Title
Localizing and Quantifying Evolutionary Selective Pressures

Author
Eames, Matthew Ashford

Publication Date
2010

Peer reviewed|Thesis/dissertation

Localizing and Quantifying Evolutionary Selective Pressures

by

Matt Eames

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Acknowledgements

I acknowledge this thesis is pretty much my two papers and my orals proposal stapled together. Thanks.

The text of this thesis in Chapter 1 is a reprint of the material as it appears in Structure. The coauthor listed in this publication directed and supervised the research that forms the basis for the dissertation.

Chapter 2 describes work that will be submitted for publication.

People I must thank: Cristina Melero, Gregg Kapp, Jeff Tabor, Ryan Ritterson, Rich Oberdorf, Dan Mandell and the rest of the Kortemme lab.

Thanks to my thesis committee Andrej Sali and Carol Gross

Big thanks to my boss Tanja Kortemme.

I'd also like to give a shout out to my super special lady-friend Bethany.

And thanks to you, the reader, alone in the darkness, clinging desperately to this thesis in wee hours of the night - I'm in your corner.

Abstract

Localizing and Quantifying Evolutionary Selective Pressures

Matt Eames

Advances in sequencing methodologies have produced a data glut; our ability to generate sequences exceeds our ability to understand the evolutionary selective pressures that shape them. The work presented in thesis consists of a two-pronged approach to study these pressures. The first approach employs an informatics-based analysis of high-throughput experimental data to detect selective pressures within protein substructures. The second uses the *lac* operon to quantify the relative impact of three selective pressures by measuring fitness effects of mutations under different induction conditions. Our results confirm our ability to detect selective pressures at the protein substructure level, as well as measure different fitness effects for different types of mutations. Furthermore, we uncovered surprising revelations about the *lac* operon itself, one of the most well-studied experimental systems in molecular biology. Finally, I present the methods used in this research as a general and robust approach for testing evolutionary models.

Table of contents

List of figures

List of tables

Introduction

While the generation of new evolutionary models has nearly kept pace with the explosive growth of biological data, validation of these models has lagged behind. To wit, our understanding of why proteins evolve at different rates, a concept that lies at the heart of evolutionary theory, remains incomplete. It seems intuitive that the degree of protein sequence conservation should depend on the protein's relative contribution to organism fitness, nonetheless, first-order approximations of functional contributions or "fitness density", including gene dispensability and number of interaction partners, have been problematic to correlate to evolutionary rates.

Surprisingly, the only parameter that has been shown to correlate with evolutionary rate is the mRNA expression level of the encoding gene. This correlation has been observed across species and has been determined to account for over 50% of the rate variance observed. It has been argued that this correlation is not based upon the innate functional importance of high-copy proteins, but instead stems from pressures to evolve sequences that are capable of folding properly despite translational errors (the "translational robustness" hypothesis). However, this concept has not yet been experimentally validated.

To resolve the mechanisms behind rate variation, one must overcome two challenges. First, to determine the scope of conservation, one must identify residues that are affected by selective pressures (i.e. whether pressures act locally or globally). For example, we might expect residues in protein cores to be most constrained, given their relative importance in protein stability. Second, one must quantify the degree to which different

selective pressures act upon organismal fitness. The goal of this thesis is to address these two challenges. A more complex understanding of pressures that drive the evolution of a protein's structural/functional characteristics will shed light on new design strategies. Particular biotechnical applications include the evaluation of system load generated by designed mutations and reengineered pathways. Knowledge of selective pressures may ultimately help to predict evolutionary trajectories, including the evolutionary divergence following gene duplication events.

The first analysis presented in this thesis attempted to detect selective pressures within structural subunits by maximizing structural information for all proteins within a single organism. This process mapped structurally characterized domains to *Saccharomyces cerevisiae*, whereupon proposed constraints could be detected within structural elements. Accordingly, we sought to resolve pressures believed to function on different scales:

i. localized structural/functional constraints (those associated with a particular structural element) imparted by the preservation of protein-protein interfaces and protein cores

ii. global constraints (those not yet associated with a structural element) imposed by expression level, abundance, and dispensability

As mentioned previously, factors including gene dispensability and promiscuity (the number of protein interaction partners), unlike mRNA expression level, do not correlate with protein evolutionary rate. A likely rationale for the failure to detect these correlations is an overestimation of the magnitude and scope with which these parameters restrict evolutionary rate. Selective pressures may not act globally, and examination of

potentially local pressures requires focusing our examination upon the residues constrained.

Protein-protein interfaces provided an ideal subset of protein structure to probe for localized selective pressures. It has been suggested that protein interfaces, which encode functionally important signaling, regulatory and structural information, are likely to contain a higher fraction of residues incapable of accepting nearly-neutral mutations relative to the remainder of the protein surface. Thus, interface residues would be more likely to be conserved than residues outside of the interface. Furthermore, residues involved in protein-protein interfaces can be easily identified from the PDB (as opposed to active site residues).

Through large-scale mapping of structurally characterized protein complexes onto *S. cerevisiae* sequences we were able to achieve residue-level discrimination of selective pressures within a single-species environment. We observed significant evolutionary differences between structurally distinct features, including buried residues, surface residues, and interface residues. Within protein-protein interfaces we were also able to detect conservation differences dependent upon degree-of-burial as well as a tendency for increased conservation of polar residues. We measured significant differences in the relative evolutionary rates of structural subsets when contrasted with such global parameters as mRNA expression level and protein abundance. Specifically, we found pressures associated with high mRNA expression level to constrain the sequences uniformly, while those associated with high protein abundance to affect specific structural subsets.

The second analysis presented here attempted to quantify the degree to which different hypothesized selective pressures act upon organismal fitness. This study was carried out using the well-characterized *lac* operon in *Escherichia coli*. The three pressures at the focus of the investigation were proposed to explain the inverse relationship between expression level and evolutionary rate. These pressures are as follows:

i. *functional loss* - suggests that mutations within a protein affecting functional residues have fitness affects proportional to the number of copies of that protein in the cell

ii. *translational efficiency* - proposes that since certain codons are translated more quickly or accurately, there is selective pressure to avoid mutations to inefficient codons, in particular for highly expressed proteins

iii. *translational robustness* - postulates that given the expression level-dependent cost of misfolding proteins, highly expressed proteins will be under selective pressure to favor sequences that fold properly despite translational errors

By determining growth rates as a measure of organism fitness under varying induction and nutrient conditions, we attempted to measure the cost and benefit of protein production. The three aforementioned pressures each make distinct predictions on the affected parameters describing the cost and benefit function, thus allowing quantification of their associated fitness affects. We quantified the pressures by making the following mutations:

i. *Functional loss*: Mutations affecting functional efficiencies (catalytic parameters).

At constant expression level, we expected a difference in benefit, but not in cost of protein production.

ii. *Translational Efficiency*: LacZ protein variants with identical protein sequences, but including synonymous mutations with different levels of translational optimization. At constant expression level we expected a difference in cost but not in benefit of protein production.

iii. *Translational Robustness*: Mutations affecting protein stability. At constant expression level, we expected both cost and benefit to be affected.

Our results revealed a surprising cost-dependence on LacY functionality that overshadowed all other cost considerations (translation/transcription costs, destabilization costs, etc). A meta-analysis of the literature suggested this to be a widespread phenomenon of permeases and membrane proteins, and not specific to the *lac* operon. Furthermore, we uncovered a possible explanation for the use of allolactose as the true inducer of the operon – as a systems check to avoid expression of LacY in a system with faulty LacZ activity (thus unable to offset high costs with the metabolic benefit of LacZ).

Chapter 1. Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance

Abstract

Genome-wide studies in *Saccharomyces cerevisiae* concluded that the dominant determinant of protein evolutionary rates is expression level, where highly-expressed proteins evolve most slowly. To determine how this constraint affects the evolution of protein interactions, we directly measure evolutionary rates of protein interface, surface and core residues by structurally mapping domain interactions to yeast genomes. We find that mRNA level and protein abundance, though correlated, report on pressures affecting regions of proteins differently. Pressures proportional to mRNA level slow evolutionary rates of all structural regions and reduce the variability in rate differences between interfaces and other surfaces. In contrast, the evolutionary rate variation within a domain is much less correlated to protein abundance. Distinct pressures may be associated primarily with the cost (mRNA level) and functional benefit (protein abundance) of protein production. Interfaces of proteins with low mRNA levels may have higher evolutionary flexibility, and could constitute the raw material for new functions.

Introduction

What are the evolutionary pressures acting on proteins building up biological protein-protein interaction networks? An intuitive expectation that has been tested by many studies [1-5] is that amino acids in protein interaction interfaces should be more conserved than protein surfaces not involved in interactions. An underlying assumption here is that interfaces, which encode functionally important signaling, regulatory and structural information, are likely to contain a lower fraction of residues capable of accepting nearly-neutral mutations [6-8] relative to the remainder of the protein surface. Hence, protein interfaces would have a higher "functional density" and evolve at slower evolutionary rates (defined as the number of nonsynonymous substitutions per site).

Several recent studies have analyzed the degree and mechanisms by which protein-protein interactions constrain the rate of protein sequence evolution [4, 9-13]. An attractive hypothesis is that proteins with many interaction partners are more likely to have a high functional density of important regions, and should hence be more conserved. While the suggested inverse dependency between a protein's evolutionary rate and the number of its interaction partners has indeed been observed, [9, 14] this effect may be complicated to separate from the strong correlation of gene expression (measured in mRNA transcripts per cell) and evolutionary rates: both evolutionary rates and number of interaction partners are related to mRNA expression level, where highly expressed proteins evolve most slowly and also have a larger number of interaction partners as detected in high-throughput studies [10, 15]. In addition, the constraints acting on protein interfaces may be difficult to detect by measuring the evolutionary rate of whole proteins because of the relatively small number of residues that constitute interface regions. To

circumvent this problem, information on the three-dimensional structures of proteins and protein complexes is needed to separately analyze conservation in protein interfaces and the remainder of the protein [4, 13]. Two recent studies used structural data to support the functional density hypothesis by showing that proteins with larger number of distinct binding interfaces [13] and smaller proportion of solvent-exposed residues [16] are indeed more conserved.

Although structural/functional constraints thus clearly slow the rate of evolution of proteins and protein interfaces [13, 16, 17], it has not been quantified how protein interfaces are affected by known determinants of protein evolutionary rates such as mRNA level and protein abundance. A recent key study proposed that gene expression level-dependent evolutionary pressures are predominantly due to constraints imposed by robustness against translational errors (translational robustness): Highly expressed proteins would be substantially optimized to still fold despite transcriptional errors to avoid potential toxic accumulation of misfolded species [18]. Accordingly, one could expect expression-level dependent pressures to be strongest for amino acids located in protein cores, where mutations should have the most dramatic effect on stability. Following this argument, it is unclear whether the same dominant pressure would persist for amino acid positions in protein interfaces that are surface-exposed in the unbound form of the protein. Due to the relatively smaller contribution of interface residues to protein stability, protein interface residues may be freed from the evolutionary constraints acting on highly expressed proteins, and show evolutionary rates to be largely determined by functional pressures. Alternatively, mRNA expression level could exert a global pressure on all residues in a protein. If such a global effect primarily reflects the

frequency of translational events but not the actual amount of protein present, differences between pressures proportional to mRNA and protein levels may be observable.

To address these questions, here we aim to characterize the combined influence of structural, functional, mRNA expression- and protein-level dependent constraints on protein interface evolution. We derive structural information from all characterized protein complexes in the Protein Data Bank (PDB) [19], allowing us to classify individual residues as members of the protein core, the surface, or a protein-protein interface. Protein-protein interface residues are further subdivided by structural features proposed to be linked to interaction affinity and specificity, such as burial in interface cores, classification as "anchor" residues that become highly buried upon complex formation [20], and polar character. Using domain definitions derived from the Pfam database of protein families [21] we map these structural characteristics to residues of all matching domains in a single genome, that of *Saccharomyces cerevisiae*. The availability of high-throughput information specific to *S. cerevisiae* and sequence information for related fungal genomes enables us to quantify evolutionary constraints on protein interfaces correlated to mRNA expression-level and protein abundance using a nonsynonymous substitution rate metric.

We find, first, that pressures proportional to mRNA expression level appear to slow the evolutionary rates of residues in *all* structural regions, including interface and other surface regions. Moreover, this global pressure significantly reduces the variability in conservation between protein cores, interfaces and surfaces in proteins with high mRNA levels. Second, while previous work has pointed to the common importance of abundance and expression level (both have been identified as the dominant determinants of

evolutionary rates), we detect a distinction between pressures correlated with mRNA-expression and those correlated with protein-abundance levels: The evolutionary rate variation between structural regions within a domain is much less correlated to protein abundance, in contrast to the observed decrease in rate variation at high mRNA level. These findings support a model where each parameter reports on distinct selective pressures that may be associated primarily with the cost (mRNA expression) and functional benefit (protein abundance) of protein production. Interfaces of proteins with low mRNA expression level (lower cost) have higher evolutionary flexibility, and may constitute the raw material for evolving new functions.

Results

We aimed to characterize selective pressures imposed upon protein-protein interfaces. We set out to directly measure evolutionary rates (quantified by the number of non-synonymous mutations per site) of residues at protein interfaces and investigated, first, structural characteristics, second, pressures proportional to mRNA expression level, third, the role of protein abundance, and fourth, functional constraints.

An outline of the computational strategy is given in Figure 1-1A. To define and characterize protein interfaces, we first required a dataset of known protein complex structures. By mapping these structures to a single, well-studied organism, *S. cerevisiae*, we aimed to integrate structural information with other high-throughput data such as mRNA expression and protein abundance. To increase structural coverage (the part of the *S. cerevisiae* genome we could assign to structurally characterized protein-protein

interfaces), we used Pfam [21] domains instead of whole proteins as the unit of structural mapping, to take advantage of the fact that a single modular domain could have occurrences in dozens of *S. cerevisiae* proteins. To identify amino acid residues present in protein-protein interfaces (Figure 1-1B), we compiled a dataset of domain-domain residue-specific contacts from the database of 3D Interacting Domains (3did) [22].

Evolutionary Rates of Domain Structural Subsets

To assess the role of structural characteristics in interface conservation, we divided domains into four structural subsets: surface residues, core residues, surface interface residues (interface residues partly exposed in the complex), and buried interface residues (Figure 1-1B and Methods). Buried interface residues were defined as residues in the interface that became buried upon complex formation but were not part of the protein core in the monomeric form of the domain. For all structural subsets, we computed evolutionary rates as the number of non-synonymous substitutions per site (dN, see Methods). Rates in this paper were calculated by aligning *S. cerevisiae* proteins to their orthologs in either *Candida albicans*, *Candida glabrata*, *Saccharomyces bayanus*, or *Saccharomyces mikatae*. Unless otherwise noted, rate data presented were taken from alignments with *C. albicans*, but similar results were obtained using comparisons to the other fungal genomes (see Table 1-S1 in Supplementary Materials).

We first confirmed that residues classified as part of the core by our mapping procedure have a slower median non-synonymous substitution rate than surface residues (0.223 *vs* 0.361 (P = 9.5e-15, Wilcoxon-signed rank test), Table 1-1 and Figure 1-2), in agreement with data from earlier studies [4, 23]. We also found buried interface and surface

interface residues to be significantly more conserved than surface residues outside of the interface (0.231 *vs* 0.361 (P = 2.6e-11), 0.273 *vs* 0.361 (P = 0.003), respectively). These results indicate that our structural mapping procedure is able to recapture expected conservation properties. It should however be noted that the surface residue dataset is likely to contain a mixture of both surface and interface residues, given the high probability that many interfaces have not yet been structurally characterized [4], see Methods Section in Supplementary Materials. Another source of error in our assignment of structural subsets is that not all structurally characterized domain interface residues compiled from multiple organisms may actually be involved in interface formation in fungal species (see Discussion).

Pressures Correlated to mRNA ExpressionLevel

It has been well documented that gene expression level measured by mRNA molecules per protein per cell is the single most important determinant of evolutionary rates in proteins, at least in unicellular organisms such as *S. cerevisiae* [14, 15, 18]. Three principle arguments have been set forth to explain the negative correlation between divergence and expression level, and the "translational robustness" hypothesis was shown to agree best with the data [18]. This model postulates that given the expression level-dependent cost of misfolding proteins, highly expressed proteins will be under selective pressure to favor amino acid sequences that fold properly despite translational errors. On account of the importance of the protein core for folding, we reasoned that gene expression level would evolutionarily constrain core residues to greater degree than surface residues. To detect this effect, we measured the difference in nonsynonymous

substitution rate between core residues and surface residues for each protein domain as a function of mRNA expression level (Figure 1-3A).

Moreover, we wanted to address the question of whether a similar selective pressure would also affect protein residues buried only upon interface formation. We imagined two different scenarios: Since protein interfaces (of transient interactions) are exposed to solvent in the uncomplexed form, interface mutations are less likely than core mutations to substantially destabilize the monomer. It follows that interface mutations may not significantly alter a protein's misfolding probability after translational errors. In this case and under the translational robustness hypothesis, we would expect the evolutionary rates of protein interface residues to be largely uncorrelated to expression level. Alternatively, mRNA expression level could exert a strong global pressure on all protein residues such that the evolutionary rates of surface, core, and buried interface residues are similarly correlated to mRNA expression level. To test for these possibilities, we computed the difference in evolutionary rates of buried interface residues and surface residues for each domain considered as a function of mRNA expression level (Figure 1-3B). We used interface residues buried upon complex formation, as opposed to all interface residues; we expected selective pressure for this set to be most easily detectable (see Figure 1-3B) as we reasoned that buried interface portions are likely to be less susceptible to alignment errors made in our structural mapping procedure.

We found that the magnitude of evolutionary rate differences for core *versus* surface ΔdN(core-surface) or buried interface *versus* surface ΔdN(interface-surface) residues was highly dependent on mRNA expression level (Figure 1-3). For proteins with medium and low expression levels, protein cores have a notable tendency to be more conserved than

the surfaces of their corresponding domains, as expected. However, we observe that the difference in evolutionary rates between surface and core residues becomes progressively small with increasing mRNA level. A similar diminished difference in evolutionary rates at high mRNA level is found for buried interface and surface residues. In contrast, at low mRNA levels buried interface residues have the freedom to evolve faster or slower than surfaces in proteins with low mRNA levels (Figure 1-3B). The significance of these fast evolving interfaces is not clear, and they could be artifacts resulting from assumptions made in our structural mapping procedure. Examination of proteins with a higher relative evolutionary rate of interface to surface residues did not reveal any biases in protein function or number of interaction partners (data not shown).

Next we asked whether the observed reduction in evolutionary rate differences between structural regions at high mRNA expression levels was primarily due to a progressively higher pressure on buried residues or whether all structural regions were affected to some extent by a pressure proportional to mRNA level. Figure 1-4A indeed shows a global reduction in evolutionary rate of all structural regions with increasing mRNA level, which is surprising given the expected stronger pressure on core residues in accord with the translational robustness hypothesis.

Because relatively few proteins shown in Figure 1-3A have very high mRNA expression levels, we investigated whether the observed differences were the result of a small sampling set for higher expression regimes. We assessed the statistical significance of evolutionary rate distributions in both the upper and lower expression extremes (17.5 percentiles). We found that the reduction in the variability of ΔdN(interface-surface)

values in highly expressed proteins relative to a lower expression set of equal size was significant ($P = 0.0004$, Figure 1-5A). This trend is independent of how the percentiles of the upper and lower expression extremes are chosen, and can even be observed by dividing the expression distribution into two halves (see Table 1-S2 in Supplementary Materials). This constraint is also observed to restrict the variability in ΔdN(core-surface) values in proteins with high mRNA levels (Figure 1-S1A in Supplementary Materials). The observed narrowing of evolutionary rate variability within a domain is not merely a consequence of an overall lower rate of proteins with high mRNA levels, as normalizing ΔdN(interface-surface) by dN of the domain also showed a significant difference between proteins with high and low mRNA levels ($P = 0.001$).

