Hierarchy of interactions in protein evolution

Approved by Supervisory Committee

.....

Dr. Rama Ranganathan (Advisor)

.....

Dr. Hongtao Yu (Chair)

.....

Dr. Joseph Takahashi

.....

Dr. Benjamin Tu

Hierarchy of interactions in protein evolution

A DISSERTATION PRESENTED
BY
VICTOR H. SALINAS
TO
THE FACULTY OF THE GRADUATE SCHOOL OF BIOMEDICAL SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS SOUTHWESTERN MEDICAL CENTER
DALLAS, TEXAS
MAY 2018

Thesis advisor: Professor Dr. Rama Ranganathan Victor H. Salinas

Hierarchy of interactions in protein evolution

Abstract

Deciphering the relationship between genotype and phenotype is complicated by the sheer number of possible cooperative interactions amongst the parts that make up biological systems. For even small systems such as individual protein domains, it has been difficult to comprehensively obtain high quality empirical data of amino acid interactions to distinguish different models for the global pattern of cooperativity. The statistical coupling analysis (SCA) - one approach for studying the co-evolution of amino acid positions in homologous sequences - provides a model for this pattern that is distinct from spatial proximity in tertiary structure, positional conservation, or even other forms of co-evolution. Here, we use an extension of deep mutational scanning to analyze nearly 50,000 single and double mutations in several homologs of a model protein - the PDZ family of protein interaction domains. Across the domains queried experimentally, the distributions of couplings between pairs of positions from all possible double mutants are well-approximated by unimodal distributions such that their average provides an estimate of the intrinsic coupling between them. Importantly, the SCA provides the best representation of this experimental pattern of couplings conserved among the homologs. These results highlight the heterogeneous pattern of couplings in protein structures and motivate the re-focus of efforts to understand protein folding and function toward the study of the origin of the co-evolving network of amino acids.

Contents

Listing of figures

Listing of tables

I dedicate this work to my parents, Victor and Ligia, and my wife, Meagen.

Acknowledgments

My mentor Dr. Rama Ranganathan will find himself acknowledged in all my future scientific pursuits. He has deeply influenced how I approach all things science: from my taste in problems and how I think about them to the very words I write or speak to communicate ideas. For all of his wisdom and support, the challenging intellectual environment to which he welcomed me, and for inculcating that which Richard Feynman called "the pleasure of findings things out," I am deeply indebted.

One of the most important ways through which Rama effected his influence was via the group of individuals with whom he populated his lab. These were individuals, comprising former and current lab members, whose scientific rigor and expertise in a wide range of disciplines was matched by a willingness to discuss anything from problems encountered in one's own work to an article in the most recent *New Yorker*. I particularly want to single out Drs. Frank Poelwijk and Michael Stiffler for their help in defending the merits of what amounted to be a crazy mutagenesis project, for their instruction on how to properly do and analyze experiments, and for their constant encouragement for me to figure out things on my own and for the guidance that kept me on the right path.

I also want to thank the scientific and administrative support staff in the lab and at the school, specially Eric Bonventre, Mike Socolich, Kathy Lyons, and Sarah Beauregard, who supported my capricious ordering and requests for reagents and insulated me from who-knows-what so that I could focus on my work; and Robin Downing and Dr. Andrew Zinn, who welcomed and helped me transition to the Medical Scientist Training Program and have advocated for my education all along. My thesis committee, comprising Drs. Hongtao Yu, Joe Takahashi, and Ben Tu, provided deep insight into each of my projects and reinforced the primacy of doing research–finding answers to scientific questions–before all else, including the desire for recognition or publication.

Various scientists outside of UT Southwestern contributed to many aspects of my projects: Dr. Taylor Johnson and his laboratory at UT Arlington, together with Ian White in the lab, for their design and construction of a continuous culture device for growing bacteria; and Drs. Susan Golden at UCSD and Takao Kondo at Nagoya University for reagents (strains and plasmids) and advice in carrying out the cyanobacteria experiments.

I would be remiss not to acknowledge my undergraduate mentor and dear friend, Dr. Loise Francisco, who directed me toward a career in research and has advised me all along, and Dr. Mark Gilbert, someone whom I hold with the highest esteem and who has been a source of hope for me and my wife.

Finally, I thank my parents, Ligia and Victor, and my beautiful wife, Meagen, for their sacrifice, unconditional love, inspiration and perpetual encouragement to pursue my passion, this time in the form of a long and challenging learning experience. I am also grateful to my sibilings, Hugo and

Karina, and my extended family for the love, support, and understanding they've provided over the years, as well as my close friends, abroad and in the lab, who were always there when I needed someone to listen or to give me an honest opinion, preferably over a beer.

0
Introduction

Its publication in 1953 challenged leading hypotheses championed by mammoths in the discipline. More than simply catapulting two scientists to academic and popular stardom, it defined a legacy intimately related to the history and course of human civilization.

The contribution of Michael Ventris and John Chadwick was the decipherment of Linear B, which irrefutably showed that the Mycenaean civilization of Crete spoke Greek and pushed back the advent of a written script for the Greek language by at least 500 years[22].

Their discovery shares similarities with a contemporary work that often overshadows it[*]. The description of the structure of deoxyribonucleic acid (DNA)[177], published by James Watson and Francis Crick in 1953, also represents a seminal finding on a subject previously beset by limited evidence and methodologies that ultimately demanded a reinterpretation of prior work and paved the way for new paths of inquiry. And while the comparison between the DNA sequence of a genome and words on a book may have already transcended the threshold from figurative to literal[25], Eric Lander best summarizes the differences between these two accomplishments of the twentieth century: "Genome. Bought the book. Hard to read."[111] The genome, unlike Linear B, remains undeciphered.

This dissertation concerns itself with the problem of how information is encoded in the genetic

[*]Similarities also exhibited by the other notable feat of 1953, the ascent to the summit of Mount Everest by Sir Edmund Hillary and Tenzing Norgay.

material of cells, the so-called 'genotype-phenotype' map[39,33]. This problem is instantiated in diverse fields of biology such as the linking of genetic variants to complex traits in genetics[91,100], the elucidation of fitness landscapes in evolutionary biology[132], and the prediction of structure and function from sequence in biochemistry and structural biology[37]. The work detailed in subsequent chapters espouses the philosophy (and corresponding approaches) that evolutionary conservation, *i.e.* the invariance through time, of genome elements and interactions among them reveal the constraints underlying the mapping between genotype and phenotype: the former being a statistical representation of functional importance[190], while the latter reflecting the fact that most traits or phenotypes are the outcome of many interacting genes[64,191,178]. This philosophy resonates with basic and common operations and experiences in biology, from the comparison of aligned sequences to the observation that genetically engineered mouse models exhibit different characteristics depending on the background on which they were generated[90].

Chapter 1 introduces the approach of co-evolution, which parses the record of extant DNA or protein sequences to infer the functional interactions of components from their statistical correlations. This approach has been successfully applied to identify groups of amino acids in proteins that specify the function and structure of the polypeptides, and it presents a model for the encoding of biological information that is experimentally tested in Chapter 2.

Complexity in biological systems emerges from a hierarchy of interactions of heterogeneous components[137], from the atoms that compose all matter, to the amino acids that constitute proteins, to groups of proteins that interact through physical contact or metabolites to perform cellular tasks. Chapter 2 explores the pattern, pervasiveness, and predictability of interactions between amino acids in a protein, whereas Chapter 3 lays the groundwork for experimentally interrogating how interactions between proteins and the environment influence the encoding of constraints on the protein sequence. The respective model systems, namely the PDZ family of protein-protein interaction domains and the *kai* oscillator of cyanobacteria, reflect the convenience they offer for addressing these questions (rich literature with abundant data, availability of experimental assays, a general understanding of their function in the context of the organism).

Many of these experimental endeavors required specific technological developments, one of which was a turbidostat for carrying out well-controlled, continuous culture experiments in bacterial cells. Chapter 4 describes the design and implementation of such a device, which was engineered by the graduate students of Dr. Taylor Johnson's laboratory at the University of Texas at Arlington in collaboration with Kristopher Ian White, a graduate student at UT Southwestern in the Ranganathan lab.

Collectively, these efforts aspire to characterize features of the genotype-phenotype map that may

lead to a simplification of the problem or its re-formulation into more fundamental questions, as well as the enablement and prescription of experiments to pursue answers to these. The dissertation culminates in Chapter 5, in which the contribution of this work is assessed and an outlook of the 'decipherment' of the genome is presented.

1

Co-evolution: a synthesis of methods and applications

How are biological properties of proteins encoded in the sequence of constituent amino acids? A global and oft-employed approach, now bolstered by ever-increasing databases of genomic data, is to parse the evolutionary record of protein families for signatures of conservation. The invariance of amino acid identities in groups of related sequences is regarded as a statistical representation of functional importance[190]. However, residues in proteins are engaged in higher-order interactions that contribute to activity and stability, which underlies the expectation that the functional contribution of an amino acid is not an independent attribute. As a result, methods to analyze amino acid co-evolution emerged with the key assumption that understanding the statistical constraints between residues would yield more informative insight into the sequence determinants of structure and function[56,99].

Over the years, a substantial body of work has lent support to this assumption, demonstrating that the information encoded in just pairwise correlations between amino acids in multiple sequence alignments (1) can identify positions throughout the protein involved in enzyme catalysis, ligand binding, allosteric networks, and novel phenotypes and (2) can be successfully harnessed for protein design[160,145]. More recently, various groups used co-evolutionary information to infer three-dimensional structures for numerous protein families[102,69], fulfilling the goal of the earliest endeav-

ors that analyzed correlated (compensatory) mutations[56,6,126].

A dichotomy thus emerges between two classes of approaches that ostensibly survey the same data (amino acid correlations in multiple sequence alignments) yet reveal either global, contiguous networks of residues linking distal sites in proteins on the one hand, versus local, physically contacting residues on the other.

Here, I argue that the dichotomy is merely superficial, reflects distinct motivations driving the analyses, and obscures their fundamental connection as parts in a hierarchy of the information content encoded in ensembles of evolutionarily-related sequences. To clarify these contentions, I begin by coarsely describing the methodologies and their respective goals before culminating in a synthesis that seeks to motivate further research in this area, which has the potential for furthering our understanding of protein structure, function, and evolution.

Analysis of correlated mutations: goals, implementations, and findings

The notion that the effect of a mutation is context-dependent was already acknowledged at the dawn of comparative sequence analysis[189]. An interaction between two amino acids that is essential to protein function or fold may constrain the number or nature of substitutions at those positions during evolution. These constraints may then be reflected in the statistics of amino acid frequencies and their correlations in multiple sequence alignments (Fig. 1.1 *A*), which several algorithms have extracted and linked to important phenotypes, including substrate specificity and protein stability in the serine proteases[62] and bacterial histidine kinase-response regulator (HK-RR) pairs[158]. Their success depended on analytical methods, heuristics, and intuition to deal with phylogenetic bias (sequence identity reflecting descent from common ancestor and small time window of sequence divergence) and sampling noise (the number of sequences being analyzed sample a miniscule fraction of the possible pairwise combinations of amino acids at any two sites). The statistical coupling analysis (SCA) algorithm employs both weighting of the amino acid covariance matrix by conservation-dependent weights to emphasize conserved correlations as well as spectral decomposition to isolate groups of residues with the strongest contribution to the correlations[99,62]. In addition, SCA and other methods based on mutual information (MI) calculations also use alignment randomization (shuffling of amino acids within a column of the multiple sequence alignment) to identify spurious correlations arising from sequence relatedness[62,158]. That these approaches work reflects one of the fundamental results from co-evolutionary analyses: the correlations are sparse, with only a subset of all amino acids engaging in statistical interactions. This yields a description of proteins in which functionally important residues are beset by unconstrained ones and motivates the experimentally tested hypothesis that these are the regions of the protein underlying evolvability and robustness,

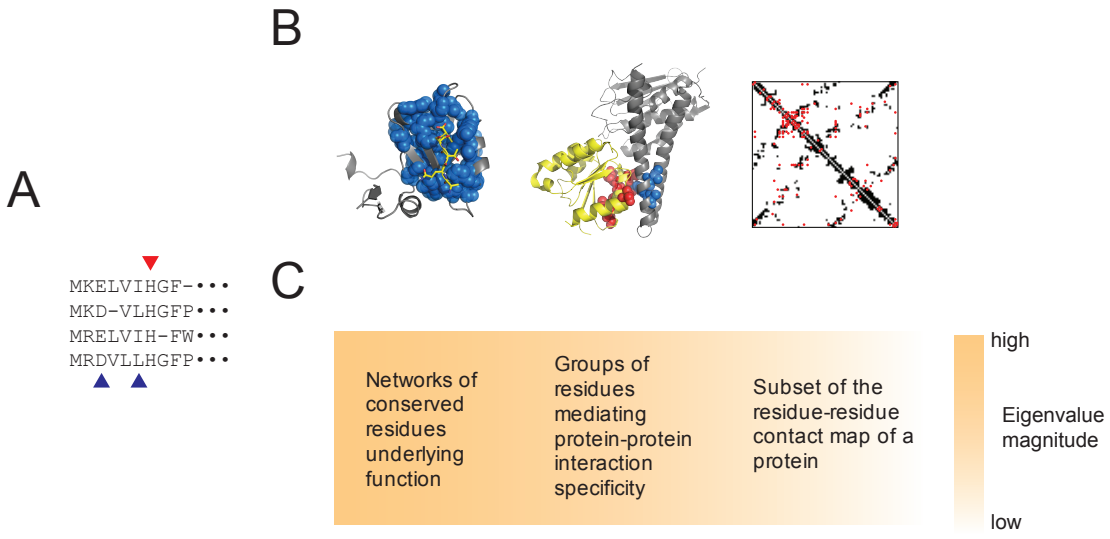

Figure 1.1: *A*, Co-evolution is computed from the co-occurrence of amino acid substitutions at pairs of positions in a multiple sequence alignment. The two positions marked by the blue arrows exhibit co-evolution: the identity of the residue on the left always covaries with the identity of the one on the right. The residue marked by a red arrow, on the other hand, shares no co-evolution with either of the previous positions as it is completely conserved. *B*, the output of co-evolutionary analysis from the SCA, MI, and DCA methods, with descriptions arranged according to the region of the eigenspectrum of the covariance matrix the methods emphasize (*C*). From *left* to *right*, the SCA algorithm identifies a conserved network of co-evolving amino acids in the PDZ domain, here depicted on the structure of a member of this protein family (PDB: 1BE9); MI extracts positions at the interface between a histidine kinase (*grey*)-response regulator (*yellow*) pair (PDB: 3DGE) that underlies the specificity of the interaction; DCA infers the contact-graph of a protein (here the PDZ domain family, Pfam: PF00595) based on a model of the direct interactions explaining the calculated covariance matrix (*red*, predicted contacts overlaid on the actual contact map of PDB 1BE9).

respectively[105].

The MI and amino acid covariance matrices between residues i and j are equivalent[30] and are computed as:

$$MI_{i,j} = \sum_{x=1}^{20}\sum_{y=1}^{20} f_{i,j}^{x,y} \log \frac{f_{i,j}^{x,y}}{f_i^x f_j^y} \tag{1.1}$$

$$C_{i,j}^{x,y} = f_{i,j}^{x,y} - f_i^x f_j^y \tag{1.2}$$

Importantly, in SCA, the $C_{i,j}^{x,y}$ 4-dimensional tensor that expresses the covariance between amino acids x and y at residues i and j is further weighted by a conservation-based function, the gradient of the Kullback-Leibler divergence of the individual amino acids at a position

$$\phi_i^x = \log \frac{f_i^x(1-q^x)}{(1-f_i^x)q^x} \tag{1.3}$$

where q^x represents the background probability of observing amino acid x as estimated from the non-redundant protein database. The weighted SCA matrix, $\tilde{C}_{i,j}^{x,y}$, is expressed as

$$\tilde{C}_{i,j}^{x,y} = \phi_i^x \phi_j^y C_{i,j}^{x,y} \tag{1.4}$$

Once again, this has the outcome of emphasizing correlations between more conserved residues.

The two analytical approaches have been used effectively on a myriad of protein families, identifying: specificity-determining residues in bacterial histidine kinase-response regulator (HK-RR) pairs[158]; residues that underlie substrate specificity and protein stability in serine proteases[62]; residues that are sensitive to mutation and that mediate specificity switching in the PDZ domain[105]; 'hotspots' for allosteric control of enzymatic function in *E. coli* dihydrofolate reductase[139]; allosteric networks of amino acids in Hsp70 chaperones[159], TonB-dependent transporters[45], G proteins[65], and G protein-coupled receptors[165]; and drug-targetable allosteric sites in cathepsin K peptidase[116] (Fig. 1.1 *B*). The greater testaments to the relevance of the correlations captured by the methods, however, were the successful demonstrations of the sufficiency of the pairwise statistics to: (1) enable the engineering of non-natural HK-RR pairs by rationally introducing mutations to switch specificity[158,19] and (2) design non-natural members of a protein family that fold and function akin to their natural counterparts[160,145,140].

Nevertheless, the two approaches exhibit different performances in predicting functional residues in distinct model systems[49,30]. I will discuss a deeper connection between MI and SCA in a subsequent section but note that the disparity in performance likely stems from the conservation of the phenotype under scrutiny: as Colwell et al.[30] point out, if the phenotype varies among sequences in the alignment, then the residues that control it ought to reflect this variation. For example, recognition specificity is likely more diverse (less conserved) in HK-RR pairs, as these are encoded by numerous genes (from 20-30 up to 300) even in a single bacterial species, and the avoidance for crosstalk is essential for organismal fitness[158]. A weighting function that reflects the conservation of the phenotype in the alignment, then, will better identify those residues that control it. These observations also bring to light the biggest limitation of co-evolution detection methods: that despite more sophisticated methods and models, for most biological systems, we lack knowledge of the fitness constraints that act on them. The poor correspondence between the computed statistical parameters and experimental measurements of function may simply be due to comparisons with phenotypes that inadequately describe the system's fitness function, or at worst, are the biochemist's

spandrels[58].

Direct Coupling Analysis: aims, approaches, and revelations

The tangibility of molecular structure naturally made it the target phenotype for earlier efforts that looked at correlated mutations. The small free energy difference between the folded and unfolded states of a protein imply that at least a fraction of the atomic interactions (van der Waals, ionic, hydrogen bonds, covalent disulfide bonds) that contribute to the folded state should be present in sequences that constitute a protein family. Statistical analysis of co-evolution, consequently, presented itself as a method to calculate and begin to understand nonadditivity in more complex systems[35].

Any two amino acid positions may be correlated through transitive interactions with other positions. Consider two amino acids, x and y, at positions i and j, respectively. Despite not directly interacting, they may exhibit correlation through interaction with amino acid z at position k, and the magnitude of the correlation would also reflect all other paths of indirect connectivity through other amino acids. The problem of distinguishing direct from indirect interactions from amino acid correlations as a way of arriving at residue-residue contacts was thus framed and tackled using Bayesian[15], statistical physics[89,179,29], and pseudolikelihood[7,81,41] approaches. The general aim of these methods is to generate a global statistical model that describes a set of interactions consistent with the statistical correlations from the sequence alignment. These efforts have succeeded at predicting cognate pairs of HK-RR in genomes as well as interface residues[15,179], providing accurate contact maps for hundreds of protein families with verifiable structures[102,109], predicting 3D structures of experimentally less tractable proteins, such as those in the membrane[69,71] and those forming multi-protein complexes[72,123], and revealing contacts not in the native state that suggest different structural conformations[108] (Fig. 1.1 *B*).

The success of these methods suggests that sequences that satisfy a given protein fold inhabit a much lower-dimensional space than suggested by examining any given protein structure with thousands of residue interactions: in the published works, the contacts used to predict folds (with C_α distance RMSD values relative to reference structures less than 5.0 Å) represented only a tiny fraction. Moreover, the approaches ignore, either explicitly or implicitly, higher-order interactions. Thus, protein structure, like protein function, is encoded in a small number of amino acids in a protein.

A deep question remains concerning the validity of the approaches and the approximations employed. Given that they are fruitful, what do they teach us about the evolution and biophysical organization of proteins? The maximum entropy model, for example, is but one of many models that could describe the statistical correlations and it may fare worse upon comparison[44]. Finally,

given the structure-function paradigm so pervasive in biology[37], how can one rationalize the apparent lack of correspondence between predictions by the above methods of residues that are important for function (as in SCA and MI) and for structure (as in DCA)?

Amino acid co-evolution: a synthesis

Certainly, a cursory reconciliation can be made by considering that methods such as DCA strive to identify only direct spatial interactions. Functionally important interactions such as those identified by SCA, on the other hand, could also take the form of spatially distant amino acids interacting to stabilize certain transitory conformations or allosteric communication between an active site and a surface patch on opposite sides of a protein through intervening networks of amino acids[99]. The protein design experiments of Russ et al.[145] and Socolich et al.[160], however, suggest a degree of overlap between the two groups of 'interacting' amino acids, in that information about the conserved and correlated amino acids was sufficient to recapitulate both folded and active proteins. Another interpretation of these results is that function itself specifies structure.

The interpretation I advocate is based on the independent theoretical findings of Cocco et al.[29] and Rivoire[142]. Therein, the authors show that the SCA and DCA methods operate on extreme ends of the eigenspectrum of the amino acid covariance matrix, with SCA emphasizing the higher modes (those associated with larger eigenvalues) and DCA the lower significant modes[29,142] (Fig. 1.1 *C*). Together with the experimental findings elucidated above, I propose that encoded in the covariance matrix is a hierarchy of information specifying protein phenotypes (activity, fold) that define the historical fitness[92] of a protein family and on which evolutionary forces (drift, selection) have acted at different magnitudes and timescales.

This decomposition is consistent with the salient conclusions of past work and provides a model for proteins with testable predictions. First, it is supported by the observation that weighting functions emphasize protein phenotypes having different degrees of divergence in sequence alignments[30]. In this hierarchy, the amino acids identified by SCA would control traits with the highest degree of conservation, succeeded by those revealed by MI, and finally the residue-residue contacts found by DCA. This not only suggests that the amino acids specifying fold exhibit the least evolutionary constraint, but that the sequence space consistent with a protein structure is potentially larger than that underlying function. Second, the statistical decomposability implies a degree of modularity (independence) among the different modes, with each still comprising only a subset of amino acids in the protein (consistent with a sparse information encoding).

The experiments I envision to test this model undertake distinct goals. I briefly outline them below.

Information content and encoding in sequences. Why might two positions exhibit a correlation signal in a multiple sequence alignment? To begin to address this question, one requires not only experimental systems for which one can measure fitness-contributing phenotypes but also the right framework with which to understand the nature of the correlations. The advent of high-throughput sequencing has fueled experimental evolution efforts to evolve novel phenotypes[11,87] or determine the functional effects of a large number of mutations[105,51,68,143]. As mutations drive the correlations in the alignments, it is a natural course of action to see how patterns of mutational effects correlate with alignment statistics. McLaughlin et al.[105] report a statistically significant correlation between SCA-identified 'sector' positions and positions that were sensitive to single mutation and mediated a specificity switch in the PDZ domain. *A priori*, a concordance between a metric of statistical coupling and single mutant effects is not expected, and indeed, the results could be equally explained by single-site (conservation) statistics[169]. To better test the significance of the statistical correlations, I propose to comprehensively measure the nonadditivity, or epistasis, between pairwise mutations in proteins, as has been recently reported for the protein G B1 domain[118]. To discriminate between patterns of nonadditivity idiosyncratic to a single member of a protein family, however, these experiments ought to be conducted in other orthologues that sample the sequence (and, if possible, known functional) diversity to compare the conservation of such patterns in a manner analogous to comparative sequence analysis. Advances in gene synthesis[86] and high-throughput methods to assess protein function[52] are expected to facilitate these endeavors.

Modularity, sparsity, and sufficiency of sequence correlations. Following Russ et al.[145] and Socolich et al.[160], protein design experiments provide the most unambiguous means by which to test the information content in sequence statistics and their sufficiency in encompassing the relevant parameters of protein function. I propose to extend efforts along this line of inquiry by designing synthetic members of protein families that satisfy, for example, the SCA, MI, and DCA constraints and characterizing the libraries of proteins using functional and structural assays. The goal is to ascertain the information specified by them: do SCA-designed sequences show marginal stability? Are DCA-sequences folded but weakly- or non-functional? An application of this approach could also be used to distinguish natural sequences from synthetic ones designed using physical potentials. The finding that the constraints of fold and function in natural sequences are sparsely encoded seems, at first glance, in contradiction to the optimization of precise packing interactions that underlies computational protein design. Along these lines, Ollikainen & Kortemme[117] recently reported that structurally-optimized sequences show patterns of amino acid covariation similar to natural sequences only when backbone flexibility in the design algorithm was relaxed.

