THE STRUCTURAL PROPERTIES OF ADAPTATION AND ALLOSTERY IN PROTEINS: A CASE STUDY IN A PDZ DOMAIN

APPROVED BY SUPERVISORY COMMITTEE

Rama Ranganathan, M.D, Ph.D.

Michael Rosen, Ph.D.

Luke Rice, Ph.D.

DEDICATION

This is dedicated to everyone who got me here.

ACKNOWLEDGMENTS

I would like to acknowledge a few key people in particular, people without whom I could not have become the person I am today. First, my parents. Thank you for giving me the gift of perseverance in the face of struggle and always being supportive of me in times of failure. You both taught me how to deal with anything that is thrown my way and it is the single most important trait that I possess today. Second, my fiancée Rasika. It isn't easy having a significant other who has chosen science as their profession. Science is unpredictable, it is (sometimes) cruel, and (mostly) doesn't buy the Ferrari. None of that mattered to you. You encouraged me to pursue what makes me happy, what gets me up in the day, what drives me to succeed and for that I'm eternally thankful. Third, my lab, in particular three people: Salman Banani, Subu Subramanian, and Bill Russ. To the three of you, when I came to UTSW, I would have never imagined meeting three such intelligent people. Now that I'm almost done, I have got to know all three of you and ask myself what more could I have really asked for in scientific colleagues. You three have taught me just as much about science as any course, teacher, or seminar. I'm very lucky to have three friends like you guys. Finally, I would like to thank Rama. The last five years have been the best five years of my life. And a large part of that is because of the (lucky) choice I made in choosing a lab. I found a mentor who is truly a mentor in all senses of the word. I didn't just grow as a scientist, I grew as a person. I learned what commitment means, what drive means, what excellence means, what clarity of thought means. I learned to appreciate the process, not the result. I learned to articulate, to convey, to think logically. I've had only a few good teachers in my

life but only after graduate school did I realize how life-altering one could be. I hope one day to adequately articulate my appreciation. But for now, I can just say thank you.

THE STRUCTURAL PROPERTIES OF ADAPTATION AND ALLOSTERY IN PROTEINS: A CASE STUDY IN A PDZ DOMAIN

by

ARJUN SWAMINATHAN RAMAN

DISSERTATION / THESIS

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 / MASTER OF SCIENCE / MASTER OF ARTS

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

May, 2017

THE STRUCTURAL PROPERTIES OF ADAPTATION AND ALLOSTERY IN PROTEINS: A CASE STUDY IN A PDZ DOMAIN

Publication No. _____

Arjun Swaminathan Raman, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, Graduation Year

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

Complex systems are present at all scales of biology spanning amino acid interactions in proteins to inter-species interactions in ecosystems. Despite their ubiquity, an understanding of how to study complex systems methodically is lacking. Fundamentally, this is because a procedure to discover the appropriate parametrization of such systems into relevant parts is absent. The Ranganathan Lab developed an evolutionary approach, Statistical Coupling Analysis (SCA), to understand how amino acid interactions within proteins give rise to basic protein characteristics such as fold and function. The result of this approach was a structural decomposition of proteins into units of coevolving residues termed protein ‘sectors’.

Is the transformation from amino acids to protein sectors a useful and relevant representation of the essence of proteins? Previous work has demonstrated that specifying co-evolution is sufficient to design synthetic natural-like proteins and encodes the information needed to execute a primary function. As evolved biological systems however, proteins must also also adapt quickly to new functions. In addition, a fundamental characteristic of many proteins is the ability to transmit information over a significant distance in the protein structure, a phenomenon known as allostery. Where in the protein do these characteristics lie? Here, we address the sequence origins and structural properties of a) the adaptive capacity of proteins and b) allosteric communication within proteins.

Understanding how proteins can adapt quickly to new function is an outstanding question in biology, marked by failed attempts at engineering new or altered function guided by structural approaches focused on the importance of active site or binding pocket positions. It is often the case that mutations located distal to the active site harbor adaptive potential, a non-obvious observation given the crystal structure of a protein. To more fundamentally understand the adaptive process in proteins, here we examine a two-step mutational path to new specificity in a model protein $PSD95^{pdz3}$ where one intermediate, a Glycine (G) to Threonine (T) mutation at position 330, removed from the binding site of the protein, maintains native function while simultaneously adapting the protein to an alternate function (termed a 'conditionally neutral' mutation) whereas the other intermediate, a Histidine (H) to Alanine (A) mutation at position 372, promotes a direct specificity switch, abrogating native function. Through a stochastic population dynamics simulation, we find the conditionally neutral intermediate promotes adaptation over a wide range of mutation rates and rates of environmental fluctuation while the direct specificity switching mutation facilitates adaptation only within a special regime of population dynamics parameters. We comprehensively identify the spatial distribution of all adaptive mutations in $PSD95^{pdz3}$, finding that direct specificity switching mutations are found exclusively at sector positions directly at the

binding site whereas conditionally neutral mutations are generally found distal to the binding site but connected to it through the protein sector. Crystal structures reveal how a mutation at position 330 creates plasticity at the binding site to create a dual function protein. Overall, these results illustrate the importance of a spatially distributed network of coupled residues for adapting in a fluctuating environment.

Revealing paths of allostery in protein structures has been a central goal of structural biology. In $PSD95^{pdz3}$, we find that the G330T mutation causes local backbone remodeling and structurally affects solely a distant conserved helix. Detecting structurally coupled residues in the protein reveals a network of connected residues spanning the 330 position to the distant helix propagating through the core of the protein sector. As a check of this allosteric path, we are able to abrogate the allosteric transmission through a mutation in the middle of the protein sector.

Both the ability to adapt to new function and the property of allostery are found almost exclusively within the protein sector. The work highlighted here in addition to previous studies in the lab thus appear to suggest that the sector is a good descriptor and model for understanding how proteins work. It will therefore be important for future studies to determine if other methods of protein design such as Direct Coupling Analysis and Rosetta provide equivalently sufficient descriptions of proteins and what additional information, if any, these approaches reveal.

Contents

List of Figures

Abbreviations

Å ..... Angstrom

amp ..... ampicillin

aTC ..... anhydroustetracycline

B2H bacterial-2-hybrid

bp ..... basepair

Cdc42 ..... cell division cycle 42

clm ..... chloramphenicol

CAT ..... chloramphenicol-acetyl-transferase

CRIPT ..... Cysteine-rich interactor of PDZ3

dox ..... doxycycline

eGFP ..... enhanced Green Fluorescent Protein

FACS ..... fluorescence activated cell aorting

GST ..... glutathione-S-transferase

HKL2000 .... crystallographic analysis software

IPTG ..... isopropyl-β-D-thiogalactopyranoside

$\mathbf{K_d}$ ..... dissociation constant

kan ..... kanamycin

KL ..... Kullback-Leibler Divergence

MSA ..... multiple sequence alignment

μg ..... microgram

mL ..... milliliter

μL ..... microliter

mM ..... millimolar

μM ..... micromolar

nM ..... nanomolar

NaCit ..... Sodium Citrate

Par6 ..... phosphorylated after Rapamycin 6

PSD95 ..... post-synaptic density 95

PDZ ..... PSD95, Discs-large, Zona-occuldens 1

PDZ3 ..... third PDZ domain of PSD95

PCR ..... polymerase chain reaction

Phenix ..... refinement software

SCA ..... statistical coupling analysis

TLS ..... translation libration screw

WT ..... wild-type

G330T ..... glycine to threonine mutation at position 330

H372A ..... histidine to alanine mutation at position 372

T_{-2} ..... threonine amino acid at the -2 position of peptide

$T_{-2}F$ ..... threonine to phenylalanine mutation at the -2 position of peptide

Chapter 1

Introduction

1.1 Complexity and Biology

We live in a world that is considered exceedingly 'complex'. In the vernacular, this term is used to connotate an unreasonable amount of difficulty, indeed bordering on futility, with respect to understanding a phenomenon. Traditionally, this is precisely where science has proven to be fruitful—take a seemingly incomprehensible process (the movement of the planets in our solar system for instance), reduce the process to a far more simplistic form (say from planets to a round object that is for all intents a very, very small planet), and study the process at this level. While actively disregarding certain characteristics of the original process, this method of attempting to understand the world around us has resulted in critical pieces of work in all realms of science—the reduction of genes to amino acids, the reduction of chemical reaction syntheses to movement of individual electrons, and, perhaps the most noteworthy example, the development of classical mechanics.

Focusing on the latter, one of the truly extraordinary results in classical mechanics has been that the reductionist representation wholly captures the original incomprehensible process: the mathematical description of an apple falling from a tree is completely sufficient to describe and predict the movement of Jupiter. Following from this description, it is not an over-exaggeration to claim that most (or all) of the engineered feats of our modern world, from electricity to the building of the architecturally impressive to modern computing, are conceptually a result of a rigorous study of extremely simplified yet representative systems. Thus, because of the empirical success of the reductionist approach, the use of this methodology in studying natural processes has reached near unanimity in science.

What happens, however, when the behavior of each individual part of the system does not mirror the whole? When the behavior of the gene deviates from the behavioral aggregate of each amino acid comprising the gene; when an examination of the physical chemistry of a single electron is insufficient to describe the mechanism of generating an organic molecule; when the study of a single apple no longer scales to the study of planets? In fact, even in the most rigorously explored areas of science such as particle physics and statistical mechanics, the insufficiency of the reductionist approach to gain understanding has been well documented. As an example, ferromag-

netism, the mechanism by which objects achieve a magnetic state, was indescribable by classical mechanical approaches treating each atom within the material as a small magnet or magnetic dipole. It wasn't until Ernst Ising treated the reduction of the magnetic material not as a sum of individual atoms or magnetic dipoles, but as the individual atoms and their local environment was a sufficient mathematical description developed to comprehend and predict the behavior of magnetic systems.[1] Why was this addition so crucial for understanding ferromagnetism? In essence, this is due to the fact that the individual magnetic dipoles influence the dipoles in their immediate surroundings. It is not only the behavior of the individual dipole, but also, the interaction between one dipole and its neighbors that generates a reduced yet representative sub-system that is amenable to physical dissection (Fig 1.1). An even more fundamental, yet mysterious, example

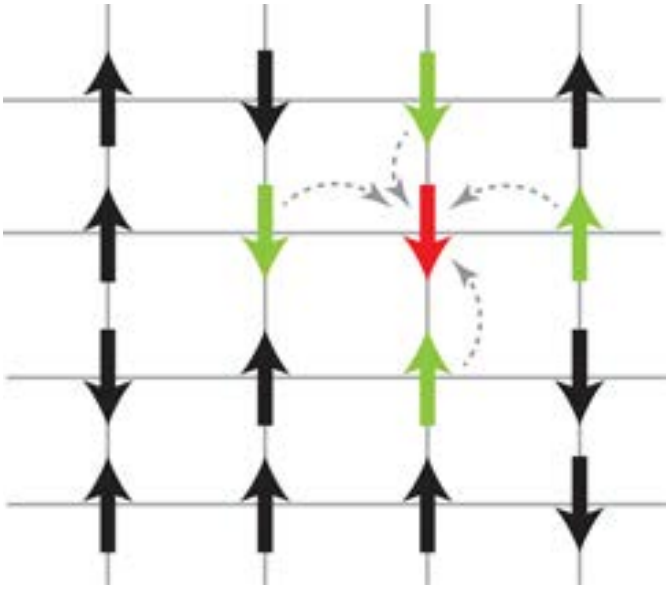

Figure 1.1: To describe the magnetic properties of materials, Ising represented the material as a set of spins that can be either up or down (shown as arrows here). With respect to understanding the behavior of the system, Ising realized the critical influence of the surrounding environment on a particular spin. For instance, the behavior of the spin in red is influenced, or is coupled to, its local environment (shown in green).

is encountered in measurements of quantum states of groups of particles—a process known as quantum entanglement. These particles exhibit what is known as non-local connectedness: the particles retain a connection made during their generation, specifically with respect to particle spin, regardless of external experimental parameters such as even how far apart in space the two particles are when the measurement is made (as Einstein described, 'spooky action at a distance').[2] Thus, in a similar way to ferromagnetism, quantum states are entangled in such a way as to render

a reduction to the individual particle useless. Using the above examples as case studies, we gain a clearer sense of a central characteristic of complex systems: the summed behavior of each part within a system does not equal the behavior of the system as a whole. In other words, the behavior of the whole system is emergent, arising from the collective behavior of its constituent parts.

As a biological example of this very characteristic, we need not look farther than one of the most fundamental discoveries in biology: the second protein structure ever seen, hemoglobin. It had been known since the early 1900's from Christian Bohr that hemoglobin was a critical biomolecule needed for survival of living beings due to its oxygen carrying capacity.[3] In his seminal studies on the statistical mechanics of oxygen binding to hemoglobin, Christian Bohr was able to show that as the partial pressure of oxygen increases, the propensity for hemoglobin to bind oxygen increases as well, but in a non-linear fashion, until hemoglobin is saturated at which point no more oxygen will bind to the protein(Fig 1.2). This non-linear behavioral dependence on reaction substrate became

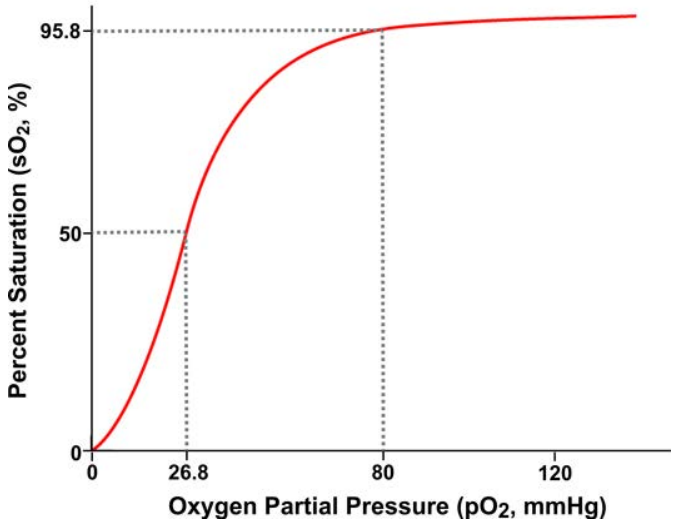

Figure 1.2: The relationship between saturation of hemoglobin and partial pressure of oxygen is measured to be nonlinear, illustrating the highly cooperative characteristic of oxygen binding to hemoglobin.

known in biochemistry as 'cooperativity'. Gilbert Adair subsequently discovered that hemoglobin is in fact composed of four subunits (and hence is termed a 'tetramer') and proposed a model where each hemoglobin subunit binds oxygen with sequentially increased affinity for each molecule.[4] Such a model would nicely explain the non-linear dependence seen in Figure 1.2. The chemist Linus Pauling took this thinking one step further, putting forth a mechanism by which oxygen affinity

Figure 1.3: Hemoglobin is a tetrameric protein, composed of four subunits capable of binding oxygen. The propensity of one subunit to bind oxygen is low. However, in the background of a bound subunit, the propensity for further binding of oxygen increases in a non-linear fashion as shown in Fig 1.2.

could increase in the background of bound oxygen through neighboring heme-heme interactions.[5] After the first protein structure had been solved by x-ray crystallography (myoglobin by John Kendrew), naturally the focus was to understand how a protein could exhibit such non-linear behavior and why nature would design protein in such a way. From the structure of Max Perutz, the basis for oxygen binding was rationalized and from further functional studies, the physiological role for hemoglobin was deduced.[6–8] However, though incredibly useful for providing insight into the function of hemoglobin, the protein structure—an atomistic snapshot of the protein—was insufficient to mechanistically understand the origins of arguably the most crucial characteristic that allows hemoglobin to support life—cooperativity—and exactly how binding a single oxygen is able to cause a holistic change in a protein characteristic. Going back to our understanding of what comprises a complex system, the reason for this is a fundamental one: biological properties emerge from the collective behavior of the parts within a system. In the case of hemoglobin, the system can be considered the protein and the parts amino acids comprising the protein. The cooperative event of binding oxygen and transmitting this information across the protein, must therefore be governed by interactions between amino acids in a directional, anisotropic fashion—a quality that can be exposed by neither the spatial distribution of amino acids within the protein structure nor any targeted perturbation experiments (such as mutations) to the system. As Feynman duly noted in what has become 'The Feynman Lectures' quoted by many a prominent scientist(s)

> "One of the great triumphs in recent times (since 1960), was at last to discover the exact spatial atomic arrangement of certain proteins... Over a thousand atoms... have been located in a complex pattern in two proteins. The first was hemoglobin. One of the sad aspects of this discovery is that we cannot see anything from the pattern; we do not understand why it works the way it does. Of course, that is the next problem to be attacked".

The nature of cooperativity, that is the cooperative action of many parts within a system generating an emergent property, is encountered at every scale of biology, ranging from the interactions

of genes to specify normal (such as height and metabolic processes)[9] and disease (chronic diseases of multifactorial origin such as cancers, diabetes, and neurodegenerative disease)[10] phenotypes, to the collective dynamics of species within an ecosystem (the collective behavior of ants to form an ant colony for example).[11] It is therefore crucial to not only acknowledge the presence of complexity in biology and our subsequent lack of comprehension with respect to the origins and mechanisms of biological phenomenon, but implement a strategy for what can be thought of as 'relevant reductionism'—a decomposition of a system into a relevant representation of the parts engaged in collective behaviors that give rise to the defining biological characteristics of the system as a whole.

1.2 Decomposing Complex Systems: A Classical Approach

In the most simplistic sense, systems are naturally decomposed by the intuition of the experimenter. To study the behavior of ants for instance, the most parsimonious choice of reductionism would be a single ant.[11] Following this, the decomposition of a complex system, where interactions are important for the global behavior of the whole, is often non-obvious and unintuitive to the human eye. How then can such systems be decomposed?

Proteins, often described as molecular machines within the cell, are examples of complex biological materials. Proteins are comprised of strings of elementary biological parts known as amino acids and from the interactions between these amino acids arise protein form, or fold, and protein function (Fig. 1.4).[12] Structural and biochemical studies of proteins (including hemoglobin as mentioned above) have clearly demonstrated the non-linear contribution of a particular amino acid to the energetics of the protein. More concretely, it was observed that perturbations (known as 'mutations') to amino acids comprising proteins exhibit a strong heterogeneous response, with mutations at certain amino acids egregiously affecting protein function while mutations at other positions retaining the functional abilities of the protein.[13–18] Thus, one way to identify the relevant parts of a protein would be to perform a systemic mutational scan of a protein, noting the positions that encode the information necessary for protein function. However, similar to the classical approach of understanding ferromagnetism, this approach fails to consider potentially important interactions or couplings between amino acids. Borrowing from genetics approaches, Alan Fersht

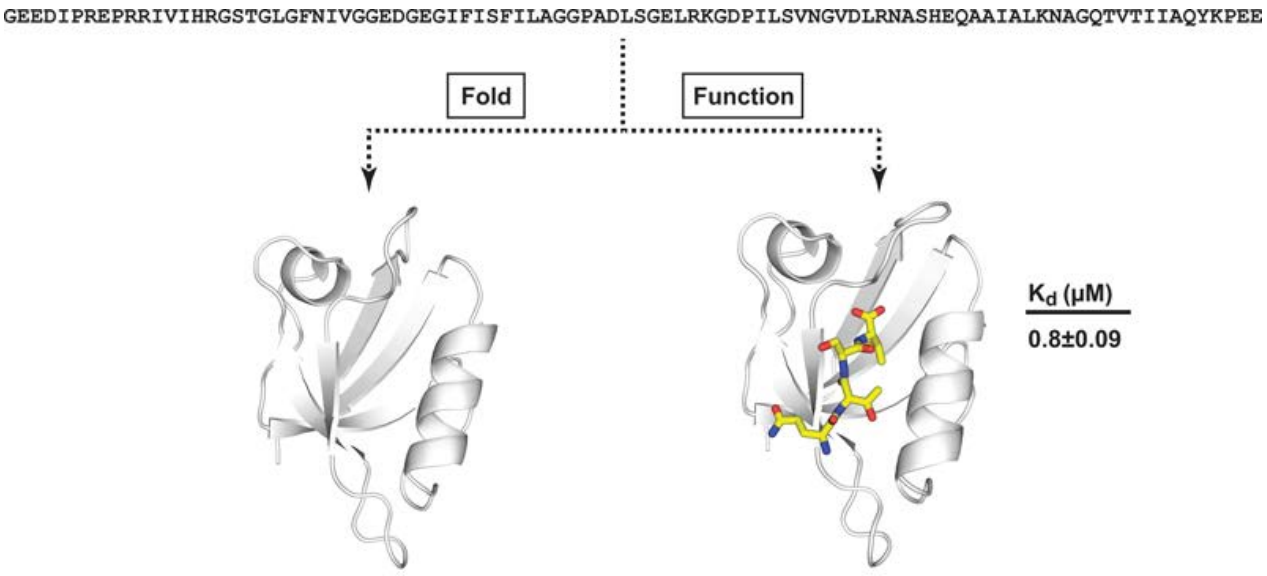

Figure 1.4: The primary sequence of a protein is composed of a string amino acids with a possibility of 20 differenct amino acids at each site. This sequence contains the information to specify a protein's fold and its function. In this figure, a particular protein, $PSD95^{pdz3}$ is shown in its native fold and bound to a peptide with high affinity.

therefore proposed a formalism known as thermodynamic mutant cycles to biochemically reveal the energetic couplings within a protein (Fig 1.5).[19,20] For example, consider attempting to measure

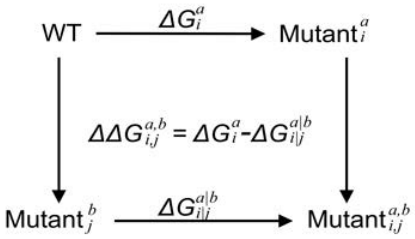

Figure 1.5: Thermodynamic mutant cycles provide a useful construct to detect couplings within a system given a set of perturbations. The coupling energy between mutant 1 and mutant 2 is simply the difference between the effect of mutant 1 and mutant 1 in the background of mutant 2. If this difference is zero, then mutant 1 and 2 are thermodynamically independent of each other; if non-zero, then the two mutations are coupled.

the energetic coupling between two mutations, a and b, at positions i and j (Fig. 1.5)). Mutation a at position i has a measurable effect ΔG_i^a upon binding a ligand for instance. Now, compare this effect to mutation a at position i but in the background of mutation b at position j, $\Delta G_{i|j}^{b,a}$. If the two effects are identical, then mutation b at position j had no influence on the effect of mutating position i and the two mutations are considered as additive (the energetic contribution of the double mutant is exactly equal to a sum of each single mutant). However, if there is a difference in the two effects, then we claim that mutation a at position i and mutation b at

position j are coupled and accordingly, we measure a coupling energy between the two mutations that adds to each mutation's individual energetic contribution. As an aside (but nevertheless important), within this formalism, it is imperative that the readout, such as free energy, is a thermodynamic state function meaning the energy difference between the beginning and end of a path is independent of the path taken. We can see from this example how such an approach could reveal the complexity of a biological system: two thermodynamically coupled mutations would be considered to be engaged in a complex interaction. From a practical standpoint however, there exists a significant limitation to a systematic scan of thermodynamic mutant cycles across an entire system. To measure the pairwise coupling of just two mutations in a protein requires four measurements in the lab: native protein (herein referred to as wild-type protein) bound to ligand, mutant protein a,i bound to ligand, mutant protein b,j bound to ligand, and the double mutant bound to ligand. For the three-way coupling between three mutations would thus require 8 measurements; the four-way coupling between four mutations would require 16 measurements and so on. For simplicity, even if only pairwise couplings are considered, the number of necessary measurements is prohibitive given a protein comprising of just one-hundred amino acids. We thus end this section where we started: how can a complex system be decomposed?

1.3 Decomposing Complex Systems: An Evolutionary Approach

All biological systems have been exposed to the evolutionary process: repeated iterations of mutation and selection over billions of years. While the human eye has historically failed to identify, much less understand, the relevant interactions within biological systems, the evolutionary process potentially provides clues as to the relevant constraints on biological systems. Consider a phylogenetic tree from which biological matter has emerged from a common ancestor. If a particular trait is important for function and survival, this trait will be conserved across organisms within the tree; similarly, if an interaction between two parts of a system is important for function and survival, this interaction will be conserved across organisms within the tree. As an extension of this concept if a part, or interaction between parts, is unimportant for survival, then evolution

would release the constraint on this part, allowing it to float or vary within the population. Thus, the identification of parts engaged in collective behaviors seems possible using the principle of conserved correlation given a representative ensemble.

Proteins can be grouped into protein families based on their defining characteristics and ancestral history, despite potential divergence in amino acid identity. For example, the globin family of proteins display less than 20% sequence identity overall while sharing a common fold and conserved function, binding of heme moiety and oxygen transport.[21] The origins of sequence divergence can be attributed to a variety of factors such as the selective pressures encountered within the idiosyncratic eco-niche of any particular protein or, perhaps a more significant cause of low sequence identity, neutral drift. As for the former, while each protein has presumably evolved from a common ancestor, there exist an array of environments that proteins within any family have faced. Thus, adaptation to any particular environment may be reflected in variation beyond the conserved interactions of the family. However, a far more likely determinant of sequence divergence is that most amino acids contribute in a minimal way to the fitness of the protein. To date, there are few systematic studies that reveal the relevant amino acids in a protein but the few that do exist demonstrate a consistent pattern: many mutations in the protein are neutral, not affecting primary function, with only a subset of residues encoding functional information about the protein.[16–18] Consistent with this observation are experiments demonstrating the sufficiency of heterologous transplantation to support organismal growth: the transplantation of a homologous protein in place of an organism's endogenous protein has been shown to confer only a minimal fitness effect in the organism.[22,23]

Thus using a multiple sequence alignment of a protein family—a construct that contains members of a protein family aligned such that direct comparison of amino acid identity can be performed with ease—as a representative ensemble of realizations, conserved interactions between can be statistically identified (a method known as 'Statistical Coupling Analysis' or 'SCA').[24,25] Consider three positions within an alignment of a protein family, i, j, k (Fig 1.6). If position i is important for function across the protein family, then this position will be conserved, with particular amino acids present well over the statistical background. If the interaction between position i and j is important for function, then not only will these two residues be conserved, but

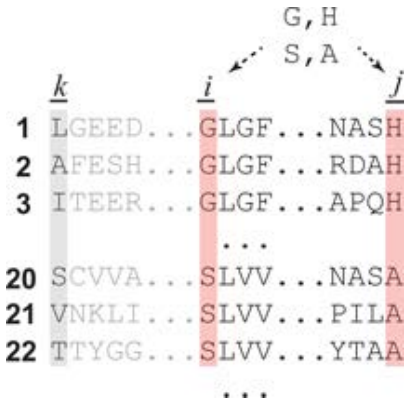

Figure 1.6: Amino acid sequences of members within a protein family are aligned so as to facilitate direct comparison of amino acid content between sequences. Positions i and j are conserved, accepting either a Gly/Ser or His/Ala respectively. Additionally, when Gly_i is mutated to Ser, His_j is mutated to Ala. If the mutual presence of $\mathrm{Gly_iHis_j}$ and $\mathrm{Ser_iAla_j}$ deviates significantly from background expectation then we consider the two positions correlated to each other.

the identity of amino acids at these positions will be correlated to each other. That is, statistically when amino acid a is at position i, amino acid b will likely be at position j. Equivalently, if a variant is found at position i, a correlated variant is found at position j. If a position is unimportant for function, for instance position k, we expect the frequency distribution of amino acids to relax to the background distribution of amino acids found in the entirety of the protein database. In essence, this means there is no evolutionary pressure to retain amino acid identity at this position. Mathematically, the deviation of the frequency distribution of a position from background as generated by neutral drift is given as

$$D_i^a = f_i^a \ln \frac{f_i^a}{q^a} + (1 - f_i^a) \ln \frac{1 - f_i^a}{1 - q^a} \tag{1.1}$$

where q^a is the background frequency of amino acid a as found over the entire protein database and f_i^a is the frequency of amino acid a at position i. Pairwise conserved correlations between positions can be represented in matrix form as

$$\widetilde{C}_{ij}^{ab} = \phi_i^a \phi_j^b (f_{ij}^{ab} - f_i^a f_j^b) \tag{1.2}$$

where ϕ_i^a and ϕ_j^b are conservation weights calculate for amino acid a at position i and amino acid b at position j and f_{ij}^{ab} is the joint frequency of finding amino acid a at position i and amino acid

b at position j.

$$\phi_i^a = \left| \frac{\partial D_i^a}{\partial f_i^a} \right| = \left| \ln \left[\frac{f_i^a(1-q^a)}{(1-f_i^a)q^a} \right] \right| \tag{1.3}$$

This four dimensional correlation tensor is compressed into a two dimensional position by position correlation matrix where the diagonal, $\widetilde{C}_{ii}$ is representative of the intrinsic conservation of position i and the off-diagonal terms illustrate the covariance (or coevolution since this signal is measured over the evolutionary record of the protein family) of two positions i and j. As an example, the conservation weighted correlation matrix (also referred to as the SCA matrix) is shown for a family of small, 100 amino acid binding domains known as PDZ domains (Fig. **??**).

The alignment contains approximately 1000 PDZ sequences spanning a variety of species thereby reflecting adequate sequence divergence to detect core features present across all PDZ domains without bias. We find that the SCA matrix holds a few key features, that lend intuition to the problem. First, the matrix is sparse with a majority of positions evolving independently of each other and by extension, only a few positions in the protein exhibit a siginficant correlation to ther positions. Second, the pattern of correlations appears to be heterogeneous, spread throughout the primary sequence of the protein with no obvious pattern. Decomposition of the SCA matrix reveals a network of correlated residues in the protein structure. For the PDZ family, this network of coevolving residues is shown on a single family member, $\mathrm{PSD95^{pdz3}}$, the third PDZ domain of a neuronal protein found in rats whose function is to bind the 9-amino acid C-terminus of a protein known as CRIPT (cysteine-rich interactor of PDZ).[26] For this particular family, the sector is composed of approximately 20% of the protein spanning the protein core, a subset of residues lining the binding pocket, and extending to surface exposed sites on the three-dimensional structure of $\mathrm{PSD95^{pdz3}}$. Topologically, the sector does not fall under known the description of known secondary structural elements such as helices and beta-sheets. Instead, what is seen is an unintuitive decomposition of the protein into a group of amino acids that coevolve with each other and form a connected network in the three-dimensional structure, despite exhibiting no obvious spatial pattern in the primary structure. Extending this analysis to other protein families, indeed we see that the existence of sectors is general with all proteins examined to-date showing a subset (near 20%) of coevolving residues within a plurality of independently evolving residues.[27]

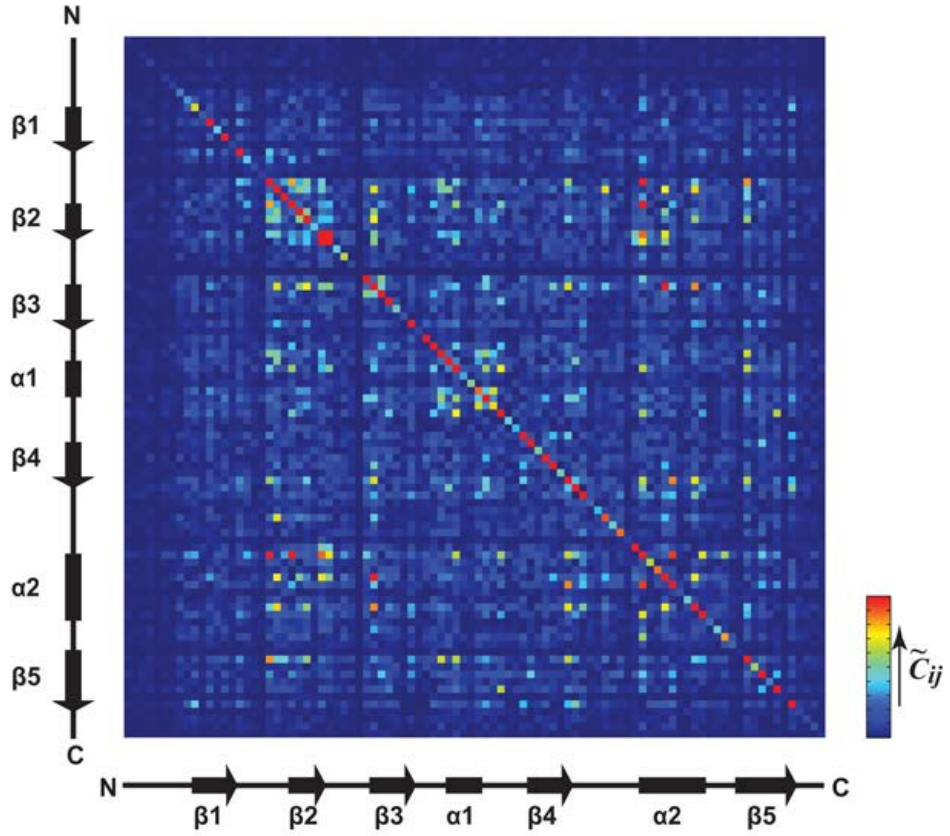

Figure 1.7: The SCA matrix, a position by position (N to C terminus) conservation weighted correlation matrix, is shown for the family of PDZ domains. We see that most positions evolve independently of each other (blue pixels indicating low SCA scores) with only a few off diagonal pixels showing significant coevolution.

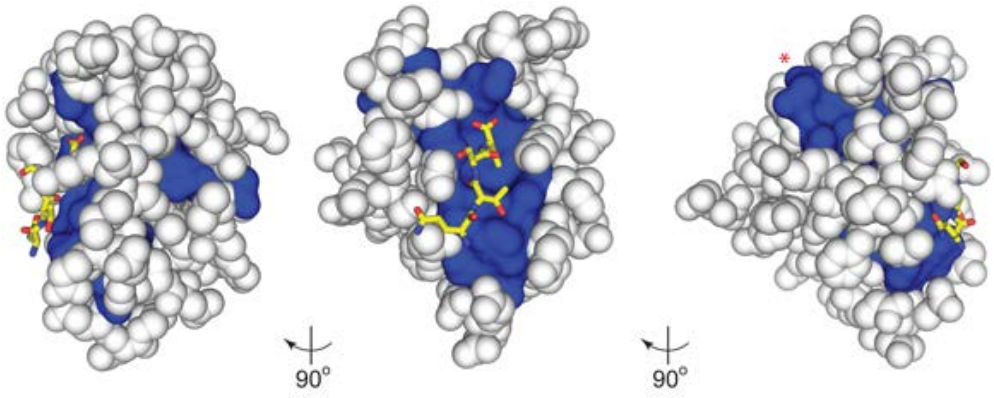

Figure 1.8: Decomposition of the SCA matrix (Fig. 1.7) reveals a network of connected coevolving residues in the protein structure termed the protein 'sector' (shown in blue). Here the sector for the PDZ family is shown on a particular member of the family, $PSD95^{pdz3}$. The sector spans the binding pocket of the domain (with ligand shown in yellow stick bonds) and extends to functionally relevant surface sites (shown in the rotation views). For example, the asterisked sector position is known to be an allosteric regulatory region of the Par6 PDZ domain.

So, using evolution as a guide, the approach of statistical coupling analysis is able to yield a transformed 'view' of the parts that comprise the protein. Revisiting the original motivation

of understanding complex systems, we have canonically described proteins as being composed of amino acids but found that understanding proteins at the level of individual amino acids may not be sufficient to explain the emergent phenomenon of function that arises from the collective behavior of amino acids. SCA thus provides a reduction of the problem, simplifying the system from what was originally for instance 100 amino acids for the PDZ domain, into non-sector and sector residues. However, it remains to be seen if this is a *relevant* reduction. In other words, can we explain fundamental behaviors of the protein–fold, function, etc.–by studying the protein sector?

1.4 The Protein Sector: A relevant reduction?

The protein sector provides a nice framework to reduce the apparent complexity in sufficiently describing a protein. But first, it is necessary to answer the naive question: is it real? This question itself contains many sub-questions, some of which have been addressed previously in the Ranganathan Lab.

1.4.1 Thermodynamic Coupling

Recall that the protein sector is a statistical entity, identified by detecting statistically coupled residues across an ensemble. Therefore, it is not a given that statistical coupling in a multiple sequence alignment maps directly to energetic coupling, as measured by thermodynamic mutant cycles, in any single individual protein. To get a sense of the coupling energies within the protein sector, mutant cycles within $PSD95^{pdz3}$ were performed over a set of sector and non-sector positions, comparing the coupling energies measured by thermodynamic mutant cycles to the statistical coupling measured by SCA.[24] Because of limitations in experimental throughput, the energetic couplings in $PSD95^{pdz3}$ were measured in the background of wild-type protein and a particular mutation H372Y found to exhibit the highest statistical coupling out of any position in the protein. Consistent with the sector description, Lockless *et al.* found that positions within the sector, both proximal and distal to H372Y, were highly coupled to H372Y. In contrast, positions outside the sector showed much less energetic coupling to H372Y. Thus, as an initial study, a good correspondence between statistical coupling and energetic coupling was established.

1.4.2 Fold and Function

The amino acid sequence of a protein specifies its native fold and function in a physiological setting. Is the information in the SCA matrix sufficient to specify the fold and function of proteins? Work by Russ *et al.* and Socolich *et al.* showed for the case of the WW domain protein family that the couplings in the SCA matrix were indeed adequate to generate synthetically designed proteins (built from a Monte-Carlo algorithm) that achieved a native-like fold and wild-type like function.[28,29] Similarly, Gosal *et al.* demonstrated that synthetic PDZ domains that retained the coupling information from the SCA matrix were functional and retained the wild-type like binding specificity preference.[30] Interestingly, these proteins, generated from Monte-Carlo heating trajectories using wild-type $PSD95^{pdz3}$, were shown to be unstable in the unliganded form but became stable when bound to native binding partner. Efforts are currently underway to solve the structure of a synthetically designed PDZ domain using X-ray crystallography. While the tests performed on the WW and PDZ domains demonstrated the sufficiency of correlations, they also addressed the necessity of placing correlations into designed sequences. Sequences that were designed solely satisfying the primary statistical constraint of conservation failed to fold and function.

Limited mutagenesis studies have shown a relationship between sector positions and mutations of functional relevance in proteins (see Halabi *et al.*) but a comprehensive single mutational scan investigating this relationship, performed by McLaughlin *et al.*, provided the first truly global picture of the functionally relevant amino acids of a protein.[17] Using $PSD95^{pdz3}$ as a model system, McLaughlin *et al.* created a quantitative high-throughput assay using a bacterial-2-hybrid system coupled to next-generation sequencing to assay all possible single mutations to the protein. McLaughlin *et al.* found that while a majority of conserved residues were functionally relevant (as is to be expected as previously discussed), the protein sector was a more accurate descriptor of the functionally relevant amino acids. It is important to note that because the sector is mathematically defined as correlations between conserved residues, the effect of conservation is naturally within the protein sector, making dichotomies such as the properties of conserved residues versus sector residues almost artificial. Previous work investigating the biological role

of the protein sector has thus shown that the information to encode protein fold and function is contained within the protein sector.

