Systems biology of energetic and atomic costs in the yeast transcriptome, proteome, and metabolome

Michael D Barton[*†] Balázs Papp[‡] Daniela Delneri[†]
Stephen G Oliver[†§] Magnus Rattray[¶] Casey M Bergman[†]

April 28, 2008

Running Title : Systems biology of energetic cost in yeast

Keywords : gene expression, flux balance analysis, metabolic cost, cellular economics, codon usage bias

Subject Categories : Metabolic and regulatory networks, Functional genomics

Character count : 50,249

[*]Corresponding Author - mail@michaelbarton.me.uk
[†]Faculty of Life Sciences, University of Manchester, Manchester, M13 9PT, UK
[‡]Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, H-6726 Szeged, Hungary
[§]Department of Biochemistry, University of Cambridge, Sanger Building, 80 Tennis Court Road, Cambridge, CB2 1GA, UK (Present Address)
[¶]School of Computer Science, University of Manchester, Manchester, M13 9PL, UK

Abstract

Proteins vary in their cost to the cell and natural selection may favour the use of proteins that are cheaper to produce. We develop a novel approach to estimate the amino acid biosynthetic cost based on genome-scale metabolic models, and directly investigate the effects of biosynthetic cost on transcriptomic, proteomic and metabolomic data in *Saccharomyces cerevisiae*. We find that our systems approach to formulating biosynthetic cost produces a novel measure that explains similar levels of variation in gene expression compared with previously reported cost measures. Regardless of the measure used, the cost of amino acid synthesis is weakly associated with transcript and protein levels, independent of codon usage bias. In contrast, energetic costs explain a large proportion of variation in levels of free amino acids. In the economy of the yeast cell, there appears to be no single currency to compute the cost of amino acid synthesis, and thus a systems approach is necessary to uncover the full effects of amino acid biosynthetic cost in complex biological systems that vary with cellular and environmental conditions.

Background

Everything in a living cell has a cost: from the energy needed to transform molecules against thermodynamic equilibria, to the raw materials needed to produce the constituents of a new cell. Natural selection may be expected to minimise such cellular costs, and evidence for adaptation to require less energy and matter may exist at the molecular or cellular level. Testing this hypothesis requires answering several questions about the meaning of cost in the cell, and how to measure it. For example, how does one assign a biochemical price to a molecule whose state is dependent on changing environmental and cellular conditions? Similarly, is it possible to separate different costs from one another, or from other molecular constraints in the cell? Moreover, are biosynthetic costs the same at different levels of the gene expression hierarchy? Knowing the answers to these questions is central to a systematic understanding of the chemical forces that shape the composition of biomolecules, and how biomolecular composition relates to gene expression at the transcriptional and post-transcriptional levels.

Craig and Weber [14] pioneered the quantitative analysis of cost at the cellular level to investigate the effects on the synthesis and evolution of a small number of *Escherichia coli* proteins. Their approach to estimate the cost of a protein is the sum of how many units of high energy phosphate bonds (e.g. ATP) and reducing hydrogen atoms (e.g. NADPH) are diverted from the available energy pool to produce each of the constituent amino acids from glucose, averaged over the length of the protein. Akashi and Gojobori [3] used a modified version of this approach to show in the chemoheterotrophic bacteria *E. coli* and *Bacillus subtilis*, that predicted gene expression levels based on codon usage bias show a negative correlation with average protein cost. This work provided the first genome-wide evidence that evolution has optimised prokaryotic cells to produce highly expressed proteins with less expensive amino acids, and established an important link between the metabolism of a cell and the evolution of its genome sequence. Heizer *et al.* [20] extended these findings to four additional prokaryotic species including photoautotrophs, demonstrating that this cost optimisation occurs whether the source of energy is organic or inorganic. More recently, Swire [33] used Craig and Weber's [14] cost values to generate a new cost measure for an amino acid based on its usage in proteins as a function of overall protein cost computed from all other amino acids, and showed that cost selection affects multiple prokaryotic, archaeal and eukaryotic genomes. Wagner [35] developed a method similar to Craig and Weber [14] that includes the energetic costs of synthesising both mRNA and protein for *Saccharomyces cerevisiae*, and showed that the cost of doubling gene expression after a gene duplication is likely to be significant enough to come under under selection pressure.

Seligmann [31] argued that, while the number of high energy molecules

is an important part of the energetic investment of synthesising an amino acid, this approach is unlikely to explain the entire investment made by a cell when producing an amino acid. Instead, Seligmann [31] used the molecular weight of an amino acid as a proxy for energetic costs, reasoning that this may take into account all the manifold effects of the complexity of producing larger amino acids. Molecular weight also has the advantage of being constant across species, and therefore can be used to test the cost selection hypothesis where the genome sequence is available but the topology of amino acid synthetic pathways is unknown. Seligmann used this to prove that on an individual protein basis, molecular weight is indeed minimised across a range of bacterial and eukaryotic genomes. Estimating the cost of an amino acid based on its molecular weight also raises the issue of the potential costs of the atomic content in biomolecules. Baudouin-Cornu *et al.* [4] showed that enzymes in pathways scavenging sulphur, carbon, or nitrogen from the environment are under-represented in terms of their composition for that particular nutrient. Further research by Bragg *et al.* [7] showed across 141 genomes that the sulphur content of the encoded protein varies widely and is associated with environmental conditions of the species. Both of these results indicate that atomic composition may also play an important role in the cost of producing a protein.

Even if all the energy required for protein synthesis can be accurately predicted, this may not represent the true cost under all cellular or environmental conditions. Just as in supply and demand economics[32], when a particular atom is scarce in the cell or environment, synthesis of molecules abundant in this atom will be more expensive, in comparison to molecules where that atom is under-represented [12]. Therefore, estimates of the cost of synthesising an amino acid should be performed in a systems biology framework where the cost of producing a molecule can be calculated for in the wider cellular context under a variety of environmental conditions. Here we show how two systems biology approaches, sensitivity analysis [18, 34] and flux balance analysis (FBA)[30], can be combined to provide a novel means to estimate the cost of amino acid synthesis in the microbial eukaryote, *S. cerevisiae*. In this report, we produce two new measures of amino acid cost: the first estimates the "relative" cost of synthesising the amino acid by perturbing the required quantity for growth, the second is derived by multiplying the first cost by the biomass requirement to obtain a per molecule "absolute" cost. We calculated each of these two cost types for four nutrient limiting conditions (glucose, nitrogen, sulphur, and phosphorus) to investigate how cost varies according to environmental conditions. As in previous studies[14, 31, 20, 33, 35], we focus on amino acid synthesis, because this allows us to analyse the effects of cost on gene expression.

Using our novel systems approach, we analysed the joint effects of energetic and atomic costs on the transcriptome, proteome and metabolome of *S. cerevisiae*. We show that biosynthetic cost and atomic composition

do indeed have a measurable relationship with gene expression, but that the effects of cost are dependent on the level at which the gene expression hierarchy is considered. Importantly, we analyse transcriptomic and proteomic data directly, whereas previous work has used codon usage bias as a proxy for gene expression [3, 15, 20], examined atomic composition for only a small set of expressed genes [8], or attempted to predict gene expression based on amino acid composition without respect to biosynthetic cost [28]. We show that our systems biology approach explains a significant proportion of variation in gene expression levels, independent of codon usage bias, tRNA gene number or atomic content. Our relative measure of cost is poorly correlated with previously reported measures of cost, but also explains a comparable amount of gene expression, suggesting that no single measure currently captures all aspects of biosynthetic cost. We also extend cost analysis to levels of free amino acid levels in the metabolome, an aspect of cellular economics that has not been considered in previous research, but which provides intimate links to protein synthesis.

Results

A systems biology approach to estimating the cost of amino acid synthesis

To estimate the cost of synthesising an amino acid in *S. cerevisiae*, we used the genome-scale metabolic model created by Duarte *et al.* [17]. For each amino acid, the required quantity for growth was altered and the effect on one of four nutrient uptake fluxes was measured. These uptake fluxes were glucose, ammonium, sulphate and phosphate. The biomass production rate for the model was fixed at a constant value, so that cost estimates were scaled to the same growth rate in each nutrient limiting condition. For each amino acid, a percentage increase and decrease in requirement for production of biomass was applied. This allowed the cost of an amino acid relative to biomass requirement ("relative cost") to be measured as the slope between the percentage change in amino acid requirement and the predicted uptake flux. A per-molecule "absolute cost" was then calculated for each amino acid by dividing relative cost by the biomass requirement (see Methods for details). Figure 1 shows previous costs measures reported in the literature, and Figure 2 shows our novel systems biology cost estimates. Figure 3 shows an agglomerative hierarchical clustering of all cost measures listed in Table 1, based on the Spearman's Rank correlation (Additional File 2).

