Phenotypic Mutation Rates and the Abundance of Abnormal Proteins in Yeast

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td><a href="http://dx.doi.org/10.1371/journal.pcbi.0030203">http://dx.doi.org/10.1371/journal.pcbi.0030203</a></td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:2625515">http://nrs.harvard.edu/urn-3:HUL.InstRepos:2625515</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Phenotypic Mutation Rates and the Abundance of Abnormal Proteins in Yeast

Martin Willensdorfer1,2*, Reinhard Bürger1,3, Martin A. Nowak1,4
1 Program for Evolutionary Dynamics, Harvard University, Cambridge, Massachusetts, United States of America, 2 Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, United States of America, 3 Department of Mathematics, University of Vienna, Vienna, Austria, 4 Department of Mathematics, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America

Phenotypic mutations are errors that occur during protein synthesis. These errors lead to amino acid substitutions that give rise to abnormal proteins. Experiments suggest that such errors are quite common. We present a model to study the effect of phenotypic mutation rates on the amount of abnormal proteins in a cell. In our model, genes are regulated to synthesize a certain number of functional proteins. During this process, depending on the phenotypic mutation rate, abnormal proteins are generated. We use data on protein length and abundance in Saccharomyces cerevisiae to parametrize our model. We calculate that for small phenotypic mutation rates most abnormal proteins originate from highly expressed genes that are on average nearly twice as large as the average yeast protein. For phenotypic mutation rates much above $5 \times 10^{-4}$, the error-free synthesis of large proteins is nearly impossible and lowly expressed, very large proteins contribute more and more to the amount of abnormal proteins in a cell. This fact leads to a steep increase of the amount of abnormal proteins for phenotypic mutation rates above $5 \times 10^{-4}$. Simulations show that this property leads to an upper limit for the phenotypic mutation rate of approximately $2 \times 10^{-3}$ even if the costs for abnormal proteins are extremely low. We also consider the adaptation of individual proteins. Individual genes/proteins can decrease their phenotypic mutation rate by using preferred codons or by increasing their robustness against amino acid substitutions. We discuss the similarities and differences between the two mechanisms and show that they can only slow down but not prevent the rapid increase of the amount of abnormal proteins. Our work allows us to estimate the phenotypic mutation rate based on data on the fraction of abnormal proteins. For S. cerevisiae, we predict that the value for the phenotypic mutation rate is between $2 \times 10^{-4}$ and $6 \times 10^{-4}$.


Introduction

Every biological organism is built according to information stored in its genome. Genomes composed of billions of base pairs are not unusual. This information has to be duplicated during cell replication. Since replication errors can have devastating effects, DNA replication needs to be very accurate. Estimates for error rates in Eukaryotes are as low as $5 \times 10^{-4}$ errors per base pair per replication [1]. But even flawless genetic information is useless if the cell is not able to synthesize functional proteins. Transcription and translation, the two processes involved in decoding DNA, have to be sufficiently accurate to allow a cell to build a reliable protein machinery. We refer to errors that occur during transcription and translation as phenotypic errors, and to errors that occur during DNA replication as genotypic errors. Most phenotypic errors are introduced during translation when ribosomes translate RNA sequences into amino acid sequences [2,3]. The accuracy of translation depends on the considered codon and context. In Escherichia coli it can range from $5 \times 10^{-4}$ to $1 \times 10^{-4}$ (see Table 1 for some examples), with $5 \times 10^{-4}$ as a commonly used estimate for the average frequency of errors per codon [4,5]. In comparison, Blank et al. [2] measured an E. coli error rate during transcription of $5 \times 10^{-4}$.

Measuring the genotypic mutation rate is easier than measuring the phenotypic mutation rate. Estimates of genotypic mutation rates exist for many organisms. The data show that the number of mutations per genome per replication is constant for a wide range of organisms [1]. This is in agreement with theoretical results that suggest that the number of errors per replication per genome have to be below a certain error threshold to avoid an error catastrophe at which the propagation of genetic information becomes impossible [6–9]. There are many theoretical approaches for studying the evolution of genotypic mutation rates [10–14], and one can tentatively claim that we have a basic understanding of what governs the evolution of genotypic mutation rates.

This is not the case for phenotypic mutation rates. Apparently, very little theoretical work has been devoted to this topic. A notable exception is Wilke and Drummond [15], who study translational robustness and the evolution of genespecific phenotypic mutation rates. Their work predicts selection for proteins that fold properly despite mistranslation and provides an explanation for the fact that highly expressed genes evolve slower. A closely related study, and the starting point for this investigation, is Bürger et al. [16]. They studied a model in which the total number of synthesis

Editor: Lauren Ancel Meyers, University of Texas Austin, United States of America

Received January 24, 2007; Accepted September 5, 2007; Published November 23, 2007

A previous version of this article appeared as an Early Online Release on September 6, 2007 (doi:10.1371/journal.pcbi.0030203.eor).

Copyright: © 2007 Willensdorfer et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: LL, log-likelihood

* To whom correspondence should be addressed. E-mail: ma.wi@gmx.at
Author Summary

A functional protein machinery, built from genetic information, is central to every living organism. Surprisingly, the decoding of genes into amino acid sequences is fairly inaccurate. Errors in this process (phenotypic mutations) are several orders of magnitude more frequent than errors during DNA replication (genotypic mutations). Many researchers have explored the evolution of genotypic mutation rates, but there are as yet few investigations into the evolutionary dynamics of phenotypic mutation rates. Here we present a mathematical model that describes the effect of phenotypic mutation on the amount of abnormal proteins in cells. We parameterize our model using data from yeast (Saccharomyces cerevisiae). We show that for phenotypic mutation rates above $5 \times 10^{-4}$ per amino acid, the error-free synthesis of large proteins becomes nearly impossible. We estimate the phenotypic mutation rate of $S. \textit{cerevisiae}$ to be between $2 \times 10^{-4}$ and $6 \times 10^{-4}$ per amino acid.

Attempts to produce sufficiently many functional proteins is limited and showed that the selection pressure to reduce the phenotypic mutation rate below a certain threshold vanishes.

In addition, empirical information is scarce. Most measurements of phenotypic mutation rates are limited to $E. \textit{coli}$ [5]. For fast-growing $E. \textit{coli}$ laboratory strains, a correlation was found between ribosomal accuracy and ribosomal kinetics [3,17]. This suggests that the (high) phenotypic mutation rates are a result of a cost–benefit tradeoff. More accurate ribosomes reduce speed of translation and are hence disadvantageous. Natural isolates, however, do not show such a correlation between ribosomal accuracy and ribosomal kinetics. They display a wide diversity of ribosomal kinetic properties and growth rates which suggests that the tradeoff between accuracy and kinetics is not limiting in natural populations [17–19]. Apparently, natural populations are not so obsessed with optimizing translation kinetics for fast growth under laboratory conditions. This is not surprising considering that the estimated doubling time of, for example, intestinal $E. \textit{coli}$ (40 h) is substantially longer than the doubling time of laboratory strains (0.5 h) [20].

Hence, it is not clear if the optimization of translation kinetics is governing the evolution of phenotypic mutation rates. In this paper we analyse phenotypic mutations from a genomic/proteomic viewpoint. In particular, we derive and analyze a model that allows us to calculate the amount of abnormal proteins in a cell as a function of the phenotypic mutation rate. We also evolve genotypic and phenotypic mutation rates in computer simulations that are based on properties of the $S. \textit{cerevisiae}$ genome/proteome. We discover that the current estimate for global phenotypic mutation rates of $5 \times 10^{-4}$ is at a value where the amount of erroneous proteins begins to increase exponentially with the mutation rate. Further, at this value we observe a change in the kind of genes that contribute the most to erroneous proteins. For phenotypic mutation rates below $5 \times 10^{-4}$, erroneous proteins from highly expressed genes are frequent. Above $5 \times 10^{-4}$, however, erroneous proteins from large genes begin to dominate. Finally, we study models in which individual proteins can decrease their phenotypic mutation rate by using preferred codons or evolve robustness against amino acid substitutions. We point out the similarities and differences between the two mechanisms and show how an increase of the amino acid substitution rate above $5 \times 10^{-4}$ affects the adaptation of highly expressed proteins.

