Increased concentration of proteins with growth rate as a result of passive resource redistribution

Uri Barenholz, Leeat Keren, and Ron Milo

Weizmann Institute of Science

February 13, 2015

Abstract

Most proteins show changes in level across growth conditions. Many of these changes seem to be coordinated with the growth rate rather than the specific environment or the protein function. Although cellular growth rates, gene expression levels and gene regulation have been at the center of biological research for decades, there are only a few models using the value of the growth rate to partially predict protein levels.

We present a simple model that predicts a widely coordinated increase in the concentration of many proteins proportionally with the growth rate. The model reveals how passive redistribution of resources, due to active regulation of only a few proteins, can have quantitatively predictable proteome wide effects. Our model provides a potential explanation for why and how such a coordinated response of a large fraction of the proteome to the growth rate arises under different environmental conditions. The simplicity of our model can also be useful by serving as a baseline null hypothesis in the search for active regulation. We exemplify the usage of the model by analyzing the relationship between growth rate and proteome composition for the model microorganism *E.coli* as reflected in two recent proteomics data sets spanning various growth conditions. We find that the cellular concentration of a large fraction of the proteins, and from different cellular processes, increases proportionally with the growth rate. Notably, ribosomal proteins are only a small fraction of this group of proteins. Despite the large fraction of proteins that display this coordinated response, this response only accounts for a relatively small fraction of the overall variability in the proteome across different growth conditions, possibly due to experimental noise. We suggest that, although the concentrations of many proteins change with the growth rate, such changes could be part of a global effect, not requiring specific cellular control mechanisms.

1 Introduction

Many aspects of the physiology of microorganisms change as a function of the growth environment they face. A fundamental system biology challenge is to predict and understand such changes, and specifically, changes in gene expression as a function of the growth environment.

Early on it was found that the expression of some genes is coordinated with growth rate, rather than with the specific environment. Classic experiments in bacteria, by researchers from what became known as the Copenhagen school, have shown that ribosome concentration (inferred from the RNA to protein ratio in cells) increases in proportion to growth rate [28]. The observed increase in concentration has been interpreted to indicate that, given that translation rates and the fraction of active ribosomes remain relatively constant across conditions, a larger fraction of ribosomes out of the proteome is needed in order to achieve faster growth [22, 8, 32]. The search for mechanisms in *E.coli* that underlie this observation yielded several candidates. Specifically, coordination between ribosome production and growth rate was attributed both to the pools of purine nucleotides [12, 9], and the tRNA pools through the stringent response [6, 2]. For a more thorough review see [23].

In the last two decades, with the development of the ability to measure genome-wide expression levels, it was found that changes in gene expression (measured through mRNA levels and promoter-reporter libraries) as a function of growth rate is not limited to ribosomes and ribosomal genes. In *E.coli*, the expression of catabolic and anabolic genes is coordinated with growth rate, and suggested to be mediated by cAMP [27]. In *S.cerevisiae*, it was shown that a surprisingly large fraction of the genome changes its expression levels in response to environmental conditions in a manner strongly correlated with growth rate [17, 10, 5, 11]. Studies examining the interplay between global and specific modes of regulation, suggested that global factors play a major role in determining the expression levels of genes [10, 19, 29, 1, 17, 11]. In *E.coli*, this was mechanistically attributed to changes in the pool of RNA polymerase core and sigma factors [18]. In *S.cerevisiae*, it was suggested that differences in histone modifications around the replication origins [26] or translation rates [10] across conditions may underlie the same phenomenon. Important advancements in *E.coli* were achieved by analyzing measurements of fluorescent reporters through a simplified model of gene expression built upon the empirical scaling with growth rate of different cell parameters (such as gene dosage, transcription rate and cell size)[19]. These studies suggest that the expression of all genes changes with growth rate, with different architectures of regulatory networks yielding differences in the direction and magnitude of these changes.

Despite these advancements, many gaps remain in our understanding of the connection between gene expression and growth rate. Primarily, it is unclear what is the scope of interconnection between gene expression and growth rate. Is it unique to specific groups of genes or is it a more global phenomenon shared across most genes in the genome? What fraction of the variability observed in gene expression patterns across different growth conditions results from active adaptation to the specific condition, and how much results from global, gene and condition-independent, response. Genome-wide proteomic data sets, such as those generated by mass-spectrometry, which probe the proteome composition at different growth rates, offer potential insights into these questions.

In this work we present a parsimonious model, which does not require condition-specific parameters, that quantitatively predicts the relationship between protein abundance and growth rate in the absence of gene-specific changes in regulation. Our model provides a baseline for the behavior of endogenous genes in conditions between which they are not differentially regulated, on top of which different regulatory aspects can be added. The model predicts an increase in protein concentration with growth rate as an emerging property that is the result of passive redistribution of resources, without need for specific regulation mechanisms. In order to exemplify and expore the scope of validity of the model, we analyzed two recently published proteomic data sets of *E.coli* under different growth conditions [31, 13]. We find a statistically significant, coordinated, positive correlation between growth rate and the protein concentration of many genes, from diverse functional groups. However, this response accounts for a relatively small fraction of the total variability of the proteome across the different growth conditions for which these data sets were obtained. Our analysis suggests that experimental noise may underly this relatively poor explanatory power, concluding that more data will be required in order to support or refute the model we present.

2 Results

2.1 Simple considerations predict passively driven increase in the concentration of proteins as a function of the growth rate

What is the simplest way to model the differences in the proteome composition of two populations of cells, one growing in a permissive environment, and the other facing a more challenging growth condition? In an attempt to parsimoniously analyze such differences, we have constructed a minimalistic model that predicts the behavior of non-differentially regulated genes across different growth conditions. Before presenting the model mathematically, we give a brief intuitive depiction.

The model assumes that, under favorable growth conditions, the cell actively down-regulates some proteins that were needed in harsher conditions but not needed in the favorable condition, as illustrated in Figure 1. As a result, the fraction of each of the rest of the proteins out of the proteome is increased compared with the harsh condition, as long as there is no gene-specific regulation. All those proteins increase their levels but are expected to show the same relative ratios among each other after the increase as they were before. The growth rate is also expected to increase in comparison with the harsh condition, as the ratio of bio-synthetic machinery to the rest of the proteome is higher, as depicted in Figure 1B. The growth rate is dependent on the amount of bio-synthesis a cell needs to perform in order to synthesize the proteins needed under its growth environment. To demonstrate the idea concretely, one could think about the down regulation of the lac operon in the presence of Glucose. This situation alleviates the need to transcribe and translate lactose metabolism genes and leads to faster growth.

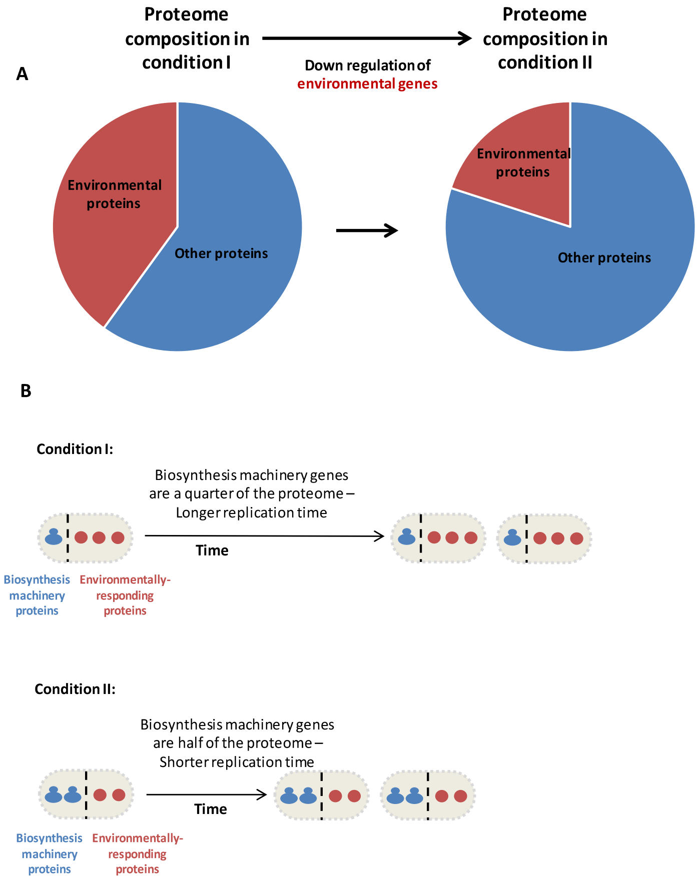
A
Proteome composition in condition I
Down regulation of environmental genes
Proteome composition in condition II
Environmental proteins
Other proteins
Environmental proteins
Other proteins
B
Condition I:
Biosynthesis machinery genes are a quarter of the proteome – Longer replication time
Time
Biosynthesis machinery proteins
Environmentally-responding proteins
Condition II:
Biosynthesis machinery genes are half of the proteome – Shorter replication time
Time
Biosynthesis machinery proteins
Environmentally-responding proteins

Figure 1: A minimalistic model predicts down regulation of environmental genes increases the concentration of other proteins (Panel A). As a result, the ratio of bio-synthesis machinery genes to the rest of the proteome increases, resulting in faster growth (Panel B).

2.1.1 The concentration of a protein is determined by both gene specific control, and global expression machinery availability

For every protein, the model separately considers the resulting concentration as the product of two control mechanisms:

1. Protein/gene specific controls such as the gene associated promoter sequence, 5'-UTRs, ribosomal binding site sequence, and factors affecting the specific expression of the gene such as transcription factors and riboswitches that react with the relevant gene. While some of these controls (such as, for example, the ribosomal binding sites) are static, and

therefore condition independent, others are dynamic and will differ under different environmental conditions (such as transcription factors state).