Ribosomal proteins are known to be both highly conserved and highly expressed. To test for bias caused by these special characteristics of ribosomal proteins, we repeated the analysis excluding ribosomal domains (Figure 1-S2 in Supplementary Materials). The difference in the distributions of evolutionary rate differences between buried interface and surface residues for highly and lowly expressed proteins was still significant (Kolmogorov-Smirnov test, $P = 0.0023$). We obtained similar statistical significance values when computing evolutionary rates relative to other fungal genomes (See Table 1-S1 in Supplementary Materials).

Pressures correlated to Protein Abundance

mRNA expression level and protein abundance are correlated, but to varying extents [24]. We hence asked whether the observed dependence of the ΔdN(interface-surface)

distributions on mRNA expression level is also observed when using protein abundance data. We found that protein abundance does not appear to be correlated to a restriction in ΔdN(interface-surface) in the same fashion that mRNA expression level is (Figure 1-5B, see Figure 1-S1B in Supplementary Materials for ΔdN(core-surface)). In contrast to what we observed with mRNA expression level, the distributions in ΔdN(interface-surface) values are not distinguishable between proteins with high and low protein abundance levels. We observe a similar differential effect of abundance and expression level when computing substitution rates of *S. cerevisiae* relative to other sequenced fungal genomes, such as *C. glabrata*, *S. bayanus* and *S. mikatae* (see Table 1-S1 in Supplementary Materials), and when normalizing ΔdN(interface-surface) by dN(domain), which did not result in a significant difference between proteins with high and low abundance (P = 0.17), but did between proteins with high and low mRNA expression (P = 0.001).

To further confirm the difference between pressures correlated to mRNA-level and protein abundance, we tested how pressures proportional to protein abundance affect evolutionary rates of core, buried interface and surface regions separately, as above. As expected, evolutionary rates of all structural regions are less affected by abundance than by mRNA level (Figure 1-4B). Residues become only slightly more conserved with increasing protein abundance level. Only the highest and lowest protein abundance bins have significantly different evolutionary rates for each of the 4 structural subsets (see Table 1-S3 in Supplementary Materials).

If pressures proportional to mRNA levels in fact limit the evolutionary rates of all protein

residues similarly regardless of the residue location on the surface or at an interface (i.e. independent of a functional constraint imposed by protein interactions), a comparison of randomly picked residue patches of similar size as the interface should show the same distribution as a function of mRNA level as depicted in Figure 1-5A. Figure 1-S3A in the Supplementary Materials confirms this suggestion. Analogously, repeating the residue patch comparison as a function of abundance (Figure 1-S3B in Supplementary Materials) should not reveal a dependence on protein levels, and yields similar results as depicted in Figure 1-5B (Figure 1-S3B in Supplementary Materials).

Combining the results of Figures 1-4 and 1-5, we conclude that first, the evolutionary rates of all structural regions are affected proportional to mRNA level, but to significantly lesser extent by protein levels; the effect proportional to protein levels could result largely from the underlying dependence of mRNA and protein levels. Second, the variability in evolutionary rates between different structural regions within a given domain should be correlated with mRNA level but be much less affected by protein abundance. Figure 6 shows that this is indeed the case, where the mean differences of evolutionary rates between surfaces and cores and their standard deviations decrease with mRNA level but stay mainly constant as a function of protein abundance.

Conservation of Polar Residues

We next sought to resolve functional pressures that constrain the evolution of protein interfaces. It has been suggested that polar residues are highly conserved in interfaces because they constitute energetic “hot spots” or contribute to interaction specificity [5]. To confirm this finding in our analysis, we subclassified our structural subsets into polar

residues (Asp, Glu, Lys, Arg, Ser, Thr, Asn, and Gln; definitions taken from Hu et al) and remainder (all residues not classified as polar, including nonpolar residues and aromatics). We found that, with the exception of the buried interface residue set, polar residues in all structural subsets evolved faster than remainder residues, with a high degree of statistical significance (Table 1-2). However, buried interface residues were unique in their equivalent conservation of polar and remainder residues (evolutionary rates of 0.370 and 0.361, respectively). These results suggest the presence of pressures favoring the conservation of polar residues in protein interfaces.

Relationship Between Degree of Burial and Evolutionary Rates

Previous research has highlighted the importance of "anchor residues" in protein-protein interactions. Anchor residues were identified as side chains that pack into structurally constrained grooves of their binding partners [20]. Such residues are exposed in the monomeric form of the protein and bury most of their surface area only upon complex formation. To assess the conservation of these anchor residues in our set of domain-domain interactions, we measured the relationship between evolutionary rate and degree of burial upon interface formation (Figure 1-7). Degree of burial was defined as the difference in number of neighboring C_β atoms within a 10Å sphere of the residue of interest between the monomeric and complexed forms of the domain. Interface residues were binned according to their degree of burial, and the median evolutionary rate of each bin was computed. We found a correlation between increase in burial upon complex formation (additional neighbors) and sequence conservation (1-6 additional neighbors, median dN = 0.302; 7-12 additional neighbors, median dN = 0.295; 13-18 additional

neighbors, median dN = 0.195; 19-24 additional neighbors, median dN = 0.031. Similar results were obtained when using the difference in solvent-accessible surface area (SASA) to define degree of burial (Figure S4 in Supplementary Materials). As defined by a Wilcoxon-signed rank test, the dN distributions for all interface residue bins are significantly different from each other ($P < 3e-7$) with the exception of the first two bins 1-6, 7-12 ($P = 0.58$) and the last two 13-18, 19-24 ($P = 0.06$). Thus we observe a significant conservation signal not only for residues buried in interfaces, which has been shown to more generally lead to proteins with larger interface area to be more conserved [13], but also specifically for representative of anchor residues that show an increase of 13-24 neighbors upon complex formation (Figure 1-7). This result was not altered when using different bin widths (data not shown). To detect other possible biases in the binned datasets, which could arise from unusual structural/functional characteristics of domains containing residues with high degrees of burial (such as domains largely buried within complexes) or from expression level biases, we also measured the evolutionary rate of surface residues of the domains in each bin. The median values for surface evolutionary rates ranged from 0.35 to 0.37. Differences between the bins were found to be statistically insignificant ($P > 0.2$), suggesting that the observed conservation signal for anchor residues is not substantially biased by expression-level dependent correlations.

Discussion

We set out to combine genome-wide structural mapping with systems-level data on mRNA expression and protein abundance to characterize pressures on the evolution of proteins and protein-protein interfaces in yeast genomes. While expression level has been established as the dominant determinant of protein evolutionary rates, we aimed to dissect the influence of functional local evolutionary pressures acting on residues in protein interfaces proportional to mRNA expression-level and protein abundance. Our analysis reveals two main new results: First, we find that mRNA expression level shows a profound influence on the evolutionary rates of protein interfaces relative to protein surfaces: for proteins with high mRNA levels, the distribution of ΔdN(interface-surface) values is substantially narrowed (Figure 1-5A). This suggests that interface residues possess more evolutionary flexibility in proteins with low mRNA expression level, allowing such interfaces to evolve at different rates compared to residues on other surfaces of the same domains. Similar results are observed for the ΔdN(core-surface) distributions. Second and in contrast, this restriction in the variation of evolutionary rates between different regions in the same protein domain is largely absent when comparing proteins with high and low abundance (Figure 1-5B). This indicates that despite some correlation between abundance and expression level, their differential effects reflect separate selective pressures.

Prior work has suggested that the selective pressure to evolve sequences robust enough to fold properly despite translational errors is the primary constraint on highly expressed proteins [18]. Following from this "translational robustness" hypothesis, we initially

reasoned that protein cores (but not necessarily protein interfaces exposed in the uncomplexed form) may be affected most dramatically by protein expression level, as non-synonymous core mutations would be expected to be most destabilizing. However, the evolutionary rates of all structural regions (core, surface and interface residues) appear to be slowed by pressures proportional to mRNA expression level (Figure 1-4A), including translational robustness. Though expression level is likely to reflect multiple selective pressures, it is apparent that one such pressure affects all residues, irrespective of their structural location.

Proteins with low mRNA expression levels are to a lesser extent affected by these global pressures, and hence variability in selective pressures targeting structural subsets becomes detectable (Figure 1-5A). In addition, our data suggest that in proteins with low mRNA levels, core residues and residues buried in the interface may not be under the same pressures. Core residues rarely evolve faster than the surface, which contrasts with the apparent freedom of residues buried in interfaces to evolve at rates both faster and slower than the surface (Figure 1-3). It is difficult to gauge the significance of such quickly evolving interfaces. They may be a result of recently evolved functionality, or merely a byproduct of caveats intrinsic to our structural mapping procedure (see below).

As mRNA expression level, protein abundance is also known to correlate with evolutionary rate [15]. This finding has been attributed to a correlation between expression level and protein abundance. Hence, the observed differences in the evolution of protein interfaces as a function of mRNA expression level and protein abundance (compare Figures 1-5A and 1-5B) are somewhat unexpected. However, our finding supports evolutionary mechanisms in which restraints are linked to the frequency of

translation events, which have a stronger correlation with expression level than with protein abundance [18, 24].

Two broad classes of selective pressures may be at play: pressures associated with production, which could be thought of as operating globally across the nucleotide sequence, on average, and pressures associated with benefit, which may more selectively target local protein structural and functional characteristics. mRNA expression level and protein abundance, which are somewhat correlated, will reflect selective pressures within both classes. However, the disparate evolutionary trends observed when contrasting expression level and protein abundance suggest that the two parameters show differing degrees of correlation within each class.

mRNA expression level may in large part reflect cost-dependent pressures. In particular, mRNA expression has been related to both codon efficiency and translational robustness measuring potential costs due to protein misfolding. The fact that evolutionary rates of all structural regions are reduced proportionally to mRNA level (Figure 1-4A) implies a correlation with pressures operating more globally upon the sequence. In contrast, protein abundance may more directly report on benefit-dependent pressures related to protein structure and function. Structural/functional constraints can act locally; indeed localized constraints, including those observed in protein interfaces, are still distinguishable by substantial variability of evolutionary rates of different structural regions in highly abundant proteins, despite pressures that may affect whole sequences. This conclusion is further supported by comparing evolutionary rate differences of buried interface and surface residues between pairs of proteins in which one member has a higher abundance and a lower mRNA expression level than its partner: For 10,000 randomly generated

pairs meeting this definition, 5,814 of those with higher abundance had a greater evolutionary rate difference between buried interface and surface residues ($P < 2.2e-16$). A more detailed dissection of expression and abundance-dependent evolutionary forces will require quantitative comparisons of the cost and benefit of protein production in mutant populations.

The evolutionary trends we have uncovered are strong enough to persist through the approximations we have made, including completeness of interface identification from known complexes in the PDB (see Results & Methods), functionality of each interface inferred from the PDB in *S. cerevisiae*, and accuracy of alignment between the PDB and *S. cerevisiae* sequences. While the preservation of interface functionality between organisms and domain family members used in this analysis is difficult to estimate, this is undoubtedly a large source of error. We can, however, measure the evolutionary rates of structural subsets in *S. cerevisiae* proteins that have been crystallized, where the data allow correct structural classification. This test, though the small number of *S. cerevisiae* structures limits the statistical significance, reveals that functional interfaces exhibit similar conservation as observed in Figure 1-2 for inferred interfaces. Other possible explanations for the disparity in selective pressures observed for highly and lowly expressed proteins could stem from biases in the sequence-structure alignments. When we investigated this possibility, we indeed found that highly expressed proteins an average had a higher sequence identity in alignments made with PDB domains. However further analysis separating highly and lowly expressed proteins into bins of comparable alignment identity showed that similar results as summarized in Figures 1-3 and 1-5 are obtained regardless of the level of sequence identity (see Methods).

From the evidence presented in this paper, we speculate that proteins with low mRNA levels have sufficiently relaxed evolutionary constraints to serve as the raw material for new genes. If this is indeed the case, it would provide insight into how proteins evolve, from the mutational trajectory of proteins following gene duplication events to modified approaches for directed evolution experimentation. Modulation of not only gene expression patterns but also gene expression levels has been suggested to contribute substantially to evolutionary changes [25]. Lowering mRNA expression but not necessarily protein abundance levels may increase the mutational flexibility in naturally occurring systems and synthetic evolution experiments.

Methods

Determination of Interface Residues

Residues were identified as participating in domain-domain interfaces as designated through either side-chain-side-chain or side-chain-main-chain contacts. Contacts were identified using the criteria established in the 3did [22], with interacting residues defined as having one or more of hydrogen bonds (N-O distances $\leq$ 3.5Å), salt bridges (N-O distances $\leq$ 5.5Å), or van der Waals interactions (C-C distances $\leq$ 5Å). Using the domain definitions and domain boundaries listed in the 3did, protein domain sequences and their corresponding interacting residues were compiled as the "PDB domain dataset" (7,471 sequences, 580 different domains; Figure 1A).

Mapping Residues to Saccharomyces cerevisiae

Domain sequences, using Pfam domain definitions, of all *S. cerevisiae* proteins

containing domains observed in the PDB domain dataset were collected as the *S. cerevisiae* domain dataset (1,570 domains from 950 proteins). Alignments between sequences in the PDB domain dataset and the *S. cerevisiae* domain dataset were made using CLUSTALW [26]. *S. cerevisiae* residues that aligned with PDB interface residues were identified as contributing to the interface, unless the position was the site of an insertion/deletion. For every *S. cerevisiae* domain, the interface was defined as the sum of all mapped interface residues from every domain-domain alignment.

Previous research addressing the relationship between sequence and interaction divergence has defined the 20-30% identity regime as the "twilight zone" at which structural similarities between sequences begin to break down, but has found a reduced identity cutoff for reliable interaction modeling using Pfam domain classification employed here [27]. To test the effects of alignment error on our results, we recomputed our data using a minimum 20% identity in PDB-*S. cerevisiae* alignments for mapping. The difference in evolutionary rates of interfaces of highly and lowly expressed proteins was retained in the recomputed results, as expected given that the vast majority of our domains align above this threshold. It has also been put forth that an 80% sequence identity should be used to guarantee true interactions [28]; our results support the conservation of interface residues, but do not address questions of interaction specificity that is most likely dependent on higher-resolution features of protein interfaces.

Determination of Buried and Surface Residues

Surface and buried residues were identified in PDB structures by tallying the number of C_β atoms of neighboring residues within a 10Å sphere around the C_β atom of the residue

of interest (≤16 C_β neighbors for surface residues, ≥25 C_β neighbors for buried residues). Buried residues were determined in both monomeric form (analyzing neighbors on the same chain, "core residues") and complexed form (analyzing neighbors in the entire structure, "residues buried in complex"). Surface residues were only determined in the monomeric form. "Buried interface residues" were identified as the intersection of "interface residues" and "residues buried in complex", but removing "core residues"; "surface interface residues" were defined as the intersection of "interface residues" and "surface residues"; the "surface residues" set excluded "buried interface residues". (For illustration of the structural sets, see Figure 1-1B).

mRNA Expression Levels and Protein Abundance

Gene expression data for *S. cerevisiae* measured in mRNA molecules per cell were taken from [29]. Protein abundance data were taken from [30]. Results using protein abundance data reported in the text excluded proteins defined as having "no detected expression" in [30]; however, inclusion of these proteins did not significantly affect the results (data not shown).

Evaluation of Evolutionary Rates

Evolutionary rates were obtained by comparing the *S. cerevisiae* domain-containing proteins sequences to their *C. albicans*, *C. glabrata*, *S. bayanus*, and *S. mikaae* orthologs. Putatively orthologous sequences were identified by means of reciprocal best hits, with the minimum length of the alignable region >80% of the longer protein, and a BlastP [31] cutoff of $E > 10^{-10}$. From this set, protein sequence alignments were made with

CLUSTALW and then mapped to their respective nucleotide sequences. Interface-containing domains, as defined in the *S. cerevisiae* domain set, were then used to compute non-synonymous substitution rates (dN) for this alignment subset using the PAML software program CODEML [32]. 396 orthologous (*cerevisiae-albicans*) pairs were used in this analysis. The average domain contained 9.1 "buried interface residues", 28.6 "surface interface residues", 67.1 "surface residues", and 72.6 "core residues". Analyses performed in this paper were replicated using curated orthologs from the Candida Genome Database (CGD) (http://www.candidagenome.org/), producing similar results (data not shown).

Sequences

Protein and nucleotide sequences were obtained from the *Saccharomyces* Genome Database (SGD, http://genome-www.stanford.edu/Saccharomyces/) for *S. cerevisiae*, from the CGD (http://www.candidagenome.org/) for *C. albicans*, downloaded from ftp://genome-ftp.stanford/pub/yeast/sequence/fungal_genomes/S_bayanus/ for *S. bayanus*, ftp://genome-ftp.stanford/pub/yeast/sequence/fungal_genomes/S_mikatae/ for *S. mikatae* and http://cbi.labri.u-bordeaux.fr/Genolevures/ for *C. glabrata*.

Statistical Analysis

The R package was used for statistical analysis [33].

Figures for chapter 1

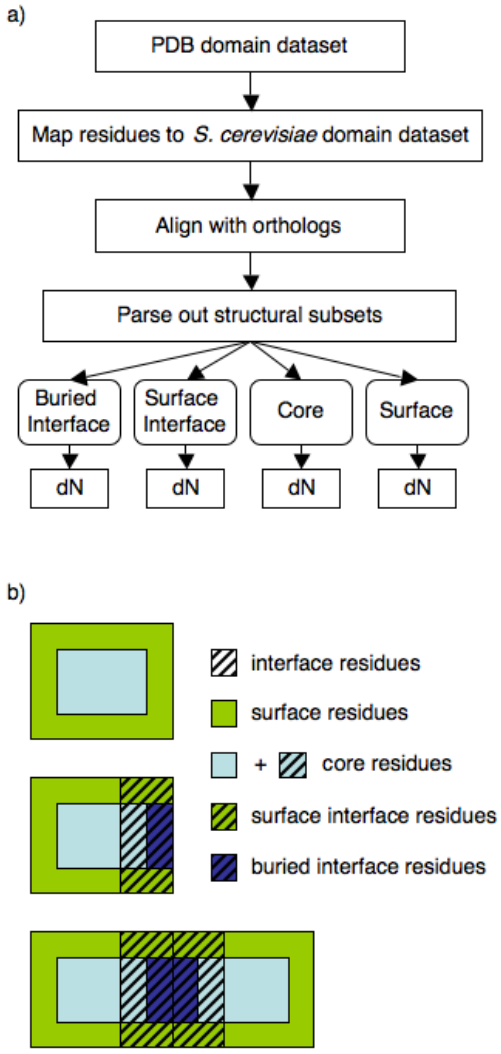


Figure 1-1. Illustration of the computational strategy. (A) Flow cart outlining the steps for genome-wide structural mapping of *S. cerevisiae* domains and determination of evolutionary rates for structural subsets. (B) Cartoon depicting the 4 structural subsets used in this analysis.

