Computational Structure-based Modeling and Analysis with Application to Rational and Evolutionary Molecular Engineering

by

Kathryn Anne Armstrong

Submitted to the Biological Engineering Division
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biological Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2007

© Massachusetts Institute of Technology 2007. All rights reserved.

Author .....
Biological Engineering Division
May 25, 2007

Certified by .....
Bruce Tidor
Professor of Biological Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Alan J. Grodzinsky
Chairman, Department Committee on Graduate Students

Thesis Committee

Accepted by .....

Catherine L. Drennan

Associate Professor of Chemistry

Chairman of Thesis Committee

Accepted by .....

Bruce Tidor

Professor of Biological Engineering and Computer Science

Thesis Supervisor

Accepted by .....

K. Dane Wittrup

J. R. Mares Professor of Chemical Engineering and Bioengineering

Thesis Committee Member

Accepted by .....

John M. Essigmann

Professor of Chemistry, Toxicology, and Biological Engineering

Thesis Committee Member

Computational Structure-based Modeling and Analysis with Application to Rational and Evolutionary Molecular Engineering

by

Kathryn Anne Armstrong

Submitted to the Biological Engineering Division
on May 25, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biological Engineering

Abstract

The design and development of new proteins and small molecules has considerable practical application in medicine, industry, and basic science. Frequently, progress in this area is made by altering an existing small molecule or protein for new function. This thesis presents methods for the analysis and design of rationally and evolutionarily designed molecules and focuses on applying these methods to make protein and small molecule changes more strategically. First, electrostatic analysis of a series of small molecule neuraminidase inhibitors was used to demonstrate that charge optimization improves the electrostatic component of the binding free energy, despite changes in binding mode and discrete chemical constraints. Additionally, chemical changes suggested by charge optimization frequently corresponded to tighter-binding inhibitors, indicating that this technique would be useful for the design of future inhibitors. Second, computational sequence and structure analysis were used to study the PDZ3–CRIPT binding interaction and a method for sequence analysis was developed to locate residues important for binding specificity. Third, computational analysis of the horseradish peroxidase active site suggested five positions as candidates for mutation, and further studies of new mutant enzymes led to ideas for the improvement of computational enzyme design procedures. Finally, both computational protein design techniques and a model of the evolutionary process were used to study the efficiency of evolution as a tool for creating new proteins in the laboratory. We identified sequences that serve as better evolutionary starting points than others and provide a general framework for considering the impact of protein structure on the allowed sequence space and therefore on the challenges that each protein presents to evolutionary protein engineering procedures.

Thesis Supervisor: Bruce Tidor
Title: Professor of Biological Engineering and Computer Science

Acknowledgments

I would like to thank my advisor, Bruce Tidor, for providing helpful suggestions and direction for my work while also letting me pursue project ideas of my own. His focus on good scientific questions and his guidance in presentation and writing style have been powerful forces in my development as a scientist.

My thesis committee (Dane Wittrup, Cathy Drennan, and John Essigmann) has also provided numerous helpful comments and insights throughout my thesis research. In particular, I would like to thank them for their helpful experimental reality checks on my research and for suggesting possible experimental applications of my work.

I would also like to thank the current and past members of the Tidor laboratory for creating a stimulating working environment and for providing endless guidance and support through all of my thesis research.

Other collaborators have also provided interesting project ideas, helpful comments, and experimental testing that guided my work. I would like to thank Alan Cheng for his guidance on the neuraminidase project and for giving me a taste of the industrial research environment. Ken Kosik and Onur Sakarya provided the sequence alignments underpinning the PDZ project and performed other experimental work on that system. Eugene Antipov and Daša Lipovšek in the Klibanov and Wittrup laboratories performed the experimental work in the horseradish peroxidase project. I would also like to thank Mark Nelson for his suggestions and enthusiasm on the possible application of my work to directed evolution.

I thank the students and faculty of the Biological Engineering Division for providing helpful comments during my student seminar talks and in many other conversations. This creative and supportive research community has made my time in graduate school extremely enjoyable. In particular I thank Doug Lauffenburger for molding this innovative and friendly department and for his career guidance and support.

Finally, I would like to thank my parents, my brother, and my boyfriend Galen for their love, support and encouragement.

Contents

1 Introduction **8**

2 Optimal charges in lead progression: A structure-based neuraminidase case study **14**

2.1 Introduction 15

2.2 Methods 17

2.2.1 Structure preparation 17

2.2.2 General electrostatic calculations 18

2.2.3 Charge optimization 18

2.3 Results and Discussion 19

2.3.1 Electrostatic distance from the optimum 22

2.3.2 The neuraminidase series case study 24

2.3.3 Application of charge optimization in lead progression 29

2.4 Conclusion 30

3 Evolution of PDZ domain binding specificity in the animal synapse **33**

3.1 Introduction 34

3.2 Results 38

3.2.1 Atomic-level conservation of the binding region 38

3.2.2 Specificity-determining residues 44

3.2.3 Co-evolving residues within the PDZ domain 46

3.2.4 A comprehensive picture of PDZ-domain binding 47

3.3 Discussion and Conclusion 49

3.4 Methods 51
3.4.1 Sequence data collection 51
3.4.2 Generation of PDZ sequence alignments 51
3.4.3 Structure preparation and analysis 52
3.4.4 Energetic calculations 53
3.4.5 Mutual information metric 53

4 Re-engineering horseradish peroxidase using active-site focused combinatorial libraries 55

4.1 Introduction 56
4.2 Results 61
4.2.1 Selection of five active-site positions for randomization 61
4.2.2 Design of a combinatorial library 62
4.2.3 Selection results for exhaustive and computational libraries . . 64
4.2.4 Characterization of mutant CD8.02 69
4.2.5 Analysis of correlations in enzyme dynamics 71
4.3 Discussion and Conclusion 71
4.4 Methods 73
4.4.1 Structure preparation 73
4.4.2 Computational sequence and structure search 74
4.4.3 Substrate docking and binding energy calculations 74
4.4.4 Construction of the exhaustive and computationally-focused active-site libraries by saturation mutagenesis 75
4.4.5 Selection of enantioselective horseradish peroxidase variants using yeast-display and fluorescence-activated cell sorting (FACS) 77
4.4.6 Normal modes analysis 77

5 Computationally mapping sequence space to understand evolutionary protein engineering 79

5.1 Introduction 80
5.2 Results 84

5.2.1 Sequence-space graph topology 84
5.2.2 Error-prone PCR simulation 88
5.2.3 Genetic recombination simulation 96
5.3 Discussion 99
5.4 Methods 103
5.4.1 Evolution simulations 103
5.4.2 Sequence-space graph generation 105
5.4.3 Measurement of graph properties 107

6 General conclusions **108**

Chapter 1

Introduction

Humans have attempted to manipulate and enhance the chemicals found in the natural environment for hundreds of years in order to transform their existence. More recently, molecules have been created by making random modifications to existing molecules, and this process has been incredibly fruitful for both the design of new therapeutic drugs and for the design of new proteins and enzymes. However, as more difficult molecular engineering problems arise, such as for cancer and HIV/AIDS treatments, traditional design techniques are becoming more difficult [1, 2], and it is becoming more important to use structure-based molecular design. Computational molecular modeling has become an important tool for understanding molecular interactions and for guiding the design process, and this thesis makes contributions to both areas.

The process of molecular design has become more "rational" over the centuries as chemical and biological knowledge has improved. Plants were used for healing in the ancient world [3, 4] (Figure 1.1A) and it has become possible to isolate their active compounds. For example, the small molecule drugs penicillin and morphine are naturally occurring compounds found in mold and opium poppies [5, 6]. Recently humans have begun altering naturally-occurring molecules for new uses, and even making entirely novel molecules (Figure 1.1B). For example a small change was made to salicylic acid (found in willow bark) to create aspirin [7] and scientists frequently mutate proteins and enzymes in the laboratory by making changes to the DNA that encodes these molecules [8]. Sometimes this is done by accelerating the evolution process and selecting for mutations with some desired new property. This technique is called "protein engineering" or "directed evolution" [9]. Structure-based molecular engineering has become more common with the growth of the Protein Data Bank [10]

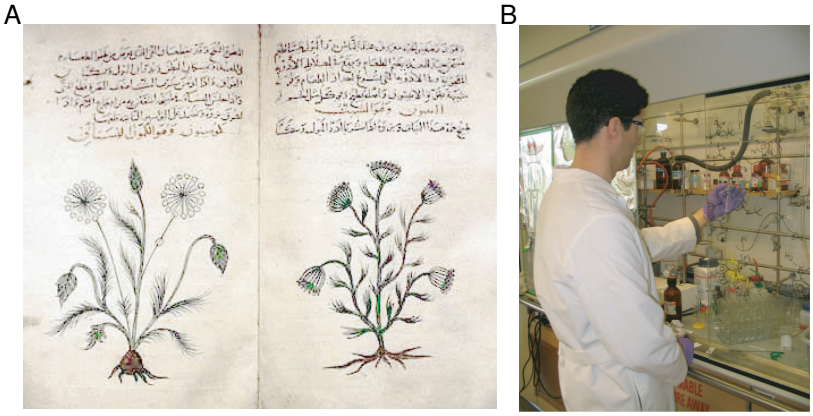

Figure 1.1: (A) Two pages from Discorides' *De Materia Medica* on cumin and dill [4]. (B) A chemist at MIT creating a fluorescent molecule in the laboratory.

and biochemical structural information has become an integral part of the molecular design process. For example, the antiretroviral drugs that inhibit HIV protease mimic the natural peptides that bind the viral enzyme [11] and many pharmaceutical and biotechnology companies use protein structures to guide the drug design process.

Increasing computational power and the development of computational molecular modeling techniques now allow molecular interactions to be studied in more detail. Scientists have been able to move beyond questions that are answerable experimentally. For example, when analyzing a binding interface, one can computationally separate the energetic contributions of individual atoms. One can also distinguish the properties of each atom that most strongly influence its contribution, and analysis of these details frequently leads to suggestions for modifications that can alter binding. Computational analysis of known molecules provides statistical information about preferred sequences and structures. This detailed understanding of molecular structures and the important factors for interaction will become more critical as traditional molecular design techniques face more difficult engineering problems. The ability to compute the energy of a molecular interaction leads naturally to the use of computational molecular modeling for molecular design.

This thesis presents several studies of molecular complexes, including both natural and man-made molecules, and explores how the information gained in these studies can be used to engineer new molecules. We studied two natural proteins and the

changes that have occurred at their binding interface over evolutionary time. We also worked with experimental collaborators to engineer several new enzymes and analyzed these new mutants computationally. Our improved understanding of plausible protein mutations and a model of the evolutionary process led us to study the efficiency of evolution as a tool for creating new proteins. Based on these models, we present new ideas for protein engineering. We also analyzed a series of small molecules bound to an enzyme active site and demonstrated that a method for molecular charge analysis can be used for small molecule engineering. A detailed analysis of the biochemical system is required in each case before considering molecular engineering, and full dissection of each result is important for further understanding and for improvement of the molecular engineering process.

Two theoretical models of biochemical systems make this detailed analysis possible. One is a description of molecules as mechanical systems and the other is an electrostatics model to speed the computation of the interactions between charged atoms. Molecular mechanics techniques describe molecules as collections of charged spherical masses (representing atoms) connected by springs (representing bonds). Careful parameterization of the spring constants between pairs, triplets, and quadruplets of atoms as well as the energetic interactions between non-connected atoms has produced several useful energetic forcefields [12–17]. These can be used to calculate the energy of a static atomic structure and even to model molecular dynamics using Newton's laws of motion. Biological molecules of interest typically exist in an aqueous environment, and modeling the numerous water molecules surrounding a given molecule is computationally very expensive. However, the behavior of this water has a significant impact on the interactions between charged atoms, called electrostatics. To model the influence of water on the interaction between charges in the system more efficiently, a second theoretical model called continuum electrostatics was developed [18, 19]. This model describes molecules as collections of charges inside a defined molecular shape, where inside the shape the dielectric constant is low and outside the shape the dielectric constant is high. The high dielectric region represents water and can also contain mobile ions to more accurately represent a biological environment.

The linearized Poisson–Boltzmann equation can be used to solve for the electrostatic potential in such a system.

To transfer these analyses of molecular systems into suggestions for the design of new molecules, we make chemical modifications computationally and compute the free energy of each change. Search algorithms are required to make structural changes and to compute new energies quickly, and two algorithms are used for this purpose in this thesis [20, 21]. In order to make chemical changes and new energy calculations efficiently, the "inverse-design" assumption is used to limit the number of sequences and structures to be searched. Instead of asking how each possible molecular modification will change the conformation of a protein or small molecule, one asks only for the impact of each molecular modification on the energy of the *existing* protein backbone structure or molecular scaffold. In the continuum electrostatic model, charge optimization techniques have been developed to evaluate the energetic impact of changes in molecular charge and to find the optimal set of charges for a given molecule [22, 23].

These models for molecular energetics and these methods for searching the space of new molecules were used in this thesis to computationally analyze and design new molecules. Four projects are presented that cover different topics and goals in the area of computational molecular modeling and analysis.

In Chapter 2 we analyzed the binding of eleven small molecule inhibitors of the enzyme neuraminidase with the goal of better understanding the role of electrostatics in molecular interaction. This enzyme is an important antiviral target for influenza and one of the molecules studied is the recently notorious tamiflu. Using the series of known binders, we asked retrospectively whether charge optimization improves the electrostatic portion of each molecule's binding affinity and whether these changes translated to improvement in the total binding affinity. Our results demonstrated that calculating the optimal molecular charges creates suggestions for successively tighter-binding molecules, indicating that this type of analysis may aid further design of new inhibitors.

In Chapter 3 we studied changes at the PDZ3-CRIPT binding interface across a series of organisms from sponge to human. This protein–protein complex is important

for the function of the synapse in the animal nervous system. The sponge does not have a nervous system, but proteins similar to PDZ3 and CRIPT were found in the recently-sequenced sponge genome, leading us to carefully study this binding interface and its changes over evolutionary time. We explored sequence and structure changes between organisms and between similar proteins to gain information about protein mutations that either alter or do not change binding function. This analysis suggested that these proteins can interact in the sponge, meaning that before they evolved into part of the synapse they could have interacted with some other, unknown function. We also pinpointed PDZ3 residues important for binding specificity.

In Chapter 4 we studied the enzyme horseradish peroxidase and known mutants of this enzyme to suggest strategies for computational enzyme design. Existing enzymes are typically very specific for their biologically-relevant substrates and there is commercial interest in developing enzymes with new substrate specificity. Horseradish peroxidase is widely produced for use as a commercial catalyst and the ability to modify this or any enzyme for a specific function would be useful for chemical synthesis. We computationally explored the sequences allowed in the enzyme active site and suggested five active site positions for mutagenesis. New mutants were selected by our collaborators in the Wittrup and Klibanov laboratories, and we studied these to analyze our earlier computational search. We also analyzed the overall dynamics of horseradish peroxidase to locate other possible fruitful sites for mutagenesis.

In Chapter 5 we modeled the process of directed evolution to evaluate the efficiency of this experimental sequence search method. Directed evolution mimics the evolution process in the laboratory, where DNA is mutated and recombined to create new sequences. The evolution is "directed" because new proteins or enzymes are selected in the laboratory from the pool of generated sequences. We used computational protein design to define sets of "allowed" sequences that might represent solutions to a protein engineering problem. We then simulated the directed evolution process, varying experimentally-tunable parameters, and asked how efficiently the allowed sequences were found. This work identified a non-intuitive relationship between the mutation rate and the number of rounds of gene replication and highlighted some

sequences that can serve as better evolutionary starting points than others.

Overall, this work presents several techniques for the analysis of biochemical systems and demonstrates the usefulness of this analysis in thoroughly understanding molecular systems. Understanding the molecular determinants of binding affinity and specificity is important for designing new molecules such as therapeutic drugs and enzymes for commercial applications, and computation is increasingly playing a role in this design process.

Chapter 2

Optimal charges in lead progression: A structure-based neuraminidase case study[1]

Abstract

Collective experience in structure-based lead progression has found electrostatic interactions to be more difficult to optimize than shape-based ones. A major reason for this is that the net electrostatic contribution observed includes a significant nonintuitive desolvation component in addition to the more intuitive intermolecular interaction component. To investigate whether knowledge of the ligand optimal charge distribution can facilitate more intuitive design of electrostatic interactions, we took a series of small-molecule influenza neuraminidase inhibitors with known protein cocrystal structures and calculated the difference between the optimal and actual charge distributions. This difference from the electrostatic optimum correlates with the calculated electrostatic contribution to binding ($r^2 = 0.94$) despite small changes in binding modes caused by chemical substitutions, suggesting that the optimal charge distribution is a useful design goal. Furthermore, detailed suggestions for chemical modification generated by this approach are in many cases consistent with observed improvements in binding affinity, and the method appears to be useful despite discrete chemical constraints. Taken together, these results suggest that charge optimization is useful in facilitating generation of compound ideas in lead optimization. Our results also provide insight into design of neuraminidase inhibitors.

[1]Reproduced with permission from:
Armstrong, K. A., Tidor, B. & Cheng, A. C. Optimal Charges in Lead Progression: A Structure-Based Neuraminidase Case Study *J. Med. Chem.*, 49 (8): 2470-2477, 2006.
Copyright 2006 Americal Chemical Society.

2.1 Introduction

Ligand-receptor affinity is affected by shape and charge complementarity as well as other factors including strain, entropy, titration effects, and conformational change. Collective structure-based drug design (SBDD) experience has found that molecular visualization and chemical intuition are often sufficient for successful optimization of shape-based hydrophobic and van der Waals interactions, while charge-based hydrogen-bond and salt-bridge interactions are more difficult to optimize in the same fashion. Davis and Teague [24] conclude from a large number of SBDD case studies that optimization of hydrogen bonds and salt bridges is unpredictable, and they advise medicinal chemists to focus on optimizing the more predictable hydrophobic interactions. They further point out that optimization of electrostatic interactions may be more of an issue with the advent of high-throughput screens, which preferentially select lipophillic compounds that can require addition of polar surface area for bioavailability. Thus, there is substantial need for approaches for discovering electrostatic enhancements to binding affinity.

The nonintuitive nature of electrostatics in ligand optimization is largely due to the net electrostatics being composed of ligand and receptor desolvation in addition to the intermolecular charge–charge interaction that can be "seen" in molecular visualization of the bound complex. While desolvation is difficult to visualize, being a difference in properties in the unbound and bound states, it can be calculated using continuum electrostatic methods such as those that solve the Poisson–Boltzmann equation [18, 19, 25–27]. Reasonably accurate affinity estimates with inclusion of desolvation can be calculated using these methods, despite the positional sensitivity noted in some docking studies [28]. In applying these methods, a library of possible analogues might be designed around a cocrystallized lead and then prioritized computationally before synthesis of a subset, thereby decreasing the number of compounds needing to be made in lead optimization. However, this is an iterative trial-and-error approach, and ideally one wants to address charge optimization in the same fashion as shape-based optimization, where one can see the ligand and binding pocket shapes and

design changes with moderate success.

Quantitative charge optimization techniques [22, 23] provide an approach to facilitate idea generation by computing the optimal charge distribution for a given scaffold taking into account the ligand and receptor desolvation as well as the intermolecular electrostatic interaction. The method has been used to design chemical modifications to improve charge interactions [29–31]. Designing synthetically feasible molecules that conform to these optimal charges can be difficult, because discrete chemical modifications will rarely produce the exact desired charges [31], and chemical modifications can result in shifts in the ligand binding mode as well as ligand and receptor conformations. The designed improvement in electrostatics must compete with possibly unfavorable changes to the other energetic components, which can result in unpredictable tradeoffs between electrostatics, van der Waals, entropy, and conformational change as compared to the starting co-complex structure used in the design analysis.

Here, we examine a series of known small molecule neuraminidase inhibitors with available co-complex structures in a simulated lead-optimization study to understand the behavior and utility of charge optimization in a drug discovery setting. We ask whether adjusting a molecular charge distribution such that it more closely matches the calculated optimum improves the electrostatic portion of the binding affinity and whether these changes translate to improvements in the experimental binding affinity. This is a nontrivial correspondence due to many factors, including the discrete constraints of chemistry and minor shifts in the ligand binding mode and receptor conformation.

Neuraminidase is an important antiviral target for influenza, with efforts to develop inhibitors starting in the late 1960s. It was only with the elucidation of the protein crystal structure in 1983 [32] and SBDD efforts over the following decade that inhibitors with therapeutically efficacious potencies were first developed. These efforts led to the marketing of the first approved drugs, oseltamivir and zanamivir, in 1998. We investigate the utility of charge-optimization within the context of neuraminidase and present analyses that may aid further optimization and design of new inhibitors.

2.2 Methods

2.2.1 Structure preparation

N9-subtype neuraminidase co-complex structures were identified through a BLAST search of proteins in the Protein Databank (PDB) [10]. All selected structures are atomic resolution structures with crystallographic resolutions in the 1.6–2.1 Å range. Eleven receptor–ligand complex structures were used as starting points for the neuraminidase calculations (PDB identifiers 1BJI, 1F8B, 1F8C, 1F8D, 1F8E, 1L7F, 1MWE, 1NNC, 2QWI, 2QWJ, 2QWK) [33–38].

Cocrystal structures were prepared using Pfizer in-house software that automates standard protein preparation tasks. Manual preparation or publicly available tools can also be used. We specify here our protein preparation procedure. Ends of protein regions with missing density were capped with either an N-terminal acetyl group or C-terminal amide group, and incomplete residues were automatically built using rotamer replacement. All cofactors were removed except the buried calcium ion that is about 11 Å from the binding site, which was included and assigned a charge state of $+2e$. Hydrogens were added to the molecules, titrated at neutral pH, and rotationally optimized in the presence of the cocrystallized ligand partner. Histidines were assigned the most favorable protonation and flip states, and glutamines and asparagines were assigned their most favorable flip states. Water molecules in the structure exposing more than 1 $Å^2$ of surface area to other waters or to solvent accessible regions were removed, and the remaining crystallographic water molecules were each rotated spherically about the oxygen atom to optimize hydrogen-bonding networks.

The charges and radii of all receptor atoms were assigned according to the CHARMM PARAM19 parameter set and all ligands were assigned CHARMM PARAM19 radii [12]. As a test of parameter sensitivity, parallel calculations were run with all charges and radii reassigned using the PARSE parameter set [39]. Reference charges for all ligands were determined using Gaussian 03 [40] and the RESP procedure [41, 42]. Ligands were first geometry-optimized in the gas phase using the

Hartree-Fock method with the 6-31G* basis set, and then their electrostatic potentials were calculated using the same level of theory. Partial atomic charges for all atom centers were then calculated using a two-stage RESP fit to these potentials. The overall charge used for each compound was zero, except for compounds **1** and **4**, which were each assigned -1e overall charge, and **3**, which was assigned a +1e overall charge [34].

2.2.2 General electrostatic calculations

The continuum electrostatic approximation was used to calculate electrostatic potentials. Calculations were made with a locally developed Poisson-Boltzmann solver (M. D. Altman and B.T., unpublished) that uses the finite-difference approach to solve the linearized Poisson-Boltzmann equation [43, 44] with a multigrid implementation [45, 46]. These potentials were calculated using a solvent dielectric of 80 with a bulk ionic strength of 0.145 M, a solute dielectric of 4, and a Stern layer of 2.0 Å [47]. A molecular surface generated with a 1.4-Å radius probe defined the dielectric boundary. The averaged results of 10 translations on a 129×129×129 cubic grid were used in all cases.

2.2.3 Charge optimization

The optimal charge distribution for each ligand was calculated by optimizing a variable point charge at each atom center of the ligand. Charge-optimization theory, described in detail elsewhere [22, 23, 29, 30, 48–50], enables calculation of the optimal charge distribution for a given ligand bound to a receptor. This distribution lies at the minimum electrostatic free energy of binding for those atom centers, which is computed by adding the unfavorable desolvation penalties of both the ligand and receptor and the favorable intermolecular interactions between the molecules. The vector of optimal charges is a function of the receptor charges, Q_R, the matrix of ligand–receptor interactions, C, and the ligand desolvation matrix, L:

$$Q_L^{opt} = -\frac{1}{2} L^{-1} C^T Q_R.$$

The individual partial atomic charges computed for each ligand were constrained using LOQO [51] to be no more than $1.0e$ in magnitude, except in the lead progression analysis, where no atomic charge constraints were used. Additionally, the total charge on each ligand was constrained to be no more than $2.0e$ in magnitude, except in the lead progression analysis, where no total charge constraints were used. These constraints limit the optimization to a reasonable charge space for medicinal chemistry. This theory has been applied to several ligand–receptor systems, such as barnase–barstar [29], chorismate mutase [30], and granulocyte colony stimulating factor [52], as well as protein kinase A [31], CDK2 [31], CDK5 [53], cathepsin B [54], PSD-95 PDZ domain [55], bovine trypsin [55], Gln-tRNA synthetase [56], and HIV-1 gp41 ectodomain [57].

