing matrix variables. The first canonical correlation between the amino acid cost and usage matrices explained 90% of variation in amino acid usage. Figure 5.1b highlights the contribution of each of the four amino acid cost estimates to this first canonical correlation. This figure suggest that $A_{glucose}$ and $A_{sulphur}$ content explain the majority of the variation in amino acid usage and that the contributions of molecular weight and ammonium are minimal.

As the canonical correlation indicates that the cost estimates explain a large degree of variation in amino acid usage, a combined cost derived from these four estimates may be a more useful metric than any single cost estimate. The four amino acid cost estimates were therefore combined into a single measure through matrix multiplication of the weights indicated in Figure 5.1b. Figure 5.1c compares this derived cost estimate of with percent amino acid usage in the *S. cerevisiae* genome.

The figure highlights that the combined cost measure shows a stronger correlation with amino acid usage using than $A_{glucose}$ or molecular weight alone (compare with Figure 3.2 on page 77). This illustrates that a combined metric of the different factors involved in amino acid synthesis may be a useful approach to explore. The different weighting of nutrients in a combined cost may also be useful in



(a)

(b)

(c)

Figure 5.1: Exploring a combined estimate of amino acid cost. Figure (a) shows a heatmap of amino acid usage across seven *Saccharomyces* species where each row is scaled by the mean. Darker grey indicates a greater amino acid usage in that species relative to the other species. Dendrograms are added to each axis to illustrate the Spearman's rank correlation in usage between species or individual amino acids. Figure (b) illustrates the contributions of each of the amino acid estimates to the first canonical correlation between amino acid usage and cost. Figure (c) compares the combination of amino acid cost estimates with amino acid usage in the *S. cerevisiae* genome. Robust linear regression and 95% confidence intervals are used to indicated trend. Spearman's rank correlation is indicated.

understanding the type of environment *S. cerevisiae* has evolved in, where in the above example glucose and sulphur are the nutrients associated with changes in amino acid usage. Further analysis of diverse species may indicate, as might be expected, that different nutrients are more costly in different environments.

Bibliography

[1] Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgård MJ: **Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox.** *Nature Protocols* 2007, **2**(3):727–738.

[2] Reed JL, Vo TD, Schilling CH, Palsson BØ: **An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR).** *Genome Biology* 2003, **4**(9):R54+.

[3] Duarte NC, Herrgård MJ, Palsson BØ: **Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model.** *Genome Research* 2004, **14**(7):1298–1309.

[4] Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK, Botstein D: **Genetic and physical maps of *Saccharomyces cerevisiae*.** *Nature* 1997, **387**(6632 Suppl):67–73.

[5] Niederberger P, Prasad R, Miozzari G, Kacser H: **A strategy for increasing an in vivo flux by genetic manipulations. The tryptophan system of yeast.** *The Biochemical Journal* 1992, **287 (Pt 2)**:473–479.

[6] Fell DA: **Metabolic control analysis: a survey of its theoretical and experimental development.** *The Biochemical Journal* 1992, **286 (Pt 2)**:313–330.

[7] Fell D: *Understanding the Control of Metabolism.* Ashgate Publishing 1996.

[8] Rossell S, van der Weijden CC, Lindenbergh A, van Tuijl A, Francke C, Bakker BM, Westerhoff HV: **Unraveling the complexity of flux regulation: A new method demonstrated for nutrient starvation in *Saccharomyces cerevisiae***. *Proceedings of the National Academy of Sciences* 2006, **103**(7):2166–2171.

[9] Daran-Lapujade P, Rossell S, van Gulik WM, Luttik MA, de Groot MJ, Slijper M, Heck AJ, Daran JM, de Winde JH, Westerhoff HV, Pronk JT, Bakker BM: **The fluxes through glycolytic enzymes in *Saccharomyces cerevisiae* are predominantly regulated at posttranscriptional levels.**

Proceedings of the National Academy of Sciences 2007, **104**(40):15753–15758.

[10] Lehar J, Krueger A, Zimmermann G, Borisy A: **High-order combination effects and biological robustness**. *Molecular Systems Biology* 2008, **4**:215+.

[11] Hopkins AL: **Network pharmacology: the next paradigm in drug discovery**. *Nature Chemical Biology* 2008, **4**(11):682–690.

[12] Hornberg JJ, Bruggeman FJ, Bakker BM, Westerhoff HV: **Metabolic control analysis to identify optimal drug targets.** *Progress in Drug Research. Fortschritte der Arzneimittelforschung. Progrès des Recherches Pharmaceutiques* 2007, **64**.