As expected if the limiting factor is the availability of atoms to create the molecule rather than energetic limitation, we find that the absolute cost of an amino acid under S and N limitation is directly proportional to the number of atoms of that nutrient in the molecule. The absolute costs of all amino acids under P limitation are zero, in accordance with the fact

that amino acids do not contain phosphate atoms. Absolute cost estimates under glucose limitation have Spearman correlation coefficients greater than 0.8 with Akashi and Gojobori's [3] energetic cost, Craig and Weber's [14] energetic cost, Wagner's [35] respiratory energetic cost, and molecular weight (Additional File 2; Figure 3). Wagner's [35] fermentative energy cost and Craig and Weber's [14] biosynthetic complexity show lower coefficients of 0.522 and 0.65, respectively, but are still significantly correlated. Figures 4A and 4B illustrate the relationship of the absolute cost under glucose limitation with two example data sets, Akashi & Gojobori energetic cost and molecular weight. These results show our absolute cost measure under glucose limitation is in good agreement with the previous manually-curated measures described in the literature, and indicate that the cost of an amino acid is most likely a function of energetic limitation in yeast.

Our relative costs are can be viewed as the absolute cost of synthesising the amino acid, scaled by its use in the proteome. For example, cysteine and methionine have the same absolute cost under sulphur limiting conditions, but in relative terms the cost of methionine is much greater because it is used more in the proteome. A similar pattern is also observed for histidine and lysine, whose rank order absolute costs switch when scaled by proteome use. As expected, phosphate limitation does not shown any effect on the relative cost of amino acid synthesis. Compared with previously reported cost measures, relative cost under glucose limitation shows no significant correlation with any previously described cost measure (all $p > 0.05$), as illustrated in Figures 4C and 4D. The highest Spearman coefficient between relative cost under glucose limitation and any other literature dataset is 0.077 ($p = 0.49$), when compared with Wagner's fermentative energetic cost, indicating that our relative cost [35] measure under glucose limitation has little in common previous descriptions of amino acid cost.

We used similar measures to calculate amino acid cost using the iJR904 model of the *E. coli* metabolic network [29] to estimate the generality of our results among species and FBA model. Absolute costs of synthesis are highly correlated between species under glucose limitation (Spearman $R = 0.94$, $p = 0$), as are relative costs under glucose limitation (Spearman $R = 0.74$, $p < 0.001$). The high correlation of absolute cost is expected given the conservation of metabolism [10], whereas the relatively lower correlation of relative costs may arise from differences in amino acid composition of the proteome across species. The estimated costs for *E. coli* are also included in Figure 4 for illustrative purposes, and demonstrate the general applicability of our method to any species with a genome-scale metabolic model.

The cost of amino acid synthesis influences the yeast transcriptome, proteome and metabolome

Transcriptome

We investigated the capacity of absolute and relative cost under glucose limited growth, as well as each previously reported cost measure, to explain then transcript expression levels in each of the four nutrient limiting environments from the *S. cerevisiae* dataset of Castrillo *et al.* [13] using multivariate regression. The expression of each transcript was modelled as a function of the codon adaptation index (CAI) of the coding sequence, average tRNA gene number, the mean energetic cost per residue of the protein, and the mean atomic composition per residue of the protein. We included CAI and tRNA gene number in the model since these factors are known to correlate with gene expression levels [22, 2], and allows us to demonstrate an independent effect of cost that controls for these factors. Each of the selected cost types, was cycled as the cost variable in the regression equation. Only the relative cost under glucose limitation was used, as this is the environment most relevant to yeast [16, 9, 13] and the other costs under P, N and S limitation are proportional to atomic composition, which is already included in the model. Table 2 shows the explanatory power for the full regression model for each cost type in predicting transcript levels. All cost models explain ∼40% of the variation in transcript levels across genes, with the difference in variation explained by the best and worst model being only 4.5%.

Using Akaike's Information Criterion (AIC) [1] the importance of the variables in the regression equation was measured by removing each in turn, then comparing the goodness of fit with the model containing all terms. Figure 5A compares the importance of each variable with other variables in the same model for each cost type. Compared to characteristics of the encoded protein, the CAI of the transcript is at least half an order of magnitude more important than the nearest explanatory variable, regardless of which cost type is included. This result supports the well-established fact that codon bias correlates with gene expression in growing yeast cells [22, 23, 11]. The dominant influence of CAI over other factors also explains why the different cost types do not yield substantially different predictive power in the model. In the molecular weight model which has the greatest explanatory power (42.2%), the most important individual variable for predicting transcript level after CAI is cost, and carbon content is the third most important. A general trend across all models is that the most important variable after CAI is either cost, carbon content or nitrogen content. The importance of tRNA gene number on transcript levels appears relatively fixed regardless of which cost is used. Finally average sulphur content appears the least predictive measure.

Proteome

The importance of cost in explaining protein levels was also modelled using multivariate regression followed by variable removal. To analyse the effect of cost on gene expression at the protein level, we used data from Ghaemmaghami *et al.* [19], measuring antibody tap-tagged protein expression, since protein expression levels from Castrillo *et al.* [13] were measured relative to a background (see Methods for details). Table 2 illustrates the explanatory power of each cost model to predict protein expression levels. As with the transcript data, each model explains approximately $\sim 40\%$ of gene expression, and the difference in explained variation between the best and worst model is very small ($\sim$0.8%), relative to the overall variance explained.

Figure 5B shows the relative importance of each factor in the multivariate regression model for protein levels. This analysis illustrates similar trends to that of the transcript data where CAI is, by an order of magnitude, the most important factor in the model. This is not unexpected given that in the original paper, Ghaemmaghami *et al.* [19] showed Spearman $R = 0.57$ for the relationship between CAI and protein expression. The best fit model uses absolute cost under glucose limited conditions, in which biosynthetic cost, carbon content and nitrogen content all have a similar importance in explaining variation. Sulphur content is again the least important variable. This is a similar trend to the transcript data where generally (i.e. across all models) biosynthetic cost, carbon content and nitrogen content all play a similar importance in explaining variation in gene expression levels, and sulphur content is the least important. However the importance of tRNA gene number, and sulphur content are more variable than in the transcript data, and in some instances their removal improves model parsimony, as indicated by a negative AIC.

Metabolome

The availability of comprehensive metabolomic data for *S. cerevisiae* from Castrillo *et al.* [13], allows us to determine if atomic and energetic costs are important in the synthesis of amino acids. Using similar multivariate regression and variable removal, we investigated the importance of each variable in explaining free amino acid levels. We used the same factors as the previous two analyses, with the exception of CAI (which is not applicable to amino acids). The main difference between analysis of the metabolomic data and either the transcriptomic or proteomic data are fewer number of data points for free amino acids versus those for transcripts and proteins. Table 2 shows how much of the variance in free amino acid level is explained by each of the multivariate models. In contrast to transcript or protein levels, cost type shows the greatest range in explaining variation in free

amino acid levels, with R^2 coefficients ranging from 76.7% for molecular weight, to 87.5% for relative cost under glucose limitation. The explanatory power of these cost models at the metabolomic level are remarkable given that CAI was not included, and are due only to the effects of energetic costs, atomic costs and genomic tRNA gene number.

Figure 5C shows the importance of each cost type in explaining free amino acid levels. The general trend across these models in that all variables appear important, though there is more variability for carbon and sulphur content. In particular under glucose limitation, carbon and sulphur content are less important and therefore the explanation of free amino acid levels can be attributed to nitrogen content, tRNA gene number, and the relative cost of synthesis.