2.3 Results and Discussion

For the set of N9-subtype neuraminidase inhibitors with known cocrystal structures, we calculated both the actual ligand atomic charges and the theoretically optimal ligand atomic charges that would lead to the most favorable electrostatic binding free energy for each ligand. Comparison of the reference and optimal charges suggests nonoptimal ligand atoms and functional groups whose modification or substitution may lead to improved potency. One way to represent the neuraminidase inhibitor binding pocket is to divide it into five regions [58] based on common inhibitor functional group interactions, as shown in Figure 2.1A. The carboxylate at position 2 makes salt bridges to three arginines (Arg118, Arg371, Arg292) in the S1 pocket and is conserved in all known inhibitors. The acetamide at position 5 making interactions in the S3 pocket is also conserved in most inhibitors, while the substituent at position 4 occupying the S2 pocket and the glycerol at position 6 occupying the S4 and S5 pockets can be substantially optimized, as can be seen in Figure 2.2.

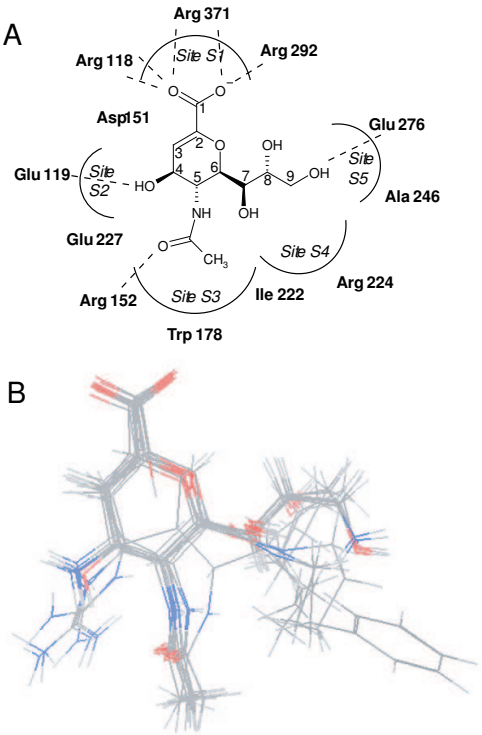

Figure 2.1: Common binding mode of neuraminidase inhibitors. (A) The neuraminidase active site broken up into five regions, following the precedence of Stoll et al. [58]. (B) Three-dimensional overlay of the 11 ligands used in this study based on crystal structure superposition. Main chain atoms of the protein backbone were superposed using MOE (version 2004.03; Chemical Computing Group, Montreal, Quebec, Canada).

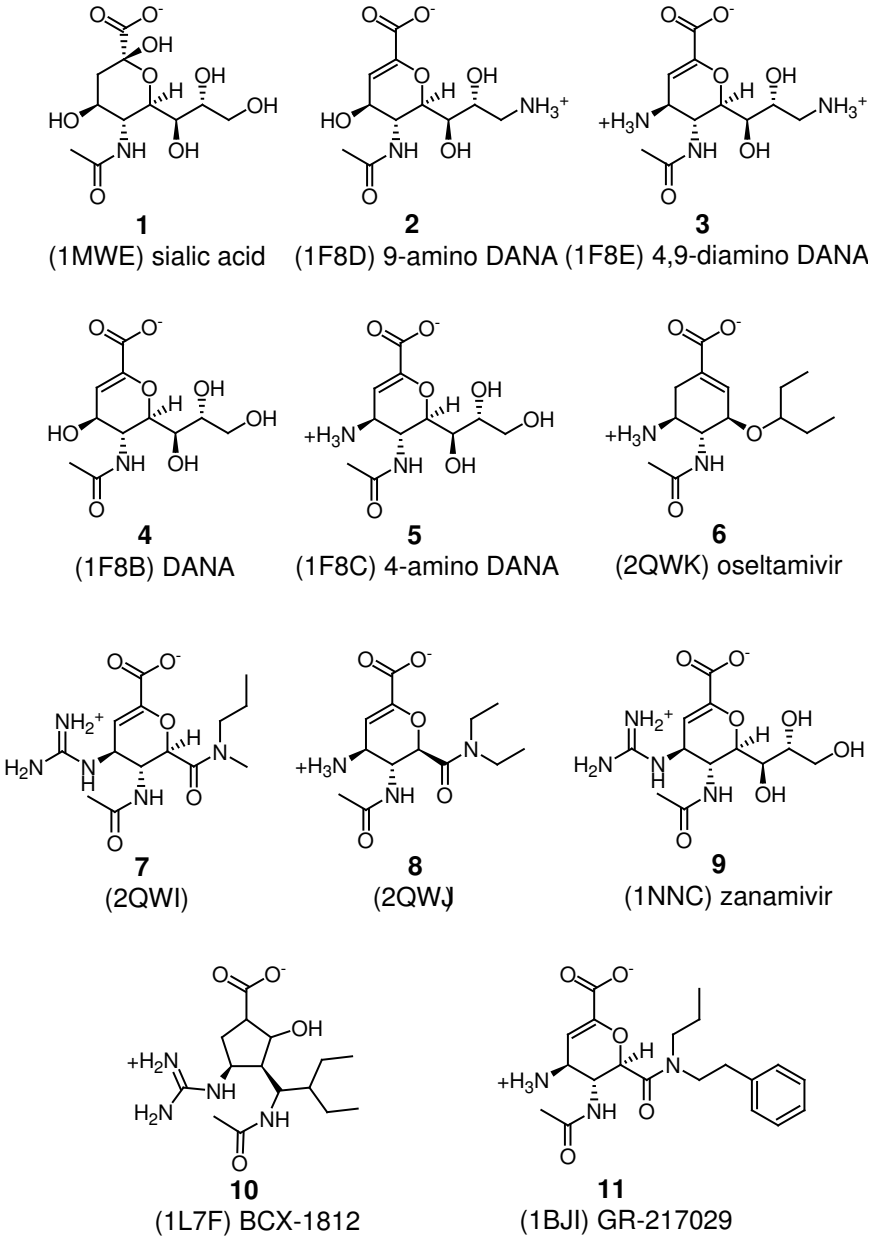

Figure 2.2: Chemical structures of neuraminidase N9-subtype small molecule ligands used in this study. The PDB IDs of the relevant cocrystal structure are given in parentheses, followed by applicable common names.

2.3.1 Electrostatic distance from the optimum

For each of the 11 ligands shown in Figure 2.2, we find that the electrostatic component of the computed binding free energy improves with chemical modifications that bring the charge distribution of each molecule closer to its optimum. This distance from the optimum, denoted $\Delta\Delta G^{opt}_{elec}$, is defined as the difference between ΔG^{ref}_{elec}, the electrostatic component of the free energy of binding the ligand with reference charges, and ΔG^{opt}_{elec}, the electrostatic component of the free energy of binding the ligand with optimal charges. Table 2.1 gives these values for each of the ligands. Experimentally measured binding affinities for the series of small molecule inhibitors were also collected from the literature [33, 34, 38, 58, 59] to facilitate comparisons and aid understanding of the tradeoffs being made in each active site between the electrostatics and other contributors to binding.

PDB ID	D^{wt}_L	D^{opt}_L	D_R	I^{wt}	I^{opt}	ΔG^{wt}_{elec}	ΔG^{opt}_{elec}	$\Delta\Delta G_{opt}$	ΔG_{expt}
1MWE	12.4	27.9	27.7	-19.0	-59.5	21.1	-4.0	-25.1	-4.1
1F8B	11.9	22.2	26.9	-15.1	-48.9	23.7	0.2	-23.5	-7.3
1F8C	16.8	24.9	27.5	-32.4	-54.5	11.9	-2.1	-14.0	-10.1
1F8D	14.2	23.8	26.4	-23.0	-50.9	17.5	-0.7	-18.2	-4.6
1F8E	25.0	25.7	27.5	-42.1	-55.4	10.3	-2.2	-12.5	-6.6
1BJI	15.8	24.7	28.3	-35.4	-53.3	8.6	-0.4	-9.0	-10.7
1L7F	16.3	29.0	28.1	-35.4	-61.6	9.1	-4.5	-13.6	-12.2
1NNC	19.5	29.7	30.8	-35.1	-63.3	15.2	-2.8	-18.0	-11.8
2QWI	17.6	25.9	27.3	-36.7	-55.9	8.1	-2.7	-10.8	-11.4
2QWJ	15.3	22.8	26.6	-35.1	-49.6	6.8	-0.3	-7.0	-9.0
2QWK	14.6	26.4	28.2	-37.0	-56.7	5.7	-2.2	-8.0	-11.8

Table 2.1: Electrostatic components computed using the CHARMM PARAM19 parameter set. All experimental values contain some reported error and some error related to our conversion using $\Delta G = -RTln(K_i$ or $IC_{50})$. D^{wt}_L, D^{opt}_L, and D_R are calculated desolvations of the wild-type ligand, the optimal ligand, and the receptor, respectively. I^{wt} and I^{opt} are the calculated electrostatic component of the interaction energy of the wild-type and optimal ligands with each receptor. ΔG^{wt}_{elec} and ΔG^{opt}_{elec} are the calculated electrostatic free energy of binding the wild-type and optimal ligands. $\Delta\Delta G_{opt}$ is the distance from the optimum as described earlier and ΔG_{expt} is the experimental binding affinity, converted into kcal/mol. The IC_{50}s for compounds 1MWE and 1L7F varied widely over different papers and trials, and so we have used a rounded average value for 1L7F and an approximate value of 1 mM for 1MWE, following the precedence of Taylor et al.

PDB ID	D_L^{wt}	D_L^{opt}	D_R	I^{wt}	I^{opt}	ΔG_{elec}^{wt}	ΔG_{elec}^{opt}	$\Delta\Delta G_{opt}$	ΔG_{expt}
1MWE	13.2	23.6	27.3	-12.3	-47.7	28.2	3.2	-25.0	-4.1
1F8B	12.3	21.7	26.6	-9.2	-43.8	29.7	4.4	-25.2	-7.3
1F8C	15.8	25.5	27.3	-31.1	-51.7	12.0	1.0	-10.9	-10.1
1F8D	13.9	21.4	27.5	-19.3	-43.3	22.1	5.5	-16.6	-4.6
1F8E	22.4	26.5	28.0	-42.6	-53.4	7.9	1.2	-6.7	-6.6
1BJI	16.5	24.5	29.8	-34.4	-50.0	11.9	4.4	-7.6	-10.7
1L7F	16.0	28.0	28.7	-32.5	-56.5	12.2	0.3	-11.9	-12.2
1NNC	18.2	29.6	32.3	-33.6	-59.8	16.9	2.1	-14.8	-11.8
2QWI	17.2	27.1	29.6	-34.3	-54.8	12.5	1.9	-10.5	-11.4
2QWJ	16.1	21.7	26.6	-32.9	-44.2	9.7	4.1	-5.6	-9.0
2QWK	14.6	25.8	28.3	-35.2	-52.3	7.6	1.7	-5.9	-11.8

Table 2.2: Electrostatic components computed using the PARSE parameter set. All experimental values contain some reported error and some error related to our conversion using $\Delta G = -RTln(K_i$ or $\mathrm{IC}_{50})$. D_L^{wt}, D_L^{opt}, and D_R are calculated desolvations of the wild-type ligand, the optimal ligand, and the receptor, respectively. I^{wt} and I^{opt} are the calculated electrostatic component of the interaction energy of the wild-type and optimal ligands with each receptor. ΔG_{elec}^{wt} and ΔG_{elec}^{opt} are the calculated electrostatic free energy of binding the wild-type and optimal ligands. $\Delta\Delta G_{opt}$ is the distance from the optimum as described earlier and ΔG_{expt} is the experimental binding affinity, converted into kcal/mol. The IC_{50}s for compounds 1MWE and 1L7F varied widely over different papers and trials, and so we have used a rounded average value for 1L7F and an approximate value of 1 mM for 1MWE, following the precedence of Taylor et al.

All 11 ligands have similar five- or six-membered ring scaffolds and bind in very similar modes, as shown in Figure 2.1B. Even though the optimal electrostatic free energies of binding vary in value between -4.5 and +0.2 kcal/mol, there is good correlation ($r^2 = 0.94$) between the distance from the optimum, $\Delta\Delta G_{elec}^{opt}$, and the reference electrostatic free energy contribution, ΔG_{elec}^{ref} (see Table 2.1 and Figure 2.3). These results from the full charge optimization suggest that the electrostatic gain in binding free energy as the charge distribution of the ligand moves closer to its optimum is significantly greater than the amount that the electrostatics of the optimum changes due to the new chemical structure and positioning in the active site. The sensitivity of these results to parameter sets has been tested by repeating all calculations using PARSE radii and partial atomic charges (Table 2.2). The trends from the PARSE results track closely with the CHARMM PARAM19 results.

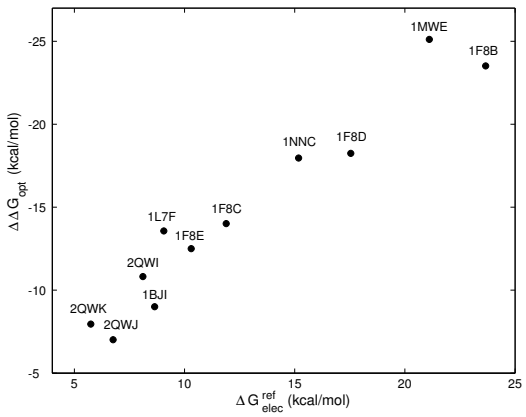

Figure 2.3: Electrostatic contributions to ligand binding relative to electrostatic differences from the optimum. Scatter plot of the differences from the optimal charge distribution, $\Delta\Delta G^{opt}_{elec}$, versus calculated electrostatic component of the binding affinity, ΔG^{ref}_{elec}, for the 11 ligands. Free energy terms are defined and discussed in the text. Positive values of ΔG^{ref}_{elec} represent unfavorable net contributions to binding affinity, while negative values represent favorable contributions.

2.3.2 The neuraminidase series case study

Full charge optimization analysis of the neuraminidase substrate sialic acid is presented in Figure 2.4A. The full optimization suggests substantial room for electrostatic improvement and indicates that sialic acid (compound **1**) is a promising starting point for inhibitor design given that its IC_{50} against neuraminidase is about 1 mM [33]. With 18 of the 39 atoms in compound **1** calculated to be greater than $0.5e$ from the optimal overall charge distribution, a natural question is where to begin applying chemical modifications in order to increase binding affinity. In lead optimization, chemical modification is often focused on a few R groups. To understand the utility of charge optimization in this setting, we focus on modifications at the position 2, 4, 7, 8, and 9 hydroxyl groups of compound **1**. The series of small molecules deriving from compound **1** increase in potency through compounds **2**, **3**, **4**, **5**, and **6**, following a simulated forward lead progression effort.

Performing charge optimization on only the R groups as a set and comparing these optimal charges to the reference partial charges on compound **1** suggests changes to improve binding. The group with the largest charge difference from the optimum is the hydroxyl at position 7, where the oxygen and hydrogen have reference charges of $-0.6e$

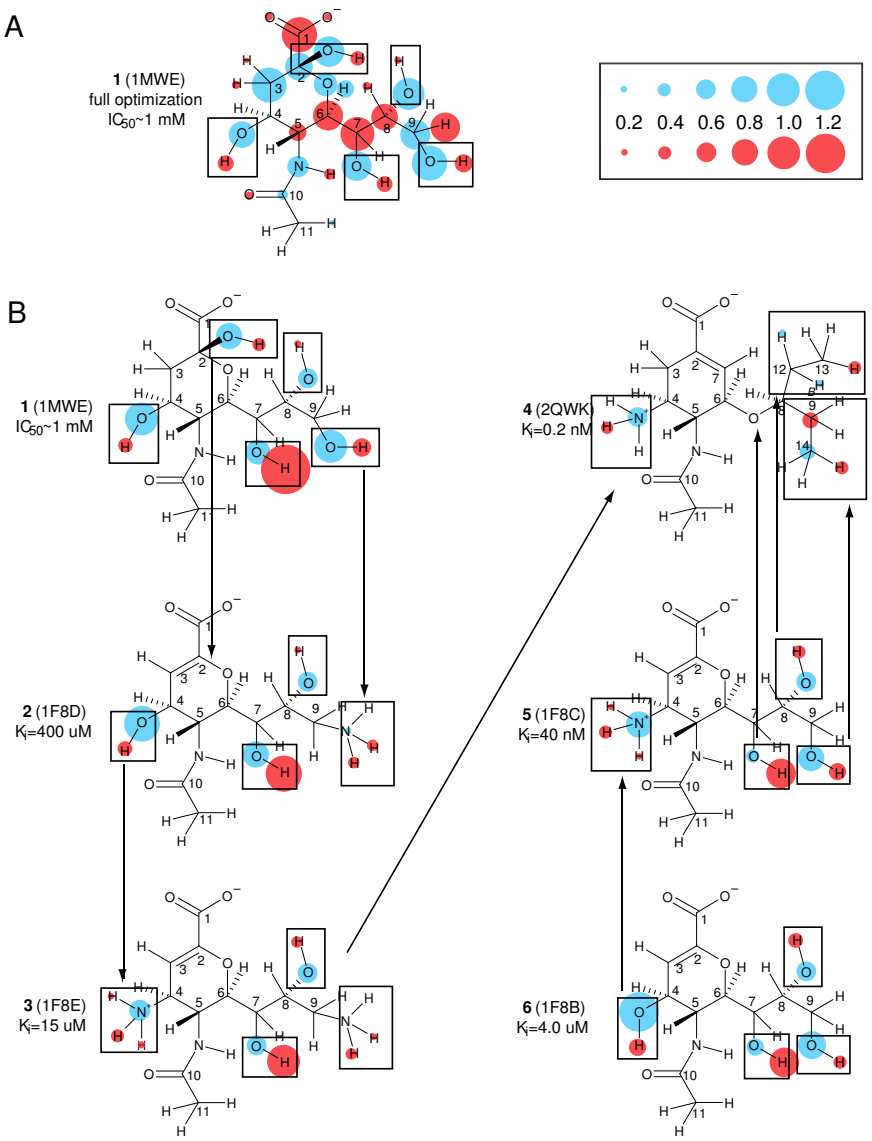

Figure 2.4: Structure-activity relationship analysis in a subset of neuraminidase ligands showing the suggested charge modifications for each compound. (A) Unconstrained optimization as described in the Methods for 1MWE. (B) Optimization of only the R groups. For each ligand, all groups in boxes were optimized together without constraints. The magnitude of the difference between the real and optimal charge for each atom is depicted by the size and color of the circle centered on each atom, as shown in the box in the upper right corner. Larger differences are indicated by larger circles, and circles are colored blue if the optimal charge is more positive and red if it is more negative. Numbers given in the key are in units of electron charge, and charge differences less than $0.15e$ are not shown.

and $+0.4e$, respectively. Comparison with calculated optimal charges of $+0.1e$ and $-1.2e$ for the oxygen and hydrogen atom positions, respectively, suggests conversion of the hydroxyl to a more negative group. The second largest charge differences are found on the hydroxyl groups at positions 2 and 4, where charge optimization suggests (1) converting the hydroxyl group at position 2 with reference charges of $-0.6e$ and $+0.4e$ for the oxygen and hydrogen atoms, respectively, to a more neutrally charged substituent with optimal charges of about $+0.2e$ at both positions and (2) substituting the hydroxyl group at position 4, which has reference charges of $-0.7e$ and $+0.4e$ for the oxygen and hydrogen atoms, respectively, with a positively charged substitutent that has optimal charges of $+0.4e$ and $-0.05e$ at the respective positions. At the position 9 hydroxyl group, which has reference charges of $-0.7e$ and $+0.4e$ for the oxygen and hydrogen atoms, respectively, the analysis suggests replacement with a slightly positive group with optimal charges of $+0.3e$ and $-0.1e$ at the respective atomic positions. Finally, neutralizing the hydroxyl group at position 8 from reference charges of $-0.7e$ and $+0.4e$ at the oxygen and hydrogen atom positions to optimal charges of $-0.1e$ and $+0.2e$, respectively, is computed to improve binding.

The suggestions at positions 2 and 9 are in some sense incorporated into compound **2**, with the replacement of the position 9 hydroxyl group with a positive amine group and deletion of the position 2 hydroxyl group. Other modifications are possible that closely match the optimal charge distribution, including, for instance, alkyl substitutions. However, the actual favorable changes are along the lines of the charge optimization suggestions, indicating that charge optimization provides a useful guide. Although the greatest charge differences between reference and optimal are not found at positions 2 and 9, modifications consistent with the optimal charge distribution at these positions do contribute to a more potent compound with a measured K_i of 400 μM. The change in bond order and resulting change in stereochemistry at the 2-position carbon is not accounted for in the calculation of the optimal charges, although comparing the two cocrystal structures (1MWE and 1F8D) shows that these changes do not substantially change the position and conformation of the molecule (see Figure 2.1B). Analysis of compound **2** suggests converting the hydroxyl group

at position 7 to a negatively charged substituent, converting the hydroxyl group at position 4 to a positive group, and converting the hydroxyl at position 8 to a neutral or slightly positive group. These three suggestions are nearly identical to those found for compound **1**. The group at position 9 is now closer to the electrostatic optimum, as can be seen in Figure 2.4, although the analysis suggests that bringing the group closer to neutral can improve binding.

Compound **3** can be derived from compound **2** by conversion of the 4-position hydroxyl group to a protonated amine as shown, appearing to satisfy one of the charge optimization suggestions and resulting in a 27-fold improvement in affinity. The amine at position 4 has an overall charge of $+0.6e$ and is very close to the calculated optimal overall charge of $+0.5e$ for the group. A closer look at the atomic charges, however, suggests that the reference charges of $-0.3e$ for the nitrogen and $+0.3e$ for each of the three hydrogens on the amine would more optimally be inverted so that the nitrogen position has a charge of $+0.4e$ and the hydrogens have charges of $-0.1e$, $+0.1e$, and $+0.1e$. This charge inversion would decrease the desolvation penalty while a similar electrostatic interaction energy would be maintained, because the overall charge would be the same but the exterior atomic positions would be less charged. Unfortunately, this change appears chemically difficult.

The hydroxyl group at position 7 is still suboptimal in compound **3**, as is, to a lesser degree, the hydroxyl group at position 8. At the position 7 hydroxyl group, substitution of reference charges of $-0.6e$ and $+0.4e$ at the oxygen and hydrogen positions, respectively, with "inverted" optimal charges of $+0.04e$ and $-0.7e$ is predicted to increase binding affinity. One idea that is consistent with the charge optimization suggestion is conversion of the hydroxyl to a carbonyl. An interesting alternative approach to realizing the optimal distribution is found in compound **6**. The hydroxyl group at position 7 is deleted, and in the cocrystal structure (PDB ID 2QWK), a crystallographic water molecule sits near where the hydroxyl group sat in compound **3**. The water hydrogen likely sits at the hydroxyl oxygen position, thereby satisfying the calculated optimal charge distribution. The ether oxygen of compound **6** probably makes a hydrogen bond to the water in 2QWK and could thus play an important role

in stabilizing the water's position. The crystallographic water is not seen in cocrystal structures with compounds **1**–**5**. This analysis together with the ether being calculated to be nearly optimal in compound **6** (data not shown) suggests a contributory role for the crystallographic water.

The 4-fold increase in K_i in moving from compound **3** to compound **4** is not explained by the charge-optimization results, suggesting that other contributions are responsible for the increase in potency. Other studies have pointed out difficulties in explaining the structure-activity relationship (SAR) between compounds **3** and **4** [60] and possible explanations include differences in binding site protonation states [34], differences between binding states in low-temperature crystallographic and room-temperature assay conditions [60], and differences in internal strain contributions to binding [61]. Compound **4** has most of the same R groups as compound **1**, with the difference being the removal of the hydroxyl group at position 2 and a change in bond order, as described previously. The difference between the reference and optimal charges at the four R groups is essentially the same as found in compound **1**, except for at the hydroxyl group at position 7, where the group is significantly closer to the optimum (see Figure 2.4B). The optimal charge distribution with $-0.7e$ and $+0.04e$ at the oxygen and hydrogen, respectively, is closer to the reference charges of $-0.6e$ and $+0.4e$, respectively, than found with compound **1**. The optimal charge distribution appears to be more sensitive than the reference charge distribution to the modifications in moving from compound **1** to **4**.

Compound **4** is electrostatically suboptimal at four positions; the hydroxyl group at position 4 should be more positive, the hydroxyl groups at positions 8 and 9 should be neutral, and the hydroxyl group at position 7 should be negative, if the optimal distribution is used as a guide. In compound **5**, an amine replaces the position 4 hydroxyl, and the modification, which is consistent with the first suggestion, results in a 100-fold increase in potency. The overall charge on the amine at position 4 is now $+0.53e$ and is very close to the overall charge of $+0.55e$ found from optimization. Optimal charges for the hydroxyl groups at positions 8 and 9 of compound **5** are both approximately neutral at $+0.1e$, while the optimal charges for the hydroxyl group

at position 7 is as described above for compound **3**. Making the hydroxyl groups at positions 8 and 9 neutral by conversion to alkyl groups along with other changes (see Figure 2.4B) results in compound **6**, which further improves the affinity 200-fold. The change at position 7, described in detail above, appears to involve a crystallographic water. The central ring is also modified, though this change was not made as much for SAR reasons as it was for chemical stability and versatility reasons [62]. Results of full optimization of compound **6** suggest that the R groups are close to their charge optimum, while further charge optimization is possible on the central scaffold.

