Genome Biology and Evolution Advance Access published October 14, 2013 doi:10.1093/gbe/evt154

GENOME BIOLOGY AND EVOLUTION

SMBE

Evaluating the Fitness Cost of Protein Expression in *Saccharomyces cerevisiae*

Katarzyna Tomala and Ryszard Korona*

Institute of Environmental Sciences, Jagiellonian University, Gronostajowa 7,

30–387 Krakow, Poland

*Author for correspondence: Ryszard Korona, address as above, phone: +48126645136, fax: +48126646912, e-mail: ryszard.korona@uj.edu.pl

Data deposition: Supplementary material

Genome Biology and Evolution

SMBE

Abstract

Protein metabolism is one of the most costly processes in the cell and is therefore expected to be under the effective control of natural selection. We stimulated yeast strains to overexpress each single gene product to approximately one percent of the total protein content. Consistent with previous reports, we found that excessive expression of proteins containing disordered or membrane-protruding regions resulted in an especially high fitness cost. We estimated these costs to be nearly twice as high as for other proteins. There was a tenfold difference in cost if, instead of entire proteins, only the disordered or membrane-embedded regions were compared with other segments. Although the cost of processing 'bulk' protein was measurable, it could not be explained by several tested protein features, including those linked to translational efficiency or intensity of physical interactions after maturation. It most likely included a number of individually indiscernible effects arising during protein synthesis, maturation, maintenance, (mal)functioning, and disposal. When scaled to the levels normally achieved by proteins in the cell, the fitness cost of dealing with one amino acid in a standard protein appears to be generally very low. Many single amino acid additions or deletions are likely to be neutral even if the effective population size is as large as that of the budding yeast. This should also apply to substitutions. Selection is much more likely to operate if point mutations affect protein structure by, for example, extending or creating stretches that tend to unfold or interact improperly with membranes.

Key words:

molecular evolution rate, protein overexpression, membrane proteins, disordered proteins, budding yeast

Introduction

Proteins constitute a major component of the dry mass of a cell. Synthesis of amino acids and subsequent assembly of polypeptides are costly. The two processes are estimated to consume about half of the ATP molecules in a growing yeast cell and involve a large fraction of its nucleic acids and ribosomal proteins (Verduyn 1991; Warner 1999). The huge cost of protein synthesis has been recognized as such for decades (Maaloe and Kjeldgaard 1966; Waldron and Lacroute 1975). More recently, it has been shown that newly assembled polypeptides are released into a crowded environment of macromolecules in which their folding is easily derailed (Ellis 2001). They often end up in a form that is not only unproductive but can also be toxic and sometimes resistant to degradation (Stefani and Dobson 2003; Winklhofer, et al. 2008). However, while it is certain that the costs and risks associated with the turnover of the total protein load are large, it remains unknown how much individual protein species differ in this respect. In theory, it is possible to calculate the cost of protein synthesis because the substrates and the process are well known. However, the required parameters are many and they have not yet been estimated with sufficient accuracy (Siwiak and Zielenkiewicz 2010; von der Haar 2008). Because the routes of folding and degradation for different polypeptides are still underway, the energy or fitness costs associated with such events are presently impossible to assess (Hartl, et al. 2011). Thus, it remains a great challenge in current research to provide analytical, experimental or computational estimates of selective pressures acting on individual proteins.

Evidence that different proteins experience different selective forces on traits other than their primary functions can be extracted from the DNA sequence. In particular, it is well established that the rate of molecular evolution differs widely between genes and that those expressed the most are the ones that change the least (Pal, et al. 2001; Sharp 1991). One

explanation could be that the highly expressed genes mutate at a lower rate, a possibility that has gained some support recently (Martincorena, et al. 2012). Most researchers, however, believe that more highly expressed genes are under stronger purifying selection. Some of the tentative explanations invoke functional arguments: importance (essentiality) of function, multiplicity of functions, centrality to metabolic networks, number of transcription factors assisting expression or enrichment for genetic and/or physical interactions (Bloom and Adami 2004; Fraser, et al. 2002; Jordan, et al. 2003; Pal, et al. 2006; Vitkup, et al. 2006; Wall, et al. 2005; Xia, et al. 2009). For each of these factors, however, correlation with the rate of evolution is much lower than that for the level of gene expression (Rocha 2006; Wang and Zhang 2009). Thus, it appears that it is the amount of protein product that matters most. This could mean that selection tends to purge mutations located in highly expressed genes because they lead to a greater waste of resources (Barton, et al. 2010; Vieira-Silva, et al. 2011). Not only efficient use of materials and energy but also a high rate of translation can be important. This could result in selection for optimal codon usage in the highly expressed genes (Akashi 2001; Plotkin and Kudla 2010). The more protein molecules, the higher the toxic effect after misfolding; therefore, misfolding-resistant sequences should especially be preserved in highly expressed genes, which would constrain their evolution (Drummond, et al. 2005; Drummond and Wilke 2008; Yang, et al. 2010). In sum, there is no lack of hypotheses for how the amount of synthesized protein could dictate the rate of molecular evolution. However, these hypotheses have been conceived through comparative analyses of DNA/protein sequences and have been verified mostly in the same way. In this paper, we report the results of a study aimed at testing these hypotheses experimentally, which has so far been addressed by only a few researchers.

The postulate of controlled alteration of selected determinants of the protein production cost has proved difficult to implement. For example, changing the actual codon

usage to a devised one alters the stability and hence the abundance of the resulting mRNA variants. The effect of mRNA abundance can be more important than the sought effect of mRNA composition (Agashe, et al. 2013; Kudla, et al. 2009). Even the seemingly straightforward task of demonstrating that overproduction of unnecessary proteins is disadvantageous has proved challenging. There must be costs associated with synthesis of redundant polypeptides, but there are also costs of their presence in the cell and their interactions with cell structures (Eames and Kortemme 2012; Plata, et al. 2010; Stoebel, et al. 2008). Our approach is based on the assumption that universal costs of protein expression do exist and can be at least partly disentangled if the number and diversity of analyzed proteins are sufficiently large. We relied on a genomic collection of yeast strains, each overexpressing a single protein. Two previous studies measured approximately how much protein was overproduced and categorized the growth effects accompanying this overproduction (Gelperin, et al. 2005; Sopko, et al. 2006). One experiment measured fitness using a quantitative assay but the level of production was not estimated and the average production could not be calculated as the applied protocol of overexpression differed from those used earlier (Yoshikawa, et al. 2011). We therefore carried out our own assays in which we stimulated genes to moderate protein overproduction, measured overexpressed protein levels quantitatively, and estimated the growth rate with high accuracy.

