EXPLORING THE FITNESS LANDSCAPE OF TEM-1 BETA-LACTAMASE:

A SURVEY OF INTRAGENIC EPISTASIS & THE FITNESS EFFECTS OF INDELS

by

Courtney E. Gonzalez

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

October 2018

Abstract

The distribution of fitness effects, or fitness landscape, of a protein offers a picture of the relationship between mutations and their effects on a broad scale. Mutations in a protein can have positive, negative, or neutral effects on its function. Mapping these effects on a large scale allows us to better understand evolutionary dynamics in nature and informs our ability to better engineer proteins in the laboratory. I used deep-mutational scanning techniques, including saturation mutagenesis, high-throughput selection, and DNA deep-sequencing, to explore two important and understudied aspects of the fitness landscape of TEM-1 beta-lactamase. First, I examined pairwise intragenic epistasis among sequential amino acid substitutions in TEM-1. Epistasis, or interactions between mutations, play a central role in shaping the fitness landscape, but a clear picture of the prevalence and patterns of epistasis has yet to emerge. This study is the first to systematically examine pairwise epistasis throughout an entire protein performing its native function in its native host. I explored the relationship between epistasis and secondary structure, solvent accessibility, distance from the active site, amino acid identity, and individual mutant effect. I found pervasive negative epistasis, particularly in highly structured regions of the protein and among buried residues, and a high frequency of negative sign epistasis among individually beneficial mutations. Second, I present a near-comprehensive analysis of the fitness effects of single amino acid insertions and deletions (InDels) in TEM-1. Short InDels are a common type of mutation in nature, often having important consequences, such as opening new pathways for adaptation. InDels also represent a useful source of variation in the protein

engineering toolbox. Despite their importance and utility, the distribution of fitness effects of InDels is vastly understudied compared to substitutions. I found InDels to be largely deleterious, but notable regions of tolerance were observed throughout the protein. I found secondary structure, weighted contact number, and evolutionary variation in class A beta-lactamases to be the most predictive of their fitness effects.

Advisor: Dr. Marc Ostermeier

Professor of Chemical and Biomolecular Engineering

Johns Hopkins University

Readers: Dr. Jeffrey Gray

Professor of Chemical and Biomolecular Engineering

Johns Hopkins University

Dr. Robert Schleif

Professor of Biology

Johns Hopkins University

Acknowledgments

I first want to acknowledge my advisor, Dr. Marc Ostermeier, for his mentorship throughout my graduate school career. Marc, thank you for supporting my growth as a researcher and encouraging me to follow my interests. Your constructive feedback and confidence in me have made me a better scientist. I would also like to thank Dr. Jeff Gray for his mentorship in teaching. The freedom and responsibility I was given as a teaching assistant for your class helped me discover my strengths and inspired me to pursue all the teaching opportunities I could during my time at Hopkins.

I also want to thank the members of the Ostermeier Lab, past and present. I cannot imagine a better group of scientists and friends. Thank you for your collaboration, valuable conversations, and shared knowledge. I especially want to thank Barrett Steinberg, Nirav Shelat, Dillon Nye (honorary lab member), Tina Xiong, and Tiana Warren for their close friendship. Thank you for the shared music and books, the late night Pictionary games, the hikes, the rock climbing, the heart-to-heart conversations, and the moral support. I could not have made it through without you.

Above all, I am profoundly grateful for my family. I would not be here without their love and unfailing support. To my grandparents, Elaine Creighton-Baca and Judge José Baca, thank you for instilling in me the value of education. To my brothers, Tim and Alex, thank you for the daily texts, the emotional support, and for literally driving across the country when I needed help. And to my parents, Sue and Joe, thank you for providing every opportunity for my education and encouraging me each step along the way. There are not words enough to thank you for everything you have done for me. I am forever grateful for your unconditional love and support.

Finally, although she will never be able to read this, I have to acknowledge my closest companion throughout graduate school: my dog, George Eliot. I know it may be unconventional to include a canine here, but where I come from, dogs are people too, and this one has been my best and most loyal friend for the past seven years.

I am deeply grateful for everyone who supported me along this journey.

Table of Contents

List of Figures

Chapter 1: Introduction and Background

Mutations and the Fitness Landscape

Mutations are the source of genetic variation in evolution. Typically, DNA sequences are copied with very high fidelity, but rare errors in replication or repair in protein-coding regions can result in various changes on the amino acid level, including substitutions, insertions, and deletions. These changes can have positive, negative, or neutral effects on the function and expression of that protein, resulting in different phenotypes. Under selection pressure, these differences can be determining factors in what genes become fixed in a population [1].

Evolutionary biologists use the term *fitness* to describe the ability of an organism to survive and procreate, and thus predict evolutionary success. Understanding the relationship between genetic variation and fitness is a fundamental objective in biology. In 1932, Sewall Wright introduced the concept of the *fitness landscape* as a way to visualize this relationship [2]. The metaphor of the fitness landscape imagines a topographical map in which genotype space is represented on the x-y plane, and fitness is mapped onto the z-axis (Figure 1). The result is a landscape of high fitness peaks and low fitness valleys corresponding to different genotypes. This 3-D representation is a highly simplistic representation of a true fitness landscape, given the vast multi-dimensionality of all possible genotypes. For example, the sequence space of all combinations of possible point mutations a single 1000 base pair gene is 4^{1000}, a number greater than the total number of particles in the universe [2].

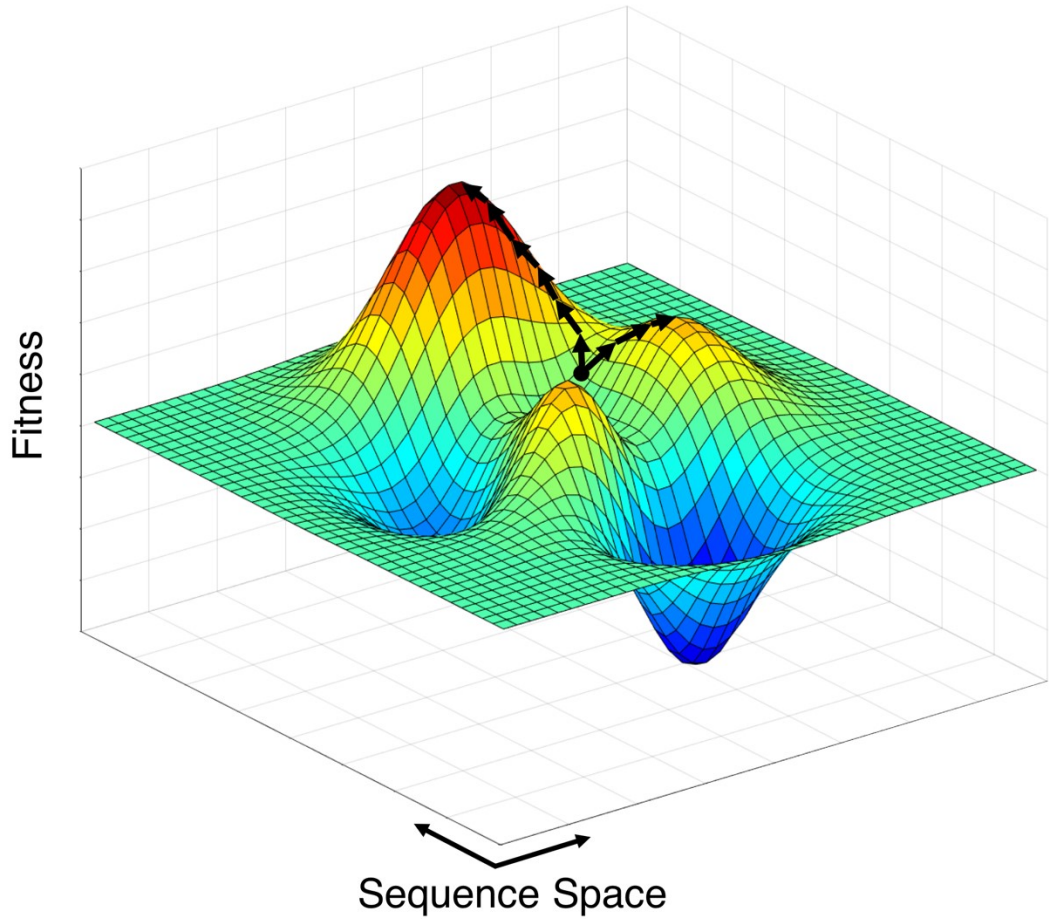

Figure 1. The Metaphor of the Fitness Landscape
The x-y plane represents possible genetic sequences and the z-axis represents the fitness conferred by that genotype. Fitness peaks are indicated in warmer colors and fitness valleys are indicated in cooler colors. Arrows on the landscape show potential adaptive walks starting from a specific point, ending at local or global maxima.

Still, the metaphor of a landscape with fitness peaks and valleys provides a useful image to conceptualize evolutionary trajectories and adaptation. In 1970, John Maynard Smith proposed the concept of an "adaptive walk" to imagine how proteins may traverse the landscape via stepwise mutational changes [3]. He uses the analogy of a word game in which one word is changed to another word, one letter at a time, provided each

intermediate is also a word (WORD --> WORE --> GORE --> GONE --> GENE). Analogously, a protein variant can accumulate variation and traverse the landscape one mutation at a time, provided each subsequent mutation results in a functional protein. Accumulation of mutations toward a fitness peak can be pictured as an uphill climb, where each subsequent mutation is one step away in sequence space and results in a higher fitness (Figure 1). One might imagine a relatively smooth landscape with a single peak, or a more rugged landscape with multiple peaks and valleys, and the ways in which these different topographies could influence the outcomes of adaptive walks. In theory, the fitness landscape shows which outcomes are fundamentally possible in evolution.

Epistasis and the Fitness Landscape

The structure of the fitness landscape is complicated by *epistasis*, or genetic interactions that result in a deviation from the additive effects of mutations[4]. In other words, epistasis occurs when the effect of a mutation is different depending on the genetic background, or context, in which it occurs. In general, epistasis can be categorized as *magnitude epistasis* or *sign epistasis*. Magnitude epistasis describes when the size of the effect varies depending on the context. This type of epistasis can affect the curvature of the fitness landscape, but does not introduce ruggedness which can constrain evolutionary trajectories. Magnitude epistasis can be positive or negative. Positive epistasis occurs when the fitness conferred by two or more mutations is higher (more beneficial) than predicted based on their individual effects; negative epistasis occurs when the combined fitness effect is lower (more detrimental) than predicted (Figure 2). Sign epistasis describes when the sign (beneficial or detrimental) of the effect of the

mutation changes depending on the genetic background. For example, negative sign epistasis occurs if mutation A is beneficial alone, but the combined fitness effect of mutation A and B is detrimental. This type of epistasis can introduce peaks and valleys in the fitness landscape that render certain pathways inaccessible.

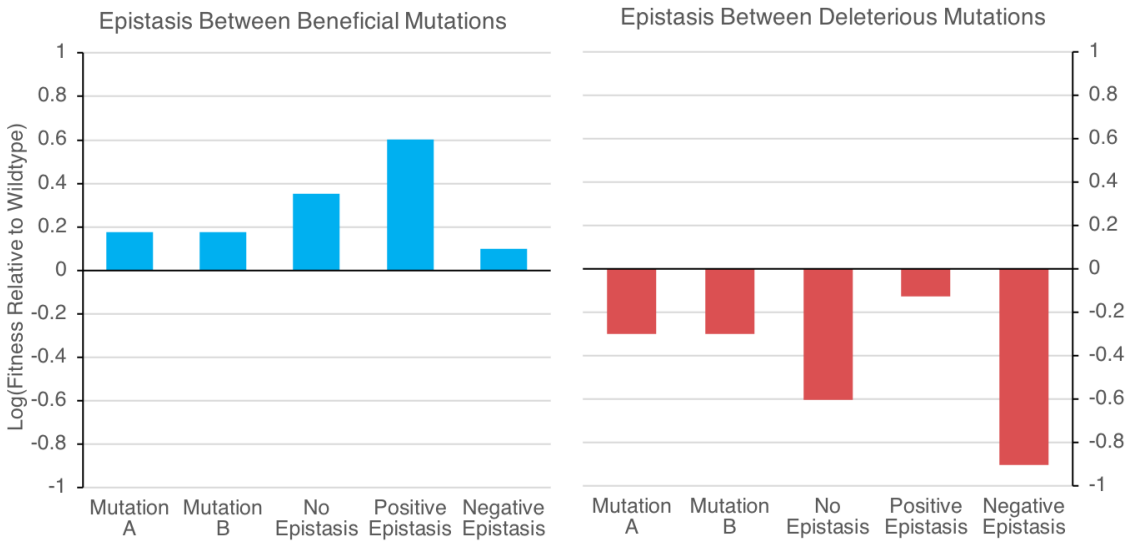

Figure 2. Magnitude epistasis between two beneficial or two deleterious mutations

Fitness Landscape Models

Numerous models have been proposed to predict the shape of the fitness landscape. Among the most popular is the NK model [5], which aims to capture pervasive sign epistasis. In this model, the shape of the landscape is tuned by the size of the genome, represented as a binary sequence of length *N*, and the number of "interaction partners" (K). Under different parameters, the NK model can result in a smooth landscape with a single peak, a maximally rugged landscape, or something in between.

When K=0, the model results in a smooth mountain-like fitness landscape. This limit is termed the “Mt. Fuji” landscape and depicts the shape of a landscape without epistasis [6]. The other extreme, where K=N-1, results in a maximally rugged landscape, termed a “House of Cards” landscape [7]. The “Rough Mount Fuji” landscape aims to capture an intermediate ruggedness, thought to be more representative of true fitness landscapes [6, 8]. While theoretical models of fitness landscapes are useful in providing a framework for what fitness landscapes may look like, they stand to benefit from empirical observations of real fitness landscapes.

Empirical Fitness Landscapes

Recently, advances in molecular biology and DNA sequencing have allowed for empirical studies of fitness landscapes. One way to study this relationship between genotype and fitness is with deep mutational scanning, a method that involves the creation of a large number of mutants, selection based on their function or fitness, and high-throughput DNA sequencing to link effects to their respective genotype. Deep mutation scanning is motivated by the question, as phrased by Fowler et al: “…what if we knew the functional consequences of every possible single amino acid change at every position in a protein?”[9] In these studies, which typically focus on a single protein, fitness is often more specifically referred to as “protein fitness” or “gene fitness”, to distinguish it from the true biological fitness of the organism, which encompasses everything involved in survival and reproduction. The link between the functional effects of mutations and their effect on organismal fitness is not completely understood, and the systems involved in deep mutational scanning studies are sometimes far removed from

biological conditions [10]. Still, these high-throughput mutational studies have the potential to advance our understanding of proteins by revealing important and often unpredictable results. They can identify thermodynamically stabilizing mutations, mutations that result in enhanced catalytic activity or improved binding, or mutations or residues that are important for structure. For example, mutations far away from the active site of the protein have been found to drastically affect enzyme activity and thermodynamic stability [11]. Together, these studies have also offered unexpected insight into the general impact of single amino acid substitutions. An analysis of 14 such studies comprising over 30,000 mutations revealed methionine to be the most tolerated amino acid substitution, while histidine and asparagine best predicted the effects of other substitutions [12].

Large scale mutational studies also allow us to look at the epistatic effects of multiple mutations on a much larger scale than ever before possible. For example, a 2016 study on the local fitness landscape of the green fluorescent protein examined over 50,000 variants containing two or more mutations, and revealed patterns of epistatic interaction including up to 30% negative epistasis (depending on the number of mutations), a low frequency of positive epistasis, and a correlation between epistasis and functional sites, solvent accessibility, and mutational proximity[13].

The sequence-function or fitness maps generated by deep mutational scanning offer insights into the nature of fitness landscapes and the patterns therein. Empirical studies also allow us to build models with parameters based on observation, in order to better understand and predict evolution. In the following studies (Chapters 2 and 3), we

utilize deep mutational scanning principles to characterize aspects of the fitness landscape of TEM-1 β-lactamase.

TEM-1 β-lactamase

TEM-1 β-lactamase is a commonly studied model in protein evolution. β-lactam antibiotics, such as ampicillin, kill bacteria by binding to the transpepdidases that catalyze cross linking of the cell wall. β-lactamases are enzymes native to bacteria that provide resistance to β-lactam antibiotics. They do so by breaking the four atom β-lactam ring found in these antibiotics, rendering them inactive. There are 4 classes of β-lactamases (A-D). Class B β-lactamases are zinc-dependent metallo-B lactamases, while the other three are characterized by a serine active site. TEM-1 is a class A β-lactamase and the most common β-lactamase found in gram negative bacteria[14]. It has been extensively studied and is a convenient model protein for molecular evolution studies because cells containing TEM-1 can be challenged to grow in the presence of ampicillin and resistance can be used as a proxy for fitness.

A number of studies have focused on the distribution of fitness effects and epistatic interactions of substitutions in TEM-1. In a landmark study on epistasis, Bershtein et al hypothesized a “threshold robustness” to deleterious mutations that destabilize the protein [15]. They found that under low selection pressure, a large fraction of mutations was initially tolerated, but after this threshold was exhausted, the fitness resulting from accumulating mutations fit an accelerated fitness decline curve indicative of negative epistasis. They theorize that initial mutations result in some stability cost on a

physico-chemical scale, but these effects are buffered until they exhaust the threshold, after which deleterious mutations result in a more-than-additive negative effect on fitness.

Three large scale distributions of fitness effects of substitutions have been reported for TEM-1. A 2013 study assessed the fitness effects of 64% of possible amino acid substitutions reachable by point mutation [16]. They used a minimum inhibitory concentration (MIC) assay to determine the fitness effects of mutations. The distribution was bimodal, with a peak around wildtype fitness levels and a peak for inactivating mutations. In the background of a known stabilizing mutation, the distribution shifted toward mutants showing no fitness effect, indicating positive epistasis between the stabilizing mutation and individually deleterious mutations. The effects on ampicillin resistance and cefotaxime resistance of all 4,997 single amino acid mutations in the mature TEM-1 protein were examined by Stiffler et al [17]. They also found a bimodal distribution for fitness effects with respect to ampicillin resistance. Interestingly, they found that mutational tolerance was dependent on the concentration of antibiotic, and that mutations that result in a new function (resistance to cefotaxime) are neutral only under low selection pressure. This suggests that adaptive mutations to new functions may not be able to accumulate in environments under strong purifying selection pressure.

A comprehensive, high-resolution map of nearly every possible amino acid substitutions (95.6%) throughout the entire protein, including the signal sequence, was presented by Firnberg, et al [18]. In agreement with the other studies, they found TEM-1 to be fairly robust to substitution mutations (Figure 3), with 53.2% of alleles maintaining at least half of the fitness of wildtype, as measured by resistance to ampicillin. They also found a small fraction (7.0%) of substitutions that conferred a fitness benefit above

wildtype. Fitness effects of synonymous mutations were marginal compared to missense mutations, and occurred most significantly at the beginning of the gene. They explore the underlying mechanisms of deleterious effect mutations and find reduced specific protein activity to be more determining than reduced protein abundance. In line with the threshold robustness theory, they hypothesize that TEM-1's high tolerance to mutation may result in part from a buffering effect with respect to cellular protein levels.