2.3.3 Application of charge optimization in lead progression

Medicinal chemistry optimization of neuraminidase inhibitors has focused on the five sites as defined in Figure 2.1A, and it is interesting to note that compound **6** has a nearly optimal charge distribution at all of these sites, suggesting that lead optimization in this case corresponds to charge optimization. Charge optimization does not provide modifications per se and partly requires chemical intuition to design actual substitutions that satisfy the computed optimal charge distribution. However, it does appear to retrospectively explain the structure-activity relationships seen in the series of neuraminidase inhibitors. Two potential lessons can be taken from this study. The first is that strong suggestions do result in substantial inhibitor potency improvements. On the other hand, slight modification of charges, such as those suggested for compound **6** R groups, may in practice be difficult to make given discrete chemical constraints. A second lesson suggested by the data here is that individual optimizations of these subsites are relatively independent from one another. A lead progression analysis, where all atoms of each ligand were optimized simultaneously, suggests changes to parts of the molecule that are "fixed" in the analysis here (data not shown). Even so, the suggested changes to the variable subsites coincide very well with those suggested in Figure 2.4B. Likewise, when each of these subsites is optimized individually instead of as a set (data not shown), the suggested changes coincide very well with those presented here.

One arising question is whether charge optimization must be applied in a sequen-

tial step basis. Comparison of compounds **1** and **6** suggests that the recommended modifications are largely applicable despite small changes such as the rotation of Glu276 [62]. In addition, because optimization of these groups is largely independent, changes could potentially have been made concurrently to improve affinity. However, the magnitudes of the charge optimization suggestions do change as we follow the series, as seen in Figure 2.4B, indicating that the subsites are not completely independent. Since each chemical modification shifts the charge distribution of the entire molecule, and because each ligand sits in the active site in a slightly different conformation, it is likely that charge optimization will be most useful when applied sequentially. In a drug discovery team setting, we envision that a computational chemist applies this method to cocrystal structures of leads and works with a medicinal chemist in finding suitable replacements according to charge optimization suggestions. Quantitative charges for R group replacements can be computed using the RESP methodology (see Methods) before synthesis. In this way, visualization and chemical intuition facilitates optimization of charge-based hydrogen-bond and salt-bridge interactions in a fashion similar to what is possible with shape-based interactions.

2.4 Conclusion

The trends associated with the neuraminidase ligands examined here indicate that in a polar active site, calculation of the electrostatic optimum of a ligand can help to suggest useful chemical modifications that generally improve the overall binding affinity. Small modifications in the neuraminidase ligands cause their charge distributions to move significantly with respect to their optima, and the modifications result in large changes in overall binding affinities. These changes are not only due to electrostatics, however. As shown in Figure 2.5, the electrostatic component of the free energy of binding does not dominate the total binding affinity. Therefore, the van der Waals, entropy, and molecular strain changes that occur with each new ligand overpower, in some cases, the electrostatic effects in determining the final binding

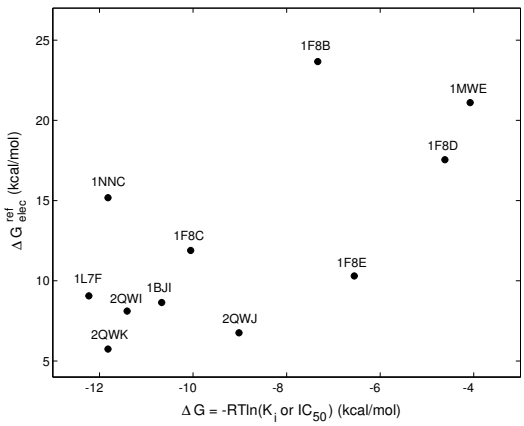

Figure 2.5: Electrostatic contributions to ligand binding relative to experimentally measured binding affinities. Scatter plot of calculated electrostatic components, ΔG^{ref}_{elec}, versus experimentally measured binding affinities, $\Delta G = -RTln(K_i$ or $\mathrm{IC}_{50})$, for the 11 ligands. Free energy terms are defined and discussed in the test. Positive values of ΔG^{ref}_{elec} and ΔG represent unfavorable net contributions to binding affinity, while negative values represent favorable contributions.

affinity. Because the active site of neuraminidase is quite polar, these effects are not expected to decrease in other targets.

Despite this, charge optimization can be a useful tool, because it can be used to make and evaluate molecular modifications. Charge optimization can suggest useful changes to a molecule in lead progression that consistently improve the electrostatic portion of the binding affinity. We note that neuraminidase is a challenging drug target, and over three decades were required to find a compound that was potent enough for physiological interference. In comparison, inhibitors binding in more hydrophobic pockets, such as those found in kinases and other "drugable" targets, can often be progressed in terms of potency without the aid of cocrystal structures, suggesting that while medicinal chemists can intuitively optimize compounds that are driven by van der Waals packing, it is much more difficult to electrostatically optimize compounds, and computational analysis such as that presented here can be particularly useful in these cases. Pharmaceutical lead progression involves optimization of properties other than potency, such as selectivity and ADME profiles, requiring multifactorial optimization and decision-making. Because charge optimization provides intuitive direction, it can be incorporated into human generation of compound ideas that balance

these various requirements.

Chapter 3

Evolution of PDZ domain binding specificity in the animal synapse[1]

Abstract

The genome of the sponge *Amphimedon queenslandica* was recently sequenced, revealing that this organism possesses homologs of several genes from the animal synapse. Present-day sponges lack complete synapses and nerve cells, possibly because they lack a few key genes, primarily the ionotropic glutamate receptors (iGluRs). We computationally analyzed two of the newly discovered sponge genes (PDZ3 and CRIPT) and their binding interaction, using the crystal structure of this complex in the rat as a starting point. We found that despite the low overall sequence conservation between the rat and sponge proteins, the binding site residues are 100% conserved. This protein–protein complex is predicted to exist in the sponge and we have studied the electrostatic interactions across this binding interface in detail. We developed a mutual information metric for detecting co-evolution between residues in a multiple sequence alignment of PDZ genes. We also used this technique to locate residues important for PDZ domain specificity.

[1]Portions of this text have been submitted as: O. Sarkarya*, K. A. Armstrong*, M. Adamska, M. Adamski, I. Wang, B. Tidor, B. M. Degnan, T. H. Oakley, and K. S. Kosik. A post-synaptic scaffold at the origin of the animal kingdom. (*These authors contributed equally to the work.)

3.1 Introduction

Sponges are a diverse group of aquatic animals that comprise the phylum Porifera. Sponges are classed as metazoans due to their internal digestive chamber, but in contrast to most other animals, sponges lack organ systems. A limited number of discrete cell types and a unique water-canal system characterize their body plan [63]. Based on molecular and morphological evidence, modern poriferans are divided into the siliceous and calcareous sponges, which are the first and second branches off the main metazoan phylogenetic tree [64–66]. Neurons are thought to have arisen in the main eumetazoan lineage after sponges diverged [64].

In the animal synapse, two neurons come together with different sets of proteins collecting near the membranes on either side of the boundary between them. The collection of proteins found to group on the signal-receiving neuron is known as the post-synaptic density. While a definitive proteomic analysis of post-synaptic components does not exist, estimates of the number of proteins involved ranges from 77 to 374 [67, 68]. These reports describe receptors, channels, adaptors, scaffolds, and proteins involved in adhesion, signaling, translation, and the cytoskeleton. Because the proteins comprising the post-synaptic density exist in all organisms with nervous systems, one can track this gene set over evolutionary time. Recently, genomes from two representative organisms at pivotal positions in nervous system evolution have become available. They are the sponge *Amphimedon queenslandica*, which lacks neurons and a nervous system, and the cnidarian *Nematostella vectensis*, which possesses a nervous system. A large number of vertebrate post-synaptic gene homologs are present in the sponge and nearly the entire gene set is present in *Nematostella* [69]. Figure 3.1 shows the post-synaptic density components and their putative arrangement near the membrane. Components are colored by the earliest organism in which they are found, indicating that the majority of post-synaptic proteins are evolutionarily quite old.

The main distinction among the post-synaptic density gene set between sponge and cnidarian are genes that encode excitatory post-synaptic receptors. Among the

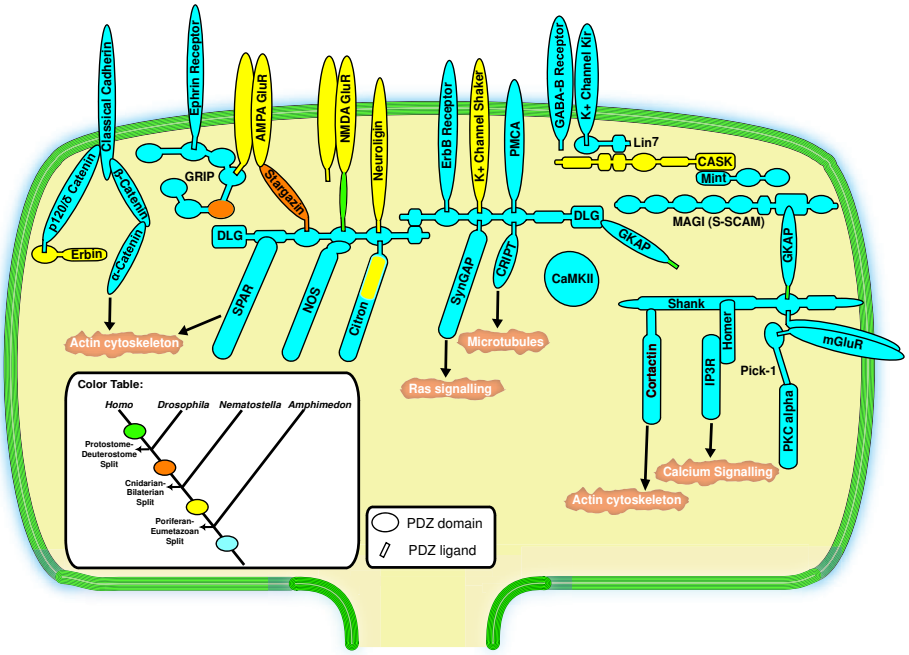

Figure 3.1: Post-synaptic density proteins at the neuron membrane. Proteins are colored by the earliest organism in which they are found. Homologs of blue proteins were found in *Amphimedon*, homologs of yellow proteins were found in *Nematostella*, orange proteins were found in *Drosophila* and green components are found in *Homo sapiens*. PDZ domains are represented by small ovals, and along the DLG backbone the first, second, and third ovals represent PDZ1, PDZ2, and PDZ3 domains. Figure created by Onur Sakarya.

post-synaptic receptor genes absent from *Amphimedon* are the Shaker-type voltage gated K^+ channel, neuroligin, and iGluRs (ionotropic glutamate receptors, such as NMDA, AMPA, and Kainate receptors) [69]. All of these genes are found in *Nematostella* and in more recent animals and are thought to be essential for linking the post-synaptic density with the extracellular region of the animal synapse.

To investigate the role of the post-synaptic density in the sponge, we focus on scaffolding genes because almost a complete set of these post-synaptic genes is found in the sponge genome. We focus on the *dlg* scaffold as a key component of the post-synaptic density, shown in Figure 3.1. The motif organization of *dlg*, with three PDZ (PSD-95/Discs-large/ZO-1) domains, a Src homology-3 (SH3) domain, an inactive guanylate kinase-like (GUK) domain, and an L27, appears to be unchanged since the origins of the animal kingdom [69]. The PDZ1 and PDZ2 domains of the *dlg* scaffolding protein are thought to anchor the iGluR NMDA to the synapse [70], and the PDZ3 domain of *dlg* interacts with several binding partners such as CRIPT, PMCA, and Citron [71–73]. We have chosen to focus on the protein–protein interaction between PDZ3 and CRIPT as a key link between the NMDA receptors and the submembrane cytoskeleton, as CRIPT is thought to recruit microtubules to the synapse [74]. The structure of this binding complex has been solved in the rat (PDB code 1BE9 [75]; 1.82 Å resolution).

It is known that PDZ domain specificity is important for function and PDZ1, PDZ2, PDZ3, and other PDZ domain types are classified based on the consensus sequences for the peptides they bind [76]. However, the mechanism for the PDZ domains' enforcement of this binding specificity is not well-understood, and studies in this area sometimes conclude that PDZ domains are in fact non-specific binders and that specificity must come from elsewhere *in vivo* [77]. Other studies have found that different PDZ domain types do in fact bind specific four-residue peptide motifs [76, 78, 79], but such studies have been limited in the number of PDZ domain types examined. It is certainly not well understood which PDZ residues are important for binding specificity or how this specificity arose in molecular evolution. Additionally, a previous experimental electrostatic analysis of PDZ domain binding focused on

the terminal carboxylate as the only major source of peptide charge [80]. Previous structural studies of PDZ domain binding have made use of the 1BE9 crystal structure and sequence alignments to locate the PDZ domain binding site residues that make van der Waals and hydrogen bond interactions with the peptide ligand [81]. However, a structural analysis of the electrostatics of this binding interaction has never been done.

Other methods have been used for finding specificity-determining residues and for detecting co-variation between residues over a multiple sequence alignment. Mutual information has been used to locate specificity-determining residues by a multiple groups [82, 83] and a mutual-information-like metric has been used to locate residues that have been under evolutionary pressure to change together [84]. However, a common problem with mutual information is a tendency to overestimate the information when data is limited [85]. Other groups have solved this problem by estimating the statistical significance of their result by generating synthetic datasets of a similar size [82, 83] while others have simply ignored positions where the mutual information is likely to be over-estimated [84].

Our detailed analysis of the PDZ3–CRIPT interaction is possible because this co-complex structure has been solved in the rat and because the sequence identity of these genes between organisms is high enough to give us high quality sequence alignments. Additionally, the recently sequenced genomes of the sponge and the cnidarian give us data for these proteins in organisms both with and without synapses. This allows us to consider differences between them as possibly important for synaptic function. Individual PDZ domains from the post-synaptic genes of a wide range of organisms were collected, aligned, and their phylogeny computed by Onur Sakarya and Ken Kosik at the University of California, Santa Barbara.

Using this data, we explored the question of residue specificity by asking which PDZ domain amino acid positions are the most predictive of PDZ domain type. Because PDZ domains are classified based on the peptides they bind, these two characteristics are assumed to be well-correlated. We also asked whether in the evolution of the PDZ domains any pairs of positions have been under co-evolutionary pressure

and might therefore have a structural or functional relationship.

We generated a homology model of the sponge PDZ3–CRIPT complex in order to perform a detailed analysis of the structural interactions important for binding. We also performed an electrostatic analysis across this protein–protein interface to locate residues that might be important for binding specificity. We developed a mutual information metric that reduces false positives in mutual information calculations in order to ask these questions. Because any positive value for this metric between a PDZ domain position and PDZ domain type is meaningful, we used this metric to generate a list of domain positions important for PDZ domain specificity. We used the same mutual information technique to explore pairwise co-evolution between residue positions and pinpointed pairs of PDZ domain residues that were likely under evolutionary pressure to mutate in concert.

3.2 Results

3.2.1 Atomic-level conservation of the binding region

To investigate conservation of the binding domain at the atomic level we compared the PDZ3–CRIPT interaction across a variety of organisms using multiple aligned sequences, a single X-ray crystal structure, and appropriate homology models. The organisms studied were *Rattus norvegius* (rat), *Danio renio* (fish), *Drosophila melanogaster* (fly), *Anopheles gambiae* (mosquito), *Strongylocentrotus* (sea urchin), *Caenorhabditis elegans* (worm), *Nematostella vectensis* (cnidarian), and *Amphimedon queenslandica* (porifera/sponge). Figure 3.2 contains an alignment of the PDZ3 sequences from the synaptic scaffolding proteins for these eight organisms. We used the rat PDZ3–CRIPT co-crystal structure (PDB code 1BE9 [75]; 1.82 Å resolution), as the starting point for our analysis of this protein–protein complex. The rat and sponge sequences are only 50% identical (over the PDZ3 domain — residues 302–402 in 1BE9), while the average identity between the rat and each of the other organisms is 66%. Despite these differences over the entire domain, the core residues in direct

Figure 3.2: Sequence alignment of the PDZ3 domain over eight organisms: rat, sponge, cnidarian, sea urchin, worm, mosquito, fish, and fly. Positions with at least 75% identity are colored according to the legend.

contact with the bound peptide are identical between the rat and the sponge, as assessed by identifying contact residues in the rat crystal complex and finding complete conservation in the aligned sequences. Among the entire set of organisms, the only mutation to a residue that directly contacts the peptide in the rat structure is I328V (all residue numbering corresponds to the 1BE9 crystal structure). We colored the crystal structure in Figure 3.3 based on our alignment of the rat and sponge sequences. Residues that are identical between the rat and the sponge sequence are colored green, highlighting that conservation is focused in the binding region.

Five separate homology models for the sponge PDZ3–CRIPT co-complex structure all demonstrated a shape that was highly similar to the rat crystal structure (average C_α RMSD 2.7 Å), with the largest deviation occurring in the N- and C-terminal regions and at the proline insertion in the sponge sequence between positions 318 and 319. To assess the effect of these rat-to-sponge sequence changes on binding, we computationally estimated the free energy of binding for the PDZ3–CRIPT interaction for the rat crystal structure and for each sponge model complex. The calculated total binding affinity varied by less than two-fold between these organisms, over all of the homology models, indicating that PDZ3 and CRIPT could have interacted prior to the appearance of a synapse.

Despite the essentially identical residues lining the binding pocket for these eight PDZ3 domains, seven of them bind a CRIPT protein whose C-terminal sequence follows the motif K/R-Q-S/T-S-V/A/T (by convention, we number this peptide from

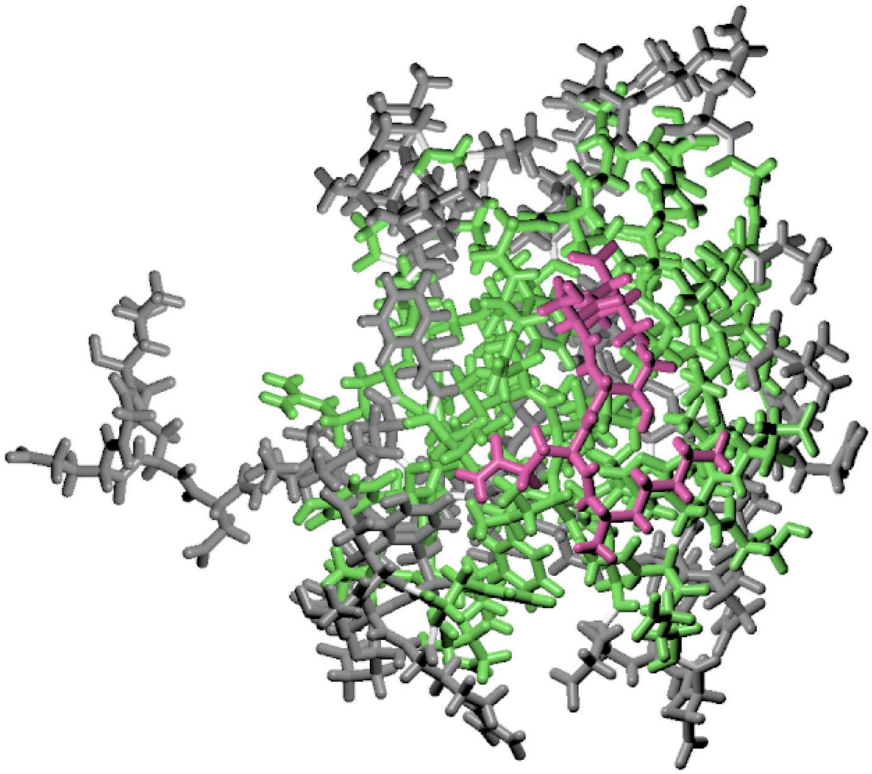

Figure 3.3: Crystal structure of rat PDZ3 binding CRIPT (PDB code 1BE9). The complex was crystalized with the last 9 residues of CRIPT (TKNYKQTSV) but the structure was resolved for only the last five residues. These CRIPT residues are shown in pink and the PDZ3 domain (residues 301–415) are shown in green and gray. The PDZ3 residues that are conserved between the rat and sponge PDZ3 sequences are colored green. Conserved residue numbers outside of 302–402 are not shown because they are designated as part of the construct in the crystal structure file. The identical residues between rat and sponge are: 311, 312, 314, 318, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 335, 336, 337, 338, 339, 341, 345, 347, 351, 353, 354, 356, 357, 359, 360, 362, 363, 364, 367, 371, 372, 373, 375, 376, 378, 379, 380, 382, 383, 385, 386, 387, 392, 393, 394, 396, and 400 (all residue numbering corresponds to the 1BE9 crystal structure).

the -4 (K/R) to the 0 (V/A/T) position), whereas sponge PDZ3 binds a CRIPT with a charge-to-hydrophobic substitution at the -4 position: V-Q-S-T-V.

Interactions between the PDZ3 domain and the CRIPT peptide were mapped in Figure 3.4 to study structural and electrostatic differences between these protein–protein complexes. All possible van der Waals, hydrogen bond, and electrostatic interactions seen in any of the homology models were included in this figure, but all reported distances are based on the rat crystal structure. Some of the sequence changes near the binding site could structurally affect the PDZ3–CRIPT interaction. For example, the valine at position 0 of the CRIPT peptide is alanine or threonine in some organisms. Additionally, the serine and threonine in the CRIPT peptide have changed locations in sponge, in comparison to the other organisms. These mutations move a methyl group from a surface boundary position to a more surface-exposed position in our homology models, but it is unknown how much this change will affect binding.

An electrostatic analysis of the binding site highlighted differences in the sponge sequences compared to the other organisms that could impact binding specificity. In Figure 3.4A we show important interactions between the CRIPT C-terminal peptide and the PDZ3 domain in all the organisms except for the sponge. All residues calculated to make van der Waals contacts, hydrogen bonds, or electrostatic interactions of greater than 0.1 kcal/mol in magnitude in any of the PDZ3 homology models are shown. For all the organisms except for sponge, the positively-charged residues at the –4 position and the interacting negative charges in the PDZ3 domain are largely conserved. In contrast, the hydrophobic residue at position –4 in the sponge CRIPT peptide no longer makes distant electrostatic interactions with the PDZ3 domain, as shown in Figure 3.4B, eliminating up to five such interactions. Additionally, fewer of the residues that interact distantly in the rat are still charged. These changes suggest that mutations may occur in a concerted way to maintain the appropriate binding function in each organism.

We explored the significance of this lysine-to-valine mutation in the CRIPT C-terminal peptide through Poisson–Boltzmann electrostatic calculations. The lysine

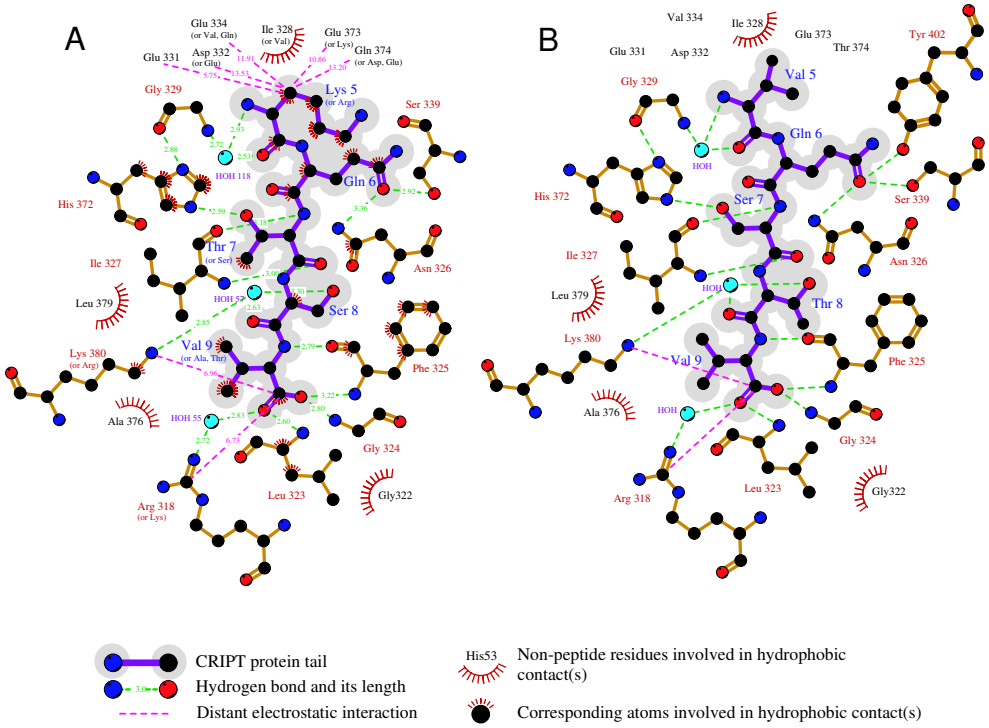

Figure 3.4: Structural interactions between (A) the C-terminus of the CRIPT peptide and the PDZ3 domain in rat, fish, fly, mosquito, cnidarian, sea urchin, and worm and (B) the C-terminus of the CRIPT peptide and the PDZ3 domain in sponge. All the distances shown in part A and the "ray" marks indicating hydrophobic contacts are based on the rat (1BE9) crystal structure. Distances to Lys5 are shown to the beta-carbon because this is the last atom whose coordinates were reported in the crystal structure. Hydrophobic contacts from residues G322, A376, and L379 were found to interact in some organisms, but in the rat they were somewhat distant. For each residue, the main label corresponds to the rat sequence, and the smaller (or "residue name") notes correspond to other organisms (not rat or sponge). Two interactions seen in the worm homology model were not included because they were unlike those seen in other organisms (G330 and P335). In the sponge diagram (part B), interactions that are possible given homology to the rat sequence are indicated, such as the hydrogen-bonds between PDZ3 and peptide. Residue numbering in the sponge structure corresponds to the aligned residue in the rat sequence. Some of the homology models suggested variant hydrogen-bonding patterns in the sponge, especially for residues Gln6 and Asn326. The figure was created using Ligplot[86] and HBPLUS[87]. The identical residues between rat and sponge are: 311, 312, 314, 318, **322**, **323**, **324**, **325**, **326**, **327**, **328**, 329, 330, 331, 335, 336, 337, 338, **339**, 341, 345, 347, 351, 353, 354, 356, 357, 359, 360, 362, 363, 364, 367, 371, **372**, 373, 375, 376, 378, 379, **380**, 382, 383, 385, 386, 387, 392, 393, 394, 396, and 400. Those residues in bold make direct contact with the CRIPT ligand in the rat crystal structure.