We first examined our data by asking whether the fitness effect of overexpression was heavily dependent on the cellular role of a tested gene. It was not, as we found by reviewing gene annotations. This was encouraging because we could assume that the effect of metabolic deregulation would not obscure the effect of carrying useless or toxic protein molecules. We thus asked which of the several protein properties could be the best predictor of fitness variation. We confirmed previous reports showing that proteins containing transmembrane (Kitagawa, et al. 2006; Osterberg, et al. 2006) and disordered (Ma, et al. 2010; Vavouri, et al.

2009) regions are especially costly to fitness when overexpressed. Crucially, we compared quantitatively these costs with the cost of expressing "normal" (well-structured cytosolic) proteins. We found that the cost of expressing well-structured cytosolic proteins is very low when scaled to one amino acid addition (and thus also substitution).

Materials and Methods

Strains

We used a previously constructed collection of single yeast ORFs, each with the same inducible promoter P_{GAL1} followed by the same tandem affinity tag (His6, HA epitope, protease 3C site, ZZ domain, 19 kDa) cloned into a multicopy plasmid (Gelperin, et al. 2005). Plasmids were hosted by the haploid yeast strain Y258. Most of the cloned genes had been tested for errors; only approximately 3% of them were likely to have an undetected mutation (Gelperin, et al. 2005).

Fitness assays

The overexpression strains were inoculated directly from plates shipped by the distributor (Open Biosystems) into 200 µl of SC with glucose but lacking uracil to stabilize the plasmid. To stimulate overexpression, we used SC with raffinose as a source of carbon and galactose as an inducer, according to a protocol described in the original study that led to moderate overexpression. We then transferred 10 µl aliquots of each culture into 190 µl of fresh glucose medium and incubated for 48 hours. From these cultures, 10 µl aliquots were transferred to 135 µl of SC with raffinose for another 48 hours. The raffinose cultures were diluted 10 times and the ODs measured. These cell suspensions were diluted again at 1:50 in SC with raffinose and galactose (2% each). In this growth/induction medium, the cultures were allowed to grow for 20 hours, at which point their ODs were determined. The ratio of the two OD

measurements, which were corrected for the dilution factor, served to calculate the number of cell doublings for each culture. All growth assays were carried out at 30 ºC.

Protein assays

Overproduction of proteins was induced by transferring cells sequentially from glucose to raffinose, and then to raffinose/galactose medium for 8 hours. The cells were then centrifuged, washed with ice-cold water and frozen. To extract proteins, the cells were beaten with glass beads in 100 µl of lysis buffer (50 mM Tris-HCl, pH 7.5, 0.5% SDS, 0.1 mM EDTA, protease inhibitors) for 4 hours at 4 ºC. Cell remnants were then spun down, and the supernatants were collected. Total protein content was determined using a BCA protocol. For a competitive ELISA assay, plates were coated overnight at 4 ºC with 0.05 µl of normal rabbit serum (Pierce) diluted in 100 µl of 0.2 M carbonate-bicarbonate buffer, pH 9.4. After washing, plates were blocked with 300 µl of 2% BSA for 24 hours. The yeast protein extracts were mixed with protein A conjugated to peroxidase (Pierce) then 100 µl of the resulting mixture was added to the blocked plate wells, for a total 10 µg of total yeast protein and 25 ng (~26 µU) of protein A per well. After 1 hour of incubation, the mixtures were discarded and the wells washed and filled with 100 µl of the TNB substrate. The reaction was terminated after 30 minutes with 100 µl of 2 M H_2SO_4, and then, the absorbance at 450 nm was measured. All washing steps were performed with 200 µl of PBS containing 0.05 % Tween 20. One of the tagged proteins (Ade2p) was purified, diluted into a gradient of known concentrations, and used as a standard to calibrate the reads.

Gene Ontology and protein properties

To analyze the GO categories (SGD, Saccharomyces Genome Database), we applied an ANOVA model in which each of the 5,084 overexpressed genes was described by the Yeast

Slim categories taking values of zero or one (absent or present). We used the 'lm' function of the R package, followed by the 'step' function (based on AIC statistics) to reduce the number of predictor variables by eliminating the non-significant ones (R Development Core Team 2010). The analyses were performed separately for the 'molecular function', 'cellular component', and 'biological process' classifications. As these classifications contained tens of terms, we did not analyze interactions between them because the latter were very numerous and usually contained too few data points to be meaningful.

Protein properties were analyzed by implementing a multiple regression model using the 'lm' function. Continuous predictor variables were log-transformed (except for gravy score and mRNA 5' folding energy); a small constant was added to those with zero values before transformation (Wall, et al. 2005). The continuous predictor variables included: mRNA abundance (Garcia-Martinez, et al. 2004), protein half-life (Belle, et al. 2006), intrinsic disorder/protein length + 0.01(Linding, et al. 2003), protein length (SGD), CAI+0.1 (SGD), gravy score (SGD), and protein abundance, that is, the number of molecules per protein species (Ghaemmaghami, et al. 2003). To calculate the energy of structures at the 5' end of mRNAs, we used the Vienna RNA Package 2.0 (Lorenz, et al. 2011) for stretches extending from the -4 to +37 nucleotide positions (Plotkin and Kudla 2010). All continuous predictor variables were standardized prior to analysis. There were also two categorical variables: physical interaction status (not hub, intermediate number of interactions, party hub, date hub) (Ekman, et al. 2006; Han, et al. 2004) and the presence of transmembrane segments (not predicted, predicted by only one study, predicted by two studies) (Krogh, et al. 2001; Persson and Argos 1994). ORFs with missing values in any of the predictor variables were excluded from this analysis. There were 2,913 ORFs with a complete set of predictors, and only those were included in the final orthogonal model. We included all 10 listed variables in the model and the first order interactions between them (except for interactions between the two

categorical variables). The entire procedure was repeated 40 times with random permutations of the order of categories in the model. The *p*-values for predictor variables were averaged over repeats (geometrically).