1.4.3 The Remaining Characteristics

If our goal is to attain a 'relevant' reduction of the system, what biological properties, then, remain to be accounted for? This thesis will focus on exploring the relationship between the structural decomposition of proteins as discovered by Lockless *et al.* (and expounded upon by Halabi *et al.*) and two fundamental biological characteristics of proteins: 1) The ability to adapt quickly to new functions and 2) The structural property of long range signal transmission, termed 'allostery'. Not only is an understanding of the structural origins of these properties lacking, but in order for the sector reduction to be a representative one, it is necessary that such a model can explain these critical characteristics of proteins. If true, the sector representation could then be considered a truly appropriate description for understanding the properties of proteins.

References

[1] B. Cipra, "An Introduction to the Ising Model," *The American Mathematical Monthly*, vol. 94, no. 10, p. 937, 1987.

[2] G. Bacciagaluppi and A. Valentini, "Quantum Theory at the Crossroads," *ArXiv*, 2009.

[3] C. Bohr, K. Hasselbalch, and A. Krogh, "Ueber einen in biologischer Beziehung wichtigen Einfluss, den die Kohlensäurespannung des Blutes auf dessen Sauerstoffbindung übt," *Skandinavisches Archiv für Physiologie*, vol. 16, no. 2, pp. 402–412, 1904.

[4] G. Adair, "A Critical Study of the Direct Method of Measuring the Osmotic Pressure of Haemoglobin," *Proceedings Royal Society of London A*, vol. 108, no. 748, pp. 627–637, 1925.

[5] L. Pauling, "The Oxygen Equilibrium of Hemoglobin and Its Structural Interpretation.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 21, no. 4, pp. 186–191, 1935.

[6] M. Perutz, W. Bolton, R. Diamond, H. Muirhead, and H. Watson, "Structure of Haemoglobin: An X-ray Examination of Reduced Horse Haemoglobin," *Nature*, vol. 201, no. 4946, pp. 1212–1213, 1964.

[7] M. F. Perutz, "Stereochemistry of cooperative effects in haemoglobin.," *Nature*, vol. 228, no. 5273, pp. 726–739, 1970.

[8] M. F. Perutz, a. J. Wilkinson, M. Paoli, and G. G. Dodson, "The stereochemical mechanism of the cooperative effects in hemoglobin revisited.," *Annual review of biophysics and biomolecular structure*, vol. 27, pp. 1–34, 1998.

[9] A. R. Wood, T. Esko, J. Yang, S. Vedantam, T. H. Pers, ..., and T. M. Frayling, "Defining the role of common variation in the genomic and biological architecture of adult human height," *Nature Genetics*, vol. In Press, no. 11, 2014.

[10] H.-Y. Chuang, M. Hofree, and T. Ideker, "A decade of systems biology.," *Annual review of cell and developmental biology*, vol. 26, pp. 721–744, 2010.

[11] D. M. Gordon, "The Ecology of Collective Behavior," *PLoS Biology*, vol. 12, no. 3, pp. 1–4, 2014.

[12] C. Anfinsen, "Principles that Govern the Folding of Protein Chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[13] T. Clackson and J. Wells, "A Hot Spot of Binding Energy in a Hormone-Receptor Interface," *Science*, vol. 267, no. 5196, pp. 383–386, 1995.

[14] K. a. Brown, 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*, vol. 90, no. 24, pp. 11753–11756, 1993.

[15] J. J. Perona, L. Hedstrom, W. J. Rutter, and R. J. Fletterick, "Structural origins of substrate discrimination in trypsin and chymotrypsin.," *Biochemistry*, vol. 34, no. 5, pp. 1489–1499, 1995.

[16] D. M. Fowler, C. L. Araya, S. J. Fleishman, E. H. Kellogg, J. J. Stephany, D. Baker, and S. Fields, "High-resolution mapping of protein sequence-function relationships.," *Nature methods*, vol. 7, pp. 741–6, Sept. 2010.

[17] R. N. McLaughlin, F. J. Poelwijk, A. Raman, W. S. Gosal, and R. Ranganathan, "The spatial architecture of protein function and adaptation.," *Nature*, vol. 491, pp. 138–42, Nov. 2012.

[18] M. A. Stiffler, D. R. Hekstra, and R. Ranganathan, "Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase," *Cell*, vol. 160, no. 5, pp. 882–892, 2015.

[19] A. Horovitz, "Double-mutant cycles: a powerful tool for analyzing protein structure and function," *Folding and Design*, vol. 1, no. 6, pp. R121–6, 1996.

[20] P. J. Carter, G. Winter, a. J. Wilkinson, and a. R. Fersht, "The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus).," *Cell*, vol. 38, no. 3, pp. 835–840, 1984.

[21] D. Bashford, C. Chothia, and A. M. Lesk, "Unique Features of the Globin Amino Acid Sequences Unique Features of the Globin Amino Acid Sequences," *Molecular Biology*, pp. 199–216, 1987.

[22] W. B. Frommer and O. Ninnemann, "Heterologous Expression of Genes in Bacterial, Fungal, Animal, and Plant Cells," *Annual Review of Plant Physiology and Plant Molecular Biology*, vol. 46, no. 1, pp. 419–444, 1995.

[23] E. Wyckoff 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*, vol. 85, no. 17, pp. 6272–6276, 1988.

[24] S. W. Lockless and R. Ranganathan, "Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families," *Science*, vol. 286, no. October, pp. 295–299, 1999.

[25] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, "Protein sectors: evolutionary units of three-dimensional structure.," *Cell*, vol. 138, pp. 774–86, Aug. 2009.

[26] M. Niethammer, J. G. Valtschanoff, T. M. Kapoor, D. W. Allison, R. J. Weinberg, A. M. Craig, M. Sheng, C. Hill, and N. Carolina, "CRIPT , a Novel Postsynaptic Protein that Binds to the Third PDZ Domain of PSD-95 / SAP90," *Neuron*, vol. 20, pp. 693–707, 1998.

[27] G. M. Süel, S. W. Lockless, M. a. Wall, and R. Ranganathan, "Evolutionarily conserved networks of residues mediate allosteric communication in proteins.," *Nature Structural Biology*, vol. 10, pp. 59–69, Jan. 2003.

[28] W. P. Russ, D. M. Lowery, P. Mishra, M. B. Yaffe, and R. Ranganathan, "Natural-like function in artificial WW domains.," *Nature*, vol. 437, pp. 579–83, Sept. 2005.

[29] M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan, "Evolutionary information for specifying a protein fold.," *Nature*, vol. 437, no. 7058, pp. 512–518, 2005.

[30] W. S. Gosal and R. Ranganathan, "Form follows function : Parsing the evolutionary rules for protein folding and function," *In Prep.*

Chapter 2

The Structural Principles of Protein Adaptation

2.1 The Adaptive Capacity of Proteins

All biological systems are subject to the forces of evolution. These systems are designed to function in their current environment but importantly must be able to adapt given an environmental change to survive the evolutionary process of mutation and selection.[1] Proteins are no different. Traditionally, the properties of protein fold and function have been evaluated in the context of a single environment or fitness pressure. If this fitness pressure changes, however, the protein has to evolve to the new challenge without suffering a fitness loss so as to render the protein dead.[1–10] In other words, proteins must be designed with the ability to adapt to new functions in a small number of sequence changes to be fit in a changing world. It is therefore understandable that this constraint is in fact quite crucial, posing a non-trivial target criteria to be satisfied when designing a natural protein.

Many have attempted to understand how to rationally change the function of a protein.[2,8,9] Consider the case of the serine protease family, two members of which are chymotrypsin and trypsin.[9] These proteins share a common tertiary structure yet differ slightly in substrate preference. Chymotrypsin activity is specific for large hydrophobic residues whereas trypsin cleaves peptides at Arginine or Lysine residues. According to Hedstrom *et al.*, the primary catalytic site sequence difference between the two proteins is at position 189, where trypsin has an aspartic acid and chymotrypsin has a serine. Interestingly however, making the D189S substitution does not switch the specificity of trypsin to chymotrypsin, but rather creates a functionally poor enzyme and would therefore be an unviable evolutionary intermediate (Fig. 2.1).[9] Hedstrom *et al.* were thus able to show that simply using the active site to rationalize adaptive mutations is wholly insufficient and incompatible with the evolutionary process. The specificity switch between the two proteins was eventually achieved by mutating two loops of residues removed from the binding pocket that exhibited sequence divergence between trypsin and chymotrypsin (Fig. 2.2).

As many studies, using a variety of enzymes, approached this very problem of engineering new or altered function, the lack of understanding with respect to the relationship between sequence and evolvability became increasingly clear. Many attempts at rationally directing protein function, mostly involving insight gleaned from the three-dimensional structure of the protein,

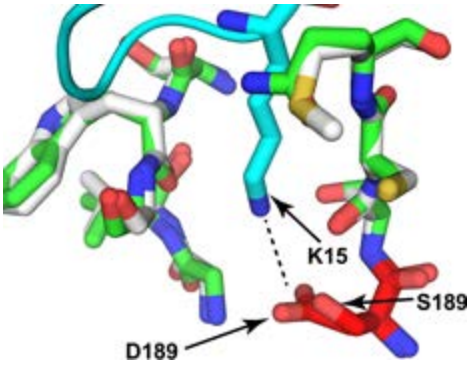

Figure 2.1: Apart from subtle differences in the active site of chymotrypsin (white) and trypsin (green), the major sequence difference between the two structures is at position 189 (highlighted in red), in contact with the major determinant of substrate specificity (lysine (shown in blue) or arginine for trypsin and bulky hydrophobic residues for chymotrypsin). Hedstrom *et al.* showed that making the D189S substitution in fact created an enzyme with poor catalysis and failed to aid in altering the specificity from trypsin to chymotrypsin.

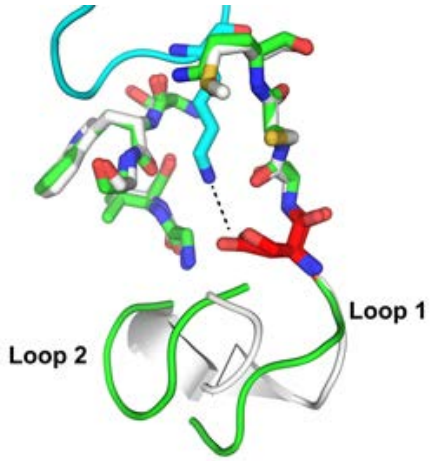

Figure 2.2: Two loops, in direct connection with the active site, control the successful specificity switch from tryptic to chymotryptic activity in the background of D189S. It is important to note that these loops were originally considered not as important for specificity as the active site and are spatially non-obvious mutations to generate given the three-dimensional structures of the two proteins. With heightened interest in understanding the rules of engineering novel function in proteins, the observation of long-range influences in catalytic activity, binding specificity, and overall function has become increasingly frequent.

have resulted in insufficient or inadequate functional alterations, either proceeding through non-functional intermediates (such as the example of trypsin to chymotrypsin) or altogether failing and resorting to more arbitrary mutational strategies.[2,8,9,11] Indeed, the obvious lack of intuition for the problem has resulted in the advent of an increasingly popular experimental strategy known as 'directed evolution'–starting with a natural seed sequence and evolving to new function through mimicking the evolutionary process through repeated rounds of mutagenesis and selection.[2] With the current technical capacity for sampling sequence space (complete single mutation scans and,

in some cases, nearly complete double mutation scans are possible)[10,12–14] and structure determination, it is natural to ask why is converting function, for something as seemingly simple as trypsin to chymotrypsin, so unintuitive and difficult? After all, nature clearly has been able to find a solution. The problem is perhaps not a technical one after all, but more fundamental. When attempting to understand the evolutionary drift of hemoglobin across species, Linus Pauling and Emile Zuckerkandl remarked:

> Perhaps the most important consideration is the following: There is no reason to expect that the extent of functional change in a polypeptide chain is proportional to the number of amino acids 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. Of course, the two aspects are not unrelated, since the functional effect of a given single substitution will frequently depend on the presence or absence of a number of other substitutions.

At the essence of this statement are two crucial insights: 1) The mapping between phenotype and sequence is extremely heterogeneous and non-linear. Crystal structures were initially thought to hold great potential in clarifying this relationship, however to a large extent one could argue that the mapping between phenotype and structure is equally as nebulous. 2) The effect of a given mutation is often context dependent, contingent on the genetic background in which the mutation is made. The latter point harkens back to the discussion of thermodynamic mutant cycles and complex interactions. Phenotype, and by extension the ability to adapt to new phenotype, exists at the level of amino acid interactions whose relevance, by nature, is not apparent at the level of protein structure. The tryptic to chymotryptic specificity switch is an example of the role of coupling with respect to adaptation: the D189S mutation was adaptive only in the genetic background of an altered distant loop. Thus, here we seek to identify, and more generally understand, the sequence determinants of adaptability and how these mutations relate to the design of natural proteins, described as a network of sparse, spatially distributed coupled positions–the sector–amongst a majority of thermodynamically independent, non-sector positions.

2.2 The Approach: A Comprehensive Single Mutation Scan of a Model Protein

To understand the adaptive process in natural proteins, we choose a model system that contains the canonical properties of proteins (with secondary structural elements that fold into a well-packed three-dimensional structure forming a hydrophobic core with solvent exposed surfaces and also performs a physiological function) but has the advantage of experimental facility with regards to biochemical and structural measurements. We chose the third PDZ domain of a protein, Post-Synaptic Density Discs Large Zona Occludens 95 ($PSD95^{pdz3}$). Generally, PDZ domains in proteins serve to promote protein-protein interactions and act as subcellular 'scaffolds', bringing together many cellular proteins thereby increasing the local concentration of particular proteins to facilitate cellular reactions.[15]

2.2.1 Structure and Biochemical Specificity of $PSD95^{pdz3}$

PDZ domains generally bind the C-termini of proteins.[15–17] Specifically the third PDZ domain of PSD95, chosen for study because of the extensive history of its use in the lab, binds the C-terminal 9 amino acids of a postsynaptic protein named CRIPT (Fig. 2.3).[18] The protein is constructed of four beta strands and two alpha helices with the binding cleft between β_2 strand and α_2 helix. The nine amino-acid peptide is situated, N to C terminus, in an antiparallel fashion in the binding pocket, with the C-terminus engaged in extensive hydrogen bonding interactions with the backbone of residues 323 through 325 in the protein, an area known as the carboxylate binding loop (CBL) containing the conserved GLGF motif in PDZ domains. PDZ domain specificity is determined by the four C-terminal amino acids of its binding partner. In general, PDZ domains are grouped into specificity classes based on these four amino acids with a majority of the specificity being driven by two positions on the ligand known as the -2 and 0 positions.[16] Class I PDZ ligands contain a sequence motif of **-X-T/S-X-**Φ where '**X**' is any amino acid and Φ is a hydrophobic residue. The CRIPT sequence is **TKNYKQTSV** thus $PSD95^{pdz3}$, for instance, binds a typical class I peptide. In contrast the C-terminal sequence of class II peptides are **-X-**Φ**-X-**Φ. We use this difference in class designation to model an adaptive challenge to $PSD95^{pdz3}$. The protein natively binds a class

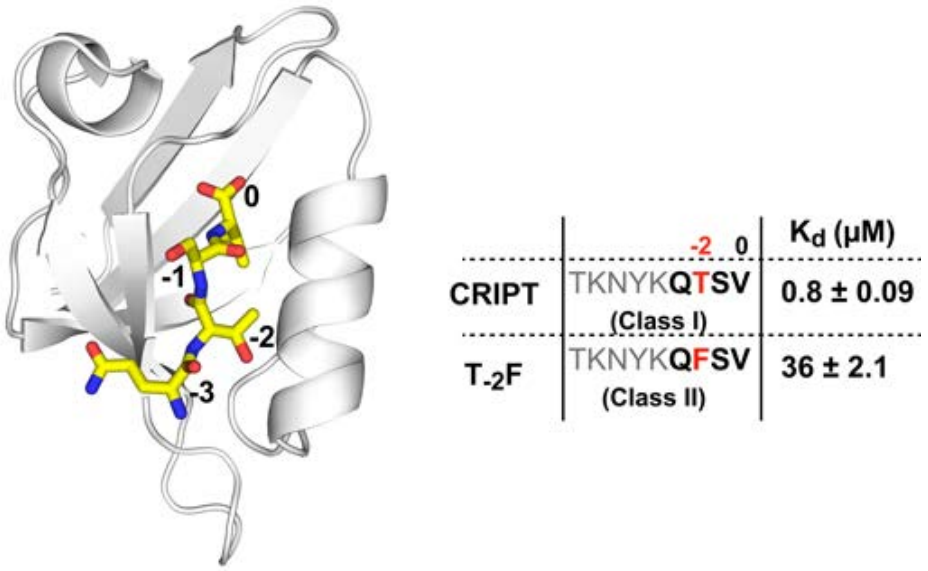

Figure 2.3: The third PDZ domain of PSD95 binds a nine amino acid peptide sequence derived from the CRIPT protein. The C-terminus of the peptide (the Val(0) position) engages in hydrogen bonding interactions with the carboxylate binding loop (CBL) of the protein. PSD95^{pdz3} binds CRIPT with a high affinity and exhibits a 45 fold specificity preference towards its native class I peptide over a synthetic, class II mutant peptide T$_{-2}$F.

I peptide in CRIPT; can the protein be evolved to bind a class II peptide that we define?

2.2.2 Adapting PSD95^{pdz3} to new function

We pose an adaptive challenge to PSD95^{pdz3} defined as binding a class II peptide with sequence **TKNYKQFSV** (T$_{-2}$F). We find that the protein binds the native peptide with 0.8 μM binding affinity, and the alternate class II peptide with 36 μM binding affinity thereby exhibiting a 45-fold preference for native peptide (Fig. 2.3).[10] To identify adaptive mutations in a copmrehensive manner, McLaughlin *et al.* developed a high-throughput quantitative bacterial-2-hybrid assay to measure the binding affinity of each single mutant of the protein against both CRIPT and T$_{-2}$F peptides. (The method for this technique will not be described here as it is fully explicated in McLaughlin *et al.*). The affinity data from this experiment is shown in Fig 2.4. To adapt the protein to the T$_{-2}$F peptide, we chose the maximally adaptive mutation towards T$_{-2}$F (the reddest pixel in the T$_{-2}$F mutant matrix), G330T. Using fluorescence polarization with a tetramethyl-rhodamine (TMR) labeled peptide, we found that this mutant is able to bind both peptides equally well despite being located away from the binding site, a result consistent with the role of

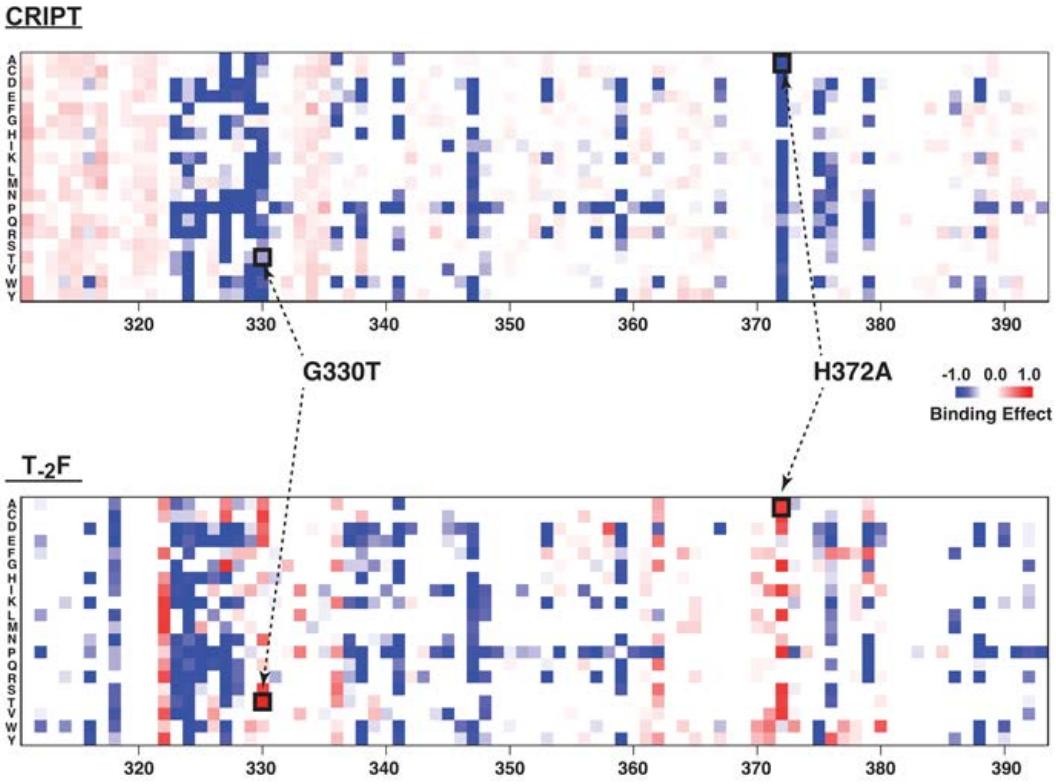

Figure 2.4: A library of all possible single mutants of $PSD95^{pdz3}$ was generated and assayed against binding to CRIPT and T-2F peptides by McLaughlin *et al.*. To adapt the protein to T-2F specificity, two mutations were chosen based on this data: G330T and H372A. G330T represents the highest gain of affinity mutation for the alternate peptide. After observing that G330T can bind both peptides equally well with high affinity, H372A was chosen to decrease affinity towards CRIPT peptide and thereby cause a specificity switch.

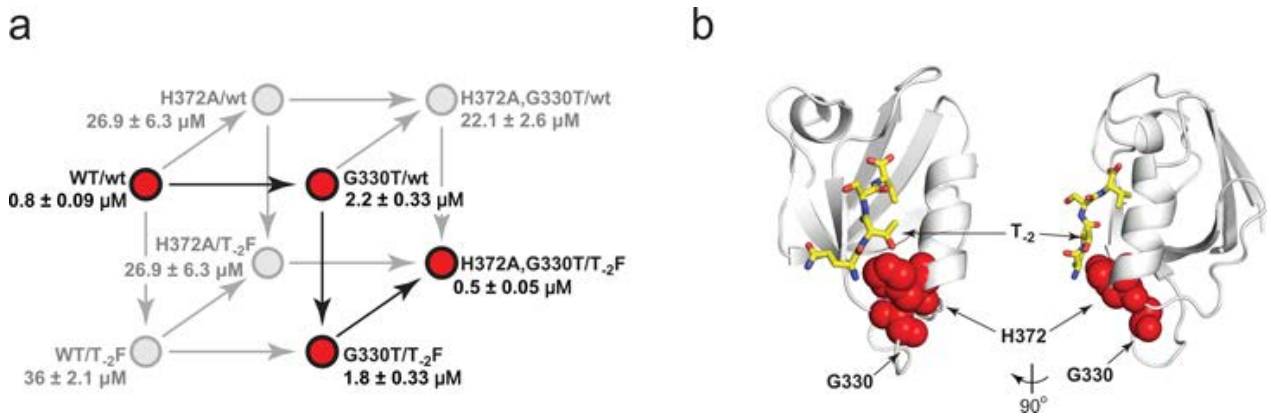

Figure 2.5: A. Of all possible paths from CRIPT specificity to $T_{-2}F$ specificity, as defined by the three mutations G330T, H372A, and $T_{-2}F$, only the G330T-first path retains a high affinity complex along adaptation to new function. If the H372A or $T_{-2}F$ mutations are made first, a significant decrease in affinity for the protein ligand complex is observed. Variants such as G330T are termed 'conditionally neutral': observed to be neutral dependent on condition of observation (in this case, the 'condition of observation' is the identity of ligand). Because mutants such as G330T are neutral for a wild-type environment but are also fit in alternate environments (such as, for instance, a hypothetical $T_{-2}F$ environment), the ability to access these mutations has been argued to be a critical characteristic of the adaptive capacity of proteins. **B.** The two mutation path is shown on the protein structure. While the H372A mutation is a structurally reasonable sequence variant, in direct contact with the T_{-2} position on the ligand, the G330T mutation is far less obvious, residing one contact shell away from the ligand. The ability of G330T to influence adaptation coupled with its structural location in the protein make this two step mutational path to new function an interesting and representative case study to deeply understand the adaptive process in natural proteins.

distant mutations in the adaptive process as demonstrated by numerous examples including the trypsin to chymotrypsin switch (Fig. 2.5a). Mutation H372A was picked next, predicted from the data in Figure (Rick figure), and subsequently shown via fluorescence polarization, to decrease binding to CRIPT peptide and maintain a high affinity for $T_{-2}F$ peptide. Just two mutations, H372A and G330T, were sufficient to switch the peptide from being specific for CRIPT to acquiring a specificity preference for the $T_{-2}F$ peptide (Fig. 2.5a).

From the biochemical affinities and structural location of the two mutations, it is clear this path contains the complexity empirically encountered in understanding protein adaptation. Amongst the three posible single variants from the wild-type protein-peptide complex, only the G330T mutation retains a high affinity for native function, and binds the $T_{-2}F$ peptide with approximately native affinity (Fig. 2.5**b**). These types of mutations (termed 'conditionally neutral' or 'cryptic genetic variants') have been shown to critically influence adaptation as such sequence variation is beneficial in new environments, thereby enabling adaptation, yet remain neutral in the native environment, retaining wild-type like function.[19–23] Additionally, though G330T promotes adaptation

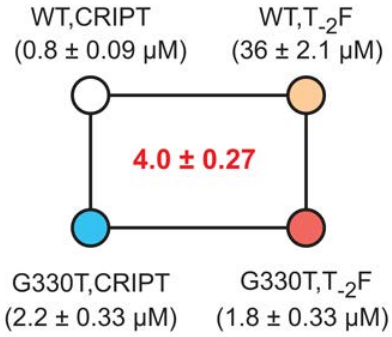

Figure 2.6: The G330T mutation, despite a contact shell removed from the peptide, is coupled to the $T_{-2}F$ mutation. Affinities of protein-peptide complexes shown here were measured by fluorescence polarization (see Methods).

and is strongly thermodynamically coupled to the $T_{-2}F$ peptide mutation (Fig. 2.6), residue 330 is situated distal to the -2 position of the ligand, making it an unclear choice of sequence variation based solely upon the protein structure. Thus the existence of this simplified yet representative path, in addition to the experimental tractability of $PSD95^{pdz3}$, provided a unique opportunity to investigate the relationship between sequence variation, protein structure, and phenotype–a critical interdependence underlying our comprehension of the adaptive process.

2.3 A Comprehensive Functional Characterization of the Adaptive Process

Thus far, the phenotype of the protein along the two-step adaptive path is defined by the affinities for simply two peptides, CRIPT and $T_{-2}F$. In reality however, the phenotype of a PDZ domain is measured by its specificity profile across all possible ligands the domain could bind. For instance, from the binding specificity of $PSD95^{pdz3}$, it would be assumed that the protein would bind class I peptides. To understand how sequence variation along the path of adaptation relates to the phenotype of the protein, we assayed the wild-type, each single mutant, and the double mutant against all possible binding partners of PDZ domains by creating a library of peptides with a fully randomized C-terminus (TKNYK**XXXX**), in total 160,000 peptides and measuring the binding effect of each peptide on the protein using a bacterial-2-hybrid high-throughput quantitative assay, modified from McLaughlin *et al.*.

2.3.1 The Bacterial-2-Hybrid System

The bacterial-2-hybrid system is a three component genetic circuit expressed in *E.coli* cells that couples PDZ-ligand affinity with a selection readout.[10] The PDZ domain, ligand, and readout are each expressed on a separate plasmid (Fig. 2.7). Post induction of the system by doxycycline, the expression of the readout, in this case the antibiotic resistance gene chloramphenicol acetyl transferase (CAT), is dependent on the affinity between PDZ domain and ligand (Fig. 2.7). If the affinity between PDZ and ligand is high, then CAT will be produced, allowing the PDZ and ligand to survive in a culture with chloramphenicol as a selection pressure. Experimental conditions were optimized so that a linear relationship exists over wide range of binding affinity (0.5 μM to 200 μM) and propensity for survival under antibiotic selection (see Methods for details of conducting the assay). The propensity for cellular survival (also termed 'enrichment') is measured as an enrichment value

$$E_i = \ln \frac{f_i^{sel}}{f_i^{uns}} - \ln \frac{f_{WT}^{sel}}{f_{WT}^{uns}} \tag{2.1}$$

where i can be either a protein or peptide and f_i is the frequency in either the unselected (uns) or selected (sel) populations. The allelic frequencies are determined by next generation sequencing on an Illumina platform (the MiSeq in-lab instrument or HiSeq core instrument) and normalized relative to a reference allele, noted as wild-type in equation 2.1. Given the linear relationship between binding affinity and enrichment, the next step was to transform in the ligand library to the bacterial-2-hybrid system and conduct four selection experiments, one for each protein.

2.3.2 Making the Ligand Library

Due to the high sequence complexity of the library, the generation of the library required unconventional methods, outlined in Figure 2.8. Two oligonucleotides were engineered, both with restriction sites (BsaI) that would self-restrict to create complementary sticky ends, and one with four sets of randomized codons made as 'NNS' (where 'N' is any nucleotide and 'S' is either a cytosine or guanine) by the company IDT. Whole plasmid PCR was performed using these oligonucleotides, thereby amplifying the entire plasmid with the randomized C-terminus. The placement of BsaI sites on both the 5' and 3' ends permitted a unimolecular ligation after restriction, a

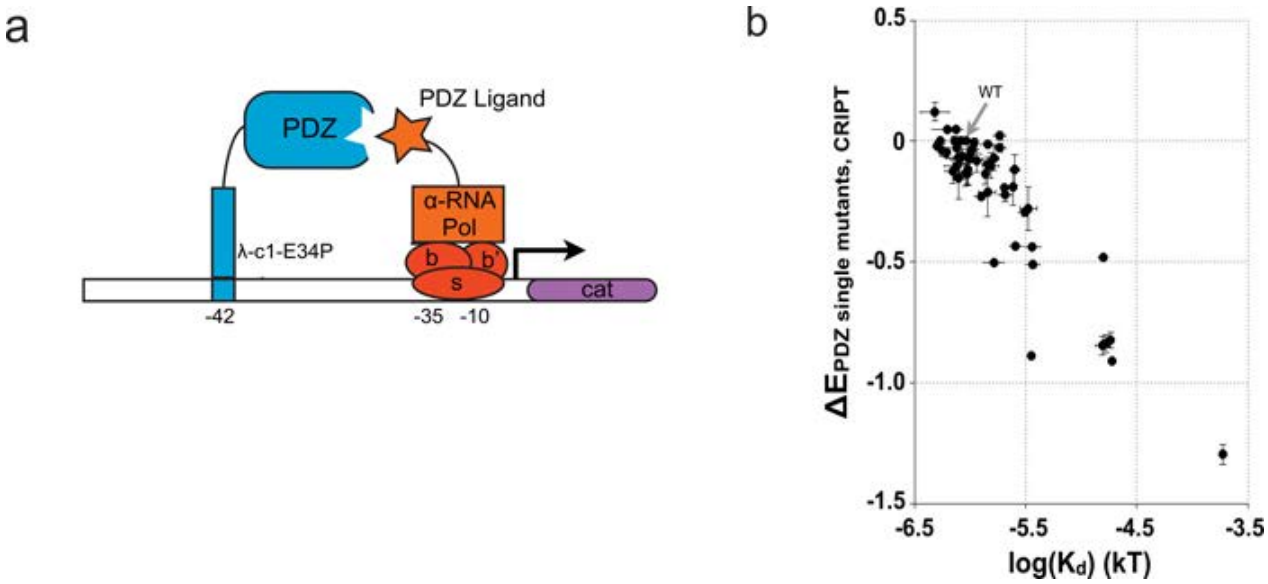

Figure 2.7: **A.** The PDZ domain is expressed as a fusion protein to the c1 protein from lambda phage inducible under IPTG, the PDZ ligand is expressed as a fusion protein to the α subunit of RNA polymerase inducible under doxycyline, and the CAT gene is expressed as a function of PDZ-ligand affinity. The three plasmids encoding each part of the bacterial-2-hybrid system are transformed into MC4100Z1 *E.coli* cells. **B.** Mutant PSD95^{pdz3} domains representing a range of binding affinities for CRIPT peptide were placed in the bacterial-2-hybrid system to test the relationship beween allelic enrichment and binding affinity. Experimental conditions were tuned such that binding affinity was linearly related to enrichment as measured by post-selection allelic frequencies.

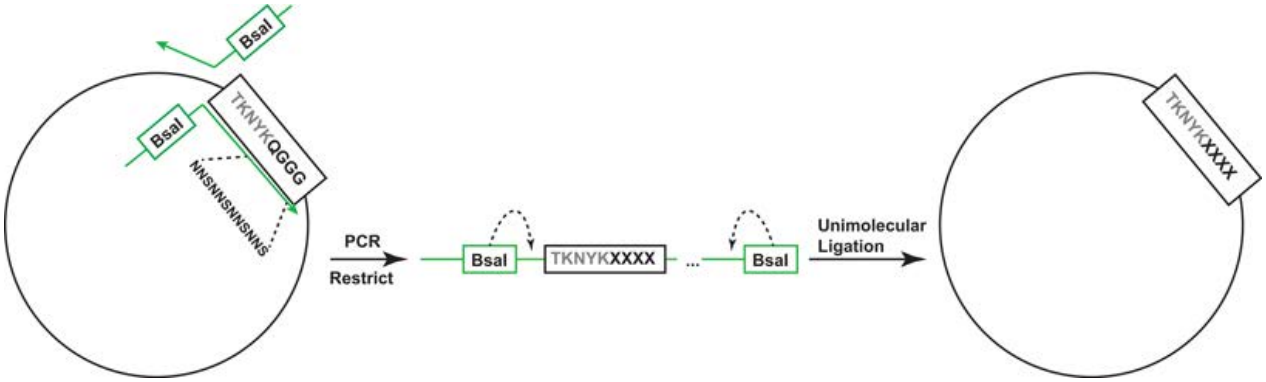

Figure 2.8: The library of all possible partners of PDZ domains is generated by randomizing the C-terminal four amino acids of the peptide sequence. The total complexity of the library is 10^5 (or 10^6 taking into account codon variants). To ensure that variants of the population are not lost during the cloning process, a highly efficient cloning strategy was necessary. We started from a plasmid that encodes TKNYKQGGG, a peptide that does not bind any PDZ domain and is a functionally dead ligand with respect to PSD95^{pdz3} binding to ensure no source of artiically positive signal in the populations post selection. Two oligonucleotides constructed with a BsaI site were used to amplify the entire plasmid and encode the ligand library by the inclusion of four consecutive degenerate codons 'NNS' (N=any nucleotide, S=cytosine or guanine) at the C-terminus of the peptide signal. The PCR product was restricted with BsaI and self-ligated (in a total ligation volume of 1mL), zymo purified into 10 μL. 10 transformations were performed into Max-DH10B cells (reported competency of 10^{13}) and pooled together. After growth overnight, the pooled populations were miniprepped.

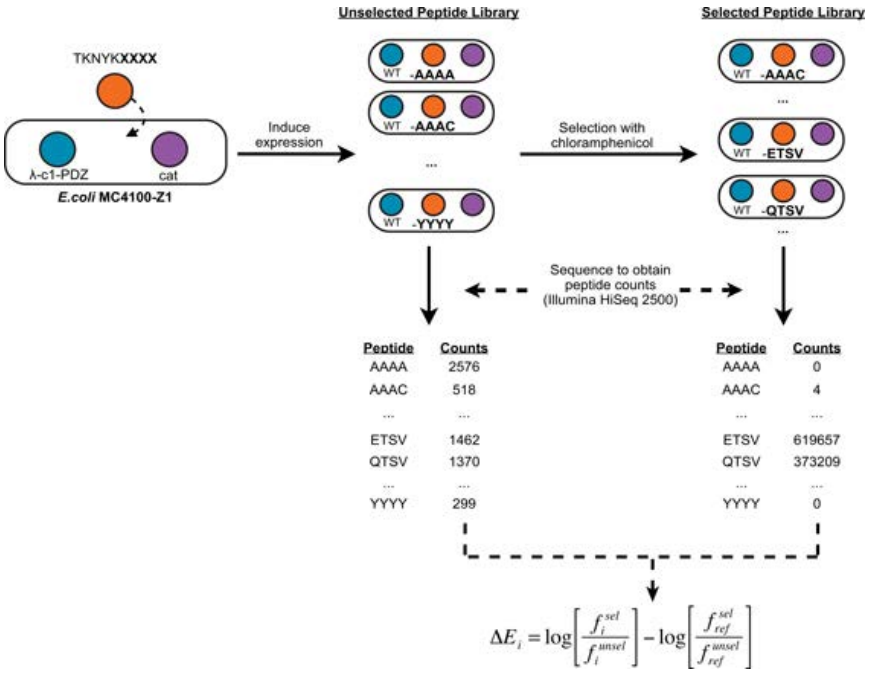

Figure 2.9: Plasmid containing ligand library was transformed into *E.coli* cells containing plasmids for the PDZ domain and CAT. The system was induced and an aliquot of the culture was prepped as a measure of the unselected population. Presumably, at this point, each peptide is equally represented in the culture as no chloramphenicol is in the culture and therefore CAT is not needed for cellular survival. Chloramphenicol was then introduced into the culture, purging the population of peptides in proportion to their affinity (as seen in the pilot experiment that set the dynamic range of the assay in Figure 2.7. Both the unselected and selected populations were sequenced and a corresponding enrichment was calculated for each peptide.

critical step to achieve a high transformation efficiency as to prevent bottlenecking of the peptide population.

2.3.3 Determining Protein Phenotype

The ligand library was transformed into MC4100Z1 *E.coli* cells already containing the plasmids encoding the λ-c1-PDZ fusion protein (the pZS plasmid) and the assay's PDZ-ligand affinity dependent readout, CAT (Fig. 2.9). The enrichment values calculated for the binding space of each protein were normalized to a high affinity peptide for that particular protein to make comparisons between specificities reasonable; the enrichment values for wild-type and G330T PSD95^{pdz3} were normalized to CRIPT whereas the H372A and H372A/G330T enrichment values were normalized to T$_{-2}$F. With respect to sampling, the ideal oversampling target would be 1000 fold. In other words, if there are 1x10^5 variants then one should hope to cover 1x10^8 sequencing reads. For each protein, a corresponding unselected population (also termed the 'input' population) was sequenced.