Materials and Methods

In the following, we develop and analyze a model regarding the evolution of phenotypic mutation rates. We use data from $S. \textit{cerevisiae}$ to parameterize our model and calculate here relevant properties of the available yeast data.

The genotypic mutation rate in $S. \textit{cerevisiae}$ is approximately $2.2 \times 10^{-10}$ mutations per base pair per replication [1]. Our model requires the number of deleterious mutations per codon per replication as the unit for the genotypic mutation rate. Since each codon is composed of three nucleotide acids and $438/576 \approx 3/4$ single site mutations are nonsynonymous [21], the mutation rate per codon is given by

$$n = 3/4 \times 10^{-10} = 4.95 \times 10^{-11}$$. Of these nonsynonymous mutations, about 10% to 60% [22] are deleterious, placing the genotypic mutation rate somewhere between $4.95 \times 10^{-11}$ and $2.97 \times 10^{-10}$ deleterious mutations per codon per replication. For simplicity, we use $1 \times 10^{-10}$. The phenotypic mutation rate in yeast appears to be similar to the mutation rates measured in $E. \textit{coli}$ [23]. The mutation rate is therefore likely to range from $1 \times 10^{-5}$ to $5 \times 10^{-5}$, with $5 \times 10^{-4}$ as an estimate for the global phenotypic mutation rate [5].

To parameterize our model, we need the length $n_i$ and abundance $y_i$ of each protein in yeast. Complete genomic sequences [24] provide the length, $n_i$, of each protein of an organism. We only consider reading frames from the Saccharomyces Genome Database [24] that have been classified as nonspurious by Ghaemmaghami et al. [25]. This leaves us with 5,675 open reading frames (proteins) that have an average length of 496 amino acids. The effective genome is $n = \sum n_i = 2.81 \times 10^6$ residues long. Information about the abundance, $y_i$, of proteins is provided by Ghaemmaghami et al. [25]. From their data we calculate that the total amount (measured in number of amino acids) of functional proteins is given by

$$y = \sum y_i n_i = 2.03 \times 10^{10}$$ and that $y = \sum y_i n_i = 1.34 \times 10^{15}$ (this expression is relevant for Equation 8).
and $K$ genes of (possibly) different length and expression level. Gene $i$ ($i = 1, \ldots, K$) is $n_i$ amino acids long. During one cell cycle, $y_i$ functional proteins have to be synthesized from gene $i$. We assume that regulation of gene expression guarantees that the gene is expressed until $y_i$ functional proteins are present. A cell with error-free transcription and translation will therefore synthesize exactly $y_i$ proteins of gene $i$. Since we are interested in variation within a population, we can normalize the fitness of such a cell to 1 and consider only relative fitnesses. Any cost of synthesizing all the functional proteins is accounted for in the fitness value of 1. Cells, however, have a phenotypic mutation rate $u > 0$, which denotes the probability (per codon) that protein synthesis is erroneous and produces a nonfunctional protein. Hence, a phenotypic mutation is a deleterious amino acid substitution that occurs during protein synthesis. We assume that phenotypic mutations are independent of each other. Hence, the probability of synthesizing a nonfunctional protein from gene $i$ is given by $u_i = 1 - (1 - u)^{n_i}$. Let $x_i$ denote the number of nonfunctional proteins that have been synthesized until $y_i$ functional proteins were made. Then a cell synthesizes $\sum x_i$ functional and $\sum x_i$ nonfunctional proteins. It uses $y = \sum n_i y_i$ amino acids to synthesize functional proteins and $x = \sum n_i x_i$ amino acids to synthesize nonfunctional proteins.

We are interested in the cost of erroneous protein synthesis in natural populations. Natural populations grow much slower than laboratory cultures. The growth of bacteria is limited by the rate of protein synthesis per ribosome. In slow-growing bacteria, the rate of protein synthesis per ribosome is, because of limiting amounts of charged tRNAs, almost 50% lower than in fast-growing bacteria [26]. We will therefore base our cost function on the effects of the phenotypic mutation rate on the availability of charged tRNAs.

Most erroneous proteins are identified as abnormal and are degraded rapidly [27]. In this case, the amino acids that have been used to synthesize the erroneous proteins can be recycled to charge new tRNAs. This constant turnover, however, will diminish the pool of charged tRNAs by an amount depending on $x$. Hence, the cost of phenotypic mutations is a function of $x$. We will use $\eta(x)$ to denote the cost of erroneous protein synthesis. Even though we have motivated $\eta(x)$ by the cost of depleting the tRNA pool, it can account for other possible costs of erroneous proteins synthesis as well. Examples include toxic effects of aggregates and misfolded proteins, the waste of metabolic energy (ATP/GTP), or usage of the ribosomal machinery to synthesize nonfunctional instead of functional proteins. Overall, the fitness of a cell that uses $x$ amino acids to synthesize nonfunctional proteins is given by $1 - \eta(x)$. It is unnecessary to explicitly consider the cost of protein synthesis of functional proteins, $c_{\mu}$, because $y$ is constant and, hence, $1 - \eta(y) = \eta(x) \cdot 1 - \frac{\eta(y)}{\eta(x)} = 1 - \eta(x)$, where $\eta(x)$ denotes the costs if the costs of synthesis of functional proteins are not included.

Besides phenotypic, we also take genotypic mutations into account. This allows us to directly compare the cost and evolution of phenotypic and genotypic mutation rates. Genotypic mutations introduce deleterious mutations at rate $\mu$ per codon, that is, gene $i$ replicates successfully with probability $(1 - \mu)^n_i$, and all genes are replicated successfully with probability $(1 - \mu)^n$, where $n = \sum n_i$ can be interpreted as the effective genome size, i.e., the total number of amino acids encoded by all $K$ genes. We assume that the population is large enough so that deleterious mutations cannot spread. We ignore mutations that recover the wild type.

In the following, the average fitness of a population with phenotypic and genotypic mutation rates $\mu$ and $\nu$ per codon will play a central role. In the absence of genotypic mutations, the mean fitness is $1 - \bar{\eta}(u)$, where $\bar{\eta}(u) = \sum \eta(x)p_u(x)$, and $p_u(x)$ denotes the probability that a cell with a phenotypic mutation rate uses $x$ amino acids to synthesize abnormal proteins. Genotypic mutations reduce this mean fitness. Because we ignore back mutations, at mutation-selection balance the mean fitness is reduced by the factor $(1 - \mu)^n$, and this factor is independent of the actual fitness of the mutants. This is a special case of Haldane’s principle for the mutation load (see the section Mutation Rates at Equilibrium). The mean fitness at mutation-selection balance is given by

$$f(\mu, u) = (1 - \mu)^n [1 - \bar{\eta}(u)].$$

As a consequence, we get the following formula for the cost $C_\mu$ of phenotypic mutations, which is defined as the difference between the expected mean fitness of a population with and without phenotypic mutations:

$$C_\mu = f(\mu, 0) - f(\mu, u) = (1 - \mu)^n \bar{\eta}(u).$$

Similarly, the cost $C_\nu$ of genotypic mutations is

$$C_\nu = f(0, \nu) - f(\mu, \nu) = [1 - (1 - \mu)^n][1 - \bar{\eta}(u)].$$

Evolving Mutation Rates

We allow the mutation rates $\mu$ and $\nu$ to evolve. Our aim is to derive approximations for the mutation rates that evolve in the long run. For this purpose, it is convenient to consider $\mu$ and $\nu$ as quantitative traits with values between 0 and 1. Both traits have the potential to evolve due to new mutations which are assumed to occur at constant rates $\pi_\mu$ and $\pi_\nu$, respectively. These mutations will primarily increase the genotypic and phenotypic mutation rates, but some will decrease them. Because the precise form of these mutation distributions does not enter our approximate formulas given here, we introduce them only in the context of our computer simulation model below. Mutation is counteracted by selection because, on average, cells with a higher mutation rate have a lower fitness. Applying Haldane’s principle, it can be shown (see the section Mutation Rates at Equilibrium) that the evolved genotypic mutation rate, $\bar{\mu}$, can be approximated as

$$\bar{\mu} \approx \frac{\nu}{\mu}/n.$$  

Similarly, the phenotypic mutation rate equilibrates to a value $\bar{\nu}$, which is obtained (approximately) by solving the equation

$$\pi_\nu = \bar{\mu}$$

for $\nu$. If, as is likely the case, $\bar{\eta}$ is monotone increasing, this solution $\bar{\nu}$ is uniquely determined. By taking the ratio of Equation 4 and Equation 5, and performing a simple rearrangement, we obtain

$$\bar{\mu} n \approx \frac{\bar{\eta}(\bar{\nu}) \pi_\nu}{\pi_\mu}.$$  

The term $\bar{\mu} n$ gives the number of mutations per genome per replication and is surprisingly constant for a wide range of organisms [1]. The rates $\pi_\mu$ and $\pi_\nu$ at which the genotypic and phenotypic mutation rates are changed will primarily
phenotypic mutations and abnormal proteins. Here, we write synthesis an abnormal protein. As a consequence, we can possible range of phenotypic mutation rates \(1 \leq c \leq 3\) until exactly \(g_i\). We assume henceforth \(g_i\) functions of the evolved genotypic and phenotypic mutation rates. This ratio depends on (i) the effective genome size, \(\sum n_i\), (ii) the average total cost of abnormal protein synthesis, \(c \sum n^*_i y_i\), and (ii) the ratio at which mutations of the two mutation rates occur. Below, we will use computer simulations to complement these analytical considerations.

**Simulating the Evolution of Mutation Rates**

We simulate the evolution of phenotypic and genotypic mutation rates based on the model introduced above. The main difference is that we use a finite (effective) population of size \(N = 10^3\) and that we specify our mutation distributions for the mutation rates. Each generation, \(N\) organisms are selected (with replacement) for reproduction with probabilities proportional to their fitness (Wright-Fisher model of drift and selection). Fitness is calculated as \(1 - \eta(u) = 1 - \bar{c}\), where \(\bar{c}\) gives the expected amount (in amino acids) of abnormal proteins. Protein length and abundances are taken from *S. cerevisiae* (see Materials and Methods). The number of expected abnormal proteins is calculated according to Equation 7. To avoid the fixation of genotypic mutants for high genotypic mutation rates at the beginning of the simulations, we set the fitness of genotypic mutants to zero. As predicted by the theory, the fitness of the genotypic mutants does not affect the equilibrium mutation rates (see the section Effect of Initial Values and Parameters on the Simulation Results).

The initial population is homogeneous with equal phenotypic and genotypic mutation rate. To allow the evolution of mutation rates, we change (mutate) \(u\) and \(\mu\) with probabilities \(\pi_u\) and \(\pi_\mu\), respectively. Unless otherwise mentioned, we assume \(\pi_u = \pi_\mu = 10^{-6}\). In the section Effect of Initial Values and Parameters on the Simulation Results, we show that changes in the initial values of \(u\) and \(\mu\) do not affect the results of our simulations and changes in \(c\) affect them as predicted by the theory. Since beneficial mutations, that is, mutations decreasing the mutation rate, are generally rare, we increase the mutation rate with a probability of 0.99 (conditional on a mutation event). In this case, we draw a new phenotypic mutation rate from a beta distribution \(B(\alpha = 1, \beta = 9)\) on \([u, 2u]\) or on \([\mu, 2\mu]\) for genotypic mutation rates. Hence, the average increase is 10% and small changes are more frequent than large ones. Similarly, in case of a decrease, we draw the new mutation rate from a reflected beta distribution \(B(\alpha = 1, \beta = 9)\) on \([0, u]\) or \([0, \mu]\). Hence, small changes are again more likely than large ones.

During a simulation run, we kept track of the ancestry of each individual. After \(4 \times 10^6\) generations, we can use this line of descent to observe the evolution of phenotypic and genotypic mutation rates. For each parameter combination, we con-

\[
S \approx u \sum n_i^2 y_i, 
\]

and \(\bar{\eta}(u) \approx uc \sum n^*_i y_i\). Therefore, Equation 6 can be rearranged to

\[
\frac{\bar{\eta}(u)}{\bar{\mu}} \approx c \sum n_i \pi_u \pi_\mu. 
\]

This illustrates nicely the factors that determine the ratio of the evolved genotypic and phenotypic mutation rates. This ratio depends on (i) the effective genome size, \(\sum n_i\), (ii) the average total cost of abnormal protein synthesis, \(c \sum n^*_i y_i\), and (iii) the ratio at which mutations of the two mutation rates occur. Below, we will use computer simulations to complement these analytical considerations.

**Simulating the Evolution of Mutation Rates**

We simulate the evolution of phenotypic and genotypic mutation rates based on the model introduced above. The main difference is that we use a finite (effective) population of size \(N = 10^3\) and that we specify our mutation distributions for the mutation rates. Each generation, \(N\) organisms are selected (with replacement) for reproduction with probabilities proportional to their fitness (Wright-Fisher model of drift and selection). Fitness is calculated as \(1 - \eta(u) = 1 - \bar{c}\), where \(\bar{c}\) gives the expected amount (in amino acids) of abnormal proteins. Protein length and abundances are taken from *S. cerevisiae* (see Materials and Methods). The number of expected abnormal proteins is calculated according to Equation 7. To avoid the fixation of genotypic mutants for high genotypic mutation rates at the beginning of the simulations, we set the fitness of genotypic mutants to zero. As predicted by the theory, the fitness of the genotypic mutants does not affect the equilibrium mutation rates (see the section Effect of Initial Values and Parameters on the Simulation Results).

The initial population is homogeneous with equal phenotypic and genotypic mutation rate. To allow the evolution of mutation rates, we change (mutate) \(u\) and \(\mu\) with probabilities \(\pi_u\) and \(\pi_\mu\), respectively. Unless otherwise mentioned, we assume \(\pi_u = \pi_\mu = 10^{-6}\). In the section Effect of Initial Values and Parameters on the Simulation Results, we show that changes in the initial values of \(u\) and \(\mu\) do not affect the results of our simulations and changes in \(\pi_u\) and \(\pi_\mu\) affect them as predicted by the theory. Since beneficial mutations, that is, mutations decreasing the mutation rate, are generally rare, we increase the mutation rate with a probability of 0.99 (conditional on a mutation event). In this case, we draw a new phenotypic mutation rate from a beta distribution \(B(\alpha = 1, \beta = 9)\) on \([u, 2u]\) or on \([\mu, 2\mu]\) for genotypic mutation rates. Hence, the average increase is 10% and small changes are more frequent than large ones. Similarly, in case of a decrease, we draw the new mutation rate from a reflected beta distribution \(B(\alpha = 1, \beta = 9)\) on \([0, u]\) or \([0, \mu]\). Hence, small changes are again more likely than large ones.

During a simulation run, we kept track of the ancestry of each individual. After \(4 \times 10^6\) generations, we can use this line of descent to observe the evolution of phenotypic and genotypic mutation rates. For each parameter combination, we con-
seven sets of simulations that use different values for $c$.