Mechanism generating sparse correlations. I envision these efforts to culminate with an attempt

to understand the generative process underlying the amino acid correlations. The sparsity and modular structure of the correlations beg a comparison to the properties of biological systems of robustness and evolvability[55]. Consistent with this and more recent theoretical work[66], I argue that an evolutionary history of fluctuating selection pressures can give rise to this observed pattern of sparse and connected network or amino acids. Novel platforms for carrying out parallel, continuous evolution experiments[43] with the capacity to control the strength and nature of the selective pressure, as well as the ability to monitor the evolving populations over time, now make this hypothesis experimentally feasible.

Outlook

A minimal model for proteins ought to describe their function, structure, dynamics, evolutionary history, and how these are all encoded in the primary sequence of amino acids. Such a minimal model is suggested by the analysis of amino acid covariation in multiple sequence alignments, which reveals a hierarchy of information levels that correspond to properties of proteins which show different spatial distributions and rates of evolutionary divergence. Experiments, now made possible by existing technology, promise to test the validity of the model.

2

A conserved pattern of energetic interactions underlying function in the PDZ domain

The precise structure and function of a protein reside in global patterns of local and nonlocal interactions between amino acids[36]. Understanding the pattern of these interactions is the implicit goal in efforts from diverse fields seeking to determine the contribution of polymorphisms to complex phenotypes (*e.g.* oncogenicity and disease risk)[100,191,178,176], to predict evolutionary trajectories of local adaptations[132,33] and the effects of mutations[113], and to inform physical models for protein engineering[146] and the prediction of protein structure directly from sequence[37,103]. However, despite our knowledge of the physical forces that govern amino acid interactions, the major obstacle confronting these efforts remains the non-additivity, or cooperativity, inherent in them[73,35].

Oft-referenced models for inferring the network of interactions emerge from the hierarchical decomposition of protein structure. Secondary structure properties like α-helical periodicity impose constraints on amino acid identities reflecting the backbone and side chain interactions that confer fold stability of the structural motif[185,110,166,167,104,129] and the polarity of the protein surface and core[130,121,40]. At the level of tertiary structure, these constraints generalize to the protein contact

graph.

The pattern of co-evolution between amino acids in homologous sequences (discussed in detail in Chapter 1) has been proposed as the structural organization, distinct from the hierarchy of primary, secondary, and tertiary structure, underlying conserved biological properties[62]. At the heart of this proposal lies the assumption that an interaction between two amino acids that is essential to protein function or fold may constrain the number or nature of substitutions at those positions during evolution. Co-evolving positions typically comprise a small fraction of the protein, are physically contiguous in structure, and represent determinants of protein function and fold[99,145,62,158,105,139,102]. From these analyses, a model emerges in which *structure and function are encoded in a subset of conserved and interacting amino acids in a protein*. Experimental data supporting this model, however, amount to studying the effects of a small number of mutations in diverse model systems[99,65,157,45,62] (inadequate because of the number of mutations analyzed) or all possible single mutations in one protein[105] (inadequate because interactions cannot be assessed by making only single mutations).

The goal of this project is to experimentally elucidate the pattern of interactions in a protein to test these models. To place these efforts in the appropriate context, first I summarize the contribution of experimental and computational efforts toward our understanding of the influence that the requirements for fold and function place on amino acids in proteins. I then review the experimental framework frequently employed to interrogate interactions between amino acids followed by insight gained from applying it in various systems.

2.1 On the sequence determinants of protein fold and function

2.1.1 Inference of constraints from structural and functional analyses

The study of the sequence determinants of protein function and structure benefited greatly from the application of site-directed mutagenesis and recombinant protein expression, which provided the raw material for detailed biophysical measurements and X-ray structure determination[1,3,125]. These experiments permitted testing of early observations made on the very first protein structures (*e.g.* polarity of the core and protein surface)[130] and corroborated by comparative sequence analysis.

A ubiquitous finding that emerged from these investigations is a periodic distribution of amino acid chemical properties in an α- helix. This periodicity manifests as a pattern of polarity and hydrophobicity of amino acids in crystal structures[128] and sequence alignments[134] and in experimentally determined patterns of mutational tolerance in various model systems[60,16,8,120]. Importantly, the periodicity is consistent with the 3.6 residues/turn of the helix, and so dominant is the pattern

that synthetic peptides with an imposed periodicity of polar and nonpolar residues adopt α-helical folds even when the helical propensities of the constituent amino acids are not skewed toward helix formation[185,110]. Moreover, while the stability of this secondary structural element was primarily attributed to hydrogen bonds between the backbone amino and carbonyl groups of residues at subsequent turns of the helix, comparisons of structures and mutational studies showed that stability also emerged from intrahelical salt bridges[166,104,167] and van der Waals interactions between side chains of residues n,$n+3$ or $n+4$ residues apart (which also places them on the same side of the helix)[129]. In fact, the absence of hydrogen bonds within the peptide backbone does not preclude formation of stable α-helices, as polyproline peptoids have been shown to form these secondary structures, presumed to be stabilized in by steric side chain interactions[183].

From a more global perspective, reconciling structure with mutagenesis experiments has been more challenging. On one end of the spectrum are observations of measurements of protein stability or function correlating with little to no structural changes, from comparisons of a temperature sensitive mutant of T4 lysozyme[61], or the catalytically-dead D27N mutant of *E. coli* dihydrofolate reductase (DHFR)[77], to their respective wildtype counterparts. On the other end are observations for which the data and existing models do not provide an explanation for the relationship between the functional effects of mutations, the chemical properties of the amino acids, and the observed structures. For example, the tolerance of gene V of bacteriophage *f1* to mutations in the core[187], as well as the pattern of acceptable mutations in the α1-helix[60] and core[97,95] of the λ repressor, are rationalized by invoking some structural plasticity in these proteins that enables the accommodation of mutations by rearrangements of the protein fold extending even to distal regions. Less speculative findings of these and other studies[32,96,14], though limited by the number of mutations that could be assessed, emphasized the context dependence of mutational effects even between adjacent residues and the diversity in the tolerance of individual amino acids to mutation[97,3,13,26]. They also highlighted how the heterogeneous pattern of functional importance of amino acids undermined the utility of three-dimensional structures: the gap between understanding structure and function conveys that structural details by themselves are not sufficient to assess the energetics of either[4].

Further evidence to refine these models would emerge from more comprehensive measurements enabled by high-throughput technologies for testing more phenotypes in more protein variants. Site-saturation mutagenesis experiments coupled to massively parallel sequencing have recently elucidated, for example, the effects of large numbers of mutations in a myriad of proteins (compiled in Table 2.1 and extensively reviewed[52,63,12,34,93]). The data emanating from such experiments largely corroborate hypotheses from aforementioned work, and they also introduced evolution as an important constraint influencing the sequence to function/structure mapping in proteins. The most

salient conclusion is, once again, of a *sparse and heterogeneous pattern in the contribution of amino acids to the structure and function of proteins*.

At a superficial level, these conclusions conflict with the observation that protein packing densities approach that of crystals of inorganic molecules, suggestive of optimized interactions among complementarily-shaped residues[97]. Instead, the experiments imply a heterogeneity also present in the pattern of interactions between amino acids, a ramification that many of these investigators acknowledged but for which they lacked the requisite evidence[13,96]. Exploration of the importance of amino acid interactions, already appearing in tandem with many of the earlier mutagenesis studies and discussed below, provided hints of this. Revealing the global pattern of interactions remains an endeavor fraught with challenges to this day and forms the motivation of the present work.

2.1.2 The thermodynamic mutant cycle

As previously discussed, the contribution of an amino acid to a measurable property of protein structure or function (*e.g.* thermal stability or binding energy) is frequently determined by making a mutation at that site to one of the other 19 amino acids. Alanine mutations are often used with the justification that they test the contribution of side chain atoms beyond the β-carbon[31], but such schemes have been replaced by saturation mutagenesis where all single mutations are probed. The contribution of the residue in the wildtype state, or more accurately of the specific mutation being made, is then parametrized as the difference between the measured property of the mutant and the wildtype protein (Fig. 2.1*A*).

Interactions or couplings, in turn, are assessed by the degree of non-additivity between measurements of the thermodynamic property of mutant variants at specific residues appearing together and in isolation using the formalism of the thermodynamic mutant cycle introduced by Fersht and coworkers[20,74,1]. In this analysis, the wildtype and all single- and n-way mutant proteins are projected as the vertices on an n-dimensional hypercube, where free energies linking the protein variants on any one face sum to zero (in accordance with the first law of thermodynamics). The essence of the approach is to determine the dependence of the effect of a mutation on the background on which it occurs, and thus is analogous to calculating 'epistasis' between two genetic loci[133]. For example, we can parametrize the effect of making two mutations as the sum of the free energies of each individual mutant minus a coupling coefficient (Fig. 2.1*B*). If the coefficient is zero, then one concludes that the mutations do not interact (or they interact 'additively'), and the effect of the double mutant is exactly equal to the sum of the individual effects. On the other hand, if the mutations do interact, the coupling coefficient is nonzero and its sign and magnitude inform us on the nature and strength of the interaction, respectively.

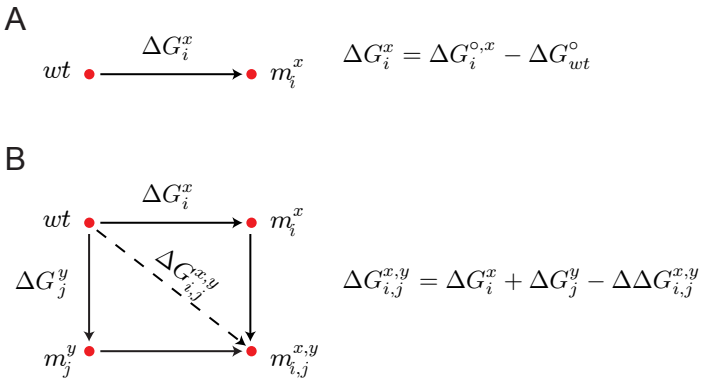

Figure 2.1: *A*, The effect of making mutant x and position i, ΔG_i^x, is parametrized as the difference between its intrinsic effect, $\Delta G_i^{\circ,x}$, and that of the wildtype, ΔG_{wt}°. *B*, Calculating the interaction between two mutations, x and y at positions i and j, respectively, requires the measurement of the individual effects and the effect of the double mutant, $\Delta G_{i,j}^{x,y}$, as in *A*. The coupling coefficient, $\Delta\Delta G_{i,j}^{x,y}$, conveys the degree to which the effect of the double mutant predicted from the sum of the individual mutant effects differs from the value obtained experimentally.

Scanned protein	*Size of mutagenized protein/region*	*Model system*	*Mutagenesis Method*	*Selection*
Fab antibody fragment[54]	50 AA	Ribosome display	Overlap extension	Ligand binding
YAP65 WW domain[51]	25 AA	T7 bacteriophage	Doped oligonucleotide	Ligand binding
E4B ubiquitin ligase[161]	102 AA	T7 bacteriophage	Doped oligonucleotide	Ubiquitination activity
T4 Lysozyme[138]	163 AA	Bacteriophage	Nonsense suppression	Plaque formation
CcdB[2]	101 AA	Escherichia coli	Megaprimer extension	Toxin activity
PSD95 PDZ domain[105]	83 AA	E. coli	Overlap extension	Ligand binding
G protein–coupled receptor[153]	376 AA	E. coli	NNN oligonucleotides	Ligand binding
lac Repressor[101]	328 AA	E. coli	Nonsense suppression	Repression of beta-galactosidase
TEM1 beta lactamase[164]	263 AA	E. coli	Overlap extension	Growth in antibiotic
Designed influenza inhibitor[181]	51 AA	Saccharomyces cerevisiae surface display	Outsourced	Ligand binding
Designed lysozyme inhibitor[135]	53 AA	S. cerevisiae surface display	Overlap extension	Ligand binding
Designed digoxigenin binder[171]	34/39 AA	S. cerevisiae surface display	Kunkel mutagenesis/doped oligonucleotide	Small molecule binding
IgG1 CH3 domain[173]	220 AA	S. cerevisiae surface display	Error-prone PCR	Ligand binding after thermal stress
Hsp90[68]	9 AA	S. cerevisiae complementation	Cassette ligation	Growth rate
Matα2 degron[83]	32 AA	S. cerevisiae fusion protein	Doped oligonucleotide	Growth rate
Ubiquitin[143]	75 AA	S. cerevisiae complementation	Cassette ligation	Growth rate
Pab1[106]	75 AA	S. cerevisiae complementation	Doped oligonucleotide	Growth rate
Neuraminidase[184]	470 AA	Mammalian cell	Error-prone PCR	Oseltamivir resistance
IgG CDRs[50]	59 AA	Mammalian cell display	Outsourced	Ligand binding
B-Raf[175]	77 AA	Mammalian cell	Cassette ligation	Vemurafenib resistance
Influenza nucleoprotein[10]	497 AA	Mammalian cell	Overlap extension	Viral infectivity
Influenza hemagglutinin[170]	564 AA	Mammalian cell	Overlap extension	Viral infectivity
IgG-binding domain of protein G (GB1)[118]	55 AA	mRNA display	Cassette ligation	In vitro binding
Gal4[84]	64 AA	S. cerevisiae complementation	Overlap extension with array-synthesized primers + megaprimer extension	Growth
p53[84]	393 AA	n/a	Overlap extension with array-synthesized primers + megaprimer extension	n/a
BLC11A enhancer[18]	4200 bp	Mammalian cell	CRISPR cleavage + NHEJ	enhancer activity (protein production)
BRCA1[48]	6 bp, 76 bp	Mammalian cell	CRISPR cleavage + HDR	Splicing detection (BRCA1 RNA-seq)

Table 2.1: Diversity of model systems, assays, and approaches to library synthesis in large-scale mutagenesis experiments. Modified from Fowler & Fields [52].

Applications of this formalism to measurements of the stability or function of protein variants give a sense of the heterogeneity of amino acid interactions, from the finding of additivity between mutations at contacting positions over many different mutant pairs in gene V protein[151,150,152] to energetic coupling between mutations at n,$n+4$ positions and additivity between residues with other periodicities in an α-helix of the ribonuclease barnase[75,156,155,9]. Meta-analyses of these types of experiments in various proteins further revealed that most mutations interact additively or nearly so, with non-additivity measured between residues in contact or in distant parts of the same protein[180,59,94]. Probing the distance dependence of coupling energy in more detail, Schreiber & Fersht[154] measured 33 double mutant cycles between 5 barnase and 7 barstar (a barnase protein inhibitor) residues and found that coupling energy decreased with increasing distance between the mutations. While this would be expected given the distance dependence of all interatomic forces, closer inspection of the data suggests a threshold (around 8.0 Å) beyond which non-additive interactions are rare, while both strong and weak coupling energies can be observed between residues (charged or uncharged) even at 7.0 Å apart. The authors thus conclude that "...structural details by themselves are not sufficient to evaluate the energetics of binding".

Cooperativity in the pairwise and higher-order interactions among amino acids in a protein underlies basic aspects of protein function, from ligand binding[67,136] and catalysis[63] to allostery[148]. But efforts to measure this cooperativity seem to outpace the development of explanatory models of the data. One reason for this is that despite these efforts, experimental quantitation of interaction energies between amino acids of sufficient breadth and quality to validate any model have been lacking (but see Olson et al.[118]), with the sheer number of measurements precluding a sufficient sampling of the combinatorial complexity of the interactions. Consider the hypothetical case of a 100 amino acid-long protein. The total number of position pairs to examine is given by the binomial coefficient $\binom{n}{k}$, or the number of ways of choosing k elements of a set of n elements; for the hypothetical protein, this amounts to 4950 possible pairs. Because the measurements may depend on the amino acid mutations being made, the least biased approach would be to interrogate all possible 19×19 double mutations between those two positions, yielding a total of 1,786,950 variants!

To overcome this combinatorial complexity, the co-evolutionary approach enables the inference of important interactions from the statistical analysis of extant sequences, themselves a record of the iterative processes of mutation and selection carried out by nature. As a model, it helped explain the existence of allosteric pathways in diverse proteins and the information it extracted sufficed to specify fold and function in at least two protein engineering endeavors[140]. Moreover, co-evolution recapitulates the findings of sparsity and heterogeneity in the encoding of function and structure into amino acids in proteins, and in the seminal paper by Lockless & Ranganathan[99], it significantly

correlated with experimentally determined thermodynamic coupling of binding free energy of 16 double mutants in a PDZ domain. This latter work specifically provides the foundation for a more comprehensive experimental test of the co-evolution model, described in the next section.

2.2 Results

2.2.1 Experimental approach

The technology enabling this endeavor is a high-throughput assay for binding previously implemented to assess the functional effects of all single mutants in a member of the PDZ family of protein interaction domains[105]. The anatomy of the assay consists of three components (Fig. 2.2*A*): a plasmid encoding the PDZ domain as a carboxy-terminal fusion to the λ-phage cI DNA binding domain, a second plasmid encoding the PDZ ligand fused to the C-terminus of the N-terminal domain of the α-subunit of RNA polymerase, and a third plasmid encoding a reporter, chloramphenicol acetyltransferase (*CAT*), which confers resistance to the antibiotic chloramphenicol, with an upstream binding site for the λ-cI domain. A binding event between the PDZ domain and its ligand recruits RNA polymerase, causing transcription of the *CAT* gene and growth of *E. coli* cells harboring this system. Importantly, assay conditions, specifically chloramphenicol and inducer concentrations, were optimized to produce a monotonic relationship between growth (more specifically, relative enrichment) and binding affinity measured *in vitro* for purified protein variants (Fig. 2.2*B*).

I leveraged this assay toward the study of comprehensive double mutations in 5 members of the PDZ family of protein interaction domains. These were selected for their ability to express in *E. coli*, their high-affinity binding to *bona fide* natural ligands, and their sequence diversity[163] (Table 2.2). I further focus the study on the α2-helix, a 9-amino acid structural element stabilized by intrahelical hydrogen bonds that directly abuts the peptide ligand and offers an experimentally tractable size (Fig. 2.3). Importantly, only four out of the nine amino acids in this helix exhibit any significant co-evolution, rendering this region of the protein as the minimal model for characterizing the pattern of thermodynamic cooperativity in proteins and for testing the model embodied by SCA.

2.2.2 The pairwise distributions and patterns of interactions

I generated libraries of all possible double mutants at all pairs of positions for each PDZ domain and tested the function of the libraries against the native ligands in the bacterial two-hybrid. In this assay, the functional effect of a mutation x at position i, ΔE_i^x, is computed as the log ratio of the frequency of each genotype after (*sel*) and before (*unsel*) chloramphenicol selection, normalized by

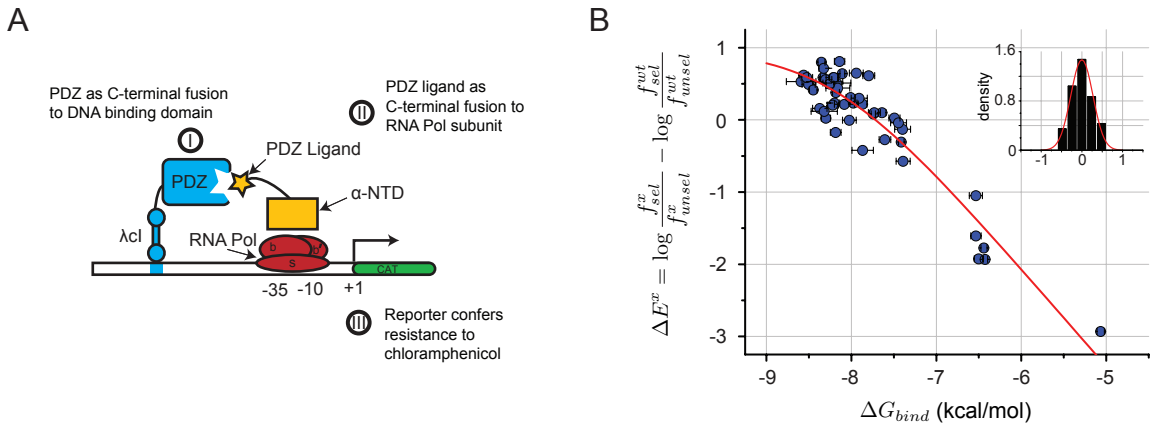

Figure 2.2: *A*, The three components of the bacterial two-hybrid, (1) a PDZ domain fused to the cI DNA binding domain of λ phage, (2) a PDZ ligand fused to the N-terminus of the α-subunit of *E. coli* RNA polymerase, and (3) a reporter gene coding for the enzyme chloramphenicol acetyltransferase, are all encoded in 3 separate plasmids and introduced into *E. coli*. A binding event between the PDZ domain and its ligand recruits RNA polymerase and results in the transcription of the reporter, leading to growth of cells in the presence of the antibiotic chloramphenicol. *B*, Binding measurements for PSD95^{pdz3} single mutant variants were obtained from fluorescence polarization experiments on individually purified proteins (error bars reflect the standard deviation from 3 replicate measurements). Optimized experimental conditions produced a monotonic relationship between between the relative enrichment values, ΔE_i^x, from the sequencing assay and free energy of binding, $\Delta G_{binding} = -RT \log K_d$, from the *in vitro* data. A simple model (*red*, detailed in section 2.4) was used to relate enrichment to fraction bound of the PDZ domain, and the normal, zero-centered distribution of the residuals (*inset*) of the model fit justifies its application.

PDZ domain	*Mutated AA*	*Ligand*	*AA sequence*	*Affinity*
PSD-95^{PDZ3}	HEQAAIALK	CRIPT	TKNYKQTSV	0.8 μM [105]
Shank3	HKQVVGLIR	Dlgap1/2/3	YIPEAQTRL	0.2 μM [163]
A1-syntrophin	HDEAVQALK	Scn5a	PDRDRESIV	1.3 μM [163]
ZO-1	HAFAVQQLR	Claudin8	SIYSKSQYV	4.6 μM [188]
PSD-95^{PDZ2}	HEDAVAALK	NMDAR2A	KMPSIESDV	3.6 μM [163]

Table 2.2: PDZ domains were selected based on knowledge of their expression in *E. coli*, availability of a high-affinity ligand, and sequence diversity in the α-helical region.

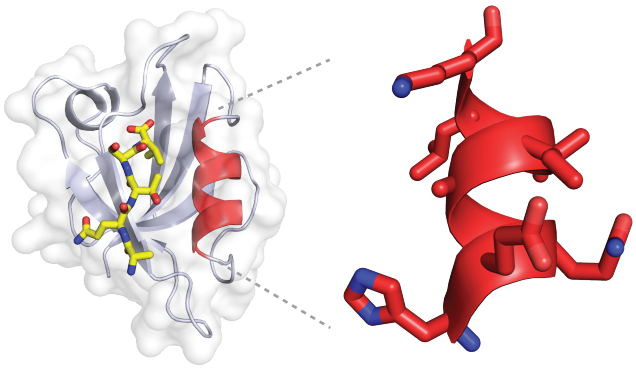

Figure 2.3: PDZ domains, ~100 amino acids in length, adopt stereotyped structures of mixed α-helices and β-strands and mediate protein-protein interactions by binding to carboxy-termini of other proteins. The binding cleft is formed by the β2 strand and the α2 helix, the latter the target of the mutagenesis experiments described in this work, and highlighted on a representative structure (PDB: 1BE9).

the same ratio of the wildtype (WT):

$$\Delta E_i^x = \log \frac{f_{i,sel}^x}{f_{i,unsel}^x} - \log \frac{f_{sel}^{wt}}{f_{unsel}^{wt}} \tag{2.1}$$

These quantities are converted to free energies (ΔG_i^x, in kcal mol^{-1}) using a simple model that relates enrichment, ΔE_i^x, to fraction bound of the PDZ-λcI fusion (see section 2.4). As discussed in the previous section, the thermodynamic mutant cycle formalism provides the framework to characterize the interaction between two mutations x and y at positions i and j, where the functional contributions of each residue are revealed via mutation and are related according to the equality

$$\Delta G_{i,j}^{x,y} = \Delta G_i^x + \Delta G_j^y - \Delta\Delta G_{i,j}^{x,y} \tag{2.2}$$

The quantity of interest in this analysis, $\Delta\Delta G_{i,j}^{x,y}$, is a measurement of the non-additivity between the two mutations.