2. The global availability of bio-synthetic resources in the cell, including availability of RNA polymerase, co-factors, Ribosomes concentration, amino-acids etc. All of these factors can potentially differ across different environmental conditions.

For simplicity, the model refers to the fraction of a specific protein out of the proteome, and not to the concentration of that protein in the biomass. The concentration of a specific protein in the biomass can be calculated given this fraction and the concentration of total protein in the biomass, which is known to be relatively constant [3, 30] (for further discussion see 4.1.2).

According to the model, every gene, under every environmental condition, is given an 'affinity-for-expression' (or 'intrinsic-strength') score that encapsulates its gene-specific control state under the condition considered. We denote the affinity of gene i under growth condition c by $w_i(c)$ (the notion of affinity for expression is not new, and was first suggested in [20]). Our model assumes that the bio-synthetic resources of the cell (Ribosomes, RNA polymerases, etc.) are distributed among the genes according to their affinities under the condition at hand. The notion of affinities can thus reduce the number of parameters needed to predict expression levels markedly.

For example, given that an organism expresses 1000 genes across 10 different growth conditions, one could imagine that characterizing the expression pattern of all genes across all conditions will require 10000 parameters, (the expression level of every gene under every condition), each of which can potentially vary continuously across some predefined range. According to our model, each gene has only a finite set of affinities, possibly only one or two, and thus the expression pattern under every condition can be characterized by only specifying which, out of the total gene-specific small set of possible affinities, each gene acquires under every condition. Moreover, given that the selection of expression level for a given gene is driven by some specific environmental cues, one needs only to know what cues are present at each condition in order to fully specify the affinities all genes acquire under that condition, and thus predict the resulting proteome composition.

The model calculates the resulting protein fraction of a gene, under a specific condition, as the specific affinity of that gene under the condition, divided by the sum of all the affinities of all of the genes under that same condition. Thus, if two genes have the same affinity under some condition, they will occupy identical fractions out of the proteome under that condition. If gene A has twice the affinity of gene B under a given condition, then the fraction protein A occupies will be twice as large as the fraction occupied by protein B under that condition, etc.

This relationship can be simply formulated as follows:

$$p_i(c) = \frac{P_i(c)}{P(c)} = \frac{w_i(c)}{\sum_j w_j(c)} \tag{1}$$

where $p_i(c)$ denotes the fraction of protein i under condition c out of the proteome, $P_i(c)$ denotes the mass of protein i under condition c per cell, $P(c)$

denotes the total mass of proteins per cell under condition c, and the sum, $\sum_j w_j(c)$, is taken over all the genes the cell has.

This equation implies that the observed fraction of a protein is determined by two factors, already, obviously, its own specific affinity that is present in the nominator, but second, and less intuitive and commonly thought of, the affinity of all of the other genes under the growth condition, as reflected by the denominator.

2.1.2 A change in growth condition triggers changes in expression of specific proteins that indirectly affect all of the proteome

Different environmental conditions require the expression of different genes in order to achieve growth. For example, comparing two growth media, one that includes amino-acids, and one that does not, it can be assumed that when amino-acids are present, no need exists for the cell to express amino-acids synthesizing enzymes, whereas when amino-acids are absent, these enzymes must be expressed. Therefore, ideally, the cell will be able to sense the presence or absence of amino-acids in the growth media and, for the amino-acids synthesizing genes, down or up regulate their affinities accordingly. If we now consider some unrelated gene i, whose specific affinity is unaltered between these two conditions, we suggest that its concentration will still change between the two conditions as the affinities of at least some of the other genes (the amino-acids synthesizing enzymes) change, changing the denominator in equation 1 and thus affecting the distribution of resources between all of the expressed genes.

Generalizing this notion, for every group of conditions, one could divide the proteins into those whose intrinsic affinity remains constant across all of the conditions, and to those whose intrinsic affinity changes (meaning their expression is actively regulated by the cell) between at least some of the conditions, as is shown in Figure 1A. An interesting consequence of the formulation in Equation 1 is that proteins whose intrinsic affinities remain constant across different growth conditions, also maintain their relative concentrations across these conditions with respect to each other.

2.1.3 The observed growth rate is an outcome of proteome composition and environmental conditions

While it is sometimes implied that different cellular components are regulated by the growth rate, here we consider the growth rate as an outcome of the environmental conditions that affect the proteome composition. Specifically, the doubling time is proportional to the ratio of the total amount of proteins per cell and the amount of bio-synthesis machinery in that cell. The larger the ratio of total proteins to bio-synthesis proteins is, the longer these bio-synthesis proteins will have to operate in order to duplicate the proteome, and thus the longer the doubling time of the cell will be.

To illustrate this assumption concretely, one could think about the synthesis of polypeptides. If a cell has R actively translating ribosomes, each of which synthesizing polypeptides at a rate of $\eta \approx 20$ amino acids per second, it follows that the cell synthesizes $\approx \eta R$ amino acids per second. If the total amount of protein in that same cell is P (measured in amino acids count), it follows that the time it will take the actively translating ribosomes to synthesize the proteins for an identical daughter cell is $\tau \approx \frac{P}{\eta R}$ (up to a ln(2) factor resulting from the fact that the ribosomes also synthesize more ribosomes during the replication process and that these new ribosomes will increase the total rate of polypeptides synthesis) as is illustrated in Figure 1B.

Theoretically, the fastest doubling time a cell may have is the doubling time achieved when all of the proteome of the cell is the bio-synthetic machinery. We denote this minimal doubling time by T_B. If the bio-synthetic machinery is only half of the proteome, the doubling time will be $2T_B$ etc.

To integrate the notion of total protein to bio-synthetic protein ratio into our model, we make the following simplifying assumption: There is a group of bio-synthetic genes (e.g. genes of the transcriptional and translational machineries) the affinities of which remain constant across different growth conditions, that is, these genes are not actively differentially regulated across different conditions. Furthermore, we assume that the machineries these genes are involved at, operate at relatively constant rates and active to non-active ratios across conditions (which is known to be true for ribosomes [3]).

Under these assumptions we can define this group of bio-synthesis genes, G_B, such that, for every gene that belongs to this group, $k \in G_B$, its affinity, $w_k(c)$ is constant regardless of the condition, c.

$$w_k(c) = w_k \tag{2}$$

To keep our notations short, we will define the (condition independent) sum over all of these bio-synthesis genes as the constant:

$$W_B = \sum_{k \in G_B} w_k$$

As these genes form the bio-synthesis machinery, and according to the assumptions presented above, it follows that the doubling time under a given condition, $\tau(c)$ will be proportional to the ratio of total protein to bio-synthesis protein under that condition, with the proportionality constant being T_B:

$$\tau(c) = T_B \frac{P(c)}{\sum_{k \in G_B} P_k(c)} = T_B \frac{\sum_j w_j(c)}{W_B} \tag{3}$$

Therefore, the model implies that for conditions that require the expression of larger amounts of non-bio-synthetic genes (i.e. higher values in the sum over w_j that are not in W_B), the resulting doubling time will be longer, i.e., the growth rate will be lower.

2.1.4 The concentration of a non-differentially regulated protein is expected to increase with the growth rate

Recalling that the connection between the growth rate and the doubling time is: $g(c) = \frac{\ln(2)}{\tau(c)}$, we now combine Equation 1 with Equation 3 to get that:

$$p_i(c) = \frac{w_i(c)}{\sum_j w_j(c)} = \frac{w_i(c)}{W_B} \frac{W_B}{\sum_j w_j(c)} = \frac{w_i(c)}{W_B} \frac{T_B}{\ln(2)} g(c) \tag{4}$$

Incorporating all the condition-independent constants (W_B, T_B, ln(2)) into one term, A, we get that the predicted fraction of protein i out of the proteome under condition c is:

$$p_i(c) = Aw_i(c)g(c) \tag{5}$$

which implies that, for every two conditions between which gene i maintains its affinity, $(w_i(c_1) = w_i(c_2))$, the fraction protein i occupies out of the proteome scales in the same way as the growth rate does between these two conditions.

To summarize, the simplified model we have constructed predicts that, under no specific regulation, the fraction a protein occupies out of the proteome should scale with the growth rate. A group of such proteins should therefore maintain their relative concentrations across conditions.

2.1.5 Protein degradation differentiates between measured growth rate and biomass synthesis rate

The model we have developed predicts that when the growth rate approaches zero, the concentration of every protein with constant affinity also approaches zero. This approach to zero applies specifically to the biosynthesis genes, that have constant affinities according to our assumptions. However, it is known that the concentration of these proteins, and specifically of ribosomal proteins does not drop to zero when the growth rate approaches zero. Expanding our model to account for the expected effects of proteome degradation affects the predicted concentration of non-differentially regulated proteins at zero growth rate.

Simplifying the analysis by assuming that protein degradation acts on all proteins in the same way, and that it is not dependent on the growth condition, the effect of protein degradation can be understood as follows: at any time, some fraction of the entire proteome is degraded. Therefore, the *observed* growth rate, g, is, in fact, the amount of proteins produced *minus* the amount of proteins degraded. To illustrate, if a cell does not grow, the implication is not that no proteins are produced, but rather that proteins are produced at exactly the same rate as they are degraded.