Rat	E373	E331	A375	A376	K380				
	-0.8	-0.3	0.1	0.2	0.5				
Cnidarian	E331	E373	E334	K380					
	-0.2	-0.1	-0.1	0.1					
Fish	E331	E373	G329	H372	Q374				
	-0.5	-0.5	-0.1	0.4	0.1				
Fly	E373	E331	H372	K380					
	-0.2	-0.1	-0.1	0.1					
Mosquito	E331	E373	H372						
	-0.2	-0.2	0.1						
Sea Urchin	E373	E331	G329	D332	H372	A375			
	-2.6	-1.2	-0.4	-0.1	0.6	0.1			
Worm	E331	D332	P335	E334	G330	H372	K373	G329	R380
	-0.9	-0.2	-0.1	-0.1	0.5	0.4	0.3	0.2	0.1
Sponge	N/A								
	N/A								

Table 3.1: Electrostatic interactions of magnitude greater than 0.1 kcal/mol between each residue in each organism's PDZ3 domain and the lysine at the –4 position in CRIPT. The top row for each organism gives the interacting position from PDZ3 and the bottom row gives the electrostatic interaction energy.

side chain at position –4 of the CRIPT peptide pays a desolvation penalty of 2.0 kcal/mol in the rat crystal structure and recovers –1.2 kcal/mol, primarily from electrostatic interaction with residues E331 and E373, as well as with smaller contributions from D332 and E334. The lysine also interacts unfavorably with A375, A376 (through backbone interactions), and K380 in the rat structure (+0.8 kcal/mol), leading to a net negligible electrostatic interaction (–0.004 kcal/mol) between this lysine and all other residues. Similarly, in the cnidarian, fish, fly, mosquito, sea urchin, and worm homology models, the lysine (or arginine) has electrostatic interactions and a desolvation penalty of roughly the same magnitude and is usually slightly less favorable than a hydrophobic isostere would be at that position. Table 3.1 shows the largest residue-by-residue electrostatic interactions this residue makes in each organism. A valine side chain at this position does not pay any desolvation cost and recovers no electrostatic interactions. This evolutionary change from valine to lysine at CRIPT position –4 appears to be net neutral for binding, but its interaction with the cooperative mutations in PDZ3 could be beneficial for interaction specificity.

3.2.2 Specificity-determining residues

Using three available crystal structures from *dlg*, a core union of 13 amino acids that are in contact with the ligand was defined. These included a PDZ1 co-crystal, a PDZ2 co-crystal, and a PDZ3 co-crystal (PDB identifiers 2I1N, 2G2L and 1BE9, respectively [10, 75, 88, 89]). These positions all made either van der Waals contacts of 3.9 Å or closer or hydrogen bonds of 3.5 Å or shorter with CRIPT in any of the three crystal structures.

We created a multiple sequence alignment of 64 sequences where the genes from human, fly, cnidarian, and sponge were collected for 16 different PDZ domain types. We used a mutual information metric, described in Methods, to study the covariance between each PDZ position and the PDZ domain type. Because PDZ domain types are defined based on their binding specificity, we used the mutual information between each position and its PDZ domain type as a signal that that residue is important for binding specificity. Their covariation with PDZ domain type could either mean that that residue has direct influence on binding or has a secondary structural effect on binding due to covariation with a binding residue or stabilization of the PDZ domain structure. It is important to note that positions that are highly conserved over all PDZ domain types, while likely important for binding, will have little to no information about the PDZ domain type. Most of the residues directly contacting the peptide fall into this highly conserved category. The 11 positions with the most mutual information with the PDZ domain type are shown in Figure 3.5. These positions are close to the binding site but most are not contacting the CRIPT peptide. Interestingly, some of the specificity-determining residues are quite far from the CRIPT peptide, in the direction that the CRIPT peptide extends past the –4 position. This may indicate that longer peptides should be considered in future studies of PDZ domain specificity.

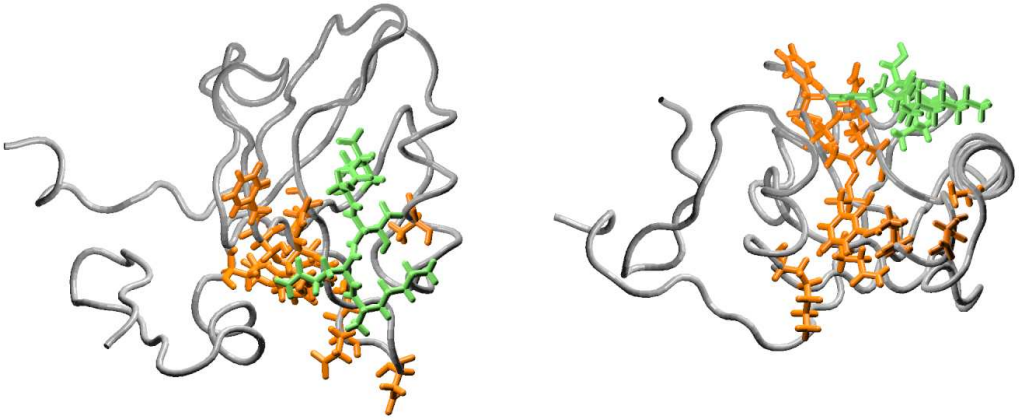

Figure 3.5: The 11 positions with the largest mutual information with PDZ domain type (with mutual information > 0.8 bits) are colored in orange. The CRIPT peptide is shown in green and the rat PDZ3 structure is shown in gray. On the left the specificity-determining residues and the CRIPT–PDZ3 interaction are shown from the front. On the right the specificity-determining residues and the CRIPT–PDZ3 interaction are shown from the bottom. The right figure is the same structure as the left figure but is rotated 90° forward into the page.

3.2.3 Co-evolving residues within the PDZ domain

We also used our mutual information metric to study whether concerted mutations have been required in the evolution of the PDZ3–CRIPT interaction. Because our alignment of both CRIPT C-terminal peptides and their corresponding PDZ-domain binding partners had only eight sequences, we could not accurately determine if pairs of residues across the binding interface co-evolved together. However, we were able to create a larger alignment of only PDZ domains where we could ask how the residues on the PDZ side of the interface have changed together. It is important to note that even though the binding site residues were conserved in our eight-organism alignment of PDZ3 domains, these binding site residues change frequently in this larger alignment, so calculating the mutual information between these binding site positions is meaningful.

We analyzed a multiple sequence alignment of 284 PDZ genes from many different organisms and studied the mutual information between pairs of positions in the alignment. In Figure 3.6 we report the mutual information metric between all pairs of positions in the alignment that are non-gaps in the rat PDZ3 sequence. All positive values in this matrix are expected to represent true mutual information because our mutual information metric was specifically designed to eliminate false positive values of the mutual information. The ten largest values from this matrix are the ten strongest pairwise co-evolutionary relationships and the residues involved in these relationships are depicted in Figure 3.7. Sets of residues with co-evolutionary relationships were found in many regions of the PDZ domain. Some residues are near the binding site, and the sets of co-evolving residues shown in blue and yellow in Figure 3.7 contain at least one member that is part of the core 13 binding residues. A pair of co-evolving residues was also found on the opposite side of the protein (shown in red in Figure 3.7, however, and the cause of their co-evolutionary relationship is unknown. It is possible that they are important for stabilizing the PDZ domain structure or that another unknown binding partner has impacted their evolution. As our mutual information metric typically gives a lower bound on the true mutual informa-

tion between positions, the positive correlation between these positions is thought to be meaningful.

3.2.4 A comprehensive picture of PDZ-domain binding

We hypothesized that positions important for binding would either be highly conserved or have strong mutual information with PDZ domain type. We therefore computed a third metric: position conservation over the 64 sequences in our first alignment. For each position in the PDZ domain we plotted three metrics in Figure 3.8A: that position's conservation, its mutual information metric with PDZ domain type, and its largest pairwise mutual information metric with another position. The 13 core binding site residues are highlighted in Figure 3.8A by filled circles and with corresponding colored residues in the inset. Each binding site residue appears to be very highly conserved, have very high mutual information with PDZ domain type, or both. Other positions also fall along the diagonal in this analysis and may also be particularly important for binding. We analyzed 9 of these other residues: those highlighted with purple arrows in Figure 3.8A. The locations of these residues with respect to the other binding site residues are shown in Figure 3.8B. They all interact with binding site residues and are removed from the binding site by one shell of amino acids. These residues could still be important for binding by a secondary mechanism such as stabilization of the binding site residues. Additionally, the mutual information metric between each position and PDZ domain type is correlated with the largest pairwise mutual information metric between that position and any other position, shown by the circle colors in Figure 3.8A (correlation coefficient 0.84). Therefore, residues important for PDZ domain binding specificity have a strong co-evolutionary relationship with at least one other position in the PDZ domain, suggesting that changes in required binding function could have provided evolutionary pressure that forced some residues to mutate in concert.

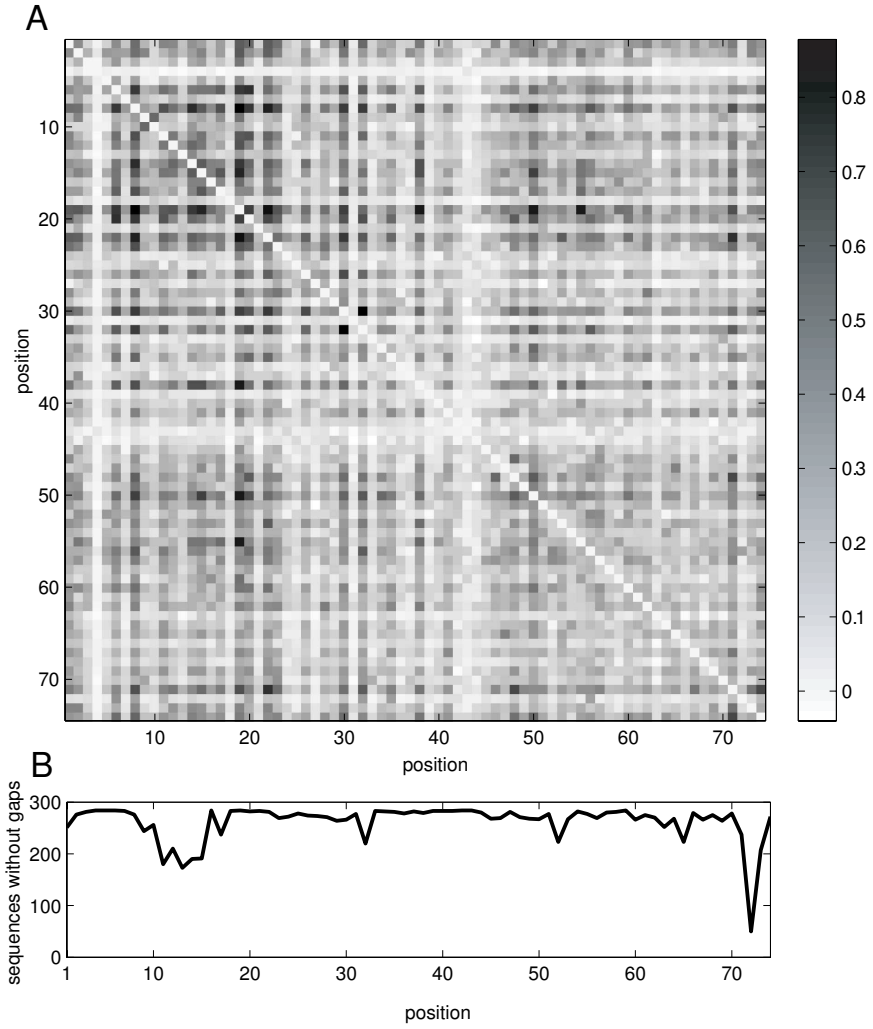

Figure 3.6: (A) Mutual information metric for the pairs of positions in the multiple sequence alignment that are non-gaps for the rat PDZ3 sequence, as described in the Methods. (B) The number of sequences at each position that do not have a gap at that position. Positions with a small number of gapless sequences may lack sufficient data in the sequence alignment to accurately assess a potential statistical relationship.

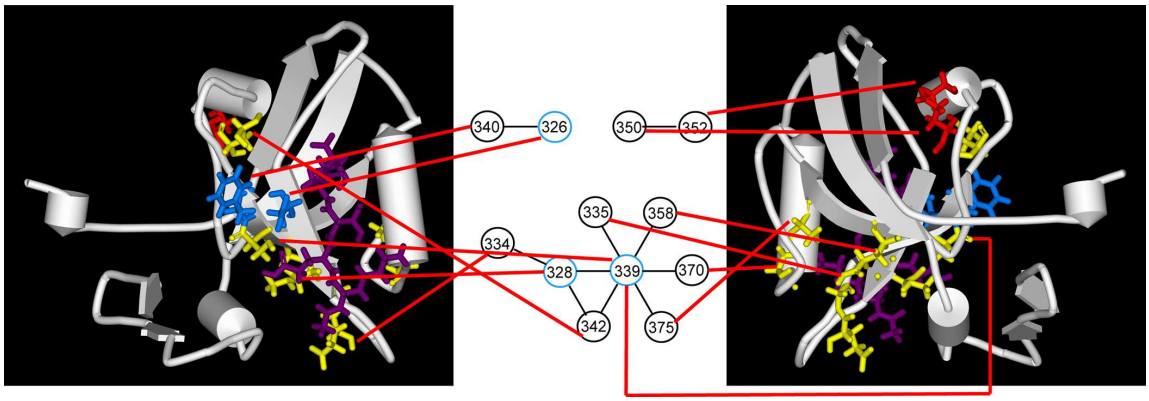

Figure 3.7: The 10 largest pairwise mutual information values in the matrix in Figure 3.6A. The circles contain the residue numbers corresponding to those 10 pairwise information values and those residues are highlighted in the front (left) and back (right) of the PDZ3–CRIPT structure, where the PDZ3 domain is shown in white and the CRIPT peptide is shown in purple. Blue circles indicate residues that are part of the core set of 13 residues that bind to the ligand in PDZ1, 2, or 3. Each group of residues found to co-evolve is colored separately in the structure (yellow, red, blue, or purple), and red lines link each residue number to its position in the structure.

3.3 Discussion and Conclusion

The data presented here and other experimental studies [69] support the presence of part of the post-synaptic scaffold in the sponge. Our calculations as well as the experimentally-discovered expression of many post-synaptic genes in the sponge flask cells are suggestive that part of the post-synaptic scaffold exists as an assembled structure in the sponge. Therefore, a portion of the synapse could have been in place before synapses appeared, meaning that only a limited number of modifications from this structure would be required to create the modern synapse, primarily the addition of iGluR receptors.

Our electrostatic analysis suggests that PDZ domain binding specificity could have changed as these proteins moved from an unknown function in the sponge to involvement in the synapse in more recent organisms. This could also be an important functional change necessary for formation of a synapse. Several PDZ domain residues were found to be important for PDZ binding specificity. That these specificity-determining residues also typically co-evolve strongly with at least one other PDZ domain position

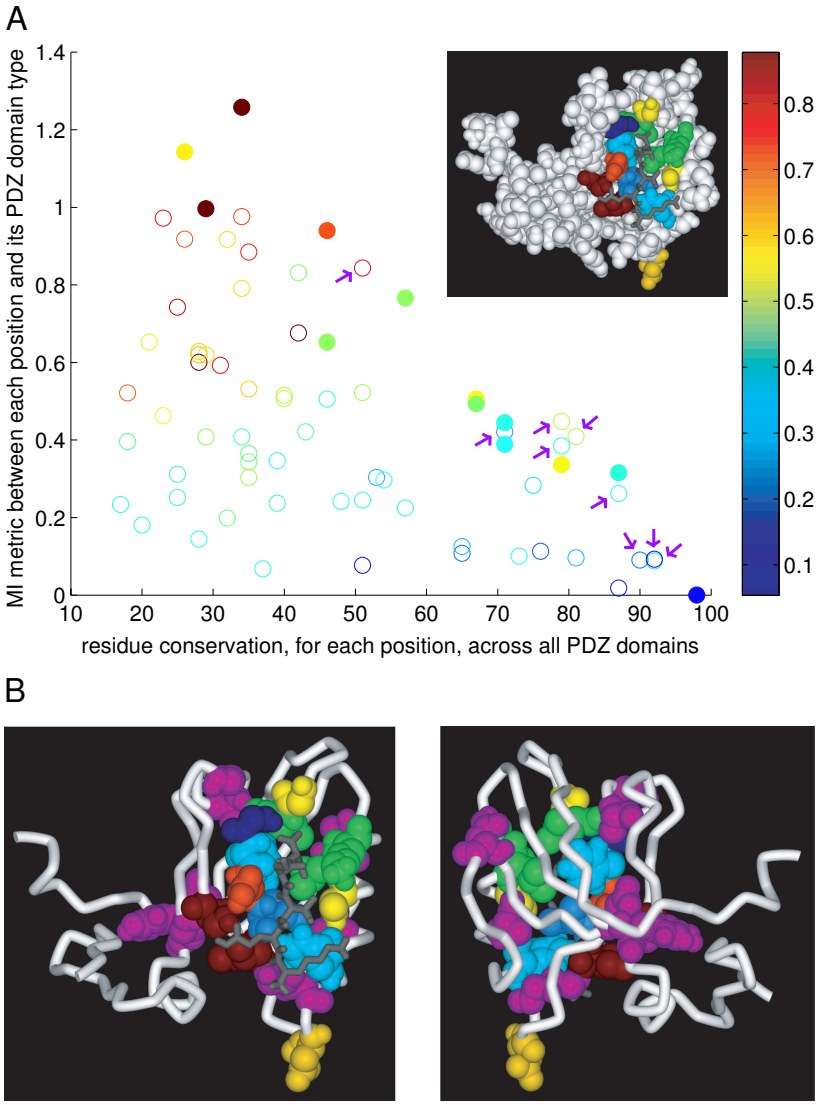

Figure 3.8: (A) Binding site residues are either highly conserved over all PDZ domains or are strong predictors of PDZ domain type. The x-axis is the residue conservation for each position over all PDZ domains and the y-axis is the corrected mutual information between each position and its PDZ domain type. Filled circles indicate positions in the core set of 13 residues that bind to the ligand in PDZ 1, 2, or 3. The color of each circle corresponds to the maximum mutual information that this position has with any other position. The color corresponding to the maximum mutual information for each of the 13 binding residues is painted onto that residue in the rat PDZ3–CRIPT crystal structure, shown in the inset. The core 13 residues are 322, 323, 324, 325, 326, 327, 328, 332, 339, 372, 376, 379, and 380. The circles with purple arrows are also either highly conserved over all PDZ domains or have strong mutual information with PDZ domain type, and correspond to residue numbers 329, 330, 345, 347, 356, 357, 363, 375, and 392. (B) These other residues are highlighted in purple in the rat PDZ3–CRIPT crystal structure, with the 13 core residues highlighted as in part (A) for comparison.

gives added support for the popular idea that locating co-evolving residues can be used as a proxy for locating residues important for protein function.

Additionally, our work points to the importance of the –4 position in the PDZ peptide ligand for electrostatic interactions with the PDZ domain. We found several specificity-determining residues in the PDZ domain in the region where the –5 to –8 peptide residues might lie. This suggests that future studies on PDZ domain binding electrostatics and specificity might be better informed by including more of the peptide ligand in the analysis.

3.4 Methods

3.4.1 Sequence data collection

Genomic traces and ESTs were generated as part a collaborative genome project with the Joint Genome Institute and are publicly available in the NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/). These genes were isolated as previously described [69].

3.4.2 Generation of PDZ sequence alignments

These sequences were selected and aligned by Onur Sakarya at the University of California, Santa Barbara according to the following protocol. For 12 selected PDZ-domain containing scaffolding genes, PDZ domain sequences were determined and numbered according to their spatial distribution on their corresponding scaffolds. Those PDZ domains of *Homo sapiens*, *Danio renio*, *Strongylocentrotus purpuratus*, *Drosophila melanogaster*, *Anopheles gambiae*, *Caenorhabditis elegans*, *Nematostella vectensis*, and *Amphimedon queenslandica* orthologs were collected by BLASTP searches on the NCBI blast server (http://www.ncbi.nlm.nih.gov/BLAST/) following a symmetrical best hit approach (SymBet) [90] with a cut-off threshold of E-value 10^{-5} or less. For each lineage, the dataset was enlarged, when possible, by the inclusion of PDZ domains from co-orthologs [91], which are two or more genes in one

lineage that are, collectively, orthologous to one or more genes in another lineage due to a lineage specific duplication(s). As an example, human SAP-97, PSD-95, SAP-102 and Chapsyn-110 are co-orthologs with fly DLG. Those PDZ domains (284 in total) were aligned using the global pair algorithm of MAFFT [92], which aligns large sets of distantly related sequences of similar length with high accuracy.

3.4.3 Structure preparation and analysis

The 1BE9 crystal structure was used for all rat calculations and as a starting point for all homology models. Hydrogen-atom positions were added to this structure using the HBUILD facility [93] and the PARAM22 parameter set [13] within the CHARMM computer program [94]. All water molecules were removed from the structure, except those making hydrogen bonds with the peptide (Figure 3.4). Missing side chains (F301, D332, and K5) were built using the IC BUILD facility, and the missing atoms of K5 (all atoms after the beta-carbon) were then minimized using the Adopted Basis Newton–Raphson (ABNR) method within the CHARMM minimization facility.

Homology models were generating using Modeller version 8v2 [95], using the rat (1BE9) crystal structure and each other sequence as aligned in Figure 3.2. Ten models were requested for each sequence, and the structure was selected that had the lowest Modeller objective function without errors in PROCHECK [96]. The five homology models referenced for sponge were also generated by Modeller.

Hydrogen bonds and hydrophobic contact distances in Figure 3.4 were computed using Ligplot [86] and HBPLUS [87]. The rat 1BE9 co-crystal structure was used to calculate all of the interaction distances in this Figure, as well as the hydrophobic contacts, except where noted in the Figure caption. Ligplot was also run on the homology models for the other organisms to determine which other hydrophobic contact and hydrogen bond interactions should be included.

3.4.4 Energetic calculations

Binding free energies were calculated as the sum of van der Waals, solvent-accessible surface area, and continuum electrostatic interaction and solvation terms. The CRIPT peptide was minimized in the bound-state structure, using ABNR in CHARMM, before calculation of the rigid-body binding energy. The van der Waals contribution to the binding free energy was calculated with the PARAM22 parameter set in CHARMM. The solvent-accessible surface area contribution was calculated with the surface area contribution to the free energy calculated as 5 calories per mole per square angstrom plus a constant of 860 calories per mole [39]. Electrostatic potentials were calculated by solution of the Poisson–Boltzmann equation, using a locally-modified version of DELPHI [19, 26, 27] with PARSE atomic radii and partial atomic charges [39], a solvent dielectric of 80 with a bulk ionic strength of 0.145 M, a solute dielectric of 4, and a 2.0-Å ion exclusion layer. The averaged result of 3 translations on a $129 \times 129 \times 129$ cubic grid was used in all cases.

3.4.5 Mutual information metric

Our mutual information metric is somewhat different from standard approaches and is designed to minimize the report of falsely high values for mutual information [85]. Thus, positive values of our mutual information metric indicate a true co-evolutionary relationship between positions or a true covariation between a position and its PDZ domain type.

We calculated a mutual information metric between all pairs of positions in the multiple sequence alignment and between each position and its PDZ domain type. The standard mutual information between positions i and j in a sequence alignment, where X is a random variable associated with the distribution of amino acids at position i and Y is a random variable associated with the distribution of amino acids at position j, is:

$$MI(i,j) = \sum_{x \in X, y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

The units of this metric is bits and this definition is standard [97].