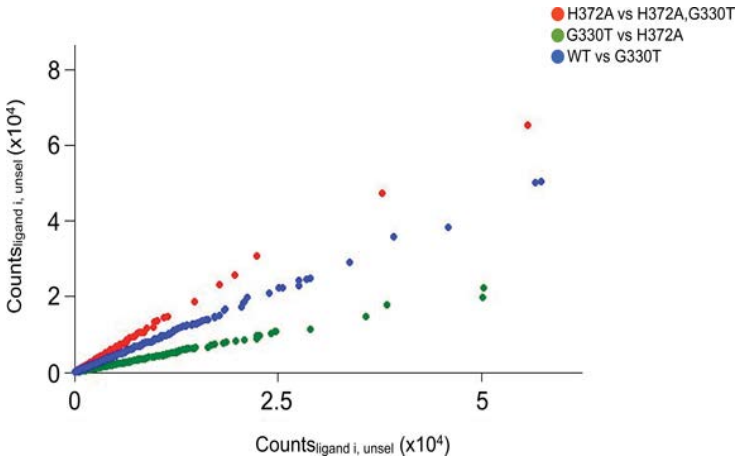

Figure 2.10: The experiments on each protein included an unselected population and a selected population. The unselected populations of each protein are similar in relation, an intuitive result given that no selection pressure is placed on the input culture. To guarantee oversampling of the unselected population, the input populations for the four different experiments were pooled into a single input population from which the enrichment calculations were made.

Figure 2.10 shows a comparison of the four input populations against each other. The relationship between the input populations was found to be linear (meaning the population frequencies in each input experiment were in proportion to each other) and therefore the populations were combined to give a single input population in order to gain adequate statistical sampling of the input. The library coverage was near exhaustive with $1.07 \text{x} 10^8$ total reads in the input population spanning approximately 155,000 out of 160,000 peptides with adequate statistics (Fig. 2.11). The selected populations were slightly undersampled relative to the total number of possible peptides with sequencing coverage for all four proteins approximately half an order of magnitude below 1000 fold oversampling of the population. Peptide enrichments for each protein were calculated and an enrichment threshold of greater than -0.8 (corresponding to 15μM from the assay calibration in Figure 2.7, an affinity empirically determined to indicate a physiologically positive PDZ ligand interaction) was applied for the peptide populations to determine the peptides that a given protein bound well(Fig. 2.12). Of the 160,000 possible peptides, 2447 bind at least one of the four proteins well. To get a sense of the specificity trend across the proteins, a subset of peptides are shown in Figure 2.13. Over the four proteins, the ligands were split by their sequence identities primarily driven by a single dimension: the amino acid identity of the -2 position of the peptide. Wild-type protein tends to bind peptides with a threonine or serine at the -2 position, adhering to class I specificity. G330T binds peptides with from both classes, consistent with the measured binding

Unselected Library Statistics

	Number of reads	Number of Ligands (> 50 counts)	% Library Coverage
Total Input Library	1.07×10^8	154,521	96.7

Selected Library Statistics

	Number of reads	Number of Ligands	Number of Ligands (> 50 counts)
WT	46,598,840	56,640	55,278
G330T	51,195,397	90,735	83,056
H372A	29,419,042	59,935	43,488
H372A/ G330T	48,989,769	123,295	86,255

Figure 2.11: An input, unselected library of ligands was generated for each protein and pooled (Fig. 2.10). The combined input library achieved 1000 fold oversampling over the complexity of the library. Each selected population was slightly undersampled, achieving approximately 500 fold oversampling over the complexity of the library.

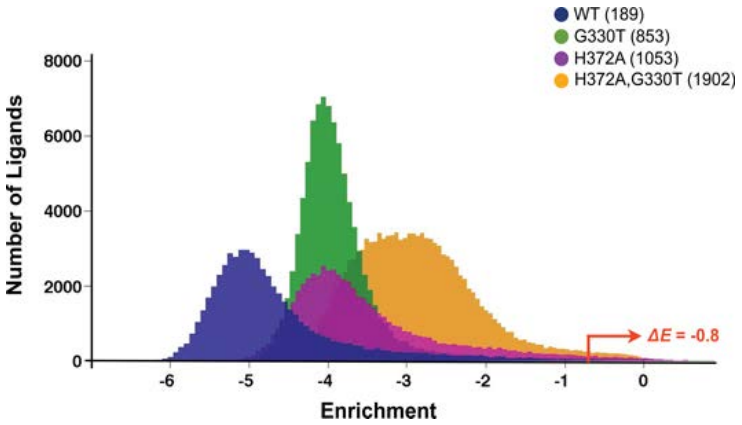

Figure 2.12: Peptides that bind a particular protein well are identified by those with an enrichment value of greater than -0.8, corresponding to an affinity of approximately 15μM. In parenthesis next to the identity of protein is the number of peptides that fall within this range. So for instance, wild-type PSD95^{pdz3} binds 189 peptides well.

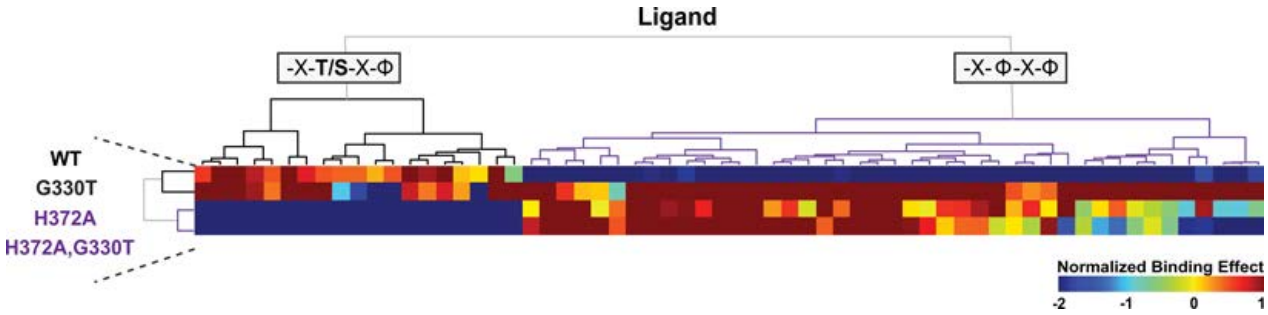

Figure 2.13: 2447 peptides bind one of the four proteins with greater than 15μM affinity. A subset of these peptides are shown here, clustered by 'cityblock' clustering as a function of enrichment values. While the clustering algorithm operates on enrichment values, we found a clear sequence distinction, the identity of the -2 position of the peptide, driving the enrichment differences. Peptides that bind wild-type protein well were observed to have a T/S at the -2 position; G330T accepts peptides with both a T/S or a hydrophobic residue at the -2 position (thus spanning both class I and class II peptides); H372A and the double mutant accept only hydrophobic residues at the -2 position, adhering to a class II specific domain.

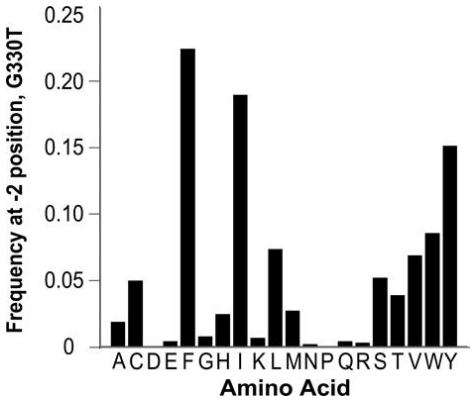

Figure 2.14: The G330T mutant shows a specificity profile spanning class I and class II specificity spaces, accepting both Thr or Ser (class I specificity) or hydrophobic residues (class II specificity).

affinities of G330T with CRIPT and T$_{-2}$F peptides (also seen in Fig. 2.14). H372A and the double mutant bind well to peptides with a hydrophobic residue at the -2 position.

Representing the data in matrix form, as in Figure 2.13, conveys the impression that the sole source of data variance across the four proteins occurs due to the -2 ligand position. For a more quantitative treatment of the data, we perform eigendecomposition of the covariance matrix of the enrichment matrix, identifying major modes of data variance. Intuitively, the covariance matrix of the data represents the extent to which two peptides share a similar binding pattern. Projecting each ligand onto a space defined by the bases of the decomposed covariance matrix can be thought of as placing two peptides close to each other if they share the same binding profile across the four proteins. The eigenspectrum of the covariance matrix is shown in Figure 2.15. Given the eigendecomposition of the covariance matrix, we were able to define a PDZ binding

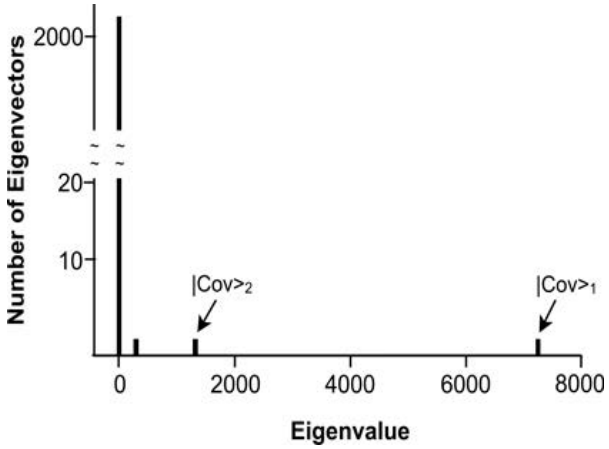

Figure 2.15: The covariance of the full enrichment matrix (Fig. 2.13) was decomposed and found to have two major modes of variance as shown by its eigenspectrum containing 93% of all data variance. These two modes are labeled Cov_1 and Cov_2. Each of the 2447 peptides are projected onto these axes to gain a visual representation of the binding space for each protein along the adaptive path.

space and project each of the 2447 peptides onto this space to visualize how sequence variation along the adaptive path can change the phenotype of the protein (Fig. 2.16). Wild-type $PSD95^{pdz3}$, consistent with its known specificity preference, follows a traditional class I binding designation, binding peptides with a Thr or Ser at the -2 position. The G330T single mutant, removed from the binding pocket and measured to bind both CRIPT and $T_{-2}F$ equally well, serves as a conduit between the two specificity classes, binding peptides with both a Thr or Ser and a hydrophobic residue at the -2 position equally well. This is in contrast to the other single mutant, H372A, located directly at the binding pocket and only able to bind class II peptides (Fig. 2.17). The double mutant similarly exhibits a complete specificity switch, showing no overlap with the wild-type binding space.

From the phenotypes of the proteins along the adaptive path, we were able to show that with just two mutations in the protein, a conversion to a physiological distinct class specificity was possible in $PSD95^{pdz3}$. Focusing on the phenotypes of the single mutants, there seemed to be potentially two different types of adaptive mutations: those that provide dual functionality (such as G330T able to bind both classes of peptides) and those that provide a direct specificity switch (such as H372A). From the binding space, it is clear that while G330T retains the ability to bind class I peptides, H372A loses this ability but provides a near complete specificity switch to class II function. A fundamental question thus arises: which of these kinds of mutations promote adaptation?

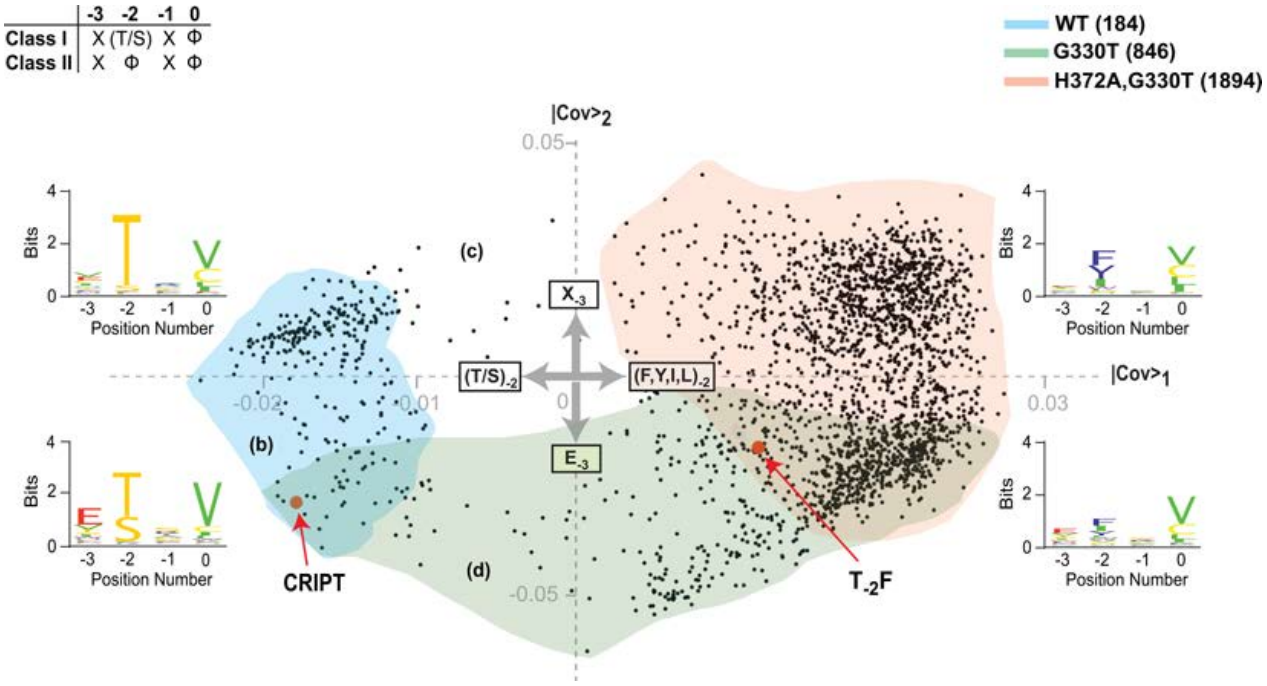

Figure 2.16: Each of the 2447 peptides is projected onto the two axes defined by Figure 2.15. Consistent with the dendogram representation (Fig. 2.13), a majority of the data variance is explained by the known class difference at the -2 position of the peptide. Peptides with a T/S_{-2} are placed on the negative side of the x-axis and those with a hydrophobic residue at the -2 position are placed on the positive side of the x-axis. A minority of the data variance is due to selection at a combination of the -2 and -3 positions of the peptide, reflected in the split of peptides along the vertical axis. Those with a selection for E at the -3 position are placed on the bottom half of the binding space whereas peptides exhibiting no selection at the -3 position are placed on the top half of the binding space. Consistent with known class designations, wild-type $PSD95^{pdz3}$ spans the class I specificity space (blue) comprising the left half of the graph. The G330T protein bridges both specificity classes. The double mutant is specificity switched, binding only class II peptides and thus situated on the right half of the graph.

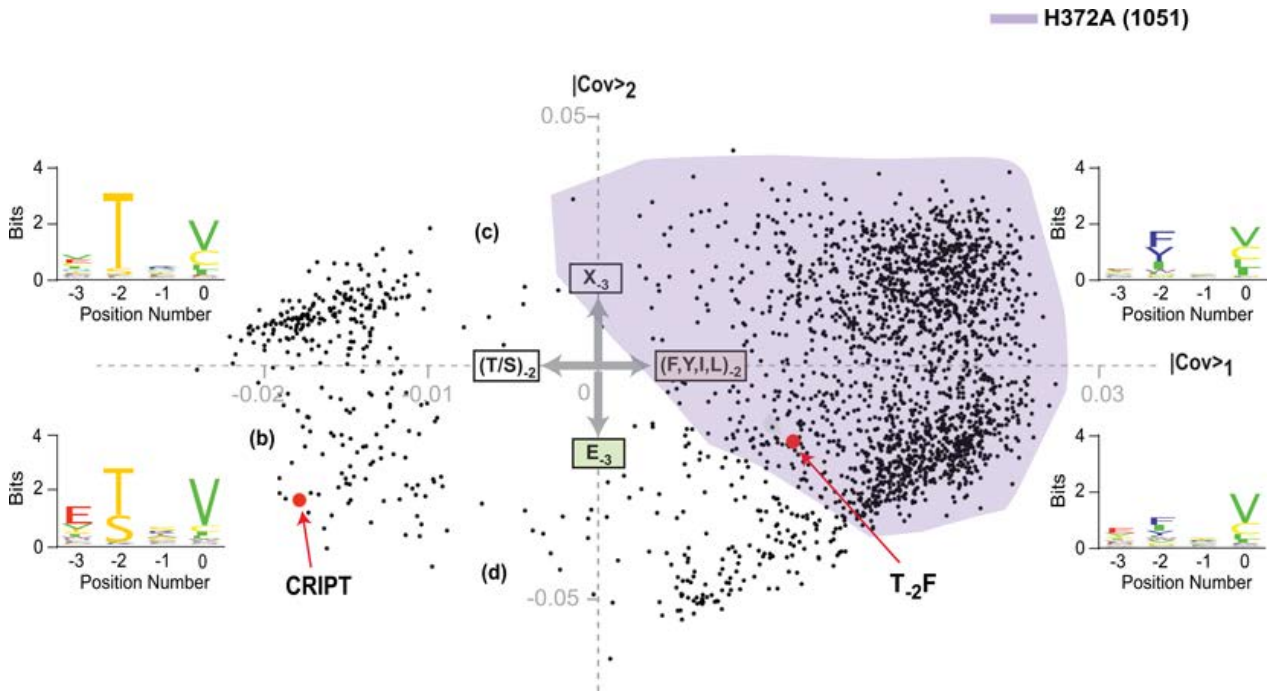

Figure 2.17: The H372A binding space shows a direct specificity switch with a single mutation. H372A PSD95^{pdz3} can only bind class II peptides with high affinity, thus abrogating native function.

The ability of the population to adapt to new function, and consequently the pathway that is followed (G330T first versus H372A first), is contingent upon the dynamics of the population, specifically the size of the population, the mutation rate, and the rate at which the environment fluctuates between a CRIPT environment and T$_{-2}$F environment. As a thought experiment, if the mutation rate were so low as to not produce any variants within the population before the identity of the environment is switched, the population would not at all be able to adapt to the new environment. However, aside from such extremes, it is not obvious which intermediate is a more favorable adaptive route. We therefore performed a simulation with a fixed population of 1000 members spanned over a wide range of mutation and environmental switch rates to statistically determine the efficacy of conditional neutral (generalists such as G330T) and direct class switching mutations in adapting to new function.

2.4 The Role of Conditionally Neutral Versus Class Switching Mutations in Aiding Adaptation

To get a sense of which types of mutations would help a population of wild-type PSD95^{pdz3} alleles adapt to the class II specificity, we performed a simulation tracking the flux of the population through the two different pathways to the double mutant as a function of mutation rate and environmental switch rate. Overall, a general formalism can be used for determining the steady state population distribution over an evolutionary trajectory in alternating environments by generating a transition matrix after each 'generation' that provides the probability of transition between alleles in the population.

2.4.1 General Formulation

For each generation in the simulation, there's a probability for each allelic transition between the four proteins (in the limit that each time step allows equilibration of the population). The set of transition probabilities between each of the four proteins can be represented by a transition matrix $\mathbf{P}$

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{21} & P_{31} & P_{41} \\ P_{12} & P_{22} & P_{32} & P_{42} \\ P_{13} & P_{23} & P_{33} & P_{43} \\ P_{14} & P_{24} & P_{34} & P_{44} \end{bmatrix}.$$

$\mathbf{P}$ is normalized such that the sum of the probabilities across a row is equal to 1.

At each generation, this matrix of transitions is applied to the state of the system $\mathbf{X}$. For instance, for the first generation

$$\mathbf{P}\mathbf{X_0} = \mathbf{X_1} \tag{2.2}$$

Generalizing this to t generations, the state of the population at generation t

$$\mathbf{P^t}\mathbf{X_0} = \mathbf{X_t} \tag{2.3}$$

What happens if there exists two environments that correspond to two different transition

matrices and the simulation alternates environments at a particular frequency? As an example, say the environment changes immediately after t environments and we want the state of the population $\mathbf{X_{t+1}}$. In this case, the transition matrix for the alternate environment (call it $\mathbf{Q}$) should be applied to the state $\mathbf{X_t}$. So

$$\mathbf{QX_t} = \mathbf{X_{t+1}} \tag{2.4}$$

Using (2),

$$\mathbf{QP^tX_0} = \mathbf{X_{t+1}} \tag{2.5}$$

So, following this, if the switch to the $\mathbf{Q}$ environment occurs after t generations and lasts for an additional t generations, the state $\mathbf{X_{t+t}}$ is

$$\mathbf{Q^tP^tX_0} = \mathbf{X_{t+1}} \tag{2.6}$$

The first eigenmode of $\mathbf{Q^tP^t}$ is the state population at generation $2t$ given that the environment switches at t.

2.4.2 Transition matrix for infinite population of WT, G330T, H372A, H372A/G330T

Assuming the population equilibrates after each time step, we can generate a transition matrix for the four proteins with a fixed mutation rate. The 4x4 transition matrix will have diagonal terms that represent the probability of persistence of a particular allele ($P_{i,i}$) and off diagonal terms that represent the probability of an allele mutating to another ($P_{i,j}$). As in the general formuation, the sum of the events that can occur for a particular allele has to equal 1 (so each of the probabilities needs to be normalized accordingly). As an example, let's take wild-type allele (which would correspond to the first row of the transtiion matrix). The probability of wild-type allele mutating to G330T or H372A is μ. The probability of wild-type mutating to the double mutant is μ^2 because two mutations need to occur in order for the double mutant to become populated. The probability of wild-type persisting is therefore $1 - 2\mu - \mu^2$. Because we are in the

limit of infinite popuation, the probability of an event occurring to the wild-type allele should be directly proportional to the fitness (assuming that mutations occur before selection) and so each transition probability is weighted by f_{WT}, the fitness of wild-type allele in the environment. For the purposes of this experiment, the fitness of each allele is determined by a binding isotherm defined by a single site:

$$\frac{[L]}{[L]+K_d} \tag{2.7}$$

where $[L]$ is the ligand concentration. In this model setup, it is important to choose a ligand concentration such that there is a range of fitness values reflective of the measured affinities. If for instance the values of $[L]$ are too large, then the K_d values carry little value. We therefore chose a $[L]$ value of 10μM, a value between the high affinity complexes of 1μM and low affinity of 36μM. The fitness matrix $\mathbf{F}$ multiplied by the mutation matrix $\mathbf{A}$ gives a transition matrix $\mathbf{P}$ (with rows 1, 2, 3, and 4 corresponding to wild-type, G330T, H372A, and H372A/G330T respectively).

$$\mathbf{FA} = \mathbf{P} \tag{2.8}$$

where

$$\mathbf{F} = \begin{bmatrix} f_1 & 0 & 0 & 0 \\ 0 & f_2 & 0 & 0 \\ 0 & 0 & f_3 & 0 \\ 0 & 0 & 0 & f_4 \end{bmatrix}.$$

and

$$\mathbf{A} = \begin{bmatrix} 1-2\mu-\mu^2 & \mu & \mu & \mu^2 \\ \mu & 1-2\mu-\mu^2 & \mu^2 & \mu \\ \mu & \mu^2 & 1-2\mu-\mu^2 & \mu \\ \mu^2 & \mu & \mu & 1-2\mu-\mu^2 \end{bmatrix}.$$

So, as in the previous section, the transition matrix raised to the t operates on the initial state of the system $\mathbf{X_0}$ to give the state of the system at generation t: $\mathbf{X_t}$. Again, an arbitrary number of transition matrices can be made for any number of given environments (as demonstrated for two environments in (5) with a single environmental switch). The eigendecomposition of the

aggregate of transition matrices across the trajectory (after $t$generations) is extremely difficult to solve analytically. Furthermore, the analytical solution provides little intuition about the steady state dynamcis of the population. Thus, a stochastic simulation seemed appropriate to tackle the problem (see Methods for MATLAB scripts to execute code).

2.4.3 The Simulation

The simulation starts with 1000 wild-type alleles with a particular mutation rate that remains fixed throughout the trajectory and poissonian environmental switching. For instance, if the environmental switch rate is set to be on average once every 2500 generations, then the environmental switching encountered in the simulation will be a poissonian distributed variable around one switch per 2500 generations. At each generation, a transition matrix is made that is the probability of seeing a transition between two alleles from computing the flux of each allele in the system. For instance, the net flux of wild-type at a particular generation t is

$$\Phi_{WT,net,t} = \Phi_{WT,+,t} - \Phi_{WT,-,t} \tag{2.9}$$

The positive flux for wild-type is equal to the amount of wild-type made by conversion of other alleles in the population from the previous generation $t-1$ plus the persistence flux of wild-type governed by the fitness value:

$$\Phi_{WT,+,t} = N_{WT,t-1} * f_{WT} + N_{G330T,t-1} * \mu + N_{H372A,t-1} * \mu + N_{H372A/G330T,t-1} * \mu^2 \tag{2.10}$$

In words, this is simply the number of wild-type that persists in the population plus the number of wild-type that gets added due to mutations (set by the mutation rate) to G330T, H372A, or the double mutant. The negative flux for wild-type is equal to the amount of wild-type in the previous generation $t-1$ converted to the other alleles in the population:

$$\Phi_{WT,-,t} = N_{WT,t-1} * \mu + N_{WT,t-1} * \mu + N_{WT,t-1} * \mu^2 \tag{2.11}$$

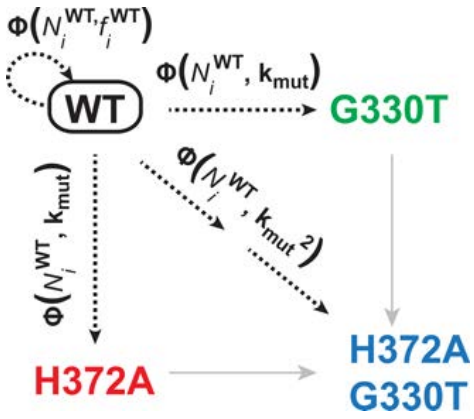

Figure 2.18: Within the simulation, the flux of each allele is calculated at a given generation. As an example of what is meant by 'flux', here is shown the outward flux of wild-type alleles. For generation $i+1$, the wild-type can either persist as itself or mutate to either G330T, H372A or the double mutant. The probability of persisting is simply a function of the number of wild-type alleles, N_i^{WT} in generation i and the fitness of wild-type in generation i, f_i^{WT}. The probability of mutating to either single mutant is a function of the number of wild-type alleles and the mutation rate k_{mut}. Since two mutations are required to mutate to the double mutant, the probability of converting to double mutant is a function of the number of wild-type alleles and k_{mut}^2. Similarly there are positive fluxes, that is the probability of conversion to wild-type, from G330T, H372A, and the double mutant. The simulation tracks the fluxes over all alleles of the population to generate a probability distribution of seeing any given allele at generation $i+1$.

So the probability of seeing wild-type in generation t is

$$P_{WT,t} = \frac{\Phi_{WT,net,t}}{\Phi_{WT,net,t} + \Phi_{G330T,net,t} + \Phi_{H372A,net,t} + \Phi_{H372A/G330T,net,t}} \tag{2.12}$$

The outgoing (or negative) flux of wild-type PSD95^{pdz3} is illustrated in Figure 2.18. We similarly have three more probabilities, $P_{G330T,t}$,$P_{H372A,t}$, and $P_{H372A/G330T,t}$. In the simulation, the *mrnd* function in matlab is applied to this probability distribution to generate the populations for each allele that sum to 1000 (the population size) at each generation, thus the stochastic nature of the simulation. The probabilities obviously change as the environment is altered due to the altered fitness values. Thus, throughout the simulation, there is constant competition between the population and depletion of any given protein as a function of mutation rate and protein fitness.

2.4.4 Relative Adaptive Utility of G330T versus H372A

The population, intially comprising of only wild-type PSD95^{pdz3} was evolved over a course of 10,000 generations with a defined mutation rate and rate of environmental fluctuation with the environment defined as either the CRIPT or T$_{-2}$F ligand. Figure 2.19 shows an example of

a single trajectory with a low mutation rate (10^{-6}) and environmental fluctuation rate (once per 2500 generations). The environment in the trajectory cycles between CRIPT and T$_{-2}$F and accordingly the population shifts to adapt to the corresponding environment. We see that for this particular pair of mutation rate and environmental switch rate, both H372A (the first CRIPT to T$_{-2}$F switch) and G330T (the second CRIPT to T$_{-2}$F switch) facilitate adaptation. (The only way the double mutant population can adapt back to the wild-type population in the T$_{-2}$F to CRIPT transition is either through a) G330T or b) a double mutation (a highly improbable event as making two mutations in a single generation scales as k_{mut}^2).) The metric that determines which protein promotes adaptation is, upon inspection, obvious: the more prevalent of the two proteins in the generation *immediately preceeding* the switch will adapt the population to the double mutant. Take, for instance, the first CRIPT to T$_{-2}$F transition (generation 2425 to 2435, Figure 2.19). Right before the switch, the number of H372A variants in the population are greater than the G330T. In contrast, for the second transition, the number of G330T variants at the switch outnumber the H372A variants. Averaging the results of this particular trajectory over many identical trajectories, we would find that the relative adaptive utility of G330T would be equal to that of H372A.

To more generally understand the population dynamical regimes in which H372A or G330T is favored to adapt the population, the simulation was performed over a range of mutation rates and environmental switch rates (Fig. 2.20). Querying a spectrum of mutation rates and rates of environmental switching, there exists a range of parameter space in which conversion to the double mutant protein is guaranteed. In the remainder of the parameter space, the rate of ligand switching is too rapid for the given mutation rate to create the double mutant and therefore the population converges to the only mutant that is fit in both environments, G330T. Over the space of guaranteed conversion to the double mutant protein, we find that for a majority of the parameter space, making the G330T mutation first is the heavily favored path for adaptation to class II specificity in comparison to H372A, which at best is able to match, but not exceed, the adaptive utility of G330T. Why is this? As the example in Figure 2.19 demonstrates, the choice of using G330T or H372A to convert to the double mutant is dependent on the relative population of either mutant at the instance of the environmental switch. In the particular trajectory shown in

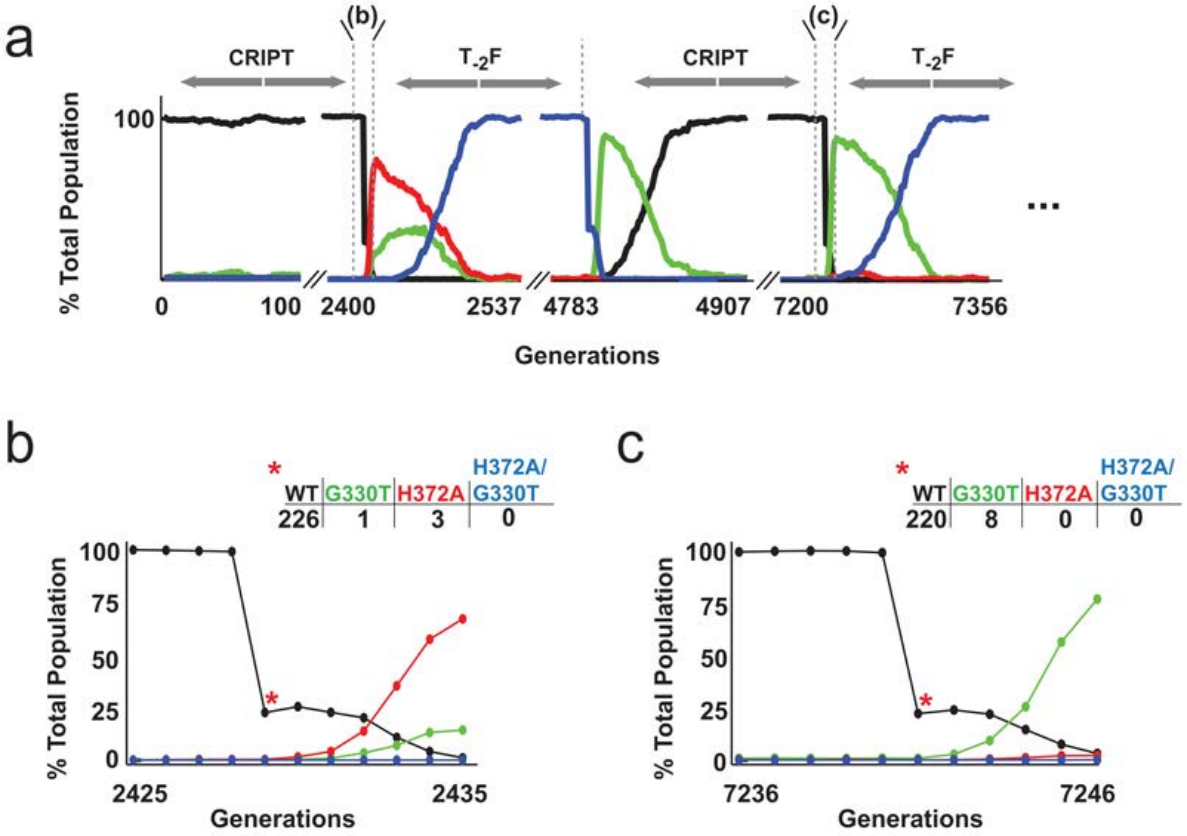

Figure 2.19: A. A single trajectory with mutation rate of $5x10^{-6}$ and environmental switch rate of every 2500 generations is shown. The populations starts out as completely wild-type (black trace) in a CRIPT environment. As the environment changes from CRIPT to $T_{-2}F$, we see adaptation occur with the population shifting to the double mutant allele (blue trace). For the first transition, H372A is the preferred path of adaptation over G330T. The environment proceeds to switch back to the CRIPT environment and G330T facilitates this adaptation. Note, for the transition from $T_{-2}F$ to CRIPT, only G330T can facilitate adaptation because the H372A mutant is unfit in the CRIPT environment. Upon the final transition from the CRIPT to $T_{-2}F$ environment (at approximately generation 7200), the G330T variant facilitates adaptation. **B.,C.** Zooming in on the two CRIPT to $T_{-2}F$ transitions, we see that the major factor governing which path (H372A vs G330T) is chosen is the relative population distribution at the generation immediately preceeding the switch (asterisked generation represents the environmental change). In the first transition, more H372A variants exist and subsequently H372A facilitates adaptation. In the second transition, only G330T variants exist and therefore G330T promotes adaptation to the double mutant.

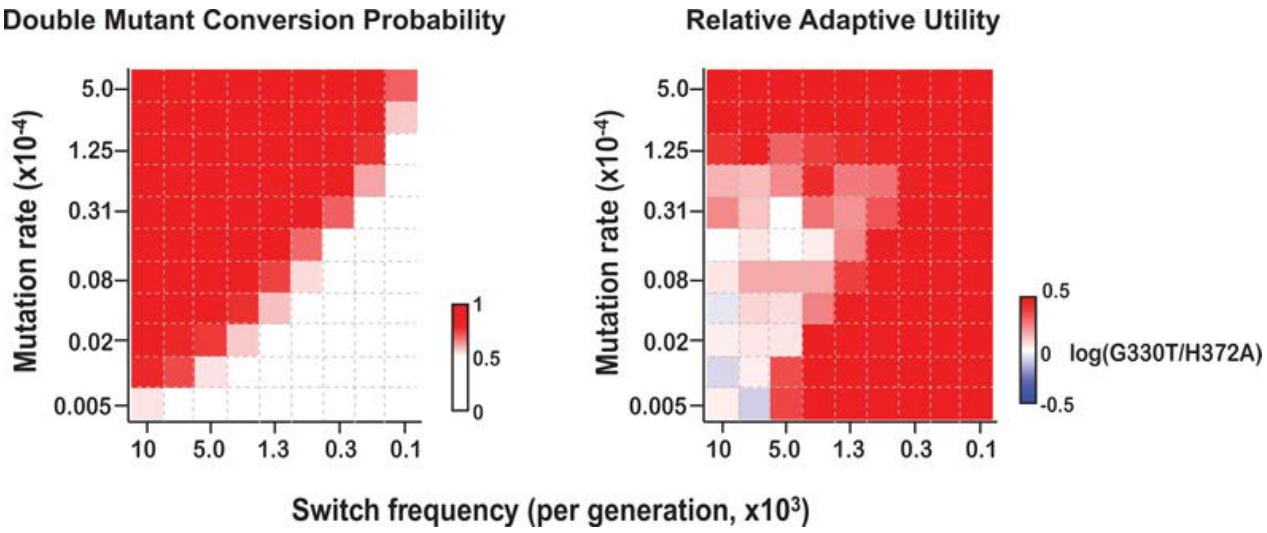

Figure 2.20: The stochastic simulation was performed over a range of mutation rates and environmental switch rates. The conversion to double mutant is guaranteed in a subspace of the total parameter space. If the rate of ligand switching is too high relative to the mutation rate, the population fixates on the G330T mutant, the variant that is good for both environments. In the regime that the population is guaranteed to reach the double mutant, G330T is mostly the dominant adaptive variant–a reasonable result given that G330T is fitter in the CRIPT environment and would therefore outnumber the H372A variant at the generation of the environmental switch. The only regime in which H372A aids in adaptation is in the so-called 'stochastic' regime where a sufficient number of mutations have not been made in the population to allow the population to adapt. In this scenario, adaptation would only occur *after* the environmental switch at which point the likelihood of G330T or H372A emerging as a result of mutation (and thereby facilitating adaptation) is mathematically equivalent. Thus the relative adaptive utility at low enough mutation rates for a given environmental switch rate for G330T and H372A should reach 50% given adequate statistical sampling.

Figure 2.19, the mutation rate is sufficiently low relative to the rate of environmental fluctuation as to render the flux through either mutant equally probable. Mathematically this has a precise meaning. Consider a single trajectory that goes through τ generations before an environmental switch, with a population size of N and a mutation rate μ. The product of these three values, $N\mu\tau$, is the number of mutations that are produced in the population within the number of generations before the environment switches. Obviously, if the number of mutations produced is high, then the population can adapt to the switch in environment. However, if $N\mu\tau$ is less than 1 (a regime known as the stochastic limit where finite population effects dictate outcomes), then this means that not a single mutation is made in the time between environmental switches and therefore the population can only adapt after the environmental switch. In this scenario, the probability that H372A or G330T facilitates adaptation is statistically random. Because the probability of mutating wild-type to either single mutant is equal (simply based on the mutation rate μ), whichever mutant just happens to appear after the environmental switch will facilitate adaptation. This regime occurs at low mutation rates and low switch rates (the lower left of the adaptive utility grids in Figure 2.20). The number of alleles in the population is 1000 and because there is some noise in the *mnrnd* function in generating the population distribution, our stochastic regime spans around the range of $N\mu\tau$ is less than 20 (meaning that if less than 20 mutations are made in population before the environmental switch, the population is in the stochastic regime and H372A is equally as probable as G330T in aiding adaptation).