To determine the uncertainty in the coupling values, I carried out 4 independent replicates of the experiment for the PSD95^{pdz3} library, which demonstrated high reproducibility of the assay. I hypothesized that most of the variance in the measurements could be explained by the sequencing counts, especially for those mutants which are rarely observed in the selected library. To test this, I modeled the experiment as a Poisson process, where the mean and variance of the number of times a mutation is observed is parametrized as the sequencing counts for that mutant. Under this model,

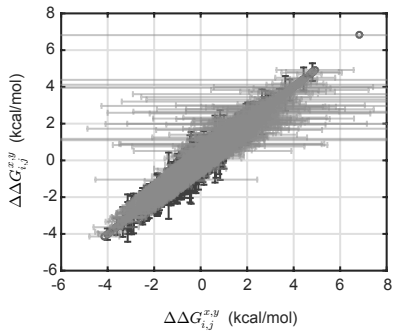

Figure 2.4: Coupling values, $\Delta\Delta G_{i,j}^{x,y}$, for over 10,000 double mutants in PSD95^{pdz3} are plotted against themselves with horizontal error bars depicting the standard deviations of the experimental values for each double mutant derived from 4 independent experiments and vertical error bars depicting the standard deviation computed from a statistical error model (see text; both depict uncertainty in the coupling measurement after error propagation). The average standard deviation of experimental epistasis across all double mutants is $0.33\,\mathrm{kcal\,mol^{-1}}$, whereas the average statistical error is $0.96\,\mathrm{kcal/mol}$ and is greatest for double mutants with low counts in either the selected or unselected populations.

the error in the enrichment measurement is calculated by propagating the errors through the calculation of $\Delta G_{i,j}^{x,y}$ using a first-order Taylor series expansion of both the ΔE_i^x and ΔG_i^x functions. When compared to the variance computed from the experimental replicates (Fig. 2.4), the Poisson error model overestimates the measurement uncertainty, especially for mutations with low enrichment values (indicating low sequencing counts). In fact, the average standard deviation for the experimental replicates across all mutant pairs was $0.33\,\mathrm{kcal\,mol^{-1}}$, lower than the $0.50\,\mathrm{kcal\,mol^{-1}}$ cutoff commonly used to distinguish significant non-additivity from biochemical measurements[154]. Thus, the increased throughput of the sequencing-based assay does not compromise the level of precision required to measure functional couplings. Moreover, the more strict Poisson model was used to filter out lower-confidence measurements based on a cdf cutoff of the error distribution. After this statistical culling, I obtained nearly 10,000 double mutants for each PDZ domain (8500 to 10,304), representing averages between 236 to 286 out of a possible 361 double mutants per position pair.

I measured the coupling between all amino acids x and y at every pair of positions i and j, $\Delta\Delta G_{i,j}^{x,y}$, to investigate how these parameters varied for different pairs and their dependence on the specific mutations used to interrogate them. A comparison of the distributions of couplings between four positions of PSD95^{pdz3} is revelatory (Fig. 2.5). Consider the position pair 379 and 380 (Fig. 2.5*A*): most double mutants appear to interact additively or nearly so, with the distribution of all values centered near $-0.07\,\mathrm{kcal\,mol^{-1}}$. In other words, mutations at either residue have little influence on

each other despite being linked by a polypeptide bond. But this non-additivity is not confined to adjacent residues in primary sequence, as couplings between mutations at positions 375 and 378 also tend toward 0 kcal mol^{-1} (Fig. 2.5*B*). Both position pairs between 376 and 379, as well as 372 and 376, on the other hand, exhibit distributions of coupling energies that are centered progressively farther from 0 kcal mol^{-1} (Figs. 2.5*C* and *D*). The main observation that these examples convey is that all of these distributions can be approximated by unimodal distributions with well-defined means, an observation applicable for all pairs of positions in PSD95^{pdz3} (Fig. 2.6) and every other PDZ domain. This phenomenon is not unique to double mutants in the PDZ domain, as pairs of positions in the GB1 domain, for which a more global and combinatorially complete dataset is available[118], also exhibit unimodal distributions of coupling values (Fig. 2.7).

Assumptions that classify mutations as either deleterious or neutral, as informed either by the distribution of single mutation effects[105,164] or the disruption or preservation of specific interatomic interactions, would predict multimodal or broader distributions of couplings. But the observed distributions are consistent with a model wherein each double mutant represents a measurement of the intrinsic coupling between two positions in addition to a δ effect related to the specific amino acid mutations being made. Thus, I propose that averaging over all possible 361 double mutant cycles provides means of dimension reduction of the data as the best estimate of positional coupling. Rather than conveying the non-additivity between mutations at specific positions, the mean represents the native coupling of the residues in the wildtype protein.

More deeply, the homolog-averaged values, $< \Delta\Delta G_{i,j}^{x,y} >_{homolog}$, are similarly distributed (Fig. 2.8), and these quantities now represent the pairwise interactions that are conserved across the 5 homologs. Comparison of the means of position pair couplings of each individual homolog to that of the average reveals examples of homolog idiosyncracies – for example, position pairs 375 and 376, as well as 372 and 373, only show significant mean non-additivity in a few homologs, notably PDZ3 –, as well as invariances, the most obvious being couplings between position pairs involving residues 372, 375, and 376 (Fig. 2.9).

Upon closer inspection, however, it is clear that the Gaussian fit for the distribution of couplings for some position pairs is inadequate. The distribution between the pair of positions 375 and 376, for example, is better approximated by a double-Gaussian fit (Fig. 2.10). Each mode of the distribution could emerge from contributions of couplings from distinct homologs or groups of position pairs within homologs each with unique means. Comparison of the coupling values and distribution shapes and means for this position pair across the PDZ homologs (Fig. 2.10) demonstrates that the latter phenomenon is the source of the bimodality. Interestingly, the two modes actually emanate from position pairs involving amino acids that are either tolerated or not in an alignment of

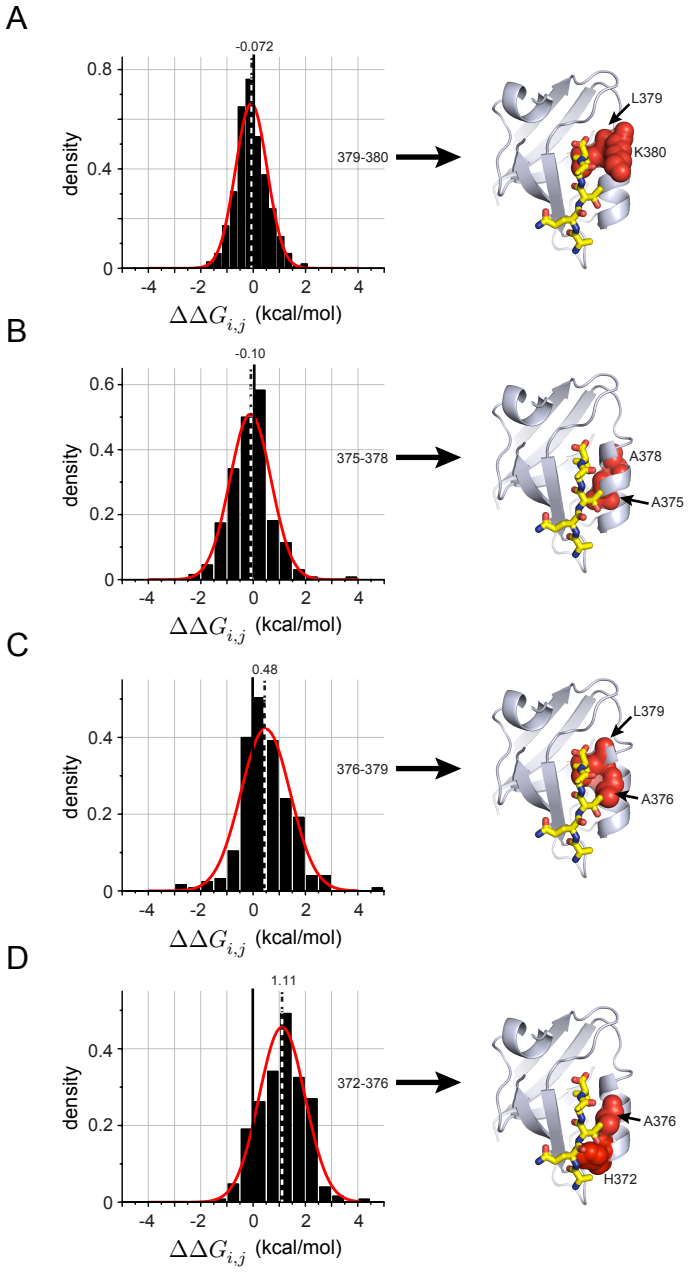

Figure 2.5: *A* and *B*, Distributions of $\Delta\Delta G^{x,y}$ for positions 378-379 and 375-378, respectively, in PSD95^{pdz3}. Though one pair involves residues adjacent in primary structure, the distributions of coupling values for both pairs are similar and centered near $0\,\mathrm{kcal\ mol^{-1}}$. Coupling energies involving positions pairs 376-379 (*C*) and 372-376 (*D*), meanwhile, tend toward higher $\Delta\Delta G$ values. All distributions are well-approximated by unimodal distributions whose means represent an estimate of the intrinsic coupling between the two positions.

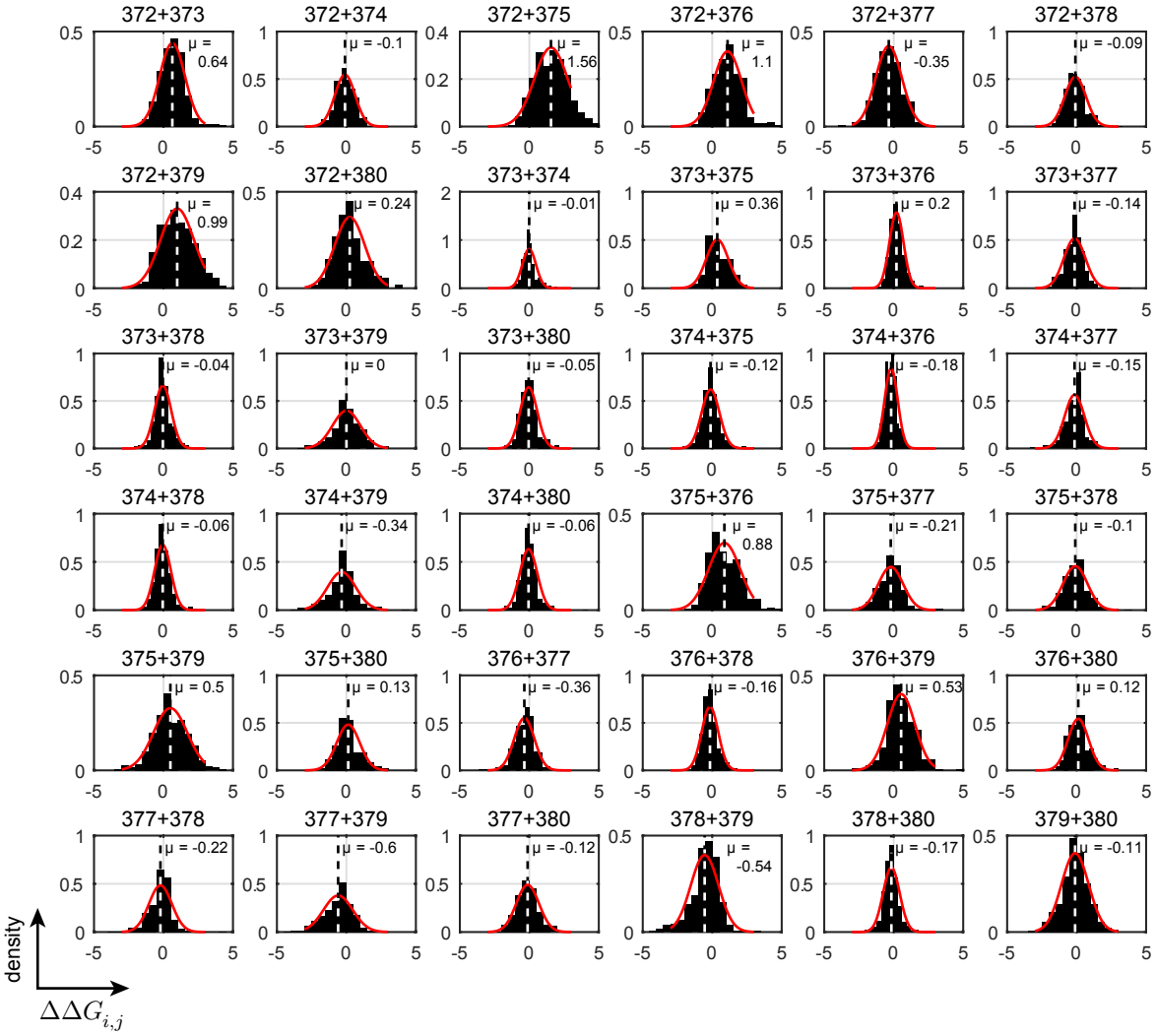

Figure 2.6: Distributions of $\Delta\Delta G_{i,j}^{x,y}$ for all 36 pairs of positions in the α2-helix of PSD95^{pdz3}, overlaid with Gaussian approximations for each. The unimodality is consistent with a model where all double mutants provide a measurement of the intrinsic coupling between two positions in addition to some effect emanating from the specific mutations being made. As a result, all such measurements tend to a mean value that best represents the coupling between the two positions.

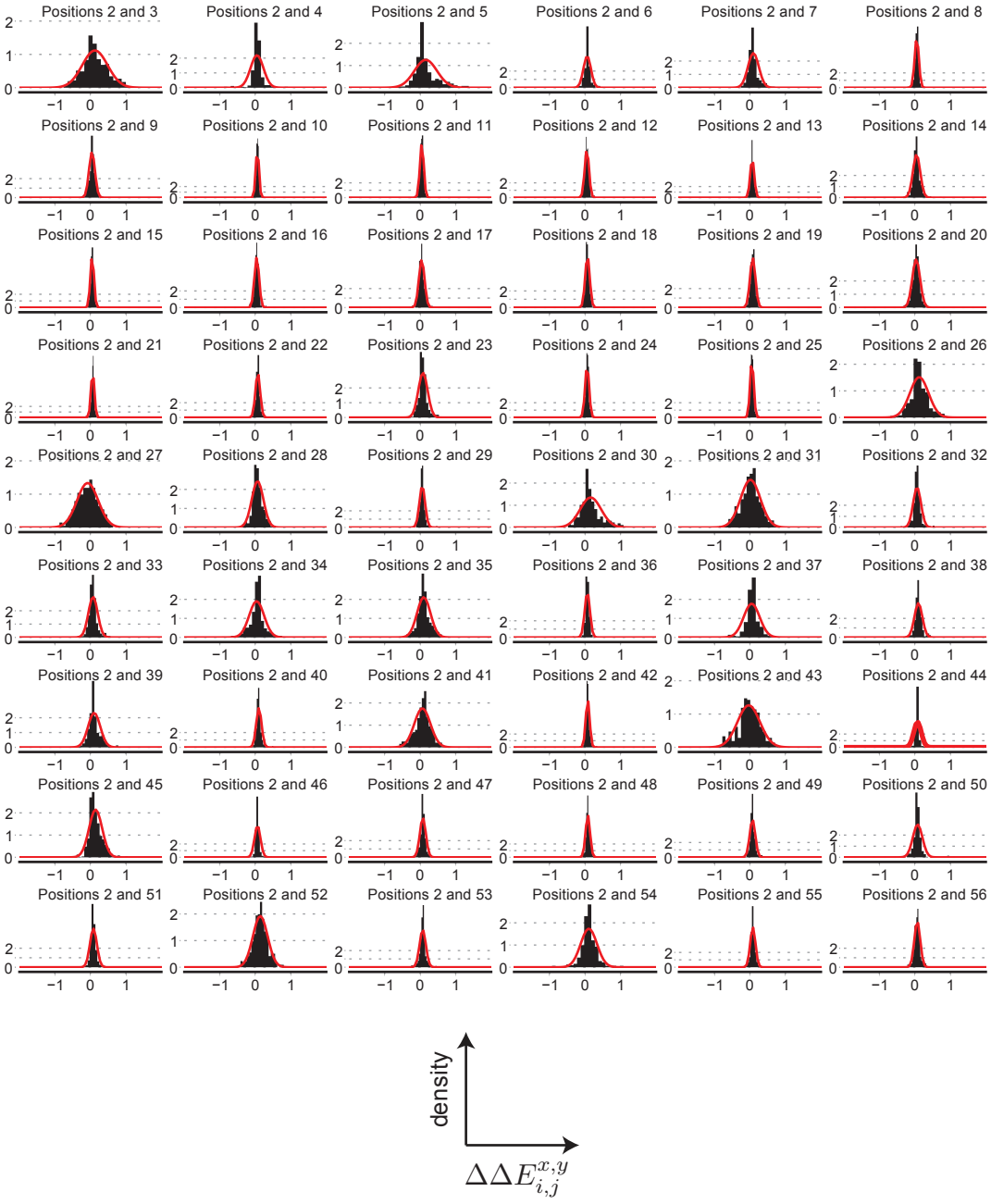

Figure 2.7: Distributions of $\Delta\Delta E_{i,j}^{x,y}$ for all 54 pairs of positions involving position 2 of the GB1 domain [118]. These distributions, sampling nearly all possible 361 double mutants per pair of positions, are also unimodal, and thus corroborate the observations made in PSD95^{pdz3}. Note that the non-additivity is computed based on enrichment rather than free energy, although both are linearly related (see Fig. S3M in Olson et al. [118]).

representative PDZ domains.

Mathematically, the averaging has the consequence of isolating the two-way couplings from higher-order interactions, and therefore better reflects the nonadditivity at the pairwise level[133]. It is precisely this data and parametrization that permit, for the first time, a comparison to the co-evolution analysis, itself an analysis of statistical interactions between amino acid positions conserved in a protein family.

2.2.3 Evaluation of statistical models of amino acid interactions

Toward this, I compared the association between the experimentally determined pattern of interactions versus those inferred from evolutionary statistics. The statistical coupling analysis (SCA) uses conservation-weighted covariance as the measure of co-evolution, and we find that it significantly overlaps with the experimentally determined set of interactions (Fig. 2.11*A*, $p = 9.29 \times 10^{-5}$ by Fisher's exact-test, conclusions robust to cutoffs used for the distributions of statistical and experimental coupling). This correlation remains significant even when the data are constrained to those amino acid mutations that appear in the alignment. The conservation weighting employed in SCA has been argued to function as a Bayesian prior, with the consequence of emphasizing correlations among more conserved positions and potentially improving the prediction of interactions involved in more invariant properties of the protein family[30]. We explored the extent to which conservation weighting influences the performance of SCA by comparing the experimental pattern of interactions to mutual information (equivalent to the unweighted covariance matrix[30]). Statistical analysis of the distributions show that they are independent (Fig. 2.11*C*, $p = 1.00$, Fisher's exact test). Finally, the direct coupling analysis (DCA)[102,81] is another algorithm for computing co-evolution used to predict three-dimensional contacts in proteins. An outstanding question is whether it can also identify functional interactions, as suggested in recent reports[70,46]. Comparing the most highly co-evolving positions by DCA with the experimental data shows that they are not correlated (Fig. 2.11*B*, $p = 7.33 \times 10^{-2}$, Fisher's exact-test). Taken together, these comparisons show a strong correlation between conservation-weighted co-evolution and the experimental pattern of energetic coupling.

2.2.4 Structural basis of the observed pattern of interactions

While secondary structure does not explain the overall pattern of conserved interactions, a periodicity consistent with the n,$n+3$ and n,$n+4$ pattern of an α-helix is noted in the pattern of significant interactions among the first, fourth, fifth, and eighth positions in the helix. This periodicity also recapitulates the distribution of residue burial and hydrophobicity, with these same positions not only

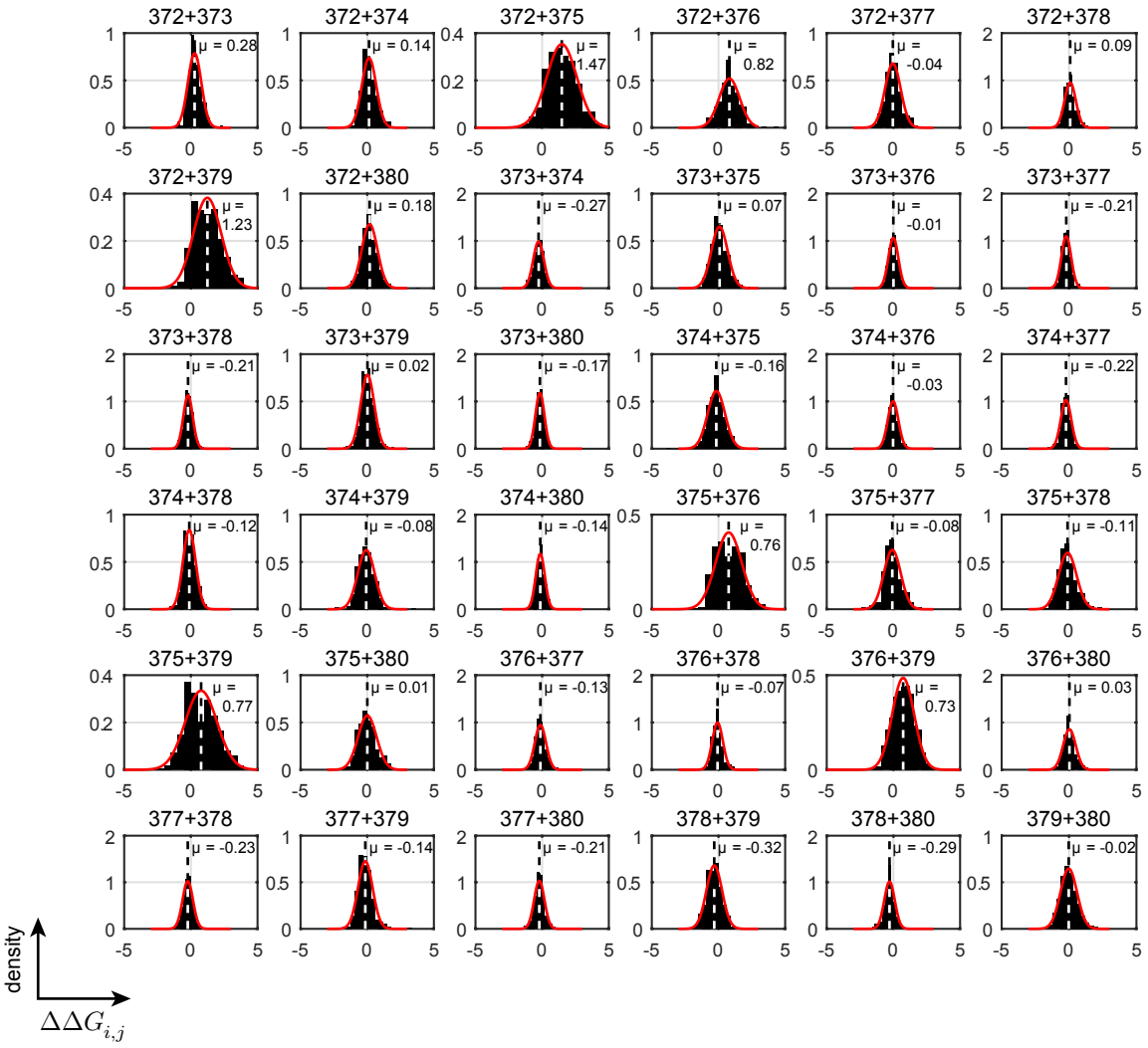

Figure 2.8: Distributions of $< \Delta\Delta G_{i,j}^{x,y} >_{homolog}$ are also unimodal and therefore amenable to the simple dimension-reduction approach proposed in the text.