Results

Fitness effects of moderate overexpression of genes are small

We found that an overproduced protein species constituted typically approximately 1% of the total protein amount (more detailed data reported below), which is much less than doses known to be severely toxic (Dong, et al. 1995; Geiler-Samerotte, et al. 2011). We measured fitness by estimating how many cell divisions occurred in single-strain liquid cultures over a period of about one day (see Methods). This included both lag and growth phases resulting in an average number of doublings of 7.75 (median 7.83) with a standard deviation of 0.45. (The cultures reached about one fourth of their final density.) Thus, variation in fitness was not high, especially given that a sizable portion of it came from differences between plates and was eliminated from all subsequent analyses by within-plate normalization (Methods). Previous studies evaluated the growth of colonies on common agar plates (Gelperin, et al. 2005; Sopko, et al. 2006) or in individual liquid cultures over a shorter time interval (Makanae, et al. 2013; Yoshikawa, et al. 2011). Those earlier estimates generally agree with ours (Supplementary Fig. 1). We sought to assay fitness in a way that would increase the role of fast growth, and thus fast protein processing, in the final measure of fitness. Importantly, we wanted to compare quantitative fitness estimates with quantitative estimates of protein overproduction for a large number of individual clones, which had not been performed in previous studies.

Fig. 1 shows the distribution of normalized fitness estimates for 5,182 strains containing a unique cloned ORF known to express a protein (SGD). The intraclass correlation

coefficient (*ICC*) calculated over four independent repeats was 0.966, indicating that repeatability of our fitness measurements was high. Good repeatability within a strain and large differences between strains (the shape of clouds) suggest that factors other than measurement errors were responsible for much of the fitness variation. Some factors, such as the average copy number of individual plasmids, could not be controlled in this experimental system. All individual records, both normalized and non-normalized, are listed in Supplementary Table 1.

Functional categorization explains little of the gene overexpression effects

As reported below in detail, the median content of overexpressed proteins was about 400 times higher than the median content of normally expressed ones (Ghaemmaghami, et al. 2003). This could potentially disturb at least some cellular functions. The overexpressed genes fell into 22 Yeast Slim GO cell component categories, 41 molecular function categories, and 100 biological process categories (we decided to reduce the biological process categories to 40 by combining some of the most similar ones). Within each of these three classifications, we first applied a linear model including all categories and then progressively simplified it by eliminating statistically non-significant categories (see Methods). We obtained a relatively low number of potentially important predictors shown in Fig. 2. There were a few categories associated with increased fitness. These suggest that speeding up turnover of nucleotides and adjusting oxidative metabolism could have a positive effect on fitness. Negative effects were more numerous and larger. They were linked to cell wall and membrane structures. Although these factors were significant on a statistical level, they had very small average effects, approximately 0.005, which is clearly less than the standard deviation of the overall distribution of normalized fitness estimates, 0.032 (Fig. 1b). The observed weak dependence of fitness effects on the functions of the overexpressed proteins

may be specific to our experimental system. Other arrangements, e.g., *E. coli* and high overexpression, have shown that unnaturally high levels of transcription factors and regulatory proteins can be toxic (Singh and Dash 2013).

To further test whether growth was indeed relatively insensitive to metabolic deregulation, we focused our analyses on enzymes alone. We revisited a study in which the molecular evolution of enzymes was considered dependent on their metabolic centrality and connectivity (Vitkup, et al. 2006). Connectivity of an enzyme had been calculated as the number of other metabolic enzymes that produce or consume the enzyme's products or reactants. In our data set, 329 of the 350 enzymes examined in the original study were included. We used the same categorization of metabolic connectivity but did not find it helpful in explaining the observed variation in the fitness response to gene overexpression ($r = –0.029$, $p = 0.6$). Apparently, the cell's metabolic network is well buffered against perturbations in the expression level of participating enzymes, at least when single enzymes are overabundant. As reported above, most cellular structures and processes were also remarkably resistant to such alterations. We therefore decided that it would be acceptable to execute the analysis of protein properties for all genes together, ignoring their cellular roles and making the statistics both simpler and more powerful.

Only a few protein properties correlate with the cost of overexpression

A review of theoretical and empirical studies disclosed 10 properties of proteins/mRNAs that were frequently examined as factors potentially affecting the rate of evolution. The dependence of fitness on the most significant factors is shown in Fig. 3a. The remaining factors are presented in Supplementary Fig. 2. These graphs illustrate how the fitness of the overexpression strains correlates with each characteristic separately. They show that although the effects of some factors (e.g., protein length) are small, they can be remarkably regular. In

a formal statistical analysis, we used a linear model, which examined jointly all single factors and selected interactions (see Materials and Methods). The results are reported more thoroughly in Supplementary Table 2. Here, in Fig. 3b, we present only summaries of statistics for individual factors. Some factors, such as protein half-life, codon adaptation index, frequency of physical interactions, abundance under normal expression, energy of 5' mRNA fold, and gravy score proved non-significant. Two of the statistically significant factors, the presence of transmembrane regions and the proportion of protein length occupied by sequences predicted to be loosely shaped (intrinsically disordered), refer to properties that become meaningful only after a protein chain is synthesized and folded. Other properties may be important at the time of synthesis. There was a negative correlation between the level of mRNA under normal expression and fitness. This could mean that overexpression of the normally common transcripts tends to deplete optimal tRNAs for production of redundant proteins and thus slow down elongation of those needed. However, the effect of high CAI on fitness, although negative, was not statistically significant. The energy of the folding of 5' mRNAs was also neutral, suggesting that transcripts with rigid spatial structures did not trap too many ribosomes (Plotkin and Kudla 2010). It thus appears that there is no shortage of ribosomes, and possibly optimal tRNAs, when one percent of translation is useless, at least under the growth conditions applied here. Finally, there was a negative correlation between protein length and fitness indicating that the amount of an overproduced protein mattered (because all overexpressed proteins had the same promoter). This relation attracted our attention especially because it appeared to be very regular over the entire range of protein lengths (Fig. 3a). We therefore decided to test experimentally whether the length of a protein is a good proxy for its amount under overexpression.