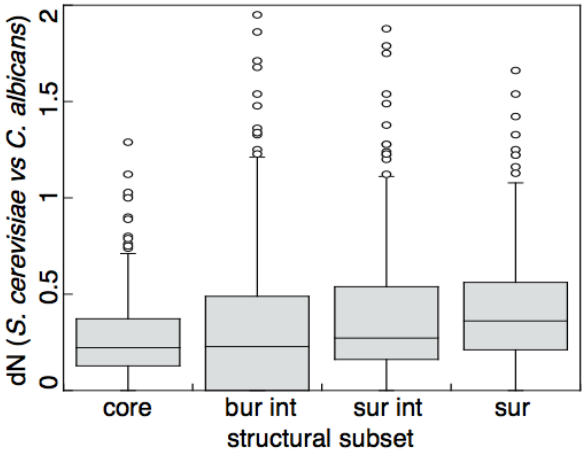


Figure 1-2. Distributions of non-synonymous substitution rates, dN, for residue subsets with different structural characteristics. Boxes enclose the first and third quartile of the distribution and display a notch at the median; whiskers extend outward to the most extreme data point no more than three times the interquartile range from the box. Data points outside this range are drawn individually. As defined by a Wilcoxon-signed rank test, the dN distributions for all structural characteristics are significantly different from each other ($P < 0.005$) with exception of "buried residues" and "buried interface residues" ($P = 0.18$, not significant). 4 outliers fall outside the upper boundary.

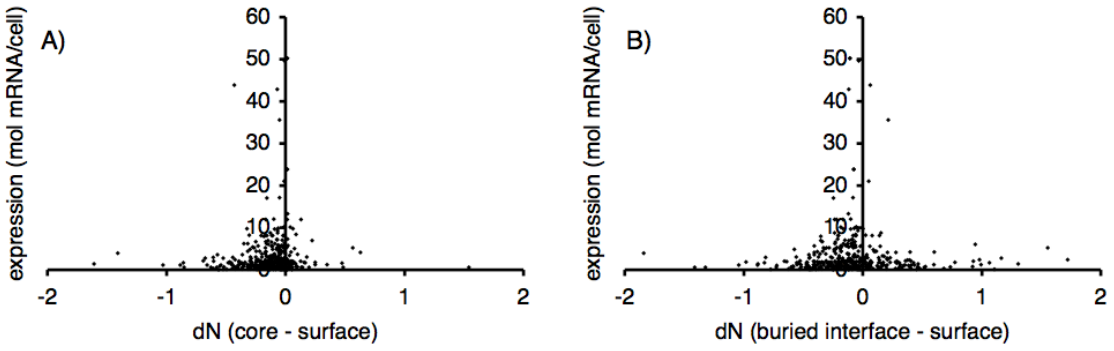


Figure 1-3. The evolutionary rate differences between structural regions within the same domain are reduced at high mRNA expression levels. Shown (A) is the difference in dN between the core residues and surface residues subsets for each domain as a function of

mRNA expression level or (B) the difference between the buried interface residues and surface residues sets. One outlier falls outside the limits in (B).

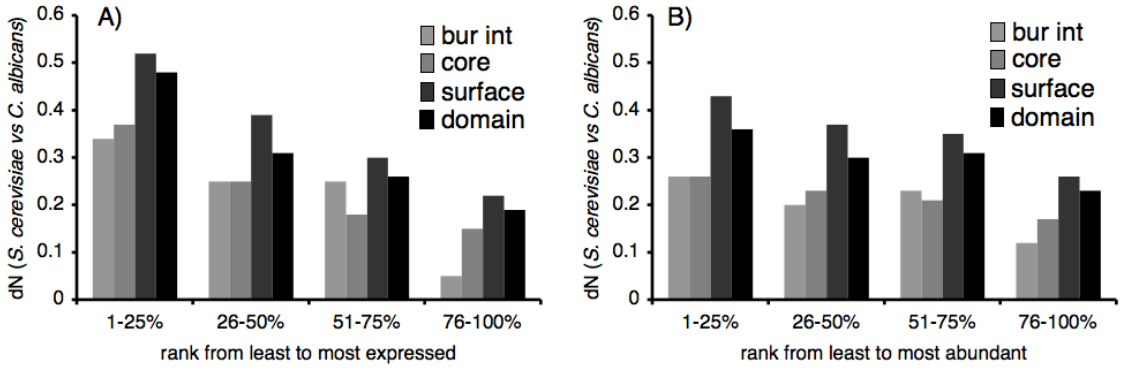


Figure 1-4. Pressures proportional to a protein's mRNA expression level affect the degree of conservation in all structural subsets, whereas differences as a function of protein abundance are substantially less pronounced. Shown are the median dN values for protein sets binned by increasing mRNA expression level (A) or protein abundance (B). As defined by a Wilcoxon-signed rank test, the dN distributions for all expression bins are significantly different from each other within each structural subset class ($P \leq 0.004$, with the exception of 26-50% *vs* 51-75% for all four subsets). The dN distributions for all abundance bins are not significantly different within each structural subset class ($P > 0.01$, with the exception of 1-25% *vs* 76-100% for all four subsets. For a list of P values, see Table S3.

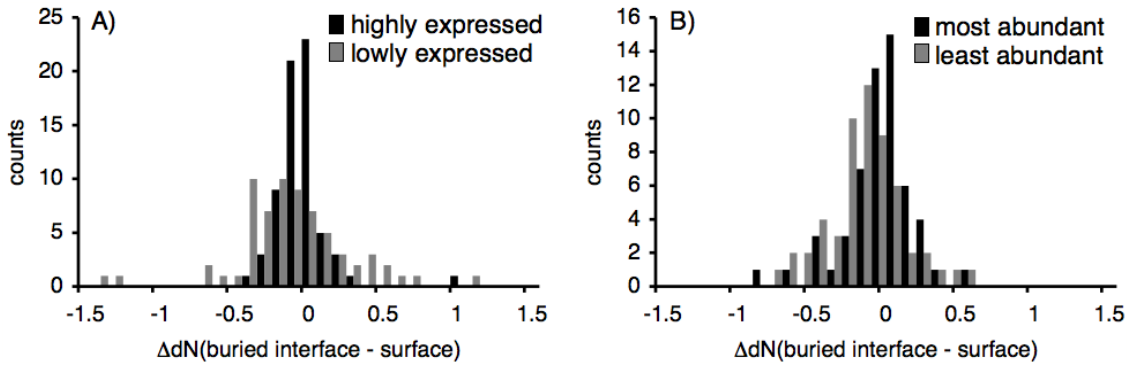


Figure 1-5. Pressures correlated to mRNA expression level reduce the evolutionary rate differences between buried interface and surface residues of the same domain, while those correlated to protein abundance do not. The distribution of differences in dN between the buried interface residues and surface residues sets as grouped by expression extremes (A) or abundance extremes (B). The extremes reflect the 17.5% highest and lowest expressed genes (A) or abundant proteins (B). Using the Kolmogorov-Smirnov test, we found a significant difference between the two expression distributions, but not between the two abundance distributions (P = 0.0004 for (A), P = 0.835 for (B)). One outlier falls outside the limits in (B).

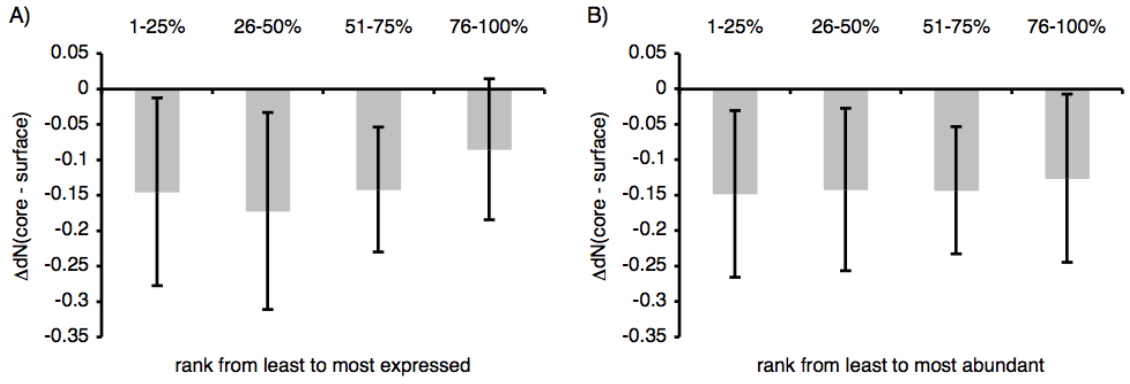


Figure 1-6. The mean difference in evolutionary rates of core and surface residues within the same domain decreases as a function of mRNA level, but not as a function of protein abundance. The mean ΔdN(core-surface) for proteins is binned by increasing mRNA

expression level (A) or protein abundance (B). The whiskers represent standard deviations within each bin.

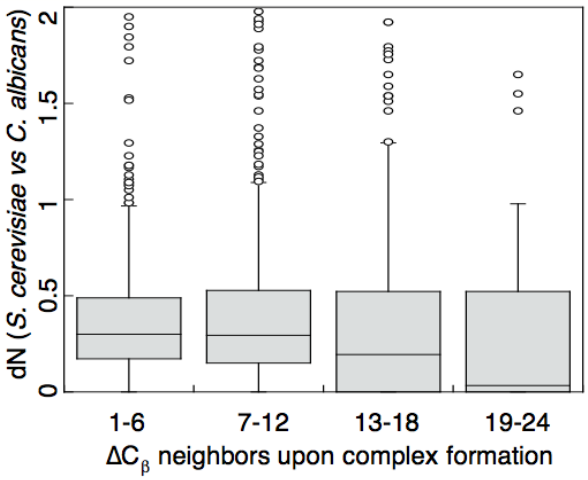


Figure 1-7. The evolutionary rates of interface residues depend on the degree of burial upon complex formation. As defined by a Wilcoxon-signed rank test, the dN distributions for all degree of burial bins are significantly different from each other ($P < 3e-7$) with the exception of the first two bins 1-6, 7-12 ($P = 0.58$) and the last two 13-18, 19-24 ($P = 0.06$). Bin 1-6 represents 22,210 residues; bin 7-12, 14,479 residues; bin 13-18, 3,229 residues; bin 19-24, 418 residues. 9 outliers fall outside the upper boundary.

Tables for chapter 1

Structural subset	total	high exp.	low exp.	high ab.	low ab.
	dN	dN	dN	dN	dN
Core	0.22	0.12	0.38	0.18	0.26
Buried interface	0.23	0.02	0.36	0.11	0.24
Surface	0.36	0.19	0.53	0.28	0.42
Surface interface	0.27	0.17	0.47	0.21	0.36

Table 1-1. The median non-synonymous substitution rates (dN) for the 4 structural subsets. "High exp" and "low exp" constitute the 17.5% highest and lowest expressed genes found in our dataset. "High ab" and "low ab" constitute the 17.5% most and least abundant proteins in our dataset.

Structural subset	Surface	Core	Surface interface	Buried interface
dN polars	0.434	0.305	0.322	0.219
dN remainder	0.311	0.209	0.260	0.184
P (Wilcoxan)	2.4e-7	1.9e-6	0.002	0.730

Table 1-2. Polar buried residues in interfaces are as conserved as all other residues. Polar residues include Asp, Glu, Lys, Arg, Ser, Thr, Asn, and Gln. All other residues are classified as the remainder. The median dN for each dataset is shown.

Supplementary materials for chapter 1

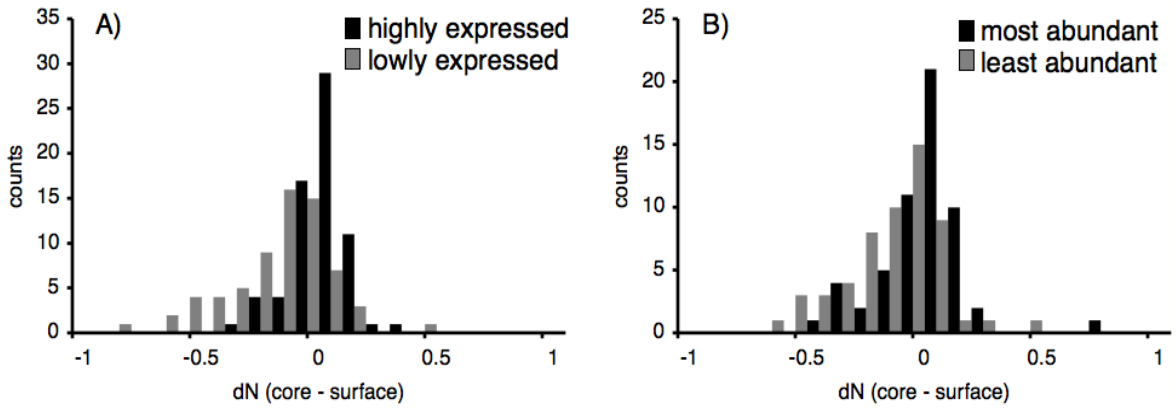


Figure 1-S1. Pressures correlated to mRNA expression level reduce the evolutionary rate differences between core and surface residues of the same domain, while protein those correlated to protein abundance do not. The distribution of differences in dN between the core residues and surface residues sets as grouped by expression extremes (A) or abundance extremes (B). The extremes reflect the 17.5% highest and lowest expressed genes (A) or abundant proteins (B). Using the Kolmogorov-Smirnov test, we found a significant difference between the two expression distributions, but not between the two abundance distributions (P = 0.005 for (A), P = 0.343 for (B)). One outlier falls outside the limits in (B).

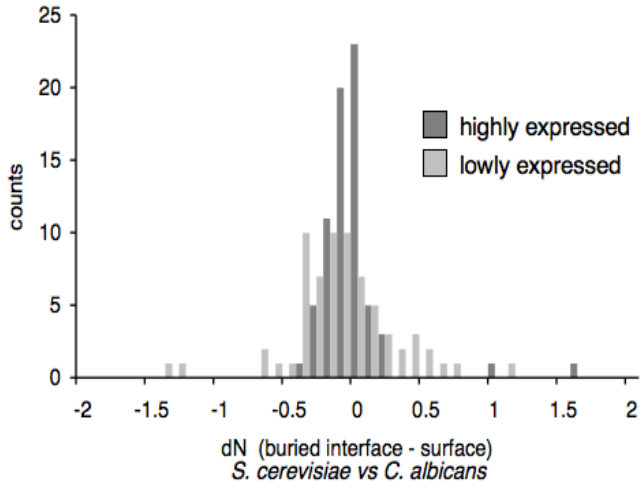


Figure 1-S2. Effect of excluding ribosomal domains. Pressures correlated to mRNA expression level reduce the evolutionary rate differences between buried interface and surface residues of the same domain, with ribosomal domains removed. A Kolmogorov-Smirnov test reveals a significant difference between the two expression distributions ($P = 0.0023$).

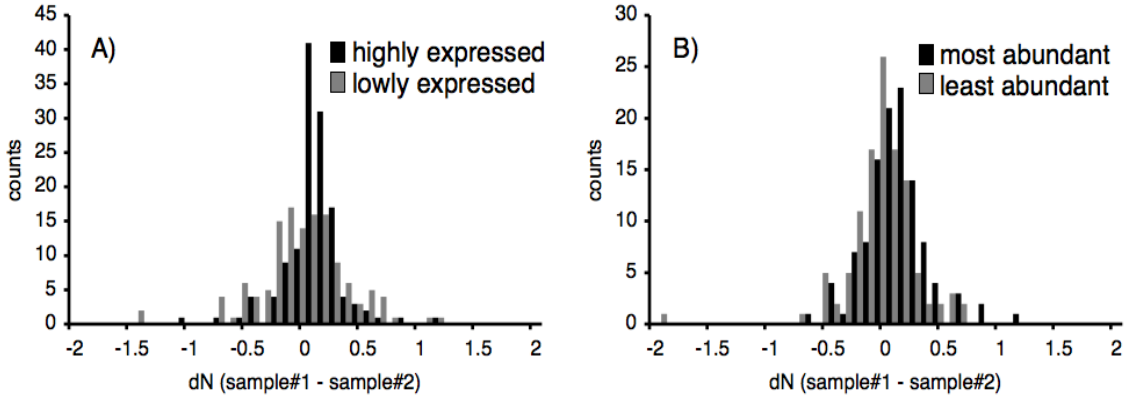


Figure 1-S3. Pressures correlated to mRNA expression level reduce the evolutionary rate differences between randomly drawn residue subsets of the same domain, while those correlated to protein abundance do not. The distribution of differences in dN between two subsets (consisting of 10 non-overlapping residues) as grouped by expression extremes

(A) or abundance extremes (B). The extremes reflect the 17.5% most and least expressed genes (A) or abundant proteins (B). Using the Kolmogorov-Smirnov test, we found a significant difference between the two expression distributions, but not between the two abundance distributions ($P = 0.009$ for (A), $P = 0.55$ for (B)).

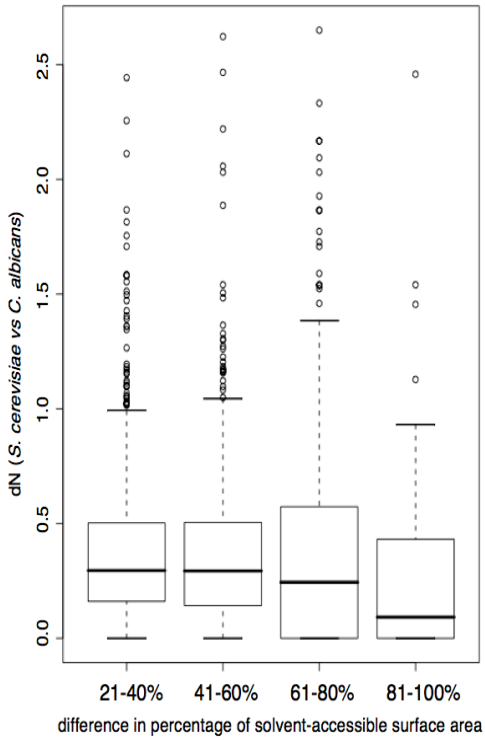


Figure 1-S4. The relationship between evolutionary rate and degree of burial for interface residues, as calculated by solvent accessible surface area (SASA). As defined by a Wilcoxon-signed rank test, the dN distributions for all interface residue bins are significantly different from each other ($P < 0.006$) with the exception of the first two bins 21-40%, 41-60% ($P = 0.55$). 2 outliers fall outside the upper boundary.

	expression	abundance
C. albicans	0.0004	0.835
C. glabrata	6.1e-6	0.092
S. mikatae	7.5e-5	0.237
S. bayanus	2.4e-5	0.006

Table 1-S1. Expression- and abundance-dependent evolutionary trends are preserved in different ortholog datasets. Kolmogorov-Smirnov tests between both upper and lower expression and abundance distributions (17.5 percentiles) in ΔdN(interface-surface), as calculated comparing *S. cerevisiae* sequences to four different ortholog datasets: *C. albicans*, *C. glabrata*, *S. mikatae*, and *S. bayanus*.

Expression extremes	KS-test statistical significance
10	0.00025
20	0.00005
30	0.00008
40	0.00104
50	0.02774
60	0.01293
70	0.02656
80	0.13790
90	0.69940

Table 1-S2. Changes in upper and lower expression regime size retain significance of expression-dependent evolutionary constraints, as calculated using *S. cerevisiae*-*S. bayanus* orthologs. *S. bayanus* was used in this analysis on account of the larger data set of *S. cerevisiae*-*S. bayanus* orthologs.

	expression				abundance			
	bur int	core	surface	domain	bur int	core	surface	domain
P(1-25,26-50)	0.003	2e-5	3e-4	1.1e-5	0.13	0.1	0.05	0.17
P(1-25,52-75)	0.001	1e-7	3e-6	4.6e-10	0.997	0.02	0.05	0.21
P(1-25,76-100)	1e-11	4e-15	2e-16	2.2e-16	0.005	0.001	4e-4	0.001
P(26-50,51-75)	0.61	0.14	0.17	0.017	0.17	0.34	0.8	0.96
P(26-50,76-100)	9e-5	5e-6	2e-7	2.2e-9	0.09	0.04	0.06	0.036
P(51-75,76-100)	8e-6	0.003	1e-4	2.4e-4	0.004	0.25	0.12	0.047

Table 1-S3. *P*-values for Figure 4 in the main manuscript, as defined by a Wilcoxon-signed rank test. The dN distributions for all expression bins are significantly different from each other within each structural subset class ($P < 0.004$, with the exception of 26-50% vs 51-75% for all four subsets). The dN distributions for all abundance bins are not significantly different within each structural subset class ($P > 0.01$, with the exception of 1-25% vs 76-100% for all four subsets.