Discussion

The principal achievements of this work are twofold. First, we developed a novel method to estimate the cost of amino acid synthesis using a systems biology approach that incorporates sensitivity analysis and flux balance analysis of genome-scale metabolic models. We compared our novel estimates of amino acid costs to six measures reported in the literature and showed that absolute cost under glucose limitation is highly correlated with previous cost measures, while relative cost under glucose limitation is not. Furthermore we showed that our systems biology approach can be applied to calculate environment-specific biosynthetic costs, which highlighted the effects of limiting elements of amino acid cost. Second, we investigated the utility of energetic cost measures in conjunction with atomic costs and other factors to analyse transcript, proteomic, and metabolomic data from *S. cerevisiae*. Our analysis shows that amino acid costs do show a relationship with gene expression, but explain only a minor component of transcript and protein levels relative to factors related to translational optimisation such as CAI. In contrast, we find that energetic and atomic costs do explain a substantial degree of the variation in levels of free amino acids in the metabolome.

No single currency for amino acid biosynthetic cost

Our systematic review and comparison of energetic cost types previously described in the literature (Table 1) shows that they are highly correlated with one another. Among previously reported measures, molecular weight is the least related (Figure 3), which is expected since the other energetic cost measures are based on manual curation of metabolic networks. This finding supports the view of Seligmann [31] that the molecular weight of an amino acid includes investments by the cell that can not easily be estimated from the metabolic network alone. Nevertheless, molecular weight and biosynthetic cost based on curated metabolic networks are highly correlated (see

also [31]). Of the two costs estimated from a glucose limited state, which is probably most relevant to yeast biology, our absolute cost measure correlates with those previously described in the literature (Figure 3 and 4), confirming previous cost measures and validating our general approach to estimating biosynthetic cost. Our absolute cost measure, like all previously reported cost measures (with the exception of Wagner's fermentative measure [35]), points to tryptophan as being the most expensive amino acid for the cell to produce (Figure 1, Table 1). Tryptophan is considered expensive because of its complex double ring structure and the number of high energy molecules required for its synthesis and is (along with methionine) unique in that it is encoded by only one codon in the genetic code.

In contrast to our absolute cost in glucose limitation, the corresponding relative cost shows little relationship with any previously described cost metric under the same conditions (Figures 3 and 4), and provides a novel perspective on how to measure the cost of amino acid biosynthesis. Under glucose limitation, relative cost shows leucine and lysine to be the most expensive amino acids and tryptophan is estimated as one of the cheapest, in contrast to other previously reported cost measures (see above). Our relative cost measure reflects the absolute cost of synthesising the amino acid, scaled to its use in the proteome. Therefore although a tryptophan molecule may be expensive to produce individually, its low usage in the proteome makes it cheaper to maintain overall at the cellular level.

To test whether absolute or relative cost may have shaped the long-term evolution of yeast genes, we compared the cost of each amino acid estimated under glucose limitation with their proportional usage in the genome (Figure 7). It is important to note that our relative costs are estimated using amino acid biomass composition in the cell, not amino acid usage in the genome, and therefore these two datasets are independent. We find a high correlation between relative cost and amino acid usage in the genome (Spearman R = 0.65, p = 0.0021), but not for absolute cost (Spearman R = -0.37, p =0.1053). This result supports the observation that certain amino acids in *S. cerevisiae* are more likely to appear in highly expressed proteins noted by Jansen & Gerstein [24], who suggested that this could be related to their cost of synthesis. An interesting exception to the relationship between glucose limited relative cost and usage in the genome is that serine does not follow the proportional use versus cost trend. Serine was previously identified as an outlier in an analysis of the relationship between cost and rates of amino acid substitution [21]. We speculate that there are biological reasons why serine may be less costly than expected relative to other amino acids based on it usage in yeast genes. Serine is involved in nucleic acid synthesis, as well as that of glycine and cysteine, therefore additional demand for serine may be buffered by the many pathways to which it is linked, and this was identified by our approach to estimating cost using sensitivity analysis of genome scale metabolic models.

While it is clear that no single measure may fully capture all aspects of the cost of amino acid synthesis, we believe our systems biology method for computing amino acid cost has a number of advantages over previous methods. The first is that, given a genome-scale FBA model, computationally generated cost measures require no manual curation and allow cost calculations that are more explicitly replicable. Moreover, use of a computational model allows costs to be calculated under a variety of nutrient conditions, permitting a more flexible approach to exploring costs under different cellular and environmental conditions. Additionally, we believe our approach takes into account the whole cellular state, including all simulated reactions and metabolites, not just those between substrate and product. The main drawback is that a species-specific FBA model must be available to perform the analysis, though our results shows that absolute cost of synthesis is conserved across species, while relative requirements may vary. Future work could address how costs vary between organisms that have evolved and adapted their proteome to markedly different environmental conditions.

Translational optimisation over cost minimisation

At the transcript and protein levels, our models explain approximately 40% of the the variation in expression (Table 2). Overall, codon usage bias is the most important factor for explaining variance in gene expression levels, and the other factors only show a limited impact on model fit. Therefore, of the variance in gene expression explained by our models, the majority is due to optimisation of the coding sequence for translation rather than cost minimisation. Nevertheless, we can demonstrate an independent, albeit small effect of amino acid cost on transcript and protein levels (Table 2, Figure 5). The correlation between cost and gene expression are complex and depend on the cost measure used (Figure 5). The small effect of the cost type used in the model may also be expected, as we have shown that they are all highly correlated, and therefore little variance occurs between each measure. The exception to this is the relative cost measure under glucose limited conditions which shows no correlation with previous measures, yet still explains a significant degree of variation in gene expression levels. This indicates that the physiological maintenance of amino acids in the the proteome, and not just their absolute cost, is an important factor in considering the cost of gene expression.

For the analysis of the metabolomic data, the variation in the R^2 values between models is much greater than observed at the transcript or protein levels. One possible explanation for this is the reduced number of data points (N = 184 [13]), compared with the large transcript (N = 36264 [13]) and protein (N = 2204 [19]) data sets used in the previous analyses. In contrast to transcript and protein levels, the model that explained the greatest variation in free amino acid levels was based on our relative cost measure,

demonstrating the value of this model for interpreting energetic investment at the metabolomic level. This is of further interest as the relative cost relates absolute cost of synthesis to its usage in the proteome, therefore indicating that free amino acids are maintained at levels in the cell proportional to that encoded in the genome, a finding which has not been demonstrated previously.

Conclusions

We have conducted a systematic investigation of the hypothesis that the cost of synthesising amino acids has shaped the evolution of protein primary structure and gene expression in yeast. Our analysis indicates that cost plays a role, but not as large as might be expected given that a predicted 80% [35] of the cellular ATP budget is devoted to protein synthesis. Instead our research shows that CAI, and therefore translational efficiency is the dominant factor in the evolution of gene expression. We believe this indicates that the optimisation of translation outweighs any benefits that would be gained from the use of cheaper amino acids. This is further illustrated by our analysis of the metabolomic data where the cost measure that shows the greatest explanatory power, is highly correlated with the usage of amino acids in the proteome.

Materials and Methods

A systems biology approach to estimating the cost of amino acid synthesis

Flux balance analysis was performed using the COBRA toolbox [5], running in the MATLAB environment. The genome scale models used were iND750 for *S. cerevisiae* [17] and iJR904 for *E. coli* [29]. In each model, the stoichiometry of the biomass reaction determines the required ratio of biomolecules used to produce a new unit of biomass. However, costs between models can be compared by fixing biomass flux to a constant value. The units used in the model are mmol of reactant, per gram of biomass, per hour. Simulation of the model, using the lpsolve library [25], is the solution of a linear programming problem to maximise or minimise the flux through a particular reaction given the topology, and upper and lower bounds on the reactions in the metabolic network of the organism.