Relating fitness cost to the amount of protein

We estimated the cellular level of overproduced protein for a large sample of strains. Repeatability of estimates obtained by competitive ELISA was high (*ICC* = 0.944, *n* = 719, $p << 0.001$) and centered on a median of 0.63 % (Fig. 4a). The relationship between the amount of overproduced protein and its length is shown in Fig. 4b; Pearson's correlation coefficient was significant ($r = 0.136$, $df = 717$, $p = 0.0002$). To find a quantitative relation between the length of a protein and its amount under overexpression, we used a data set without the outliers seen in Fig. 4b (see Supplementary Methods for details). We found that when the length of a protein doubles, its amount under overexpression increases by about half (the slope of a linear regression with both axes log-transformed was 0.47). We could then assign to every protein its expected amount under overexpression as a function of its length. From the common model of multiple regression, we found the relationships between the length of a protein (and its amount), the presence of transmembrane regions, and the presence of disordered regions, the three factors jointly effecting fitness (Supplementary Table 3). This information is summarized in Table 1, which lists the cost of expressing different proteins per 1% of total protein mass and per amino acid. To get the latter estimates, we assumed that the total mass of proteins in the yeast cell is 6.0×10^{-12} g (Sherman 2002). Knowing the number of molecules (Ghaemmaghami, et al. 2003) and their molecular weights, we could calculate the total weight of every protein. The contribution of special regions was calculated from the proportions of the transmembrane or disordered regions calculated for every individual protein species (Krogh, et al. 2001; Linding, et al. 2003; Persson and Argos 1994). One implicit assumption that could introduce only a minimal bias to our estimates is the assumption that the per amino acid weight of the transmembrane, disordered, and other regions was equal (See Supplementary Methods for more details regarding calculations).

Table 1 shows that the average effect of having a disordered region or a transmembrane domain is remarkable but not excessively large. On average, disordered regions nearly doubled the fitness cost of the entire protein. Similarly, the membrane proteins were substantially more costly than were the cytosolic ones. The costs expressed per amino acid show the relative fitness changes of expanding some regions at the expense of other regions. They may also serve to compare fitness costs of proteins expressed at different levels. The yeast proteins are represented by very different numbers of molecules per cell under natural expression, from ten to one million (Ghaemmaghami, et al. 2003).

In the analyses described above, either some of the characteristics borrowed from other studies or our own measurements were lacking for a number of genes. We asked which of our results would hold if a single analysis were performed for those genes only for which both the fitness estimate, as well as the protein overexpression level, and all other variables were known. There were only 423 such genes. Detailed results are presented in Supplementary Table 4. Briefly, the presence of transmembrane domains remained the most significant factor. Three factors pertaining to protein abundance—the measured level, the reported half-life, and the predicted length—were also significant or nearly significant. This latest finding is yet another indication that it is not only the structural properties of a redundant protein but also its amount that contributes to toxicity.

Discussion

We found that overexpression of single genes in *S. cerevisiae* generally leads to moderate but variable effects on growth. This variation is partly explained by the properties of the over-expressed protein molecules and the roles they play in cellular metabolism. Cell growth also correlated to the amount of over-expressed protein, indicating that synthesis and processing of useless polypeptides lowers the efficiency of cell growth. This particular cost was relatively

small, which explains why it has not been convincingly demonstrated in former studies. Proteins with disordered or intra-membrane regions were especially damaging to fitness when overexpressed. Based on these findings, we propose that an addition, or exchange, of a single amino acid is of little consequence for fitness unless it extends or creates protein regions forming critical structures.

There are two possible explanations why the disordered and transmembrane regions are especially damaging to fitness when overexpressed. One of them concentrates on overload, the other on toxicity. Considering overload, we note that the summed mass of all membrane proteins is 15% of the total protein content in a yeast cell. Similarly, the disordered stretches of polypeptides make up approximately 12% of total protein. Therefore, the same weight of an extra 1% of protein constitutes a considerably higher overload in terms of proportion added to the proteins that are in membranes or are disordered. The costs associated with transmembrane proteins can include membrane piercing, interfering with other membrane proteins, or engaging membrane-specific folding pathways. Similarly, if maintaining the total pool of loosely structured proteins poses some special cost to the cell, then every overexpressed member of this group adds a higher proportion to this cost. Generally, the costs of overload could result from expressing those proteins that are more expensive/risky to keep in the cell even if they function as expected. A type of overload hypothesis has been proposed in which malfunctioning of membranes occurs in response to the overexpression of a membrane protein (Eames and Kortemme 2012). On the contrary, the cost of toxicity means that over-expressed protein chains acquire new and unwanted functions. It is possible that both the disordered and membrane proteins are especially likely to undergo such transformation. The ‘disordered’ or ‘unstructured’ regions have important functions in signaling, control, and regulation (Dunker, et al. 2008). Proteins with such regions interact with one another and with unrelated proteins, which leads to misfolding and

aggregation (Olzscha, et al. 2011; Uversky, et al. 2008; Vavouri, et al. 2009). Aggregates tend to expose hydrophobic surfaces and therefore tend to illegitimately penetrate and damage cellular membranes (Kourie and Henry 2002; Stefani 2008). Even the programmed formation of transmembrane domains can be sensitive to crowding and non-prescribed interactions with other regions of polypeptides (Chakrabarti, et al. 2011; Levine, et al. 2005; Mackenzie 2006; Skach 2009). In sum, there are good hypothetical explanations why transmembrane and disordered proteins are especially likely to be overloaded or driven into toxicity when overexpressed. However, substantial efforts would be needed to find which of the two possible mechanisms is actually occurring when a particular protein is over-expressed.