However, if the mutation rate is increased and the timescale of selection within the population is faster than the rate of change in environment, then the more prevalent single mutant at the instance of the CRIPT to $T_{-2}F$ transition will be the variant that is more fit in the CRIPT environment and therefore allowed to float at low levels in the population. Indeed for a trajectory with an increased mutation rate, we find that CRIPT to $T_{-2}F$ transitions are mediated by only the G330T variant. (Fig. 2.21). When the mutation rate is raised, more of the wild-type alleles are converted to either G330T or H372A. In the CRIPT environment, the H372A variants suffer a great fitness cost relative to the G330T variants and therefore are purged from the population. The G330T variants, on the other hand, can persist in the CRIPT environment and are available once the environment switches to the T_{-2} environment.

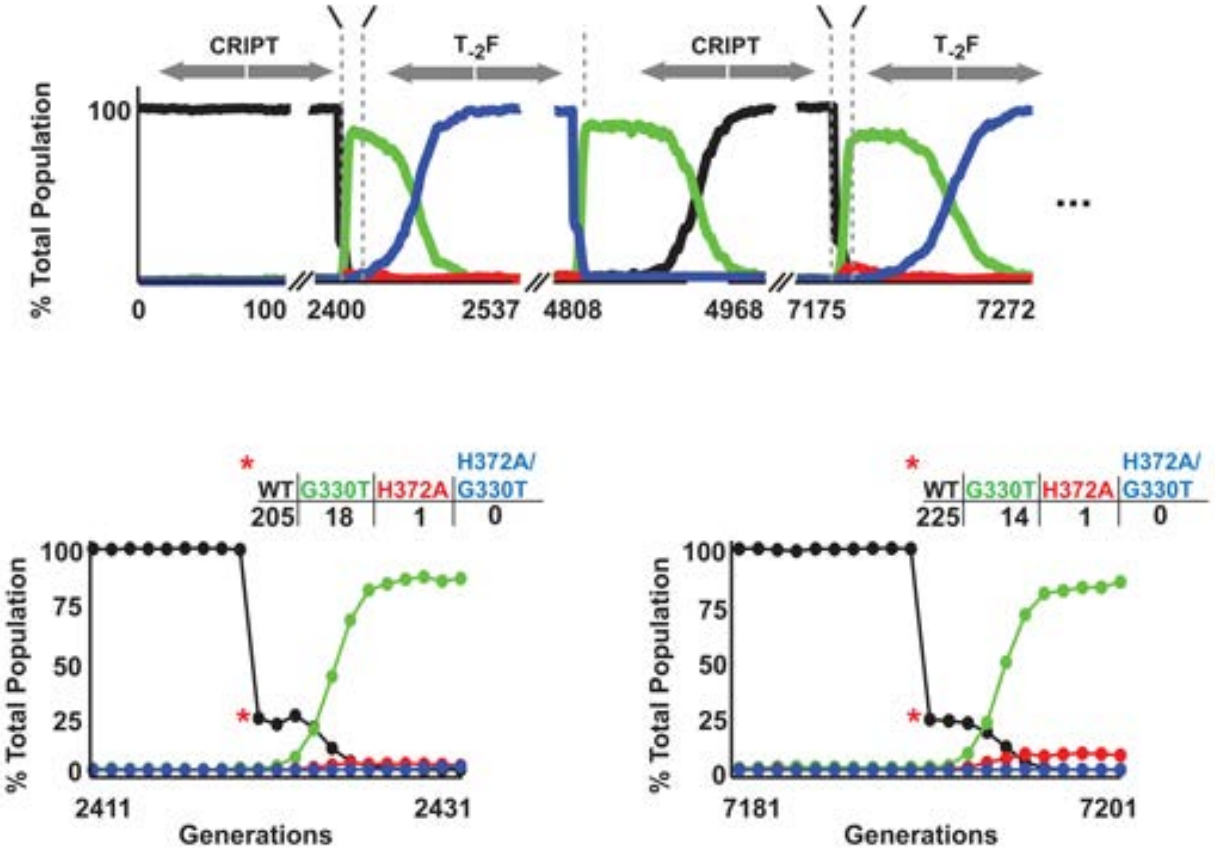

Figure 2.21: Given a higher mutation rate (with an environmental switch rate of every 2500 generations), more mutations are introduced into the population at each generation. While the generation of G330T and H372A is equally probable, H372A is unfit in the CRIPT environment and is therefore deselected relative to the G330T variant which suffers only a slight decrease in fitness. As a result, the G330T variant is allowed to float in the population, greatly outnumbering the H372A variant at the time of the environmental switch. This phenomenon is precisely why conditionally neutral mutations or cryptic genetic variation are argued to be particularly useful for adapting to new function.

Overall, we concluded from the simulation that the G330T variant (and by extension, variants that can perform both the native and target functions) is in general more useful for adaptation than H372A. The only regime in which the direct class switching mutation (H372A) is able to facilitate adaptation is in the stochastic regime where mutation rates are sufficiently low as to retard adaptation until after the environmental switch. This result is consistent with previous literature arguing the importance of conditional neutrality and cryptic genetic variation in protein adaptation. As opposed to direct class-switching variants, conditionally neutral mutants do not suffer a fitness decrase in the wild-type environment and therefore can persist to enable the evolutionary process. Fundamentally this result can be attributed to the fact that in order for class-switching variants to be more generally useful, two events, protein sequence variation *and* a change in environmental fitness pressure, must happen simultaneously–an anathema to the evolutionary principle of single variations driving adaptation.

2.5 Adaptive Mutants and the Design of Natural Proteins

We thus far identified a two-step mutational path to new function in $PSD95^{pdz3}$ and showed that for most of relevant population dynamics parameters, mutations that are conditionally neutral are particularly useful for adaptation to new function. To relate the capacity to adapt quickly to the design of natural proteins, we asked where do such mutations reside in the protein structure? To answer this question, we revisited the single mutation scan of $PSD95^{pdz3}$ assayed against CRIPT and $T_{-2}F$ performed by McLaughlin *et al.* to identify mutations that are neutral for CRIPT but adaptive towards $T_{-2}F$.

2.5.1 Identifying Conditionally Neutral Mutations in $PSD95^{pdz3}$

From the single mutation scan on $PSD95^{pdz3}$, mutations that are gain of function for $T_{-2}F$ but neutral for CRIPT can be identified. For instance, Figure 2.22 highlights two positions, 322 and 362, for which most variants made at these positions are neutral for CRIPT (white pixel) and adaptive to $T_{-2}F$ (red pixel). We plotted all single mutations in $PSD95^{pdz3}$ for CRIPT and $T_{-2}F$ against each other, defining neutrality for CRIPT binding as any enrichment value greater than 16μM and adaptive for $T_{-2}F$ as above 15μM. Thus, using this plot, we can define both conditionally

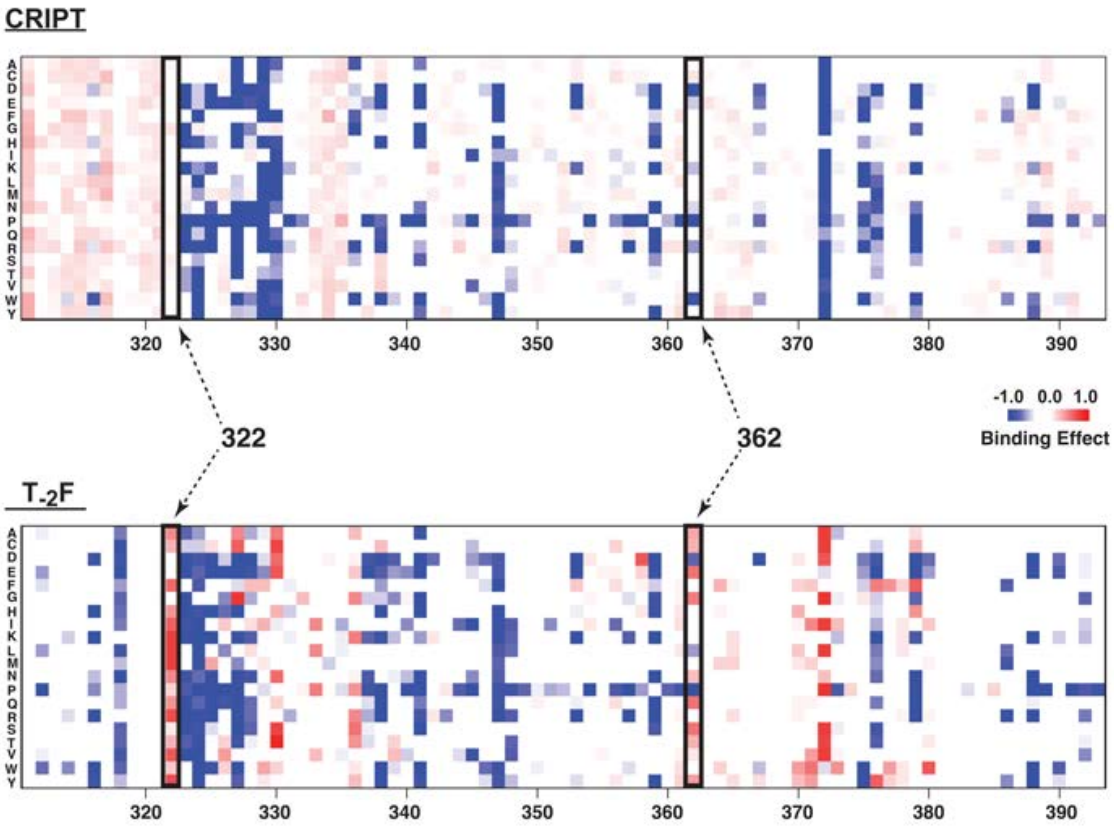

Figure 2.22: The saturation single mutation scan provides valuable information with respect to identifying conditionally neutral positions in PSD95^{pdz3}. The mutational effect of all possible mutations at positions in particular are highlighted here: position 322 and 362. Both of these positions, when mutated, are mostly neutral for the CRIPT peptide but promote adaptation to the T$_{-2}$F peptide. Thus, functionally mutations at these positions are equivalent to the G330T mutation.

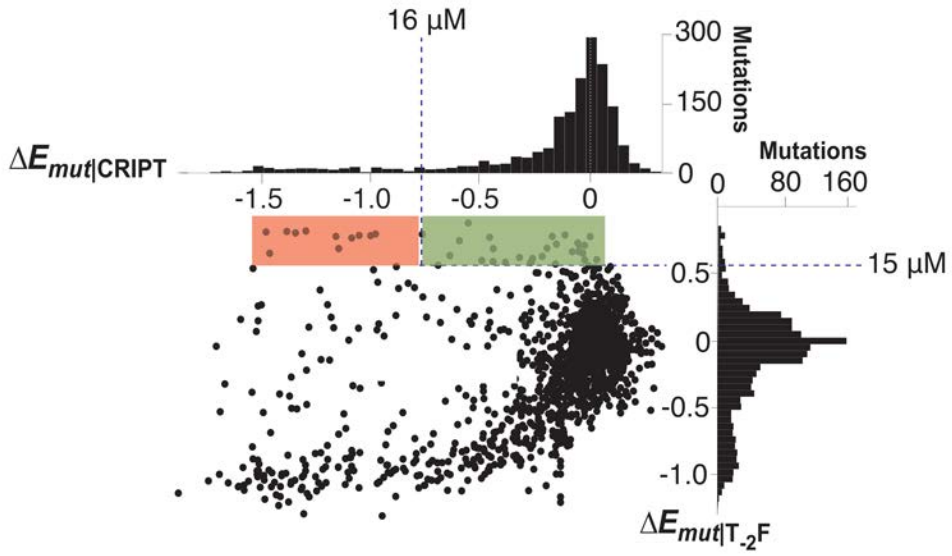

Figure 2.23: To comprehensively identify the conditionally neutral mutations in the protein, a scatter plot of the functional effects of all single mutants assayed against CRIPT and $T_{-2}F$ peptides are plotted against each other. The explicit K_d values shown are measured affinity values by either isothermal titration calorimetry or fluorescence polarization and correspondingly associated with an enrichment value as measured from the bacterial-2-hybrid system. The blue line for each histogram delineates cutoffs that are used to determine neutrality for CRIPT (less than 16μM) and adaptive for $T_{-2}F$ (less than 15μM). In green shading are all mutants that are considered as conditionally neutral such as G330T (neutral for CRIPT peptide but adaptive for the $T_{-2}F$ peptide. In red shading are all mutants that are considered direct class switching mutations such as H372A (decrease binding affinity towards CRIPT peptide and adaptive towards the $T_{-2}F$ peptide.

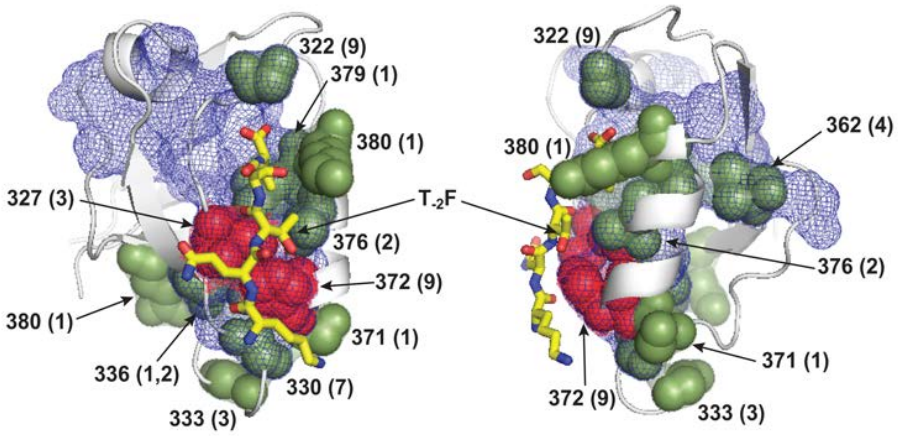

Figure 2.24: Class switching (red) and conditionally neutral mutations (green) are shown in spheres on the cartoon of $PSD95^{pdz3}$ with the protein sector shown in blue mesh with the number of adaptive mutations at these sites in parenthesis. The gradation of red or green reflects the number of adaptive mutations the position harbors. Class switching mutations were found to cluster around the -2 position of the peptide (residues 372 and 327). Conditionally neutral mutations were exhibited a spatially heterogeneous pattern with most of such mutations located several contact shells away from the -2 position. Both types of mutations were found almost exclusively within the sector.

neutral mutations as well as class-switching mutations (mutations that are deleterious for CRIPT binding yet adaptive to $T_{-2}F$, similar to H372A).

Figure 2.24 shows the positions that contain conditionally neutral mutants and class-switching mutants. The class switching positions (red) are clustered in direct contact with -2 position of the peptide, the location of the adaptive challenge on the protein. In contrast, the conditionally neutral mutations are spread throughout the protein, with a majority of such mutations located multiple contact shells removed from the T_{-2} position such as residues 322 and 362 as noted in Figure 2.22. Interestingly, regardless of whether the adaptive mutation is designated class-switching or conditionally neutral, we found that adaptive mutations in general are highly enriched within the protein sector (blue mesh). The main distinction between the two classes of adaptive mutations is their spatial distribution within the protein sector. An alternate view of this result is seen in Figure 2.25. Here, positions are plotted as contact shells removed from the -2 position of the peptide. Class switching mutations are mostly found in direct contact with the -2 position such as position 372 harboring 9 such mutations and located 2.7Å from the site of adaptive challenge. Conditionally neutral mutations are mostly located at positions removed at least one contact shell from the -2 position such as positions 330 (7 conditionally neutral mutations), 362 (4 mutations),

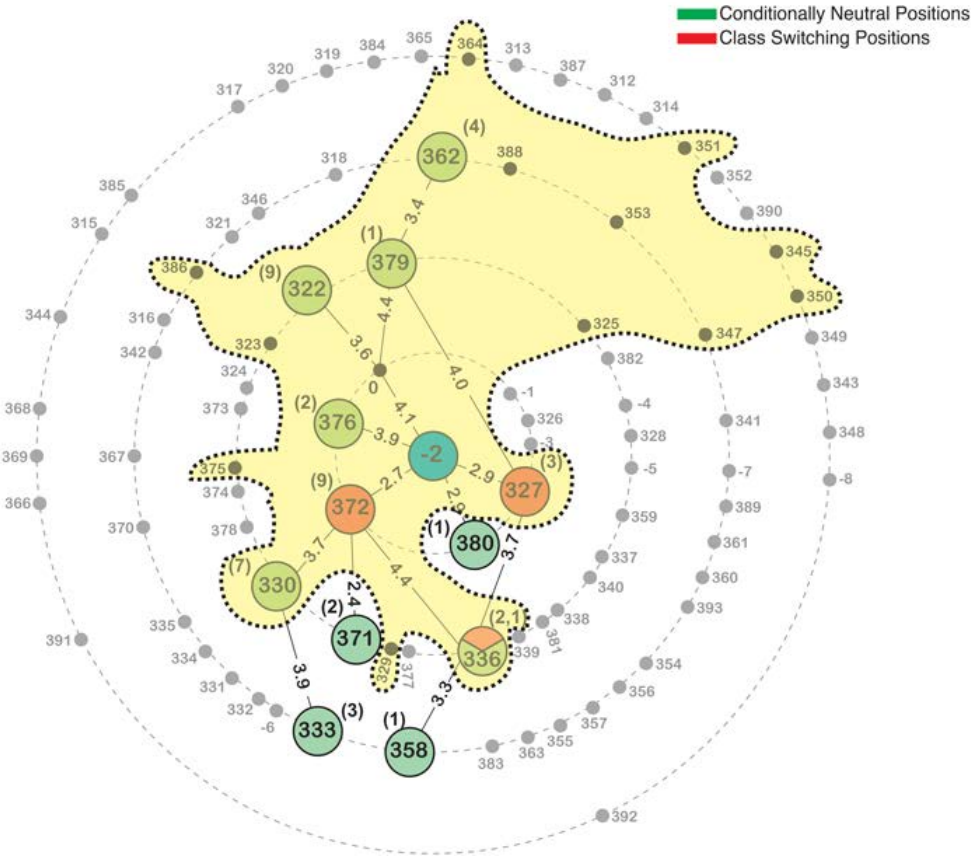

Figure 2.25: Positions in the protein are plotted as a function of contact shell away from the -2 position of the peptide with number of adaptive mutations at the position shown in parenthesis and the distance between positions delineated in the position bonds. For instance, the first shell of residues emanating outwards from the center are in direct contact with the -2 position. Positions within the protein sector are outlined in yellow surface. Class switching mutations (red) are mostly found in direct contact with the -2 position while a majority of conditionally neutral mutations (green) are at least a single contact shell removed from the -2 position. For instance, positions 322 and 330, removed from the -2 position by a single contact shell, harbor 9 and 7 conditionally exaptive mutations respectively. As a more extreme example, residue 362–a sector position 12Å and three contact shells away from the -2 position of the ligand–harbors 4 conditionally neutral mutations. While their spatial distribution may be different, we found that both conditionally neutral and exaptive mutations are highly enriched within the protein sector thus arguing that the protein sector generally encodes the ability to adapt to new function.

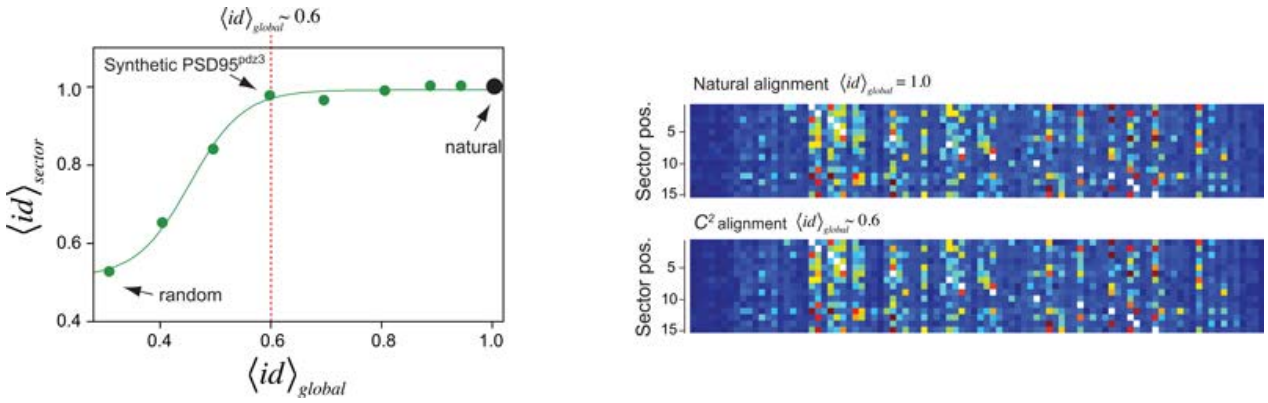

Figure 2.26: Using a Monte-Carlo design algorithm, annealing on the correlations of the SCA matrix, a number of synthetic PDZ domains were made using wild-type PSD95^{pdz3} as a seed sequence. The wild-type sequence was 'heated' until the correlations outside the protein sector were lost, thereby preserving the identity of the sector but mutating many positions outside the sector. Thus, while the synthetic sequences have a global identity of 60%, the sector identity is 100% as evidenced by the correlation matrix of the natural alignment and the synthetic alignment (labeled as 'C^2'). (Figure courtesy of Walraj Gosal.)

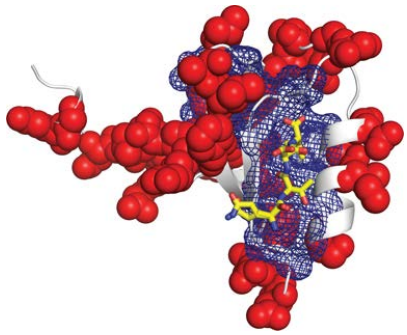

Figure 2.27: Shown as red spheres are the positions on PSD95^{pdz3} that are mutated in the synthetic PDZ domain, designed to maintain the protein sector (blue mesh).

and 322 (9 mutations). Of all possible adaptive mutations, 32 are found to be conditionally neutral and 13 class switching. All class switching mutations and 25 out 32 conditionally neutral mutations are found within the protein sector (blue outline) with the conditionally neutral mutations highly enriched at sector positions at least one contact shell removed from the -2 peptide position. These results thus suggest that the ability to adapt to new function is exclusively found within the sector. An experimental corollary to this statement would be that mutations made outside the sector should not influence the phenotype of PSD95^{pdz3}. To address the uniqueness of the protein sector with regard to protein phenotype, we designed and measured the full binding space of a synthetic PDZ3 protein designed by a Metropolis Monte-Carlo simulated heating algorithm with 60% overall sequence identity to wild-type PSD95^{pdz3}(Fig. 2.26, synthetic protein designed and cloned by Walraj Gosal). While the two sector mutations, G330T and H372A, caused a drastic reshaping

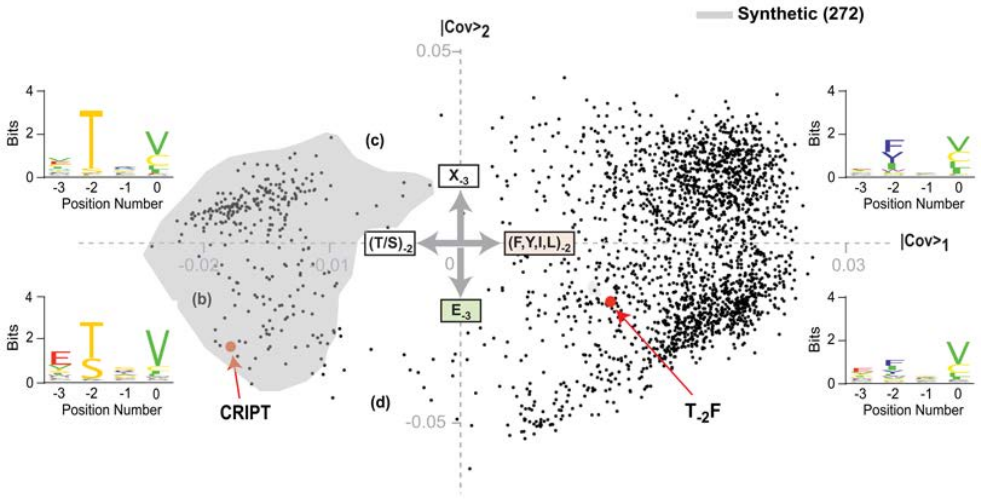

Figure 2.28: The binding space of the synthetically designed PDZ domain was determined using the bacterial-2-hybrid ligand library screen. The synthetically designed protein with a 100% sector identity maintains class I specificity, binding 100 more peptides than wild-type protein, despite being 60% identical to the wild-type protein.

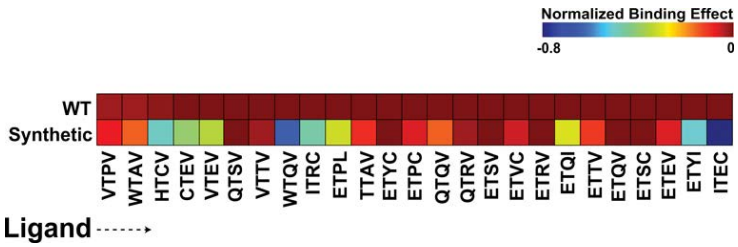

Figure 2.29: The synthetic PDZ domain binds class I peptides similar to wild-type PSD95^{pdz3} (Fig. 2.28) but notably exhibits differences in the detailed binding specificity within the class I designation. For example, the 25 ligands with the highest affinity for wild-type protein do not bind the synthetic domain with equally high affinity.

of the PSD95^{pdz3} binding space, remarkably the synthetic protein only mildly differed from the binding space of wild-type protein, accepting approximately 100 additional ligands localized to the class I specificity designation of the full binding space (Fig. 2.28). A noted difference, however, was that the detailed preference of peptides within the synthetic protein binding space differs from wild-type (Fig. 2.29). For instance, we observe that the best 25 binders to wild-type protein (greater than 0.8μM) exhibit a spectrum of affinities for the synthetic protein ranging from 15μM to greater than 0.8μM. With regard to adaptation to a significantly different selection pressure however, the result from the binding space of the synethetically designed PDZ domain in addition to the spatial distribution of adaptive mutations in the protein suggests that the adaptive capacity of PSD95^{pdz3}, with regard to overall class specificity, is contained uniquely within the protein sector.

In summary, the spatial distribution of adaptive mutations are found within the protein sector and can be categorized into two classes: class-switching and conditionally neutral. The mutations in direct contact to the adaptive challenge are found to be class switching, unable to promote adaptation without abrogating native function–a characteristic of binding pocket mutaitonal intolerance well-documented and in some cases comprehensively studied. Conditionally neutral mutations, such as G330T, are located distal to the adaptive challenge yet still within the protein sector, presumably affecting phenotype through thermodynamic coupling to the peptide. Following these results, we were able to propose an approximate structural decomposition constructed upon the well-known and seemingly paradoxical relationship between robustness and evolvability. Non-sector mutants are largely unconditionally neutral, conferring the property of robustness without contributing to the adaptive capacity of proteins. Mutations in sector positions residing away from the adaptive challenge are conditionally neutral–robust to native function but permitting the acquisition of new function.

2.6 A Mechanistic Understanding of Conditional Neutrality

The spatial pattern of conditionally neutral mutations in PSD95^{pdz3} follows previously observed trends when studying the problem of protein adaptation: mutations distant from the active site, binding site, or site of challenge heavily influence the capacity of the protein to adapt to new function. How is this? How does mutant like G330T cause the protein to be able to bind both class I and class II peptides equally well? The two-step mutational path (G330T and H372A) provided us the opportunity to answer this question. Crystal structures of wild-type and G330T PSD95^{pdz3} unbound and bound to both CRIPT and T$_{-2}$F peptides were solved to a high resolution (less than 2Å crystal conditions and refinement procedure details provided in Methods). All structures were isomorphous (identical space group and within 5% unit cell dimensions) meaning that the structural changes in the crystal structure reflect the mutations made in the protein and not external factors such as differences due to differing space groups (Fig 2.30). The wild-type protein bound to native peptide shows well-known signatures of the PSD95^{pdz3}-CRIPT peptide

Data Set	WT Apo	WT CRIPT	WT-$T_{(-2)}$F	G330T Apo	G330T CRIPT	G330T $T_{(-2)}$F
Collection/Refinement						
Source	APS	UTSW	UTSW	APS	APS	APS
Resolution	39.92-1.85	40.15-1.901	36.56-1.90	50-2.05	27.2-1.571	40.1-1.76
Space Group	P4132	P4132	P4132	P4132	P4132	P4132
Unit Cell	89.5 89.5 89.5 90 90 90	89.8 89.8 89.8 90 90 90	89.6 89.6 89.6 90 90 90	89.5 89.5 89.5 90 90 90	90.2 90.2 90.2 90 90 90	89.7 89.7 89.7 90 90 90
Unique Reflections	10988 (1069)	10255 (1002)	10141 (977)	8143 (791)	18111 (1772)	12710 (1231)
Redundancy	11.3 (11.8)	17.5 (16.7)	10.9 (10.1)	10.1 (10.2)	22.8 (21.5)	27.4 (27.4)
Completeness (%)	99.90 (100.0)	100.0 (100.0)	99.73 (99.29)	99.91 (100.0)	100.0 (100.0)	99.92 (100.0)
Wilson B-Factor	18.64	24.71	26.60	38.29	15.91	31.4
Average B-Factor	28.60	32.80	33.10	41.6	25.1	44.9
R-factor	0.1838	0.1611	0.1797	0.2154	0.1737	0.19
R-free	0.2252	0.1828	0.2045	0.2437	0.1853	0.24
Number of atoms	1102	1298	1240	1725	1361	1221
Number of waters	99	143	113	64	143	90
Protein residues	119	128	128	112	128	126
RMS (bonds)	0.013	0.015	0.015	0.017	0.020	0.008
RMS (angles)	1.36	1.68	1.62	1.20	1.83	1.07
Ramachandran favored (%)	99.2	95.1	97.0	97.5	94.0	98
Ramachandran outliers (%)	0.0	3.5	0.70	0.0	2.60	1.40

Figure 2.30

binding event. For instance, the H372 sidechain is situated within hydrogen bonding distance of the T_{-2} position on the peptide–an interaction thought to significantly contribute to the high affinity protein-peptide state (Fig. 2.31). Mutating the T_{-2} position to phenylalanine induces a spatially anisotropic propagation of structural effects. The $T_{-2}F$ mutation clashes with the H372 position causing the histidine sidechain to occupy a split rotameric conformation pointed away from the peptide. The two conformers of residue 372 interact with position 330, inducing an equivalently occupied dual conformation of the 330 residue and surrounding loop region thus causing a decreased binding affinity of the protein-peptide complex and explicitly revealing the mechanism of energetic coupling between the -2 position of the peptide and the G330 residue as occurring through position 372 (Fig. 2.31). Making the G330T mutation locks the loop into an alternate conformation creating space near the H372 region (Fig. 2.32). When bound to CRIPT peptide, the G330T protein binds with high affinity due to the H372/T_{-2} hydrogen bonding interaction (Fig. 2.31). However, when bound to $T_{-2}F$ peptide, the H372 rotamer is permitted to rotate away from the peptide without steric clash with the loop region, thereby creating a high affinity complex (Fig. 2.31). Thus the G330T mutation couples to residue H372 to create a plastic binding pocket, able to accept peptides with Thr/Ser or hydrophobic residues at the -2 position thereby generating a conditionally neutral mutant. Presumably, this mechanism of long range coupling to the binding site is a method by which a distant sector position such as 362, several contact shells away from the -2 position, can affect protein phenotype. The structural results demonstrate the importance of being able to engage in coupling with other residues, lending further evidence to our proposal that non-sector mutations are unconditionally neutral. That is, mutations at positions outside the sector generally do not engage in coupled interactions and therefore cannot influence the phenotype of the protein from afar.

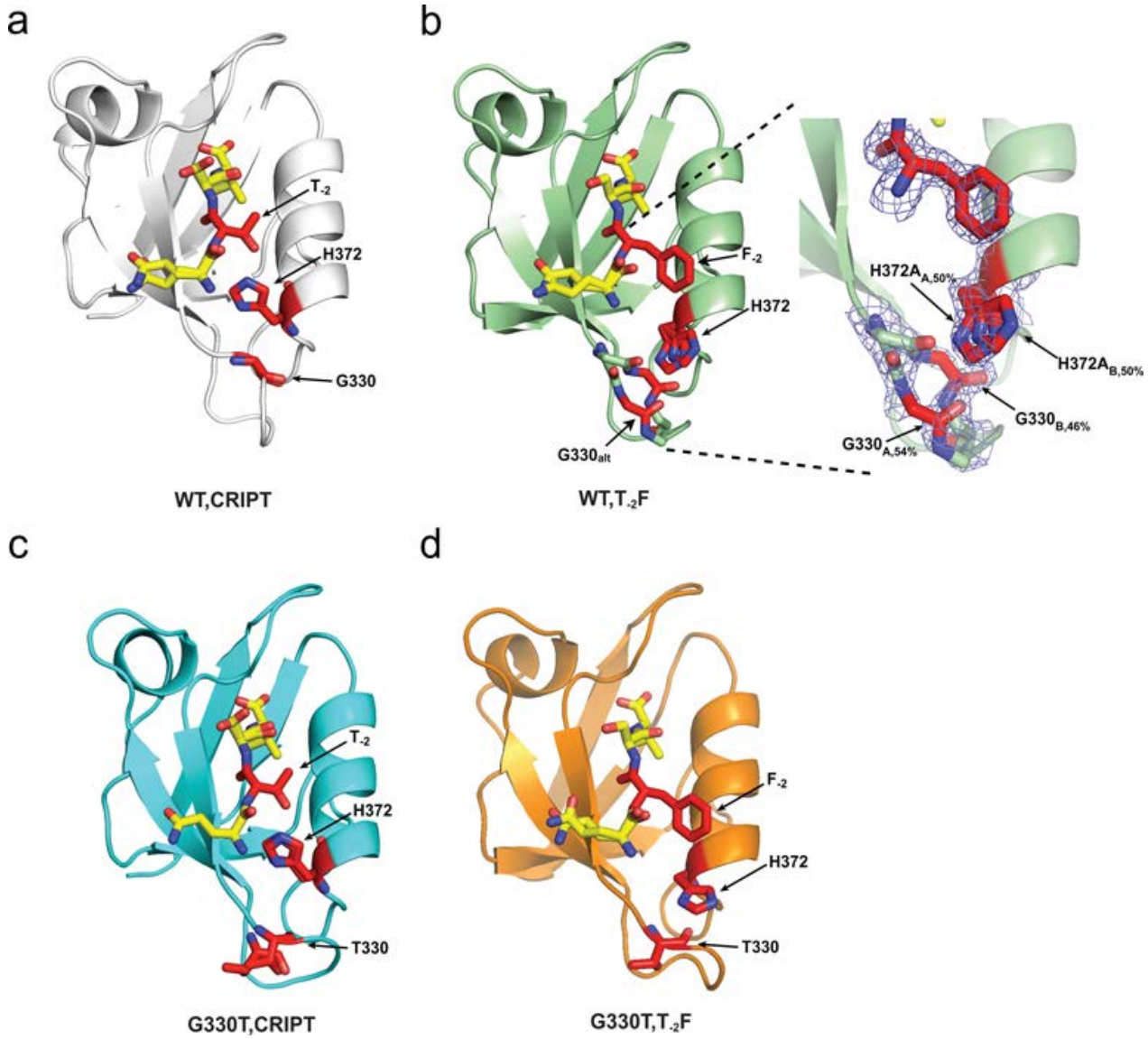

Figure 2.31: **A.** Wild-type PSD95^{pdz3} bound to CRIPT: H372 is within hydrogen bonding distance to the T_{-2} position on the ligand, an interaction thought to contribute significantly to the high affinity of this complex. **B.** The $T_{-2}F$ mutation on the peptide clashes with the sidechain of H372, causing residue 372 to rotate away from the peptide and secondarily clash with residue 330 thereby forcing the loop in which 330 resides into an alternate conformatin (a 64:46 occupancy split). The series of clashes caused by the $T_{-2}F$ mutation, presumably causes the lower binding affinity of the wild-type $T_{-2}F$ complex. This complex provides concrete physical evidence of the energetic coupling seen between the G330 residue and the T_{-2} position on the peptide. **C.** Making the G330T mutation causes the local structural environment to shift into the alternate conformation revealed by the $T_{-2}F$ mutation (Fig. 2.32. In the G330T-CRIPT complex, the H372 sidechain is within hydrogen bonding distance just as the wild-type-CRIPT complex. **D.** In the G330T-$T_{-2}F$ complex however, the H372 residue is free to rotate away from the peptide without clashing with the 329-330 region. Thus the G330T mutation permits the protein to bind both classes of peptides with high affinity.

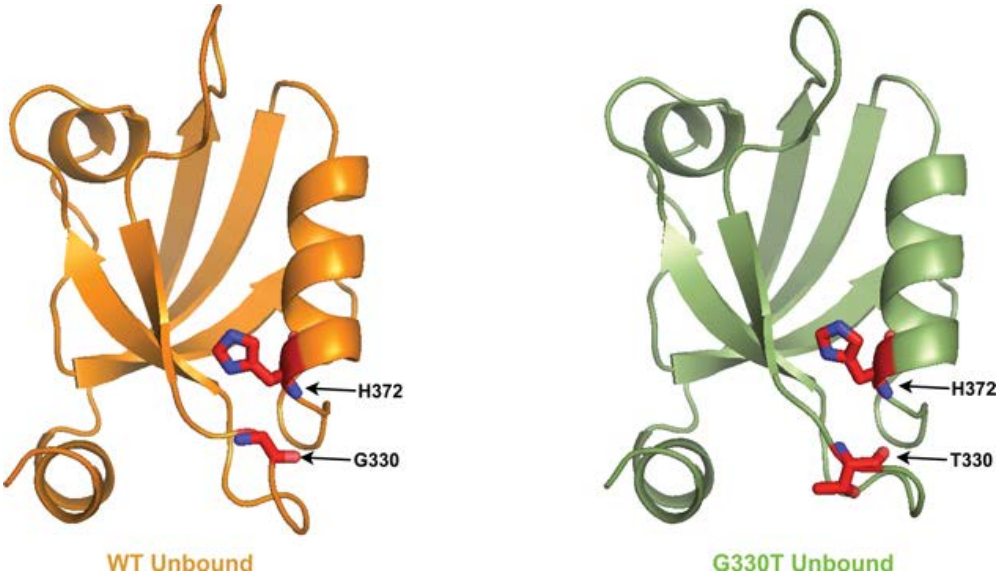

Figure 2.32: The G330T mutation causes the local structural shift, as seen by the comparison of wild-type PSD95^{pdz3} and G330T unbound.