[13] Teusink B, Passarge J, Reijenga CA, Esgalhado E, van der Weijden CC, Schepper M, Walsh MC, Bakker BM, van Dam K, Westerhoff HV, Snoep JL: **Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry.** *European Journal of Biochemistry* 2000, **267**(17):5313–5329.

[14] Steuer R: **Computational approaches to the topology, stability and dynamics of metabolic networks.** *Phytochemistry* 2007, **68**(16-18):2139–2151.

[15] Feist AMM, Herrgård MJJ, Thiele I, Reed JLL, Palsson BØ: **Reconstruction of biochemical networks in microorganisms.** *Nature Reviews Microbiology* 2008, :129–143.

[16] Chang A, Scheer M, Grote A, Schomburg I, Schomburg D: **BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009**. *Nucleic Acids Research* 2008, :D588–D592.

[17] Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M: **KEGG Atlas mapping for global analysis of metabolic pathways.** *Nucleic Acids Research* 2008, **36**(Web Server issue):423–426.

[18] Sun J, Sayyar B, Butler JE, Pharkya P, Fahland TR, Famili I, Schilling CH, Lovley DR, Mahadevan R: **Genome-scale constraint-based modeling of *Geobacter metallireducens*.** *BMC Systems Biology* 2009, **3**.

[19] Feist AM, Palsson BØ: **The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli***. *Nature Biotechnology* 2008, **26**(6):659–667.

[20] Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Molecular Systems Biology* 2007, **3**:121+.

[21] Duarte NCC, Becker SAA, Jamshidi N, Thiele I, Mo MLL, Vo TDD, Srivas R, Palsson BØ: **Global reconstruction of the human metabolic network based on genomic and bibliomic data.** *Proceedings of the National Academy of Sciences* 2007, **104**(6):1777–1782.

[22] Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Buthgen N, Borger S, Costenoble R, Heinemann M, Hucka M, Le Novere N, Li P, Liebermeister W, Mo ML, Oliveira AP, Petranovic D, Pettifer S, Simeonidis E, Smallbone K, Spasie I, Weichart D, Brent R, Broomhead DS, Westerhoff HV, Kurdar B, Penttila M, Klipp E, Palsson BØ, Sauer U, Oliver SG, Mendes P, Nielsen J, Kell DB: **A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology**. *Nature Biotechnology* 2008, **26**(10):1155–1160.

[23] Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novère N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.** *Bioinformatics* 2003, **19**(4):524–531.

[24] Le Novère N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, Nielsen P, Sauro H, Shapiro B, Snoep JL, Spence HD, Wanner BL: **Minimum information requested**

in the annotation of biochemical models (MIRIAM). *Nature Biotechnology* 2005, **23**(12):1509–1515.

[25] Coles SJ, Day NE, Murray-Rust P, Rzepa HS, Zhang Y: **Enhancement of the chemical semantic web through the use of InChI identifiers**. *Organic and Biomolecular Chemistry* 2005, **3**(10):1832–1834.

[26] Famili I, Förster J, Nielsen J, Palsson BØ: ***Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network.** *Proceedings of the National Academy of Sciences* 2003, **100**(23):13134–13139.

[27] Segrè D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks.** *Proceedings of the National Academy of Sciences* 2002, **99**(23):15112–15117.

[28] Burgard AP, Pharkya P, Maranas CD: **Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization**. *Biotechnology and Bioengineering* 2003, **84**(6):647–657.

[29] Shlomi T, Berkman O, Ruppin E: **Regulatory on/off minimization of metabolic flux changes after genetic perturbations.** *Proceedings of the National Academy of Sciences* 2005, **102**(21):7695–7700.

[30] Mahadevan R, Schilling CH: **The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.** *Metabolic Engineering* 2003, **5**(4):264–276.

[31] Varma A, Boesch BW, Palsson BØ: **Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates.** *Applied Environmental Microbiology* 1993, **59**(8):2465–2473.

[32] Wagner A: **Energy constraints on the evolution of gene expression**. *Molecular Biology and Evolution* 2005, **22**(6):1365–1374.

[33] Craig CL, Weber RS: **Selection costs of amino acid substitutions in ColE1 and ColIa gene clusters harbored by *Escherichia coli*.** *Molecular Biology and Evolution* 1998, **15**(6):774–776.

[34] Akashi H, Gojobori T: **Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis.*** *Proceedings of the National Academy of Sciences* 2002, **99**(6):3695–3700.