Following the study by Firnberg et al, Steinberg et al examined the fitness landscapes of TEM-17, TEM-19, and TEM-15 [19]. TEM-17 and TEM-19 are each one single amino acid mutation away from TEM-1. Together in TEM-15, the two mutations confer resistance to cefotaxime rather than ampicillin. Thus, the landscapes represent an adaptive pathway for the evolution of cefotaxime resistance. They examined epistasis along this adaptive pathway and found that the prevalence of epistasis depended on the background mutation. Epistasis was observed in 8% of mutations with TEM-17 and 53% with TEM-19. They found the epistatic landscape of TEM-19 best predicted the final TEM-15 epistatic landscape.

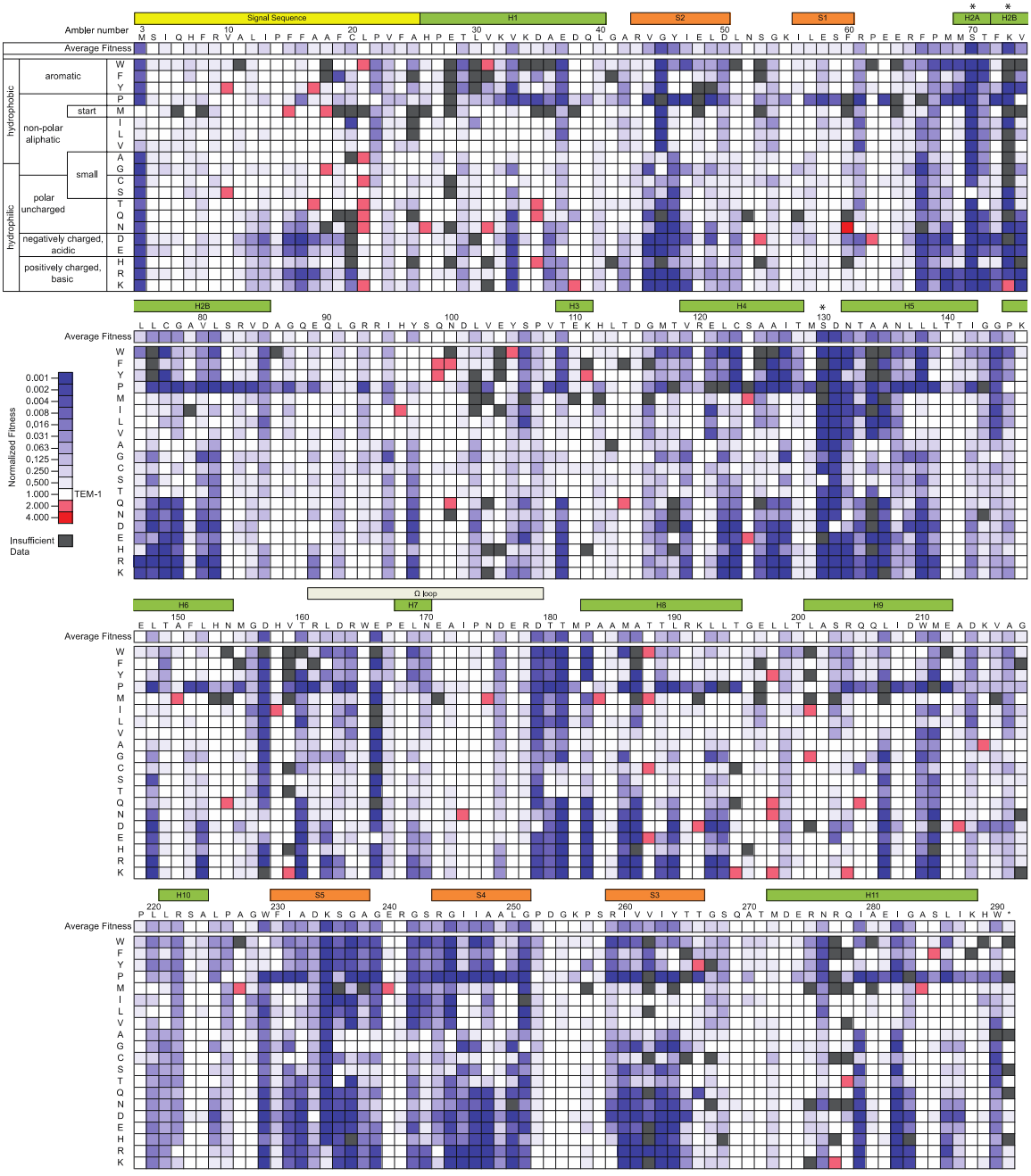

Figure 3. The sequence-function landscape of substitutions in TEM-1 [18]

Library Creation Methods

Each of these studies [18, 19], and the work presented in the following chapters, involved the creation of saturation mutagenesis libraries, high-throughput selection for protein fitness, and deep-sequencing via next-generation sequencing technology. Firnberg et al and Steinberg et al created their libraries with a novel oligonucleotide-directed mutagenesis technique called PFunkel mutagenesis [20]. PFunkel allows for the creation of saturation mutagenesis libraries in a single tube. Based on Kunkle mutagenesis [21], it relies on uracil-containing DNA and PCR cycling with kinased mutagenic oligos. Though time-efficient and convenient, PFunkel has the limitation of having the frequency of library members dependent on how well the mutation-containing oligonucleotide works in the mutation generating reaction. This aspect biases the frequency of library members, making some mutations harder to study because they are not present or are scarce in the library.

Another convenient method for creating high-throughput site-directed mutations in a 96-well format is inverse PCR (Figure 4). It is particularly attractive because individual reactions can be monitored, resulting in a less biased library once the reactions are pooled. In this method, pairs of oligos are designed to linearize the plasmid at each desired site. The forward oligo can be designed to create a substitution, insertion, or deletion of one or multiple nucleotides (Figure 4b). The individual reactions can be visualized on a gel to verify their success, and pooled to create a library. The method's drawbacks are that it requires a separate PCR reaction for each codon mutagenized (i.e. it is more labor intensive) and it potentially has a higher spurious mutation rate, since the

method uses multiple cycles of PCR instead of the single extension reaction that is used in PFunkel.

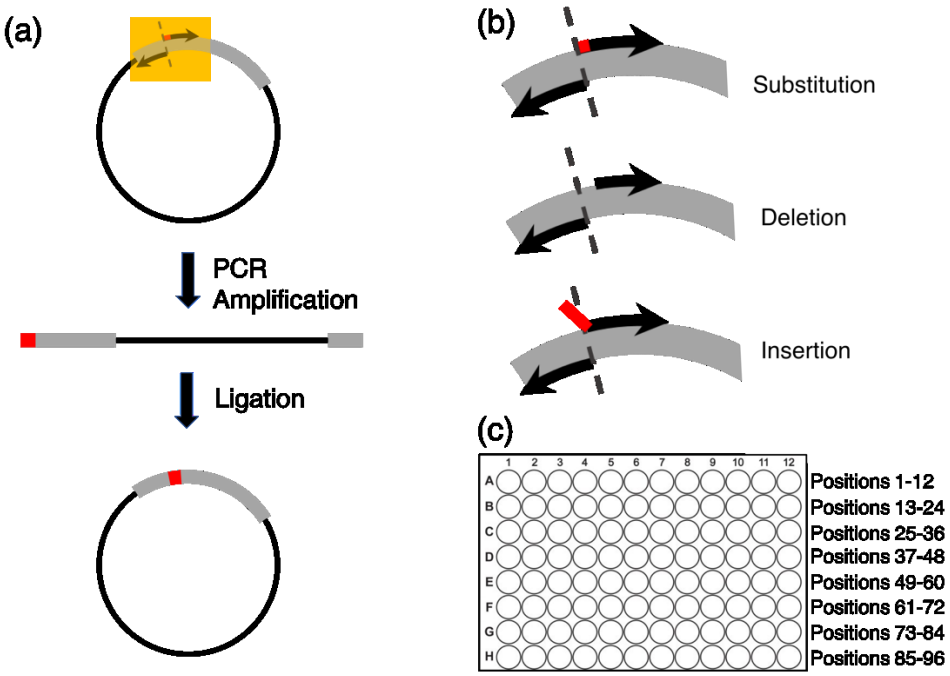

Figure 4. Schematic of Inverse PCR
(a) At each desired site, a PCR reaction with mutagenic primers is performed to linearize the plasmid and introduce the mutation. After ligation, the recircularized plasmid contains the mutation. (b) The yellow highlighted area in (a) is magnified to show potential oligo primer designs to create a substitution, deletion, or insertion mutation. (c) PCR reactions can be performed in 96-well format and pooled to create a site-saturation mutagenesis library.

Proxies for TEM-1 Fitness

Many studies examining fitness effects of mutation use growth competition experiments to measure fitness [22]. In the case of TEM-1, the combined population of alleles is challenged to grow in the presence of Amp. Sequencing the population before and after the growth competition allows for calculation of an enrichment value, which can be used as a proxy for fitness. However, growth competition experiments with TEM-

1 suffer from some limitations, including the dependence of enrichment values on the concentration of Amp and the inability to measure low fitness values at high resolution. Minimum inhibitory concentration (MIC) assays are also commonly used to measure the level at which an allele is able to confer resistance to the antibiotic, which can be used as a proxy for fitness [16]. However, standard MIC assays rely solely on a positive selection for antibiotic resistance, which makes low-fitness alleles difficult to isolate. To overcome the limitations of both growth competition experiments and standard MIC assays for TEM-1, Firnberg et al and Steinberg et al used a synthetic biology approach that measures Amp resistance in a MIC-like manner.

Bandpass selection for Amp resistance

In contrast to typical positive selection for ampicillin resistance, in which cells grow only if they exhibit sufficient beta-lactamase activity to degrade the amount of ampicillin present, this system also restricts growth of cells with excessive beta-lactamase activity relative to the ampicillin concentration (Figure 5) [23]. This system allows for selection of cells exhibiting low or intermediate beta-lactamase activity. The system relies on a genetic circuit between the *ampR* gene and *tetC* gene, which allows for a user-specified selection range based on the antibiotics added to the media (Figure 5a). In the absence of sufficient beta-lactamase activity to hydrolyze ampicillin, ampicillin compromises cell wall synthesis, which inhibits cell growth (Figure 5b). The breakdown of the cell wall results in the intermediate, aM-pentapeptide (aM-Pp), the accumulation of which induces the *ampC* promoter for the expression of TetC, which confers tetracycline resistance. This accumulation of AM-Pp (and induction of the *ampC* promoter) happens

even at levels of ampicillin that are too low to prevent cell growth. However, if cells have too much beta-lactamase activity, ampicillin is rapidly degraded below the level that causes Am-Pp to accumulate, leaving the cells sensitive to tetracycline. Thus, in the presence of tetracycline, cells expressing beta-lactamase require that the ampicillin concentration be in a particular, narrow range that is set by the level of ampicillin resistance that beta-lactamase provides (Figure 5bc). This system allows a library of alleles to be plated on different concentrations of Amp and parsed into multiple sublibraries ranging from low fitness variants to high fitness variants, as determined by resistance to Amp. To afford control over beta-lactamase expression, it is regulated under the *tac* promoter through IPTG-induction, which in the absence of IPTG is repressed by LacI.

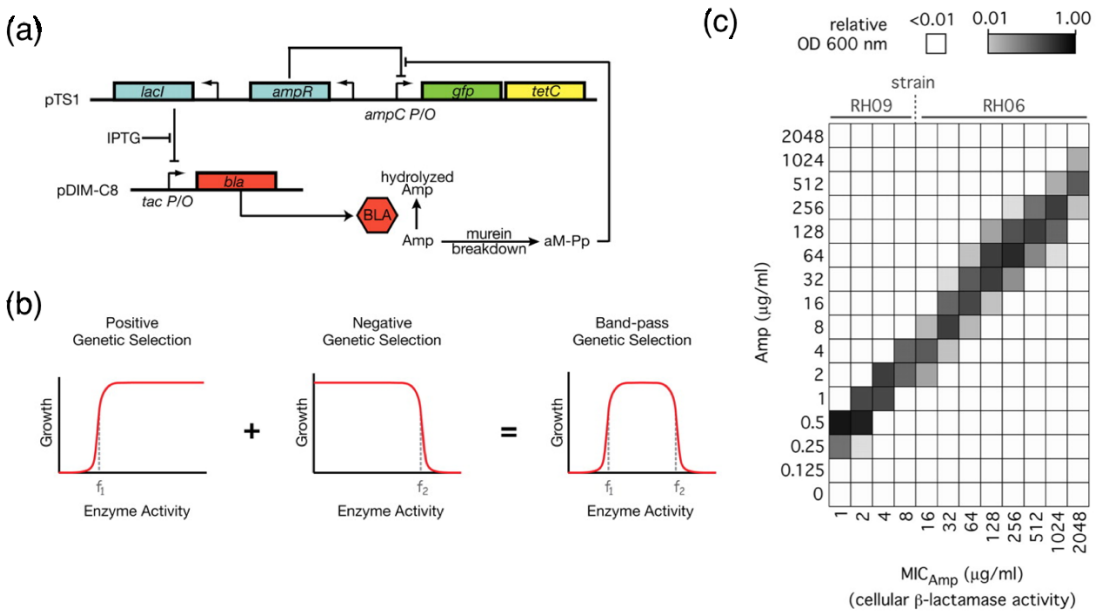

Figure 5. Bandpass selection for Amp resistance
Figure adapted from [23] (a) Schematic of the genetic circuit between the ampR and tetC genes. (b) Depiction of the bandpass effect resulting from combining a positive and negative selection. (c) The growth of cells as a function of Amp concentration (y axis) and cellular β -lactamase activity (x axis).

Deep-Sequencing of Libraries

Recent advances in DNA sequencing technology have been one of the most important boons for the progress of the study of fitness landscapes [24]. Next-generation DNA sequencing allows for the sequence identification of millions of gene variants. Variants can be linked to their respective phenotypes by adding identifying barcodes via PCR prior to sequencing. For example, Firnberg et al and Steinberg et al use 13 unique barcodes to represent the level of Amp at which each sublibrary of mutated alleles grew. Two of the most common deep-sequencing platforms are Illumina and PacBio (Pacific Biosciences) [25], each of which has strengths and limitations. Illumina offers the highest number of reads (currently routinely over 20 million), however, the read length available on most Illumina platforms is currently only 2x300 bp. Alternatively, PacBio offers longer read lengths (~10,000 bp), but about 100-fold fewer total reads.

Limitations of previous studies

Previous studies of mutational effects in TEM-1 have focused almost exclusively on substitution mutations. Insertions and deletions (InDels) represent another important, yet understudied, source of genetic variation in nature and the engineering. Furthermore, studies of epistasis in TEM-1 are limited to mutation accumulation studies (ref) and epistasis with respect to a small number of anchor mutations (ref). In the following studies, we utilize deep mutational scanning principles to further characterize the fitness landscape of TEM-1 by systematically examining epistasis between double mutants and investigating variation beyond substitutions. In chapter 2, we describe the patterns of

epistatic interactions among sequential amino acid substitutions throughout the protein, and in chapter 3, we describe the effects of amino acid insertions and deletions (INDELs). This work constitutes a significant piece of the puzzle in the emerging picture of the distribution of fitness effects and epistasis, both for TEM-1 specifically, and for protein landscapes more broadly.

Chapter 2: Pervasive pairwise intragenic epistasis among sequential mutations in TEM-1 β-lactamase

Summary

Interactions between mutations play a central role in shaping the fitness landscape, but a clear picture of intragenic epistasis has yet to emerge. To further reveal the prevalence and patterns of intragenic epistasis, we present a survey of epistatic interactions between consecutive mutations in *TEM-1* β-lactamase. We measured the fitness effect of ~12,000 pairs of consecutive amino acid substitutions and used our previous study of the fitness effects of single amino acid substitutions to calculate epistasis for over 8,000 pairs. We found widespread negative epistasis, especially in beta-strands and a high frequency of negative sign epistasis among individually beneficial mutations. In general, we found secondary structure and solvent accessibility to be better predictors of epistasis than mutant amino acid identity or distance from the active site. This study is the first to systematically examine pairwise epistasis throughout an entire protein performing its native function in its native host.

Introduction

Understanding the fitness effects of mutations is fundamental to the study of molecular evolution. Mutations can have different effects depending on the genetic background in which they occur. For example, a mutation that is beneficial in one context may become deleterious in another, limiting mutational trajectories or yielding evolutionary dead-ends. This interaction between two or more mutations, called epistasis,

plays a central role in evolution. Epistasis affects speciation [26, 27], the benefits of recombination and sex [28], genetic robustness [15, 29], and the predictability of evolution [30].

Genetic interactions can manifest in various ways. When two or more mutations interact such that their combined effect is more beneficial than predicted from their individual effects, it is termed positive epistasis. Alternatively, negative epistasis occurs when the combined effect is more deleterious than predicted. The magnitude of epistasis can have important consequences for the dynamics of evolution by affecting the curvature of the fitness landscape [6]. Sign epistasis occurs when a mutation is deleterious in one context, but beneficial in the presence of an additional mutation(s). The opposite is termed negative sign epistasis. A particular case of sign epistasis is reciprocal sign epistasis, in which two or more individual mutations are deleterious individually, but their combined effect is beneficial. This type of epistasis is particularly consequential in shaping the topography of the fitness landscape, causing local ruggedness and rendering certain peaks inaccessible [31].

Despite its theoretical importance in evolution, epistasis is understudied empirically and its contribution to evolution is not well understood. Empirical studies have aimed to elucidate aspects of epistasis in various ways. One way is by explicitly quantifying the functional or fitness effects of two or more mutations within a gene. Studies of intragenic epistasis have found it to be widespread [32-34] or rare [35, 36], mutational interactions to be typically strong [37] or weak [38], and sign epistasis to occur at a wide range of frequencies [34, 39]. The lack of consensus reflects the variety of molecules studied, differences in measuring function or fitness, modes of analysis, and

fundamental limitations of multi-mutant studies. Recent studies of epistatic interactions in RNA molecules, which are attractive due to their typically shorter gene lengths and fewer possible combinations of mutations, reveal a predominance of negative epistasis [40]. While it is possible to characterize nearly all combinations of two point mutations in a small RNA molecule, capturing the full landscape of every pair of amino acid substitutions in an average size protein is currently beyond our limits. Intragenic epistasis studies of proteins necessarily compensate by looking at combinations of a small subset of mutations, focusing on a small region, or surveying a small fraction of the possible pairs.

Many studies have focused on combinations of a small set of mutations, or random mutations in the background of a few "anchor mutations". For example, a study by Schenk et al [39] looked exclusively at combinations of beneficial mutations, quantifying epistasis in sets of four single mutations that had a known "large effect" or "small effect" on improving antibiotic resistance. They found significant negative epistasis in both landscapes and pervasive negative sign epistasis, especially among large effect mutations. Parera and Martinez (2014) tested epistasis by introducing a known deleterious amino acid substitution into various backgrounds of a protease and measuring catalytic efficiency compared to wildtype [37]. Significant epistasis was observed in 50 of the 56 backgrounds tested. A study by Bank at el (2014) analyzed more than 1,000 double mutants comprised of 7 point mutation backgrounds of neutral to slightly deleterious effect and found common negative epistasis (46%) and rare positive epistasis [32]. While these studies show important patterns in epistasis among a few known mutations, or among random mutations in the background of a few anchor mutations,

they may be limited in their ability to capture larger epistatic trends. We previously reported epistatic landscapes along an evolutionary pathway [19] wherein ~12,500 single amino acid mutants were analyzed in the background of the mutations that make up an adaptive pathway from TEM-1 to TEM-15 β-lactamase. The anchor mutation in each landscape was found to be a determining factor in the patterns of epistasis observed. For instance, while epistasis was rare in one background (8%), it was observed for 53% of mutants in another. This suggests that the use of anchor mutations to capture general trends in epistasis may bias the conclusions.

