A GLOBAL EXPERIMENTAL ANALYSIS OF PROTEIN FUNCTION:
A CASE STUDY IN THE PDZ DOMAIN

APPROVED BY SUPERVISORY COMMITTEE

_____
Rama Ranganathan, M.D., Ph.D.

_____
Johann Deisenhofer, Ph.D.

_____
Michael Rosen, Ph.D.

_____
Leon Avery, Ph.D.

Dedication

This thesis is dedicated to my wife and all of my family.

Acknowledgements

The past years have been an exciting adventure, both personally and intellectually. I have many people to thank for all the ways they have contributed to my adventure. I want to thank my mentor Dr. Rama Ranganathan for providing me with an environment that allows freedom of thought without sacrificing the details of technical process, and for allowing me to work on a project that truly challenged and inspired me. His ability to identify and distill the important problems in biology brought me to the lab, and I hope I leave with a bit more of this ability than when I started. He also showed me that some of the most important developments can stem from scientific discussion over a stiff cocktail. I thank him for providing me a place to do creative, fun, and important research and developing in me the skills necessary to tackle fundamental problems in biology.

I do not have enough space to thank all of the individuals at UT Southwestern that have contributed to the success of my work here. I thank my thesis committee – Dr. Mike Rosen, Dr. Hans Deisenhofer, and Dr. Leon Avery – for their invaluable constructive criticism in the development and execution of this research and their attempts to keep me focused over the years. In the lab, Bill Russ has been my most trusted confidant, be it through lengthy discussions of data, expeditions to hygienically-questionable food spots, new life experiences, or wearing a suit when I asked him to. I thank him for convincing me to join this lab and for all his support since. Tina 'lab boy' Vo served as my lab mom and provided innumerable bits of support through her years here, be it pouring plates, cooking me dinner, or helping at my wedding; I send much love her way. Shan Mishra played an important role in my early days in the lab, providing discussions of science and all other matters and serving as a model of dedication. I thank Najeeb Halabi, Alan Poole, Heather Mishra, Walraj Gosal, Franciscus Jacobus Poelwijk, Subu Subramanian, my first year crew, and many others for their extensive support in so many ways through my time here. I also thank Dr. Tom Wilkie, Dr. David Mangelsdorf, Dr. Ward Wakeland, Dr. Ryan Potts, Angela Mobley, and Elizabeth Curry for their invaluable academic and scientific support.

My interest in biological research was solidified during my undergraduate years by a group of brilliant teachers at Trinity University who taught me the value and excitement of basic science research. For this, I thank Dr. David Ribble, Dr. Bob Blystone, and Dr. Jim Shinkle. Also during my undergraduate years, Dr. Donald Everett taught me the value of

pursuing a career in education based upon what you love to do; he also taught me the value of a well-made TandT. I think of his words of encouragement and wisdom more often than I ever thought I would; for this I recognize him in memory.

I have also been blessed with an amazing group of friends who have supported me through the years in all that I do, even if it includes going to school until the age of twenty-nine. I particularly thank Dr. Derek Thomas for teaching me molecular biology, the art of climbing, and the importance of a pint to a deep discussion of biology, and Travis Haley for pretending to be interested in my work and never hesitating to throw down.

Lastly, my family has undoubtedly been the most supportive and sustaining group in my life over the past years. I thank my parents for always encouraging my academic and personal pursuits and providing love and support in so many ways; my mother for always sacrificing herself to help her children in every way and teaching us to never stop laughing; my dad for showing me the importance of dedication to one's work and the value of a well-deserved beer after hard work. I thank my grandparents for always being proud of what I do; my grandfather for encouraging me to follow in his footsteps of an academic career focused on a subject about which you feel passionate; my grandmother for teaching me innate curiosity about the world, especially outside the confines of organized education. My sister Amanda has provided unending support as roommate, encourager, and a model of utter dedication and hard work. I thank her deeply for all her encouragement and wholeheartedly endorse her newfound love for Cell. I thank my extended family, the Reeds and the Hauensteins for their continuous encouragement and nourishment of my soul and my belly. I particularly thank Carmen for making me forget everything when we have tea.

In the end, absolutely none of this would have been possible without out my amazing, wonderful, beautiful, brilliant wife. My defense of this thesis comes almost ten years after we started dating, and every one of those years has been marked by her carrying me through adversity and us having lots of fun together. Her dedication to making our relationship and now marriage work from separate cities has been amazing. She has encouraged me to pursue a career that I love, not necessarily one that brings us riches, and for this and all of the infinite ways she makes me happy, I love her and I thank her.

A GLOBAL EXPERIMENTAL ANALYSIS OF PROTEIN FUNCTION: A CASE STUDY IN THE PDZ DOMAIN

By

Richard Noel McLaughlin Jr.

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

August, 2010

A GLOBAL EXPERIMENTAL ANALYSIS OF PROTEIN FUNCTION:
A CASE STUDY IN THE PDZ DOMAIN

Publication No. _____

Richard McLaughlin, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2010

Supervising Professor: Rama Ranganathan, M.D., Ph.D.

A complete understanding of the energetic architecture of a protein can be achieved only with a comprehensive description of the interaction of every amino acid with every other amino acid in a protein. Many efforts to understand the apparent complexity of protein function have attempted to address this problem with limited mutagenesis studies. A global computational description of amino acid interactions, Statistical Coupling Analysis has shown the existence of a contiguous subset of positions within a protein that displays significant co-evolution, termed protein sectors. Limited mutagenesis studies have shown sectors to be networks of higher-order interaction crucial for protein function; however, a theory of such global scope requires validation with a global experiment. Here, we design an assay system that measures the cellular function of a PDZ domain in a high-throughput and quantitative manner. We perform a comprehensive single amino acid mutagenesis experiment to show that most positions in the protein are robust to most mutations, and the set of positions that shows sensitivity to mutation is enriched for sector positions. Further, we perform a global pairwise epistasis experiment in which we measure the way in which every amino acid mutation in the PDZ domain feels the effect of a second mutation at a key specificity and affinity determining position in the peptide ligand of the PDZ domain. We find that those positions that show strong non-additivity in the context of the peptide mutation are all contained within the PDZ sector. Further, these sector positions that display strong

non-additivity all display the property of rapidly changing specificity upon mutation. That is, any mutation at these sector positions has a negative functional effect in the context of the endogenous peptide. However, these positions appear to be spring-loaded for change since these same mutations enhance function in the context of an alternative peptide. We hypothesize that proteins are robust as shown by their insensitivity to general mutation. However, proteins are simultaneously fragile as shown by their sensitivity to specific mutagenesis at sector positions. This fragility, however, is strongly coupled to evolvability as shown by the enhancement of alternative function endowed by these endogenously detrimental mutations.

Table of Contents

List of Figures

Abbreviations

Å	Angstrom
aTC	anhydro-tetracycline
bp	basepair
cAMP	cyclic adenosine monophosphate
Cdc42	cell division cycle 42
CRIPT	cysteine-rich interactor of PDZ3
DHFR	dihyrofolate reductase
DLG	discs large
DNA	deoxyribonucleic acid
dNTP	deoxy-nucleotide triphosphate
DSC	differential scanning calorimetry
eGFP	enhanced green Fluorescent Protein
FACS	fluorescence-activated cell sorting
GAL4	Galactose metabolism 4
GK	guanylate kinase
GPCR	G protein coupled receptor
GRIP-1	glutamate receptor-interacting protein 1
GST	glutathione S-transferase
hGH	human growth hormone
hGH-R	human growth hormone binding protein
IPTG	isopropyl-β-D-thiogalactopyranoside
ITC	Isothermal titration calorimetry
kcal	kilocalorie
K_d	Dissociation constant
MAGUK	membrane-associated guanylate kinase
MAPKK	mitogen activated protein-kinase kinase
NMCAA	next most common amino acid
µg	micrograms
ml	milliliters
µl	microliters
mM	millimolar

μM	micromolar
mol	mole
nM	nanomolar
NMR	nuclear magnetic resonance
O_R	repressor operator
P_{-2}	refers to -2 position of peptide (P_0 refers to carboxy terminal residue)
Par-6	phosphorylated after Rapamycin 6
Pbs2	Polymyxin B sensitivity 2
PCR	polymerase chain reaction
PDZ	PSD95, Discs-large, Zo-1
PDZ3	third PDZ domain from PSD95
PSD	post-synaptic density
RNA	ribonucleic acid
RTA	real-time analysis
SAP90	synapse associate protein 90, alternative name for PSD95
SCA	Statistical Coupling Analysis
SH3	Src homology 3 domain
Sho1	synthetic high osmolarity sensitive 1
T_m	melting temperature
TMR	tetra-methyl rhodamine
WT	wild-type

Chapter 1: Introduction

Every protein has purpose. Natural proteins have evolved to perform a specific function and retain this function through perturbation. Due to the nature of evolution, changes in a protein that lead to increases in fitness result in a higher prevalence of organisms containing that modified protein – natural selection. This feedback between the environment and random sequence perturbations at the protein level results in system changes chosen specifically for their interaction with the environment [1]. This iterative, long-timescale design process produces systems that differ fundamentally from engineered systems. Instead of a discrete design solution, evolving systems must change with their environment, or remain constant if their environment remains constant. Natural systems must therefore be built in such a way so as to facilitate this adaptation when necessary, yet preserve functionality when the environment remains constant.

The diversity of extant life and the fossil and phylogenetic records demonstrate that, through evolution, biological systems have retained the capacity to adapt to environmental changes – be they changes to a currently populated environment or expansion to a distinct environment. The ability of biological systems to adequately adapt to the current environment yet retain the ability to subsist through environmental changes underlies these observations of diversity. For example, the globin protein family contains sequences with as little as 16% sequence identity [2-3]. Despite this significant divergence, all these sequences preserve the core fold and heme-binding/oxygen transport function of the globins. Differences between the environments in which each of the globins functions certainly demand adaptive sequence changes, so some of this functional divergence can be attributed to the idiosyncrasies of each

protein's environment, independent of the conserved function of the family. For example, functional characterization of the hemoglobin protein encoded by the mammoth genome shows a small number of mutations that enable hemoglobin to offload oxygen at low temperatures [4]. These sequence changes likely resulted in a direct fitness advantage to the mammoth, but would provide no benefit to mesophilic animals.

Though some sequence change may be driven by selective pressure, much of the difference between the globins is likely neutral sequence change. This large sequence space that still retains globin function suggests that these proteins may be functionally robust to sequence changes. Importantly, it is difficult to demonstrate whether only a small number of amino acids at particular positions (16% in this case) is sufficient to encode globin function or whether there are many solutions to designing a sequence that functions as a globin. Heterologous complementation studies in a number of protein families have shown that a homologous protein from another, sometimes distantly related organism can support normal laboratory growth when swapped for an organism's endogenous allele [5-6]. This and other experimental data suggest that many proteins can tolerate significant sequence change while retaining function.

While phylogenetic and experimental data show the tolerance of many proteins to random mutations, these same data also demonstrate that proteins are particularly susceptible to functional modification by mutations at particular positions. For example, a single mutation to the beta chain of hemoglobin causes human sickle-cell anemia (an allele with adaptive function in the context of plasmodium infection) [7]. Further, a number of human diseases are known to be inherited in a Mendelian fashion, showing that they result from modification of a single gene [8]. This idea of functional heterogeneity across the primary structure was observed globally from the first comparison of globin sequences by Pauling and Zuckerkandl:

"There is no reason to expect that the extent of functional change in a polypeptide chain is proportional to the number of amino acid substitutions in the chain. Many such substitutions may lead to relatively little functional change, whereas at other times the replacement of one single amino acid residue by another may lead to a radical functional change," and "the functional effect of a given single substitution will frequently depend on the presence or absence of a number of other substitutions [9]*."*

These examples clearly illustrate the fact that the patterns of functional effects from amino acid mutations are highly heterogeneous; however, the mechanisms and underlying design principles that enable these characteristics remain poorly understood. The basis of all these observations is the fact that amino acids interact in an energetically heterogeneous manner, but it is the particular distribution and pattern of such effects that underlies the robust and adaptive properties of natural proteins – results of the iterative selection of evolution over very long timescales. In order to understand these patterns of heterogeneity, and eventually the nature of the processes that built these patterns, one needs a method to comprehensively analyze the contribution of each amino acid of a protein to the fitness conferred by that protein.

Motivated by these observations of proteins, the goals of my thesis research are to globally measure the functional contribution of every amino acid through single and pairwise mutation analysis, as well as describe the patterns and distributions of these effects and their correlation with the patterns of statistical co-evolution in the protein. These goals are achieved through the following specific aims:

1) Develop a high-throughput, quantitative measurement of protein function in a cellular environment that recapitulates, as closely as possible, the fitness constraints on that protein in the context of an entire organism.

2) Perform a perturbation analysis to observe the dependence of measured cellular function effects on the measured biophysical properties of subtle mutations at every position of a protein
3) Understand the patterns of cellular function effects in proteins through a comprehensive measurement of single mutation effects and the non-additive effects of a global pairwise mutation analysis.
4) Correlate the positions important for cellular function, as evidenced by single mutation effects and non-additive pairwise mutation effects, with the positions in the protein family that display strong statistical co-evolution with other positions.

Previous methods aimed at understanding how the atomic properties of the protein, manifested as the amino acid sequence, determine the function of the protein include structure determination, targeted and random mutagenesis, and computational analysis of protein sequences. The remainder of this introduction presents a critical review of the existing literature in these areas.

Structural studies reveal the apparent homogeneity of proteins

Structural analysis of biological macromolecules has revolutionized the field of molecular biology by providing detailed insight into the basis of the molecular mechanisms of macromolecular function. The first structural studies of hemoglobin revealed an assembly of well packed amino acids, each interacting with several other amino acids, the exclusion of polar amino acids from the core, the alternating polar - non-polar nature of helices, and the helix-breaking nature of proline [2]. Later studies would show that the mean density of proteins approached that of organic crystals of free amino acids [10-11]. Further, each amino acid makes about the same number of contacts with other amino acids, and the packing density in the core of

many proteins is uniformly high [12-14]. These observations together suggest a homogeneous architecture of amino acids in which each amino acid contributes to the fold and presumably the function of the protein (if one assumes that structure begets function).

Despite these first-order observations of protein form, the basic principles of allostery and the studies discussed above and in more detail in the following sections clearly demonstrate a heterogeneous pattern of functional constraints. This heterogeneity is evident in the diversity of sequences and positional conservation in the globin alignments, the importance of higher-order amino acid interactions as shown in mutagenesis studies, and countless structural and mutational examples of protein allostery. However, the general ideas of the spatial proximity model, in which the energetic interaction of two residues scales with the distance between them, persist in the experimental, computational, and protein design fields. For example, most *ab initio* protein design and mutation-effect prediction algorithms work on the principle of optimizing the packing interactions of amino acids [15-17]. These perspectives are likely not from a lack of appreciation of allosteric effects in proteins, but probably stem from the apparent complexity inherent to higher-order and long-range interactions. Experimental studies arguing for the unique importance of the local environment suffer from similar limitations when they attempt to make global conclusions from limited datasets. In one striking example, a survey of a limited set of thermodynamic couplings measures a significant interaction between two positions separated by many angstroms, but this interaction is labeled as an outlier of the data [18]. Though these local-environment approaches may be tractable for protein design and standard experimental strategies, they do not recapitulate the design principles of natural proteins built through the iterative process of evolution.

Biophysical studies reveal the complexity of amino acid interactions

Heterogeneity of the local environment

For proteins, X-ray crystallography and NMR structure determination have shown the precise coordination of chemistry at an active site or ligand-binding pocket necessary to engage a specific substrate or catalyze reactions. Such studies in a variety of enzymes and binding-proteins have revealed those residues that participate directly in the catalytic mechanism or engagement of ligand [19-21]. However, the crystallographic proximity of an amino acid side chain to substrate or ligand does not necessarily implicate that amino acid as functionally important. More directed examinations of the interaction interface of two proteins clearly demonstrate this non-uniform distribution of functional effects. Mutation studies of the interaction of human growth hormone (hGH) with its receptor (hGHbp) revealed such a non-uniform pattern of energetic contributions of individual amino acids to the function of the protein [22]. Despite the interaction of nearly thirty side chains from each protein, the majority of the binding energy is contributed by two tryptophan residues within a hydrophobic patch of hGHbp. On the other side of the interface, hGH exhibits a similar heterogeneity with striking complementarity to the tryptophans of the receptor. In fact, mutation of complementary positions on hGH could rescue the interaction with hGHbp containing a deletion of one of the critical tryptophans [23]. Critically, this architecture in which only a few amino acids contribute strongly to the interaction of two proteins is not obvious from the structure of the complex [24]. The structural picture in which many amino acids interact at the interface suggests a more homogeneous architecture in which every interaction contributes a small amount to the net interaction energy.

Interaction at a distance

While mutagenesis studies guided by the proximity of a given amino acid to the active site provide information about the functional importance of a subset of amino acids, these approaches neglect the functional contribution of long range effects from amino acids located outside of the active site. Studies of the structural changes induced by mutation are powerful for identifying long range effects in proteins since they have the potential to visualize the entire molecule. For example, structures of RNase-S in which mutations are introduced into position 13 of the S-protein subunit showed a small number of significant structural changes relative to the wild-type complex. However, one of the most dramatic structural changes occurred 25Å away from the site of mutation [25]. This long-range structural coupling was observed due to the visualization of the complete structure in the wild-type and mutant backgrounds. This coupling may be observable using mutagenesis studies, but there would be no *a priori* motivation for investigating the coupling these two particular regions. High resolution structure determination allows the identification of such long-range couplings.

Though advancements have been made towards higher throughput [26-27], high-resolution structure determination still requires significant purified protein and labor to screen crystal conditions. Even if such techniques were not limited by throughput, the critical limitation to structure determination as a method to understand the interaction of amino acids is that energy cannot be seen in the structure. That is, even though distant structural changes may occur as a result of a certain mutation, the residues important for effecting that structural change cannot necessarily be discerned from crystal structures. This is due largely to the fact that the free energy of folding is the sum of many forces across the structure which act with a steep distance

dependence and sum to close to zero. In addition, structural changes do not necessarily imply functional changes which must be measured with a complementary technique.

Mutagenesis reveals the long-range interaction of amino acids

In order to understand the contribution of particular amino acids to particular aspects of a protein's function, many groups utilize the strategy of site-directed mutagenesis. This approach consists of creating a mutation at a particular position of interest in a protein and observing the effect of that mutation according to some measurable characteristic of the protein, relative to the wild-type protein. Many studies utilize this approach to attempt to understand the importance of a particular amino acid residue to the function of a particular protein. These studies typically sample only a small number of mutations in or around the active site or binding site of the protein of interest. This strategy of spatial proximity-based mutagenesis has been used in many experiments, including the hGH-hGHbp studies discussed above, that demonstrate the heterogeneous contribution to function of the amino acids that constitute the binding site in the structure [28]. A recent examination of the contribution of individual residues to the specificity and affinity of PDZ domains mutated any residue within 4.5Å of the peptide ligand in the structure of nine different PDZ domains as a means to determine the conserved aspects of peptide binding [29]. Some groups even attempt to show that the magnitude of the effect of a mutation on another position decreases with distance from the site of mutation, suggesting that the best predictor of the energetic effect of a mutation is its proximity to the active site of the protein [18].

While it is true that many of the positions that contribute to the function of a protein are located in or near the active site, there are many examples of functionally important positions located at a significant distance from the active site, and a complete understanding of the energetic architecture of a protein requires the description of all the energetically important interactions amongst amino acids. For example, in an analysis of the interaction of residues in the activation gate and the selectivity filter of the *shaker* voltage-gated potassium channel, the authors observe significant cooperativity between these two regions of the protein that exist at significant distance from each other in the crystal structure [30]. In the T-cell receptor variable domain, a portion of the receptor crucial for recognizing MHC-bound antigens, amino acids from two domains located more than 20Å apart display energetic cooperativity for ligand binding [31]. In DHFR, mutation of a critical residue in the substrate binding pocket (D27S) significantly reduces catalytic activity. This effect can be partially restored by a mutation to a surface residue 15Å away from position 27 [32]. Serine proteases [33], hemoglobin [34], and antibody Fab fragments [35] also contain functionally important positions at significant distance from the active site. In every one of these examples, these positions could not be predicted as functionally important from their structural proximity to the active site. Overall, the binding site of a molecule and those amino acids that comprise or pack against the binding site are crucially important for function. In fact, on average this portion of the molecule is most likely to contain functionally important positions. However, in order to comprehensively understand the interaction of amino acids that comprises a functional protein, one cannot be concerned with only the average patterns of functional importance – every interaction that contributes significantly to the function of a protein must be understood, whether that interaction is within packing distance or many angstroms away.

Large scale mutagenesis of proteins

In addition to studying the effect of specific mutations in order to understand the functional mechanism of a specific protein, other studies employ more widespread mutagenesis with the goal of understanding the general biophysical properties of proteins. These experimental strategy attempt to explain: 1) the molecular-level physical properties that influence protein function including stability, binding, and expression level; 2) the atomic-level physical properties of the protein that influence protein function including secondary structure, surface exposure, and conservation; 3) the observation that proteins are robust to mutations; 4) the design principles of natural proteins created through evolution.

Mutagenesis to understand the distribution of functional effects

The robustness of natural proteins to random mutagenesis

As the object of constant mutagenesis and selection for function, natural proteins must be able to buffer the effects of mutation in order to maintain molecular function and support organismal fitness. For a protein functioning within an organism, there is a range of function levels within which the fitness of the organism is unaffected. This function level could be the flux through a metabolic pathway, the catalytic rate of an enzyme, or the affinity of a binding protein for ligand. On the molecular scale for a single protein, robustness means that the long term change in organismal fitness as a function of acquiring mutations is close to zero. This could mean that most mutations have an effect small enough to maintain the protein function within the neutral range of function levels, or that beneficial mutations occur at a significant

frequency such that they offset the deleterious mutations and over time fitness remains relatively unchanged.

Experimentally, introducing many mutations into a protein and measuring the distribution of function effects could provide the demonstration and explanation of robustness in natural proteins. Such studies have been performed in a number of model systems including, most prominently, TEM-1 β -lactamase, barnase, staphylococcal nuclease, T4 lysozyme, and hen lysozyme.

TEM-1 β-lactamase

Due to the role of antibiotic resistance in human disease, β-lactamase has attracted significant attention as a model system for understanding the relationship between amino acid sequence and protein function [36-37]. The Palzkill laboratory performed experiments in which sets of three contiguous codons were mutated across the entire primary structure of TEM-1 β-lactamase in an effort to understand comprehensively the contribution of individual amino acids to the protein's penicillinase activity [38]. Libraries of mostly double and single mutants were selected on solid medium containing 1mg/ml ampicillin, and approximately 1000 colonies were sequenced spanning the entire primary structure. Of 263 positions, only 43 tolerated no other amino acid besides the wild-type residue. Many of the intolerant positions lie in the active site or pack against catalytic residues. About 25% of the intolerant positions are found in the core or in the hydrophobic region at the interface of the two domains of TEM-1, but many similarly buried positions do tolerate mutation. Further, a number of positions scattered throughout the structure – not buried, not near the active site – show no functional substitutions. Highly conserved

positions tend to be less substitutable in the experiment, but the overall contingency of residue variance on conservation is weak.

Despite the large sampling of mutations across all positions in β-lactamase, this study does little more than show in this model system the same principles that have been shown in other systems including bacteriophage T4 lysozyme [39], *E. coli* lac repressor [40], and bacteriophage f1 gene V [41-42]: the immutable nature of a small percentage of residues, typically those residues around the active site, and the correlation of immutable positions with buried and conserved residues. Further, the technical approach of mutating three codons simultaneously significantly limits the interpretation of any individual mutation.

One strong aspect of this paper is the choice of model system. As discussed more below, the selection for protein function occurs *in vivo*, and the function of β-lactamase can be directly related to organismal fitness for bacteria growing in the presence of ampicillin. Though, more comprehensive definitions of function might measure the specificity effects of these mutations, since the ability to hydrolyze molecules besides ampicillin with only a few mutations might be an important aspect of fitness.

Barnase

Barnase is a 110 amino acid protein secreted by *Bacillus amyloliquefaciens* that possesses ribonuclease activity [43]. Presumably, barnase functions as a means to scavenge nucleotides from free RNA in the environment, but may also play a role as a defense against other soil organisms [44]. The bacterium inhibits the toxic ribonuclease activity of barnase by producing the barstar protein intracellularly which binds very tightly to barnase ($K_d = 10^{-14}$ M), occluding its active site [45]. Extensively studied as a model system for protein folding, several groups

have characterized a significant number of mutants for their effect on folding as well as their effect on the stability and ribonuclease activity of barnase. Mutations made across the structure show a broad range of stability effects as measured by urea denaturation, with more buried residues tending to have a greater effect on stability [46-47]. Substitution of hydrophobic sidechains with alanine tended to have a large stability effect, but the effect of hydrogen-bond disruption was highly variable. Overall, the average effect of a mutation was highly variable, but the vast majority of mutations assayed were destabilizing.

When assayed for ribonuclease activity, instead of stability, using a screen with a very low activity threshold (defined as greater than the RNA hydrolysis rate of non-enzyme catalysts), barnase displays significant robustness to mutation [48], similar to that seen in other proteins such as hen lysozyme [49]. Of the 110 positions in barnase, only 15 positions display any mutation that abolishes function to below the threshold level. Assuming the mutagenesis to be comprehensive, at 95 positions in barnase, mutation to any other amino acid preserves some ribonuclease activity. Those substitutions that abolish function either modify the active site, introduce a non-hydrophobic amino acid into the core, or create unfavorable backbone constraints. This picture of resistance to complete inactivation differs from the profile of stability disruption in that a very small percentage of mutations abolish function (5% of all substitutions) relative to the percentage of mutations that destabilize barnase (almost all substitutions) (**Figure 1.1**).

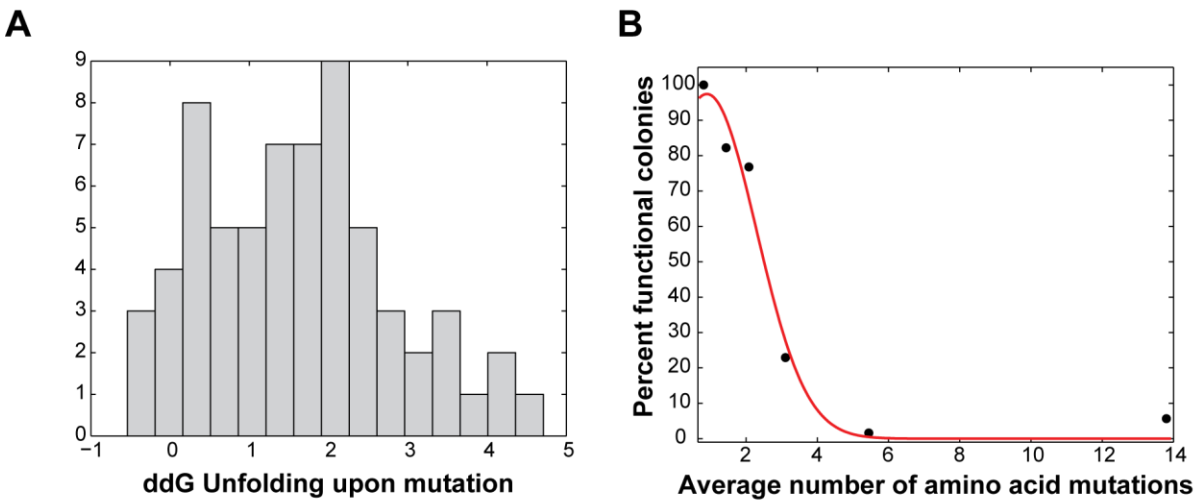


Figure 1.1 Stability versus Functional Effects of Mutations
A) In barnase, most of the sixty-five mutations made across the primary structure result in a decrease the free energy of unfolding of the protein as measured by reversible urea denaturation. B) When the fraction of clones that produce lytic halos in populations of randomly mutated hen lysozymes is measured, almost all proteins retain function with one mutation, but only twenty percent remain functional with four mutations on average. Figure created with data adapted from [42, 45].

Several important differences need to be noted between these studies. First, the stability studies generate a relatively small number of mutations, 64 total mutations, relative to the number of mutations sampled in the activity studies, estimated as almost all possible single mutations. This discrepancy in sample size also results directly from the technical constraints of measuring stability – protein expression, purification, and urea denaturation – versus the technical constraints of measuring ribonuclease activity – PCR with mutagenic oligos, ligation, transformation, and plating. Also, the mutations sampled in the stability studies were chosen due to their apparent importance in the structure; that is, they are buried or form interactions that appear likely to stabilize the tertiary structure.

State of the mutagenesis literature

The mass mutagenesis studies from the literature contain important data regarding the distribution of functional effects in many protein model systems and the idiosyncratic physical constraints that may play an important role in individual instances of large mutational effects. However, the design of each study contains some limitation that prevents clear and complete analysis of the data with the intent of generally describing the energetic architecture of a protein.

The proper description of the energetic architecture of a protein entails the measurement of the interaction of every amino acid with every other amino acid. The approach of previous studies has been to make many mutations at many positions and measure the effect of each mutation. However, no experiment has measured the individual function of every mutation at every amino acid. Even with the design of an experiment that permits the measurement of every single-mutation in a given protein, two additional crucial aspects of protein function must be considered. First, the true measure of the effect of a mutation would be the quantification of the fitness effect at the organismal level created by a given single mutation at the protein level. Second, due to the interaction of amino acids, a mutation may exhibit a different functional effect in the background of a mutation relative to its effect in the wild-type background. These, non-additive energetic effects of more than one mutation are referred to as pairwise mutational effects [50]. In this thesis, we refer to any pairwise or higher energetic interaction of amino acids as higher-order. As demonstrated through structural and energetic studies, the higher-order interaction of amino acids is crucially important for the function of proteins [30, 51]. As such, any single-mutant analysis of proteins fails to describe the complete energetic constraints of a protein. A truly complete description would contain higher-order mutation effects as well. Last, all previous attempts to describe the energetic architecture of a protein have been largely

exploratory, that is, not based upon a model of the expected energetic architecture of proteins designed through the iterative process of natural selection.

Previous work on the sector model of protein function:

Previous work from our lab sought a general description of the functionally important amino acid interactions in a protein by computing the patterns of co-variation of positions in large and diverse sequence alignments. Based upon the logic that the most important amino acid interactions should be conserved in a protein family, Statistical Coupling Analysis (SCA) calculates the degree of co-variation between two positions in a diverse alignment of homologous protein sequences. SCA reveals a surprisingly sparse and heterogeneous network of spatially contiguous amino acids in any given protein family (**Figure 1.2**). That is, at the level of the protein family most positions evolve independently while a limited subset of all positions displays significant co-evolution. The importance of these conserved networks of coevolving amino acids, termed protein sectors, has been shown in an array of protein families by limited mutagenesis of sector and non-sector positions [52-56]. The more in-depth analysis of the function of three independent sectors in the serine protease family led to the hypothesis that sectors represent the core contributors to biological function [57].

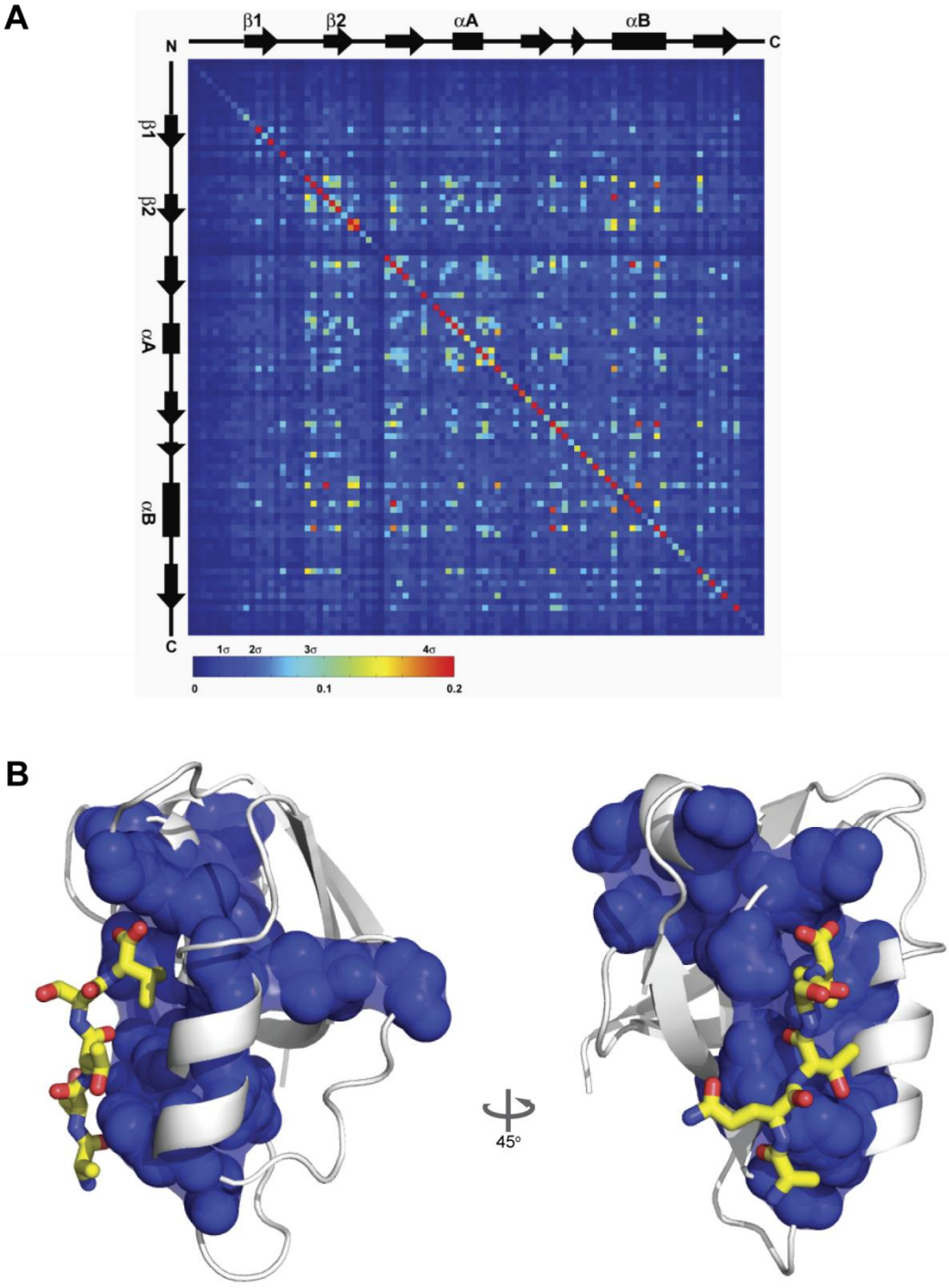


Figure 1.2 Statistical Coupling Analysis of the PDZ Family

A) The statistical coupling matrix displays the covariation between all positions in a diverse alignment of PDZ domains. Highly covarying positions, a small fraction of all positions, are represented by hot pixels. Independently varying positions, most of the positions, are displayed in cool colors. B) Those positions that display high covariation are represented as spheres on the cartoon representation of PDZ3 (PDB ID, 1BE9). The highly coupled positions form a contiguous network that spans the peptide binding surface and extends to three surfaces on the back side of the domain.

In addition to the functional importance of sector positions, the general pattern of highly covarying positions in an array of protein families begs the question of how and why natural proteins have been built to display these patterns. As discussed above, the properties of an evolving system include the ability to maintain a constant output in the face of perturbation (robustness), while with the same design architecture retaining the ability to generate novel functionality in a short mutational walk when the selective pressure of the environment demands (evolvability). The sparse and heterogeneous pattern of amino acid co-evolution suggests an interesting hypothesis for the origin and existence of sectors in natural proteins. The statistical independence of most positions may represent the decoupling of these positions from engaging in higher-order interactions, thereby decreasing the likelihood that a mutation to one of these positions would disrupt function. Some positions may have large intrinsic effects, but the effect of a mutation at an independently-evolving site should not be propagated to other positions. Since the majority of the protein displays such independence, the average functional change in response to a random mutation would be small, giving rise to the observation of protein robustness. In contrast, the network of highly-coevolving positions may represent the means to achieve a significant functional change with a limited number of mutations. The effect of a mutation would propagate to the positions that display covariance with a given position, thereby producing a larger magnitude effect than a mutation at a decoupled position. The first step in validating this model would be to demonstrate that the pattern of functional effects in a protein correlates with the sector positions. Further, the sector positions should be specifically enriched for higher-order interactions of functional importance.

Conclusions

This thesis describes a research project designed to comprehensively measure the interaction of the amino acids of a protein in order to more clearly understand the patterns of design constraints in natural proteins as a means to eventually understand the evolutionary processes that generate such patterns.

Overall, this thesis describes experiments and analysis encompassing:

1) The development an in-vivo assay for PDZ function that as closely as possible recapitulates the organismal context of the protein with the capability of quantitatively measuring the cellular function of thousands of PDZ mutants.
2) The measurement of the effect on stability, binding, and cellular function of subtle single-mutations in PSD95-PDZ3 in order to describe the individual importance of biophysical parameters to the cellular function of PSD95-PDZ3.
3) The measurement of the effect on cellular function of all single-mutations in PSD95-PDZ3.
4) The measurement of the coupling of all positions in PSD95-PDZ3 to a key specificity- and affinity-determining residue of the protein's peptide ligand.
5) A demonstration, for the first time, of the connection between the statistically coevolving positions in the PDZ family and the functionally important positions in a single PDZ family member and a comprehensive experimental validation of the sector model of protein function.

References

1. Hartwell, L.H., et al., *From molecular to modular cell biology.* Nature, 1999. **402**: p. C47-52.
2. Perutz, M.F., J.C. Kendrew, and H.C. Watson, *Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence.* Journal of Molecular Biology, 1965. **13**(3): p. 669-678.
3. Bashford, D., C. Chothia, and A.M. Lesk, *Determinants of a protein fold : Unique features of the globin amino acid sequences.* Journal of Molecular Biology, 1987. **196**(1): p. 199-216.
4. Campbell, K.L., et al., *Substitutions in woolly mammoth hemoglobin confer biochemical properties adaptive for cold tolerance.* Nat Genet, 2010. **42**(6): p. 536-540.
5. Frommer, W.B. and O. Ninnemann, *Heterologous Expression of Genes in Bacterial, Fungal, Animal, and Plant Cells.* Annual Review of Plant Physiology and Plant Molecular Biology, 2003. **46**(1): p. 419-444.
6. Wyckoff, E. and T.S. Hsieh, *Functional expression of a Drosophila gene in yeast: genetic complementation of DNA topoisomerase II.* Proceedings of the National Academy of Sciences of the United States of America, 1988. **85**(17): p. 6272-6276.
7. Aidoo, M., et al., *Protective effects of the sickle cell gene against malaria morbidity and mortality.* The Lancet, 2002. **359**(9314): p. 1311-1312.
8. Antonarakis, S.E. and J.S. Beckmann, *Mendelian disorders deserve more attention.* Nat Rev Genet, 2006. **7**(4): p. 277-282.
9. Zuckerkandl, E.P., L, *Evolutionary divergence and convergence in proteins*, in *Evolving genes and proteins*, V.V. Bryson, HJ, Editor. 1965, Academic Press: New York. p. 357-417.
10. Richards, F.M., *The interpretation of protein structures: Total volume, group volume distributions and packing density.* Journal of Molecular Biology, 1974. **82**(1): p. 1-14.
11. Chothia, C., *Structural invariants in protein folding.* Nature, 1975. **254**(5498): p. 304-308.
12. Sharma, R., *Logic and Mechanism of Evolutionarily Conserved Interaction in PDZ Domains*. 2004, University of Texas Southwestern Medical Center.
13. Richards, F.M. and W.A. Lim, *An analysis of packing in the protein folding problem.* Quarterly Reviews of Biophysics, 1993. **26**(04): p. 423-498.
14. Tsai, J., et al., *The packing density in proteins: standard radii and volumes.* Journal of Molecular Biology, 1999. **290**(1): p. 253-266.
15. Bradley, P., K.M.S. Misura, and D. Baker, *Toward High-Resolution de Novo Structure Prediction for Small Proteins.* Science, 2005. **309**(5742): p. 1868-1871.
16. Kuhlman, B., et al., *Design of a Novel Globular Protein Fold with Atomic-Level Accuracy.* Science, 2003. **302**(5649): p. 1364-1368.
17. Schymkowitz, J., et al., *The FoldX web server: an online force field.* Nucl. Acids Res., 2005. **33**(suppl_2): p. W382-388.
18. Chi, C.N., et al., *Reassessing a sparse energetic network within a single protein domain.* Proceedings of the National Academy of Sciences, 2008. **105**(12): p. 4679-4684.
19. Zhang, Z.-Y., *Protein-Tyrosine Phosphatases: Biological Function, Structural Characteristics, and Mechanism of Catalysis.* Critical Reviews in Biochemistry and Molecular Biology, 1998. **33**(1): p. 1-52.
20. Perona, J.J. and C.S. Craik, *Structural basis of substrate specificity in the serine proteases.* Protein Science, 1995. **4**(3): p. 337-360.
21. Cowburn, D., *Peptide recognition by PTB and PDZ domains.* Current Opinion in Structural Biology, 1997. **7**(6): p. 835-838.
22. Clackson, T. and J. Wells, *A hot spot of binding energy in a hormone-receptor interface.* Science, 1995. **267**(5196): p. 383-386.

23. Atwell, S., et al., *Structural Plasticity in a Remodeled Protein-Protein Interface.* Science, 1997. **278**(5340): p. 1125-1128.
24. Jain, R.K. and R. Ranganathan, *Local complexity of amino acid interactions in a protein core.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(1): p. 111-116.
25. Varadarajan, R. and F.M. Richards, *Crystallographic structures of ribonuclease S variants with nonpolar substitution at position 13: packing and cavities.* Biochemistry, 1992. **31**(49): p. 12315-12327.
26. Lesley, S.A., et al., *Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(18): p. 11664-11669.
27. Heinemann, U., G. Illing, and H. Oschkinat, *High-throughput three-dimensional protein structure determination.* Current Opinion in Biotechnology, 2001. **12**(4): p. 348-354.
28. Clackson, T.W., JA, *A hot spot of binding energy in a hormone-receptor interface.* Science, 1995. **267**(5196): p. 383-6.
29. Tonikian, R., et al., *A Specificity Map for the PDZ Domain Family.* PLoS Biol, 2008. **6**(9): p. e239.
30. Sadovsky, E. and O. Yifrach, *Principles underlying energetic coupling along an allosteric communication trajectory of a voltage-activated K+ channel.* Proceedings of the National Academy of Sciences, 2007. **104**(50): p. 19813-19818.
31. Moza, B., et al., *Long-range cooperative binding effects in a T cell receptor variable domain.* Proceedings of the National Academy of Sciences, 2006. **103**(26): p. 9867-9872.
32. Brown, K.A., E.E. Howell, and J. Kraut, *Long-range structural effects in a second-site revertant of a mutant dihydrofolate reductase.* Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**(24): p. 11753-11756.
33. Perona, J.J., et al., *Structural origins of substrate discrimination in trypsin and chymotrypsin.* Biochemistry, 1995. **34**(5): p. 1489-1499.
34. Perutz, M.F., et al., *The Stereochemical Mechanism of the Cooperative Effects in Hemoglobin Revisited.* Annual Review of Biophysics and Biomolecular Structure, 1998. **27**(1): p. 1-34.
35. Patten, P.A., et al., *The Immunological Evolution of Catalysis.* Science, 1996. **271**(5252): p. 1086-1091.
36. Majiduddin, F.K., I.C. Materon, and T.G. Palzkill, *Molecular analysis of beta-lactamase structure and function.* International Journal of Medical Microbiology, 2002. **292**(2): p. 127-137.
37. Frère, J.-M., *Beta-lactamases and bacterial resistance to antibiotics.* Molecular Microbiology, 1995. **16**(3): p. 385-395.
38. Huang, W., et al., *Amino Acid Sequence Determinants of β-Lactamase Structure and Activity.* Journal of Molecular Biology, 1996. **258**(4): p. 688-703.
39. Rennell D, B.S., Hardy LW, Poteete AR, *Systematic mutation of bacteriophage T4 lysozyme.* J Mol Biol, 1991. **222**(1): p. 67-88.
40. Markiewicz, P., et al., *Genetic Studies of the lac Repressor. XIV. Analysis of 4000 Altered Escherichia coli lac Repressors Reveals Essential and Non-essential Residues, as well as "Spacers" which do not Require a Specific Sequence.* Journal of Molecular Biology, 1994. **240**(5): p. 421-433.
41. Terwilliger, T.C., et al., *In Vivo Characterization of Mutants of the Bacteriophage f1 Gene V Protein Isolated by Saturation Mutagenesis.* Journal of Molecular Biology, 1994. **236**(2): p. 556-571.
42. Bowie, J., et al., *Deciphering the message in protein sequences: tolerance to amino acid substitutions.* Science, 1990. **247**(4948): p. 1306-1310.