[35] Heizer EM, Raiford DW, Raymer ML, Doom TE, Miller RV, Krane DE: **Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis**. *Molecular Biology and Evolution* 2006, **23**(9):1670–1680.

[36] Seligmann H: **Cost-Minimization of Amino Acid Usage**. *Journal of Molecular Evolution* 2003, **56**(2):151–161.

[37] Pál C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD: **Chance and necessity in the evolution of minimal metabolic networks**. *Nature* 2006, **440**(7084):667–670.

[38] Becker S, Palsson BØ: **Three factors underlying incorrect in silico predictions of essential metabolic genes**. *BMC Systems Biology* 2008, **2**.

[39] Papp B, Pál C, Hurst LD: **Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast.** *Nature* 2004, **429**(6992):661–664.

[40] Wang Z, Zhang J: **Why is the correlation between gene importance and gene evolutionary rate so weak?** *PLoS Genetics* 2009, **5**:e1000329+.

[41] Förster J, Famili I, Palsson BØ, Nielsen J: **Large-scale evaluation of in silico gene deletions in *Saccharomyces cerevisiae.*** *OMICS* 2003, **7**(2):193–202.

[42] Harrison R, Papp B, Pál C, Oliver SG, Delneri D: **Plasticity of genetic interactions in metabolic networks of yeast**. *Proceedings of the National Academy of Sciences* 2007, **104**(7):2307–2312.

[43] Deutscher D, Meilijson I, Schuster S, Ruppin E: **Can single knockouts accurately single out gene functions?** *BMC Systems Biology* 2008, **2**.

[44] Berkelaar M, Eikland K, Notebaert P: *lpsolve : Open source (Mixed-Integer) Linear Programming system.*

[45] Hartigan JA, Wong MA: **Algorithm AS 136: A k-means clustering algorithm**. *Applied Statistics* 1979, **28**:100–108.

[46] Pál C, Papp B, Lercher MJ: **An integrated view of protein evolution.** *Nature reviews. Genetics* 2006, **7**(5):337–348.

[47] Tekaia F, Yeramian E: **Evolution of proteomes: Fundamental signatures and global trends in amino acid compositions**. *BMC Genomics* 2006, **7**:307+.

[48] Mazel D, Marlière P: **Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins.** *Nature* 1989, **341**(6239):245–248.

[49] Baudouin-Cornu P, Surdin-Kerjan Y, Marliere P, Thomas D: **Molecular evolution of protein atomic composition**. *Science* 2001, **293**(5528):297–300.

[50] Bragg JG, Thomas D, Baudouin-Cornu P: **Variation among species in proteomic sulphur content is related to environmental conditions.** *Proceedings of the Royal Society B* 2006, **273**(1591):1293–1300.

[51] Rispe C, Delmotte F, van Ham RCHJ, Moya A: **Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids**. *Genome Research* 2004, **14**:44–53.

[52] Goodarzi H, Torabi N, Najafabadi HS, Archetti M: **Amino acid and codon usage profiles: adaptive changes in the frequency of amino acids and codons.** *Gene* 2008, **407**(1-2):30–41.

[53] Swire J: **Selection on synthesis cost affects interprotein amino acid usage in all three domains of life**. *Journal of Molecular Evolution* 2007, **64**(5):558–571.

[54] Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic I, Yamada T, Paccanaro A, Jensen LJ, Snyder M, Bork P, Gerstein MB: **Quantifying environmental adaptation of metabolic pathways in metagenomics**. *Proceedings of the National Academy of Sciences* 2009, **106**(5):1374–1379.

[55] Kanaya S, Yamada Y, Kudo Y, Ikemura T: **Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis.** *Gene* 1999, **238**:143–155.

[56] Raiford DW, Heizer EM, Miller RV, Akashi H, Raymer ML, Krane DE: **Do amino acid biosynthetic costs constrain protein evolution in *Saccharomyces cerevisiae*?** *Journal of Molecular Evolution* 2008, :621–630.

[57] Bragg JG, Wagner A: **Protein carbon content evolves in response to carbon availability and may influence the fate of duplicated genes.** *Proceedings of the Royal Society B* 2007, **274**(1613):1063–1070.

[58] Sanchez-Perez G, Mira A, Nyiro G, Pasic L, Rodriguez-Valera F: **Adapting to environmental changes using specialized paralogs**. *Trends in Genetics* 2008, **24**(4):154–158.