Materials and Methods) and because we assume $p_l = p_u$, Equation 9 predicts $\bar{u}/\bar{v} = 2.10 \times 10^{-7}/c$. For $c = 10^{-10}$, this equals $2.10 \times 10^2$ and is indeed very close to the observed value of $\bar{u}/\bar{v} = 2.00 \times 10^{-9}$, which is very close to $C_u$.

As expected, a decrease in the cost of abnormal proteins, $c$, results in an increase of the phenotypic mutation rate. But, apparently, there is an upper limit for phenotypic mutation rates above which a further decrease of $c$ does not increase it much more (compare brown and grey curves in Figure 1). Also, for very small $c$, Equation 9 becomes inaccurate. For example, if $c = 10^{-13}$, we get $\bar{u}/\bar{v} = 8.74 \times 10^{-7}/1.01 \times 10^{-9} = 0.87 \times 10^6$ instead of $\bar{u}/\bar{v} = 2.11 \times 10^6$.

We can explain both observations by considering the average number $x$ of amino acids required to synthesize nonfunctional proteins (Equation 7) and the total number $\sum x_i$ of nonfunctional proteins. For brevity and in distinction to the number of nonfunctional proteins, $\sum x_i$, we henceforth refer to $\bar{x} = \sum x_i\bar{x}_i$ as the amount of abnormal proteins. (We use the number of amino acids as the unit for the amount of proteins.) Figure 2 displays these quantities as functions of $u$ (solid line for $\bar{x}$, dash-dotted line for $\sum x_i$) as well as the linear approximation (Equation 8) to $\bar{x}$ (dashed line). As one can see, the linear approximation becomes inaccurate if $u > 5 \times 10^{-4}$, and $\bar{x}$ begins to grow exponentially with $u$. Consequently, even a small increase in $u$ will cause a tremendous increase in $\bar{x}$. Even if abnormal proteins are not very costly, the rapid increase in $\bar{x}$ prevents a further increase of $u$.

We used $(1 - u)^{n_i} - 1 = u n_i$ to linearize $\bar{x}$. This approximation is only accurate if $u n_i$ is sufficiently small. For a protein length of about 890 amino acids (see below for an explanation why we chose 890) and a phenotypic mutation rate of $5 \times 10^{-4}$, we have $u n_i = 0.445$, which is apparently too large for the approximation to be accurate. For $n_i = 890$ and $u = 5 \times 10^{-4}$, the exact value for $(1 - u)^{n_i} - 1$ is 0.56; this illustrates the difference between the true value of $\bar{x}$ and its linear approximation at $u = 5 \times 10^{-4}$. It is important to emphasize that the observed upper bound for $u$ is a consequence of the protein length distribution and the expression profile of the organism, as we will see below.

Components of $\bar{x}$

It would be interesting to know the length and expression level of the genes that contribute most to the amount of abnormal proteins in a cell. For a given phenotypic mutation rate, it is easy to calculate $n_i\bar{x}_i$, the amount (in amino acids) of erroneous proteins that were produced from gene $i$. To determine which gene lengths and expression levels are most important for $\bar{x}$, we calculate weighted averages of $n_i$ and $\bar{x}_i$. As weights, we use the amount of erroneous proteins that stem from gene $i$, that is, $n_i\bar{x}_i$. Hence, we have

$$\sum_i n_i \bar{x}_i / \sum_i n_i \bar{x}_i$$

and

$$\sum_i \gamma_i n_i \bar{x}_i / \sum_i n_i \bar{x}_i$$

as indicators of the average protein length and expression level, respectively, that are most important for the amount of abnormal proteins in the cell. Since the expected number of abnormal proteins, $\bar{x}$, depends on the phenotypic mutation rate, the weighted averages are functions of $u$ as well. How they change as a function of $u$ is shown in Figure 3. Interestingly, the weighted average of $n_i$ for very small $u$ is about 890 and nearly twice as large as 496, the average protein length in yeast. Even more interestingly, this value
Phenotypic Mutations and Abnormal Proteins

Figure 3. Weighted Averages of Protein Length (10) and Expression Level (11)

To determine which protein and expression levels are most relevant for the amount of abnormal proteins in a cell, we calculated weighted averages of protein length (solid line) and expression level (dashed line). As weights, we used the amount of expected abnormal proteins, $n_i x_i$. For small phenotypic mutation rates, highly expressed proteins are most relevant for the amount of abnormal proteins in a cell. This begins to change at $u = 5 \times 10^{-4}$, when lowly expressed, large proteins begin to dominate. Inaccurate protein synthesis makes it practically impossible to synthesize these larger proteins error-free.

doi:10.1371/journal.pcbi.0030203.g003

Figure 4. Relative Contribution of Each Gene to the Amount of Erroneous Proteins in Yeast if $u = 5 \times 10^{-4}$

The solid line shows the cumulative contribution of each gene to $\bar{x}$. A steep increase, as seen here, indicates that few genes are responsible for most of the abnormal proteins in a cell. We also plot the cumulative distribution for the three components of $n_i x_i$: protein length, $n_i$ (dashed line), number of functional proteins, $y_i$ (dotted line), and expected amount of erroneous proteins to synthesize one error-free protein, $u_i/(1 - u_i) = (1 - (1 - u)^n)/(1 - u)^n$ (dashed-dotted line). The curvature of the solid line is similar to the curvature of the dotted line. Hence, most of the abnormal proteins stem from few genes because these genes are also expressed at a very high level.

doi:10.1371/journal.pcbi.0030203.g004

increases suddenly as $u$ increases beyond $5 \times 10^{-4}$. For high values of $u$, phenotypic mutations are so frequent that it becomes essentially impossible to synthesize large proteins accurately. These large proteins dominate the cost of phenotypic errors. This can also be seen in the change of the weighted average of the expression level. For low mutation rates, the amount of erroneous proteins is dominated by highly expressed genes with a weighted average expression level of $2.67 \times 10^5$ proteins per cell. This value begins to decrease at $u = 5 \times 10^{-4}$ because large proteins, instead of highly expressed proteins, begin to increasingly contribute to the amount of abnormal proteins in the cell.

For $u = 5 \times 10^{-4}$, the bulk of the amount of erroneous proteins comes from proteins that are noticeably larger than the average protein and are highly expressed. How much these proteins contribute to $\bar{x}$ compared with the rest of the genome can be seen in Figure 4. The solid line shows $\sum_{i=1}^{x_i} n_i x_i/\bar{x}$ for $k = 1, 2, \ldots, 5675$, that is, the cumulative contribution of each gene to $\bar{x}$. Genes are sorted decreasingly by their contribution to $\bar{x}$. It is obvious that only few genes contribute to most of the abnormal proteins in the cell. In fact, $5\%$ ($10\%$) of the genes contribute to $78.6\%$ ($87.5\%$) of the abnormal proteins in a yeast cell. The average length of these proteins is 927 (818) which confirms the conclusion from above that genes that contribute most to the amount of abnormal proteins are much larger than the average gene.

From Equation 7 we know that $n_i \bar{x}_i = n_i y_i (1 - (1 - u_i)^n) / (1 - u_i)^n$ and can distinguish three components: (i) the protein length, $n_i$, (ii) the expression level, $y_i$, and (iii) the expected number of erroneous proteins that have to be synthesized to get one error-free protein, $u_i/(1 - u_i)$, with $u_i = 1 - (1 - u)^n$. Which of these components is primarily responsible for the fact that only few genes contribute to most of the abnormal proteins in a cell? To answer this question, we can compare the dashed, dotted, and dash-dotted lines in Figure 4 which show how unevenly genes contribute to each of the three components (for $u = 5 \times 10^{-4}$). For example, the dotted line shows $\sum_{i=1}^{x_i} n_i y_i / \sum_{i=1}^{x_i} y_i$ for $k = 1, 2, \ldots, 5675$. From the three components, only the expression level (dotted line) shows a curvature similar to the solid line. Hence, the fact that few genes contribute to most of the abnormal proteins in a cell is due to differences in expression levels rather than differences in protein length.