Studies that looked at random pairs of mutations often focused their scope to a small domain within a protein. Often the domain has been excised from its native protein, necessitating the characterization of interactions affecting a biophysical property, such as binding, in a non-native context. These studies are instrumental in revealing local epistatic interactions involved in a particular biophysical property. For instance, Araya et al calculated epistasis for ~5000 variants in a 34-amino acid WW binding domain using phage display [38]. They found epistasis to be rare, with values small in magnitude, and no population tendency toward positive or negative epistasis. In a 2014 study, Olson et al quantified the effects of all double mutations between all positions in the IgG- binding domain of protein G (GB1), using in vitro mRNA display [35]. They reported notable instances of both positive and negative epistasis, as well as sign epistasis, but overall observed that epistasis was rare. Likewise, Melamed et al (2013) analyzed double mutants within a 90 amino acid RNA recognition motif in a poly(A)-binding protein and found that only 3.6% exhibited negative epistasis and 1.0% exhibited positive epistasis [36]. They also found that pairs of mutations zero to five residues apart along the primary

sequence exhibited a significantly higher frequency of both positive and negative epistasis than pairs further apart. Bank et al (2016) examined epistasis among all possible combinations of 13 amino acid mutations at 6 sites in the heat shock protein, Hsp90 [32]. They found a prevalent pattern of negative epistasis and ruggedness in their local landscape, concluding that predicting fitness landscapes from the effects of individual mutations is made exceedingly difficult by genetic interactions.

Few studies have examined interactions between random pairs of mutations throughout an entire protein. A 2016 study of the fitness landscape of the green fluorescent protein defined fitness as the level of fluorescence in *E. coli* [13]. The authors sampled ~2% of all possible pairs of mutations, representing 30% of pairs of positions in the protein, and found that less than 5% exhibited epistasis. They observed pairs exhibiting epistasis to be located at sites across the gene, but slightly closer together than random. They found that pairs containing weak-effect mutations exhibited epistasis more often than pairs containing strong effect mutations, and suggest that the combined effect of weak mutations exhausts a stability threshold. Finally, they observed both strong and weak epistasis more prevalently among pairs of two buried sites, compared to pairs containing at least one solvent exposed site. Overall, they conclude that pairwise epistasis is more common at sites important to function.

Existing studies lack a survey of pairwise intragenic epistasis of a protein performing its native function in its native host in which the mutations are not limited to a particular domain or involve a small set of anchor mutations. Here, we examine pairwise epistasis throughout TEM-1 β-lactamase, a 286 amino acid antibiotic resistance protein native to *E. coli*. Informed by the observation that epistasis is more prevalent in pairs

close together in primary sequence [36], we asked a specific question: how does epistasis present in pairs of consecutive amino acid substitutions throughout the protein? Previously, we quantified the fitness effect of nearly all (95.6%) possible single amino acid substitutions in TEM-1 [18]. We use this data set to compare individual effects of mutations to the fitness effects of over 8,000 sequential double mutants. We find widespread negative epistasis (especially in beta-strands), with negative epistasis (52%) occurring 7.6 times as frequently as positive epistasis (6.8%).

Results and Discussion

TEM-1 is a convenient model for the study of gene/protein evolution, as it confers an easily identifiable and quantifiable phenotype – resistance to penicillin antibiotics, such as ampicillin (Amp). Although growth competition experiments in the presence of Amp can be used to measure enrichment of various alleles as a proxy for fitness, the values obtain depend on the concentration of Amp used [17]. In addition, the relative growth rate of cells with different alleles will change over time as the Amp in the culture is degraded, so the fitness values obtained are not precise relative growth rate comparisons. In addition, growth competition experiments have low resolution of low fitness alleles. As an alternative, minimum inhibitory concentration (MIC) assays can be used as a proxy for fitness, quantifying the ability of the allele to confer resistance to the antibiotic [16] [41], but MIC assays are not high throughput. Here, we use our previously described synthetic biology approach to quantify Amp resistance in a MIC-like fashion as a proxy for fitness [18, 23]. This method overcomes the limitations of growth competition experiments and standard MIC assays, as the fitness measures are ampicillin

concentration independent and low fitness values are as precisely measured as high fitness values. Our fitness values measure the level of ampicillin resistance conferred by the gene and are predictive of fitness values measured by growth competition experiments in the presence of a range of ampicillin concentrations (Figure 6).

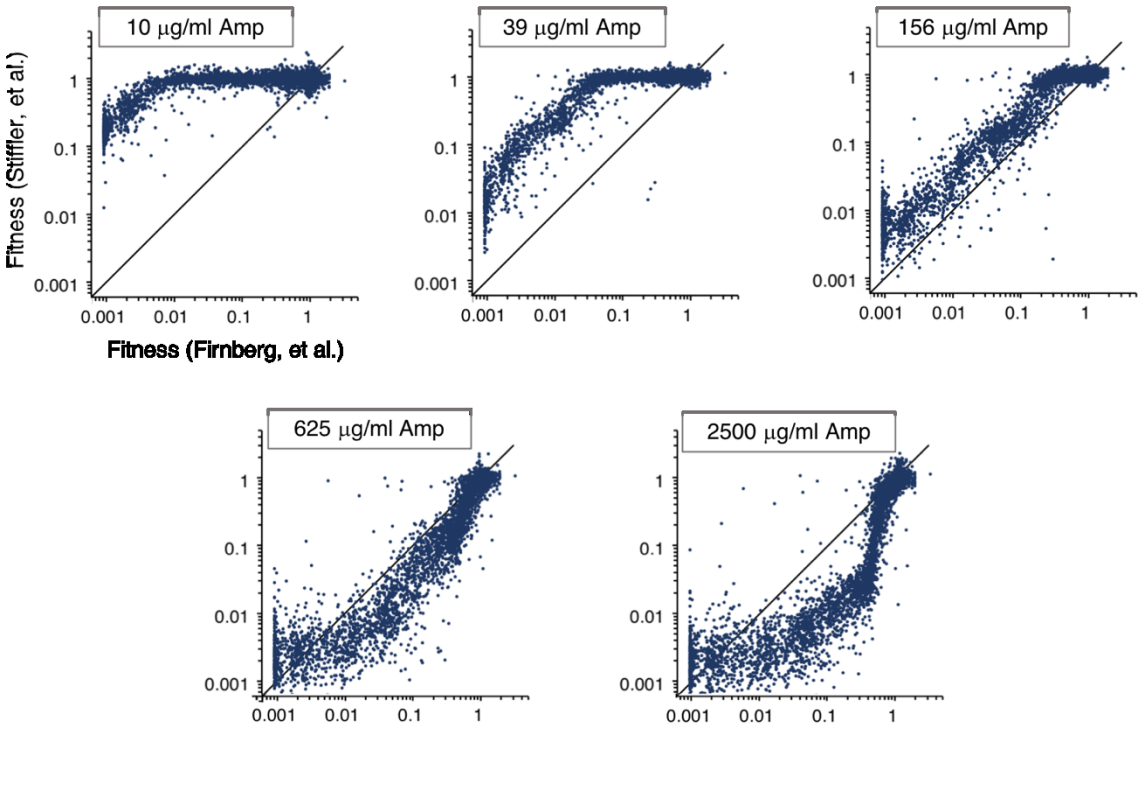

Figure 6. Fitness values for amino acid substitutions in TEM-1 measured by growth competition compared to fitness values measured by our bandpass MIC-like method Stiffler et al [17] performed the growth competition experiments in liquid LB media (with different concentrations of Amp as indicated) with DH10B *E. coli* cells containing TEM-1 under its native, constitutive promoter on plasmid pBR322. The fitness value associated with a mutation was measured by calculating the change in allele frequency relative to wildtype between before and after the growth competition. We performed our experiments on LB-agar plates with SNO301 *E. coli* cells containing TEM-1 under the IPTG-inducible *tac* promoter on a lower-copy *p15A* origin plasmid [18]. Fitness was measured as the resistance of cells carrying the mutation relative to wildtype using the bandpass system. This fitness measurement does not depend on the ampicillin concentration in the media (i.e. the fitness measurement for Firnberg et al is the same in all five graphs). The line is x=y.

We created a library of ~30,000 sequential double mutants in TEM-1 using inverse PCR using abutting, degenerate primers in which the 5'-end of one primer had the sequence $(NNN)_2$. [42]. We created separate libraries for each third of the gene to be compatible with the read length of the Illumina MiSeq 2x300 deep sequencing platform. We plated transformed SN0301 *E. coli* cells with the double mutant library on plates containing tetracycline and 13 different Amp concentrations ranging from 0.25 μg/ml to 1024 μg/ml. Whereas Amp prevents growth if the Amp concentration is too high relative to the amount of Amp resistance conferred, tetracycline prevents growth if the concentration of Amp is too low relative to the amount of Amp resistance conferred. As a result, a particular allele will confer growth only in a narrow range of Amp concentrations – a behavior that results from the band-pass synthetic gene circuit in SNO301 cells (see Firnberg et al for a detailed explanation [18]). We recovered the resulting sublibraries from the plates, PCR-amplified the appropriate third of the gene with Illumina MiSeq compatible barcodes, and deep sequenced the amplicons to determine how often each allele appeared on each plate. Sequencing reads of alleles containing synonymous codons were grouped together. The reported fitness is the calculated Amp concentration at which the mutant allele appeared most frequently relative to the same value calculated for wildtype allele. We calculated fitness values only for double amino acid mutants with 20 or more sequencing counts (see Materials and Methods for a more detailed explanation).

We next applied an adjustment to these fitness measurements to account for potential experimental differences between the two sets of fitness measurements. Our

epistasis calculations rely on consistent fitness measurements between our previous fitness measurements of single mutants [18] and the measurements of double mutants presented here. Thus, we took measures to ensure that the fitness values were consistent between the two experiments. We hypothesized that small differences in plating, incubation temperature, or other experimental factors may affect a cell's propensity to form a colony on each plate, perhaps resulting in a slight shift higher or lower in the Amp concentrations that favor growth. Such phenomena would result in systematic shifts in fitness values between the two experiments, which could be different for different ranges of fitness values.

To examine this possibility, we compared single mutant fitness values measured in each experiment. Our double mutant library creation technique also produced alleles containing one amino acid substitution and a synonymous wildtype mutation. We assumed that all observed synonymous mutations were neutral, consistent with our previous observations that the vast majority of synonymous mutations in *TEM-1* are neutral [18]. We compared the fitness values for the 1,470 such alleles in our experiment with the corresponding single mutant fitness values from Firnberg et al. We observed small offsets in fitness values that were sometimes different for different fitness value ranges. For example, fitness values less than ~0.125 were uniformly ~30% higher in the double mutant data set than the single mutant data set, whereas fitness values nearer to the wildtype value had a much smaller offset. Based on this observation, we adjusted the double mutant fitness measurements set to account for these differences. We judge this cross-experiment normalization procedure to be the most justifiable way to compare the

two sets of data. However, we also analyzed the data without the fitness value adjustments, and the overall trends presented in this study remained the same.

We obtained fitness values for 12,374 alleles of unique double mutant pairs, with an average of 30 pairs per position. This number represents 12.0% (12,374/102,855) of all possible consecutive double mutants. The distribution of fitness values of the double-mutants shows a shift toward lower fitness values (Figure 7b), compared to the distribution of fitness values of the single mutants (Figure 7a). The bi-modal shape of the single mutant distribution, with one peak around wildtype fitness and one at low fitness, is nearly lost. Only 89 double mutations resulted in fitness values significantly higher than wildtype. Nearly half (49.9%) of double mutations resulted in a near-complete loss of function (w<0.05). This shift toward low fitness is expected and in agreement with other mutation accumulation studies [13, 15, 36].

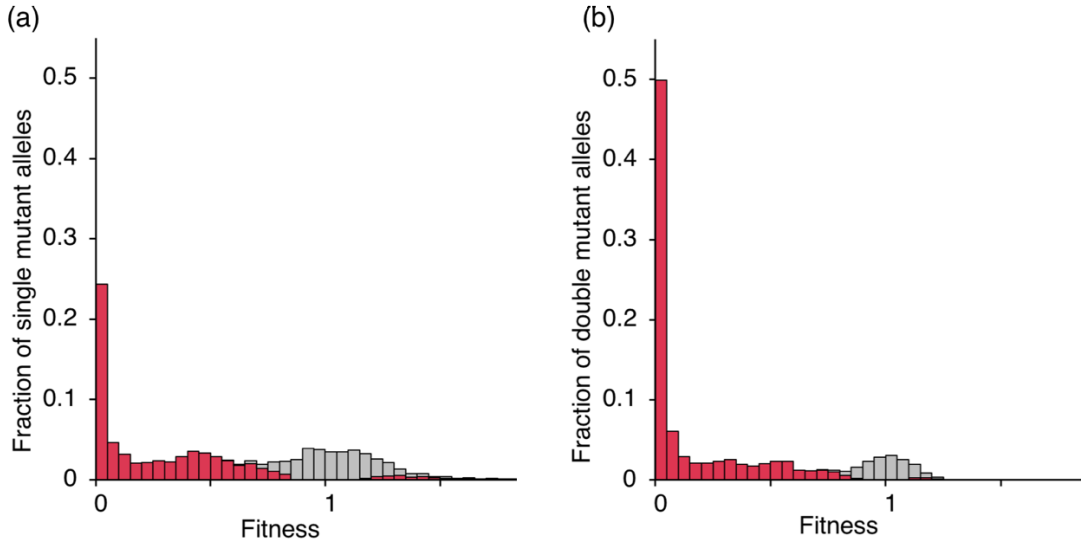

Figure 7. Distribution of mutational fitness effects of single and double mutants
(a) Distribution of 5460 single mutant fitness values [18]. (b) Distribution of 12,374 sequential double mutant fitness values. The single mutant distribution has a very small number of fitness values >1.8 that are not shown. Bars are stacked to show total fractions. Fitness values are normalized to that of wildtype TEM-1 beta-lactamase. Fitness values that are significantly different from 1.0 are indicated in red.

We define pairwise epistasis as occurring when the product of the fitness values of two individual mutations differs from the fitness of the combined pair. Epistasis (ε) between mutation A with fitness w_A and mutation B with fitness w_B is calculated as:

$$\varepsilon_{AB} = \log_{10}\left(\frac{w_{AB}w_o}{w_A w_B}\right) \quad (2.1)$$

where w_o is the fitness of wildtype TEM-1 and w_{AB} is the fitness of the double mutant.

We calculated epistasis for 8.1% (8,302/102,885) of all possible pairs of sequential amino acid substitutions. For our epistasis analysis, we exclude pairs containing mutations with individual fitness values less than 0.02 to avoid the lower limit in fitness measurements causing high epistasis values by artifact. Over half (58%) of all double mutants analyzed exhibited significant epistasis (Figure 8). The high prevalence of epistasis suggests a significant increase in epistasis among sequential mutations, lending support to a previous observation of this trend [36]. It may also reflect differences in the prevalence of epistasis with regard to fitness (here the ability of the allele to confer Amp resistance to live cells), compared to epistasis with regard to a less complex biophysical property, as hypothesized by Sackman and Rokyta [34]. The distribution of epistasis values was skewed toward negative values, with a mean epistasis of -0.32 and a median of -0.18, indicating that the combined fitness effect of two mutations is often more deleterious than predicted in the absence of epistasis. Negative epistasis (51%) occurred 7.5 times as frequently as positive epistasis (6.8%).

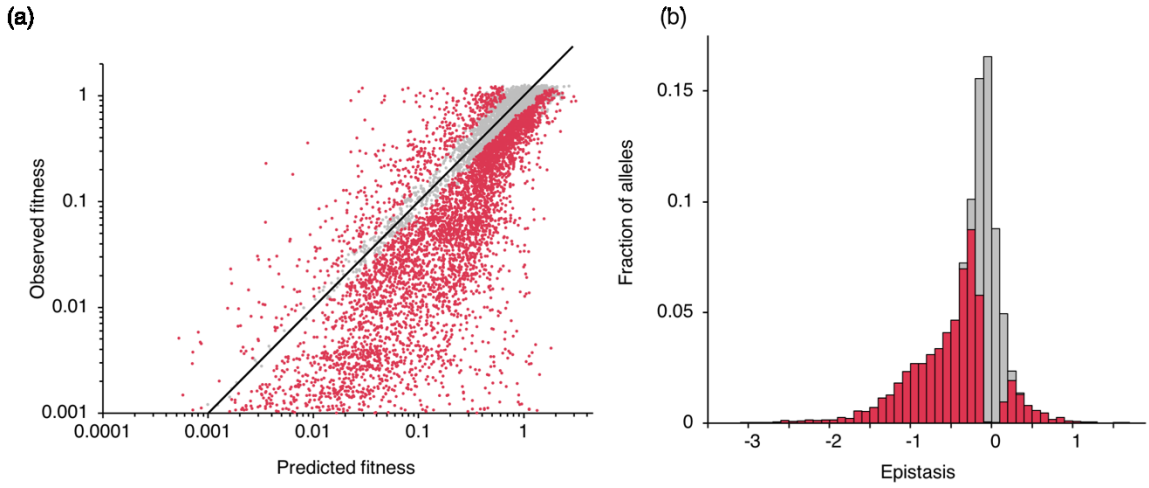

Figure 8. Distribution of epistasis values among sequential mutations
(a) Observed fitness versus predicted fitness for 8,302 double mutant alleles. (b) The distribution of epistasis values among 8,302 double mutant alleles. Bars are stacked to show total fractions. Significant epistasis values are indicated in red.

We found that the product of single mutant fitness values (i.e. the predicted fitness in the absence of epistasis) predicted double mutant fitness values with a Pearson's R^2 of 0.71. This is within the range of the correlations found in other epistasis studies, which had R^2 values ranging from 0.67 [38] to 0.76 [36].

Examining epistasis among sequential double mutant pairs allowed us to map median epistasis at each position and look at trends within secondary structures (Figure 9). Although negative epistasis dominates, there were 19 pairs of positions with positive median epistasis values, indicating hot spots for synergistic potential (Figure 9a). Interestingly, we note a particularly high median epistasis at positional pair 221-222. This median was calculated from a total of 21 observations. With the exception of one pair, the double mutants at this position were combinations of deleterious single mutations (median fitness of 0.052). Residues 221 and 222 make up the first two amino acids of a

four-residue helical element (helix 10). Positive epistasis, indicating a higher than expected fitness between individually deleterious mutations at this positional pair suggests hot spot for compensatory interactions, possibly buffering structural disruptions in the helix. Positive epistasis occurred 3 times more frequently in the signal sequence (17.8%) than the mature protein (5.82%) ($P<0.0001$, Fisher's exact test) (Figure 9c). The signal sequence is a 23 amino acid peptide that directs export of the protein to the periplasmic space of *E. coli*. The signal sequence is removed in the periplasm and is not part of the mature protein. However, mutations within the signal sequence can change protein abundance and therefore affect fitness. Over half (52%) of the occurrences of positive epistasis in the signal sequence were between one beneficial and one deleterious mutation, with the remaining 48% being between mutations that are deleterious individually. Positive epistasis in this region suggests detrimental mutations are easily partially compensated by mutations at adjoining positions.