[59] Rocha EP, Danchin A: **An analysis of determinants of amino acids substitution rates in bacterial proteins.** *Molecular Biology and Evolution* 2004, **21**:108–116.

[60] Castrillo JI, Zeef LA, Hoyle DC, Zhang N, Hayes A, Gardner DC, Cornell MJ, Petty J, Hakes L, Wardleworth L, Rash B, Brown M, Dunn WB, Broadhurst D, O'Donoghue K, Hester SS, Dunkley TP, Hart SR, Swainston N, Li P, Gaskell SJ, Paton NW, Lilley KS, Kell DB, Oliver SG: **Growth control of the eukaryote cell: a systems biology study in yeast.** *Journal of Biology* 2007, **6**(2):4+.

[61] Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185–193.

[62] Ghaemmaghami S, Huh WKK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS: **Global analysis of protein expression in yeast.** *Nature* 2003, **425**(6959):737–741.

[63] Coghlan A, Wolfe KH: **Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*.** *Yeast* 2000, **16**(12):1131–1145.

[64] Akashi H: **Translational selection and yeast proteome evolution.** *Genetics* 2003, **164**(4):1291–1303.

[65] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[66] Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423**(6937):241–254.

[67] Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: **Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting.** *Science* 2003, **301**(5629):71–76.

[68] Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW: **Functional genomic analysis of the rates of protein evolution.** *Proceedings of the National Academy of Sciences* 2005, **102**(15):5483–5488.

[69] Mitsuteru NG, Goto N, Nakao MC, Kawashima S, Katayama T, Kanehisa M: **BioRuby: open-source bioinformatics library** 2003.

[70] Yang Z: **PAML 4: Phylogenetic analysis by maximum likelihood**. *Molecular Biology and Evolution* 2007, **24**(8):1586–1591.

[71] Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.** *Molecular Biology and Evolution* 2001, **18**(5):691–699.

[72] Ikemura T: **Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs.** *Journal of Molecular Biology* 1982, **158**(4):573–597.

[73] Akaike H: **A new look at the statistical model identification**. *Automatic Control, IEEE Transactions on* 1974, **19**(6):716–723.

[74] Jansen R, Bussemaker HJ, Gerstein M: **Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models.** *Nucleic Acids Research* 2003, **31**(8):2242–2251.

[75] Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappé MS, Short JM, Carrington JC, Mathur EJ: **Genome streamlining in a cosmopolitan oceanic bacterium.** *Science* 2005, **309**(5738):1242–1245.

[76] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, Mcdade KE, Mckenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**(7057):376–380.

[77] Shendure J, Ji H: **Next-generation DNA sequencing**. *Nature Biotechnology* 2008, **26**(10):1135–1145.

[78] Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis S, Guigo R: **Identifying protein-coding genes in genomic sequences**. *Genome Biology* 2009, **10**:201+.

[79] Juretic N, Bureau TE, Bruskiewich RM: **Transposable element annotation of the rice genome**. *Bioinformatics* 2004, **20**(2):155–160.

[80] Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 genes.** *Science* 1996, **274**(5287).

[81] Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, Tsai IJ, Bergman CM, Bensasson D, O'Kelly MJ, van Oudenaarden A, Barton DB, Bailes E, Nguyen AN, Jones M, Quail MA, Goodhead I, Sims S, Smith F, Blomberg A, Durbin R, Louis EJ: **Population genomics of domestic and wild yeasts.** *Nature* 2009, **458**(7236):337–341.

[82] Friedberg I: **Automated protein function prediction–the genomic challenge.** *Briefings in Bioinformatics* 2006, **7**(3):225–242.

[83] Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW: **Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy.** *Nature Genetics* 1996, **14**(4):450–456.

[84] Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Güldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kötter P, Labonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M: **Functional profiling of the *Saccharomyces cerevisiae* genome.** *Nature* 2002, **418**(6896):387–391.

[85] Dudley AM, Janse DM, Tanay A, Shamir R, Church GM: **A global view of pleiotropy and phenotypically derived gene function in yeast**. *Molecular Systems Biology* 2005, **1**.

[86] Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, Nislow C, Giaever G: **Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast.** *Genetics* 2005, **169**(4):1915–1925.

[87] Delneri D, Hoyle DC, Gkargkas K, Cross EJM, Rash B, Zeef L, Leong HS, Davey HM, Hayes A, Kell DB, Griffith GW, Oliver SG: **Identification and characterization of high-flux-control genes of yeast through competition analyses in continuous cultures.** *Nature Genetics* 2008, **40**:113–117.