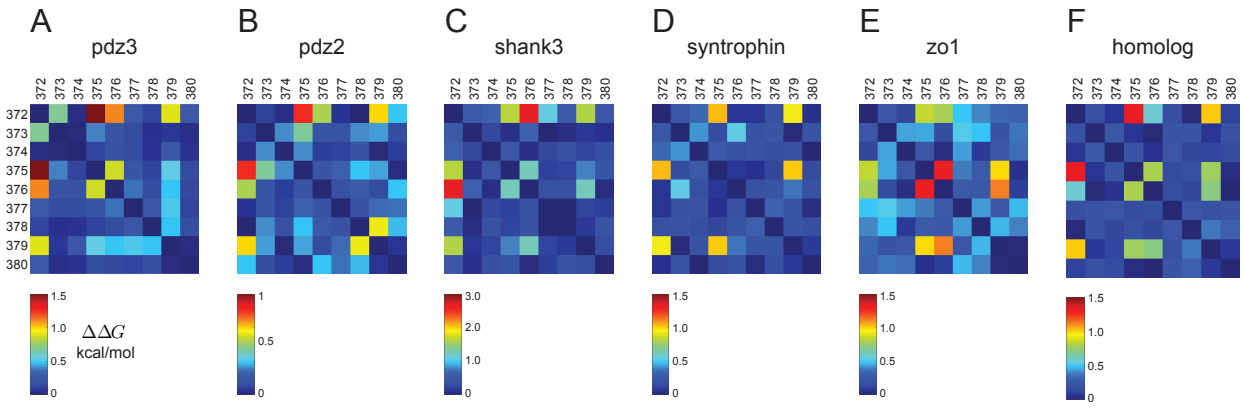

Figure 2.9: Coupling energies for pairs of positions for each homolog (*A*-*E*) were averaged over all amino acid mutations to arrive at single values conveying the native non-additivity between them in the native state. The pixels in each matrix represent these mean quantities (rather, the absolute value of these) in units of kcal mol^{-1}. *F*, the interaction energies between pairs of mutations were averaged across the five homologs, and the means of these distributions represent the homolog-averaged positional couplings.

possessing the lowest values of relative solvent accessibility (Table 2.3), but also being located on the side of the helix facing the ligand. As discussed in a previous section, prior work has demonstrated the contribution of non-additive intrahelical interactions between n,$n+4$ pairs toward helical stability, as well as additivity between n,$n \neq 4$ pairs[166,104,166,129,156,75,9]. This periodicity has also been observed in studies of mutational tolerance and sensitivity[16,8], patterns of polarity[130] and hydrophobicity[40], and co-evolution[134] in many model proteins. Thus, the data obtained here, emerging from an analysis of mutational effects on binding affinity rather than stability itself, reflect fundamental biophysical constraints related to the structure of the alpha helix.

The concept of spatial proximity pervades efforts that strive to describe the structure-function relationship in proteins, and it stems from the distance dependence of interatomic forces. However, no significant correlation between mean coupling, $< \Delta\Delta G_{i,j}^{x,y} >_{x,y}$, and inter-residue distance (minimal all-atom distance between two positions in PSD95$^{\mathrm{pdz3}}$, PDB: 1BE9) exists (Fig. 2.12), in contradiction to earlier work[154] described in the previous section. The discrepancy may lie in the smaller number of double mutants studied therein or the distance range that was covered, though the distance of 10 Å sampled within the helix should be sufficient to examine the gamut of intermolecular forces (from hydrogen bonds and salt bridges to weak and strong van der Waals and electrostatic interactions). The relationship between coupling and distance for the GB1 domain mutagenesis data[118] provides a model that helps to reconcile these findings (Fig. 2.13). Residue pairs in

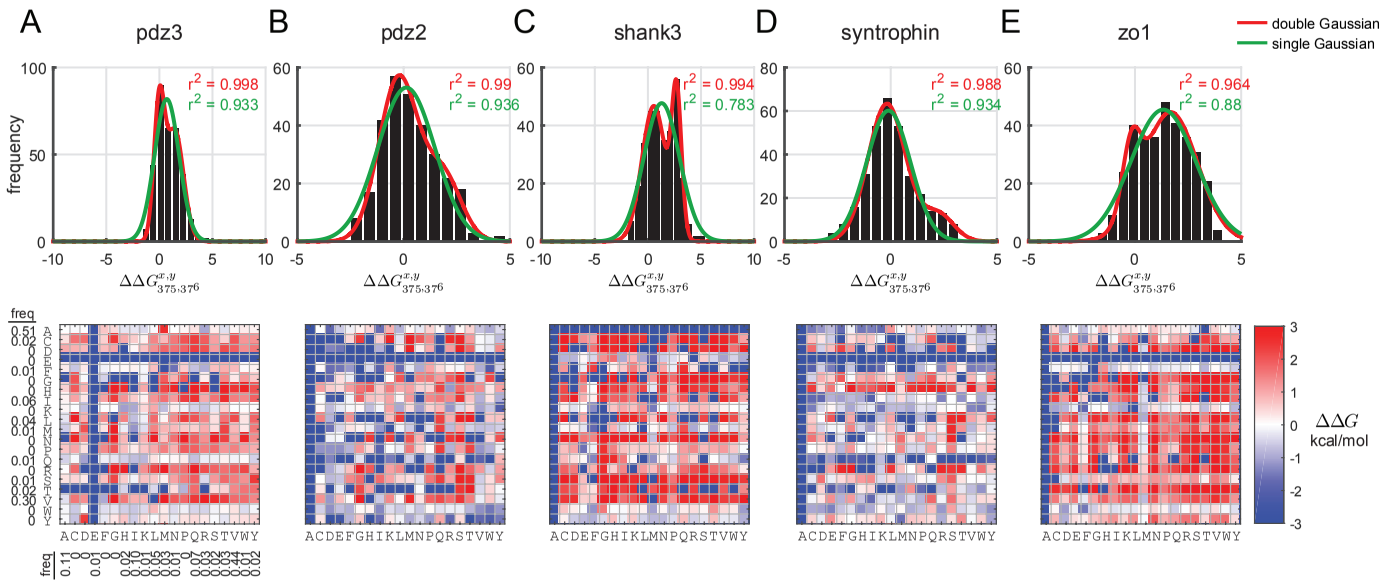

Figure 2.10: Distributions (*top row*) of the coupling energies between all mutations at positions 375 and 376, overlaid with single- and double- Gaussian fits for each PDZ homolog. In all cases, the double-Gaussian fits better approximate the data (r^2 values are adjusted for number of parameters). Matrices (*bottom row*) conveying all the coupling energies for every double mutation at those positions in units of $\mathrm{kcal\ mol^{-1}}$. Adjacent to each amino acid are their respective frequencies in an alignment of PDZ sequences, showing that the bimodality actually emanates from distinct sets of residues that are hydrophobic and statistically conserved.

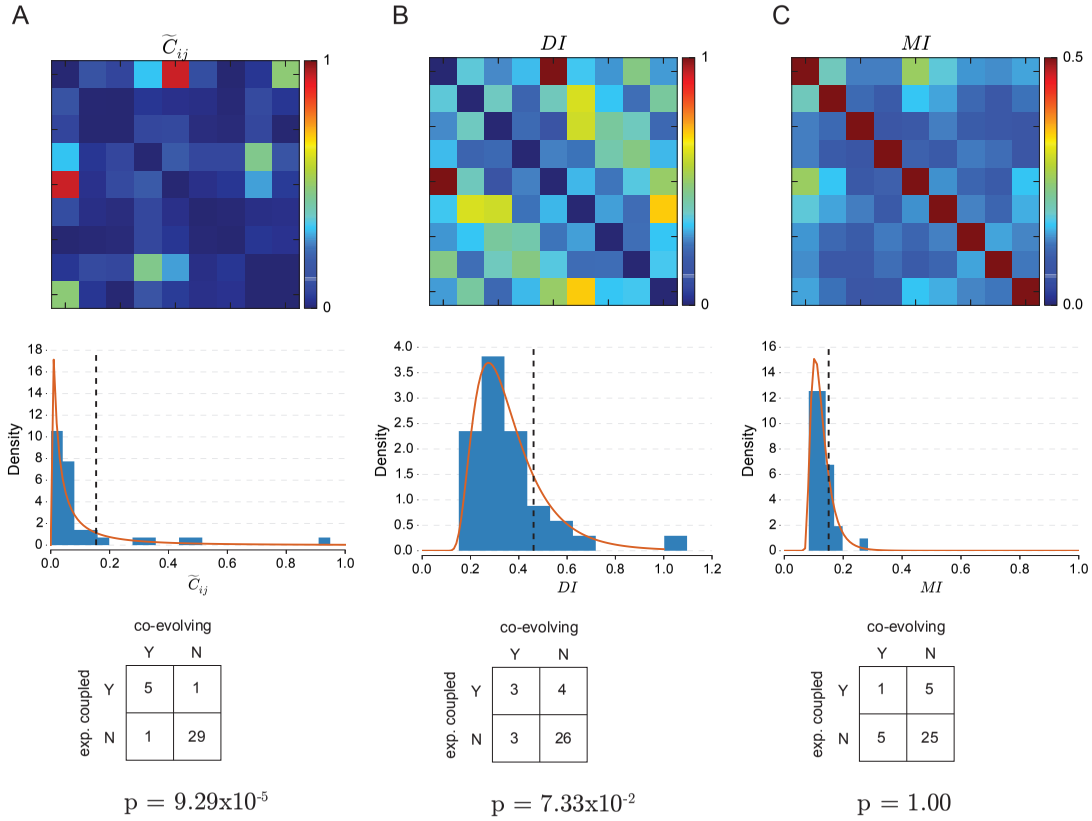

Figure 2.11: Matrices (*top row*) showing coupling values between pairs of positions and corresponding distributions (*middle row*) of couplings inferred from SCA (*A*), DCA (*B*), and MI (*C*), respectively (each overlaid with a fit to a lognormal distribution to distinguish significant interactions from all others). Statistical comparisons between the co-evolutionary methods and the experimentally determined couplings were carried out using contigency tables (*bottom row*).

Residue	*Ratio (SA to "random coil")*	*I/O*
HIS372	12.5	I
GLU373	84.0	O
GLN374	69.8	O
ALA375	0.0	I
ALA376	6.0	I
ILE377	49.1	n/a
ALA378	17.8	I
LEU379	0.0	I
LYS380	54.1	O

Table 2.3: Solvent accessible surface area calculations [53] for the α2-helix of PSD95^{pdz3} (PDB: 1BE9). Residues are considered buried (*I*) if their side-chain surface area to "random coil" values are less than 20% or exposed (*O*) if the ratios are greater than 50%.

this protein (PDB: 1PGB) span a range of 30 Å, and a comparison between distance and coupling shows a threshold (at around 10 Å) beyond which positions exhibiting high coupling strength are less probable. However, even the short distance sampled by the α-helix contains both weak interactions between adjacent (less than 5 Å) residue pairs and strong interactions between residues around 10 Å apart. Consequently, the interpretation advocated by the GB1 analysis is that a distance dependence of coupling strength within proteins exists (representing perhaps the likelihood of energetic propagation through the protein), but that the local physical environment around each residue is heterogeneous (with some physical interactions contributing more to function or stability than others).

2.2.5 Consequences of the unimodality of non-additivity

The finding that the distributions of non-additivity of all double mutants between pairs of positions can be approximated by unimodal distributions suggests that the data can be subsampled and still yield an accurate estimate of the distribution mean, or the intrinsic coupling between positions. Practically, the subsampling technique must be unbiased with respect to the mutations it probes and also be experimentally tractable. I examined whether error-prone polymerase chain reaction (PCR), a frequently-used technique to generate mutant libraries through the incorporation of incorrect nucleotides during DNA amplification [124], would be suitable for this pursuit. Though most amino acid mutations sampled represent those that are accessible through single nucleotide changes, the method is no more complicated than setting up a single PCR reaction. In addition, the error

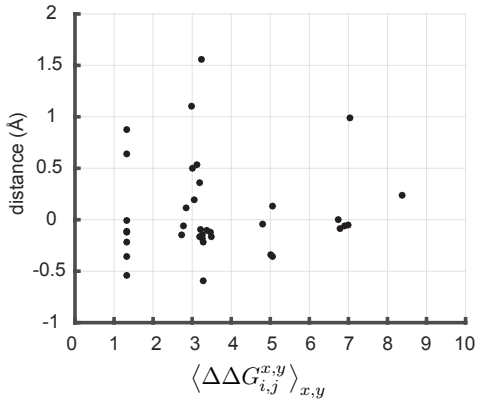

Figure 2.12: No discernible correlation exists between mutant-averaged coupling, $< \Delta\Delta G_{i,j}^{x,y} >_{x,y}$, and inter-residue distance, calculated as the minimum all-atom distance (van der Waals radii + 20%) between two amino acids.

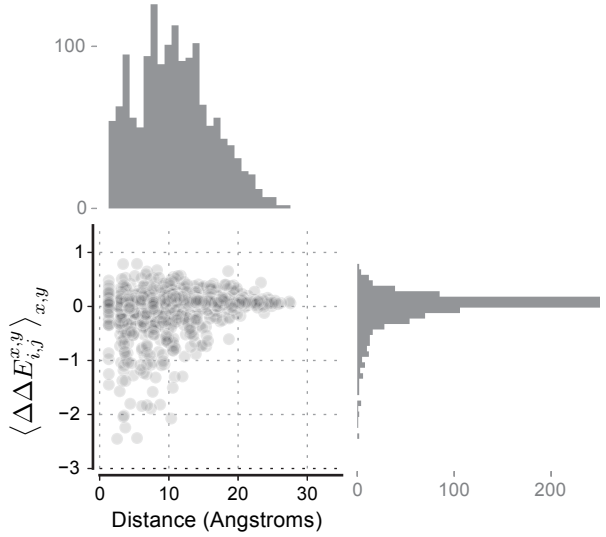

Figure 2.13: A scatter plot of mutant-averaged coupling, $< \Delta\Delta E_{i,j}^{x,y} >_{x,y}$, and inter-residue distance for the 1485 position pairs in the GB1 domain is juxtaposed with histograms showing the density of the data points along each axis. The data show that beyond a distance of around 10 Å, strong couplings are rare, but that within this range (equivalent to the range covered by the α2-helix in the PDZ domain), no clear correlation is seen between these variables. This is consistent with the analysis in Fig. 2.12.

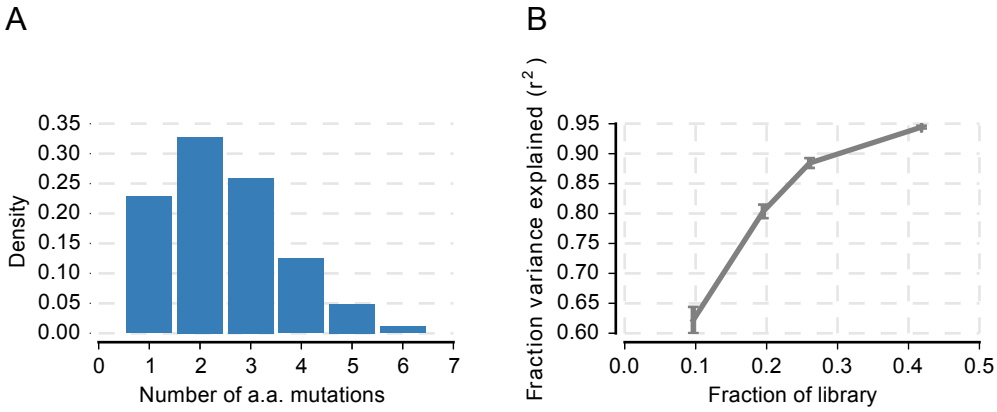

Figure 2.14: *A*, Error-prone PCR was simulated on the α2-helical region of PDZ3 with an error rate of 12%, which was shown to maximize the number of double mutations in the resulting libraries. *B*, Ten runs of the simulation at 4 population bottlenecks (all amenable to next generation sequencing-based experiments) were conducted and the sequences sampled in each library were used to compute estimates for $< \Delta\Delta G_{i,j}^{x,y} >_{x,y}$, which were subsequently compared to the estimates obtained from the more complete library assayed experimentally. The mean and standard error of the correlation coefficients (r^2) from 10 estimates are plotted at each population bottleneck, clearly showing that a small subsampling is sufficient to recapitulate a significant proportion of the original data.

rate of incorrect nucleotide incorporation can be modulated by varying buffer components, types of polymerases, template amount, and number of amplification cycles, enabling fine-tuning of the distribution of mutations in the library.

I simulated error-prone PCR *in silico* with an error rate that maximized the number of amino acid double mutants (Fig. 2.14*A*) and bottlenecked the resulting libraries to examine the degree to which the more complete dataset can be subsampled and still reproduce the positional coupling measurements. Remarkably, a 10% subsampling of the dataset could explain over 60% of the original data, while a 40% sampling reproduces 95% (Fig. 2.14*B*).

2.3 Conclusion

Approaches to protein design that seek to optimize packing interactions often succeed in producing functional and folded polypeptides[88]. This experimental data, on the other hand, argue that most residues in a protein contribute independently to fold and function, with only a small number of the total possible interactions identified as functionally significant; these observations are consistent with the model of proteins suggested by co-evolutionary analysis. I thus propose that insight into the sequence-function relationship can be acquired by reparametrizing the problem as both the

origin of and the physics underlying the architecture of co-evolution. The answer should describe the process consistent with the evolutionary forces of mutation, drift, and selection that generates a physical pattern of sparse and heterogeneous interactions. This work provides the most compelling motivation for this pursuit, and methods like the one described here and in development (Hekstra et al., submitted) promise to bring us closer to an answer.

Moreover, a major hurdle confronting efforts to elucidate the global pattern of amino acid interactions in proteins is the requirement for accurate and precise measurements of protein fold or function of thousands of variants. The unimodal distributions of non-additivity between mutations at position pairs, first described in this work, suggest that only a fraction of all possible 361 double mutants needs to be measured to estimate positional coupling using the approach of the thermodynamic mutant cycle. Error-prone PCR represents a simple approach to generate libraries that achieve an adequate compromise between experimental efficiency and accuracy, and offers a prescription for conducting future experiments to reveal the pattern and pervasiveness of interactions in proteins.

2.4 Methods

2.4.1 Bacterial two-hybrid

The bacterial two-hybrid assay (Fig. 2.2*A*) used in this work is based on the system of McLaughlin et al.[105] and modified by Frank Poelwijk. The PDZ-λcI fusion, whose expression in under the control of the *lac* promoter, is encoded by the PZS22 plasmid with a low-copy SC101 origin of replication and a trimethoprim (Tm) resistance selection cassette. The PDZ ligand-RNA polymerase α-subunit fusion, on the other hand, is under a tetracycline-inducible promoter and is encoded by the PZA31 plasmid (with a low-copy p15A origin of replication and kanamycin (Km) resistance cassette). In this implementation of the assay, the eGFP reporter has been replaced by the antibiotic resistance gene *CAT*, which codes for the enzyme chloramphenicol acetyltransferase. This reporter is encoded by the pZE1RM plasmid, with the medium-copy number ColE1 origin of replication and an ampicillin (Am) resistance cassette. Upstream of the reporter, the second binding site for the λcI domain has been removed to improve the stability of the system, which was previously compromised by recombination between the two sites and transcription of the reporter in the absence of PDZ-PDZ ligand interactions.

Optimal chloramphenicol concentrations for selection were deduced from replicate experiments assaying the binding of a library of single amino acid variants of $PSD95^{pdz3}$ toward the CRIPT ligand. A range of chloramphenicol concentrations was scanned, and the lowest concentration which yielded the maximum dynamic range and correlation between binding free energy, $\Delta G_{binding}$, and

enrichment (see section 2.2.1) was selected (Fig. 2.15). The working concentration in all experiments was $150\,\mu g\,mL^{-1}$.

The assay itself begins with the transformation of PDZ variants into electrocompetent pZE1RM$^+$pZA31$^+$ MC4100Z1 cells that harbor chromosomal copies of the *lac* repressor lacIq and the *tet* repressor TetR[168]. After recovery in 1 mL SOC, serial dilutions from $5\,\mu L$ of the recovery culture are plated onto Km$^+$Am$^+$Tm$^+$ plates to determine transformation efficiency; the remainder of the culture is used to inoculate 50 mL of LB media supplemented with $50\,\mu g\,mL^{-1}$ of Am, $40\,\mu g\,mL^{-1}$ of Km, and $20\,\mu g\,mL^{-1}$ of Tm in a flask incubated overnight at 37°. After ~12 hours, a $1:1000$ dilution of the culture is done into fresh LB media with the three antibiotics and allowed to grow at 37 °C to bring the cells into exponential growth ($OD_{600} = 0.1$), at which point another $1:100$ dilution is made into LB supplemented with the three antibiotics in addition to $50\,ng\,mL^{-1}$ of doxycycline hydrochloride (dox). Cells are then incubated at 25 °C for 2 log-orders of growth (~6.7 doublings), at which point protein expression is expected to have reached steady-state[131]. Growth at this reduced temperature also helps to constrain the effects of mutations to binding affinity and minimize minor destabilizing effects. After induction, the cells undergo another $1:100$ dilution into fresh LB media supplemented with Am, Km, Tm, and dox, in addition to chloramphenicol at a final concentration of $150\,\mu g\,mL^{-1}$. Cells are grown at 25 °C and are harvested when the culture reaches an optical density of $0.1 OD_{600}$ units, after which plasmids are purified. Two PCR reactions are performed to amplify the region of interest to be sequenced and to append Truseq barcodes and Illumina adapters. Samples are multiplexed and sequenced on either the Illumina Miseq or Hiseq2500 instruments (University of Texas Southwestern Medical Center Genomics and Microarray Core). Allele counts are extracted from sequencing files using basic UNIX bash text parsing commands for downstream analysis in Matlab or Python.

2.4.2 Relating sequencing counts to free energy measurements of binding affinity

We begin by defining a parameter, ΔE^x, that conveys the enrichment of an allele x in the selected *versus* unselected populations relative to the wildtype,

$$\Delta E^x = \log \frac{f^x_{sel}}{f^x_{unsel}} - \log \frac{f^{wt}_{sel}}{f^{wt}_{unsel}} \tag{2.3}$$

where f^x_{sel} and f^x_{sel} represent the frequencies of allele x in the selected and unselected populations, respectively.

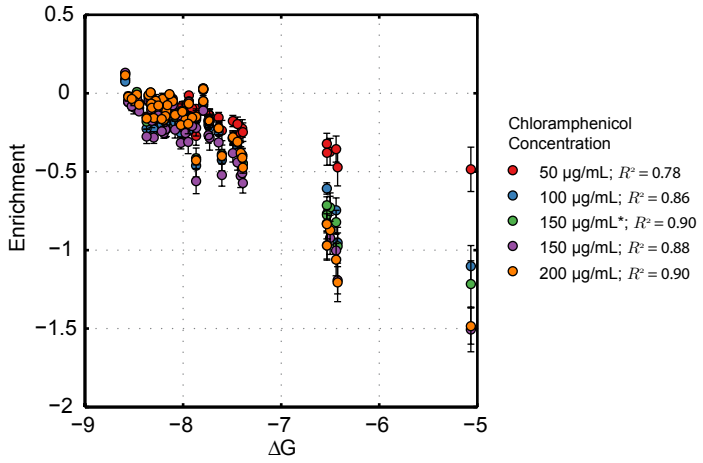

Figure 2.15: A library of mutants with characterized binding affinities were assayed in triplicate in the two hybrid under different concentrations of the antibiotic chloramphenicol, the selection agent. Plotted are the mean and standard deviation of the enrichment values of each mutant against the binding free energy, $\Delta G = RT \log K_d$. Most conditions produced high correlations between the two parameters, but the concentration of $150\,\mu\mathrm{g\,mL}^{-1}$ was the lowest concentration that also yielded the highest dynamic range. The asterisk denotes a set of replicates carried out with $150\,\mu\mathrm{g\,mL}^{-1}$ of chloramphenicol but in the absence of inducer during the selection, which led to a shorter period of selection (as reporter expression is not maintained) and a decreased dynamic range.

Allele enrichment reflects cell growth in the presence of an antibiotic, itself the outcome of a binding event between the PDZ domain and its ligand that results in recruitment of the RNA polymerase holoenzyme and transcription of an antibiotic resistance gene. We devised a model that relates enrichment to the fraction of PDZ domains bound to the ligand. Formally,

$$K_d = \frac{[L]_{free}[R]_{free}}{[RL]} \tag{2.4}$$

where K_d represents the dissociation constant of the PDZ domain-ligand interaction, and $[L]$, $[R]$, and $[RL]$ denote the cellular concentrations of ligand, PDZ domain, and PDZ domain-ligand complex, respectively.

Rearrangement of equation (2.4) leads to an expression for the fraction of PDZ domains bound

$$f_B = \frac{[RL]}{[R]_{total}} = \frac{[L]_{free}}{K_d + [L]_{free}} \tag{2.5}$$

Finally, the relationship between f_B and ΔE^x is expressed as

$$\Delta E^x \propto \log f_B \tag{2.6}$$

We fit a set of 48 measurements of ΔE^x and $\Delta G^x - \Delta G^{wt}_{PDZ3}$ to the model*

$$\Delta E^x = a \cdot \log f_B + C = a \cdot \log \frac{[L]_{free}}{[L]_{free} + e^{\frac{\Delta G^x - \Delta G^{WT}_{PDZ3}}{R \cdot T}}} + C \tag{2.7}$$

to derive values for a, C, and $[L]_{free}$, which would then allow for estimates of ΔG values for each mutant.