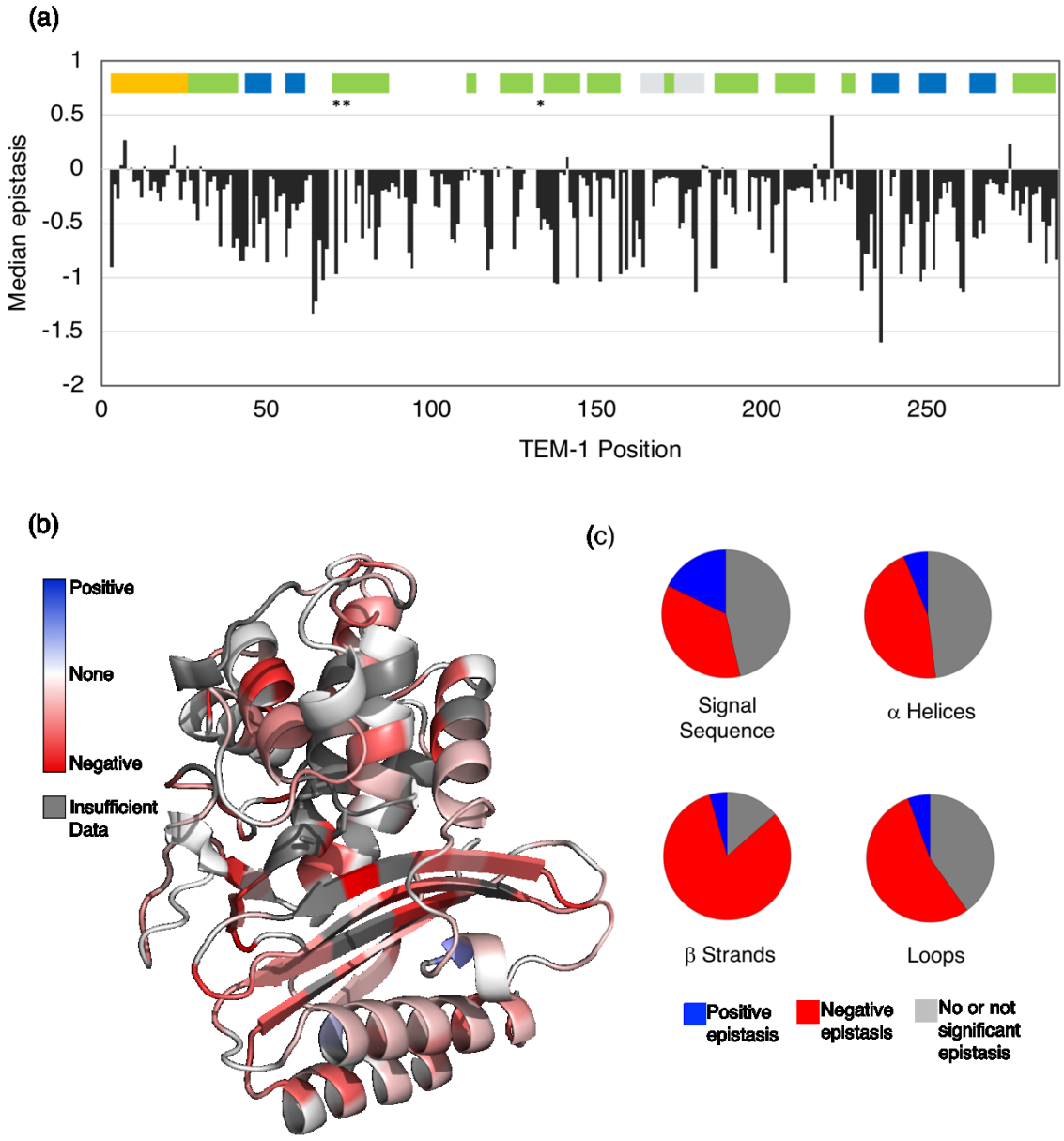

Figure 9. The relationship between protein sequence, structure, and epistasis
(a) Median epistasis values across the TEM-1 primary sequence. Median values were calculated only for position pairs with 5 or more epistasis values. Median epistasis for a mutation pair is plotted at the first position of that pair. Colored bars indicate regions that code for the signal sequence (yellow), alpha helices (green), beta strands (blue), and the omega loop (grey). Asterisks indicate the location of important catalytic residues. (b) Median epistasis values mapped onto the TEM-1 structure. Active site residues are indicated in green. (c) Frequency of positive epistasis (blue), negative epistasis (red), and no or not significant epistasis (grey) in the signal sequence and secondary structure elements. Data are categorized by the structural identity of the first mutation.

In the mature protein, negative epistasis occurred most often in beta-strands (Figure 9c), indicating that the interaction between two sequential mutations within these structures is often more detrimental than the combination of their individual effects. A majority (68%) of mutations occurring in beta-strands were individually deleterious. These findings suggest that the threshold robustness to additional deleterious mutations [15] is more quickly exhausted in beta-strands, presumably because the complexity of the structure has more constraints on the amino acids at each position.

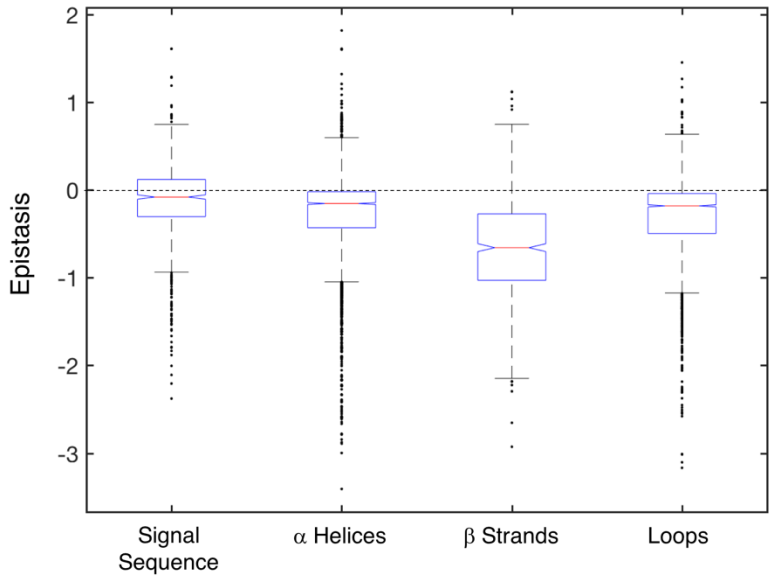

Figure 10. Epistasis values for the signal sequence and secondary structures
The central line indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles. The whiskers extend to the most extreme data points not considered outliers, which are represented by circles.

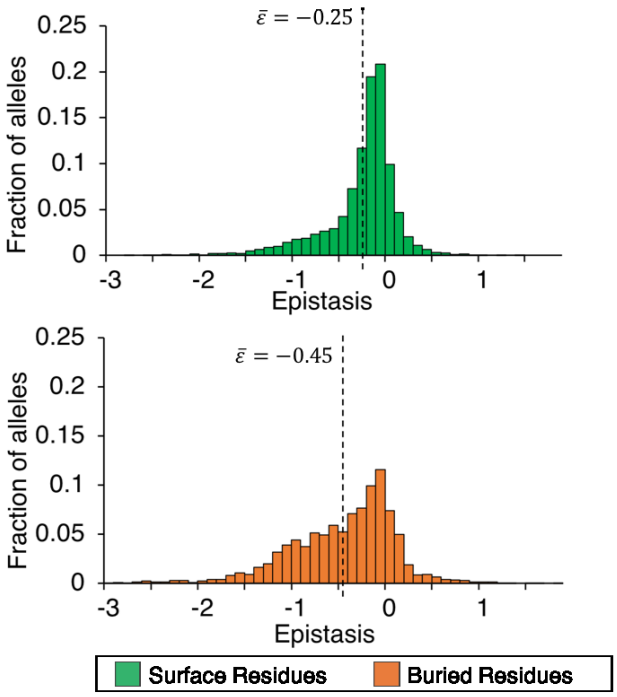

Figure 11. Epistasis distributions for buried and surface residues
The median value of the distribution is indicated.

We also examined epistasis among surface residues versus buried residues. We define surface residues as those with >20% solvent accessibility, and buried residues as those with <20% solvent accessibility. On average, buried residue pairs exhibited lower epistasis values than surface residue pairs (P <0.0001, by Student's t-test), suggesting that multiple mutations at internally oriented residues are more likely to interact antagonistically (Figure 11). Epistasis values for buried residues also had a broader distribution of values than epistasis values for solvent accessible residues (P<0.0001 by Brown–Forsythe test). We find that position and structure is more important in predicting epistasis than the identity of the amino acid pair substituted. We observed no obvious pattern in epistasis between different pairs of amino acids, however we note that the lowest two median epistasis values occurred between pairs of two cysteines and pairs of

two aspartic acids (Figure 12). We found no correlation between epistasis and the distance from the active site (Figure 13).

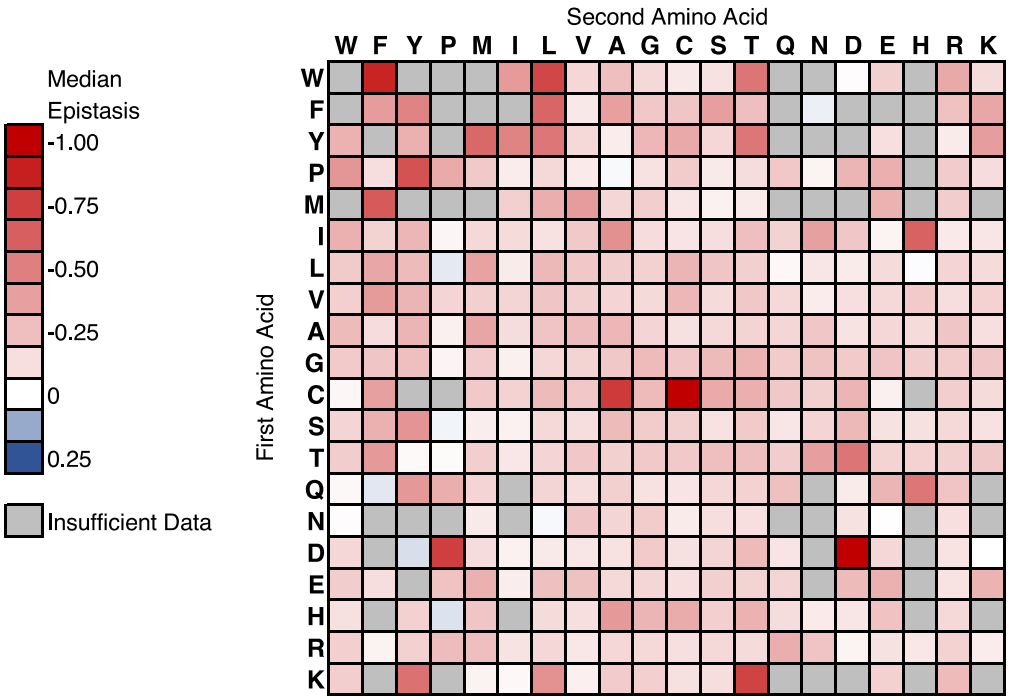

Figure 12. Median epistasis between pairs of mutant amino acids
The heat map indicates median epistasis values for mutant amino acid pairs that occurred throughout the protein. Median values are presented only for pairs with five or more epistasis values.

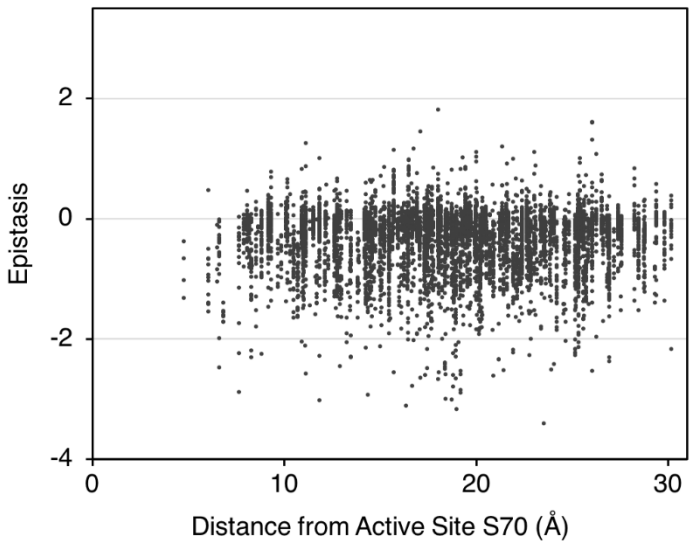

Figure 13. Epistasis versus distance from the active site S70
Distance was calculated from the first amino acid substituted.

Previous studies have noted differences in epistasis among individually beneficial versus deleterious mutations [32, 39]. Additionally, it has been posited that the effect size of the mutation may influence its epistatic effect in the context of another mutation [43]. To probe this further, we examined epistasis versus the effect size of the individual mutations contained in the pair. We define a mutation as deleterious if its fitness is more than two times its error below wildtype fitness and beneficial if its fitness is more than two times its error above wildtype.

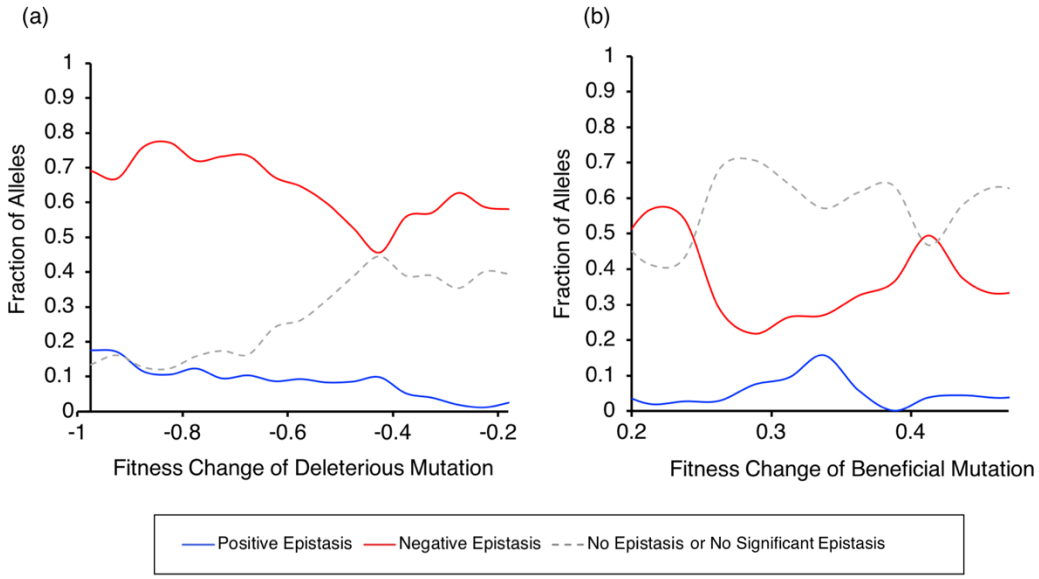

Figure 14. The effect of size and nature of the mutational effect on the frequency of positive and negative epistasis
(a) Frequency among mutation pairs with at least one deleterious mutation and (b) Frequency among mutation pairs with at least one beneficial mutation. The deleterious or beneficial mutation must have a statistically significant effect on fitness, but the other mutation in the pair may be deleterious, beneficial, or neutral. Boxcar smoothing was applied to the data to improve visualization of trends.

In general, epistasis was more frequently observed in pairs containing at least one deleterious mutation, whereas pairs containing at least one beneficial mutation more often displayed additive interactions (Figure 14). Epistasis was especially prevalent among large effect deleterious mutations (w<0.1), with nearly 90% of all pairs containing a large effect deleterious mutation exhibiting either positive or negative epistasis. In particular, pairs containing large effect deleterious mutations have a higher frequency of positive epistasis than pairs containing small effect deleterious mutations, suggesting that the fitness cost of highly deleterious mutations can be somewhat dampened by the presence of an additional mutation.

We also examined sign epistasis for 11,679 double mutant alleles for which we had corresponding single mutant fitness values. Sign epistasis is solely determined by the sign of fitness measurements (beneficial or deleterious). Unlike magnitude epistasis, it is not calculated from the product or ratio of two fitness values. Therefore, we included pairs containing single mutants with $w<0.02$ in the analysis of sign epistasis. By definition, positive sign epistasis can only occur for pairs containing at least one deleterious mutation and negative sign epistasis can only occur for pairs containing at least one beneficial mutation. We observe positive sign epistasis in only 13 out of 9673 pairs containing a deleterious mutation. The low frequency of positive sign epistasis indicates a scarcity of paths to climb above wildtype fitness in a single step next to deleterious mutations. Negative sign epistasis is much more prevalent, occurring in 55.4% of pairs containing a beneficial mutation. This indicates a moderately rugged landscape for sequential double mutants that is dominated by fitness valleys. We examined the relationship between negative sign epistasis and individual mutation effect size, but found the frequency to be >50% across all effect sizes. Thus, for beneficial mutations, the magnitude of the fitness effect does not predict the likelihood of surrounding fitness valleys. We found no cases of reciprocal sign epistasis, suggesting that many peaks may be accessible on the TEM-1 fitness landscape.

Conclusions

The picture of epistasis in protein evolution is still emerging. Our study examines pairwise intragenic epistasis in TEM-1 beta lactamase in the context of it performing its native function (antibiotic resistance) in its native host (*E. coli*). We specifically

examined pairwise epistasis between sequential amino acid substitutions across the entire length of the primary sequence. We postulated that consecutive double mutants represent a subset of possible mutational pairs that are more likely to exhibit epistatic effects due to spatial proximity and direct physical link in the backbone. Indeed, we find widespread negative epistasis in consecutive mutations throughout the protein, particularly in beta-strands, where amino acid orientation of sequential residues is important to structure fidelity. Our results lend support to the emerging landscape of pervasive negative epistasis and the threshold robustness hypothesis, the connection between individual mutant effect and epistatic patterns, and the importance of solvent accessibility in predicting the magnitude of epistasis. Together with other studies of epistasis in proteins in their native context, and compared with studies of epistasis with regard to biophysical properties, our findings lend support to the hypothesis that epistasis may be pervasive with regard to fitness, while reflecting underlying additive biophysical phenotypes.

Materials and Methods

Library Creation

The *TEM-1* gene was expressed on pSkunk3, a 4.36 kb plasmid containing spectinomycin resistance and the *p15* origin of replication, under the IPTG-inducible *tac* promotor in *E. coli*. We used inverse PCR with primers (IDT) designed to create every possible sequential double mutant in TEM-1, using NNN-NNN degenerate nucleotide oligos and a compatible reverse primer designed for each position. PCR products were visualized using gel electrophoresis, to confirm the creation of a linearized plasmid product at each of the 286 positions. We pooled the PCR products, isolated the ~4 kb

band from an agarose electrophoresis gel, phosphorylated the DNA at 37°C (NEB T4 PNK), and ligated it overnight at 16°C. NEB 5-alpha F' lacIq *E. coli* were transformed with the ligation product and plated on LB-agar plates containing 50 μg/ml spectinomycin and 2% glucose. At least 500,000 transformants were obtained for each third.

We recovered each library from the plate in LB media and isolated the plasmid library. We transformed electrocompetent SNO301 *E. coli* cells with each library and plated on LB-agar plates containing 50 μg/ml spectinomycin, 50 μg/ml chloramphenicol, and 2% glucose. At least 80,000 transformants were obtained from each third. We recovered each library from the plate in LB media and made glycerol stocks. The library sizes were greater than the number of sequences we could analyze by deep sequencing. Thus, we prepared a smaller sublibraries of each library by plating ~10,000 CFU from each library on LB-agar plates with 50 μg/ml spectinomycin, 50 μg/ml chloramphenicol, and 2% glucose (i.e. permissive growth conditions), recovering those cells, and creating final frozen sublibrary stocks for selection.