The true probability distributions needed for an accurate solution are not known, and so they are estimated from the sequence alignment by counting the number of occurrences of each amino acid individually (to estimate $p(x)$ and $p(y)$) and in pairs (to estimate $p(x, y)$). This approximation often overestimates the true mutual information because limited sampling creates a biased estimate of the pair probability distribution that might be interpreted as mutual information. We calculate an estimate of the probable bias due to a lack of data, eliminating many falsely large estimates for the mutual information. Without changing the composition of amino acids, we randomly shuffle the distribution of amino acids at each position in the alignment N times, and calculate $MI(i, j)$ for each of the N shufflings. Only the $p(x, y)$ estimate changes during this shuffling, which is run until convergence of the mean $MI(i, j)$ over these random permutations (called $MI_{\text{correction}}(i, j)$). This term represents the spurious mutual information that is only due to the number of observed sequences being insufficient to truly estimate $p(x, y)$. Our mutual information metric is therefore:

$$MI_{\text{metric}}(i, j) = MI(i, j) - MI_{\text{correction}}(i, j).$$

This correction term should approach zero as more sequence data is included in the calculation. In this work, we consider all 20 amino acids and the gap ('-') to contain information at each position. This mutual information metric approximates a lower bound for the mutual information between pairs of positions.

Chapter 4

Re-engineering horseradish peroxidase using active-site focused combinatorial libraries[1]

Abstract

We have computationally analyzed the enzyme horseradish peroxidase, both at the active site and at distal locations, with the goal of improving enzyme re-engineering methodology. This enzyme is naturally used for the biosynthesis of plant hormones but is most widely produced for commercial uses. We used computational protein design techniques to explore the sequence space available to the enzyme active-site and found a focused region of sequence space that is small enough to be searched completely by high-throughput experimental techniques and that is predicted to accept mutation. We also computationally selected a subset of this space predicted to be compatible with native folding. Experimental collaborators in the Wittrup and Klibanov laboratories at MIT created combinatorial libraries representing both of these spaces. New enzymes were found in both libraries, allowing us to analyze our computational predictions and gain insight for future enzyme engineering studies. We also explored the role of distal mutations in the context of dynamics and correlated motions. The residues whose motions are correlated with that of the active-site offer useful suggestions for targeted mutagenesis experiments.

[1]Portions of this text have been submitted as:
Lipovšek, D.; Antipov, E.; Armstrong, K. A.; Olsen, M. J.; Klibanov, A. M.; Tidor, B.; & Wittrup, K. D. Selection of horseradish peroxidase variants with enhanced enantioselectivity by yeast surface display

4.1 Introduction

Many properties of natural enzymes make them appealing to industrial chemists and academic scientists alike. They can catalyze interesting chemical transformations, are highly selective to their substrates, and are generally environmentally friendly as they result in few byproducts [98, 99]. However, they all have functional limitations such as substrates accepted, reactions catalyzed, stereoselectivity, and operating conditions. For these reasons, there is great interest in re-engineering enzymes to give them novel activities. Progress in this area has generally made use of a parent enzyme sequence that is used to seed an evolutionary search in sequence space [100–102]. Until recently it was thought that to successfully modify enzyme activity, mutations must be made far from the active site in order to influence binding and catalysis without destroying activity. Indeed, distal mutations appear often in directed evolution experiments that modify catalysis [103–106]. However, the mechanism of influence for those distal mutations has not been well understood. More recently it has been proposed that when mutations in the active-site are possible, they are able to more effectively influence catalysis [107]. However, the problem still remains of how to find the appropriate mutations. The sequence space available for enzyme re-engineering is phenomenally large, and so limiting that space before experimental screening or selection should be useful. Computational techniques are able to focus the sequence-space search by optimizing the stabilization energy of a fold or binding interaction. Experimental methods can limit the space by focusing mutation on a small number of residues. However, the best way to choose these residues to limit the space has not been fully explored. The methodology of enzyme re-engineering would be well-served by a better understanding of the availability of new function within the active-site and of the mechanisms of influence of distal mutations.

Existing techniques for re-engineering enzymes include computational design, evolutionary engineering, homology-driven design, and catalytic antibodies. Computational methods for enzyme design include structure-based modeling of detailed electrostatics, packing, and other interactions, as well as statistics-based methods [108–111].

Although computational methods cannot currently model precise catalytic processes, some groups have met with success by stabilizing around different reasonable-looking fixed positions of catalytic residues [108–110]. Enzymes designed in this manner have typically been less efficient than those created by directed evolution, however, perhaps because changes in catalysis are coupled with subtle changes in backbone structure and protein dynamics [112]. Evolutionary engineering procedures typically use a sequence randomization procedure such as error-prone PCR or genetic recombination to create a library of variable sequences for screening or selection. These techniques have been used to engineer many new enzymes (see references 85, 86, 87 and references therein). Other groups have used a hybrid of computation and evolutionary engineering, using computation as the sequence variation step for evolutionary engineering approaches [113, 114] and to select crossover sites in genetic recombination experiments [115]. In homology-driven design, enzyme active-sites are engineered to create new mechanism or function or to fit different substrates by grafting desired properties of one enzyme onto another. This method has achieved some successes [116–118], but the efficiency of enzymes designed in this manner is sometimes poor compared to that of natural enzymes [112]. In catalytic antibody-based design, a transition state in a catalytic pathway is identified and a transition-state analogue is created which can be injected into an organism with an immune system (a mouse, for example) to create antibodies that bind to this transition state. These binding antibodies are then screened for catalytic activity. This method has also achieved success [103, 116] but again the catalytic efficiency of the resulting antibodies has been low [112].

Studies of distal residues typically rely on an analysis of sequence alignments or molecular dynamics simulations. Studies of positional covariation in sequence alignments have been used to located interesting distal residues that may be important for protein function [119], but it is unclear whether these residues can influence function from a distance. Studies of molecular motions have more directly studied this idea, as distal locations may affect the active-site by influencing global motions of the protein. These global motions may influence catalysis as promoting motions which decrease

the activation free energy barrier, ΔG, or as dynamical motions which influence the transmission coefficient for the rate constant of a chemical reaction [106, 120]. In dihydrofolate reductase (DHFR) for example, the catalytic rate is significantly affected by mutations far from the active site [103] and one group has done computer simulations to show that these mutations affect the correlated motions of the C_α atoms in the protein in a way that could affect the reaction rate and the hydride-transfer step [121]. Such dynamic information from computational simulations and NMR experiments has been used to choose a target region of the enzyme for mutagenesis [103]. Another recent computational study demonstrated that molecular dynamics simulations could be used to locate distal residues with motions correlated with those of the active-site residues [122].

Our work has focused on the enzyme horseradish peroxidase — a plant enzyme containing a ferric protoporphyric IX (protoheme) in its active site. Although its natural biological function is the biosynthesis of plant hormones, it is most widely produced because of its commercial uses, for example, as a component of clinical diagnostic kits and for immunoassays [124]. The catalytic residues in the horseradish peroxidase active site include a histidine coordinated to the iron below the heme as well as a catalytic histidine and an arginine in the active site. The structure of horseradish peroxidase and the arrangement of the active site residues are shown in Figure 4.1. Horseradish peroxidase catalyzes a three-step reaction in which the enzyme is first oxidized by a peroxide (such as hydrogen peroxide), producing an oxy-ferryl heme intermediate (called compound I) which is then reduced in two separate electron-transfer steps to compound II and then the native state. The mechanism for compound I formation is described elsewhere [125]. One in-vivo substrate of horseradish peroxidase is ferulic acid, a phenolic compound found in plant cell walls. Beyond this, however, very little is known about the function of horseradish peroxidase inside the horseradish plant. The reduction of compound I can catalyze eight different types of reactions, depending on the substrates. These are electron transfer, disproportionation, sulfoxidation, epoxidation, demethylation, dehydrogenation, hydroxylation, and α-oxidation [124].

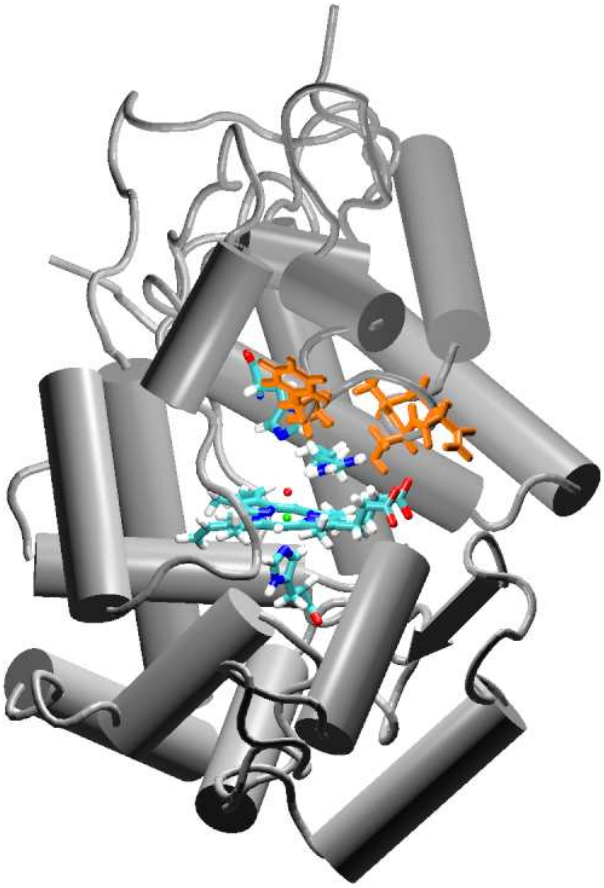

Figure 4.1: Horseradish peroxidase structure (PDB code 1HCH; 1.57 Å resolution [10, 123]) with the active site residues F68, G69, N72, S73, and A74 highlighted in orange. The heme group, the histidine coordinated to the iron (H170) and the catalytic arginine (R38) and histidine (H42) are shown in atom colors.

The enzyme design problem is most difficult because the size of the sequence space to be searched is huge ($\approx 10^{130}$ possible sequences for a 100-residue enzyme). Although searching the entire set of sequences available to an enzyme is not possible, various methods for focusing the search exist. Evolutionary engineering studies have had the most success when they choose a starting sequence with function very close to their goal sequence [126], effectively focusing on a region of sequence space near their desired function. One could focus the search to only a few residues, such as those at the active site. This allows an exhaustive combinatorial search of active site sequences and should reveal any active site mutations are at those positions. This strategy has been successful in several studies, with results summarized in this review [127]. Computation can also avoid examining the entire set of available sequences by using algorithms such as dead-end elimination and A* or Monte-Carlo to pare down the search space. This is the strategy taken by most hybrid computation and evolutionary engineering techniques. Locating important distal residues without trying all possibilities would be another useful strategy for reducing the search space.

We therefore ask whether active-site mutations can be found that alter the substrate enantioselectivity of horseradish peroxidase for L- versus D-tyrosinol. We selected a small number of active-site positions to search so that we could answer this question definitively using high-throughput experimental screening. Additionally, we ask whether computational protein design can be used to locate these new enzymes and how iterations of computation and experiment can be used to improve computational design methodology. We also explore whether there are residues far from the horseradish peroxidase active-site that could influence catalysis from a distance.

We have computationally analyzed the active-site of horseradish peroxidase and selected five residues to target for mutagenesis. The library resulting from saturated mutagenesis at these five positions has $19^5 = 2.5 \times 10^6$ members (all amino acids except cysteine allowed at each position). We also computationally selected a reduced portion of this library which was predicted to be enriched in sequences compatible with the native fold. The Wittrup laboratory used yeast-display and a fluorescent assay along with high-throughput selection with flow cytometry to completely search

the exhaustive and the computationally-focused libraries of active-site sequences, and the Klibanov laboratory used advanced enzymology to thoroughly study the kinetic properties of the variant enzymes. Substrate docking of L- and D-tyrosinol was done to further analyze the new enzymes. We have also explored the roles of distal mutations in the context of dynamics and correlated motions. This joint theoretical and experimental approach of computational library design along with experimental and computational analysis has improved our understanding of this enzyme and of enzyme design in general.

4.2 Results

4.2.1 Selection of five active-site positions for randomization

Groups of five residue positions were chosen based on structural proximity to the active site, and computational protein design techniques (described in Methods) were used to determine allowed sequences compatible with the native backbone structure. All sequences calculated to be within 15 kcal/mol of the wild type energy were included in this initial library. These calculations highlighted multiple sets of candidate positions that were calculated to allow many sequences. Further analysis of the built structures showed that they did not fill the active site with heavy atoms or consist of many charged residues (results not shown). We then used three other metrics to choose positions 68, 69, 72, 73, and 74 for randomization: mutual information between positions in our sequence alignments, amino acid frequency in known genes, and ease of synthesis. Mutual information was used to seek interactive positions, which are of special interest. The mutual information [97] between positions highlights pairs that might be structurally dependent on each other and therefore might be forced to mutate in unison. A sequence alignment of horseradish peroxidase genes was taken from Pfam [128], and our wild type horseradish peroxidase gene sequence was aligned by eye with the 309 seed alignment sequences to create a sequence alignment of 310 sequences. Highly conserved positions in this alignment were considered

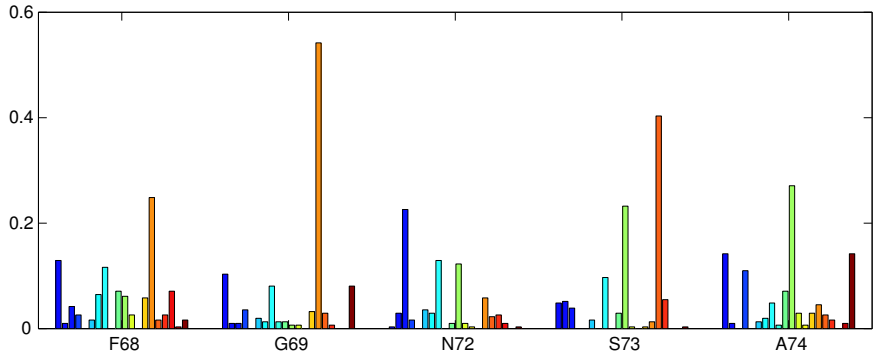

Figure 4.2: Amino acid frequencies for positions 68, 69, 72, 73, and 74 in the Pfam sequence alignment. The height of each colored bar corresponds to the frequency of that amino acid, and amino acids are listed in alphabetical order by three-letter code from left to right.

to be risky for randomization. Positions 68, 69, 72, 73, and 74 have low conservation in our sequence alignment (Figure 4.2), moderate mutual information (Figure 4.3), and their close proximity in sequence makes them relatively easy to modify using standard molecular biology techniques. Figure 4.4 shows the amino acid frequencies for positions 68, 69, 73, and 74 from our computationally-generated sequence library. In earlier rounds of computational design we found that position 72 was not restricted in sequence by the protein structure and was allowed to be any amino acid. We therefore excluded it from the results presented in Figure 4.4. Cysteine was not considered as an option in these calculations because we did not plan to use it in the experimental work. These five active-site positions were then targeted for mutagenesis, both exhaustively (with all amino acids except cysteine) and with a computationally-reduced set of amino acids at each position.

4.2.2 Design of a combinatorial library

The described computational design at positions 68, 69, 72, 73, and 74 produced a list of 121,104 sequences with energies on the native backbone predicted to be within 15 kcal/mol of the wild type energy. Because it would be unreasonable to do site-directed mutagenesis of each sequence individually, we then designed a combinatorial library with the goal of generating as many of these sequences predicted to be foldable as

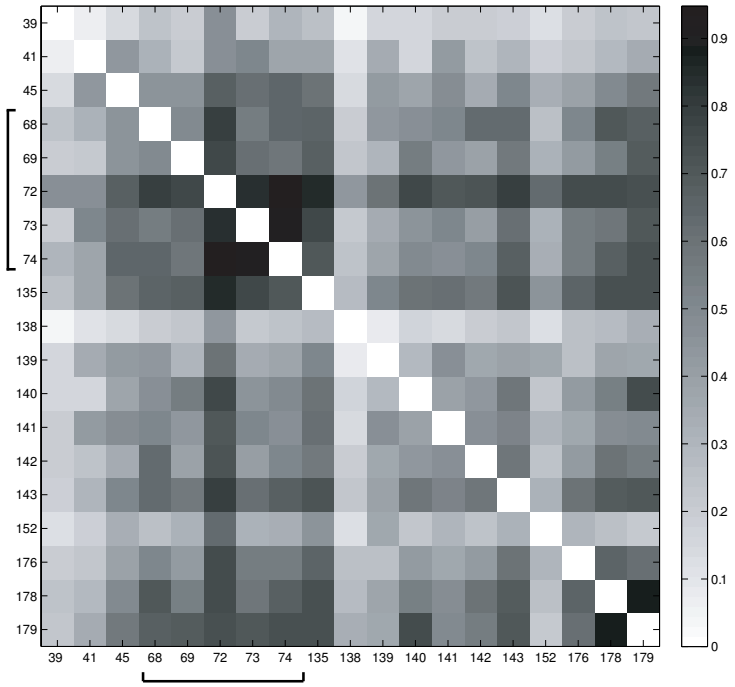

Figure 4.3: Mutual information (bits) between pairs of active site positions in the Pfam sequence alignment. Positions 68, 69, 72, 73, and 74 are highlighted.

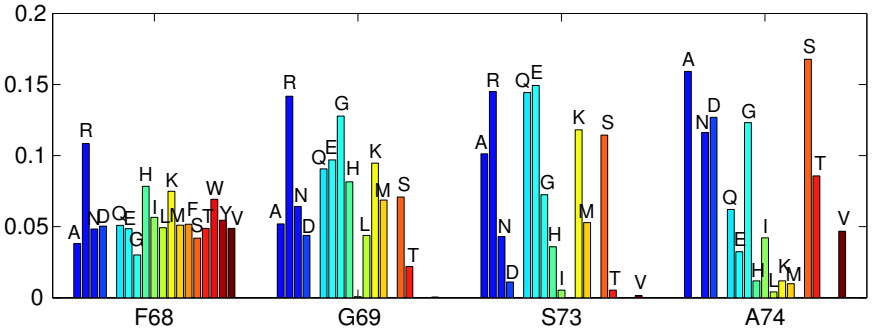

Figure 4.4: Amino acid frequencies for positions 68, 69, 72, 73, and 74 in our computationally-generated allowed sequence list. The height of each colored bar corresponds to the frequency of that amino acid, and amino acids are listed in alphabetical order by three-letter code from left to right. Position 72 is not restricted in sequence by the protein structure and our calculations allow it to be any amino acid. Cysteine was not considered as an option in these calculations because we did not plan to use it in the experimental work.

wild type	allowed amino acids
F68	19 amino acids (all except C)
G69	R N Q E G H K M S
N72	19 amino acids (all except C)
S73	R Q E K S
A74	A N D G S

Table 4.1: Amino acids allowed at each position in the combinatorial library. The wild type amino acid at each position is given before the position number.

possible while also focusing the library as much as possible. We chose a set of residues at each of the five positions, such that both the number of sequences computed to be foldable and the fraction of sequences computed to be foldable in the resulting combinatorial library are maximized. The tradeoff between these two factors is presented in Figure 4.5, where the most-frequent residues were selected at each position. To select the eventual set of combinatorial library amino acids, we required that at least 1,000 sequences be foldable and used a Monte-Carlo procedure to select the optimal number of residues at each position. In each Monte-Carlo step, the next most or least frequent amino acid in the list of sequences computed to be foldable could be added or subtracted from the combinatorial library. We also enforced that the wild type residues be included in the combinatorial library, and because all amino acids were allowed at positions 68 and 72, those positions were completely randomized to all amino acids except cysteine. The final computationally-focused combinatorial library is presented in Table 4.1, representing 0.025 of the possible sequence space. The fraction of these sequences that came from our list of sequences computed to be foldable is 0.77. Our experimental collaborators were able to synthesize this combinatorial library exactly using triphosphoramidite mixtures as described in Methods.

4.2.3 Selection results for exhaustive and computational libraries

Selection of enantioselective horseradish peroxidase variants was performed by Daša Lipovšek according to the following protocol. Each library underwent two parallel selections by fluorescence-activated cell sorting (FACS), one for enantioselectivity for

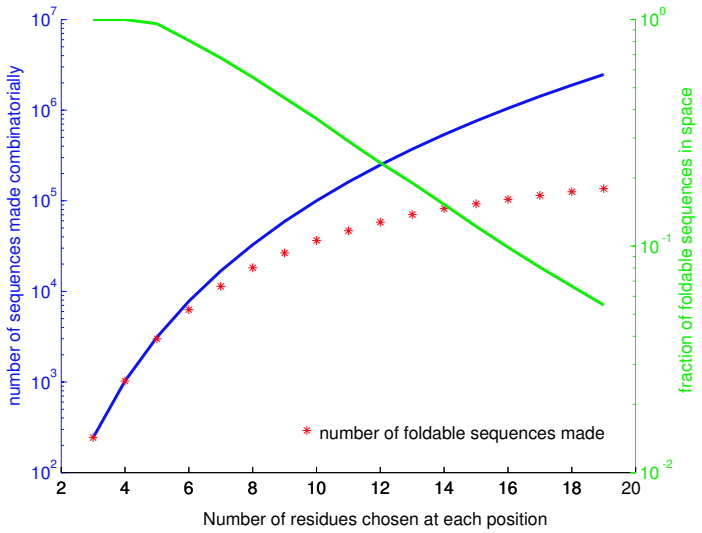

Figure 4.5: The tradeoff in combinatorial library design between the number of foldable sequences made and the fraction of foldable sequences in the combinatorial library. When only a few residues are chosen at each position for the combinatorial library, only a few foldable sequences are made but the fraction of foldable sequences in the resulting library is nearly 1. As more residues are chosen at each position, the number of foldable sequences made increases, but they form a smaller fraction of the total combinatorial library. Based on the screening capacity of an experimental system, this type of analysis could be used to determine the appropriate combinatorial library size.

D-tyrosinol over L-tyrosinol ($E_{D/L}$), and one for L-tyrosinol over D-tyrosinol ($E_{L/D}$), with alternating rounds of positive and negative selection. A minimum level of expression was also required in each selection. Between 21 and 24 clones from each selected population were sequenced and the enantioselectivity of all the clones that appeared in the selected population more than once was determined.

The enantioselectivities of selected horseradish peroxidase variants were determined by Eugene Antipov according to the following protocol. Yeast cells transformed with each variant of interest were labeled in parallel with L- and D-tyrosinol as described above for 0 to 4 min. Each of four time-point samples was analyzed using a Coulter Epics XL flow cytometer. The mean fluorescence of Alexa-488 was plotted against time to determine the initial reaction rates with each substrate, and the enantioselectivity was calculated as $E_{D/L}$ = (initial rate with D-tyrosinol)/(initial rate with L-tyrosinol) and $E_{L/D}$ = (initial rate with L-tyrosinol)/(initial rate with D-tyrosinol). The enantioselectivity for the AD8.01 clone from the computationally-focused library was calculated differently: the rate was taken over a 20-minute window instead of a 4-minute window. Therefore, this number cannot be directly compared to the enantioselectivities of the other mutants.

Results of both the $E_{D/L}$ and $E_{L/D}$ selections are presented in Tables 4.2 and 4.3. The selection for $E_{D/L}$ from the computationally-focused library and selections for both $E_{D/L}$ and $E_{L/D}$ in the exhaustive library yielded a single horseradish peroxidase variant that was enriched more than any other selected clone, and whose enantioselectivity exceeded that of wild type horseradish peroxidase. Variant CD8.02 (F68E, G69P, N72R, S73A), whose sequence was found in 76% of the clones in the final population selected for $E_{D/L}$ from the exhaustive library, has a 3.4-fold preference for D- over L-tyrosinol, which is a 3.8-fold improvement over wildtype horseradish peroxidase. Clone CL8.01 (F68L, G69A, N72E, S73L, A74Y), whose sequence was found in 52% of the clones in the final population selected for $E_{L/D}$ from the exhaustive library, has a 9-fold preference for D- over L- tyrosinol, which is a 7.5-fold improvement over wild type horseradish peroxidase. Variant AD8.01 (G68A, N72Q) was found in 96% of the clones in the final population selected for $E_{D/L}$ from the computationally-

$E_{D/L}$				$E_{L/D}$			
Clone	Sequence	Freq.	$E_{D/L}$	Clone	Sequence	Freq.	$E_{L/D}$
wtHRP	FGNANSA		0.9±0.1	wtHRP	FGNANSA		1.2±0.1
CD8.02	EP..KA.	76%	3.4±0.2	CL8.01	LA..ELY	52%	9±2
CD8.14	RP..HWT	10%	0.6±0.1	CL8.09	WA..AM.	17%	1.9±0.1
CD8.01	WV..FWS	5%	NA	CL8.02	.A..VVT	13%	3.3±0.7
CD8.07	MV..PMG	5%	NA	CL8.03	HA..ARG	13%	1.4±0.1
CD8.11	HS..GM.	5%	NA	CL8.16	RH..WTT	4%	NA

Table 4.2: Horseradish peroxidase variants from the exhaustive library, selected for enantioselectivity for D-tyrosinol ($E_{D/L}$) or L-tyrosinol ($E_{L/D}$). The sequence from residue 68 to 74 is shown and mutations from the wild type horseradish peroxidase sequence are recorded. A dot "." represents no mutation from the wild type. The reported frequency is the percent of the 21-24 sequenced clones with that sequence. Enantioselectivities were calculated as described in the text.

reduced library, has a 1.3-fold preference for D- over L-tyrosinol, which under this different measure of enantioselectivity (rates are calculated over 20 minutes) is a 3.7-fold improvement over wild type horseradish peroxidase. Other mutants were selected from the exhaustive library as well which were not found to be as enantioselective after characterization. In $E_{D/L}$ selection, CD8.01 (F68W, G69V, N72F, S73W, A74S), CD8.07 (F68M, G69V, N72P, S73M, A74G), CD8.11 (F68H, G69S, N72G, S73M), and CD8.14 (F68R, G69P, N72H, S73W, A74T) were found. In $E_{L/D}$ selection, CL8.02 (G69A, N72V, S73V, A74T), CL8.03 (F68H, G69A, N72A, S73R, A74D), CL8.09 (F68W, G69A, N72A, S73M), and CL8.16 (F68R, G69H, N72W, S73T, A74T) were found.