Integrating this notion into the model means that, where the equations previously referred to the cellular growth rate, g, as the indicator of protein synthesis rate, they should in fact refer to the cellular growth rate plus the degradation rate, as that is the real rate of protein synthesis. Therefore, if we denote by α the degradation rate (assuming for now equal degradation rates for all genes and under all conditions), Equation 5 should be rewritten as:

$$p_i(c) = Aw_i(c)(g(c) + \alpha) \tag{6}$$

This equation predicts linear dependence of the concentration of unregulated proteins on the growth rate, with an intercept with the horizontal axis occurring at minus the degradation rate. Degradation can thus explain why concentrations of non-differentially regulated proteins do not drop to zero when the growth rate is zero.

2.1.6 Slower biological processes rates at slower growth affect the relation between proteome composition and growth rate

The simplified model assumes that the doubling time is proportional to the ratio of total protein to bio-synthetic protein. This assumption fails if the rate at which each biosynthetic machine operates changes across conditions. Replacing this assumption by a dependence of bio-synthesis rate with growth rate (such that, the faster the growth, the faster the synthesis rates, per machine), will affect the resulting predictions as well. Slower bio-synthesis rates under slower growth rates imply that, compared with the model prediction, higher fraction of bio-synthesis proteins is needed to achieve a given growth rate. Thus, lower synthesis rates under slower growth rates will be reflected by a lower slope and higher interception point for non-regulated proteins than those predicted by the constant-rate version of the model.

2.2 Analysis of proteomic data sets

To assess the extent to which the predictions of our model are reflected in actual proteome compositions, we analyzed two published proteomics data sets of *E.coli*, [31] and [13]. These data sets use mass spectrometry to evaluate the proteomic composition of *E.coli* under 5 different growth rates using a chemostat, in [31], and 19 different growth conditions, spanning both different carbon sources and chemostat-controlled growth rates, in [13]. The data set from [13] contains more conditions than those analyzed below, see section 4.1.3 for further details.

2.2.1 A large fraction of the proteome is positively correlated with growth rate

In each data set, the growth rate and proteome composition were measured for several conditions. We calculated the Pearson correlation of every protein with the growth rate, conducting the analysis separately for each data set. A histogram of the distribution of the correlations is shown in Figure 2. We find that more than a third of the proteins (628 out of 1656 measured in the data set from [13], hereafter referred to as H, and 378 out of 919 in the data set from [31], hereafter referred to as V) have a strong positive (> 0.5) correlation with the growth rate. Further discussion of the choice of threshold for defining strong correlation with the growth rate is in section 6.1. Further comparison and analysis of the causes underlying the differences between the two data sets as reflected in Figure 2 are in section 6.2. Notably, in both data sets, the proteins that have a high correlation with the growth rate are involved in many and varied cellular functions and span different functional groups (See tables S1 and S2).

Previous studies already found that ribosomal proteins are strongly positively correlated with growth rate [25, 15, 18]. Our analysis agrees with these findings as we find the concentration of the vast majority of the 56 ribosomal proteins to be strongly positively correlated with growth rate. However, we also find that the group of proteins strongly positively correlated with growth rate includes many more proteins than the ribosomal proteins. Importantly, the proteins that we find to be strongly positively correlated with growth rate are not generally expected to be co-regulated, and their behavior does not seem to be the result of any known transcription factor or regulation cluster response.

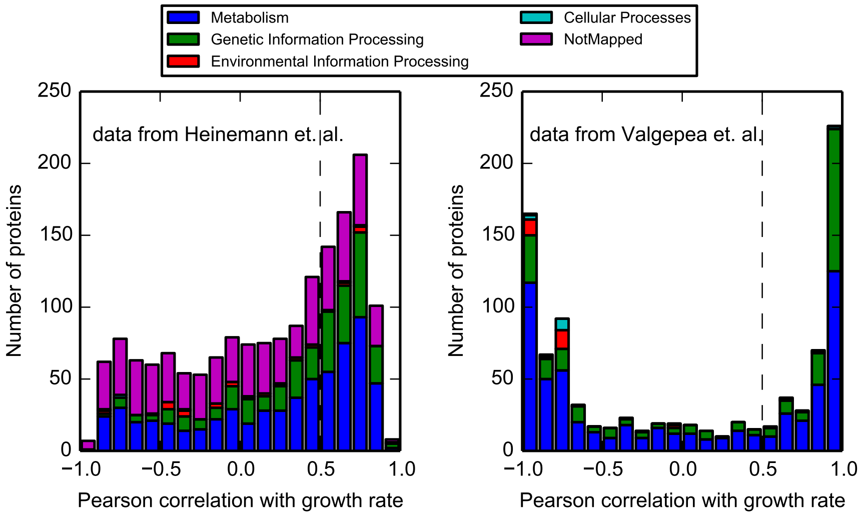

Figure 2: A strong positive Pearson correlation of the concentration with the growth rate is observed for a large fraction of the proteins in the two data sets analyzed (thresholds used for high correlation are marked in dashed lines and further discussed in 6.1). These proteins span many functional groups.

2.2.2 Proteins positively correlated with growth rate share a similar response

Following the identification of the group of proteins strongly positively correlated with growth rate, we examined how similar is the behavior with growth rate for these different proteins. We note that similar correlation with growth rate for different proteins does not imply that such proteins share the same scaling with growth rate, that is, they may have very different slopes or fold changes with an increasing growth rate.

In order to compare the responses of different proteins across conditions, we therefore, for every protein, divided its concentration under every condition by its average concentration across all of the conditions (see 4.1.1 for further details). This normalized concentration across conditions represents the concentration of the specific protein under every condition, relative to its mean concentration across all conditions. We note that, under this metric, sharing similar responses among a group of proteins implies that proteins in that group maintain their relative ratios, ratios that are determined by the average concentration of each of these proteins across the different environmental conditions. We refer to proteins that share a similar normalized response across different conditions as being *coordinated* or *coordinately regulated*. Note that our model suggests a mechanism for this coordinated expression changes that is not based on shared transcription factors but rather is a result of passive redistribution of resources.

To assess the coordination between the proteins that were found to be strongly positively correlated with growth rate we therefore calculated the slope of a linear regression line for the normalized concentration vs. the growth rate for every one of these proteins and plotted the result in Figure 3. The resulting distribution reveals that, not only is a significant fraction of the proteome strongly positively correlated with the growth rate, but that this response is also coordinated between the different proteins.

Quantitatively, a protein with a normalized slope of 0.5 will change in concentration from $\frac{7}{8}$ of its mean concentration at the slowest growth rate measured ($\mu \approx 0.1$), to $\frac{9}{8}$ of its mean concentration at the fastest growth rate ($\mu \approx 0.6$), whereas a protein with a normalized slope of 2 will have concentrations in the range $\frac{1}{2}$ to $\frac{3}{2}$ of its mean concentration across the same range of growth rates. We note that such changes are relatively small compared with the known levels of noise in MS whole proteome measurements. Therefore, the ratio between proteins with such slopes of 0.5 and 2 lies in the relatively narrow range of $\frac{3}{4}$ to $\frac{7}{4}$ of the ratio between their mean concentrations, implying their relative amounts will change by at most just over 2-fold over the range of growth rates measured.

Our results, showing that a large number of proteins maintain their relative concentrations across different growth conditions thus extend the scope of similar results obtained for *S.cerevisiae* in [17] and for expression levels in *E.coli* under stress conditions in [16].

Next we examined how the response of the strongly correlated proteins relates to the well-studied response of ribosomes concentration. To that end, we performed the same analysis of slopes, restricting it to ribosomal proteins alone, as is shown by the stacked green bars in Figure 3. We find that, on average, strongly correlated proteins scale in the same way as ribosomal proteins do (see also Figure S5), implying that the observed response of ribosomal proteins to growth rate is not unique and is coordinated with a much larger fraction of the proteome, thus encompassing many more cellular components.

To investigate the effect of noise in determining the range of slopes observed, we calculated, for every protein, the standard error with respect to the regression line that best fits its concentrations. Given these standard errors we generated the expected distribution of slopes that would result by conducting our analysis on proteins that share a single, identical slope, but with the calculated noise in measurement. The expected distribution is shown in gray line in Figure 3.

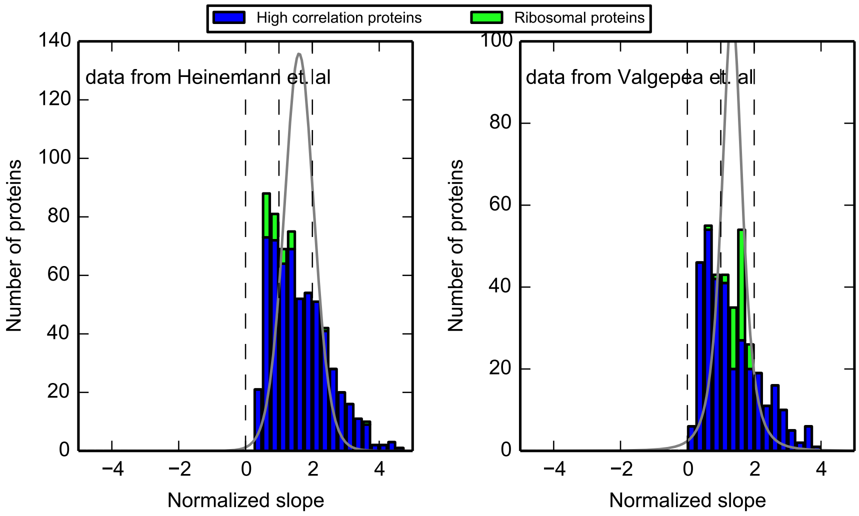

Figure 3: Histogram of the slopes of regression lines for every protein that is highly correlated with growth, for the two data sets analyzed (628 and 378 proteins in the left and right panels respectively). Ribosomal proteins are stacked in green on top of the non ribosomal proteins, marked in blue. Proteins concentrations were normalized to account for differences in slopes resulting from differing average concentrations (See text and section 4.1.1). The expected distribution of slopes given the individual deviations of every protein from a linear regression line, assuming all proteins are coordinated, is plotted in gray. Left panel - data from [13], right panel - data from [31]. High correlation proteins share similar normalized slopes, implying they are coordinated, maintaining their relative ratios across conditions (see text for further details). Ribosomal proteins, shown in green, scale with growth rate in a manner similar to the rest of the high correlation proteins (see also Figure S5).