Selection and Sequencing

High-throughput selection for resistance to ampicillin (Amp) was performed using a band-pass genetic circuit, described previously [18]. Briefly, *E. coli* SNO301 cells containing the double mutant library were plated on LB-agar plates containing 20 μg/ml tetracycline and 13 different Amp concentrations, ranging from 0.25 μg/ml to 1024 μg/ml, in 2-fold increments. Plates were incubated for 21 hours at 37°C. The library was plated in triplicate on each Amp concentration and the CFUs from each plate were

counted to determine the frequency of colonies appearing on each plate. Based on these counts, a proportional amount of barcoded PCR amplicon from each plate was deep sequenced. Amplicons were prepared by recovering the cells from each selection plate, isolating the plasmid DNA, and performing PCR with appropriate primers as described previously [18, 19]. Barcodes to identify each plate and adapters compatible with Illumina MiSeq platform were added in this PCR step. Amplicons were pooled and sequenced using Illumina MiSeq with 300 base pair, paired-end reads.

Data Analysis

The de-multiplexed MiSeq reads were analyzed using custom MATLAB scripts. Paired-end reads were trimmed and concatenated to yield full length reads. Each read was then aligned to *TEM-1* using a Smith-Waterman algorithm with a gap opening penalty of 100. Reads with an alignment score lower than 300 were filtered out and only reads containing two sequential codon substitutions were used for analysis. Fitness was calculated for each unique double amino acid mutant based on the counts from each plate (Amp concentration). Synonymous codons were grouped together and total counts were used to calculate the single amino acid fitness. First, counts were adjusted based on the number of sequencing reads obtained from each plate relative to the CFUs observed on that plate, as described previously [19]. Detailed description of the fitness calculation can be found in our previous studies [18, 19], with a few differences. In this study, we excluded alleles with fewer than 20 counts and alleles with a maximum single plate count less than 1/3 the total count. For each allele (i) that passed these criteria, the plate with the highest adjusted counts and the four plates on either side (i.e. two plates with higher

Amp and two plates with lower Amp) were used to calculate an unnormalized fitness value, representing the midpoint resistance to Amp:

$$f_i = \frac{\sum_{p=1}^{13} c_{i,p} \log_2(a_p)}{\sum_{p=1}^{13} c_{i,p}} \quad (2.2)$$

where $c_{i,p}$ is the adjusted count of allele i on plate p, and a_p is the Amp concentration on plate p (in μg/ml). The reported fitness values are normalized to wildtype *TEM-1*:

$$w_i = \frac{2^{f_i}}{2^{f_{TEM-1}}} \quad (2.3)$$

Wildtype fitness was calculated in the same way (i.e. using adjusted sequencing counts) and verified separately by separately plating cells expressing wildtype TEM-1 in triplicate during the bandpass selection step. Both colony counts of the wildtype plates and wildtype sequencing counts revealed a midpoint Amp resistance of ~185 μg/ml (186.1μg/ml, 184.8 μg/ml, and 182.3μg/ml for each of the thirds, and 187.4 μg/ml for the colony counts).

We adjusted the fitness measurements based on a comparison between fitness values for 1,470 single amino acid substitutions containing a synonymous wild type mutation and the corresponding single amino acid fitness values from Firnberg et al. We calculated a ratio of the two fitness values across different fitness value ranges. Based on the offset of this value from 1, we determined adjustment factors for each range of fitness values, which ranged from 0.52 to 0.97. We multiplied the calculated double mutant fitness values by these adjustment factors and used these cross-experiment normalized fitness values for all subsequent analysis.

Error in fitness (σ_{w_i}) was estimated via Eqs 2.4 and 2.5, using our previously determined correlation between sequencing counts (n_i) and the standard deviation of the difference in fitness between synonymous alleles [18, 19].

$$\sigma_{w_i} = w_i \times e_i \tag{2.4}$$

where e_i, the upper-level estimate of the fraction error in fitness, is given by:

$$e_i = 0.667 {n_i}^{-0.387} \tag{2.5}$$

Fitness values were determined to be significantly different than 1 if they were greater or less than 1 by twice the error estimate.

Epistasis was calculated using Eq 2.1. To determine epistasis values that were significantly different than 0, upper and lower limits were calculated using Eqs 2.6 and 2.7:

$$\epsilon_{AB,U} = \log_{10}\left[\frac{w_{AB}w_0}{w_A w_B}\right]\left(1 + \sqrt{{e_A}^2 + {e_B}^2 + {e_0}^2 + {e_{AB}}^2}\right) \tag{2.6}$$

$$\epsilon_{AB,L} = \log_{10}\left[\frac{w_{AB}w_0}{w_A w_B}\right]\left(1 - \sqrt{{e_A}^2 + {e_B}^2 + {e_0}^2 + {e_{AB}}^2}\right) \tag{2.7}$$

Epistasis values were determined to be significantly positive or significantly negative based on Eq 2.8 and 2.9, respectively:

$$\varepsilon_{AB} - 2\left(\varepsilon_{AB} - \epsilon_{AB,L}\right) > 0 \tag{2.8}$$

$$\varepsilon_{AB} - 2\left(\varepsilon_{AB} - \epsilon_{AB,L}\right) < 0 \quad (2.9)$$

Sign epistasis was determined based on fitness measurements of the individual mutations and double mutant pair. Positive sign epistasis was defined as occurring when at least one of the mutants was individually deleterious (less than twice the error below 1), and the double mutant was beneficial (greater than twice the error above 1). Likewise, negative sign epistasis was defined as occurring when at least one of the mutants was individually beneficial, and the double mutant was deleterious. Reciprocal sign epistasis required both mutants to be individually deleterious, while the double mutant was beneficial. Negative reciprocal sign epistasis was the inverse.

Acknowledgements

This research was supported by the National Science Foundation (DEB-1353143, CBET-1402101, and MCB-1817646 to M.O.) and by the National Institutes of Health under a Ruth L. Kirschstein National Research Service Award (F31GM101941) to C.E.G.

Author contributions: C.E.G performed the experiments, and C.E.G. and M.O. conceived and designed the experiments, analyzed the data, and wrote the paper.

Chapter 3: Fitness effects of single amino acid insertions and deletions in TEM-1 β-lactamase

Summary

Short insertions and deletions (InDels) are a common type of mutation found in nature and a useful source of variation in protein engineering. InDel events have important consequences in protein evolution, often opening new pathways for adaptation. Yet much less is known about the effects of InDels compared to point mutations and amino acid substitutions. In particular, deep mutagenesis studies on the distribution of fitness effects of mutations have focused almost exclusively on amino acid substitutions. In this chapter, we present a near-comprehensive analysis of the fitness effects of single amino acid InDels in TEM-1 β-lactamase. While we found InDels to be largely deleterious, partially overlapping deletion-tolerant and insertion-tolerant regions were observed throughout the protein, especially in unstructured regions and at the end of helices. The signal sequence of TEM-1 tolerated InDels more than the mature protein. Most regions of the protein tolerated insertions more than deletions, but a few regions tolerated deletions more than insertions. We examined the relationship between InDel tolerance and a variety of measures to help understand its origin. These measures included evolutionary variation in β-lactamases, secondary structure identity, tolerance to amino acid substitutions, solvent accessibility, and side-chain weighted contact number. We found secondary structure, weighted contact number, and evolutionary variation in class A beta-lactamases to be the most predictive of InDel fitness effects.

Introduction

Insertions and deletions (InDels) are an important source of genetic variation in nature. They occur nearly as frequently as point mutations in some genomes [44, 45], and can result in dramatic effects on the properties of a protein and how it evolves [45-48]. Within the metaphor of an adaptive walk across a fitness landscape [49], InDels can be thought to represent a "leap" across sequence space rather than a step [50]. As such, InDels have the potential to open up new pathways for adaptation. For example, amino acid substitutions appear to be enriched around the site of InDel events in evolving proteins, either because InDel events actively trigger amino acid substitutions [50], or because substitutions enable subsequent InDels to accumulate in their vicinity via "neutral roaming" [45]. This suggests that how the surrounding protein region changes during selection may be substantially impacted by InDels. In the human genome, 15-21% of polymorphisms can be attributed to short InDels [51]. In-frame InDels are known to be the cause of diseases such as cystic fibrosis and are implicated in numerous types of cancer [52, 53].

InDels also represent a potentially underutilized source of variation in protein engineering. Though routine engineering of backbone modifications has been challenging, InDels have long been recognized as important tools for altering protein structure and properties [45, 54]. Because insertions and deletions add or remove atoms from the polypeptide backbone, they can cause major structural modifications not available through substitutions alone. They may be particularly important when seeking to dramatically change active-site structure, as they have been found to propagate long-

range effects on catalytic activity [54]. However, despite their importance in nature and the laboratory, InDels remain understudied compared to substitutions.

The fitness effects of point mutations and substitutions have been extensively studied in recent years [12, 25]. Previously, the Ostermeier lab comprehensively characterized the fitness effects of single amino acid substitutions in TEM-1 β-lactamase [18]. Other large-scale mutagenesis studies have been reported for over 14 proteins, characterizing the effects of single amino acid substitutions on function or fitness [12]. Such studies have advanced our understanding of the genetic code, protein structure, epistasis, and predictive models. However, we lack a similar systematic, large-scale analysis on the fitness effects of InDels.

Multiple studies have offered insight into the effects of deletions on a smaller scale. For example, a 2007 study of TEM-1 β-lactamase assayed 53 single amino acid deletions occurring throughout the protein, and found that 13 (24.5%) of the variants were inactive, while the remaining variants varied in activity, including four that retained wild-type levels, as measured by a minimum inhibitory concentration (MIC) assay [55]. The majority of debilitating deletions occurred in secondary structure elements and buried/core residues. Similarly, a 2014 study on enhanced green fluorescent (EGFP) protein characterized the tolerance to 87 random single amino acid deletions throughout the protein [56]. They found that the majority of tolerated deletions occurred in loops, while the rest were found equally distributed in helices and β-strands, with the termini of β-strands being more tolerant than the middle. Computational analysis of the EGPF found that structural properties such as relative solvent accessibility and packing density can be used to predict tolerance to deletions [57].

Insertion studies are even more limited, generally examining only a few rationally chosen insertion sites in a protein. For example, 2006 study in TEM-1 assessed the impact of random peptide insertion into three loops and found that tolerance depended largely on the insertion site [55]. Based on their findings, they also suggested that tolerance to insertions was not well-correlated to tolerance to substitutions in the same region.

While these studies provide important insights into the effects of InDels, they are limited by their scale. Here, we present a near-comprehensive analysis of the fitness effects of single amino acid insertions and deletions in TEM-1 β-lactamase, a widely studied antibiotic resistance protein. We find that while InDels are largely deleterious compared to substitutions, partially overlapping regions of tolerance to insertions and deletions exist throughout the protein.

Results and Discussion

TEM-1 β-lactamase is a commonly studied protein and convenient model for protein evolution experiments. It confers high resistance to penicillin antibiotics, such as ampicillin, which can be used as a proxy for protein fitness [16-18]. We use our band-pass, MIC-like approach for measuring antibiotic resistance in a high-throughput, high-resolution manner, as described in the previous chapter.

We focused on in-frame insertions or deletions of three nucleotides. We did not study insertions or deletions that are one or two nucleotides in length, as such mutations are frame-shifting mutations with drastic changes to protein sequence and nearly always inactivate proteins. We did not study three-nucleotide insertions or deletions that are out

of frame, as they cause substitutions in the amino acid sequence in addition to the amino acid insertion or deletion. We wanted to be able to isolate the effect of the single amino acid insertion or deletion away from any substitution effects.

We used inverse-PCR to create a plasmid library designed to code for every possible single amino acid insertion (5,720 variants) and every possible single amino acid deletion (286 variants) in TEM-1. For insertions, we used degenerate primers in which the 5'-end of the forward primer had an additional (NNN) sequence. For deletions, we used primers in which the 5'-end of the forward primer had a 3 base pair deletion. We transformed SNO301 *E. coli* cells with each library of InDel alleles and plated on tetracycline and 13 different Amp concentrations, as described in the previous chapter. We recovered the 13 sublibraries and performed deep-sequencing to determine how often each allele appeared on each plate. Sequencing reads of alleles containing synonymous codon insertions were grouped together, with the exception of the stop codons. The amber (UAG) stop codon exhibits nonsense suppression in SNO301 *E. coli* via the *supE44* tRNA allele, which results in glutamine incorporation at UAG codons with variable efficiency depending on the nucleotides immediately flanking UAG [18]. To avoid convolution, we included only non-amber stop codons in our analysis. The reported fitness values are calculated as the Amp concentration at which the mutant allele appeared most frequently relative to the wildtype allele (see Material and Methods for a more detailed description).

We obtained fitness values for 77.9% (4457/5720) of possible amino acid insertions and 97.9% (280/286) of possible amino acid deletions in TEM-1 (Figure 15). As expected, we find that insertions and deletions are largely deleterious. Over half of

insertions (51%) and deletions (59%) resulted in at least a 100-fold decrease in fitness relative to TEM-1. In contrast, only 9.8% of insertions and 11% of deletions retained 50% of wild-type fitness, though close to half (40.9%) of these were in the signal sequence, which is cleaved and not part of the mature protein. Though we measured 74 InDels alleles with fitness values greater than 1, only 27 were significantly different than 1. Visual examination of the heatmap depicting the fitness landscape (Figure 15) suggests a higher tolerance to InDels outside of secondary structures. It also suggests that the fitness effect of an insertion depends more on the site of the insertion than on the amino acid identity. To examine this quantitatively, we looked at the distribution of mean fitness values per position and compared it to the distribution of mean fitness values grouped by amino acid (Figure 16). We found that the mean fitness values per position have a wider distribution of values than the mean fitness values grouped by the amino acid inserted (P=0.009, Brown-Forsythe test).

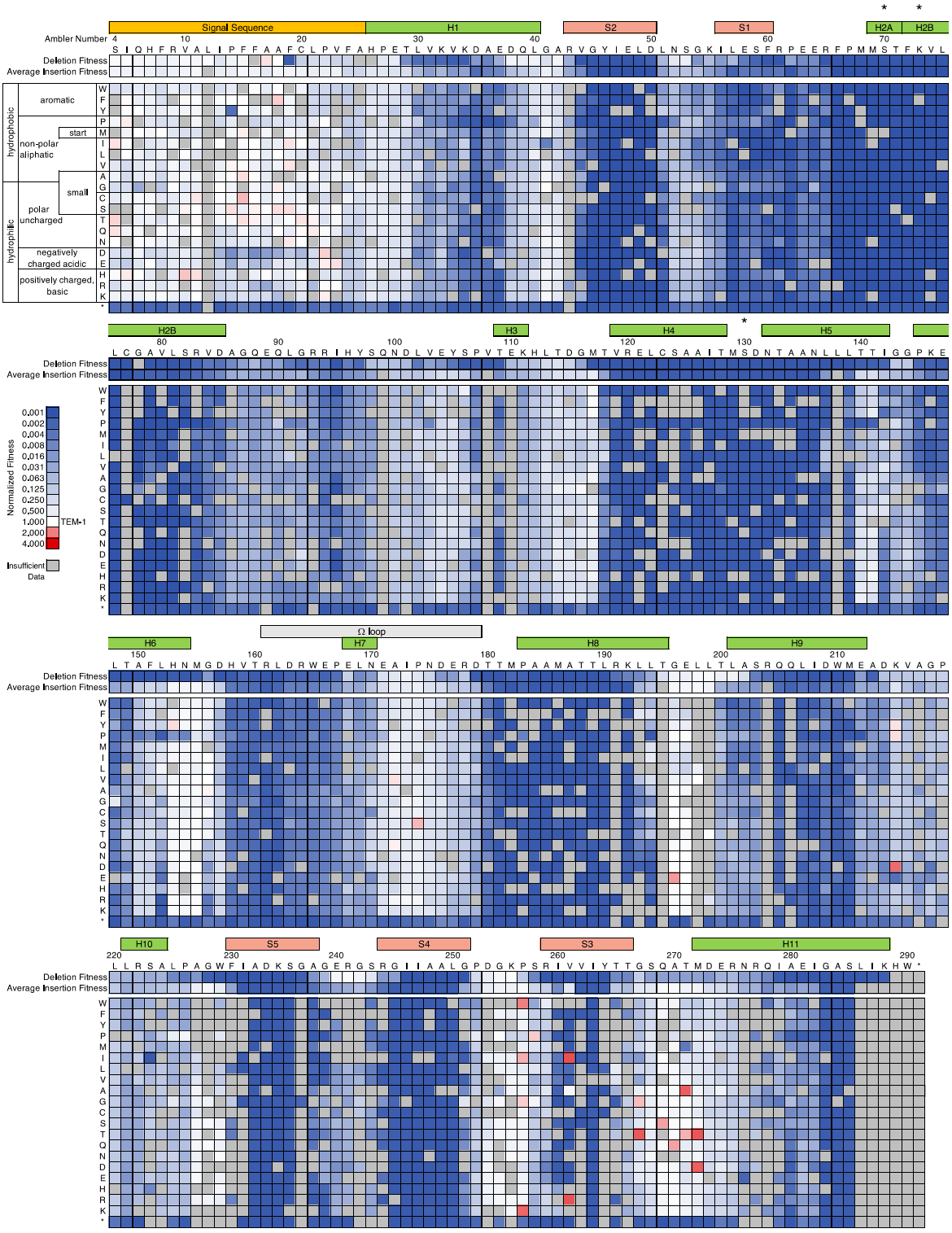

Figure 15. The sequence-function landscape of amino acid insertions and deletions in TEM-1

The heat map indicates relative fitness values as calculated based on ampicillin resistance. Insertion position is defined by the new position of the inserted amino acid

(e.g. an insertion denoted at position 50 was inserted between residues 49 and 50 in TEM-1). Ambler consensus numbering for beta-lactamases is used. The signal sequence (yellow), α helices (green), β strands (orange), Ω loop (grey), and active sites (*) are indicated.

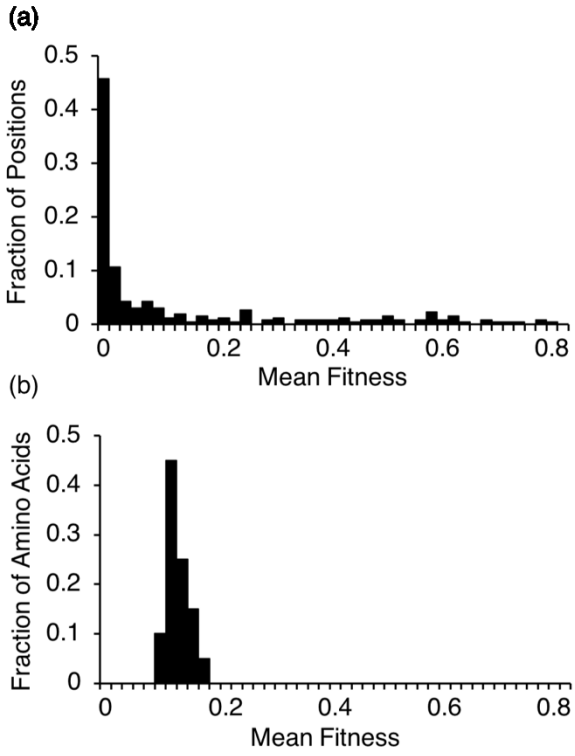

Figure 16. Distribution of mean fitness values of insertions by position and amino acid
(a) Mean fitness was calculated for each position in TEM-1 with >4 insertion fitness values. The distribution shows the fraction out of 270 positions. (b) A mean fitness was calculated for each amino acid insertion (regardless of position). The distribution shows the fraction out of 20 amino acids.