There are two other properties of proteins that correlated with the cost of over-expression: the length of the polypeptide and the abundance of the cognate mRNA under normal expression. As explained in the Results, we believe the two traits are simply correlated with the amount of useless protein and that this unnecessary burden is the real cause of fitness decrease. We base our assumption on the remarkable regularity of the relationship between polypeptide length and fitness loss, as well as on a statistically significant relation between polypeptide length and an actual abundance of overexpressed protein in the cell. We considered two alternative hypotheses. One assumes that long proteins are disproportionally more likely to misfold and thus over-exploit molecular chaperones. To test this, we asked whether the over-expression of proteins known to interact with molecular chaperones had more substantial effects on fitness. We do not report these tests because we did not find any relationship between the fitness cost and the frequency of interactions with single chaperones (Bogumil, et al. 2012), sets of chaperones revealed in large-scale studies (Gong, et al. 2009), or smaller but carefully confirmed chaperone assemblages (Hartl, et al. 2011). These results are in accord with a report suggesting that chaperones are efficient enough to handle a load of misfolded proteins that is substantially higher than 1% (Vabulas and Hartl 2005). Another

alternative explanation, that long proteins have more domains and thus are more damaging to the cellular regulatory mechanisms, has been tested and rejected (see Results). We therefore propose that our observed negative effect of protein length on fitness reflects the general cost of protein processing, which includes all expenses involved in protein synthesis, maturation, maintenance, and disposal.

Our results can be used to address the question of whether natural selection is strong enough to prevent a single amino acid being added or exchanged for another one. The efficiency with which genomes and proteomes are purged of mutations depends not only on the strength of their effects but also on population size (Fernandez and Lynch 2011; Lynch and Conery 2003). Natural selection operates when $2N_es>1$, where N_e stands for effective population size and s for the selection coefficient. It is effective when the quotient is ten times higher. The effective population size of a species closely related to *S. cerevisiae*, *S. paradoxus*, was estimated at 8.6×10^6 (Tsai, et al. 2008). We found that the average cost of processing one amino acid is approximately 4×10^{-11} (Table 1), so this would be the cost of adding one unnecessary amino acid to one polypeptide and need to be multiplied by the number of affected molecules. It follows that to be non-neutral ($2N_es>1$), a mutation of this type must hit a protein represented by more than 1,453 molecules per cell. In *S. cerevisiae*, some two thirds of proteins meet this weaker criterion but only a small minority the stronger one (Ghaemmaghami, et al. 2003). Thus, selection can possibly act on a single amino acid only if the effective population size is as large as in yeast and only if proteins are sufficiently abundant. The entire cost of this size would be at stake if an amino acid were to be deleted or inserted. Substitution would most likely still be less costly and thus more often neutral. In many organisms the effective population size is much smaller, even by three orders of magnitude (Charlesworth 2009; Gossmann, et al. 2012), making selection still less effective. Our empirical findings generally agree with the results of a former computational study.

Expediting single atoms of the main components of yeast biomass (such as carbon or nitrogen) has been found selectively non-neutral for just approximately 1% of proteins (those most abundantly expressed). Only under starvation for those rarer, such as sulfur, a wasteful use of one atom (or an amino acid in which it resides) can be significant for a substantial proportion of proteins (Bragg and Wagner 2009).

Considering the factors that could control the evolution of protein sequence, it is remarkable that the fitness costs associated with amino acids residing within the disordered or transmembrane regions were so much higher. It appears justifiable to speculate that natural selection would operate most intensely on mutations creating new or extending existing regions of danger. Not only mutations making misfolding or misinteraction unavoidable would be selected against (Yang, et al. 2012) but also any changes in the DNA sequence that could increase the rate of transcriptional and translational errors resulting in alterations of the spatial structure of proteins (Drummond, et al. 2005; Drummond and Wilke 2008). Such changes could result in selection coefficients that were higher by several orders of magnitude than those arising from amino acid substitutions in standard protein regions. This is because any unwinding of a polypeptide can involve dozens of amino acids, each being ten times more costly than it was in a safe structure. There is some evidence to suggest that selection preventing structural aberration can be strong (Chiti and Dobson 2006; Geiler-Samerotte, et al. 2011), but further work is clearly needed to show that much or perhaps most of the variation in the rate of protein evolution can be attributed to selection, minimizing the danger of protein misfolding and toxicity.

Acknowledgments

This work was supported by the Foundation for Polish Science (a "Mistrz" grant to R.K.), a National Science Centre grant no. 2011/01/B/NZ8/00042 (to K.T.) and an IESc grant DS/WBiNoZ/INoS/762/2011-2012 (to both).

References

Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ 2013. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. Mol Biol Evol 30: 549-560.

Akashi H 2001. Gene expression and molecular evolution. Curr Opin Genet Dev 11: 660-666.

Barton MD, Delneri D, Oliver SG, Rattray M, Bergman CM 2010. Evolutionary systems biology of amino acid biosynthetic cost in yeast. PLoS One 5: e11935.

Belle A, Tanay A, Bitincka L, Shamir R, O'Shea EK 2006. Quantification of protein half-lives in the budding yeast proteome. Proc Natl Acad Sci U S A 103: 13004-13009.

Bloom JD, Adami C 2004. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: response. BMC Evol Biol 4: 14.

Bogumil D, Landan G, Ilhan J, Dagan T 2012. Chaperones divide yeast proteins into classes of expression level and evolutionary rate. Genome Biol Evol 4: 618-625.

Bragg JG, Wagner A 2009. Protein material costs: single atoms can make an evolutionary difference. Trends in Genetics 25: 5-8.

Chakrabarti O, Rane NS, Hegde RS 2011. Cytosolic aggregates perturb the degradation of nontranslocated secretory and membrane proteins. Mol Biol Cell 22: 1625-1637.

Charlesworth B 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10: 195-205.

Chiti F, Dobson CM 2006. Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem 75: 333-366.