2.2.3 Changes in the proteome across environmental conditions are dominated by proteins that are positively correlated with growth rate

Lastly, we assessed the significance of the positive correlation of proteins with growth rate, out of the total change in proteome composition across conditions. To that end, we summed the concentrations of all of the proteins that are strongly correlated with growth rate across the conditions measured and plotted their total concentration against the growth rate in Figure 4. Both data sets show that the concentration of these proteins change ≈ 2 fold across an ≈ 5 fold change in the growth rate under the different growth conditions. Moreover, most of the variability of the total concentration of these proteins can be explained by the growth rate (R^2 of 0.8 in the data set from [13] and > 0.99 in the data set from [31]). For further analysis of the differences between the two data sets see section 6.2. Importantly, the strongly correlated proteins form a large fraction of the proteome, exceeding 50% of the proteome, mass-wise, at the higher growth rates measured. Thus, when considering the changes in proteome composition across conditions, we find that, at higher growth rates, more than 50% of the proteome composition is affected by the coordinated response of the same group of proteins with growth rate.

However, despite the magnitude of this phenomena, when calculating the fraction of the total variability in the proteome that is accounted for by this linear response, we observe that only $\approx$ 9% of the change in the proteome composition across conditions results from linear scaling with growth rate of the proteins that share a coordinated, positive response with the growth rate in the data set from [13] and this fraction is even lower in the data set from [31] as can be seen in Figure S1. A lot of this seeming difference results from the fact that a single linear response captures only a fraction of the variability of these proteins across the different growth conditions, possibly due to measurement noise. Further discussion of the fraction of variability explained can be found in 6.1. The noise in current whole proteome measurement techniques make it difficult to distinguish between proteins that scale coordinately, as is predicted by our model, and proteins that scale differentially, but within measurement uncertainty. Thus, it is unclear to what extent the effect we predict affects actual protein concentrations versus their possible individual up regulation with growth rate. We expect future improvements in the accuracy of whole proteome measurements to quantitatively reveal the importance of passive coordinated scaling with growth rate in shaping the proteome composition. These coming improvements in accuracy will enable better testing of the scope and validity of the model presented here.

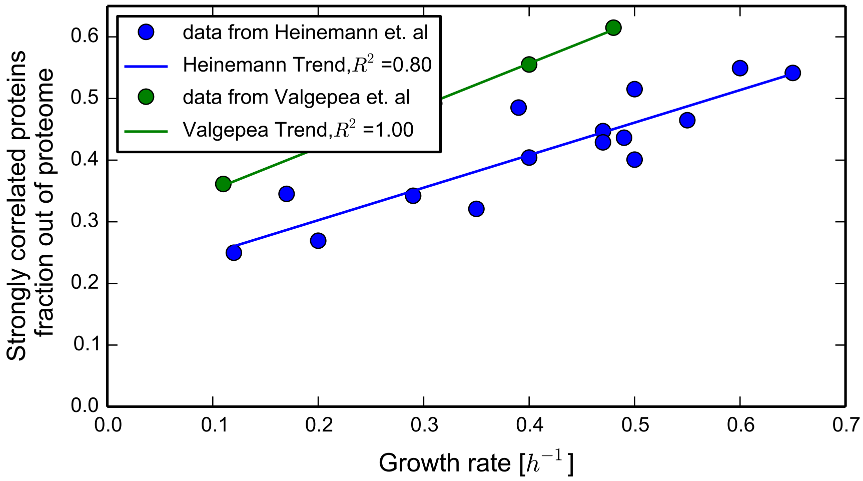

Figure 4: Fraction of the proteome occupied by proteins that are strongly positively correlated with growth rate. The accumulated sum of the proteins that are strongly positively correlated with growth rate (defined as having a correlation above 0.5), as a fraction out of the proteome, with linear regression lines is shown. These proteins form a large fraction ($\geq 50\%$) out of the proteome at higher growth rates. The accumulated concentration of the strongly correlated proteins doubles as the growth rate changes by about 5-fold. Assuming constant degradation rates, the trend lines correspond to protein half life times of ≈ 1.7 hours.

2.2.4 The statistical features we find do not naturally rise in randomized data sets

We performed two tests to verify that the trends we find, namely, the large fraction of proteins with a strong correlation with growth rate, the coordination among these proteins, their large accumulated fraction out of the proteome and the fraction of variability explained by a single linear regression approximation of their concentrations are all non-trivial characteristics of the data set that do not naturally rise in randomly generated data but that do arise if our model is correct. To this extent we repeated our analysis on two simulated data sets:

- A data set at which the amount of every protein was shuffled across the different conditions.
- A synthetic data set assuming half the proteome being perfectly coordinated and linearly dependent on growth rate, with the parameters we find in our analysis, and the other half having no correlation with growth rate, and with a simulated normally distributed measurement noise of 20%.

We find that in the shuffled sets the number of proteins being significantly positively correlated with growth rate is much smaller than found in the real data sets (43 vs. 628 in the data set from [13] and 152 vs. 378 in the data set from [31]) as is shown in Figure 5. As a consequence, these proteins now occupy a much smaller fraction out of the proteome mass-wise ($< 5\%$ and 20% on average across conditions vs. 35% and 50% in the data sets from [13] and [31] respectively) as is shown in Figure 6. Finally, the fraction of variability in the proteome that can be explained by a single linear regression to these proteins is smaller for the data set from [13] than that obtained for the real data set (1% vs. 9% for a threshold of $R \geq 0.5$), as is seen in Figure S6.

We find that the simulated (second) set does display similar characteristics to those we find in the real data, confirming that if, indeed, our model is valid, experimental measurements would overlap with those that we obtained.

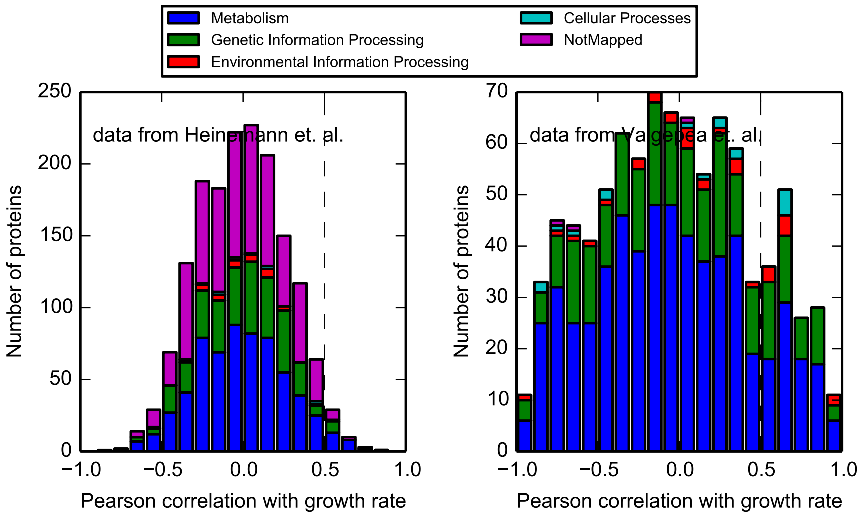

Figure 5: The Pearson correlation with growth rate of the concentration of proteins for a shuffled concentration across conditions. In each data set, the amount of every protein was shuffled across the different growth conditions. The shuffling procedure creates a data set that has much fewer proteins that are significantly positively correlated with growth rate, compared with the original data sets.

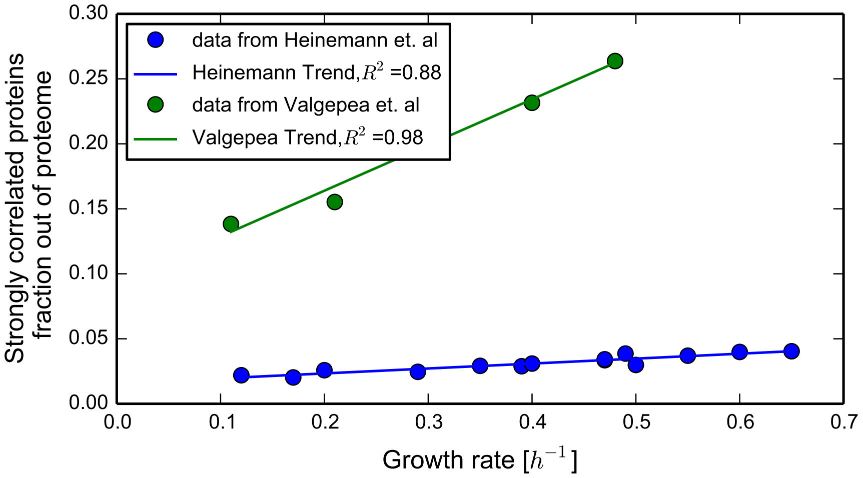

Figure 6: The fraction out of the proteome of the proteins that are highly correlated with growth rate in shuffled data sets. In each data set, the amount of every protein was shuffled across the different growth conditions. As much fewer proteins are strongly positively correlated with the growth rate in the shuffled data set, their total fraction out of the proteome is also much smaller compared with the original data sets.