43. Nishimura, S.N., M., *Ribonuclease of Bacillus subtilis.* Biochimica et biophysica acta, 1958. **30**(2): p. 430-1.
44. Hartley, R.W., *Barnase and barstar: two small proteins to fold and fit together.* Trends in Biochemical Sciences, 1989. **14**(11): p. 450-454.
45. Schreiber, G. and A.R. Fersht, *Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering.* Biochemistry, 1993. **32**(19): p. 5145-5150.
46. Serrano, L., et al., *The folding of an enzyme : II. Substructure of barnase and the contribution of different interactions to protein stability.* Journal of Molecular Biology, 1992. **224**(3): p. 783-804.
47. Serrano, L., A.G. Day, and A.R. Fersht, *Step-wise Mutation of Barnase to Binase : A Procedure for Engineering Increased Stability of Proteins and an Experimental Analysis of the Evolution of Protein Stability.* Journal of Molecular Biology, 1993. **233**(2): p. 305-312.
48. Axe, D.D., N.W. Foster, and A.R. Fersht, *A Search for Single Substitutions That Eliminate Enzymatic Function in a Bacterial Ribonuclease.* Biochemistry, 1998. **37**(20): p. 7157-7166.
49. Kunichika, K., Y. Hashimoto, and T. Imoto, *Robustness of hen lysozyme monitored by random mutations.* Protein Eng., 2002. **15**(10): p. 805-810.
50. Serrano, L., et al., *Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles.* Biochemistry, 1990. **29**(40): p. 9343-9352.
51. Hidalgo, P. and R. MacKinnon, *Revealing the architecture of a K+ channel pore through mutant cycles with a peptide inhibitor.* Science, 1995. **268**(5208): p. 307-310.
52. Hatley, M.E., et al., *Allosteric determinants in guanine nucleotide-binding proteins.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(24): p. 14445-14450.
53. Lockless, S.W. and R. Ranganathan, *Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families.* Science, 1999. **286**(5438): p. 295-299.
54. Shulman, A.I., et al., *Structural Determinants of Allosteric Ligand Activation in RXR Heterodimers.* Cell, 2004. **116**(3): p. 417-429.
55. Smock, R.R., O; Russ, WP; Swain, JF; Leibler, S; Ranganathan, R; Gierasch, LM, *An Interdomain Sector Mediating Allostery in Hsp70 Molecular Chaperones.* Molecular Systems Biology, 2010. **In Press**.
56. Suel, G.M., et al., *Evolutionarily conserved networks of residues mediate allosteric communication in proteins.* Nat Struct Mol Biol, 2003. **10**(1): p. 59-69.
57. Halabi, N., et al., *Protein Sectors: Evolutionary Units of Three-Dimensional Structure.* Cell, 2009. **138**(4): p. 774-786.

Chapter 2: Development of a high-throughput assay for protein function

In order to generally explore the patterns of functional constraints in proteins, we needed to design an assay system that could quantitatively measure the function of many mutations in a single protein. The optimal assay system would allow measurement of hundreds or more mutant forms of a single protein in a quantitative and high-throughput fashion. The protein studied should be small enough to allow a comprehensive investigation, but not so small as to raise concerns over generality of the system. Importantly, the assay should not focus on a limited aspect of the function of the chosen protein, such as stability or structure alone, as many previous studies have done. Instead, the assay should as closely as possible recapitulate all of the constraints on the protein's function in the context of an organism.

Choice of a protein model system

A variety of considerations went into the choice of a particular model system for these experiments. We wanted to choose a system in which we were minimally limited by the generalization we could make to all proteins through studying a single system. At the same time, our choice of model system was intimately linked to the methods available for assaying the function of that protein model system. The following section outlines the benefits and detriments of the candidate model systems and our rationale for choosing the PDZ domain.

β-lactamase

β-lactamase enzymes hydrolyze β-lactam antibiotics including ampicillin and penicillin. In the context of human health, these enzymes represent the major mode of antibiotic resistance in *E. coli* and other bacteria [1]. The fact that antibiotic resistance genes directly determine the fitness of a bacterium in the presence of an antibiotic makes this protein family one of only a few in which the function of the protein can be directly related to the fitness of the organism. This represents a significant advantage in that we could make direct statements about fitness, defined in this case as the instantaneous reproductive rate, without an assumption about how the function of a particular protein might contribute to the fitness of the organism. However, even with antibiotic resistance genes like TEM-1 β-lactamase, there is no clear understanding of the relevance of these genes in natural environments [2], so the extension of laboratory results to an ecological context would be tenuous.

Despite this important advantage of directly impacting the bacterium's reproductive rate, β-lactamase has been extensively investigated in mutational studies, probably due to the ease of screening for functional variants. As such, further mutation studies in this system would be measured against these extensive and often difficult-to-interpret studies. Last, the significant size of TEM-1 β-lactamase, greater than 260 amino acids, would require a large number of mutants to sample even a single mutation at each position.

SH3 domains

As small polyproline-binding proteins, SH3 domains are involved in a variety of cell signaling cascades [3]. These domains are composed of five anti-parallel β-strands comprising

two β-sheets [4]. The yeast genome contains 27 SH3 domains, one of which, Sho1, plays a direct role in activating the high-osmolarity stress response pathway [5]. Much of the work involving this particular SH3 domain has focused on understanding the specificity of signaling pathways. For example, of the 27 yeast SH3 domains only the Sho1-SH3 domain binds to its in-vivo target Pbs2, a MAPKK. However, several non-yeast SH3 domains bind Pbs2 to varying extents, suggesting significant negative selection for cross-reactivity amongst the SH3 domains within a single organism but not between distantly-related organisms [6].

This high-osmolarity response assay system could function as a model system for assaying many SH3 mutants since the screen for function is growth on high-salt medium of a yeast cell expressing a given SH3 variant. Notably, this system also represents a good measure of the contribution of protein function to organismal fitness, where affinity of the Sho1 SH3 domain for Pbs2 correlates positively with growth rate on high-salt medium [6]. This group also studied the effect of introducing cross-talk into pathway by measuring the growth of a Sho1 mutant with increased affinity but decreased selectivity for Pbs2. This mutant displayed a growth disadvantage on high-salt medium despite its increased affinity for Pbs2. This result suggests that the Sho1-Pbs2 system may be a good model for looking at system-level properties like negative selection and selectivity. However, in the most basic sense, the function of an SH3 domain could be distilled to its affinity for its cognate peptide, and this assay system may not produce a one-to-one relationship between SH3 binding and yeast growth. Also, working with yeast presents a disadvantage due to their slow growth rate and more labor intensive molecular biology techniques. In addition, screening growth rates on solid medium would not be feasible for the scale of mutants necessary for these studies, and preliminary work in the lab suggested

that the relationship between SH3-Pbs2 affinity and growth rate in liquid culture was less reproducible than growth on solid medium (Russ W., unpublished).

PDZ domains

Like SH3 domains, PDZ domains are protein-protein interaction modules utilized extensively as part of multi-domain signaling proteins in a range of organisms. These approximately 100 amino acid domains were named after the first three protein in which they were found: PSD-95 (postsynaptic density 95) [7], DLG (discs large) [8-9], and ZO-1 (zona occludens 1) [10]. PDZ domain-containing proteins function primarily in protein-protein interactions that direct the subcellular localization of signaling complexes for cell polarity processes including the organization of neuronal postsynaptic densities and cell adhesion [11-13]. This subcellular localization of the components of a signaling pathway serves to increase the local concentration of specific molecules and enhance the fidelity of signaling through minimization of cross-talk through the spatial inclusion of only certain molecules [11, 14]. PDZ domains typically bind the C-terminus of other proteins, but have also been shown to bind lipids and internal regions of other proteins [15].

As a model system, PDZ domains present many of the same advantages as SH3 domains including a small size and extensive existing knowledge of the function of particular domains. Unlike SH3 domains, there was no clear assay system in which PDZ function could be correlated with organismal fitness. The Sho1-Pbs2 system does not directly correlate the isolated function of one protein-protein interaction and the fitness of an organism due to the additional constraint of off-target interactions of the domain of interest with endogenous ligands. Though the

property of negative selection deserves significant study in its own right, in order to make correlations between a protein's sequence and the function of that polypeptide, we chose to limit our measurement as much as possible to the interaction of one protein and one ligand. As such, we decided to develop a novel system for quantifying protein function that measures the interaction of one protein and one ligand in a cellular context; the assay should integrate properties of the nucleotide and protein sequence such as codon optimization, GC content, RNA secondary structure, expression level, folding rate, degradation rate, stability, and affinity for peptide.

While both SH3 and PDZ domains represent tractable proteins for building an assay system, the majority of the work in our lab directed at understanding the statistical and mechanical basis of amino acid co-evolution in protein families has been performed in the PDZ family and specifically the third PDZ domain of rat PSD95 [16-18]. Further, a variety of structural and mutagenesis studies have been performed in the context of the PDZ family [19-23], and while not generally considered allosteric proteins, an allosteric network has been described in the Par6-PDZ domain [24-25]. Our goal was to develop an assay system generally applicable to any protein-protein interaction, but for these reasons we chose the PDZ domain as the primary model system for building a high-throughput, quantitative assay of protein function. In the end, we could have chosen any one of the model proteins described in the previous sections, but the diversity of the family and the abundance of complementary data in existence and in development in the lab strongly motivated our decision. Importantly, the motivating questions of this work are general to proteins; these same studies could be performed with any of these systems, and we would expect similar results. In fact, similar studies with β-lactamase are currently underway in the lab.

Biology of the third PDZ domain of PSD95

Structural and functional data exist for a variety of PDZ domains, but due to the extensive previous research in our lab on the domain, we chose to use PSD95-PDZ3 of the Norway rat, *Rattus norvegicus*, for our primary model protein. Hereafter referred to as PDZ3, PSD95-PDZ3 is one of three PDZ domains (as well as an SH3 and GK domain) in PSD95, also known as SAP90, the archetypal member of the membrane-associated guanylate kinase (MAGUK) family. PSD95 plays a role in a variety of neuronal processes including the clustering, recruitment, and regulation of several receptors [26-28], as well as roles in synapse plasticity and organization [29-30]. The third PDZ domain of PSD95 has been shown to interact specifically with the postsynaptic protein CRIPT (cysteine-rich interactor of PDZ three) [31]. CRIPT also binds microtubules and functions to localize PSD95 to the synapse, though the function of this localization on receptor biology is unclear [32].

The structural basis of peptide binding in the third PDZ domain of PSD95

The crystal structure of PSD95-PDZ3 from rat shows a six-stranded β-sandwich with two additional α-helices [19] (**Figure 2.1 A**). The domain engages the four terminal peptide residues in a cleft between the β2 strand and α2 helix, though some have claimed from non-crystallographic evidence that additional peptide positions contribute to binding [33]. Binding of the peptide extends the β-sheet of the β2 strand, and the terminal position of the peptide (P_0) binds in a cavity and is engaged by the carboxylate-binding loop and hydrogen bonds with L323, G324, and F325 in the canonical PDZ 'GLGF' motif of the carboxylate-binding loop between the β1 and β2 strands (**Figure 2.1 B**). The terminal valine of the peptide also interacts with

R318, the first residue of the carboxylate-binding loop, through an ordered water molecule, and forms sidechain interactions with L323, F325, and L379 [19]. Position P_{-1} does not contact the PDZ domain, but points away from the binding pocket. However, positions P_{-2} and P_{-3} both interact with the PDZ domain. Upon peptide binding, H372 displays a 180° rotation of its imidazole side-chain to form a hydrogen-bond link between P_{-2} and G329 [19] (**Figure 2.1 C**). Position P_{-3} forms hydrogen bonds with N326 on β1 and S339 on β3 (**Figure 2.1 B**). No interactions with PDZ3 and only partial density were observed for P_{-4}.

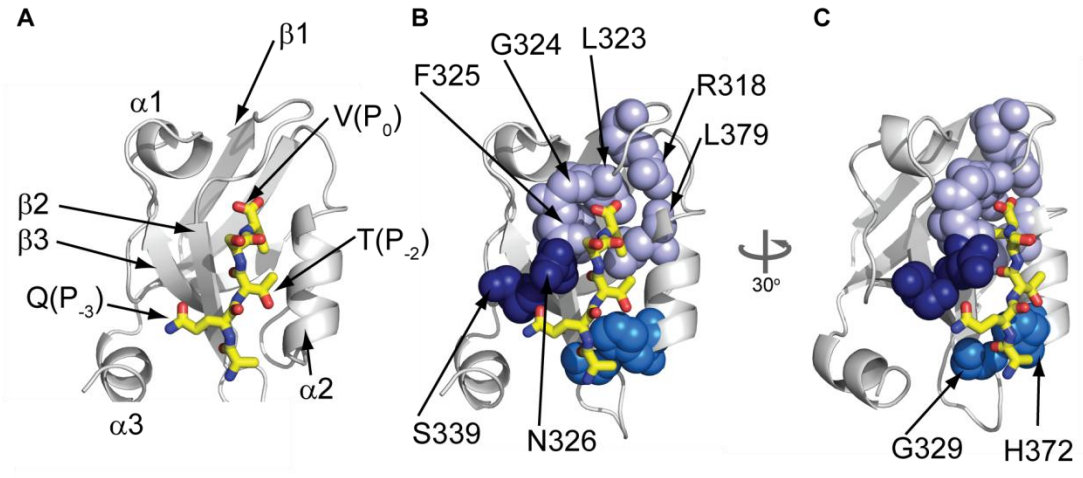


Figure 2.1 The Structure of Peptide Binding in PSD95-PDZ3
The crystal structure of PSD95-PDZ3 shows a B-sandwich configuration in which the CRIPT peptide binds in a pocket between the α2 helix and the β2−strand, clamped from above by the β1-β2 loop (A). Positions 0, -2, and -3 of the CRIPT peptide form crystal contacts with the PDZ domain as portrayed as light, medium, and dark blue spheres respectively (B,C). Image created from PDB ID 1BE9.

Building an assay for PDZ3 function

With the goal of being able to comprehensively measure the interaction of all amino acids in PDZ3, we wanted to design an assay that met the following criteria:

1) The measure of function should be an integrated parameter of not just binding or stability, but as many as possible of the known constraints imposed on the protein *in-vivo* in an attempt to measure a functional parameter that correlates to a protein's function in the context of an organism not just in the context of a test tube of buffer.
2) The assay should quantitatively and reproducibly measure this function of a PDZ domain in order to accurately compare the function of variants.
3) The assay should have throughput capable of measuring thousands of individual PDZ3 mutants in order to sample enough variants to make statements about the global patterns in the protein.

Previous measures of function for protein-protein interactions

Before designing a novel untested assay for PDZ function, we examined the existing methods for measuring protein-protein interactions. Most of these methods are either limited in throughput due to the need for purified proteins or non-quantitative. For example, fluorescence polarization provides a quantitative and reproducible method to measure the binding of a protein to a fluorophore-labeled ligand [34]. There have even been methods developed combining protein microarrays and fluorescence polarization to measure the interaction of all 157 mouse PDZ domains with each of 217 candidate peptides in the mouse genome [35-36]. Though these methods measure many interactions, they still require protein expression and purification and the

chemical synthesis of candidate peptides – all low-throughput methods. In this work, we utilize fluorescence polarization to directly measure the affinity of about one hundred PDZ3 variants for CRIPT peptide [36], but even this scale of measurements required months to complete. In addition, these methods measure the affinity of a purified PDZ domain for a given peptide in the context of some buffer in a test tube or on a glass slide, and since a fixed concentration of purified protein is used in each assay, this approach does not integrate other constraints of the sequence such as stability and expression level.

Another approach to measure protein-protein interactions, the yeast two-hybrid assay utilizes fusion proteins to detect the interaction of two candidate proteins in the context of a yeast nucleus [37-38]. In the original configuration, the GAL4 N-terminal DNA-binding domain is fused to one candidate protein while the GAL4 C-terminal transcription-activating domain is fused to the other candidate protein. If the two candidate proteins interact and bring the two domains of GAL4 into proximity, transcription of the downstream reporter gene occurs [37]. This approach permits the screening of a large number of potential interaction partners for a given protein; in fact, a typical approach is to screen a protein of interest against an entire cDNA library using a selective marker as the reporter gene for interaction [39]. Upon induction of the bait and prey fusion proteins the entire library is plated on a selective medium, and those colonies that grow should contain an interaction partner of the target protein (or a false positive). While this approach allows the screening of tens of thousands of proteins, the function of an individual protein is only scored in a binary fashion; a given bait-prey combination either permits cell growth on a selective medium or does not. There is no quantitative resolution to the measure of function for an individual protein.

Development of a quantitative bacterial two-hybrid assay

Because the two -hybrid assay provides a genetically encoded screen with high-throughput, we decided to build a quantitative two -hybrid assay for PDZ domains which maintained the throughput of standard two -hybrid assays but also produced a quantitative measure of function over the range of natural PDZ function levels. Due to the slow growth rate and more labor intensive transformation process for yeast, we decided to pursue a two-hybrid system in the context of *E. coli*. In addition, bacterial PDZ domains appear to be much more limited number than eukaryotic domains [40]. As a result, the bacterial cytoplasm may present fewer off-target ligands than a eukaryotic cytoplasm.

The Hochschild bacterial 2-hybrid system

Like the yeast 2-hybrid assay, the original bacterial 2-hybrid system couples a protein-protein interaction of interest to a transcriptional readout [41]. The assay requires fusion of one bait-protein to a DNA-binding domain and another prey-protein to one of the subunits of the *E. coli* RNA polymerase. In the presence of a plasmid containing the appropriate DNA-binding motif upstream of a reporter gene, activation of transcription occurs if the protein fused to the DNA-binding domain interacts with the protein fused to the RNA-polymerase subunit. This transcriptional activation by binding near a promoter and interacting with one of the subunits of RNA polymerase is also used by bacterial transcription-activating proteins [42]. Here, typical assay setups utilize the phage-λ cI protein for its DNA-binding N-terminal domain, the N-terminal domain of the α-subunit of RNA polymerase, and β-galactosidase or β-lactamase as a reporter gene [43]. The endogenous phage-λ promoter, P_{RM}, contains three binding sites,

operators, for λ-cI, two upstream of and one overlapping with the polymerase binding site [44]. The λ-cI protein dimerizes through its N-terminal domain and, as a dimer, can bind to any of these three operators, but displays the highest affinity for O_R1. Due to the proximity and orientation of O_R1 and O_R2, O_R2 fills very quickly after O_R1 binding due to the interaction between λ-cI dimmers [45]. As a result, λ-cI likely occupies both sites, O_R1 and O_R2, situated at -42bp and -62bp respectively. Binding of λ-cI to either O_R1 or O_R2 activates transcription if fused to a protein that interacts directly or through another arbitrary protein-protein interaction with RNA polymerase [41] (**Figure 2.2**). In the version of the promoter we use, O_R3, the operator site that overlaps with the promoter and thereby represses transcription when occupied, has been mutated to abolish binding and repression of transcription from P_{RM}. In addition to the activation of P_{RM} transcription through interaction with the polymerase a-subunit, the N-terminal domain of λ-cI can also directly interact with the α-subunit of RNA polymerase and activate transcription independent of the λ-cI C-terminal fusion when bound close to the polymerase holoenzyme at O_R2 [46]. This α-subunit-independent transcriptional activation will be further discussed in chapter 3.

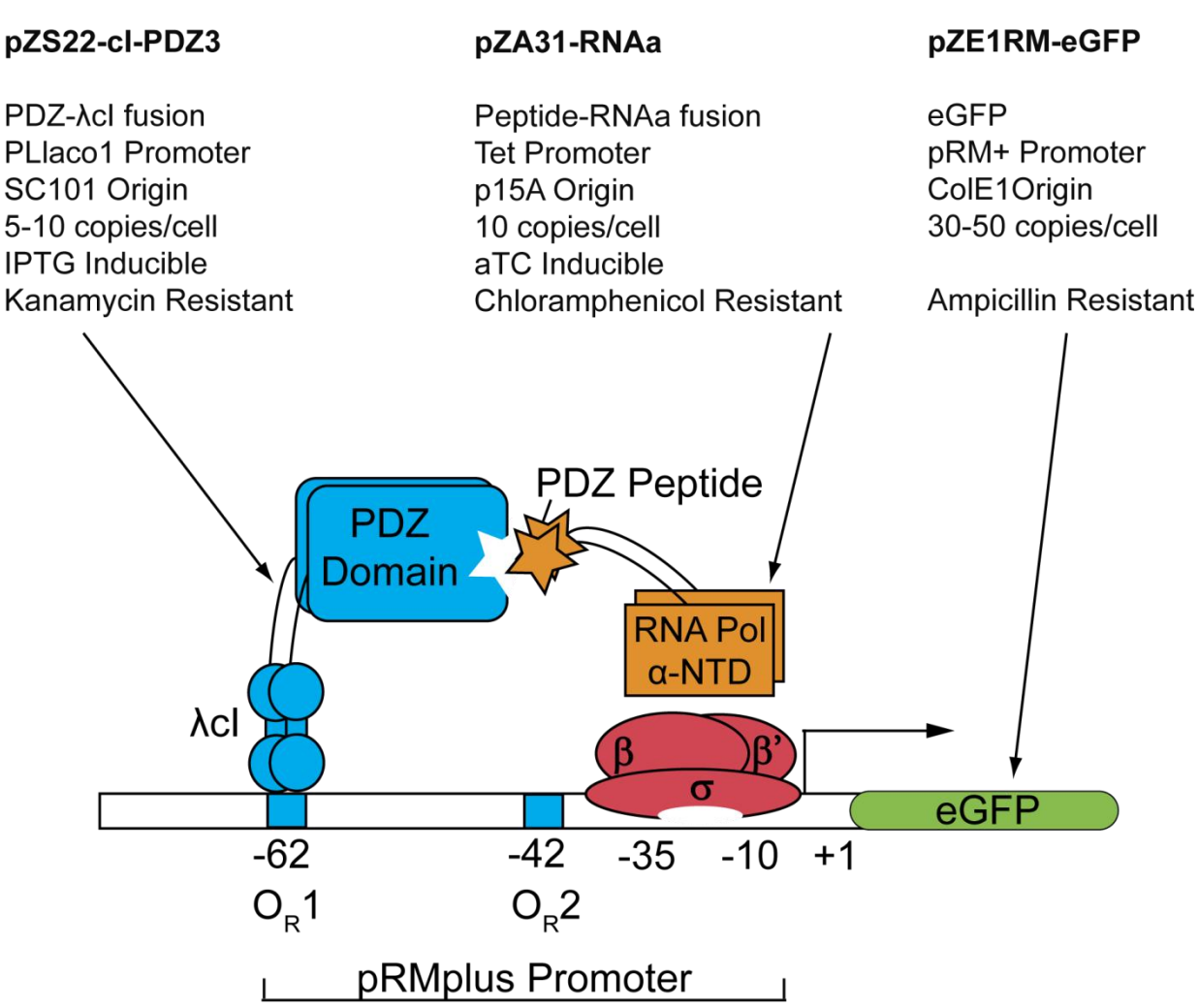


Figure 2.2 Bacterial 2-hybrid Plasmids and Assay Schematic
In the bacterial 2-hybrid assay, PSD95-PDZ3 is expressed from the pZS22 plasmid as a fusion protein with the λ-cI protein, containing a DNA-binding domain and a dimerization domain, under the control of a lactose-responsive promoter; the endogenous PDZ3 peptide, CRIPT, is expressed from the pZA31 plasmid as a C-terminal fusion with the alpha subunit of the *E. coli* RNA polymerase N-terminal domain under the control of a tetracycline-responsive promoter. The reporter plasmid contains the λ-phage P_{RM} promoter, comprising two intact λ-cI binding sites upstream of the eGFP gene (O_R1 and O_R2).

Initial construction of the bacterial 2-hybrid assay

The Hochschild bacterial 2-hybrid assay presents a useful framework for screening a large number of protein-protein interactions in a genetically amenable organism [41]. Though some correlation between interaction strength and reporter gene expression level has been shown [41, 47-48], the assay is typically used as a binary measurement to screen for proteins that interact with a protein of interest [38]. For our purposes, we needed an assay that combined the high-throughput nature of the bacterial 2-hybrid assay with the quantitative and reproducible nature of thermodynamic binding assays such as fluorescence polarization and isothermal titration calorimetry. A number of variables in the bacterial 2-hybrid can be manipulated, and we reasoned that the appropriate combination of parameters should be able to produce a characteristic relationship between the steady state thermodynamics of the bait and prey proteins and the level of transcription of the reporter gene.

The genetic components of the 2-hybrid assay exist on three plasmids (**Figure 2.2**). The pZS22 plasmid expresses the bait protein, in our case PSD95-PDZ3, fused to the λ-cI protein under the control of an IPTG-inducible promoter, $P_L$lac01, and contains a low copy number origin of replication, pSC101, and a kanamycin resistance cassette. The pZA31 plasmid expresses the prey protein, here the nine-amino acid CRIPT peptide, fused to the N-terminal domain of the α-subunit of *E. coli* RNA polymerase. This plasmid contains a tetracycline sensitive promoter, the low copy number p15A origin of replication, and a chloramphenicol resistance cassette. The final plasmid contains the P_{RM} promoter, with a mutated O_R3, upstream of the fast maturing, high quantum yield enhanced-GFP gene (GFP F64L S65T), the medium copy number ColE1 origin, and the ampicillin resistance gene, β-lactamase.

In order to decrease background reporter gene expression, the fusion protein plasmids are maintained in a repressed state by the constitutive expression of the lac- and tet-repressors. The genes for these repressors, *lacI* and *tetR*, were integrated into the genome of the *E. coli* K12 derivative MC4100 to create the MC4100-Z1 cell line [49]. In the absence of IPTG and aTC, expression of the fusion proteins and any resulting eGFP should remain low. Upon addition of both inducers, if the PDZ domain interacts with the expressed ligand, an increase in eGFP should be observed within tens of minutes, the timescale of transcription, translation, and eGFP maturation.

Choice of a reporter gene for the bacterial 2-hybrid assay

The 2-hybrid setup can utilize any gene as the readout for bait and prey interaction; the most commonly used are β-lactamase, β-galactosidase, or a fluorescent protein [43]. We wanted the transcriptional readout that provided the optimal dynamic range, reproducibility, and ease of measurement for both individual mutants and complex populations of variants. Ampicillin resistance and lactose utilization are both measured as enhancement of growth rate, a true fitness measure. However, we worried that the exponential nature of microbial growth might present a very stringent selection – enriching exponentially for the most beneficial mutations and depleting exponentially for deleterious mutations – perhaps to the point of being unable to resolve deleterious mutations. As an additional constraint made moot by our use of allele quantification by sequencing (see Chapter 3), growth rates are not amenable to highly parallel measurement unless measured in very small volume multi-well plates, and in our experience, microbial growth rates display high variance in such a format. Population measurements could be achieved with a

method of highly reproducible growth, but thousands of individual measurements would be highly laborious. There are methods of single cell growth measurement, but these methods require specialized microscopy equipment and are limited in throughput [50]. Though neither of these limitations is prohibitory, we did not pursue a growth-based assay readout based upon these disadvantages and the perceived advantages of eGFP as a reporter gene.

Our choice of a fluorescent protein was motivated several factors. First, GFP is easy to measure both on a single cell and population level. Second, fluorescence has a large dynamic range and is limited only by the minimal number of GFP molecules necessary for visualization. Third, cells displaying a particular intensity of fluorescence can be specifically enriched using fluorescence-activated cell sorting, which would correspond to the isolation of functional clones in our assay. For these reasons, we proceeded to optimize the bacterial 2-hybrid assay using eGFP as the reporter gene.

Optimization of the assay using characterized PDZ3 mutants

As a means to characterize and eventually optimize the dynamic range of the assay, we cloned several previously characterized PDZ3 mutants that display a range of dissociation constants for CRIPT peptide spanning 1-200 μM, as measured by ITC [17-18]. These domains represent an appropriate 'standard curve' since these mutants cover the range of K_d's sampled by natural PDZ domains.

Initially, we measured the fluorescence produced by the PDZ-CRIPT peptide interaction using a population level measurement of eGFP fluorescence over time. eGFP fluorescence and growth rate were simultaneously measured in 48-well plates in a Victor2 fluorescence plate

reader (Perkin Elmer). This measurement system allowed us to observe the profile of eGFP fluorescence and growth for mutants expressing tightly-binding PDZ3 mutants relative to those expressing weakly-binding PDZ3 mutants. Initially we screened a variety of aTC and IPTG concentrations to find the combination that produced the largest difference in the maximum growth-normalized eGFP fluorescence between WT, I336A, and I40A – a greater than 50-fold K_d range. We found 100ng/µl aTC and 100µM IPTG to produce the largest dynamic range for these three mutants. When measured with the optimized induction conditions at 18^{o}C, we observed no difference in the growth kinetics of cells expressing different PDZ3 mutants, demonstrating the absence of any growth disadvantage even at the highest levels of eGFP expression when grown at this low temperature. Further, the initial rate of eGFP fluorescence increase and the maximum level of eGFP fluorescence correlated with the K_d of a given PDZ3 mutant (**Figure 2.3**).

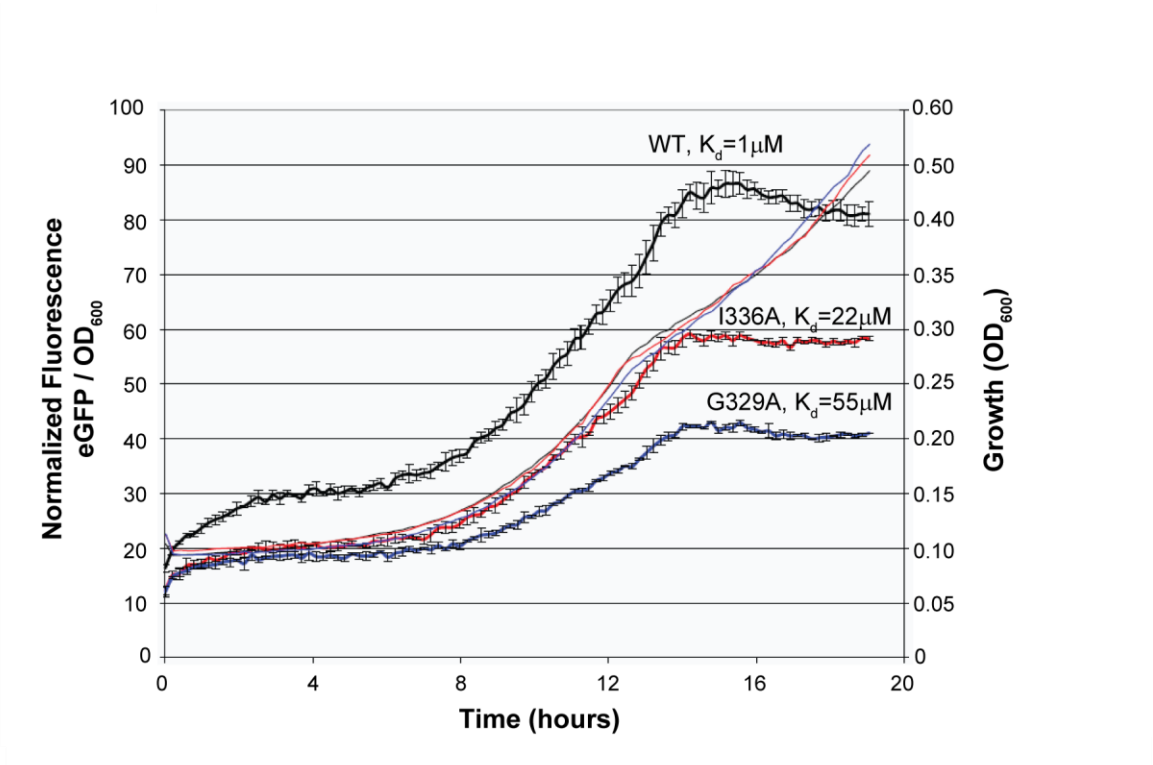


Figure 2.3 Bacterial 2-hybrid Population Level Fluorescence Measurements
Single colonies of MC4100-Z1 expressing PDZ3 WT, I336A, or G329A were grown overnight and inoculated at low density into 500 µl of inducing medium in triplicate wells of a 48-well plate. Culture growth at 18°C as a function of absorbance at 600 nm and eGFP fluorescence were measured every 15 minutes. The eGFP fluorescence at saturation for a given PDZ domain correlates with its K_d for peptide.

When grown at 30°C or 37°C, the reproducibility of eGFP fluorescence over time across replicates from the same inoculating culture decreased, but the most significant problem with the plate reader measurements came from the variance between colonies from a single transformation. We observed a significant difference in the time-course of fluorescence increase and the maximum fluorescence when multiple colonies from a single transformation of pZS22-PDZ3-WT were grown overnight, induced, and measured for eGFP fluorescence (**Figure 2.4**).

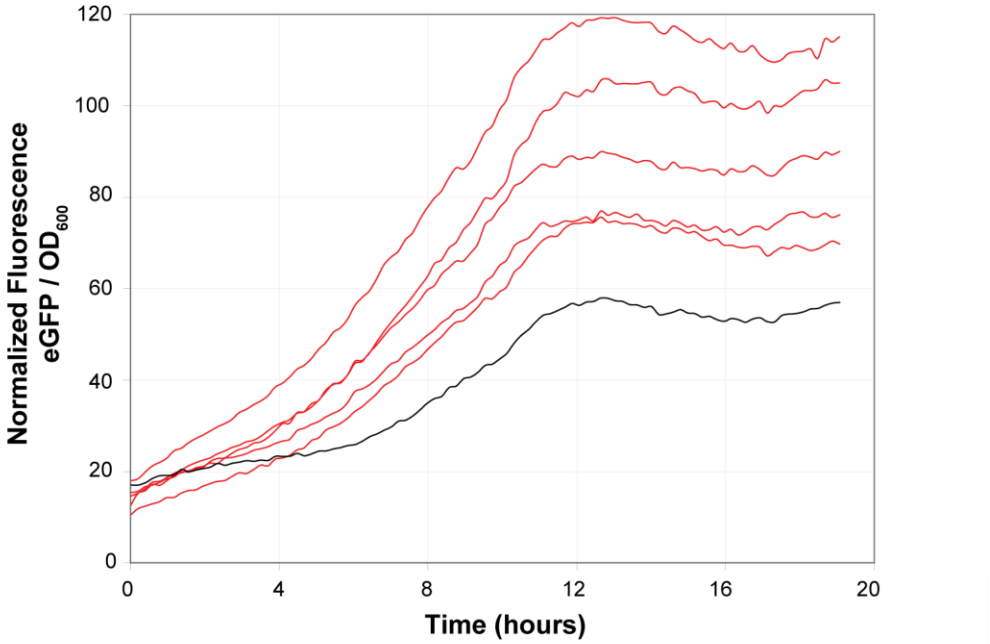


Figure 2.4 Bacterial 2-hybrid Colony Reproducibility
From a single transformation of I336A, 5 single colonies (red lines) and one streak (black line) were assayed for eGFP fluorescence from the bacterial 2-hybrid assay over time. The streaks display significant variability, especially in their maximum eGFP fluorescence.

With the help of the laboratory of Michael Elowitz at Cal Tech, we measured the variability in eGFP fluorescence within an induced population on the single cell level [51]. For this measurement, we sent cells and induction protocols to the Elowitz lab where they used fluorescence microscopy (**Figure 2.5 A**) and automated image processing software to quantify the distribution of eGFP fluorescence in cells expressing PDZ3-WT or PDZ3-H372Y, which binds CRIPT peptide with a K_d of 206 µM. Consistent with our observations of variability between colonies, we observed a broad distribution of fluorescence in the WT-expressing population (**Figure 2.5 B**). As a result, we decided to pursue a fluorescence measurement

technique in which we could visualize the distribution of eGFP fluorescence at the single cell level, not just as a collective measurement of many cells.

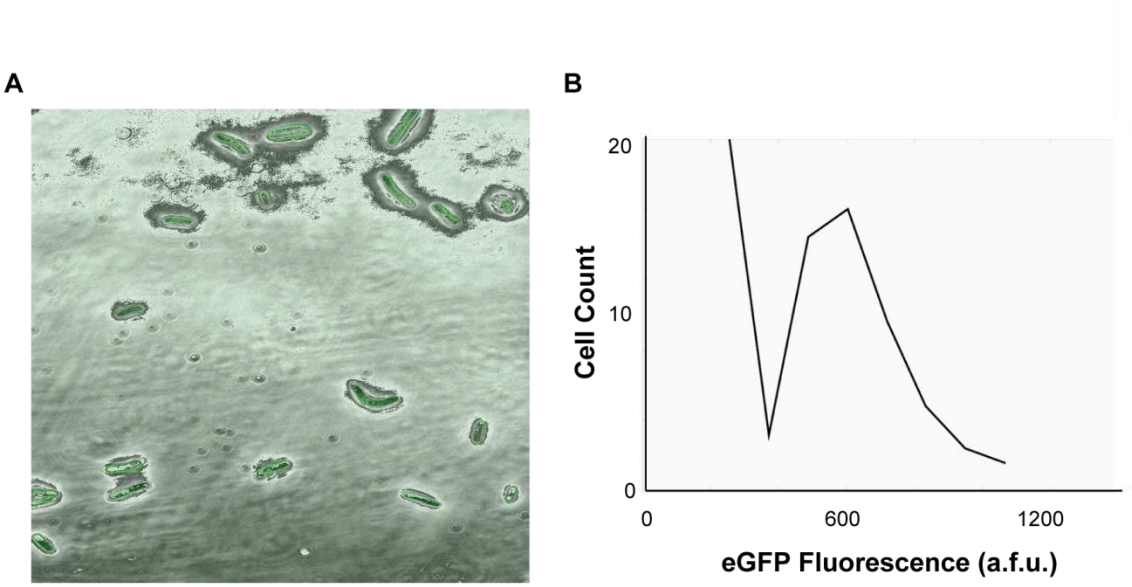


Figure 2.5 Single Cell Fluorescence Microscopy of eGFP Expression in the Bacterial 2-hybrid Assay
Single cell quantification of eGFP expression in cells expressing wild-type PDZ3 shows significant cell-to-cell variability (A) and a broad distribution of fluorescence intensities across the population (B).

Fluorescence-activated cell sorting for eGFP quantification in the bacterial 2-hybrid assay

Flow cytometers pass cells through an illuminated region at very high rates, hundreds to thousands of cells per second, and detectors measure the quantity of light scattered by each cell at small or large angles relative to the plane of illumination. In our case, the cells are illuminated with a narrow wavelength laser tuned near the excitation frequency of eGFP (488 nm), and the detector measures the intensity of the light emitted from eGFP at 509 nm and scattered at small angles. Originally developed for mammalian cells, flow cytometry has now been adapted to the

smaller cell size of bacteria for characterization of physiological parameters, as measured by changes in scattering parameters, and measurement of fluorescence from expressed fluorescent proteins or exogenous antibodies or small molecules [52-54] (**Figure 2.6**).

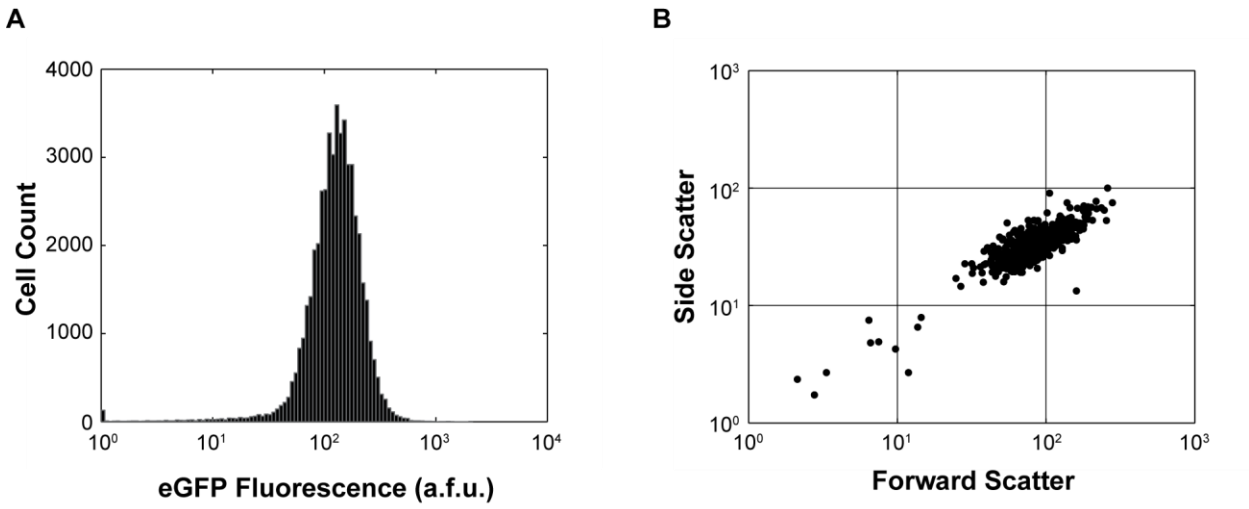


Figure 2.6 Flow Cytometry of eGFP from the Bacterial 2-hybrid Assay
Flow cytometry of induced cells from the bacterial 2-hybrid assay shows a single dominant peak of eGFP expression (A). The same cells show a single population of cells when measured for scatter at small angles and large angles relative to the path of illumination (forward scatter and side scatter respectively), demonstrating that these cells are not aggregated or morphologically heterogeneous.

For the bacterial 2-hybrid assay, cells expressing WT or one of five PDZ3 mutants displaying an affinity for CRIPT peptide between 400 nM and 206 μM were used to characterize the relationship of K_d for CRIPT and eGFP production. The initial protocol called for transformation of each mutant, overnight growth in LB at 37°C, induction at 18°C for 4 hours, and measurement with single-color flow cytometry. For each mutant, we observed a single distribution of eGFP intensity. However, the mean of the eGFP distribution for a single mutant

displayed significant variability. This variability existed between separate transformations of the same PDZ domain and within separately grown populations from a single transformation. To minimize the variability of eGFP distributions for a single mutant and maximize the dynamic range of the assay, we performed a thorough screen of growth and induction conditions. We found that increasing the number of cell doublings between transformation and induction, growing the transformed cells in a non-inducing medium, and decreasing the length of induction to be the factors most important for reproducibility and maximization of the dynamic range. The optimized protocol, as detailed in the methods section at the end of this chapter, involves 18 hours of exponential growth before induction in selective non-inducing medium (ZYM-505, not LB) and a subsequent 2 hour induction.

With the optimized protocol, we found a linear correlation between the mean of the eGFP distribution and the K_d for CRIPT peptide for a series of PDZ3 mutants (**Figure 2.7**). Though the complete parameterization of each eGFP distribution would contain much more information than simply the mean of the distribution, we found the mean to be highly reproducible and predictive of K_d. This relationship of linearity spanned the range of affinities found in natural PDZ domains and suggested our assay reported the equilibrium thermodynamic properties of the PDZ-peptide interaction. Despite this correlation between the binding energy and mean eGFP fluorescence for a small selection of PDZ domain, we reasoned that additional biophysical parameters could be constraining cellular function. In order to more thoroughly characterize the bacterial 2-hybrid assay and the biophysical constraints on cellular function, we decided to measure the binding, stability, and cellular function of a single subtle mutation at every position in the alignable region of PDZ3.

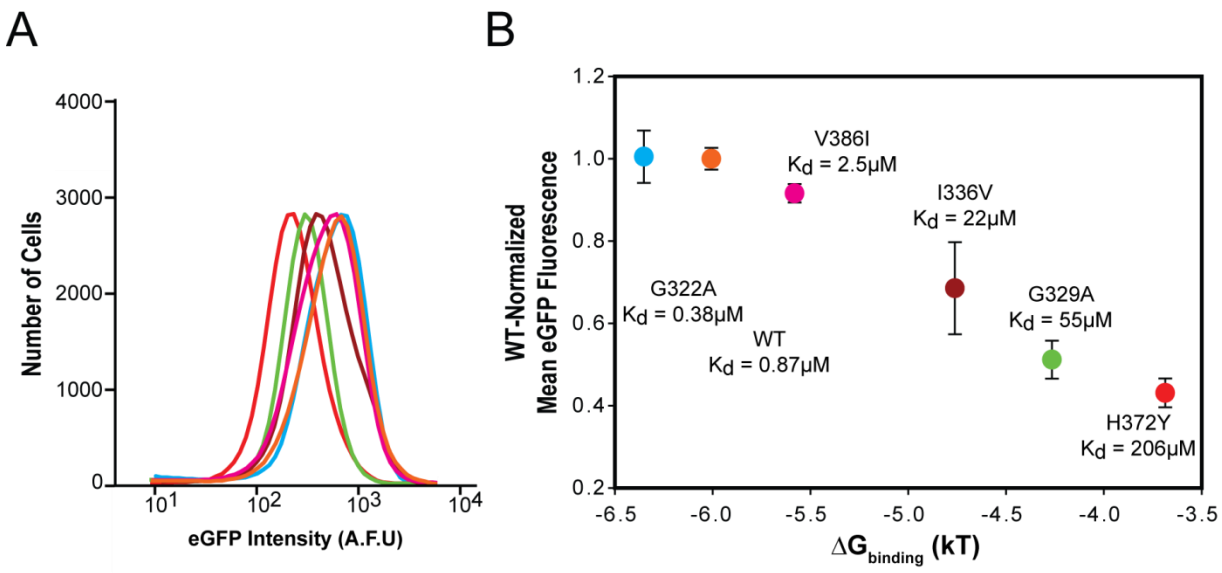


Figure 2.7 The Relationship of Peptide Affinity and Bacterial 2-hybrid eGFP Production

Wild-type and five single mutants of PDZ3 displaying a range of affinities for CRIPT peptide from 380nM to 206uM were used to optimize the bacterial 2-hybrid assay. The eGFP distributions for these proteins shift leftward relative to wild-type with decreasing affinity (A). When measured in triplicate, the mean eGFP of each PDZ domain scales linearly with its affinity for CRIPT peptide (B).