Several of the mutants found in the exhaustive library were not found in the computationally-reduced library. Because the yeast eukaryotic secretory system is expected to express only folded proteins on its surface, we investigated reasons why these sequences were not included in our combinatorial library of sequences computed to be foldable. Possible explanations include the rigid backbone approximation, the discrete rotamer library approximation, and the elimination of sequences during generation of the combinatorial library. We found that CD8.11 and CL8.09 were included in our original library of sequences computed to be foldable, but because of the low frequency of methionine at position 73 in that library, this residue was not included

$E_{D/L}$				$E_{L/D}$			
Clone	Sequence	Freq.	$E_{D/L}$	Clone	Sequence	Freq.	$E_{L/D}$
wtHRP	FGNANSA		0.35	wtHRP	FGNANSA		2.8
AD8.01	A...Q..	96%	1.3	wtHRP	.....	100%	3.8
wtHRP	.....	4%	0.35				

Table 4.3: Horseradish peroxidase variants from the computationally-reduced library, selected for enantioselectivity for D-tyrosinol ($E_{D/L}$) or L-tyrosinol ($E_{L/D}$). The sequence from residue 68 to 74 is shown and mutations from the wild type horseradish peroxidase sequence are recorded. A dot "." represents no mutation from the wild type. The reported frequency is the percent of the 21-24 sequenced clones with that sequence. Enantioselectivities were calculated as described in the text. Note that this enantioselectivity is calculated differently from the enantioselectivity in Table 4.2. Here, the rates were taken over 20 minutes whereas in Table 4.2 they were taken over 4 minutes.

in the combinatorial library. CD8.02, CD8.07, and CD8.14 all have a proline residue at position 69 or 72. The native backbone fold does not tolerate a proline at either of these positions, and angle movements of up to 60 degrees would be required to accommodate them. Therefore, the rigid backbone assumption prevented these residues from being included in our library of sequences computed to be foldable. The remaining five sequences were all computed to be very high in energy on the native backbone fold and do not contain proline residues. To investigate whether the rigid backbone assumption or the discrete rotamer assumption were to blame for their not being included, we minimized each side chain on the rigid backbone, mimicking the effect of a continuous rotamer library. Comparing these new minimized energies with the minimized energy of the wild type sequence we found that CD8.01, CL8.02, CL8.03, and CL8.16 would have been included in our library if we had used continuous rotamers. CL8.01 still would not have been included in our library in this case, so we assume that the rigid backbone approximation must be relaxed in order to include this residue. While these assumptions cannot currently be relaxed during the dead-end elimination and A* sequence and structure search, minimizing each sidechain after the dead-end elimination design is certainly possible. More testing of this idea, using both positive and negative data, would be useful future work.

4.2.4 Characterization of mutant CD8.02

Each new mutant selected from the exhaustive and computationally-focused libraries was found to have multiple mutations from the wild type sequence. However, due to the nature of the selection process, some of those point mutations might not be required for the altered enantioselectivity. Therefore, we chose one of the multiple mutants (CD8.02) and analyzed the necessity of each mutation for the new function by experimentally making reversion mutations back to the wild type residues. Mutations were made by Daša Lipovšek and enantioselectivities were determined by Eugene Antipov. We also made these mutations computationally and used substrate docking and binding energy calculations to explore whether these calculations correlated with the influence of each point mutation. For these calculations we assumed that the substrate binding energy and the binding mode would be the most important determinants of catalysis, with the distance between the phenolic hydroxyl and the oxygen linked to the heme iron being the most important characteristic of the binding mode. Given that no attempt was made to correlate these properties with real enantioselectivities before making "predictions" here, these computational results should only be considered as possible explanations for the observed enantioselectivities and as an analysis of the computational assumptions.

The calculated and experimental enantioselectivities of CD8.02-based reversion mutants are shown in Table 4.4. Mutant Rev68 (CD8.02(E68F)) remained active but its $E_{D/L}$ was half that of CD8.02 (1.7 compared to 3.4). Computationally, L-tyrosinol was predicted to bind more tightly to the horseradish peroxidase active site and the hydroxyl of L-tyrosinol was predicted to approach the heme oxygen more closely, and so L-tyrosinol was incorrectly predicted to react slightly more easily. The Rev69 (CD8.02(P69G)) and Rev72 (CD8.02(K72N)) mutants lost activity against both L- and D-tyrosinol to the degree that precluded accurate determination of their enantioselectivities. Computationally, the hydroxyl groups of neither D- nor L-tyrosinol was able to closely approach the heme oxygen in Rev69, and so this mutant was correctly predicted to be inactive. However, L-tyrosinol was able to get close to the heme

Clone	sequence	ΔG^D_{bind}	ΔG^L_{bind}	dist,D	dist,L	calc$_{D/L}$	E$_{D/L}$
wtHTP	`FGNANSA`	-23.1	-23.5	2.4 Å	2.2 Å	<1	0.9±0.1
CD8.02	`EP..KA.`	-17.8	-18.9	2.4 Å	2.3 Å	<1	3.4±0.2
rev68	`.P..KA.`	-19.4	-23.0	4.5 Å	2.8 Å	<1	1.7±0.1
rev69	`E...KA.`	-25.3	-21.3	4.9 Å	5.26 Å	inactive	inactive
rev72	`EP...A.`	-21.4	-22.9	6.0 Å	3.42 Å	<1 or inactive	inactive
rev73	`EP..K..`	-16.9	-22.9	2.0 Å	5.0 Å	>1	5.5±0.4

Table 4.4: Experimental and computational analysis results for variant CD8.02. Binding energies for D- and L-tyrosinol were computed as described in Methods. "dist,D" and "dist,L" correspond to the distance between the phenolic hydroxyl oxygen and the oxygen linked to the heme iron in the docked structure with the lowest computed binding energy.

oxygen in Rev72 and so this mutant was incorrectly predicted to be enantioselective for L-tyrosinol. Mutant Rev73 (CD8.02(A73S)) had an even higher enantioselectivity than its parent clone, $E_{D/L} = 5.5$, which corresponds to a 6-fold improvement over wild type horseradish peroxidase. Computationally, we found that D-tyrosinol was able to closely approach the heme oxygen while L-tyrosinol was not, and so correctly predicted this mutant to be enantioselective for D-tyrosinol.

That a single mutation from the selected CD8.02 sequence back to the wild type at positions 68, 69, or 72 abolishes activity or reduces enantioselectivity suggests that, in the context of the selected CD8.02 sequence (including Ala 73), Glu 68, Pro 69, and Lys 72 are all required for catalytic activity and high $E_{D/L}$. In principle, mutant Rev73, which had a higher enantioselectivity than CD8.02, should have been encoded in the exhaustive library and selected in the selection for $E_{D/L}$. Two possible explanations are that a 5-fold over-sampling of the theoretical sequence space in the physical library was not sufficient to include a copy of each possible sequence, and that Rev73 had other properties that were selected against, such as a lower expression level in yeast. Computationally, the fair agreement between the computational and experimental results was interesting, given that the catalytic reaction was not modeled in detail. The distance between the phenolic hydroxyl and the reactive center of the heme does appear to be an important factor in the ability of horseradish peroxidase to abstract a hydrogen atom from the substrate. However, this computational model

was not able to predict the enantioselectivity in all cases, indicating that other factors must be at play.

4.2.5 Analysis of correlations in enzyme dynamics

We have also explored the role of distal mutations in the context of dynamics and correlated motions. Correlated motions can be elucidated from a single structure by normal modes analysis, and simulation studies have suggested that a small number of modes characterize the overall dynamics of a protein [129]. When motions of atoms far from the active site are correlated with those of the catalytic or active site residues, these distal locations may be important for enzymatic catalysis and mutation at these distal positions may be able to influence catalysis from a distance. We performed a normal modes analysis of horseradish peroxidase as described in Methods, producing a matrix of correlation coefficients between each pair of C_α atoms in the enzyme. Figure 4.6 shows this matrix, and one can see that some pairs of residues do indeed have strong positive or negative correlated motions. The motion of the active site (roughly between residues 60 and 80) is correlated with residues in several other parts of the protein, which are highlighted with different colors in Figure 4.6. The residues whose motions are correlated with that of the active site offer useful suggestions for locations that could be targeted for mutagenesis when designing new mutant libraries for testing.

4.3 Discussion and Conclusion

Our active-site-directed libraries focused attention on only five positions: 68, 69, 72, 73, and 74. This approach allowed us to discover the highly enantioselective mutant CD8.02, where three simultaneous changes from the wild type sequence were likely required to achieve the measured enantioselectivity. The cooperativity between the amino acids allowed at these positions might be especially strong because of the proximity of these residues to the enzyme active-site. In contrast, error-prone PCR would have a very low probability of locating this three-residue mutant, making our

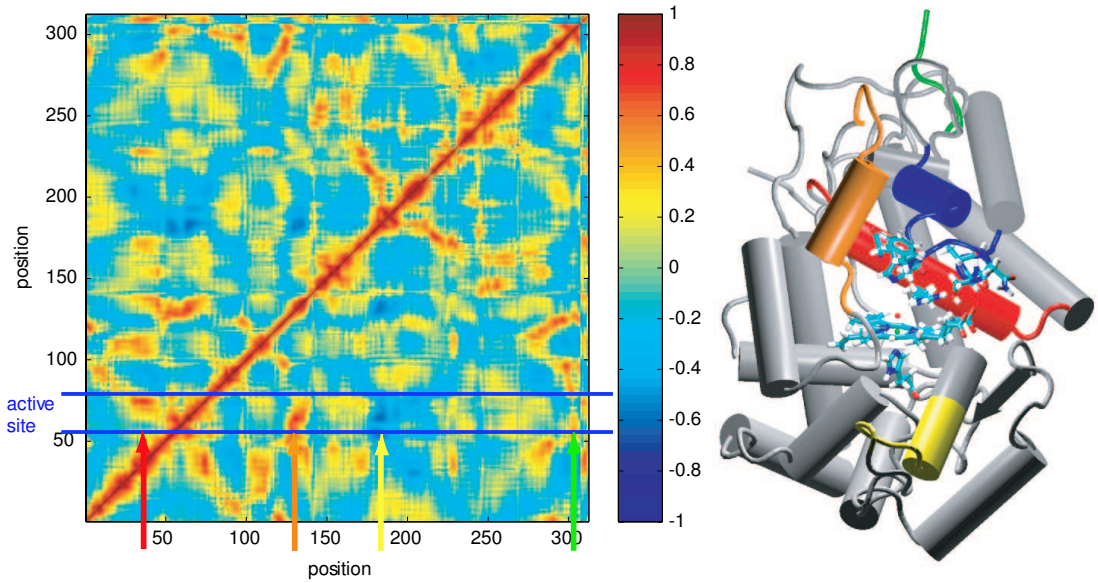

Figure 4.6: Correlation coefficients for each pair of C_α atoms in horseradish peroxidase, determined by normal modes analysis. The correlation coefficient for each pair is indicated by color, ranging from blue (strong negative correlation) to red (strong positive correlation). The active site loop residues are highlighted in blue, and the regions with strong positive or negative correlation are highlighted by arrows whose colors correspond to colored regions in the horseradish peroxidase structure. The heme, catalytic residues, and residues 68, 69, 72, 73, and 74 are shown in atom colors.

approach very attractive for future enzyme design projects where the enzyme structure and active-site location are known. The results of selecting mutants from both the exhaustive and computationally-reduced libraries also helped us better understand the limitations of our computational model and the impact of our assumptions. This work provides valuable insights for those attempting other enzyme engineering projects.

This joint theoretical and experimental approach has produced a better understanding of this enzyme's motions, allowed sequence space, and active-site plasticity. Together this set of studies provides a unique view of the sequence space available for enzyme modification and explores a potentially powerful combination of theory and experiment for re-engineering enzyme function. The techniques developed for this problem may also be generally applicable to other hard protein design problems, involving other functions, in the future.

4.4 Methods

4.4.1 Structure preparation

The horseradish peroxidase structure was taken from the Protein Data Bank (PDB code 1HCH; 1.57 Å resolution [10, 123]). This structure contains a heme group with an oxy-ferryl oxygen linked to the heme iron. All water molecules were removed from the structure, except the 17 waters that made at least three hydrogen bonds with the protein. Hydrogen bonds with other waters were also counted when both waters were completely buried. The flip states for asparagine, glutamine, and histidine residues were changed when the local hydrogen bonding pattern favored the alternative conformation. Hydrogen-atom positions were added to this structure using the HBUILD facility [93] and the PARAM22 parameter set [13] within the CHARMM computer program [94].

4.4.2 Computational sequence and structure search

The computational search over all possible amino acids allowed at the five active-site positions was done using the dead-end elimination and A* algorithms [20, 21], along with the appropriate approximations to make this "inverse design" possible. The two key approximations are that the enzyme backbone remains rigid during the sequence and structure search and that only discrete conformations of the amino acids are allowed. We used the dead-end elimination algorithm to find the global minimum energy sequence and structure for the five active-site positions, then used the A* algorithm to create a list of sequences and structures at those positions, rank-ordered by their energies. A backbone-independent rotamer library was used for these calculations [130, 131] expanded by +/− ten degrees in both the Chi_1 and Chi_2 angles. The pairwise energy function for these calculations was the sum of van der Waals, solvent-accessible surface area [39], and a Coulombic elecstrostatic term with a dielectric constant of 4 times the distance between each pair of atoms. We node that re-evaluation of these energies using the more accurate (and more computationally expensive) Poisson-Boltzmann continuum electrostatic model did not significantly change the distribution of amino acids at each position in the designed library. More details about this implementation of the dead-end elimination and A* algorithms are given elsewhere [132, 133].

4.4.3 Substrate docking and binding energy calculations

Poses of L- and D-tyrosinol docked into the horseradish peroxidase active site were generated using the Induced-Fit Docking procedure [134] within the Maestro environment from Schrödinger (New York, NY). Mutations were made to the 1HCH structure within Maestro (F68E, G69P, N72K, S73A; all 4 mutations and all combinations of 3 mutations), and then multiple docked conformations were generated using Induced-Fit Docking where the active-site histidines (H42, H170) were not allowed to move, and residues F68 and F69 were removed during the initial docking. The SP procedure was used for the initial docking and the XP procedure was used for the final

docking into the rebuilt structure, and 2–17 docked conformation were generated for each ligand into each mutant.

Binding free energies were calculated using CHARMM for each of the generated docked conformations. These energies were computed as the sum of van der Waals, solvent-accessible surface area, and continuum electrostatic interaction and solvation terms. The substrate was rigidly minimized in the bound-state structure, using ABNR in CHARMM, before calculation of the rigid-body binding energy. The van der Waals contribution to binding free energy was calculated with the PARAM22 parameter set in CHARMM. The solvent-accessible surface area contribution was calculated with the surface area contribution to the free energy calculated as 5 calories per mole per square Angstrom plus a constant of 860 calories per mole [39]. Electrostatic potentials were calculated by solution of the Poisson-Boltzmann equation, using a locally-modified version of DELPHI [19, 26, 27] with PARSE atomic radii and partial atomic charges [39], a solvent dielectric of 80 with a bulk ionic strength of 0.145M, a solute dielectric of 4, and a 2.0-Å ion exclusion layer. The averaged result of 3 translations on a 129x129x129 cubic grid was used in all cases.

4.4.4 Construction of the exhaustive and computationally-focused active-site libraries by saturation mutagenesis

The exhaustive and computationally-reduced libraries were generated by Daša Lipovšek according to the following protocol. The libraries were constructed by replacing the BstBI-EagI fragment of the wild type HRP gene with a synthetic DNA fragment randomized at amino acid positions 68, 69, 72, 73, and 74. The synthetic DNA fragment was assembled from two defined oligonucleotides, c1 (GTTGTGACGCATCGATCTTGTTAGACAACACAACATCATTTC-GAACAGAGAAAGATGCG) and c2 (CTGCGCAGGATACAGTTCTTGGGCAT-GCACTCTCCACGGCCGCCTTCATTCTGTCAATCACAGGAAATCCGC), and one randomized oligonucleotide, rC (CATCATTTCGAACAGAGAAAGATGCG11A ACGCA111CGCGGATTTCCTGTGATTGACAGAATG), where "1" stands for ei-

ther an equimolar mixture of codons encoding for the 19 non-cysteine amino acid residues (in the exhaustive case) or a mixture of codons representing the computationally-selected amino acids. The triphosphoramidite codon mixture was purchased from Glen Research (Sterling, VA), and the randomized oligonucleotide was synthesized manually by Trilink (San Diego, CA). The oligonucleotides were assembled using KOD Hot Start Polymerase (Novagen). First, 20 pmol of oligonuclotide c2 and 10 pmol of oligonuclotide rC were combined with 1 U of KOD Hot Start Polymerase in 50 μL of KOD Mix (1 $\times$ KOD buffer, 0.2 mM dNTP mix, 1 mM MgSO4, 1 M betaine, and 3% DMSO). The oligonucleotides were denatured for 2 min at 95°C; subjected to ten cycles of 30 sec at 94°C, 30 sec at 58°C, and 1 min at 68°C; and, finally, incubated for 10 min at 68°C. Twenty pmol of c1 was added to the mixture in 2 μL, and the thermocycling program was repeated as above. The resulting double-stranded DNA fragment was ethanol-precipitated, and 2 μg of the product were amplified 10-fold by limiting the amounts of the PCR primers pc1 (GTTGTGACGCATCGATCTTGTTAGAC) and pc2 (CTGCGCAGGATACAGTTCTTGGGC), and using 20 cycles of the program described above with 30 U KOD Hot Start Polymerase in 1.5 mL of KOD mix. The amplified DNA fragment ("HRP-C insert") was again ethanol-precipitated and resuspended in dH2O at 0.6 μg/μL. The pCT2-HRP plasmid missing the BstBI-EagI fragment was prepared by a sequence of restriction digests (EagI, BsssHII, and BstBI), followed by purification on Qiagen PCR-Purification columns and ethanol precipitation. The gapped pCT2-HRP plasmid and HRP-C insert, which overlapped in sequence with the ends of the gapped pCT2-HRP plasmid by 41 nucleotides both upstream of the BstBI restriction site and downstream of the EagI site, were cotransformed into EBY100 following the established protocol [135]. Twenty μg of gapped pCT2-HRP and 30 μg of HRP-C insert were transformed into 1 mL of electro-competent EBY100, yielding a yeast-surface-display library of estimated 9.0 $\times$ 10^7 independent transformants, which is larger than the 2.5 $\times$ 10^6 possible sequence permutations permitted by library design. Of the 24 clones from library HRP-C that were sequenced, 22 conformed to the library design and 2 showed protein truncations due to frameshift mutations.

4.4.5 Selection of enantioselective horseradish peroxidase variants using yeast-display and fluorescence-activated cell sorting (FACS)

Enantioselective variants were selected by Daša Lipovšek according to the following protocol. Double-labeled yeast cells were sorted on a Becton Dickinson Aria high-speed cell sorter with 488-nm and 635-nm lasers at the rates of 6,000-10,000 cells per second. Gates were adjusted to collect the yeast cells positive for Alexa-633 signal that also had the highest Alexa-488 signal (for the positive selection rounds 1, 3, 5, and 7) or the lowest Alexa-488 signal (for the negative selection rounds 2, 4, 6, and 8). Of the Alexa-633-positive cells, the 1% of the cells with the highest Alexa-488 signal was collected in round 1, whereas 0.5% of the cells with the highest Alexa-488 signal was collected in rounds 3, 5, and 7. Conversely, of the Alexa-633-positive cells, 3% of the cells with the highest Alexa-488 signal was excluded in round 2, and 10% of the cells with the highest Alexa-488 signal was excluded in rounds 4, 6, and 8. Selected cells were collected in 0.5 mL of SD-CAA, pH 4.5, containing 50 μg/mL kanamycin, 100 U/mL penicillin G, and 200 U/mL streptomycin, then grown to saturation in 5 mL of the same media by shaking them at 30°C for 2 days before being induced and labeled for the next round of sorting.

4.4.6 Normal modes analysis

In a regular molecular dynamics simulation, the correlation coefficient for two particle coordinates $\vec{r}$ and $\vec{s}$ over N time points is given by:

$$\mathrm{corr}_{i,j} = \frac{\sum_{i=1}^{N}(\vec{r}_i - \vec{r}_{\mathrm{mean}}) \cdot (\vec{s}_i - \vec{s}_{\mathrm{mean}})}{\sqrt{\sum_{i=1}^{N}(\vec{r}_i - \vec{r}_{\mathrm{mean}})^2 \sum_{i=1}^{N}(\vec{s}_i - \vec{s}_{\mathrm{mean}})^2}}$$

Normal mode analysis uses a minimized structure and makes a harmonic approximation to obtain vibrational motions at different frequencies, which can then be superimposed to obtain a picture of the molecule's global fluctuations. Therefore, this

technique does not require a simulation of the protein's motion in time. This effect has been studied in bovine pancreatic trypsin inhibitor and hen egg white lysozyme [136] and in human immunodeficiency virus (HIV)-1 protease, TRP repressor, calmodulin, and troponin C [137, 138]. To obtain the particle correlation coefficients in a normal mode analysis, the second derivative of the minimized potential with respect to the Cartesian coordinates is taken, yielding a matrix. The eigenvalues and eigenvectors of this matrix are then computed, which correspond to the normal mode vibration frequencies and directions, respectively. This matrix must be computed when the potential is at a minimum for this second-derivative information to be meaningful, and so the structure was first completely minimized using the Adopted Basis Newton-Raphson (ABNR) method within the CHARMM minimization facility with a final minimization of 200,000 steps of size of 10^{-9} with an energy tolerance of 0.0. The correlation coefficient between two particles r_i and r_j over N normal modes is then given by:

$$\mathrm{corr}_{i,j} = \frac{\sum_{k=1}^{N} (r_{ix} r_{jx} + r_{iy} r_{jy} + r_{iz} r_{jz})}{\sqrt{\sum_{k=1}^{N} (r_{ix}^2 + r_{iy}^2 + r_{iz}^2) \sum_{k=1}^{N} (r_{jx}^2 + r_{jy}^2 + r_{jz}^2)}}$$

where each r_{ix} represents $\frac{a_k^{ix}}{\omega_k^2}$, the x-displacement for particle i (the x-component of the eigenvector of particle i in normal mode k) divided by the square of the frequency of normal mode k (which is the square root of the absolute value of the eigenvalue).

Chapter 5

Computationally mapping sequence space to understand evolutionary protein engineering[1]

Abstract

Evolutionary protein engineering has been dramatically successful, producing a wide variety of new proteins with altered stability, binding affinity, and enzymatic activity. However, the success of such procedures is often unreliable, and the impact of the choice of protein, engineering goal, and evolutionary procedure is not well understood. We have created a framework for understanding aspects of the protein engineering process by computationally mapping regions of feasible sequence space for three small proteins using structure-based design protocols. We then tested the ability of different evolutionary search strategies to explore these sequence spaces. The results indicate that sequences compatible with a given protein fold are close together in sequence space but are non-uniformly distributed. The evolutionary relationships among feasible sequences reveal hub-like sequences that serve as particularly fruitful starting sequences for evolutionary search and simulation results point to a non-intuitive relationship between the error-prone PCR mutation rate and the number of rounds of replication. Moreover, genetic recombination procedures were examined, and trade-offs relating sequence diversity and search efficiency were identified. This framework allows us to consider the impact of protein structure on the allowed sequence space and therefore on the challenges that each protein presents to error-prone PCR and genetic recombination procedures.

[1]This text has been submitted as: K. A. Armstrong and B. Tidor. Computationally mapping sequence space to understand evolutionary protein engineering.