Adaptation of Highly Expressed Proteins

If most of the abnormal proteins in a cell are synthesized by few, highly expressed proteins, the cell could reduce the cost of phenotypic mutation rates considerably by decreasing the phenotypic mutation rate for these few genes. In fact, highly expressed genes are special in many ways. They use preferred codons more frequently than “normal” genes [28] and evolve more slowly [29]. The usage of preferred codons conveys several advantages, among them is a more efficient [30–32] and accurate [33–35] translation. As argued by Drummond et al. [29], the slow rate of evolution of highly expressed genes might be the result of selection for translational robustness, that is, the ability of proteins to work properly despite amino acid substitutions [15,29]. The effect of preferred codon usage and translational robustness are conceptually very different. The usage of preferred codons reduces the phenotypic
mutation rate by reducing the amino acid substitution rate, while an increase in translational robustness reduces the phenotypic mutation rate by improving a protein’s ability to withstand the effect of amino acid substitutions. Let us first consider translational robustness.

**Selection for translational robustness**

Let \( u_{aa} \) denote the amino acid substitution rate. Together with the robustness of a protein against amino acid substitutions it determines the protein’s phenotypic mutation rate \( u \). According to Bloom et al. [36], the probability that a protein retains its wild-type structure after \( m \) amino acid substitutions is given by

\[
P_f(m) = 1 - \sum_{n=0}^{m} \binom{n}{m} (1 - u_{aa})^m u_{aa}^n,
\]

where \( v \) denotes the average neutrality to amino acid substitutions (m-neutrality), that is, the average probability that a protein will be unaffected by (“neutral” to an additional amino acid substitution. In the following we make the conservative assumption that a protein is functional if it is able to fold into its wild-type structure and that the wild-type sequence always folds into its wild-type structure, i.e., that \( P_f(0) = 1 \). Consequently, \( P_f(m) = v^m \) is the probability that a protein exposed to \( m \) amino acid substitutions is functional.

A protein contains \( m \) amino acid substitutions with probability

\[
p(m) = \binom{n}{m} (1 - u_{aa})^{n-m} u_{aa}^m,
\]

and is therefore functional with probability

\[
p_F(i) = \sum_{m=0}^{\infty} p(m) P_f(m) = \sum_{m=0}^{\infty} \binom{n}{m} (1 - u_{aa})^{n-m} u_{aa}^m v^m.
\]

The probability to synthesize a nonfunctional protein from gene \( i \) is given by

\[
u_i = 1 - p_F(i).
\]

For protein \( i \), the average number of nonfunctional proteins is given by

\[
\bar{x}_i = \frac{y_i}{1 - u} = y_i (p_F^{-1} - 1),
\]

and the overall average number of nonfunctional proteins in a cell is given by

\[
\bar{x} = \sum_i n_i \bar{x}_i = \sum_i n_i y_i (p_F^{-1} - 1) = \sum_i n_i y_i \left( \sum_{m=0}^{\infty} \binom{n}{m} (1 - u_{aa})^{n-m} u_{aa}^m v \right)^{-1} - 1.
\]

In comparison with Equation 7, we note that \((1 - u)^{-n}\), the term that is responsible for the rapid, nonlinear increase of \( \bar{x} \), has been replaced by a sum over the number of amino acid substitutions. Since \( n_i \) is usually much larger than the number of amino acid substitutions, we have \( n_i \approx n_i - m \) for relevant \( m \) values and can approximate the sum very accurately by

\[
(1 - u_{aa})^{-n} \sum_{m=0}^{m} \binom{n}{m} (u_{aa} v_i)^m.
\]

We see that the term that caused the dramatic increase in the previous model also appears in this model, which considers translational robustness. The phenotypic mutation rate \( u \) is replaced by \( u_{aa} \) and the term \((1 - u)^{-n}\) multiplied by a factor that depends on the protein’s m-neutrality \( v_i \).

Analogous to our previous observations, we can expect a nonlinear increase of \( \bar{x} \) for \( u_{aa} > 5 \times 10^{-4} \).

In theory, but not in practice, it is possible to reduce the phenotypic mutation rate to zero by increasing \( v_i \) to 1 for all proteins. In practice, an upper limit for \( v_i \) is given by the function and stability of the protein. In our simulations, this upper limit is set by a prior distribution for the values of \( v_i \). Very high values for \( v_i \) will be possible but unlikely. Given this (soft) upper limit for m-neutrals, we can, for a given amino acid substitution rate, ask which proteins will be selected for translational robustness (large m-neutrals) and what amount of abnormal proteins can be expected.

We present simulations in which m-neutrals are drawn from a beta distribution, \( B(v_i | a, b) \propto v^{-a - 1} (1 - v)^{b - 1} \) with \( a = 16.95 \) and \( b = 20.72 \), which has variance 0.0064 and mean 0.45, reflecting the mean and variance of m-neutrals of seven proteins estimated by Bloom et al. [36]. We want to emphasize that the quantitative results, in particular the relative changes as a function of \( u_{aa} \), are not affected when other (reasonable) prior distributions are used. We obtained similar results for a beta distribution with mean 0.5 and variance 0.02 and for corresponding normal distributions (truncated to the interval [0, 1]). Even though the absolute value of \( \bar{x} \) is smaller for prior distributions that allow larger m-neutrals, the relative changes remain the same.

Mutations generate m-neutrals \( v_i \) for each protein from the prior distribution. Selection determines if an m-neutrality reaches fixation and subsequently the eventual distribution of \( v_i \) after many generations of selection. By how much selection has caused a protein’s m-neutrality to deviate from the prior distribution can be expressed in terms of the log likelihood (LL) of the \( v_i \) values under the prior distribution, \( \log B(v_i | a, b) = (a - 1) \log (v_i) + (b - 1) \log (1 - v_i) \). If selection leads to a significant increase of a protein’s m-neutrality, then the LL of this m-neutrality will be very low. The average LL for a cell’s proteins is given by

\[
\log \left( \prod_{i=1}^{K} B(v_i | a, b) \right)^{1/K} = \frac{1}{K} \sum_{i=1}^{K} \log [B(v_i | a, b)]
\]

and can be used to quantify the extent of selection for translational robustness that the proteins of a cell were exposed to. As we will see, the LL of the m-neutrals (after selection) decreases substantially for \( u_{aa} > 5 \times 10^{-4} \) and indicates the intensified selection for translational robustness.

We use a drift-selection model based on the fixation probability in a Moran process to determine the \( v_i \) values at mutation–selection balance (the post-selection \( v_i \) distribution). For a given amino acid substitution rate, we initialize \( v_i \) for all proteins by setting it equal to the mean of the prior distribution and calculate \( \bar{x} \). For every protein (gene) \( i \), we draw a new \( v_i \) from the prior distribution and calculate the new \( \bar{x} \) that reflects this change, \( \bar{x}_{\text{new}} = \bar{x} + n_i y_i (p_{F,\text{new}}^{-1} - p_{F}^{-1}) \). For computational reasons and since the binomial distribution has most of its weight at small values of \( m \), we truncate the summation over \( m \) in Equation 14 to the smallest integer larger than \( n_i u_{aa} + 4 \sqrt{n_i u_{aa}(1 - u_{aa})} \) (i.e., \( m \)-values that are more than four standard deviations away from the mean are ignored; the probability of getting \( m \)’s above this threshold is less than \( 2 \times 10^{-5} \)). We accept the new \( v_i \) with probability \( (1 - 1/ r) / (1 - 1/b^3) \), which corresponds to the fixation probability in a...
The simulations conducted in this section implicitly assume that the population is homogeneous and that one mutant appears at a time and either goes extinct or gives rise to its corresponding LL decreases more rapidly than for the other two groups of proteins. Hence, proteins with large yin respectively. As expected, proteins with large m-neutralities than large or highly expressed proteins and, accordingly, have the smallest LL. With increasing $u_{aa}$, large proteins contribute more to the amount of abnormal proteins and the corresponding LL decreases more rapidly than for the other two groups of proteins. In Figure 6 we show the m-neutralities of individual proteins for three amino acid substitution rates. It is also evident by the decline of the LL. The lower panel in Figure 5 shows the change in the average LL of three sets of 100 proteins. The three sets of proteins are given by the 100 proteins with the largest $n_i$, the 100 proteins with the largest $y_i$, and the 100 proteins with the largest $y_i n_i^2$. For 100 proteins, the 5% (0.1%) quantile for the average LL is $-25.117$ $(-25.215)$. The lower panel shows that the extent of selection for translational robustness increases nonlinearly for large proteins whereas it increases approximately linearly in the other two groups of proteins.