[88] Giaever G, Shoemaker DD, Jones TW, Liang H, Winzeler EA, Astromoff A, Davis RW: **Genomic profiling of drug sensitivities via induced haploinsufficiency**. *Nature Genetics* 1999, **21**(3):278–283.

[89] Baetz K, Mchardy L, Gable K, Tarling T, Rebérioux D, Bryan J, Andersen RJ, Dunn T, Hieter P, Roberge M: **Yeast genome-wide drug-induced haploinsufficiency screen to determine drug mode of action.** *Proceedings of the National Academy of Sciences* 2004, **101**(13):4525–4530.

[90] Hughes TR: **Yeast and drug discovery.** *Functional & Integrative Genomics* 2002, **2**(4-5):199–211.

[91] Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K, Oliver SG: **A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations.** *Nature Biotechnology* 2001, **19**:45–50.

[92] Xu Z, Sun X, Yu S: **Genome-scale analysis to the impact of gene deletion on the metabolism of *E. coli*: constraint-based simulation approach.** *BMC Bioinformatics* 2009, **10 Suppl 1**.

[93] Notebaart RA, Teusink B, Siezen RJ, Papp B: **Co-regulation of metabolic genes is better explained by flux coupling than by network distance**. *PLoS Computational Biology* 2008, **4**:e26+.

[94] Loewe L: **A framework for evolutionary systems biology**. *BMC Systems Biology* 2009, **3**:27+.

[95] Hong EL, Balakrishnan R, Dong Q, Christie KR, Park J, Binkley G, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Krieger CJ, Livstone MS, Miyasato SR, Nash RS, Oughtred R, Skrzypek MS, Weng S, Wong ED, Zhu KK, Dolinski K, Botstein D, Cherry JM: **Gene**

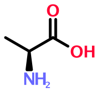
Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Research* 2007, **36**(Database):D577–D581.

[96] Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, **57**:289–300.

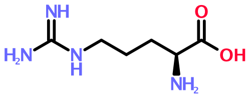
[97] Pál C, Papp B, Hurst LD: **Highly expressed genes in yeast evolve slowly.** *Genetics* 2001, **158**(2):927–931.

Appendices

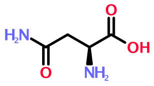
A
Amino Acids



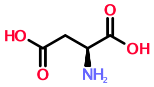
Alanine
(Ala / A)



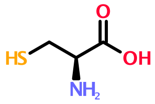
Arginine
(Arg / R)



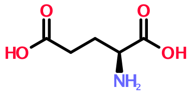
Asparagine
(Asn / N)



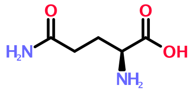
Aspartic Acid
(Asp / D)



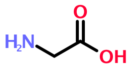
Cysteine
(Cys / C)



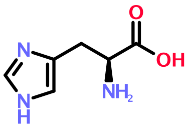
Glutamic Acid
(Glu / E)



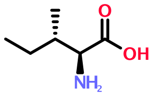
Glutamine
(Gln / Q)



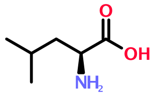
Glycine
(Gly / G)



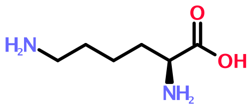
Histidine
(His / H)



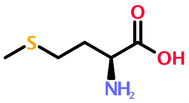
Isoleucine
(Ile / I)



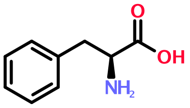
Leucine
(Leu / L)



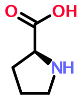
Lysine
(Lys / K)



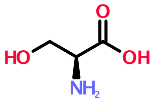
Methionine
(Met / M)



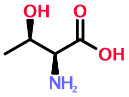
Phenylalanine
(Phe / F)



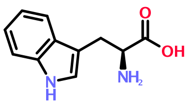
Proline
(Pro / P)



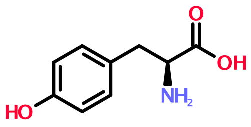
Serine
(Ser / S)



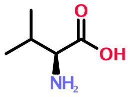
Threonine
(Thr / T)



Tryptophan
(Trp / W)



Tyrosine
(Tyr / Y)



Valine
(Val / V)

The twenty standard amino acids

B

Research Online

Discussing research online

During this PhD the work presented in this thesis was discussed online using a blog at *www.michaelbarton.me.uk*. There is a tradition in science of discussing pre-publication research at conferences using posters and presentations; the aim of the blog was to provide a similar avenue for discussion of preliminary findings. Previous research blog posts can be found at the above link, but a previously written essay discussing the usefulness of presenting research online (described as Open Notebook Science) is presented below.