To use this model to calculate ΔG for mutants in other PDZ homologs, two corrections are required. First, the estimates must be derived relative to the wildtype state of the other homologs. Secondly, because the ΔE^x values are normalized relative to just mutants in a specific homolog as each library was assayed independently, they must be further scaled to the ΔE^x values of PDZ3. The scaling factor is derived from the expected f_B, f'_B, for a PDZ homolog given its affinity relative to PDZ3 (both of which are known values).

$$f'_B = \frac{[L]_{free}}{[L]_{free} + \frac{K_d^{homolog}}{K_d^{PDZ3}}} \tag{2.8}$$

The correction factor, δ, is given by the expression

$$\delta = a \cdot \log \frac{f'_B}{f_B} + C \tag{2.9}$$

where f_B is computed from the its relationship to ΔE^x using the sequencing counts in the homolog library.

Thus, to arrive at estimates of ΔG for each homolog mutant, first we calculate the ΔG of the homolog wildtype,

$$\Delta G^{WT}_{homolog} = R \cdot T \Big[\log \left([L] - [L] \cdot e^{\frac{\Delta E^{WT}_{homolog} + \delta - C}{a}} \right) - \frac{\Delta E^{WT} - C}{a} \Big] + \Delta G^{WT}_{PDZ3} \tag{2.10}$$

*aided by the equality $\Delta G = R \cdot T \log K_d$, where $R \cdot T$ is $0.5922\,\text{kcal mol}^{-1}$

and use this measurement to appropriately scale the mutant ΔG values

$$\Delta G^{x}_{homolog} = R{\cdot}T\Big[\log\left([L] - [L]\cdot e^{\frac{\Delta E^{x}_{homolog}+\delta-C}{a}}\right) - \frac{\Delta E^{x}_{homolog} - C}{a}\Big] + \Delta G^{WT}_{PDZ3} - \Delta G^{WT}_{homolog} \tag{2.11}$$

2.4.3 Library generation

The library was generated using a set of primers encoding all possible pairs of codons of the 9-amino acid α-helix randomized as NNS to produce all possible single and double mutants. Each of the 36 primers was used in a PCR reaction with the wildtype PDZ allele as template, amplifying the entire plasmid but introducing the mutations only on one primer so as to reduce biasing and extra wildtype generation as seen in other techniques, such as overlap-extension PCR[80]. Importantly, upstream of both forward and reverse primers are BsaI sites, which, upon restriction digestion, leaves complementary (and scarless) overhangs. This allows for a convenient method of generating the double mutant library by combining all of the 36 PCR reactions and setting up a one-pot digestion-ligation reaction[42].

2.4.4 Co-evolution algorithms

An curated alignment of 1489 PDZ domains (courtesy of Alan Poole) was used in these analyses, with only minor differences obtained from using the Pfam alignment (Pfam: PF00595) that do not affect any of the conclusions in the work. More specifically, the Pfam alignment contains sequences with missing or circularly permuted regions that cause a large fraction of the positions analyzed to be ignored because of the prevalence of gaps.

The SCA6.0 package was downloaded from `http://reynoldsk.github.io/pySCA/index.html` and run with default parameters. Code for computing mutual information was written by Olivier Rivoire and obtained from the SCA5.0 package. A pseudolikelihood-optimization implementation of DCA[81] was downloaded (`http://gremlin.bakerlab.org`) and run locally with default parameters.

3

Biochronicity in cyanobacteria: a model for understanding the genotype to fitness map

Biological systems are challenged by the need to perform sophisticated computations in an environment that is highly variable. This demand, in turn, has led to the evolution of systems capable of efficiently and faithfully processing environmental fluctuations in relevant parameters such as nutrients and light into appropriate physiological responses. Consider the case of the cyanobacterial circadian clock: the central oscillator comprises a network of proteins that displays an intrinsic free-running 24-hour period, is robust to temperature fluctuations in the environment, and is capable of being entrained to match the cycle of the solar day. The clock regulates both the structure and transcriptional activity of the organism's genome in a diurnal rhythm, leading to circadian oscillations in the expression patterns of roughly two-thirds of all mRNA transcripts[174]. Moreover, the timekeeping mechanism of cyanobacteria functions to control the execution of cellular processes in accordance with the metabolic state of the cell[147]. In this way, for example, proteins sensitive to oxygen will be expressed in the nighttime when photosynthesis, the major ecological contribution of these organisms, is not being carried out[98].

Furthermore, this apparently complex behavior is mediated by a simple core system within which well-characterized biochemical reactions are taking place. Indeed, the *in vitro* recapitulation of a temperature-compensated clock with a 24-hour free-running period using only the purified proteins

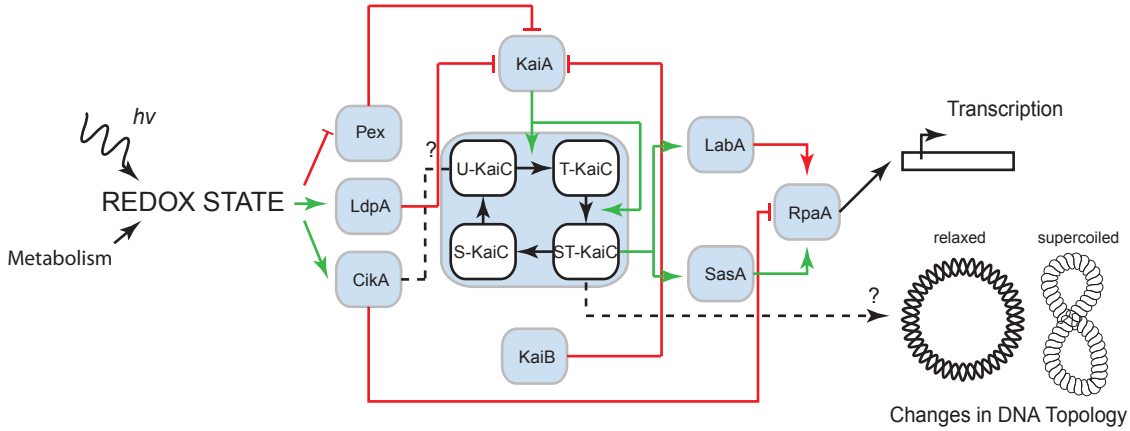

Figure 3.1: The core circadian oscillator of the model organism *Synechococcus elongatus* PCC 7942, a species of cyanobacteria, consists the three Kai proteins: KaiC, the enzyme that undergoes 24-hour cycles of phosphorylation at residues S431 (S-KaiC) and T432 (T-KaiC); KaiA, which stimulates the autophosphorylation of KaiC; and KaiB which counteracts this stimulation. These proteins are embedded in a larger network of proteins that transduce input from the environment (such as Pex, LdpaA, and CikA) and communicate the phosphostate of KaiC to the transcription factors (mainly RpaA) that control global gene expression.

KaiA, KaiB, and KaiC helped define this post-transcriptional oscillator as the one that drives the circadian rhythm *in vivo*[112] (Fig. 3.1). Specifically, this clock's outputs are the phospho-states of KaiC, which are determined by the autokinase and autophosphatase activities of the same protein and that are themselves modified by associations with KaiA and KaiB. As predicted, mutations in the *kai* genes that alter the period of the circadian oscillator as measured with an *in vivo* bioluminescent reporter similarly shift the period of the isolated protein clock[112].

More importantly, however, the regulation mediated by the clock is adaptive. Mixed-strain competition experiments in which cyanobacteria with wildtype clocks were co-cultured with mutants bearing clocks with altered periods demonstrate that matching of the clock's free-running period with the period of the experimentally applied light-dark cycles confers a fitness (growth) advantage[122].

The system's experimental tractability and biological and evolutionary significance* render it an ideal model system for studying the long sought-after links among genotype, phenotype, and fitness. How genotypic variation manifesting as changes in the biochemical activities of the Kai proteins maps to phenotype (here parametrized mainly as alterations in the period of the circadian clock), and how phenotype in turn maps to organismal fitness are fundamental questions that demand

further experimental exploration.

3.1 Results

3.1.1 Understanding the link between circadian clock function and fitness

My first foray into these questions aims to tackle the matter of the adaptive value of the circadian clock. Prior work alluded to above demonstrated that resonance between the periods of the circadian clock and that of the environmental light-dark cycles resulted in enhanced survival of the organisms[122]. The main issue with these studies lies in the authors' interpretation of the data, mainly that this enhanced survival is mediated by interactions among the strains in the competition culture wherein those with dissonant clocks had their growth inhibited by those with resonant ones[182,57,144]. A simpler proposal, however, would be that resonance between the circadian clock and the environment enhances the growth rates of bacteria. The authors discarded this hypothesis on the basis of growth experiments of isolated strains, but the data are insufficient to detect the small growth rate differences that could lead to divergence of populations in the time scales of these experiments (greater than 30 days).

In fact, using the same mathematical abstraction of the circadian clock previously employed to model the competition experiments[57,144], I simulated a competition between strains with 28- and 24-hour periods in a 24-hour light-dark cycle, removing any terms in the model involving growth inhibition. Growth in this model is parametrized as the overlap between specific phases of the environmental and circadian cycles of each strain (Fig. 3.2*B*), an assumption consistent with the observation that growth of *S. elongatus* is restricted to the time window when the cell's 'subjective day' coincides with actual daylight[57,144]. The results of the simulation clearly show that growth enhancement of the strain with a clock whose period matches that of the environmental light-dark cycle (Fig. 3.2*C* and *D*). need not invoke interactions between strains. Instead, resonance between the endogenous clock and the imposed light-dark cycle enables strains to grow for longer periods of time.

To experimentally test the model predictions, I set up the experimental system from Ouyang et al.[122] and Woelfle et al.[182] with some modifications to enable sensitive measurements of strain frequencies in the population using high-throughput sequencing and permit competition of more than two strains simultaneously. I generated 3 mutants of KaiC used by Ouyang et al.[122] and Woelfle et al.[182] whose characterized circadian clocks exhibit 22-hour, 28-hour, and arrhythmic periods (Table 3.1). These were barcoded and inserted into the neutral site I (NSI) site of the *S. elongatus* chromosome by standard methods[28].

[*] But see[38].

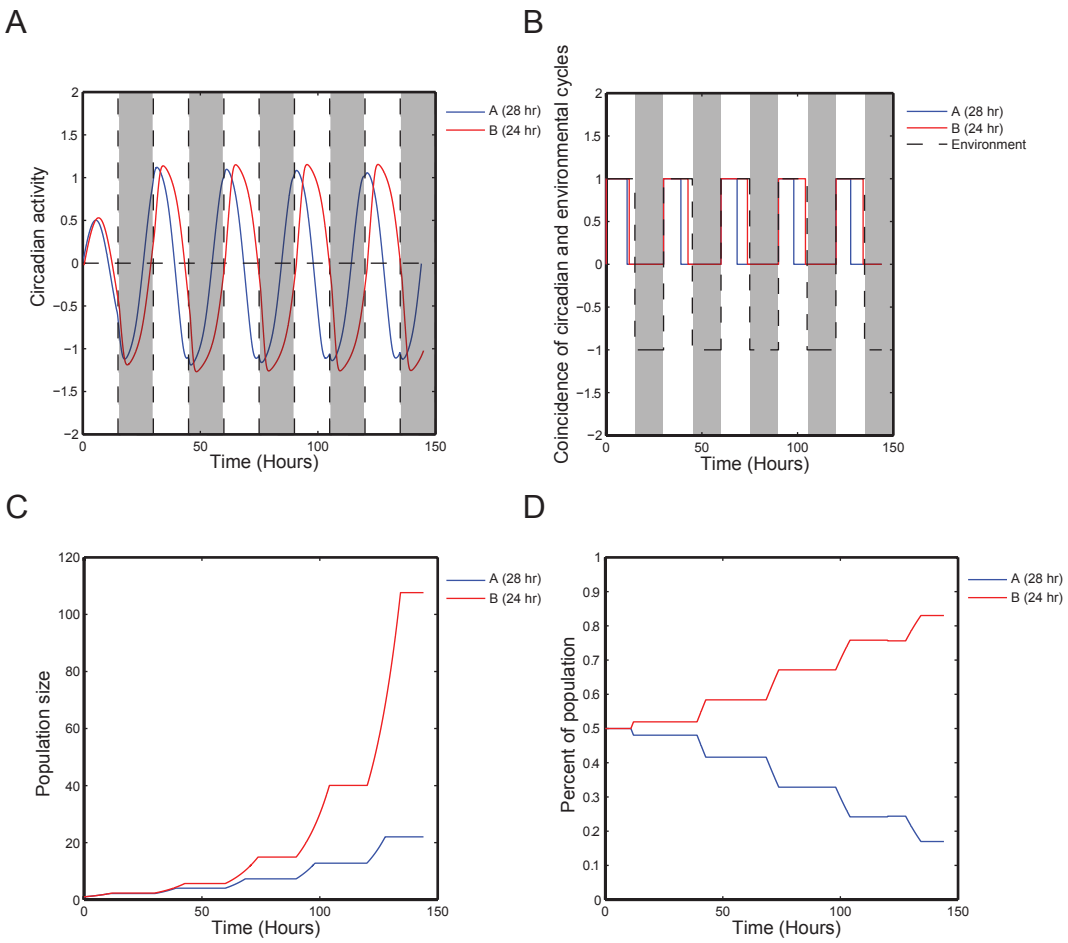

Figure 3.2: *A*, A simulated competition experiment was carried out between strains with 28-hour (A) and 24-hour (B) circadian clocks. Shown are the circadian activities of each strain subjected to a 24-hour period light-dark cycle. *B*, The coincidence of the positive phase of the circadian activity of each strain with the positive (white) phase of the environmental cycle are plotted (values oscillate between 0 and 1). The integral of the traces for each strain is proportional to their growth during the experiment, as the cells only grow when their subjective sense of light (positive phase of circadian activity) co-occurs during the actual light (white) phase. *C* and *D*, Composition of the competition culture over time, represented as the absolute number or fraction of each strain in the culture, respectively.

Mutant	*Period (Hours)*	*AA Mutation*
C1	22	A87>V
C4	28	P236>S
C13	Arrhythmic	G460>E

Table 3.1: Single amino acid mutations in KaiC produce cyanobacterial strains with altered period.

Competition under constant light led to the co-survival of the arrhythmic and wildtype (24-hr period) strains after 32 days (Fig. 3.3*A*). The strains with 22- and 28-hr periods, however, declined gradually in frequency over this same time period. In contrast, under a daily illumination schedule of 12 hours of light followed by 12 hours of darkness, the wildtype strain is the dominant strain in the culture by the end of the experiment; all other strains exhibit a gradual decline (Fig. 3.3*B*). Statistical analysis of the fractions of the population contributed by each strain shows that while the wildtype and arrhythmic strains showed increases in frequency in the constant light environment, only the strain with a 24-hour period exhibits growth in the 12 hour light:12 hour dark illumination protocol (Fig. 3.3*C*). These results are consistent with the previously reported phenomenon[122,182] of the cyanobacterial clock conferring increased fitness only in rhythmic environments.

The experiments of Ouyang et al.[122] and Woelfle et al.[182] also demonstrated that specific environments could enhance the growth of certain strains if the periods of the circadian clock and environmental light-dark cycle matched. To test this, I competed the strains under illumination protocols with periods of 22 and 30 hours (11 and 15 hours of light and dark), which should favor the growth of the 22- and 28-hour period strains, respectively. In neither of these competitions did the expected strain outcompete all others. Under the 22-hour light:dark cycle, while the 22 hour strain appeared to show an initial increase in frequency, by the end of the experiment its growth relative to the other strains declined and the wildtype (24-hour period) strain became dominant (Fig. 3.4*A*). On the other hand, under the 30-hour period environmental light-dark cycle, the 28-hour period strain exhibited gradual decline, while the culture at the end of the experiment consisted of both the wildtype and arrhythmic strains (Fig. 3.4*B*).

Collectively, these results supported only one conclusion of the previously published mixed-strain competition experiments: that the circadian clock has adaptive value only in rhythmic environments. The support is only partial, however, as two of the strains (those with 22-hour and 28-hour periods) exhibited poor growth under all illumination protocols, including those that were expected to enhance their survival.

Fundamental differences in the experimental setups could have contributed to these discrepancies. For example, both Ouyang et al.[122] and Woelfle et al.[182] employed mutants strains generated by chemical mutagenesis[85,78], which despite having the *kai* locus sequenced could have borne mutations at other sites in the genome as well. On the other hand, the strains tested here expressed the Kai proteins from a neutral site on the chromosome and were generated by site-specific mutagenesis. Ascertaining the dependence of the circadian phenotypes on expression from an endogenous locus or other mutations requires comparing the circadian activities of all strains, a task only enabled by automated methods that sample cultures over long periods of time in environments that support

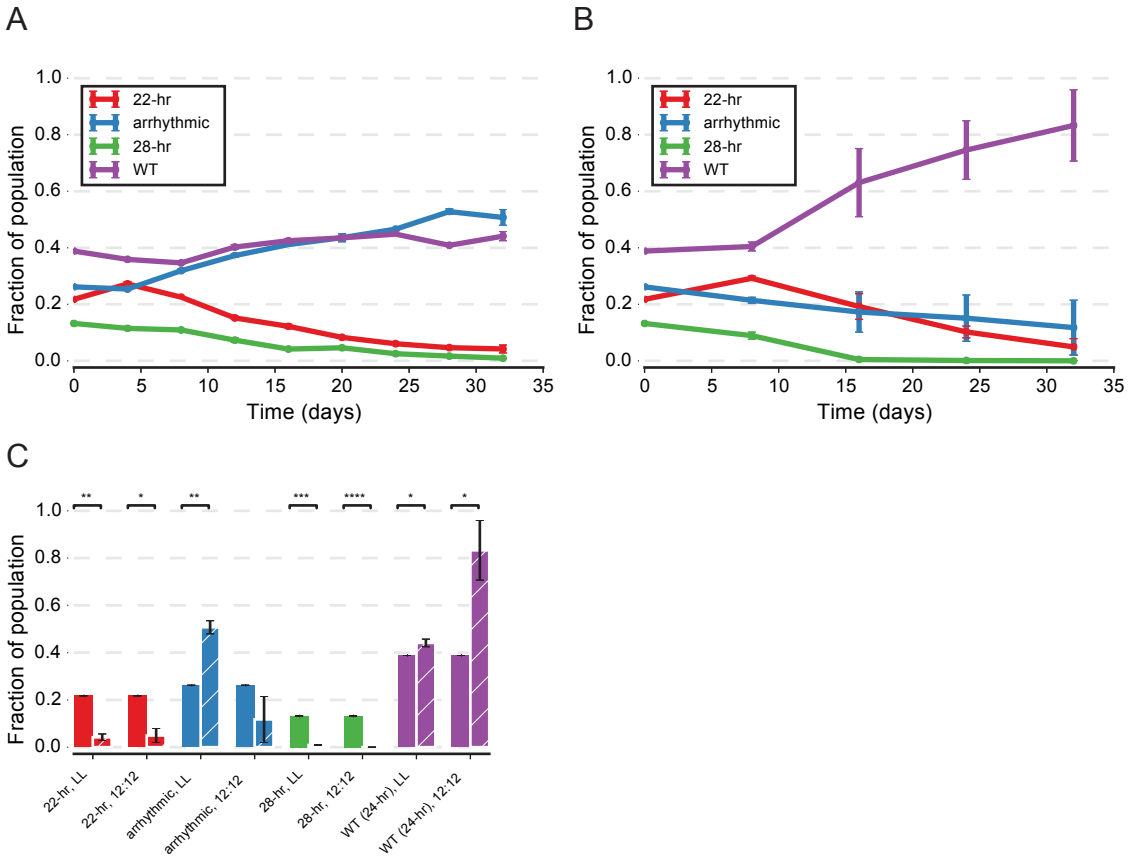

Figure 3.3: The four barcoded strains were mixed, underwent synchronization, and were competed in enviroments with constant light *A* or one in which light and dark alternated every 12 hours *B*. In the former, both the wildtype and arrhythmic strains grew slightly in frequency, whereas in the latter, the wildtype emerged as the dominant strain in the culture by the end of the experiment. Points and error bars represent mean plus standard deviation of two replicates. *C*, Initial (*solid*) and end (*striped*) frequencies of each strain in the population in both environments. Error bars represent the standard deviation between two replicates, while asterisks denote p-values (****, $p < 0.0001$; ***, $p < 0.001$; **, $p < 0.01$; *, $p < 0.05$; two-sided t-test).

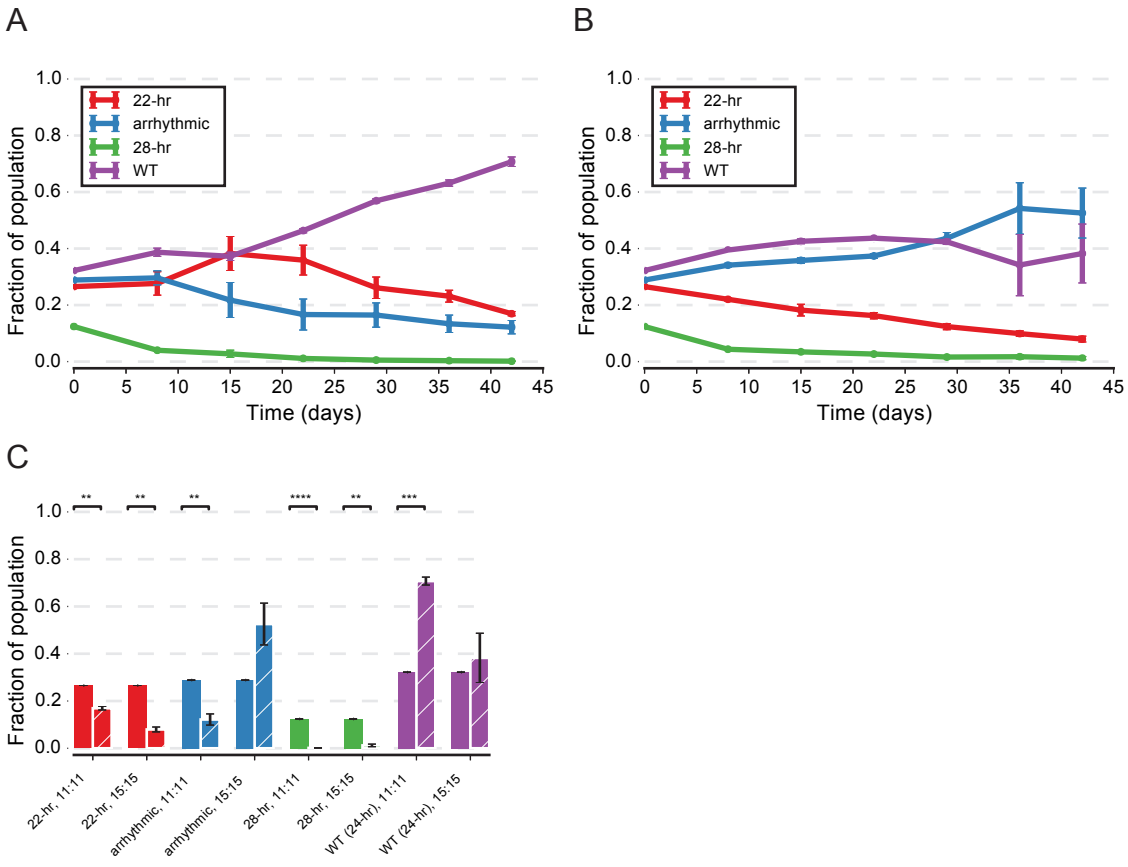

Figure 3.4: The four barcoded strains were competed in environments that favor the growth of the 22-hour, *A*, or the 28-hour, *B*, period mutants. Unexpectedly, the period-matched strain did not outcompete all others. *C*, Initial (*solid*) and end (*striped*) frequencies of each strain in the population in both environments. Error bars and asterisks as in Fig. 3.3.

the fastidious growth of these organisms. A continuous culture device was designed and built for this purpose and is described in Chapter 4.