doi:10.1371/journal.pcbi.0030203.g005

Figure 5. Selection for Translational Robustness as a Function of Amino Acid Substitution Rate

The top panel shows the average amount of abnormal proteins, $\bar{x}$, after 500,000 cycles of mutation and selection (see text for more detail). The mid-panel shows the average log-likelihood (LL) of the m-neutralities after selection. For $u \geq 1.92 \times 10^{-3}$, the average LLs are significantly lower than expected by chance. The $5\%$ (0.1%) quantile is given by $-25.024$ ($-25.040$). The total LL decreases noticeably for $u_{aa} > 5 \times 10^{-4}$. The lower panel shows the average LL of three sets of 100 proteins. We consider the 100 largest proteins, the 100 most highly expressed proteins, and the 100 proteins with the largest $y_i n_i^2$. For 100 proteins, the 5% (0.1%) quantile for the average LL is $-25.117$ ($-25.215$). The lower panel shows that the extent of selection for translational robustness increases nonlinearly for large proteins whereas it increases approximately linearly in the other two groups of proteins.
another homogeneous population. Hence, every organism in these simulations represents a homogeneous population of size $N$. This organism is of course also the MRCA of this population. We can compare its fitness, $$f = e^{c\bar{x}}/C_0$$ with the fitness, $$f' = 1/C_0$$ for $c \bar{x} \approx 0.005$, of the MRCA in our previous simulations to speculate on what would happen if we allowed $u_{aa}$ to change here as well. Here, where $u_{aa}$ was held constant, the fitness of the organism converged to fairly small values compared to the equilibrium values of $f' \approx 0.995$ in our previous simulations which allowed changes in $u_{aa}$. For $u_{aa} = 1 \times 10^{-5}$, where selection for higher m-neutralities is insignificant, $f = 0.938$, and $f$ is much lower for larger phenotypic mutation rates (e.g., $f = 0.090$ for $u_{aa} = 5 \times 10^{-5}$). This suggests that if $u_{aa}$ is able to evolve freely to equilibrium values of $f \approx 0.995$, then selection for translational robustness will be insignificant. Simulations in which we mutated $u_{aa}$ as described in the previous section confirmed this expectation. No significantly elevated m-neutralities evolved (unpublished data).

This was not the case in simulations with few (e.g., ten) genes, where the LL of highly expressed proteins converged to significantly lower values. Apparently, if there are many genes and if $u_{aa}$ is in mutation-selection balance, the m-neutralities of individual proteins do not contribute enough to allow selection for higher m-neutralities. But as we have seen above, if $u_{aa}$ is constant, selection for larger m-neutralities can reduce $\bar{x}$ to some extent. Hence, if $u_{aa}$ is above its mutation-selection-balance value, then significantly higher m-neutralities will evolve and decrease the phenotypic mutation rate by decreasing the effect of amino acid substitutions.

### Selection for Preferred Codons

Besides increasing the translational robustness of certain proteins, a cell can also use preferred codons to decrease the phenotypic mutation rate $u$. This would actually decrease the amino acid substitution rate and is therefore conceptually different from translational robustness, which reduces the effect of amino acid substitutions but not their occurrence. Considering codon usage, the amino acid substitution rate $u_{aa}$ has two components, a ribosomal component $u_r$ and a codon-based component, $u_c$. We assume that $u_{aa} = u_r u_c$ for preferred codons and that $u_{aa} = u_r$ for nonpreferred codons. Preferred codons are more accurate than nonpreferred codons, hence, $u_c < 1$. In this section we ignore translational robustness, i.e., $u = u_{aa}$. A protein of length $n_i$ that uses $\hat{n}_i$ preferred codons synthesizes a functional protein with probability $(1 - u_r)^{n_i - \hat{n}_i}(1 - u_r u_c)^{\hat{n}_i}$. The average amount of abnormal proteins is given by

---

**Figure 6.** m-neutralities, $v_i$. After Selection for Three Different Amino Acid Substitution Rates $u_{aa}$

The data points are sorted according to the length of the proteins. The increased effectiveness of selection for higher m-neutralities in large proteins for high amino acid substitution rates is clearly visible.

doi:10.1371/journal.pcbi.0030203.g006
After changing $\tilde{n}_i$, we calculate $\tilde{x}$ and accept the new $\tilde{n}_i$ as described in the section Selection for Translational Robustness. We report the average of 20 simulations for each $u_r$.

Each simulation was terminated after $2 \times 10^7$ sequential mutations (not necessarily fixation) of $\tilde{n}_i$.

Figure 7 shows the results of our simulations. Analogous to Figure 6, we plot the equilibrium fraction of preferred codons, $p_i = \tilde{n}_i/n_i$, for each gene for three ribosomal amino acid substitution rates $u_r$. For $u_r = 1.37 \times 10^{-4}$, only few genes evolve a major codon bias of $p_i > 0.6$. The gene with the largest codon bias of about 0.8 encodes for the protein with the largest $\tilde{x}$ and contributes 5.7% to the total amount of functional proteins and 7.9% to the total amount of abnormal proteins.

For $u_r > 5 \times 10^{-4}$, large proteins begin to contribute more to the amount of abnormal proteins and selection increases the codon bias of large proteins. Similar to our observation in Figure 2, the codon bias cannot prevent the drastic increase of $\tilde{x}$ for $u_r > 5 \times 10^{-4}$. As before, no significant codon bias evolved if we allowed $u_r$ to change as well.

Comparing Figure 6 with Figure 7, we notice that selection for translational robustness results in a more distinct bias in $\tilde{x}$ than what we observe for $p_i$ after selection for preferred codons. In the next section we compare the two mechanisms to identify the source of this difference.

Similarities between Preferred Codons and Translational Robustness

In our simulations, the two mechanisms differ in the way $p_{F,i}$, the probability of synthesizing a functional protein, is calculated and in the way it is mutated.