Dong H, Nilsson L, Kurland CG 1995. Gratuitous overexpression of genes in Escherichia coli leads to growth inhibition and ribosome destruction. J Bacteriol 177: 1497-1504.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A 102: 14338-14343.

Drummond DA, Wilke CO 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341-352.

Dunker AK, Silman I, Uversky VN, Sussman JL 2008. Function and structure of inherently disordered proteins. Curr Opin Struct Biol 18: 756-764.

Eames M, Kortemme T 2012. Cost-benefit tradeoffs in engineered lac operons. Science 336: 911-915. doi: 336/6083/911

Ekman D, Light S, Bjorklund AK, Elofsson A 2006. What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? Genome Biol 7: R45.

Ellis RJ 2001. Macromolecular crowding: obvious but underappreciated. Trends Biochem Sci 26: 597-604.

Fernandez A, Lynch M 2011. Non-adaptive origins of interactome complexity. Nature 474: 502-505.
Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW 2002. Evolutionary rate in the protein interaction network. Science 296: 750-752.
Garcia-Martinez J, Aranda A, Perez-Ortin JE 2004. Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. Mol Cell 15: 303-313.]
Geiler-Samerotte KA, et al. 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proc Natl Acad Sci U S A 108: 680-685.
Gelperin DM, et al. 2005. Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. Genes Dev 19: 2816-2826.
Ghaemmaghami S, et al. 2003. Global analysis of protein expression in yeast. Nature 425: 737-741.
Gong Y, et al. 2009. An atlas of chaperone-protein interactions in Saccharomyces cerevisiae: implications to protein folding pathways in the cell. Mol Syst Biol 5: 275.
Gossmann TI, Keightley PD, Eyre-Walker A 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol Evol 4: 658-667.
Han JD, et al. 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430: 88-93.
Hartl FU, Bracher A, Hayer-Hartl M 2011. Molecular chaperones in protein folding and proteostasis. Nature 475: 324-332.
Jordan IK, Wolf YI, Koonin EV 2003. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3: 1.
Kitagawa M, et al. 2006. Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. DNA research 12: 291-299.
Kourie JI, Henry CL 2002. Ion channel formation and membrane-linked pathologies of misfolded hydrophobic proteins: the role of dangerous unchaperoned molecules. Clin Exp Pharmacol Physiol 29: 741-753.
Krogh A, Larsson B, von Heijne G, Sonnhammer EL 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567-580.
Kudla G, Murray AW, Tollervey D, Plotkin JB 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science 324: 255-258.
Levine CG, Mitra D, Sharma A, Smith CL, Hegde RS 2005. The efficiency of protein compartmentalization into the secretory pathway. Mol Biol Cell 16: 279-291.
Linding R, Russell RB, Neduva V, Gibson TJ 2003. GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res 31: 3701-3708.
Lorenz R, et al. 2011. ViennaRNA Package 2.0. Algorithms for Molecular Biology 6: 26.
Lynch M, Conery JS 2003. The origins of genome complexity. Science 302: 1401-1404.]
Ma L, Pang CN, Li SS, Wilkins MR 2010. Proteins deleterious on overexpression are associated with high intrinsic disorder, specific interaction domains, and low abundance. J Proteome Res 9: 1218-1225.
Maaloe O, Kjeldgaard NO. 1966. Control of Macromolecular Synthesis. New York: W. A. Benjamin Inc. .

Mackenzie KR 2006. Folding and stability of alpha-helical integral membrane proteins. Chem Rev 106: 1931-1977.
Makanae K, Kintaka R, Makino T, Kitano H, Moriya H 2013. Identification of dosage-sensitive genes in Saccharomyces cerevisiae using the genetic tug-of-war method. Genome research 23: 300-311.
Martincorena I, Seshasayee AS, Luscombe NM 2012. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature 485: 95-98.
Olzscha H, et al. 2011. Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions. Cell 144: 67-78.
Osterberg M, et al. 2006. Phenotypic effects of membrane protein overexpression in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 103: 11148-11153.
Pal C, Papp B, Hurst LD 2001. Highly expressed genes in yeast evolve slowly. Genetics 158: 927-931.
Pal C, Papp B, Lercher MJ 2006. An integrated view of protein evolution. Nat Rev Genet 7: 337-348.
Persson B, Argos P 1994. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. J Mol Biol 237: 182-192.
Plata G, Gottesman ME, Vitkup D 2010. The rate of the molecular clock and the cost of gratuitous protein synthesis. Genome Biol 11: R98.
Plotkin JB, Kudla G 2010. Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics 12: 32-42.
R Development Core Team 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
Rocha EP 2006. The quest for the universals of protein evolution. Trends Genet 22: 412-416.
Sharp PM 1991. Determinants of DNA sequence divergence between Escherichia coli and Salmonella typhimurium: codon usage, map position, and concerted evolution. J Mol Evol 33: 23-33.
Sherman F 2002. Getting started with yeast. Methods Enzymol 350: 3-41.
Singh GP, Dash D 2013. Electrostatic Mis-Interactions Cause Overexpression Toxicity of Proteins in E. coli. PLoS One 8: e64893.
Siwiak M, Zielenkiewicz P 2010. A comprehensive, quantitative, and genome-wide model of translation. PLoS Comput Biol 6: e1000865.
Skach WR 2009. Cellular mechanisms of membrane protein folding. Nat Struct Mol Biol 16: 606-612.
Sopko R, et al. 2006. Mapping pathways and phenotypes by systematic gene overexpression. Mol Cell 21: 319-330.
Stefani M 2008. Protein folding and misfolding on surfaces. Int J Mol Sci 9: 2515-2542.
Stefani M, Dobson CM 2003. Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. J Mol Med (Berl) 81: 678-699.
Stoebel DM, Dean AM, Dykhuizen DE 2008. The cost of expression of Escherichia coli lac operon proteins is in the process, not in the products. Genetics 178: 1653-1660.
Tsai IJ, Bensasson D, Burt A, Koufopanou V 2008. Population genomics of the wild yeast Saccharomyces paradoxus: Quantifying the life cycle. Proc Natl Acad Sci U S A 105: 4957-4962.
Uversky VN, Oldfield CJ, Dunker AK 2008. Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys 37: 215-246.