5.1 Introduction

The observation that proteins can be evolved to fulfill different functional roles has been harnessed as a fundamental technology by protein engineers and has had considerable practical application in medicine, industry, and basic science. Laboratory protocols have been developed for evolutionary experiments in which rounds of sequence variation are followed by selecting or screening for desired activity, sometimes in an iterative fashion [100, 101, 139]. While this class of approaches has produced change-of-function variants with important properties across a variety of application areas, success is by no means guaranteed. The same design goal is generally not performed repeatedly in the same laboratory or between laboratories either with the same or different procedures, so it is difficult to judge the uniqueness of the results and the robustness of the procedures. Nevertheless, there is a sense that the success of these procedures varies between projects and design goals, and it is unclear the extent to which this variability reflects differences in the challenges attempted, in the technology applied, in the sizes or properties of the space of protein sequences searched, or something about the fundamental adaptability or designability of the parent protein. The mutation rate and protocol in the variation procedure influence the distribution of both DNA and protein sequences created, and the screening or selection capabilities limit the number of sequences that can be examined. These factors combine to impact the efficiency of the sequence space search and the likelihood of discovery of a new protein variant with desired properties.

Methods for protein engineering include computational approaches and experimental techniques, as well as hybrids of the two. Computational protein design methods include structure-based modeling of detailed electrostatics, packing, and other interactions, as well as data mining in sequence space [110, 140–142]. Computation has also been used to address the designability question, where both atomistic and lattice models have been used to estimate the number of sequences that will fold to a given protein structure [143, 144]. While this metric does tell us something about designability, it does not address the accessibility of each of these sequences

by standard experimental techniques. Computation has also been used to simulate evolutionary engineering experiments, such as error-prone PCR and gene shuffling, and has provided valuable insights into the effects of experimental parameter variation on the efficiency of the sequence space search [145–147]. The current work combines the computational identification of feasible sequences with simulations of evolutionary engineering procedures to address the relative effectiveness of techniques for experimental search of sequence space.

Experimental laboratory evolution approaches, at least in principle, can be used to engineer any new function for which there is a screen or selection. They have been applied successfully to modify protein stability, binding affinity and specificity, and enzymatic properties such as catalytic rates, substrates accepted, reactions catalyzed, stereoselectivity, and operating conditions [100, 101, 139, 148, 149]. Applied to proteins, different laboratory evolution techniques use one or more DNA sequences as substrate for a partial randomization procedure that results in a large library of sequences. The DNA-level genotype of each sequence is physically linked to the protein-level phenotype to enable sequencing of the gene associated with the newly isolated protein activity. Experimental platforms for the selection or screening step of laboratory evolution include cell surface display (on yeast or bacteria) [150–152], phage display [153, 154], mRNA display [155], ribosome display [156], and in vitro compartmentalization [157]. Cells have also been used for in vivo compartmentalization when the goal protein function is linked to cellular metabolism [113, 158] or fluorescence [159–161]. Each of these systems has been used as the foundation for an assay for screening randomized sequence libraries that have been generated by error-prone PCR, genetic recombination, saturation mutagenesis, or by computational protein design. Although the number of unique sequences that can be screened by each assay is not well established, currently not more than 10^{15} sequences can be examined in one experiment, and typical effective experimental capacity is probably significantly lower, perhaps 10^6–10^8. Error-prone PCR randomization strategies generally make low numbers of DNA mutations because a low mutation rate has been shown to typically maximize the number of functional results [9, 162, 163]. A range

of rates is possible experimentally [164], however, and some laboratories have found higher mutation rates to be beneficial [165, 166]. Genetic recombination, focused mutagenesis, and computational design techniques can have higher mutation rates and can locate design solutions more distant from wild type [158, 167–171].

In hybrid computational–experimental studies, computational methods have been used as the sequence variation step for evolutionary engineering approaches [113, 172] and to select crossover sites for recombination in gene shuffling experiments [173, 174]. They have also been used to choose residue positions most likely to accept mutation [114, 175]. Our work serves as a new and complementary approach, using computationally designed sequences to guide evolutionary experiments toward feasible portions of the sequence space.

A useful framing for studying different evolutionary protein engineering approaches identifies a search space (for example, all the sequences explicitly considered), which can be characterized by its size as well as other properties; the framing also identifies the mechanism of effecting the search in this space. It could be critical to evaluate the thoroughness of a search (e.g., the fraction of the potential search space actually examined) as well as its diversity (e.g., the relative differences among the sequences searched; the distribution of goal sequences found within the search space). Each search technique includes a starting sequence or multiple starting sequences and eventually produces a set of protein sequences for screening or selection. The search mechanism biases the set of sequences created, and because no more than 10^{15} sequences can be screened (much less than the possible $20^{100} \approx 10^{130}$ sequences possible for a protein with 100 residues), this bias influences the choice of the final engineered protein. A vanishingly small fraction of these 10^{130} sequences will have the new desired function, and so the distribution of desired protein variants in the sequence space is another important factor that is, in general, hard to quantify but relates to the difficulty of the engineering experiment.

We therefore ask how the biases of the randomization procedures influence the generation of diverse functional libraries and whether the expected locations of desired proteins in the search space point toward beneficial experimental sequence space

search procedures. We analyze simulated error-prone PCR and genetic recombination experiments and explore how changes in experimental parameters can influence the efficiency of the sequence-space search for allowed sequences.

We develop and apply a model for evolutionary search to understand better the key factors affecting the search capabilities and outcome of evolutionary engineering experiments. This model includes a definition of the allowed sequences for a protein design problem, and we test the accessibility of these sequences in the search space by simulating different protein engineering experiments. Allowed sequences are defined in our model as those computed to be stable on the wild-type crystal structure backbone fold. Simulated evolutionary experiments start from any allowed sequence, including the wild type, and we test the ability of each simulation to find these allowed sequences. Our representation of these foldable sequences is a "sequence-space graph" that depicts evolutionary paths between sequences. A node in this type of graph represents a protein sequence and for the present study an edge between two nodes indicates that only a single DNA mutation is required to interconvert encodings for those two protein sequences. Because protein folding is a prerequisite for function, an evolutionary search technique that locates many sequences computed to be foldable should also successfully generate sequences with new function. This principle of designing for a stable fold as a prerequisite for locating new function has been used successfully in the past, both computationally to design a beta-lactamase inhibitor and experimentally to design a cytochrome P450 enzyme family [113, 176].

We have computationally explored the sequence space computed to be foldable for three small proteins, allowing us to design, simulate, and evaluate strategies for protein engineering via directed evolution. The different backbone conformations of these proteins provide interesting constraints on the available sequences, influencing the possible evolutionary paths through the sequence space. Computer simulations of error-prone PCR and genetic recombination were used to explore these sequence spaces and demonstrated that using our sequence-space graphs to guide the choice of starting sequence can help generate more diverse libraries of foldable sequences. We explore the influence of mutation rate and number of rounds of mutation on error-

prone PCR experiments and show that the effective diversity of the sequence library produced increases when the simulation starts from a sequence with more allowed proteins close by in mutational space (or sequence-space graph "hubs"). We also explore the relationship between the computed protein energy and its evolvability. Additionally, our model shows that foldable mutants with several changes from the starting sequence are more easily found by genetic recombination than by error-prone PCR but that the similarity between the allowed sequences for our chosen folds limit the possible number of mutations achievable by recombination. Finally, our work provides a framework for understanding and analyzing protein engineering techniques through the effect of structural and stability constraints and for further exploring the variable success of protein design projects.

5.2 Results

5.2.1 Sequence-space graph topology

Sequence-space graph topologies vary by protein structure. Computational protein design techniques (described in Methods) were used to generate an exhaustive list of low-energy sequences for three small proteins: the Pin1 WW domain (WW-domain), bovine pancreatic trypsin inhibitor (BPTI), and an immunoglobulin G-binding domain from streptococcal protein G (B-domain), whose crystal structures are shown in Figure 5.1A [10, 177–179]. The 3,000 lowest energy sequences computed to be stable for each of the three protein backbones were considered an estimate of the set of allowed (or "foldable") sequences for that structure. We then tested the ability of different sequence-space search procedures to find the most diverse sets of these sequences. These sequences were transformed into sequence-space graphs, providing a graphical depiction of the allowed sequence spaces, shown in Figure 5.1B. Each node in this type of graph represents a protein sequence. An edge between two nodes indicates that a single DNA mutation will interconvert encodings for the two proteins corresponding to those nodes. Surface positions were excluded from the calculations

	WW-domain	BPTI	B-domain
number of components	144	320	39
average component size	20.76	9.29	76.82
largest component size	613	1251	2854
average degree	4.18	3.28	5.64
average clustering coefficient	0.094	0.064	0.087
average of shortest paths	8.94	9.69	10.52
diameter	32	26	31

Table 5.1: Properties of the sequence-space graphs in Figure 5.1B. A component is a set of connected nodes, and the size of a component is the number of nodes in that connected set. The degree of a node is the number of neighbors it has, or the number of links touching that node. The clustering coefficient describes local topology: the fraction of each node's neighbors that are neighbors of each other. The shortest path is calculated between every pair of nodes that are connected to each other in the network. The maximum of these shortest paths is defined as the graph diameter. The average of these path lengths is also recorded.

to focus attention on interrelated buried positions most critical for stability.

The structural constraints of each protein fold influence the type of sequences allowed for its core and therefore the topology of its sequence-space graph. Graph properties such as component sizes, degree distributions, and clustering coefficients can help to characterize the sequence space through which evolutionary routes are taken. Multiple, separated components exist in each sequence-space graph, as shown in Figure 5.1B and quantified in Tables 5.1 and 5.2. This means that single DNA mutations can explore within a component, but that multiple, simultaneous mutations are necessary to cross between components. The sizes of these separated sequence-space graph components differ from one to 2,854, indicating a very wide range. In the B-domain sequence-space graph, nearly all the protein sequences computed to be foldable are part of one giant component, indicating that only single DNA mutations are required to move throughout almost the entire sequence space. The BPTI and WW-domain sequence-space graphs are more fractured, and single DNA mutations would only allow mutation within smaller subsets of these graphs.

Sequence-space graph nodes are well-connected but with a variable number of neighbors, making evolution between foldable protein sequences efficient but non-uniform. Within each sequence-space graph component, each protein sequence is connected to

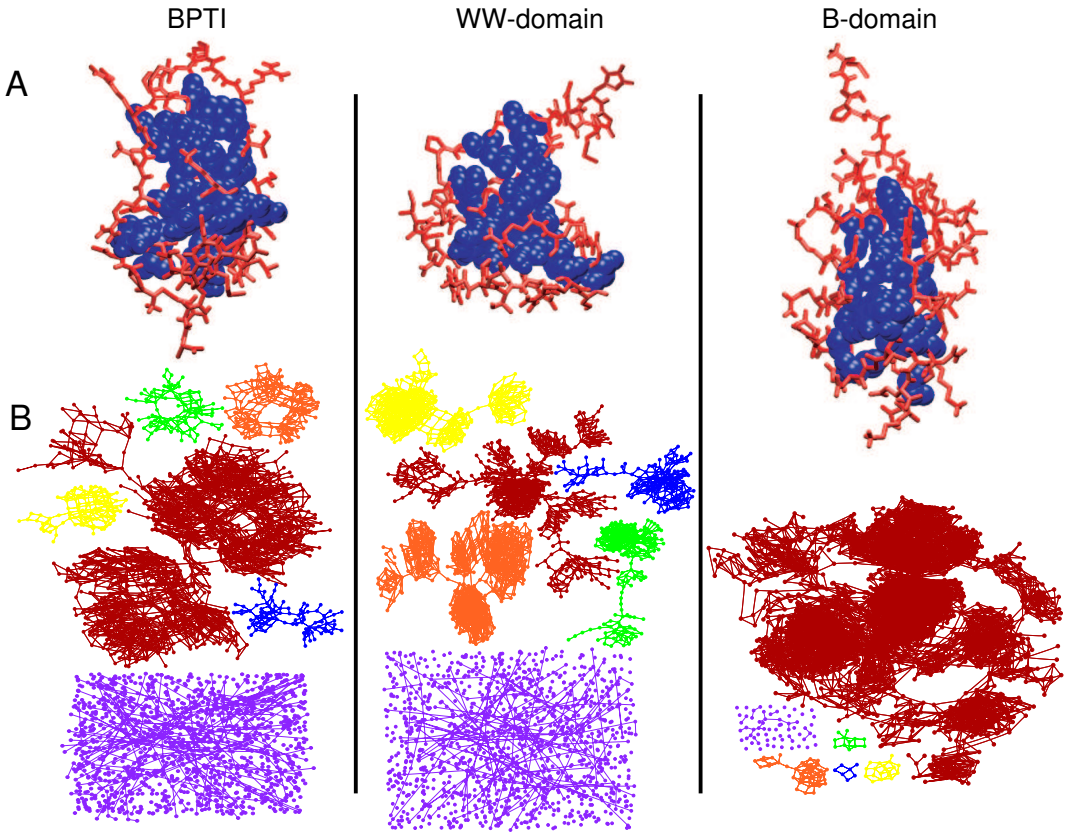

Figure 5.1: (A) Crystal structures of the bovine pancreatic trypsin inhibitor (BPTI), pin1 WW domain (WW-domain), and an immunoglobulin G binding domain from streptococcal protein G (B-domain) (PDB codes 1BPI, 1F8A, and 1IGD at 1.09, 1.84, and 1.10 Å resolution, respectively [10, 177–179]). Core residues allowed to mutate are colored blue. (B) Sequence-space graphs for BPTI, WW-domain, and B-domain with the five largest components colored in red, orange, yellow, green, and blue in decreasing size order. The remaining vertices are colored purple. Component sizes are given in Table 5.2.

	BPTI	WW-domain	B-domain
red	1251	613	2854
orange	158	492	44
yellow	102	437	19
green	94	205	12
blue	92	149	7
purple	1303	1214	64

Table 5.2: Sizes of the graph components in Figure 5.1B. The largest five sequence-space graph components are colored red, orange, yellow, green, and blue, in decreasing size order, and the remaining vertices that are not part of the largest five components are colored purple.

on average 3.3 to 5.6 other protein sequences. This node degree is the number of sequences computed to be foldable that are within one DNA mutation of that sequence. In the full space of all possible protein sequences, with no requirement for fold stability, the average node degree would be between 66 and 119 because each amino acid can reach on average 6 to 7 other protein sequences with a single DNA mutation, and our cores varied in size from 11 to 17 residues. The low observed degree in our graphs indicates that the vast majority of single non-silent DNA mutations away from each native core protein sequence would not yield a protein sequence computed to be foldable.

However, the observed node degrees in our sequence-space graphs are large compared to the expected node degree of zero when 3,000 random sequences of 11 to 17 residues are chosen. This points to the uneven distribution of our foldable sequences in sequence space caused by the selected backbone structures. We quantify this bias by computing a "density" of sequences in a region of the search space. The fraction of sequence space expected to fold to a specific structure is tiny. For example, studies of hydrophobic/polar patterned lattice models indicate that approximately 10^{-9} of the possible sequences for a protein of 100 residues will fold to the same structure [180]. If these sequences were spread evenly throughout the search space, multiple mutations would be required to move between every pair of sequences. We created sets of 3,000 random sequences of length 11, 12, and 17 to compare to the B-domain, WW-domain, and BPTI sequence spaces. In each case, we define the "corner" of the sequence space occupied by that set of sequences as the product over all positions of the number of amino acids allowed at that position. The sequence density in this corner of sequence space is then 3,000 divided by the size of the corner. The size of the corner is 3.0×10^{11}, 4.3×10^{9}, and 3.1×10^{9} for BPTI, WW-domain, and B-domain respectively. In contrast, for the sets of random sequences the corners are of size 1.3×10^{22}, 4.1×10^{15}, and 2.0×10^{14} for sequences of length 17, 12, and 11, respectively. Therefore, our 3,000 sequences computed to be foldable are clustered together in a corner of the sequence space with a density 10^{5} to 10^{10} times greater than that for random sequences. Although the density of foldable sequences in their

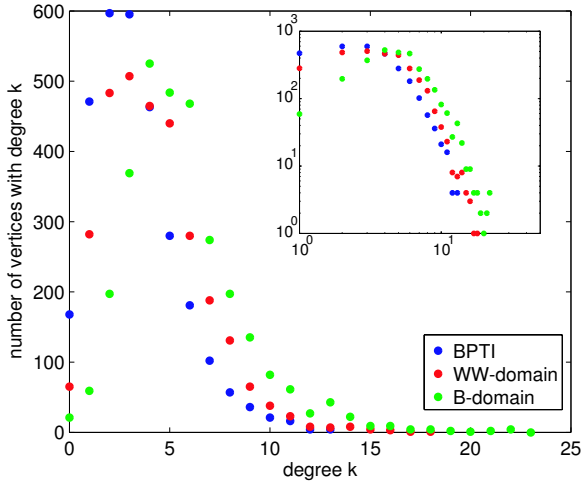

Figure 5.2: Degree distributions for the sequence space graphs shown in Figure 5.1B. The degree of a node is the number of edges touching that node in its sequence-space graph.

corner of the search space is low compared to the density of all sequences, they are close enough together to be connected to one another in our graphs, making evolution between them by single DNA mutations more efficient.

The node degrees in our sequence-space graphs are variable, as shown in Figure 5.2. Single DNA mutations from protein sequences with many neighbors (called "hubs") yield more sequences computed to be foldable than they do from others. There are more of these hub nodes than one would expect from a random graph with the same number of nodes and edges; however, there are fewer hubs than one would expect in a pure scale-free graph.

5.2.2 Error-prone PCR simulation

The number of rounds of error-prone PCR and the screening limit influence search success more strongly than does the mutation rate. We performed computational simulations of evolutionary experiments. The essential idea is that multiple rounds of a protocol to introduce sequence variation were applied to a random starting sequence. The sequences need not correspond to allowed foldable sequences but in this work the starting sequence was always one of those computed to be foldable. A set of rounds of sequence variation was then followed by selection. Only functional sequences pass

the selection, and foldability is a prerequisite for functionality in this model. In an initial set of computations, we simulated 9, 12, 15, and 18 rounds of error-prone PCR without selection and studied the resulting sequence distributions as a function of the mutation probability. The resulting protein sequence distributions were characterized by the number of mutations in the final population, the number of unique foldable sequences in the final population, and by counting the number of functional sequences (based on pre-selected arbitrary criteria) in the final population. Here, sets of 30, 300, and 1000 sequences were considered functional; thus, the fraction of functional sequences ranged from 1/100 to 1/3 of the foldable space.

The distribution of the number of DNA sequences at the end of 9, 12, 15, and 18 rounds of error-prone PCR approximately follow Poisson distributions, as described in Methods. The expected number of mutations per sequence is proportional to the mutation probability and the number of rounds of error-prone PCR, where the mutation probability (or mutation rate) is defined as the probability of mutation of each DNA base-pair per generation. As shown in Figure 5.3, diffusion away from the wild-type sequence increased with both increasing mutation rate and increasing rounds of error-prone PCR.

Because the node degrees are low compared to what they would be in an exhaustive sequence-space graph, most of these generated sequences leave the graph at some point during their mutation from the starting sequence in our error-prone PCR simulations. We explored the likelihood of this occurring in our 10,000 simulations of 15 rounds of error-prone PCR at a mutation probability of 0.01 per base-pair per generation. The probability of leaving the foldable sequence space at some point during the simulations was 0.94, 0.87, and 0.82 for BPTI, WW-domain, and B-domain, respectively. The probability of ending the simulation within the foldable space having left it at some point during the simulation was 0.002, 0.004, and 0.006 for BPTI, WW-domain, and B-domain, respectively. Therefore, a typical error-prone PCR simulation experiment produced a small fraction of sequences computed to be foldable, and most of those produced remained on the sequence-space graph throughout the simulation.

Using the same simulations of error-prone PCR at varying mutation rates and

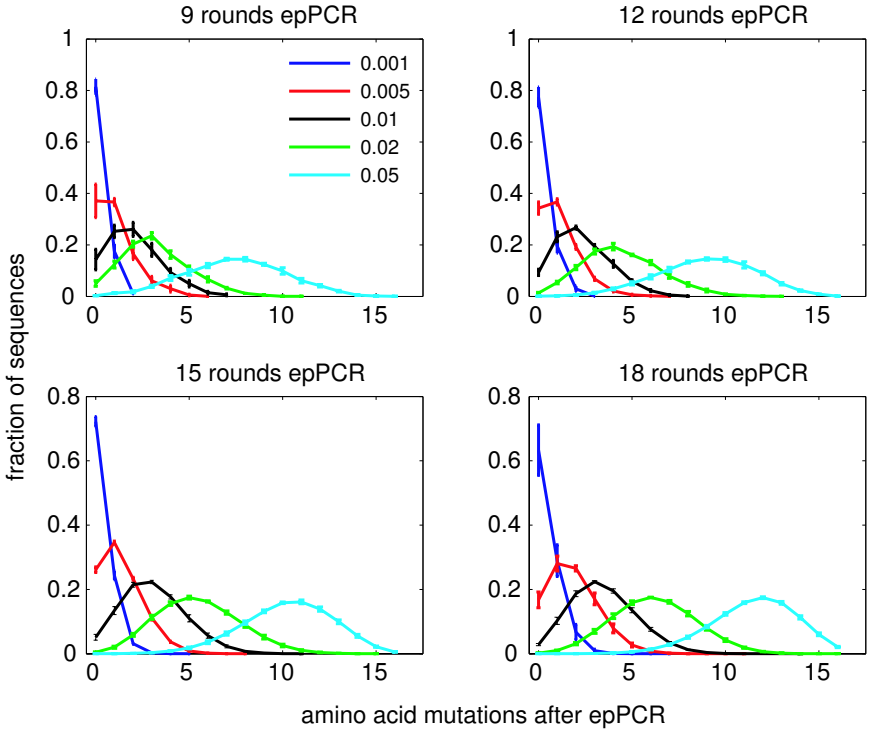

Figure 5.3: Histograms of the number of amino acid mutations per sequence, after 9, 12, 15, or 18 rounds of error-prone PCR. Five different DNA mutation probabilities were used and are probabilities per base-pair per generation. The mean and standard error bars are shown for simulations on 10,000 starting sequences with 17 amino-acid residues.

rounds of replication, we analyze the impact of a screening limit on the diversity of foldable sequences created. The number of unique, foldable sequences produced increased with the number of rounds of error-prone PCR (Figure 5.4A). This is because in each round more of the sequence space was explored as more sequences were produced. However, the fraction of foldable sequences in the screening pool decreased drastically with the number of rounds of error-prone PCR (Figure 5.4B). This is because the number of deliterious mutations increased more rapidly than did the number of allowed mutations. The significance of this observation is underscored by imposing a screening restriction (here of 10^5 sequences) in Figure 5.4C. This cutoff eliminated screening of 62% of the sequences after 18 rounds of error-prone PCR, allowing the large fraction of non-folding sequences in the search space to dominate the screening and making 15 rounds of error-prone PCR preferable to 18 rounds. If this cutoff were made at 10^4 sequences, 12 rounds of error-prone PCR would be preferable and 96% of the sequences from the 18-round experiment would be eliminated. This analysis highlights the tradeoff between increasing diversity and decreasing folding fraction with additional rounds. If the entire pool cannot be screened, the additional rounds may be detrimental.

Figure 5.4 also highlights that the number of rounds of error-prone PCR has more impact on the number of unique foldable sequences generated than does the mutation probability. A two-fold change in the mutation probability yielded a change of no more than 1.17-fold in the number of unique foldable sequences produced and a ten-fold change in the mutation probability yielded a change of no more than 2.5-fold in the number of unique foldable sequences produced. However, a two-fold change in the number of rounds of error-prone PCR caused up to a 7-fold change in the number of unique foldable sequences produced. This is interesting because the number of mutations per sequence is influenced equally by both the mutation probability and the number of rounds of error-prone PCR (Figure 5.3). At a certain number of rounds of error-prone PCR, a lower mutation probability was always more effective at producing more foldable sequences. However, as the mutation rate increases, the number of sequences created does not increase but the generated mutants are more

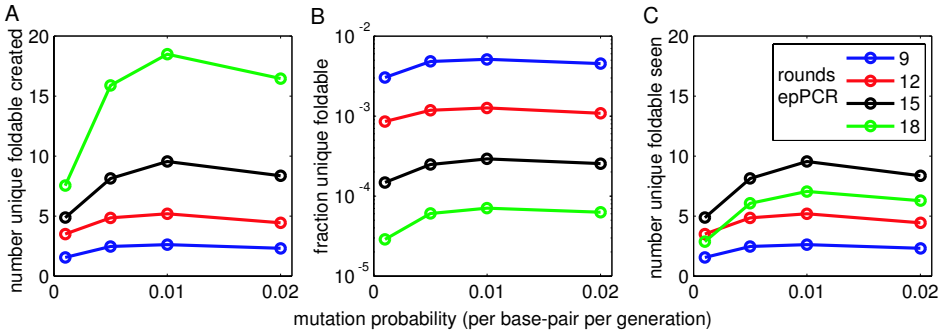

Figure 5.4: Variation of the number of unique foldable sequences identified by search as a function of the mutation rate and the number of rounds of error-prone PCR. All simulations started with sequences computed to be foldable for the WW-domain. Error-prone PCR simulations were performed from random single starting sequences computed to be foldable, at 4 different mutation rates and for 9, 12, 15, and 18 rounds of error-prone PCR. All results are the mean per trial over 10,000 trials. (A) The number of unique foldable protein sequences generated. (B) The fraction of unique foldable protein sequences within the total number of sequences generated. (Peaks still exist in these curves at a mutation rate of 0.01 but are not visible due to the log scale.) (C) The number of unique foldable protein sequences created after a screening limit of 10^5 sequences was imposed.

distant from the starting sequence. However, the requirement for *unique* foldable sequences made more intermediate mutation rates most effective.