Determination of Interface Residues

A minimum of five residues was used to define an interface in order to filter out small non-biological interfaces (such as crystal contacts), although it is possible such contacts still exist in the dataset. Domain-domain interactions within the same protein (intra-molecular contacts) were not included in this analysis. For every *S. cerevisiae* domain,

the interface was defined as the sum of all mapped interface residues from every domain-domain alignment. For example, two structures were available to determine the interface of the Profilin domain. 1HLU and 2BTF had 12 and 15 interface residues, respectively. All of the contacts in 1HLU were found in 2BTF, and the additional 3 mapped contacts of 2BTF were summed into the *S. cerevisiae* Profilin domain interface.

Mapping Residues to Saccharomyces cerevisiae

Short repeat sequences, though prevalent in the *S. cerevisiae* genome, were not used in this analysis. Sequence repeats are known to have unusual structural and evolutionary properties [34], and their prevalence in *S. cerevisiae* could bias the results: Tetratrico peptide repeats (TPR), for example, are known to possess a consensus sequence of 8 amino acids, while the remaining residues exhibit a high degree of sequence diversity [35]. We attempted to filter sequences such as the TPR from the dataset by requiring a minimum domain length of 40 amino acids.

Determination of Buried and Surface Residues

Each domain was required to have at least 1 residue in the following groups: “surface residues”, “core residues”, “surface interface residues”, and “buried interface residues”. Domains possessing structural subsets that aligned entirely with an insertion/deletion were not used in this study.

The surface residue dataset is likely to contain a mixture of both surface and interface residues, given the high probability that many interfaces have not yet been structurally characterized. Accordingly, we can infer that the set of true surface residues is likely to

evolve faster than found in our analysis. To test for the effects of incorrectly assigning true interface residues to the surface set, we randomly moved residues from the buried interface set to the surface residue set and monitored the significance of the separation between their respective evolutionary distances. We found that although the average substitution rate of the buried interface increases, the separation between the two sets is significant even with half of the original buried interface residues moved to the surface residue dataset (data not shown).

Chapter 2. Quantitative analysis of fitness pressures reveals unexpected cost/benefit epistasis in the *lac* operon

Abstract

Cost and benefit of protein production are key parameters that shape the fitness of microbial organisms. Accordingly, the economics of protein expression should pose constraints reflected both in the choice of regulatory mechanisms as well as the evolution of the proteins themselves. Here we use the *lac* operon of *Escherichia coli* to experimentally quantify the selective pressures proposed to explain the well-documented relationship between expression level and protein evolutionary rate. Though our findings provide insight into the relative strengths these pressures, they most importantly yield surprising implications about the *lac* operon itself. We present results that cannot be explained with the current model linking bacterial growth rate with the cost associated solely with the process of protein production. Instead, through 27 redesigned *lac* operon constructs, we characterize as the dominant determinant of cost the level of functional permease LacY. We suggest "within operon epistasis" as a regulatory mechanism, by which the penalties of expressing a costly protein, such as a permease, are only incurred when they can be balanced by a concurrent fitness benefit *via* the metabolic enzyme LacZ. The model also provides an explanation for the choice of the product of a LacZ-catalyzed side reaction, allolactose, as the lac operon inducer, working as a "systems check" for functional LacZ.

Introduction

The *lac* operon of *Escherichia coli* encodes one of the most well studied metabolic control devices in biology [36, 37]. It has served as a paradigm for genetic regulation [38], bistable behavior [39] and operon structure [40]. The proteins expressed by the operon, in particular the metabolic enzyme β-galactosidase (LacZ) and the transporter LacY, as well as the operon regulator LacI, have been important model systems for characterizing principles of protein allostery [41], enzyme kinetics [42] and membrane protein structure and function [43]. More practically, the lac operon induction mechanism is widely used for controllable protein expression systems.

Because its regulatory mechanism is known in detail and can be manipulated in predictable ways, the *lac* operon has emerged as an ideal experimental system to quantify the impact of protein expression on cell physiology and organism fitness [44, 45]. By determining a quantitative relationship between *lac* operon expression and growth rate, Dekel and Alon [44] derived a cost and benefit model that predicts the optimal level of protein expression from the *lac* operon under given conditions. They moreover demonstrated experimentally that expression could be rapidly tuned to the predicted optimum by evolution over several 100 generations. Stoebel *et al.* built upon these efforts by attempting to determine the underlying source of the cost of *lac* operon expression [46]. Based on these studies, it appeared increasingly clear not only that the cost and benefit of protein expression reflect fundamental constraints on organism fitness, but furthermore, that these constraints can be quantified by using the *lac* operon as a model system.

Through their impact on organism fitness, the costs and benefits of protein production should also pose constraints on the evolution of the proteins themselves. In support of this link, there is a pronounced negative correlation between protein expression levels and protein evolutionary rates [14 , 15, 18, 25, 47-50] observed across a broad spectrum of species [51]. While an explanation for this relationship has been proposed and substantiated through computational analyses by Drummond, Wilke and co-workers [18], direct experimental evidence has been absent.

Assuming that expression-level dependent effects on fitness could be determined using the *lac* operon, our work set out to use the operon as a well-studied model system to directly experimentally test and quantify the relative importance of evolutionary pressures proposed to explain why the most highly expressed proteins evolve the slowest [18]. The Drummond and Wilke model [18] outlined three selective pressures: (i) translational efficiency, which states that increases in expression level lead to selection for optimized sequences that translate faster or more accurately; (ii) functional efficiency, which proposes that mutations altering the function of a protein would be more strongly selected against in highly-expressed proteins, assuming that the fitness effect caused by a functional perturbation is proportional to the cellular concentration of the protein in question; and (iii) translational robustness, which poses that highly expressed proteins are constrained to sequences with increased robustness against translational errors (i.e. the protein would still fold even if it incurs errors during its cellular synthesis). The authors favored a model where the expression level-rate correlation is not based on the innate

functional importance of high-copy proteins, but instead stems from pressures to evolve sequences that can fold properly despite translational errors (translational robustness)[18]. While this model is intriguing and consistent with many observations [51], there exists no direct experimental demonstration of the proposed effect of translational robustness. Moreover, it is quite feasible that all proposed effects play some role in contributing to evolutionary rates; hence, a key question becomes to experimentally quantify the relative contribution of the three selective pressures.

To dissect and quantify these contributions and their dependence on protein expression levels, we made defined nucleotide perturbations to the β-galactosidase gene *lacZ* in the *lac* operon of *E. coli*, designed to modulate primarily one of these selective pressures. For translational efficiency we targeted exclusively synonymous substitutions, thereby preserving protein sequences but varying the levels of translational optimization. For functional efficiency we targeted catalytic residues; through single point mutations we aimed to largely preserve translational properties and protein stability, while altering functionality. For translational robustness we targeted residues identified as structurally important, simultaneously avoiding catalytic residues and minimizing translational efficiency effects. We then measured the effects of these perturbations on both expression levels of the lac operon proteins and on bacterial growth rate as a proxy for fitness.

Our initial comparison of the three investigated pressures suggested that LacZ functionality – the amount of LacZ produced and its functional efficiency - has the dominant effect on organism fitness. The measured benefit of *lac* operon expression, as

might be expected for a metabolic system, is directly proportional to LacZ levels. However, the incurred benefits were offset by substantial fitness penalties that could not be explained using the expected relationship between LacZ levels and cost of expression.

Instead, analysis 27 designed operon strains revealed surprising results about the *lac* operon itself: most notably, we can attribute the vast majority of the cost of expressing the operon to the function of the *lac* permease, a link that had previously been missed. Furthermore, through operon organization, the cell appears to have evolved a means by which to incur this cost only when a corresponding metabolic benefit will be reaped: Our results suggest both a new rationale for pressures to co-express proteins from a single operon and explain the benefit of evolving allolactose, not lactose, as the natural inducer of the *lac* operon. In particular, we suggest "within-operon epistasis" as a mechanism for beneficial proteins to offset the fitness penalties of costly proteins.

Results

Experimental Strategy for Cost/Benefit Analysis

The *lac* operon as a model system (Figure 2-1A) enables dissecting the cost and benefit of expression of the *lac* proteins in each designed strain in the presence and absence of the natural substrate of LacZ, lactose, as outlined in [44]. In each lactose response experiment (Figure 2-1B), we maximize expression of the operon with saturating conditions of the artificial inducer IPTG. Because the operon is always fully induced, we

decouple regulatory effects from determining the cost and benefit of expression, and also avoid complications arising from bistability at low inducer concentrations [52]. We then measure the growth rate at different external lactose concentrations relative to uninduced cells. The corresponding change in growth rate reflects both the cost of protein production and the benefit of lactose metabolism. In this experiment, mutations modulating translational efficiency, functional efficiency and translational robustness should have different effects: For translational efficiency we expect a difference in cost of protein production due to translational rate differences and expression level changes, but not to benefit per protein produced, given the identical amino acid sequence. For functional efficiency we expect differences in benefit, but minimal changes to cost of production if protein stability is not affected. For translational robustness both cost and benefit may be affected due to the maintenance of destabilized proteins and their potentially reduced enzymatic activity.

Maintaining Wild-type Expression Characteristics

Given that we expect to measure fitness effects dependent on protein expression-levels close to those occurring in a naturally occurring system, it is desirable that our mutated *lacZ* genes retain the same copy number and promoter as wild-type *lacZ* (though mutations to the open reading frame might result in altered expression profiles). Consequently, a plasmid-based expression system is not ideal, as even low copy number plasmids can yield non-physiological levels of expression. Furthermore, the repressor *lacI* or the exclusion of downstream genes could also alter expression levels. We adopted

a cloning strategy (see Methods) whereby following a knockout of the entire *lac* operon, we reinserted the operon, after making the desired mutations, back into the chromosome. Our resultant knock-in of the wild-type *lac* operon (*KIlac*) successfully recaptured native cost-benefit profiles (Figure 2-2A).

Translational Efficiency Strains

To estimate effects of translational efficiency on fitness, we first measured the cost and benefit of a strain containing a translationally optimized *lacZ* sequence (*Oplac1***;** Table 2-1A*)*; as in Figure 2-1B, we define cost as the relative reduction in growth rate due to operon expression and benefit as the relative increase in growth rate in the presence of lactose. The sequence *Oplac1* was synthesized using preferred codons at every position (GENEART, see Table 2-S1 for sequence data), thus optimizing the codon adaptation index. Surprisingly, upon measuring the lactose response curve (Figure 2-2B) of *Oplac1*, we observed both a reduced cost and benefit of protein production, indicative of reduced expression. To verify reduced expression, we measured the relative LacZ activity of *Oplac1* and *KIlac* strains (Table 2-2). Confirming our lactose response curve predictions, we found *Oplac1* to have reduced β-galactosidase activity. To investigate the source of reduced protein expression, we also assayed the *lacZ* mRNA levels of *Oplac1* relative to *KIlac* (Table 2-2). As the *lacZ* mRNA level of *Oplac1* was unchanged compared to that of *KIlac*, the reduced LacZ activity of *Oplac1* is most likely the result of a diminished translational rate.

It has been suggested that translationally optimal sequences may require usage of non-optimal codons; for example, inter-domain regions may be used to slow translation to limit interference of domain folding [53]. For our next synthesized sequence, we adopted a different optimization strategy (developed by DNA 2.0) that incorporated codons for a subset of amino acids read by tRNAs that are the most highly charged during amino acid starvation (*Oplac2,* see Table 2-S1 for sequence data) [54]. We observed a reduction in benefit and cost of a strain carrying *Oplac2* that was comparable to *Oplac1* (Figure 2-2B). Measuring the β-galactosidase activity of *Oplac2* confirmed that the reduced protein expression levels were also similar to those of *Oplac1*. In addition, *Oplac2* exhibited reduced mRNA levels relative to *KIlac* and *Oplac1* (Table 2-2), suggesting a reduced transcription rate or increased mRNA degradation, though the translation rate of *Oplac2* per mRNA is between levels of *KIlac* and *Oplac1*.

During our analysis, we uncovered two additional strains with novel cost-benefit phenotypes. These strains were erroneously synthesized missing the first 6 nucleotides, corresponding to the first 2 N-terminal amino acid residues (Methionine and Threonine), and instead beginning at the Methionine at position 3. These deletions were introduced to *Oplac1* and *Oplac2*, producing *Oplac1Δ6* and *Oplac2Δ6*, respectively (Table 2-1A). Analysis of these two strains revealed a drastically altered lactose response curve (Figure 2-2B). The cost and benefit of protein production were significantly reduced. These phenotypes suggested substantially lowered expression compared to *Oplac1* and *Oplac2*, which was confirmed by measuring LacZ enzymatic activity and mRNA levels (Table 2-2).

Recent work by Kudla *et al.* [55] found evidence that the stability of mRNA folding near the ribosome binding site (RBS) played a more significant role in expression than codon bias. Indeed, the computationally predicted folding energy from nucleotide -4 to +37 relative to start was significantly larger for *Oplac1* and *Oplac2* than *KIlac* (*KIlac* ΔG = -6.7 kcal/mol, *Oplac1* ΔG = -9.5 kcal/mol, *Oplac2* ΔG = -9.1 kcal/mol) [56]. To test whether the mRNA secondary structure of our codon redesigned sequences affected expression, we substituted the first 37 nucleotides of *Oplac1* and *Oplac2* with those of *KIlac*, producing *Oplac1(37)-KI* and *Oplac2(37)-KI*, respectively (Table 2-1A). We also generated the reciprocal constructs by substituting the first 37 nucleotides of *KIlac* with those of *Oplac1*and *Oplac2* (Table 2-1A). While *Oplac1(37)-KI* and *Oplac2(37)-KI* retained the same mRNA levels of *Oplac1* and *Oplac2* (Table 2-2), the same was not true of the reciprocal construct, where both *KI(37)-Oplac1* and *KI(37)-Oplac2* had their mRNA levels reduced by about half compared to *KIlac*. We also observed changes to LacZ activity in our redesigned sequences. In particular in the *Oplac1*-based constructs, these differences did not simply parallel the mRNA level changes. Most strikingly, *Oplac1(37)-KI* had much increased LacZ levels over *Oplac1* (Table 2-2; see Figure 2-S1 for the lactose response curves). In contrast, the *Oplac2*-based constructs only showed small differences in both LacZ protein and mRNA levels. The differences in the lactose response curves of these strains will be further analyzed below.

Functional Efficiency Strains

To measure effects of altered functional efficiency, we introduced point mutations into LacZ that had been previously determined to alter enzyme kinetics. We selected 6 mutations from the literature representing a range of catalytic activities (Table 2-1B)[57-60]. As stated above, we expected to observe changes to protein benefit, but not to the cost of production, as previous work had shown that most selected point mutations have minimal impact on LacZ stability [57]. Indeed, our results revealed a gradual loss *of lacZ* benefit corresponding to the reduced catalytic activity (Figure 2-2C). Surprisingly, in the *E537Q* strain encoding a severely catalytically comprised LacZ mutant, we measured a cost that increases as a function of lactose concentration (for analysis of this cost, see Lactose-Dependant Growth Penalty, below). To confirm that our point mutation did not affect expression levels through changes to protein stability or translational efficiency, we assayed protein levels of gfp-tagged strains of *KIlac* and *E537Q*. We observed that protein expression levels were unaffected by the mutation (Figure 2-S2).

Translational Robustness Strains

The translational robustness hypothesis posits that genes have evolved to minimize the number of proteins that misfold after translational errors [18]. To estimate the effects of translational robustness, we introduced point mutations specifically selected to destabilize LacZ. We chose two mutations previously identified for their reduced stability (though also observed to have reduced catalytic performance) [61], and three mutations predicted by structure-based modeling using ROSETTA [62] to have large destabilization effects (Table 2-1C). The lactose response curves (Figure 2-2D) revealed similar phenotypes as

observed for catalytically compromised strains. We could not measure any significant cost differences between mutations affecting catalytic performance and mutations affecting both stability and catalytic performance.

Cellular costs incurred by the cell for production of destabilized protein may include increased activation of degradation pathways. To observe these effects we introduced the ssrA tag to the C-terminus of LacZ as described in [46], thus targeting the protein for degradation by the ClpXP and ClpAP proteases [63]. The tag reduced LacZ activity by a third (Table 2-1D), but did not alter the cost or benefit of protein production significantly (Figure 2-3A), as observed previously [46].

To further examine the effects of destabilization and corresponding potential for misfolding, we tested a strain encoding the aggregation-prone fusion protein preS2-LacZ, which had the preS2 region of the hepatitis B virus surface antigen attached to the N-terminus of LacZ [64, 65]. Surprisingly, this fusion also did not affect the cost (at minimum lactose) or benefit (at 1mM external lactose) of protein production (Figure 2-3A), though it differed in response to intermediate lactose levels, consistent with other strains exhibiting reduced LacZ activity (Figures 2-2A-C). To confirm that preS2-LacZ indeed aggregates under our conditions, we measured the presence of protein in soluble and insoluble fractions (see Figure 2-S3). We observed the majority of the expressed protein preS2-LacZ in the insoluble fraction, though some activity was measured in the soluble fraction (Table 2-2). However, the insoluble fraction is known to still retain LacZ

activity [64, 65], suggesting the tag was merely responsible for preS2-LacZ accumulation in the insoluble fraction, but not the complete inactivation of LacZ.

Lactose-Dependent Growth Penalty

Our results above revealed two components of protein expression cost through two different methods of reducing LacZ protein activity. The first method, the reduction of protein activity by reducing protein levels through codon redesigned sequences, resulted in reduced cost of expression. The second, the reduction of protein activity through mutations to catalytic residues, did not alter the cost at saturating IPTG levels in the absence of lactose but did result in a "hidden penalty" phenotype (Figure 2-2C). This phenotype was characterized by increasing cost with increasing lactose concentration. Adding a kinetically dead mutation (E537Q) to an expression-reduced construct (*Oplac2Δ6*) generated a strain with both reduced cost and the hidden penalty (*Oplac2Δ6E537Q*, Figure 2-3B). We were unable to engineer a codon-redesigned sequence with sufficiently low activity as to observe the hidden penalty without the E537Q mutation.

We next wanted to investigate possible sources of the observed lactose-dependent penalty. Previous work in the literature has described a phenomenon known as "lactose killing", in which cells taken from a lactose-limited chemostat die when plated on lactose minimal medium [66]. It had been hypothesized that this behavior results from the action of the *lac* permease, which imports a proton with every lactose molecule, and upon

exposure to lactose drives a toxic reduction in the proton-motive force and membrane potential. To examine this phenomenon as a potential source of the lactose-dependent hidden penalty we introduced a mutation (S56L) to *lacY* known to disrupt permease functionality but not affect LacY levels [67] in both *KIlac* and *E537Q*. Neither strain that contained the S56L mutation displayed the hidden penalty phenotype (Figure 2-3C). In addition, both strains demonstrated a substantial reduction in the cost of expression. This was unexpected, as activity of LacZ, assumed to be responsible for the majority of the observed cost, remained unaffected in the *S56L* strain (Table 2-2). To more clearly define the role of Lac*Y* in the hidden penalty, we knocked out *lacZ* and *lacA* from the operon, generating a *ΔZA* strain. Measuring the benefit curves of *ΔZA* and *ΔZA-S56L*, we observed only *ΔZA* to maintain the hidden penalty phenotype (Figures 2-3B,C). Taken together, these results suggested that the hidden penalty phenotype is linked only with functional LacY and disappears when Lac*Y* is still expressed, but not functional (for analysis of the reduced cost of the *ΔZA* strain, see Cost of Expression, below).