3 Discussion

We construct a parsimonious model connecting protein concentration levels and the growth rate as an outcome of the limited bio-synthesis resources of cells. We re-introduce the notion of intrinsic affinity for expression, first presented in [20], and rarely used ever since, as a key determinant for the differences in expression of different proteins under a given growth condition. We show that integrating the notion intrinsic affinity for expression with the limited bio-synthesis capacity of cells results in a simple mechanism predicting increased concentration of many proteins with the growth rate, without assuming the existence of specific transcription factors regulation.

The framework we present emphasizes the importance of accounting for global factors, that are reflected in the growth rate, when analyzing gene expression and proteomics data. Specifically, we suggest that the default response of a protein (that is, the change in the observed expression of a protein, given that no specific regulation was applied to it) is to linearly increase with growth rate. We point out that, as non-differentially regulated proteins maintain their relative abundances, one can deduce the parameters of the linear increase with growth rate of any non-differentially regulated protein by observing the scaling of other such proteins and fixing the ratio between the protein of interest and the reference proteins, as is demonstrated in Figure S4.

We analyze two recent whole proteome data sets to explore the scope and validiy of our model. We characterize a coordinated response in *E.coli* between many proteins and the growth rate. This response spans proteins from various functional groups and is not related to the specific medium of growth. A similar phenomena is observed for *S.cerevisiae* as was reported in [17] and may thus be conserved across various organisms and domains of life. Our analysis suggests that, while changes in the proteome composition may seem complex, for a large number of proteins and under many conditions, they can be attributed to a linear, coordinated, increase with growth rate, at the expense of other, down-regulated proteins. The well studied scaling of ribosomes concentration with growth rate can be considered one manifestation of the more general phenomena we describe here. We find that this response is not unique to ribosomal proteins but is, in fact, shared with many other proteins spanning different functional groups.

Interestingly, our model suggests that a linear correlation between ribosomal proteins and the growth rate might be achieved without special control mechanisms. Nonetheless, many such mechanisms have been shown to exist [23]. We stress that the existence of such mechanisms does not contradict the model. Mechanisms for ribosomal proteins expression control may still be needed to achieve faster response under changing environmental conditions or a tighter regulation to avoid unnecessary production and reduce translational noise. Furthermore, such mechanisms may be crucial for synchronizing the amount of rRNA with ribosomal proteins as the two go through different bio-synthesis pathways. Nevertheless, the fact that many non-ribosomal proteins share the same response as ribosomal proteins do, poses interesting questions regarding the scope of such control mechanisms, their necessity and the trade-offs involved in their deployment.

3.1 Relation to previous studies

The findings in this study support and broaden the findings in other recent studies. Specifically, for *S.cerevisiae* a few recent studies found that the concentration of the majority of the proteins is coordinated across conditions [17, 10, 2] and increases with growth rate. In principle, the model we suggest here can be applied to any exponentially growing population of cells and may thus also serve as a potential explanation for the phenomena observed in these studies and others.

Other recently published studies in *E.coli* have suggested different models and in some cases have results and predictions that do not coincide with those presented in this study. Notably, in [19] the opposite behavior for unregulated genes is predicted. A few differences can explain this seeming discrepancy. The modeling in [19] relies on data collected under different growth rates than those observed in our work. The predictions of the model are based on the deduced dependence of various bio-synthesis process rates and physiological properties of the cells on the growth rate, properties that are, in turn, used to calculate the expected protein concentration for unregulated proteins under the different growth rates. This approach is markedly different than the approach we take, which assumes relatively small changes in bio-synthetic rates as a function of growth rate and focuses on the limited bio-synthesis resources as the main driver of changes in the resulting concentration of proteins. As the model in [19] was only tested against a handful of proteins, it is impossible to decide which of the two models better describes the global effects of growth rate on the proteome composition.

Many studies monitored the ribosome concentration in cells and its interdependence with growth rate [29, 28, 4, 32, 3]. While in all of these studies a linear dependence of ribosome concentration with growth rate was observed, in some cases different parameters were found to describe this linear dependence, compared with the observations in our study. A discussion of various reasons that may underlie these differences is given in section 6.4. Conducting similar analysis on the data sets used in this study reveals that, while a linear relation exists, it is not unique to ribosomal proteins but is in fact shared among many more genes. Furthermore, the linear dependence slope and explained variability of concentration levels of proteins explained by linear correlation with growth rate is similar among the ribosomal proteins versus all the proteins with high correlation with the growth rate as is shown in Figure S5.

The expected availability of increasing amounts of whole proteome data sets, with higher accuracy levels, will enable further investigation of the details of cellular resource distribution. The analysis of such future data sets will shed more light on the relative roles of carefully tuned response mechanisms versus global, passive effects in shaping the proteome composition under different growth environments.

4 Materials and Methods

4.1 Data analysis tools

All data analysis was performed using custom written software in the Python programming language. The data analysis source code is available through github at: http://github.com/uriba/proteome-analysis Analysis was done using SciPy [24], NumPy [7] and the Pandas data analysis library [21]. Charts where created using the MatPlotLib plotting library [14].

4.1.1 Normalizing protein concentrations across conditions

Our analysis aims at identifying proteins that share similar expression patterns across the different growth conditions. For example, consider two proteins, A and B measured under two conditions, c_1 and c_2. Assume that the measured fractions out of the proteome of these two proteins under the two conditions were 0.001 and 0.002 for A under c_1 and c_2 respectively, and 0.01 and 0.02 for B under c_1 and c_2 respectively. These two proteins therefore share identical responses across the two conditions, namely, they double their fraction in the proteome in c_2 compared with c_1.

The normalization procedure scales the data so as to reveal this identity in response. Dividing the fraction of each protein out of the proteome by the average fraction of that protein across conditions yields the normalized response. It the example, the average concentration of A across the different conditions is 0.0015 and the average concentration of B is 0.015. Thus, dividing the concentration of every protein by the average concentration across conditions of that same protein yields:

$$A'_{c_1} = \frac{A_{c_1}}{\bar{A}} = \frac{0.001}{0.0015} = \frac{2}{3} = \frac{0.01}{0.015} = \frac{B_{c_1}}{\bar{B}} = B'_{c_1}$$

for c_1 and:

$$A'_{c_2} = \frac{A_{c_2}}{\bar{A}} = \frac{0.002}{0.0015} = \frac{4}{3} = \frac{0.02}{0.015} = \frac{B_{c_2}}{\bar{B}} = B'_{c_2}$$

for c_2 showing A and B share identical responses across c_1 and c_2.

The general normalization procedure thus divides the concentration of protein i under condition c, $p_i(c)$ by the average concentration of protein i across all of the conditions in the data set, $\bar{p}_i$, to give the normalized concentration under condition c, $p'_i(c) = \frac{p_i(c)}{\bar{p}_i}$.

This normalization procedure has been applied prior to calculating the slopes of the regression lines best describing the change in fraction out of the proteome of every protein as a function of the growth rate. Furthermore, when analyzing the variability explained by linear regression on the sum of concentrations of all proteins presenting a high correlation with the growth rate, the same normalization procedure was made in order to avoid domination by the high abundance of a few proteins in that group.

4.1.2 Calculation of protein concentration

In this study, we use the mass ratio of a specific protein to the mass of the entire proteome, per cell, as our basic measure for the bio-synthetic resources a specific protein consumes out of the bio-synthetic capacity of the cell. We find this measure to be the best representation of the meaning of a fraction a protein occupies out of the proteome. However, we note that if initiation rates are limiting (e.g. if RNA polymerase rather than ribosomes become limiting), and not elongation rates, then using molecule counts ratios (the number of molecules of a specific protein divided by the total number of protein molecules in a cell) rather than mass ratios may be a better metric. We compared these two metrics and, while they present some differences in the analysis, they do not qualitatively alter the observed results.

There are different, alternative ways to assess the resources consumed by a specific protein out of the resources available in the cell. On top of the measures listed above, one could consider either the total mass or molecule count of a specific protein out of the biomass, rather than the proteome, or out of the dry weight of the cell, both of which vary with the ratio of total protein to biomass or dry weight which was neglected in our analysis. Moreover, one can consider specific protein mass or molecule count per cell, thus reflecting changes in cell size across conditions. Our analysis focuses on the relations between different proteins and resource distribution inside the proteome, and thus avoids such metrics.

4.1.3 Filtering out conditions from the Heinemann data set

The [13] data set contains proteomic data measurements under 19 different environmental conditions. However, some of these conditions violate some of the assumptions we make in our model, assumptions that are at the heart of the connection between the proteome composition and growth rate. Specifically, our model assumes constant ribosome translation rate (and bio-synthesis rates in general) which are known to vary with temperature. We therefore excluded the $42°C$ condition from our analysis. Additionally, our model assumes exponential growth, implying that measurements taken at stationary phase are expected to differ from simple extrapolation of the model to zero growth rate, the two measurements of stationary phase proteomics were thus also excluded.

Out of the conditions measured in the [13] data set, growth in LB media presented a much faster growth rate than the rest of the conditions measured (1.6 vs a range of $0.12 - 0.65$ for the other conditions). This asymmetry in the distribution of growth rates caused LB growth to dominate the analysis due to its effect on the skewness of the distribution of growth rates ($\gamma_1 = -0.4$ for the growth rates excluding LB vs. $\gamma_1 = 2.4$ with LB) reducing the statistical power of the other conditions. While including the data on growth in LB does not qualitatively change the observed results, such analysis is much less statistically robust. We have therefore omitted LB growth data in the main analysis. We present the analysis with growth data on LB in section 6.3.