2.7 The Protein Sector Encodes for the Adaptive Capacity of Proteins

We sought to answer the question of how are proteins designed to adapt to new functions quickly. We examined this problem in a single case-study involving a short cooperative path to new function in PSD95^{pdz3} with a generalist intermediate, G330T, removed from the binding site that was able to perform both the native and target function equally well and a direct class-switching intermediate, H372A, in direct structural contact with the adaptive challenge (Fig. 2.5). Using a high-throughput peptide screen, we found that the G330T mutant is able to serve as a conduit between functions whereas the H372A mutant abrogates native class specificity (Fig. 2.16). Which of these mutants is the favorable pathway for promoting adaptation? Using a stochastic simulation, the utility of conditionally neutral mutations such as G330T over direct class-switching mutations such as H372A was evident with G330T promoting adaptation over all possible mutation rates and environmental switch rates. The H372A mutation was found to aid in adaptation only in a particular regime of mutation rate and environmental switch rate: the stochastic regime in which adaptation can occur only after the environment changes from CRIPT to T$_{-2}$F (Fig. 2.20). Thus the results from the simulation, in addition to previous studies, motivated a protein-wide search

for conditionally neutral mutants to understand their spatial distribution in PSD95^{pdz3}. Using the saturation mutagenesis data from McLaughlin *et al.*, we found that the location of conditionally neutral mutations is within the protein sector but removed from the site of the adaptive challenge, facilitating adaptation from afar and suggesting the adaptive importance of the structural architecture of the protein sector–a contiguous network of coupled residues extending from the binding pocket to surface sites (Fig. 2.25). In addition, direct class-switching mutants are also found within the protein sector but localized to direct contact with the -2 position of the peptide (Fig. 2.25). Identifying the spatial distribution of adaptive mutants in the protein suggested to us that the sector encodes the information for the adaptive capacity of proteins. More rigorously testing this idea, a synthetic PDZ domain, designed to retain the 100% sector identity but only 60% overall identity, was designed and assayed for function. This synthetic domain performed similarly to wild-type PSD95^{pdz3}, with only subtle differences in minor specificity determinants within the class I specificity space (Fig. 2.28). The spatial distribution of conditionally neutral mutations located distal to the binding site but within the sector motivated a mechanistic examination of these adaptive mutants. We solved high-resolution structures of wild-type and G330T bound to native and target peptides and demonstrated how distant mutations can open up conformational plasticity at the binding site thereby allowing the protein to accomodate both classes of peptide equally well.

Overall, these results illustrate how a structural design of connected, spatially distributed coupling in the protein enables rapid adaptation through the ability to access conditionally neutral mutations. Going back to the motivation of this study, a major area of active research related to these results involves understanding how to rationally evolve proteins to different specificities. However, attempts to alter specificity based purely on structural intuition have largely employed trial and error (take for instance, the case of trypsin and chymotrypsin specificity),[9] lacking fundamental principles. Given that sectors are found in every protein familiy studied to date with a conserved topology of spatially heterogeneous connected networks with the protein structure, it is possible that the structural properties discussed here can serve as a guide for the identification of adaptive paths to new function–a conceptual contribution of potentialy utility for the engineering of altered or novel function.

Together with previous results in the lab, the protein sector, a statistical proxy for the complex interactions present within a protein, has been shown to encode the biochemical information for fold, for function, and for the ability to adapt to new function. A critical biophysical charateristic whose sequence origins have remained unexplained is the property of allostery–efficient long-range information transmission in the protein structure; a property intimately related to the emergent traits of fold and function. Revisiting the structural and functional properties of hemoglobin, an outstanding structural biology question originating from the original structure of Max Perutz and persisting until today is which amino acids are responsible for allosteric propagation of signal transmission. In other words, can we 'see', or somehow detect, allostery in individual proteins? Chapter 3 of this thesis will address this question.

References

[1] M. Kirschner and J. Gerhart, "Perspective Evolvability," vol. 95, no. July, pp. 8420–8427, 1998.

[2] P. a. Romero and F. H. Arnold, "Exploring protein fitness landscapes by directed evolution.," *Nature reviews. Molecular cell biology*, vol. 10, pp. 866–876, Dec. 2009.

[3] J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold, "Protein stability promotes evolvability," *PNAS*, vol. 2006, no. 103, pp. 5869–5874, 2006.

[4] J. a. Draghi, T. L. Parsons, G. P. Wagner, and J. B. Plotkin, "Mutational robustness can facilitate adaptation.," *Nature*, vol. 463, pp. 353–5, Jan. 2010.

[5] E. Ortlund, J. T. Bridgham, M. R. Redinbo, and J. W. Thornton, "Crystal Structure of an Ancient Protein: Evolution by Conformational epistasis," *Science*, vol. 317, no. September, pp. 1544–1549, 2007.

[6] E. Nimwegen, J. Crutchfield, and M. Huynen, "Neutral evolution of mutational robustness," *PNAS*, vol. 96, no. August, pp. 9716–9720, 1999.

[7] S. Bershtein, M. Segal, R. Bekerman, N. Tokuriki, and D. S. Tawfik, "Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein.," *Nature*, vol. 444, pp. 929–32, Dec. 2006.

[8] P. E. O'Maille, A. Malone, N. Dellas, B. Andes Hess, L. Smentek, I. Sheehan, B. T. Greenhagen, J. Chappell, G. Manning, and J. P. Noel, "Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases.," *Nature chemical biology*, vol. 4, pp. 617–23, Oct. 2008.

[9] L. Hedstrom, L. Szilagyi, and W. J. Rutter, "Converting Trypsin to Chymotrypsin : The Role of Surface Loops," *Science*, vol. 255, pp. 1249–1253, 1991.

[10] R. N. McLaughlin, F. J. Poelwijk, A. Raman, W. S. Gosal, and R. Ranganathan, "The spatial architecture of protein function and adaptation.," *Nature*, vol. 491, pp. 138–42, Nov. 2012.

[11] K. a. Reynolds, R. N. McLaughlin, and R. Ranganathan, "Hot spots for allosteric regulation on protein surfaces.," *Cell*, vol. 147, pp. 1564–75, Dec. 2011.

[12] D. M. Fowler, C. L. Araya, S. J. Fleishman, E. H. Kellogg, J. J. Stephany, D. Baker, and S. Fields, "High-resolution mapping of protein sequence-function relationships.," *Nature methods*, vol. 7, pp. 741–6, Sept. 2010.

[13] D. Melamed, D. L. Young, C. E. Gamble, C. R. Miller, and S. Fields, "Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein.," *RNA (New York, N.Y.)*, vol. 19, pp. 1537–51, 2013.

[14] M. A. Stiffler, D. R. Hekstra, and R. Ranganathan, "Evolvability as a Function of Purifying Selection in TEM-1 Beta-Lactamase," *Cell*, vol. 160, no. 5, pp. 882–892, 2015.

[15] H.-J. Lee and J. J. Zheng, "PDZ domains and their binding partners: structure, specificity, and modification.," *Cell communication and signaling : CCS*, vol. 8, p. 8, 2010.

[16] Z. Songyang, A. Fanning, C. Fu, J. Xu, S. Marfatia, A. Chishti, A. Crompton, A. Chan, J. Anderson, and L. Cantley, "Recognition of Unique Carboxyl-Terminal Motifs by Distinct PDZ Domains," *Science*, vol. 275, no. January, pp. 73–76, 1997.

[17] M. A. Stiffler, V. P. Grantcharova, M. Sevecka, and G. MacBeath, "Uncovering Quantitative Protein Interaction Networks for Mouse PDZ Domains Using Protein Microarrays," *JACS*, vol. 128, pp. 5913–5922, Mar. 2006.

[18] M. Niethammer, J. G. Valtschanoff, T. M. Kapoor, D. W. Allison, R. J. Weinberg, A. M. Craig, M. Sheng, C. Hill, and N. Carolina, "CRIPT , a Novel Postsynaptic Protein that Binds to the Third PDZ Domain of PSD-95 / SAP90," *Neuron*, vol. 20, pp. 693–707, 1998.

[19] C. Waddington, "Genetic Assimilation of an Acquired Character," *Evolution*, vol. 7, no. 2, pp. 118–126, 1953.

[20] E. J. Hayden, E. Ferrada, and A. Wagner, "Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme.," *Nature*, vol. 474, pp. 92–5, June 2011.

[21] J. Masel, "Cryptic genetic variation is enriched for potential adaptations.," *Genetics*, vol. 172, pp. 1985–91, Mar. 2006.

[22] A. Aharoni, L. Gaidukov, O. Khersonsky, S. McQ Gould, C. Roodveldt, and D. S. Tawfik, "The 'evolvability' of promiscuous protein functions.," *Nature genetics*, vol. 37, pp. 73–6, Jan. 2005.

[23] A. Le Rouzic and O. Carlborg, "Evolutionary potential of hidden genetic variation.," *Trends in ecology & evolution*, vol. 23, pp. 33–7, Jan. 2008.

Chapter 3

Revealing Pathways of Allostery in Proteins

3.1 Revealing Allosteric Pathways in Proteins

Hemoglobin has provided the benchmark for observing and attempting to understand allostery.[?,1–6] Following the results of Bohr and Adair, biologists naturally posed questions about the mechanism of allostery in hemoglobin, specifically the sequence origins of cooperativity.[7] How does such behavior emerge from the interactions between amino acids of the protein? It was the desire to answer this question that motivated Max Perutz to solve the crystal structure of hemoglobin.[1,2] It became clear soon after however, that the crystal structure in itself was insufficient to explain the allosteric behavior of the protein. Richard Feynman's quote, highlighted in the introduction of this thesis, astutely makes this very point. While many subsequent structural studies on hemoglobin were performed and aided our understanding of how allostery works, these findings were largely anecdotal to hemoglobin[6] and the essence of the problem still reimained: which amino acids in the protein structure are responsible for the property of allostery?

A concrete example of this point came from Johann Deisenhofer's laboratory in 2007.[8] FecA is an iron binding protein in bacteria beloning to the TonB receptor family of proteins. With respect to the problem of allostery, Ferguson *et al.* found that when FecA binds siderophore on the cytoplasmic side of the protein, the only observed structural change is in the so-called periplasmic pocket, directly on the opposite side of the protein with no observed structural shift in between (Fig. 3.1). Ferguson *et al.* thus addressed the question, how does signal from binding siderophore on the cytoplasmic side propagate to the periplasmic side? Statistical Coupling Analysis (SCA) was developed in 1999 by Lockless *et al.* as an approach to address precisely these kinds of questions–a statistical, evolutionary approach invoking no mechanism.[9] Using SCA on the family of TonB dependent receptors for instance, Ferguson *et al.* found that the protein sector connects the cytoplasmic and periplasmic sides, outlining a sparse connected network of residues (Fig. 3.2). This structural pattern has been observed in many protein families using the approach of SCA resulting in corroboration or prediction of functionally relevant allosteric sites (Fig. 3.3).[10] For instance, in the case of the PDZ family, SCA revealed a surface exposed coevolving cluster that was later demonstrated to the precise area of allosteric modulation of the Par6-PDZ domain function through Cdc42 binding.[11]

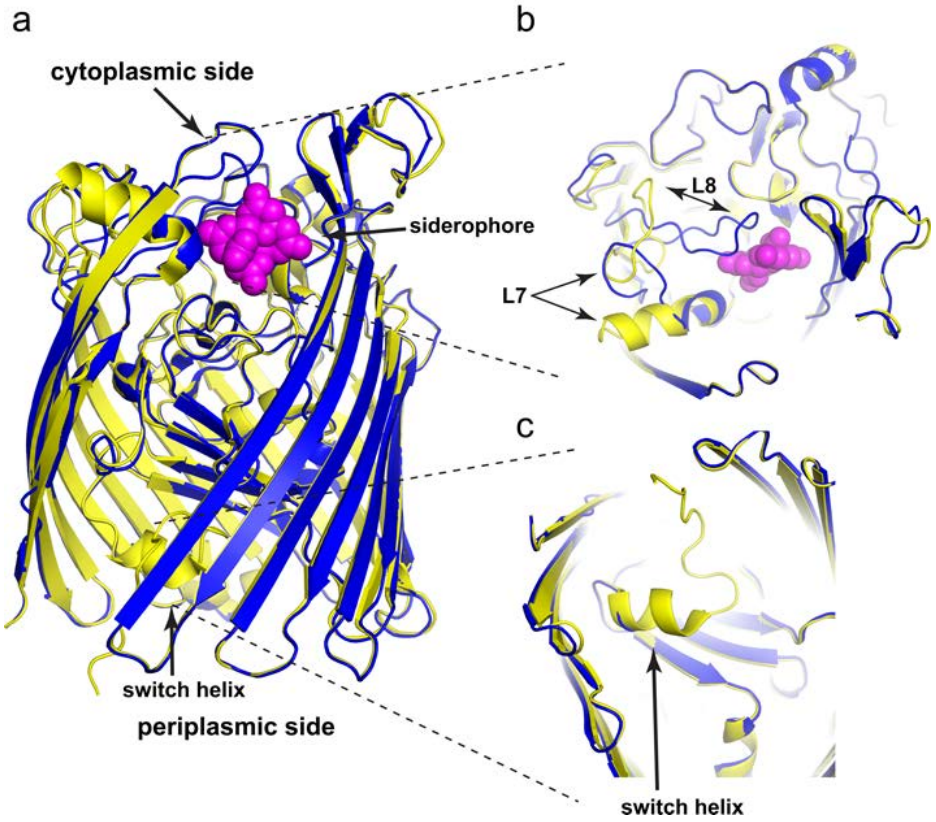

Figure 3.1: The laboratory of Johann Deisenhofer found evidence for long range communication in the bacterial iron binding protein, FecA. **A.** Upon binding to siderophore, the binding event on the cytoplasmic side of the protein elicits a structural response at a distant periplasmic helix termed the switch helix. No other structural effect is observed in the protein upon siderophore binding. **B.,C.** Zoom-ins of both the cytoplasmic and periplasmic side are shown. In panel **(B)**, local rearrangements due to siderophore binding are noticed in the L7 and L8 loops of the protein. In panel **(C)**, the distant helix rearrangement on the perisplasmic side of the protein is highlighted.

Thus using the statistics of coevolution has seemed to at least direct our thinking with respect to identifying the amino acids engaged in allosteric interactions. However, given that the protein sector is fundamentally a statistical entity, identified by analysis over an ensemble, the approach suffers a similar limitation as crystallography–we still cannot 'see' allostery in a protein. Though SCA provides a potential reduction in the sequence complexity encoding allostery, explicitly revealing the allosteric interactions within a protein is beyond the limits of this approach. The lab has thus tackled the problem of revealing allosteric pathways in proteins using two structural techniques–a global mutational scan of NMR shifts (see Alan Poole's thesis) and probing higher order structural interactions within the protein sector–the latter of which is the subject of this chapter.

As described previously, the protein sector of PDZ domains comprises a spatially heterogeneous

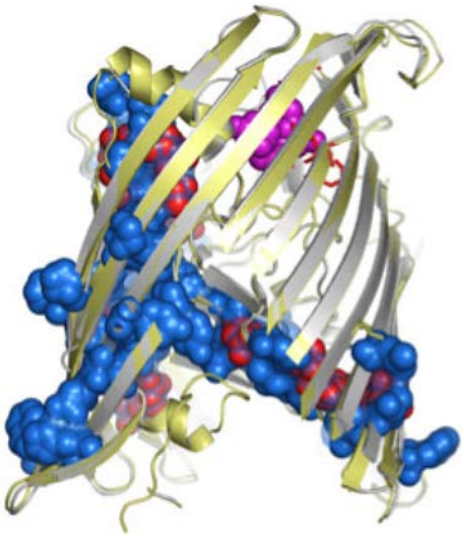

Figure 3.2: The network of statistically coupled residues for the TonB receptor family is plotted on the structure of FecA. These residues form a contiguous network of amino acids spanning the extracellular pocket, the region of siderophore binding, to the periplasmic side, the site of structural perturbation in response to the binding event.

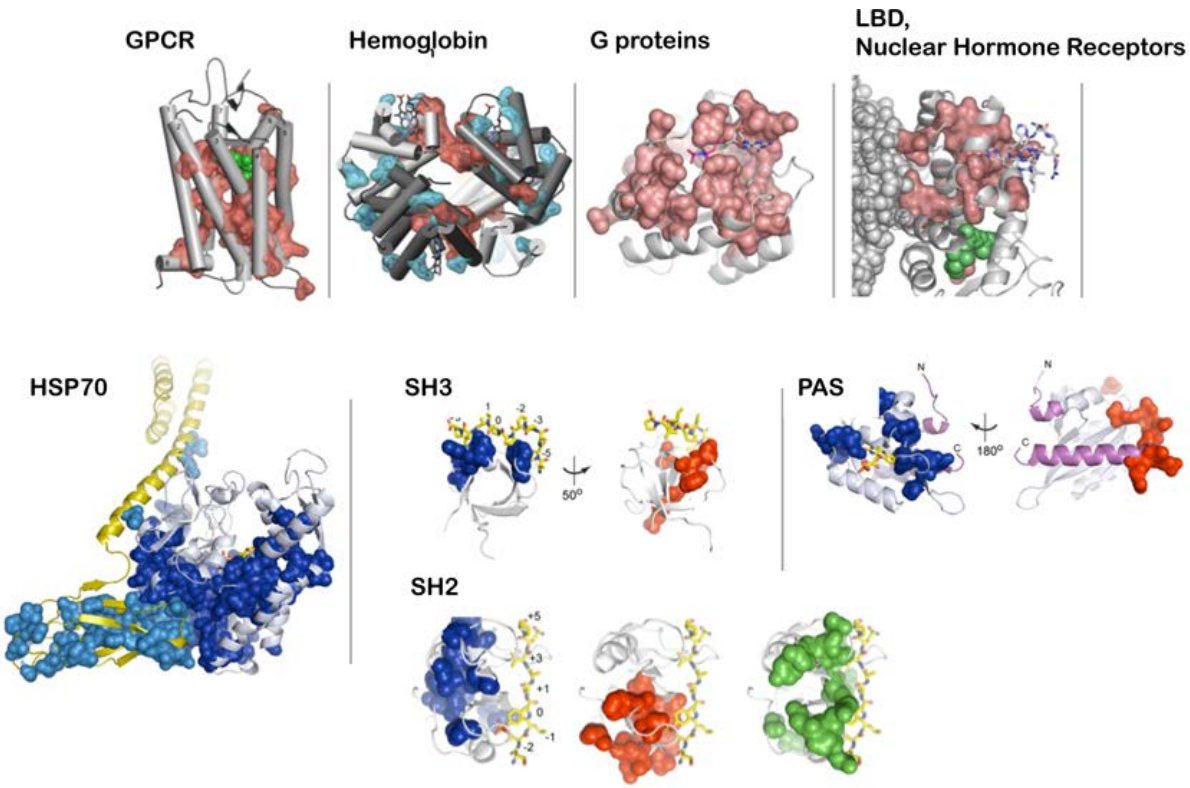

Figure 3.3: Protein sectors have generally been found in many protein families studied to date with a common structural distribution, connecting distant sites of proteins through core or binding site residues. The structural pattern of statistical couplings empirically observed suggests that the approach of statistical coupling analysis (SCA) is potentially useful in understanding the nature of allosteric communication in proteins.

subset of the binding site and extends to three surface exposed areas distant to the binding site–the β_2-β_3 loop, the α_1 helix, and residues below the β_4 strand on the backside of the protein (Fig. 3.4). Though the function of PSD95^{pdz3} has not been shown to be dependent on allosteric regulation,

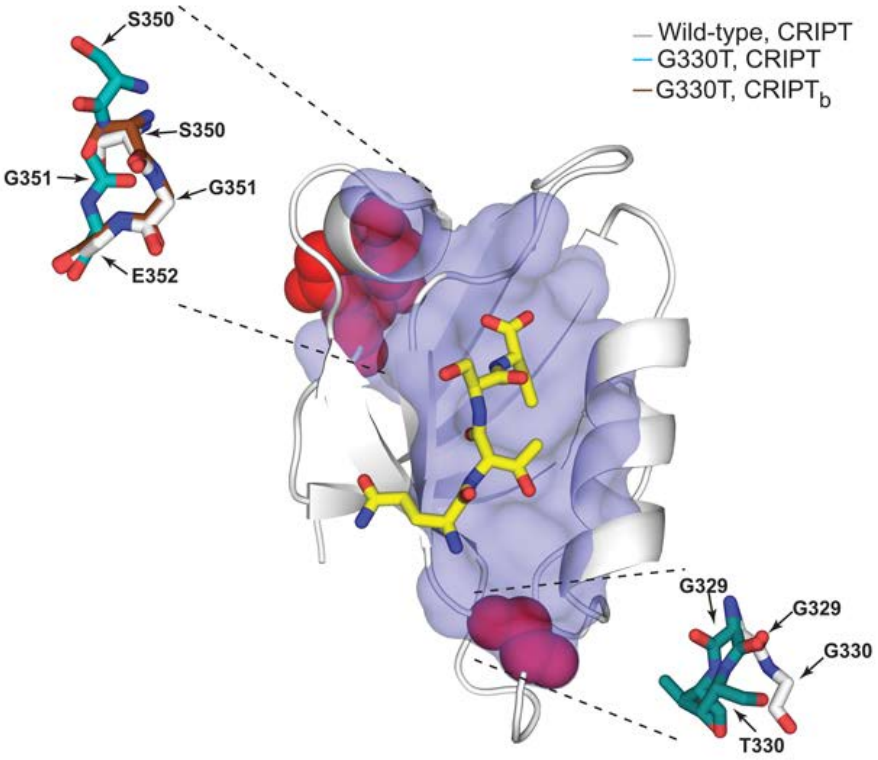

Figure 3.4: The G330T mutation, located in the β_2-β_3 loop causes no structural rearrangement in the protein except for a shift in a distant, conserved helix known as the α_1 helix. Both the β_2-β_3 loop and α_1 helix are found within the protein sector of PDZ domains (blue surface). As shown here, the sector is a spatially heterogeneous structural decomposition of the protein, spanning the β_2-β_3 loop extending through the core and binding site of the protein to the α_1 helix and backside surface exposed residues (not shown).

mutagenesis studies and sequence analysis identifying areas of conservation demonstrate that these subsets of the sector distant to the peptide binding modulate the function of the protein.[12] The structural result that motivated using this PDZ domain as a model for understanding allostery came from mutating position 330, a residue in the β_2-β_3 loop, from a glycine to a threonine. Similar to the case of FecA, upon making the G330T mutation no structural change was observed in the entire protein except for the α_1 helix located approximately 20Å away from the β_2-β_3 loop (Fig. 3.4). The G330T mutation somehow induces an alternate conformation of the α_1 helix, revealing a 70% alternate 30% wild-type backbone conformation. We therefore asked the question, can we experimentally reveal the path of allosteric communication from the β_2-β_3 loop to the α_1

helix?

3.2 Structural Coupling in the Protein Sector

The physical foundation of cooperativity lies in interactions between amino acids in the protein structure. Given that a) both the perturbation (G330T) and the observed effect (the movement of the α_1 helix) were located within the sector and b) the sector is defined as those residues coevolving with one another, the experimental approach we chose was to measure structural coupling within the protein sector to link the α_1 helix with the β_2-β_3 loop.

3.2.1 Measuring Structural Couplings in Proteins: An Overview

The measure of structural coupling is, in practice, similar to that of energetic coupling and has been used before to probe complex physical interactions in proteins.[13] The basic premise is to detect non-additive structural displacements (analogous to non-additive energetic contributions) of residues in the context of varying genetic backgrounds (Fig. 3.5). For instance, assume we have a wild-type structure and a structure with a mutation at position *i*. We can measure the structural shift of a residue *a* due to the mutation at position *i* (a His as shown in Fig. 3.5). We can also measure the structural shift of that same position due to mutation *i* made in the background of mutation *j*. The difference between the two structural shifts, $\Delta\Delta r^a_{i,j}$, is the measured non-additivity or coupling between the two mutations *i* and *j* and position *a*. A $\Delta\Delta r^a_{i,j}$ value of zero means that the movement of residue *a* is context independent with respect to mutation *i* or *j* and is therefore not structurally coupled to the interaction between the mutations (Fig. **??**). In contrast, a positive $\Delta\Delta r^a_{i,j}$ value is a measure of the context dependence of structural movement at position *a* (Fig. 3.6). Note, just like thermodynamic mutant cycles, structural mutant cycle analysis requires solving four unique crystal structures: wild-type, mutant *i*, mutant *j*, and the double mutant *i,j*. Thus, to utilize this approach, we picked two mutations within the protein sector and identified the residues engaged in structural coupling within the protein.

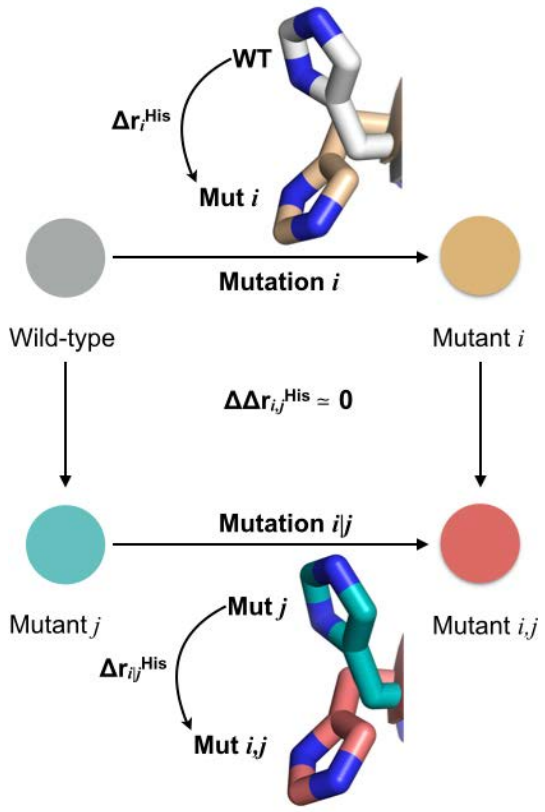

Figure 3.5: The calculation of structural coupling between two mutations requires four structures: the wild-type, both single mutants, and the double mutant. Shown here is an example of structural additivity over the mutant cycle. If the structural shift due to mutation i is the same as that observed due to mutation i in the background of mutation j, the residue is not structurally coupled to the interaction between mutation i and mutation j.

3.2.2 Measuring Structural Couplings in the Proteins: The Calculation

Structural shifts were calculated using the three-dimensional atomic coordinates from two refined protein structures. For atom a in mutant structures i and j, the structural shift due to a mutation is

$$\Delta r_{i,j}^{a} = \sqrt{(x_j^a - x_i^a)^2 + (y_j^a - y_i^a)^2 + (z_j^a - z_i^a)^2} \tag{3.1}$$

The structural shift for a whole residue was calculated simply by averaging the atomic shifts over all constitutent atoms of the residue. The $\Delta\Delta r$ value is simply the difference between the structural shift vectors (Fig. 3.7). An inherent limitation of using this calculation judge the significance of an atomic shift is that the temperature factor (also termed the B-factor) for a protein is, in generally inhomogeneous. The B-factor is a metric used to judge thermal motion or intrinsic motion due to

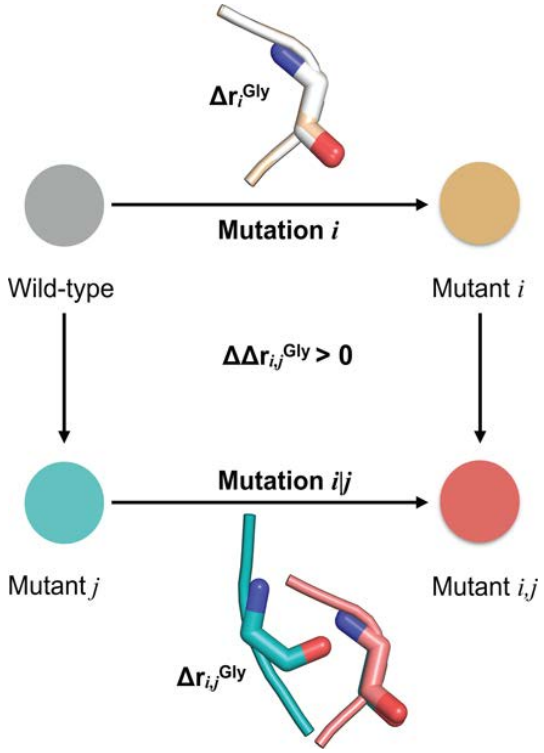

Figure 3.6: As an example of structural coupling, the glycine shown here does not move in response to making the single mutation i but moves when i is mutated in the background of mutation j. In this case, we consider this glycine to be structurally coupled to the interaction between mutation i and mutation j.

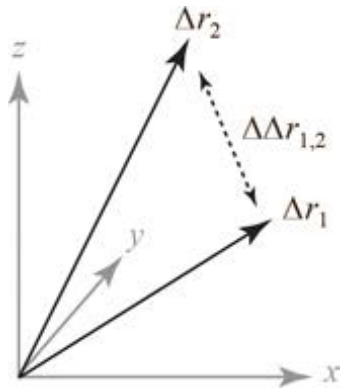

Figure 3.7: The non-additive structural shift of an atom is calculated as the difference between the structural shift due to mutation 1 and mutation 1 in the background of mutation 2.

thermal fluctuations. Certain areas of the protein structure such as surface exposed loops, may have high B-factor values indicating a disordered or 'floppy' region. Thus given a wild-type and mutant structure, the observed structural shift has to be normalized to account for B-factors: if a large structural is observed in an area with high B-factors, then this shift may be due simply to thermal motion and not a result of the mutation. Equivalently, observed structural shifts could be due to resolution differences between structures. Straud and Fauman addressed this very point,

empirically determining the normalization values for the significance of structural shifts between structures as a function of B-factor and resolution.[14] Because the β_2-β_3 loop exhibits higher B-factors than the rest of the protein, we use this method of B-factor and resolution normalization to judge the significance of a structural shift (Fig. 3.8).

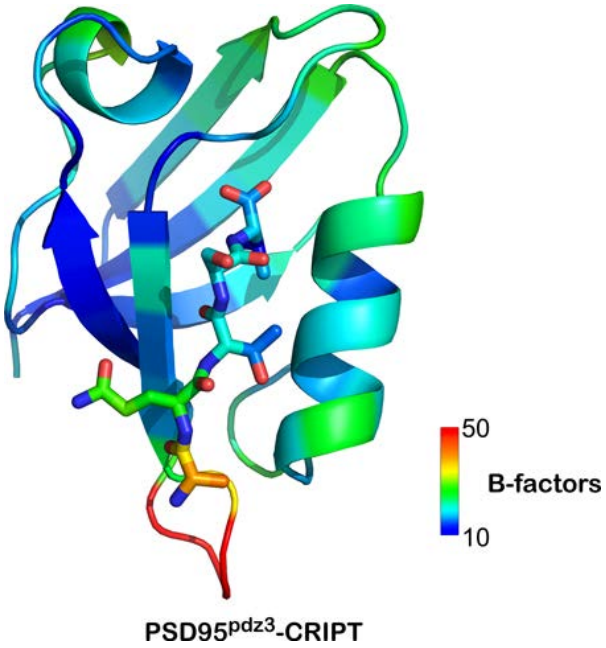

Figure 3.8: PSD95^{pdz3} is mostly well ordered with a majority of the protein having low b-factors. The only area of the protein that is inherently unstable is the β_2-β_3 oop, thus necessitating the use of b-factor normalization (see cite straud and fauman) when calculating structural shifts.

3.2.3 Measuring Structural Couplings in Proteins: The Chosen Mutations

The two mutations chosen to probe the structural coupling in PSD95^{pdz3} were the G330T mutation (influencing the backbone conformation of the α_1 helix) and the T$_{-2}$F mutation on the peptide. The $_{-2}$F mutation was chosen for two reasons: 1) the -2 position of the peptide coevolves with the PDZ protein sector (as determined by a concatenated alignment of PDZ family members with their respective specificity profiles)[15] (Figs. 3.9, 3.10) and 2) the ligand structurally bridges the α_1 helix with the β_2-β_3 loop.

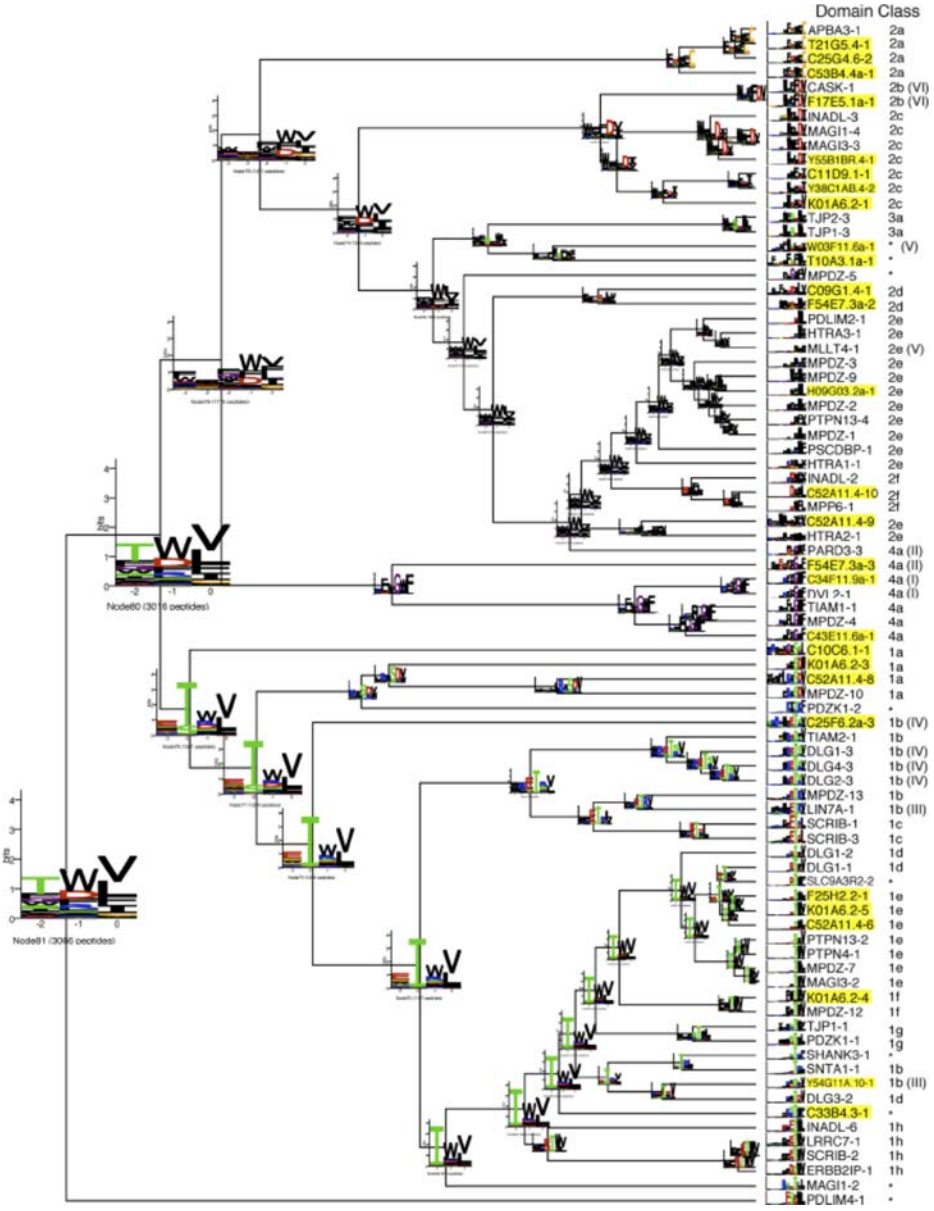

Figure 3.9: A concatenated alignment of 80 PDZ domains with their specificity profiles was constructed from Dave Sidhu's PDZ specificity data set by Bill Russ in the Ranganathan Lab. This alignment was used to measure the degree of coevolution between positions in the PDZ protein and peptide.

3.2.4 Measuring Structural Couplings in Proteins: The Results

The two chosen mutations, G330T and T$_{-2}$F define a cycle of structures shown in Figure 3.11: wild-type PSD95^{pdz3} bound to CRIPT, G330T PDZ3 bound to CRIPT, wild-type bound to T$_{-2}$F, and G330T bound to T$_{-2}$F. The structural coupling pattern observed as a result of these mutations is heterogeneous, with most positions in the protein not moving in response to the perturbations (Fig. 3.12). Of note, some positions are additive over the cycle such as residue H372 (Fig. 3.13).

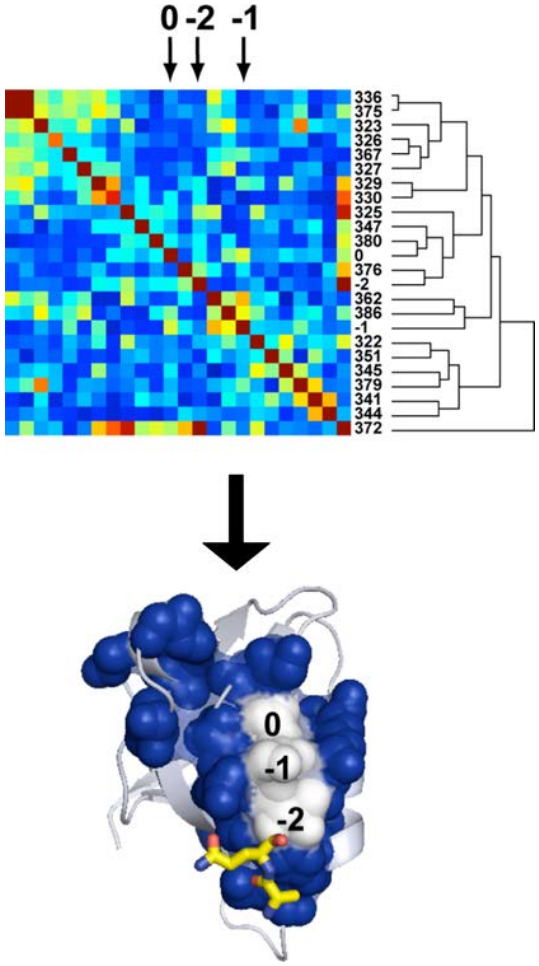

Figure 3.10: A SCA matrix was calculated and analyzed for the alignment shown in Figure 3.9 by Bill Russ. The results showed that three positions in the peptide, position zero, -1, and -2, all coevolve with the PDZ family sector.