Cost of Expression

Previous work assumed the cost of expressing the *lac* operon, as measured by relative reduction in growth, to be correlated with the expression level of *lacZ* [44, 46]. Upon plotting cost against LacZ activity for our strains, we found this correlation to represent an incomplete picture of the system (Figure 2-4A). We observed the cost of expression per protein to decrease as the LacZ activity increased, contradicting the expectation that costs should increase as more cellular resources are redirected towards higher expression

of a given protein. Furthermore, a *ΔYA* strain (knocking out *lacY* and *lacA*, Figure 2-3C) and a *S56L* strain (reducing LacY functionality) produced essentially wild-type levels of LacZ at a minimal cost to the cell (Table 2-2). These strains continue to have fully induced *lac* operon expression under our conditions of saturating IPTG, as IPTG import is not dependent on LacY activity in *E. coli* [68].

The elimination of growth cost in the *ΔYA* and *S56L* strains mentioned above suggested a role for functional *lac* permease in the relationship between cost and expression. To determine levels of LacY activity, we measured import of radiolabeled lactose (see Methods) in all strains with altered LacZ expression (Table 2-2). Interestingly, we found a pronounced linear relationship between LacY activity and the relative reduction in growth rate (Figure 2-4B). This expression polarity of *lacZ* and *lacY*, i.e. that the attenuation of *lacZ* expression levels affects the expression of downstream genes, including *lacY*, has been reported previously [69] (for further analysis on the relationship between LacZ and LacY under our conditions see Figure 2-S4). Furthermore, it has been observed that at higher induction levels, *lacY* mRNA levels plateau while *lacZ* mRNA levels continue to linearly increase, consistent with our measured trends for LacZ and LacY activity levels (Table 2-2, Figure 2-S4) [69].

Benefit of Expression

We do observe a clear relationship between benefit and LacZ activity. Plotting the difference in relative growth rate between maximum and minimum lactose concentrations

against measured LacZ activity (Figure 2-4C, plot includes all strains found in Table 2-2), we see the benefit increase proportionate to LacZ activity until activity levels reach approximately 30% of *KIlac*, at which point no additional benefit is conferred for higher activity. This observation agrees with our earlier results indicating that *lacZ* can be fully expressed at little to no cost to the cell, as wild-type cells exceed this maximum benefit threshold by a factor of 2 (Table 2-2). Such excessive production would likely not be tolerated if it were at great expense to the cell.

Evolutionary Defense Against the Hidden Penalty

Given the high cost of expressing *lacY*, particularly at high lactose concentrations, we found the relative ease surprising with which we could engineer mutations that killed LacZ activity while maintaining expression of costly Lac*Y* protein (Figure 2-2C,D). However, nature may have evolved an elegant solution to test whether LacZ is functional before inducing the operon and exposing the cell to the cost of *lacY* expression. The natural inducer of the operon is not lactose, but allolactose, which is generated by the transglycosylation activity of LacZ [70]. If LacZ is not functional, for example because it has been destabilized, then it will most likely also not be able to catalyze this secondary reaction to produce the inducer of the operon.

To test this idea, we measured the cost of expression in 1mM lactose without IPTG in strains possessing the hidden penalty phenotype under our previous conditions. As expected, none of these strains incurred the hidden penalty in the absence of IPTG

(Figure 2-5). Measurement of expression in Gfp-tagged strains of *E537Q* confirmed that lactose alone failed to induce operon expression in catalytically compromised strains (Figure 2-S5).

Discussion

The goal of this work was to quantify the relative influence of three selective pressures, those associated with translational and functional efficiency and translational robustness, upon the *lac* operon; our results proved more revealing about the operon itself. First, we found the cost of expressing the operon to be primarily determined by and linearly dependent on the activity of the LacY. This result revises previous models, where this cost increased nonlinearly as production diverted other resources away from the cell, and models suggesting the process of producing LacZ as the primary source of the expression cost. Second, we propose a role for allolactose as the *lac* operon inducer as a "systems check" to protect the cell from expressing an operon containing a faulty LacZ enzyme. Third, we suggest a new functional role for operon context - "within-operon epistasis" - as a means to avoid expression of costly proteins unless their costs can be balanced by the function of the other co-expressed proteins in the operon. Finally, we found functional loss to be the dominant pressure within the *lac* operon, though we caution against interpreting this result to be a universal evolutionary trend.

The finding that the function of the *lac* permease drives the observed cost and not the process of translation or transcription contradicts previous suggestions [46]. At least part

of these opposing conclusions can be attributed to differences in measured permease expression levels of the *ΔZA* strain. We find that the activity of LacY in a *ΔZA* strain is significantly reduced compared to the *KIlac* and *wild-type* strains (Table 2-2). As seen in our *Oplac1Δ6* and *Oplac2Δ6* strains, even small changes in 5' sequences can result in substantial changes in expression level. Similarly, substantial expression level changes in downstream genes such as *lacY* could be expected upon deletion of upstream genes in the operon, as in the *ΔZA* strain. Given that the sequence of *lacZ* deletion, including flanking regions, is identical in the two studies (see Methods), we believe we can limit the source of the different measured expression to two most likely explanations: media differences (we used glycerol and lactose as a carbon source, Stoebel *et al.* used glucose, maltose, and succinate) or methodology differences, in particular for measuring permease levels. The approach used by Stoebel *et al.* compared the α-PNPG translocation rate of a strain constitutively expressing the operon, a constitutive *ΔlacZ* strain, and a strain carrying at least five copies of the *lac* operon, and made the assumption that the *lac* permease was rate limiting for transport. However, it may be important to control for α-PNPG translocation of a *ΔlacY* strain (i.e. LacY independent α-PNPG translocation), particularly in light of the α-PNPG translocation activity of the melibiose transport system in *E. coli* [71], though it is noted that melibiose transporter should be inactive at 37° [46, 72]. While overestimation of permease levels would not account for their observed cost in a *ΔlacY* strain, it would lead to an underestimation of permease cost in the *ΔlacZ* strain. The media differences, on the other hand, might lead to a decrease in LacY transport activity (and thus the cost) through the inducer exclusion mechanism [73]. It has been demonstrated that the addition of glucose 6-phosphate results in a strong

inhibition of lactose transport mediated by LacY [73]. Hence, the cost associated with active LacY may not be observable or significantly attenuated under nutrient conditions that engage inducer exclusion. In this case, mechanisms such as inducer exclusion that affect permease activity directly (instead of levels) may constitute another strategy evolved to control the cost of an active permease.

Our results also affect models of *lac* operon expression. While many such models exist, we will focus here on that presented by Dekel and Alon, as it was the foundation for our work. Dekel and Alon presented a model of the *lac* operon, where cost was determined by LacZ levels and resource limits. The most salient feature of this model is its nonlinearity; that increasing production of *lac* operon proteins becomes increasingly costly to the cell on a per-protein basis, due to reduced capacity to manufacture other essential proteins. However, our model of cost, at least under physiological ranges of operon expression, is largely independent of both LacZ levels and resource limits, and more importantly, exerts a linear relationship with the activity of LacY (see Figure 2-S2 for a comparison of LacZ and LacY activities for the strain shown in Table 2-2). We believe the likely source of this discrepancy to be the means by which we generated our reduced-expression strains. We altered the codon usage of *lacZ* while Dekel and Alon used reduced concentrations of IPTG, which might result in different degrees of expression polarity in downstream genes, in particular *lacY*.

Furthermore, our results may shed light on the induction mechanism of the *lac* operon. The use of allolactose (produced through the transglycosylation side reaction activity of

LacZ) instead of lactose as the inducer of the operon would appear to be a circuitous means of regulation. However, this approach may be explained by the inability of strains encoding functionally compromised LacZ to express the operon and costly LacY in the presence of lactose. These same strains, induced by IPTG, incur substantial growth penalties in the presence of lactose (Figure 2-2C). We propose that *E. coli* evolved the use of allolactose as the inducer as a sort of "systems check" to avoid this potential growth penalty.

Our initial comparison of the three selective pressures highlighted the relative importance of LacZ functionality. While destabilizing LacZ also produced the hidden penalty phenotype, this penalty was incurred through loss of function; no additional penalty could be measured for destabilization-related effects. However, we may not have expressed a sufficient amount of destabilized LacZ to reveal effects predicted by the translational robustness hypothesis, as our assay was not sensitive enough to detect smaller effects. As mentioned, the hidden penalty phenotype is only observed in the presence of the artificial inducer, IPTG. Thus, because of the allolactose-based "system check" induction mechanism, the relative pressure proportional to LacZ functionality (including, importantly, the cost of loss of function) is diminished in natural environments.

As the hidden penalty phenotype is unlikely to occur in nature, the relative importance of translational efficiency emerges. While we could not measure the effects of altered codon usage on a cost-of-translation-per-protein basis, we could measure the cost of altered expression levels. Surprisingly these changes in cost were not dependent on changes in

lacZ expression, but instead upon those of *lacY* leading to generation of active *lac* permease, the expression of which was affected by upstream alternative codon usage.

Given the high cost of expressing functional *lacY*, the grouping of the *lac* genes into an operon is not surprising. The strict coexpression of the *lac* genes must be maintained for the growth benefit of LacZ to offset the growth penalties associated with functional LacY importing lactose into the cell. This organization may also suggest a broader role for operons [74] exhibiting "within-operon epistasis" – a means by which to link expression of costlier proteins with more beneficial ones.

Why is the functional permease costly to the cell? Our results appear to agree with ideas advanced behind the "lactose killing" hypothesis [66], where the buildup of protons imported into the cell with each lactose molecule leads to a toxic reduction in the proton-motive force. This is consistent with the increasing growth penalty observed in higher lactose concentrations. However, we also measure growth penalties at minimal lactose concentrations. This may suggest the *lac* permease to be inherently leaky, allowing protons into the cell even in the absence of lactose – though further experimentation, including the measurement of proton transport in the absence of lactose, is required to explore this possibility.

Are the penalties we uncovered specific to the *lac* permease or are they part of a more general phenomenon? Taken together with our results, a few recent publications provide evidence for the latter. In work by Osterberg *et al.* [75] it was revealed that

overexpression of membrane proteins in *Saccharomyces cerevisiae* resulted in a significantly reduced growth rate in 93% of the 617 membrane proteins assayed. A comparable overexpression study by Sopko *et al.* [76] covering over 5,000 genes of *S. cerevisiae* (not specific to membrane proteins) found only 15% of overexpressed genes reduced the growth rate. Of this subset, they found an overrepresentation of transport proteins. Furthermore, studies limited to single systems demonstrate parallels to results presented here. Recently published work by Palmer *et al.* showed that *E. coli* strains resistant to tetracycline develop a fitness disadvantage as tetracycline degrades and induces expression of the costly TetA efflux pump [77].

Interestingly, in a genomic analysis, we would not expect dominant evolutionary pressures to maintain permease functionality: Only active permeases would incur fitness costs and these costs are generally balanced by LacZ-related benefits. Rather, we would expect the burden of pressures to fall upon both regulatory mechanisms of permease expression, and their corresponding downstream pathways that might offset their growth penalties. In the context of permease expression costs, another observation is of note: When gene regulation in *E. coli* is rewired using new promoter-transcription factor combinations, permeases are significantly down-regulated after few rounds of selection under stress conditions [78]. This finding is consistent with a significant selective pressure on permease expression when normal regulatory mechanisms are disrupted. More generally, pressures of costly proteins being balanced by beneficial ones may account for clustering of functionally related genes within genomes [79] and the role of flux coupling in co–regulated metabolic genes [80]. Future experimentation could

attempt to tease apart the cost/benefit contributions of functional proteins co-occurring in the same pathway to elucidate the source of similar fitness epistasis effects as uncovered here in the *lac* operon paradigm.

Methods

Strains and Media

The *E. coli* strain MG1655 was used as background for all constructs. We constructed knock-out strains using the approach of Datsenko and Wanner [81]. MG1655 was transformed with pKD46 (carrying phage λ Red recombinase) and a PCR fragment encoding a kanamycin resistance gene with flanking nucleotide sequences homologous to those flanking the *lac* operon. Upon recombination and selection for resistant strains, the kanamycin-resistant gene was eliminated following transformation with pCP20 (encoding the FLP recombinase), which is subsequently cured by growth at 30°C. The deletion in our Δ*ZYA* strain started 20bp upstream of *lacI* and ran 40bp downstream of *lacA*.

We constructed all knock-ins strains using the approach of McKenzie and Craig [82]. The knock-in contained the entire *lac* operon starting 75bp upstream of *lacI* and ending 100bp downstream of *lacA*. The Δ*ZYA* strain was transformed with pGRG37. The plasmid carried both the *lac* operon and the transposon genes *tnsABCD*, which allow for a site-specific insertion at *attTn7*. Given the size of the pGRG37 vector (~19kb), we performed all mutagenesis and cloning on the smaller pENTR plasmid, where the *lac* operon was also sequenced in full for all strains. Mutagenesis was performed using QuikChange XL

(Stratagene). To transfer the *lac* operon from pENTR to pGRG37 we used the Gateway cloning system (Invitrogen). Transfer of the *lac* operon into pGRG37 and into the chromosome of the Δ*ZYA* strain were both confirmed by PCR. For the codon-redesigned sequences, DNA 2.0 or GENEART performed gene synthesis. All growth experiments were carried out in M9 medium consisting of M9 salts, 1mM $MgSO_4$, 0.1mM $CaCl_2$, 0.05% casamino acids, 0.1% glycerol, and indicated amounts of IPTG and lactose.

Growth Assays

Following the method of Dekel *et al.* [44], strains were grown in M9 medium overnight, then diluted 1:250 and arrayed in "checkerboard" distribution with each strain (one induced with 500μM IPTG and in the presence of specified concentrations of lactose, one uninduced) in alternating wells of a 96-well plate. The plates were incubated at 37°C with shaking. O.D. measurements (600nm) were taken every 2 minutes. To calculate cost-benefit effects, the differences in exponential growth rates between adjacent wells were measured, giving 172 (non-independent) growth rate differences (See Figure 2-S6 for more details). Experiments using Neidhardt's medium, which differs from our medium primarily in the substitution of individual amino acids instead of Casamino acids and the addition of uracil, showed similar results when comparing induced and uninduced *KIlac* strains (data not shown).

Cell Fractionation

Cells were grown overnight in M9 medium and 500μM IPTG, then diluted (1:100) and grown 4 hours in fresh medium before harvesting. Cells were pelleted and resuspended in

1.0mL of TE buffer (50mM Tris/HCl, 2mM EDTA, pH 8.0), sonicated, and centrifuged to separated soluble and insoluble protein as described by Zhang *et al.* [83]. Fractions were analyzed on 4-20% Tris-Glycine gels.

LacZ Activity Measurement

Cells were grown overnight in M9 medium and 500μM IPTG, then diluted (1:100) and grown 4 hours in fresh medium before harvesting. Cells were resuspended in TE buffer and sonicated. Cells were centrifuged for 20 minutes at 12,000 rpm to separate soluble and insoluble fractions. Activity was determined by assaying soluble fractions using a β-Galactosidase Assay Kit (Stratagene, see Figure 2-S7). Results were normalized to the *KIlac* strain experiments performed in parallel.

qPCR

Cells were grown overnight in M9 medium and 500μM IPTG, then diluted (1:100) and grown 4 hours in fresh medium. mRNA was isolated using an RNeasy Protect Bacteria mini kit (QIAGEN) with the addition of RNase-Free DNase (QIAGEN). qPCR was performed on a Stratagene Mx3000P with Brilliant II SYBR Green QRT-PCR Master Mix (Stratagene). The dehydrogenase *gapA* was used as a reference gene. Results were normalized to *KIlac* strain experiments performed in parallel.

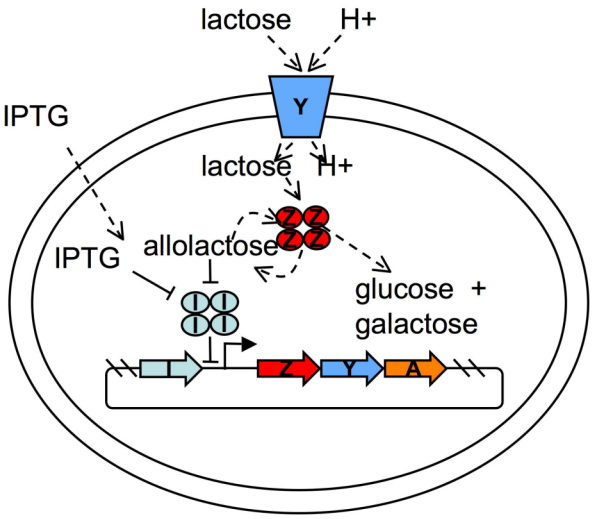
LacY Activity Assay

Permease activity assays used the approach of Dryselius *et al.* [84]. Cells were grown overnight in LB and 500μM IPTG, then diluted (1:100) and grown 2 hours in fresh

medium before harvesting. 1 ml of cells were incubated on ice for 10 minutes with 150 μg/ml chloramphenicol. ^{14}C-lactose (10 μM final concentration) was added and the cells were incubated at 37°C for 30 minutes. Cells were applied to a mixed cellulose filter (Millipore, pore size of 0.45 μm) and were washed immediately with PBS. Radioactivity on the filters was measured by liquid scintillation counting. Results were normalized to the *KIlac* strain experiments performed in parallel, and measured LacY activity levels correlated well with *lacY* mRNA levels measured by qPCR (see Table 2-S2 and Figure 2-S8).

Figures for chapter 2

(A)



(B)

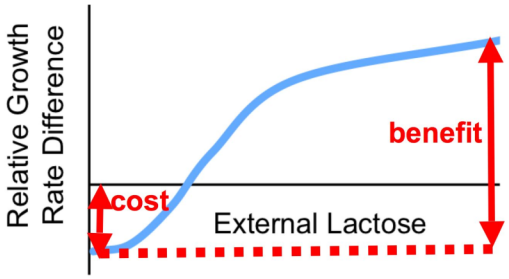


Figure 2-1. The *lac* operon as a model system to quantify fitness effects of operon expression. (A) A cartoon representation of the *lac* operon. (B) An example of the lactose response curve. The cost is represented by the relative growth rate difference between strains uninduced and induced in the presence of minimal lactose. The benefit is the relative growth difference of strains uninduced and induced in 1mM external lactose measured against their respective costs.