Including LB growth results in a much smaller set of proteins with a strong positive correlation with growth, as many of the proteins in that group in the slower conditions get down-regulated in LB, significantly reducing their Pearson correlation with growth rate. For example, the Pearson correlation with growth rate of gapA, involved in glycolisys, drops from 0.73 to 0.35 when LB is included. Another such example is glyA, involved in serine and threonine metabolism, that has a correlation with growth rate of -0.12 when LB is included in the analysis vs. a correlation of 0.7 without it.

On the other hand, the proteins that remain strongly positively correlated with growth rate when LB is included in the analysis show a higher correlation compared with the analysis shown without LB. Furthermore, despite the decrease in the number of proteins that are strongly positively correlated with growth when LB is included in the analysis (532 vs. 628), these proteins occupy $> 50\%$ of the proteome under LB due to the increase in their concentration with growth rate.

5 Acknowledgments

We kindly thank Matthias Heinemann for letting us use an unpublished proteomics data set. We thank Uri Alon, Arren Bar Even, Avi Flamholz, Stephan Klumpp, Elad Noor and Eran Segal for fruitful discussions and enlightening comments on the manuscript.

References

[1] Sara Berthoumieux, Hidde de Jong, Guillaume Baptist, Corinne Pinel, Caroline Ranquet, Delphine Ropers, and Johannes Geiselmann. Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Molecular systems biology*, 9(634):634, jan 2013.

[2] M. J. Brauer, C. Huttenhower, E. M. Airoldi, R. Rosenstein, J. C. Matese, D. Gresham, V. M. Boer, O. G. Troyanskaya, and D. Botstein. Coordination of Growth Rate, Cell Cycle, Stress Response, and Metabolic Activity in Yeast, 2008.

[3] H Bremer and PP Dennis. Modulation of chemical composition and other parameters of the cell by growth rate. *Escherichia coli and Salmonella ...*, 1987.

[4] Hans Bremer and Patrick P Dennis. Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate. (122).

[5] Juan I Castrillo, Leo A Zeef, David C Hoyle, Nianshu Zhang, Andrew Hayes, David C J Gardner, Michael J Cornell, June Petty, Luke Hakes, Leanne Wardleworth, Bharat Rash, Marie Brown, Warwick B Dunn, David Broadhurst, Kerry O'Donoghue, Svenja S Hester, Tom P J Dunkley, Sarah R Hart, Neil Swainston, Peter Li, Simon J Gaskell, Norman W Paton, Kathryn S Lilley, Douglas B Kell, and Stephen G Oliver. Growth control of the eukaryote cell: a systems biology study in yeast. *Journal of biology*, 6(2):4, 2007.

[6] Dipankar Chatterji and Anil Kumar Ojha. Revisiting the stringent response, ppGpp and starvation signaling, 2001.

[7] NumPy Community. NumPy Reference. *October*, 1(October):1–1146, 2011.

[8] Patrick P Dennis, Mans Ehrenberg, and Hans Bremer. Control of rRNA Synthesis in Escherichia coli : a Systems Biology Approach Control of rRNA Synthesis in Escherichia coli : a Systems Biology Approach †. 68(4), 2004.

[9] T Gaal, M S Bartlett, W Ross, C L Turnbough, and R L Gourse. Transcription regulation by initiating NTP concentration: rRNA synthesis in bacteria. *Science (New York, N.Y.)*, 278(5346):2092–2097, 1997.

[10] A P Gasch, P T Spellman, C M Kao, O Carmel-Harel, M B Eisen, G Storz, D Botstein, and P O Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell*, 11(12):4241–4257, 2000.

[11] Luca Gerosa, Karl Kochanowski, Matthias Heinemann, and Uwe Sauer. Dissecting specific and global transcriptional regulation of bacterial gene expression. *Molecular systems biology*, 9(658):658, 2013.

[12] R L Gourse, T Gaal, M S Bartlett, J A Appleman, and W Ross. rRNA transcription and growth rate-dependent regulation of ribosome synthesis in Escherichia coli. *Annual Review of Microbiology*, 50(1):645–677, 1996.

[13] Matthias Heinemann and A. Schmidt. Under review. 2015.

[14] John D Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, 9(3):90–95, 2007.

[15] J L Ingraham, O Maalø e, and F C Neidhardt. *Growth of the bacterial cell*. Sinauer Associates, 1983.

[16] Kunihiko Kaneko, Chikara Furusawa, and Tetsuya Yomo. Universal relationship in gene-expression changes for cells in steady-growth state. page 7, jul 2014.

[17] Leeat Keren, Ora Zackay, Maya Lotan-Pompan, Uri Barenholz, Erez Dekel, Vered Sasson, Guy Aidelberg, Anat Bren, Danny Zeevi, Adina Weinberger, Uri Alon, Ron Milo, and Eran Segal. Promoters maintain their relative activity levels under different growth conditions. *Molecular systems biology*, 9(701):701, 2013.

[18] Stefan Klumpp and Terence Hwa. Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 105(51):20245–50, dec 2008.

[19] Stefan Klumpp, Zhongge Zhang, and Terence Hwa. Growth rate-dependent global effects on gene expression in bacteria. *Cell*, 139(7):1366–75, dec 2009.

[20] Ole Maaloe. An analysis of bacterial growth. *Dev Biol Suppl*, 3:33–58, 1969.

[21] Wes McKinney. pandas: a Foundational Python Library for Data Analysis and Statistics. In *Python for High Performance and Scientific Computing*, pages 1–9, 2011.

[22] Frederick C. Neidhardt. Bacterial Growth: Constant Obsession with dN/dt. *J. Bacteriol.*, 181(24):7405–7408, dec 1999.

[23] Masayasu Nomura, Richard Gourse, and Gail Baughman. REGULATION OF THE SYNTHESIS OF RIBOSOMES AND RIBOSOMAL COMPONENTS. *Annual review of biochemistry*, pages 75–117, 1984.

[24] Travis E Oliphant. SciPy: Open source scientific tools for Python. *Computing in Science and Engineering*, 9:10–20, 2007.

[25] S Pedersen, P L Bloch, S Reeh, and F C Neidhardt. Patterns of protein synthesis in E. coli: a catalog of the amount of 140 individual proteins at different growth rates. *Cell*, 14(1):179–190, 1978.

[26] Birgitte Regenberg, Thomas Grotkjaer, Ole Winther, Anders Fausbø ll, Mats Akesson, Christoffer Bro, Lars Kai Hansen, Sø ren Brunak, and Jens Nielsen. Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in Saccharomyces cerevisiae. *Genome biology*, 7(11):R107, 2006.

[27] Alok J Saldanha, Matthew J Brauer, and David Botstein. Nutritional homeostasis in batch and steady-state culture of yeast. *Molecular biology of the cell*, 15(9):4089–4104, 2004.

[28] M Schaechter. Dependency on medium and temperature of cell size and chemical composition during balanced growth of Salmonella typhimurium. *Journal of general ...*, 19:592–606, 1958.

[29] Matthew Scott, Carl W Gunderson, Eduard M Mateescu, Zhongge Zhang, and Terence Hwa. Interdependence of cell growth and gene expression: origins and consequences. *Science (New York, N.Y.)*, 330(6007):1099–102, nov 2010.

[30] Matthew Scott, Stefan Klumpp, Eduard M Mateescu, and Terence Hwa. Emergence of robust growth laws from optimal regulation of ribosome synthesis. *Molecular systems biology*, 10(8):747, jan 2014.

[31] Kaspar Valgepea, Kaarel Adamberg, Andrus Seiman, and Raivo Vilu. Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. *Molecular BioSystems*, 9(9):2344–58, sep 2013.

[32] Alon Zaslaver, Shai Kaplan, Anat Bren, Adrian Jinich, Avi Mayo, Erez Dekel, Uri Alon, and Shalev Itzkovitz. Invariant distribution of promoter activities in Escherichia coli. *PLoS computational biology*, 5(10):e1000545, oct 2009.

6 Supplementary figures and data

6.1 Threshold selection for defining strong correlation with growth rate

The data we use includes the concentrations of proteins under different growth conditions, and the growth rate for every condition. We select a threshold correlation with growth rate to define the group of highly positively correlated with growth rate proteins.

We calculate the explained variability by the growth rate, given a threshold, by taking the difference between the total variability of the group of proteins with a correlation higher than the threshold, and the variability remaining, when assuming these proteins scale with the growth rate according to the calculated linear response. Dividing the explained variability by the total variability of the entire data set quantifies what fraction of the total variability in the proteome is explained by considering a coordinated linear scaling with growth rate for all the proteins with a correlation with growth rate higher than the threshold.

The choice of threshold is thus influenced by two contradicting factors. Choosing a low threshold results in defining many proteins as being highly positively correlated with growth rate. In this case, the correlation with growth rate of these proteins spans a large range. Therefore, applying a linear regression trend to the sum of these proteins only accounts for a small fraction of the variability of them and, as a consequence, only accounts for a small fraction of the total variability of the proteome.

On the other hand, choosing a high correlation threshold results in defining only a small number of proteins as being highly positively correlated with growth rate. A common linear regression line may thus explain a large fraction of the variability for the chosen proteins but, as their number is small, will only account for a small fraction of the total variability of the proteome.

For simplicity, we chose a threshold value of 0.5 for the two data sets analyzed in this study. Figure S1 shows how the choice of threshold affects the fraction of explained variability in the proteome by the linear dependence on growth rate of the proteins that have a correlation with growth rate that is higher than the threshold (blue line). The figure also shows the fraction of proteins that have a correlation with growth rate that is higher than the threshold out of the proteome (red line), and the fraction of explained variability by linear regression for these proteins (green line).