For each of the twenty amino acids, which are all included in the biomass reaction, we altered in turn the requirement of each for the production of biomass. This ranged from a 0.0002% increase in requirement, to a -0.0002% decrease, at 0.0001% intervals. For each alteration biomass production flux was fixed at 0.05, and the model solved to maximise one of the four input

fluxes glucose, ammonium, sulphate, and phosphate. Maximising uptake flux is the equivalent of finding the solution with the minimal flux molecule into the cell. The other uptake fluxes were set -1000, effectively unlimited. The aim of this was to simulate the expense of the amino acid under a given nutrient limitation, but also scaling each cost to the same growth rate. The relative cost for a given amino acid, under a given nutrient limitation, was then estimated as the slope between amino acid requirement and the corresponding nutrient uptake flux. As the relative cost is estimated from a percentage change in amino requirement, this can be scaled to an absolute per molecule cost by multiplication of $x_0/100$, where x_0 is the biomass requirement of that amino acid. The proofs for this relationship are shown below. The code used in this analysis is available in the supplementary materials.

$$\begin{aligned} \text{Absolute Cost} &= \left.\frac{dG}{dx}\right|_{x=x_0} \\ &= G'(x_0) \end{aligned}$$

$$\begin{aligned} \text{Relative Cost} &= \left.\frac{d}{dx} G\left(x_0\left(1+\frac{x}{100}\right)\right)\right|_{x=0} \\ &= \frac{x_0}{100}\, G'(x_0) \end{aligned}$$

Determination of transcript, protein, and amino acid characteristics

The Codon Adaptation Index (CAI) for each *S. cerevisiae* gene was taken from Wall *et al.* 2005 [36], tRNA gene number was taken from Akashi [2]. Previously reported amino acid energetic costs were obtained from Craig & Weber [14], Akashi & Gojobori [3], Wagner [35], and Seligmann [31]. For each gene, the average tRNA gene number, energetic cost, or atomic cost was computed as the sum of the count or cost over the encoded protein, divided by the length, excluding stop codons. Prior to analysis, each these variables was transformed by the natural logarithm, then scaled to have the the same mean and variance. This was to reduce any over-variation and heteroscedasticity biasing model estimation. Scaling was performed by subtracting the mean, then dividing by the root mean square for each data set. For the metabolomic data set, a small constant (0.0001) was added to sulphur content so that this variable could be logged.

Determining explanatory power of factors in transcript, protein, and metabolite data

Multiple regression was used to measure the importance of atomic and energetic cost on transcript and protein expression using the R statistical computing language [27]. For each data set, a multiple regression model was fitted. The measured quantities of the transcript, protein, or metabolite was treated as the response variable, and atomic cost, energetic cost, and the codon adaptation index (if applicable) where used as explanatory variables. Atomic cost consisted of three independent variables: carbon, nitrogen and sulphur content. Experimental conditions that differed among replicates in the datasets were treated as fixed effects in the model, and included as interaction terms. Initially, all possible interaction terms were considered and automated step-wise regression used to remove superfluous interaction terms based on a penalised log. likelihood score, Akaike's Information Criterion (AIC) [1].

To estimate the importance of each of the equation parameters, the data set was modelled without the variable in question, and then compared to the model containing all terms, again using AIC. For example, to estimate the importance of nitrogen in the Castrillo *et al.* 2007 Castrillo2007 data set, the data were first modelled using all factors - environment, dilution rate, CAI, tRNA gene count, energetic cost, nitrogen, carbon and sulphur content. The importance of nitrogen was then determined by repeating the data modelling with the same variables, except nitrogen content. The contribution of nitrogen content to explaining the variation in models was then estimated from the difference in the model without nitrogen with the model containing all terms. This process was performed for all factors in the equation, and then repeated for all energetic cost estimates as the cost variable in the equation.

Experimental data

Experimental transcriptomic, proteomic and metabolomic data used in this analysis are from Castrillo *et al.* 2007 [13], and an additional proteomic dataset is from Ghaemmaghami *et al.* 2003 [19]. Briefly, the Castrillo *et al.* 2007 [13] experiments continuously cultured *S. cerevisiae* using a chemostat under four nutrient limiting conditions and three (two for protein data) dilution rates, for a total of twelve (eight for protein) different experimental conditions. The transcript data produced from replicate microarray analysis of total RNA, which were processed by robust multi-array (RMA) quantile normalisation [6]. Proteomic data was produced using Isotope Tags for Relative and Absolute Quantification (iTRAQ) LC-MS/MS and standardised relative to a standard pool sample and normalised by median absolute deviation. Metabolomic data was obtained by GC/TOF-MS, and also nor-

malised using median absolute deviation, missing values were inferred from replicates in the same conditions.

As the protein data from Castrillo *et al.* [13] measured up/down regulation of a protein against a background, which is not suitable as a measure of absolute protein expression levels, we instead used data from Ghaemmaghami *et al.* 2003 [19] for our analyses of cost on protein expression. This reasoning was borne out by the small explanatory power ($R^2 < 3\%$) for any cost measure model using the protein data from Castrillo *et al.* [13]. Protein expression data from Ghaemmaghami *et al.* 2003 [19] is based on tandem affinity purification (TAP) of TAP-tagged *S. cerevisiae* ORFs. Expression levels for each protein were determined using antibody-tag based quantification. These data were converted to absolute protein molecules per cell using a purified *E. coli* INFA-TAP construct standardised against the range of yeast TAP tag protein observations.

For the model analysis, metabolite levels were mean averaged in each experimental condition to prevent pseudo-replication of observations. Protein and metabolite levels were logged then scaled. Transcript levels were scaled, but not logged as they were logged already in the original processing. The reasons for this are the same as above, as is the scaling method.

Authors contributions

MDB performed the research. MDB, MR and CMB analysed the data. BP contributed ideas and methods to the development of the systems biology cost measures. DD and SGO provided supervision. MDB and CMB wrote the manuscript. All authors read, revised and approved the final manuscript.

Acknowledgements

We thank Nick Gresham, Simon Oliver and Markus Herrgard for help implementing the COBRA toolbox; Nick Gresham and Adam Huffman for support of the University of Manchester Faculty of Life Sciences Bioinformatics Beowulf cluster; Evangelos Simeonidis for discussion of flux balance analysis; Leo Zeef, Juan Castrillo, Pinar Pir and Andy Hayes for helpful discussion of the transcriptomic, proteomic and metabolomic datasets; and Hans Westerhoff, Sam Griffiths-Jones, Simon Whelan and Laurence Hurst for their critical comments on this work. This work is funded by NERC Ph.D. studentship NER/S/R/2005/13609 to MDB, NERC advanced fellowship NE/B500190/1 to DD, NER/T/S/2001/00343 to SGO, and BBSRC grant BB/C008219/1 to SGO, MR, and CMB and others. This is a contribution from the Manchester Centre for Integrative Systems Biology [26]

Figures

Figure 1 - Ranked amino acid costs across literature data sets.

Each data set was centred on its median, and the variance scaled using median absolute deviation. Amino acid order is based on the median of the median-normalised cost of the amino acid across data sets.

Figure 2 - Relative and absolute derived amino acid costs on three nutrient uptake fluxes in *S. cerevisiae*

Estimated effect on exchange fluxes for each of amino acid. The amino acids are ordered by median normalised rank as in Figure 1. Phosphate limitation was not included, as costs calculated under this limitation were effectively 0.

Figure 3 - Hierarchical clustering of amino acid biosynthetic cost estimates.

The clustering method was complete agglomerative, using Spearman's Rank correlation distance (Additional File 2) between data sets.

Figure 4 - Comparison of the genome scale model derived cost data sets.

Comparison of estimated amino acid cost with number of ATP and NADPH molecules used in synthesis (left), and molecular weight (right). On the y axis are the amino acid costs estimated using flux balance analysis. Both *S. cerevisiae* and *E. coli* measures are included to illustrate variance between species. Estimated cost values have been been rescaled around their mean value to allow comparisons across species. The trends in each plot are drawn using 'loess' smoothing.

Figure 5 - Comparison of models and variable explanatory effects for transcript, protein and metabolite data.

Carbon, nitrogen, sulphur and cost are characteristics of the encoded protein, while CAI is a characteristic of the transcript. A multiple regression model was fitted to explain transcript, protein and metabolite levels. Each variable was then removed from the model and effect on model explanatory power was measured using Akaike's Information Criterion. Eight different cost effects were used as the cost explanatory variable.

Figure 6 - Comparison of amino acid estimated absolute and relative cost, and the percentage use in the yeast genome.

'Loess' smoothing is used to indicate trend.

Tables

Table 1 - Predicted amino acid cost estimates

Datasets taken from the literature are indicated with reference. The Akashi & Gojobori, Craig & Weber energy, and the two Wagner data sets are based on the curation of the number of high-energy molecules used during synthesis, where a defined ratio is used to convert them into a single measures: usually ATP. The Craig & Weber 'steps' measure [14] is based on the number of the number of biosynthetic steps between central metabolism and the produced amino acid. Molecular weight is in Daltons. Our cost measures are the first order derivatives of the relationship between the amino acid requirement for growth and nutrient uptake flux.

Table 2 - Adjusted R^2 coefficients for multiple regression models

The R^2 describes the data for the model with tRNA gene count, and all atomic, and energetic factors. CAI is also included in the model for transcript and protein data. Each row represents the specific cost factor used in that model.