The biophysics and function of mutations in PDZ3

Our motivation for a global perturbation analysis of PDZ3 was two-fold. First, such an analysis would provide a first-level understanding of the functional contribution of each amino acid in PDZ3. Second, the creation of a large number of mutations and subsequent characterization for stability, affinity for CRIPT peptide, and bacterial 2-hybrid cellular function would provide a means to more thoroughly map the relationship of the bacterial 2-hybrid cellular function parameter and the biophysics of PDZ3 mutants. Importantly, this data also allows us to

make statements about the relative contribution of individual biophysical parameters to the cellular function of a protein.

By mutating each position in PDZ3, not to alanine, but to the next-most-conserved amino acid (NMCAA) in the PDZ alignment, we hoped to make subtle mutational perturbations to the protein. A subtle perturbation should most accurately reflect the function of a position in the protein since it should not introduce non-conservative mutations that may have significant pleiotropic effects. The use of mutations commonly sampled through evolution should also reveal the degree to which a position contributes to the functional diversity of a protein family on average. For example, if a position contributes little to the function of most members of a protein family, the NMCAA should have little functional effect in PDZ3. Alternatively, if a position is responsible for significant functional diversity in the protein family, the NMCAA would likely have a significant functional effect in the context of PDZ3.

For measurement with the bacterial 2-hybrid assay, the 83 construct NMCAA library was created by oligo-directed mutagenesis of the PDZ3 sequence of pZS22-cI-PDZ3. PCR products containing the specified mutation were ligated into the pZS22 backbone and sequence verified with Sanger sequencing. For stability and binding studies, the NMCAA library was obtained in the form of each PDZ3 mutant in the pGEX-4T-1 GST-expression vector (courtesy of Alan Poole, Ranganathan Lab, UTSW). As detailed in the methods section, the mean eGFP intensity for each mutant was measured. For biophysical measurements, each mutant was expressed as a GST-fusion and batch purified with glutathione-sepharose beads. Each mutant yielded a single, high intensity band on an SDS gel (**Figure 2.8 A**). Purified proteins were assayed for binding to TMR-labeled CRIPT peptide by fluorescence polarization (**Figure 2.8 B**), and for thermal stability by differential scanning calorimetry (DSC) (**Figure 2.8 C**).

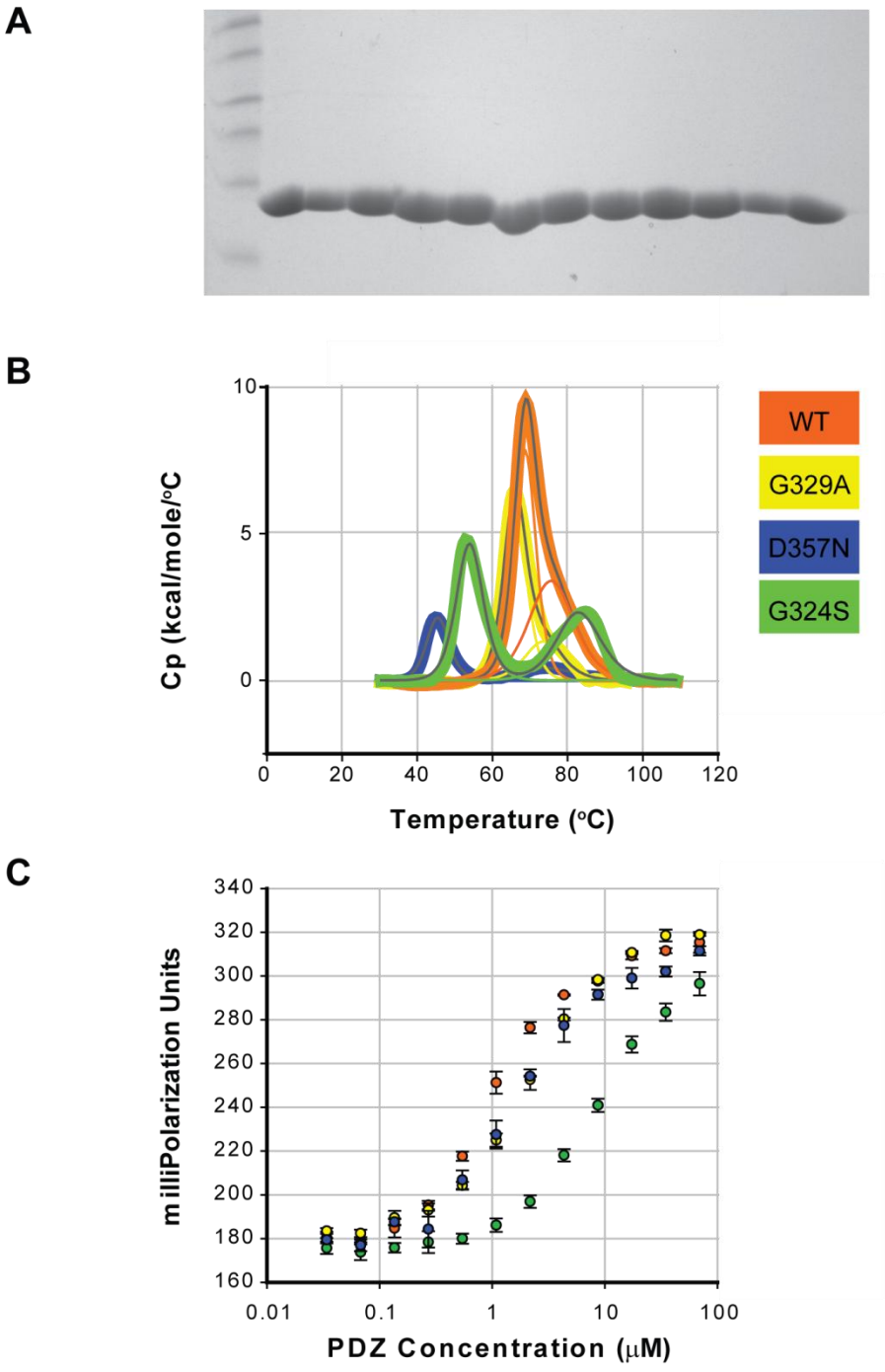


Figure 2.8 Expression and Purification, Differential Scanning Calorimetry, and Fluorescence Polarization of PDZ3 Mutants

For a library of 83 single phylogenetically-subtle subtle mutations in PDZ3, proteins were expressed as GST-fusions and cleaved during purification to yield the untagged protein (A). Purified proteins were analyzed for Tm using differential scanning calorimetry (B) and affinity for TMR-labeled CRIPT peptide using fluorescence polarization (C).

The non-two-state denaturation of PDZ3

When we used DSC to measure the thermal denaturation profile of wild-type or mutant PDZ3 proteins, we did not observe a standard two-state folding equilibrium characteristic of most monomeric, single-domain globular proteins. As detailed in the methods section, the denaturation profile of PDZ3 and almost every mutant fit a two-peak, non-two-state model. For the wild-type and three tested mutants, denaturation was reversible as demonstrated by the preservation of the T_m through three cycles of heating and cooling. Further, the T_m was not concentration dependent, suggesting that the protein does not proceed through an oligomeric denaturation. The second ΔCp peak is most likely not due to contamination of the protein preparation since wild-type protein passed over a size-exclusion chromatography column retained the same denaturation profile. We conclude that the PDZ3 protein samples a third-state upon denaturation. This conclusion is in concordance with other work in PDZ domains showing the conservation of a multi-transition denaturation in several members of the PDZ family [55-56].

For the purposes of this project, we are interested in the biophysical parameters that contribute to the function of the PDZ domain. To observe which state of the protein represents the functional state of the protein, we melted G322A and H372Y PDZ3 in the presence of increasing concentrations of peptide. Since peptide binding stabilizes the protein, we should see a stabilization of the denaturation peak that corresponds to the functional state that binds protein. We observed a distinct, peptide concentration-dependent stabilization of the first T_m (denoted T_m 1) and no change in T_m2 even at high peptide concentrations (**Figure 2.9**). For G332A, which binds peptide with slightly higher affinity than wild-type, we observed a stabilization of T_m1 by 5.9°C versus 0.2°C for T_m2 in the presence of 200 μM peptide. For H372Y, which binds peptide

with very low affinity, we observed no change in T_m1 or T_m2 in the presence of 200 μM peptide. From this result, we concluded that T_m1 is the functionally relevant T_m for PDZ3. This assumption is further strengthened by the data later in this chapter showing that the mutations that highly destabilize T_m1 produce a correlative decrease in cellular function, and those mutations that destabilize T_m2 show no such correlation (**Figure 2.10 B, C**).

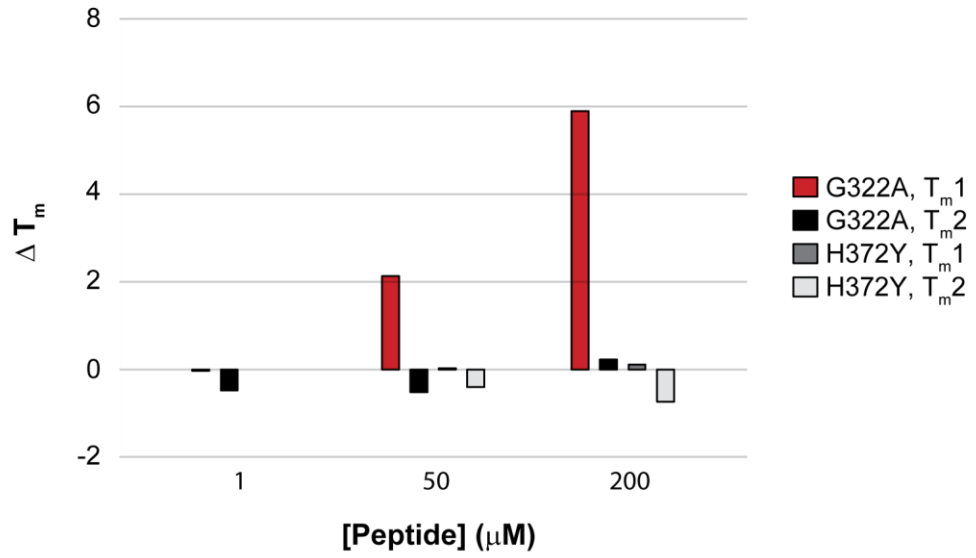


Figure 2.9 The Stabilization of T_m1 upon Peptide Binding
When the melting temperatures of G322A and H372Y are measured in the presence of increasing concentrations of peptide, T_m1 of G322A, a high affinity mutant, shows an increase proportional to the peptide concentration, while T_m2 shows no such increase. In the weakly binding H372Y mutant, neither T_m show any change upon addition of peptide.

<u>The distribution of binding and stability effects in PDZ3</u>

As expected, subtle mutation at most positions in PDZ3 had little effect on peptide binding. As seen in the histogram of binding energies (**Figure 2.9**), most mutations fall within

the body of the distribution near -8.25 kcal/mol, the binding energy of wild-type PDZ3 for CRIPT peptide. However, a clear tail of positions that decrease binding exists. In fact, 19 positions have a greater than 2-fold effect on peptide binding. This still represents less than 25% of the positions in the mutated region of PDZ3. Structurally, many of these positions lie in the binding pocket along β2 and the first half of α2. One would expect such residues with proximity to peptide to have significant binding effects; however, this pattern of spatial proximity to the peptide is not strictly predictive of significant binding effects. Some residues, such as I328 that pack against peptide show small effects on peptide binding upon mutation. Interestingly, several residues located significantly distant in the three-dimensional structure from the peptide show significant effect on peptide binding: V362 and V386 pack against each other and lie on the opposite side of the protein from the peptide-binding pocket.

In contrast to peptide binding, almost all mutations have a mild, on average 5°C destabilization effect on T_m. This is clearly observed in the broad distribution of stabilities shifted to the left of the wild-type stability value, 69°C. Unlike the binding measurements, few mutations significantly destabilize PDZ3 (**Figure 2.9**). Only 11 mutations decrease the T_m by at least 10°C, whereas 31 mutations destabilize the protein by at least 5°C. As expected, the majority of the positions that decrease the stability are buried in the protein structure. Though several position show greater than average destabilization, positions 356 and 357 have the most dominant effects on stability (18°C and 23°C respectively). The glycine at position 356 packs against the unconserved α3 helix. This mutation may cause this terminal helix to no longer pack against the conserved body of PDZ3, resulting in a significant destabilization of the protein. D357 forms a salt bridge with the guanidinium of R312, and the mutation to asparagine preserves only the hydrogen-bond potential, not the charge of the aspartic, likely resulting in this

significant destabilization Mutation of R312 to lysine is less destabilizing, probably due to compensation of the salt bridge by lysine.

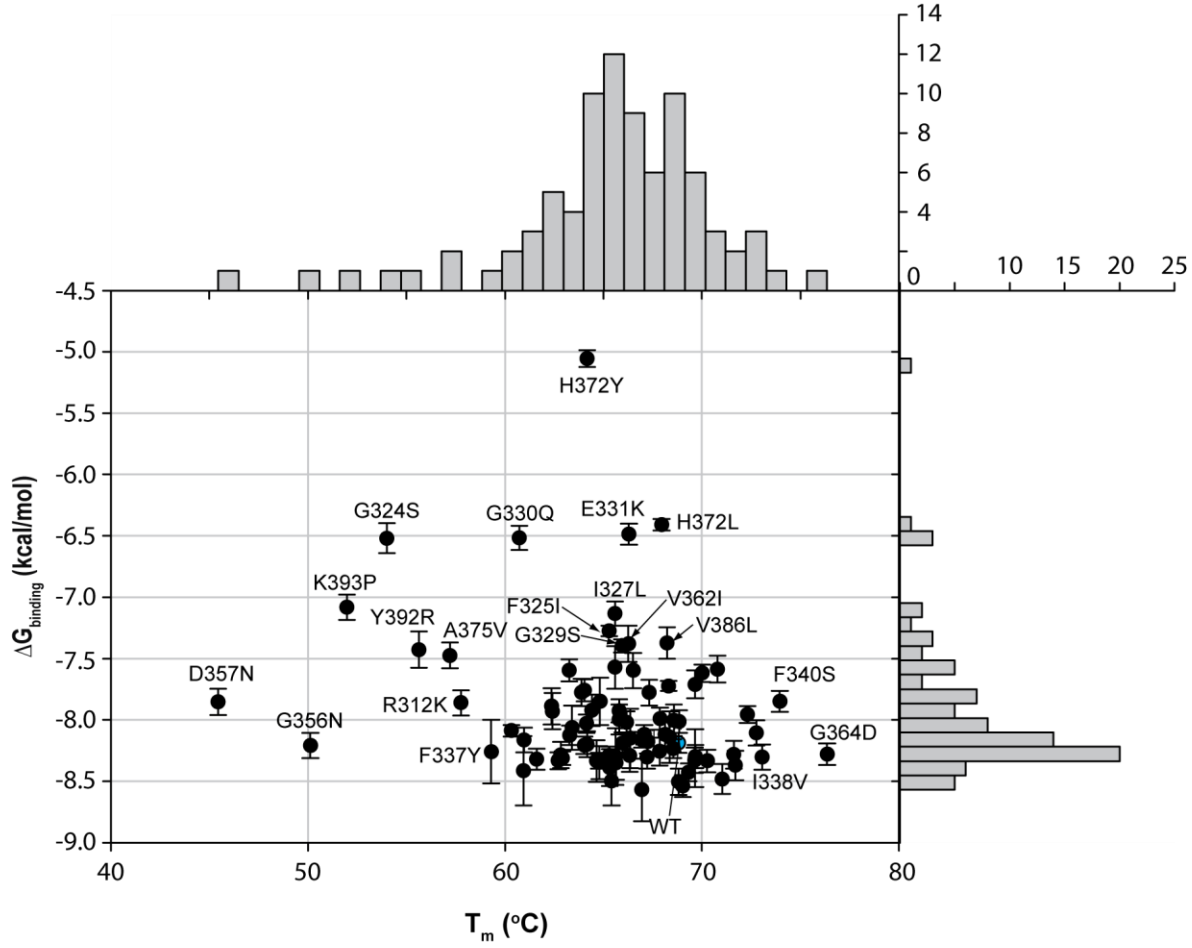


Figure 2.10 The Distribution and Correlation of PDZ3 Peptide Binding and Stability
The histogram of melting temperatures shows that most mutations have a small but significant effect on stability, while a small proportion of mutations have significant stability effects. The histogram of peptide binding energy shows that most mutations have no effect on peptide binding, but a significant number of positions have a two-fold or greater effect. There is no correlation between those mutations that decrease stability and those mutations that decrease affinity for peptide. In fact, the most destabilizing mutations have almost no effect on peptide affinity (G356N and D357N), and the mutant that most-weakly binds peptide has only a slight effect on stability (H372Y).

Interestingly, there is no correlation between those positions at which mutations destabilize the protein and those at which mutations decrease peptide binding affinity (**Figure**

2.9). In fact, those positions with the largest effects on either peptide binding or stability have no effect on the other parameter, suggesting that these two functional parameters need not be correlated. For example, the D357N mutation destabilizes the protein by 23°C but maintains wild-type-like peptide binding affinity. Similarly, the H372L mutation decreases peptide binding affinity by 1.78 kcal/mol but maintains wild-type-like stability. Three positions, A375, Y392, and K393, have significant effects on both peptide binding and stability. These positions may exhibit weakened peptide binding due to destabilization, but may also exhibit weakened peptide binding unrelated to the stability of the protein.

The distribution of functional effects in PDZ3

When the 83 subtle PDZ3 mutants are assayed for their cellular function in the bacterial 2-hybrid assay, most mutations had small effects relative to the cellular function of wild-type. The distribution of functional effects is wider than that of the stability or binding effects, but displays a similar pattern with a minority of positions producing a significant effect. Most of the positions that display cellular function effects display peptide binding effects, with the exception of the two most destabilizing mutation – G356N and D357N. Cellular function is largely robust to changes in stability unless the domain is highly destabilized, at which point a destabilization produces a correlative functional change. This relationship suggests that for single mutations, peptide binding is the most important aspect of the function of PDZ3 (**Figure 2.10**).

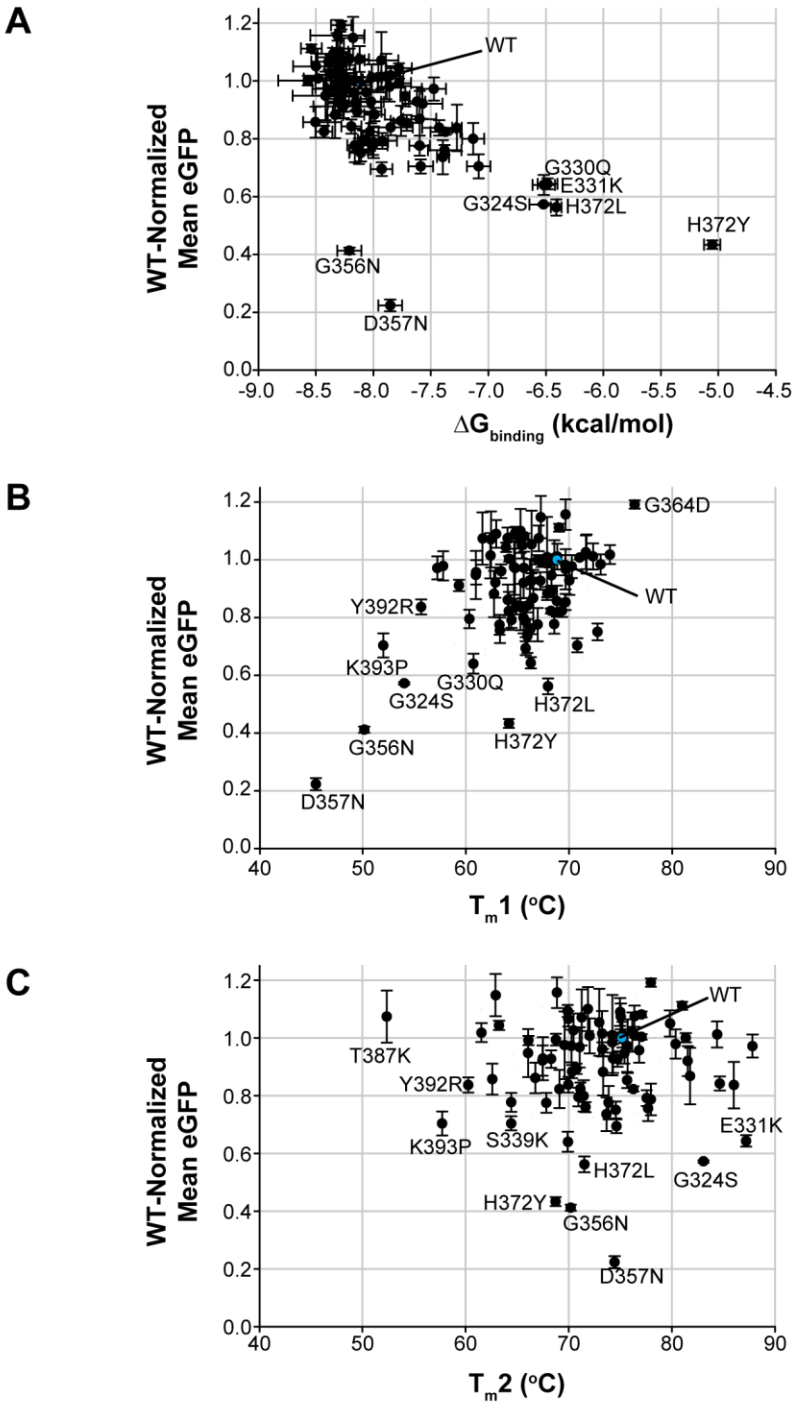


Figure 2.11 The Correlation of PDZ3 Peptide Binding and Stability with Bacterial 2-hybrid Mean eGFP

The WT-normalized mean eGFP value from the bacterial 2-hybrid assay shows a strong dependence on the binding energy of a given mutant (A). The two mutants that show a significant deviation from this relationship are G356N and D357N which show the largest destabilization (B). The melting point of the second peak in the PDZ3 denaturation curve shows no correlation to the cellular function (C).

Conclusions

In this chapter, I detail the motivation, technical development, and validation of a high-throughput method for quantifying PDZ function in the context of the *E. coli* cytoplasm. This measurement integrates the cellular constraints on function including expression, peptide-binding, and stability. With a throughput limited only by the ease of transformation and cell growth, the bacterial 2-hybrid assay system produces a quantitative measure of cellular function for hundreds of PDZ3 mutants. The functional measurements of the assay system were validated by biophysical characterization of 83 single subtle mutations in PDZ3. These biophysical measurements of peptide binding affinity and protein stability show significant heterogeneity in the distribution of stability and binding energy across the primary structure. Importantly, these stability and biding studies also show that the 2-hybrid assay reports the in-vivo function in a way that integrates both binding and stability. Though affinity for CRIPT peptide was found to be the dominant factor in the determination of the cellular function, highly destabilized mutant showed a correlative decrease in cellular function, suggesting stability contributes to the cellular function in a threshold manner while peptide affinity contributes to the cellular function in a graded manner.

Many of the effects of mutations to specific positions as measured by the bacterial 2-hybrid assay are corroborated by structural and mutagenesis studies showing the importance of the positions near the active site for protein function. However, even this first-order global analysis reveals heterogeneity in the functional importance of residues near the peptide-binding pocket and heterogeneity in the functional importance of residues distant from the peptide-binding pocket. These studies firmly establish the heterogeneous distribution in the three-dimensional structure of stability, peptide-binding affinity, and overall protein function.

Importantly, these studies also validate the invention of a novel function assay for PDZ domains that allows the more in-depth analysis of the contribution of individual amino acids to the function of the PDZ domain. The following chapter presents an improved throughput implementation of the bacterial 2-hybrid assay that permits the quantification of the function of thousands of PDZ3 mutants, thereby enabling the experiments of chapter 4 – massive single- and pairwise-mutagenesis experiments aimed at comprehensively understanding the interaction of amino acids in PDZ3 and the evolutionary implications of the underlying architecture.

Methods

The quantitative bacterial 2-hybrid assay:

Cell growth and induction

Electrocompetent MC4100-Z1 cells [49] already containing pZE1RM-eGFP and pZA31-RNAalpha-CRIPT [51] plasmids are transformed with 1 µl of 20 ng/µl pZS22-PDZ3-WT plasmid, recovered 1 hour in ZYM-505. To quantify library complexity, 1µl recovered transformation mixture is plated on LB + kanamycin (30 µg/ml). The entire 1 ml transformation is then added to 10 ml ZYM-505 [57] + kanamycin (30 µg/ml) + ampicillin (50 µg/ml) + chloramphenicol (25 µg/ml) in a 50 ml beveled flask and grown 6 hours at 37°C at 225 rpm. These 6 hour growths are diluted to 10 µl culture per 10 ml ZYM-505 + antibiotics and grown 12 hours at 37°C at 225 rpm. 12 hour growths are measured for optical density at 600 nm. A 35 µl aliquot of each culture is added to one well of a 48-well plate containing 500 µl LB +antibiotics +aTC (100 ng/µl) +IPTG (100 µM) for a final OD_{600} = 0.4. The plate is then incubated at 18°C, 150 rpm, for 2 hours for induction. Induced cells are diluted to 30 µl cells per 1 ml filter-sterilized M9 + 0.4 % glucose for cytometry. Before analysis or sorting, cells are passed through a 30-gauge needle for disaggregation to single cells.

Flow cytometry for eGFP quantification and cell sorting

All flow cytometry is performed with standardized settings on the BD FACScan. Cells are measured for eGFP fluorescence (488 nm excitation filter, 535/15 emission filter). Cell Sorting

is performed on the BD FACSAria by technicians in the UTSW cytometry core. For library selections, flow-cytometry gates are placed relative to the fluorescence distribution of WT-PDZ3 to control for systematic assay-to-assay variability. When sorting a complex library of PDZ3 mutants, a positive cell population numbering greater than 1000-times the complexity of the library was collected. Cells were sorted into chilled rich medium (ZYM-505 without antibiotics, 4°C), and the collection tube was kept chilled in the cytometer during sorting to maximize cell viability. Typical viability of sorted cell populations were >70% when plated on solid selective medium.

Expression and purification of PDZ3 mutants:

Glycerol stocks of BL21-DE3 cells were transformed with pGEX-4T-1, containing PDZ3-WT or PDZ3-mutants cloned into the BsmBI/XbaI sites, and grown overnight on LB+ampicillin plates. MDG minimal media cultures were inoculated with streaks of the freshly-transformed expression cells and grown overnight. Expression was performed using the autoinducing media ZYM-5052+ampicillin [57]. 1L cultures were inoculated with 1ml starter culture and grown at 37°C until OD_{600}~0.5 at which time cultures were cooled on ice and induced at 20°C until growth plateaued, usually 16-18 hours. Cells were harvested at 3000 rpm for 15 minutes and resuspended in 50ml conical to 35ml with NMR buffer (25mM $KHPO_4$, 50mM NaCl, 1mM EDTA, pH 7.0) + 1mM PMSF + 10 µg/ml leupeptin + 2 µg/ml pepstatin and frozen in liquid nitrogen for storage at -80°C until lysis.

For lysis, frozen pellets were slow thawed in an ice and water bath. While in an ice bath, cells were sonicated with a 0.75'' dual tip (10 sec on, 5 sec off, 5 min total). Lysed cells were transferred to a 50ml centrifuge tube and cooled before spinning 20k rpm for 1 hour in as SS-34 rotor. Supernatant was batch bound for 1 hour at 4°C on a nutator to 2ml PBS-washed and NMR buffer-equilibrated 66% GST resin (GE-Amersham). GST beads were pelleted and washed 3x with 50ml PBS and 3x with 50ml NMR buffer. Beads were then resuspended to 1.8ml total volume and transferred to a 2ml eppendorf tube with 20U thrombin and rotated 12 hours at RT or until cleavage reached ~75%. Resin was transferred to a disposable column and 200µl elutions were taken until OD_{280}<0.4. Elutions were combined and bound to 20µl benzamidine sepharose for 30 minutes at 4°C. A disposable column was used to elute the cleaved, thrombin-free PDZ protein. Purified proteins all showed a single intense band on SDS-PAGE. The concentration of each PDZ domain was determined using BCA (Pierce) and normalized to a WT-PDZ3 preparation that had been analyzed using Amino Acid Analysis (UC Davis Proteomics Core).

Differential Scanning Calorimetry and determination of thermodynamic parameters

Each purified PDZ domain in NMR buffer was diluted to 75 µM with filter-sterilized buffer from the same batch used for purification of the PDZ domains. Triplicate melts of each PDZ domain were carried out with NMR buffer as the solvent reference. The heat capacity of the protein sample relative to the buffer sample was measured at a rate of 1 °C/minute from 7°C to 120°C with a 16 second filter period. For each Cp versus temperature plot, the buffer-buffer baseline was subtracted, and the linear portions of the pre- and post-transition baselines were fit and subtracted from the transition region using a progress baseline. The resulting ΔCp versus temperature curves were fit with a two-peak, non-two state model to determine the T_m, and given the protein concentration and calorimeter cell volume, the area under the baseline subtracted curve was used to calculate the calorimetric enthalpy (ΔH_{cal}, Joules/mole of protein) and the van't Hoff enthalpy using the van't Hoff equation (ΔH_{vH}, Joules/mole of cooperative unit).

Fluorescence polarization of purified PDZ3 mutants for peptide binding:

TMR-labeled CRIPT peptide ('TMR'-TKNYKQTSV) was synthesized by the UT Southwestern Protein Chemistry Core and reconstituted to 100 nM peptide in NMR Buffer + 0.5% BSA + 5mM DTT and equilibrated to pH 7.0. Each purified PDZ protein preparation was diluted to 100 µM. In triplicate, serial dilutions of each PDZ domain were made in an untreated 96-well plate for a total of 50 µl of 8 concentrations of PDZ domain spanning 100 µM to 781 nM. In a black, clear bottom untreated 384-well, 40 µl of each PDZ dilution is mixed with 10 µl TMR-labeled peptide solution, incubated at room temperature for 1 hour, and measured for fluorescence polarization (531 excitation, 590 emission, 1 second integration) using the Perkin Elmer Victor^{3}V. The log(PDZ concentration) vs millipolarization units curve for the three triplicate assays was fit using the saturation binding model in Graphpad prism and used to extract the dissociation constant.

References

1. Frère, J.-M., *Beta-lactamases and bacterial resistance to antibiotics.* Molecular Microbiology, 1995. **16**(3): p. 385-395.
2. Martinez, J.L., *The role of natural environments in the evolution of resistance traits in pathogenic bacteria.* Proceedings of the Royal Society B: Biological Sciences, 2009. **276**(1667): p. 2521-2530.
3. Zarrinpar, A., R.P. Bhattacharyya, and W.A. Lim, *The Structure and Function of Proline Recognition Domains.* Sci. STKE, 2003. **2003**(179): p. re8-.
4. Nguyen, J.T., et al., *Exploiting the Basis of Proline Recognition by SH3 and WW Domains: Design of N-Substituted Inhibitors.* Science, 1998. **282**(5396): p. 2088-2092.
5. Seet, B.T. and T. Pawson, *MAPK Signaling: Sho Business.* Current Biology, 2004. **14**(17): p. R708-R710.

6. Zarrinpar, A., S.-H. Park, and W.A. Lim, *Optimization of specificity in a cellular protein interaction network by negative selection.* Nature, 2003. **426**(6967): p. 676-680.
7. Kim, E., et al., *Clustering of Shaker-type K+ channels by interaction with a family of membrane-associated guanylate kinases.* Nature, 1995. **378**(6552): p. 85-88.
8. Azim, A.C., et al., *DLG1: Chromosome Location of the Closest Human Homologue of the Drosophila Discs Large Tumor Suppressor Gene.* Genomics, 1995. **30**(3): p. 613-616.
9. Woods, D.F. and P.J. Bryant, *The discs-large tumor suppressor gene of Drosophila encodes a guanylate kinase homolog localized at septate junctions.* Cell, 1991. **66**(3): p. 451-464.
10. Willott, E., et al., *The tight junction protein ZO-1 is homologous to the Drosophila discs-large tumor suppressor protein of septate junctions.* Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**(16): p. 7834-7838.
11. Sheng, M. and C. Sala, *PDZ Domains and the Organization of Supramolecular Complexes.* Annual Review of Neuroscience, 2001. **24**(1): p. 1-29.
12. Kim, S.K., *Polarized signaling: basolateral receptor localization in epithelial cells by PDZ-containing proteins.* Current Opinion in Cell Biology, 1997. **9**(6): p. 853-859.
13. Feng, W. and M. Zhang, *Organization and dynamics of PDZ-domain-related supramodules in the postsynaptic density.* Nat Rev Neurosci, 2009. **10**(2): p. 87-99.
14. Ranganathan, R. and E.M. Ross, *PDZ domain proteins: Scaffolds for signaling complexes.* Current Biology, 1997. **7**(12): p. R770-R773.
15. Nourry, C., S.G.N. Grant, and J.-P. Borg, *PDZ Domain Proteins: Plug and Play!* Sci. STKE, 2003. **2003**(179): p. re7-.
16. Halabi, N., et al., *Protein Sectors: Evolutionary Units of Three-Dimensional Structure.* 2009. **138**(4): p. 774-786.
17. Lockless, S.W. and R. Ranganathan, *Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families.* Science, 1999. **286**(5438): p. 295-299.
18. Sharma, R., *Logic and Mechanism of Evolutionarily Conserved Interaction in PDZ Domains*. 2004, University of Texas Southwestern Medical Center.
19. Doyle, D.A., et al., *Crystal Structures of a Complexed and Peptide-Free Membrane Protein Binding Domain: Molecular Basis of Peptide Recognition by PDZ.* 1996. **85**(7): p. 1067-1076.
20. Fuentes, E.J., C.J. Der, and A.L. Lee, *Ligand-dependent Dynamics and Intramolecular Signaling in a PDZ Domain.* Journal of Molecular Biology, 2004. **335**(4): p. 1105-1115.
21. Petit, C.M., et al., *Hidden dynamic allostery in a PDZ domain.* Proceedings of the National Academy of Sciences, 2009: p. -.
22. Skelton, N.J., et al., *Origins of PDZ Domain Ligand Specificity. STRUCTURE DETERMINATION AND MUTAGENESIS OF THE ERBIN PDZ DOMAIN.* J. Biol. Chem., 2003. **278**(9): p. 7645-7654.
23. Tonikian, R., et al., *A Specificity Map for the PDZ Domain Family.* PLoS Biol, 2008. **6**(9): p. e239.
24. Garrard, S., et al., *Structure of Cdc42 in a complex with the GTPase-binding domain of the cell polarity protein, Par6.* EMBO J., 2003. **22**(5): p. 1125-33.
25. Peterson, F.C., et al., *Cdc42 Regulates the Par-6 PDZ Domain through an Allosteric CRIB-PDZ Transition.* Molecular Cell, 2004. **13**(5): p. 665-676.
26. Morabito, M.A., M. Sheng, and L.-H. Tsai, *Cyclin-Dependent Kinase 5 Phosphorylates the N-Terminal Domain of the Postsynaptic Density Protein PSD-95 in Neurons.* J. Neurosci., 2004. **24**(4): p. 865-876.
27. Schnell, E., et al., *Direct interactions between PSD-95 and stargazin control synaptic AMPA receptor number.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(21): p. 13902-13907.

28. Xia, Z., et al., *A Direct Interaction of PSD-95 with 5-HT2A Serotonin Receptors Regulates Receptor Trafficking and Signal Transduction.* Journal of Biological Chemistry, 2003. **278**(24): p. 21901-21908.
29. Hata, Y. and Y. Takai, *Roles of postsynaptic density-95/synapse-associated protein 90 and its interacting proteins in the organization of synapses.* Cellular and Molecular Life Sciences, 1999. **56**(5): p. 461-472.
30. Yao, W.-D., et al., *Identification of PSD-95 as a Regulator of Dopamine-Mediated Synaptic and Behavioral Plasticity.* Neuron, 2004. **41**(4): p. 625-638.
31. Niethammer, M., et al., *CRIPT, a Novel Postsynaptic Protein that Binds to the Third PDZ Domain of PSD-95/SAP90.* Neuron, 1998. **20**(4): p. 693-707.
32. Passafaro, M., et al., *Microtubule binding by CRIPT and its potential role in the synaptic clustering of PSD-95.* Nat Neurosci, 1999. **2**(12): p. 1063-1069.
33. Saro, D., et al., *A Thermodynamic Ligand Binding Study of the Third PDZ Domain (PDZ3) from the Mammalian Neuronal Protein PSD-95.* Biochemistry, 2007. **46**(21): p. 6340-6352.
34. Jameson, D.M. and G. Mocz, *Fluorescence Polarization/Anisotropy Approaches to Study Protein-Ligand Interactions*. 2005. p. 301-322.
35. Kaushansky, A., et al., *Quantifying protein-protein interactions in high throughput using protein domain microarrays.* Nat. Protocols, 2010. **5**(4): p. 773-790.
36. Stiffler, M.A., et al., *Uncovering Quantitative Protein Interaction Networks for Mouse PDZ Domains Using Protein Microarrays.* J. Am. Chem. Soc., 2006. **128**(17): p. 5913-5922.
37. Fields, S. and O.-k. Song, *A novel genetic system to detect protein-protein interactions.* Nature, 1989. **340**(6230): p. 245-246.
38. Legrain, P. and L. Selig, *Genome-wide protein interaction maps using two-hybrid systems.* FEBS Letters, 2000. **480**(1): p. 32-36.
39. Legrain, P., J. Wojcik, and J.-M. Gauthier, *Protein-protein interaction maps: a lead towards cellular functions.* Trends in Genetics, 2001. **17**(6): p. 346-352.
40. Ponting, C.P., *Evidence for PDZ domains in bacteria, yeast, and plants.* Protein Science, 1997. **6**(2): p. 464-468.
41. Dove, S.J., JK; Hochschild, A., *Activation of prokaryotic transcription through arbitrary protein-protein contacts.* Nature, 1997. **386**: p. 627-630.
42. Busby, S. and R.H. Ebright, *Promoter structure, promoter recognition, and transcription activation in prokaryotes.* Cell, 1994. **79**(5): p. 743-746.
43. Dove, S.L. and A. Hochschild, *A Bacterial Two-Hybrid System Based on Transcription Activation*. 2004. p. 231-246.
44. Li, M., H. Moyle, and M. Susskind, *Target of the transcriptional activation function of phage lambda cI protein.* Science, 1994. **263**(5143): p. 75-77.
45. Ptashe, M., *A Genetic Switch: Phage Lambda Revisited*. 3rd Edition ed. 2004, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
46. Kuldell, N. and A. Hochschild, *Amino acid substitutions in the -35 recognition motif of sigma 70 that result in defects in phage lambda repressor-stimulated transcription.* J. Bacteriol., 1994. **176**(10): p. 2991-2998.
47. Giesecke, A.V. and J.K. Joung, *The Bacterial Two-Hybrid System as a Reporter System for Analyzing Protein-Protein Interactions.* Cold Spring Harb Protoc, 2007. **2007**(6): p. pdb.prot4672-.
48. Serebriiskii, I.G., et al., *A Combined Yeast/Bacteria Two-hybrid System: Development and Evaluation.* Mol Cell Proteomics, 2005. **4**(6): p. 819-826.

49. Lutz, R. and H. Bujard, *Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements.* Nucl. Acids Res., 1997. **25**(6): p. 1203-1210.
50. Umehara, S., et al., *On-chip single-cell microcultivation assay for monitoring environmental effects on isolated cells.* Biochemical and Biophysical Research Communications, 2003. **305**(3): p. 534-540.
51. Elowitz, M.B. and S. Leibler, *A synthetic oscillatory network of transcriptional regulators.* Nature, 2000. **403**(6767): p. 335-338.
52. Davey, H. and D. Kell, *Flow cytometry and cell sorting of heterogeneous microbial populations: the importance of single-cell analyses.* Microbiol. Rev., 1996. **60**(4): p. 641-696.
53. Tombolini, R. and J.K. Jansson, *Monitoring of GFP-Tagged Bacterial Cells*. 1998. p. 285-298.
54. Dhandayuthapani, S., et al., *Green fluorescent protein as a marker for gene expression and cell biology of mycobacterial interactions with macrophages.* Molecular Microbiology, 1995. **17**(5): p. 901-912.
55. Calosci, N., et al., *Comparison of successive transition states for folding reveals alternative early folding pathways of two homologous proteins.* Proceedings of the National Academy of Sciences, 2008. **105**(49): p. 19241-19246.
56. Jemth, P. and S. Gianni, *PDZ Domains: Folding and Binding.* Biochemistry, 2007. **46**(30): p. 8701-8708.
57. Studier, F.W., *Protein production by auto-induction in high-density shaking cultures.* Protein Expression and Purification, 2005. **41**(1): p. 207-234.

Chapter 3: Solexa Sequencing-based Quantification of Function in Complex Populations of Proteins

Though the mean eGFP intensity quantifies the cellular function of an individual PDZ domain as described in Chapter 2, this approach is laborious in that each domain must be individually transformed and assayed. In addition, eGFP distribution shapes change from mutant to mutant suggesting that the mean value may not adequately describe the function parameter measured by the 2-hybrid assay. We reasoned that the most sensitive way to quantify the function of a given mutant would be to grow many mutants together and measure the enrichment or depletion of a given mutant upon selection for eGFP intensity. The bacterial 2-hybrid assay provided eGFP intensity as a selectable trait representative of the cellular function of the PDZ domain, and fluorescence-activated cell sorting could be used to isolate cells with a desired level of eGFP intensity. The major limiting factor to the development of such an enrichment assay has been the lack of a means to quantitatively characterize highly diverse libraries by measuring allele frequencies in a mixed population. The development of next-generation sequencing methodologies has enabled the cost-efficient sequencing of millions of individual sequences. Utilizing this technology in combination with the bacterial 2-hybrid assay, we developed a high-throughput sequencing method to directly quantify the frequency of mutants in the unselected and selected populations. Conceptually, this is a simple experiment; simply count the number of occurrences of each mutant before and after selection. However, this approach required two significant developments: 1) the design of a high-throughput selection for PDZ domain function; 2) the design of a method to quantify the change in frequency of every mutant upon selection.

Flow cytometry as a selection for cellular function

One major advantage of growth-based assays for measuring the function of proteins or genes is the fact that many variants can be grown together and the fitness of each variant can be assigned by some measurement of the growth rate of individual variants. For example, this has been done by labeling variants with different fluorescent proteins [1-2]. In a single assay, many of the sources of variation inherent to sequential trials of individual measurements of individuals are minimized because all of the measurements of fitness are performed simultaneously under uniformly controlled conditions. Also, the growth rate of a particular variant is recorded as the average growth rate of a large number of cells with that genotype in which the fitness of a variant now becomes relative to the composition of the population within which it grows. For example, if one variant has a very high growth rate within a population of very slowly growing variants, its measured fitness will be very large. Alternatively, if that variant of high growth rate exists within a population of several other variants with similarly high growth rates, the measured fitness of each will be much lower. The inclusion of the same variant, usually wild-type, in each experiment controls for this dependency on the composition of the population under consideration in each measurement, and the resulting parameter is usually termed 'relative fitness.'

In the bacterial 2-hybrid assay, the readout for protein function is not growth, but transcription of the eGFP gene, measured as fluorescence intensity. The definition of natural selection is tightly intertwined with reproduction; reproduction is in itself a selection since the act of reproducing enriches for the genotype that can reproduce and depletes the genotype that cannot reproduce. With the bacterial 2-hybrid assay as the front-end of our measurement of function, we used fluorescence-activated cell sorting as an artificial selection; that is, eGFP

production provides no inherent connection to enrichment in the absence of some outside force that allows high-intensity cells to persist preferentially. In the growth-linked assays discussed in the previous paragraph, we can appropriately characterize the measurement as 'fitness' since we measure the reproductive success of a certain genotype. In the case of the PDZ bacterial 2-hybrid assay, we refer to the measurement of enrichment by FACS as 'cellular function' to avoid any confusion with the classical and correct use of the term 'fitness' to mean reproductive rate, and to emphasize the fact that this measurement encompasses many constraints on the protein necessary to function in the in-vivo context of a bacterial cell.