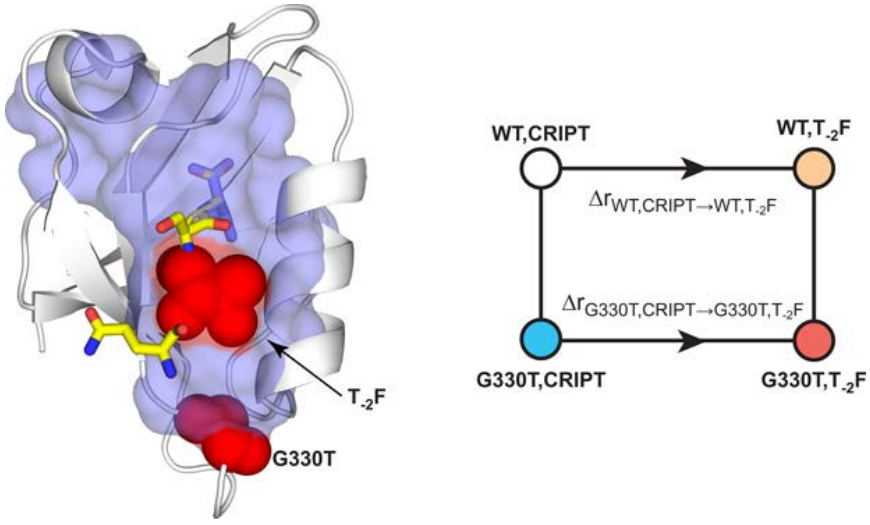

Figure 3.11: The two mutations chosen to probe structural coupling in PSD95^{pdz3} are G330T and $T_{-2}F$ (positions highlighted in red spheres on the protein structure with the protein sector in blue surface).

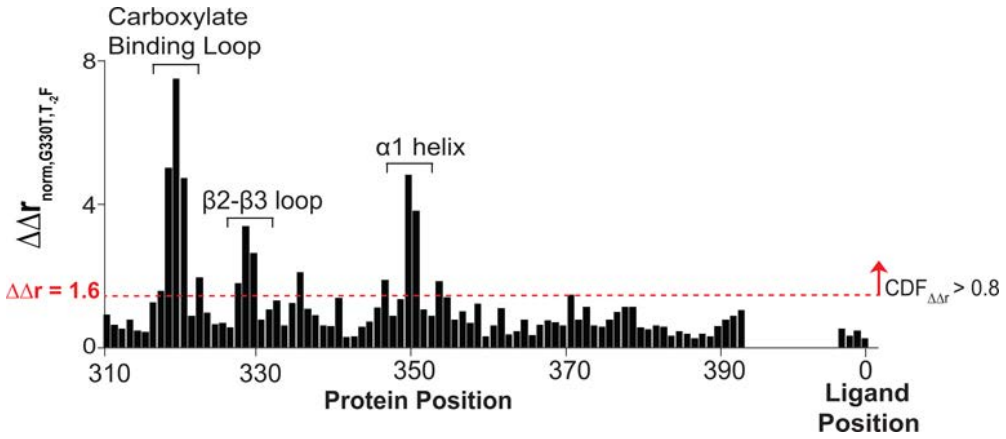

Figure 3.12: The spectrum of structural coupling in the protein exhibits sparsity and spatial heterogeneity, comprising the carboxylate binding loop, the β_2-β_3 loop, and α_1 helix, with most of the protein not exhibiting structual coupling.

In the T$_{-2}$F background, the histidine sidechain is forced into an alternate conformation pointed

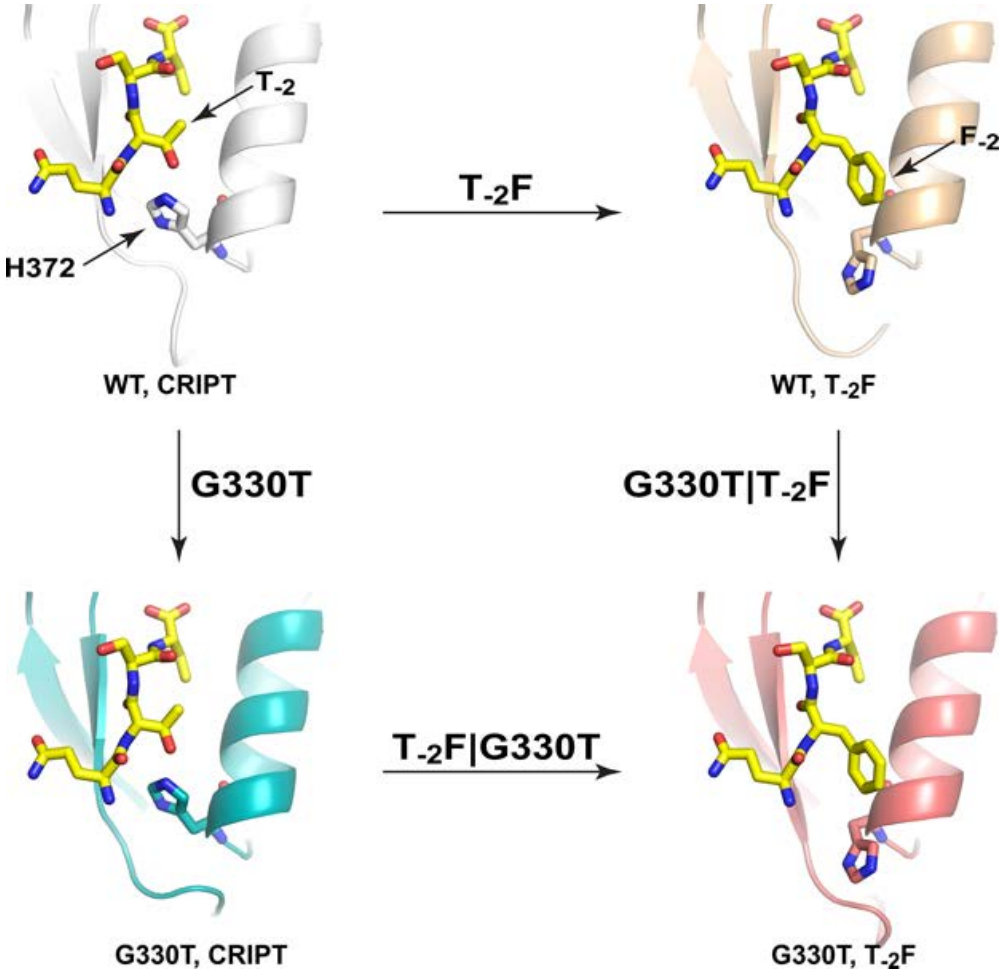

Figure 3.13: The H372A position in PSD95^{pdz3} over the structural cycle defined by the G330T and T$_{-2}$F mutations exhibits structural additivity. Upon making the T$_{-2}$F mutation, the H372 position rotates away from the peptide due to clash with the F$_{-2}$ sidechain. This effect is independent of Gly330 or Thr330; the H372 position shows the same structural effect due to T$_{-2}$F regardless of the genotype of position 330. Thus, over the cycle, the structural effect of H372 is additive.

away from the peptide (as described in the previous chapter). This structural shift, however, is not context dependent and is replicated in both wild-type and G330T genetic backgrounds.

The three areas of the protein that are structurally coupled to the G330T and $T_{-2}F$ mutations are 1) the β_2-β_3 loop, 2) the α_1 helix, and 3) the carboxylate binding loop–an area of the protein that is known to engage in extensive backbone hydrogen bonding interactions with the C-terminal zero position of the peptide sequence (Figs. 3.12, 3.14, 3.15). The β_2-β_3 loop non-additivity

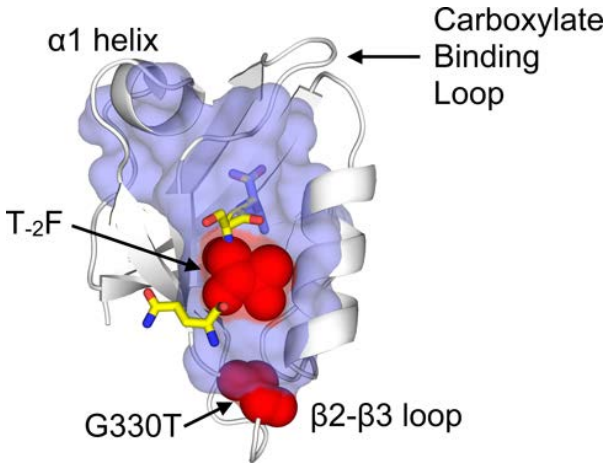

Figure 3.14: The three areas of significant structural coupling in the protein, as measured by the top 20% of structurally coupled residues in the $\Delta\Delta$r spectrum seen in Figure 3.12, are delineated here on the protein structure.

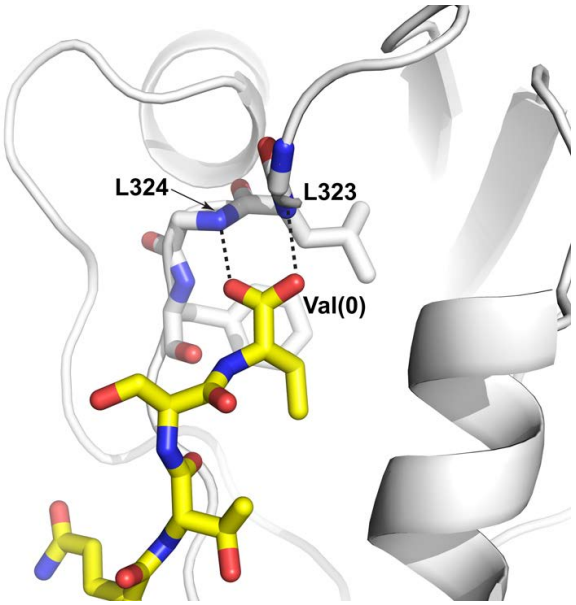

Figure 3.15: The C-terminus of the peptide engages in hydrogen bonding interactions with the backbone of a conserved GLGF motif. Shown here are the hydrogen bonding interactions between Val(0) on the peptide and L323/G324.

originates directly from the coupling of the $T_{-2}F$ and G330T mutations. The $T_{-2}F$ mutation in

the wild-type background forces the H372 position into a conformation that clashes with the β_2-β_3 loop causing two conformations fo the loop to be present (Fig. 3.16a). However, the G330T mutation itself causes the loop to occupy the alternate conformation uniquely, abrogating the native conformation. Thus, while there are two conformations of the loop upon mutating the -2 peptide position to a phenylalanine, there is only one conformation of the loop in the background of the G330T mutation (Fig. 3.16a). The non-additive structural displacement in the α_1 helix

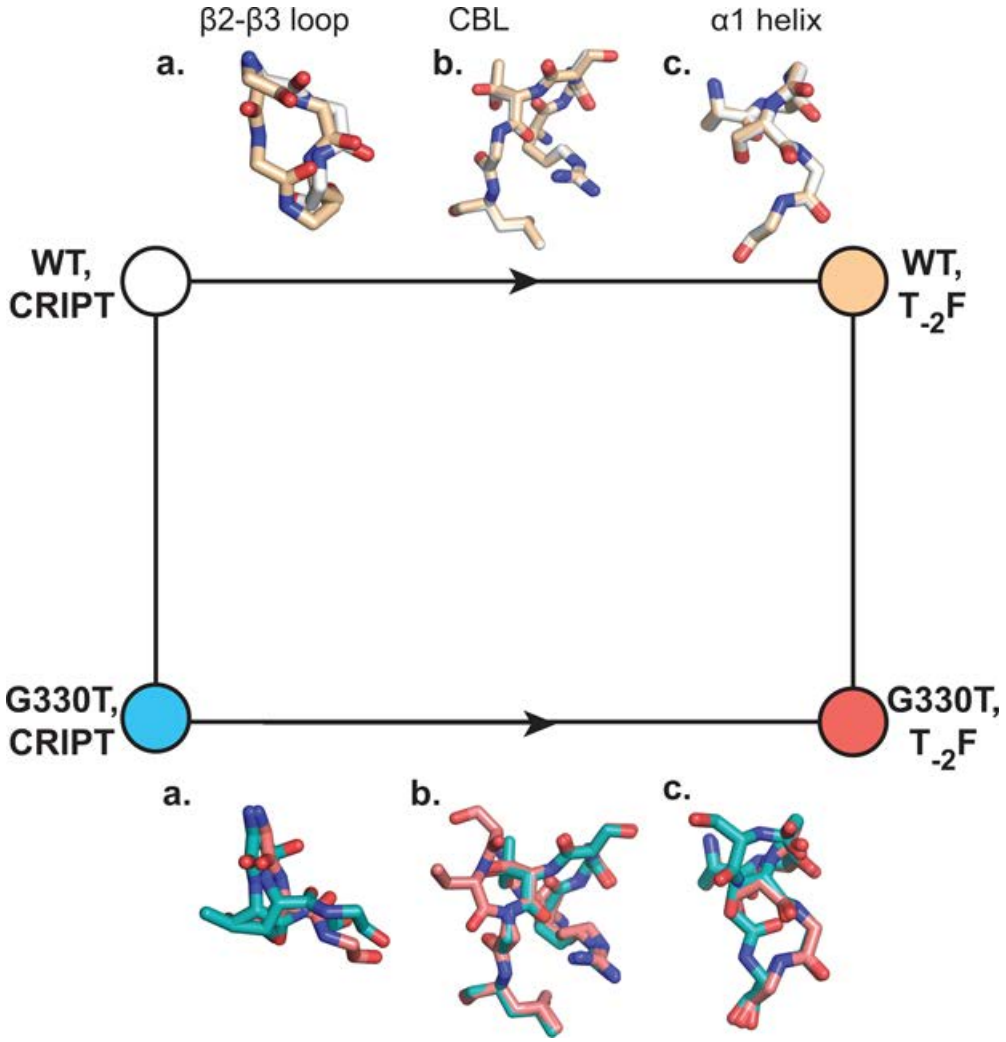

Figure 3.16: The three areas of structural coupling are highlighted here. **a.** The T_{-2} mutation causes the occupancy of an alternate conformation of the β_2-β_3 loop. Making the G330T mutation however locks this loop into a single conformation. **b.** The $T_{-2}F$ mutation has no effect on the structure of the carboxylate binding loop. However, in the background of the G330T mutation, the $T_{-2}F$ mutation reveals an alternate conformation of the loop. **c.** Similarly, the $T_{-2}F$ mutation has no effect on the α_1 helix. However, the G330T mutation in the CRIPT peptide background causes an this area to take an alternate conformation. Making the $T_{-2}F$ mutation in the background of the G330T mutation abrogates this long-range structural shift.

originates from making the G330T mutation in the CRIPT peptide background (Fig. 3.16b).

However, when the T$_{-2}$F mutation is introduced in the background of the G330T mutation, we observe that the alternate backbone conformation of the α_1 helix caused by the G330T mutation disappears. In other words, the T$_{-2}$F peptide mutation abrogates the communication from the β_2-β_3 loop to the α_1 helix (Fig. 3.16c). The structural non-additivity of the carboxylate binding loop involves indirect interactions between the T$_{-2}$F mutation, the C-terminal Val(0) of the peptide, and residues 319 through 323 (the carboxylate binding loop (CBL) itself.) The CBL is in an inherently mobile area of the protein structure, observed in an 'unclamped' conformation with high B-factors in the apo state and a 'clamped' conformation with low b-factors in the liganded state (Fig. 3.17) (see Rohit Sharma thesis). The stability of the carboxylate binding loop is induced

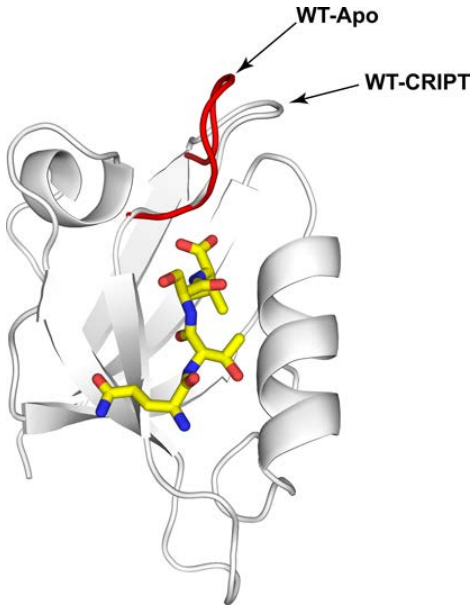

Figure 3.17: The conformation of the carboxylate binding loop is dependent on the stability of the C-terminal valine of the peptide. In the apo state, the carboxylate binding loop is in an 'open' conformation (red) with high b-factors. The presence of peptide clamps the carboxylate binding loop (white). Thus, structural fluctuations in this area can be due to structural differences in the peptide.

by backbone hydrogen bonds with the C-terminal Val(0) of the peptide (Fig. 3.15). In the G330T T$_{-2}$F structure, the Val(0) position has high B-factors indicating that the interaction between G330T and T$_{-2}$F propagates to the Val(0) position (Fig. 3.18). This C-terminal destabilization in turn causes the CBL to occupy both a clamped and unclamped conformation. Thus, while there is only one conformation of the CBL when mutating the -2 position of the peptide to phenylalanine, an alternate conformation of the CBL is revealed upon the T$_{-2}$F mutation in the background of G330T. This example thus shows how interactions between amino acids can fracture

anisotropically through the structure: a mutation at G330T and T$_{-2}$F propagates to the Val(0) peptide residue which in turn propagates to the carboxylate binding loop (CBL).

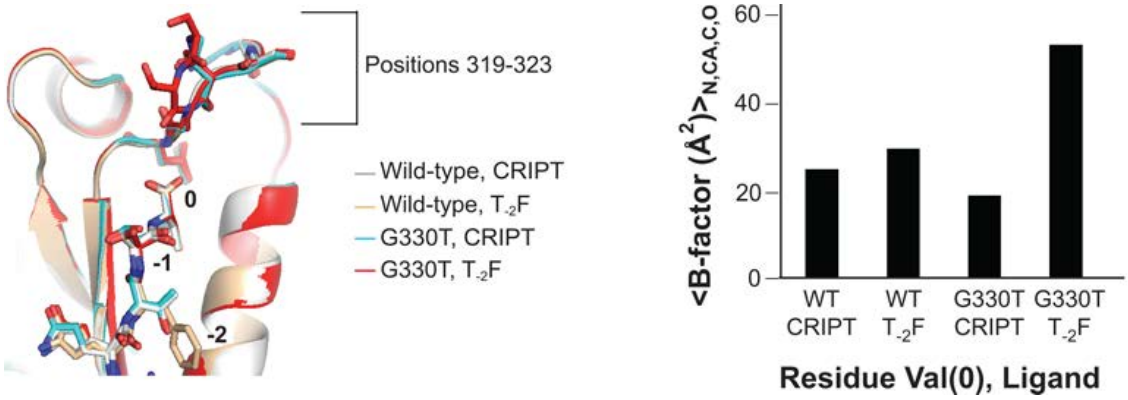

Figure 3.18: The interaction between the G330T and T$_{-2}$F mutations propagate to the carboxylate binding loop. An overlay of all four structures defined by the structural cycle shows that th eCBL of the two single mutants is identical to wild-type whereas the CBL occupes an laternate conformation in the double mutant state. The structural shift of the CBL is due to the increased b-factors of the Val(0) peptide position. This result is consistent with previous apo and liganded PDZ structures demonstrating the coupling between the CBL and the stability of the C-terminal residue of the peptide.

We define the structurally coupled residues in the protein as those above the 80% value according to the cumulative distribution (although all results discussed are robust to structural coupling cutoff). Plotting these positions on the protein structure reveals a pathway that connects the β_2-β_3 loop to the α_1 helix through the peptide and the carboxylate binding loop, a path that is located in the core of the protein sector (Fig. 3.19). While non-sector positions are also structurally coupled, these positions are non-randomly distributed in the protein structure, clustered around the protein sector such as residues 319, 320, 321 surrounding the carboxylate binding loop. This result suggests that the protein sector contains the capacity to propagate information (such as a sequence perturbation like G330T) over a long distance mediated by coupled physical interactions in the protein. Given this result, it is paramount to somehow check whether this pathway is a path of communication between distant sites in PSD95^{pdz3}.

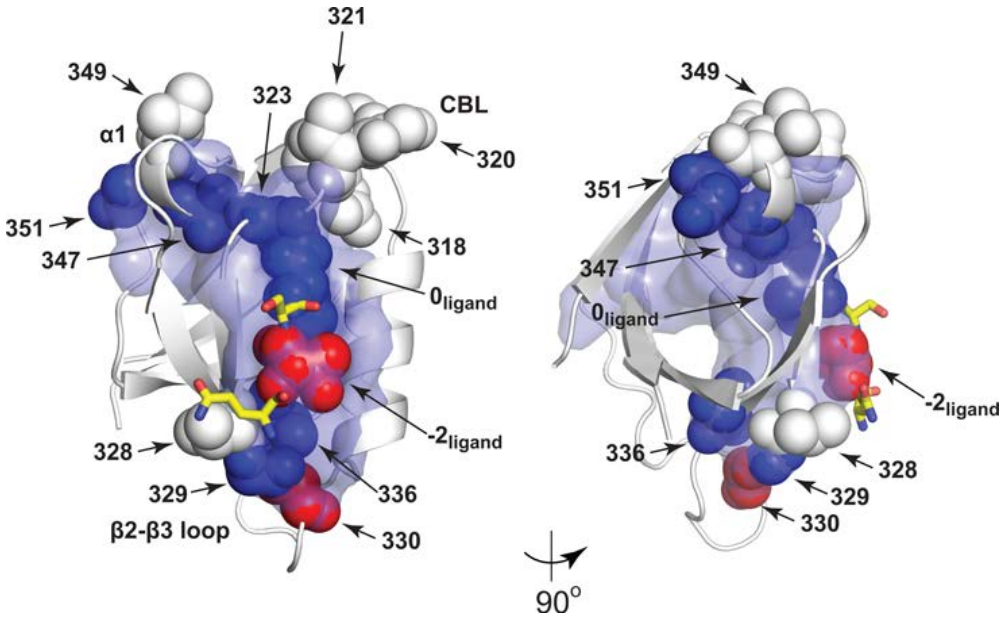

Figure 3.19: Shown here in spheres are the top 20% of structurally coupled residues in PSD95^{pdz3} as revealed by the G330T and T$_{-2}$F mutations. The blue spheres are structurally coupled sector residues; white are structurally coupled non-sector residues; and in red are the G330 and T$_{-2}$ positions. The structurally coupled residues in the protein form a continuous path of positions spanning the β_2-β_3 loop, through a few positions in the core, to the carboxylate binding loop (CBL), and to the α_1 helix.

3.3 Breaking the Coupling Between the β_2-β_3 loop and α_1 helix

Recall that the structurally coupled residues in the protein were identified by making two mutations, G330T and T$_{-2}$F. Thus, a way we could check the validity of the path is by making a mutation in between the 330 and -2 positions. If allosteric transmission is dependent on the interaction between these positions, then a mutation in between the positions could break the interaction, resulting in an abrogation of the allosteric signal. Residue H372 is situated directly between positions 330 and -2. We mutated the H372 position to A372 and solved the same set of structures (wild-type and G330T bound to both peptides) but now with all complexes in the genetic background of A372 (Fig. 3.20(a,b)). The structural coupling resulting from this cycle of structures would thus indicate which residues are structurally coupled to the G330T T$_{-2}$F interaction in the background of H372A.

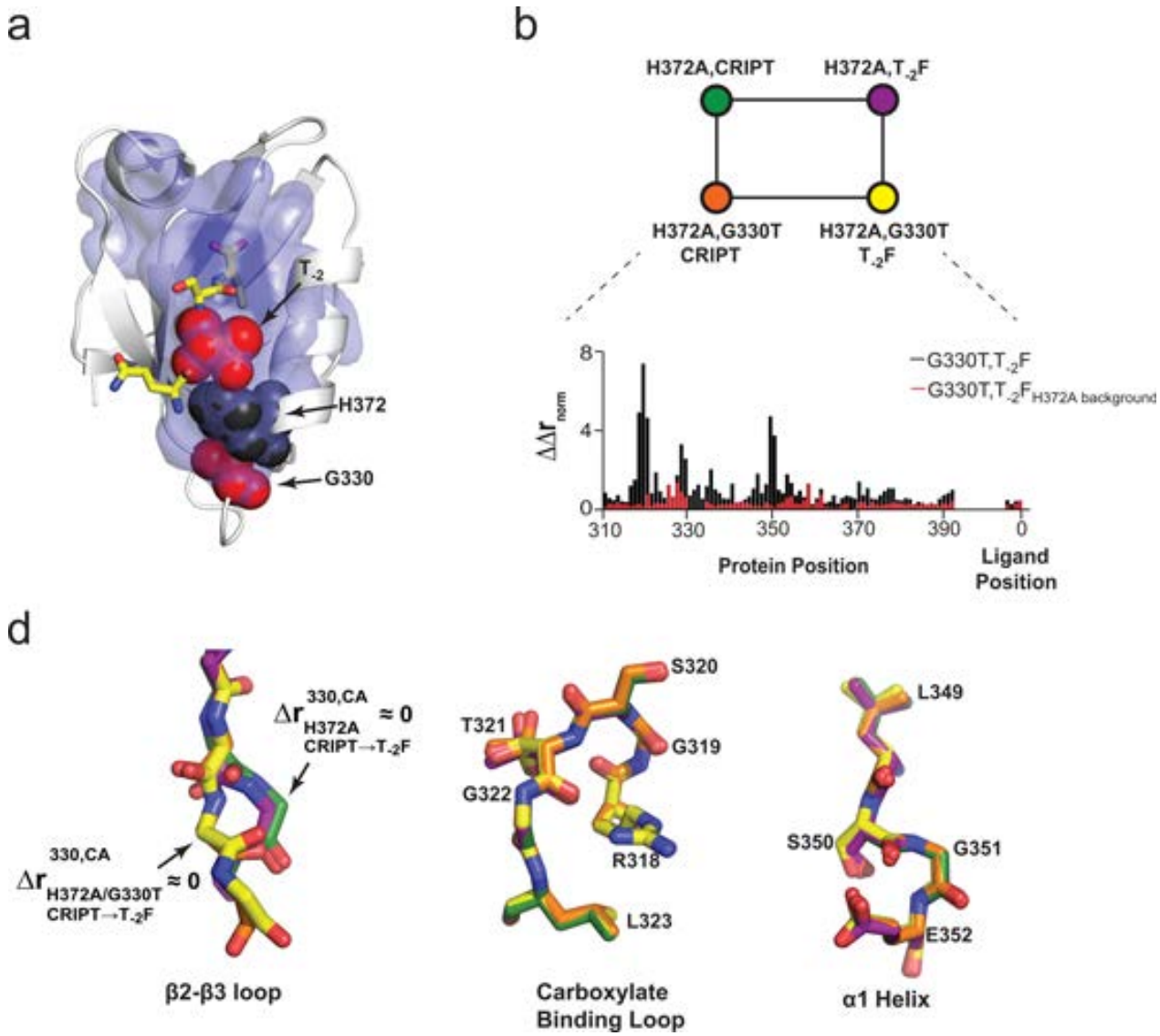

Figure 3.20: a. The H372 position lies directly between the G330 and T_{-2} positions thereby providing a good target for perturbation to check the validity of the identified adaptive path. **b.** The four structures that define the closed path of structural coupling in the background of H372A are shown in the mutant cycle. The structural coupling of the H372A mutant cycle (red) is overalyed over the wild-type structural coupling (black). The structural coupling that was present in the wild-type protein is quenched in the background of H372A. **c.** The β_2-β_3 loop is now additive over the cycle and the α_1 helix and carboxylate binding loop exhibit no structural movement.

3.3.1 The Structural and Energetic Effect of H372A

The structural coupling spectrum of the G330T,$T_{-2}F_{H372A}$ cycle is overlayed on the wild-type structural coupling spectrum in Figure 3.20(b). Globally, the H372A mutation quenched the structural coupling signal, with no area in the protein exhibiting strong structural coupling to the interaction between the G330T and $T_{-2}F$ mutations. The β_2-β_3 loop exhibits additive displacement over the cycle while the CBL and the α_1 helix show no displacement as was exposed in the wild-

type background (Fig. 3.20**c**). Though this result is consistent with the identified path of allostery, it is possible that the H372A mutation is globally distrubing the protein, nonspecifically affectings its allosteric propensity instead of selectively disturbing the communication between T330 and T_{-2}. Ideally, the H372A would specifically be 'breaking' the allostery of the protein. Revealing the energetic pattern of coupling to the $T_{-2}F$ mutation in the background of H372A would inform us about the effect of the mutation on the protein. If, for instance the H372A mutation was acting in a more local way, selectively affecting the proximal pattern of ocupling, then while the β_2-β_3 loop may no longer be coupled to the $T_{-2}F$ mutation, the α_1 helix would exhibit the same coupling pattern to peptide as in the wild-type protein. However, if H372A was acting in a more global, nonspecific fashion, then the coupling pattern across all positions in the protein may be affected. We already get a clue about the energetic effect of H372A on the coupling between mutations G330T and $T_{-2}F$ from the binding affinities of each of the proteins for CRIPT and $T_{-2}F$ peptides as measured by fluorescence polarization (Fig. 3.21). The G330T and $T_{-2}F$ mutations are highly

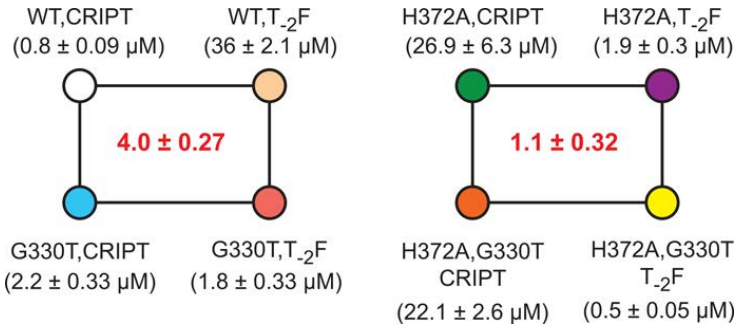

Figure 3.21: The coupling energy between G330T and $T_{-2}F$ is measured in the wild-type and H372A backgrounds. The coupling values were calculated as described in the 'Introduction' section. In the wild-type protein, G330T is highly coupled to the $T_{-2}F$ mutation. However, the H372A mutation decreases this coupling.

coupled in the wild-type protein–a result consistent with the structural effect of the $T_{-2}F$ mutation on the backbone position of the β_2-β_3 loop. However, in the background of the H372A mutation, this coupling goes away. So, at least in the case of the G330T and $T_{-2}F$ mutations, the H372A mutation appears to have a marked effect on coupling energies.

As mentioned, in the introduction, the coupling beween two mutations in the protein is calculated by measuring the conditional effect of mutation i in the background of mutation j. As McLaughlin *et al.* showed, to get a global measure of the thermodynamic coupling of the protein to the $T_{-2}F$ mutation, all single mutations of $PSD95^{pdz3}$ were made and assayed against binding

to CRIPT and $T_{-2}F$ peptides.[12] The conditional binding effect of any particular mutation in the background of $T_{-2}F$ was then calculated as the coupling of that mutation with the $T_{-2}F$ mutation. Similarly, to measure the energetic coupling in the protein to the $T_{-2}F$ mutation but with an H372A genetic background, we constructed a library of all possible single mutations of H372A $PSD95^{pdz3}$ and assayed this library against the CRIPT and $T_{-2}F$ peptides using the bacterial-2-hybrid system with eGFP serving as the transcriptional readout of the assay (Figs. 3.22, 3.23, 3.24). In the

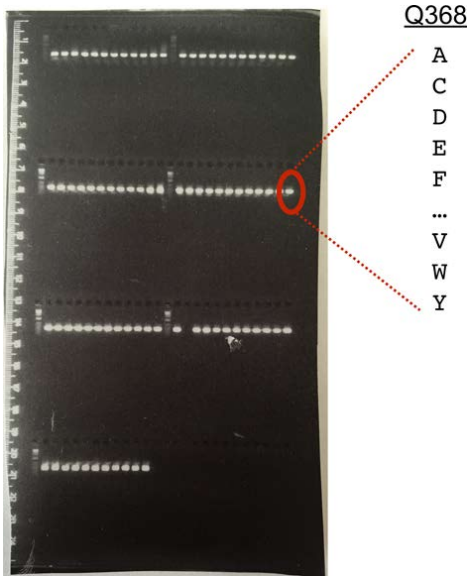

Figure 3.22: $PSD95^{pdz3}$ contains 100 amino acids, each of which can be mutated to 19 other amino acids. This gel represents all possible single mutations in the background of H372A $PSD95^{pdz3}$ with each band representing a library of all possible single mutations at a particular position. For instance, Q368 can be mutated to alanine (A), cysteine (C), etc. Note that the gel column for residue 372 is empty. This is because in this library, residue 372 is set to an alanine and not allowed to vary.

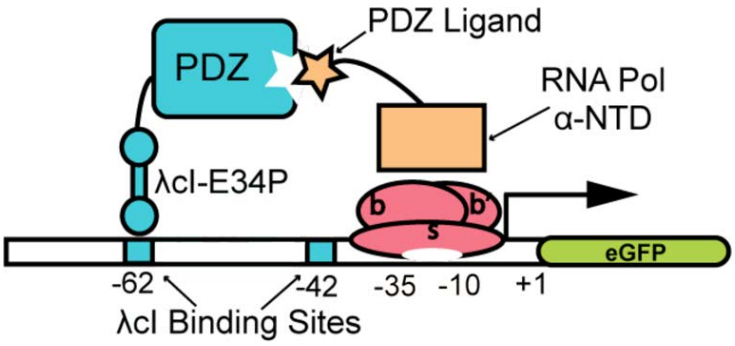

Figure 3.23: Shown here is the bacterial-2-hybrid assay exactly as designed by Rick McLaughlin for assaying a library of PDZ domains against binding peptide. The PDZ-peptide interaction is coupled to transcription of eGFP. McLaughlin *et al.* showed a linear relationship between K_d and fluorescence for a set of induction conditions and assay parameters.

context of both peptides, we find that the mutational sensitivity of H372A $PSD95^{pdz3}$ is sparse

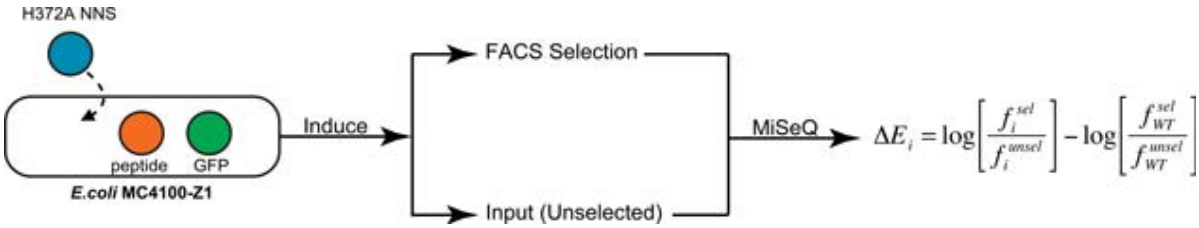

Figure 3.24: A library of PDZ mutants is transformed into the MC4100Z1 *E.coli* cell line already containing the peptide and GFP plasmids. The system is induced and split into the input or unselected population and the population to undergo selection. Selection is performed by FACS sorting (core facility, UTSW). Both the input and selected populations are mini-prepped and prepared for sequencing on the MiSeq Illumina platform. Enrichment values for each mutant PDZ allele are then calculated based on the frequency of appearance in the selected population relative to the unselected population normalized by the frequency of wild-type allele.