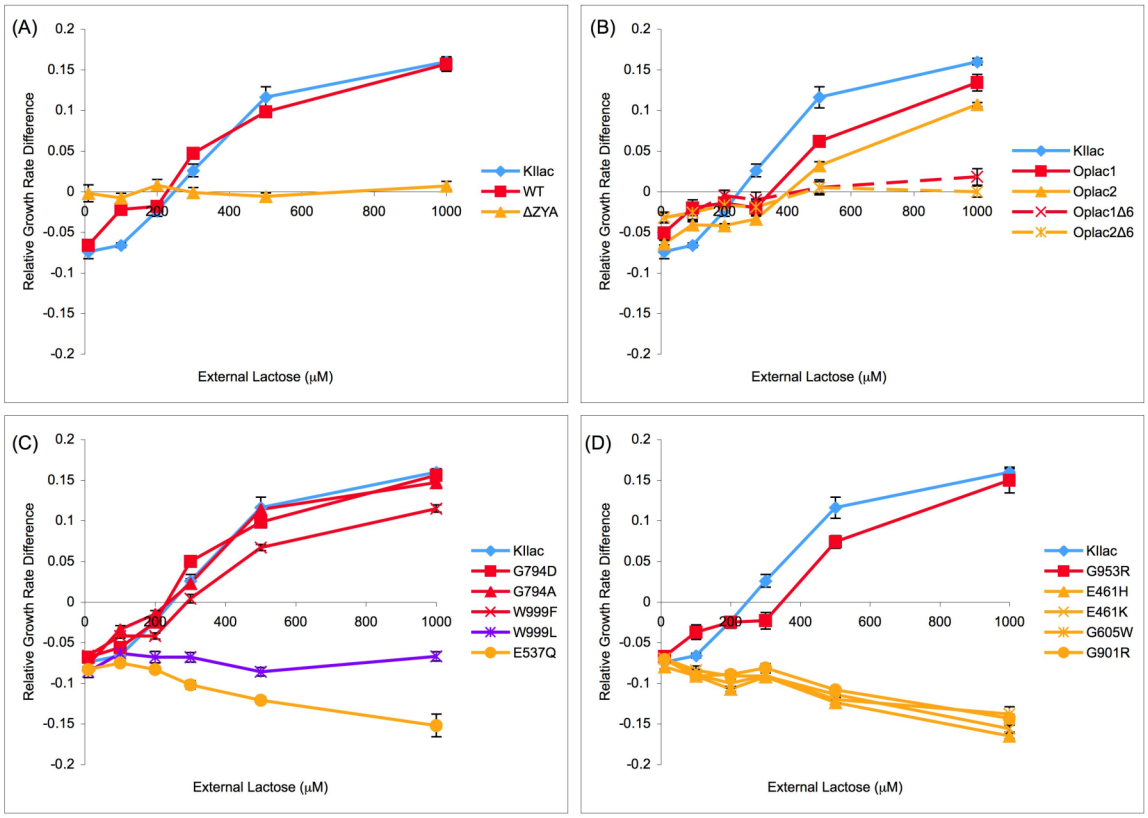


Figure 2-2. Cost/Benefit tradeoff upon modulating expression and LacZ functionality. (A) Recapturing the lactose response curve for the wild-type (*WT*) strain with a strain containing a knock-in of the *lac* operon (*KIlac*) at the *attTn7* chromosomal locus. The lactose response curve of the *ΔZYA* strain is shown as reference. (B) Lactose response curves of strains of varying translational efficiency (Table 2-1A). Dashed lines represent the *Oplac1Δ6* and *Oplac2Δ6* strains which are identical to the *Oplac1* and *Oplac2* strains with the exception of 6 missing nucleotides at the starting position. (C) Lactose response curves for strains carrying point mutations affecting catalytic functionality (Table 2-1B). (D) Lactose response curves for strains carrying mutations designed to destabilize LacZ (Table 2-1C). In each panel, the lactose response curve of *KIlac* is shown for reference.

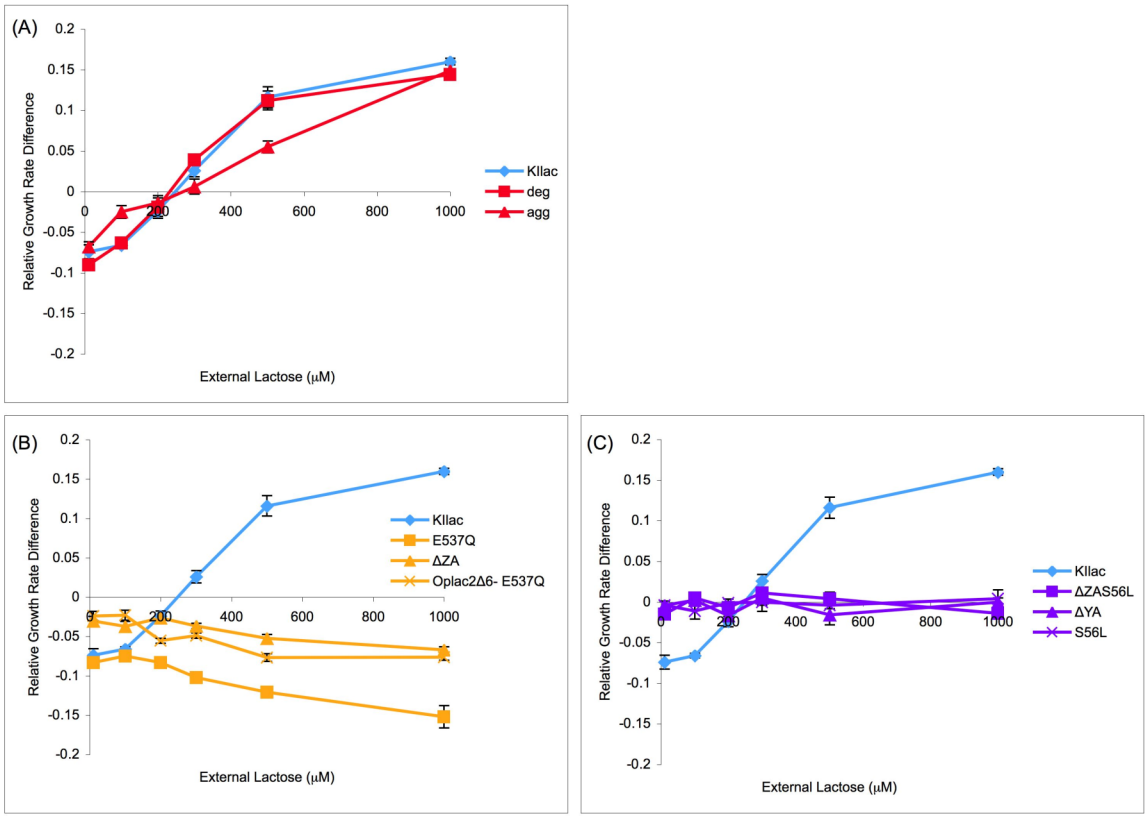


Figure 2-3. Origin of the growth penalty. (A) Lactose response curves for strains with a degradation tag (*deg*) and an aggregation-prone tag (*agg*). (B) Lactose response curves for strains with minimal LacZ activity. (C) Lactose response curves for strains with minimal LacY activity. In each panel, the lactose response curve of *KIlac* is shown for reference.

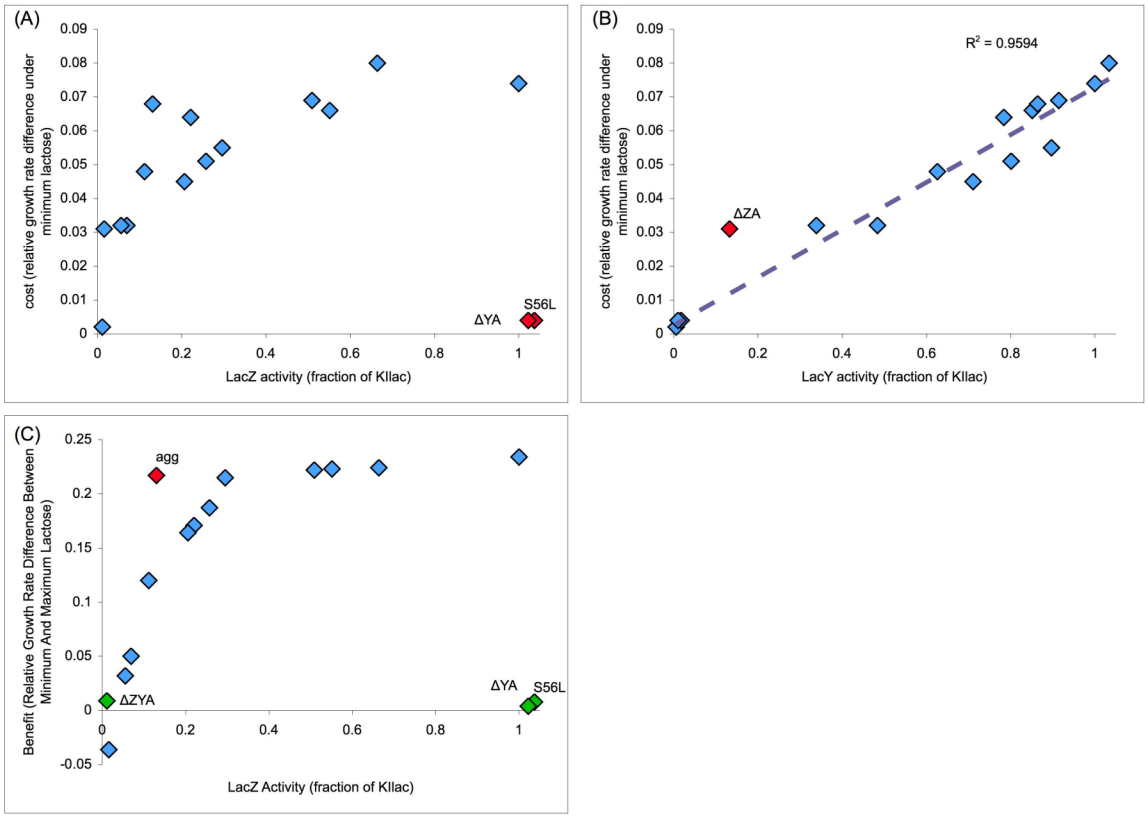


Figure 2-4. The observed cost and benefit tradeoffs lead to a model of lactose response. (A) Plot of cost of expression (growth rate difference under minimal lactose, comparing an induced and uninduced strain) against LacZ activity for strains in Table 2-2. The points in red have minimal LacY activity. (B) Plot of cost of expression against LacY activity for the same strains as in (A). The point in red, *ΔZA*, has minimal LacZ activity, which may result in a lower measurement of LacY activity as more ^{14}C-lactose is exported instead of metabolized. Fitting a linear relationship to the data (not including *ΔZA*) yields the equation: cost = 0.0686(LacY activity) + 0.003, R^2=0.9547. (C) Plot of the benefit (relative difference in growth rate between maximum and minimum lactose concentrations) against LacZ activity for the same strains as in (A). The points in green have minimal LacY activity. The point in red is the aggregation-tagged strain (*agg*).

While this plot only represents LacZ activities measured in the soluble fraction, it has been reported that *agg* possesses significant LacZ activity in the insoluble fraction [85], which may result in an underestimation of LacZ activity in our assays that use the soluble fraction.

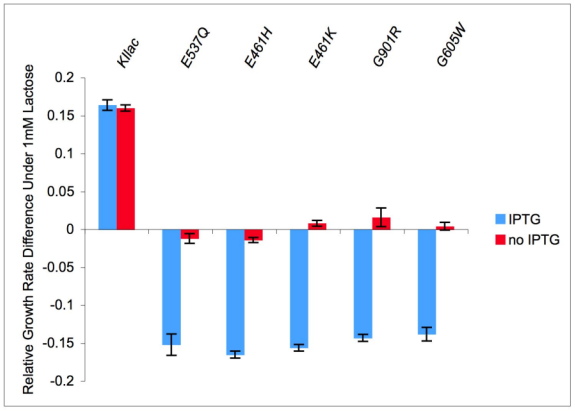


Figure 2-5. The role of allolactose as a systems check. Strains do not exhibit the hidden penalty in the absence of IPTG. Shown is a comparison of the relative growth rates of strains grown in 1mM lactose and 0.5mM IPTG and those grown in just 1mM lactose. There is no growth difference for *Klac* under the two conditions, but all strains exhibiting the hidden penalty when induced with IPTG loose all growth costs when induced only in 1mM lactose. In this case, the "systems" check is effective: no allolactose is produced when LacZ is defective, and the operon is not induced under conditions where the LacY-associated penalty cannot be offset by LacZ-related metabolic benefit.

Tables for chapter 2

Table 2-1. Strains used in this study. (A) The architecture of translational efficiency strains. The hatched pattern represents the six missing nucleotides of *Oplac1Δ6* and *Oplac2Δ6*. (B) Functional efficiency strains, with previously published kinetic constants. (C) Translational robustness strains. (D) Alternative strains designed to test the origin of the growth penalty resulting from operon expression.

Table 2-1 (A) Translational efficiency strains

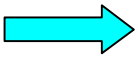
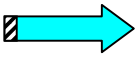
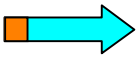
Strain	*lacZ* coding region
KIlac	
Oplac1	
Oplac2	
Oplac1Δ6	
Oplac2Δ6	
KI(37)-Oplac1	
KI(37)-Oplac2	
Oplac1(37)-KI	
Oplac2(37)-KI	

Table 2-1 (B) Functional efficiency strains

Strain	k_{cat} [s^{-1}]	K_m [mM]	Source
WT	600*	0.12	Juers (1994). Huber (2003)
G794D	285**	0.16	Martinez-Bilbao (1991)
G794A	189	0.19	Juers (2003)
W999F	54	0.26	Huber (2003)
W999L	20	3.3	Huber (2003)
E537Q	0.003	0.12	Yuan (1994)

*The k_{cat} of wild-type *E. coli* β-galactosidase has also been measured as 480s^{-1} by Yuan *et al.* [57] and 750s^{-1} Cupples *et al.* [61]. ** G794D was also measured to have a higher k_{cat} than wild-type when the kinetic analysis was carried out using lactose instead of *o*NPG [58].

Table 2-1 (C) Translational robustness strains

Strain	Source
E461H	Cupples (1990)
E461K	Cupples (1990)
G605W	ROSETTA predicted
G901R	ROSETTA predicted
G953R	ROSETTA predicted

Table 2-1 (D) Alternative strains

Strain	Description
deg	Degradation tagged LacZ
agg	Aggregation tagged LacZ
ΔZA	*lacZ*, *lacA* deletion
Oplac2Δ6-E537Q	*Oplac2Δ6* containing the E537Q mutation
S56L	Impaired LacY function
ΔZA-S56L	*ΔZA* containing the S56L mutation in *lacY*
ΔYA	*lacY*, *lacA* deletion

Table 2-2. The measured LacZ activity, *lacZ* mRNA, LacY activity, cost and benefit of translational efficiency strains with their standard errors. ND, not determined.

Strain	LacZ activity (fraction of KIlac)	LacZ mRNA (fraction of KIlac)	LacY activity (fraction of KIlac)	Cost	Benefit
KIlac	1	1	1	1	1
ΔZYA	0.01 +/- 0.01	0.00 +/- 0.00	0.01 +/- 0.00	0.00 +/- 0.01	0.01 +/- 0.01
WT	0.55 +/- 0.07	0.59 +/- 0.02	0.85 +/- 0.03	0.066 +/- 0.056	0.223 +/- 0.007
Oplac1	0.26 +/- 0.02	1.02 +/- 0.17	0.80 +/- 0.01	0.051 +/- 0.061	0.187 +/- 0.008
Oplac2	0.22 +/- 0.07	0.59 +/- 0.07	0.78 +/- 0.00	0.064 +/- 0.041	0.171 +/- 0.003
Oplac1Δ6	0.07 +/- 0.02	0.51 +/- 0.09	0.48 +/- 0.03	0.032 +/- 0.006	0.050 +/- 0.008
Oplac2Δ6	0.06 +/- 0.03	0.22 +/- 0.04	0.34 +/- 0.03	0.032 +/- 0.007	0.032 +/- 0.007
Oplac1(37)-KI	0.51 +/- 0.07	1.01 +/- 0.11	0.91 +/- 0.03	0.069 +/- 0.010	0.222 +/- 0.009
Oplac2(37)-KI	0.21 +/- 0.4	0.53 +/- 0.07	0.71 +/- 0.03	0.045 +/- 0.006	0.164 +/- 0.007
KI(37)-Oplac1	0.11 +/- 0.01	0.49 +/- 0.05	0.63 +/- 0.06	0.048 +/- 0.007	0.120 +/- 0.011
KI(37)-Oplac2	0.30 +/- 0.02	0.58 +/- 0.08	0.90 +/- 0.05	0.055 +/- 0.009	0.215 +/- 0.010
agg tag	0.13 +/- 0.02	0.59 +/- 0.10	0.86 +/- 0.07	0.068 +/- 0.006	0.217 +/- 0.008
deg tag	0.66 +/- 0.10	1.14 +/- 0.08	1.03 +/- 0.02	0.080 +/- 0.004	0.224 +/- 0.005
ΔYA	1.02 +/- 0.22	0.87 +/- 0.08	0.01 +/- 0.00	0.004 +/- 0.005	0.004 +/- 0.005
ΔZA	0.02 +/- 0.01	ND	0.13 +/- 0.01	0.031 +/- 0.006	-0.036 +/- 0.005
S56L	1.04 +/- 0.07	ND	0.02 +/- 0.01	0.004 +/- 0.004	0.008 +/- 0.008

Supplementary materials for chapter 2

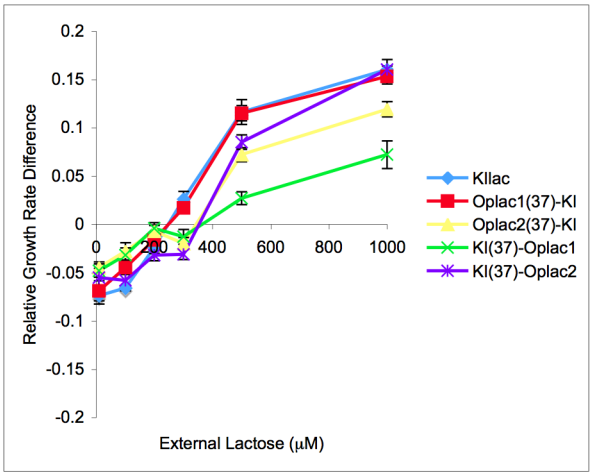


Figure 2-S1. Lactose response curves for strains with codon redesigned sequences in either the first 37 nucleotides of *lacZ* or the remainder of the *lacZ* coding region. LacY and LacZ activity levels and mRNA levels were altered and are given in Table 2-2 in the main manuscript. The rank order of cost and benefit of the different strains is consistent with the relationships shown in Figures 2-4B,C in the main manuscript, which includes the 4 designed strains shown here.

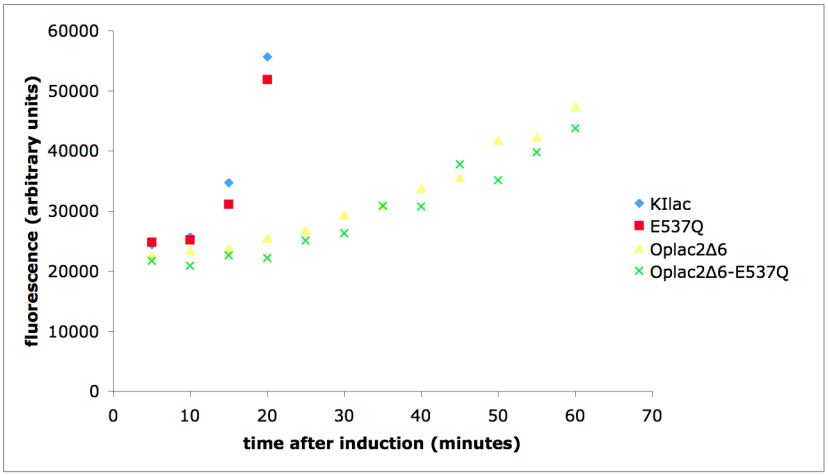


Figure 2-S2. Measuring fluorescence of gfp-tagged strains following induction. The E537Q functional loss mutation did not affect expression in either the wild-type codon context or in the Oplac2 strain.

Fluorescence Measurements

Prior to chromosomal knock-in, *GFPmut2* [86] was attached to the C-terminus of *LacZ* with a Gly-Gly-Ser-Gly-Gly-Ser linker. Strains were grown overnight, then diluted (1:100) and grown 4 hours in fresh M9 medium before induction with IPTG. Cultures were then arrayed in triplicate on 96-well plates and incubated in a fluorimeter at 37°C with shaking. Fluorescence measurements (excitation 488nm, emission 509nm) were taken every 2 minutes.