3.2 Conclusion and Outlook

3.2.1 Characterizing the core unit of the cyanobacterial circadian clock

Though the isolatable unit of the cyanobacterial clock has been identified, it still finds itself embedded in a larger network comprising proteins that transduce environmental signals to the Kai oscillator and proteins that mediate transcriptional regulation. To what extent does the clock form a modular (transplantable) unit *in vivo* and what implications does this have for the clock's evolutionary history as seen in the record of extant sequences of clock genes? Preliminary analysis of the kai genes from 100 cyanobacterial species has revealed a substantial degree of sequence conservation (see section 3.2.2), an expected observation if one adheres to the assumption that the constancy of the solar cycles throughout the evolution of cyanobacteria was the major influence in the evolution of the clock genes. The experiments I propose to characterize the core unit of the clock entail making barcoded libraries of clock operons with genes from 5 different cyanobacterial species and assaying their ability to restore rhythmicity to the circadian system of the Kai operon-null mutant of the model strain *Synechococcus elongatus* PCC 7942. I predict a high degree of functionality among *kai* genes of different species. However, cyanobacteria as a phylogenetic group have evolved for almost 2 billion years and inhabit a range of environments characterized by different temperatures and day lengths (fraction of daylight in 24-hours). I, in turn, argue that adaptations to these environments are primarily mediated by the peripheral proteins that are carry out light and redox sensing and gene expression changes, a hypothesis supported by the recent recapitulation of the *S. elongatus* circadian clock in *E. coli*, whose success was attributed to the co-transplantation of other genes in the clock network[23]. Testing this requires carrying out the previous experiments in the background of strains in which some of the genes encoding the peripheral proteins from other species, among these the quinone-sensing CikA and the two component signaling pair SasA-RpaA, are heterologously expressed. These experiments, too, are enabled by advances in high-throughput sequencing, gene synthesis[86], and gene library cloning methods[24].

3.2.2 Revealing the influence of evolutionary constraints on the distribution of the effects of mutations

The experiments proposed here set the foundation for further work that will pursue two paths. The first concerns itself with the genotype-phenotype map, and it aims to reveal the contribution

of amino acid residues in the Kai proteins to the function of the circadian clock. Experimentally, this path entails saturation mutagenesis of the Kai proteins and selection in environments with varying periods and day lengths of light-dark cycles to determine the effects of mutations on the period of the circadian clock. The high degree of single-site conservation of the Kai proteins (Fig. 3.5) suggests a pattern of mutational sensitivity distinct from other proteins studied in a similar manner[105,164]. The second path concerns itself with identifying those amino acid residues in the Kai proteins that contribute to the system's adaptability. The proposed experiment subjects the Kai proteins to rounds of directed or continuous evolution challenging the wildtype proteins to environments with distinct periods to observe the pattern of changes that occur as the proteins evolve for optimal function in these new environments. Together, these paths seek to reveal the features of the cyanobacterial circadian clock that underlie its robustness and adaptability over evolutionary timescales.

3.3 Methods

3.3.1 Growth and maintenance of cyanobacterial cultures

The vast diversity of ecological niches occupied by cyanobacteria, the outcome of their long evolutionary history, complicated early attempts to isolate, purify, and grow these organisms[141]. Approaches to these efforts were refined over decades, and now many institutes carry a vast collection of axenic strains, chief among them the Pasteur collection of Cyanobacterial Cultures (PCC), which currently boasts more than 470 in its database (`http://cyanobacteria.web.pasteur.fr/`). Though strains ordered through the PCC are accompanied by detailed conditions for the growth and maintenance of cultures, it is worth discussing the important parameters that determine them.

As the majority of cyanobacterial species are photoautotrophs, the provision of adequate light of the appropriate spectral output and intensity is an absolute requirement. Both properties are largely defined by the predominant photopigments employed in the photosynthetic machinery of the organisms. In general, cool or warm white light emitted by fluorescent lamps, which have a spectral profile mostly in the visible light regime, suffice for the growth of many species, even those with a single predominant photopigment[21]. The endogenous photopigment repertoire of an organism also determines the optimal light intensity for culturing it. Consequently, whereas light of insufficient intensity may not support photosynthesis for some species, too much light may lead to photoinhibition, or the destruction of photopigments and ensuing reduction in growth for others[21,47]. Optimal light intensities for growing cyanobacteria range from 20 to 200 microeinsteins per square meter per second ($\mu E\, m^{-2}\, s^{-1}$)[5]. Light intensities for any organism may have to be tuned within this breadth,

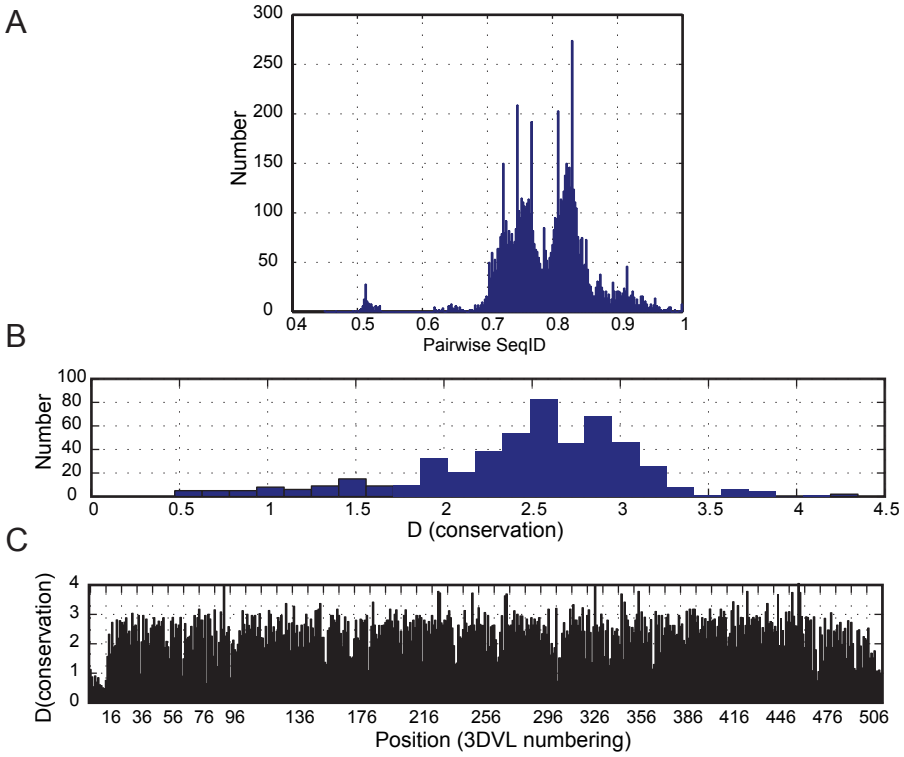

Figure 3.5: KaiC sequences were culled from the NCBI Refseq and Genome databases from organisms that also harbor the 2 other Kai proteins, and subsequently aligned using PROMALS3D [127] with manual curation. *A*, the distribution of pairwise sequence identities between all pairs of sequences in the alignment shows a degree of sequence similarity beyond that seen in typical protein families analyzed by methods such as the Statistical Coupling Analysis (15-50%) [140]. *B* and *C*, the distribution of positional conservation, computed as the Kullback-Leibler relative entropy D [62,140], and the magnitude of conservation as a function of amino acid position, respectively, both conveying the high degree of sequence conservation in this protein.

but this regime has been experimentally validated to provide enough photons for photosynthesis and maintain a reasonable growth rate (for example, *Synechococcus elongatus* PCC 7942 achieves a doubling time of $3.8\,\mathrm{d}^{-1}$ at 30 °C with an irradiance of $100\,\mu\mathrm{E\,m^{-2}\,s^{-1}}$ [122].

Though the circadian clock is functionally robust within the range of environmental temperatures experienced by any particular organism, temperature remains an important parameter that may have to be modulated for the culture of different cyanobacteria. The dependence of growth rates on temperature for many cyanobacteria has been well-documented and may vary by at least an order of magnitude between organisms isolated from copiotrophic and oligotrophic environments [21]. Optimal temperatures are thus determined by both the strain and its natural habitat, such that while thermophiles like *Thermosynechococcus elongatus* BP-1 grow optimally at 57 °C [119], experiments with the marine-dwelling *S. elongatus* PCC 7942 are regularly carried out at 30 °C.

Cyanobacteria derive their carbon source not from carbohydrates like glucose, but rather from dissolved CO_2 gas or carbonate salts. Gas exchange with the environment via incompletely closed caps on culture vessels provides sufficient CO_2 for the organism's energy demands, but the variability of gas transfer via this method may introduce variability in growth among clonal cultures. Bubbling in CO_2 gas or constantly infusing media with carbonate are strategies for standardizing the delivery of the carbon source, but steps must be taken so as to account for the alteration of the pH of the growth media as some organisms may be more pH-sensitive than others [21]. A plethora of media recipes exists with variable suitability for culturing cyanobacteria [21]. Concentrations of various salts and trace metals, type of water (sea or pure), pH, and presence of nitrates are all tailored for the optimal growth of various groups of organisms. Modified BG-11(N^+) media [17] is a universal media distinguished by its low phosphate concentration that can be used to culture nitrogen fixing and non-nitrogen fixing (including *S. elongatus* PCC 7942) cyanobacteria by omitting nitrate and ammonium during the preparation.

The experimental setup for culturing and conducting experiments with cyanobacteria involves a cool white Sylvania fluorescent lamp installed inside a temperature-controlled shaker. The light intensity experienced by culture flasks as measured by a luminometer (VWR) is about 4500 lux (roughly $60\,\mu\mathrm{E\,m^{-2}\,s^{-1}}$). The lamp itself is regulated by a digital outlet controller (Fisher Scientific) that permits us to program cycles of darkness and illumination at various periods.

3.3.2 Isolation of genomic DNA from cyanobacteria

Genomic DNA was prepared for genotyping and PCR amplification of barcoded regions for high-throughput sequencing of the competition experiments. and for sequencing of cyanobacterial genomes (both described below). I used the MasterPure Complete DNA and RNA Purification

kit (Epicenter) to isolate up to 50 μg of DNA of suitable length and purity for direct applications. Briefly, 20 ml of culture (at $0.1OD_{750}$) are spun into a pellet, lysed in a chaotropic salt solution, separated from an organic protein phase, and isopropanol-precipitated before being resuspended to usable concentrations.

3.3.3 Strain barcoding and competition experiments

Six-base pair barcodes were appended after the stop codon of KaiC in the vector AMK1 (a variant of the neutral site I (NSI)-targeting vector pAM2314 into which the kaiABC operon was cloned) by PCR. Plasmids encoding barcoded mutants and the wildtype were transformed into *S. elongatus* strain AMC669 using standard techniques[28,27]. Strain AMC669 bears a chromosomally-integrated copy of the *luxABCDE* system at NSII for monitoring of circadian activity via bioluminescence, and its endogenous *kai* operon was replaced by a kanamycin-resistance cassette using plasmid pAM4252. Transformations involve the incubation of 10^8 cells overnight in the dark with 100 ng of plasmid DNA and subsequent plating onto BG-11(N^+) agar supplemented with appropriate antibiotics. Because of the presence of multiple chromosomes in these cells[79], the *sacB* gene coding for the levansucrase enzyme was inserted into the NSI site to function as a negative selection marker and permit selection of homozygous transformants which have replaced the *sacB* gene with the barcoded *kai* operon (levansucrase metabolizes sucrose into toxic by-products[131]).

Competition experiments began with a 2-day period of entrainment to the experimental light-dark cycle protocol. The strains were co-cultured either as groups of 4 or 2 with near equal frequencies at the start of the experiment. Every 4 or 8 days (depending on the environmental period), an aliquot of the cells was removed for genomic DNA extraction and amplification of the barcoded region for sequencing-based determination of the strain frequencies in the culture. The cells were diluted 1:100 into fresh BG-11(N^+) media at every such interval to prevent the cell density from exceeding $0.2OD_{750}$ units.

Isolated genomic DNA from each time point was subjected to two rounds of PCR to add Truseq barcodes that correspond to the particular time point and Illumina adapters. The resulting amplicons were sequenced in a Miseq instrument, the output of which was processed using UNIX bash text parsing commands and analyzed in Matlab.

4

A versatile continuous-culture platform for experimental evolution

Laboratory experiments aimed at studying microbial processes such as the evolution of antibiotic resistance require methods to carefully manipulate environments and to sensitively measure meaningful, albeit often subtle, changes in the growth of the organisms[76]. These experiments are typically conducted in a bioreactor where a coupled waste and nutrient delivery system allow for the maintenance of cells in liquid culture for long periods of time. The chemostat, a manifestation of a continuous culture apparatus, was developed in the 1950s[115,107] to study growth dynamics of bacteria and was subsequently adopted for the culturing of these organisms close to a steady state described by a rate of growth of the cells, r, equal to the dilution rate of media, D[149],

$$\frac{dN}{dt} = rN - DN \tag{4.1}$$

where N is the cell density of the organisms in culture. This allowed investigators to measure the effects of environmental perturbations independent of confounding growth effects observed when bacterial cultures were allowed to experience growth rate fluctuations. Importantly, the method of continuous culture in the laboratory is also analogous to growth in natural environments[76]. Recent experiments have exploited this similarity to design evolution experiments where bacteria are

grown in the presence of increasing concentrations of antibiotics as they gradually evolve resistance to them[172]. By keeping the evolutionary pressure constant, these efforts not only identified genes responsible for the underlying resistance, but also allowed accurate comparisons between patterns of genetic changes observed in other identical cultures and cultures with different antibiotics. The ultimate aim is to understand the genotype and fitness map and predict, for example, the genetic changes that confer antibiotic resistance and design strategies to prevent them in the clinic.

The implementation of a basic continuous culture device depends on a controller that can monitor changes in culture density and, as dictated by set threshold parameters, turn on pumps that deliver new media or antibiotic drug and remove excess culture. Experiments using such devices are also typically carried out for long periods of time (greater than a week, or in one instantiation, several decades[11]) and are executed in independent vessels to permit assessment of reproducibility. In practice, however, construction of these devices and their actual use has proven to be technically challenging, and central to this has been the design of reliable and robust hardware and software to execute monitoring and feedback, all the while allowing for modular control of many vessels. Among the specific problems encountered are: crashing of the command execution and data acquisition and storage systems, which also puts into question the reliability of data that depends on precise timing of events; failure of a feedback signal to execute, resulting in the overflow of liquid in vessels; and lack of independent control of vessels, hindering the investigator's ability to sample many experimental conditions.

Here I describe the use of a continuous culture platform ("Verivital") designed to overcome these limitations. Its design and construction are the work of Dr. Taylor Johnson and his laboratory at the University of Texas at Arlington, while the specifications and testing were carried out primarily by Ian White (a graduate student in the Ranganathan lab at the University of Texas Southwestern Medical Center) and myself. As the hardware and software of this system have been described in great detail elsewhere[114], this discussion is intended to serve as a general guide for future users, wherein I outline the process of setting up an experiment, show a simple analysis of actual data, and demonstrate a unique use of the system in a synthetic biology experiment.

4.1 Experimental setup

4.1.1 Hardware description

The Verivital system consists of three modules (Fig. 4.1): (1) a Beaglebone Black running Xenomai, an development framework for carrying out all real-time tasks (optical density (OD) sensing, valve and air pump actuating); (2) a sensor board connected to the GPIO pins of the Beaglebone for re-

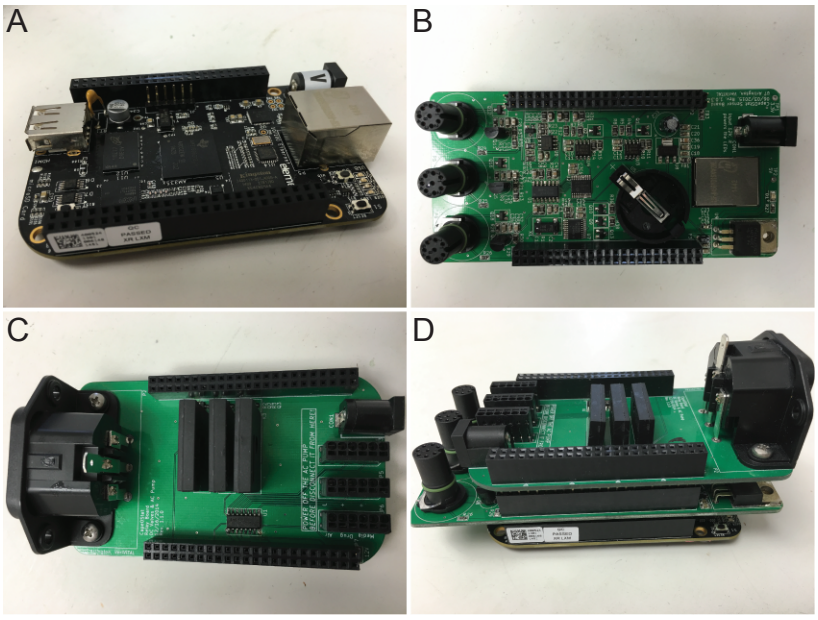

Figure 4.1: The Verivital system comprises three modules: *A*, the Beaglebone Black that contributes the processing power for the real-time tasks and data acquisition; *B*, a sensor board to which the IR LEDs are connected; and *C*, a relay board that houses connections for the media valves and air pumps. A complete unit (*D*) can support the growth of 3 cultures controlled independently.

ceiving signal from infrared (IR) detectors for measuring OD; and (3) a relay board to which the media valves and air pumps are connected. Putting together the boards and peripherals entails nothing more than connecting wires or leads to color-coded or labeled pins.

The operation of the turbidostat is illustrated in Fig. 4.2: an IR emitter/detector pair measuring turbidity in the culture senses the density passing a defined threshold, activating the PID controller to actuate the relay to turn on the media valve. Media flows from a reservoir located higher than the culture bottle for a few seconds, thereby diluting the culture. After a short delay, the media valve is turned off, the air pump relay is actuated, and air coming from the pump provides the positive pressure to move excess media into a waste container.

4.2 Protocol

4.2.1 Running the turbidostat

Use of the Verivital requires minimal knowledge of Linux or Mac Terminal. The protocol assumes that all tubing and culture vials have been sterilized and a calibration has been carried out to relate

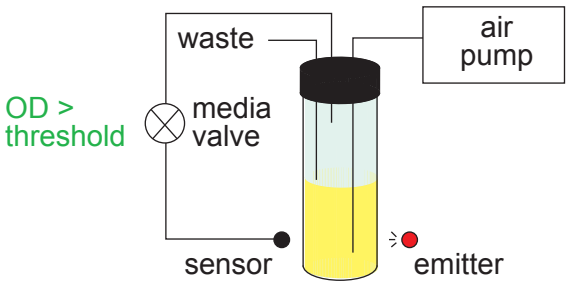

Figure 4.2: In a turbidostat, culture density is sensed by an IR sensor-emitter pair (that actually detect light scattering) of the cells and particulate matter in the culture bottle. If the detected optical density exceeds a specified threshold, a media valve is opened for a brief time and media flows into the bottle, thereby diluting the culture. An air pump mixes the culture and provides the positive pressure to push excess cells and media to a waste container.

voltage with optical density.

1. Autoclaved media reservoir [2 L bottle with spout] 250 mL bottle with cap. 2L of Sterile media(depends on length of experiment).
   (a) When you autoclave media reservoir, take off the rubber stopper and wrap with tin foil.
   (b) Make sure that all 4 tubes are capped: the tube from the media reservoir and the 3 tubes coming from the 250 mL bottle.
2. Use a stripette or other sterile device to transfer 150 mL of media into 250 mL bottle and 2 L into the media reservoir.
3. Connect the only tubing from the media reservoir into the top part of the tubing connected to the valve. [complete this step while the media reservoir is below the valve on the bench to avoid a spill.]
4. Raise the media reservoir to the top shelf.
5. Connect the 3 tubes in the cap.
   (a) The tube that has the greatest depth in the 250 mL bottle is the air tube, connect this tube to the air pump and ensure that the inner tube is as close to the bottom of the bottle as possible.
   (b) The tube that has middle depth is the outflow tube. Place the end of this tube into a 2L plastic container. Place tin foil over the top of the plastic 2L container, and poke a hole to insert the end of the tube. Ensure that the tube is resting on the top of the foil and that it is secure.
   (c) The tube that has the least depth is the media tube. Connect this tube into the bottom

aspect of the valve.

6. Check the 4 screws on the bottle-cap to check for tightness. Also, screw down cap.
7. `ssh` into the BeagleBone through wifi or USB as `root`. The password is `vital_uta`.
8. Use the command `cd /home/debian/work/hg-repo/systems/BBB/xeno_src/` to move into the correct folder.
9. Use the command `cp base_configuration_set # base_configuration` to copy the correct base configuration over. Replace the # with either 1, 2, or 3 depending on the system that will be used.
10. Use the command `vi base_configuration` to open the "vi" editor and edit the base_configuration file.
11. The parameters that could be changed are as follows:
    (a) Upper Control Limit (mV): This is the clamping voltage. When the culture reaches this limit, the system will dilute with media. A higher Upper Control Limit corresponds to a higher clamped OD. The most commonly used range is from 500-1000.
    (b) TSENSE (ms): This is the wait-time between dilution decisions. A shorter T-SENSE can be used in conjunction with a smaller MM_ON_TIME to achieve a closer clamping to a specific OD. A constraint on T_sense is that T_sense < AIR_OFF_LIGHTS_ON_TIME + AIR_OFF_LIGHTS_OFF_TIME
    (c) Tcontrol: This parameter should be 2000 ms. Do not change.
    (d) MM_ON_TIME: The amount of time the media motor is on in seconds.
    (e) DM_ON_TIME: Only used in morbidostat mode. This is the amount of time the drug pump will be on.
    (f) DM_WAIT_TIME: Only used in morbidostat mode. This the amount of time the drug motor will wait before sending drug.
    (g) AIR_OFF_LIGHTS_ON: This parameter is used to allow bubbles to recede for an accurate OD measurement. Should be 5-10 seconds.
    (h) AIR_OFF_LIGHTS_OFF: This parameter is used to turn off an LED strip (for growth of cyanobacteria) prior to making a measurement upon which a dilution decision is based as it interferes with the IR sensor. Leave at 1000 ms.
    (i) POT_1: This is the pot value that is chosen after a calibration has been done. In practice, POT_Values tend to be from 8-150, with 100 as convention.
    (j) Do not change other parameter values.
12. Run the command `./verivital2` to start program.
13. MONITOR for approximately 5 minutes to ensure the presence of no leaks, that air is mak-

ing bubbles in culture and that the outflow (waste) is being pushed quickly into the waste container.

4.2.2 Stopping/cleaning the turbidostat

1. Use the `cd` command to move into the folder `log_files`.
2. Use `ls` to show all folders in log_files.
3. Find the most recent folder.
   (a) Run the command `cat configuration`
      i. If it corresponds to the configuration file of the system you want to stop, continue.
      ii. Else, repeat step 3 with next most recent folder.
4. Use `cat configuration` to read the configuration files.
5. croll to the bottom to find the PID (process ID) number.
6. Type `pgrep verivital2` to verify that the PID number of interest is running.
7. Type the command `kill #` where # denotes the PID number.
8. Confirm that the process is no longer running by typing `pgrep verivital2`.
9. Obtain ~1 L of water, 1 L of water with 20-30% ethanol, and a 2 L empty container. Place a cap in the 2 L container and obtain 20 ml syringe.
10. Flush each tube by rinsing with 60 ml of water, then 60 ml of ethanol solution, then 60 ml of water.
11. Find a new 250 ml bottle and place it in black bottle-holder to make sure it fits. After cap is left in ethanol solution, loosely screw cap onto new bottle.
12. Dispose of remaining cells from old bottle and overflow.
13. Rinse the large media reservoir one time with water, and two times with distilled water. Make sure that water is allowed to travel through tubing.
14. Open the media valve by after connecting via `ssh` into Beaglebone.
   (a) Look in the relevant configuration file and find the pin number of the specific media valve you want to turn on.
   (b) Go to the pin directory: `cd /sys/class/gpio`
   (c) Command to export pin: `echo # > export`
   (d) Command to check pin direction (`OUT` or `IN`): `cat direction`
   (e) If direction is `IN`, change to `OUT`: `echo out > direction`
   (f) Verify change of direction.
   (g) Check value, should be 0 : `cat value`

(h) Turn on pump: `echo 1 > value`

(i) A click will be heard confirming valve has been opened.

15. Clean this tube by using same procedure as before. 60 ml water, 60 ml ethanol solution, 60 ml water.
16. If strong contamination suspected, rinse with a 5% bleach solution as well. Order is now water, bleach, water, ethanol, water.
17. Close the valve: echo 0 > value
18. Return to original directory, `cd /sys/class/gpio`, and deactivate pin `echo # > unexport`.
19. Re-cap all tubes.