Mechanics of sorting *E. coli* according to eGFP intensity

In contrast to the laborious approach of creating and quantifying the function of each mutant individually by the mean value of the eGFP distribution (as in Chapter 2), the use of cell sorting as a selection for PDZ cellular function allows the separation of those cells exhibiting a fluorescence above some cutoff. To separate cells of a chosen fluorescence intensity, a mixture of PDZ3 mutants are measured with the bacterial 2-hybrid as detailed in Chapter 2. The eGFP distribution of a mixture of mutants takes the weighted shape of the sum of the individual eGFP distributions of the variants in the population, where the weights indicate the representation of each variant. *E. coli* cells tend to aggregate even when grown in suspension. As a result, if a particle comprised of an aggregate of a few cells passes through the cytometer, the eGFP intensity of the particle is recorded as the sum of the individual eGFP intensities of the constituent cells. This is especially problematic since low-intensity cells can be sorted as high-intensity when aggregated with other cells. Though initially problematic, *E. coli* cells can be

dissociated to a single-cell suspension before sorting by passing the cells through the small aperture of a 30-gauge needle [3]. Ideally, cells sorted for high eGFP intensity could be regrown and assayed again to assess the efficiency of enrichment. Due to the nature of the bacterial 2-hybrid protocol, plasmid must be harvested and re-transformed into fresh expression cells. The genetically unstable *recA+ endA+* genotype of the MC4100-Z1 cell lines presents an additional challenge. When transformed and grown in MC4100-Z1 cells and then harvested by miniprep, even a pure preparation of pZS22 plasmid expressing the λ-cI-wild-type PDZ3fusion produces a heterogeneous population. This suggests that propagation of plasmids in MC4100-Z1 cells results in significant mutation accumulation. As a result, any repeated cycles of assay and selection in MC4100-Z1 cells requires the extraction of the PDZ sequence via PCR after selection and ligation into fresh pZS22 vector backbone. When pZS22-WT is grown in MC4100 cells overnight and the PDZ domain is subsequently extracted with PCR and ligated into the pZS backbone, the newly ligated population recapitulates the eGFP distribution of the original wild-type transformation. This demonstrates the ability of this approach to propagate the PDZ sequence with fidelity.

Thresholds of function in cell sorting

Libraries of PDZ3 mutants typically contain variants with a range of cellular functions. In order to separate the genotypes that encode high eGFP intensity, an appropriate stringency of selection must be defined. If the selection threshold is set too low, every cell will pass the selection and no functional information can be discerned. Alternatively, if the selection threshold is set too stringently, only a few genotypes will pass the filter and the information

about the majority of the population will be lost. In FACS, this selection threshold is determined by the placement of the gate during the sorting process. In flow-cytometry, a droplet containing a cell passes through the illumination zone where the fluorescence intensity and scattering parameters of each droplet are measured. Unique to cell-sorting, each droplet then passes through an electrical charging ring which imparts a charge onto each droplet according to one of the light measurements [4]. For our purposes, the gate sets the eGFP intensity above which a droplet is imparted a charge that directs it into a specific container as it passes through the electrostatic deflection device [5].

For a system in which each genotype produces a discrete state of the measured property, the purity of a filtered population is determined by the efficiency of the chosen gate. In FACS, this purity of sorting can be calculated empirically by measuring the fluorescence of the cells that have been sorted as positive and observing the percentage of the sorted population that actually produces an eGFP intensity above the gate. In reality, a single PDZ sequence produces a distribution of eGFP, not a single discrete value. As a result, the probability of a genotype being included in the selected population is a function of the percentage of that genotype's eGFP distribution that falls within the chosen gate and the purity of the sort (**Figure 3.1**). This situation is optimal because it produces a continuum of functional values determined by the percentage of a single mutant's eGFP distribution within the gate relative to the portion of all other distributions in the population that fall within the gate. Unlike the previous quantification of cellular function based upon the mean of the eGFP distribution, this approach allows a distribution shape-dependent quantification of cellular function without any complex fitting of the shape of the eGFP distribution.

This protocol of sorting cells for PDZ cellular function based on eGFP intensity provides a means to separate those cells that contain PDZ sequences which function, but provides no process for quantifying the enrichment or depletion of a given mutant upon sorting. With the progression to mainstream of high-throughput next-generation sequencing methods, we designed a method to sequence the input library and the sorted population to quantify the prevalence of each variant before and after selection utilizing the short read length Illumina Solexa sequencing technology.

High-throughput sequencing to quantify allele frequencies in complex populations

Utilized mainly for whole-genome sequencing, next-generation sequencing technologies employ one of two main sequencing methodologies – sequencing by synthesis using fluorescent reversible terminators or single molecule sequencing – to generate massive quantities of sequencing data for a fraction of the cost of traditional Sanger sequencing [6]. Previous mutagenesis studies were limited in scope and conclusion by several factors, including prominently the inability to sample a significant portion of sequence space. Here we utilize this technological development to sequence millions of PDZ sequences using Solexa sequencing. When coupled with the bacterial 2-hybrid assay, this technology allows the functional characterization of a massive number of variants, limited only by the scale of Solexa sequencing. This section details the previous uses of next-generation sequencing, the mechanics of the Solexa technology, and our piggybacking of Solexa sequencing with the bacterial 2-hybrid assay to create a quantitative, massively parallel assay for PDZ function.

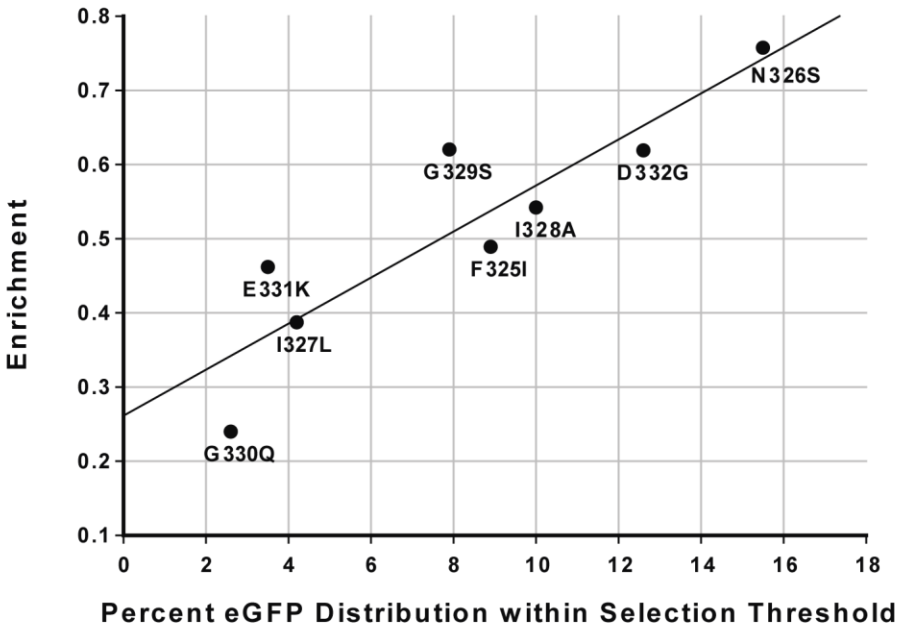


Figure 3.1 The correlation of enrichment and eGFP distribution overlap with a gate
Using the bacterial 2-hybrid assay, the eGFP distribution of NMCAA mutants located in the B2-B3 region of PDZ3 were measured individually. These mutants were mixed in equal concentrations, and the portion of the eGFP distribution of the mutant mixture corresponding to the top 15% of the eGFP distribution of wild-type PDZ3 was sorted by FACS. The input and selected populations were sequenced and the allele counts for each mutant were used to calculate the enrichment of each allele. We observe a strong correlation between the percentage of an allele's distribution that falls within the selection gate and its enrichment. This supports our assertion that this method produces a distribution shape-independent measure of PDZ cellular funtion.

Previous implementations of high-throughput sequencing for functional measurements

The major implementations of high-throughput sequencing technologies have focused on whole genome sequencing, targeted re-sequencing of genomes for single nucleotide polymorphism identification, and quantification of RNA expression levels [6-8]. Ultra-deep sequencing and targeted re-sequencing studies attempt to characterize polymorphisms within an individual or a population [9-11]. They have some similarity to our application in that the sequenced fragments are typically a single or a few regions, and as a result, are highly non-diverse.

In addition to the standard applications, several studies have implemented high-throughput sequencing to quantify the relative fitness of yeast strains containing a single gene deletion. These studies essentially repeat the previous construction and characterization with standard growth assays of the yeast comprehensive gene deletion library [12-13], but instead count a unique barcode present in each deletion strain to quantify the relative growth rate of many individual strains [14]. This highly parallel approach produces more reproducible data than the single strain measurements of growth and allows more advanced experimental setups such as a comparison of the lethal deletions in two closely related yeast species [15]. After the catalogue of lethal gene knockouts was determined with single strain growth measurements, strains of the deletion library shown to play a role in the unfolded protein response were crossed to create double knockout strains. These strains were used to measure the interaction of pairs of genes on the unfolded protein response [2]. In a similar progression from the single gene studies, a group published a gene interaction study for five genes in *Streptococcus pneumonia*. However, while the previous pairwise study used a candidate approach to find interacting genes, the massive throughput of the barcode sequencing method permitted the measurement of the interaction of each of these five genes with every other gene in the genome [16]. Both of these strategies provide the opportunity to explore a variety of pathways in an in-vivo and quantitative context.

With a technique similar to that employed in this thesis, a recent paper describes the interaction of the RNA polymerase and the transcriptional activator CRP with the sequence of the lac promoter [17]. A 75 basepair region of the sequence of the lac promoter upstream of the GFP gene was mutated to yield a library of sequences containing on average 9 mutations. This promoter library was cloned into a GFP expression plasmid and induced with or without the

addition of cAMP, which titrates the expression of CRP. The resulting GFP distribution was sorted with FACS into 5 or 10 separate bins spanning the range of eGFP intensities. The bins were individually barcoded and sequenced using 454 pyrosequencing. The output data consisted of which sequences fell where in the GFP distribution (corresponding to the degree of transcriptional activation) in the presence and absence of varying concentrations of CRP. Using mutual information, the authors extracted the coupling of the nucleotide at each position in the promoter sequence to the binding of the polymerase or CRP. This produced a list of all the functionally important nucleotide positions in the promoter, analogous on a basic level to the analysis we perform in the context of the PDZ domain. This group presented an elegant methodology, but provides no novel insight into any basic biology; they mostly recapitulated the findings of more classical experiments. However, this methodologically-parallel approach highlights the ability of next-generation sequencing to provide quantitative insight of the mapping of genotype and phenotype in a variety of systems.

A survey of next generation sequencing technologies

Next-generation, high-throughput sequencing technologies are all based on imaging some aspect of nucleotide incorporation. The two main variations use either fluorescently labeled nucleotides (Solexa, SOLiD, Helicos) or couple the release of inorganic phosphate upon nucleotide addition to bioluminescence (454) [6]. Helicos images single polymerase/template molecules, but every other technique requires the amplification of template to enable a population-amplified fluorescence signal. SOLiD differs significantly from the other technologies in that it does not use a polymerase to synthesize a complementary strand to the

template, but instead uses a ligase to ligate complementary fluorescently-labeled nucleotides that hybridize to the template molecule.

The main parameters that differ amongst the technologies are read length, cost per basepair sequenced, and total sequence yield per run. 454 boasts the highest average read length, 330 basepair, while Solexa can produce 150-200 basepair reads with paired-end reads. Helicos and SOLiD produce significantly shorter reads, 32 and 50 basepair respectively. Every technology except 454 can produce on the order of 30 gigabases of sequence per run. Despite its long reads, a 454 run produces only 0.45 gigabases per run and as such is also the most expensive technique. Solexa produces more than 400,000 basepair per dollar, while 454 produces only 12,500 basepair per dollar [18]. Our interest in Solexa was based initially upon the necessity to achieve the maximum number of reads in order to completely cover complex libraries. As discussed below we devised a strategy to be able to cover the entire 249 basepair PDZ3 sequence with the 150 basepair paired-end reads of Solexa, since 454 offered optimal read length, but far too few reads. Ultimately, the intermediate read length, large number of reads, price, and convenience of an on-site facility prompted us to choose the Solexa platform for our sequencing experiments.

Illumina's Solexa sequencing by synthesis

Almost all next-generation sequencing technologies involve the visualization of the addition of a fluorescently-labeled nucleotide to a single DNA molecule or a population of clonal DNA molecules [6, 19]. Solexa sequencing implements sequencing-by-synthesis with cyclic reversible termination of DNA polymerization (**Figure 3.2**) [8, 20]. Initially, universal adaptors

are ligated or attached via PCR to each end of the DNA fragments to be sequenced, with unique adaptors on each end of the fragment. These fragments are hybridized at low concentration to a glass slide, surface-coated with a high concentration of oligonucleotides complementary in sequence to the adaptors previously added to the DNA fragments of interest. Because each end of the fragment contains sequence complementary to one of the surface-attached oligos, each molecule of DNA binds to the slide in a bridge-conformation. With single bound DNA fragments distributed randomly across the area of the slide, the double-stranded bridges are denatured and a polymerase and unlabeled dNTPs are added to amplify each fragment. The concentration of surface-attached oligonucleotides is high, and each newly-polymerized DNA strand contains the adaptor sequence complementary to these oligos. As a result, each new strand binds the slide close to the site of amplification. This amplification and capture technique creates clusters of clonal DNA fragments.

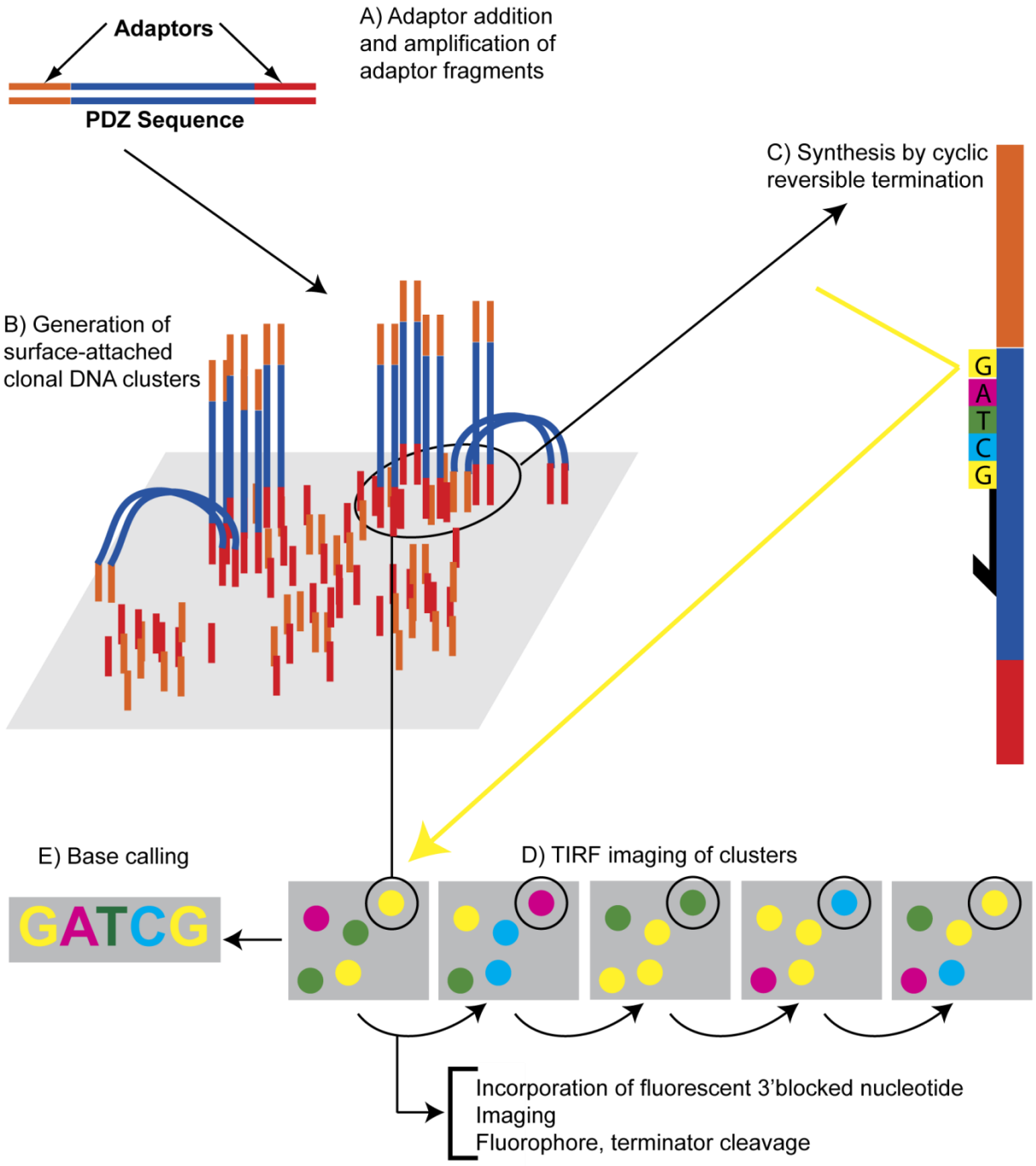


Figure 3.2 Illumina's Solexa Sequencing by synthesis with cyclic reversible termination
As described in detail in the text, Solexa sequencing requires the addition of adaptors to the template to be sequenced (A). Adaptor-containing fragments anneal to surface-immobilized oligos, and are amplified to create physically discrete clusters of clonal DNA (B). These double stranded molecules are denatured and fluorescent 3'blocked reversible terminators are added one at a a time to create the complementary strand (C). Each addition is imaged using TIRF (D), and the template sequence is extracted from the progression of colors at a single spot on the surface (E).

With spatially isolated clonal clusters of DNA immobilized on the surface of a slide, DNA polymerase, a sequencing primer, and the four dNTPs, each uniquely-labeled with a fluorescent terminator, are added to the slide. The polymerase binds the DNA fragments, primed by the sequencing oligo, and adds a single fluorescently-labeled 3'reversibly-blocked nucleotide. The two main technological advancements of this technique are the development of reversibly-terminated fluorescent nucleotides and a polymerase that effectively incorporates them. With the addition of a single base, the unincorporated labeled-dNTPs are washed away, and the dNTP incorporated at each cluster is measured with total internal reflection fluorescence imaging using two lasers and four emission filters. Imaging is followed by the chemical cleavage of both the fluorophore and the 3'blocking group of the terminator. This process is repeated to generate a series of images depicting the identity of the nucleotide incorporated at each cluster on the slide through many iterations of incorporation, imaging, and cleavage. In addition to sequencing from a single end of the template fragment, paired-end sequencing utilizes a unique sequencing primer-annealing region on each end of the template to enable two rounds of sequencing. The first round of synthesis sequences from one end of the template. The second round of synthesis sequences from the opposite end of the template, and the two reads can be associated as mate-pairs due to their co-localization on the sequencing surface. Most commonly, 36 or 72 cycles are completed to generate a 36 or 72 basepair single-end read or a 72 or 144 basepair paired-end read.

The final step in the Solexa pipeline is the generation of a sequence from the series of images. This base-calling step begins with the identification of clusters in each of the 8 lanes of the slide. A lane is simply a spatial subdivision of the chip that permits sequencing of 8 separate samples in a single run. Clusters are labeled with a 2-dimensional coordinate, and the

fluorescence intensity at each cluster in each of the four filters is quantified for each cycle of the synthesis. The resulting data is an 8-lanes x Number of spots x 4-channels x Number of cycles matrix, which represents a significant decrease in the size of the data compared to the high-resolution TIFF images of each color at each cycle. For a given spot at a given cycle, the base is called as the most intense channel, and the quality of that base call is reported as a function of the ratio of the fluorescence intensity in the called channel relative to the other three channels. That is, if two channels measure a significant intensity in cycle 15 at a particular spot, that base call receives a low quality score. The final data structure is a text file of the cluster's unique position identifier, the base calls for each cycle at that position, and the quality scores for each called base.

Subgroups of PDZ3 mutant libraries enable the use of 75 basepair sequencing reads

Our main hesitation in using Solexa sequencing was the technology's inability to produce a sequencing read that covered the entire PDZ domain sequence. We initially wanted to design an assay system that could be used to quantify the frequency of single and higher-order mutants in the conserved 249 basepair region of the PDZ domain. Short read lengths present two major problems. First, there is no established way to designate that two individual short reads were produced from the same template molecule. This means that if a PDZ variant contains more than one mutation spaced across the sequence in such a way that a single read of 75 basepair would never encompass both mutations, the sequencing would be unable to measure the existence of this double mutant. This would prevent any quantification of conditional effects of double mutants; one could create a position-by-position frequency measurement from short reads, but

this approach could only quantify the conditional effect of a second mutation that occurs locally in the primary structure. Second, Solexa sequences by synthesis from the ends of DNA fragments, so with a single PCR product of the 249 basepair region, the middle 99 basepair would never be sequenced. This can be remedied with random fragmentation of the template before adapter addition. However, if we consider a library of only single mutations and on average three reads are required to cover a single 249 basepair PDZ sequence, only one-third of all sequences would contain a mutation. This would significantly increase the number of reads necessary to adequately quantify a high diversity library.

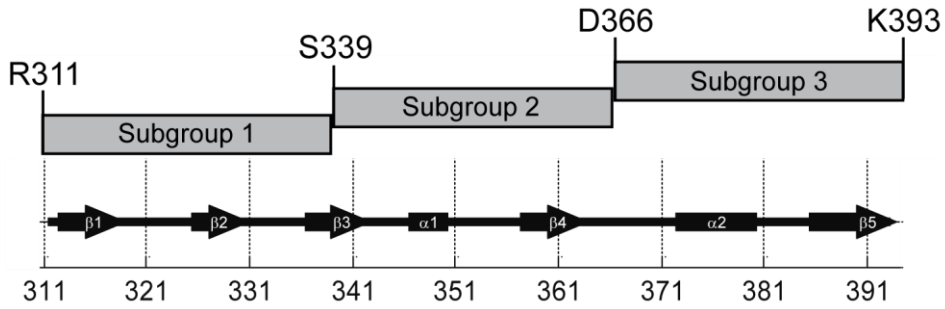


Figure 3.3 Division of PDZ3 into subgroups for Solexa sequencing
In order to cover the entire 249 basepair PDZ3 sequence with 75 basepair paired-end Solexa sequencing, the sequence was divided into three subgroups. Each subgroup contains 27 or 28 contiguous amino acid positions in the sequence. For library construction and selection, the mutants within each subgroup were combined into a single subgroup library. Sequencing could then be targeted to a particular region of the PDZ3 sequence to maximize the percentage of reads that contain a mutation.

To address the technical limitations of short read lengths, we decided to pursue a strategy in which we divided the PDZ3 sequence into three subgroups. For the NMCAA library, this meant subgroup 1 contained the mutants at positions 311-338, subgroup 2 contained the mutants

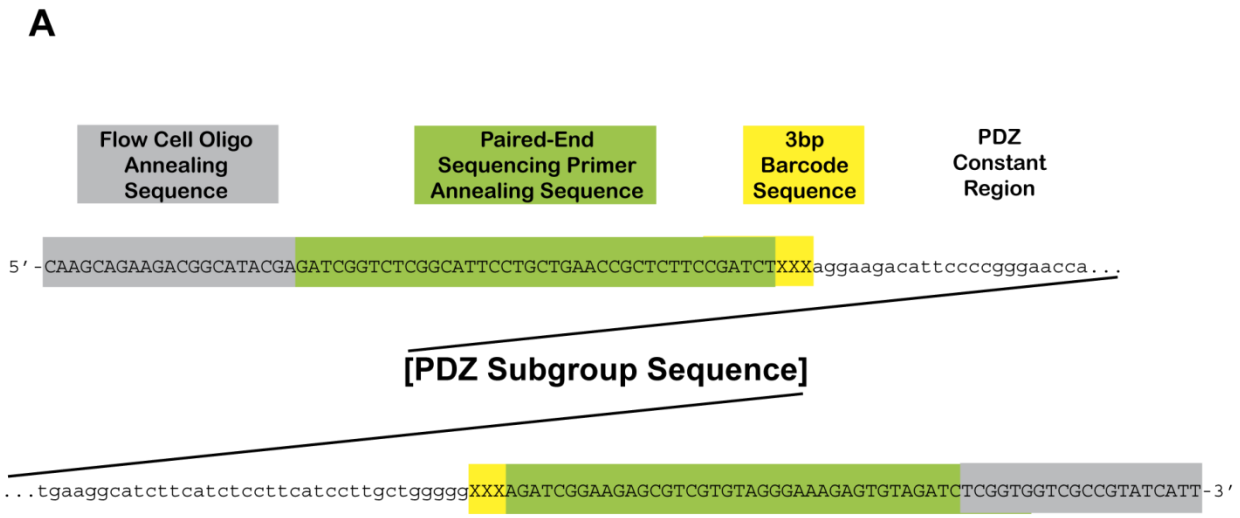
at positions 339-365, and subgroup 3 contained the mutants at positions 366-393 (**Figure 3.3**). For the bacterial 2-hybrid and cell sorting the mutants of each subgroup plus wild-type were transformed and sorted together. When we proceeded to sequence each library, we could specifically PCR the region of the sequence encompassed by the subgroup. Each fragment was 130 or 127 basepair in length which included a barcode, a constant region for amplification, and the PDZ sequence for a subgroup. With a fragment of this length, 75 basepair paired-ends should overlap in the middle of the sequence by 20 basepair, providing complete coverage of the sequence and an error checking region at the intersection of the two reads. In addition, the error rate of a base call tends to increase in the later cycles of a sequencing run [18]. This overlap allowed for stringent trimming of low-quality bases with a high probability of covering the entire sequence. This strategy of dividing libraries into subgroups and focusing the Solexa sequencing on a short stretch of the DNA sequence resulted in almost all 75 basepair paired-end reads containing a mutation; wild-type sequences only occurred with the frequency at which they were added to the library. Importantly, the addition of wild-type to each subgroup allows the mutants in different subgroups to be compared by making the measurement of each mutant relative to the measurement of wild-type.

Sample preparation and barcoding of PDZ sequences for Solexa sequencing

As described above, sequencing of a DNA fragment with Solexa requires the addition of two sequences – an adaptor sequence to attach the template fragment to the surface of the slide and an annealing sequence for the sequencing oligo to prime the PCR reaction. The standard Illumina sample preparation kit uses blunt-ending of template fragments, ligation of the adaptor

sequence, and PCR amplification of the adaptor-ligated fragments to create an amplified pool of template [21]. Though a technically simple process, the exact sequence of the adaptor sequences was kept proprietary by Illumina, forcing users to pay the $300 per sample price of a sample preparation kit. Since the defined nature of the sequenced region permitted us to add the adaptor sequence using only PCR, without the use of the more complex process of ligating adaptors, we only needed the sequence of the adaptor regions to bypass the commercial kits. The sequence of these adaptors was determined by directly sequencing the adaptors from the kit, and this information circulated on online sequencing forums. Illumina eventually released the adaptor sequences for all of their kits.

Our sample preparation involved amplification of the PDZ sequence from the pZS22 plasmid of sorted cells (**Figure 3.4**). In this first amplification a barcode and the annealing region of the adaptor were added. The surface-binding sequence was added in a second round of amplification. We utilized a 3 or 4 basepair barcode that denoted the details of each sample: which library a sample originated from, whether a sample came from the input or the selected library, and what selection gate was used. This two-round amplification was used to save money on oligonucleotides, since the complete oligo is very long (84 bp) and thus very expensive and would have to be ordered for every unique barcode. The two-round method allowed us to order short oligos containing the PDZ annealing region, a barcode, and a portion of the sequencing primer annealing region. The second oligo was universal for all barcodes, annealed to the sequence priming region, and added the rest of the adaptor.



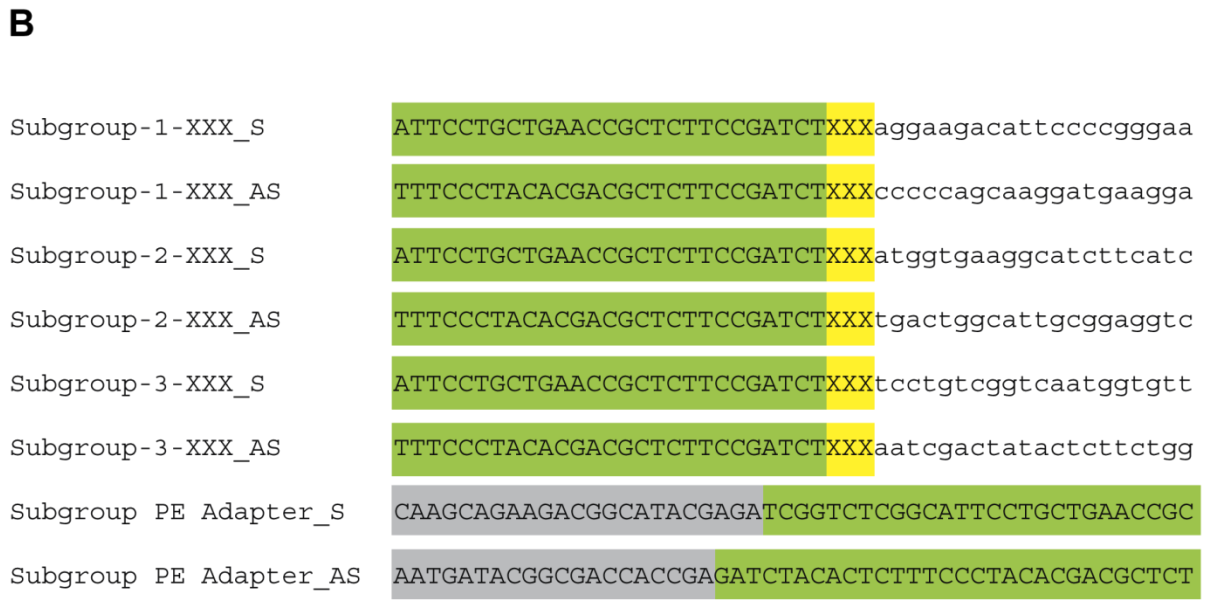


Figure 3.4 PDZ3 Solexa adapter addition

The adaptor sequences for Solexa sequncing contain a region complementary to the surface-immobilized oligosa and a region complementary to the sequencing primer. In addition, the PDZ3 sequences contain a 3 basepair barcode immediately downstream from the sequencing primer annealing site so that this barcode is the first sequence produced (A). For the addition of the barcode and first portion of the adaptor sequence to the subgroups of PDZ3, a unique oligo set is used each subgroup and anneals to a constant, non-mutagenized region in the PDZ3 sequence. A second round of amplification uses the same oligo for all subgroups and adds the remainder of the adaptor sequence (B).

When amplifying a library for sequencing, PCR can introduce significant skew into the frequency of alleles in a population. To avoid this problem, we used a large template concentration and a small number of amplification cycles in each PCR reaction, two factors shown to significantly reduce skewed amplification [22]. If template DNA is in large excess, it is more likely that in each cycle, a molecule of the original template will serve as the amplification template, and the amplification product from previous cycles will be less likely to be used as a template for additional cycles of amplification. The details of the sample preparation can be found in the methods section of this chapter. After development of the in-house sample preparation protocol, we were able to achieve a decrease in cost to approximately $10 per sample.

Application of Solexa for sequencing low diversity, high complexity libraries of PDZ3 sequences

Due to the designed application of sequencing diverse genomes, certain parts of the base-calling image analysis method discussed above rely on the sample containing a balanced composition of all four bases. In our samples, we sequence only three regions (three subgroups), resulting in an unbalanced base composition; that is, every base is not present at about equal concentration at every sequencing cycle. This presents a problem for one main aspect of the data processing. The Solexa machine uses two lasers to excite four separate fluorophores, and the emission spectra of these four molecules overlap significantly. The frequency cross-talk matrix deconvolves the bleed-through of each of the fluors into the other three emission channels. This deconvolution matrix is constructed from the first few images of a sequencing run. When our

samples do not contain every base at every one of the first few cycles, this matrix over- or under-compensates and produces a matrix that results in unusable base calls [23]. This can be easily remedied by building the crosstalk matrix using a control lane containing a genome with balanced-base composition such as the PhiX phage genome. This strategy increases the cost of each lane since a control lane must be included with every run, but this inclusion is necessary to produce quality data.

Coupling of Solexa sequencing to the bacterial 2-hybrid assay to quantify the enrichment of PDZ3 variants

The fluorescence-activated cell sorting of complex eGFP-expressing libraries produces a cell population highly enriched for those genotypes that encode a highly functional PDZ domain that produces a high eGFP intensity. A catalogue and count of the alleles present in the unselected input population and the alleles present in the intense-eGFP sorted population requires the technical ability to sequence many-times more variants than exist in the input population. The coverage of the complexity of the input library determines the sensitivity of the assay. If the input library can only be sequenced to the extent that each variant is sampled only a few times, many variants may not be present in the selected population, thereby depleting the ability of the assay to measure domains of poor function. Solexa sequencing permitted us to gather tens of millions of sequences (a typical run produced $6x10^7$ reads). However, we were left with the informatics challenge of dealing with millions of sequences, hoping to extract data from these sequences in a non-traditional manner not supported by available software. Through

a combination of commercially available and newly designed software, we were able to quantify the frequency of alleles in the input and selected populations.

The newest version of Illumina's base-calling algorithm calls bases in real-time, not all at the end of the run, in order to minimize the amount of data stored for a run. Sequences from this RTA base-caller record each read as the sequence of bases from that read, a quality score for every base in that read, and a unique name for each read that indicates its position on the slide. For trimming of low-quality bases from these sequences, we utilized a commercial software suite called CLC Genomics Workbench. This program implements the Modified-Mott algorithm from Phred to trim the raw sequencing reads according to quality scores [24-25]. The algorithm uses the values from Phred, determined empirically for Sanger sequencing, to assign an error probability to each called base. We used a cutoff value of 0.05 and a minimum read length of 49 basepair. This means the algorithm finds the longest stretch of sequence in each read that has a cumulative error probability less than the specified cutoff. If the trimmed sequence contains fewer than 49 basepair, it is discarded. Reads were then sorted into experimental groups according to the 3 or 4 basepair barcode segment, and exported as two sets of paired '.fastq' files for further analysis in MATLAB. These paired reads represent the forward and reverse sequences of the paired-end module from a single spot. Each input or selected group contained approximately 500,000 reads.

In MATLAB, the paired reads are combined into a single read by finding the overlap between the reads. If the overlap region of the two reads contains any conflicting base calls, the reads are discarded. The full-length reads are then sorted into subgroups by a constant leader sequence present in each subgroup sequence. A read is kept if it contains zero or one mutation, and for these reads, the codons in each read are counted and compiled into a table for each group

of sequences – input, gate 1, gate 2. The complete MATLAB code can be found in Appendix I. With counts for each mutant from the input and the selected populations, the enrichment of each mutant can be calculated as:

$$Enrichment_i^x = \log\left[\frac{f_{i,selected}^x}{f_{i,input}^x}\right]$$

Where *i* is an amino acid, *x* is the position in PDZ3, and f_i^x is the frequency of amino acid *i* at position x. This can be made into a relative enrichment measurement by subtracting the enrichment of wild-type in each selection:

$$Relative\ Enrichment_i^x = \log\left[\frac{f_{i,selected}^x}{f_{i,input}^x}\right] - \log\left[\frac{f_{WT,selected}}{f_{WT,input}}\right]$$

With a method for sorting complex populations of PDZ3 sequences based upon their cellular function, we decided to implement one more improvement to the design of the system. In its current state, the difference between the mean eGFP of wild-type PDZ3 and weakest binding H372Y mutant is only 2.5-fold. The next section details the logic and implementation of a modification that significantly increases this dynamic range.

A λ-cI mutation reduces background and increases dynamic range of the bacterial 2-hybrid assay

As mentioned in Chapter 2, in addition to the activation of P_{RM} transcription by interaction with the polymerase α-subunit through an arbitrary protein-protein interaction, the N-

terminal domain of λ-cI can also directly interact with the σ-subunit of RNA polymerase and activate transcription independent of the λ-cI C-terminal fusion when bound close to the polymerase holoenzyme at O_R2 [26] (**Figure 2.2**). This α-subunit-independent transcriptional activation is mediated through two amino acid contacts revealed through mutagenesis experiments and a structure of the λ-cI - RNA polymerase σ-subunit complex (**Figure 3.5**) [27-29].

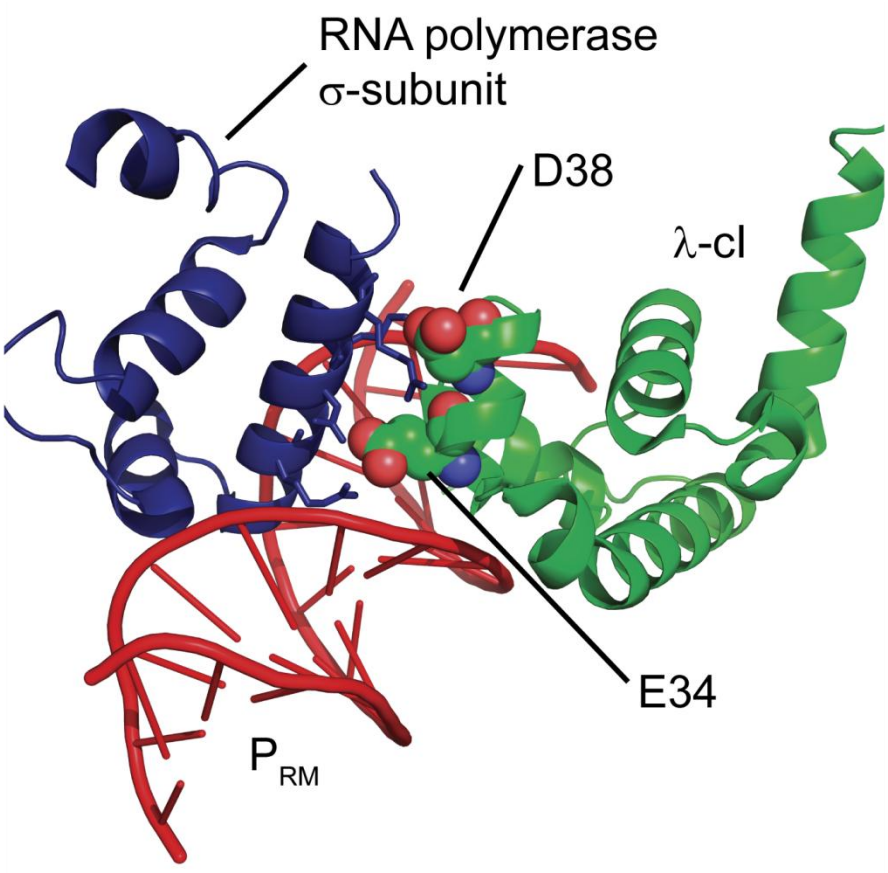


Figure 3.5 The role of λ-cI E34 in the ternary complex with PRM and the RNA polymerase σ-subunit

The crystal structure of the repressor - σ-subunit complex on the P_{RM} promoter shows the interaction of E34 with the polymerase. Only one other amino acid interacts directly with the polymerase, D38. Mutation of either of these amino acids disrupts the α-independent activation of the polymerase by λ-cI [24]. Figure created from PDB 1RIO.

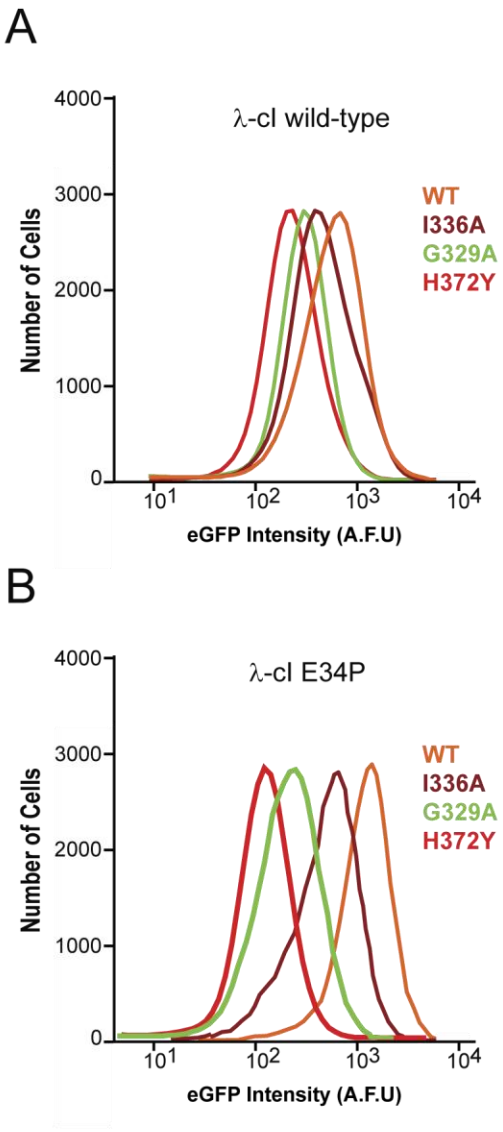


Figure 3.6 The expanded dynamic range of λ-cI E34P in the bacterial 2-hybrid assay
The spread of PDZ3 mutants spanning the range of affinities sampled by the single-mutant library increases significantly in the E34P background (B) relative to wild-type λ-cI (A). The mean eGFP of PDZ3 wild-type is 8.5-fold higher than that of H372Y in the E34P background, relative to 2.5-fold in the wild-type background.

In our lab, a limited screen of mutations at one of these sites found the E34P variant to significantly reduce P_{RM} eGFP expression in the absence of a protein-protein interaction (F.

Poelwijk and W. Gosal, personal communication). We hypothesized that the majority of the eGFP expression from weakly interacting domains may be due to non-specific eGFP expression. Therefore, we expected this mutation to reduce the signal from weakly binding PDZ3 mutants preferentially, thereby increasing the range of eGFP levels between wild-type and weakly-binding mutants. The mutation actually decreased the eGFP expression from all PDZ3 variants, suggesting that PDZ-peptide-independent transcriptional activation contributed significantly to the eGFP distribution of all mutants. Luckily, the E34P mutation decreased the background signal in a non-linear manner, with weakly binding variants displaying decreased eGFP expression to a greater extent. This increased the difference between the mean eGFP of wild-type PDZ3 and the H372Y mutant from the previous 2.5-fold to 8.5-fold (**Figure 3.6**).

Enrichment as a measure of cellular function for diverse libraries

When the cellular function of the NMCAA mutants is quantified using the E34P-variant bacterial 2-hybrid and Solexa sequencing as detailed above, we find a significantly increased sensitivity of the assay. Specifically, we find an 8.5-fold difference in the mean eGFP of the weakest and strongest binding variants, which corresponds to almost 2-log-orders difference in enrichment (**Figure 3.6**). The pattern of effects observed with this version of the assay is highly consistent with the pattern observed in the previous chapter; the major difference is the amount of work necessary to produce the two data sets and the scale of mutagenesis enabled by the Solexa-coupled assay. Using this version of the assay, the number of mutations that can be measured for function is limited only by three factors – the construction of high-diversity libraries, the efficiency of transformation of a library, and the number of obtainable sequencing

reads. With tens of millions of reads produced per lane of sequencing and transformation efficiencies greater than 10^6/ µg of DNA for MC4100-Z1 cells, it should be feasible to measure thousands of mutants and still maintain counts of several hundred per allele from a single lane of Solexa sequencing data. This enables the quantification of function for many more alleles than just a single subtle mutation at each position. Though we observe patterns of heterogeneity consistent with the statistical observations of the sector hypothesis, a complete analysis of the interaction of amino acids in the PDZ domain requires every mutation at every position. Even this comprehensive single mutant scan ignores the potentially important contribution of higher-order amino acid interactions. This Solexa-quantitated version of the PDZ function assay enables such comprehensive single mutant measurements and even limited higher-order mutant measurements.

Conclusions

This chapter details the development of a high-throughput implementation of the bacterial 2-hybrid assay that permits the quantification of the function of thousands of PDZ3 mutants. While the quantification of the mean of the GFP distribution for a single mutant provides a measure of cellular function that correlates with the biophysical effects of that mutation, this approach of Chapter 2 is highly laborious. Assaying complex libraries of PDZ3 variants as a group produces an eGFP distribution that is not interpretable quantitatively. However, this cumulative distribution can be enriched for high-intensity eGFP cells using fluorescence-activated cell sorting.