Additional Files

Additional file 1 — Amino acid costs

The amino acid cost data sets used in the analysis.

Additional file 2 — Amino acid cost correlations

Spearman's rank correlations between cost data sets.

Additional file 3 — Tablulated transcript data set

The transcript data from Castrillo *et al.* 2007 tabulated with cost, atomic composition, tRNA gene number and CAI.

Additional file 4 — Tabulated protein data set

The protein data from Ghaemmaghami *et al.* 2003 tabulated with cost, atomic composition, tRNA gene number and CAI

Additional file 5 — Tabulated metabolite data set

The metabolite data from Castrillo *et al.* 2007 tabulated with cost, atomic composition, and tRNA gene number.

References

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

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

[3] Hiroshi Akashi and Takashi Gojobori. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc Natl Acad Sci U S A*, 99(6):3695–3700, Mar 2002.

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

[5] S A Becker, A M Feist, M L Mo, G Hannum, B Ø Palsson, and M J Herrgard. Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox. *Nat Protoc*, 2(3):727–738, 2007.

[6] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, January 2003.

[7] J G Bragg, D Thomas, and P Baudouin-Cornu. Variation among species in proteomic sulphur content is related to environmental conditions. *Proc Biol Sci*, 273(1591):1293–1300, May 2006.

[8] J G Bragg and A Wagner. Protein carbon content evolves in response to carbon availability and may influence the fate of duplicated genes. *Proc Biol Sci*, 274(1613):1063–1070, Apr 2007.

[9] M. J. Brauer, A. J. Saldanha, K. Dolinski, and D. Botstein. Homeostatic adjustment and metabolic remodeling in glucose-limited yeast cultures. *Mol Biol Cell*, 16(5):2503–2517, May 2005.

[10] M J Brauer, J Yuan, B D Bennett, W Lu, E Kimball, D Botstein, and J D Rabinowitz. Conservation of the metabolomic response to starvation across two divergent microbes. *Proc Natl Acad Sci U S A*, 103(51):19302–19307, Dec 2006.

[11] R Brockmann, A Beyer, J J Heinisch, and T Wilhelm. Posttranscriptional expression regulation: what determines translation rates? *PLoS Comput Biol*, 3(3), Mar 2007.

[12] R P Carlson. Metabolic systems cost-benefit analysis for interpreting network structure and regulation. *Bioinformatics*, 23(10):1258–1264, May 2007.

[13] J I Castrillo, L A Zeef, D C Hoyle, N Zhang, A Hayes, D C Gardner, M J Cornell, J Petty, L Hakes, L Wardleworth, B Rash, M Brown, W B Dunn, D Broadhurst, K O'Donoghue, S S Hester, T P Dunkley, S R Hart, N Swainston, P Li, S J Gaskell, N W Paton, K S Lilley, D B Kell, and S G Oliver. Growth control of the eukaryote cell: a systems biology study in yeast. *J Biol*, 6(2):4–4, 2007.

[14] C L Craig and R S Weber. Selection costs of amino acid substitutions in cole1 and colia gene clusters harbored by *Escherichia coli*. *Mol Biol Evol*, 15(6):774–776, Jun 1998.

[15] Sabyasachi Das, Subhagata Ghosh, Archana Pan, and Chitra Dutta. Compositional variation in bacterial genes and proteins with potential expression level. *FEBS Letters*, 579(23):5205–5210, September 2005.

[16] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, October 1997.

[17] N C Duarte, B Ø Palsson, and P Fu. Integrated analysis of metabolic phenotypes in *Saccharomyces cerevisiae*. *BMC Genomics*, 5(1):63–63, Sep 2004.

[18] D Fell. *Understanding the control of metabolism*. Portland Press, 1997.

[19] S Ghaemmaghami, W K Huh, K Bower, R W Howson, A Belle, N Dephoure, E K O'Shea, and J S Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959):737–741, Oct 2003.

[20] Esley M Jr Heizer, Douglas W Raiford, Michael L Raymer, Travis E Doom, Robert V Miller, and Dan E Krane. Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. *Mol Biol Evol*, 23(9):1670–1680, Sep 2006.

[21] Laurence D. Hurst, Edward J. Feil, and Eduardo P. C. Rocha. Protein evolution: Causes of trends in amino-acid gain and loss. *Nature*, 442(7105):E11–E12, August 2006.

[22] T. Ikemura. 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. *J Mol Biol*, 158(4):573–597, July 1982.

[23] R Jansen, H J Bussemaker, and M Gerstein. 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 Res*, 31(8):2242–2251, Apr 2003.

[24] Ronald Jansen and Mark Gerstein. Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucl. Acids Res.*, 28(6):1481–1488, March 2000.

[25] lp_solve. http://sourceforge.net/projects/lpsolve.

[26] The Manchester Centre For Integrative Systems Biology. http://www.mcisb.org/.

[27] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[28] G P Raghava and J H Han. Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein. *BMC Bioinformatics*, 6:59–59, 2005.

[29] J L Reed, T D Vo, C H Schilling, and B Ø Palsson. An expanded genome-scale model of *Escherichia coli* k-12 (ijr904 gsm/gpr). *Genome Biol*, 4(9), 2003.

[30] C H Schilling, J S Edwards, D Letscher, and B Ø Palsson. Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol Bioeng*, 71(4):286–306, 2001.

[31] H Seligmann. Cost-minimization of amino acid usage. *J Mol Evol*, 56(2):151–161, Feb 2003.

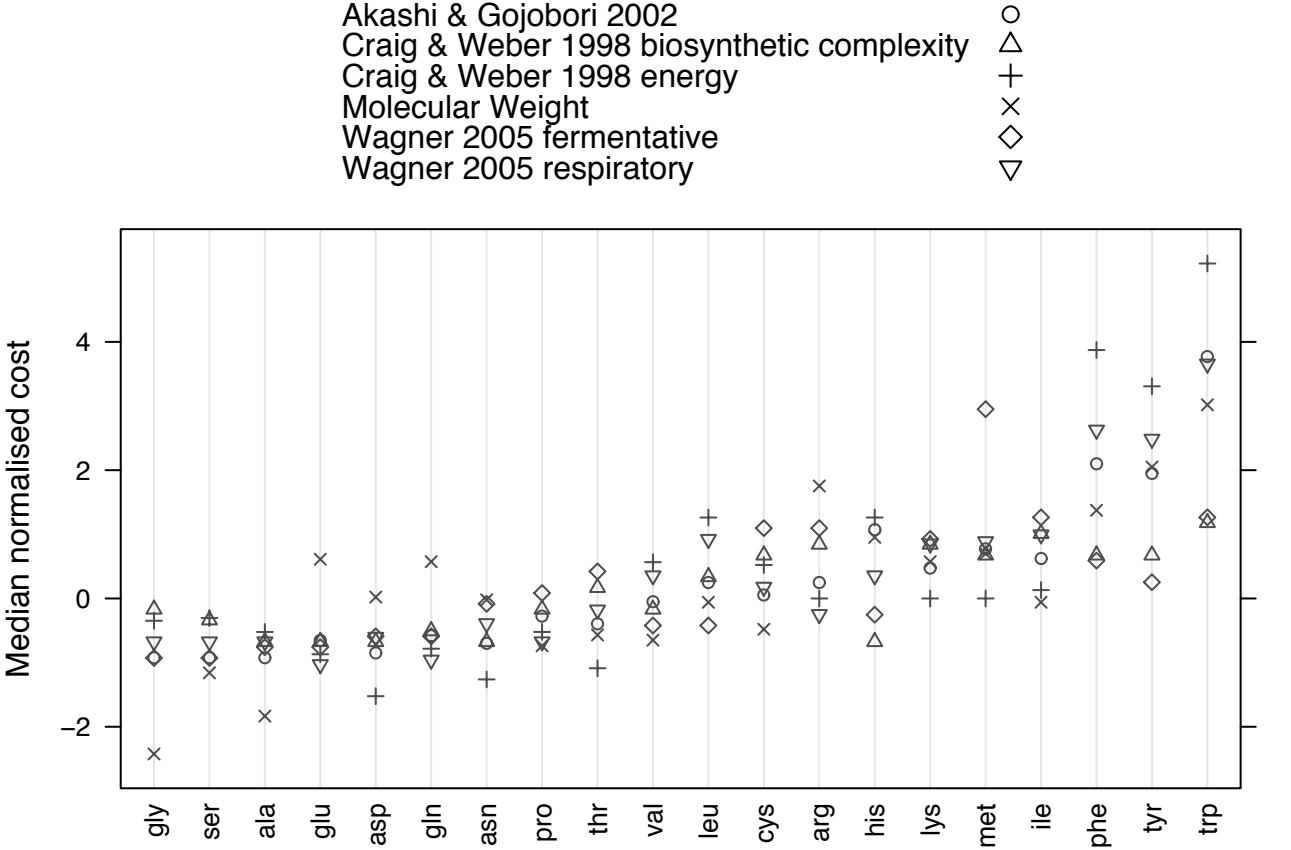
[32] Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W Strahan and T Cadell, 1776.