Examining the median fitness of alleles containing insertions and the fitness of alleles containing deletions across TEM-1, we observed “hot spots” of tolerance for InDels in the gene (Figure 17). The pattern suggests some correlation between where insertions and deletions are tolerated, and indicates higher tolerance in the signal

sequence and in unstructured regions of the protein. Higher tolerance to InDels in loops compared to helices and strands is widely observed across many families of proteins [58]. Our results also agree with previous observations in TEM-1 in particular. For example, visual examination of Figure 1 and Figure 3 suggests a notable tolerance to insertions in the loop connecting the final β-strand to the C-terminal helix, which is a location previously found to be broadly tolerant to random sequences of insertions [55].

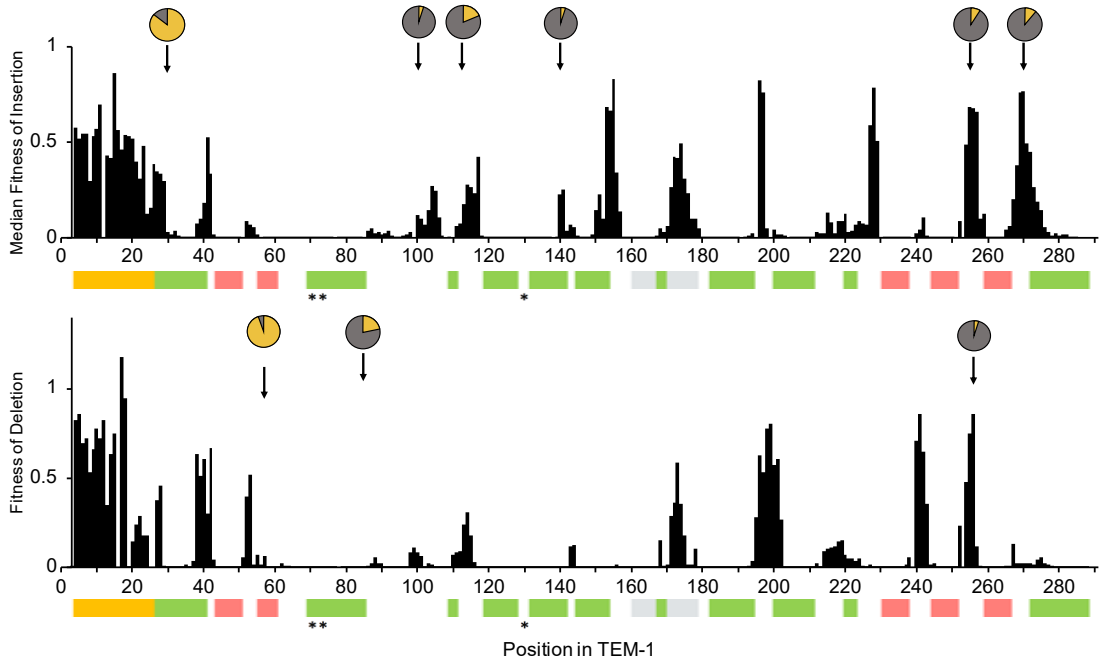

Figure 17. Fitness of TEM-1 containing InDels as a function of primary sequence. Median fitness values are presented for insertions. Arrows indicate positions at which other class A β-lactamases contain an insertion or deletion (based on a multiple sequence alignment of 156 class A β-lactamase and TEM-1). Pie charts indicate in yellow the fraction of sequences out of 156 that contain an insertion (top chart) or deletion (bottom chart) at that position. Fractions less than 3% are omitted. The colored bars indicate the signal sequence (yellow), α helices (green), β strands (pink), Ω loop (grey), and active sites (*).

We also examined the relationship between evolutionary variations in class A β-lactamases and the patterns we find in experimentally determined fitness of TEM-1. We aligned a published set of 157 class A β-lactamase sequences (including TEM-1) [59] by progressive multiple alignment using a Gonnet scoring matrix in MATLAB. We identified the positions at which other sequences contained an insertion or deletion relative to TEM-1. We find that these positions generally overlap insertion-tolerant regions in TEM-1, but several regions in TEM-1 that tolerate insertions and especially deletions are not observed in natural class A β-lactamases, at least in our dataset (Figure 17).

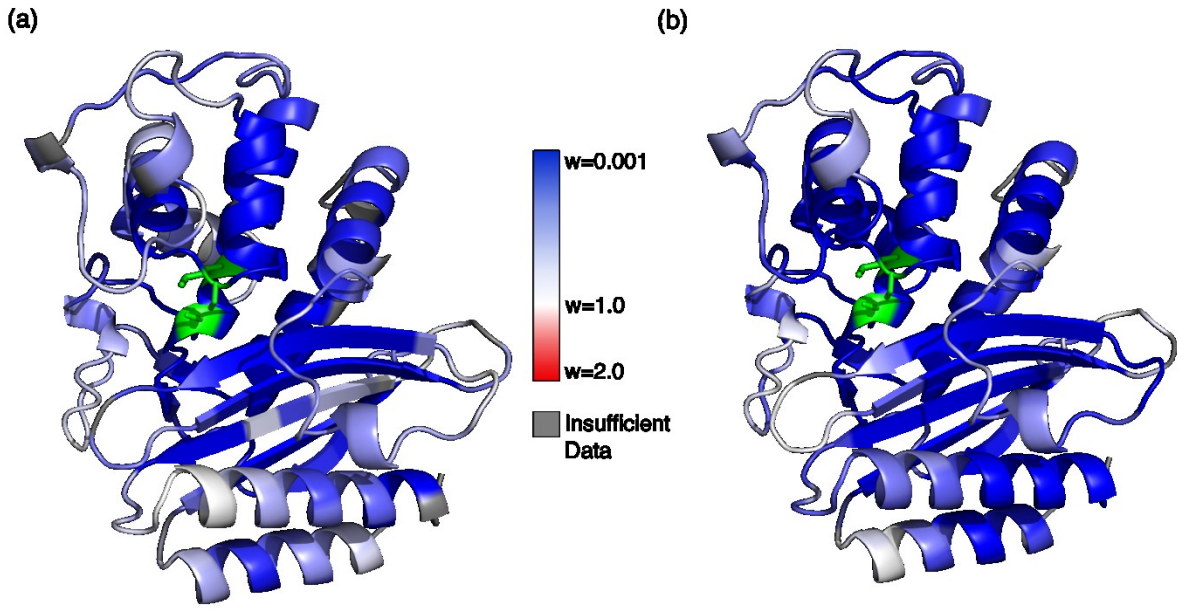

Figure 18. InDel fitness mapped onto TEM-1 structure
(a) TEM-1 secondary structure colored by mean fitness of insertions. No mean fitness values > 1 are observed. Positions for which we obtained fewer than 4 fitness values are indicated in grey. (b) TEM-1 secondary structure colored by fitness of deletions. In both figures, the active site residues are colored in green.

We find that the 23 amino acid signal sequence is the most InDel-tolerant region in TEM-1 (Figure 19). This sequence directs TEM-1's export to the periplasm via the Sec export pathway. The signal peptide is removed upon export to the periplasm and is not part of the mature protein. Presumably, mutations in the signal sequence affect fitness through changes of TEM-1's export efficiency to the periplasm. The signal sequence is also the most tolerant region to missense mutation. This tolerance reflects the loose sequence constraints for Sec-dependent signal sequences and its lack of secondary structure elements [60].

In the mature protein, helices and strands are the least tolerant to InDels. For both insertions and deletions, the mean fitness of mutant alleles in loop regions is higher than in secondary structure elements ($P<0.0001$ for insertions, $P<0.001$ for deletions, Student's t-test). This is not surprising given that backbone modifications can cause structured regions to fold incorrectly and have dramatic effects on the protein [61]. However, we found some exceptions to this overall pattern. For example, the loop region between β-strand S1 and α-helix H2A, shows no tolerance for insertions or deletions. We also found that 2.9% (51/1765) of insertions in α-helices, often at the ends of the structure element, resulted in less than a 50% decrease in fitness.

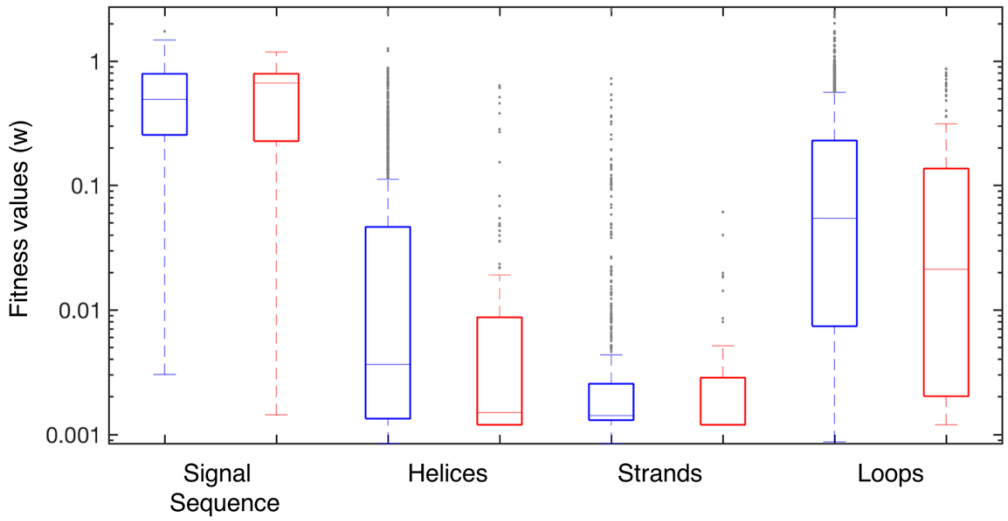

Figure 19. Relationship between InDel fitness and secondary structure
Box plots of fitness values for insertions (blue) and deletions (red) are shown for the signal sequence and secondary structure elements. The central line indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles. The median fitness value for deletions in strands is at the 25th percentile, and therefore not visible on the plot. The whiskers extend to the most extreme data points not considered outliers, which are represented by circles. Outliers are defined as values more than 1.5 times the interquartile range away from the top or bottom of the box.

To more specifically examine the difference between tolerance to insertions versus deletions, we calculated the ratio of the mean fitness of alleles with insertions to the fitness of an allele with a deletion at each position across TEM-1 (Figure 20). Overall, we find more regions where insertions are preferred over deletions, but a few regions where deletions are preferentially tolerated. For example, the C-terminal α-helix is dominated by preference to insertions, while the N-terminal α-helix contains positions where deletions are relatively preferred.

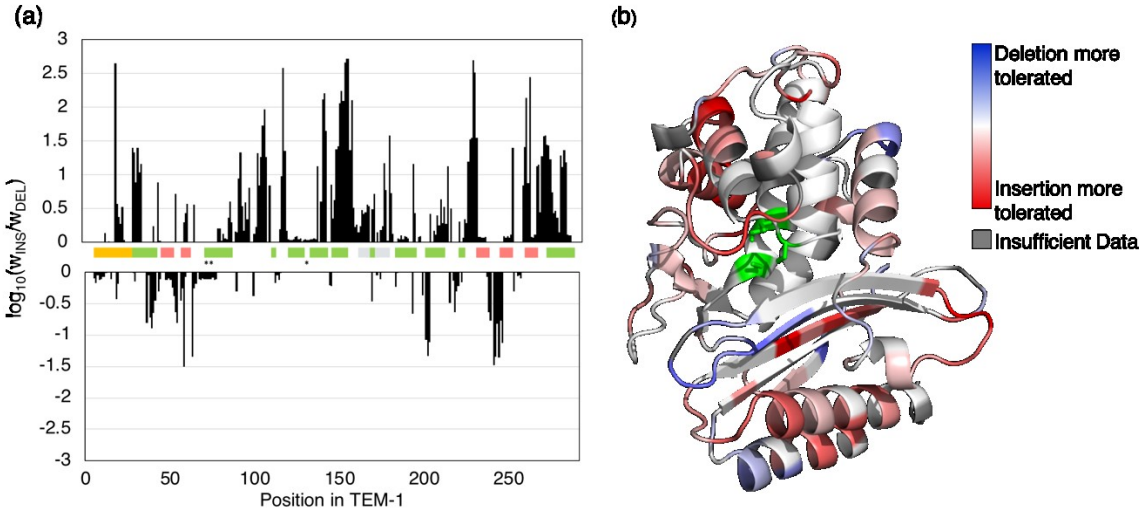

Figure 20. Differences in tolerance to insertions and deletions across TEM-1
(a) The $\log_{10}$ of the ratio between mean fitness of insertions and the fitness of a deletion at each position across TEM-1. The colored bars indicate the signal sequence (yellow), α helices (green), β strands (pink), Ω loop (grey), and active sites (*). (b) TEM-1 structure colored by the same ratio values. Blue indicates positions with higher tolerance to deletions, white indicates the same tolerance to both insertions and deletions, and red indicated higher tolerance to insertions.

We also examined the fitness effects of InDels compared to substitutions (measured in our previous study [18]). Unsurprisingly, we found a higher fraction of alleles containing InDels than alleles containing substitutions to be strongly deleterious. (Figure 21). The distributions of insertions and deletion fitness values are similar. The mean fitness of alleles containing an insertion is not significantly different than the mean fitness of alleles containing a deletion.

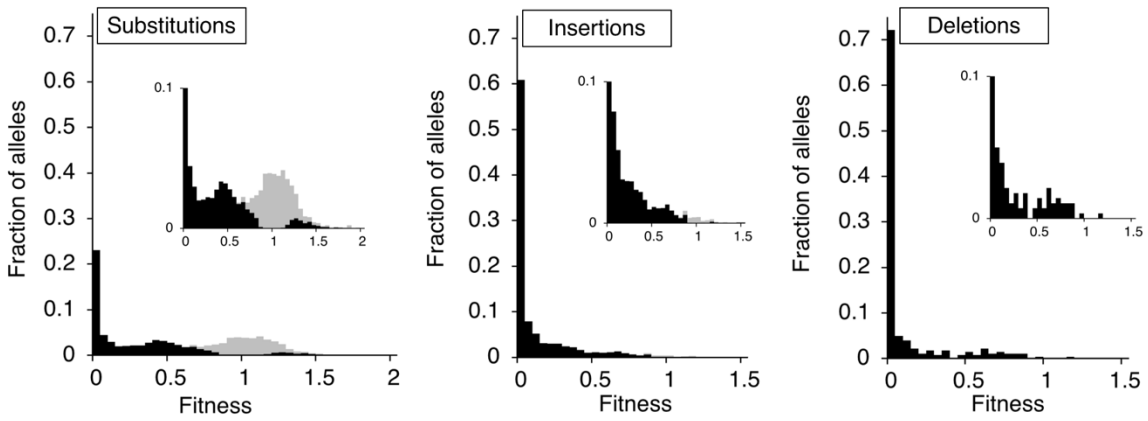

Figure 21. Distribution of Fitness Values for Substitutions and InDels
Distributions depict fitness values for 5460 alleles containing substitutions [18], 4457 alleles containing insertions, and 280 alleles containing deletions. The inset graphs show the same distributions that were truncated at a y-axis value of 0.1 to better show the distribution among higher fitness values. Grey bars indicate values that are not significantly different than 1.

To explore the comparison between insertions and deletions further, we examined the correlation between the mean fitness of alleles with an insertion at a given position and the fitness of an allele with a deletion at the corresponding position (Figure 22a) and found a weak correlation (R^2=0.32). We also compared the mean fitness change of an insertion of a given amino acid against the mean fitness change of a deletion of the same amino acid and found almost no correlation (R^2=0.07) (Figure 22c). This further indicates that the location of the InDel is more predictive than the identity of the amino acid inserted or deleted.

Next, we examined the correlation between fitness values when comparing insertions and substitutions. Specifically, we wondered if the fitness effect of an amino acid inserted before position N would correlate to the fitness effect of having position N mutated to the same amino acid. In this comparison, we included only fitness values of

insertions at positions with a mean fitness ≥ 0.1. We do this to account for the predominance of insertions that result in complete loss of function. By excluding those positions, we instead ask the question: where insertions are tolerated to some degree, what is the correlation between the effects of insertions and substitutions? We find very little correlation when we compare insertions to substitutions at the corresponding position (R^2=0.07) (Figure 22b); however, the mean fitness change of an amino acid substitution is somewhat predictive of the mean fitness effect of the same amino acid insertion (R^2=0.39) (Figure 22d). For example, the two least tolerated amino acid insertions (Pro and Trp) are also the least tolerated substitutions and the two most tolerated insertions (Ser and Thr) are among the most tolerated substitutions (Figure 22d).

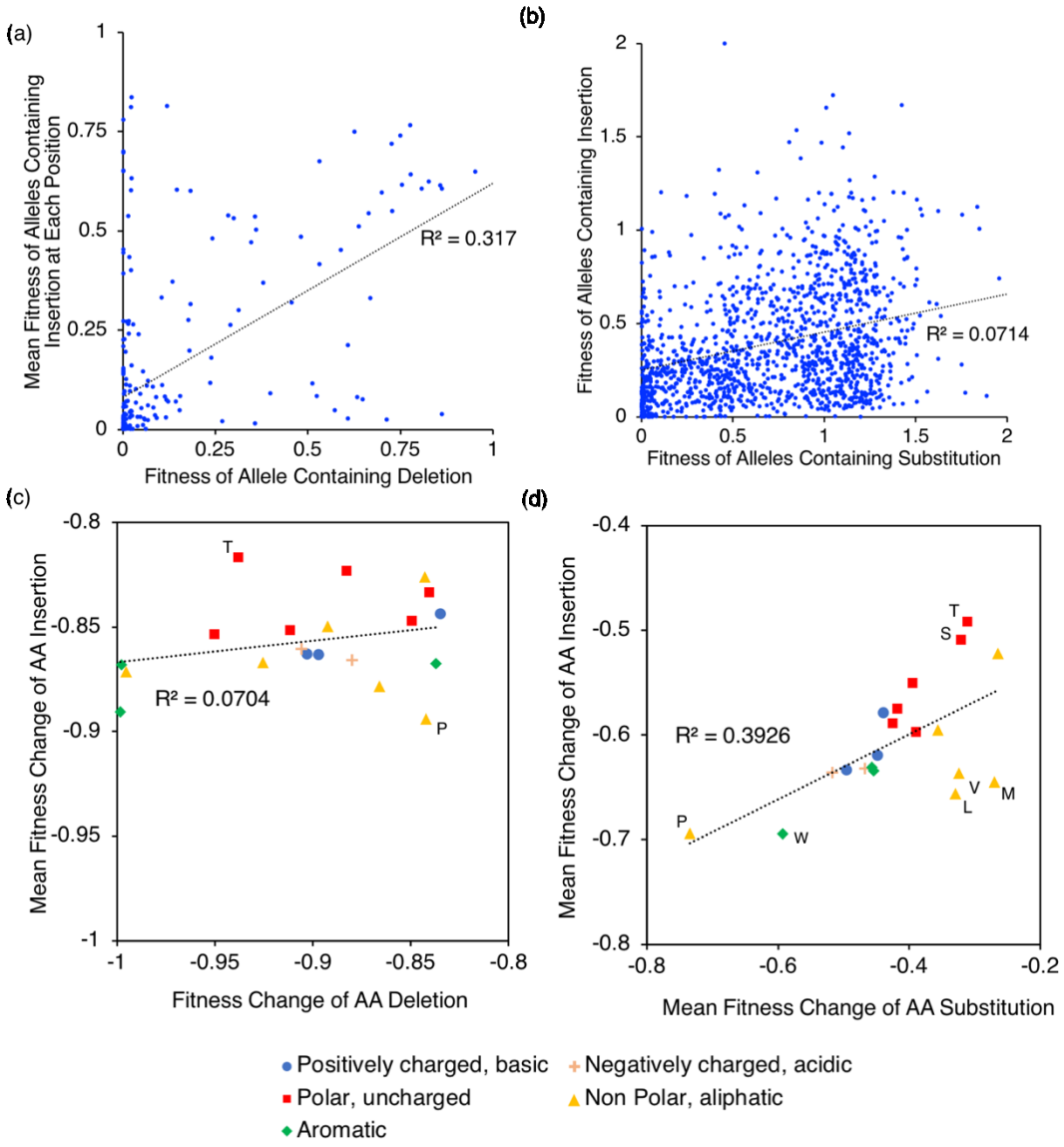

Figure 22. Comparison of the fitness effects of insertions, substitutions, and deletions

(a) Mean fitness of alleles containing insertions compared to the fitness of an allele containing a deletion at the corresponding position. (b) Fitness of alleles containing insertion compared to the fitness of alleles containing the corresponding substitution [18] (c) Mean fitness change of an amino acid inserted versus deleted. (d) Mean fitness change of an amino acid inserted versus substituted. Particular amino acids of interest are labeled. For (b) and (d) only insertion fitness values at positions with a mean fitness ≥0.1 are included.