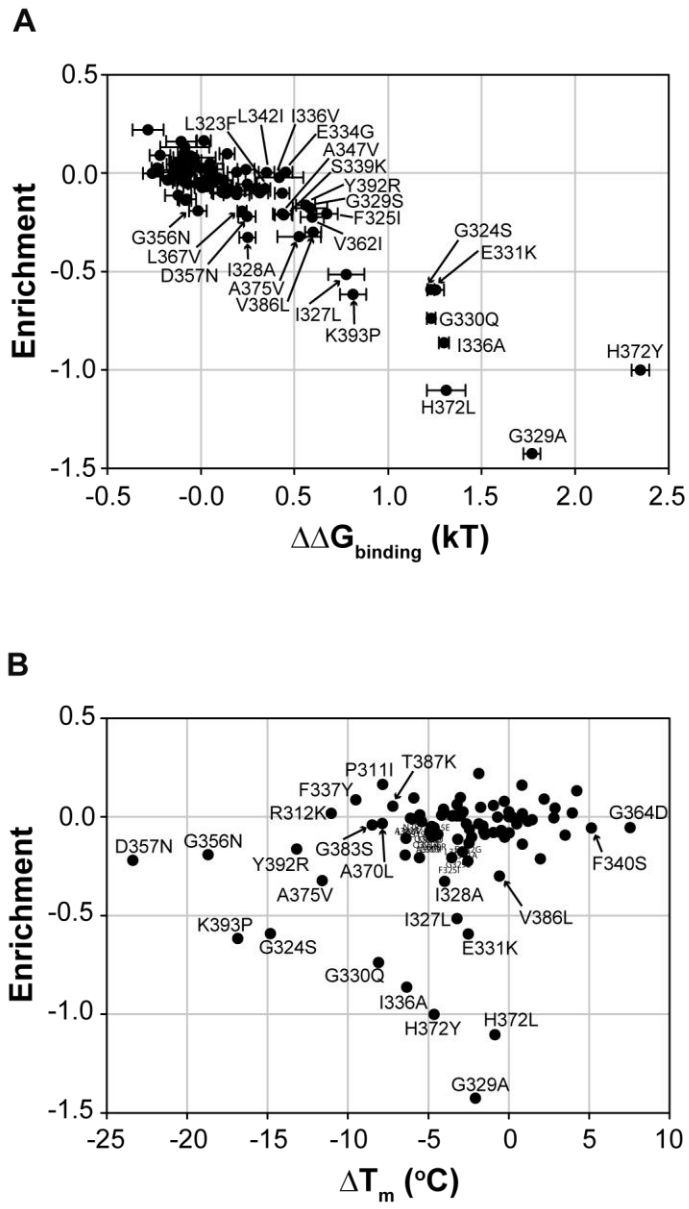


Figure 3.7 The correlation of enrichment with binding and stability in the context of λ-cI E34P

In agreement with the increase in the difference in the mean eGFP between the best and worst functioning PDZ3 variants, there is a 2-log-order span in the enrichment values for the single mutant library. The enrichment values show a strong correlation with free energy of binding (A). In contrast to the previous data, this version of the assay shows little dependence on destabilization, even for the most destabilizing mutations (B).

The second half of this chapter demonstrates that high-throughput sequencing can be used to quantify the enrichment of individual PDZ3. The increase or decrease in the prevalence of a mutant upon selection (enrichment) quantifies the cellular function of that allele. As with the mean eGFP measurements of the previous chapter, this technique produces measurements of cellular function that are validated by their correlation to the biophysical effect of individual mutations. However, the measurement of mean eGFP for the NMCAA library required several months to assay every one of the 83 constructs. This implementation of the measurement of cellular function permits the measurement of many more mutants than 83 in the span of a few weeks. The original premise of this thesis is to understand the architecture of amino acid interactions in the protein on a global scale. While the measurement of a single subtle mutation at every position shows a heterogeneous architecture consistent with the sector model of statistical co-variation, the ability to make statements about the global architecture and global models of protein function requires the implementation of a global experiment. This chapter outlines the technical developments that make such experiments possible. In the next chapter, we describe comprehensive mutagenesis experiments to observe the patterns of single mutation effects and the patterns of non-additive mutational effects in PDZ3. These experiments, enabled themselves by the technical advancements described in this chapter, permit the evaluation of the function of sector positions and their potential role in protein evolution and evolutionary-timescale functional constraints on proteins.

Methods

PDZ3 Subgroups:

In order to maximize the number of mutations sequenced per lane of sequencing and permit sequencing if a ~300bp fragment with 75bp reads, we split the PDZ3 sequence into 3 subgroups of 28, 27, and 28 positions respectively. This means that for library constructions and

selections, positions 311-338 constituted Subgroup 1, positions 239-365 constituted Subgroup 2, and 366-393 constituted Subgroup 3. The NNS-mutagenesis products for the positions of each subgroup were mixed and ligated as a single library. For selection using flow cytometry, each subgroup was sorted individually, and the sorted/input libraries were amplified individually for sequence preparation. As a result of this subgroup strategy, every read should contain a mutation or a wild-type sequence produced from the subgroup-specific NNS mutagenesis (that is the WT amino acid will be sampled at some frequency because a portion of the NNS oligos will code for the WT amino acid). If a fragmentation-type approach were used, only around one-third of all reads would contain a mutation (fragment length / read length).

Preparation of barcoded library samples for Solexa sequencing:

Sorted cell populations were diluted into ZYM-505 + kanamycin and grown 12 hours at 37°C at 250 rpm. Overnight cultures were pelleted and miniprepped (Promega Wizard Plus SV Minprep Kit). Purified DNA was quantified (Nanodrop ND-1000 Spectrophotometer), and 200 ng of plasmid DNA per 50 µl PCR reaction was used as template for the first round of adaptor addition. In order to preserve the ratio of template alleles, we used a large template concentration and few amplification cycles (16 cycles). This first PCR reaction adds the Solexa Paired-End Sequencing oligo annealing site as well as a 3-bp barcode that indicates the origin of the sample (input or selected library, selection gate). The second PCR reaction adds the remainder of the sequencing oligo annealing site and the annealing site for the flow cell oligo. All oligos were purchased from IDT as 100 nM syntheses with standard purification. Each PCR reaction included 5% DMSO and produced a single intense band.

The second round PCR products were purified (ZYMO DNA Clean and Concentrator-5 Kit) and eluted in 20µl dH2O. Purified PCR products were quantified (Invitrogen Quant-IT Picogreen dsDNA Quantification Kit) in triplicate using lambda-DNA as a standard. PCR products were diluted to 10 nM and 8 picomoles were loaded onto a Solexa v4 PE-flow cell in the UT Southwestern Solexa Sequencing Core which yielded 250,000-300,000 clusters per lane. Due to the unbalanced nature of the first bases of each PCR product, a PhiX control lane was used for matrix and phasing calculations.

References

1. Keymer, J.E., et al., *Computation of mutual fitness by competing bacteria.* Proceedings of the National Academy of Sciences, 2008. **105**(51): p. 20269-20273.
2. Jonikas, M.C., et al., *Comprehensive Characterization of Genes Required for Protein Folding in the Endoplasmic Reticulum.* Science, 2009. **323**(5922): p. 1693-1697.
3. Nebe-von-Caron, G., et al., *Analysis of bacterial function by multi-colour fluorescence flow cytometry and single cell sorting.* Journal of Microbiological Methods, 2000. **42**(1): p. 97-114.
4. Herzenberg, L.A. and R.G. Sweet, *Fluorescence-activated cell sorting.* Sci Am, 1976. **234**(3): p. 108-17.
5. Shapiro, H.M., *Practical Flow Cytometry*. 4th Edition ed. 2005: John Wiley & Sons.
6. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.

7. Marioni, J.C., et al., *RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays.* Genome Research, 2008. **18**(9): p. 1509-1517.
8. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-59.
9. Obbard, D.J., et al., *Quantifying Adaptive Evolution in the <italic>Drosophila</italic> Immune System.* PLoS Genet, 2009. **5**(10): p. e1000698.
10. De Grassi, A., et al., *Ultradeep Sequencing of a Human Ultraconserved Region Reveals Somatic and Constitutional Genomic Instability.* PLoS Biol, 2010. **8**(1): p. e1000275.
11. Turner, T.L., et al., *Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils.* Nat Genet, 2010. **42**(3): p. 260-263.
12. Tong, A.H.Y., et al., *Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants.* Science, 2001. **294**(5550): p. 2364-2368.
13. Giaever, G., et al., *Functional profiling of the Saccharomyces cerevisiae genome.* Nature, 2002. **418**(6896): p. 387-391.
14. Smith, A.M., et al., *Quantitative phenotyping via deep barcode sequencing.* Genome Research, 2009. **19**(10): p. 1836-1842.
15. Dowell, R.D., et al., *Genotype to Phenotype: A Complex Problem.* Science, 2010. **328**(5977): p. 469-.
16. van Opijnen, T., K.L. Bodi, and A. Camilli, *Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms.* Nat Meth, 2009. **6**(10): p. 767-772.
17. Kinney, J.B., et al., *Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence.* Proceedings of the National Academy of Sciences, 2010. **107**(20): p. 9158-9163.
18. Rodrigue, S., et al., *Unlocking Short Read Sequencing for Metagenomics.* PLoS ONE, 2010. **5**(7): p. e11840.
19. Wold, B. and R.M. Myers, *Sequence census methods for functional genomics.* Nat Meth, 2008. **5**(1): p. 19-21.
20. Fuller, C.W., et al., *The challenges of sequencing by synthesis.* Nat Biotech, 2009. **27**(11): p. 1013-1023.
21. *Preparing sample for paired-end sequencing: Protocol for paired-end sample preparation kit*. 2008, Illumina.
22. Polz, M.F. and C.M. Cavanaugh, *Bias in Template-to-Product Ratios in Multitemplate PCR.* Appl. Environ. Microbiol., 1998. **64**(10): p. 3724-3730.
23. *Genome Analyzer Pipeline Software User Guide, Version 1.0*. 2008, Illumina.
24. Ewing, B. and P. Green, *Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities.* Genome Research, 1998. **8**(3): p. 186-194.
25. Ewing, B., et al., *Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment.* Genome Research, 1998. **8**(3): p. 175-185.
26. Kuldell, N. and A. Hochschild, *Amino acid substitutions in the -35 recognition motif of sigma 70 that result in defects in phage lambda repressor-stimulated transcription.* J. Bacteriol., 1994. **176**(10): p. 2991-2998.
27. Jain, D., et al., *Structure of a Ternary Transcription Activation Complex.* Molecular Cell, 2004. **13**(1): p. 45-53.
28. Whipple, F.W., M. Ptashne, and A. Hochschild, *The activation defect of a lambda-cI positive control mutant.* Journal of Molecular Biology, 1997. **265**(3): p. 261-265.
29. Bushman, F.D., C. Shang, and M. Ptashne, *A single glutamic acid residue plays a key role in the transcriptional activation function of lambda repressor.* Cell, 1989. **58**(6): p. 1163-1171.

Chapter 4: Comprehensive single mutant functional analysis of PSD95-PDZ3 reveals a heterogeneous functional architecture and the role of sectors in cooperative amino acid interactions

The results of perturbation analyses form the basis of much of our biological knowledge. If we want to understand the working of a biological system, we introduce a perturbation to the system and rationalize the outcome in the context of what we know about the system. For proteins, this perturbation analysis typically occurs through the introduction of a mutation into a protein and an analysis of the resulting change in biochemical features of the protein or optimally the in phenotype of an organism. When we carry out these mutational studies, we make the assumption of *ceteris paribus* – that any change in phenotype we observe as a result of the mutation is due directly to the functional effect of that amino acid at that position, all other things being equal in the protein besides the mutation. However, proteins are dynamic, robust objects capable of local reorganization and compensation.

The interpretation of the functional effect of a mutation is made difficult by two factors. First, we often make mutations to alanine and assume that this substitution results in a loss of the function contributed by that position. While an alanine mutation may remove a wild-type side-chain, even this apparent loss-of-function mutation can create cavities or have local pleiotropic effects. As a result, it is very difficult to make statements about the functional importance of an amino acid position with limited mutagenesis studies. For example, in the PDZ domain different mutations at the same position can have different functional effects [1-2]. The second difficulty in interpreting mutational studies is that when we measure the effect of a mutation, we often choose some measurable aspect of a protein's perceived importance in the context of a biological

system. Too often, these functional parameters are limited to just stability or just binding affinity [3-6], and rarely consider the additional constraints inherent to functioning in the context of a cell or an organism. In the preceding chapters we presented the development of an assay that permits the measurement of the function of many PDZ3 mutants in the context of the *E. coli* cytoplasm, a measurement that integrates many constraints on protein function. In this chapter, we return to the biological questions that motivated this technical innovation. What are the patterns of functional importance of amino acids in the PDZ domain? What positions have intrinsic functional importance, and what positions have functional importance through their interaction with other positions? What correlation do we see between the PDZ-family level patterns of amino acid coevolution and the PDZ3-level patterns of functional effects? What can we say about the evolutionary-timescale constraints on protein function from these patterns?

Previous studies in several protein model systems have demonstrated the importance of sector positions in protein structure and function [2, 7-13]. Like many other mutational analyses, these studies measure the effect of a limited number of mutations to sector positions. For example, mutation of one-third of the sector positions in the bovine GPCR $G_{s\alpha}$ shows the importance of sector positions, but not non-sector positions, in the coupling of nucleotide-binding to effector-binding [8]. However, only a fraction of all the positions in the protein are characterized, and most of the characterized positions are mutated to alanine. Similar small-scale alanine-dominated mutation studies in other proteins demonstrated the general importance of sector positions for protein function [7, 10]. Unlike random or biophysically-directed mutagenesis studies, these studies make mutations to test an evolutionary model of amino acid function in proteins – that sector positions represent the core determinants of protein function. However, the global nature of the sector hypothesis requires a global experiment for validation.

Here we report the results of a comprehensive single mutagenesis study of PDZ3 in which we measure the functional effect of mutating every position to every amino acid. As a step towards more completely understanding the interaction of all amino acids in PDZ3, we carry out a second comprehensive mutagenesis in the context of a mutation to the PDZ3 peptide ligand. This allows us to observe those positions that contribute to specificity changes of PDZ3 and to identify those positions that engage in cooperative interactions with the mutated peptide position. This thesis concludes with a comparison of the experimental patterns of functional effects from single and pairwise mutations in PDZ3 and the statistical patterns of amino acid coevolution in the PDZ family.

Comprehensive Analysis of the function of single mutations in PDZ3

As a first-order approach to measure the functional importance of every amino acid in PDZ3, we created a library that encompassed the entire sequence space with a Hamming distance of one relative to the wild-type PDZ3 sequence [14]. This comprehensive local sequence space walk should display the range of functional effects possible at every site in the protein, not just the effect of a single alanine mutation at a few sites. It seemed likely that a number of mutations would be highly non-conservative – introducing polar amino acids into the core, placing hydrophobes on the surface, substituting prolines in the middle of helices. However, the measurement of the functional effect of every mutation at each position should reveal the average nature of mutations to a position.

Construction and measurement of all single amino acids mutations

To build the library of all single amino acid mutations in PDZ3, we employed oligonucleotide-directed mutagenesis of each position individually, as detailed in the methods section. The basic strategy was to use a biased-randomization approach in which each position in the conserved region of PDZ3 was replaced with every codon encoded by 'NNS,' where 'N' is any nucleotide and 'S' is guanine or cytosine. This randomization strategy limits the complexity of the library for each position to 32 out of the 64 possible codons, while sampling all twenty amino acids and only one stop-codon. Similar to the approach of Chapter 3, the 'NNS' PCR products for all of the positions of a subgroup were mixed together in equimolar ratios and ligated into the pZS plasmid to create subgroup libraries of every amino acid mutation at every position within that subgroup (**Figure 4.1**). For each subgroup library, we carried out the bacterial 2-hybrid assay as detailed previously. Each library transformation yielded greater than 10^6 colonies. The top 10% and 25% of each library eGFP distribution were sorted using FACS, and the input and selected populations were sequenced using Solexa 75-basepair paired-end sequencing.

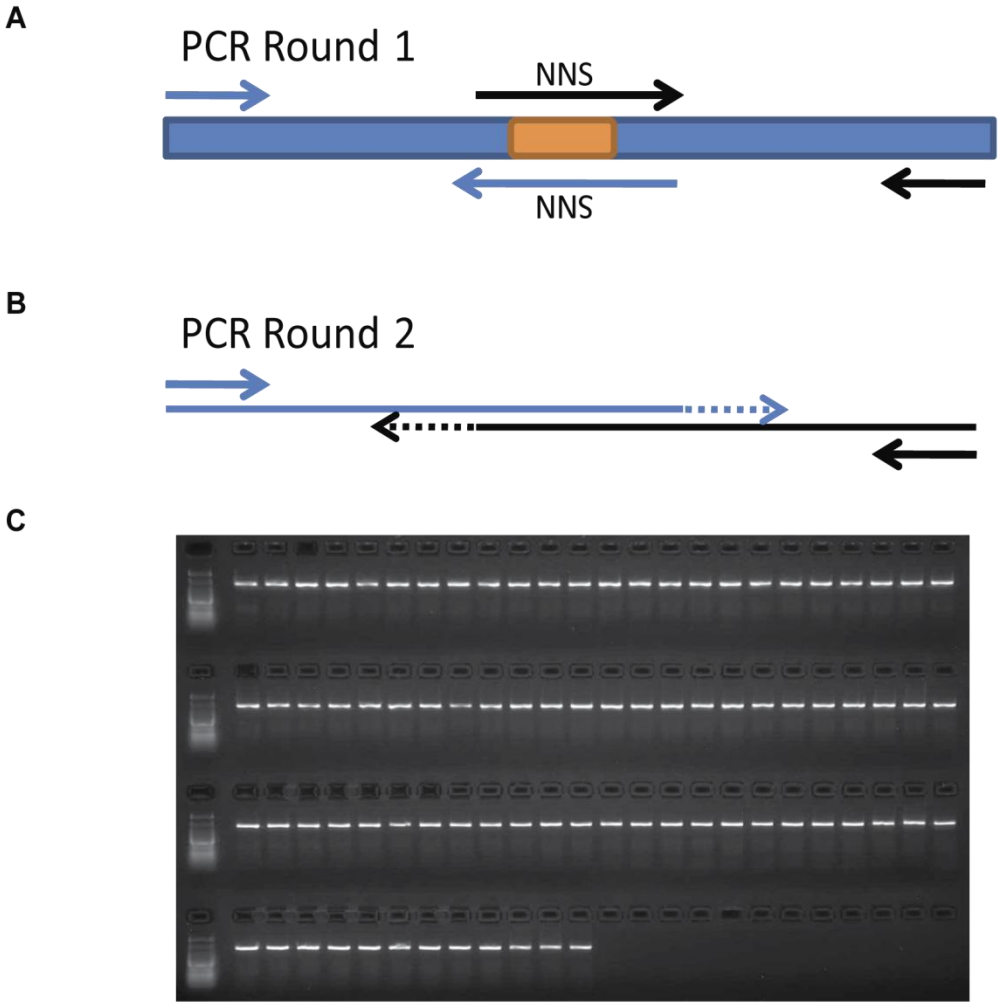


Figure 4.1 Construction of PDZ3 comprehensive single mutation libraries
Single mutant libraries were constructed with a two-round PCR protocol as detailed in the methods section. The first round consisted of two reactions, as shown with the black and blue primer sets. Each first round reaction utilized an oligo containg the 'NNS' codon and 15 complementary bases on either side of the randomized position (A). The use of 'NNS' samples all twenty amino acids and only one stop codon using 32 codons. The second round combined the two first round products and amplified to full length each variant (B). The size and purity of each PCR product was validated by electrophoresis, digested, and ligated into the pZS22 plasmid (C).

The distribution of effects of all mutations at all positions

The sequence data from each of the libraries was processed as detailed in Chapter 2 and Appendix I. Since we observed a strong correlation between the enrichment values for the top 10% and top 25% sorts, we averaged the two values to create a 20 amino acid x 83 position matrix of average enrichment values for every mutation at every position in the conserved region of PDZ3 (**Figure 4.2A**). To calculate a value for the enrichment of the wild-type allele, we averaged the enrichment values for every synonymous mutation in the PDZ3 sequence. The distribution of these individual synonymous enrichment values was rather tight ($\mu = 0.189 \pm 0.77$) (**Figure 4.2B**, dark gray histogram). To calculate the relative enrichment of each mutation, we subtracted this wild-type average from every enrichment value.

The matrix clearly shows the heterogeneity of relative enrichment effects at positions in PDZ3, both the differential effect of amino acids at a position and the differential effect of mutations at different positions. From the average effect of each amino acid at all positions, it is clear that some amino acids such as aspartic acid, glutamic acid, proline, arginine, and tryptophan tend to disrupt function independent of their position in the primary sequence. Proline has the largest deleterious effect on relative enrichment across positions, which makes sense in light of its unique conformational rigidity and ability to disrupt secondary structure elements [15]. In support of this, proline also has the second highest average penalty in the BLOSUM statistical substitution matrix [16]. Though some rows of amino acid effects appear more homogeneous than others, each amino acid has a heterogeneous effect across the primary structure.

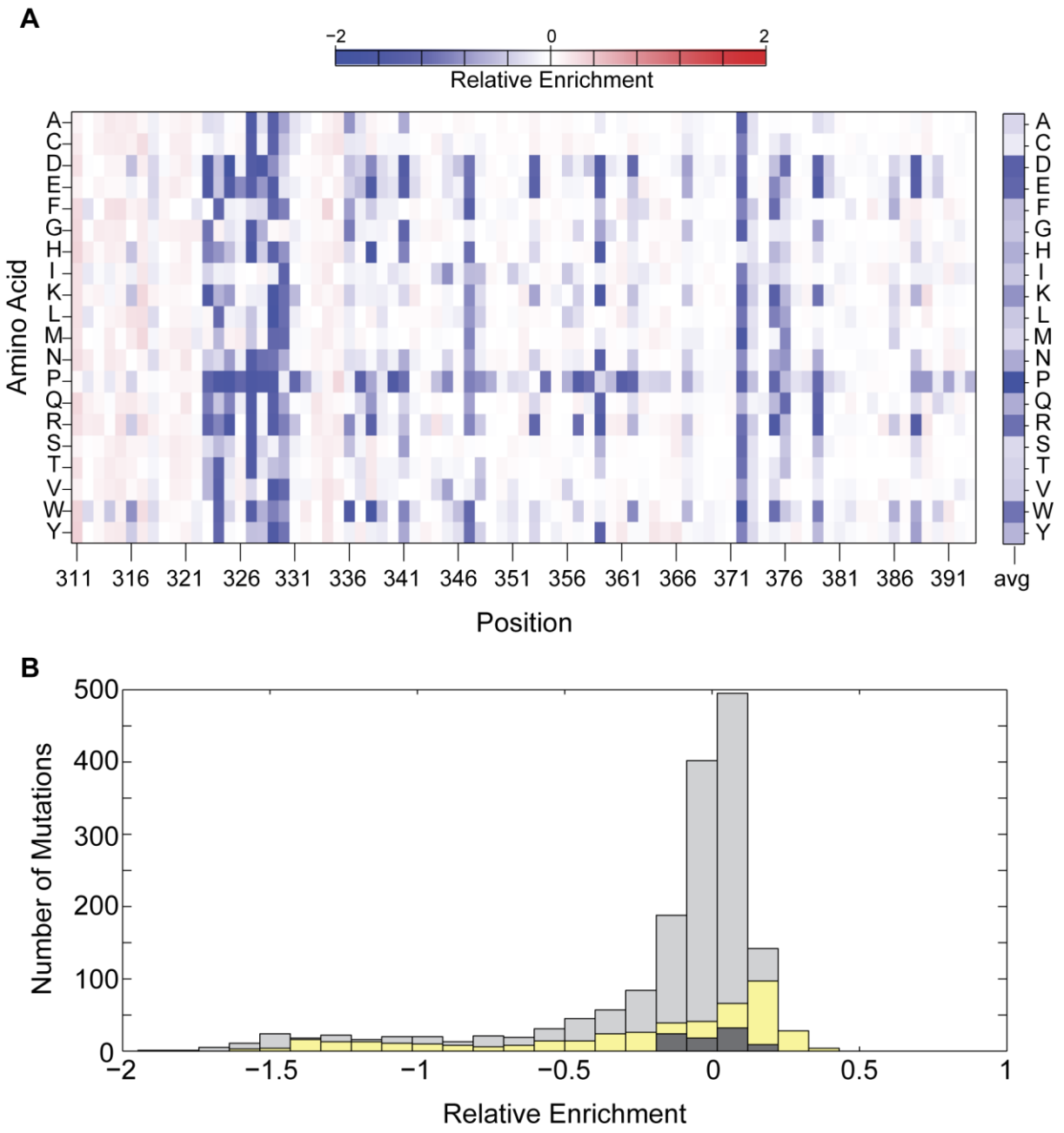


Figure 4.2 Functional effects of PDZ3 comprehensive single mutagenesis
The matrix of all single amino acid relative enrichment values at all 83 positions displays significant heterogeneity, both amongst amino acids at a positions and amongst positions for a given amino acid (A). In the histogram of the relative enrichment values in the matrix (B, light gray), significant robustness is evident since most mutations display values around zero; yet a dominant leftward tail is clear in the distribution . The distribution of all 83 wild-type enrichment values is shown in dark gray. Non-conservative mutations, those that do not occur at all in the Lockless PDZ alignment, are shown in yellow. Though non-conservative mutations make up a significant portion of the tail, there are also mutations in the tail that are found in members of the PDZ family.

The histogram of single mutation effects shows that most mutations have little or no effect on the function of PDZ3 (**Figure 4.2A**, light gray histogram). In fact, the main body of the distribution is centered around zero, with a width that includes a number of mutations that slightly decrease function as well as a number of positions that slightly increase function; yet, most of these effects are within the spread of effects we observe for wild-type (**Figure 4.2A**, dark gray histogram). Beyond the body of the distribution, there is a prominent leftward tail of mutations that produce a large decrease in PDZ3 function. One *a priori* expectation is that non-conservative mutations would be the most likely mutations to produce negative functional effects. While we do find the leftward tail of function-decreasing mutations to be enriched for mutations that do not occur in the PDZ alignment (**Figure 4.2B**, yellow histogram), this tail also contains a significant number of substitutions that are sampled by some member of the PDZ family.

Spatially, we observe that positions with many amino acid substitutions that effect function tend to be located in or near the peptide-binding pocket of the protein, as exemplified by position 372 (**Figures 4.3 and 4.4)**; however, we observe heterogeneity within the binding pocket – positions like 322 and 326 contact peptide, but show small effects for most substitutions – and at positions distant from the binding pocket such as position 359.

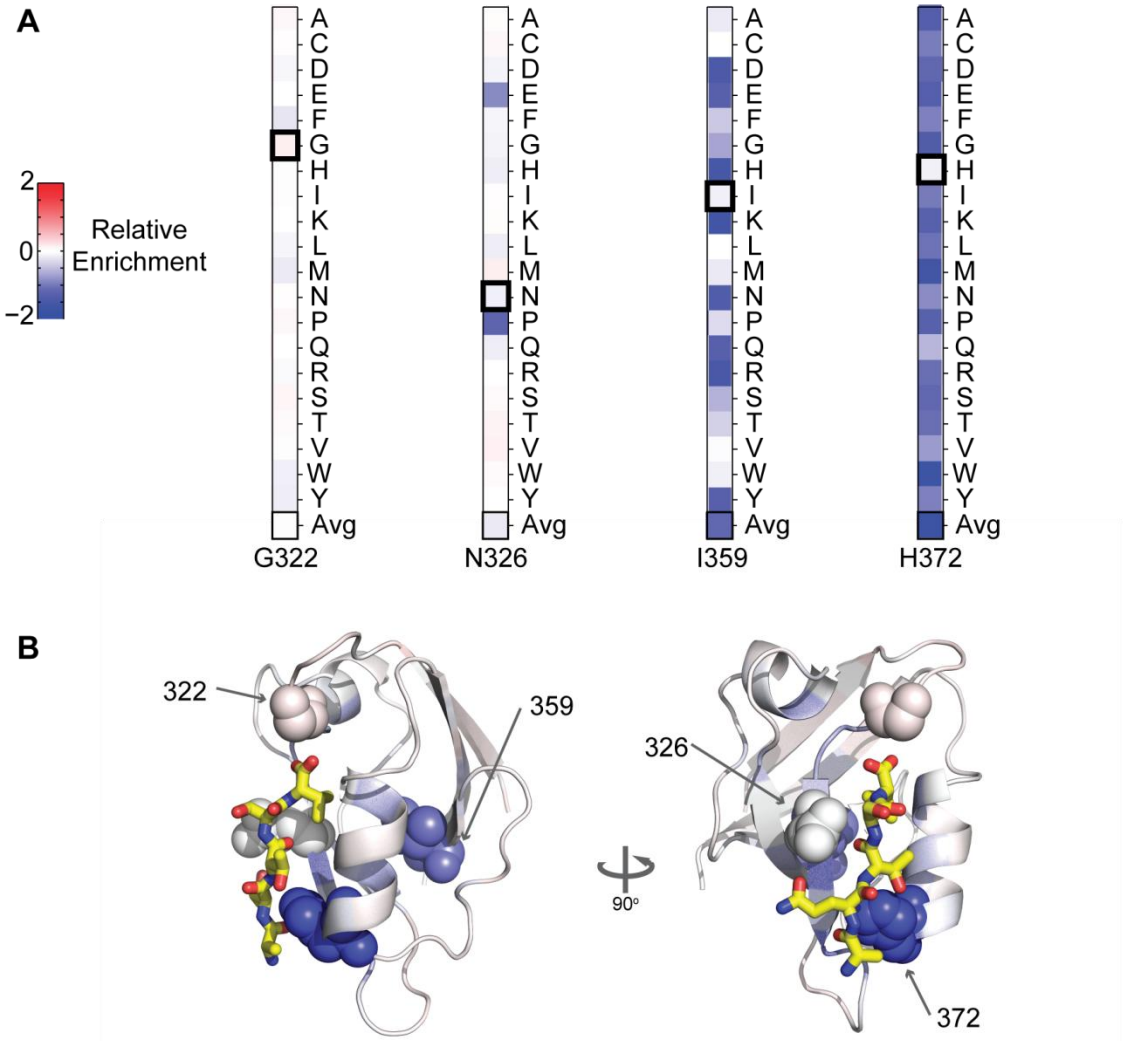


Figure 4.3 Spatial heterogeneity of functional effects in the PDZ3 structure
Positions in PDZ3 show a range of effects across the primary sequence, including some positions that have small effects for most substitions like positions 322 and 326 and positions that have large effects for all substitutions such as position 372 (A). The spatial distribution of functional effects in the PDZ3 structure shows positions in the peptide-binding pocket that contact peptide and have small functional effects and positions distant from the peptide that have large effects (B).

The positional effect of mutations

In order to more generally evaluate the functional role of each position in PDZ3, we summed the individual relative enrichment values for each position to calculate the cumulative

mutational effect of all mutations at each position, Ψ_x. Using the relative enrichment values for each mutation at each position, we calculated Ψ_x for each position 311-393:

$$\Psi_x = \sum_{i=1}^{20} Relative\ Enrichment_i^x$$

where i = an amino acid, and x = a position in PDZ3.

The distribution of positional effects shows a pattern similar to that of the individual mutational effects (**Figure 4.2B**, **Figure 4.4B**). This pattern of many neutral sites with a leftward tail comprised of a small number of positions with large cumulative effects demonstrates that the pattern of robustness in the domain is not a property of the particular amino acid substitution; that is, as discussed above, it could be that we see small effects for most mutations, but every position can significantly impact function with some particular mutation (**Figure 4.2A**). With the exception of a few amino acids, namely proline as discussed above, this is not the case. The actual pattern is one in which most positions show small effects for almost all substitutions; yet, there are a small number of positions at which most mutations result in significant functional effects (**Figure 4.4B**).

Further, from a first-order interpretation of the protein structure in accordance with the spatial proximity model [1], one might expect these positions of consistently large functional effect to be those positions that contact the peptide. The pattern of functional heterogeneity we observe in this experiment does not conform to this expectation of the importance of spatial proximity to the peptide (**Figure 4.4C)**. On the whole, we do observe functional effects upon mutation of the positions that line the peptide-binding pocket. However, the pattern is not homogeneous. Instead, we see positions that contact the peptide in the crystal structure such as

N326, S339, and K380 but have little effect on function. In fact, two of these positions pack against the P_{-2} position of the peptide, a key specificity and affinity determinant in PDZ3. In addition, the spatial distribution of the positions with significant effects is not limited to the peptide binding pocket, but extends to the backside of the molecule to V362, more than 10 Å away in the crystal structure.

Consistency of relative enrichment values with previous studies of PDZ

Though the literature is limited in studies investigating the effect of specific mutations in PDZ3 or PDZ domains in general, several published datasets are consistent with our results in PDZ3. As discussed in detail in Chapter 2, the crystal structure of PDZ3 in complex with the CRIPT peptide has been solved. In our experiment, we find a number of positions in the tail of the functional distribution that were shown to engage peptide in the structural analysis, including the canonical L323, G324, and F325 of the β2 strand that form hydrogen bonds with the terminal V at the P_0 position of the peptide [17]. In addition, positions 329 and 372 engage in a hydrogen bond with T(P_{-2}) and are found to have large effects on function in our assay. As mentioned previously, we also see a number of positions that contact the peptide yet show little functional effect upon mutation. This discordance emphasizes the inability of structural approaches to visualize energetically-important amino acid interactions.

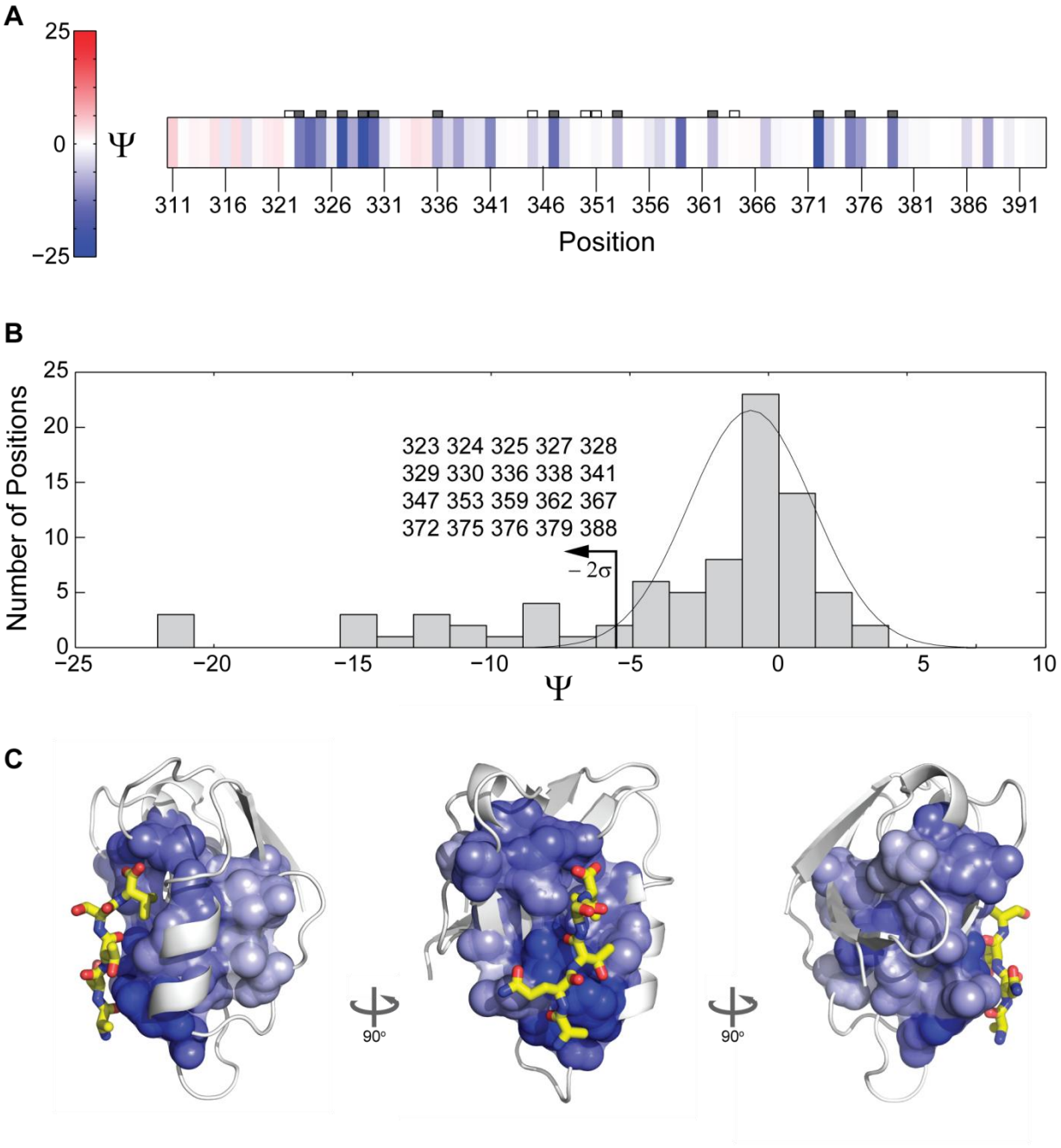


Figure 4.4 Positional relative enrichment effects of PDZ3 comprehensive single mutagenesis

The Ψ-value represents the sum of the relative enrichment values of all mutations at a position, and displays significant heterogeneity across the primary structure (A). Sector positions are shown as boxes above the Ψ-value vector, and sector positions included in the tail are shown as filled boxes (A). The distribution of positional values shows many positions of small effect and a tail of largely deleterious positions (B). When mapped on the structure, these positions in the tail are mainly located near the peptide; however, there are positions that contact peptide that show small effects and positions distant from the peptide that display large effects (C).

In a separate phage display-based binding assay, these same positions from the PDZ3 structure that engage the P_0 and P_{-2} positions were shown to be important specificity determinants in the Erbin PDZ domain [18]. The functional impact of mutations at these positions for binding the wild-type peptide was not measured. However, in a previous study 44 positions in and around the binding pocket of the Erbin PDZ domain were mutated to alanine and measured with phage-display for their ability to bind the endogenous peptide [19]. This study identified 17 positions that displayed the most significant decreases in peptide binding upon mutation. Of these positions, 13 have a homologous position in PDZ3, and 10 of these 13 positions are found in the tail of our relative enrichment distribution in the context of CRIPT peptide. This result is somewhat surprising given the fact that these studies employ two different PDZ domains, especially given Erbin PDZ's unusual preference for tryptophan at P_{-1} [20]. However, the independent approaches display significant consistency, and suggest a general conservation of functional mechanism in PDZ domains.

The relation of conservation and functional effects

The conservation of positions in multiple sequence alignments has been used in many studies to indicate the functional importance of a position in an individual protein [21-22]. With the present experiments we have a unique ability to be able to test the correlation between the conservation of the PDZ family and the functional conservation in PDZ3. We use the positional relative enrichment Ψ-values, as the correlative to global conservation of a position in the PDZ family (D_{global}). Calculated as previously reported, D_{global} represents the divergence of the observed amino acid frequency distribution from the background frequency of amino acids as

observed in the non-redundant sequence database [7]. In a plot of the values of the relative enrichment and conservation for all 83 PDZ3 positions, there is no strong correlation between the two values, only a general trend of decreased relative enrichment with increased conservation (**Figure 4.5**).

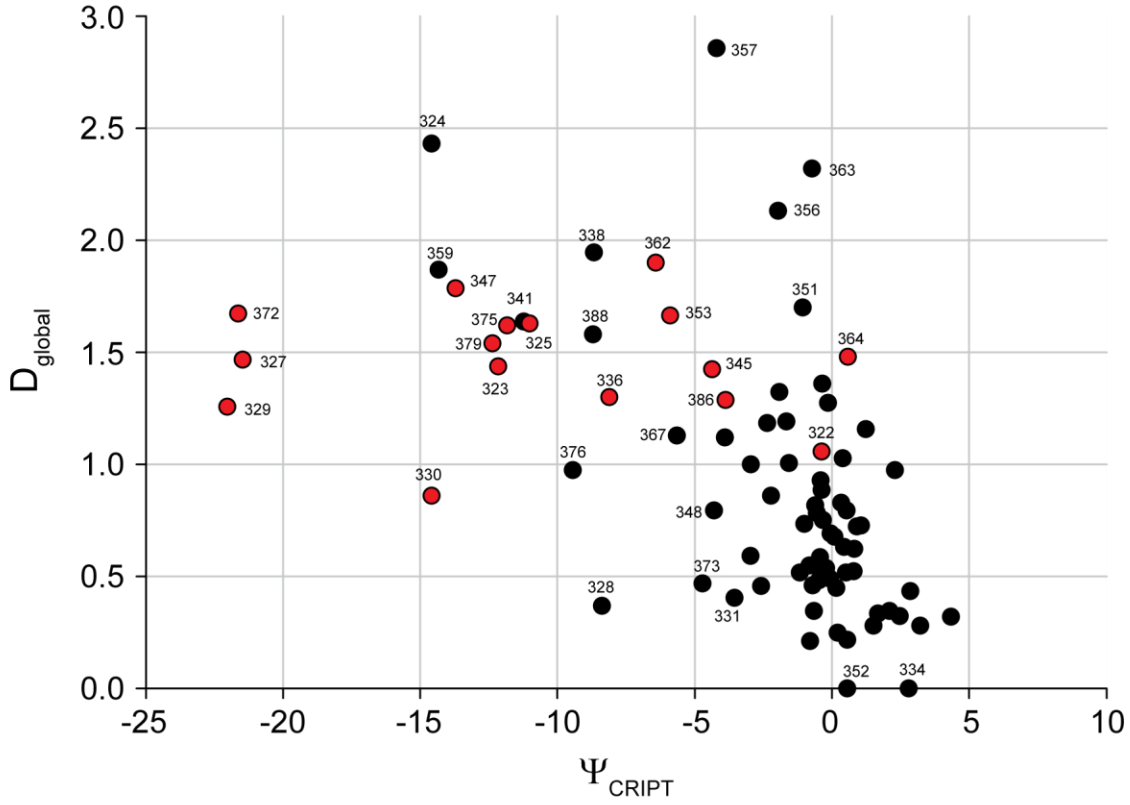


Figure 4.5 The relation of conservation in the PDZ family and positional relative enrichment

A comparison of the positional conservation in the PDZ alignment and empirically determined positional relative enrichment , Ψ, reveals a general trend of decreasing Ψ with increasing conservation. There is, however, no strong correlation between the two values. Three of the most highly coupled positions in the PDZ alignment (327, 329, and 372) display a much lower Ψ-value than expected from the general trend of the data. This may be due to the fact that these are functionally very important positions, but in the context of their co-conservation with other positions. Sector positions are shown as red circles.

Interestingly, three of the most highly covarying residues in the PDZ family (positions 327, 329, and 372) display larger functional effects than any other position with a similar degree of conservation. This underestimated functional importance from the sequence statistics could be a result of the fact that these positions display significant covariation, and as a result, no first-order conservation measurement can capture the selective constraints on these positions. However, overall we do not observe a general deviation from correlation between relative enrichment and conservation for sector positions (**Figure 4.5**, red circles). In addition, the lack of correlation between these metrics could be due to experimental analysis of relative enrichment in only one member of the PDZ family or the fact that it is unlikely that our assay captures all of the functional constraints of the natural fitness function of these proteins. Additionally, some of the positions that display a low relative enrichment but a high conservation like 357 could be artifacts of the phylogeny, conserved by chance and not by functional constraint.

The correlation of sectors and positions of significant functional effects

Previous studies in the lab have made the correlation between sector positions and positions of functional effects through mutagenesis of a selection of sector and non-sector positions [2, 7-8, 10, 13, 23]. The general finding was that sector positions were much more likely to impart a functional effect upon mutation than the control non-sector mutations. The characterization of the functional effects of mutation in the present study presents a unique data set for testing the association of sectors with the positions of most significant function in the PDZ domain. A number of previous mutagenesis studies in other protein systems (discussed in Chapter 1) have suggested a correlation between functionally important positions and conserved

positions or structurally buried positions [4, 24-31]. We reasoned that a simple statistical test of the ability of the degree of surface exposure, conservation, or statistical coupling to predict the most functionally important positions would reveal the respective ability of each of these computable parameters to predict functionally important positions. We used Fisher's exact test to calculate the probability that the extreme values of each classification (surface exposure, conservation, and statistical coupling) were associated with the tail of the relative enrichment distribution by chance [32]. Solvent-accessible surface area was calculated using the GETAREA algorithm for the PDZ3 structure (PDB ID, 1BE9) [17, 33]; conservation was calculated as previously described in this chapter [7]; the degree of statistical coupling was calculated as the maximum projection of a position upon the first or second independent component of the statistical coupling matrix [11]. To more generally quantify the significance of the association between relative enrichment and each of the classifiers, we calculated the *p*-value for the association of the positions in the tail of the relative enrichment distribution and a set of positions in rank order from each of the classifiers ranging in size from 10 to 80 positions (**Figure 4.6A, B**). In other words, the *p*-value for each classifier at the 10-position calculation represents the *p*-value for the association of the top ten positions in that classifier (e.g. the 10 most conserved positions) with the list of positions in the relative enrichment tail.