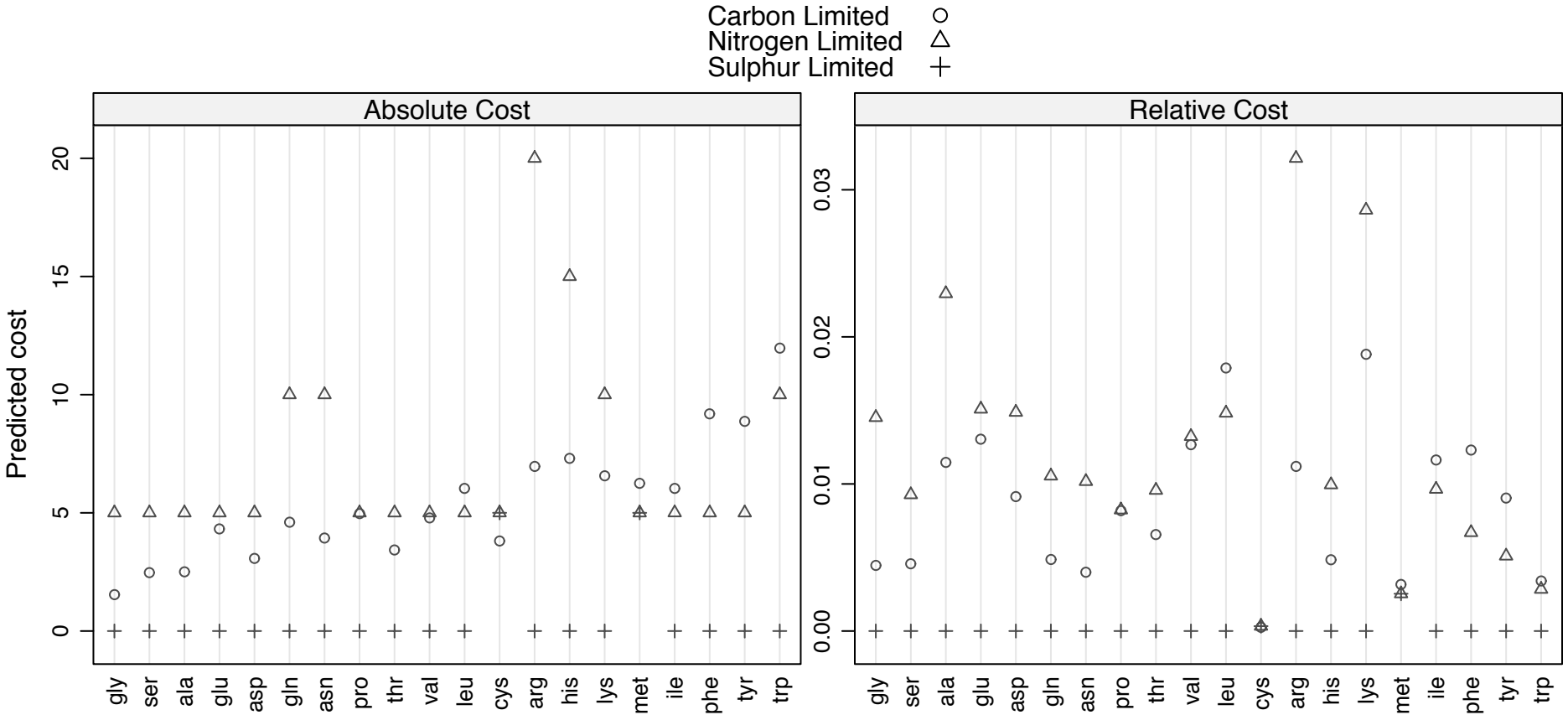
[33] J Swire. Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. *J Mol Evol*, 64(5):558–571, May 2007.

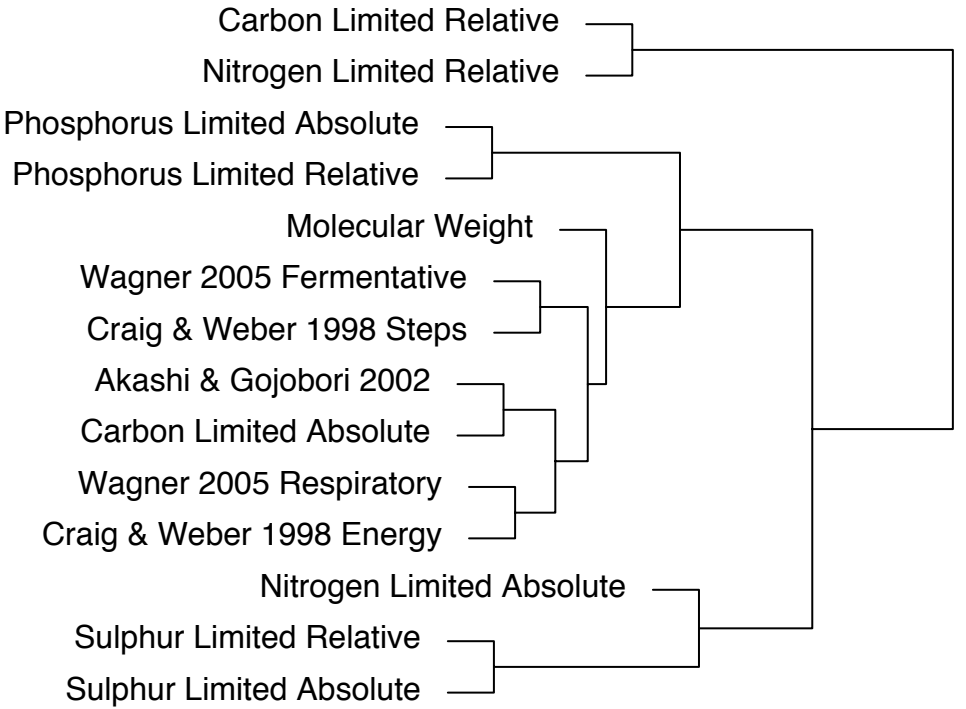
[34] B H ter Kuile and H V Westerhoff. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett*, 500(3):169–171, Jul 2001.