The optimal threshold is defined as the threshold maximizing the fraction of total variability explained (maximum of the blue line). As can be seen in Figure S1, our choice of threshold of 0.5 is relatively close to the optimum value that is 0.25 for the data set from [13], and 0.8 for the data set from [31]. Moreover, as Figure S1 illustrates, the different plotted statistics do not change markedly due to this sub-optimal choice of threshold and thus this choice does not affect our results significantly.

As different proteins have very different average concentrations, the aforementioned calculation may be biased towards proteins with higher average concentrations. To avoid this effect, the analysis presented was performed on the normalized concentrations as defined in 4.1.1.

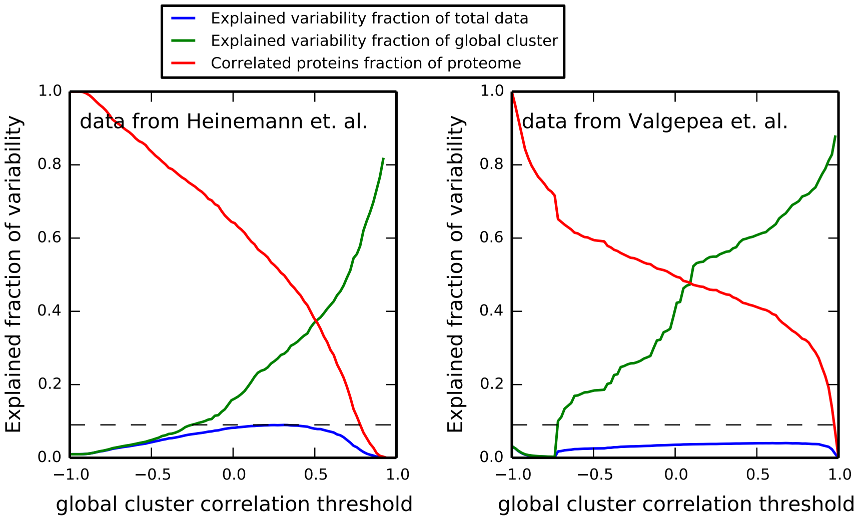

Figure S1: Statistics on the explained variability in the normalized data set as a function of the threshold used for defining strong correlation with growth rate. An optimal threshold is a threshold that maximizes the fraction of explained variability in the proteome by linear regression on proteins that have a correlation with growth rate that exceeds the threshold (blue line). The maximal explained variability is 10% for the data set from [13] and is obtained given a threshold of 0.25. For the data set from [31] the maximal explained variability is 5% and it is obtained by choosing a threshold of 0.8.

6.2 Differences between the correlations found in the two data sets

The lower correlation and higher variability found in the data set from [13] partially results from the variability in the conditions it contains as well as the higher number of conditions measured across a similar range of growth rates. Specifically, as this data set includes measurements under different carbon sources, as opposed to the data set from [31], that uses the same carbon source on all measurements, a larger variability in expression patterns is expected. Restricting the analysis of the data set from [13] only to chemostat conditions supports this suggestion and shows much less variability as is shown in Figure S2.

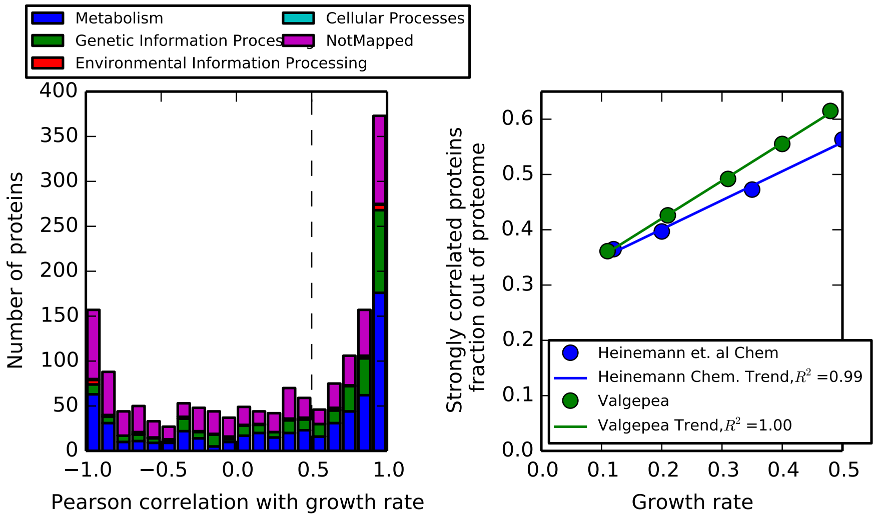

Figure S2: Restricting the analysis of the Heinemann data set to chemostat conditions yields similar results to those of the Valgepea data set.

6.3 Analysis including LB condition

Due to the fast growth rate under LB, compared with the other conditions measured in the data set from [13] it was not included in our primary analysis as was noted in section 4.1.3. Figure S3 shows the implications of including LB in the analysis. As can be seen, many proteins are now less correlated with growth rate due to down regulation under LB. However, despite having fewer proteins being strongly positively correlated with growth (525 vs. 628) and despite the accumulated fraction of these proteins being lower under the slower growth conditions ($\approx 20\%$ vs. $\approx 25\%$), these proteins do occupy $> 50\%$ out of the proteome under fast growth in LB.

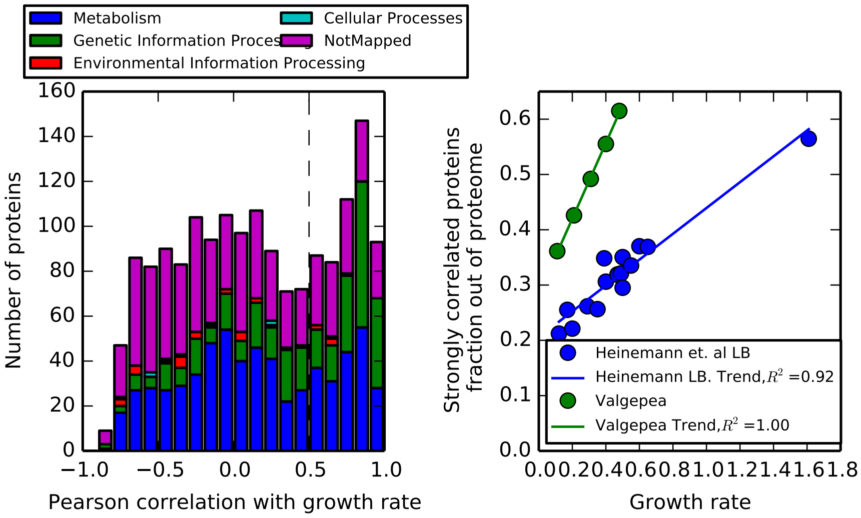

Figure S3: Including growth in LB media in the analysis of the data set from [13]. Fewer proteins are not strongly positively correlated with growth but these proteins form more than 50% of the proteome in LB growth.

6.4 Discussion of reasons for differing ribosome concentration relation to growth rate

Differences in ribosome concentration across growth rates as reported in different studies can result from a few factors:

1. Different growth rates and conditions monitored.
2. Usage of different strains.
3. In many studies the amount of ribosomes is deduced by measuring the RNA to protein ratio, assuming a relatively fixed portion of the RNA is rRNA. In our study, in contrast, ribosomal proteins are used as a proxy for estimating ribosomes concentration and, moreover, the RNA to Protein ratio is assumed to be constant. Therefore, and as it is known that ribosomes can operate even in the absence of some ribosomal proteins, such differences in manner of inference can account for some of the differences encountered.

6.5 The concentration of proteins that are not differentially regulated between conditions can be predicted by referencing other such proteins

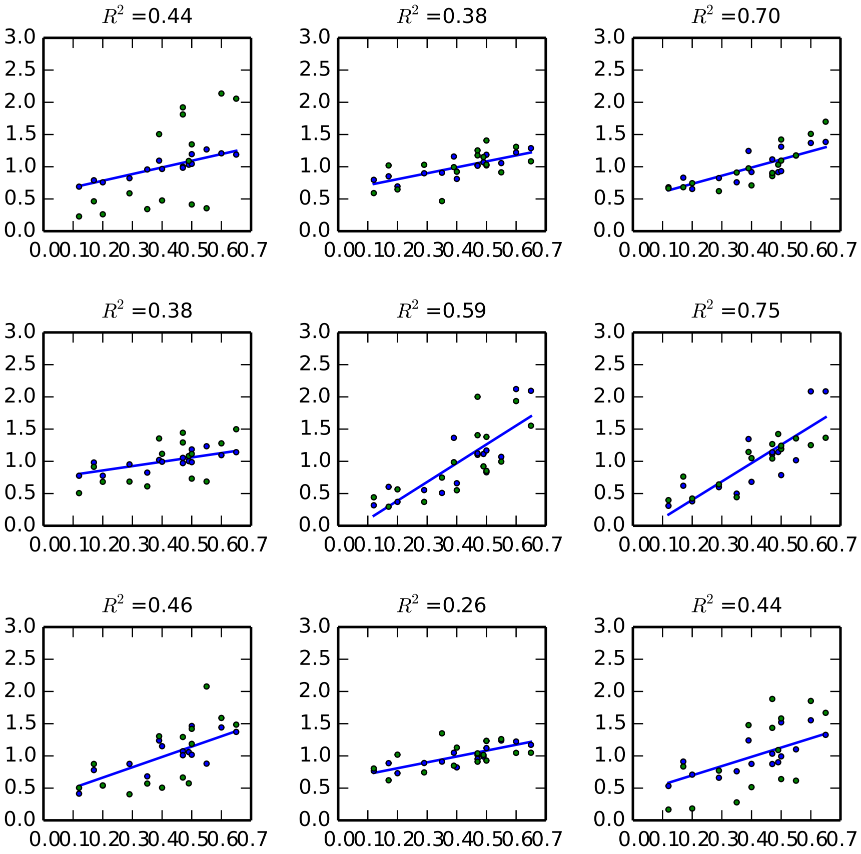

Figure S4: A selection of random predictions of protein concentrations from the highly correlated with growth rate fraction, taken from the data set of [13]. Each panel shows the average concentration of 10 random proteins that are highly correlated with growth (blue dots), a regression line that best fits the data, and the concentration of a different random protein (green dots). The R^2 value for the trend line and the different protein is given.