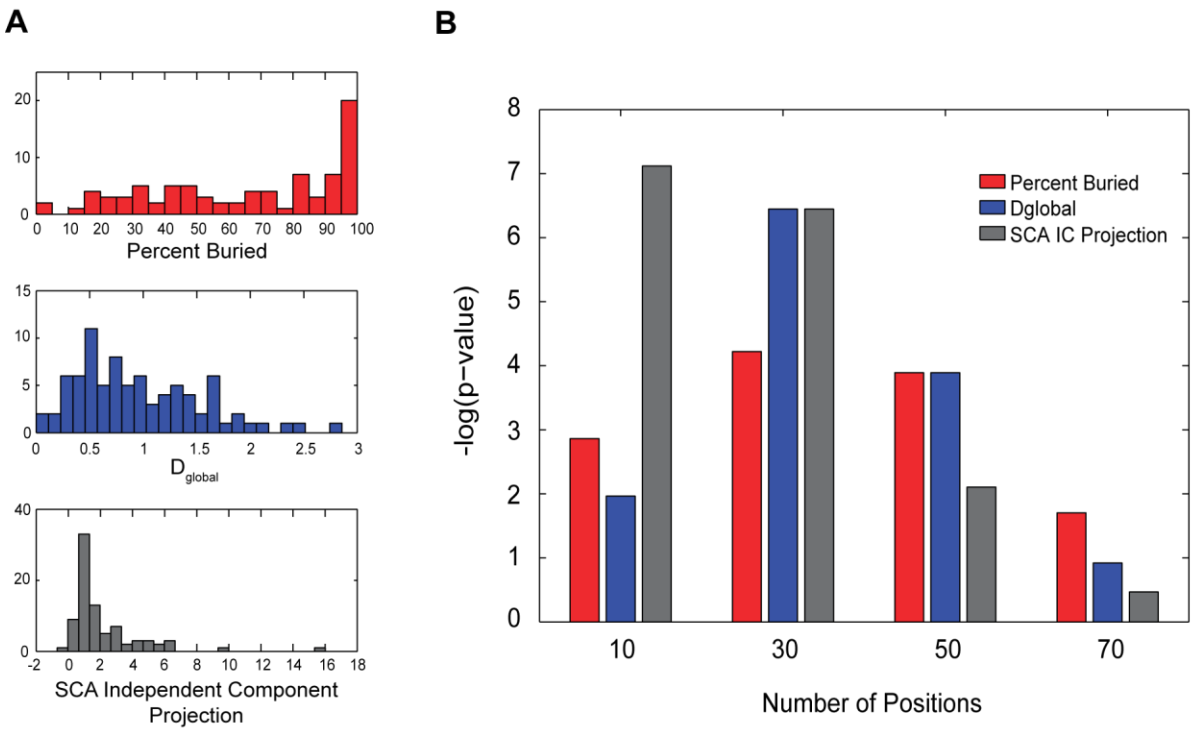


Figure 4.6 The statistical association of positions of functional effect with solvent exposure, conservation, and statistical coupling

The solvent exposure, conservation, and degree of statistical coupling were calculated as described in the text for all positions in the alignable region of PDZ3 (A). Fisher's Exact Test was used to calculate the improbability of the association of the tail of the enrichment distribution with each of the classifiers using a range of bin sizes for each classifier (B). Solvent exposure is not predictive of the enrichment tail at the 30-position bin size where all completely buried residues are included. Conservation displays a similar pattern of heterogeneity in which the most conserved residues are not the most predictive of the functionally important positions. Only the magnitude of a position's projection along IC1 or IC2 of the SCA matrix shows a relationship in which the most significantly coupled residues best predict the functionally important residues with a significance not reached by any other bin of any classifier.

With each classifier we see a general trend, as expected, of a decrease in the significance of the association as the number of positions included in the statistic increases. However, the *p*-values for the points with the fewest positions in the classifier shows the degree of statistical coupling to be a better predictor of functional importance in PDZ3 than conservation or surface exposure. This means that the most conserved positions are not necessarily the most functionally

important positions – there is heterogeneity within this group. Similarly, only some of the positions in PDZ3 that are completely buried are functionally important. However, those positions that display the greatest degree of covariation in the PDZ family are the best predictors of functionally important positions in the single homolog PDZ3.

One important point for this interpretation is that the properties of conservation, surface exposure, and covariation are not independent properties. Conserved positions tend to be buried, and positions in the core tend to have large functional effects in part due to their dense packing. Further, statistically coupled positions are, by mathematical definition, moderately conserved. In spite of this interdependence of these classifiers, we still demonstrate the superior ability of the degree of statistical coupling to predict functionally important positions over conservation and surface exposure.

In the final analysis, we do not see a perfect correlation of sector positions with those functionally important positions our assay. The tail of the relative enrichment distribution is enriched for sector positions, but also contains a number of non-sector positions. We expected this idiosyncrasy to some extent. For example, the positions that display large functional effects should be those positions that have a large intrinsic effect upon mutation and those positions that engage in cooperative interactions. We reasoned that these types of positions should be distinguishable through a pairwise mutation analysis; an experiment made possible by our high-throughput functional assay.

Comprehensive second-site mutation analysis of PDZ3

The distribution of effects of all mutations at all positions in the T(P_{-2})F background

Mutations with large functional effects in the context of binding CRIPT peptide could either generally disrupt function, independent of the peptide itself, or could specifically disrupt function in the context of binding CRIPT. We decided to carry out a massive thermodynamic mutant cycle analysis, more than an order of magnitude larger than any study, in which we measured 1660 couplings – the difference in the functional effect of every single amino acid mutation in the wild-type background relative to the functional effect of each mutation in the background of a second mutation. For the second mutation site, we chose a mutation at a position in the peptide which we know to have specificity- and affinity- effects upon mutation [18, 23]. We decided that a mutation of large effect would give us the best chance of resolving coupling over the noise inherent to any in-vivo measurement. We made all single-amino acid mutations in the background of a threonine to phenylalanine mutation at the P_{-2} position of the peptide (**Figure 4.7A**). This converts the peptide from a Class I to a Class II PDZ ligand and decreases the affinity of wild-type PDZ3 by 50-fold [23].

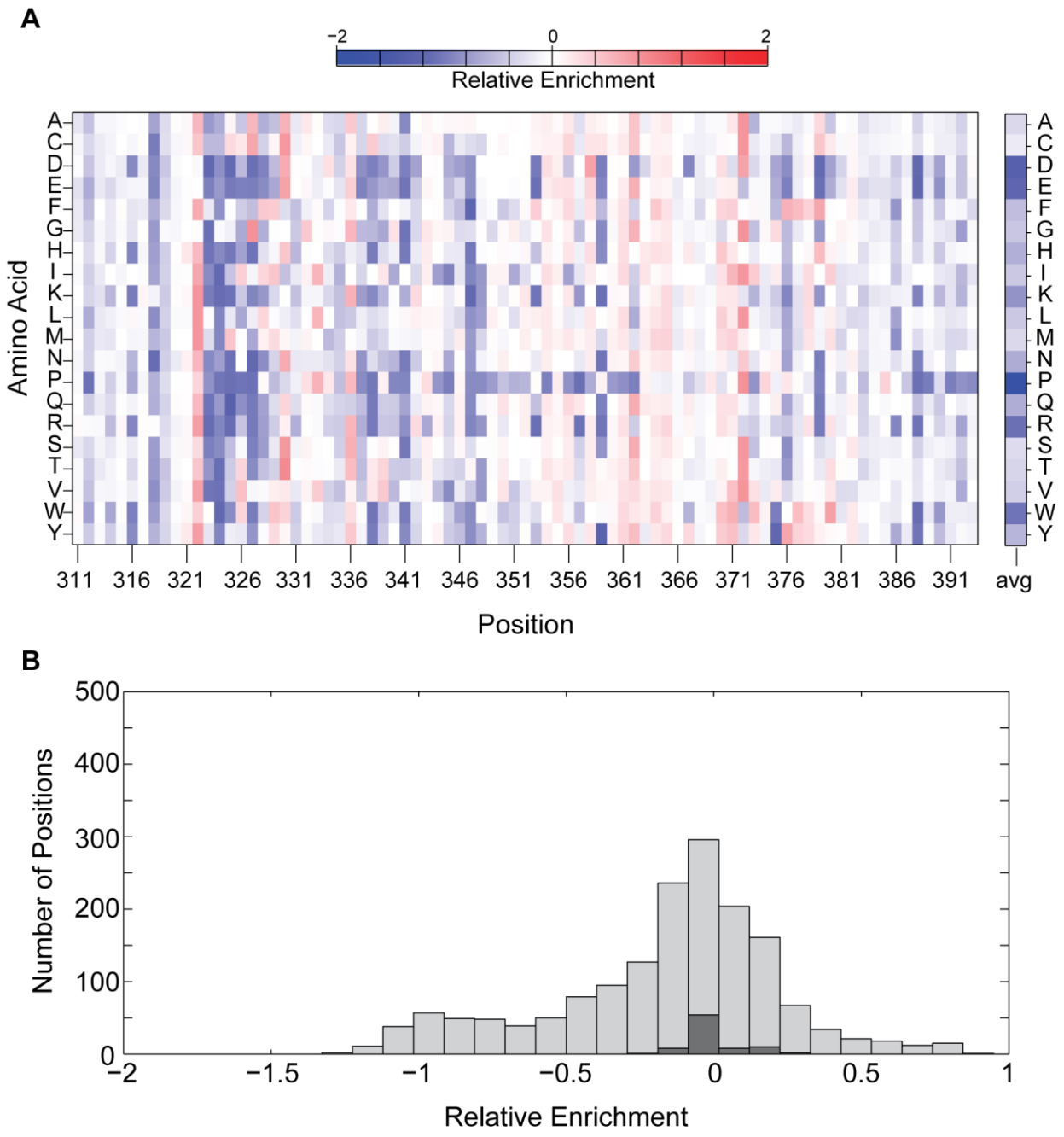


Figure 4.7 Functional effects of PDZ3 comprehensive single mutagenesis in the T(P_{-2})F background

Similar to the relative enrichment matrix in the CRIPT background, the T(P_{-2})F matrix shows the heterogeneity of PDZ3 in the patterns across positions and across amino acids (A). Again, the histogram of values shows that most mutations have enrichment similar to that of wild-type (B, dark gray histogram). In contrast to the relative enrichment effects in the wild-type backgroud, the T(P_{-2})F background displays a two-tailed distribution in which a few mutations display significantly increased enrichment (B, light gray histogram).

The distribution of single mutation effects in the background of T(P_{-2})F looks significantly different from the distribution of effects in the wild-type background (**Figure 4.7B**). Both distributions are centered around zero relative enrichment, but the T(P_{-2})F distribution exhibits significantly more spread. Importantly, the T(P_{-2})F distribution also shows a number of mutations that improve function as well as a number of positions that decrease function, in contrast to the lack of mutations in the wild-type background that largely improve function. This is likely a result of the fact that the affinity of PDZ3 for the T(P_{-2})F peptide is much lower than its affinity for CRIPT. As a result of the optimization of the PDZ3-CRIPT interaction, it may be that there are no single-mutation paths to improving wild-type PDZ3 function in the context of CRIPT binding. Overall, we still observe a pattern in which most mutations produce small functional effects, while a minority of positions produces the tails of increased and decreased function in the context of T(P_{-2})F.

The positional effect of mutations in the T(P_{-2})F background

In the distribution of Ψ-values for positional effects in the background of T(P_{-2})F, we see a two-tailed distribution, as with the distribution of single mutation effects (**Figure 4.7B, Figure 4.8B**). This is again in contrast to the distribution of positional terms for mutations in the wild-type background. As with the wild-type background, we observe a number of positions in the peptide binding pocket that display significant Ψ-values, both enriching and depleting. Of the positions in the binding pocket, most produce a decrease in function; however, the area around the site of mutation (**Figure 4.8C**, indicated with a labeled arrow) including position 372, and position 322 in the top of the peptide binding pocket both display rescuing functional effects in

the context of the T(P_{-2})F mutation. As with the wild-type background, we observe a heterogeneous spatial distribution of the positions with significant Ψ-values. Many are located in the peptide binding pocket, but a group of positions extending from the peptide binding pocket and all the way to the back side of the protein also display significant functional effects (**Figure 4.8C**).

Consistency of relative enrichment values for T(P_{-2})F with previous studies of PDZ specificity determinants

In the phage-display screen for mutations that affect specificity of the Erbin PDZ domain, the only positions identified as directly controlling specificity in the context of the P_{-2} position were the α2-1 and α2-5 positions; in PDZ3 these are positions 372 and 376, both of which we find to affect binding of T(P_{-2})F [18]. The Sidhu group found that positions in β2-β3 did not change the specificity of the PDZ domain in question, but we find these positions to significantly decrease function in the context of T(P_{-2})F. Since phage-display only identifies tightly-binding ligands, it makes sense that these positions were identified as non-contributors. In this respect, the quantitative bacterial 2-hybrid represents a significant improvement since it permits the direct measurement of mutations with both positive and negative functional effects. The fact that the Sidhu group was unable to identify positions outside of the binding pocket that impact specificity was likely due entirely to their choice to mutate only those positions that contact peptide in the crystal structure.

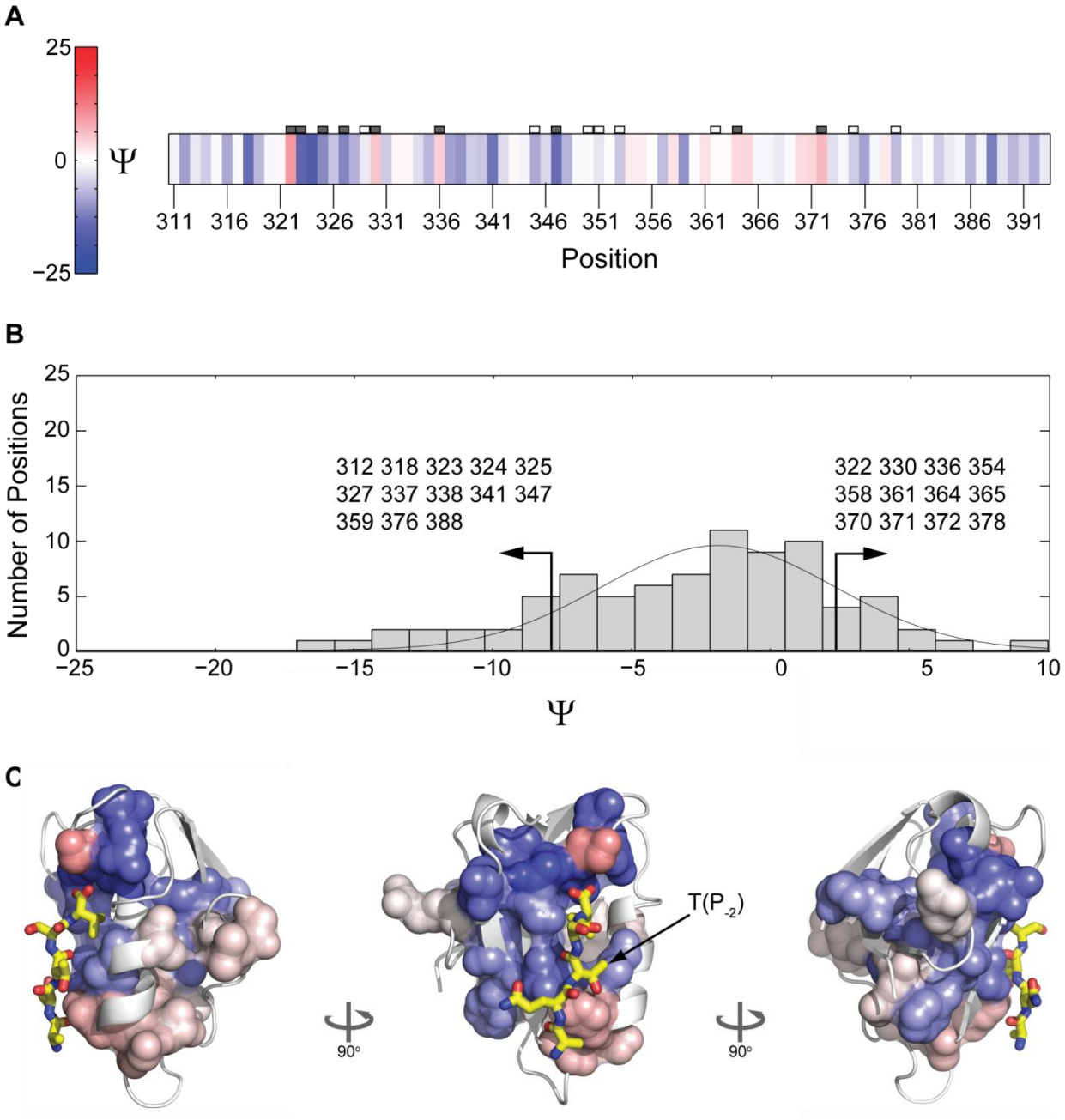


Figure 4.8 Positional relative enrichment effects of PDZ3 comprehensive single mutagenesis in the context of T(P_{-2})F

The Ψ-value displays significant heterogeneity across the primary structure (A). Again, sector positions are shown as boxes with sector positions in the tail shown as filled boxes. The histogram of Ψ-values shows two dominant tails, in contrast to the CRIPT distribution (B). In the structure, we observe a number of positions close to the position of mutation (indicated by the labeled arrow) that show large effects, but not all positions near T(P_{-2})F show significant effects. Also, positions with substanital effects are found distant from the site of mutation - at the top of the peptide-binding pocket and on the back side of the protein (C).

Non-additivity of amino acid interactions in PDZ3

The nature of the SCA calculation suggests that those positions identified as part of the sector should be functionally important, but this functional importance may only be measurable in the context of higher-order mutations. Depending on its architecture, a highly coupled network of interacting amino acid could produce large effects upon perturbation of any member of the network. Alternatively, the network could provide significant robustness to single perturbations and only exhibit functional effects upon mutation of two or more coupled positions, as is the case on the genomic scale with synthetic lethality [34]. These considerations led us to believe that while the group of positions that produce significant functional effects may contain a number of non-sector positions, sector positions may become more selectively important in the context of higher-order mutations. In order to calculate the global coupling of the T(P_{-2})F mutation to the rest of the protein, we used thermodynamic mutant cycle analysis to calculate the extent to which every mutation at every position is coupled to the T(P_{-2})F mutation [35]. Previous analyses of a limited set of couplings have produced conflicting results. Some studies have demonstrated the existence of significant long-range couplings [36-39], while others suggest that couplings with the local environment of an amino acid are the dominant couplings [1, 27, 40]. One study found that positions in the core of staphylococcal nuclease showed no significant higher-order couplings beyond pairwise couplings, but again this was in the context of stability measurements [41]. Other research into the function of statistically coupled residues in the *Shaker* voltage-gated potassium channel has demonstrated that the significance of energetic couplings amongst coupled residues increases with the order of coupling measured; that is, the higher-order couplings are the most energetically strong [42]. Yet, every one of these studies is based upon an interpretation of a very limited set of measurements, in some cases only

of positions in the core of the protein [27]. Our measurement of the effect of every single amino acid mutation in the wild-type background and in the T(P_{-2})F background allowed us to calculate couplings of every mutation at every positions with the T(P_{-2})F mutation in a comprehensive manner not previously demonstrated.

The distribution of positional couplings in PDZ3

The coupling of every position with the T(P_{-2})F mutation was calculated by subtracting the Ψ-value of a position in the wild-type background from the Ψ-value of a position in the T(P_{-2})F background (**Figure 4.9)**:

$$\Delta\Psi_x = \Psi_x^{T(P-2)F} - \Psi_x^{PWT}$$

Overall, many of the same positions fall into the tails of the CRIPT and T(P_{-2})F Ψ-value distributions (**Figure 4.10A)**. As a result, these positions display additivity. For example, positions 324, 359, and 347 all display large functional decreases (Ψ-values less than -10) in both backgrounds and as a result fall near zero in the ΔΨ distribution, indicative of an additive interaction with T(P_{-2})F. One might expect that these positions of large yet additive effects would be those positions that engage the peptide and when mutated, indiscriminately disturb function. This is not the pattern we observe; while position 324 does contact peptide, position 359 lies greater than 9 Å from any peptide amino acid.

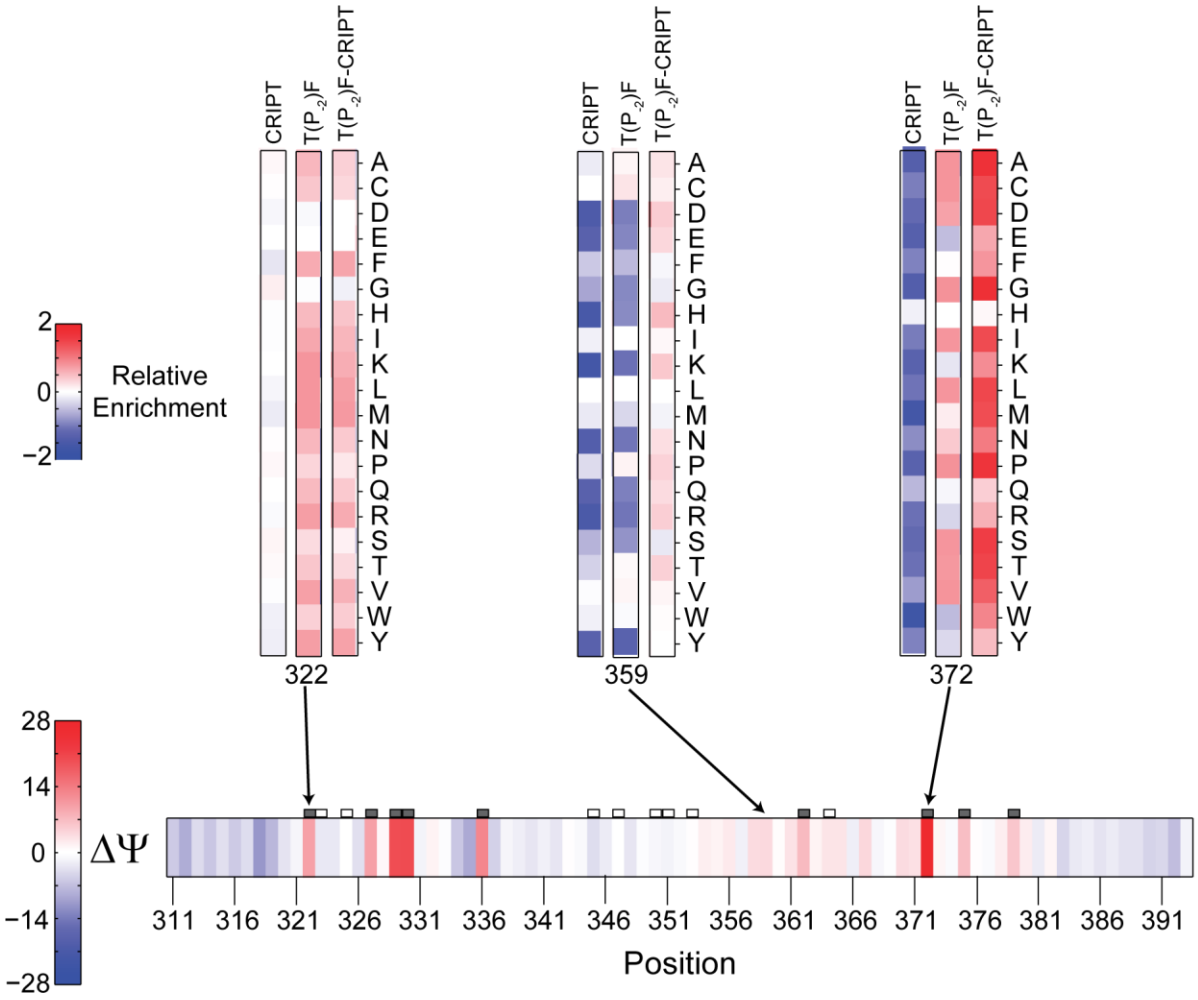


Figure 4.9 The origin of non-additivitive functional effects
The positions of strong non-additivity arise from positions dominated by functional effects of opposite sign in the context of CRIPT and the Class II, T(P_{-2})F peptide, respectively. occur across the primary structure (A). As shown in the filled boxes, not all sector positions display non-additivity with the P_{-2} mutation, but every position that exhibits strong non-additivity is a sector position.

The prominent feature of the ΔΨ distribution is the rightward tail of non-additivity. These positions project strongly away from the body of the additivity distribution. Nine positions fall in this strong rightward tail of the additivity histogram (**Figure 4.10A**). These positions are located structurally at the base of the peptide binding pocket near the site of mutation (**Figure 4.10B**, labeled arrow), at the top of the peptide binding pocket in the carboxylate-binding loop, and extend to the back of the domain (**Figure 4.10B**). Displaying one

of the largest degrees of non-additivity, position 372 sits at the base of the α2-helix and packs against the site of mutation. It makes sense that this packing interaction might contribute non-additively to the T(P_{-2})F mutation since the local environment seems the most structurally-likely place to interact non-additively with an amino acid; however not all of the positions that pack against the P_{-2} position in the structure display this non-additivity, demonstrating significant heterogeneity even in the local environment of the mutation. As observed with the single mutation effects, positions at significant distance from the site of mutation also display strong non-additivity. Position 362 lies on the opposite side of PDZ3 relative to the peptide binding pocket, almost 10 Å from any peptide amino acid. Position 322 lies at the top of the peptide binding pocket, far from contacting P_{-2}, and displays strong coupling to the T(P_{-2})F mutation presumably through the peptide itself. This through-peptide coupling has also been demonstrated through statistical and NMR studies (Bill Russ and Alan Poole, Ranganathan Lab, UTSW, unpublished).

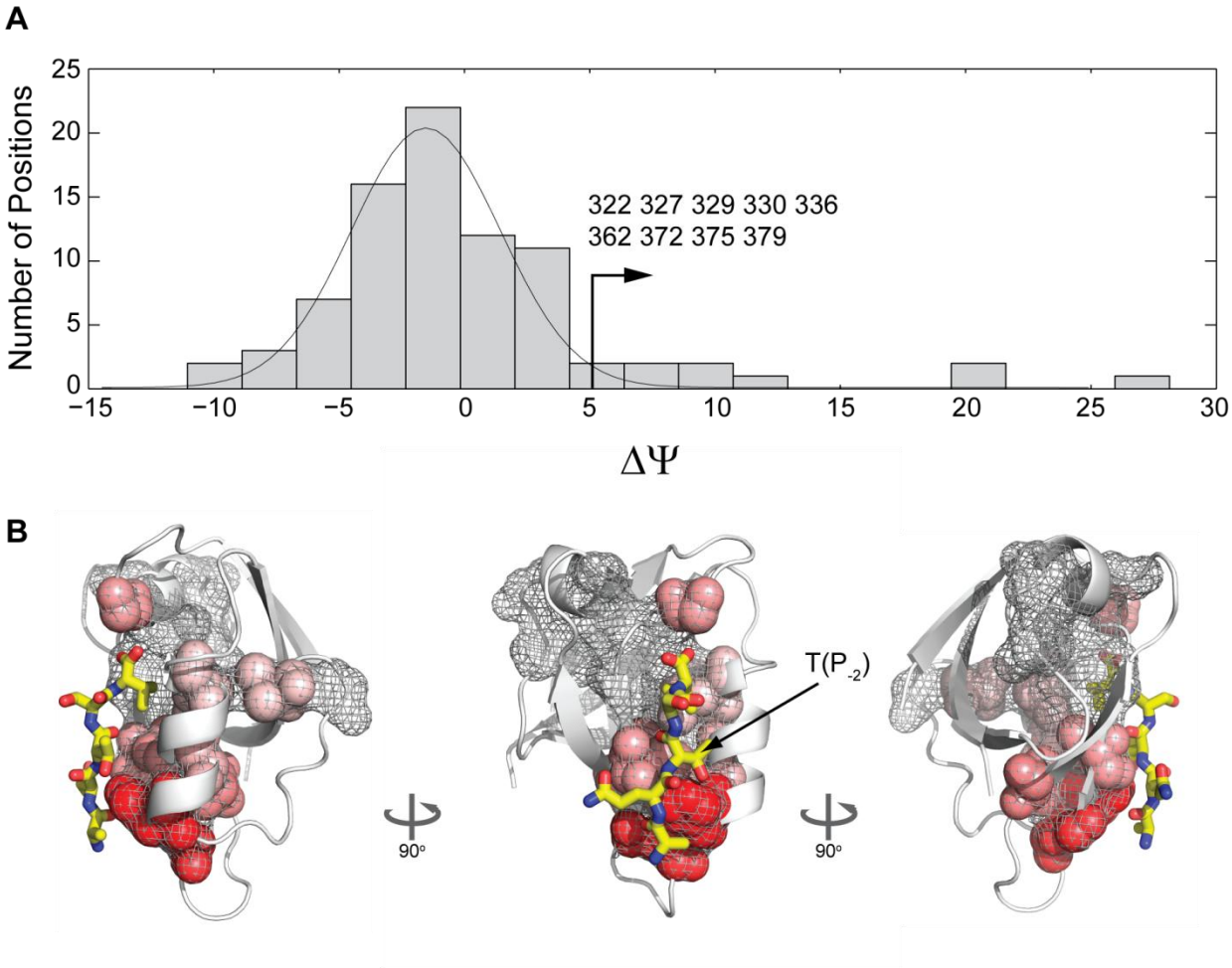


Figure 4.10 The distribution of non-additivite functional effects and their relation to sectors

The histogram of additivities shows a clear rightward tail composed of nine positions that display significant non-additivity (A). These non-additive positions occur close to the site of the second mutation (indicated by the labeled arrow), but also extend to the carboxylate binding loop and to the back of the protein, two sites implicated as important for allosteric pathways in the PDZ family (B). As shown in the wire mesh on the structure (B), every position that exhibits strong non-additivity is also a sector position.

Interestingly, two of the positions shown to exhibit strong coupling to T(P_{-2})F have been implicated in the function of allosteric pathways in other PDZ domains [43-44]. When in complex with Cdc42, the Par-6 PDZ domain has a 13-fold increased affinity for peptide compared to Par-6 alone. In Par-6, mutation of P171 was shown to disrupt this allosteric coupling. Intriguingly, in PDZ3, position 322 aligns to position 171 of Par6. This suggests that

the cooperativity of this position may be conserved in the PDZ family. In addition, another non-additive position distant from the peptide, position 362 has been implicated as a site of changes in dynamics upon peptide binding [45] and a possible site of peptide binding in the GRIP-1 PDZ domain [46]. These consistencies across PDZ homologs suggest that these observations of non-additivity are unlikely to be idiosyncratic features of PDZ3, but instead conserved features of the PDZ family.

The correlation of sectors and positions of non-additive functional effects

When we compared the tail of the additivity distribution to the sector positions in the PDZ family, we find that all of the highly non-additive positions in PDZ3 are sector positions (**Figure 4.9 and Figure 4.10B**). This is a striking result given the potential complexity of the interactions in the PDZ domain. This result suggests that SCA successfully identifies those positions that participate in higher-order interactions. Further, the fact that all of the experimentally coupled positions belong to the sector suggests that all of the pairwise interactions of $T(P_{-2})F$ with the rest of the positions in PDZ3 are evolutionarily conserved. If this were not the case, we would observe experimental couplings, idiosyncratic to PDZ3 that were not reflected in the statistical coupling analysis. One caveat to these results is that only a portion of the sector is recapitulated by the pattern of non-additivity in PDZ3. This could be because we only look at the coupling with $T(P_{-2})F$. Every position in the sector may be found to participate in a higher-order interaction with other sector positions in the protein if this type of analysis were carried out.

The coupling of fragility and evolvability at sector positions

The measurement of non-additivity at a position can arise through two scenarios – the effect of a mutation could be neutral in one background and functionally significant in the other background, or the effect of a mutation could be functionally beneficial in one background and functionally deleterious in the other (**Figure 4.11**). Each of these designs entails distinct functional characteristics. If an organism exists in an uncertain environment, it would be beneficial for a protein to be built with an architecture in which many mutations increased promiscuous functions but also were neutral for the current function. We observe this pattern in the functional effects of position 322. Mutations at position 322 in the context of the CRIPT peptide are largely neutral, yet these same mutations in the context of the T(P_{-2})F peptide are functionally beneficial. This pattern is consistent with the previous hypothesis that this position is important for evolvability of the PDZ domain [23]. This type of specificity-decreasing feature of the protein would however be detrimental in the context of a stable environment since non-specific interactions might entail fitness costs.

In contrast to the pattern of position 322, several positions in the tail of the additivity distribution display a pattern of effects in which most mutations decrease function in the context of CRIPT, but increase function in the context of T(P_{-2})F. These positions – 330, 336, 362, and 372 – produce the strongest coupling values since their effects in each background are of opposite sign. This pattern implies that when a mutation occurs at one of these positions, it is likely to decrease function. However, that same mutation that decreases function for the current peptide also increases function for a different peptide, thereby decreasing specificity and enabling a promiscuous, potentially adaptive interaction. This feature may have evolved to make

the most of the inevitable fragility of the complex system by coupling this fragility to evolvability.

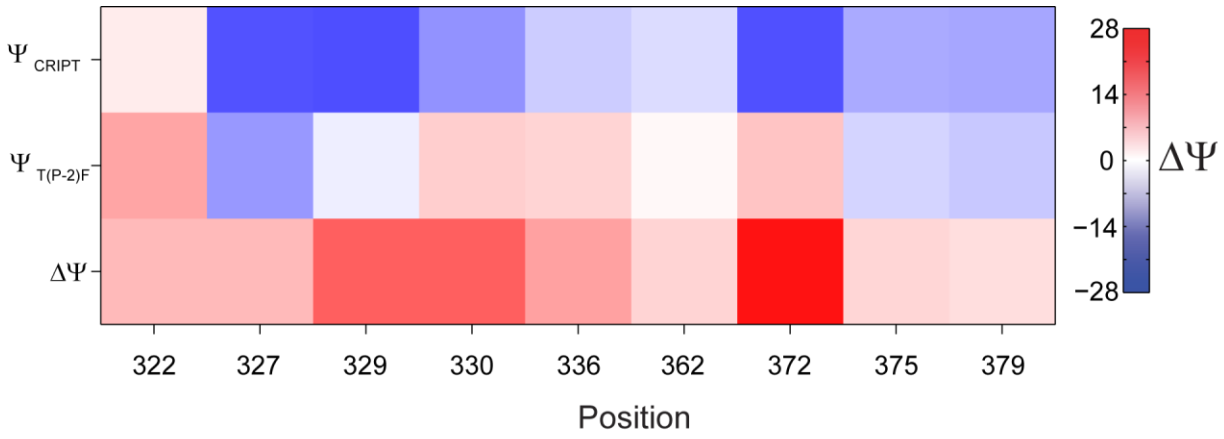


Figure 4.11 Relative enrichment of highly non-additive positions
The relative enrichment values in each background and the additivity of those measurements for the nine positions found in the tail of the additivity distribution show three patterns. Mutations at position 322 in the CRIPT background are largely neutral, while mutations in the T(P-$_2$)F background are beneficial. Mutations at this position would decrease specificity, but potentially increase evolvability. Positions 327, 329, 375, 379 have strongly buffered negative effects in the T(P$_{-2}$)F background, but still decrease relative enrichment in both backgrounds. Positions 330, 336, 362, adnd 372 have deleterious effects in the CRIPT background and enriching effects in the T(P$_{-2}$)F background. This pattern of effects results in the coupling of the susceptible portions of the protein to evolvability, since mutations at these positions are likely to decrease the current function but enhance alternative function.

Conclusions

The picture of sensitivity that emerges from these experiments is one in which most mutations at most positions have small functional effects; in other words, PDZ3 is robust to mutations. However, we observe a limited subset of positions that when mutated significantly abrogate the function of the protein. We demonstrate that those positions that elicit large functional effects when mutated are better predicted by their degree of coevolution than by their degree of conservation or by the degree to which they are buried. This result in and of itself provides significant credence to the sector hypothesis due to the fact that this conclusion comes

from a global study of all mutations at all positions in PDZ3; these are the patterns in the protein, not a likely model constructed from a limited set of mutations.

One could argue that this definition of robustness of the PDZ domain is incomplete as a result of the fact that the distribution of single mutation effects may not be reflective of the functional effects of higher-order mutations. That is, even though the additive functional effect of two mutations may be small, if most mutational effects are non-additive, the average functional effect of a second mutation could be large and functionally impairing. In stark contrast to this picture, we show that only a very limited subset of all amino acids engage in significantly non-additive interactions with a mutation at the P_{-2} position of the peptide. Importantly, every position that displays significant coupling to $T(P_{-2})F$ is also a position in the PDZ sector. This comprehensive analysis of single- and pairwise mutations clearly demonstrates the crucial functional role of sector positions as residues involved in cooperative, functionally-important interactions in PDZ3. Further, studies of two positions shown to be non-additive in PDZ3 in this study have implicated these positions as important for allosteric processes in the other PDZ domains, suggesting a conserved role of these positions in cooperative interactions in the PDZ family.

One possibility is that this result, as measured by non-additivity with $T(P_{-2})F$, is not general for patterns of non-additivity throughout the protein. Preliminary data on the non-additive interactions of position 372 suggest that this pattern of highly selective non-additivity limited to sector positions may be general, but a conclusive statement requires more experiments. This limited non-additivity adds another layer of robustness to the construction of the protein, since most functional effects of mutation are likely to be additive and should not display a high

probability of crippling the function of the protein. The extension of this pattern of general additivity to higher than pairwise mutations remains to be seen.

Finally, from the analysis of the functional effect of every single amino acid mutation for binding CRIPT peptide, we observe that the set of positions that significantly decreases function is highly enriched for sector positions. In contrast, with respect to the class II T(P_{-2})F peptide the set of positions that significantly increases function is highly enriched for sector positions. This gives rise to the non-additivity we observe at some sector position, but it also has important implications for the architecture of the protein. This pattern implies that if a mutation occurs at a sector position a significant decrease in the function with respect to CRIPT binding will occur, but overall the probability is small of a random mutation decreasing function. However, the very same positions that when mutated cause decreased CRIPT function also cause increased class II peptide function. This demonstrates that the architecture of the protein directly couples evolvability to the fragility of the robust system and suggests an optimized architecture in which it is unlikely that a mutation has a significant functional effect, but if it does that same mutation is likely to enhance a novel function.

Methods

Construction of single mutation PDZ3 libraries

Single mutant libraries were constructed using oligonucleotide-directed mutagenesis of PSD95-PDZ3. In order to mutate each alignable position in PDZ3 (positions 311-393), two mutagenic oligos (one sense, one antisense) were ordered (IDT) that contain complementary sequence to 15 basepairs on either side of the targeted position. For the targeted position, the oligos contain 'NNS' codons, where 'N' is a mixture of 'A', 'T', 'C', and 'G', and 'S' is a mixture of 'G' and 'C'. This biased randomization results in 32 codons with all 20 amino acids sampled – a

significant decrease in library complexity without loss of amino acid complexity. One round of PCR was carried out with either the sense or antisense oligo and a flanking antisense or sense oligo. A second PCR round using a combination of the first round products and both flanking primers produces the full-length double stranded product. For the 83 positions randomized here, this constituted 83 x 2 first round PCR reactions and 83 second round reactions, for a total of 249 PCR reactions. All reactions yielded a single intense band as visualized on an agarose gel. PCR product concentrations were measured using Picogreen (Invitrogen), pooled in equimolar ratios, purified, digested, and ligated into the bacterial 2-hybrid λ-cI-fusion expression vector. Each ligation was purified, eluted into 7ul dH2O, and transformed into MC4100-Z1 containing the eGFP reporter plasmid, pZE1RM-eGFP, and the RNAα-CRIPT peptide fusion expression vector. Each transformation yielded $>10^6$ variants.

1. Chi, C.N., et al., *Reassessing a sparse energetic network within a single protein domain.* Proceedings of the National Academy of Sciences, 2008. **105**(12): p. 4679-4684.
2. Lockless, S.W. and R. Ranganathan, *Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families.* Science, 1999. **286**(5438): p. 295-299.
3. Bloom, J.D., et al., *Protein stability promotes evolvability.* PNAS, 2006. **103**(15): p. 5869-5874.
4. Tokuriki, N., et al., *The Stability Effects of Protein Mutations Appear to be Universally Distributed.* Journal of Molecular Biology, 2007. **369**(5): p. 1318-1332.
5. Bershtein, S., et al., *Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein.* Nature, 2006. **444**(7121): p. 929-932.
6. Stiffler, M.A., et al., *PDZ Domain Binding Selectivity Is Optimized Across the Mouse Proteome.* Science, 2007. **317**(5836): p. 364-369.
7. Halabi, N., et al., *Protein Sectors: Evolutionary Units of Three-Dimensional Structure.* Cell, 2009. **138**(4): p. 774-786.
8. Hatley, M.E., et al., *Allosteric determinants in guanine nucleotide-binding proteins.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(24): p. 14445-14450.
9. Russ, W.P., et al., *Natural-like function in artificial WW domains.* Nature, 2005. **437**(7058): p. 579-583.
10. Shulman, A.I., et al., *Structural Determinants of Allosteric Ligand Activation in RXR Heterodimers.* Cell, 2004. **116**(3): p. 417-429.
11. Smock, R.R., O; Russ, WP; Swain, JF; Leibler, S; Ranganathan, R; Gierasch, LM, *An Interdomain Sector Mediating Allostery in Hsp70 Molecular Chaperones.* Molecular Systems Biology, 2010. **In Press**.
12. Socolich, M., et al., *Evolutionary information for specifying a protein fold.* Nature, 2005. **437**(7058): p. 512-518.
13. Suel, G.M., et al., *Evolutionarily conserved networks of residues mediate allosteric communication in proteins.* Nat Struct Mol Biol, 2003. **10**(1): p. 59-69.
14. Hamming, R., *Error detecting and error correcting codes.* Bell System Tech. J., 1950. **29**: p. 147-160.

15. Creighton, T., *Proteins: Structures and Molecular Properties*. 2nd ed. ed. 1992: W.H. Freeman and Company.
16. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proceedings of the National Academy of Sciences of the United States of America, 1992. **89**(22): p. 10915-10919.
17. Doyle, D.A., et al., *Crystal Structures of a Complexed and Peptide-Free Membrane Protein Binding Domain: Molecular Basis of Peptide Recognition by PDZ.* 1996. **85**(7): p. 1067-1076.
18. Tonikian, R., et al., *A Specificity Map for the PDZ Domain Family.* PLoS Biol, 2008. **6**(9): p. e239.
19. Skelton, N.J., et al., *Origins of PDZ Domain Ligand Specificity.* Journal of Biological Chemistry, 2003. **278**(9): p. 7645-7654.
20. Laura, R.P., et al., *The Erbin PDZ Domain Binds with High Affinity and Specificity to the Carboxyl Termini of δ-Catenin and ARVCF.* Journal of Biological Chemistry, 2002. **277**(15): p. 12906-12914.
21. Koonin, E.G., MY, *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. 2003: Kluwer Academic Publishers.
22. Boffelli, D., M.A. Nobrega, and E.M. Rubin, *Comparative genomics at the vertebrate extremes.* Nat Rev Genet, 2004. **5**(6): p. 456-465.
23. Sharma, R., *Logic and Mechanism of Evolutionarily Conserved Interaction in PDZ Domains*. 2004, University of Texas Southwestern Medical Center.
24. Huang, W., et al., *Amino Acid Sequence Determinants of β-Lactamase Structure and Activity.* Journal of Molecular Biology, 1996. **258**(4): p. 688-703.
25. Shortle, D., *Probing the determinants of protein folding and stability with amino acid substitutions.* J. Biol. Chem., 1989. **264**(10): p. 5315-5318.
26. Shortle, D., W.E. Stites, and A.K. Meeker, *Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease.* Biochemistry, 1990. **29**(35): p. 8033-8041.
27. Chen, J. and W.E. Stites, *Energetics of Side Chain Packing in Staphylococcal Nuclease Assessed by Systematic Double Mutant Cycles†.* Biochemistry, 2001. **40**(46): p. 14004-14011.
28. Axe, D.D., N.W. Foster, and A.R. Fersht, *A Search for Single Substitutions That Eliminate Enzymatic Function in a Bacterial Ribonuclease.* Biochemistry, 1998. **37**(20): p. 7157-7166.
29. Serrano, L., A.G. Day, and A.R. Fersht, *Step-wise Mutation of Barnase to Binase : A Procedure for Engineering Increased Stability of Proteins and an Experimental Analysis of the Evolution of Protein Stability.* Journal of Molecular Biology, 1993. **233**(2): p. 305-312.
30. Serrano, L., et al., *The folding of an enzyme : II. Substructure of barnase and the contribution of different interactions to protein stability.* Journal of Molecular Biology, 1992. **224**(3): p. 783-804.
31. Guo, H.H., J. Choe, and L.A. Loeb, *Protein tolerance to random amino acid change.* PNAS, 2004. **101**(25): p. 9205-9210.
32. Agresti, A., *A Survey of Exact Inference for Contingency Tables.* Statistical Science, 1992. **7**(1): p. 131-153.
33. Fraczkiewicz, R. and W. Braun, *Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules.* Journal of Computational Chemistry, 1998. **19**(3): p. 319-333.
34. Hartman, J.L., IV, B. Garvik, and L. Hartwell, *Principles for the Buffering of Genetic Variation.* Science, 2001. **291**(5506): p. 1001-1004.
35. Carter, P.J., et al., *The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus).* Cell, 1984. **38**(3): p. 835-840.
36. Shortle, D. and A.K. Meeker, *Mutant forms of staphylococcal nuclease with altered patterns of guanidine hydrochloride and urea denaturation.* Proteins: Structure, Function, and Genetics, 1986. **1**(1): p. 81-89.