4.2.3 Analyzing/Monitoring Data

All data is stored in the `/home/debian/work/hg-repo/systems/BBB/xeno_src/log_files` directory as comma-separated text files labeled by a timestamp denoting the start of an experiment. Below I demonstrate the analysis of an 18-hour turbidostat run carried out on DH10B *Escherichia coli* cells in LB media at 25 °C. The analysis is done using Matlab, although many programs can handle the format of the data files.

```
%% load data
fID=fopen('XX.csv'); % XX.csv is the data file name
data=textscan(fID,'%s %s %s %s %s %s %s %s %s %s %s %s %s ...
    %s','Delimiter',',');
fclose(fID);

%% organize data
time = data{2}(2:end,:);
adc = data{4}(2:end,:);

ix=1;
for i=1:length(time)
    if any(adc{i})
        time2(ix)=str2double(time{i});
        adc2(ix)=str2double(adc{i});
        ix=ix+1;
    end
end

%% zero and convert time variable
time2=time2-time2(1);
time2=time2/1000/60/60;

```

```
%% find peaks & troughs
[pks,locs]=findpeaks(adc2);
[tr,locstr]=findpeaks((-1.*adc2));
plot(time2,log2(adc2),'-o'); hold on; ...
    plot(time2(locstr),log2(-1.*tr),'ro'); hold ...
    on;plot(time2(locs),log2(pks),'go');

%% compute growth rates (doubling times)
gr=[];
plot(time2,log2(adc2),'-o'); hold on;
for i=2:length(locs)
    tmp=polyfit([time2(locstr(i):locs(i))],[log2(adc2(locstr(i):locs(i)))],1);
    tmp_y=polyval(tmp,[time2(locstr(i):locs(i))]);
    plot([time2(locstr(i):locs(i))],tmp_y,'-r');hold on;
    gr=[gr 1./tmp(1)];
end

%% plot data
subplot(2,1,1)
plot(time2,log2(adc2),'-o'); hold on; ...
    plot(time2(locstr),log2(-1.*tr),'ro'); hold ...
    on;plot(time2(locs),log2(pks),'go');
xlabel('Time (Hours)','FontSize', 16);ylabel('log_{2} OD','FontSize', 16)

subplot(2,1,2)
plot(time2(locs(2:end)),gr);xlabel('Time (Hours)','FontSize', ...
    16);ylabel('Doubling time (hour^{-1})','FontSize', 16)
xlim=get(gca,'xlim');
hold on;
plot([xlim(1) xlim(2)],[trimmean(gr,50) ...
    trimmean(gr,50)],'--r','LineWidth',2);
% print('growth','-depsc')
```

The output of this analysis clearly shows that the turbidostat successfully clamped the culture around a set value (Fig. 4.3*A*) at an average doubling rate of $0.58\,\mathrm{h}^{-1}$ (Fig. 4.3*B*). Fluctuations in the growth rate trace result from its calculation from two OD values, which is improved by increasing the measurement sampling rate.

4.3 Growth competition experiment in a turbidostat

I carried out a bacterial two-hybrid experiment as described in Chapter 2 except that each step in the protocol was conducted in the turbidostat. A continuous culture system offers the benefits of keeping the optical density, and therefore physiological state of the cells, constant throughout the experiment as well as demanding no input from the experimenter to carry out dilutions, (therefore allowing experiments to proceed in perpetuity).

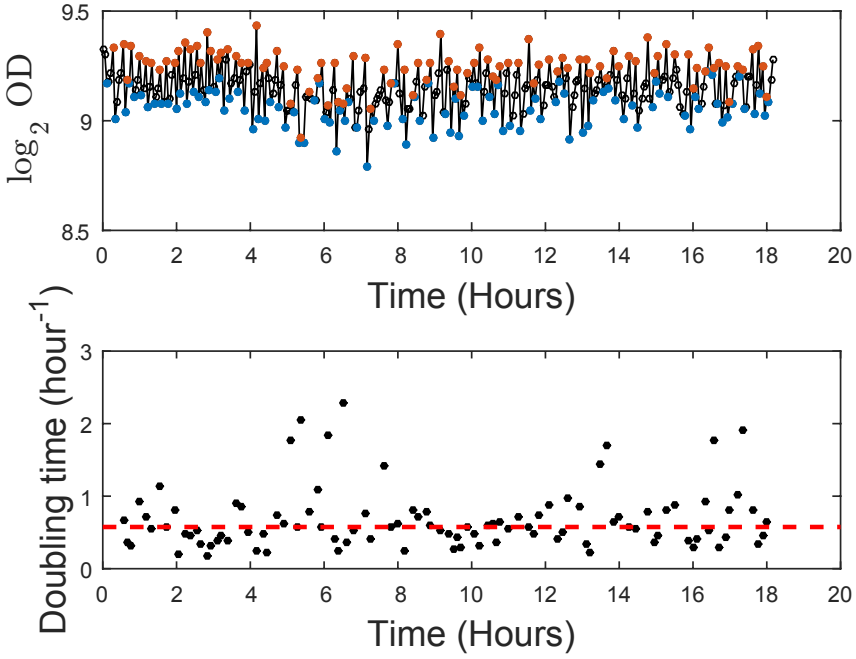

Figure 4.3: *A*, a trace of $\log_2$ -normalized OD (in units of mV) versus time (h) demonstrates the ability of the turbidostat to maintain the optical density of a culture of *E. coli* at a constant value. The magnitude of the fluctuations around the threshold OD are set by the measurement frequency and the time over which the media valve is turned on. *B*, growth rates are calculated from the slopes of linear fits connecting all points between successive troughs and peaks of the log-transformed OD trace (*blue* and *red* cicles in *A*, respectively). Large fluctuations in growth rate measurements result from fits from few data points. The red horizontal line indicates the average growth rate of the culture (ignoring the large fluctuations) over this time period.

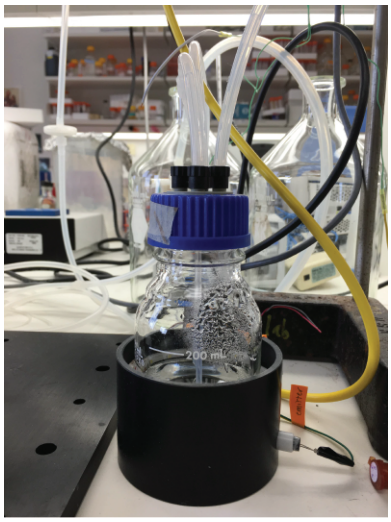

Figure 4.4: A 250 ml bottle with cells and media is housed in a holder into which the IR sensor and emitter are connected. The tubing lines on the bottle lid connect the air pump, waste line, and media valve/reservoir to the culture.

In this competition culture, cells harboring PSD95^{pdz3} single point mutants with previously characterized binding affinities to the CRIPT ligand were grown together in a flask and induced with 50 ng/ml doxycycline in a shaker at 25 °C (see Chapter 2). After induction, the culture was transferred to a 250 ml bottle (Fig. 4.4) and connected to the turbidostat set to maintain the culture at $0.15 OD_{600}$. Chloramphenicol was added to a final concentration of 150 ug/ml to begin selection, and 10 ml samples of the culture were taken at various timepoints for assessment of allele frequencies of the mutants by high-throughput sequencing.

The experimental results reproduced the linear relationship between relative enrichment of each mutant and free energy of binding, ΔG, showing an improved correlation over time (Fig. 4.5). The time course could potentially be used to fit growth rates[139], which may better reflect the *in vitro* binding measurements.

Continuous evolution experiments[43] require the monitoring and maintenance of cells over long periods of time. Demonstrations of the Verivital system have shown its robustness and versatility in enabling various types of experiments. Most importantly, however, it represents an improvement over an existing continuous-culture device in our laboratory based on the design of Toprak et al.[172]. The latter system confined cultures to 15 ml vials (low population sizes) and low-nutrient media (slow peristaltic pumps used to deliver media set the maximum growth rate that organisms could be clamped at), and because the data acquisition, data storage, and pump actuation were all overseen

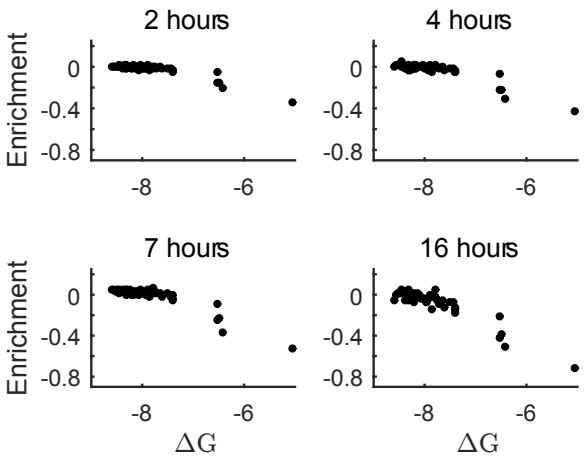

Figure 4.5: Mutants with characterized binding affinities were selected in the bacterial two-hybrid in the turbidostat, with the culture sampled at the indicated time points. The relationship between allele enrichment and binding free energy, ΔG, shows increasing linearity over time.

by software-based control within the same loop, the system required daily restarting because of imminent failures that stalled data logging and led to media overflows. These are all addressed by the Verivital system at a fraction of the cost.

5
Conclusion

The interpretation of the genome represents one of the grand challenges in the life sciences today. The philosophy driving the approaches described in this body of work is that the key, or "Rosetta Stone" to hark back to the analogy of decipherment of the Introduction, to the genotype-phenotype map lies in understanding the pattern of interactions between components of a biological system, be they amino acids as in the PDZ domain of Chapter 2, or proteins as in the cyanobacterial circadian clock of Chapter 3. This philosophy is rooted in intuition, as most biological phenomena involve or emerge from such interactions, and it is motivated by the success of a model of statistical co-evolution to explain a wide range of experimental data (see Chapter 1).

The experiments in Chapter 2 provide a direct test of the statistical model and reveal that the information encoded in the co-evolutionary statistics represents background-averaged coupling or epistasis[133]. This revelation was only made possible by interrogating the pattern of amino acid interactions in many homologs of a protein family, distinguishing this work from other efforts that focused on a single protein. However, the correlation between the experimental data and co-evolution is not perfect. The reason why may be as trivial as a sampling issue, as data of only 5 homologs are being averaged. But while it is easy to advocate more data collection, the interpretation I propose concerns the implicit assumption that the binding measurements that were made represent the 'fitness function' of these proteins. Other constraints, such as thermodynamic stability[12] and specificity (avoidance deleterious cross-talk[186]), may have likely influenced the evolution of these proteins[162].

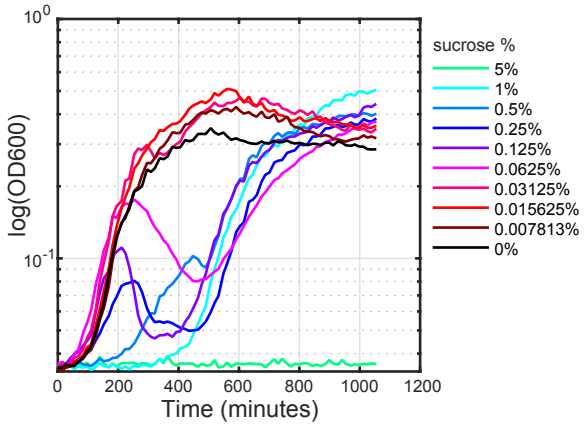

Figure 5.1: I replaced the antibiotic resistance gene with *sacB* as the reporter in the two-hybrid and tested the growth of cells harboring plasmids encoding wildtype PSD95^{pdz3} and its cognate ligand CRIPT in the presence of increasing concentrations of sucrose. The results show that the modification allows for tunable selection against high-affinity binding that opens up investigations into the role of negative selection in PDZ domain evolution.

In fact, to enable testing the role of specificity in PDZ domain function and evolution, I developed and tested a modification of the bacterial two-hybrid wherein the *CAT* antibiotic resistance gene was replaced with *sacB*, which codes for the enzyme levansucrase that produces toxic metabolites and kills cells in the presence of sucrose (Fig. 5.1). While the role of negative selection in PDZ domain evolution remains relatively unexplored, that the correlation between co-evolution and coupling with respect to binding affinity is already as significant suggests that affinity may have indeed been the more dominant selection pressure.

Cognizant of the risk associated with such an assumption, the cyanobacterial clock was adopted as a model system because of our understanding of the evolutionary and ecological forces that have and continue to shape the system. The constancy of the solar day is even reflected in the sequence statistics of the Kai proteins, as they exhibited a degree of conservation seen only in a fraction of so-called "essential" genes. The work described in Chapter 3 represents the foundation for a fruitful line of inquiry investigating: the influence of a constant environmental pressure on the distribution of mutation effects, how interactions between proteins shape the tolerance of amino acids to mutation and the pattern of amino acid interactions, and the definition of modules as groups of proteins that communicate strongly with one another and more weakly with the rest of the proteome.

One of the most surprising contributions of this work is that one need not sample all possible pairwise mutations to reveal the intrinsic coupling between amino acid positions in a protein (see Chapter 2). This, too, is consistent with the sequence analysis, as evolutionary couplings are com-

puted from a small fraction of all possible pairs of substitutions; indeed, most of the signal is contributed by the most frequent amino acid at each site[62,159]. This, coupled with the apparent sparsity of significant interactions in a protein, provides further evidence that the encoding of information in the amino acid sequences of proteins inhabits a lower-dimensional space than is suggested by the vast number of interactions of various orders into which the problem may be parametrized. The most unequivocal demonstration of this hypothesis remains the successful design of folded and functional sequences built with the constraints of the co-evolution information alone[145,160].

The most pressing question motivated by this work concerns the mechanisms by which such sparse encoding is effected. Fluctuations in selection pressures have been shown to lead to architectures of strongly coupled components ("modules") beset by weakly interacting ones in simulated systems[82,66]. This very architecture is also thought to endow organisms with the ability to tolerate genetic and external perturbations and adapt efficiently to novel environments[55]. The experiments envisioned center around continuous evolution: precisely controlling selection pressures and keeping track of the genetic changes accruing in the population. The evolved products can then be subjected to the experimental analyses described in Chapter 2 to reveal the patterns of energetic interactions within them. These patterns can then be correlated to parameters of the environment (magnitude, duration, and frequency of switching of selection pressures) to derive a precise mapping between the two. One might posit, for example, that a constant selection pressure would lead to a more dense pattern of significant interactions (where all amino acids are coupled to one another) as they evolve toward an optimal solution to the environment; such solutions might, as a consequence, exhibit increased sensitivity to mutations that would disrupt the highly coupled system. The power of these forward evolution experiments is that they require no inferences about evolutionary history, rendering these experiments crucial to testing our hypotheses about the important constraints on protein evolution. That the technology exists[43] to conduct them only intensifies the anticipation for answers to these questions.

References

[1] Ackers, G. K. & Smith, F. R. (1985). Effects of site-specific amino acid modification on protein interactions and biological function. *Annu Rev Biochem*, 54, 597–629.

[2] Adkar, B. V., Tripathi, A., Sahoo, A., Bajaj, K., Goswami, D., Chakrabarti, P., Swarnkar, M. K., Gokhale, R. S., & Varadarajan, R. (2012). Protein model discrimination using mutational sensitivity derived from deep sequencing. *Structure*, 20(2), 371–81.

[3] Alber, T. (1989). Mutational effects on protein stability. *Annu Rev Biochem*, 58, 765–98.

[4] Alber, T., Sun, D. P., Wilson, K., Wozniak, J. A., Cook, S. P., & Matthews, B. W. (1987). Contributions of hydrogen bonds of thr 157 to the thermodynamic stability of phage t4 lysozyme. *Nature*, 330(6143), 41–6.

[5] Allen, M. B. & Arnon, D. I. (1955). Studies on nitrogen-fixing blue-green algae. i. growth and nitrogen fixation by anabaena cylindrica lemm. *Plant Physiol*, 30(4), 366–72.

[6] Altschuh, D., Lesk, A. M., Bloomer, A. C., & Klug, A. (1987). Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol*, 193(4), 693–707.

[7] Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S. I., & Langmead, C. J. (2011). Learning generative models for protein fold families. *Proteins*, 79(4), 1061–78.

[8] Beard, W. A., Stahl, S. J., Kim, H. R., Bebenek, K., Kumar, A., Strub, M. P., Becerra, S. P., Kunkel, T. A., & Wilson, S. H. (1994). Structure/function studies of human immunodeficiency virus type 1 reverse transcriptase. alanine scanning mutagenesis of an alpha-helix in the thumb subdomain. *J Biol Chem*, 269(45), 28091–7.

[9] Blaber, M., Baase, W. A., Gassner, N., & Matthews, B. W. (1995). Alanine scanning mutagenesis of the alpha-helix 115-123 of phage t4 lysozyme: effects on structure, stability and the binding of solvent. *J Mol Biol*, 246(2), 317–30.

[10] Bloom, J. D. (2014). An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol Biol Evol*, 31(8), 1956–78.

[11] Blount, Z. D., Barrick, J. E., Davidson, C. J., & Lenski, R. E. (2012). Genomic analysis of a key innovation in an experimental escherichia coli population. *Nature*, 489(7417), 513–8.

[12] Boucher, J. I., Bolon, D. N., & Tawfik, D. S. (2016). Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. *Protein Sci*.

[13] Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A., & Sauer, R. T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, 247(4948), 1306–10.

[14] Bowie, J. U. & Sauer, R. T. (1989). Identifying determinants of folding and activity for a protein of unknown structure. *Proc Natl Acad Sci U S A*, 86(7), 2152–6.

[15] Burger, L. & van Nimwegen, E. (2010). Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol*, 6(1), e1000633.

[16] Burstein, E. S., Spalding, T. A., Hill-Eubanks, D., & Brann, M. R. (1995). Structure-function of muscarinic receptor coupling to g proteins. random saturation mutagenesis identifies a critical determinant of receptor affinity for g proteins. *J Biol Chem*, 270(7), 3141–6.

[17] Bustos, S. A. & Golden, S. S. (1991). Expression of the psbdii gene in synechococcus sp. strain pcc 7942 requires sequences downstream of the transcription start site. *J Bacteriol*, 173(23), 7525–33.

[18] Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., Chen, D. D., Schupp, P. G., Vinjamur, D. S., Garcia, S. P., Luc, S., Kurita, R., Nakamura, Y., Fujiwara, Y., Maeda, T., Yuan, G. C., Zhang, F., Orkin, S. H., & Bauer, D. E. (2015). Bcl11a enhancer dissection by cas9-mediated in situ saturating mutagenesis. *Nature*, 527(7577), 192–7.

[19] Capra, E. J., Perchuk, B. S., Skerker, J. M., & Laub, M. T. (2012). Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. *Cell*, 150(1), 222–32.

[20] Carter, P. J., Winter, G., Wilkinson, A. J., & Fersht, A. R. (1984). The use of double mutants to detect structural changes in the active site of the tyrosyl-trna synthetase (bacillus stearothermophilus). *Cell*, 38(3), 835–40.

[21] Castenholz, R. W. (1988). Culturing methods for cyanobacteria. *Methods in Enzymology*, 167, 68–93.

[22] Chadwick, J. (1958). *The decipherment of linear B*. Cambridge Eng.: University Press.

[23] Chen, A. H., Lubkowicz, D., Yeong, V., Chang, R. L., & Silver, P. A. (2015). Transplantability of a circadian clock to a noncircadian organism. *Sci Adv*, 1(5).

[24] Cheng, A. A., Ding, H., & Lu, T. K. (2014). Enhanced killing of antibiotic-resistant bacteria enabled by massively parallel combinatorial genetics. *Proc Natl Acad Sci U S A*, 111(34), 12462–7.

[25] Church, G. M., Gao, Y., & Kosuri, S. (2012). Next-generation digital information storage in dna. *Science*, 337(6102), 1628.

[26] Clackson, T. & Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196), 383–6.

[27] Clerico, E. M., Cassone, V. M., & Golden, S. S. (2009). Stability and lability of circadian period of gene expression in the cyanobacterium synechococcus elongatus. *Microbiology*, 155(Pt 2), 635–41.

[28] Clerico, E. M., Ditty, J. L., & Golden, S. S. (2007). Specialized techniques for site-directed mutagenesis in cyanobacteria. *Methods Mol Biol*, 362, 155–71.

[29] Cocco, S., Monasson, R., & Weigt, M. (2013). From principal component to direct coupling analysis of coevolution in proteins: low-eigenvalue modes are needed for structure prediction. *PLoS Comput Biol*, 9(8), e1003176.

[30] Colwell, L. J., Brenner, M. P., & Murray, A. W. (2014). Conservation weighting functions enable covariance analyses to detect functionally important amino acids. *PLoS One*, 9(11), e107723.

[31] Cunningham, B. C. & Wells, J. A. (1989). High-resolution epitope mapping of hgh-receptor interactions by alanine-scanning mutagenesis. *Science*, 244(4908), 1081–5.

[32] Daopin, S., Alber, T., Baase, W. A., Wozniak, J. A., & Matthews, B. W. (1991). Structural and thermodynamic analysis of the packing of two alpha-helices in bacteriophage t4 lysozyme. *J Mol Biol*, 221(2), 647–67.

[33] de Visser, J. A. & Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. *Nat Rev Genet*, 15(7), 480–90.

[34] DePristo, M. A., Weinreich, D. M., & Hartl, D. L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet*, 6(9), 678–87.

[35] Dill, K. A. (1997). Additivity principles in biochemistry. *J Biol Chem*, 272(2), 701–4.

[36] Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D., & Chan, H. S. (1995). Principles of protein folding–a perspective from simple exact models. *Protein Sci*, 4(4), 561–602.

[37] Dill, K. A. & MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*, 338(6110), 1042–6.

[38] Dobzhansky, T. (2013). Nothing in biology makes sense except in the light of evolution. *The american biology teacher*, 75(2), 87–91.

[39] Dowell, R. D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., Bernstein, D. A., Rolfe, P. A., Heisler, L. E., Chin, B., Nislow, C., Giaever, G., Phillips, P. C., Fink, G. R., Gifford, D. K., & Boone, C. (2010). Genotype to phenotype: a complex problem. *Science*, 328(5977), 469.

[40] Eisenberg, D., Weiss, R. M., & Terwilliger, T. C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Natl Acad Sci U S A*, 81(1), 140–4.

[41] Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M., & Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Phys Rev E Stat Nonlin Soft Matter Phys*, 87(1), 012707.

[42] Engler, C. & Marillonnet, S. (2013). Combinatorial dna assembly using golden gate cloning. *Methods Mol Biol*, 1073, 141–56.

[43] Esvelt, K. M., Carlson, J. C., & Liu, D. R. (2011). A system for the continuous directed evolution of biomolecules. *Nature*, 472(7344), 499–503.

[44] Feinauer, C., Skwark, M. J., Pagnani, A., & Aurell, E. (2014). Improving contact prediction along three dimensions. *PLoS Comput Biol*, 10(10), e1003847.

[45] Ferguson, A. D., Amezcua, C. A., Halabi, N. M., Chelliah, Y., Rosen, M. K., Ranganathan, R., & Deisenhofer, J. (2007). Signal transduction pathway of tonb-dependent transporters. *Proc Natl Acad Sci U S A*, 104(2), 513–8.

[46] Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O., & Weigt, M. (2016). Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. *Mol Biol Evol*, 33(1), 268–80.

[47] Figueroa, F., Salles, S., & Aguilera, J. (1997). Effects of solar radiation on photoinhibition and pigmentation in the red alga porphyra leucosticta. *Oceanographic Literature Review*, 10(44), 1174.

[48] Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C., & Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. *Nature*, 513(7516), 120–3.

[49] Fodor, A. A. & Aldrich, R. W. (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, 56(2), 211–21.

[50] Forsyth, C. M., Juan, V., Akamatsu, Y., DuBridge, R. B., Doan, M., Ivanov, A. V., Ma, Z., Polakoff, D., Razo, J., Wilson, K., & Powers, D. B. (2013). Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing. *MAbs*, 5(4), 523–32.

[51] Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D., & Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. *Nat Methods*, 7(9), 741–6.

[52] Fowler, D. M. & Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nat Methods*, 11(8), 801–7.

[53] Fraczkiewicz, R. & Braun, W. (1998). Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *Journal of Computational Chemistry*, 19(3), 319–333.

[54] Fujino, Y., Fujita, R., Wada, K., Fujishige, K., Kanamori, T., Hunt, L., Shimizu, Y., & Ueda, T. (2012). Robust in vitro affinity maturation strategy based on interface-focused high-throughput mutational scanning. *Biochem Biophys Res Commun*, 428(3), 395–400.

[55] Gerhart, J. & Kirschner, M. (2007). The theory of facilitated variation. *Proc Natl Acad Sci U S A*, 104 Suppl 1, 8582–9.

[56] Gobel, U., Sander, C., Schneider, R., & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins*, 18(4), 309–17.

[57] Gonze, D., Roussel, M. R., & Goldbeter, A. (2002). A model for the enhancement of fitness in cyanobacteria based on resonance of a circadian oscillator with the external light-dark cycle. *J Theor Biol*, 214(4), 577–97.