For the translational robustness model, we calculate $p_{F,i}$ as
\[
p_{F,i} = \sum_{n=0}^{n_i} \binom{n_i}{m} (1 - u_{aa})^{n_i-m} u_{aa}^m v_i^n \approx (1 - u_{aa})^{n_i} \sum_{n=0}^{n_i} \binom{n_i}{m} (u_{aa}v_i)^n.
\]

The approximations are accurate as long as proteins with many amino acid substitutions are rare. For the preferred codon model, we have
\[
p_{F,i} = (1 - u_i)^{n_i - n_i} (1 - u_i u_i)^{n_i} \\
= (1 - u_i)^{n_i} [1 + u_i]^{n_i} (1 - u_i) \\
= (1 - u_i)^{n_i} \sum_{n=0}^{n_i} \binom{n_i}{m} [u_i p_i (1 - u_i)]^n,
\]
where we used \(1 + u_i(1 - u_i)^{n_i} 1 + u_i \) and \([1 + u_i(1 - u_i)]^n \approx 1 + u_i p_i (1 - u_i)\), which are reasonable approximations if \(u_i\) is small. The analogy between Equation 21 and Equation 22 is obvious. In theory, the m-neutrality, \(v_i\), can range from 0 to 1. In our simulations, for the chosen prior distribution, \(v_i\)'s larger than 0.8 are rare. The m-neutrality \(v_i\) is analogous to the term \(p_i(1 - u_i)\) in the preferred codon model. Since \(p_i\) can range from 0 to 1, the two mechanisms can reduce the amount of abnormal proteins equally well if \(v_{max} = 1 - u_{max}\), where \(v_{max}\) denotes the upper limit for m-neutrality. In our simulations, we have \(v_{max} \approx 0.8 < 1 - u_{max} = 0.9\). Hence, we would expect lower \(v_i\) values in the preferred codon model.

This is not the case. For example, for \(u_{aa} = 10^{-7}\), \(\bar{x}\) converged to \(6.4 \times 10^7\) in the translational robustness model, whereas \(\bar{x}\) converged to \(7.3 \times 10^7\) in the preferred codon model. Hence, we have to consider the way in which the \(p_{F,i}\)'s are mutated to understand this result.

In the translational robustness model, \(v_i\) is sampled from a prior distribution. The new \(v_i\) value is independent of the previous one. Hence, large changes of \(v_i\) and, consequently, of \(p_{F,i}\) are possible. In the preferred codon model, the number of preferred codons can only change in increments of one, and corresponding changes of \(p_{F,i}\) and \(\bar{x}\) are small. The small changes in \(p_i\) and, therefore, in \(p_{F,i}\) allow the evolution of noticeable codon biases only in genes that produce large amounts of abnormal proteins. In the translational robustness model, large changes in the \(p_{F,i}\)'s are possible, and they have a higher fixation probability.

Take, for example, the protein with the largest value of \(v_i n_i^2\). It is 918 amino acids long, and a change of \(n_i\) from 459 to 460 increases \(p_i\) from 0.5 to 0.501 (by 0.2%). This small change reaches fixation only if the costs of abnormal proteins from this gene are very large. Changes larger than this are frequent in the translational robustness model and have a higher probability of fixation.

**Discussion**

A functional protein machinery, built from genetic information, is central to every living organism. Surprisingly, the decoding of genes into amino acid sequences is fairly inaccurate. Errors (phenotypic mutations) occur several orders of magnitude more frequently than during DNA replication. The frequency of errors depends on the codon and its context (see Table 1).

In this paper, we have explored the evolution of pheno-
typic mutation rates. In our model, a cell maintains protein synthesis until a certain number of functional proteins are present. Depending on the phenotypic mutation rate, \(u\), a certain number of amino acids, \(x\), are “wasted” in erroneous proteins and reduce the fitness of the organism by \(\eta(x)\). For simplicity, we used a linear cost function \(\eta(x) = cx\). With genomic and proteomic data from *S. cerevisiae* [24,25], we discover (a) an effective upper bound for the phenotypic mutation rate, (b) that most of the abnormal proteins stem from genes that are highly expressed and substantially larger than the average yeast protein, (c) that an average phenotypic mutation rate of \(u = 5 \times 10^{-4}\) is at a value where \(x \) begins to increase dramatically as a function of \(u\) and large, lowly expressed genes begin to contribute substantially to the amount of abnormal proteins, and (d) that an increased codon bias or translational robustness in highly expressed genes can reduce the amount of abnormal proteins but cannot stop the dramatic increase for amino acid substitution rates above \(5 \times 10^{-4}\).

To what extent do our results depend on the assumption that \(\eta(x)\) is linear and that gene expression is maintained until a certain number of functional proteins are present? Dekel and Alon [37] found a convex increase of the cost of protein synthesis with the amount of proteins synthesized. Considering this and that aggregates of misfolded proteins are, in a concentration-dependent way, toxic to cells [27,38], we can expect the cost of erroneous proteins to increase faster than linear with the amount of erroneous proteins produced. A nonlinear \(\eta(x)\), however, would only affect the position of the upper bound for \(u\), which we observed in our simulations (see Figure 1). For a nonlinear cost function, we would expect this upper bound to be lower than what we have observed here, because of the nonlinear increase of \(x\) on top of the nonlinear increase of the costs of \(x\). Results (b)–(d) are not affected by the shape of the cost function.

Let us now consider our assumption about the regulation of gene expression. The largest protein in the yeast genome is Mdn1p, a dynein-related AAA-type ATPase [24,39]. It is 4,910 amino acids long. For \(u = 5 \times 10^{-4}\), only 12.8% of the synthesized proteins are error-free. To get the required number of 0.538 \(\times 10^9\) error-free proteins [25], the cell has to synthesize 6 \(\times 10^9\) proteins. This is not a tremendous burden considering that about 46,600 \(\times 10^9\) functional proteins are synthesized in total. However, this number increases rapidly if \(u\) increases. Doubling or quadrupling \(u\) would require the synthesis of 72.7 \(\times 10^9\) or 10,000 \(\times 10^9\) proteins, respectively. It is unrealistic to assume that a cell will synthesize 100 proteins. This is not a tremendous burden considering that about 46,600 \(\times 10^9\) functional proteins are synthesized in total. However, this number increases rapidly if \(u\) increases. Doubling or quadrupling \(u\) would require the synthesis of 72.7 \(\times 10^9\) or 10,000 \(\times 10^9\) proteins, respectively. It is unrealistic to assume that a cell will synthesize 100 proteins. This is not a tremendous burden considering that about 46,600 \(\times 10^9\) functional proteins are synthesized in total. However, this number increases rapidly if \(u\) increases. Doubling or quadrupling \(u\) would require the synthesis of 72.7 \(\times 10^9\) or 10,000 \(\times 10^9\) proteins, respectively. It is unrealistic to assume that a cell will synthesize 100 proteins. This is not a tremendous burden considering that about 46,600 \(\times 10^9\) functional proteins are synthesized in total. However, this number increases rapidly if \(u\) increases. Doubling or quadrupling \(u\) would require the synthesis of 72.7 \(\times 10^9\) or 10,000 \(\times 10^9\) proteins, respectively. It is unrealistic to assume that a cell will synthesize 100 proteins. This is not a tremendous burden considering that about 46,600 \(\times 10^9\) functional proteins are synthesized in total. However, this number increases rapidly if \(u\) increases. Doubling or quadrupling \(u\) would require the synthesis of 72.7 \(\times 10^9\) or 10,000 \(\times 10^9\) proteins, respectively. It is unrealistic to assume that a cell will synthesize 100 proteins. This is not a tremendous burden considering that about 46,600 \(\times 10^9\) functional proteins are synthesized in total. However, this number increases rapidly if \(u\) increases. Doubling or quadrupling \(u\) would require the synthesis of 72.7 \(\times 10^9\) or 10,000 \(\times 10^9\) proteins, respectively. It is unrealistic to assume that a cell will synthesize 100 proteins.
cost of abnormal proteins is largely irrelevant and the upper bound for $u$ primarily a result of the protein-length distribution and not of the cost of abnormal proteins. Furthermore, in such a situation there is little selection pressure to reduce the phenotypic mutation rate much below this upper bound [16].