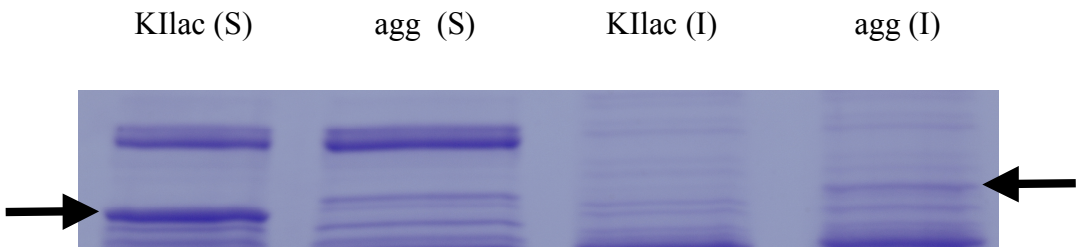


Figure 2-S3. Separation of cell extracts of *KIlac* and *agg* strains into soluble (S) and insoluble (I) fractions. The β-galactosidase band (black arrows) runs higher in the *agg* strain because of the additional size of the tag itself.

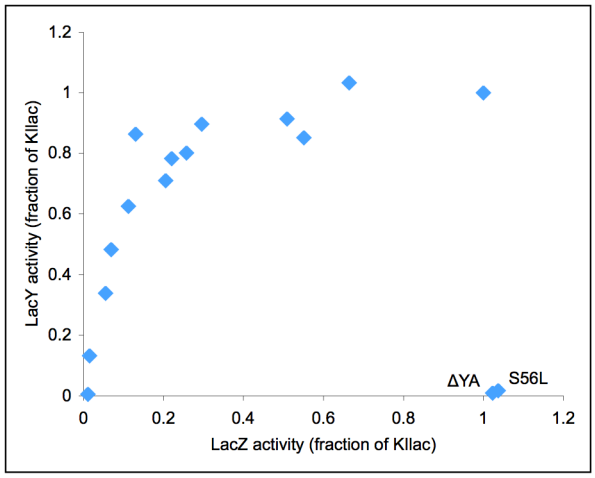


Figure 2-S4. Polarity in the lac operon. Measured LacY activity plotted against LacZ activity. LacZ activity continues to increase after LacY has reached maximal activity levels. This relationship is consistent with other observations of *lac* operon polarity [69, 84].

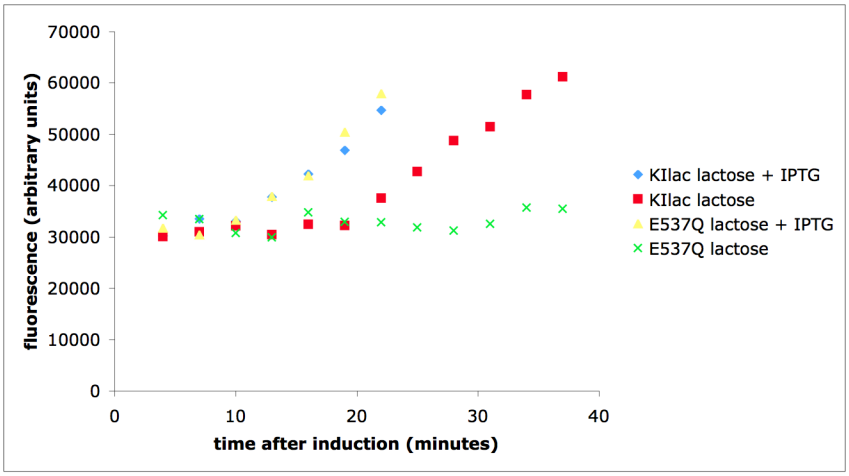


Figure 2-S5. Induction of strains with and without IPTG. The delay in induction for *KIlac* in lactose only conditions is likely attributed to the required conversion of lactose to allolactose.

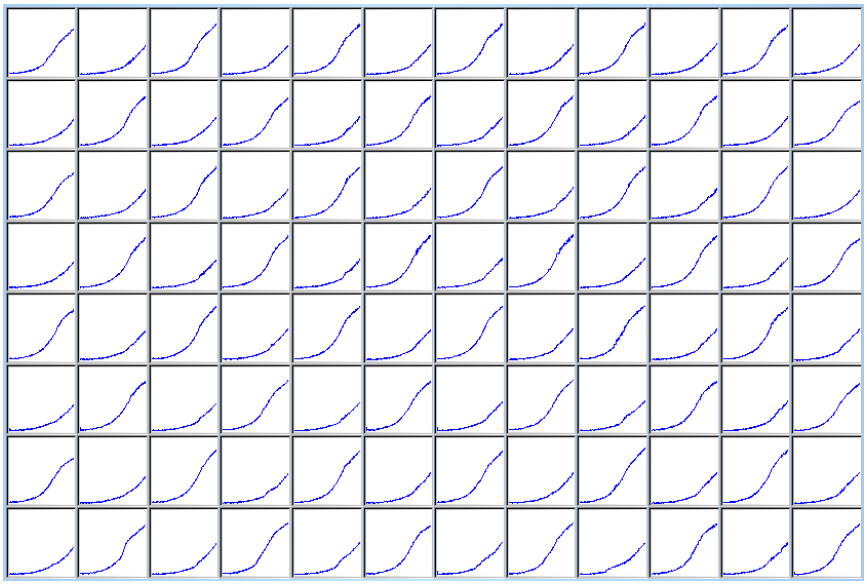


Figure 2-S6. Sample of growth measurements from a 96-well plate. The two growth conditions (induced with IPTG, uninduced) are arrayed in a "checkerboard" fashion.

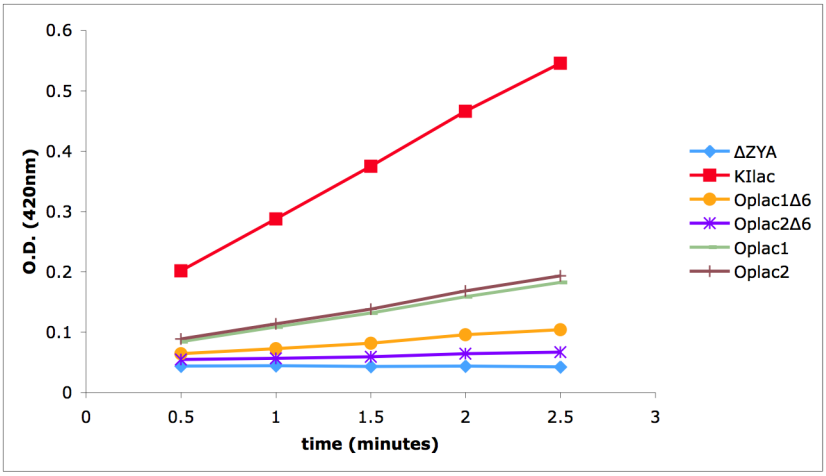


Figure 2-S7. Examples of LacZ activity measurements from different strains.

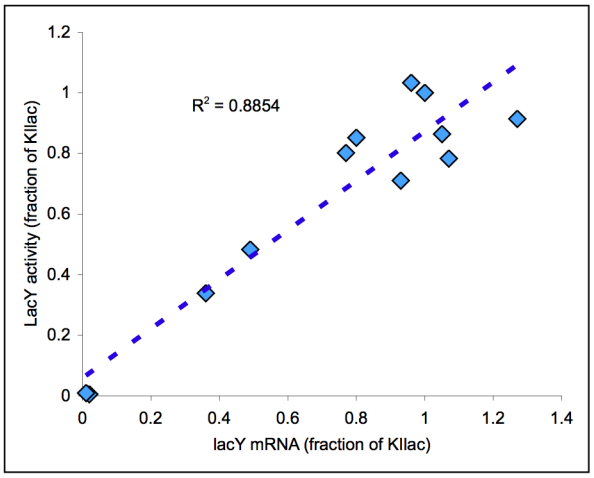


Figure 2-S8. The relationship between LacY activity and *lacY* mRNA levels.

Table 2-S1. Sequences of codon-redesigned strains.

Oplac1:

ATGACCATGATTACCGATAGCCTGGCCGTGGTGCTGCAGCGTCGTGATTGGG

AAAATCCGGGCGTGACGCAGCTGAATCGTCTGGCCGCGCATCCGCCGTTTGC

GAGCTGGCGTAACAGCGAAGAAGCGCGTACCGATCGTCCGAGCCAGCAGCT

GCGTAGCCTGAACGGCGAATGGCGTTTTGCGTGGTTTCCGGCACCGGAGGCG

GTTCCGGAAAGCTGGCTGGAATGCGATCTGCCGGAAGCGGATACCGTGGTGG

TGCCGAGCAACTGGCAGATGCATGGCTATGATGCGCCGATTTATACCAACGT

GACCTATCCGATTACCGTGAATCCGCCGTTCGTGCCGACCGAAAATCCGACC

GGCTGCTATAGCCTGACCTTTAACGTGGATGAATCTTGGCTGCAGGAAGGCC

AGACCCGTATTATTTTTGATGGCGTGAACAGCGCGTTTCATCTGTGGTGCAAT

GGCCGCTGGGTGGGCTATGGCCAGGATAGCCGTCTGCCGAGCGAATTTGATC

TGAGCGCGTTTCTGCGTGCGGGCGAAAACCGTCTGGCCGTGATGGTGCTGCG

TTGGTCTGATGGCAGCTATCTGGAAGATCAGGATATGTGGCGTATGAGCGGC

ATTTTTCGTGATGTGAGCCTGCTGCATAAACCGACCACCCAGATTAGCGATTT

TCATGTGGCGACCCGTTTTAACGATGATTTTAGCCGTGCGGTGCTGGAAGCGG

AAGTGCAGATGTGCGGCGAACTGCGTGATTATCTGCGTGTGACCGTGAGCCT

GTGGCAGGGCGAAACGCAGGTTGCGTCTGGCACCGCGCCGTTTGGCGGCGAA

ATTATTGATGAACGTGGCGGCTATGCGGATCGTGTGACCCTGCGTCTGAACGT

GGAAAATCCGAAACTGTGGAGCGCGGAAATTCCGAACCTGTATCGTGCGGTG

GTGGAACTGCATACCGCGGATGGCACCCTGATTGAAGCGGAAGCGTGTGATG

TGGGCTTTCGTGAAGTGCGTATTGAAAACGGCCTGCTGCTGCTGAACGGCAA

ACCGCTGCTGATTCGTGGCGTGAACCGTCATGAACATCATCCGCTGCATGGC

CAGGTGATGGATGAACAGACCATGGTGCAGGATATTCTGCTGATGAAACAGA

ACAACTTTAACGCGGTGCGTTGCAGCCATTATCCGAACCATCCGCTGTGGTAT

ACCCTGTGCGATCGTTATGGCCTGTATGTGGTGGATGAAGCGAACATTGAAA

CCCATGGCATGGTGCCGATGAACCGTCTGACCGATGATCCGCGTTGGCTGCC

GGCCATGAGCGAACGTGTGACCCGTATGGTGCAGCGTGATCGTAACCATCCG

AGCGTGATTATTTGGAGCCTGGGCAACGAAAGCGGCCATGGCGCGAACCATG

ATGCGCTGTATCGTTGGATTAAAAGCGTGGATCCGAGCCGTCCGGTGCAGTA

TGAAGGCGGCGGTGCGGATACCACCGCGACCGATATTATTTGCCCGATGTAT

GCGCGTGTGGATGAAGATCAGCCGTTTCCGGCGGTGCCGAAATGGTCTATTA

AAAAATGGCTGTCTCTGCCGGGCGAAACGCGTCCGCTGATTCTGTGCGAATA

TGCGCATGCCATGGGCAACAGCCTGGGCGGCTTTGCGAAATATTGGCAGGCG

TTTCGTCAGTATCCGCGTCTGCAGGGCGGCTTTGTGTGGGATTGGGTGGATCA

GAGCCTGATCAAATATGATGAAAACGGCAATCCGTGGAGCGCGTATGGCGGC

GATTTTGGCGATACCCCGAACGATCGTCAGTTTTGCATGAACGGCCTGGTGTT

TGCGGATCGTACCCCGCATCCGGCGCTGACCGAAGCGAAACATCAGCAGCAG

TTTTTTCAGTTTCGTCTGAGCGGCCAGACCATTGAAGTGACCAGCGAATACCT

GTTTCGTCATAGCGATAACGAACTGCTGCATTGGATGGTGGCGCTGGATGGT

AAACCGTTAGCATCTGGCGAAGTGCCGCTGGATGTTGCGCCGCAGGGCAAAC

AGCTGATTGAACTGCCGGAACTGCCGCAGCCGGAAAGCGCGGGTCAGCTGTG

GCTGACCGTGCGTGTGGTGCAGCCGAATGCGACCGCGTGGAGCGAAGCGGG

CCATATTAGCGCGTGGCAGCAGTGGCGTCTGGCCGAAAACCTGAGCGTTACC

CTGCCGGCAGCGAGCCATGCGATTCCGCATCTGACCACCTCTGAAATGGATT

TTTGCATCGAACTGGGCAACAAACGTTGGCAGTTTAACCGTCAGAGCGGCTT

TCTGAGCCAGATGTGGATTGGCGATAAAAAACAGCTGCTGACCCCGCTGCGT

GATCAGTTTACCCGTGCGCCGCTGGATAACGATATTGGCGTGAGCGAAGCGA

CCCGTATTGATCCGAACGCGTGGGTGGAACGTTGGAAAGCGGCCGGTCATTA

TCAGGCGGAAGCGGCGCTGCTGCAGTGTACCGCGGATACCCTGGCCGATGCG

GTGCTGATTACCACCGCCCATGCGTGGCAGCATCAGGGCAAAACGCTGTTTA

TTAGCCGTAAAACCTATCGTATTGATGGCAGCGGCCAGATGGCGATTACCGT

GGATGTGGAAGTGGCGAGCGATACCCCGCACCCGGCACGTATTGGCCTGAAC

TGCCAGCTGGCCCAGGTGGCGGAACGTGTGAACTGGCTGGGCCTGGGCCCGC

AGGAAAACTATCCGGATCGTCTGACCGCGGCGTGCTTTGATCGTTGGGATCT

GCCGCTGTCTGATATGTATACCCCGTATGTGTTTCCGAGCGAAAACGGTCTGC

GTTGCGGCACCCGTGAACTGAACTATGGCCCGCATCAGTGGCGTGGCGATTT

TCAGTTTAACATCAGCCGTTATAGCCAGCAGCAGCTGATGGAAACCAGCCAT

CGTCATCTGCTGCATGCGGAAGAAGGCACCTGGCTGAACATTGATGGCTTTC

ATATGGGCATTGGCGGTGATGATTCTTGGAGCCCGAGCGTGAGCGCGGAATT

TCAGCTGTCTGCGGGCCGTTATCATTATCAGCTGGTGTGGTGCCAGAAATAA

Oplac2:

ATGACCATGATCACCGATTCTCTGGCTGTGGTATTGCAGCGTCGCGACTGGGA

GAACCCTGGTGTCACCCAACTGAATCGTTTGGCAGCGCACCCACCGTTCGCA

TCTTGGCGTAACTCGGAGGAGGCTCGTACCGACCGTCCGTCCCAGCAACTGC

GTAGCCTGAATGGTGAGTGGCGTTTCGCATGGTTTCCAGCACCGGAAGCCGT

TCCGGAATCCTGGCTGGAGTGTGATCTGCCAGAGGCTGACACTGTTGTTGTTC

CGTCCAACTGGCAGATGCACGGTTATGACGCACCGATCTACACGAACGTGAC

GTACCCGATCACCGTGAATCCACCGTTTGTGCCGACCGAGAATCCTACCGGTT

GCTATAGCCTGACCTTCAATGTCGACGAGTCTTGGTTGCAGGAGGGTCAGAC

GCGTATCATCTTTGATGGTGTGAACTCTGCATTCCATCTGTGGTGCAATGGTC

GTTGGGTGGGTTATGGCCAAGACTCTCGCCTGCCTTCGGAGTTCGATCTGTCG

GCTTTCTTGCGTGCAGGTGAAAACCGTCTGGCTGTCATGGTACTGCGCTGGTC

GGATGGTTCGTACTTGGAAGACCAAGATATGTGGCGTATGTCGGGTATCTTTC

GCGACGTTAGCTTGCTGCACAAACCGACTACCCAAATCAGCGATTTCCATGT

CGCGACCCGTTTCAACGATGACTTTAGCCGTGCGGTTCTGGAAGCGGAAGTG

CAAATGTGTGGCGAATTGCGTGATTACTTGCGTGTGACGGTTAGCCTGTGGCA

GGGTGAGACGCAAGTGGCGTCCGGTACTGCACCTTTCGGTGGCGAGATCATC

GACGAACGTGGTGGTTACGCTGATCGCGTGACCCTGCGCTTGAATGTGGAGA

ATCCGAAACTGTGGAGCGCTGAGATTCCGAACTTGTACCGTGCAGTTGTGGA

GCTGCACACCGCAGACGGCACCTTGATTGAGGCTGAAGCCTGCGATGTGGGT

TTTCGCGAAGTCCGTATCGAGAATGGTCTGTTGCTGTTGAACGGTAAACCGCT

GCTGATTCGTGGCGTGAATCGTCACGAGCACCACCCACTGCATGGTCAAGTG

ATGGATGAACAAACGATGGTGCAAGACATTCTGCTGATGAAGCAGAACAACT

TCAACGCAGTTCGCTGTAGCCATTACCCGAATCACCCTCTGTGGTACACTCTG

TGCGACAGATACGGTCTGTACGTTGTCGACGAGGCGAACATTGAAACCCATG

GTATGGTCCCGATGAATCGCCTGACGGACGACCCACGTTGGTTGCCAGCTAT

GAGCGAACGCGTTACCCGTATGGTTCAACGCGACCGTAATCATCCGAGCGTG

ATCATTTGGTCCCTGGGCAACGAGTCGGGTCACGGTGCAAATCACGACGCGT

TGTATCGTTGGATCAAAAGCGTCGATCCATCCCGTCCGGTGCAGTATGAGGG

TGGTGGTGCGGACACCACCGCTACGGACATCATCTGTCCGATGTATGCACGC

GTCGACGAAGATCAACCGTTTCCTGCGGTTCCGAAATGGAGCATCAAGAAAT

GGTTGTCGCTGCCTGGTGAAACCCGTCCACTGATCTTGTGTGAGTACGCACAT

GCTATGGGCAATTCTCTGGGTGGCTTCGCAAAGTACTGGCAGGCGTTTCGCCA

GTATCCACGCTTGCAAGGTGGTTTCGTTTGGGACTGGGTGGACCAATCTTTGA

TCAAGTACGACGAAAACGGCAATCCATGGTCGGCATACGGTGGCGATTTCGG

TGATACCCCAAACGACCGTCAGTTCTGCATGAACGGCTTGGTATTCGCTGACC

GTACCCCTCATCCTGCGTTGACGGAAGCGAAACATCAACAACAGTTCTTCCA

ATTCCGTCTGTCGGGTCAGACCATTGAAGTGACCTCCGAGTACCTGTTCCGTC

ACAGCGACAACGAGTTGCTGCACTGGATGGTCGCGTTGGACGGCAAGCCGTT

GGCGTCGGGTGAGGTGCCGTTGGACGTTGCACCGCAAGGCAAGCAACTGATC

GAACTGCCTGAACTGCCACAACCGGAAAGCGCTGGCCAACTGTGGCTGACCG

TTCGTGTGGTGCAGCCGAATGCAACTGCGTGGAGCGAGGCTGGTCACATTTC

TGCCTGGCAGCAGTGGCGTCTGGCAGAGAACTTGTCCGTTACTCTGCCTGCTG

CGTCGCATGCGATTCCGCATTTGACCACCTCGGAGATGGACTTCTGTATTGAG

CTGGGCAACAAGCGCTGGCAATTCAATCGCCAGAGCGGTTTTCTGTCGCAAA

TGTGGATTGGCGACAAGAAGCAACTGCTGACCCCTTTGCGTGACCAATTCAC

CAGAGCACCGCTGGACAATGACATTGGTGTTAGCGAGGCGACCAGAATTGAC

CCAAATGCGTGGGTGGAACGTTGGAAGGCTGCTGGCCACTATCAAGCGGAAG

CTGCCCTGCTGCAATGTACCGCAGATACGCTGGCTGACGCTGTGCTGATCACC

ACTGCCCACGCGTGGCAGCATCAAGGCAAGACCCTGTTCATCTCGCGTAAGA

CCTATCGTATTGACGGCTCGGGTCAAATGGCCATTACCGTTGACGTGGAAGT

GGCCAGCGATACCCCTCACCCAGCACGCATCGGTCTGAATTGTCAATTGGCT

CAGGTTGCGGAACGTGTGAACTGGTTGGGTCTGGGTCCGCAAGAAAACTATC

CGGACCGCTTGACCGCTGCATGCTTCGACCGTTGGGACCTGCCACTGTCTGAC

ATGTACACGCCATACGTTTTCCCGAGCGAGAATGGTCTGCGCTGCGGTACTCG

CGAACTGAACTACGGTCCACATCAATGGAGAGGCGACTTCCAGTTCAACATC

TCCAGATACAGCCAGCAACAGTTGATGGAGACGAGCCATCGCCATCTGCTGC

ATGCTGAGGAAGGTACCTGGTTGAACATTGACGGCTTCCATATGGGCATTGG

TGGTGACGATAGCTGGAGCCCATCGGTGTCTGCCGAGTTCCAACTGTCTGCTG

GTCGCTACCACTATCAATTGGTTTGGTGCCAGAAGTAA

Strain	lacY mRNA
KIlac	1
ΔZYA	0.02 +/- 0.02
WT	0.80 +/- 0.31
Oplac1	0.77 +/- 0.41
Oplac2	1.07 +/- 0.35
Oplac1Δ6	0.49 +/- 0.21
Oplac2Δ6	0.36 +/- 0.26
Oplac1(37)-KI	1.28 +/- 0.84
Oplac2(37)-KI	0.93 +/- 0.15
agg	1.05 +/- 0.50
deg	0.96 +/- 0.33
ΔYA	0.01 +/- 0.01

Table 2-S2. *lacY* mRNA levels. See Figure 2-S8 for a comparison between *lacY* mRNA levels and LacY activity.