6.6 Breakdown by function of proteins strongly correlated with growth rate

Function	Number of proteins	% of proteome	Correlated proteins	Correlated % of proteome
Transcription	68	2.5	25	1.07
Carbohydrate Metabolism	129	18.52	44	4.67
Folding, Sorting and Degradation	96	6.8	40	2.95
Amino Acid Metabolism	99	9.05	66	6.71
Membrane Transport	74	7.68	14	0.16
Nucleotide Metabolism	61	4.85	37	3.25
Translation	112	10.74	82	9.37
NotMapped	645	24.43	171	8.5
Energy Metabolism	35	3.85	19	1.19
Lipid Metabolism	25	2.05	7	0.69
Metabolism of Other Amino Acids	21	1	12	0.76
Cytoskeleton	5	0.5	3	0.36
Metabolism of Cofactors and Vitamins	67	0.95	28	0.35
DNA maintenance	55	1.18	23	0.43
Cell Motility	5	0.73	0	0
Signal Transduction	33	0.98	7	$6.01 \cdot 10^{-2}$
Other enzymes	82	3.46	28	0.52
Glycan Biosynthesis and Metabolism	15	0.35	3	$4.78 \cdot 10^{-2}$
Xenobiotics Biodegradation and Metabolism	4	0.13	3	0.12
Not mapped	9	0.11	5	$1.48 \cdot 10^{-2}$
Metabolism of Terpenoids and Polyketides	16	0.16	11	0.11

Table S1: Breakdown by function of strongly positively correlated with growth rate proteins in the data set from [13]

Function	Number of proteins	% of proteome	Correlated proteins	Correlated % of proteome
Translation	102	20.49	86	17.2
DNA maintenance	33	1.62	5	0.29
Signal Transduction	32	0.75	3	$4.53 \cdot 10^{-2}$
Amino Acid Metabolism	102	11.79	67	9.54
Carbohydrate Metabolism	130	24.71	30	6.16
Membrane Transport	88	8.88	17	2.01
Nucleotide Metabolism	61	6.48	38	4.75
Transcription	40	1.98	12	1.02
Other enzymes	57	2.23	16	0.64
Metabolism of Cofactors and Vitamins	51	1.93	25	1.1
Folding, Sorting and Degradation	88	6.95	39	2.66
Metabolism of Other Amino Acids	22	1.14	5	0.51
Glycan Biosynthesis and Metabolism	12	0.5	3	0.14
Energy Metabolism	38	4.5	14	1.16
NotMapped	3	0.24	1	$7.95 \cdot 10^{-2}$
Cytoskeleton	5	0.36	1	$3.84 \cdot 10^{-3}$
Cell Motility	8	0.91	1	0.39
Xenobiotics Biodegradation and Metabolism	4	$9.25 \cdot 10^{-2}$	1	$5.98 \cdot 10^{-2}$
Lipid Metabolism	29	3.08	6	0.78
Metabolism of Terpenoids and Polyketides	8	0.15	6	0.12
Vesicular transport	4	1.03	1	0.14
Signaling Molecules and Interaction	1	$1.77 \cdot 10^{-3}$	0	0
Cell Growth and Death	1	0.19	1	0.19

Table S2: Breakdown by function of strongly positively correlated with growth rate proteins in the data set from [31]

6.7 Ribosomal proteins scale similarly to non-ribosomal proteins that are strongly positively correlated with growth rate

Comparing the normalized sum of ribosomal proteins to the normalized sum of the positively correlated with growth rate proteins that are non-ribosomal shows that these two groups scale in the same way with the growth rate, as is seen in Figure S5

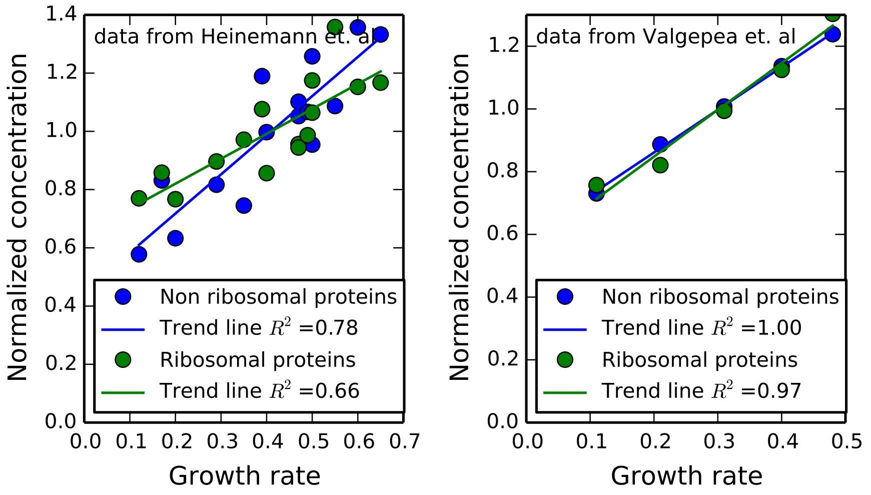

Figure S5: The scaling with growth rate of ribosomal proteins and non-ribosomal, but highly correlated with growth rate proteins is shown. Trend lines for the two groups of proteins are plotted. The scaling with growth rate is similar between the two groups of proteins.

6.8 Additional figures of simulated and randomized data sets

The maximal explained variability in data sets with shuffled protein abundances is significantly smaller than in the real data sets as is seen in figure S6.

A simulated data set, assuming half of the proteins scale linearly with growth rate with normalized intercept at 0.5, similar to the intercept found in the data analysis, and with simulated normally distributed noise levels of 25%, result in distributions similar to those found in the original data analysis (Figure S7

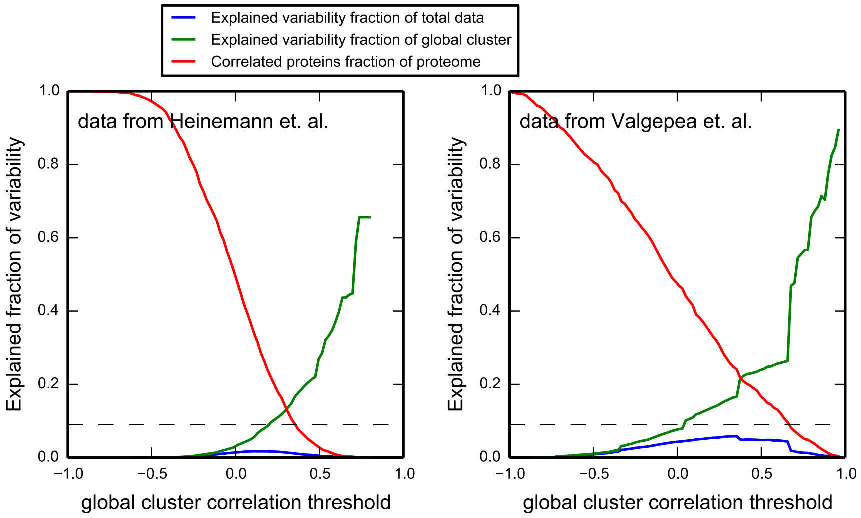

Figure S6: Fraction of explained variability by linear regression on the group of strongly positively correlated with growth rate proteins for the shuffled data sets.

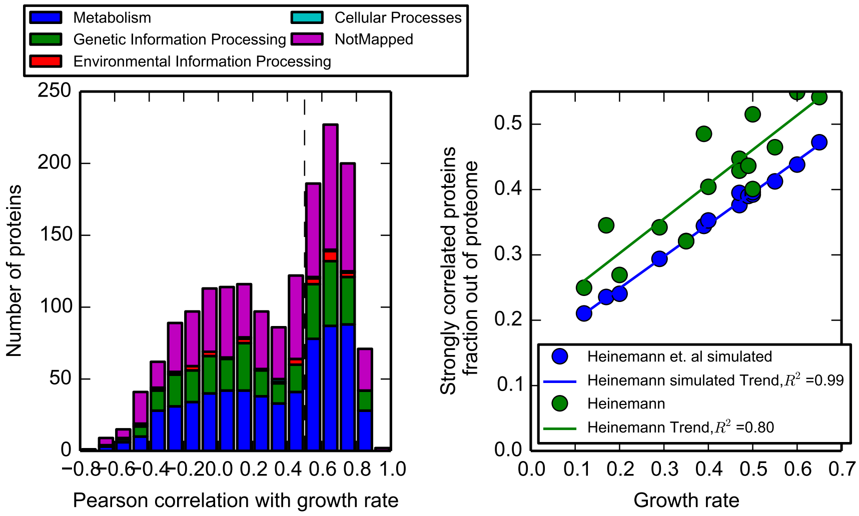

Figure S7: A simulated data set, assuming half of the proteins are prefectly correlated with growth rate and half are fixed, with simulated noise level of 25%. Average protein concentrations, growth rates and normalized slope of the correlated proteins are based on the data set from [13]. The normalized intercept of the correlated proteins was set to 0.5 in accordance with the intercept found in the original data analysis. The results are similar to those obtained for the real data set, showing that, given the experimental noise, identical coordination with growth rate of half of the proteins would result in similar outcomes to those observed in the data sets we use.