We further explored TEM-1's tolerance to insertions by determining the effective number of amino acid insertions at each position. An analogous measure of tolerance (k^*) derives from information-theoretical entropy and was originally proposed to quantify the variability at a given position in a set of aligned sequences [62]. As we showed previously, k^* can be adapted to quantify the tolerance of substitutions based on measured fitness values[18]. For substitutions, a k^* value of 1 indicates a position at which all missense mutations result in complete inactivation of the protein, and a k^* value of 20 indicates that all amino acid substitutions result in the same fitness as wildtype. Here, we define a similar measure for insertions (k^*_{INS}) which includes the possibility of no insertion (i.e. wild type) in the distribution of protein fitness values at each position (Eqs 3.1-3.4)

$$k^*_{INS} = \frac{21k^*_{0,INS}}{n} \tag{3.1}$$

$$k^*_{0,INS} = 2^S \tag{3.2}$$

$$S = -\sum_{i=1}^{k} p_i \log_2 p_i \tag{3.3}$$

$$p_i = \frac{w_i}{\sum_{j=1}^{n} w_j} \tag{3.4}$$

A k^*_{INS} value of 1 indicates a position at which no amino acid insertion is tolerated (i.e. the fitness values of all amino acid insertions are zero) and a k^*_{INS} value of 21 indicates a position at which all insertions retain wild-type fitness values.

Over 30% of positions do not tolerate a single amino acid insertion of any kind ($k^*_{INS} < 2.0$) (Fig 23a). The peak in the distribution of k^*_{INS} values between 17 and 20 indicates that there is a fraction of positions (19.3 %) for which most insertions are well-tolerated. However, there are no positions for which every inserted amino acid retains wildtype fitness ($k^*_{INS} = 21$). Some positions in the signal sequence tolerated insertions after them more than substitutions at them.

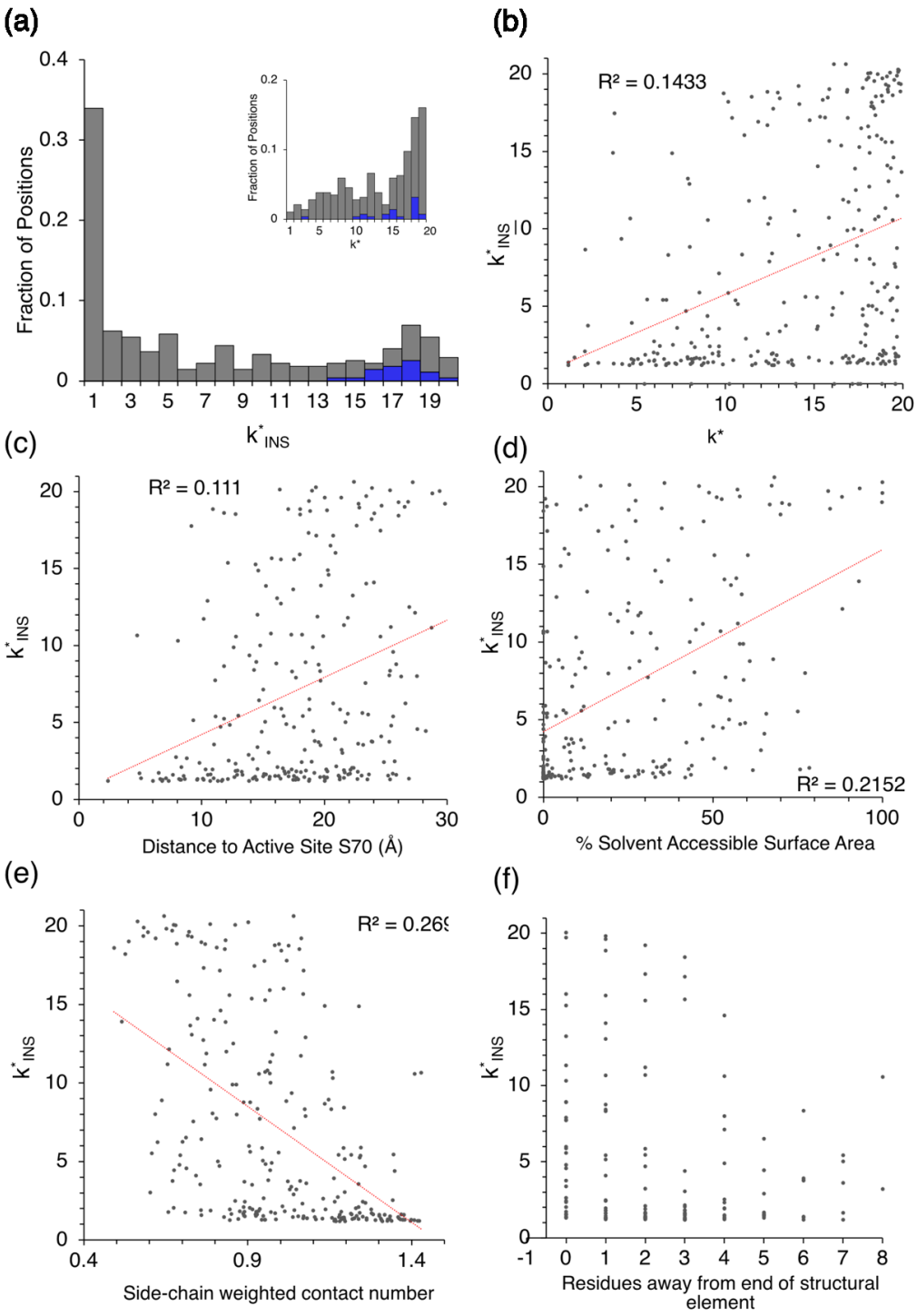

Figure 23. Determinants of tolerance of TEM-1 to amino acid insertions
(a) The distribution of k^*_{INS} values in TEM-1. k^*_{INS} values for the mature protein are colored in grey and k^*_{INS} values for the signal sequence are colored in blue. The inset shows the corresponding distribution of k^* values for substitutions [18]. (b) Correlation of k^*_{INS} with k^* of substitutions. [18] (c) Correlation of k^*_{INS} with distance from the

active site. (d) Correlation of k^*_{INS} with percent solvent accessibly surface area. (e) Correlation of k^*_{INS} with side-chain weighted contact number (WCN). (f) Correlation of k^*_{INS} with distance into secondary structure for positions within helices or strands. Distance is measured as residues away from the nearest end.

All 23 positions in the signal sequence had k^*_{INS} values above 13, but five positions had a k^* for substitutions less than 13 (Fig 23a). In the entire protein, a position's tolerance for insertion, as measured by k^*_{INS}, weakly correlated its tolerance for substitutions (Fig 23b). We found that tolerance to insertions correlates weakly with distance from the active site (Fig 23c). Positions less than 10 Å away from the active site are almost completely unaccepting of insertions. We observed a slightly stronger correlation between k^*_{INS} and percent solvent accessible surface area, with buried residues being less amenable to insertions (Fig 23d). We found that side-chain weighted contact number (WCN), a measure of how densely packed a residue is [63], best predicts how well an insertion is tolerated (Fig 23e). WCN is also the single best predictor of whether a deletion is tolerated in eGFP [57]. Within α-helices or β-strands, the ends of structural elements are more accepting of insertions than positions deeper into the center of the structure (Fig 23f).

Conclusions

Our analysis of InDels in TEM-1 provides the first systematic and near-comprehensive study of their fitness effect on a single protein and insight into a common yet understudied source of genetic variation. We found InDels to be largely deleterious, though regions of tolerance were observed, particularly in unstructured regions of the protein and at the ends of helices and strands. While regions of tolerance to insertions and

deletions partially overlapped, we found that most regions of the protein tolerated insertions more than deletions. Of the measures we examined, we found secondary structure, weighted contact number, and evolutionary variation in class A beta-lactamases to be somewhat predictive of InDel fitness effects. A broader understanding the fitness effects of InDels and how they relate to structural properties should allow for more informed protein engineering strategies, more robust computational prediction of protein structure, and a deeper understanding of the role that different types of mutations play in protein evolution.

Materials and Methods

Insertion Library Creation

The *TEM-1* gene was expressed on pSkunk2, a 4.36 kb plasmid containing spectinomycin resistance and the *p15* origin of replication, under the IPTG-inducible *tac* promotor in *E. coli*. We used inverse PCR with oligo primers (IDT) designed to create every possible single amino acid insertion in TEM-1, using primers with a degenerate nucleotide (NNN) sequence on the 5' end of the forward primer and a compatible reverse primer designed for each position. PCR products were visualized using gel electrophoresis, to confirm the creation of a linearized plasmid product at each of the 286 positions. We were unable to create a product for a small number of positions, despite troubleshooting efforts. We pooled the PCR products, creating a library for each third of the gene, to be compatible with Illumia MiSeq 2x300 bp sequencing. We isolated the ~4 kb band from an agarose electrophoresis gel for each third, phosphorylated the DNA at 37°C (NEB T4 PNK), and ligated it overnight at 16°C. NEB 5-alpha F' lacIq *E. coli* were

transformed with the ligation product and plated on LB-agar plates containing 50 μg/ml spectinomycin and 2% glucose. At least 500,000 transformants were obtained for each library (i.e. each third of the gene).

We recovered each library from the plate in LB media and isolated the plasmid library. We transformed electrocompetent SNO301 *E. coli* cells with each library and plated on LB-agar plates containing 50 μg/ml spectinomycin, 50 μg/ml chloramphenicol, and 2% glucose. At least 100,000 transformants were obtained from each third. We recovered each library from the plate in LB media and made glycerol stocks.

Deletion Library Creation

The deletion library was made in the same way as the insertion library with a few exceptions. The forward primer for inverse-PCR contained a 3-bp deletion on the 5' end, to create a deletion at every position in TEM-1. The same reverse primers were used. The deletion library was not created in thirds, as it was subsequently sequenced using PacBio, which can accommodate longer reads.

Selection and Sequencing

High-throughput selection for resistance to ampicillin (Amp) was performed using a band-pass genetic circuit, described in the previous chapter, and in previous work [18]. Briefly, *E. coli* SNO301 cells containing each library were plated on LB-agar plates containing 20 μg/ml tetracycline and 13 different Amp concentrations, ranging from 0.25 μg/ml to 1024 μg/ml, in 2-fold increments. Plates were incubated for 21 hours at 37°C. Each library was plated in triplicate on each Amp concentration and the CFUs from each

plate were counted to determine the frequency of colonies appearing on each plate. Based on these counts, a proportional amount of DNA from each plate was deep sequenced. For the insertion library, barcoded amplicons were prepared by recovering the cells from each selection plate, isolating the plasmid DNA, and performing PCR with appropriate primers as described previously [18, 19]. Barcodes to identify each plate and adapters compatible with Illumina MiSeq platform were added in this PCR step. Amplicons were pooled and sequenced using Illumina MiSeq with 300 base pair, paired-end reads. For the deletion library, we recovered cells from each selection plate, isolated the plasmid DNA, linearized it with the SphI restriction enzyme, and separately sequenced each of the 13 linearized plasmid libraries using PacBio.

Data Analysis

The de-multiplexed MiSeq reads and the PacBio reads were analyzed using custom MATLAB scripts. For MiSeq reads, paired-end reads were trimmed and concatenated to yield full length reads. Each read was then aligned to *TEM-1* using a Smith-Waterman algorithm with the lowest possible gap opening penalty of 1 and a gap extending penalty of 0.1. Reads with an alignment score lower than 100 were filtered out and only reads containing a single amino acid insertion (or deletion) were used for analysis. Fitness was calculated for each unique InDel mutant based on the counts from each plate (Amp concentration). For insertions, synonymous codons were grouped together and total counts were used to calculate the single amino acid fitness. Amber codons (UAG) were excluded from the stop codon analysis.

For each allele, counts were first adjusted based on the number of sequencing reads obtained from each plate relative to the CFUs observed on that plate, as described previously [19]. Detailed description of the fitness calculation can be found in our previous studies [18, 19], with a few differences. For the insertion library, we excluded alleles with fewer than 20 counts and alleles with a maximum single plate count less than 1/3 the total count. For the deletion library, we excluded alleles with fewer than 10 counts.

For each allele (*i*), the plate with the highest adjusted counts and the four plates on either side (i.e. two plates with higher Amp and two plates with lower Amp) were used to calculate an unnormalized fitness value, representing the midpoint resistance to Amp:

$$f_i = \frac{\sum_{p=1}^{13} c_{i,p} \log_2(a_p)}{\sum_{p=1}^{13} c_{i,p}} \tag{3.5}$$

where $c_{i,p}$ is the adjusted count of allele i on plate p, and a_p is the Amp concentration on plate p (in μg/ml). The reported fitness values are normalized to wildtype *TEM-1*:

$$w_i = \frac{2^{f_i}}{2^{f_{TEM-1}}} \tag{3.6}$$

Wildtype fitness was calculated in the same way (i.e. using adjusted sequencing counts) and verified separately by plating wildtype in triplicate during the bandpass selection step. Both colony counts and sequencing counts revealed a midpoint Amp resistance of ~215 μg/ml.

Error in fitness (σ_{w_i}) was estimated via Eqs 3.7 and 3.8, using our previously determined correlation between sequencing counts (n_i) and the standard deviation of the difference in fitness between synonymous alleles [18, 19].

$$\sigma_{w_i} = w_i \times e_i \tag{3.7}$$

where e_i, the upper-level estimate of the fraction error in fitness, is given by:

$$e_i = 0.667 {n_i}^{-0.387} \tag{3.8}$$

Fitness values were determined to be significantly different than 1 if they were greater or less than 1 by twice the error estimate.

Chapter 4: Conclusions and Future Directions

Summary of Work

In this work, we characterized the fitness landscape of InDels and the epistatic effects of sequential double mutants in TEM-1 β-lactamase. Each project involved a high-throughput, deep-mutational scanning approach resulting in the sequence identification of many thousands of mutations linked to their corresponding fitness effects. A near-comprehensive analysis of the distribution of fitness effects of single amino acid InDels and a systematic survey of epistatic effects throughout an entire protein performing its native function in its native host represent two thorough explorations of important, yet understudied aspects of the fitness landscape.

Future Directions

Adding to the InDel DFE Dataset

Fitness landscapes of single amino acid substitutions have allowed for general observations to be made about the impact of such mutations across various proteins. InDels have been studied far less comprehensively, in part because they are not part of the standard mutagenesis toolbox. We found that inverse PCR is a reliable and efficient way to create comprehensive libraries of single codon insertions and deletions in a gene. With this knowledge, future comprehensive studies of InDels may be possible in many other proteins, creating a dataset that allows for more general observations about their effects. This knowledge would be especially useful in protein engineering, where amino acid insertions are routinely used to introduce backbone flexibility and probe for new functions.

Computational Studies

Computational studies are a useful complement to large-scale mutagenesis studies. In particular, our InDel study provides the most comprehensive dataset for this type of variation. This dataset can be used to test the results of predictive computational models. For example, a dataset of 87 deletions in the green fluorescent protein was used by another group to test various computational prediction of the deletion tolerance of proteins [56, 57]. We propose that our near-comprehensive dataset of both insertions and deletions in TEM-1 offers a trove of information for similar analyses.

Expanding the Exploration of Epistasis

The study of epistasis is still largely in its infancy. Continuing advances in DNA deep sequencing technology will make increasingly larger studies possible. Our study of intragenic epistasis among sequential single amino acid substitutions represents one of many ways we imagine to better understand epistatic effects. For example, a comparison of the patterns of epistasis we observed in these sequential double mutants could be compared to that of double mutants randomly distributed throughout the protein. This type of study for genes lengths on the order of *TEM-1* is currently complicated by the read-length limits of deep-sequencing, but we expect read lengths to continue to increase as the technology advances. Another realm of epistasis involves studying higher-level epistatic interactions between more than two mutations. Again, these types of studies become exponentially complicated by the number of combinations and interactions involved in more than two mutations, and are limited in part by the technology. However,

even small fractions of multi-dimensional epistatic landscapes could further elucidate patterns of mutational interactions, including providing more insight into the nature of the robustness threshold. In addition to interactions between mutations within a gene, epistasis encompasses interaction between mutations in two or more different genes. Using comprehensive mutagenesis techniques to probe interactions between an anchor mutation in one gene and every possible amino acid substitution in another interacting gene could be one way of studying this.

Epistasis of InDels

Epistasis involving InDels is another potentially interesting avenue of exploration. Understanding mutational interactions could be key to understanding phenotypic effects, which would be especially useful in complex genetic diseases where this type of mutation is implicated. For example, while cystic fibrosis is known to be caused by a single deletion in the CF transmembrane conductance regulator protein, the protein acts within a network of components referred to as the "CFTR Functional Landscape" which influence synthesis, stability, and function [64]. A high-throughput systematic study of mutational interactions could provide insight into this network and inform more personalized therapeutics.

Pleiotropy and Collateral Fitness Effects

Another related field of inquiry to epistasis is that of pleiotropy, wherein a single mutant has multiple phenotypic effects. One way to study pleiotropic effects of mutations could be to perform large-scale mutagenesis studies on genes known to affect two

phenotypes. Another way to categorize mutations is by their primary effects versus so-called “collateral effects”. In the work presented here, we study primary effects of mutations on TEM-1, i.e. how mutations affect the ability of the protein to confer resistance to antibiotics. Ongoing work aims to examine collateral effects of mutations in TEM-1. Collateral effects may include those that cause misfolding and aggregation, or disrupt interactions between proteins. A similar saturation mutagenesis, high-throughput fitness measurement experiment in the absence of antibiotic will be the first to systematically study these effects.