Conclusions

This work represents the opposite ends of the molecular evolution research spectrum. At one end is the bioinformatics analysis of high-throughput experimental data, at the other, experimental quantification of evolutionary models. Surprisingly, the most interesting results from this research had little to with molecular evolution, but rather revealed new insights into a long-studied experimental model system, the *lac* operon.

Regarding the goals of this project from the outset – the localization and quantification of selective pressures – this work can be described as a partial success. Our attempts to quantify the relative importance of three selective pressures were overshadowed by the costs of expressing the *lac* permease. While we were able to determine that functional loss indirectly accounts for these observed costs we could not generalize this result to other systems. Furthermore, the relative size of the permease-based costs prevented us from teasing out subtler effects, including translational costs-per-protein for codon-redesigned sequences and destabilization costs, which may be present at the threshold of experimental noise. It is not clear whether a different experimental system could resolve these questions any better.

For our bioinformatics analysis, we were able to identify structural elements under unique selective pressures, though these differences were small (but significant). It seems likely that the differences would increase with use of a curated structural dataset instead the rough approximations developed through the mapping protocol. However, in light of the aforementioned experimental work, it would seem prudent to raise several caveats regarding the utility of these analyses. Namely, how much insight do we gain knowing small differences in evolutionary rate? Certainly no conservation patterns could have

possibly revealed the interplay and relative importance of the proteins in the *lac* operon. Even our understanding of gross differences in evolutionary rate remains limited; the leading hypothesis relating expression level to evolutionary rate [18] has yet to find experimental validation. When taking in these analyses, we should be mindful of the complexity and multitude of known and unknown factors affecting selection - and of the shortcomings of our one-size-fits-all models.

The initial goals of this work aside, I found this research to be immensely successful. Like all good projects, this work raised more questions than it could answer. Moreover, I believe the experimental design presented in chapter 2 to be a robust approach to validating or discrediting evolutionary models. These types of experiments – simple comparisons of mutant strains representing different models – could be applied to other evolutionary ideas, including theories of duplicate gene evolution, co-evolution, and co-regulation, to name a few. While I believe the discovery of the permease-dependent penalty will attract greater interest, I hope that the experimental design becomes legacy of this work.

References

1. Teichmann, S.A., *The constraints protein-protein interactions place on sequence divergence.* J Mol Biol, 2002. **324**(3): p. 399-407.
2. Nooren, I.M. and J.M. Thornton, *Structural characterisation and functional significance of transient protein-protein interactions.* J Mol Biol, 2003. **325**(5): p. 991-1018.
3. Caffrey, D.R., et al., *Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?* Protein Sci, 2004. **13**(1): p. 190-202.
4. Mintseris, J. and Z. Weng, *Structure, function, and evolution of transient and obligate protein-protein interactions.* Proc Natl Acad Sci U S A, 2005. **102**(31): p. 10930-5.
5. Hu, Z., et al., *Conservation of polar residues as hot spots at protein interfaces.* Proteins, 2000. **39**(4): p. 331-42.
6. Kimura, M., *Evolutionary rate at the molecular level.* Nature, 1968. **217**(5129): p. 624-6.
7. King, J.L. and T.H. Jukes, *Non-Darwinian evolution.* Science, 1969. **164**(881): p. 788-98.
8. Ohta, T., *Slightly deleterious mutant substitutions in evolution.* Nature, 1973. **246**(5428): p. 96-8.
9. Fraser, H.B., et al., *Evolutionary rate in the protein interaction network.* Science, 2002. **296**(5568): p. 750-2.

10. Bloom, J.D. and C. Adami, *Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets.* BMC Evol Biol, 2003. **3**: p. 21.

11. Fraser, H.B., D.P. Wall, and A.E. Hirsh, *A simple dependence between protein evolution rate and the number of protein-protein interactions.* BMC Evol Biol, 2003. **3**: p. 11.

12. Bloom, J.D. and C. Adami, *Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: response.* BMC Evol Biol, 2004. **4**: p. 14.

13. Kim, P.M., et al., *Relating three-dimensional structures to protein networks provides evolutionary insights.* Science, 2006. **314**(5807): p. 1938-41.

14. Pal, C., B. Papp, and L.D. Hurst, *Highly expressed genes in yeast evolve slowly.* Genetics, 2001. **158**(2): p. 927-31.

15. Drummond, D.A., A. Raval, and C.O. Wilke, *A single determinant dominates the rate of yeast protein evolution.* Mol Biol Evol, 2006. **23**(2): p. 327-37.

16. Lin, Y.S., et al., *Proportion of solvent-exposed amino acids in a protein and rate of protein evolution.* Mol Biol Evol, 2007. **24**(4): p. 1005-11.

17. Choi, S.S., E.J. Vallender, and B.T. Lahn, *Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes.* Mol Biol Evol, 2006. **23**(11): p. 2131-3.

18. Drummond, D.A., et al., *Why highly expressed proteins evolve slowly.* Proc Natl Acad Sci U S A, 2005. **102**(40): p. 14338-43.

19. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

20. Rajamani, D., et al., *Anchor residues in protein-protein interactions.* Proc Natl Acad Sci U S A, 2004. **101**(31): p. 11287-92.

21. Finn, R.D., et al., *Pfam: clans, web tools and services.* Nucleic Acids Res, 2006. **34**(Database issue): p. D247-51.

22. Stein, A., R.B. Russell, and P. Aloy, *3did: interacting protein domains of known three-dimensional structure.* Nucleic Acids Res, 2005. **33**(Database issue): p. D413-7.

23. Bloom, J.D., et al., *Structural determinants of the rate of protein evolution in yeast.* Mol Biol Evol, 2006. **23**(9): p. 1751-61.

24. Greenbaum, D., et al., *Comparing protein abundance and mRNA expression levels on a genomic scale.* Genome Biol, 2003. **4**(9): p. 117.

25. Pal, C., B. Papp, and M.J. Lercher, *An integrated view of protein evolution.* Nat Rev Genet, 2006. **7**(5): p. 337-48.

26. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.

27. Aloy, P., et al., *The relationship between sequence and interaction divergence in proteins.* J Mol Biol, 2003. **332**(5): p. 989-98.

28. Yu, H., et al., *Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.* Genome Res, 2004. **14**(6): p. 1107-18.

29. Holstege, F.C., et al., *Dissecting the regulatory circuitry of a eukaryotic genome.* Cell, 1998. **95**(5): p. 717-28.

30. Ghaemmaghami, S., et al., *Global analysis of protein expression in yeast.* Nature, 2003. **425**(6959): p. 737-41.

31. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

32. Yang, Z., *PAML: a program package for phylogenetic analysis by maximum likelihood.* Comput Appl Biosci, 1997. **13**(5): p. 555-6.

33. Ihaka, R. and R. Gentleman, *R: A Language for Data Analysis and Graphics.* Journal of Computational and Graphical Statistics, 1996. **5**(3): p. 299-314.

34. Smith, T.F., et al., *The WD repeat: a common architecture for diverse functions.* Trends Biochem Sci, 1999. **24**(5): p. 181-5.

35. D'Andrea, L.D. and L. Regan, *TPR proteins: the versatile helix.* Trends Biochem Sci, 2003. **28**(12): p. 655-62.

36. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* J Mol Biol, 1961. **3**: p. 318-56.

37. Santillan, M. and M.C. Mackey, *Quantitative approaches to the study of bistability in the lac operon of Escherichia coli.* J R Soc Interface, 2008. **5 Suppl 1**: p. S29-39.

38. Wilson, C.J., et al., *The lactose repressor system: paradigms for regulation, allosteric behavior and protein folding.* Cell Mol Life Sci, 2007. **64**(1): p. 3-16.

39. Santillan, M., M.C. Mackey, and E.S. Zeron, *Origin of bistability in the lac Operon.* Biophys J, 2007. **92**(11): p. 3830-42.

40. Jacob, F., et al., *[Operon: a group of genes with the expression coordinated by an operator.].* C R Hebd Seances Acad Sci, 1960. **250**: p. 1727-9.

41. Monod, J., J.P. Changeux, and F. Jacob, *Allosteric proteins and cellular control systems.* J Mol Biol, 1963. **6**: p. 306-29.

42. Serebriiskii, I.G. and E.A. Golemis, *Uses of lacZ to study gene function: evaluation of beta-galactosidase assays employed in the yeast two-hybrid system.* Anal Biochem, 2000. **285**(1): p. 1-15.

43. Abramson, J., et al., *Structure and mechanism of the lactose permease of Escherichia coli.* Science, 2003. **301**(5633): p. 610-5.

44. Dekel, E. and U. Alon, *Optimality and evolutionary tuning of the expression level of a protein.* Nature, 2005. **436**(7050): p. 588-92.

45. Klumpp, S., Z. Zhang, and T. Hwa, *Growth rate-dependent global effects on gene expression in bacteria.* Cell, 2009. **139**(7): p. 1366-75.

46. Stoebel, D.M., A.M. Dean, and D.E. Dykhuizen, *The cost of expression of Escherichia coli lac operon proteins is in the process, not in the products.* Genetics, 2008. **178**(3): p. 1653-60.

47. Koonin, E.V. and Y.I. Wolf, *Evolutionary systems biology: links between gene evolution and function.* Curr Opin Biotechnol, 2006. **17**(5): p. 481-7.

48. Pal, C., B. Papp, and L.D. Hurst, *Genomic function: Rate of evolution and gene dispensability.* Nature, 2003. **421**(6922): p. 496-7; discussion 497-8.

49. Herbeck, J.T. and D.P. Wall, *Converging on a general model of protein evolution.* Trends Biotechnol, 2005. **23**(10): p. 485-7.

50. Rocha, E.P. and A. Danchin, *An analysis of determinants of amino acids substitution rates in bacterial proteins.* Mol Biol Evol, 2004. **21**(1): p. 108-16.

51. Drummond, D.A. and C.O. Wilke, *Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution.* Cell, 2008. **134**(2): p. 341-52.

52. Novick, A. and M. Weiner, *Enzyme Induction as an All-or-None Phenomenon.* Proc Natl Acad Sci U S A, 1957. **43**(7): p. 553-66.

53. Thanaraj, T.A. and P. Argos, *Ribosome-mediated translational pause and protein domain organization.* Protein Sci, 1996. **5**(8): p. 1594-612.

54. Welch, M., et al., *Design parameters to control synthetic gene expression in Escherichia coli.* PLoS One, 2009. **4**(9): p. e7002.

55. Kudla, G., et al., *Coding-sequence determinants of gene expression in Escherichia coli.* Science, 2009. **324**(5924): p. 255-8.

56. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Res, 2003. **31**(13): p. 3406-15.

57. Yuan, J., M. Martinez-Bilbao, and R.E. Huber, *Substitutions for Glu-537 of beta-galactosidase from Escherichia coli cause large decreases in catalytic activity.* Biochem J, 1994. **299 (Pt 2)**: p. 527-31.

58. Martinez-Bilbao, M., et al., *A highly reactive beta-galactosidase (Escherichia coli) resulting from a substitution of an aspartic acid for Gly-794.* J Biol Chem, 1991. **266**(8): p. 4979-86.

59. Huber, R.E., et al., *Trp-999 of beta-galactosidase (Escherichia coli) is a key residue for binding, catalysis, and synthesis of allolactose, the natural lac operon inducer.* Biochemistry, 2003. **42**(6): p. 1796-803.

60. Juers, D.H., et al., *Structural basis for the altered activity of Gly794 variants of Escherichia coli beta-galactosidase.* Biochemistry, 2003. **42**(46): p. 13505-11.

61. Cupples, C.G., J.H. Miller, and R.E. Huber, *Determination of the roles of Glu-461 in beta-galactosidase (Escherichia coli) using site-specific mutagenesis.* J Biol Chem, 1990. **265**(10): p. 5512-8.

62. Kortemme, T. and D. Baker, *A simple physical model for binding energy hot spots in protein-protein complexes.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14116-21.

63. Gottesman, S., et al., *The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system.* Genes Dev, 1998. **12**(9): p. 1338-47.

64. Lee, S.C., Y.C. Choi, and M.H. Yu, *Effect of the N-terminal hydrophobic sequence of hepatitis B virus surface antigen on the folding and assembly of hybrid beta-galactosidase in Escherichia coli.* Eur J Biochem, 1990. **187**(2): p. 417-24.

65. Vasina, J.A. and F. Baneyx, *Expression of aggregation-prone recombinant proteins at low temperatures: a comparative study of the Escherichia coli cspA and tac promoter systems.* Protein Expr Purif, 1997. **9**(2): p. 211-8.

66. Dykhuizen, D. and D. Hartl, *Transport by the lactose permease of Escherichia coli as the basis of lactose killing.* J Bacteriol, 1978. **135**(3): p. 876-82.

67. Green, A.L., E.J. Anderson, and R.J. Brooker, *A revised model for the structure and function of the lactose permease. Evidence that a face on transmembrane*

segment 2 is important for conformational changes. J Biol Chem, 2000. **275**(30): p. 23240-6.

68. Hansen, L.H., S. Knudsen, and S.J. Sorensen, *The effect of the lacY gene on the induction of IPTG inducible promoters, studied in Escherichia coli and Pseudomonas fluorescens.* Curr Microbiol, 1998. **36**(6): p. 341-7.

69. Yarchuk, O., et al., *Interdependence of translation, transcription and mRNA degradation in the lacZ gene.* J Mol Biol, 1992. **226**(3): p. 581-96.

70. Jobe, A. and S. Bourgeois, *lac Repressor-operator interaction. VI. The natural inducer of the lac operon.* J Mol Biol, 1972. **69**(3): p. 397-408.

71. Ottina, K., J. Lopilato, and T.H. Wilson, *Membrane transport of p-nitrophenyl-alpha-galactoside by the melibiose carrier of Escherichia coli.* J Membr Biol, 1980. **56**(2): p. 169-75.

72. Putzrath, R.M. and T.H. Wilson, *Transport of alpha-p-nitrophenylgalactoside by the lactose carrier of Escherichia coli.* J Bacteriol, 1979. **137**(2): p. 1037-9.

73. Hogema, B.M., et al., *Inducer exclusion by glucose 6-phosphate in Escherichia coli.* Mol Microbiol, 1998. **28**(4): p. 755-65.

74. Price, M.N., A.P. Arkin, and E.J. Alm, *The life-cycle of operons.* PLoS Genet, 2006. **2**(6): p. e96.

75. Osterberg, M., et al., *Phenotypic effects of membrane protein overexpression in Saccharomyces cerevisiae.* Proc Natl Acad Sci U S A, 2006. **103**(30): p. 11148-53.

76. Sopko, R., et al., *Mapping pathways and phenotypes by systematic gene overexpression.* Mol Cell, 2006. **21**(3): p. 319-30.

77. Palmer, A.C., E. Angelino, and R. Kishony, *Chemical decay of an antibiotic inverts selection for resistance.* Nat Chem Biol, 2010. **6**(2): p. 105-7.

78. Isalan, M., et al., *Evolvability and hierarchy in rewired bacterial gene networks.* Nature, 2008. **452**(7189): p. 840-5.

79. Overbeek, R., et al., *The use of gene clusters to infer functional coupling.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2896-901.

80. Notebaart, R.A., et al., *Co-regulation of metabolic genes is better explained by flux coupling than by network distance.* PLoS Comput Biol, 2008. **4**(1): p. e26.

81. Datsenko, K.A. and B.L. Wanner, *One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products.* Proc Natl Acad Sci U S A, 2000. **97**(12): p. 6640-5.

82. McKenzie, G.J. and N.L. Craig, *Fast, easy and efficient: site-specific insertion of transgenes into enterobacterial chromosomes using Tn7 without need for selection of the insertion event.* BMC Microbiol, 2006. **6**: p. 39.

83. Zhang, Y., et al., *Expression of eukaryotic proteins in soluble form in Escherichia coli.* Protein Expr Purif, 1998. **12**(2): p. 159-65.

84. Dryselius, R., et al., *Variable coordination of cotranscribed genes in Escherichia coli following antisense repression.* BMC Microbiol, 2006. **6**: p. 97.

85. Tagourti, J., et al., *Membrane docking of an aggregation-prone protein improves its solubilization.* Gene, 2008. **426**(1-2): p. 32-8.

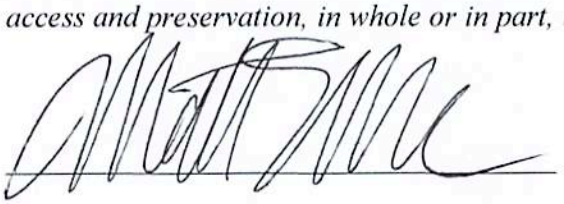
86. Cormack, B.P., R.H. Valdivia, and S. Falkow, *FACS-optimized mutants of the green fluorescent protein (GFP).* Gene, 1996. **173**(1 Spec No): p. 33-8.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

5/11/10

Date