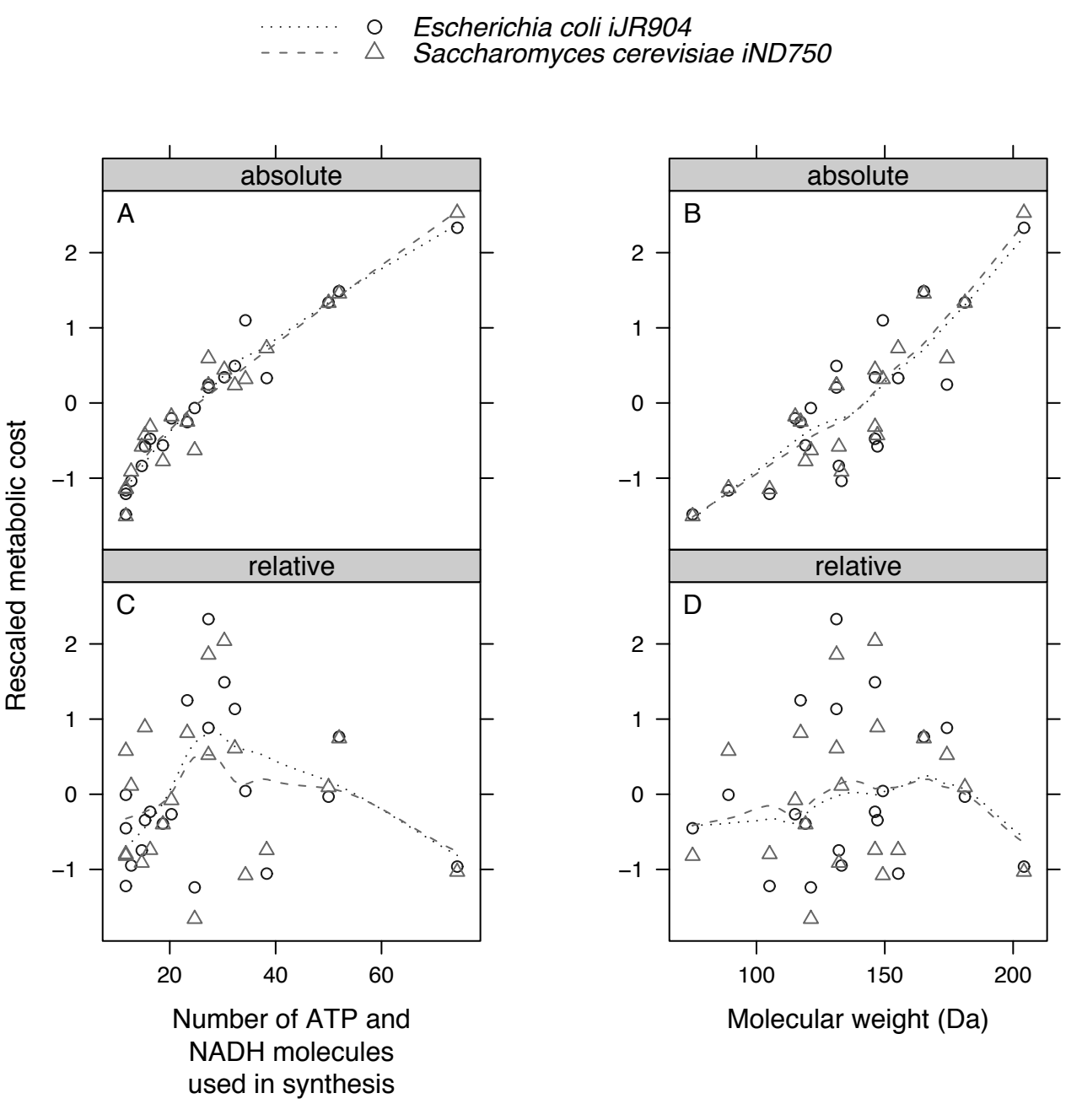
[35] A Wagner. Energy costs constrain the evolution of gene expression. *J Exp Zoolog B Mol Dev Evol*, 308(3):322–324, May 2007.

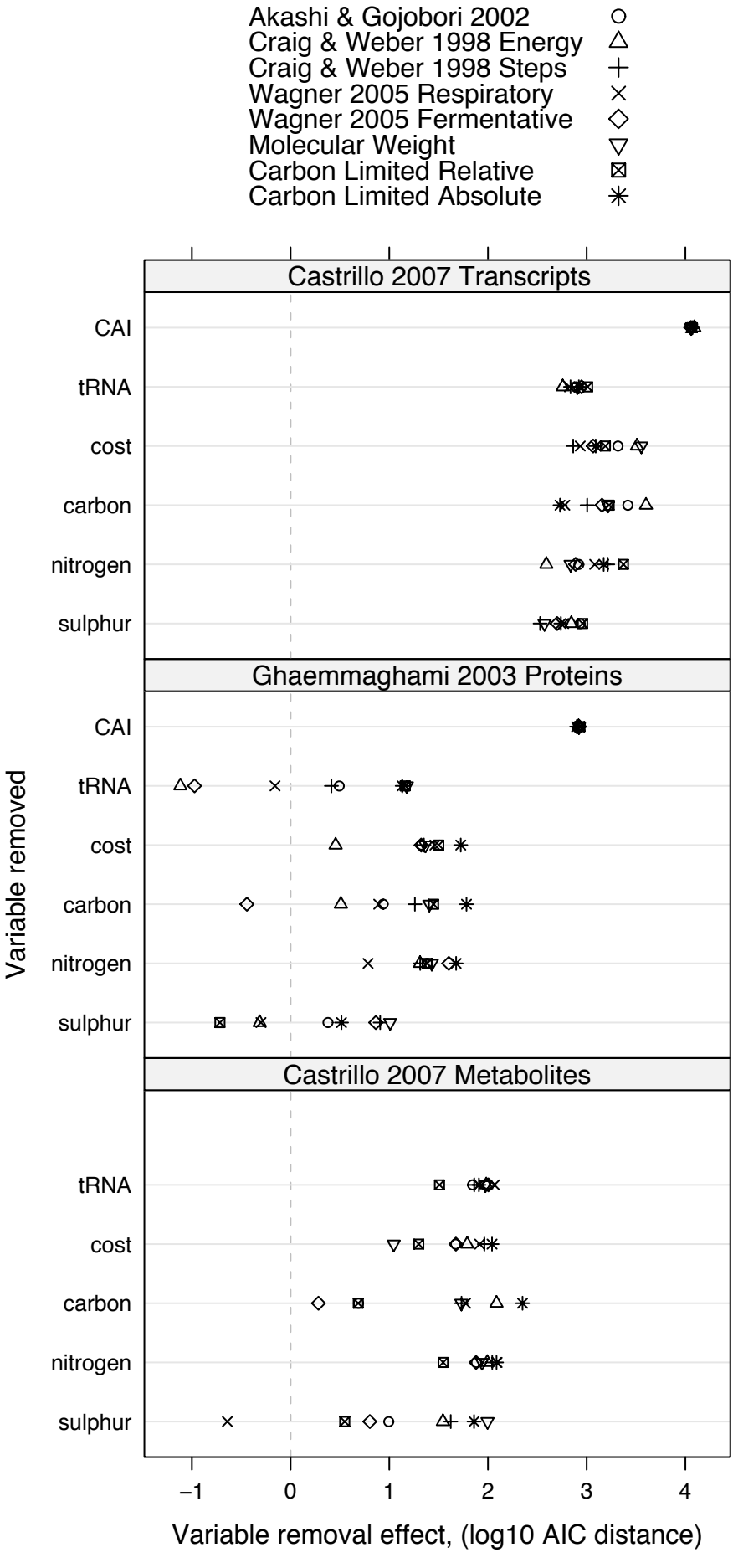
[36] D P Wall, A E Hirsh, H B Fraser, J Kumm, G Giaever, M B Eisen, and M W Feldman. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci U S A*, 102(15):5483–5488, Apr 2005.