[58] Gould, S. J. & Lewontin, R. C. (1979). The spandrels of san marco and the panglossian paradigm: a critique of the adaptationist programme. *Proc R Soc Lond B Biol Sci*, 205(1161), 581–98.

[59] Gregoret, L. M. & Sauer, R. T. (1993). Additivity of mutant effects assessed by binomial mutagenesis. *Proc Natl Acad Sci U S A*, 90(9), 4246–50.

[60] Gregoret, L. M. & Sauer, R. T. (1998). Tolerance of a protein helix to multiple alanine and valine substitutions. *Fold Des*, 3(2), 119–26.

[61] Grutter, M. G., Hawkes, R. B., & Matthews, B. W. (1979). Molecular basis of thermostability in the lysozyme from bacteriophage t4. *Nature*, 277(5698), 667–9.

[62] Halabi, N., Rivoire, O., Leibler, S., & Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4), 774–86.

[63] Harms, M. J. & Thornton, J. W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat Rev Genet*, 14(8), 559–71.

[64] Hartman, J. L. t., Garvik, B., & Hartwell, L. (2001). Principles for the buffering of genetic variation. *Science*, 291(5506), 1001–4.

[65] Hatley, M. E., Lockless, S. W., Gibson, S. K., Gilman, A. G., & Ranganathan, R. (2003). Allosteric determinants in guanine nucleotide-binding proteins. *Proc Natl Acad Sci U S A*, 100(24), 14445–50.

[66] Hemery, M. & Rivoire, O. (2015). Evolution of sparsity and modularity in a model of protein allostery. *Phys Rev E Stat Nonlin Soft Matter Phys*, 91(4), 042704.

[67] Hidalgo, P. & MacKinnon, R. (1995). Revealing the architecture of a k+ channel pore through mutant cycles with a peptide inhibitor. *Science*, 268(5208), 307–10.

[68] Hietpas, R. T., Jensen, J. D., & Bolon, D. N. (2011). Experimental illumination of a fitness landscape. *Proc Natl Acad Sci U S A*, 108(19), 7896–901.

[69] Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C., & Marks, D. S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7), 1607–21.

[70] Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Springer, M., Sander, C., & Marks, D. S. (2015a). Quantification of the effect of mutations using a global probability model of natural sequence variation. *ArXiv e-prints*, 1510.04612.

[71] Hopf, T. A., Morinaga, S., Ihara, S., Touhara, K., Marks, D. S., & Benton, R. (2015b). Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nat Commun*, 6, 6077.

[72] Hopf, T. A., Scharfe, C. P., Rodrigues, J. P., Green, A. G., Kohlbacher, O., Sander, C., Bonvin, A. M., & Marks, D. S. (2014). Sequence co-evolution gives 3d contacts and structures of protein complexes. *Elife*, 3.

[73] Horovitz, A. (1987). Non-additivity in protein-protein interactions. *J Mol Biol*, 196(3), 733–5.

[74] Horovitz, A. & Fersht, A. R. (1990). Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins. *J Mol Biol*, 214(3), 613–7.

[75] Horovitz, A., Serrano, L., Avron, B., Bycroft, M., & Fersht, A. R. (1990). Strength and co-operativity of contributions of surface salt bridges to protein stability. *J Mol Biol*, 216(4), 1031–44.

[76] Hoskisson, P. A. & Hobbs, G. (2005). Continuous culture–making a comeback? *Microbiology*, 151(Pt 10), 3153–9.

[77] Howell, E. E., Villafranca, J. E., Warren, M. S., Oatley, S. J., & Kraut, J. (1986). Functional role of aspartic acid-27 in dihydrofolate reductase revealed by mutagenesis. *Science*, 231(4742), 1123–8.

[78] Ishiura, M., Kutsuna, S., Aoki, S., Iwasaki, H., Andersson, C. R., Tanabe, A., Golden, S. S., Johnson, C. H., & Kondo, T. (1998). Expression of a gene cluster kaiabc as a circadian feedback process in cyanobacteria. *Science*, 281(5382), 1519–23.

[79] Jain, I. H., Vijayan, V., & O'Shea, E. K. (2012). Spatial ordering of chromosomes enhances the fidelity of chromosome partitioning in cyanobacteria. *Proc Natl Acad Sci U S A*, 109(34), 13638–43.

[80] Jain, P. C. & Varadarajan, R. (2014). A rapid, efficient, and economical inverse polymerase chain reaction-based method for generating a site saturation mutant library. *Anal Biochem*, 449, 90–8.

[81] Kamisetty, H., Ovchinnikov, S., & Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A*, 110(39), 15674–9.

[82] Kashtan, N. & Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci U S A*, 102(39), 13773–8.

[83] Kim, I., Miller, C. R., Young, D. L., & Fields, S. (2013). High-throughput analysis of in vivo protein stability. *Mol Cell Proteomics*, 12(11), 3370–8.

[84] Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S., & Shendure, J. (2015). Massively parallel single-amino-acid mutagenesis. *Nat Methods*, 12(3), 203–6, 4 p following 206.

[85] Kondo, T., Tsinoremas, N. F., Golden, S. S., Johnson, C. H., Kutsuna, S., & Ishiura, M. (1994). Circadian clock mutants of cyanobacteria. *Science*, 266(5188), 1233–6.

[86] Kosuri, S., Eroshenko, N., Leproust, E. M., Super, M., Way, J., Li, J. B., & Church, G. M. (2010). Scalable gene synthesis by selective amplification of dna pools from high-fidelity microchips. *Nat Biotechnol*, 28(12), 1295–9.

[87] Kryazhimskiy, S., Rice, D. P., Jerison, E. R., & Desai, M. M. (2014). Microbial evolution. global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science*, 344(6191), 1519–22.

[88] Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649), 1364–8.

[89] Lapedes, A. S., Giraud, B. G., Liu, L., & Stormo, G. D. (1999). Correlated mutations in models of protein sequences: phylogenetic and structural effects. *Lecture Notes-Monograph Series*, (pp. 236–256).

[90] Le Magnen, C., Dutta, A., & Abate-Shen, C. (2016). Optimizing mouse models for precision cancer prevention. *Nat Rev Cancer*, 16(3), 187–96.

[91] Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nat Rev Genet*, 14(3), 168–78.

[92] Leibler, S. & Kussell, E. (2010). Individual histories and selection in heterogeneous populations. *Proc Natl Acad Sci U S A*, 107(29), 13183–8.

[93] Liberles, D. A., Teichmann, S. A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L. J., de Koning, A. P., Dokholyan, N. V., Echave, J., Elofsson, A., Gerloff, D. L., Goldstein, R. A., Grahnen, J. A., Holder, M. T., Lakner, C., Lartillot, N., Lovell, S. C., Naylor, G., Perica, T., Pollock, D. D., Pupko, T., Regan, L., Roger, A., Rubinstein, N., Shakhnovich, E., Sjolander, K., Sunyaev, S., Teufel, A. I., Thorne, J. L., Thornton, J. W., Weinreich, D. M., & Whelan, S. (2012). The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci*, 21(6), 769–85.

[94] LiCata, V. J. & Ackers, G. K. (1995). Long-range, small magnitude nonadditivity of mutational effects in proteins. *Biochemistry*, 34(10), 3133–9.

[95] Lim, W. A., Farruggio, D. C., & Sauer, R. T. (1992). Structural and energetic consequences of disruptive mutations in a protein core. *Biochemistry*, 31(17), 4324–33.

[96] Lim, W. A. & Sauer, R. T. (1989). Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature*, 339(6219), 31–6.

[97] Lim, W. A. & Sauer, R. T. (1991). The role of internal packing interactions in determining the structure and stability of a protein. *J Mol Biol*, 219(2), 359–76.

[98] Liu, Y., Tsinoremas, N. F., Golden, S. S., Kondo, T., & Johnson, C. H. (1996). Circadian expression of genes involved in the purine biosynthetic pathway of the cyanobacterium synechococcus sp. strain pcc 7942. *Mol Microbiol*, 20(5), 1071–81.

[99] Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438), 295–9.

[100] Mackay, T. F. (2014). Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat Rev Genet*, 15(1), 22–33.

[101] Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S., & Miller, J. H. (1994). Genetic studies of the lac repressor. xiv. analysis of 4000 altered escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J Mol Biol*, 240(5), 421–33.

[102] Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., & Sander, C. (2011). Protein 3d structure computed from evolutionary sequence variation. *PLoS One*, 6(12), e28766.

[103] Marks, D. S., Hopf, T. A., & Sander, C. (2012). Protein structure prediction from sequence variation. *Nat Biotechnol*, 30(11), 1072–80.

[104] Marqusee, S. & Baldwin, R. L. (1987). Helix stabilization by glu-...lys+ salt bridges in short peptides of de novo design. *Proc Natl Acad Sci U S A*, 84(24), 8898–902.

[105] McLaughlin, R. N., J., Poelwijk, F. J., Raman, A., Gosal, W. S., & Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. *Nature*, 491(7422), 138–42.

[106] Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R., & Fields, S. (2013). Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly(a)-binding protein. *RNA*, 19(11), 1537–51.

[107] Monod, J. (1950). La technique de culture continue, theorie et applications. *Annales d'Institute Pasteur*, 79, 390–410.

[108] Morcos, F., Jana, B., Hwa, T., & Onuchic, J. N. (2013). Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc Natl Acad Sci U S A*, 110(51), 20533–8.

[109] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*, 108(49), E1293–301.

[110] Munoz, V. & Serrano, L. (1994). Elucidating the folding problem of helical peptides using empirical parameters. *Nat Struct Biol*, 1(6), 399–409.

[111] Nadis, S. (2003). Spoof nobels take researchers for a ride. *Nature*, 425(6958), 550.

[112] Nakajima, M., Imai, K., Ito, H., Nishiwaki, T., Murayama, Y., Iwasaki, H., Oyama, T., & Kondo, T. (2005). Reconstitution of circadian oscillation of cyanobacterial kaic phosphorylation in vitro. *Science*, 308(5720), 414–5.

[113] Ng, P. C. & Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet*, 7, 61–80.

[114] Nguyen, L. V., Nelson, E. J., Vengurlekar, A., Zhang, R., White, K. I., Salinas, V., & Johnson, T. T. (2014). Model-based design and analysis of a reconfigurable continuous-culture bioreactor.

[115] Novick, A. & Szilard, L. (1950). Description of the chemostat. *Science*, 112(2920), 715–6.

[116] Novinec, M., Korenc, M., Caflisch, A., Ranganathan, R., Lenarcic, B., & Baici, A. (2014). A novel allosteric mechanism in the cysteine peptidase cathepsin k discovered by computational methods. *Nat Commun*, 5, 3287.

[117] Ollikainen, N. & Kortemme, T. (2013). Computational protein design quantifies structural constraints on amino acid covariation. *PLoS Comput Biol*, 9(11), e1003313.

[118] Olson, C. A., Wu, N. C., & Sun, R. (2014). A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol*, 24(22), 2643–51.

[119] Onai, K., Morishita, M., Itoh, S., Okamoto, K., & Ishiura, M. (2004). Circadian rhythms in the thermophilic cyanobacterium thermosynechococcus elongatus: compensation of period length over a wide temperature range. *J Bacteriol*, 186(15), 4972–7.

[120] O'Neil, K. T., Wolfe, H. R., J., Erickson-Viitanen, S., & DeGrado, W. F. (1987). Fluorescence properties of calmodulin-binding peptides reflect alpha-helical periodicity. *Science*, 236(4807), 1454–6.

[121] O'Shea, E. K., Rutkowski, R., & Kim, P. S. (1989). Evidence that the leucine zipper is a coiled coil. *Science*, 243(4890), 538–42.

[122] Ouyang, Y., Andersson, C. R., Kondo, T., Golden, S. S., & Johnson, C. H. (1998). Resonating circadian clocks enhance fitness in cyanobacteria. *Proc Natl Acad Sci U S A*, 95(15), 8660–4.

[123] Ovchinnikov, S., Kamisetty, H., & Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife*, 3, e02030.

[124] Packer, M. S. & Liu, D. R. (2015). Methods for the directed evolution of proteins. *Nat Rev Genet*, 16(7), 379–94.

[125] Pakula, A. A. & Sauer, R. T. (1989). Genetic analysis of protein stability and function. *Annu Rev Genet*, 23, 289–310.

[126] Pazos, F., Helmer-Citterich, M., Ausiello, G., & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J Mol Biol*, 271(4), 511–23.

[127] Pei, J. & Grishin, N. V. (2014). Promals3d: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol Biol*, 1079, 263–71.

[128] Penel, S., Morrison, R. G., Mortishire-Smith, R. J., & Doig, A. J. (1999). Periodicity in alpha-helix lengths and c-capping preferences. *J Mol Biol*, 293(5), 1211–9.

[129] Perutz, M. F. & Fermi, G. (1988). Stereochemistry of salt-bridge formation in alpha-helices and beta-strands. *Proteins*, 4(4), 294–5.

[130] Perutz, M. F., Kendrew, J. C., & Watson, H. C. (1965). Structure and function of haemoglobin .2. some relations between polypeptide chain configuration and amino acid sequence. *Journal of Molecular Biology*, 13(3), 669–78.

[131] Poelwijk, F. J., de Vos, M. G., & Tans, S. J. (2011). Tradeoffs and optimality in the evolution of gene regulation. *Cell*, 146(3), 462–70.

[132] Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M., & Tans, S. J. (2007). Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126), 383–6.

[133] Poelwijk, F. J., Krishna, V., & Ranganathan, R. (2015). The context-dependence of mutations: a linkage of formalisms. *ArXiv e-prints*, 1502.00726.

[134] Pollock, D. D., Taylor, W. R., & Goldman, N. (1999). Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol*, 287(1), 187–98.

[135] Procko, E., Hedman, R., Hamilton, K., Seetharaman, J., Fleishman, S. J., Su, M., Aramini, J., Kornhaber, G., Hunt, J. F., Tong, L., Montelione, G. T., & Baker, D. (2013). Computational design of a protein-based enzyme inhibitor. *J Mol Biol*, 425(18), 3563–75.

[136] Ranganathan, R., Lewis, J. H., & MacKinnon, R. (1996). Spatial localization of the k+ channel selectivity filter by mutant cycle-based structure analysis. *Neuron*, 16(1), 131–9.

[137] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586), 1551–5.

[138] Rennell, D., Bouvier, S. E., Hardy, L. W., & Poteete, A. R. (1991). Systematic mutation of bacteriophage t4 lysozyme. *J Mol Biol*, 222(1), 67–88.

[139] Reynolds, K. A., McLaughlin, R. N., & Ranganathan, R. (2011). Hot spots for allosteric regulation on protein surfaces. *Cell*, 147(7), 1564–75.

[140] Reynolds, K. A., Russ, W. P., Socolich, M., & Ranganathan, R. (2013). Evolution-based design of proteins. *Methods Enzymol*, 523, 213–35.

[141] Rippka, R. (1988). Isolation and purification of cyanobacteria. *Methods Enzymol*, 167, 3–27.

[142] Rivoire, O. (2013). Elements of coevolution in biological sequences. *Phys Rev Lett*, 110(17), 178102.

[143] Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D., & Bolon, D. N. (2013). Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J Mol Biol*, 425(8), 1363–77.

[144] Roussel, M. R., Gonze, D., & Goldbeter, A. (2000). Modeling the differential fitness of cyanobacterial strains whose circadian oscillators have different free-running periods: comparing the mutual inhibition and substrate depletion hypotheses. *J Theor Biol*, 205(2), 321–40.

[145] Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B., & Ranganathan, R. (2005). Natural-like function in artificial ww domains. *Nature*, 437(7058), 579–83.

[146] Russ, W. P. & Ranganathan, R. (2002). Knowledge-based potential functions in protein design. *Curr Opin Struct Biol*, 12(4), 447–52.

[147] Rust, M. J., Golden, S. S., & O'Shea, E. K. (2011). Light-driven changes in energy metabolism directly entrain the cyanobacterial circadian oscillator. *Science*, 331(6014), 220–3.

[148] Sadovsky, E. & Yifrach, O. (2007). Principles underlying energetic coupling along an allosteric communication trajectory of a voltage-activated k+ channel. *Proc Natl Acad Sci U S A*, 104(50), 19813–8.

[149] Saldanha, A. J., Brauer, M. J., & Botstein, D. (2004). Nutritional homeostasis in batch and steady-state culture of yeast. *Mol Biol Cell*, 15(9), 4089–104.

[150] Sandberg, W. S. & Terwilliger, T. C. (1989). Influence of interior packing and hydrophobicity on the stability of a protein. *Science*, 245(4913), 54–7.

[151] Sandberg, W. S. & Terwilliger, T. C. (1991). Energetics of repacking a protein interior. *Proc Natl Acad Sci U S A*, 88(5), 1706–10.

[152] Sandberg, W. S. & Terwilliger, T. C. (1993). Engineering multiple properties of a protein by combinatorial mutagenesis. *Proc Natl Acad Sci U S A*, 90(18), 8367–71.

[153] Schlinkmann, K. M., Honegger, A., Tureci, E., Robison, K. E., Lipovsek, D., & Pluckthun, A. (2012). Critical features for biosynthesis, stability, and functionality of a g protein-coupled receptor uncovered by all-versus-all mutations. *Proc Natl Acad Sci U S A*, 109(25), 9810–5.

[154] Schreiber, G. & Fersht, A. R. (1995). Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J Mol Biol*, 248(2), 478–86.

[155] Serrano, L., Bycroft, M., & Fersht, A. R. (1991). Aromatic-aromatic interactions and protein stability. investigation by double-mutant cycles. *J Mol Biol*, 218(2), 465–75.

[156] Serrano, L., Horovitz, A., Avron, B., Bycroft, M., & Fersht, A. R. (1990). Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. *Biochemistry*, 29(40), 9343–52.

[157] Shulman, A. I., Larson, C., Mangelsdorf, D. J., & Ranganathan, R. (2004). Structural determinants of allosteric ligand activation in rxr heterodimers. *Cell*, 116(3), 417–29.

[158] Skerker, J. M., Perchuk, B. S., Siryaporn, A., Lubin, E. A., Ashenberg, O., Goulian, M., & Laub, M. T. (2008). Rewiring the specificity of two-component signal transduction systems. *Cell*, 133(6), 1043–54.

[159] Smock, R. G., Rivoire, O., Russ, W. P., Swain, J. F., Leibler, S., Ranganathan, R., & Gierasch, L. M. (2010). An interdomain sector mediating allostery in hsp70 molecular chaperones. *Mol Syst Biol*, 6, 414.

[160] Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H., & Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. *Nature*, 437(7058), 512–8.

[161] Starita, L. M., Pruneda, J. N., Lo, R. S., Fowler, D. M., Kim, H. J., Hiatt, J. B., Shendure, J., Brzovic, P. S., Fields, S., & Klevit, R. E. (2013). Activity-enhancing mutations in an e3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc Natl Acad Sci U S A*, 110(14), E1263–72.

[162] Stiffler, M. A., Chen, J. R., Grantcharova, V. P., Lei, Y., Fuchs, D., Allen, J. E., Zaslavskaia, L. A., & MacBeath, G. (2007). Pdz domain binding selectivity is optimized across the mouse proteome. *Science*, 317(5836), 364–9.

[163] Stiffler, M. A., Grantcharova, V. P., Sevecka, M., & MacBeath, G. (2006). Uncovering quantitative protein interaction networks for mouse pdz domains using protein microarrays. *J Am Chem Soc*, 128(17), 5913–22.

[164] Stiffler, M. A., Hekstra, D. R., & Ranganathan, R. (2015). Evolvability as a function of purifying selection in tem-1 beta-lactamase. *Cell*, 160(5), 882–92.

[165] Suel, G. M., Lockless, S. W., Wall, M. A., & Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol*, 10(1), 59–69.

[166] Sundaralingam, M., Drendel, W., & Greaser, M. (1985). Stabilization of the long central helix of troponin c by intrahelical salt bridges between charged amino acid side chains. *Proc Natl Acad Sci U S A*, 82(23), 7944–7.

[167] Sundaralingam, M., Sekharudu, Y. C., Yathindra, N., & Ravichandran, V. (1987). Ion pairs in alpha helices. *Proteins*, 2(1), 64–71.

[168] Tan, C., Marguet, P., & You, L. (2009). Emergent bistability by a growth-modulating positive feedback circuit. *Nat Chem Biol*, 5(11), 842–8.

[169] Tesileanu, T., Colwell, L. J., & Leibler, S. (2015). Protein sectors: statistical coupling analysis versus conservation. *PLoS Comput Biol*, 11(2), e1004091.

[170] Thyagarajan, B. & Bloom, J. D. (2014). The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *Elife*, 3.

[171] Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L., & Baker, D. (2013). Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*, 501(7466), 212–6.

[172] Toprak, E., Veres, A., Michel, J. B., Chait, R., Hartl, D. L., & Kishony, R. (2012). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat Genet*, 44(1), 101–5.

[173] Traxlmayr, M. W., Hasenhindl, C., Hackl, M., Stadlmayr, G., Rybka, J. D., Borth, N., Grillari, J., Ruker, F., & Obinger, C. (2012). Construction of a stability landscape of the ch3 domain of human igg1 by combining directed evolution with high throughput sequencing. *J Mol Biol*, 423(3), 397–412.

[174] Vijayan, V., Zuzow, R., & O'Shea, E. K. (2009). Oscillations in supercoiling drive circadian gene expression in cyanobacteria. *Proc Natl Acad Sci U S A*, 106(52), 22564–8.

[175] Wagenaar, T. R., Ma, L., Roscoe, B., Park, S. M., Bolon, D. N., & Green, M. R. (2014). Resistance to vemurafenib resulting from a novel mutation in the brafv600e kinase domain. *Pigment Cell Melanoma Res*, 27(1), 124–33.

[176] Wang, X., Fu, A. Q., McNerney, M. E., & White, K. P. (2014). Widespread genetic epistasis among cancer genes. *Nat Commun*, 5, 4828.

[177] Watson, J. D. & Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356), 737–8.

[178] Wei, W. H., Hemani, G., & Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nat Rev Genet*, 15(11), 722–33.

[179] Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., & Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A*, 106(1), 67–72.

[180] Wells, J. A. (1990). Additivity of mutational effects in proteins. *Biochemistry*, 29(37), 8509–17.

[181] Whitehead, T. A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S. J., De Mattos, C., Myers, C. A., Kamisetty, H., Blair, P., Wilson, I. A., & Baker, D. (2012). Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat Biotechnol*, 30(6), 543–8.

[182] Woelfle, M. A., Ouyang, Y., Phanvijhitsiri, K., & Johnson, C. H. (2004). The adaptive value of circadian clocks: an experimental assessment in cyanobacteria. *Curr Biol*, 14(16), 1481–6.

[183] Wu, C. W., Sanborn, T. J., Huang, K., Zuckermann, R. N., & Barron, A. E. (2001). Peptoid oligomers with alpha-chiral, aromatic side chains: sequence requirements for the formation of stable peptoid helices. *J Am Chem Soc*, 123(28), 6778–84.

[184] Wu, N. C., Young, A. P., Dandekar, S., Wijersuriya, H., Al-Mawsawi, L. Q., Wu, T. T., & Sun, R. (2013). Systematic identification of h274y compensatory mutations in influenza a virus neuraminidase by high-throughput screening. *J Virol*, 87(2), 1193–9.

[185] Xiong, H., Buckwalter, B. L., Shieh, H. M., & Hecht, M. H. (1995). Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proc Natl Acad Sci U S A*, 92(14), 6349–53.

[186] Zarrinpar, A., Park, S. H., & Lim, W. A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. *Nature*, 426(6967), 676–80.

[187] Zhang, H., Skinner, M. M., Sandberg, W. S., Wang, A. H., & Terwilliger, T. C. (1996). Context dependence of mutational effects in a protein: the crystal structures of the v35i, i47v and v35i/i47v gene v protein core mutants. *J Mol Biol*, 259(1), 148–59.

[188] Zhang, Y., Yeh, S., Appleton, B. A., Held, H. A., Kausalya, P. J., Phua, D. C., Wong, W. L., Lasky, L. A., Wiesmann, C., Hunziker, W., & Sidhu, S. S. (2006). Convergent and divergent ligand specificity among pdz domains of the lap and zonula occludens (zo) families. *J Biol Chem*, 281(31), 22299–311.

[189] Zuckerkandl, E. & Pauling, L. (1965a). Evolutionary divergence and convergence in proteins. *Evolving genes and proteins*, 97, 97–166.

[190] Zuckerkandl, E. & Pauling, L. (1965b). Molecules as documents of evolutionary history. *J Theor Biol*, 8(2), 357–66.

[191] Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*, 109(4), 1193–8.

This thesis was typeset with LaTeX, using a template from Jordan Suchow under the AGPL license (github.com/suchow/Dissertate)