Understanding the Molecular Foundations of Fitness

Finally, one of the most fundamental challenges in the study of fitness effects is “bridging the physical scales” of biology [65]. Fitness is a complex biological trait that involves biophysical properties that affect structure, stability, expression, catalytic activity, and/or binding on the molecular scale. Some mutational studies probe specifically for function, such as binding affinity, while others such as the ones we present here, select for fitness as it relates to cell viability and growth. This complicates the ability to combine findings from various studies into a broader picture of fitness and epistatic landscapes. For example, observations of the effect of multiple mutations on physio-chemical properties tend to reveal mostly additive interactions [4, 66], while many studies of epistatic interactions with respect to fitness, as we study here, reveal pervasive non-additive effects. Thus, it appears that mutations often interact additively on one scale, but those properties interact non-additively to produce epistasis on another

scale. Better understanding how these scales interact would enhance our ability to unite the conclusions from different types of studies into a broader picture.

Appendix

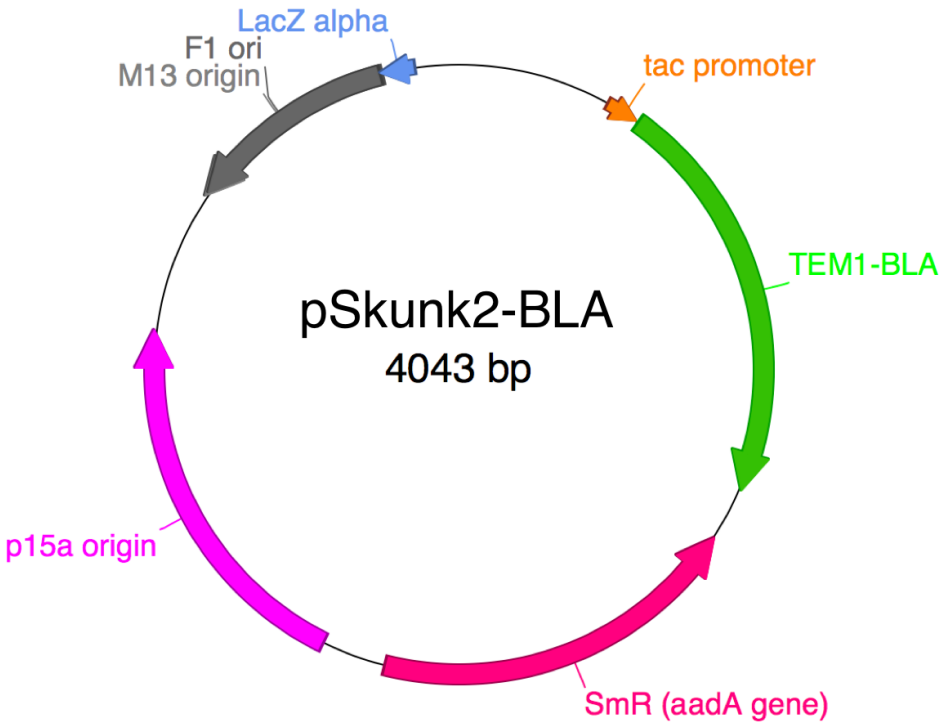

Figure 24. Plasmid map of pSkunk2-BLA.
TEM-1 is under the expression of the tac promoter. The aadA gene confers resistance to streptomycin and spectomycin antibiotics. The p15a origin gives a low copy number (10-12) plasmid. pSkunk3-BLA differs only in the non-coding region between TEM-1 and SmR (for improved compatibility in Sanger sequencing).

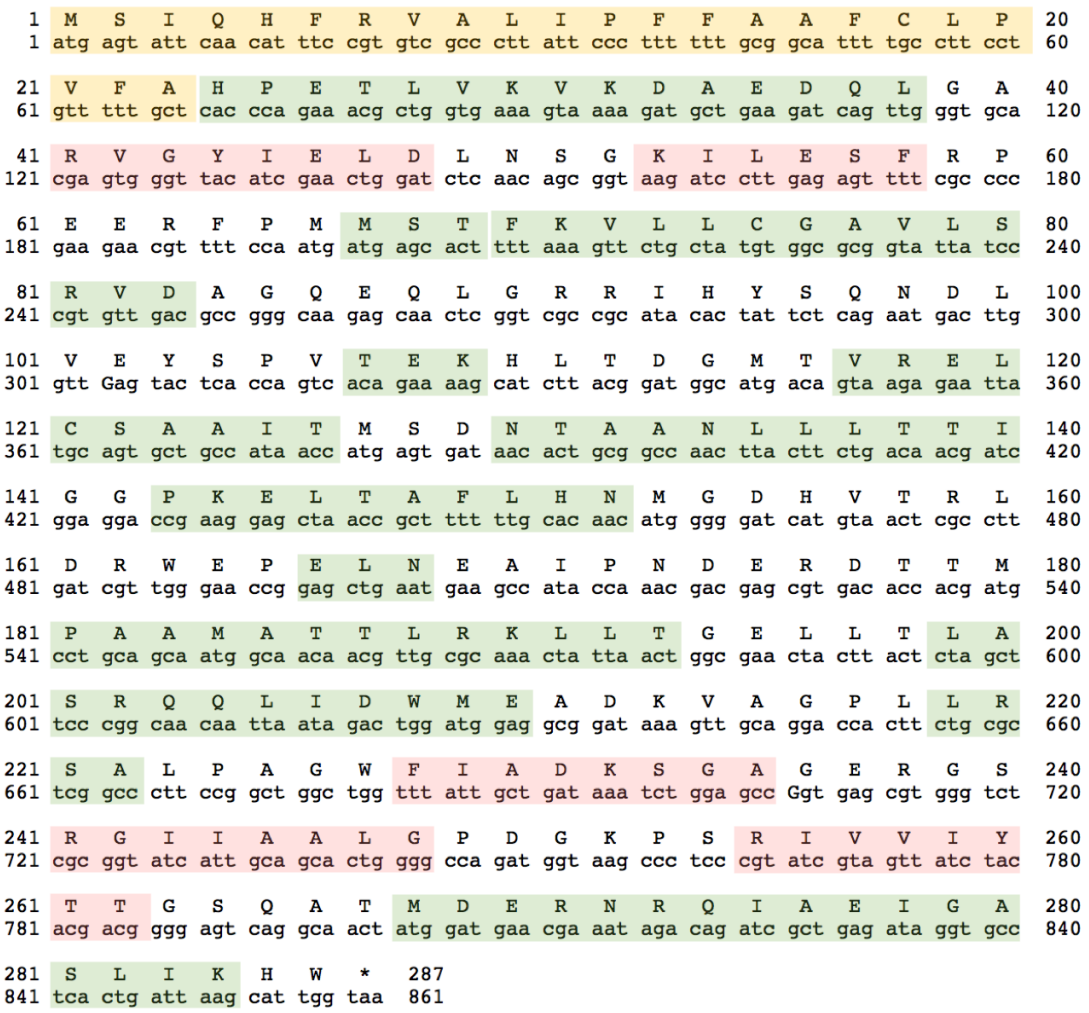

Figure 25. Amino acid and codon sequence for TEM-1.
Highlighted regions indicated sequences that code for the signal sequence (yellow), α-helices (green), and β-strands (pink).

References

1. Graur, D.L., W., *Fundamentals of Molecular Evolution*. Second Edition ed. 2000: Sinauer Associates, Inc.

2. Wright, S., *The roles of mutation, inbreeding, crossbreeding and selection in evolution.* Proceedings of the sixth international congress of genetics 1932.

3. Smith, J.M., *Natural selection and the concept of a protein space.* Nature, 1970. **225**(5232): p. 563-4.

4. Starr, T.N. and J.W. Thornton, *Epistasis in protein evolution.* Protein Science : A Publication of the Protein Society, 2016. **25**(7): p. 1204-1218.

5. Kauffman, S.A. and E.D. Weinberger, *The NK model of rugged fitness landscapes and its application to maturation of the immune response.* Journal of Theoretical Biology, 1989. **141**(2): p. 211-245.

6. Ivan, G.S., et al., *Quantitative analyses of empirical fitness landscapes.* Journal of Statistical Mechanics: Theory and Experiment, 2013. **2013**(01): p. P01005.

7. Kingman, J.F.C., *A simple model for the balance between selection and mutation.* Journal of Applied Probability, 2016. **15**(1): p. 1-12.

8. Neidhart, J., I.G. Szendro, and J. Krug, *Adaptation in Tunably Rugged Fitness Landscapes: The Rough Mount Fuji Model.* Genetics, 2014. **198**(2): p. 699.

9. Fowler, D.M. and S. Fields, *Deep mutational scanning: a new style of protein science.* Nature methods, 2014. **11**(8): p. 801-807.

10. Boucher, J.I., D.N.A. Bolon, and D.S. Tawfik, *Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature.* Protein Science, 2016. **25**(7): p. 1219-1226.

11. Freeman, A.M., et al., *Action at a Distance: Amino Acid Substitutions That Affect Binding of the Phosphorylated CheY Response Regulator and Catalysis of Dephosphorylation Can Be Far from the CheZ Phosphatase Active Site.* Journal of Bacteriology, 2011. **193**(18): p. 4709.

12. Gray, V.E., R.J. Hause, and D.M. Fowler, *Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions.* Genetics, 2017. **207**(1): p. 53-61.

13. Sarkisyan, K.S., et al., *Local fitness landscape of the green fluorescent protein.* Nature, 2016. **533**(7603): p. 397-401.

14. Cooksey, R., et al., *Patterns and mechanisms of beta-lactam resistance among isolates of Escherichia coli from hospitals in the United States.* Antimicrobial Agents and Chemotherapy, 1990. **34**(5): p. 739.

15. Bershtein, S., et al., *Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein.* Nature, 2006. **444**(7121): p. 929-32.

16. Jacquier, H., et al., *Capturing the mutational landscape of the beta-lactamase TEM-1.* Proceedings of the National Academy of Sciences, 2013. **110**(32): p. 13067.

17. Stiffler, Michael A., Doeke R. Hekstra, and R. Ranganathan, *Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase.* Cell, 2015. **160**(5): p. 882-892.

18. Firnberg, E., et al., *A comprehensive, high-resolution map of a gene's fitness landscape.* Mol Biol Evol, 2014. **31**(6): p. 1581-92.

19. Steinberg, B. and M. Ostermeier, *Shifting Fitness and Epistatic Landscapes Reflect Trade-offs along an Evolutionary Pathway.* J Mol Biol, 2016. **428**(13): p. 2730-43.

20. Firnberg, E. and M. Ostermeier, *PFunkel: Efficient, Expansive, User-Defined Mutagenesis.* PLOS ONE, 2012. **7**(12): p. e52031.

21. Kunkel, T.A., *Rapid and efficient site-specific mutagenesis without phenotypic selection.* Proceedings of the National Academy of Sciences, 1985. **82**(2): p. 488.

22. Merritt, J. and J.S. Edwards, *Assaying gene function by growth competition experiment.* Metabolic Engineering, 2004. **6**(3): p. 212-219.

23. Sohka, T., et al., *An externally tunable bacterial band-pass filter.* Proceedings of the National Academy of Sciences, 2009. **106**(25): p. 10135.

24. Boucher, J.I., et al., *Viewing Protein Fitness Landscapes Through a Next-Gen Lens.* Genetics, 2014. **198**(2): p. 461-471.

25. Gupta, K. and R. Varadarajan, *Insights into protein structure, stability and function from saturation mutagenesis.* Current Opinion in Structural Biology, 2018. **50**: p. 117-125.

26. Gavrilets, S., *Fitness landscapes and the origin of species*. Monographs in population biology. 2004, Princeton, N.J.: Princeton University Press. xviii, 476 p.

27. Dettman, J.R., et al., *Incipient speciation by divergent adaptation and antagonistic epistasis in yeast.* Nature, 2007. **447**(7144): p. 585-8.

28. de Visser, J.A.G.M. and S.F. Elena, *The evolution of sex: empirical insights into the roles of epistasis and drift.* Nature Reviews Genetics, 2007. **8**: p. 139.

29. Wagner, A., *Robustness and Evolvability in Living Systems*. 2005: Princeton University Press.

30. de Visser, J.A.G.M. and J. Krug, *Empirical fitness landscapes and the predictability of evolution.* Nature Reviews Genetics, 2014. **15**: p. 480.

31. Weinreich, D.M., R.A. Watson, and L. Chao, *Perspective: Sign epistasis and genetic constraint on evolutionary trajectories.* Evolution, 2005. **59**(6): p. 1165-74.

32. Bank, C., et al., *A Systematic Survey of an Intragenic Epistatic Landscape.* Molecular Biology and Evolution, 2015. **32**(1): p. 229-238.

33. Bank, C., et al., *On the (un)predictability of a large intragenic fitness landscape.* Proceedings of the National Academy of Sciences, 2016. **113**(49): p. 14085.

34. Sackman, A.M. and D.R. Rokyta, *Additive Phenotypes Underlie Epistasis of Fitness Effects.* Genetics, 2018. **208**(1): p. 339.

35. Olson, C.A., N.C. Wu, and R. Sun, *A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain.* Current biology : CB, 2014. **24**(22): p. 2643-2651.

36. Melamed, D., et al., *Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein.* RNA, 2013. **19**(11): p. 1537-1551.

37. Parera, M. and M.A. Martinez, *Strong Epistatic Interactions within a Single Protein.* Molecular Biology and Evolution, 2014. **31**(6): p. 1546-1553.

38. Araya, C.L., et al., *A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function.* Proceedings

of the National Academy of Sciences of the United States of America, 2012. **109**(42): p. 16858-16863.

39. Schenk, M.F., et al., *Patterns of Epistasis between beneficial mutations in an antibiotic resistance gene.* Mol Biol Evol, 2013. **30**(8): p. 1779-87.

40. Bendixsen, D.P., B. Ostman, and E.J. Hayden, *Negative Epistasis in Experimental RNA Fitness Landscapes.* J Mol Evol, 2017. **85**(5-6): p. 159-168.

41. Schenk, M.F., et al., *Patterns of Epistasis between Beneficial Mutations in an Antibiotic Resistance Gene.* Molecular Biology and Evolution, 2013. **30**(8): p. 1779-1787.

42. Ochman, H., A.S. Gerber, and D.L. Hartl, *Genetic applications of an inverse polymerase chain reaction.* Genetics, 1988. **120**(3): p. 621.

43. Pumir, A. and B. Shraiman, *Epistasis in a Model of Molecular Signal Transduction.* PLoS Computational Biology, 2011. **7**(5): p. e1001134.

44. Denver, D.R., et al., *High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome.* Nature, 2004. **430**: p. 679.

45. Tóth-Petróczy, Á. and D.S. Tawfik, *Protein Insertions and Deletions Enabled by Neutral Roaming in Sequence Space.* Molecular Biology and Evolution, 2013. **30**(4): p. 761-771.

46. Hashimoto, K. and A.R. Panchenko, *Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(47): p. 20352-20357.

47. Cooley, R.B., D.J. Arp, and P.A. Karplus, *Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices enhancing protein functionality.* Journal of molecular biology, 2010. **404**(2): p. 232-246.

48. Britten, R.J., *Transposable element insertions have strongly affected human evolution.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(46): p. 19945-19948.

49. Kauffman, S. and S. Levin, *Towards a general theory of adaptive walks on rugged landscapes.* Journal of Theoretical Biology, 1987. **128**(1): p. 11-45.

50. Leushkin, E.V., G.A. Bazykin, and A.S. Kondrashov, *Insertions and deletions trigger adaptive walks in Drosophila proteins.* Proceedings of the Royal Society B: Biological Sciences, 2012. **279**(1740): p. 3075-3082.

51. Mullaney, J.M., et al., *Small insertions and deletions (INDELs) in human genomes.* Human Molecular Genetics, 2010. **19**(R2): p. R131-R136.

52. Falini, B., et al., *Cytoplasmic Nucleophosmin in Acute Myelogenous Leukemia with a Normal Karyotype.* New England Journal of Medicine, 2005. **352**(3): p. 254-266.

53. Ye, K., et al., *Systematic discovery of complex insertions and deletions in human cancers.* Nature Medicine, 2015. **22**: p. 97.

54. Shortle, D. and J. Sondek, *The emerging role of insertions and deletions in protein engineering.* Current Opinion in Biotechnology, 1995. **6**(4): p. 387-393.

55. Mathonet, P., et al., *Active TEM-1 β-lactamase mutants with random peptides inserted in three contiguous surface loops.* Protein Science, 2009. **15**(10): p. 2323-2334.

56. Arpino, James A., et al., *Random Single Amino Acid Deletion Sampling Unveils Structural Tolerance and the Benefits of Helical Registry Shift on GFP Folding and Structure.* Structure(London, England:1993), 2014. **22**(6): p. 889-898.

57. Jackson, E.L., S.J. Spielman, and C.O. Wilke, *Computational prediction of the tolerance to amino-acid deletion in green-fluorescent protein.* PLoS ONE, 2017. **12**(4): p. e0164905.

58. Pascarella, S. and P. Argos, *Analysis of insertions/deletions in protein structures.* Journal of Molecular Biology, 1992. **224**(2): p. 461-471.

59. Marciano, D.C., N.G. Brown, and T. Palzkill, *Analysis of the plasticity of location of the Arg244 positive charge within the active site of the TEM-1 β-lactamase.* Protein Science, 2009. **18**(10): p. 2080-2089.

60. Crane, J.M. and L.L. Randall, *The Sec System: Protein Export in Escherichia coli.* EcoSal Plus, 2017. **7**(2): p. 10.1128/ecosalplus.ESP-0002-2017.

61. Kim, R. and J.-t. Guo, *Systematic analysis of short internal indels and their impact on protein folding.* BMC Structural Biology, 2010. **10**: p. 24-24.

62. Shenkin, P.S., B. Erman, and L.D. Mastrandrea, *Information-theoretical entropy as a measure of sequence variability.* Proteins: Structure, Function, and Bioinformatics, 1991. **11**(4): p. 297-313.

63. Marcos, M.L. and J. Echave, *Too packed to change: side-chain packing and site-specific substitution rates in protein evolution.* PeerJ, 2015. **3**: p. e911.

64. Amaral, M.D. and W.E. Balch, *Hallmarks of therapeutic management of the cystic fibrosis functional landscape.* Journal of Cystic Fibrosis, 2015. **14**(6): p. 687-699.

65. Bershtein, S., A.W.R. Serohijos, and E.I. Shakhnovich, *Bridging the physical scales in evolutionary biology: from protein sequence space to fitness of organisms and populations.* Current Opinion in Structural Biology, 2017. **42**: p. 31-40.

66. Horovitz, A., *Double-mutant cycles: a powerful tool for analyzing protein structure and function.* Folding and Design, 1996. **1**(6): p. R121-R126.

Curriculum Vitae

Courtney Elaine Gonzalez was born in El Paso, Texas on March 29, 1988. She graduated third in her class from Coronado High School (El Paso, Texas) in 2006. She attended the University of Utah (Salt Lake City, Utah) on a non-resident full academic scholarship from 2006 to 2010, where she did research under Dr. Leonard F. Pease on a novel technique to detect eosinophilic esophagitis. She graduated *cum laude* with an Honors Degree of Bachelor of Science in Chemical Engineering in May 2010. In the same year, she began her doctoral studies in Chemical and Biomolecular Engineering at Johns Hopkins University (Baltimore, Maryland) with a Whiting School of Engineering Dean's Fellowship. She joined the lab of Dr. Marc Ostermeier in November 2010. In 2012, she was awarded the NIH Ruth L. Kirschstein National Research Service Award Predoctoral Fellowship. In 2013, she was awarded the George M.L. Sommerman Engineering Graduate Teaching Assistant Award for the Whiting School of Engineering.