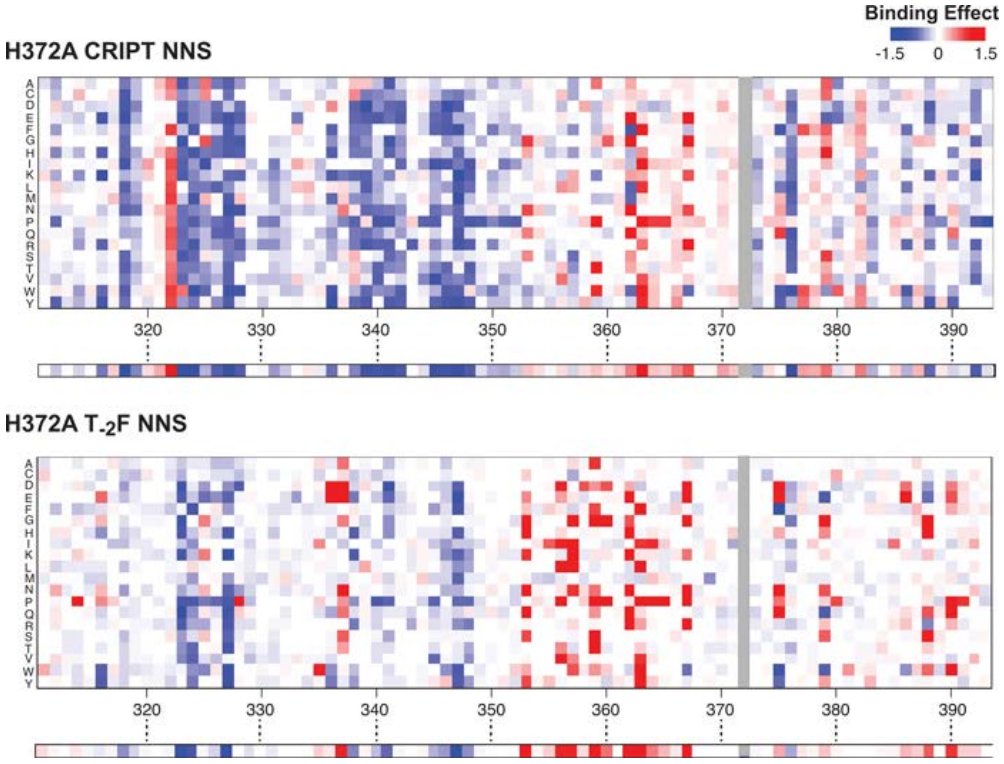

Figure 3.25: A library of all possible single mutations in the background of H372A PSD95^{pdz3} was assayed against the CRIPT and T$_{-2}$F peptides. **a.,b.** The mutational sensitivity of the H372A protein shows sparsity with respect to binding either peptide. H372A has a 27μM binding affinity towards CRIPT peptide. Many mutations in the H372A background at positions 320 through 330, and 340 through 350 (the β_2-β_3 loop through the α_1 helix respectively) show extreme sensitivity to mutation towards binding CRIPT peptide. In contrast, there exists much more neutrality to mutation for binding the T$_{-2}$F peptide. Only a few positions show mutational sensitivity (blue pixels) or gain of function (red pixels) with a vast majority of mutations exhibiting no effect (white pixels) towards T$_{-2}$F. (Note, the 372 position is grayed out because this position is set to an alanine).

with a subset of positions affecting binding affinity of the protein for peptide (Fig. 3.25). Figures 3.26 and 3.27 show the mutational sensitivity of the α_1 helix and β_2-β_3 loops when assayed for binding CRIPT or T$_{-2}$F peptides. Zooming into the two areas of interest, the α_1 helix and the

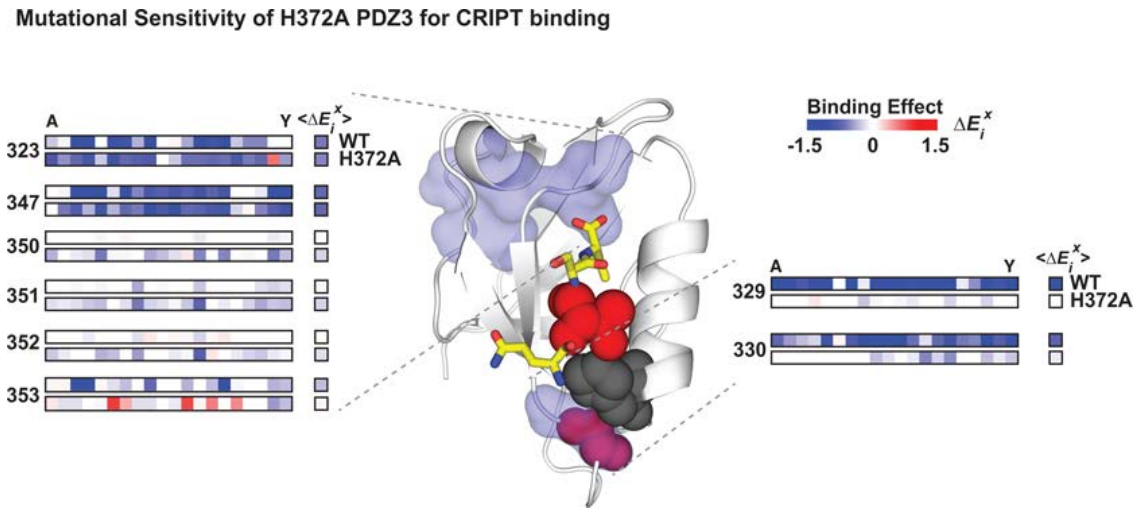

Figure 3.26: The α_1 helix shows a relatively similar pattern of mutational sensitivity to CRIPT binding in both wild-type and H372A backgrounds. Positions 323, 347, and 353 are sensitive to mutation whereas positions 350 through 352 are generally robust to variation in the wild-type protein. This same trend is observed in the H372A protein. In contrast, the β_2-β_3 loop, while sensitive to mutation in the wild-type background, shows no effect upon mutation in the H372A background.

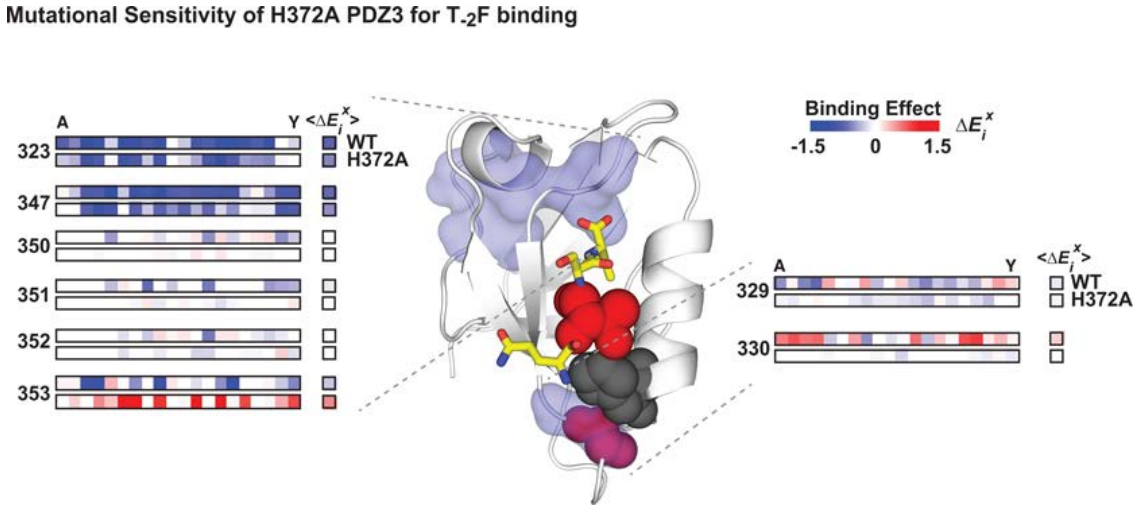

Figure 3.27: Similar to the CRIPT peptide case, here the α_1 helix retains a relatively similar pattern of mutational sensitivity to T$_{-2}$F binding in both wild-type and H372A backgrounds. However, the β_2-β_3 loop is insensitive to mutation.

β_2-β_3 loop, we see a similar trend in both the CRIPT and T$_{-2}$F case. The α_1 helix generally shows the same pattern of mutational sensitivity to binding peptide in the background of wild-type and the H372A mutation. Positions 323, 347, and 353 are sensitive to mutation whereas 350 through 352 are mostly neutral in both wild-type and H372A PSD95$^{\text{pdz3}}$. However, the β_2-β_3 loop loses sensitivity to mutation in the background of H372A, with almost all mutations exhibiting an effect in the wild-type background but no mutations conferring an increase or decrease in binding affinity in the H372A background (Fig. 3.26). The same trend is observed in the T$_{-2}$F case (Fig. 3.27).

The energetic coupling to the T$_{-2}$F mutation in the wild-type and H372A background for the α_1 helix and β_2-β_3 loop is shown in Figure 3.28). The coupling of all single mutations to the T$_{-2}$F

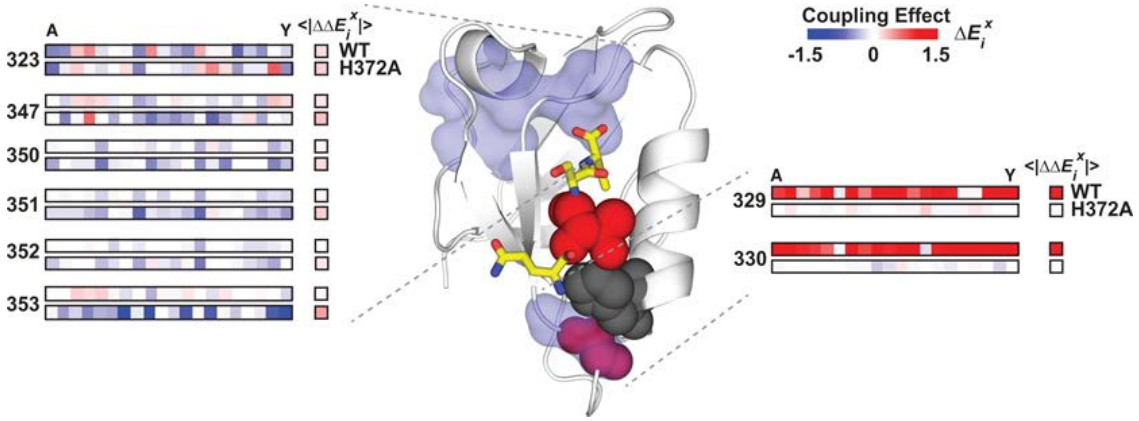

Figure 3.28: The energetic coupling of the protein to the T$_{-2}$F mutation is calculated by subtracting the binding effect matrices of CRIPT from T$_{-2}$F, the same procedure as employed in McLaughlin *et al.*. The α_1 helix remains coupled to the T$_{-2}$F peptide mutation in the background of H372A whereas the β_2-β_3 loop is disengaged and uncoupled from the T$_{-2}$F mutation.

mutation was calculated by subtracting the two binding effect matrices. In sum, these results demonstrate that the H372A mutation selectively disengages the β_2-β_3 loop but does not affect the ability of the α_1 helix to communicate to the peptide. Additionally, we are able to directly observe the importance of local couplings in the protein structure for long-range propagation of signal across a protein. In this particular case, without the histidine at position 372, allosteric propagation from the β_2-β_3 loop is not possible.

3.3.2 Conclusion and Future Directions

The work outlined in this chapter focused on the fundamental and unsolved problem in structural biology of 'seeing' allostery in proteins. The empirical result that the G330T mutation in PSD95^{pdz3} causes no significant structural shift except at a distant, conserved helix (the α_1 helix) in the protein provided for us a unique opportunity to address this problem in a well-defined and experimentally tractable model system. Because both of these positions are in the protein sector of the PDZ family, and the sector is a network of statistically coupled residues, the experimental approach that we took to reveal the path of allostery in the protein was to probe structural couplings within the protein as revealed by two mutations in the protein sector: G330T and T$_{-2}$F. We found that three areas of the protein exhibit strong structural coupling to these mutations: the

β_2-β_3 loop, the carboxylate binding loop (CBL), and the α_1 helix. Interestingly, all three of these areas were found to be within the protein sector of the PDZ family. Additionally, the non-sector structurally coupled residues were positioned directly around the protein sector, distributed in a non-random fashion throughout the protein. Finally, we showed that by mutating a residue situated directly between the G330 and T_{-2} positions, we could break the allosteric communication from the β_2-β_3 loop to the α_1 helix.

Overall, these results a) describe a method by which allosteric pathways in proteins can be accessed experimentally and b) provide the first structural evidence for protein sectors in individual proteins. With regard to (a), recently other methods have been deveoped to address the problem of detecting allosteric networks such as the work from James Frasier (QFit and room temperature crystallography),[16,17] or ensemble refinement procedures in the Phenix software package,[18] and NMR based methods such as CHESCA[19] and detecting correlated shifts in NMR spectrum due to perturbation (Alan Poole, Ranganathan Lab). To our knowledge, our results are the first to use higher order measurements (i.e. pairwise couplings) and x-ray crystallography to address the problem. It will be important in the future to test the generality of our result. For instance, perhaps the sufficient experiment to see allostery is a pairwise structural cycle within the protein sector of a member within a protein family. Only further tests in more protein families can test this assertion.

With regard to directly observing sectors, the sector is by definition a statistical entity defined by analysis over a multiple sequence alignment. Given this fact, there have historically been two questions commonly asked about the protein sector: 1) What is the extent to which the nature of the ensemble is captured within any particular member of the family and 2) How can sectors be exposed in individual proteins? Our results, along with the results of McLaughlin *et al.*, would suggest that a good portion of information contained within the ensemble is also within any individual member of the family. In retrospect, this should not be all that surprising given that the sector is calculated as *conservation weighted* correlation. That is, the conserved properties of the alignment are heavily weighted as those whose interactions are important. Thus, one would almost expect that each member of the family contain a good portion of what is revealed by the ensemble. However, revealing the sector in an individual protein could be a truly difficult

experiment. We have shown here, through a crystallographic approach, that pairwise or higher order interactions within the protein can reveal a large part of the protein sector. McLaughlin *et al.* also showed the same result through probing functional epistasis in the protein to the $T_{-2}F$ mutation. In both cases, single order measurements (for instance, the structural shift due to a single mutant or the mutational sensitivity of all single mutants to binding CRIPT peptide) did not provide sufficient evidence to 'see' the sector or the sector was better identified by probing higher order, coupled observables.[12] Again, this fact is not surprising given that the sector is representative of correlations in the protein family and therefore any experiment meant to reveal the sector should, in principle, selectively probe the interactions between amino acids.

The protein sector has provided a useful parametrization to describe the decompostion of energetic coupling in the protein, at least between pairwise coupled and uncoupled residues. Can the sector description be similarly useful in understanding the material decomposition of a protein? In a more fundamental formulation of this question, what kind of material are proteins? Are they a homogeneous material or, in the similar vein of the observed pattern of energetic coupling, a heterogeneous, hierarchically organized, complex material? Does the sector provide a framework to understand the composition of proteins as a substance? These are a subset of the natural questions that arise from our results of being able to physically expose the sector in a single protein and are a foray into more generally understanding the physical properties of complex systems.

References

[1] H. Muirhead and M. F. Perutz, "Structure of Haemoglobin. a Three-Dimensional Fourier Synthesis of Reduced Human Haemoglobin At 5-5 a Resolution.," *Nature*, vol. 199, pp. 633–638, 1963.

[2] M. F. Perutz, "Stereochemistry of cooperative effects in haemoglobin.," *Nature*, vol. 228, no. 5273, pp. 726–739, 1970.

[3] A. Szabo and M. Karplus, "A mathematical model for structure-function relations in hemoglobin.," *Journal of molecular biology*, vol. 72, no. 1, pp. 163–197, 1972.

[4] A. W.-m. Lee and M. Karplus, "Structure-specific model of hemoglobin cooperativity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 80, no. December, pp. 7055–7059, 1983.

[5] W. a. Eaton, E. R. Henry, J. Hofrichter, and A. Mozzarelli, "Is cooperative oxygen binding by hemoglobin really understood?," *Nature Structural Biology*, vol. 17, no. 1-2, pp. 147–162, 2006.

[6] Q. Cui and M. Karplus, "Allostery and cooperativity revisited.," *Protein science : a publication of the Protein Society*, vol. 17, no. 8, pp. 1295–1307, 2008.

[7] J. Monod, J. Wyman, and J. P. Changeux, "On the Nature of Allosteric Transitions: a Plausible Model.," *Journal of molecular biology*, vol. 12, no. 1, pp. 88–118, 1965.

[8] A. D. Ferguson, C. a. Amezcua, N. M. Halabi, Y. Chelliah, M. K. Rosen, R. Ranganathan, and J. Deisenhofer, "Signal transduction pathway of TonB-dependent transporters.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 2, pp. 513–518, 2007.

[9] S. W. Lockless and R. Ranganathan, "Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families," *Science*, vol. 286, no. October, pp. 295–299, 1999.

[10] G. M. Süel, S. W. Lockless, M. a. Wall, and R. Ranganathan, "Evolutionarily conserved networks of residues mediate allosteric communication in proteins.," *Nature Structural Biology*, vol. 10, pp. 59–69, Jan. 2003.

[11] F. C. Peterson, R. R. Penkert, B. F. Volkman, and K. E. Prehoda, "Cdc42 regulates the Par-6 PDZ domain through an allosteric CRIB-PDZ transition," *Molecular Cell*, vol. 13, no. 5, pp. 665–676, 2004.

[12] R. N. McLaughlin, F. J. Poelwijk, A. Raman, W. S. Gosal, and R. Ranganathan, "The spatial architecture of protein function and adaptation," *Nature*, 2012.

[13] R. K. Jain 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*, vol. 101, no. 1, pp. 111–116, 2004.

[14] R. M. Stroud and E. B. Fauman, "Significance of structural changes in proteins: expected errors in refined protein structures.," *Protein science : a publication of the Protein Society*, vol. 4, no. 11, pp. 2392–2404, 1995.

[15] R. Tonikian, Y. Zhang, S. L. Sazinsky, B. Currell, J. H. Yeh, B. Reva, H. a. Held, B. a. Appleton, M. Evangelista, Y. Wu, X. Xin, A. C. Chan, S. Seshagiri, L. a. Lasky, C. Sander, C. Boone, G. D. Bader, and S. S. Sidhu, "A specificity map for the PDZ domain family," *PLoS Biology*, vol. 6, no. 9, pp. 2043–2059, 2008.

[16] J. S. Fraser, H. van den Bedem, a. J. Samelson, P. T. Lang, J. M. Holton, N. Echols, and T. Alber, "Accessing protein conformational ensembles using room-temperature X-ray crystallography," *Proceedings of the National Academy of Sciences*, vol. 108, no. 39, pp. 16247–16252, 2011.

[17] J. S. Fraser, M. W. Clarkson, S. C. Degnan, R. Erion, D. Kern, and T. Alber, "Hidden alternative structures of proline isomerase essential for catalysis.," *Nature*, vol. 462, no. 7273, pp. 669–673, 2009.

[18] B. Tom Burnley, P. V. Afonine, P. D. Adams, and P. Gros, "Modelling dynamics in protein crystal structures by ensemble refinement," *eLife*, vol. 2012, no. 1, pp. 1–29, 2012.

[19] R. Selvaratnam, S. Chowdhury, B. VanSchouwen, and G. Melacini, "Mapping allostery through the covariance analysis of NMR chemical shifts.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 15, pp. 6133–6138, 2011.

Chapter 4

Conclusion and Future Directions

4.1 Conclusion and Future Directions

4.2 The Protein Sector: A Relevant Reduction?

In the broadest sense, this thesis in combination with previous work in the lab has addressed the sufficiency of the sector description with respect to understanding how proteins work. Specifically, this thesis has focused on understanding the origins of two basic characteristics of proteins–their evolvability and allosteric properties–with the goal of testing whether the protein sector is a good representation of these two characteristics.

4.2.1 The Structural Properties of Adaptation in Proteins

Proteins have the ability to adapt in just a few sequence mutations to new pressures yet how they are built to satisfy this characteristic has remained unclear. To more fully understand how the design of proteins encodes for this characteristic, we examined a two-step mutational path in PSD95^{pdz3} to new function. We found this path to have two routes to new function as judged by a comprehensive functional study of each protein along the adaptive path: 1) through a general, conditionally neutral intermediate (G330T) and 2) through a direct functionally switching mutation (H372A). Structurally, the G330 position was indirectly connected to the binding site through H372 which directly contacted the peptide of the protein. Thus, this path gave us a simplified yet representative anecdote in which to study the adaptive process.

We found that mutations such as G330T, conditionally neutral mutations, are useful over a range of population dynamics parameters (mutation rate and the rate of environmental fluctuation). Mutations such as H372A can only facilitate adaptation in a special regime of parameters known as the stochastic regime in which the generation of mutations cannot keep up with the timescale of environmental fluctuation. Fundamentally this result is due ot the fact that the only way H372A can be as helpful as G330T is if two events, protein sequence variation *and* environmental fluctuation, occur simultaneously. In essence this is why mutations such as G330T are so useful; conditionally neutral variaition is able to function in the current environment while, at the same time, aiding adaptation in the event of an altered environment.

To address the adaptive potential of the protein in a global sense, given that mutations such

as G330T were important for adaptation we elucidated all possible single mutations in the protein that were conditionally neutral (adaptive towards target function while retaining native function). We found that all direct functional switching mutations (adaptive towards target function but abrogate native function) are within sector positions directly in contact with the binding site (residues 327 and 372 in the protein structure). In contrast, most neutral adaptive mutations are also within the protein sector but located several contact shells away from the adaptive challenge (residues 364 and 372 for example).

To understand how such distant mutations can work, we solved crystal structures of wild-type and G330T bound to both native and target peptide. The structures revealed that distant mutations couple to the binding site to promote adaptation. Specifically, the G330T mutation creates plasticity at the binding pocket, allowing the H372 residue to accommodate both classes of peptides equally well. The nature of the structural result demonstrates the importance of coupling in promoting adaptation, consistent with our finding that a majority of neutral adaptive mutations occur within the protein sector.

Following these results, we proposed an approximate structural decomposition constructed upon the well-known and seemingly paradoxical relationship between robustness and evolvability. Non-sector mutants are unconditionally neutral, conferring the property of robustness without contributing to the adaptive capacity of proteins. In contrast, sector positions residing away from the adaptive challenge are largely conditionally neutral and therefore evolutionarily advantageous–robust to native function but permitting the acquisition of new function.

4.2.2 Revealing Paths of Allostery in Proteins

Seeing allostery in proteins has been a central problem of structural biology since its inception. We tackled this problem in PSD95^{pdz3} where we knew from the structural studies above that the G330T mutation caused local remodeling of an area termed the β_2-β_3 loop and induced a structural shift at a distant, conserved helix known as the α_1 helix. Because both the β_2-β_3 loop and α_1 helix reside within the protein sector of the PDZ family, we hypothesized that allosteric transmission between these sites was occuring through coupled physical interactions in the protein. Thus, the approach we took was to identify structurally coupled residues in PSD95^{pdz3}.

To measure structural coupling, we chose to solve the structural cycle comprised of two mutations, G330T and $T_{-2}F$. The T_{-2} position of the ligand was shown previously to coevolve with the protein sector and spatially lies between the β_2-β_3 loop and α_1 helix, thus making it an appropriate choice of residue perturbation. The structural coupling spectrum reveals three areas in the protein that exhibit non-additive structural shifts: the β_2-β_3 loop, the carboxylate binding loop (CBL), and the α_1 helix. Within the structural cycle, we were able to explicitly reveal how the interaction between the G330T and $T_{-2}F$ mutations propagate to distant areas of the protein such as the carboxylate binding loop through a chain of amino acid interactions. We found that the pattern of structural coupling revealed a path of connected amino acids propagating through the core of the protein sector spanning the β_2-β_3 loop, through the binding site of the protein, and extending to the α_1 helix. The non-sector positions demonstrating structural non-additivity were non-randomly distributed in the protein structure, forming shells of residues around the protein sector, particularly near the α_1 helix and the carboxylate binding loop.

To verify whether this path was truly allosteric by nature, we solved the same cycle of structures as defined by the G330T and $T_{-2}F$ mutations, but in the background of a His372 to Ala mutation, a residue directly between the G330 and T_{-2} positions. This mutation quenched structural coupling causing an abrogation of the long-range effect of G330T. We examined how this mutation breaks allostery by measuring the pattern of energetic coupling to the $T_{-2}F$ mutation in the H372A $PSD95^{pdz3}$ protein. Biochemical analysis using a bacterial-2-hybrid approach shows that H372A effectively 'cuts' the allostery from the β_2-β_3 loop. Positions G330 and G329, coupled to the $T_{-2}F$ mutation in the wild-type protein, are no longer coupled to the $T_{-2}F$ mutation in the background of H372A. The α_1 helix however retains the ability to couple to the $T_{-2}F$ mutation. These results suggested that we were able to find the path of allosteric communication from the β_2-β_3 loop to the α_1 helix using the approach of probing structural coupling within the protein sector. These results motivate further experiments to test a) the generality of this result in other proteins and b) the material decomposition of the protein as a complex material.

4.3 Other Models of Protein Design

The two characteristics of proteins explicated in this thesis are simply two of the may qualities that define the essence of proteins. Thus the more fundamental matter into which this thesis fits is the goal of a relevant reduction of the protein as a complex biological system. Is the protein sector a good model to describe proteins?

Studies in the Ranganathan Lab have been able to attribute many definite protein characteristics ot the protein sector such as fold, the ability to execute a primary task, and now, the ability to rapidly adapt to new environments and propagate allosteric signal across the protein structure.[1–3] Given these qualities, it indeed appears that a statistical decomposition of the protein via the evolutionary record of the protein family is a good approach to identify the parts of relevance within an individual protein. In the most definitive demonstration of this concept, Russ *et al.*, Socolich *et al.* and Gosal and Ranganathan were able to design synthetic proteins annealed upon the information in the SCA matrix. What characteristic then, if any, is *not* able to be described by the approach of SCA?

While the protein sector encodes information about biochemical function, adaptation, and in a broader sense, the constraints placed on sequence variation as a result of the evolutionary pressures faced by the protein, the sector does not obviously reveal the pattern of contacts in the protein structure. Weigt *et al.* have used an alternate decomposition of evolutionary signal to reveal the contact graph of the protein known as Direct Coupling Analysis (DCA).[4] The premise of this approach is to use correlations in the multiple sequence alignment to deduce an interaction matrix using statistical physical models. Specifically, the interaction matrix is a construct that provides information about which amino acids interact with each other. Weigt and coworkers have found that decomposition of the interaction matrix as derived from the pure, unweighted correlation matrix, identifies the contact graph of proteins to a large degree. Here therefore is an example of an alternate design principle, the use of pure correlation is from a large multiple sequence alignment, revealing a different characteristic of proteins. Additionally, David Baker's laboratory (University of Washington) offers another alternate approach founded on the premise of local electrostatic field optimization through ab initio calculations to design proteins.[5,6] This method

in principle spans an alternate conceptual space, explicitly ignoring evolutionary information but rather operating on fundamental physical principles. Indeed Rosetta has been able to design active enzymes through emphasis on active site topology and physical chemistry. Interestingly however, these proteins are far from natural enzymes with respect to catalytic activity, often requiring the use of directed evolution to achieve high performance function.

Thus, an outstanding question exists with respect to the information gleaned by DCA, Rosetta, and SCA. Do DCA designed sequences perform like natural sequences? To be more concrete, the contact graph as calculated from DCA is a property of the ensemble (similar to the sector in SCA). However, single sequences designed from the SCA approach have demonstrated to be natural-like, folding into well-packed three-dimensional structures and functioning in a similar fashion to wild-type protein. If instead of SCA, DCA is used as the target criteria for design, do individual sequences fold, function, and adapt readily to new function? Equivalently, do Rosetta designed proteins do the same? To date, single sequence design using methods other than SCA have not been implemented. It will be imperative in the future to understand the difference between these approaches and specifically their ability to potentially distinguish different protein characteristics.

4.4 The Energetic Architecture of Proteins as a Function of their Environment

Identifying where in the sequence essential protein characteristics such as function, adaptive capacity, and allostery reside, and perhaps more fundamentally, understanding how a material can be built with these characteristics, has escaped most who have longed for a deeper understanding of the design of natural proteins. Why is this? Current physical theory contains little information with regards to the functional importance of interactions between parts. Indeed, man-made systems, while sometimes non-linear in their behavior or response, are generally engineered in a non-complex fashion, where the term ‘complex’ suggests the non-additive contribution of parts within the system to the whole.

With respect to the functional importance of complexity, engineered systems are made for the here and now. That is, they are designed to function well in a ‘niche’. Our data on the adaptive

capacity of proteins however, suggests that while biological materials function in a particular environment, their design of sparse and distributed coupling enables them to rapidly adapt given an environmental fluctuation. For instance, in the case of $PSD95^{pdz3}$, many conditionally neutral variants exist that could potentiate adaptation from the CRIPT to the $T_{-2}F$ environment. Our structural studies suggest the biophysical underpinnings of rapid adaptation can be traced to the characteristic of spatially heterogeneous, distributed coupling in the protein. Of course, the design of natural protein as we currently see it is a result of the cumulative fitness pressure faced by the protein through evolution; the sector architecture is a result of many generations of selection imposed upon the protein. Thus, a natural next step is to ask whether the protein sector is indeed fundamental. Revisiting the stochastic simulation presented in Chapter 1, we are able to clearly demonstrate the utility of distributed coupling given environmental fluctuations. But, what would happen if the protein evolved under a totally constant environment? In the world of the simulation, what if $PSD95^{pdz3}$ only faced a CRIPT environment for all of its existence? The protein, in such a scenario, may have little incentive to evolve a complex network of coupling used to withstand fluctuations in selection pressures and instead, optimally tune binding solely for CRIPT peptide to maximize fitness. Perhaps it is in fact the constraint of survival in fluctuating environments that drives the emergence of complex interactions in systems. It will therefore be important in the future to address the role of the environment, in particular the statistical history of environmental variation, in shaping the energetic complexity of proteins.

References

[1] W. P. Russ, D. M. Lowery, P. Mishra, M. B. Yaffe, and R. Ranganathan, "Natural-like function in artificial WW domains.," *Nature*, vol. 437, pp. 579–83, Sept. 2005.

[2] M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan, "Evolutionary information for specifying a protein fold.," *Nature*, vol. 437, no. 7058, pp. 512–518, 2005.

[3] W. S. Gosal and R. Ranganathan, "Form follows function : Parsing the evolutionary rules for protein folding and function," *In Prep.*

[4] F. Morcos, a. Pagnani, B. Lunt, a. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, "PNAS Plus: Direct-coupling analysis of residue coevolution captures native contacts across many protein families," *Proceedings of the National Academy of Sciences*, vol. 108, no. 49, pp. E1293–E1301, 2011.

[5] D. Röthlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. a. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, and D. Baker, "Kemp elimination catalysts by computational enzyme design.," *Nature*, vol. 453, no. 7192, pp. 190–195, 2008.

[6] L. Jiang, E. a. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas, D. Hilvert, K. N. Houk, B. L. Stoddard, and D. Baker, "De novo computational design of retro-aldol enzymes.," *Science (New York, N.Y.)*, vol. 319, no. 5868, pp. 1387–1391, 2008.

Chapter 5

Methods

5.1 Methods

5.2 Methods for Chapter 1

5.2.1 Sector Identification

SCA and sector identification were performed using SCA v5.0 (Matlab platform). The software and script for performing the calculations are available from the laboratory website (`http://systems.swmed.edu/rr_lab`). The sector positions identified in the PDZ family were: 322, 323, 325, 327, 329, 330, 336, 345, 347, 350, 351, 352, 353, 359, 362, 364, 372, 375, 376, 379, 386, 388.

5.2.2 Fluorescence Polarization Affinity Measurements

$PSD95^{pdz3}$ mutants were expressed in BL21(DE3) textitE.coil cells as glutathione S-transferase (GST) fusions and purified with affinity chromatography and cleavage of the GST tag with thrombin. TMR (tetramethylrhodamine) labeled peptide was synthesized by the UTSW core facility. Binding affinities were measured by tracking fluorescence polarization measurements over a range of protein concentrations on a Victor plate reader.

5.2.3 Peptide Library Construction

A comprehensive peptide library was generated using oligonucleotide-directed mutagenesis of the C-terminus of the peptide. To construct a library of four fully randomized consecutive C-terminal amino acids two oligonucleotides were synthesized (IDT), each containing a BsaI restriction site designed such that restriction creates complementary overhangs of the 5' and 3' ends, and one containing four consecutive 'NNS' codons in which N and S are equimolar mixtures of A,C,T,G and C,G respectively encoding all possible amino acids at each C-terminal position. A single round of PCR was conducted, amplifying the entire plasmid carrying the ligand sequence and subsequently restricted with BsaI. Importantly, the template used in this PCR reaction encoded the negative control ligand sequence (TKNYKGGG) so as to minimize artificial positive enrichment signal post sequencing. A unimolecular ligation (1mL) was incubated overnight at 16 degrees and zymo purified into 10 μL. 10 individual transformations into MaxDH10B were grown overnight after

recovery and prepped so as to minimize any possible bottlenecking effect of the peptide population. Transformation of the prepped libraries yielded greater than 10^8 transformants.

5.2.4 Bacterial-2-Hybrid Assay with Antibiotic Resistance Readout

The bacterial-2-hybrid, previously developed by Rick McLaughlin,[1] was modified to adequately handle libraries of large complexity (such as the peptide library, in total a 160,000 unique peptides). With a fluorescence readout, as implemented in the previous version of the assay, viability of the cells decreased upon FACS sorting due to the time taken to sort an adequate number of cells. Coupling the PDZ-peptide interaction with transcription of an antibiotic resistance marker solved this problem for larger mutant populations. The bacterial-2-hybrid system is a three-plasmid assay consisting of 1) pZS22-PDZ3 (trimethoprim (trm) resistant) providing IPTG inducible expression of λ-cI bacteriophage protein fused to PSD95^{pdz3}, 2) pZA31-ligand (kanamycin (kan) resistant) providing doxycycline induced expression of α subunit of RNA polymerase fused to the nine-amino acid peptide binding partner of PSD95^{pdz3}, and 3) pzE1RM (ampicillin (amp) resistant) containing the target promoter driving the expression of chloramphenicol acetyl transferase (cat). For assaying the library of peptides against a PDZ domain, MC4100-Z1 *E.coli* cells containing the pZS and pZE plasmids were transformed with the peptide library and growin in LB media for two hours post transformation recovery with 20μg/mL trm, 50 μg/mL kan, 100 μg/mL amp to an O.D.$_{550}$ of 0.04. The culture was then induced using 50 ng/mL doxycyline plus antibiotics for 3 hours to an O.D.$_{550}$ of 0.1. 10mL of the induction culture was used as inoculum for the selection culture while the remainder of the induction culture was miniprepped, PCR amplified, and sequenced as the input population (see below for sequencing details). Selection was carried out using 150 μg/mL chloramphenicol for 6 hours. During selection, care was taken to make sure the O.D.$_{550}$ of the culture never exceeded 0.1. The culture was then washed thoroughly using LB media lacking chloramphnicol and grown overnight at 37 degrees C. The culture was miniprepped and amplified by PCR to prepare as the selected population for deep sequencing. Deep sequencing of the input and selected populations of the pZA31 ligand sequence were performed using a single-end Hi-Seq 2500 sequencing run (University of Texas Southwestern genomics core) and analyzed using self-coded software shell scripts. Custom oligonucleotides were designed to prepare miniprepped

populations for deep sequencing using two rounds of PCR (round 1 primers were designed to anneal specifically to the pZA plasmid, round 2 primers are common to all sequencing runs, containing the necessary indices for a high-throughput sequencing run and were designed and ordered by Bill Russ). Sequencing data is available electronically (excel format).