A short essay on Open Notebook Science

Posted February 01, 2008

As you might expect from the name, open notebook science (ONS) has similarities with open source software. The clearest likeness between the two is the belief that by sharing and collaborating more can be achieved than through secrecy and competition. An open approach to software development is proven to be successful: the greatest achievement is the development and increasing adoption of the Linux operating system. On this foundation other applications like the Apache web server, MySQL database, and the PHP scripting language have been built, and the combination of the four is the engine running many websites. If ONS can enjoy a fraction of the success open source software does, then science can only benefit.

ONS didn't occur spontaneously, but is a step in the liberalisation of science by the freedom that the Web allows. An early example is the arXiv.org server started in 1991 as a repository for the physics community to share manuscripts prior to publication, 17 years later it now contains ∼450,000 articles. Another often overlooked example of openness are the free biological databases such as EMBL and GenBank which allow unrestricted access to the genomes for all sequenced organisms. More recently, many journals are adopting open science policies, whereby all research is freely available upon publication, where previously the reader had to pay a fee. Now, increasingly research funding bodies are also stipulating, as a condition, that any articles resulting from the research are freely available at least 6 months after publication, examples being NIH and BBSRC.

When you work in science, many ideas come from reading papers, attending talks, and speaking to colleagues in the pub. So I think it's fair to say that we will profit from further sharing on websites, such blogs and wikis, and the more everybody is open, the more the community benefits as a whole. Of course, being open creates questions on how scientists can still be recognised for their work, as well as how research can be commercialised. Most importantly, peer-review is still the best arbiter of research quality, and raw results must be viewed with this in mind.

One of the earliest adopters of complete openness is Jean-Claude Bradley, where his own and students' laboratory notebooks are stored on a wiki, and freely available for anyone to read while also updated as new results are being produced. Jean-Claude also first discussed the term "Open Notebook" in relation to this, where he defined it as the researcher's notebook being open to the world and there is no insider knowledge. From Jean-Claude's example, a small but growing number of researchers have followed: using blogs, wikis, and project management systems to make their research available. Examples of people using blogs to share research are Cameron Neylon and Rosie Redfield whose research groups use blogs either as the primary lab book or as a forum for describing and discussing results. In addition to Jean Claude Bradley, other projects using wikis for ONS are 1CellPK and Maldi. Pedro Beltrao and Jeremiah Faith use software management systems, where many tools useful for tracking software development, are applicable to bioinformatics research.

There are questions that this kind of openness generates. For instance, what do the journals think about publishing research that has already appeared on a blog? For most journals informally posting your research online is considered in the same light as giving a talk at a conference. A few exceptions exist though, such as Cell and Lancet, but on the whole publishers like NPG, BMC and PLoS are happy with kind of sharing, though it is always worth checking. Another question worth asking is what is your university's policy towards intellectual property: does it belong to the researcher or the institution? Which leads into another point, in the last few years researchers have been increasingly expected to consider how their work can be commercialised, but any work disclosed on a blog or wiki cannot be patented, which should be borne in mind when you post new ideas or methods. Finally, there is common sense - how do your collaborators feel about early sharing of research? Or could the work you're posting online be considered politically sensitive - involving animals or embryonic stem cells?

If after reading this, and looking at ONS researcher websites, you think that ONS can be useful to you, where do you begin? In my experience a blog is safe and easy place to start. You can discuss already published research, and if you feel confident you could begin to mention the results you've been producing. Then depending on how you feel, move towards making your notebook entirely open, using a wiki. Services like wordpress.com and blogger.com, offer easy blog creation for free, while wikispaces.com can be used to create a wiki. At the moment there is no single standard application for ONS, so a good idea is to experiment and see what suits you and your research.

So what is the future for Open Notebook Science? At present, proposals have been created for an ONS network, and a session at Pacific Symposium on Biocomputing 2009. There is a small, but increasing number of scientists who are adopting open practices into their research, while a further few follow the mantra of "no insider information" and are completely open. Returning to my point at the start of this article, the creator of Linux said, when talking about open source software, "Many eyes make all problems shallow" and if Open Notebook Science can benefit from similar principles, this will be to the advantage of the individual as well as the community as a whole.

The original article can be found at
www.michaelbarton.me.uk/2008/02/a-short-essay-on-open-notebook-science