37. Hidalgo, P. and R. MacKinnon, *Revealing the architecture of a K+ channel pore through mutant cycles with a peptide inhibitor.* Science, 1995. **268**(5208): p. 307-310.
38. LiCata, V.J. and G.K. Ackers, *Long-Range, Small Magnitude Nonadditivity of Mutational Effects in Proteins.* Biochemistry, 1995. **34**(10): p. 3133-3139.
39. Robinson, C.R. and S.G. Sligar, *Electrostatic stabilization in four-helix bundle proteins.* Protein Science, 1993. **2**(5): p. 826-837.
40. Schreiber, G. and A.R. Fersht, *Energetics of protein-protein interactions: Analysis ofthe Barnase-Barstar interface by single mutations and double mutant cycles.* Journal of Molecular Biology, 1995. **248**(2): p. 478-486.
41. Chen, J. and W.E. Stites, *Higher-Order Packing Interactions in Triple and Quadruple Mutants of Staphylococcal Nuclease†.* Biochemistry, 2001. **40**(46): p. 14012-14019.
42. Sadovsky, E. and O. Yifrach, *Principles underlying energetic coupling along an allosteric communication trajectory of a voltage-activated K+ channel.* Proceedings of the National Academy of Sciences, 2007. **104**(50): p. 19813-19818.
43. Peterson, F.C., et al., *Cdc42 Regulates the Par-6 PDZ Domain through an Allosteric CRIB-PDZ Transition.* Molecular Cell, 2004. **13**(5): p. 665-676.
44. Garrard, S., et al., *Structure of Cdc42 in a complex with the GTPase-binding domain of the cell polarity protein, Par6.* EMBO J., 2003. **22**(5): p. 1125-33.
45. Fuentes, E.J., C.J. Der, and A.L. Lee, *Ligand-dependent Dynamics and Intramolecular Signaling in a PDZ Domain.* Journal of Molecular Biology, 2004. **335**(4): p. 1105-1115.
46. Feng, W., et al., *PDZ7 of Glutamate Receptor Interacting Protein Binds to Its Target via a Novel Hydrophobic Surface Area.* Journal of Biological Chemistry, 2002. **277**(43): p. 41140-41146.

Conclusions of this thesis

Biological systems display design properties unique from those of engineered systems. Primarily, these features arise as a result of and in response to the long-timescale process of evolution. For example, an engineered system is often designed for a discrete function and a discrete lifetime; biological systems achieve propagation by the ability to buffer perturbations and maintain stability while simultaneously retaining the ability to adapt or responding to changes in environmental conditions. Life exhibits a wide range of mechanisms to ensure stability and adaptability of different timescales. Short duration, unstable environmental conditions like food availability can be adapted to by movement or changes in metabolic state orchestrated by epigenetic processes like transcription or protein localization. Infrequent, large magnitude environmental changes may be beyond the scope of epigenetic adaptation and may require genetic variation to achieve adaptation. On the molecular scale, this genetic variation occurs constantly as random mutation. In order to maintain the stability necessary to support organismal fitness, most of these mutations have no phenotypic effect. This stability is evident in the diversity of protein sequences with conserved function. However, the fact that some mutations have very drastic functional effects on a protein and the general diversity of life demonstrate that proteins are simultaneously capable of buffering perturbations and eliciting functional change. The mechanisms by which these properties exist in proteins is not well understood. A number of groups have demonstrated the pattern of general robustness and limited susceptibility in a variety of protein model systems. However, each of these studies presents significant limitations to a complete description of the patterns of functional effects in proteins, let alone a model for the patterns. Most notably, these studies tend to make general conclusions with a limited subset of mutations, and the techniques used to measure the functional

effects of mutations rarely take into consideration the selective constraints of a protein's function in the context of an organism – the only correct context to make statements about the effect of a mutation. These are difficult problems since most techniques for measuring function are low-throughput, and the techniques that are available even to measure small samples focus on an isolated set of functional constraints, usually only stability or binding.

Our interest in these problems came from a basic desire to understand the information encoded by the amino acid sequence of a protein. Previous work in the lab described the existence of a limited set of highly-covarying amino acids in a number of protein families. These groups of coevolving positions, termed sector, form a contiguous network within the protein that often spans functional surfaces. A number of mutagenesis studies have shown the importance of these positions, and most importantly, novel sequences designed using the coupling information display fold and function similar to that of natural proteins, while sequences designed with the conservation information alone do not fold or function. The general importance of sectors for protein function was clear, but the sparseness of these residues suggested an additional hypothesis about the selective pressures that might evolve such an architecture. By evolving a small subset of positions that control function, a mutation might produce a significant functional effect only if it occurred in a sector positions. Most mutations would occur in non-sector, independently-evolving positions and exhibit small effect. At the same time, sector may play a role in evolving new functions by being loaded for functional change. We decided to comprehensively test the distribution of function across positions in a protein and how this distribution related to the patterns of amino acid coevolution.

In order to address these questions, we needed an assay system that permitted the measurement of function for thousands of variants of a single protein. Crucially, this assay

system must recapitulate as closely as possible the *in-vivo* functional constraints on a protein, including binding affinity, stability, expression level, tendency to aggregate, amongst others. We chose the PDZ domain PSD95-PDZ3 as our model system due to the extensive structural, functional, and statistical data available on this protein from our lab and others. To test the function of PDZ3 in a cellular context, we built a quantitative bacterial 2-hybrid assay which expressed eGFP as a function of a PDZ3 variant binding the CRIPT peptide. We characterized the free energy of binding and stability for a library of single subtle mutations at all 83 positions in PDZ3 to show that the bacterial 2-hybrid assay produced eGFP at a concentration that correlated well with the in-vitro biophysics of PDZ3 variants. Limited by the throughput of individual assays, we developed a high-throughput assay in which fluorescent activated cell sorting was used as a selection for high eGFP intensity, and the relative enrichment of a given PDZ3 variant was quantified using next-generation Solexa sequencing. We found that the enrichment of a given PDZ3 variant provided a sensitive measure of its cellular function as judged by the correlation of the enrichment of a protein and its biophysical parameters.

The development of this technology allowed us to eliminate the caveats of previous mutagenesis studies limited by a few mutations at a few positions, measured by some *in-vitro* technique. With the Solexa-quantified bacterial 2-hybrid assay we could address in an unbiased fashion the functional effect of every mutation at every position in PDZ3. The distribution of effects shows that most mutations have subtle effects on function in PDZ3, demonstrating the significant robustness of this domain, but a small percentage of mutations produces large decreased in function. We find that this group of positions of large functional effect is enriched for sector positions. We show that statistically, the degree to which a position is coupled is a better predictor of its functional effect than either conservation or the solvent exposure of that

position. This demonstrates from an unbiased, comprehensive mutagenesis that sector positions are the most functionally important positions in a protein.

Even though we observe the strong statistical association of sectors positions with functionally important positions, we still observe non-sector positions in the set of functionally important positions. We expected that these sets of positions could be differentiated by measuring the non-additivity of each positions interaction with other positions in higher-order couplings. We expected that sector positions should display strong couplings, while the positions of strong but intrinsic effects should not. To measure coupling, we performed a thermodynamic mutant cycle in which we measured the functional effect of every single-amino acid mutation in the wild-type PDZ3 background and in the background of a threonine to phenylalanine mutation at the -2 position of the peptide. We observe a two-tailed distribution of effects for mutations in the mutant peptide background in which most mutations display small effects, but a portion of mutants either strongly decrease or increase function.

When we analyze the non-additivity of every mutation at every position with the T to F mutation at the -2 position of the peptide, we observe a distribution of effects with a single dominant rightward tail of highly non-additive positions. These positions of non-additivity occur both near the site of mutation and distant from the site of mutation. Interestingly, two of the positions of non-additivity have been implicated as important in allostery in the PDZ family. Strikingly, we find that every position that falls in the tail of the additivity distribution is also a sector position. This demonstrates the role of sector positions as functionally important and the main constituents of higher-order functional interactions in the PDZ domain. Further, we observe two patterns of non-additivity in the sector positions. One position, 322, produces minor functional effects upon mutation in the CRIPT background, but produces significant increases in

function in the mutant peptide background. This pattern could provide an advantage in a variable environment since it maintains the endogenous function while relaxing specificity and enabling potentially advantageous promiscuous interactions. Another set of four positions in the additivity tail displays a pattern in which mutations decrease function substantially in the CRIPT background, but increase function in the context of the mutant peptide; this pattern effectively couples the fragile positions of the protein to evolvability of the protein.

In the end, enabled by the significant technical advancements of the Solexa-quantified bacterial 2-hybrid assay these data produce a comprehensive picture of the functional effects of mutations to individual amino acids and the functional effect of pairwise mutations. The distributions of functional effects show significant robustness in PDZ3 with concomitant sensitivity at a set of positions. These residues of functional effect are a small percentage of all positions, and an even smaller number of positions engage in functionally important higher-order interactions. This limited coupling within the protein extends the degree of robustness, since most mutations of functional effect exhibit additive effects. Lastly, we observe an elegant feature of the highly non-additive sector positions in which positions enable new functions without sacrificing the current function, or directly couple a large functional decrease to a significant enhancement of a novel, potentially fitness-enhancing function.

Future Directions

The major motivation of this project has been the hope that by studying the products of evolution with sufficient detail at the appropriate levels, we can come to understand mechanistically the selective pressures that created the design of evolved systems and the logic for the way in which evolved systems function. The patterns of SCA represent a compelling suggestion that this research approach holds merit. This thesis work has sought to take the system-level observation of sparse coupled networks of functional residues and examine their role in depth in a single model system. Ultimately, the work provides the first comprehensive description of the function of amino acids individually and in pairs, and demonstrates the crucial role of sectors in protein function as positions loaded for functional change and engaging in pairwise functional interactions. Here I describe several areas of investigation made possible by the work described in this thesis or complementary to the conclusions of this thesis.

The importance of higher order interactions in the protein

While this study demonstrates the participation of nine sector positions in non-additive pairwise functional interactions, the scale of functionally important coupling may extend to much higher order than pairwise. Targeted studies of fourth and fifth way couplings in *Shaker* by the Yifrach group suggest that these higher order couplings are the most energetically important couplings in the channel. The methods described in this thesis make these sorts of analyses possible in a range of protein families, certainly first in PDZ domains. While we do observe a highly significant representation of sector positions in the group of positions with the greatest effect on function, there are still positions that either show small individual effects or

small pairwise effects. It could be that cooperativity is a common property of all sector positions, but these couplings only manifest with higher than pairwise measurements.

The most feasible approach would be the targeted approach to measuring higher order couplings in which a number of positions close in the primary sequence and likely to display coupling plus a few others are mutated in all possible combinations and measured as a library with the Solexa-quantitated bacterial 2-hybrid assay. However, the recent release of Solexa reagents capable of spanning the entire PDZ sequence with 150 basepair paired ends makes an unbiased approach also possible. In such an approach, error-prone PCR would be used to randomly introduce mutations in the PDZ domain. Though, even with the ability to sequence complete PDZ sequences, the complexity of such an n-way library would quickly exceed the capabilities of the functional assay and sequencing methods.

The comparison of the evolutionary timescale properties of natural and designed domains

The strongest test of the sufficiency of the sector hypothesis has been the design of proteins that fold and function like natural proteins using only the coupling and conservation patterns. An important extension of this sufficiency test would be to demonstrate that SCA-designed proteins also demonstrate evolutionary-timescale properties like robustness and evolvability comparable to natural proteins. The technology developed in this work presents a powerful system for being able to measure and compare these properties in SCA-designed and natural proteins. For example, one could make randomly mutated libraries of SCA and natural PDZ domains, each library containing an increasing number of mutations, and measure the loss of overlap of each library's eGFP distribution with that of WT. This could serve as a measure of

the collective loss of function in the library. In addition, with 300 basepair Solexa sequencing it becomes possible to quantify the function of each sequence within the library to probe whether the probability of observing a functional decrease in a protein is related to the location of mutations in the primary sequence, specifically the mutations to sector positions. Another approach to this comparison could be to perform an analysis similar to the one described in this thesis on SCA-designed domains. A recapitulation of the robustness and importance of sector positions in first and second order functional effects in these designed domains would demonstrate their similarity to natural proteins in their functional response to mutation.

Discerning specificity changes from increased promiscuity

One hypothesis in the literature, largely from the Tawfik group, suggests that functional transitions in proteins proceed through a promiscuous intermediate. In addition, they suggest that this promiscuity exists latently in proteins and provides the raw material for changes in function since if that promiscuous functions becomes beneficial, the protein provides an advantage to the organism. Additional mutations that optimize that promiscuous function could then be selected for to create a protein specific for the new activity. The data described in this thesis provide suggest that there are a number of ways to change the specificity of PDZ3; whether these mutations actually make the domain less specific for CRIPT and more specific for T(P_{-2})F remains to be seen. The patterns of effects in the non-additive positions suggest that both mechanisms may occur. Position 322 shows no change in CRIPT function, but a significant increase in T(P_{-2})F function. This is an increase in promiscuity without a sacrifice of the endogenous function. Other positions show the pattern of strongly decreasing CRIPT function and increasing T(P_{-2})F function. In the context of only these two peptides, these mutations have

changed the specificity. However, these mutations could also be increasing function in the context of a number of peptides, just not CRIPT. In order to tease out the precise specificity changes, one could purify a select group of mutants from the libraries that show each pattern of interest and perform a peptide display assay similar to that of the MacBeath group. Further, based upon preliminary work I performed, an assay is under development in the lab to use the quantitative bacterial 2-hybrid assay to measure the specificity profiles of individual PDZ domains. This approach could be used in this sort of an experiment to measure the specificity profile of a number of variants that show functional changes for CRIPT and T(P_{-2})F. One weakness of these two methods is their low throughput. Specifically for this question, one could design a bacterial 2-hybrid assay that incorporated a selection for specificity in the context of multiple peptides. This could be achieved by expressing the T(P_{-2})F peptide in cells with eGFP and selection of other peptides in cells expressing another fluorescent protein or a toxic gene product. This would allow the selection of a population of variants that has high function in the context of T(P_{-2})F but low function in the context of other peptides.

The use of molecular dynamics-based algorithms for the prediction of mutational effects

The use of algorithms in the literature to predict the functional effect of mutations has become unfortunately widespread. One algorithm, FoldX, is used with frequency to make statements about the distribution of mutational effects in different proteins. The data from my graduate work presents an unprecedented data set for analyzing the efficiency of this algorithm to predict the functional change of thousands of mutants, or even the stability change of dozens of mutants. Preliminary analysis of the correlation between the energetic effect predicted by

FoldX and either the enrichment value or change in stability of the NMCAA mutants shows a poor prediction of either parameter by the algorithm. This general trend has been reported previously by the Schreiber group as well, yet the protein evolution field continues to use this algorithm extensively.

The importance of stability in the function of proteins

Another prominent idea in the protein evolution field states that stability effects are the most prominent effects due to the fact that most mutations have an effect on stability. We demonstrate that in our pool of NMCAA mutants, most mutations have small effects on stability, while a few mutations have large stability effects. Further, the functional effect of these stability changes when measured in a cellular context is minimal. This data suggest that for single mutations, most mutations have a small destabilizing effect, but these stability effects do not translate to cellular function effects. One possibility is that if we measured the global distribution of higher order mutations, we may see significantly less robustness in the domain and it is possible that this loss of function would result from the additive effect of stability changes. A paper by Bershtein and Tawfik measured the portion of a randomly mutated population of beta-lactamases that retain function at increasing levels of mutations. They found the protein to be robust if it contained up to approximately 4 mutations, after which the fitness decreased exponentially. They term this 'threshold epistasis.' The conclusions of this paper are difficult to rationalize due to their extensive use of FoldX (see previous section) to measure the fitness of mutants. However, one can imagine this threshold epistasis as simply the manifestation of the probability of mutating a functionally important (perhaps coupled) residue. Further, the non-additive functional effects of the subset of sector positions described in the main

text may contribute to the steep slope of the fitness versus number of mutations curve described in the Tawfik paper. From the pairwise experiment above, we may be able to write an equation that approximates Tawfik's curve but is based upon the distribution of the magnitude of single and double mutant effects in the protein. In the end, it is important to use the appropriate methods to be able to make statements about fitness effects in proteins, a motivating factor of this thesis work.

One additional consideration is the fact that stability and function are very different parameters by their nature. Evolution selects for function, and evolution selects for stability to the extent that it contributes to the function of a protein. From SCA and this work, we observe a subset of amino acids that engage in higher order interactions that constitute the core functional positions in the protein. If we look at the distribution of stability in the protein, we see that the core residues tend to have larger stability effects than the surface residues, but there the distribution of stability effects throughout the protein is almost homogeneous. This is due to the fact that stability of a protein is determined by a fine balance of forces across the entire protein – a collective effect. Function is distributed in a heterogeneous fashion and appears to result from the interaction of a subset of positions – an exclusive yet cooperative effect. One will never find a global suppressor of functional mutation effects due to the nature of the distribution of function across the protein. As Bloom and Arnold show, such global suppressors of stability effects do exist due to the nature of stability as the net sum of subtle forces acting across the entire protein. One possible reason stability effects are less important in protein evolution is that while destabilizing effects occur frequently, stabilizing effects may also occur frequently. Due to the nature of stability a destabilizing mutation can be rescued by a large number of different stabilizing mutations. There is a much lower probability that a rescuing second mutation exists

for a mutation of functional effect due to the heterogeneous, cooperative nature of the network of amino acids controlling function in the protein. One experimental approach to demonstrate this contrast would be to perform a comprehensive mutagenesis in the background of a mutation with a large stability effect and observe the pattern of positively epistatic mutations in this background relative to the background of a functionally deleterious mutation.

The design of a true fitness system

Fitness is the number of offspring produced by an organism. We are particularly interested in understanding how changes in the sequence of a protein equate to changes in organismal fitness. In this project we attempted to extend the scale of the measured functional constraints on protein function beyond just *in vitro* binding and stability as is so common in previous studies. Yet, the system we created is still arbitrary in that we include many factors such as the biophysics of eGFP that are not part of the fitness of a PDZ domain, and we exclude many factors that are experimentally difficult to encapsulate such as negative selection. Ultimately, we don't know every functional constraint on the PDZ domain that contributes to organismal fitness, so we are forced to do our best to recapitulate the endogenous constraints on the protein while minimizing simplifying assumptions and preserving our ability to perform the experiments in a way that we learn something.

The only inclusive way to measure the relationship between the sequence of a protein (genotype) and the fitness of the encoded phenotype is to measure this relationship in the natural environment of some organism. Unfortunately, most organisms are not amenable in generation time or the availability of genetic tools to this sort of an analysis. Organisms such as bacteria and yeast may be amenable to such an analysis; however, viruses probably represent the best opportunity to follow sequence changes and fitness of individual variants in a environment

similar to their natural environment. For example, the polio virus enters its host through the oral or intestinal epithelium, moves into the blood stream, and eventually gains access to the central nervous system. This progression requires the virus to encounter a highly variably environment of host defenses, different tissue types, and the passage between tissues. It is likely that distinct variants of the virus are required for each step of infection, so sequence variability is particularly important for the fitness of these viruses. The mouse provides a highly used model system and presents the opportunity to study the sequence changes in the quasispecies that underlie each step in the progression of the infection. Using high throughput sequencing, one could survey this time- and environment-dependent progression of the sequence and study the correlation of sequence changes to patterns of coevolution in the virus. This particular system would not consider factors such as transmission of the virus, but such a system would come closest of any laboratory system to comprehensively recapitulating the natural fitness constraints of a protein.

Appendix I: MATLAB code for processing paired-end Solexa sequencing data

```
function [data] = PDZ3_Subgroups_PE_Allele_Count(fastq, keep_sequences_flag,
H372Y_flag)

%This m-file takes the quality-trimmed reads from CLC Genomics Workbench,
%sorts the reads into subgroups, joins the paired reads into a single read,
%and counts the codons at each position in each subgroup read.

%data consists of a 1x3 cell in which cell(1,1) corresponds to the data for
%subgroup 1, (1,2) subgroup 2, and (1,3) subgroup 3.

%Each subgroup cell contains the single reads, paired reads, mismatches,
%number of WT-sequences, and a codon matrix for that subgroup.

%The single reads and paired reads are large files, so use
%keep_sequences_flag ==1 if you want to output this data.

%The H76Y_flag should be set to one if the dataset uses the H372Y
%background.

%This code is written for parallelization, so engage the matlabpool before
%running this code.



%%
%Read in the filtered sense and antisense matched pairs from CLC-exported
%fastaq files.

%fastq is a string specifying the fastaq file containing the paired reads
%for a given barcode.  This is all three subgroups together.

%Files should be of the .fastq format:
%@HWI-EAS392_0001:5:50:6:1362#0/1
%TTCAGCATTGTGGGCGGCGAGGAT
%+HWI-EAS392_0001:5:50:6:1362#0/1
%ABBB?BBCBCB??9B?;@@B@1>B

tic

fid = fopen(fastq);
reads_1 = textscan(fid, '%*s %s %*s %*s %*s %*s %*s %*s', 'delimiter', '\n');
fclose(fid);
clear fid

fid = fopen(fastq);
reads_2 = textscan(fid, '%*s %*s %*s %*s %*s %s %*s %*s', 'delimiter', '\n');
fclose(fid);
clear fid
```

```
display('reads')
display(toc)

%%

%Sort paired reads into subgroups.

seqs = char(reads_2{1});
seqs_as = char(reads_1{1});


match1 = 'AGGAAGACATTCCCCGGGAA';
match1as = 'CCCCCAGCAAGGATGAAGGA';

match2 = 'ATGGTGAAGGCATCTTCATC';
match2as = 'TGACTGGCATTGCGGAGGTC';

match3 = 'TCCTGTCGGTCAATGGTGTT';
match3as = 'AATCGACTATACTCTTCTGG';

[sub1, sub1as] = subgroups(seqs, seqs_as, match1, match1as);
[sub2, sub2as] = subgroups(seqs, seqs_as, match2, match2as);
[sub3, sub3as] = subgroups(seqs, seqs_as, match3, match3as);

display('subgroups')
display(toc)

%%

%Combine matched pairs into a single read of the specified length for each
%subgroup.

wt_sub1 =
'AGGAAGACATTCCCCGGGAACCAAGGCGGATCGTGATCCATCGGGGCTCCACCGGCCTGGGCTTCAACATTGTGGG
CGGCGAGGATGGTGAAGGCATCTTCATCTCCTTCATCCTTGCTGGGGG';
wt_sub2 =
'ATGGTGAAGGCATCTTCATCTCCTTCATCCTTGCTGGGGGTCCAGCCGACCTCAGTGGGGAGCTACGGAAGGGGGA
CCAGATCCTGTCGGTCAATGGTGTTGACCTCCGCAATGCCAGTCA';
wt_sub3 =
'TCCTGTCGGTCAATGGTGTTGACCTCCGCAATGCCAGTCACGAACAGGCTGCCATTGCCCTGAAGAATGCGGGTCA
GACGGTCACGATCATCGCTCAGTATAAACCAGAAGAGTATAGTCGATT';

if H372Y_flag == 1
    wt_sub3 =
'TCCTGTCGGTCAATGGTGTTGACCTCCGCAATGCCAGTTACGAACAGGCTGCCATTGCCCTGAAGAATGCGGGTCA
GACGGTCACGATCATCGCTCAGTATAAACCAGAAGAGTATAGTCGATT';
end

[pair_sub1, sub1_mismatch] = jseqs(sub1, sub1as, length(wt_sub1));
[pair_sub2, sub2_mismatch] = jseqs(sub2, sub2as, length(wt_sub2));
[pair_sub3, sub3_mismatch] = jseqs(sub3, sub3as, length(wt_sub3));

display('pairs')
```

```
display(toc)

%%

%Filter joined reads for those sequences which contain only a single
%mutation and count codons for these single mutant sequences.


[sub1_codons, sub1_num_wt] = ccount_singles_PE(wt_sub1, pair_sub1, 3);
[sub2_codons, sub2_num_wt] = ccount_singles_PE(wt_sub2, pair_sub2, 3);
[sub3_codons, sub3_num_wt] = ccount_singles_PE(wt_sub3, pair_sub3, 3);

display('singles')
display(toc)

%%
%Combine outputs into data array.
%Comment in/out desired outputs

if keep_sequences_flag ==1

    data.sub1 = sub1;
    data.sub1as = sub1as;
    data.sub2 = sub2;
    data.sub2as = sub2as;
    data.sub3 = sub3;
    data.sub3as = sub3as;

    data.pair_sub1 = pair_sub1;
    data.sub1_mismatch = sub1_mismatch;
    data.pair_sub2 = pair_sub2;
    data.sub2_mismatch = sub2_mismatch;
    data.pair_sub3 = pair_sub3;
    data.sub3_mismatch = sub3_mismatch;
end

data.sub1_codons = sub1_codons;
data.sub2_codons = sub2_codons;
data.sub3_codons = sub3_codons;

data.sub1_num_wt = sub1_num_wt;
data.sub2_num_wt = sub2_num_wt;
data.sub3_num_wt = sub3_num_wt;

display(toc)

    end


%%

function [sub, subas] = subgroups(seqs, seqs_as, match, matchas)
    sub = zeros(size(seqs,1), size(seqs,2));
    subas = sub;
```

```
    parfor j = 1:size(seqs,1)

        if seqs(j,1:20) == match
            if seqs_as(j,1:20) == matchas

                sub(j,:) = seqs(j,:);
                subas(j,:) = seqs_as(j,:);
            end
        end
    end

    del = sub(:,1)==0;

    sub(del,:) = [];
    subas(del,:) = [];

    sub = char(sub);
    subas = char(subas);
end

%%

function [pairs, mismatch] = jseqs(sub, subas, wt_length)


pairs = zeros(size(sub,1), wt_length);
mismatch = zeros(size(sub,1),1);

parfor s = 1:size(sub,1)

    sense = sub(s,isletter(sub(s,:)));
    antisense = seqrcomplement(subas(s,isletter(subas(s,:))));


    if length(sense)+length(antisense)>= wt_length
        joined = joinseq(sense, antisense);
        if length(joined) == wt_length
            pairs(s,:) = joined;
        else
            mismatch(s,1) = s;
        end

    else
        sense((length(sense)+1):(length(sense)+(wt_length-length(sense)-
length(antisense)))) = repmat('-',1,(wt_length-length(sense)-
length(antisense)));
        joined2 = joinseq(sense, antisense);
        if length(joined2) == wt_length
            pairs(s,:) = joined2;
        end
    end
end
```

```
del = pairs(:,1)==0;
pairs(del,:) = [];

delm = mismatch(:,1)==0;
mismatch(delm,:) = [];

pairs = char(pairs);
mismatch = char(mismatch);

end



%%

function [codon_mat, num_wt] = ccount_singles_PE(wt, seqs, frame)

trans = nt2aa(wt, 'FRAME', frame);
wt_codons = zeros(length(trans),5,5,5);
wti = nt2int(wt,'unknown',5,'acgtonly',true);

for count = frame:3:length(wti)-2
    wt_codons((count+(3-frame))/3, wti(count),wti(count+1),wti(count+2)) =
wt_codons((count+(3-frame))/3, wti(count),wti(count+1),wti(count+2)) + 1;
end

wt_sum = nnz(wt_codons);
num_wt=0;

codons = zeros(length(trans),5,5,5);


parfor i = 1:size(seqs,1);
    dnai = nt2int(seqs(i,:),'unknown',5,'acgtonly',true);
    seq_codons = zeros(length(trans),5,5,5);


for count = frame:3:length(dnai)-2
    seq_codons((count+(3-frame))/3, dnai(count),dnai(count+1),dnai(count+2))
= seq_codons((count+(3-frame))/3, dnai(count),dnai(count+1),dnai(count+2)) +
1;
end

if nnz(wt_codons.*seq_codons) == (wt_sum-1)
    codons = codons + seq_codons;
elseif nnz(wt_codons.*seq_codons) == wt_sum
    num_wt = num_wt+1;
end

%if rem(i, 5000) == 0;
%disp(i);
%end
```

```
end

parfor t = 1:length(codons)
labelcount = 0;
validcodons = 0;
labels = cell(1,125);
for first = 1:5
for second = 1:5
for third = 1:5
codon = int2nt([first second third]);
labelcount = labelcount + 1;
labels{labelcount} = codon;
output(t).(codon) = codons(t,first,second,third);
validcodons = validcodons + codons(t,first,second,third);
end
end
end
end
c_mat = struct2cell(output);
c_mat = cell2mat(c_mat);
c_mat = squeeze(c_mat);
%Convert 125 row matrix into 65 row matrix
%Row 65 = any codon with a '-'
codon_mat = zeros(65,length(trans));
index = [1  2   3   4   65  5   6   7   8   65  9   10  11  12  65  13  14
15  16  65  65  65  65  65  65  17  18  19  20  65  21  22  23  24  65  25
26  27  28  65  29  30  31  32  65  65  65  65  65  65  33  34  35  36  65
37  38  39  40  65  41  42  43  44  65  45  46  47  48  65  65  65  65  65
65  49  50  51  52  65  53  54  55  56  65  57  58  59  60  65  61  62  63
64  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65  65
65  65  65  65  65  65  65  65  65  65  65  65  65];
index = transpose(index);
index = [index ones(size(index))];
for i = 1:length(trans)
codon_mat(:,i) = accumarray(index, c_mat(:,i));
end
end
```

Appendix II: Tables of NMCAA, NNS Measured and Calculated Parameters

Table Appendix II.1 Biophysical Properties of NMCAA Mutants and the Conservation and Structural Properties of PSD95-PDZ3

The Kd for CRIPT peptide was measured by fluorescence polarization and the thermodynamic unfolding properties were measured with DSC as described in the text with purified protein. The global conservation term was calculated using an alignment of more than 1500 PDZ sequences as described in the text (alignment provided by Alan Poole, Ranganathan Lab, UTSW). Solvent-exposed surface area was calculated using the 1BE9 structure with the GETAREA program (http://curie.utmb.edu/getarea.html) as described in the text. The number of contacts per position was calculated using the cutoff of the number of positions with the sum of the Van der Waals radii of the two atoms plus 20% in the 1BE9 structure (MATLAB code courtesy of Rama Ranganathan, UTSW).

	WT	P311I	R312K	R313T	I314V	V315E	I316L	H317E
log(K_d) (kT)	-6.0339	-6.0179	-5.7932	-6.1692	-6.2517	-6.143	-6.1388	-6.1436
log(Kd), Standard Error	0.0822	0.1016	0.1025	0.119	0.1225	0.1678	0.1043	0.0909
Bacterial 2-hybrid Normalized Mean eGFP (A.F.U)	1	0.9571	0.9791	1.0209	1.0079	0.9747	0.9819	0.9774
Standard Deviation, 2-hybrid		0.0417	0.0507	0.0678	6.15E-03	0.1174	0.0356	0.0397
T_m1 (°C)	68.82	60.98	57.78	71.69	71.02	64.64	69.63	70.27
ΔH_{cal}1 (kcal/mol)	58010	34250	65360	91090	95260	57610	104700	110100
ΔH_{vH}1 (kcal/mol)	123300	87180	106200	99720	97670	101300	81710	90080
T_m2 (°C)	75.16	76.82	80.34	76.15	76.41	69.6		
ΔH_{cal}2 (kcal/mol)	50550	25780	8585	56660	43680	111400		
ΔH_{vH}2 (kcal/mol)	60820	42210	-97040	58580	60670	38050		
D_{global}		0.32	0.5109	0.2813	1.1578	0.4335	1.1853	0.2807
Random Coil Ratio Solvent-exposed Surface Area		56.6	26.5	67.6	0.1	38.5	1.3	72.2
Number of Contacts in 1BE9		2	4	3	4	5	4	7

	R318K	G319P	S320G	T321G	G322S	L323F	G324S	F325I
log(K_d) (kT)	-6.0859	-5.9427	-6.1171	-6.0684	-6.141	-5.6921	-4.8047	-5.3608
log(Kd), Standard Error	0.113	0.1798	0.2306	0.0762	0.0636	0.0409	0.123	0.0419
Bacterial 2-hybrid Normalized Mean eGFP (A.F.U)	1.0098	0.961	0.9725	1.0077	0.9719	0.9469	0.5727	0.8368
Standard Deviation, 2-hybrid	0.139	0.0238	0.0568	0.0612	0.0924	0.0354	4.15E-03	0.08
T_m1 (°C)	67.86	63.41	65.61	68.55	64.77	68.3	54	65.28
$\Delta H_{cal}1$ (kcal/mol)	57560	41790	44700	17410	48560	43270	44490	55570
$\Delta H_{vH}1$ (kcal/mol)	110300	100200	123800	120300	119100	133700	88890	132000
T_m2 (°C)	74.22	73.26	70.27	72.01	75.71	75.38	83.04	86
$\Delta H_{cal}2$ (kcal/mol)	69190	75810	50590	28230	58160	51810	36260	17470
$\Delta H_{vH}2$ (kcal/mol)	56880	38260	49440	55690	46080	58690	64310	94970
D_{global}	1.0022	0.7284	0.3475	0.324	1.0571	1.4373	2.4316	1.6276
Random Coil Ratio Solvent-exposed Surface Area	6.1	40.7	98.4	75.1	62.4	7.2	15.8	1.1
Number of Contacts in 1BE9	4	8	4	5	10	5	9	6

	N326S	I327L	I328A	G329S	G330Q	E331K	D332G	G333N
log(K_d) (kT)	-6.264	-5.257	-5.7844	-5.4504	-4.8038	-4.7799	-5.5996	-5.9846
log(Kd), Standard Error	0.1985	0.0966	0.1918	0.0528	0.0987	0.0861	0.1436	0.0806
Bacterial 2-hybrid Normalized Mean eGFP (A.F.U)	1.05	0.7996	0.8387	0.7359	0.6403	0.6435	0.8682	1.0749
Standard Deviation, 2-hybrid	0.0451	0.0544	0.0273	0.0584	0.0345	0.0195	0.0975	0.0448
T_m1 (°C)	65.41	65.59	64.83	65.95	60.73	66.29	66.51	67.07
$\Delta H_{cal}1$ (kcal/mol)	55220	52850	49290	49120	48850	55640	68480	64330
$\Delta H_{vH}1$ (kcal/mol)	120100	103800	113500	113200	103100	124500	114400	99770
T_m2 (°C)	79.8	71.44	69.94	73.65	69.9	87.18	81.76	75.03
$\Delta H_{cal}2$ (kcal/mol)	22210	29630	84560	18010	43070	17510	27410	48270
$\Delta H_{vH}2$ (kcal/mol)	41590	72040	41190	70720	48890	64790	-50080	46890
D_{global}	0.8601	1.4679	0.3703	1.2567	0.8597	0.4046	0.7844	0.3349
Random Coil Ratio Solvent-exposed Surface Area	66.8	10.1	27.1	10.8	8.3	89	55.4	51.7
Number of Contacts in 1BE9	10	4	13	3	2	3	4	10

	E334G	G335P	I336V	F337Y	I338V	S339K	F340S	I341V
log(K_d) (kT)	-5.5809	-6.3164	-5.6158	-6.0872	-6.1202	-5.5918	-5.7849	-5.9066
log(Kd), Standard Error	0.1733	0.2555	0.0724	0.26	0.1033	0.1088	0.0839	0.0912
Bacterial 2-hybrid Normalized Mean eGFP (A.F.U)	0.9204	1.0006	0.9283	0.9117	0.9843	0.7043	1.0185	1.0116
Standard Deviation, 2-hybrid	0.0517	0.0163	0.0501	0.0198	0.0353	0.0243	0.0326	0.0445
T_m1 (°C)	65.59	66.95	70.01	59.31	73.05	70.79	73.95	68.84
$\Delta H_{cal}1$ (kcal/mol)	68290	104700	56170	23640	65420	61010	35140	47910
$\Delta H_{vH}1$ (kcal/mol)	115200	64820	97170	110400	98320	93220	86590	98910
T_m2 (°C)	81.52	81.33	74.28		74.22	64.4	61.53	75.21
$\Delta H_{cal}2$ (kcal/mol)	48960	45560	59320		39850	-11860	36770	36450
$\Delta H_{vH}2$ (kcal/mol)	43940	52290	49170		56310	61720	33590	61670
D_{global}		0.976	1.3005	1.1209	1.9474	0.5921	0.4591	1.6383
Random Coil Ratio Solvent-exposed Surface Area	29.6	1.2	0	0.7	0	16.5	58.6	15.6
Number of Contacts in 1BE9	6	13	8	10	8	5	5	3

	L342I	A343P	G344D	G345S	P346A	A347V	D348A	L349R
log(K_d) (kT)	-5.6832	-6.0363	-6.0157	-5.8642	-6.2673	-5.5984	-6.0267	-5.8389
log(Kd), Standard Error	0.1128	0.0871	0.0654	0.0666	0.1075	0.0894	0.0974	0.1286
Bacterial 2-hybrid Normalized Mean eGFP (A.F.U)	0.8541	0.8421	0.7765	1.0125	0.8572	0.7756	1.148	0.7921
Standard Deviation, 2-hybrid	0.0277	0.0244	0.0577	0.0454	0.0535	0.0345	0.0733	0.0274
T_m1 (°C)	69.65	65.97	66.96	72.31	68.82	63.26	67.25	64.41
$\Delta H_{cal}1$ (kcal/mol)	35690	37220	72100	129400	71470	25140	35230	93540
$\Delta H_{vH}1$ (kcal/mol)	120100	123600	82600	73050	91240	137900	110500	82650
T_m2 (°C)	75.64	84.63	73.85	84.37	62.57	67.82	62.91	77.54
$\Delta H_{cal}2$ (kcal/mol)	69620	12640	22800	68200	32500	31860	49830	-11940
$\Delta H_{vH}2$ (kcal/mol)	40720	81050	68550	57890	27860	32510	33800	82080
D_{global}	0.346	0.7954	1.3608	1.4248	1.0075	1.7874	0.7938	0.5382
Random Coil Ratio Solvent-exposed Surface Area	64.1	100	59.1	6.9	22.7	0	41.3	74.2
Number of Contacts in 1BE9	2	2	5	4	12	12	12	7

	S350D	G351N	E352R	L353I	R354Q	K355V	G356N	D357N
log(K_d) (kT)	-6.0429	-5.9108	-5.8941	-5.919	-6.1532	-5.9747	-6.0496	-5.7881
log(Kd), Standard Error	0.106	0.0884	0.124	0.0763	0.1345	0.1019	0.1043	0.105
Bacterial 2-hybrid Normalized Mean eGFP (A.F.U)	1.0042	0.9283	0.7875	0.8231	1.0939	0.7508	0.4122	0.2234
Standard Deviation, 2-hybrid	9.68E-03	0.0747	0.0549	0.0658	0.0208	0.0295	0.0109	0.0204
T_m1 (°C)	64.18	66.18	65.8	64.14	64.79	72.76	50.13	45.45
$\Delta H_{cal}1$ (kcal/mol)	119700	19670	25530	28880	46520	20580	42550	17400
$\Delta H_{vH}1$ (kcal/mol)	116400	137300	126900	113700	106900	157500	84200	99750
T_m2 (°C)	77.06	67.48	77.96	69.08	69.89	74.55	70.18	74.43
$\Delta H_{cal}2$ (kcal/mol)	61600	64830	10600	33520	16270	47530	33100	9078
$\Delta H_{vH}2$ (kcal/mol)	48010	55060	44090	56320	87390	76810	32990	-46020
D_{global}	0.8289	1.7023		1.6649	0.6907	0.6329	2.1313	2.8573
Random Coil Ratio Solvent-exposed Surface Area	5.4	78	53.4	0	50.8	33.3	0	0.2
Number of Contacts in 1BE9	6	9	5	3	4	4	9	9

	Q358R	I359L	L360V	S361A	V362I	N363D	G364D	V365T
log(K_d) (kT)	-5.8425	-5.8884	-6.1077	-6.109	-5.4391	-5.8993	-6.1024	-6.109
log(Kd), Standard Error	0.0947	0.089	0.1074	0.1341	0.1463	0.1271	0.0886	0.0695
Bacterial 2-hybrid Normalized Mean eGFP (A.F.U)	0.6945	0.8833	0.922	1.0541	0.7604	0.7773	1.1923	1.1002
Standard Deviation, 2-hybrid	0.0237	0.0229	0.0522	0.1162	0.0174	0.0331	0.0144	0.0764
T_m1 (°C)	65.8	67.87	62.84	66.34	66.27	68.57	76.35	65.29
$\Delta H_{cal}1$ (kcal/mol)	37320	29340	29920	45290	37530	67980	37210	41450
$\Delta H_{vH}1$ (kcal/mol)	121400	118300	116500	121300	126800	81540	105100	107400
T_m2 (°C)	74.63	70.32	67.46	72.96	71.63	64.42	77.94	71.84
$\Delta H_{cal}2$ (kcal/mol)	17340	73590	79290	95150	47730	24280	27150	43840
$\Delta H_{vH}2$ (kcal/mol)	56250	46670	40470	40150	45730	36460	61300	44770
D_{global}	0.7363	1.869	1.2754	0.7531	1.9002	2.3195	1.4809	0.6226
Random Coil Ratio Solvent-exposed Surface Area	12.3	0	17.2	16.3	1.7	31.3	84	68.1
Number of Contacts in 1BE9	7	5	7	2	5	11	7	9

	D366S	L367V	R368E	N369G	A370L	S371T	H372L	E373D
log(K_d) (kT)	-6.1259	-5.8145	-5.8441	-5.7317	-6.2025	-6	-4.724	-6.0075
log(Kd), Standard Error	0.0561	0.1485	0.1479	0.1047	0.2836	0.2109	0.0466	0.2438
Bacterial 2-hybrid Normalized Mean eGFP (A.F.U)	1.0901	1.0151	1.0712	0.9916	0.948	0.8937	0.5622	0.9269
Standard Deviation, 2-hybrid	0.0483	0.0785	0.0973	0.016	0.0838	0.0181	0.0276	0.0272
T_m1 (°C)	62.92	62.37	62.42	67.32	60.95	68.36	67.95	66.36
$\Delta H_{cal}1$ (kcal/mol)	44350	45670	62620	41940	26020	18190	47620	41870
$\Delta H_{vH}1$ (kcal/mol)	103200	107800	98300	63620	98880	140100	92400	97460
T_m2 (°C)	74.97	73.23	71.26	66.08	66.05	70.76	71.5	74.72
$\Delta H_{cal}2$ (kcal/mol)	19590	23780	21810	8271	14820	37200	28170	8183
$\Delta H_{vH}2$ (kcal/mol)	68070	45640	-62700	154100	42370	70350	78320	98160
D_{global}	0.7237	1.1287	0.5178	0.929	0.5882	1.0276	1.6743	0.4696
Random Coil Ratio Solvent-exposed Surface Area	53.1	3.3	59.6	77.9	7.6	71	48.6	84
Number of Contacts in 1BE9	5	10	9	8	10	9	9	6

	Q374E	A375V	A376V	I377E	A378L	L379I	K380R	N381A
log(K_d) (kT)	-6.1266	-5.5092	-6.2942	-6.104	-6.1838	-6.1556	-6.121	-5.9889
log(Kd), Standard Error	0.2349	0.1054	0.0896	0.1137	0.1498	0.1344	0.0924	0.0801
Bacterial 2-hybrid Normalized Mean eGFP (A.F.U)	1.1573	0.9721	1.1119	1.0266	1.0664	1.0816	0.9276	0.7566
Standard Deviation, 2-hybrid	0.0527	0.0398	0.0136	0.058	0.0269	0.0102	0.0301	0.0439
T_m1 (°C)	69.65	57.22	69.02	71.61	65.32	65.64	67.2	63.28
$\Delta H_{cal}1$ (kcal/mol)	71630	53770	70070	66960	16820	34160	5672	33480
$\Delta H_{vH}1$ (kcal/mol)	74680	115200	126300	89630	127400	120500	193000	115200
T_m2 (°C)	68.86	87.8	80.98	70.47	69.96	77.05	68.3	77.68
$\Delta H_{cal}2$ (kcal/mol)	5465	20800	45130	40400	18280	35620	49930	21500
$\Delta H_{vH}2$ (kcal/mol)	193400	97870	64540	52100	42540	52050	65610	43020
D_{global}	0.8873	1.6208	0.9756	0.4834	0.6775	1.5424	1.1922	0.4606
Random Coil Ratio Solvent-exposed Surface Area	69.8	0	28.2	49.1	17.8	5.5	83.1	66.1
Number of Contacts in 1BE9	3	6	5	11	4	2	5	6

	A382T	G383S	Q384G	T385E	V386L	T387K	I388L	I389V
log(K_d) (kT)	-6.1408	-5.9606	-6.0484	-6.1163	-5.4342	-6.1323	-6.2119	-5.9842
log(Kd), Standard Error	0.0676	0.0472	0.0711	0.0955	0.1308	0.085	0.0717	0.0978
Bacterial 2-hybrid Normalized Mean eGFP (A.F.U)	0.8823	0.7954	1.0754	0.9677	0.8235	1.0744	0.8242	0.9938
Standard Deviation, 2-hybrid	0.0802	0.0322	0.0365	0.0693	8.92E-03	0.0906	0.0212	0.0202
T_m1 (°C)	62.71	60.33	64.03	69.66	68.23	61.6	69.32	68.13
$\Delta H_{cal}1$ (kcal/mol)	25890	25280	46640	78160	55680	32150	81040	15450
$\Delta H_{vH}1$ (kcal/mol)	107300	96440	91080	95070	113000	80840	91770	130000
T_m2 (°C)	73.32	70.94	76.37	71.05	76.24	52.36	71.12	68.75
$\Delta H_{cal}2$ (kcal/mol)	12490	11320	14660	43270	77290	3776	31540	65250
$\Delta H_{vH}2$ (kcal/mol)	52610	57330	60820	51300	56080	98730	55810	55810
D_{global}	0.5486	0.4914	0.4485	0.2187	1.2881	0.5228	1.5815	0.5168
Random Coil Ratio Solvent-exposed Surface Area	0	48.5	80.2	54.6	0	33.9	0	33
Number of Contacts in 1BE9	12	4	5	10	9	5	7	12

	A390V	Q391A	Y392R	K393P	H372Y	G329A	I336A
log(K_d) (kT)	-5.719	-5.7319	-5.4742	-5.2208	-3.7245	-4.265	-4.735
log(Kd), Standard Error	0.0904	0.1115	0.1468	0.103	0.0698	0.045	0.0267
Bacterial 2-hybrid Normalized Mean eGFP (A.F.U)	0.8616	1.0436	0.8374	0.704	0.4335		
Standard Deviation, 2-hybrid	0.0529	0.0165	0.0264	0.0415	0.0152		
T_m1 (°C)	64.04	63.89	55.64	51.98	64.1656	66.74	62.47
$\Delta H_{cal}1$ (kcal/mol)	23780	34670	20370	19090	33743.948		
$\Delta H_{vH}1$ (kcal/mol)	148900	78300	102500	85780	108375.63		
T_m2 (°C)	66.77	63.23	60.25	57.73	68.7126		
$\Delta H_{cal}2$ (kcal/mol)	27210	4672	6658	5221	23272.661		
$\Delta H_{vH}2$ (kcal/mol)	63930	160600	97330	73040	73880.532		
D_{global}	1.3249	0.2486	0.817	0.2112	1.6743		
Random Coil Ratio Solvent-exposed Surface Area	0	19.6	0.9	35.9	48.6		

Table Appendix II.2 The Enrichment of NMCAA Mutants

The enrichment for each NMCAA mutant at two gates was measured as described in the text. Briefly, the count of each allele was determined in the input and the flow-cytometry-selected populations by Solexa sequencing. The enrichment represents the logarithm of the ratio of the frequency of each allele in the selected relative to the input population.