5.2.5 Stochastic Simulation

The population dynamics simulation followed the flux of the total population numbering 1000 members through a fluctuating environment (either CRIPT or T$_{-2}$F) over 10000 generations. Conditions were intialized with 1000 members of wild-type PSD95^{pdz3}. The probability of a given allele occurring in generation i is defined to be dependent on the relative flux of the allele. For example to calculate the probability of wild-type occuring in generation i

$$\Phi_{positive,i,WT} = (N_{i-1})_{WT} f_{WT} + (N_{i-1})_{G330T} k_{mut} + (N_{i-1})_{H372A} k_{mut} + (N_{i-1})_{H372A/G330T} k_{mut}^2 \quad (5.1)$$

$$\Phi_{negative,i.WT} = 2(N_{i-1})_{WT} k_{mut} + (N_{i-1})_{WT} k_{mut}^2 \quad (5.2)$$

$$P_{i,WT} = \frac{\Phi_{positive,i,WT} - \Phi_{negative,i,WT}}{\Phi_{net,i,WT} + \Phi_{net,i,G330T} + \Phi_{net,i,H372A} + \Phi_{net,i,(H372A,G330T)}} \quad (5.3)$$

where

$$\Phi_{net,i,WT} = \Phi_{positive,i,WT} - \Phi_{negative,i,WT} \quad (5.4)$$

Similarly, probabilities can be calculated at each generation for each allele and the resultant population is drawn from a multinomial distribution defined by the calculation probabilities. The fitness of allele in a a particular environment is defined by a single site binding isotherm and thereby is determined by the affinity of protein for ligand.

$$f_i^x = \frac{[L]}{[L] + K_d^x} \quad (5.5)$$

where i is the identity of the allele and x is the environment in which the affinity is measured. *[L]* was defined to be 10μM to ensure a reasonable dynamic range of fitness values based on the measured affinity values ranging from 0.8 (wild-type CRIPT complex) to 36 μM (wild-type T$_{-2}$F

complex). Population dynamics statistics for each pair of mutation and environmental switch rates were averaged over 1000 trajectories to generate probabilities of conversion to double mutant and relative adaptive utility of G330T and H372A. The simulation was written and conducted using MATLAB with simulation jobs submitted to the nucleus cluster at UTSW. The MATLAB code is shown below

```
%Need to load fitness values into Matlab. In our case, these values are determined by the protein peptide
    interaction.
%The two fitness `landscapes' for CRIPT and T(-2)F are variabes `f_val_CRIPT' and `f_val_T7F'
%For a single trajectory with mutation rate of kmut and
%environmental switch rate of a switch every N_iterations with a population size of pop_size, run evo_sim.m
function [pop_freq,x] = evo_sim(pop_size,kmut,switch_freq,N_iterations,f_val_CRIPT,f_val_T7F);
env = 0;
pop_freq = zeros(4,N_iterations);
probs = zeros(N_iterations,4);
pop_freq(1,1) = pop_size;
pop_freq(2:4,1) = 0;
for i=2:N_iterations;
    x(i) = env;
    probs(i,:) = gen_prob(pop_freq(:,(i-1)),kmut,env,f_val_CRIPT,f_val_T7F);


    pop_freq(:,i) = (mnrnd(pop_size,probs(i,:),1))';

    %disp(i);
    %disp(switch_freq);
    %disp(mod(i,switch_freq));
    if mod(i,switch_freq) == 0;
        if mod(i/switch_freq,2) == 0; %even
            env = 0; %CRIPT
        elseif mod(i/switch_freq,2) == 1; %odd
            env = 1; %T7F
        end


    end
end

pop_freq = single(pop_freq);
x = single(x);

%Below are commented lines for plotting
% figure;
% subplot(2,2,1); plot(pop_freq(1,:),'k','LineWidth',1.5); title('WT','FontWeight','bold');
% set(gca,'Xtick',0:N_iterations/4:N_iterations,'Ytick',0:pop_size/4:pop_size,...
%'FontSize',14,'FontWeight','bold','FontName','Arial','LineWidth',2,'TickLength',[0.025,0.025]);
% ylim([-0.01*pop_size 0.1*pop_size+pop_size]); box off
%
% subplot(2,2,2); plot(pop_freq(2,:),'g','LineWidth',1.5); title('G330T');
% set(gca,'Xtick',0:N_iterations/4:N_iterations,'Ytick',0:pop_size/4:pop_size,'FontSize',14,...
%'FontWeight','bold','FontName','Arial','LineWidth',2,'TickLength',[0.025,0.025]);
% ylim([-0.01*pop_size 0.1*pop_size+pop_size]); box off
%
```

```
47 % subplot(2,2,3); plot(pop_freq(3,:),'r','LineWidth',1.5); title('H372A');
48 % set(gca,'Xtick',0:N_iterations/4:N_iterations,'Ytick',0:pop_size/4:pop_size,'FontSize',14,...
49 %'FontWeight','bold','FontName','Arial','LineWidth',2,'TickLength',[0.025,0.025]);
50 % ylim([-0.01*pop_size 0.1*pop_size+pop_size]); box off
51 %
52 % subplot(2,2,4); plot(pop_freq(4,:),'b','LineWidth',1.5); title('H372A,G330T');
53 % set(gca,'Xtick',0:N_iterations/4:N_iterations,'Ytick',0:pop_size/4:pop_size,'FontSize',14,...
54 %'FontWeight','bold','FontName','Arial','LineWidth',2,'TickLength',[0.025,0.025]);
55 % ylim([-0.01*pop_size 0.1*pop_size+pop_size]); box off
56 %
57 % figure;
58 % plot(pop_freq(1,:),'k','LineWidth',1.5); hold on; plot(pop_freq(2,:),'g','LineWidth',1.5); hold on; ...
59 %plot(pop_freq(3,:),'r','LineWidth',1.5), hold on; plot(pop_freq(4,:),'b','LineWidth',1.5);
60 % ylim([-0.01*pop_size 0.1*pop_size+pop_size]);
61 % xlabel('Iteration','FontSize',16,'FontWeight','bold');
62 % ylabel('Population Size','FontSize',16,'FontWeight','bold');
63 % title('Mut: 0.1, No Switch', 'FontSize',20);
64 % set(gca,'Xtick',0:N_iterations/4:N_iterations,'Ytick',0:pop_size/4:pop_size,'FontSize',14,...
65 %'FontWeight','bold','FontName','Arial','LineWidth',2,'TickLength',[0.025,0.025]);
66 % box off;
67
68
69 %The above code should give a 5x10000 matrix for the population
70 %dynamics as defined by the mutation rate and environmental switch rate.
71
72 %To reduce an entire trajectory to single numbers indicating
73 %the probability of conversion from wild-type to double mutant
74 % and the adaptive utility of either G330T or H372A
75 %during the environmental switch, run gen_Dbmut_pop.m
76
77 function [kmut,switch_freq,Dbmut_avg_frac,G330T_avg_frac,H372A_avg_frac] = gen_Dbmut_pop(pop_size,kmut,
       switch_freq,N_iterations,f_val_CRIPT,f_val_T7F);
78
79 [pop_freq,env_vector] = evo_sim(pop_size,kmut,switch_freq,N_iterations,f_val_CRIPT,f_val_T7F);
80
81
82
83 %Given a trajectory, do all WT molecules go to double mutant?
84 %The only types of trajectories I currently want to analyze are ones in
85 %which the wt-->dbmut transition is total.
86
87
88 if all(env_vector == env_vector(1)) == 1;
89     error('no switch you dumbass');
90 else
91     counter = 0;
92     %within the switched environment, do we see a pop_size of dbmut?
93     for i=2:length(env_vector);
94
95         if i < length(env_vector);
96             if env_vector(i) - env_vector(i-1) == 1; %environment has switched
97                 counter = counter + 1;
98
99                 if i+(switch_freq-1) <= length(env_vector);
100                     Dbmut_pop(counter,:) = pop_freq(4,i:i+(switch_freq-1)); %selects dbmut population in t7f
                            environment before switch back to CRIPT
101                     Dbmut_pop_frac(counter) = max(Dbmut_pop(counter,:))/(pop_size);
102                     G330T_pop(counter,:) = pop_freq(2,i:i+(switch_freq-1));
103                     H372A_pop(counter,:) = pop_freq(3,i:i+(switch_freq-1));
```

```
104                     G330T_pop_frac(counter) = max(G330T_pop(counter,:))/(pop_size);
105                     H372A_pop_frac(counter) = max(H372A_pop(counter,:))/(pop_size);
106
107                 end
108
109
110
111             end
112         end
113     end
114 end
115
116 clear pop_freq;
117 clear env_vector;
118
119 Dbmut_avg_frac = mean(Dbmut_pop_frac);
120 G330T_avg_frac = mean(G330T_pop_frac);
121 H372A_avg_frac = mean(H372A_pop_frac);
122
123 %To make a grid of many mutation rates and switch rates,
124 % run gen_param_grid.m
125
126
127 function [Dbmut_avg_frac,G330T_avg_frac,H372A_avg_frac] = gen_param_grid(pop_size,N_iterations,kmut_vector,
        env_switch_vector,f_val_CRIPT,f_val_T7F);
128
129 %kmut_vector is a vector of mutation rates
130 %env_switch_vector is a vector of environmental switch frequencies
131 %Output is a length(kmut_vector) x length(env_switch_vector) grid
132
133 counter_forloop = 0;
134 for j=1:length(env_switch_vector);
135     parfor i=1:length(kmut_vector);
136         [Dbmut_avg_frac(i,j),G330T_avg_frac(i,j),H372A_avg_frac(i,j)] = gen_Dbmut_pop(1000,kmut_vector(i),
              env_switch_vector(j),N_iterations,f_val_CRIPT,f_val_T7F);
137     end
138
139     counter_forloop = counter_forloop+1;
140     disp(counter_forloop);
141 end
142
143 Dbmut_avg_frac = single(Dbmut_avg_frac);
144 G330T_avg_frac = single(G330T_avg_frac);
145 H372A_avg_frac = single(H372A_avg_frac);
146
147
148 %To make N_iterations_ensemble of such grids and average over them to
149 %get good statistics, run gen_ensemble.m
150
151 function [Dbmut_avg_frac_total,G330T_avg_frac_total,H372A_avg_frac_total] = gen_ensemble(
        N_iterations_ensemble,pop_size,N_iterations,kmut_vector,env_switch_vector,f_val_CRIPT,f_val_T7F);
152
153 %Does gen_param_grid many times then takes the mean of the grid values for
154 %a given mutation rate and environmental switch rate
155
156 parfor i=1:N_iterations_ensemble;
157     [Dbmut_avg_frac(:,:,i),G330T_avg_frac(:,:,i),H372A_avg_frac(:,:,i)] = gen_param_grid(pop_size,
          N_iterations,kmut_vector,env_switch_vector,f_val_CRIPT,f_val_T7F);
158 end
```

```
159
160  Dbmut_avg_frac_total = mean(Dbmut_avg_frac,3);
161  G330T_avg_frac_total = mean(G330T_avg_frac,3);
162  H372A_avg_frac_total = mean(H372A_avg_frac,3);
163  Dbmut_avg_frac_total = single(Dbmut_avg_frac_total);
164  G330T_avg_frac_total = single(G330T_avg_frac_total);
165  H372A_avg_frac_total = single(H372A_avg_frac_total);
166
167
168 %At the end of running all m files, should have a grid of
169 % values indicating the adaptive utility of G330T vs H372A
170 % and the predilection for adapting to the double mutant given a set
171 % of mutation rates and environmental switch rates
```

5.2.6 Crystallography of PSD95^{pdz3}

For structural studies, PSD95^{pdz3} wild-type and variants were expressed in BL21(DE3) *E.coli* cells as glutathione S-transferase (GST) fusions and purified as described previously. The purified proteins were concentrated to 40-50μg/mL and either flash froen or used immediately for crystal trials. Concentrations of protein used for crystal trials ranged from 7 to 11 mg/ml with most crystals exhibiting optimal diffraction at 9 mg/ml. CRIPT and T$_{-2}$F peptides used for crystallography were synthesized at the UT Southwestern peptide core facility to a purity of greater than 95% as determined by HPLC. For liganded complexes, a 2:1 molar excess of ligand was used to ensure binding of ligand to PDZ variant. Crystal conditions included variations around 1.0 M NaCitrate and pH 7.0 using the hanging drop method of 1.5μL protein/ligand solution and 1.5 μL crystal condition incubated at 16 degrees C. Crystals belonged to space group P4$_1$32. Crystals were cryoprotected by serial soaking in mother liquor plus increasing glycerol concetrations (5%, 10%, and 15%) and flash frozen in liquid nitrogen. The table of crystal conditions for each structure is shown in Figure 5.1.

5.2.7 Model Refinement

X-ray data were indexed, integrated, and scaled with HKL2000.[2] The structures of all wild-type and mutant PDZ3 domains were solved using phenix.automr with apo PDZ3 as a search model (PDB accession code 1BFE).[3] Models with riding hydrogen atoms were generated using phenix.autobuild,[4] subjected to an initial round of rigid body refinement and Cartesian simulated annealing, and further built and refined over a number of additional cycles, either manually in

Crystal	[NaCitrate] (M)	pH	[Protein]
WT(apo)	1.0	6.9	9
WT(cript)	1.0	7.0	9
WT($t_{-2}f$)	1.125	7.1	9
G330T(apo)	1.05	7.0	7
G330T(cript)	1.2	7.0	8
G330T($t_{-2}f$)	1.2	6.8	9
H372A(apo)	0.95	7.0	9
H372A(cript)	1.05	7.0	7
H372A($t_{-2}f$)	1.05	7.0	7
Dbmut(apo)	1.0	7.0	13
Dbmut(cript)	1.25	7.0	9
Dbmut($t_{-2}f$)	1.2	6.75	7

Figure 5.1: All proteins in ranges of NaCitrate and pH with the same space group and until cell dimensions within 5% of each other. All protein-peptide complexes required seeding with wild-type except for the wild-type apo and CRIPT crystal. Wild-type apo was used for seeding apo crystals, wild-type-CRIPT crystals were used for seeding liganded protein complexes. All crystals were grown at 16 degrees C.

Coot or automatically with phenix.refine and Buster.[5] Translation/libration/screw (TLS) groups were determined using the TLSMD web server and incorporated in the B-factor model towards the end of the refinement process. Structures were validated using MolProbity.[6]

5.3 Methods for Chapter 2

5.3.1 Measuring Structural Coupling

All models of structures were aligned using least squares minimization in Coot. The calculation used to measure structural shifts between two structures is shown below (MATLAB) written by Rohit Sharma.

```
function [str1,str2] = dr(str1,str2,alt)
% This function accepts two structures and finds the difference of the two:
%           str1 --> str2
% It returns the x-displacement (.dx), y-displacement (.dy), z-displacement (.dz),
% displacement (.dr), normalized displacement(.drnorm).  Using the
% direction of the vector (calculated in spherical coordinates) and the
% magnitude of the normalized displacement, the function also calculates
```

```
8  % the normalized x, y, and z vectors.
9
10 % If alt = 1, for atoms that have alternate conformations the difference is calculated
11 % using a weighted average of its position and propagated positional
12 % errors. If alt = 0 then the second conformation is not used.
13
14 % Example usage: [ww2,aw27] = dr(ww2,aw27,0);
15
16 % Input: two models in structure arrays.
17 % Output:  additional fields in each structure:
18 %       1) .dr      raw disp
19 %       2) .drn     normalized displacement
20 %       3) .dx      raw displacement in x
21 %       4) .dy      raw displacement in y
22 %       5) .dz      raw displacement in z
23 %       6) .dxn     normalized displacement in x
24 %       7) .dyn     normalized displacement in y
25 %       8) .dzn     normalized displacement in z
26 %       9) .drsets  list of sets used for calculation
27 %       10).drmsg   'alternate conformations used/not used'
28
29 counter=0;
30
31 [numchainA,blah] = size(find(strcmp(str1.chainid,'A')));        % number of atoms with chainid A
32 [numchainP,blah] = size(find(strcmp(str1.chainid,'B')));        % number of atoms with chainid B
33 total = numchainA + numchainP;                                  % number of atoms in model excluding waters
34
35 for n = 1:total                                                 % Go through all protein and peptide atoms in
       model 1
36
37                                                                % find indices and number of occurrences for
                                                                      label1 (res,#,atomid) and label2(#,
                                                                      atomid)
38      ind1_1 = find(strcmp(str1.label,str1.label(n)));          % indices of label1 (res,#,atomid) in str1
39      [num1_1,blah] = size(ind1_1);                             % number of occurrences of label1 in str1 (
            could be 1,2)
40      mc = strcmp(str1.atomid(n),'N')|strcmp(str1.atomid(n),'CA')|strcmp(str1.atomid(n),'C')|strcmp(
            str1.atomid(n),'O');  % is it mainchain atom?
41
42      ind1_2 = find(strcmp(str2.label,str1.label(n)));          % indices of label1 in str2
43      [num1_2,blah] = size(ind1_2);                             % number of occurrences of label1 in str2 (
            could be 0,1,2)
44      ind2_2 = find(strcmp(str2.label2,str1.label2(n)));        % indices of label2 in str2  (label2 only has
             #,atomid - can be used to check for mutation)
45      [num2_2,blah] = size(ind2_2);                             % number of occurrences of label2 (#,atomid)
            in str2 (could be 0,1,2)
46
47      if (alt==0) & strcmp(str1.ac(n),'B')
48          occupiedin1 = 0;
49      else
50          occupiedin1 = str1.occ(n);
51      end
52
53      occupiedin2=0;
54      switch num1_2                                             % is the atom found in structure 2?
55          case 0
56              if (num2_2==0)                                    % if label1 is not found and label2 is not
                    found then it must not be modeled in str2
57                  occupiedin2 = 0;
```

```
58              elseif ((num2_2==1)&mc)                                 % if label1 is not found but label2 is found
                    it must be a mutated position (only residue names don't match).
59                  occupiedin2 = str2.occ(ind2_2(1));                  % Only count as occupied if mainchain atom;
                        let occupiedin2 be occupancy of that mc atom (should be 1).
60              end
61          case 1
62              occupiedin2 = str2.occ(ind1_2(1));                      % if label1 is found once then let
                    occupiedin2 be whatever the occupancy of that atom.
63          case 2
64              occupiedin2 = 1;                                        % if label2 is found twice then it must be
                    modeled as alternate confs in str2.
65      end
66
67      occupiedinboth = occupiedin1 & occupiedin2;                     % is it occupied in both structures?
68
69      if occupiedinboth                                               % if the atom is in both structures...
70
71                                                                      % str1 atom will be assigned coordinates,pe.
72          if ((num1_1==2) & alt)                                      % If atom is found twice and alt conf = 'on'
                = 1, then weight.
73              x1 = str1.x(ind1_1(1))*str1.occ(ind1_1(1)) +  str1.x(ind1_1(2))*str1.occ(ind1_1(2));
74
75              disp(ind1_1(1));
76              disp(ind1_1(2));
77
78              disp(str1.x(ind1_1(1)));
79              disp(str1.x(ind1_1(2)));
80              disp(str1.occ(ind1_1(1)));
81              disp(str1.occ(ind1_1(2)));
82              disp(str1.x(ind1_1(1))*str1.occ(ind1_1(1)));
83              disp(str1.x(ind1_1(2))*str1.occ(ind1_1(2)));
84              disp(x1);
85              counter=counter+1;
86              disp(counter);
87
88              y1 = str1.y(ind1_1(1))*str1.occ(ind1_1(1)) +  str1.y(ind1_1(2))*str1.occ(ind1_1(2));
89              z1 = str1.z(ind1_1(1))*str1.occ(ind1_1(1)) +  str1.z(ind1_1(2))*str1.occ(ind1_1(2));
90              pe1 = ((str1.occ(ind1_1(1))*str1.poserr(ind1_1(1)))^2 + (str1.occ(ind1_1(2))*str1.poserr(ind1_1
                    (2)))^2)^0.5;
91
92              disp(y1);
93              disp(z1);
94
95          else                                                        % Otherwise, x1,y1,z1,pe1 are simply its
                corresponding values.
96              x1 = str1.x(ind1_1(1));
97              y1 = str1.y(ind1_1(1));
98              z1 = str1.z(ind1_1(1));
99              pe1 = str1.poserr(ind1_1(1));
100         end
101                                                                     % str2 atom will be assigned coordinates,pe,
                                                                            and index.
102         if ((num1_2(1)==2) & alt)                                   % If the atom is found twice and alt confs
                are 'on' then take weighted ave
103             x2 = str2.x(ind1_2(1))*str2.occ(ind1_2(1)) +  str2.x(ind1_2(2))*str2.occ(ind1_2(2));
104             disp(x2);
105             y2 = str2.y(ind1_2(1))*str2.occ(ind1_2(1)) +  str2.y(ind1_2(2))*str2.occ(ind1_2(2));
106             z2 = str2.z(ind1_2(1))*str2.occ(ind1_2(1)) +  str2.z(ind1_2(2))*str2.occ(ind1_2(2));
```

```
107             pe2 = ((str2.occ(ind1_2(1))*str2.poserr(ind1_2(1)))^2 + (str2.occ(ind1_2(2))*str2.poserr(ind1_2
                    (2)))^2)^0.5;
108             i2 = ind1_2(1);
109             disp(y2);
110             disp(z2);
111         else                                                % Otherwise, coords and error are
                corresponding values.
112             x2 = str2.x(ind2_2(1));
113             y2 = str2.y(ind2_2(1));
114             z2 = str2.z(ind2_2(1));
115             pe2 = str2.poserr(ind2_2(1));
116             i2 = ind2_2(1);
117         end
118                                                             % Calculate displacements
119         str1.dx(n) = x2 - x1;
120         str1.dy(n) = y2 - y1;
121         str1.dz(n) = z2 - z1;
122         str1.dr(n) = (str1.dx(n)^2 + str1.dy(n)^2 + str1.dz(n)^2)^0.5;
123
124         str1.drn(n) = str1.dr(n)/(pe1^2 + pe2^2)^0.5;
125         [theta, phi, r] = cart2sph (str1.dx(n),str1.dy(n),str1.dz(n));
126         [str1.dxn(n), str1.dyn(n), str1.dzn(n)] = sph2cart (theta, phi, str1.drn(n));
127
128         str2.dx(i2) = str1.dx(n); str2.dy(i2) = str1.dy(n); str2.dz(i2) = str1.dz(n); str2.dr(i2) = str1.dr(
                n);
129         str2.dxn(i2) = str1.dxn(n); str2.dyn(i2) = str1.dyn(n); str2.dzn(i2) = str1.dzn(n); str2.drn(i2) =
                str1.drn(n);
130
131     else                                                    % if atom is not found in str1 and str2 in
            above conditions then set values to 0.
132         str1.dx(n) = 0; str1.dy(n) = 0; str1.dz(n) = 0; str1.dr(n) = 0; str1.dxn(n) = 0; str1.dyn(n) = 0;
                str1.dzn(n) = 0; str1.drn(n) = 0;
133     end
134 end
135
136 str1.dx = str1.dx';
137 str1.dy = str1.dy';
138 str1.dz = str1.dz';
139 str1.dr = str1.dr';
140 str1.dxn = str1.dxn';
141 str1.dyn = str1.dyn';
142 str1.dzn = str1.dzn';
143 str1.drn = str1.drn';
144
145 str2.dx = str2.dx';
146 str2.dy = str2.dy';
147 str2.dz = str2.dz';
148 str2.dr = str2.dr';
149 str2.dxn = str2.dxn';
150 str2.dyn = str2.dyn';
151 str2.dzn = str2.dzn';
152 str2.drn = str2.drn';
153
154 str1.dr_sets = [str1.set str2.set];
155 str2.dr_sets = [str1.set str2.set];
156
157 if alt == 1
158     str1.drmsg = 'alternate conformations used';
159     str2.drmsg = 'alternate conformations used';
```

```
160 else
161     str1.drmsg = 'alternate conformations not used';
162     str2.drmsg = 'alternate conformations not used';
163 end
```

The calculation used to calculate structural couplings ($\Delta\Delta$r values) is shown below (MATLAB)

```
1  function [str1, str2, str3, str4] = ddr(str1, str2, str3, str4, alt)
2  % ddr calculates the structural coupling vector for 4 structures in a
3  % structure cycle with format:
4  %       str1 -------> str2
5  %         |             |
6  %         |             |
7  %       str3 -------> str4
8  % If alternate conformations exist they will be treated as weighted
9  % averages if alt=1; if alt=0 then second conformation is not used.
10
11 % Input: 4 models in structure arrays; alternate conformation flag
12
13 % Output: 4 structures with additional fields for ddr and ddrnorm:
14 %   .ddr            raw structural coupling
15 %   .ddrn           normalized coupling
16 %   .ddx            x component of ddr
17 %   .ddy            y component of ddr
18 %   .ddz            z component of ddr
19 %   .ddx            x component of ddrn
20 %   .ddy            y component of ddrn
21 %   .ddz            z component of ddrn
22 %   .ddr_sets       list of sets used for calculations
23 %   .ddrmsg         'alternate conformations used/not used'
24
25 [numchainA,blah] = size(find(strcmp(str1.chainid,'A')));          % number of atoms with chainid A
26 [numchainP,blah] = size(find(strcmp(str1.chainid,'B')));          % number of atoms with chainid P
27 total = numchainA + numchainP;                                    % number of atoms in model excluding
       waters
28
29
30 for n = 1:total
31     ind1_1 = find(strcmp(str1.label(n),str1.label));             % where atom label1 (res, #, atomid)
           occurs in str1 (1,2)
32     [num1_1,blah] = size(ind1_1);                                % how many times label1 occurs in str1
           (1,2)
33     ind2_1 = find(strcmp(str1.label2(n),str1.label2));           % where atom label2 (#,atomid) occurs in
           str1 (1,2)
34     mc = strcmp(str1.atomid(n),'N')|strcmp(str1.atomid(n),'CA')|strcmp(str1.atomid(n),'C')|strcmp(str1.atomid
           (n),'O');  % is it mainchain atom?
35     if (alt==0)&(strcmp(str1.ac(n),'B'))
36         occupiedin1 = 0;
37     else
38         occupiedin1 = str1.occ(n);
39     end
40
41     ind1_2 = find(strcmp(str1.label(n),str2.label));             % where atom label1 occurs in str2
           (0,1,2)
42     [num1_2,blah] = size(ind1_2);                                % how many times label1 occurs in str2
```

```
43      ind2_2 = find(strcmp(str1.label2(n),str2.label2));              % where atom label2 occurs in str2
            (0,1,2) - should be superset of ind1_x
44      [num2_2,blah] = size(ind2_2);                                   % how many times label2 occurs in str2
45      occupiedin2=0;
46      switch num1_2                                                   % is the atom found in structure 2?
47          case 0
48              if (num2_2==0)                                          % if label1 is not found and label2 is
                    not found then it must not be modeled in str2
49                  occupiedin2 = 0;
50              elseif ((num2_2==1)&mc)                                 % if label1 is not found but label2 is
                    found it must be a mutated position (only residue names don't match).
51                  occupiedin2 = str2.occ(ind2_2(1));                  %     Only count as occupied if mainchain
                        atom; let occupiedin2 be occupancy of that mc atom (should be 1).
52              end
53          case 1
54              occupiedin2 = str2.occ(ind1_2(1));                      % if label1 is found once then let
                    occupiedin2 be whatever the occupancy of that atom.
55          case 2
56              occupiedin2 = 1;                                        % if label2 is found twice then it must
                    be modeled as alternate confs in str2.
57      end
58
59      ind1_3 = find(strcmp(str1.label(n),str3.label));                % where atom label1 occurs in str3
            (0,1,2)
60      [num1_3,blah] = size(ind1_3);                                   % how many times label1 occurs in str3
61      ind2_3 = find(strcmp(str1.label2(n),str3.label2));              % where atom label2 occurs in str3
            (0,1,2)
62      [num2_3,blah] = size(ind2_3);                                   % how many times labels occurs in str3
63      switch num1_3                                                   % is the atom found in structure 2?
64          case 0
65              if (num2_3==0)                                          % if label1 is not found and label2 is
                    not found then it must not be modeled in str2
66                  occupiedin3 = 0;
67              elseif ((num2_3==1)&mc)                                 % if label1 is not found but label2 is
                    found it must be a mutated position (only residue names don't match).
68                  occupiedin3 = str3.occ(ind2_3(1));                  %     Only count as occupied if mainchain
                        atom; let occupiedin2 be occupancy of that mc atom (should be 1).
69              end
70          case 1
71              occupiedin3 = str3.occ(ind1_3(1));                      % if label1 is found once then let
                    occupiedin2 be whatever the occupancy of that atom.
72          case 2
73              occupiedin3 = 1;                                        % if label2 is found twice then it must
                    be modeled as alternate confs in str2.
74      end
75
76      ind1_4 = find(strcmp(str1.label(n),str4.label));                % where atom label1 occurs in str4
            (0,1,2)
77      [num1_4,blah] = size(ind1_4);                                   % how many times label1 occurs in str4
78      ind2_4 = find(strcmp(str1.label2(n),str4.label2));              % where atom label2 occurs in str4
            (0,1,2)
79      [num2_4,blah] = size(ind2_4);                                   % how many times label2 occurs in str4
80      switch num1_4                                                   % is the atom found in structure 2?
81          case 0
82              if (num2_4==0)                                          % if label1 is not found and label2 is
                    not found then it must not be modeled in str2
83                  occupiedin4 = 0;
84              elseif ((num2_4==1)&mc)                                 % if label1 is not found but label2 is
                    found it must be a mutated position (only residue names don't match).
```

```
85                  occupiedin4 = str4.occ(ind2_4(1));                  %     Only count as occupied if mainchain
                        atom; let occupiedin2 be occupancy of that mc atom (should be 1).
86              end
87          case 1
88              occupiedin4 = str4.occ(ind1_4(1));                      % if label1 is found once then let
                    occupiedin2 be whatever the occupancy of that atom.
89          case 2
90              occupiedin4 = 1;                                        % if label2 is found twice then it must
                    be modeled as alternate confs in str2.
91      end
92
93      if ((num1_2>=1) & (num1_3>=1) & (num1_4>=1))                    % if label1 occurs at least once in each
            then it is common
94          common = 1;
95      elseif (mc & (num2_2>=1) & (num2_3>=1) & (num2_4>=1))           % otherwise, only if it is a mainchain
            atom common to all
96          common = 1;                                                 % will it be considered common. (ie
                mutated position)
97      else
98          common = 0;
99      end
100
101     occupiedinall = occupiedin1 & occupiedin2 & occupiedin3 & occupiedin4;  % is the atom found in all 4
            structures?
102
103     if occupiedinall                                                % if atom is in all or if mainchain
            common to all...
104                                                                     % assign str1 coordinates and pe
105         if ((num1_1 == 2) & alt)                                    % if atom is found twice in str1 and alt
                is on, coords and pe are weighted
106             x1 = str1.x(ind1_1(1))*str1.occ(ind1_1(1)) + str1.x(ind1_1(2))*str1.occ(ind1_1(2));
107             y1 = str1.y(ind1_1(1))*str1.occ(ind1_1(1)) + str1.y(ind1_1(2))*str1.occ(ind1_1(2));
108             z1 = str1.z(ind1_1(1))*str1.occ(ind1_1(1)) + str1.z(ind1_1(2))*str1.occ(ind1_1(2));
109             pe1 = ((str1.occ(ind1_1(1))*str1.poserr(ind1_1(1)))^2 + (str1.occ(ind1_1(2))*str1.poserr(ind1_1
                    (2)))^2)^0.5;
110         else                                                        % if atom found once, then coords and pe
                are same
111             x1 = str1.x(ind1_1(1));
112             y1 = str1.y(ind1_1(1));
113             z1 = str1.z(ind1_1(1));
114             pe1 = str1.poserr(ind1_1(1));
115         end
116                                                                     % assign str2 coordinates and pe
117         if ((num1_2 == 2) & alt)                                    % if atom is found twice and alt is on
                then weight
118             x2 = str2.x(ind1_2(1))*str2.occ(ind1_2(1)) + str2.x(ind1_2(2))*str2.occ(ind1_2(2));
119             y2 = str2.y(ind1_2(1))*str2.occ(ind1_2(1)) + str2.y(ind1_2(2))*str2.occ(ind1_2(2));
120             z2 = str2.z(ind1_2(1))*str2.occ(ind1_2(1)) + str2.z(ind1_2(2))*str2.occ(ind1_2(2));
121             pe2 = ((str2.occ(ind1_2(1))*str2.poserr(ind1_2(1)))^2 + (str2.occ(ind1_2(2))*str2.poserr(ind1_2
                    (2)))^2)^0.5;
122         else                                                        % otherwise, assign to values for index
                of label2, first element
123             x2 = str2.x(ind2_2(1));
124             y2 = str2.y(ind2_2(1));
125             z2 = str2.z(ind2_2(1));
126             pe2 = str2.poserr(ind2_2(1));
127         end
128         i2 = ind2_2(1);
129                                                                     % assign str3 coordinate and pe
```

```
130         if ((num1_3 == 2) & alt)                                    % if atom is found twice and alt is on
                then weight
131             x3 = str3.x(ind1_3(1))*str3.occ(ind1_3(1)) + str3.x(ind1_3(2))*str3.occ(ind1_3(2));
132             y3 = str3.y(ind1_3(1))*str3.occ(ind1_3(1)) + str3.y(ind1_3(2))*str3.occ(ind1_3(2));
133             z3 = str3.z(ind1_3(1))*str3.occ(ind1_3(1)) + str3.z(ind1_3(2))*str3.occ(ind1_3(2));
134             pe3 = ((str3.occ(ind1_3(1))*str3.poserr(ind1_3(1)))^2 + (str3.occ(ind1_3(2))*str3.poserr(ind1_3
                    (2)))^2)^0.5;
135         else                                                        % otherwise, assign to values for index
                of label2, first element
136             x3 = str3.x(ind2_3(1));
137             y3 = str3.y(ind2_3(1));
138             z3 = str3.z(ind2_3(1));
139             pe3 = str3.poserr(ind2_3(1));
140         end
141         i3 = ind2_3(1);
142                                                                     % assign str4 coordinate and pe
143         if ((num1_4 == 2) & alt)                                    % if atom is found twice and alt is on
                then weight
144             x4 = str4.x(ind1_4(1))*str4.occ(ind1_4(1)) + str4.x(ind1_4(2))*str4.occ(ind1_4(2));
145             y4 = str4.y(ind1_4(1))*str4.occ(ind1_4(1)) + str4.y(ind1_4(2))*str4.occ(ind1_4(2));
146             z4 = str4.z(ind1_4(1))*str4.occ(ind1_4(1)) + str4.z(ind1_4(2))*str4.occ(ind1_4(2));
147             pe4 = ((str4.occ(ind1_4(1))*str4.poserr(ind1_4(1)))^2 + (str4.occ(ind1_4(2))*str4.poserr(ind1_4
                    (2)))^2)^0.5;
148         else                                                        % otherwise, assign to values for index
                of label2, first element
149             x4 = str4.x(ind2_4(1));
150             y4 = str4.y(ind2_4(1));
151             z4 = str4.z(ind2_4(1));
152             pe4 = str4.poserr(ind2_4(1));
153         end
154         i4 = ind2_4(1);
155                                                                     % calculate ddr values
156         str1.ddx(n) = (x2 - x1) - (x4 - x3);
157         str1.ddy(n) = (y2 - y1) - (y4 - y3);
158         str1.ddz(n) = (z2 - z1) - (z4 - z3);
159         str1.ddr(n) = (str1.ddx(n)^2 + str1.ddy(n)^2 + str1.ddz(n)^2)^0.5;
160         prop_pe = (pe1^2 + pe2^2 + pe3^2 + pe4^2)^0.5;
161         str1.ddrn(n) = str1.ddr(n)/prop_pe;
162 %         str1.ddxn(n) = str1.ddx(n)/prop_pe;
163 %         str1.ddyn(n) = str1.ddy(n)/prop_pe;
164 %         str1.ddzn(n) = str1.ddz(n)/prop_pe;
165
166         [theta, phi, r] = cart2sph (str1.ddx(n),str1.ddy(n),str1.ddz(n));
167         [str1.ddxn(n), str1.ddyn(n), str1.ddzn(n)] = sph2cart (theta, phi, str1.ddrn(n));
168                                                                     % assign to appropriate indices in other
                                                                            structures
169         str2.ddx(i2) = str1.ddx(n); str3.ddx(i3) = str1.ddx(n); str4.ddx(i4) = str1.ddx(n);
170         str2.ddxn(i2) = str1.ddxn(n); str3.ddxn(i3) = str1.ddxn(n); str4.ddxn(i4) = str1.ddxn(n);
171         str2.ddy(i2) = str1.ddy(n); str3.ddy(i3) = str1.ddy(n); str4.ddy(i4) = str1.ddy(n);
172         str2.ddyn(i2) = str1.ddyn(n); str3.ddyn(i3) = str1.ddyn(n); str4.ddyn(i4) = str1.ddyn(n);
173         str2.ddz(i2) = str1.ddz(n); str3.ddz(i3) = str1.ddz(n); str4.ddz(i4) = str1.ddz(n);
174         str2.ddzn(i2) = str1.ddzn(n); str3.ddzn(i3) = str1.ddzn(n); str4.ddzn(i4) = str1.ddzn(n);
175         str2.ddr(i2) = str1.ddr(n); str3.ddr(i3) = str1.ddr(n); str4.ddr(i4) = str1.ddr(n);
176         str2.ddrn(i2) = str1.ddrn(n); str3.ddrn(i3) = str1.ddrn(n); str4.ddrn(i4) = str1.ddrn(n);
177     else                                                            % otherwise set to zero
178         str1.ddx(n) = 0; str1.ddy(n) = 0; str1.ddz(n) = 0; str1.ddr(n) = 0;
179         str1.ddrn(n) = 0; str1.ddxn(n) = 0; str1.ddyn(n) = 0; str1.ddzn(n) = 0;
180     end
181 end
```

```
182
183 str1.ddx = str1.ddx'; str1.ddy = str1.ddy'; str1.ddz = str1.ddz';
184 str1.ddxn = str1.ddxn'; str1.ddyn = str1.ddyn'; str1.ddzn = str1.ddzn';
185 str1.ddr = str1.ddr'; str1.ddrn = str1.ddrn';
186
187 str2.ddx = str2.ddx'; str2.ddy = str2.ddy'; str2.ddz = str2.ddz';
188 str2.ddxn = str2.ddxn'; str2.ddyn = str2.ddyn'; str2.ddzn = str2.ddzn';
189 str2.ddr = str2.ddr'; str2.ddrn = str2.ddrn';
190
191 % str3.ddx = str3.ddx';
192 % str3.ddy = str3.ddy';
193 % str3.ddz = str3.ddz';
194 % str3.ddxn = str3.ddxn';
195 % str3.ddyn = str3.ddyn';
196 % str3.ddzn = str3.ddzn';
197 % str3.ddr = str3.ddr';
198 % str3.ddrn = str3.ddrn';
199 %
200 % str3.ddx = str3.ddx';
201 % str3.ddy = str3.ddy';
202 % str3.ddz = str3.ddz';
203 % str3.ddxn = str3.ddxn';
204 % str3.ddyn = str3.ddyn';
205 % str3.ddzn = str3.ddzn';
206 % str3.ddr = str3.ddr';
207 % str3.ddrn = str3.ddrn';
208
209 str1.ddr_sets = [str1.set str2.set str3.set str4.set];
210 str2.ddr_sets = [str1.set str2.set str3.set str4.set];
211 str3.ddr_sets = [str1.set str2.set str3.set str4.set];
212 str4.ddr_sets = [str1.set str2.set str3.set str4.set];
213
214 if alt == 1
215     str1.ddrmsg = 'alternate conformations used'; str2.ddrmsg = 'alternate conformations used';
216     str3.ddrmsg = 'alternate conformations used'; str4.ddrmsg = 'alternate conformations used';
217 else
218     str1.ddrmsg = 'alternate conformations not used'; str2.ddrmsg = 'alternate conformations not used';
219     str3.ddrmsg = 'alternate conformations not used'; str4.ddrmsg = 'alternate conformations not used';
220 end
```

Note, the protein model files (.pdb) made as a result of refinement in phenix (and saved by Pymol and Coot) are of a different format that necessary to input to the above code. Below are three lines of bash scripting that modify the pdb files for appropriate input submission into the matlab files above.

```
1 awk -F 'FS' 'BEGIN{FS="\t"}{for (i=1; i<=NF-1; i++) if(i<3 || i>5) {printf $i FS
      };{print $NF}}'
2
3 cut -c-72,78-
4
```

```
5 awk -F"," 'BEGIN{a="";for(i=1;i<=1;i++){a=" "a}} { print $0""a}'
```

5.3.2 Bacterial-2-Hybrid: GFP readout

The bacterial-2-hybrid assay used to determine the affinities of all possible single mutants of H372A PSD95^{pdz3} was implemented directly as described in McLaughlin *et al.*. The library of PDZ domains was transformed into MC4100 *E.coli* cells already containing the pZA and pZE plasmids. These cells were recovered for 1 hour in ZYM 505 medium and the transformation efficiency was quantified by plating 1μL of recovery onto a plate with trimethoprim, chloramphenicol, and ampicillin. The entire 1 mL transformation was added to 10 mL ZYM 505 + kan(30μg/mL)+amp(50μg/mL)+clm(25μg/mL) culture in a 50 mL beveled flask and grown 6 hours at 37 degrees at 225 rpm. 6 hour growths are then diluted to 10 μL culture per 10 mL of ZYM 505 plus antibiotics and grown for 12 hours at 37 degrees C at 225 rpm shaking. A 35 μL aliquot of the culture is added to one well of a 48 well plate containing 500 μL LB + antibiotics + anhydrous tetracycline (aTC, the inducer for the peptide, 100 ng/mL) + IPTG (the inducer for the PDZ domain, 100 μM). Note, it is important here to use as fresh of aTC as possible. The plate is then induced at 18 degrees shaking at 150 rpm for 2 hours. Induced cells are diluted to 30 μL of cells per 1 mL of filter-sterilized M9 + 0.4% glucose for cytometry. The cells are passed through a 30 gauge needle for disaggregation into single cells. The cultures are sorted using fluorescence activated cell sorting (MoFlo instrument, UTSW cytomtetry core facility, Angela Mobley). The cells were sorted with a 10% gate into M9 medium (a deviation from the protocol of McLaughlin *et al.*. McLaughlin *et al.* sorted into ZYM505, but in my experience, ZYM505 gave high background counts whereas M9 was much cleaner of medium). The sorted cells are then poured into a 10 mL culture of ZYM505 with antibiotics. These cells are grown to full density then prepped along with the input population. The preps are then sequenced with a MiSeq Illumina platform. Files generated from the MoFlo device (.fcs files) are available electronically for the experiments performed here.

References

[1] R. N. McLaughlin, F. J. Poelwijk, A. Raman, W. S. Gosal, and R. Ranganathan, "The spatial architecture of protein function and adaptation.," *Nature*, vol. 491, pp. 138–42, Nov. 2012.

[2] Z. Otwinowski and W. Minor, "Macromolecular Crystallography Part A," *Methods in Enzymology*, vol. 276, no. January 1993, pp. 307–326, 1997.

[3] D. a. Doyle, A. Lee, J. Lewis, E. Kim, M. Sheng, and R. MacKinnon, "Crystal structures of a complexed and peptide-free membrane protein- binding domain: Molecular basis of peptide recognition by PDZ," *Cell*, vol. 85, no. 7, pp. 1067–1076, 1996.

[4] P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, and P. H. Zwart, "PHENIX: A comprehensive Python-based system for macromolecular structure solution," *Acta Crystallographica Section D: Biological Crystallography*, vol. 66, no. 2, pp. 213–221, 2010.

[5] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan, "Features and development of Coot," *Acta Crystallographica Section D: Biological Crystallography*, vol. 66, no. 4, pp. 486–501, 2010.

[6] V. B. Chen, W. B. Arendall, J. J. Headd, D. a. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: All-atom structure validation for macromolecular crystallography," *Acta Crystallographica Section D: Biological Crystallography*, vol. 66, no. 1, pp. 12–21, 2010.