Vabulas RM, Hartl FU 2005. Protein synthesis upon acute nutrient restriction relies on proteasome function. Science 310: 1960-1963.
Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B 2009. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. Cell 138: 198-208.
Verduyn C 1991. Physiology of yeasts in relation to biomass yields. Antonie Van Leeuwenhoek 60: 325-353.
Vieira-Silva S, Touchon M, Abby SS, Rocha EP 2011. Investment in rapid growth shapes the evolutionary rates of essential proteins. Proc Natl Acad Sci U S A 108: 20030-20035.
Vitkup D, Kharchenko P, Wagner A 2006. Influence of metabolic network structure and function on enzyme evolution. Genome Biol 7: R39.
von der Haar T 2008. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. BMC Syst Biol 2: 87.
Waldron C, Lacroute F 1975. Effect of Growth Rate on the Amounts of Ribosomal and. Journal of Bacteriology 122: 855-865.
Wall DP, et al. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A 102: 5483-5488.
Wang Z, Zhang J 2009. Why is the correlation between gene importance and gene evolutionary rate so weak? PLoS Genet 5: e1000329.
Warner JR 1999. The economics of ribosome biosynthesis in yeast. Trends Biochem Sci 24: 437-440.
Winklhofer KF, Tatzelt J, Haass C 2008. The two faces of protein misfolding: gain- and loss-of-function in neurodegenerative diseases. EMBO J 27: 336-349.
Xia Y, Franzosa EA, Gerstein MB 2009. Integrated assessment of genomic correlates of protein evolutionary rate. PLoS Comput Biol 5: e1000413.
Yang JR, Liao BY, Zhuang SM, Zhang J 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci U S A 109: E831-840.
Yang JR, Zhuang SM, Zhang J 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. Mol Syst Biol 6: 421.
Yoshikawa K, et al. 2011. Comprehensive phenotypic analysis of single-gene deletion and overexpression strains of Saccharomyces cerevisiae. Yeast 28: 349-361.

Genome Biology and Evolution

SMBE

Figure legends

Fig. 1. The effects of single gene overexpression on growth. The number of cell divisions in single-strain cultures was estimated four times independently. The estimates were divided by the median values of relevant replications to obtain normalized values. **(a)** The repeatability of the individual normalized fitness estimates and **(b)** the frequency distribution of strains' means. The vertical dashed line marks the slowest growing 91 strains. These were removed from all of the following statistical analyses to make the distribution symmetric and closer to normal. (This exclusion was unlikely to affect our analyses. For example, we correlated fitness with ten properties of proteins for all data and those lacking the 91 data points. For data analyzed in this way, pairs of Pearson's coefficients were themselves very much correlated: Pearson's r=0.988, Spearman's r_s=1.)

Fig. 2. Gene Ontology categories as predictors of the overexpression cost. The graph shows the highest and most statistically significant deviations of the Yeast Slim category means from the grand mean (not fitness gains or losses when compared with a strain with no overexpression).

Fig. 3. Protein properties and the fitness cost of overexpression. (a) Examples of fitness predictors (only the most significant predictors are shown; the remaining ones are in Supplementary Fig. 2). Moving averages are shown as red lines for continuous variables. **(b)** Results of multifactorial analysis. Statistical significance of positive (green) and negative (red) effects is shown.

Fig. 4. The level of protein overexpression. (a) Frequency distribution of the amount of protein at the normal (empty bars) and overexpressed (grey bars) levels. Normal protein levels were taken from a previous study (Ghaemmaghami, et al. 2003) and overexpression estimates were obtained in this study using a competitive ELISA assay. **(b)** The relationship between protein length and protein overexpression level (see Supplementary Methods).

Table1. Fitness cost of protein expression.

Protein type[1]	1% of total protein[2] (mean±SE)	Special region fraction (mean±SD)	Cost per single aa[3] (mean±SE)
Standard	0.023±0.005	-	$(7.32\pm1.63)\cdot10^{-11}$
Disordered (added)	0.017±0.004	0.11±0.08	$(6.76\pm1.47)\cdot10^{-10}$
Trans-membrane (added)	0.012±0.002	0.13±0.10	$(4.78\pm0.82)\cdot10^{-10}$

[1] Proteins were standard (that is, cytosolic and well structured), contained disordered regions, and were located in membranes. The proportion of protein length taken by the disordered or transmembrane regions is shown in the middle column.

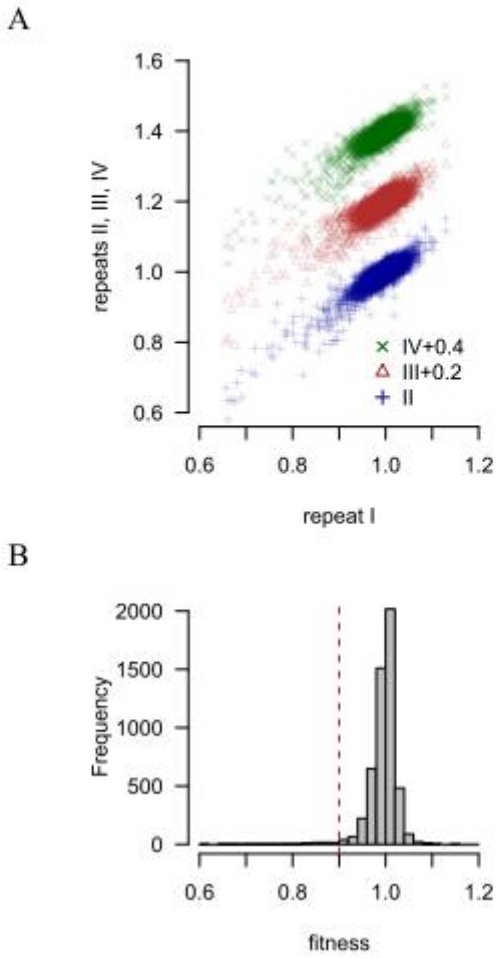
[2] The fitness cost of producing 1% of superfluous polypeptide (standard), plus the costs added by the presence of disordered or trans-membrane regions.