	10% Enrichment	25% Enrichment
WT	0.19	0.17
P311I	0.33	0.37
R312K	0.18	0.22
R313T	0.25	0.20
I314V	0.27	0.27
V315E	0.16	0.22
I316L	0.38	0.30
H317E	0.16	0.18
R318K	0.30	0.18
G319P	0.16	0.16
S320G	0.26	0.23
T321G	0.28	0.24
G322S	0.26	0.18
L323F	0.08	0.13
G324S	-0.44	-0.38
F325I	-0.07	0.02
N326S	0.21	0.17
I327L	-0.30	-0.37
I328A	-0.21	-0.09
G329S	-0.03	0.04
G330Q	-0.55	-0.56
E331K	-0.44	-0.39
D332G	0.05	0.11
G333N	0.18	0.28
E334G	0.18	0.20
G335P	0.42	0.38
I336V	0.17	0.15
F337Y	0.34	0.20
I338V	0.34	0.28

	10% Enrichment	25% Enrichment
S339K	-0.05	-0.01
F340S	0.11	0.14
I341V	0.05	0.15
L342I	0.17	0.20
A343P	0.24	0.17
G344D	0.15	0.14
G345S	0.05	0.13
P346A	0.22	0.19
A347V	-0.05	0.00
D348A	0.12	0.10
L349R	0.09	0.09
S350D	0.11	0.15
G351N	0.16	0.13
E352R	0.31	0.25
L353I	0.04	0.12
R354Q	0.21	0.21
K355V	0.19	0.21
G356N	-0.16	0.14
D357N	-0.06	-0.02
Q358R	0.20	0.17
I359L	0.10	0.10
L360V	0.17	0.17
S361A	0.03	0.06
V362I	-0.08	0.00
N363D	0.04	0.12
G364D	0.11	0.15
V365T	0.21	0.16
D366S	0.30	0.26
L367V	-0.05	0.03

	10% Enrichment	25% Enrichment
R368E	0.02	0.13
N369G	0.10	0.09
A370L	0.21	0.09
S371T	0.10	0.13
H372L	-0.89	-0.95
E373D	0.15	0.09
Q374E	0.23	0.17
A375V	-0.22	-0.06
A376V	0.23	0.13
I377E	0.21	0.14
A378L	0.27	0.11
L379I	0.08	0.05
K380R	0.13	0.14
N381A	0.22	0.17
A382T	0.20	0.15
G383S	0.17	0.11
Q384G	0.14	0.12
T385E	0.07	0.01
V386L	-0.14	-0.10
T387K	0.30	0.17
I388L	0.14	0.15
I389V	0.20	0.16
A390V	0.06	0.10
Q391A	0.09	0.11
Y392R	-0.01	0.05
K393P	-0.44	-0.43
H372Y	-0.66	-0.98
G329A	-1.11	-1.38
I336A	-0.65	-0.72

Table Appendix II.3 The Enrichment Values of Comprehensive NNS Mutagenesis in the Context of CRIPT

The enrichment for each mutant represents the average of the enrichment at the 10% and 25% gates as described in the text. If an amino acid has more than one codon encoded by 'NNS,' the codon counts are summed for all codons of an amino acid.

Amino Acid	Position 311	312	313	314	315	316	317	318	319	320	321	322	323	324
A	0.228	0.125	0.239	0.305	0.273	0.291	0.218	-0.005	0.204	0.236	0.265	0.217	-0.128	-0.103
C	0.295	0.120	0.262	0.283	0.296	0.260	0.386	-0.020	0.161	0.274	0.273	0.175	0.073	-0.041
D	0.359	0.113	0.160	0.226	0.226	-0.365	0.280	-0.181	0.065	0.231	0.291	0.043	-1.355	-0.329
E	0.246	0.142	0.229	0.198	0.202	-0.003	0.179	-0.185	0.228	0.294	0.226	0.115	-1.384	-0.320
F	0.398	-0.108	0.137	0.362	0.373	0.247	0.288	-0.122	0.235	0.150	0.150	-0.084	0.115	-1.124
G	0.419	0.102	0.287	0.278	0.287	0.146	0.287	0.032	0.242	0.256	0.272	0.272	-0.923	0.245
H	0.444	0.299	0.207	0.275	0.263	-0.325	0.279	-0.002	0.151	0.193	0.273	0.111	-0.830	-0.673
I	0.358	-0.045	0.072	0.000	0.220	0.000	0.321	-0.035	0.082	0.245	0.259	0.102	-0.202	0.000
K	0.357	0.212	0.250	0.045	0.232	-0.347	0.418	0.252	0.192	0.222	0.206	0.135	-1.391	-0.453
L	0.354	-0.081	0.175	0.228	0.267	0.355	0.366	-0.136	0.151	0.224	0.234	0.020	0.219	-1.018
M	0.333	0.146	0.110	0.286	0.171	0.300	0.415	0.050	0.213	0.276	0.272	-0.038	0.236	-0.429
N	0.394	0.300	0.169	0.309	0.283	0.144	0.232	-0.146	0.245	0.215	0.326	0.175	-0.263	-0.163
P	0.198	-0.066	0.204	-0.057	0.253	-0.213	0.179	0.079	0.170	0.236	0.238	0.214	-1.099	-1.430
Q	0.418	0.268	0.218	0.094	0.299	0.305	0.246	-0.009	0.205	0.182	0.329	0.151	-0.840	-0.314
R	0.374	0.280	0.299	0.166	0.364	-0.274	0.364	0.258	0.182	0.243	0.230	0.079	-1.118	-0.797
S	0.236	0.173	0.185	0.259	0.303	0.239	0.238	0.011	0.233	0.291	0.255	0.233	-0.123	-0.398
T	0.348	0.158	0.238	0.293	0.262	0.250	0.175	0.043	0.165	0.274	0.261	0.206	-0.432	-1.189
V	0.247	0.155	0.245	0.283	0.305	0.198	0.255	-0.041	0.151	0.262	0.248	0.110	-0.425	-1.236
W	0.444	-0.157	0.212	-0.305	0.264	-0.711	0.390	-0.166	0.210	0.202	0.186	-0.001	0.175	-1.152
Y	0.463	0.135	0.208	0.298	0.302	-0.274	0.288	-0.048	0.155	0.177	0.252	-0.019	0.131	-1.077

	325	326	327	328	329	330	331	332	333	334	335	336	337	338
A	0.096	0.179	-1.409	-0.134	-1.232	-0.561	-0.093	0.005	0.115	0.247	0.308	-0.669	-0.096	0.224
C	0.190	0.220	-1.333	0.008	-1.153	-0.325	-0.049	0.157	0.269	0.288	0.342	-0.061	-0.035	0.276
D	-1.752	0.006	-1.325	-1.455	-0.723	-0.188	0.162	0.184	0.195	0.264	0.178	-0.327	-0.202	-0.766
E	-1.360	-0.755	-1.259	-0.662	-1.329	-0.062	0.000	0.096	0.243	0.176	0.258	-0.229	-0.392	-1.412
F	0.000	0.032	-0.120	-0.145	-1.102	-0.828	0.132	0.196	0.222	0.428	0.171	-0.614	0.000	-0.267
G	-0.079	0.012	-1.350	-0.471	0.195	0.224	-0.015	0.092	0.212	0.198	0.263	-0.857	-0.003	0.031
H	-0.363	-0.017	-1.523	-0.247	-1.185	-0.699	0.044	0.192	0.253	0.255	0.307	-0.154	0.189	-1.563
I	-0.013	0.176	0.133	0.229	0.000	-1.163	0.155	0.082	0.229	0.312	0.134	0.000	0.028	0.000
K	-0.622	0.176	0.000	-0.022	-1.538	-0.971	-0.401	0.123	0.263	0.211	0.270	-0.253	0.011	0.000
L	0.007	-0.014	-0.322	0.028	-1.578	-0.978	-0.014	0.178	0.156	0.271	0.224	-0.021	0.143	0.244
M	0.240	0.270	-0.296	0.221	-1.068	-1.040	0.064	0.075	0.251	0.275	0.205	-0.007	0.213	0.150
N	-0.624	0.000	-1.303	-0.962	-0.919	-0.508	0.069	0.182	0.242	0.192	0.241	-0.135	-0.091	0.060
P	-1.477	-1.075	-1.451	-1.394	-1.363	-0.022	-1.098	-0.475	-0.023	0.232	0.413	0.011	-1.181	-0.549
Q	-0.817	-0.021	-1.388	0.046	-1.234	-0.544	0.094	0.065	0.323	0.230	0.198	0.087	-0.163	-0.658
R	-1.450	0.145	-1.361	0.021	-1.288	-0.833	-0.328	0.120	0.291	0.241	0.308	-0.752	0.003	-1.184
S	-0.213	0.199	-1.391	-0.189	0.016	-0.464	-0.003	0.070	0.254	0.279	0.261	-0.361	0.022	0.315
T	-0.070	0.241	-1.387	-0.064	-0.353	-0.423	0.065	0.163	0.211	0.289	0.166	0.027	-0.110	0.302
V	-0.007	0.262	-0.182	0.205	-1.386	-1.051	0.050	0.026	0.234	0.334	0.293	0.172	-0.083	0.325
W	-0.188	0.202	-1.277	-0.467	-0.838	-0.631	0.104	0.286	0.158	0.341	0.160	-1.386	0.153	-1.614
Y	0.070	0.127	-0.351	-0.334	-1.371	-0.926	0.099	0.213	0.153	0.324	0.177	0.000	0.280	0.000

	339	340	341	342	343	344	345	346	347	348	349	350	351	352
A	0.182	0.167	-0.514	0.173	0.182	0.178	0.157	0.222	0.198	0.125	0.150	0.202	0.140	0.185
C	0.062	0.047	0.036	0.102	0.156	0.246	-0.040	0.182	0.074	0.010	0.157	0.190	0.186	0.154
D	-0.137	-0.008	-1.337	-0.087	0.199	0.162	-0.095	-0.323	-0.920	0.000	0.185	0.143	0.063	0.174
E	-0.165	-0.083	-1.217	-0.143	0.214	0.107	-0.090	0.014	-0.954	0.094	0.176	0.207	0.147	0.000
F	-0.005	0.000	0.123	0.211	0.134	0.208	0.058	-0.104	-1.292	0.028	0.139	0.052	0.049	0.127
G	0.015	0.181	-1.109	0.033	0.212	0.242	0.195	0.131	-0.015	0.150	0.063	0.206	0.175	0.160
H	0.096	0.166	-0.810	0.172	0.183	0.159	0.052	0.086	-0.969	0.139	0.150	0.227	0.124	0.179
I	0.090	-0.099	0.000	0.201	0.102	-0.208	-0.720	0.034	-0.396	-0.383	0.218	0.113	-0.274	-0.020
K	-0.016	0.058	-0.174	0.295	0.129	0.180	0.022	0.130	-0.755	-0.151	0.166	0.126	0.175	0.274
L	-0.102	0.124	0.181	0.179	0.190	0.236	0.042	-0.016	-0.661	-0.255	0.188	0.145	0.077	0.218
M	0.026	0.098	0.145	0.223	0.146	0.075	0.100	0.132	-0.801	0.015	0.131	0.169	0.126	0.253
N	-0.188	0.004	-0.506	-0.007	0.108	0.105	-0.198	0.192	-0.629	-0.005	0.198	0.210	0.162	0.251
P	-0.063	-1.382	-0.958	0.101	0.222	-0.352	-0.790	0.194	-0.824	-0.722	-0.484	0.026	-0.139	-0.101
Q	-0.022	0.139	-0.200	0.051	0.124	0.197	0.122	0.140	-0.590	0.139	0.128	0.181	0.160	0.259
R	-0.021	0.047	-0.376	0.154	-0.108	0.340	-0.056	0.120	-0.793	-0.283	0.108	0.130	0.137	0.296
S	0.190	0.141	-0.608	0.120	0.181	0.147	0.105	0.212	0.091	0.171	0.125	0.148	0.209	0.188
T	0.023	0.035	-0.120	0.058	0.180	0.002	0.062	0.029	0.149	-0.311	0.146	0.216	0.186	0.174
V	0.008	0.051	0.115	0.131	0.176	-0.145	-0.587	0.097	-0.010	-0.395	0.146	0.196	-0.057	0.110
W	-0.333	0.156	-0.778	-0.093	0.148	0.161	-0.103	-0.356	-1.067	0.159	0.131	-0.004	-0.097	0.096
Y	-0.027	0.161	-0.542	0.058	0.243	0.187	-0.011	-0.107	-0.967	-0.229	0.152	0.041	-0.027	0.159

	353	354	355	356	357	358	359	360	361	362	363	364	365	366
A	0.232	0.170	0.183	0.202	0.098	0.134	-0.031	0.142	0.065	0.189	0.104	0.142	0.149	0.198
C	0.252	0.188	0.192	0.259	0.195	0.220	0.144	0.151	0.109	0.261	0.206	0.069	0.144	0.172
D	-1.123	0.222	0.104	-0.203	0.000	-0.308	-1.322	-0.058	0.042	-1.068	0.094	0.143	0.088	0.000
E	-1.184	0.224	0.173	0.023	0.047	0.188	-1.169	0.165	0.089	-0.309	0.261	0.207	0.163	0.121
F	0.231	0.139	0.197	0.164	0.095	0.173	-0.303	0.141	0.273	-0.183	-0.151	0.209	0.281	0.134
G	-0.046	0.209	0.173	0.200	-0.182	0.127	-0.564	-0.126	0.128	-0.138	0.173	0.189	0.196	0.200
H	-0.172	0.176	0.147	0.097	0.186	0.202	-1.409	0.171	0.249	0.000	0.173	0.253	0.186	0.212
I	0.099	0.217	0.183	-0.306	-0.015	0.138	0.000	0.203	0.260	-0.027	-0.069	0.064	0.250	0.102
K	-0.564	0.204	0.000	0.107	-0.257	0.091	-1.474	0.244	0.182	-0.351	0.036	0.104	0.144	0.123
L	0.177	0.203	0.138	-0.268	-0.106	0.186	0.119	0.181	0.270	-0.095	0.049	0.184	0.240	0.121
M	0.152	0.172	0.170	0.238	0.236	0.156	-0.044	0.184	0.108	0.041	0.193	0.254	0.218	0.092
N	-0.239	0.213	0.190	0.005	-0.024	0.201	-1.234	0.191	0.179	-0.292	0.162	0.182	0.138	0.191
P	0.175	-0.947	0.189	-0.494	-1.284	-0.984	-0.162	-0.396	-1.250	-1.089	-0.160	-0.208	-0.234	0.181
Q	-0.174	0.228	0.152	0.119	0.249	0.000	-1.169	0.165	0.182	-0.128	0.239	0.173	0.162	0.186
R	-1.095	0.214	0.150	0.109	-0.986	0.202	-1.358	0.259	0.273	-0.434	0.064	0.206	0.223	0.285
S	0.079	0.205	0.171	0.165	0.045	0.162	-0.451	0.103	0.190	0.166	0.156	0.136	0.191	0.276
T	0.135	0.179	0.190	0.060	0.153	0.190	-0.237	0.146	0.165	0.169	0.108	0.128	0.202	0.214
V	0.146	0.170	0.217	-0.224	0.097	0.149	0.098	0.186	0.217	0.134	0.079	0.121	0.191	0.190
W	-0.207	0.029	0.100	0.185	-0.246	0.187	0.000	0.191	0.226	-0.755	0.067	0.281	0.182	0.165
Y	-0.181	0.115	0.003	0.179	0.079	0.162	-1.177	0.199	0.302	0.068	0.074	0.325	0.276	0.330

	367	368	369	370	371	372	373	374	375	376	377	378	379	380
A	0.013	0.039	0.082	0.147	0.151	-1.215	-0.102	0.140	0.092	0.167	0.137	0.134	-0.133	-0.056
C	-0.060	0.039	0.120	0.175	0.142	-0.848	-0.084	0.127	-0.021	-0.180	0.066	0.147	0.118	0.112
D	-0.603	0.035	0.059	-0.021	0.094	-1.017	0.120	0.071	-0.354	-0.986	0.130	0.070	-1.161	-0.218
E	-0.499	0.073	0.044	0.171	0.152	-1.208	0.080	0.197	-0.767	-0.327	0.177	0.169	-1.376	-0.194
F	0.067	0.055	0.160	0.128	0.165	-0.809	-0.276	0.182	-0.993	-0.134	0.046	0.106	0.084	0.135
G	-0.356	0.106	0.092	0.072	0.110	-1.254	-0.074	0.137	-0.035	-0.055	0.152	0.161	-0.595	0.059
H	0.036	0.082	0.135	0.079	0.137	0.000	-0.102	0.133	-0.430	-0.223	0.197	0.183	-0.112	0.104
I	0.051	0.104	0.204	0.158	0.128	-0.870	-0.153	-0.038	-0.177	-0.087	0.088	0.187	0.068	0.124
K	-0.324	0.198	0.158	0.200	0.152	-1.164	-0.241	0.071	-1.242	-0.420	-0.010	0.232	-0.957	0.000
L	0.111	0.079	0.115	0.148	0.191	-0.923	-0.094	0.167	-0.826	-0.583	0.144	0.187	0.081	0.123
M	-0.003	-0.007	0.119	0.085	0.216	-1.466	-0.116	0.127	0.021	-0.791	0.109	0.139	-0.015	0.151
N	-0.259	0.057	0.085	0.068	0.077	-0.742	0.014	0.119	-0.551	-0.668	0.068	0.218	-0.519	0.103
P	-0.528	-0.080	0.073	0.059	0.196	-1.167	-0.348	0.128	-1.058	-0.395	-0.096	-0.174	-1.019	-0.171
Q	0.041	0.081	0.100	0.110	0.194	-0.422	0.028	0.000	-0.416	-1.017	0.102	0.068	-1.177	0.059
R	-0.332	0.209	0.174	0.128	0.188	-0.952	-0.251	0.122	-1.414	-0.398	0.136	0.151	-1.278	0.135
S	-0.081	0.056	0.154	0.117	0.133	-1.029	-0.095	0.121	0.094	-0.318	0.166	0.145	-0.387	0.025
T	0.019	0.062	0.147	0.103	0.112	-0.957	-0.075	0.086	-0.068	-0.372	0.159	0.190	-0.127	0.137
V	-0.012	0.053	0.124	0.100	0.178	-0.633	-0.089	0.115	-0.141	0.179	0.157	0.123	-0.034	0.142
W	-0.422	0.127	0.009	0.053	0.110	-1.560	-0.103	0.122	-0.974	-0.307	-0.054	0.138	-0.841	0.086
Y	0.077	0.046	0.018	0.071	0.150	-0.821	-0.164	0.070	0.000	0.054	0.251	0.102	-0.408	0.068

	381	382	383	384	385	386	387	388	389	390	391	392	393
A	0.192	0.142	0.160	0.164	0.114	0.156	0.188	0.065	0.234	0.138	0.102	0.122	0.076
C	0.155	0.189	0.063	0.118	0.173	0.087	0.202	0.190	0.253	0.154	0.170	0.080	0.119
D	0.095	0.068	0.091	0.024	0.147	-0.189	-0.015	-1.284	0.049	-0.199	0.170	0.081	0.115
E	-0.002	0.123	0.118	0.060	0.042	-0.354	0.087	-1.246	0.014	-0.251	0.177	0.170	0.154
F	0.152	-0.036	0.125	0.278	0.054	-0.490	0.254	0.280	0.186	0.094	0.269	0.199	0.116
G	0.163	0.156	0.161	0.132	0.149	0.105	0.259	-0.041	0.194	0.139	0.077	-0.003	0.113
H	0.107	0.065	0.171	0.093	0.190	0.082	0.209	-0.211	0.206	0.018	0.221	0.196	0.109
I	0.002	0.037	0.114	0.224	0.239	-0.033	0.115	0.000	0.000	0.038	0.285	-0.026	0.016
K	0.193	0.096	0.158	0.156	0.188	-0.185	0.236	-0.455	0.369	0.110	0.058	0.084	0.122
L	0.082	0.002	0.111	0.186	0.176	-0.120	0.212	0.145	0.256	0.135	0.187	0.135	0.125
M	0.084	0.092	0.087	0.131	0.148	-0.215	0.162	0.083	0.144	0.087	0.240	0.031	0.086
N	0.130	0.115	0.062	0.077	0.162	0.122	0.144	-0.331	0.235	0.108	0.276	0.270	0.139
P	-0.066	0.055	0.110	0.178	0.181	0.147	0.088	-0.650	-0.645	-0.038	-0.563	-0.025	-0.435
Q	0.134	0.234	0.157	0.000	0.155	0.065	0.119	-0.145	0.243	-0.168	0.000	0.234	0.166
R	0.066	0.057	0.086	0.145	0.292	-0.303	0.244	-1.311	0.331	0.118	0.166	0.017	0.133
S	0.171	0.182	0.139	0.102	0.135	0.182	0.199	-0.174	0.240	0.169	0.156	0.126	0.163
T	0.124	0.174	0.119	0.141	0.141	0.139	0.134	-0.035	0.234	0.184	0.151	0.101	0.132
V	0.056	0.041	0.164	0.129	0.230	0.076	0.183	0.112	0.180	0.080	0.099	-0.036	0.059
W	0.048	-0.081	0.222	0.210	0.165	-0.465	0.186	-0.518	0.165	-0.263	0.309	0.156	0.108
Y	-0.008	0.061	0.099	0.194	0.057	-0.100	0.163	-0.590	0.211	0.017	0.235	0.062	0.170

Table Appendix II.4 The Enrichment Values of Comprehensive NNS Mutagenesis in the Context of T(P_{-2})F

The enrichment for each mutant represents the average of the enrichment at the 10% and 25% gates as described in the text. If an amino acid has more than one codon encoded by 'NNS,' the codon counts are summed for all codons of an amino acid.

Amino Acid	Position 311	312	313	314	315	316	317	318	319	320	321	322	323	324
A	-0.204	-0.468	-0.075	-0.069	-0.022	0.001	-0.006	-0.705	-0.268	-0.004	0.105	0.547	-0.759	-0.624
C	-0.093	-0.311	-0.149	-0.099	0.014	-0.072	0.072	-0.975	-0.239	-0.067	0.102	0.442	-0.529	-0.519
D	-0.021	-0.447	-0.152	-0.077	-0.155	-0.878	-0.161	-0.880	-0.429	-0.049	0.013	-0.057	-1.164	-0.837
E	-0.224	-0.633	-0.184	-0.288	-0.116	-0.094	-0.081	-0.911	-0.345	-0.088	-0.012	0.003	-0.952	-0.757
F	-0.161	-0.511	0.024	-0.187	0.074	-0.193	-0.163	-0.616	-0.284	-0.073	-0.011	0.647	-0.178	-0.823
G	-0.156	-0.352	-0.072	-0.169	-0.035	-0.211	0.009	-0.766	-0.042	-0.073	0.016	-0.004	-1.078	-0.048
H	-0.005	-0.273	-0.062	-0.333	0.023	-0.921	-0.067	-0.734	-0.374	0.049	0.210	0.544	-1.033	-0.892
I	-0.043	-0.372	-0.171	0.000	-0.089	0.000	-0.031	-0.690	-0.347	-0.025	-0.005	0.672	-0.869	-1.146
K	0.018	-0.316	-0.236	-0.546	-0.010	-0.977	0.108	-0.098	-0.401	0.063	0.101	0.796	-0.978	-1.101
L	-0.035	-0.639	-0.185	-0.122	-0.047	0.096	-0.042	-0.852	-0.351	-0.041	-0.043	0.807	-0.018	-0.923
M	0.059	-0.375	-0.175	-0.191	0.015	-0.091	-0.070	-0.549	-0.417	0.028	0.008	0.797	-0.254	-0.897
N	-0.067	-0.360	-0.037	-0.059	-0.110	-0.556	-0.017	-1.099	-0.375	-0.154	0.149	0.560	-0.888	-0.734
P	-0.048	-1.062	-0.093	-0.362	-0.070	-0.668	-0.093	-0.591	-0.265	0.036	0.052	0.333	-0.817	-1.029
Q	0.033	-0.390	-0.167	-0.268	-0.014	-0.097	0.067	-0.600	-0.339	0.074	0.055	0.533	-1.045	-0.754
R	0.089	-0.046	-0.051	-0.507	0.095	-0.907	0.058	-0.016	-0.271	-0.105	0.098	0.748	-1.102	-0.908
S	-0.178	-0.313	-0.110	-0.176	-0.037	-0.148	-0.005	-0.741	-0.211	-0.060	0.103	0.292	-0.940	-0.869
T	-0.051	-0.421	-0.160	-0.093	0.031	-0.284	-0.017	-0.758	-0.452	-0.042	0.005	0.444	-0.798	-1.092
V	-0.061	-0.457	-0.075	-0.084	-0.028	-0.127	-0.010	-0.832	-0.405	-0.104	0.085	0.730	-1.022	-1.051
W	0.078	-0.710	0.002	-0.666	0.072	-0.954	0.001	-0.741	-0.290	-0.030	0.115	0.363	0.036	-1.146
Y	-0.024	-0.401	-0.043	-0.383	-0.064	-0.868	0.010	-0.571	-0.319	-0.079	0.195	0.739	-0.197	-0.922

	325	326	327	328	329	330	331	332	333	334	335	336	337	338
A	0.120	-0.435	0.558	-0.600	-0.485	0.582	-0.182	0.209	0.075	-0.268	-0.303	0.415	-0.320	0.046
C	0.299	0.203	0.671	-0.346	0.120	0.791	-0.125	0.110	0.034	-0.181	-0.436	0.141	-0.328	0.400
D	-1.118	-0.521	-1.071	-0.926	-0.517	0.640	0.051	-0.003	-0.151	0.254	-0.356	0.245	-0.834	-0.950
E	-0.986	-0.940	-0.971	-0.871	-0.667	0.708	0.000	0.105	-0.134	-0.018	-0.215	0.358	-0.930	-0.773
F	0.000	-0.388	-0.496	0.408	0.397	-0.263	-0.477	-0.125	0.077	-0.336	-0.019	0.195	0.000	-0.613
G	-0.122	-0.643	0.829	-0.606	-0.017	0.026	-0.417	0.405	-0.013	0.093	-0.031	0.547	-0.518	-0.541
H	-1.007	-0.714	-1.041	-0.224	-0.078	0.447	-0.546	-0.015	0.002	-0.261	-0.392	-0.129	-0.228	-0.978
I	-0.448	0.372	-0.085	0.320	0.448	-0.457	-0.169	0.144	0.571	-0.129	-0.588	0.000	-0.327	0.000
K	-1.059	-0.423	-0.918	-0.769	-0.304	0.023	-0.481	-0.092	-0.092	-0.304	-0.474	0.604	-0.740	-0.879
L	-0.176	-0.664	-0.137	0.322	-0.023	-0.136	-0.251	0.060	0.564	0.039	-0.466	-0.419	-0.127	-0.264
M	0.462	0.049	-0.661	0.152	0.287	0.116	-0.098	0.128	0.034	-0.520	-0.392	-0.004	-0.101	-0.401
N	-1.060	0.000	-1.130	-0.853	-0.336	0.653	-0.185	-0.147	-0.151	-0.221	-0.350	0.234	-0.806	-0.258
P	-1.071	-1.040	-1.058	0.088	-0.517	-0.283	-0.122	0.252	0.591	0.203	-0.500	0.536	-0.868	-1.003
Q	-1.287	-0.835	-0.996	-0.718	-0.252	0.338	-0.262	0.078	-0.040	-0.259	-0.342	-0.373	-0.499	-0.989
R	-1.096	-0.454	-0.988	-0.727	-0.379	0.145	-0.600	0.002	-0.042	-0.364	-0.396	0.571	-0.600	-0.997
S	-0.489	-0.061	-0.952	-0.743	-0.172	0.765	-0.108	0.137	-0.008	0.006	-0.376	0.617	-0.458	-0.007
T	-0.404	0.329	-0.872	-0.465	-0.358	0.897	0.004	0.081	-0.032	-0.004	-0.162	0.529	-0.513	0.187
V	-0.353	0.495	-0.072	-0.041	0.167	0.194	-0.179	0.136	0.398	-0.357	-0.471	0.133	-0.408	0.204
W	-0.862	0.332	-0.975	-0.338	0.455	0.345	-0.408	0.051	-0.055	-0.274	-0.238	0.278	-0.033	-1.027
Y	-0.133	-0.182	-0.798	-0.021	0.286	0.112	-0.360	-0.057	-0.184	-0.231	-0.175	0.408	-0.098	-0.976

	339	340	341	342	343	344	345	346	347	348	349	350	351	352
A	0.155	-0.109	-0.882	0.043	0.108	-0.007	0.096	0.103	0.111	0.013	0.019	-0.002	0.053	0.026
C	-0.269	-0.227	-0.387	-0.105	0.101	0.143	-0.561	-0.165	-0.265	-0.316	0.021	0.104	0.007	0.020
D	-0.784	-0.389	-0.946	-0.708	0.058	0.063	-0.642	-0.732	-0.816	0.000	0.013	0.033	0.028	0.037
E	-0.632	-0.659	-1.021	-0.422	0.118	-0.003	-0.504	-0.302	-0.823	-0.057	0.017	0.004	-0.033	0.000
F	-0.222	0.000	-0.103	-0.306	0.033	0.070	-0.336	-0.512	-1.210	-0.101	-0.026	-0.415	-0.270	0.082
G	-0.664	0.239	-1.040	-0.300	0.215	0.121	0.131	-0.279	-0.237	0.048	-0.290	0.086	0.155	-0.165
H	-0.256	-0.115	-1.000	-0.143	0.172	0.009	-0.165	-0.184	-1.007	-0.025	0.002	0.079	-0.026	0.100
I	0.204	-0.648	0.000	-0.302	0.155	-0.731	-0.932	-0.196	-0.910	-0.780	0.129	0.125	-0.648	-0.241
K	-0.560	-0.030	-0.612	0.287	0.010	0.062	-0.147	0.086	-1.142	-0.752	0.051	-0.168	0.076	0.202
L	-0.203	-0.014	0.105	0.149	0.110	0.096	-0.176	-0.298	-0.823	-0.736	0.141	0.047	-0.237	0.133
M	-0.300	-0.044	-0.074	-0.097	0.086	0.091	-0.209	0.176	-0.758	-0.321	0.063	-0.070	0.028	0.164
N	-0.485	-0.563	-0.980	-0.404	0.075	0.121	-0.662	-0.011	-0.793	-0.097	-0.031	0.166	0.032	0.046
P	0.139	-0.900	-0.917	-0.072	0.271	-0.787	-0.964	0.112	-0.872	-0.864	-0.746	-0.461	-0.592	-0.555
Q	-0.386	-0.397	-0.700	-0.145	0.130	0.095	-0.304	0.148	-0.880	-0.088	0.078	0.238	0.067	0.170
R	-0.425	-0.443	-0.843	0.295	-0.671	0.152	-0.331	0.147	-0.909	-0.794	0.075	-0.180	0.021	0.161
S	0.138	0.058	-1.014	-0.162	0.149	0.001	-0.225	0.101	-0.254	-0.007	-0.004	0.120	0.021	0.016
T	0.339	-0.398	-0.409	-0.313	0.132	-0.196	-0.283	-0.354	0.121	-0.767	-0.020	0.176	0.033	-0.088
V	0.416	-0.531	-0.033	-0.094	0.096	-0.514	-0.874	-0.147	-0.410	-0.774	0.041	0.220	-0.431	-0.176
W	-0.499	-0.175	-0.971	-0.265	0.038	-0.010	-0.353	-0.707	-0.867	0.064	-0.068	-0.465	-0.363	-0.119
Y	-0.400	0.024	-0.890	-0.321	0.026	0.127	-0.397	-0.495	-0.853	-0.577	-0.031	-0.260	-0.350	-0.011

	353	354	355	356	357	358	359	360	361	362	363	364	365	366
A	0.124	0.167	0.092	0.197	-0.249	0.166	0.114	0.020	0.136	0.460	-0.136	0.140	0.198	-0.023
C	0.121	0.147	0.144	0.342	-0.180	0.231	0.222	0.101	0.200	0.437	0.071	0.148	0.126	0.028
D	-0.995	0.274	0.167	-0.440	0.000	0.705	-0.967	0.054	0.068	-0.762	-0.320	0.157	0.190	0.000
E	-1.074	0.176	0.236	0.250	-0.109	0.332	-0.907	0.005	0.173	0.609	0.049	0.148	0.112	-0.156
F	0.388	0.069	0.135	0.301	-0.159	0.250	-0.528	0.119	0.241	-0.346	0.006	0.421	0.247	-0.117
G	-0.031	0.242	0.041	0.151	-0.490	0.082	-0.892	-0.060	0.255	0.601	-0.006	0.147	0.191	-0.041
H	-0.459	0.129	-0.025	0.232	-0.126	0.269	-0.863	0.063	0.268	0.000	0.190	0.258	0.227	-0.021
I	-0.021	0.277	0.165	-0.709	-0.198	0.171	0.000	0.101	0.304	-0.365	0.223	0.079	0.147	-0.001
K	-1.026	0.180	0.000	0.212	-0.786	0.021	-1.067	0.102	0.084	-0.413	0.076	0.163	0.304	-0.181
L	0.159	0.191	0.076	-0.152	-0.120	0.203	0.014	0.118	0.271	-0.641	0.144	0.218	0.299	-0.115
M	-0.011	0.190	0.129	0.216	0.117	0.178	-0.304	0.143	0.206	-0.232	0.172	0.342	0.355	-0.143
N	-0.173	0.162	0.037	0.043	-0.176	0.173	-1.033	0.134	0.121	0.399	0.121	0.190	0.225	-0.020
P	0.070	-0.950	0.185	-0.826	-1.069	-0.635	0.138	-0.755	-0.835	-0.940	0.129	-0.097	-0.138	-0.052
Q	-0.476	0.208	0.096	0.305	0.202	0.000	-0.952	0.024	0.185	0.631	0.198	0.165	0.173	-0.059
R	-1.066	0.149	0.120	0.211	-1.053	0.175	-1.020	0.099	0.253	-0.390	0.061	0.218	0.314	-0.047
S	0.047	0.152	-0.004	0.173	-0.374	0.176	-0.810	0.092	0.169	0.623	-0.004	0.160	0.223	0.116
T	0.030	0.232	0.145	0.190	-0.268	0.178	0.087	0.144	0.207	0.424	0.034	0.177	0.138	0.068
V	0.129	0.134	0.253	-0.123	-0.178	0.165	0.109	0.097	0.255	0.284	0.160	0.221	0.209	-0.050
W	-0.266	0.048	0.222	-0.081	-0.492	0.033	-0.048	0.106	0.305	0.207	0.158	0.345	0.149	-0.014
Y	-0.511	0.178	0.059	0.272	-0.134	0.207	-1.272	0.148	0.334	0.500	-0.147	0.214	0.305	0.024

	367	368	369	370	371	372	373	374	375	376	377	378	379	380
A	0.042	-0.123	-0.045	-0.035	0.202	0.823	-0.571	-0.038	-0.039	-0.110	-0.155	-0.056	0.306	-0.432
C	-0.108	-0.122	-0.109	0.088	-0.004	0.815	-0.166	-0.076	-0.200	-0.542	-0.121	0.065	0.447	0.198
D	-0.900	-0.144	-0.031	0.275	-0.265	0.706	-0.291	-0.111	-0.676	-0.860	-0.145	0.112	-1.121	-0.707
E	0.029	-0.168	-0.141	0.154	0.164	-0.490	-0.051	-0.050	-0.615	-0.952	-0.138	0.026	-1.074	-0.628
F	-0.034	-0.314	0.061	0.338	0.368	0.053	0.098	-0.059	-0.527	0.604	0.514	0.345	0.684	-0.003
G	-0.100	-0.035	-0.090	-0.293	0.078	0.834	-0.477	0.107	0.282	-0.527	-0.171	-0.031	-0.752	0.042
H	0.195	-0.126	0.013	0.427	0.214	0.000	-0.148	-0.004	-0.025	-0.509	0.034	0.102	0.526	0.127
I	0.001	0.042	-0.142	0.247	0.576	0.802	0.404	-0.148	0.179	-0.597	-0.119	0.286	0.026	0.380
K	0.155	-0.087	-0.037	0.196	0.203	-0.201	-0.567	-0.184	-0.394	-0.904	-0.226	0.292	-0.800	0.000
L	-0.013	-0.068	0.023	0.367	0.200	0.804	-0.085	-0.157	-0.268	-0.674	-0.222	0.270	-0.074	0.169
M	0.154	-0.129	0.039	0.292	0.349	0.178	0.025	-0.117	-0.078	-0.658	-0.267	0.218	-0.109	0.259
N	0.104	-0.111	-0.064	0.154	-0.167	0.409	-0.201	0.087	0.154	-0.711	-0.124	0.007	-0.969	0.091
P	-0.432	-0.236	0.209	-0.225	0.071	0.837	-0.810	0.339	-0.126	0.203	0.065	-0.160	-0.926	0.049
Q	0.179	-0.073	-0.020	0.238	0.244	-0.076	-0.251	0.000	0.166	-0.679	-0.227	0.047	-0.986	-0.069
R	0.111	-0.068	0.010	0.264	0.184	-0.316	-0.380	-0.108	-0.366	-0.425	-0.190	0.237	-0.988	0.239
S	0.016	-0.132	-0.013	0.050	-0.011	0.797	-0.461	-0.008	0.047	-0.850	-0.154	0.038	-0.263	-0.167
T	-0.157	-0.113	-0.048	0.162	-0.004	0.784	-0.077	-0.132	-0.178	-0.783	-0.153	0.047	-0.164	0.276
V	-0.028	-0.022	-0.123	0.130	0.496	0.813	0.256	-0.165	0.174	-0.291	-0.105	0.074	-0.024	0.213
W	-0.577	-0.038	-0.074	0.489	0.620	-0.511	0.525	0.129	-0.963	0.045	0.510	0.273	0.100	0.728
Y	0.300	-0.092	0.025	0.397	0.496	-0.296	0.210	-0.104	-1.187	0.682	0.345	0.286	0.143	0.090

	381	382	383	384	385	386	387	388	389	390	391	392	393
A	-0.001	-0.041	-0.235	-0.128	-0.114	-0.088	0.019	-0.340	-0.067	-0.114	-0.165	-0.357	-0.007
C	-0.051	0.100	-0.136	-0.097	-0.068	-0.057	-0.026	-0.221	-0.004	-0.132	-0.127	-0.279	-0.012
D	-0.340	0.110	-0.119	-0.202	-0.119	-0.393	-0.164	-1.143	-0.271	-0.715	-0.080	-0.468	0.085
E	-0.235	0.181	-0.303	-0.232	-0.221	-0.322	-0.078	-1.184	-0.184	-0.826	-0.137	-0.461	-0.083
F	0.007	0.151	-0.144	-0.101	-0.199	-0.795	-0.006	-0.442	-0.086	-0.434	-0.047	-0.073	-0.310
G	0.128	0.082	-0.093	-0.145	-0.114	-0.394	0.058	-0.425	-0.071	-0.187	-0.230	-0.746	-0.049
H	-0.072	0.074	-0.244	-0.084	-0.054	-0.349	-0.041	-0.319	-0.095	-0.411	-0.106	-0.046	-0.032
I	-0.215	-0.181	-0.148	-0.033	0.007	-0.433	-0.116	0.000	0.000	-0.187	-0.207	-0.538	-0.206
K	-0.235	0.050	-0.414	-0.202	-0.144	-0.117	-0.105	-0.931	-0.077	-0.543	-0.216	-0.574	-0.087
L	-0.166	-0.290	-0.280	-0.083	-0.113	-0.716	-0.003	-0.066	-0.006	-0.141	-0.201	-0.171	-0.011
M	-0.339	-0.015	-0.213	-0.095	-0.059	-0.841	-0.027	-0.277	-0.083	-0.152	-0.308	-0.290	-0.118
N	-0.114	0.159	-0.265	-0.084	-0.066	-0.223	-0.080	-0.944	-0.121	0.052	-0.258	0.002	0.025
P	-0.354	0.030	-0.509	-0.043	0.342	0.013	-0.427	-1.079	-0.983	-0.269	-0.955	-0.815	-0.862
Q	-0.113	0.124	-0.342	0.000	-0.050	-0.001	-0.017	-0.482	-0.164	-0.411	0.000	-0.134	-0.067
R	-0.141	0.049	-0.317	-0.155	0.055	-0.906	-0.021	-1.037	-0.047	-0.334	-0.182	-0.694	-0.110
S	0.054	0.202	-0.189	-0.140	-0.126	-0.135	0.015	-0.352	-0.138	-0.092	-0.119	-0.389	-0.043
T	0.064	-0.032	-0.322	-0.147	-0.028	-0.053	-0.045	-0.210	-0.031	-0.147	-0.223	-0.472	-0.084
V	-0.281	-0.256	-0.095	-0.090	0.029	-0.183	-0.026	-0.195	-0.044	-0.136	-0.101	-0.653	0.019
W	-0.203	0.190	-0.030	0.009	-0.134	-0.160	0.041	-1.097	-0.175	-0.774	-0.136	-0.136	-0.148
Y	-0.013	0.202	-0.151	-0.001	-0.081	-0.657	-0.157	-1.022	-0.044	-0.508	-0.177	-0.058	-0.096