If the synthesis of large proteins is such a problem, why does the cell not synthesize many smaller proteins and assemble them after successful production? An intermediate check for proper folding (which equals proper function for most amino acid substitutions) would prevent the incorporation of nonfunctional subunits and reduce the probability of assembling a nonfunctional complex. In yeast, proteins with a length of about 1,000 amino acids are quite common. This suggests that the complexation of proteins much smaller than 1,000 amino acids constitutes a considerable challenge. For so many large proteins, it might be impossible to get the same biological function from a complex of smaller proteins. According to our model, an upper bound of 1,000 for the yeast protein length does not reduce the drastic increase by removing all proteins from the dataset that are larger than 1,000 amino acids, we can still observe a rapid increase in $\bar{x}$ at $u = 5 \times 10^{-4}$; doubling (quadrupling) $u$ would lead to a 2.4 (7.3)-fold increase in the amount of abnormal proteins. Therefore, partitioning extremely large proteins into protein complexes is not sufficient to avoid the negative effects of an increasing phenotypic mutation rate.

Instead of complexing large proteins, evolution could reduce the phenotypic mutation rate of individual proteins. The phenotypic mutation rate of individual proteins could be reduced by using preferred codons [33–35] or by increasing the translational robustness of proteins [15,29,40]. Our analysis shows that these two mechanisms have nearly the same potential to minimize $\bar{x}$ if $u_i$ is sufficiently small (i.e., if preferred codons are sufficiently more accurate than non-preferred codons). One big difference between preferred codons and translational robustness is the way in which the trait is mutated. For preferred codon usage, it seems reasonable to assume that the number of preferred codons changes in increments of one, which leads to very small changes in the amount of abnormal proteins.

Considering translational robustness, little is known about how mutations change the translational robustness of a protein. In our simulations, we mutate the translational robustness of a protein by sampling it from a prior distribution, which allows for large changes. Alternatively, one can use models that allow only small changes in a protein’s translational robustness. More empirical data on the translational robustness spectrum of proteins is necessary to develop a satisfying model.

The effect of incremental changes of the number of preferred codons on the amount of abnormal proteins is fairly small. An increase in the number of preferred codons by one increases the probability of synthesizing a functional protein only by a factor of $(1 - ur) / (1 - u_i)$. For $u = 5 \times 10^{-4}$ and $u_i = 0.1$, this equals 1.00045. Since only few genes contribute much to the amount and number of abnormal proteins, this will lead to very small changes of $\bar{x}$ for most proteins.

As mentioned previously, preferred codons are also able to increase the rate of translation. Selection for faster translation (or higher expression level) could be responsible for the observed codon biases. Since the time it takes to synthesize $\gamma_i$ functional proteins is proportional to $\gamma_i n_i$ and the amount of erroneous proteins is approximately proportional to $\gamma_i n_i^2$, it is possible to distinguish between the two sources of codon bias by comparing the observed codon bias in yeast with the predicted codon bias if selective forces were proportional to $\gamma_i n_i$ or $\gamma_i n_i^2$.

Further, a refined version of our preferred codon model that considers the genetic code and the actual amino acid sequence of each yeast protein could be used to estimate the cost of abnormal proteins and the amino acid substitution rate. For a given amino acid substitution rate, $u_i$, an increase of the cost of abnormal proteins, $c$, increases the extent of codon bias but does not affect its distribution with respect to the protein length (the points in the top panel of Figure 7 would all move upward by an amount that is independent of $n_i$, since $n_i \bar{x}$ remains unchanged for constant $u_i$). For given $c$, an increase of $u_i$ changes the extent of codon bias as well as the codon bias distribution with respect to the protein length (as seen in Figure 7, if $u_i$ increases, the codon bias of large proteins changes to a greater extent than the codon bias of small proteins since $\bar{x}$ will increase more for genes with large $n_i$). Hence, by choosing different values for $c$ and $u_i$ and by comparing the resulting extent and distribution (with respect to $n_i$) of codon biases with the extent and distribution of codon bias found in yeast, one can estimate the two parameters.

To experimentally measure the rate of amino acid substitutions during protein synthesis is notoriously difficult. Abnormal proteins are difficult to detect and usually degraded within minutes [27]. Experiments are usually limited to measuring the rate of specific substitutions at specific sites (see Table 1). One exception is work by Ellis and Gallant [4], who measured the rate of substitution of charged amino acids by uncharged amino acids. For many proteins such substitutions are detectable as satellite spots after 2-D gel electrophoresis. However, their method might fail to detect rapidly degraded abnormal proteins and is dependent on the number of codons at which charge substitutions can occur [4].

It would be highly desirable to be able to calculate the actual frequency of phenotypic mutations, that is, the frequency of deleterious amino acid substitutions during protein synthesis as opposed to the frequency of all (detrimental or not) amino acid substitutions. We can use our model together with data on the fraction of proteins that are abnormal and degraded rapidly [41,42] to calculate this. Schubert et al. [41] and Princiotta et al. [42] measured that in human cells about 33% and 25%, respectively, of newly synthesized proteins are rapidly degraded. The proteins are degraded mainly because of their inability to achieve a functional state [27]. Since these are values for human cells and might also include proteins that could not achieve a functional state despite error-free protein synthesis, we will use 15%–35% as the range for the fraction of proteins that are nonfunctional due to phenotypic mutations. In our model, $y$ and $x$ give the amount of functional and nonfunctional proteins synthesized, respectively. Hence the fraction of nonfunctional proteins synthesized due to phenotypic errors is given by $x / (y + x)$. According to our model (Equation 7) and the data from yeast (see Materials and Methods), $2.4 \times$
10⁻⁴ to 6.1 × 10⁻⁴ deleterious amino acid substitutions per codon would result in the synthesis of 15% to 35% nonfunctional proteins. Better estimates of the fraction of abnormal proteins in yeast would allow a narrowing of the calculated range.

Mutation Rates at Equilibrium

Here, we derive our main analytical results on the magnitude of the genotypic and phenotypic mutation rates stated in the section The Model. We start by recalling Haldane’s principle for an asexually reproducing population. This population is assumed to be sufficiently large so that random genetic drift can be ignored. The only evolutionary forces considered are selection and mutation. We assume that there is an optimal type (wild type) in this population. Its fitness is denoted by \( W \) and parameters, we conducted additional simulations analogous to the simulations presented in Figure 1. For Figure 1, we used \( p_u = p_u = 10^{-4} \) and \( 10^{-7} \) as initial values of \( u \) and \( \mu \); genotypic mutations were lethal. The blue and violet lines in Figure 8 show that the initial values for \( u \) and \( \mu \) and the fitness of the genotypic mutant do not change the equilibrium mutation rates at mutation–selection balance. The genotypic and phenotypic mutation rates will converge to the same equilibrium mutation rates as long as (a) the initial value for \( u \) is low enough so that \( f = 1 - \eta(\bar{X}_0) \geq 0 \), and (b) the initial value for \( \mu \) is low enough (or genotypic mutations deleterious enough) so that a fixation of genotypic mutants does not occur.

We conducted simulations with different values for \( p_u \) and \( \mu \). The green and cyan lines in Figure 8 show the evolution of \( u \) and \( \mu \) for \( p_u = p_u = 10^{-3} \) and \( p_u = p_u = 10^{-5} \), respectively. As expected, higher (lower) \( p_u \) and \( \mu \) lead to faster (slower) evolution of \( u \) and \( \mu \) and to increased (decreased) equilibrium values. The magnitude of this change is smaller than predicted by theory, e.g., Equation 9. This can be attributed to the finite population size, \( N = 10^7 \). In finite populations, selection is inefficient for costs \( (C_u, C_\mu) \) below a certain threshold. Note that from Equations 23 and 2 we have \( f(\mu, u) \approx f(\mu, 0)(1 - p_u) \) and \( f(\mu, u) \approx f(\mu, 0)(1 - C_u) \), respectively.

Acknowledgments

We thank two anonymous reviewers for their detailed and constructive comments.

Author contributions. MW, RB, and MAN conceived and designed the experiments and wrote the paper. MW performed the experiments. MW and RB analyzed the data.

Funding. MW was supported by a Merck-Wiley fellowship. Support from the NSF/NIH joint program in mathematical biology (NIH grant...
References