[3] The fitness cost of expressing one amino acid in one protein molecule if the amino acid is located in standard or special regions.

Genome Biology and Evolution

SMBE

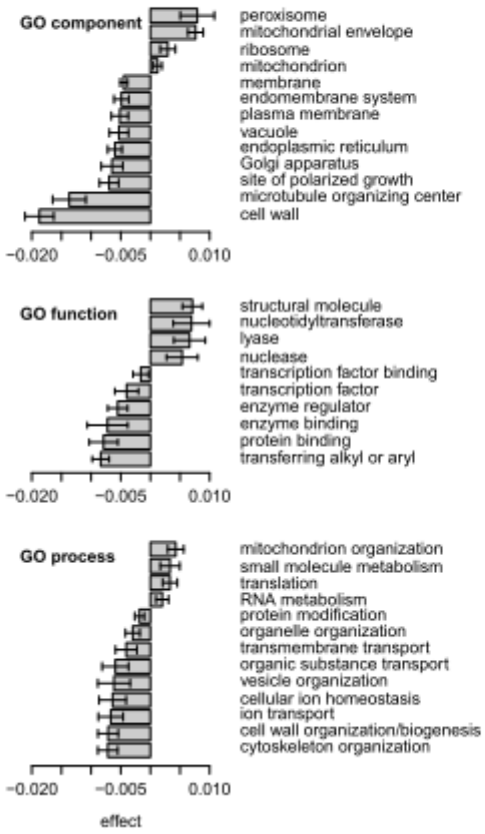
Fig. 1



Genome Biology and Evolution

SMBE

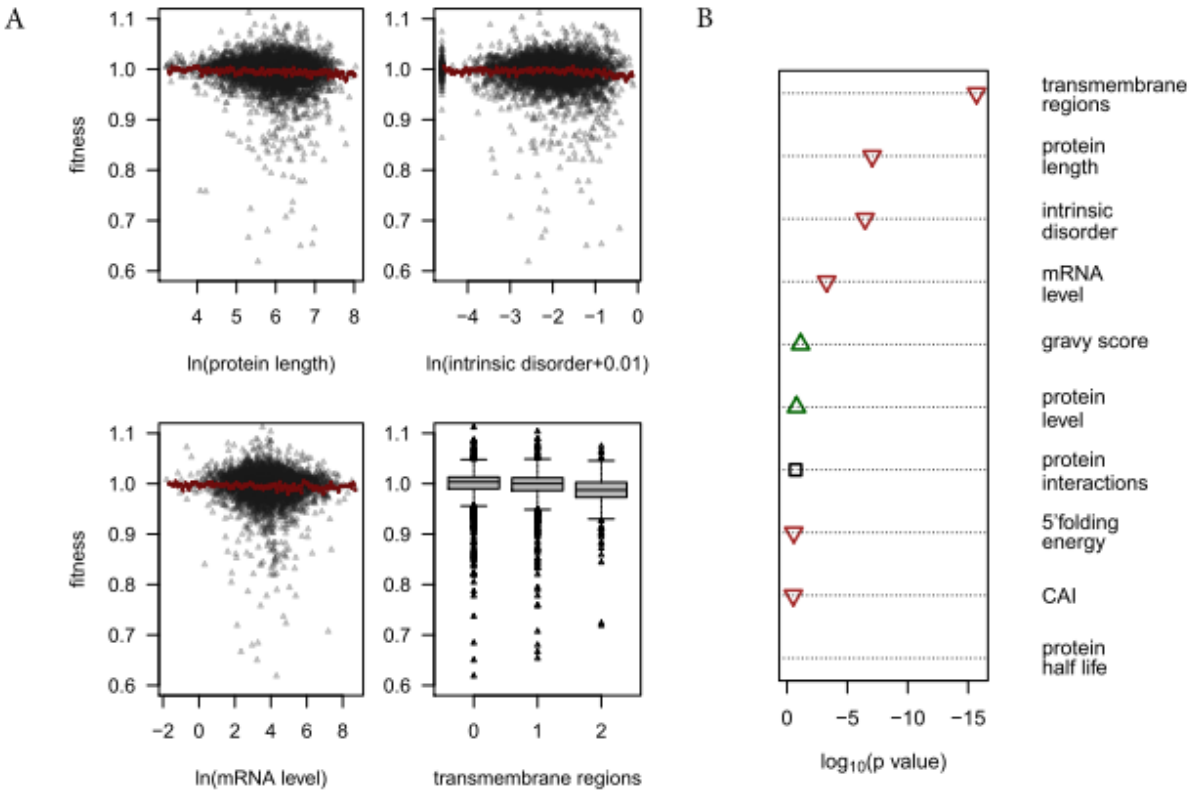
Fig. 2



Genome Biology and Evolution

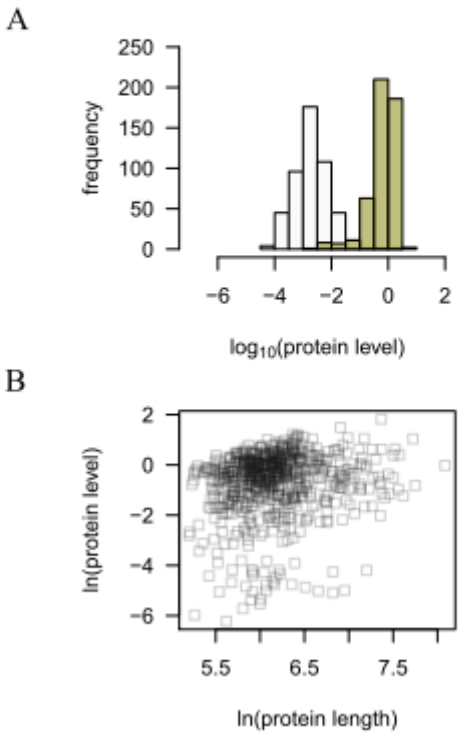
SMBE

Fig. 3



Genome Biology and Evolution

Fig. 4



SMBE