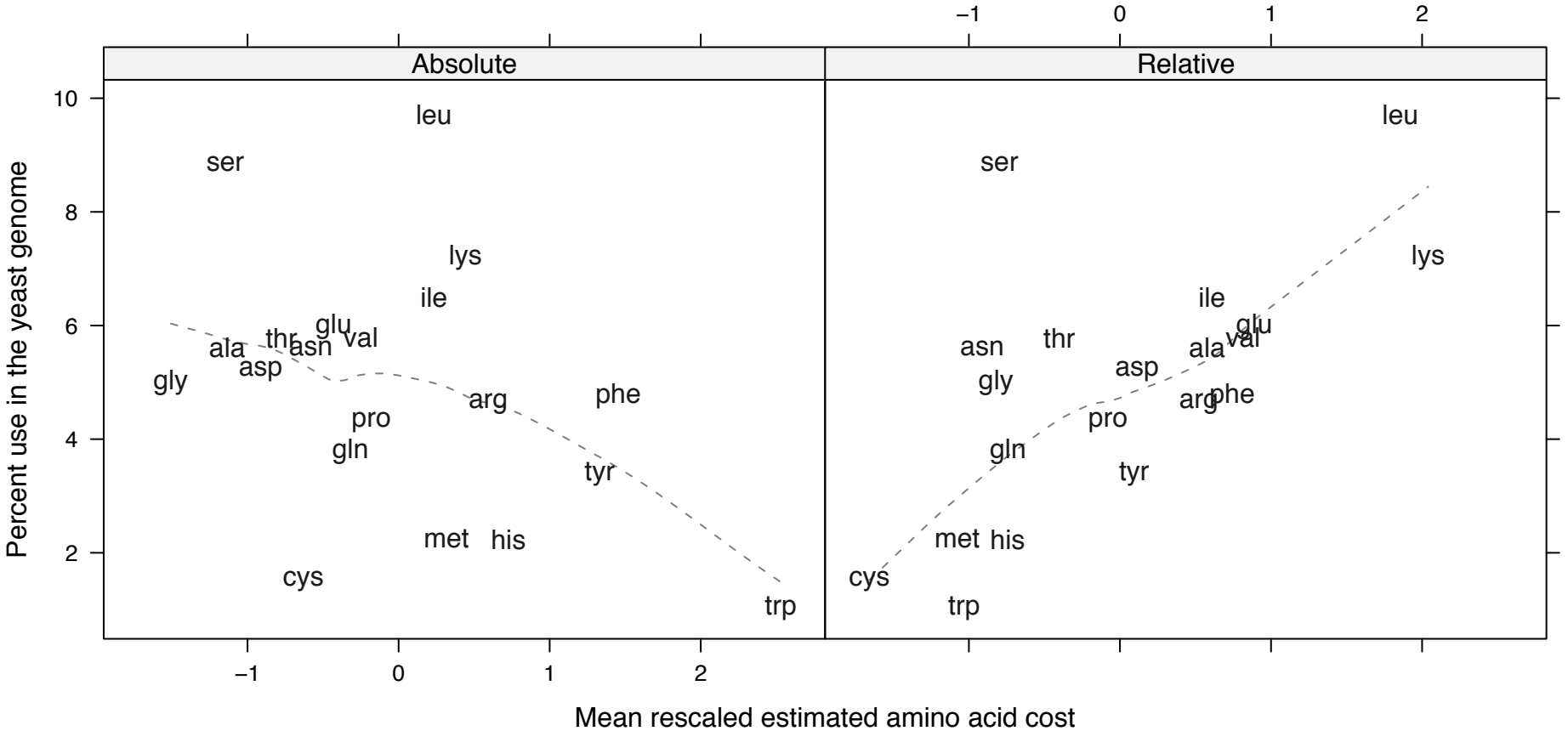

Akashi & Gojobori 2002
Craig & Weber 1998 biosynthetic complexity
Craig & Weber 1998 energy
Molecular Weight
Wagner 2005 fermentative
Wagner 2005 respiratory
Median normalised cost
4
2
0
−2
gly
ser
ala
glu
asp
gln
asn
pro
thr
val
leu
cys
arg
his
lys
met
ile
phe
tyr
trp


Carbon Limited
Nitrogen Limited
Sulphur Limited
Absolute Cost
Relative Cost
Predicted cost
0
5
10
15
20
0.00
0.01
0.02
0.03
gly
ser
ala
glu
asp
gln
asn
pro
thr
val
leu
cys
arg
his
lys
met
ile
phe
tyr
trp


Carbon Limited Relative
Nitrogen Limited Relative
Phosphorus Limited Absolute
Phosphorus Limited Relative
Molecular Weight
Wagner 2005 Fermentative
Craig & Weber 1998 Steps
Akashi & Gojobori 2002
Carbon Limited Absolute
Wagner 2005 Respiratory
Craig & Weber 1998 Energy
Nitrogen Limited Absolute
Sulphur Limited Relative
Sulphur Limited Absolute


Escherichia coli iJR904
Saccharomyces cerevisiae iND750
absolute
A
B
relative
C
D
Rescaled metabolic cost
Number of ATP and NADH molecules used in synthesis
Molecular weight (Da)


Akashi & Gojobori 2002
Craig & Weber 1998 Energy
Craig & Weber 1998 Steps
Wagner 2005 Respiratory
Wagner 2005 Fermentative
Molecular Weight
Carbon Limited Relative
Carbon Limited Absolute
Castrillo 2007 Transcripts
Ghaemmaghami 2003 Proteins
Castrillo 2007 Metabolites
CAI
tRNA
cost
carbon
nitrogen
sulphur
Variable removed
−1
0
1
2
3
4
Variable removal effect, (log10 AIC distance)


Absolute
Relative
Percent use in the yeast genome
Mean rescaled estimated amino acid cost
leu
ser
lys
ile
glu
ala
thr
asn
val
asp
gly
arg
phe
pro
gln
tyr
met
his
cys
trp
asn
thr
ala
val
asp
gly
arg
phe
pro
gln
tyr
met
his
cys
trp
10
8
6
4
2
−1
0
1
2

Amino Acid	Akashi & Gojobori (2002)	Craig & Weber (1998) Energy	Craig & Weber (1998) Steps	Wagner (2005) Fermentative	Wagner (2005) Respiratory	Molecular Weight	*S.cerevisiae* Glucose Limited Absolute	*S.cerevisiae* Glucose Limited Relative	*S.cerevisiae* Nitrogen Limited Absolute	*S.cerevisiae* Nitrogen Limited Relative	*S.cerevisiae* Phosphorus Limited Absolute	*S.cerevisiae* Phosphorus Limited Relative	S S L A
ala	11.7	12.5	1	2	14.5	89.1	2.50005	0.0114702	4.99999	0.02294	-1.3846E-05	-6.3524E-08	
arg	27.3	18.5	10	13	20.5	174.2	6.96484	0.0111925	20.0005	0.0321407	-9.0067E-06	-1.4474E-08	
asn	14.7	4	1	6	18.5	132.1	3.92904	0.00399583	10	0.01017	-8.9739E-05	-9.1265E-08	
asp	12.7	1	1	3	15.5	133.1	3.07215	0.00913963	4.99996	0.0148749	-9.4532E-05	-2.8123E-07	
cys	24.7	24.5	9	13	26.5	121.2	3.80496	0.00025113	5.00167	0.00033011	-0.0018141	-1.1973E-07	
gln	16.3	9.5	2	3	10.5	146.2	4.60794	0.00485677	10.0004	0.0105404	0.00021018	2.2153E-07	
glu	15.3	8.5	1	2	9.5	147.1	4.32146	0.0130422	4.99999	0.01509	-0.0001145	-3.4556E-07	
gly	11.7	14.5	4	1	14.5	75.1	1.53572	0.00445974	5.00001	0.01452	-4.4995E-06	-1.3067E-08	
his	38.3	33	1	5	29	155.2	7.30842	0.00484548	14.9999	0.00994496	-0.00018253	-1.2102E-07	
ile	32.3	20	11	14	38	131.2	6.0362	0.0116318	5.00411	0.00964293	5.6333E-06	1.0855E-08	
leu	27.3	33	7	4	37	131.2	6.03577	0.01789	4.99981	0.0148194	1.8922E-05	5.6086E-08	
lys	30.3	18.5	10	12	36	146.2	6.57164	0.018808	10	0.0286201	5.3382E-06	1.5278E-08	
met	34.3	18.5	9	24	36.5	149.2	6.25041	0.00316896	4.99997	0.00253498	0.00011023	5.5885E-08	
phe	52	63	9	10	61	165.2	9.18754	0.0123021	5.00032	0.00669543	0.00011065	1.4816E-07	
pro	20.3	12.5	4	7	14.5	115.1	4.96506	0.00817745	5.00012	0.00823519	3.8203E-05	6.2921E-08	
ser	11.7	15	3	1	14.5	105.1	2.46805	0.00457577	4.99991	0.00926983	-4.6624E-05	-8.644E-08	
thr	18.7	6	6	9	21.5	119.1	3.42864	0.00656241	5.00001	0.00957003	6.9529E-05	1.3308E-07	
trp	74.3	78.5	12	14	75.5	204.2	11.9646	0.00339795	9.99939	0.00283983	0.00017554	4.9854E-08	
tyr	50	56.5	9	8	59	181.2	8.86861	0.00904599	4.99996	0.00509996	4.9074E-05	5.0055E-08	
val	23.3	25	4	4	29	117.2	4.78575	0.0126631	4.99997	0.0132299	-2.5679E-05	-6.7946E-08	

Cost type	Castrillo et al 2007 Transcripts	Castrillo et al 2007 Proteins	Castrillo et al 2007 Metabolites
S. cerevisiae Absolute	0.388749709	0.405764672	0.781732446
S. cerevisiae Relative	0.383017877	0.407540155	0.875189923
Akashi & Gojobori (2002)	0.3977329	0.404755457	0.804773529
Craig & Weber (1998) Energy	0.416365042	0.399785666	0.834703171
Craig & Weber (1998) Steps	0.374764611	0.403776254	0.866232929
Wagner (2005) Respiratory	0.381891849	0.404720924	0.822070533
Wagner (2005) Fermentative	0.377009278	0.406075083	0.850457048
Molecular Weight	0.422381883	0.405313457	0.766590579