Effective library diversity depends on graph topology and the location of the starting sequence. For further error-prone PCR simulations, we held the mutation probability constant at 0.01 per base-pair per generation and the number of rounds of error-prone PCR constant at 15. We then asked how the effective library diversity varies with the local topology of each sequence in our sequence-space graphs. Our measure of diversity is the number of unique foldable sequences generated during the error-prone PCR simulation. In each error-prone PCR simulation, one foldable protein sequence was chosen at random, and a random DNA sequence encoding that protein sequence was chosen as a starting point for error-prone PCR. After fifteen rounds of simulated error-prone PCR, foldable and functional sequences were selected. Averages of these results over 10,000 randomly chosen starting sequences are described.

Figure 5.5A shows the number of unique foldable protein sequences created after 15 rounds of error-prone PCR, plotted separately by the degree of the starting node. Error-prone PCR simulations that started from nodes with a small degree typically resulted in low-effective diversity sequence libraries, whereas simulations started from nodes with larger degrees typically resulted in higher-effective diversity libraries. This is because simulations from larger degree nodes were more likely to remain on the sequence-space graph throughout the simulation, with that probability increasing by 0.012 per degree on average. To put this number in context, the overall probability of staying on the graph was 0.06, 0.13, and 0.18 for BPTI, WW-domain, and B-domain. Therefore, starting from a node with a degree about 10 larger than an average node should approximately double the probability of staying on the sequence-space graph, compared to an average node. Simulations from larger degree nodes were also slightly more likely to return to the graph after having left it at some point in their simulation, with that probability increasing by 0.0001 per degree. In Figures 5.5B-D the results of the same trials are plotted again by the degree of the starting sequence, where at the end of 15 rounds of error-prone PCR only functional sequences were selected, and where 30, 300, or 1000 functional sequences were randomly pre-selected from the 3,000

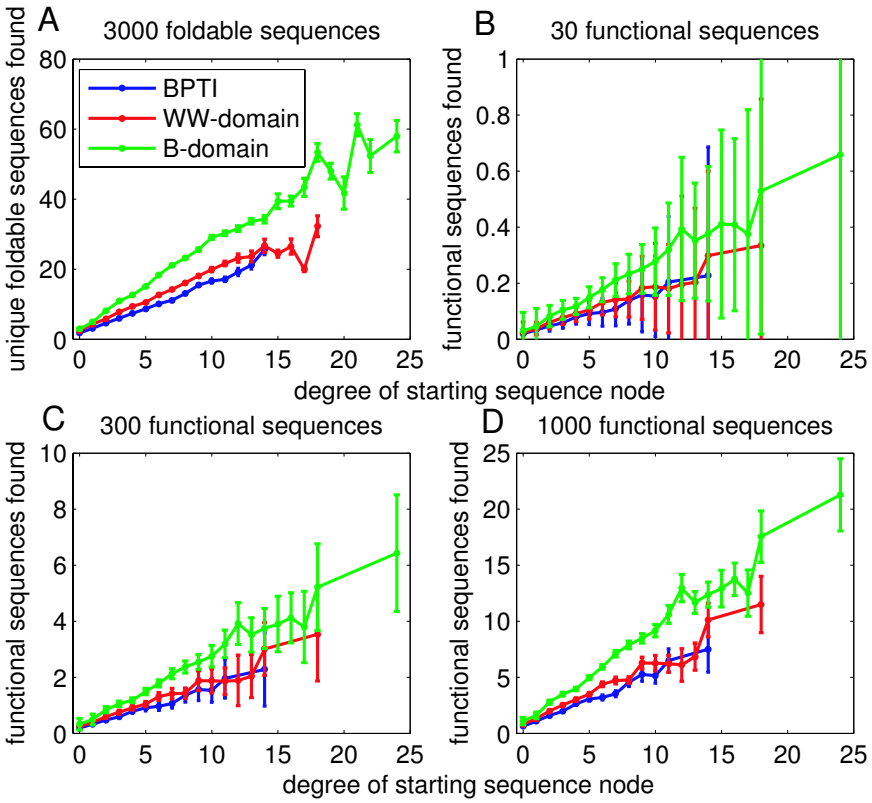

Figure 5.5: (A) The number of unique foldable protein sequences generated after 15 rounds of error-prone PCR simulation with the results recorded separately by the degree of the starting sequence in its sequence-space graph. Standard error bars over 10,000 trials are shown. (B-D) The number of unique functional protein sequences generated after 15 rounds of error-prone PCR simulation. The fraction of the foldable sequence space that was chosen to be functional was 1/100, 1/10, and 1/3 for plots B, C, and D, respectively. The results were recorded separately by the degree of the starting sequence in its sequence-space graph, and standard deviation bars over 1,000 random selections of functional sequences over 10,000 error-prone PCR trials are shown.

sequences computed to be foldable. The standard error for these results was small, ranging from 0.001 to 0.079 functional sequences found, meaning that the average is well established. Nevertheless, the standard deviation bars plotted for these results show very large variance when function is rare in the sequence space. The implication of these results is that any particular set of evolutionary experiments may produce results far from the average when function is rare, and that exhaustive work may be necessary to experimentally determine average behavior to compare methodologies.

These results highlight the principle that the local graph topology surrounding

the starting sequence can be a significant influence on the effectiveness of protein evolution experiments. The degree of a node is influenced both by the number of similar sequences that are also computed to be foldable and by the codon bias of the genetic code. For example, neither negatively-charged amino acid (D or E) can be reached from the amino acids F, L, I, M, S, P, T, C, W, or R by a single DNA mutation, regardless of which codons are used to specify them. Neither positively-charged amino acid (R or K) can be reached from the amino acids N, F, V, or A. A goal protein function requiring multiple mutations could be found by successive single mutations, but excessive library sizes might be required to ensure a high probability of creating such a mutant.

It has been suggested that increasing a protein's stability may also increase its evolvability because preliminary stabilizing mutations may allow more potentially destabilizing mutations that improve function [181]. This principle has been used successfully to design a family of cytochrome P450 enzymes [176]. To explore this concept further we asked whether the computed stability of our protein sequences correlated with the size of the foldable sequence space available. We compared pairs of sequences from our sequence-space graphs where the sequences were identical except for a mutation at a particular position i. For example, one sequence might have an alanine at that position and another might have a valine. We then asked whether the lower-energy sequence from this pair was more or less evolvable at the remaining positions. The number of sequences in the sequence-space graph where position i is constrained (to alanine, for example) is the evolvability of that sequence in the pair. For every core position in BPTI, WW-domain, and B-domain we counted the number of sequence pairs where the lower-energy sequence was more evolvable and the number of pairs where the higher-energy sequence was more evolvable. As shown in Figure 5.6, more frequently the lower-energy sequence was more evolvable (at 23 of 38 positions), although this was not the outcome at every position. The positions where the lower-energy sequence was more evolvable were more often the less-buried positions in the protein structures, but this also was not a requirement for the result. These findings are consistent with a picture in which a stabilizing mutation that

is uncoupled (or only loosely coupled) to functionally important residues is likely to enlarge the feasible space [182, 183], but a stabilizing mutation that strongly interacts with functionally important residues may have more varied effects.

5.2.3 Genetic recombination simulation

Genetic recombination can make multiple mutations and the efficiency of this process depends on the identity between the starting sequences. Making multiple mutations from a starting sequence may be important for evolutionary protein engineering projects when more drastic functional changes are desired. Indeed, multiple mutations have even been shown to be required for new function in some cases, where subsets of the functional mutations were shown to be non-functional [172, 184, 185], including the work presented in Chapter 4. Each genetic recombination experiment started with two or three parent sequences randomly chosen from our sequence-space graph, and two or six new sequences were created in each recombination event. These new sequences could each be a new foldable protein sequence, a copy of a parent sequence, or a protein sequence not computed to be foldable. We ran 10,000 trials of genetic recombination in each sequence-space graph. Figure 5.7 depicts the distribution of amino acid mutations away from the starting sequence for both the error-prone PCR and there genetic recombination experiments. A larger fraction of unique foldable multiple mutants is created by genetic recombination than by error-prone PCR, and the distribution of mutations away from the starting sequences illustrates that achieving three or more mutations in one trial was more likely in the genetic recombination simulations. Because of the similarity between the sequences computed to be foldable (the mean of the pairwise amino acid distances between foldable sequences is 6.3 amino acids with variance 3.5 amino acids) there is relatively limited diversity among sequences selected for recombination. With larger sets of positions varied and greater variability among them, it is likely that larger gains could be seen for genetic recombination.

Figure 5.8A depicts the relationship between the relative similarity of recombined sequences and the success of generating a new foldable sequence. These results point

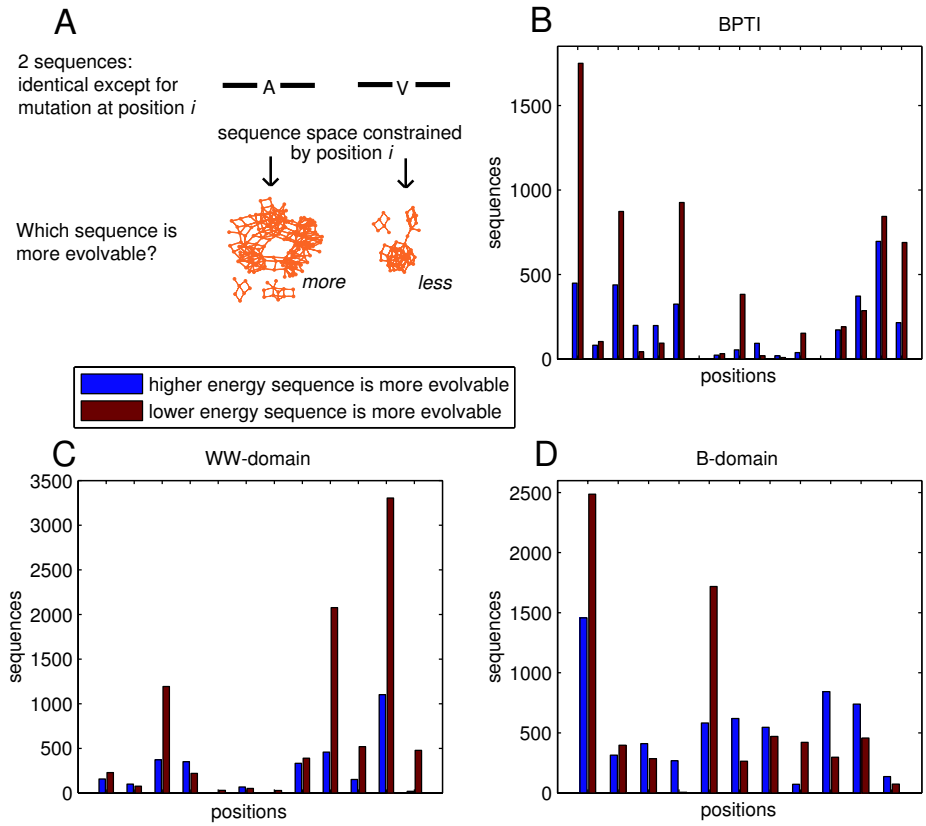

Figure 5.6: Evolvability of sequence pairs with one mutation and different energies. (A) For pairs of sequences with a mutation at only one position, we asked whether the sequence with the lower or higher computed energy was more evolvable. This evolvability is defined as the number of sequences remaining in the sequence-space graph when the variable position in the pair is constrained. This was done for all pairs of sequences found in the sequence-space graphs that had only a single mutation between them, and the results are separated by each core position in order by the residue number in the structure. (B) In BPTI the core positions are 4, 6, 10, 18, 20, 21, 22, 23, 24, 25, 33, 35, 36, 43, 44, 45, and 47. (C) In WW-domain the core positions are 9, 11, 15, 17, 20, 23, 27, 28, 29, 36, 37, and 38. (D) In B-domain the core positions are 8, 10, 12, 14, 31, 35, 39, 44, 48, 57, and 59.

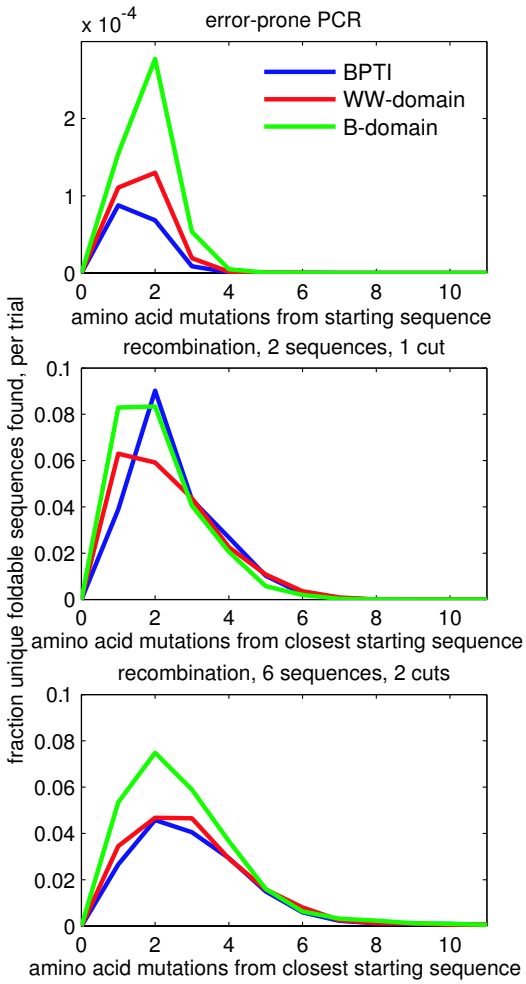

Figure 5.7: The distribution of unique foldable sequences created in each evolutionary simulation as a function of the number of mutations from the starting sequence. (A) Fraction of unique foldable sequences found in each simulation of 15 rounds of error-prone PCR with a mutation probability of 0.01 per base-pair per generation. Each simulation started with one sequence and after each trial the distance of each unique foldable resulting sequence to this one sequence was computed. (B) Fraction of unique foldable sequences found in each genetic recombination simulation started with 2 randomly-chosen foldable sequences. These sequences were cut once and recombined to produce 2 sequences. For each of the two resulting sequences, the floor of the mean distance from the resulting sequence to each of the two starting sequences was computed. (C) Fraction of unique foldable sequences found in each genetic recombination simulation started with 3 randomly chosen foldable sequences. These sequences were cut twice at random and recombined such that 6 sequences were produced. These 6 sequences contained all possible combinations of the three starting sequences such that some fragment from each starting sequence is part of each resulting sequence. The minimum distance from each resulting sequence to the closest starting sequence was computed.

to a region of optimal DNA distance between the starting sequences for recombination where new, foldable sequences were most likely to be generated by a recombination event. When the two parent sequences were very similar to one another, recombination most frequently resulted in one or two copies of the parent sequences; when the two parent sequences were very different from one another, recombination most often produced one or two sequences not computed to be foldable. We repeated these simulations with a different recombination procedure, starting with three randomly-chosen foldable sequences which were recombined around two points to produce six sequences containing fragments from all three starting sequences in all combinations. The results from these simulations were very similar and point to a nearly idential optimal region of DNA distances between the parent sequences (between 0.14 and 1.29 bases per residue for WW-domain).

Figure 5.8B shows the fraction of sequences generated by each recombination event that were both computed to be foldable and are located in a separate component of the sequence-space graph from that of either parent. These results were again separated by the DNA distance between the parent sequences, indicating that a larger difference in sequence identity was necessary to hop to a separate component of sequence space. A new sequence in a new sequence-space graph component was generated with probability 0.05, 0.04, and 0.002 in the BPTI, WW-domain, and B-domain sequence spaces, respectively. Genetic recombination appears to be an excellent method for making large jumps in sequence space with direction toward the portions of sequence-space computed to be foldable. Using an optimal DNA distance between the starting sequences further improves the success of this technique.

5.3 Discussion

Here we have applied protein structural modeling to estimate the space of foldable sequences and carried out simulations of accelerated evolution in the entire sequence space to understand better this important technology. Fundamental to the analysis is the concept that stable folded structure is a prerequisite for function; this is

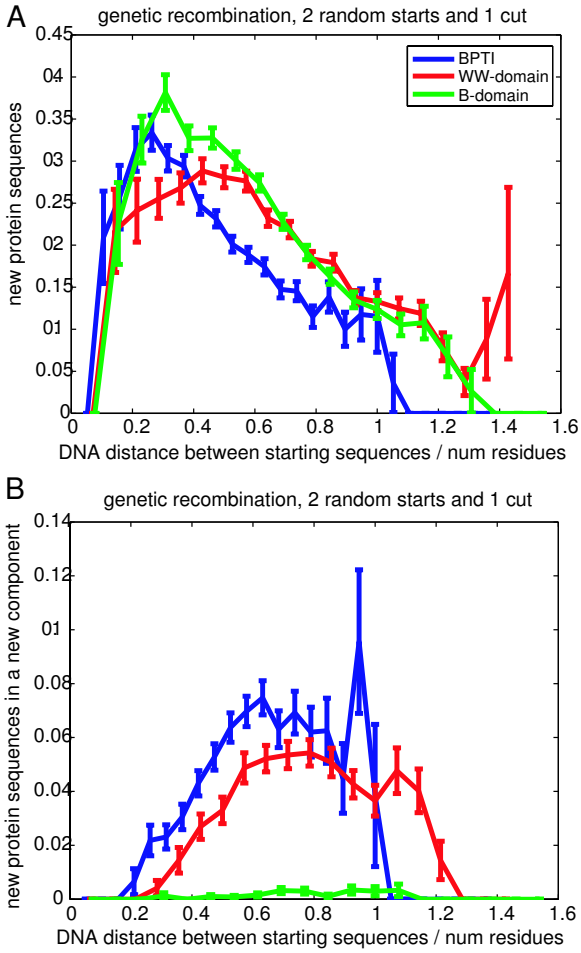

Figure 5.8: Genetic recombination of two sequences, with results separated by the distance between parent sequences. Each genetic recombination trial started with two parent sequences computed to be foldable and the sequences were cut once at a random location and recombined. The number of foldable sequences created per trial were plotted at each DNA distance per residue between the parent sequences. Standard error bars over 10,000 trials are shown. (A) Sequences that are computed to be foldable and are different from either parent sequence are included. (B) Only sequences that are computed to be foldable, are different from either parent sequence, and are in a different graph component from that of either parent sequence are included.

usually but not always the case for human proteins. The results indicate that the foldable sequences are non-uniformly distributed in sequence space, as if clustered in one corner (with a sequence density 10^5 to 10^{10} times greater than that for random sequences of the same length). Nevertheless, the sets of foldable sequences are sparsely distributed, such that many point mutations lead to non-foldable sequences. A relatively small number of sequences exist as hubs such that they are connected to a relatively large number of foldable neighbors via point mutations. These hubs were found to be especially productive starting points for error-prone PCR simulated evolutionary experiments. Evolutionary simulations also indicate that once a sequence leaves the set of foldable sequences, it is very rare for its descendents to re-enter the foldable space. The simulations also demonstrate a tradeoff between the diversity and screening (or selection) capacity. Increases to the mutation rate do not increase the diversity of foldable sequences as effectively as corresponding increases in the number of rounds of error-prone PCR. This is due partially to the observation that greater mutation rate leads to increased probability of leaving the set of foldable sequences (which is generally irreversible); moreover, simulations with greater mutation rate tend to more sparsely sample the space of mutant sequences as compared to simulations with a lower mutation rate and correspondingly more rounds of error-prone PCR. However, the exponential increase in the number of sequences with more rounds of error-prone PCR decreases the fraction of unique foldable sequences. That is, a greater number of unique foldable sequences results from long evolution; a greater fraction can be obtained with shorter evolution. Thus, experiments that can access the full number of unique foldables by screening or selecting the entire "culture" should use long evolution; but with a smaller screening capacity, one can only access the fraction foldable, and so shorter evolution is preferred. A hybrid strategy may be better, in which evolution is carried out up to the screening capacity, non-foldables are eliminated (because they are unlikely to evolve into foldables), and evolution is continued on to the remainder up to the screening capacity. This would produce the diverse foldable sequence library associated with longer continuous evolution schemes without the corresponding increase in selection or screening capacity

that would otherwise be necessary. The genetic recombination simulations demonstrated that recombination efficiently makes multiple mutations with direction toward the foldable portions of the sequence space and that the sequence identity between the parent sequences influences this outcome. Sequence identities between approximately 10% and 40% produced new foldable sequences with the highest probability. However, our procedure is biased toward frequent recombination, and so parent sequences with less similarity may also be useful experimentally. The methods presented here assume that functional sequences will be evenly distributed throughout the foldable space. If one had more information about the locations of the foldable sequences, one could bias these procedures appropriately. For example, if one knew that several mutations would be necessary for a specific function, one could bias the search with a larger mutation rate.

Because node degree and the location of starting sequences in our sequence-space graphs influenced the results of our evolutionary simulations, protein design calculations such as those described here might be used to generate sequence-space graphs to aid selection of sequences for evolutionary design. However, generating an entire sequence-space graph currently requires tremendous computational resources for proteins much larger than those analyzed here. However, simpler procedures could be used to obtain the relevant information. The number of neighbors (node degree) could be estimated by computationally exploring only the area of the sequence-space graph local to a chosen starting sequence. Such an exploration of all single mutations for a protein is computationally inexpensive compared to exploring the global graph properties, as was done in this study. Locating the well-separated sequence necessary for recombination could also be done using computational design of a small number of sequences. BLAST [186] could also be used to search for new sequences with the desired separation.

Lists of real foldable and functional sequences (and non-foldable and non-functional sequences) for a real protein and goal function would be useful for validating the conclusions presented in this article. Lacking such data, we have generated allowed protein sequences computationally, and used multiple techniques to do so in

order to reduce the influence of computational error in our conclusions. However, due to approximations in the search algorithm, these lists are lower bounds on the true sequences allowed for each protein; the rigid backbone and the discrete rotamer library used during the sequence and structure search disallow some sequences that would be able to fold similarly to the wild-type sequence. Although some sequences are missing from our foldable sequence spaces, the sequences chosen are expected to be stable on their respective backbone structures. Some previous studies of allowed protein sequence spaces have used lattice models to determine folding stability, and our work with crystal structures and atomic-level models is comparatively a more detailed model which produces chemically-realistic sets of allowed sequences. Additionally, our use of the dead-end elimination and A* algorithms to locate the protein sequences computed to be stable offers an interesting comparison to the related atomic-level work by Koehl et al. [144] where they defined the allowed sequences for a given fold using an iterative mean-field approach.

This work provides a framework for thinking about evolutionary protein engineering and could serve as a testing platform for other new ideas and techniques for sequence space search.

5.4 Methods

5.4.1 Evolution simulations

Evolution simulations were performed using our own code in the C++ programming language. Multiple rounds of error-prone PCR were simulated from one starting DNA sequence in each trial. We ran 10,000 trials of each error-prone PCR experiment from randomly-chosen starting sequences to generate good estimates of the reported statistics for each combination of parameters. The mutation rate (the probability of mutating each DNA base to any of the four bases in each round of error-prone PCR) was varied from 0.001 to 0.05 per base-pair per generation in Figure 5.3 and from 0.001 to 0.02 per base-pair per generation in Figure 5.4, and this probability was fixed

at 0.01 per base-pair per generation in all subsequent trials. The number of rounds of error-prone PCR was 9, 12, 15, or 18 in Figures 5.3 and 5.4 and was fixed to 15 in all subsequent trials. The starting DNA sequence coding for a given protein sequence was chosen at random when multiple codons were allowed. In genetic recombination simulations, the two or three starting parent sequences were chosen at random from the list of allowed sequences, one or two cut positions between amino acid residues were chosen at random, and the sequences were recombined around these points. The the order of the cut fragments was maintained during the recombination. The random() function in C++ was used for random number generation, with varying random number seeds.

We focused on the number of unique allowed protein sequences seen in the last round of error-prone PCR as the output of each simulation. This was highlighted over other possible metrics (such as the maximum number of mutations found or the total number of allowed sequences found) because creating the largest number of different foldable sequences (or the most diverse possible foldable library) should allow the greatest chance of locating functional sequences during subsequent library screening or selection.

In our simulations of error-prone PCR and genetic recombination, we sought to capture the most fundamental characteristics of these experiments. In error-prone PCR experiments, the mutation probability and the sequence replication process at the DNA level were the fundamental elements of the process captured in our model. We did not consider the DNA base mutation bias that differs between organisms because experimentalists can counter this effect for the most part by adding the appropriate dNTP's to bias the experiment toward the ideal in our model: an equal mutation probability between all DNA bases [164]. In our genetic recombination experiments, the process of cutting the parent sequences at a random position and recombining them to yield variable new sequences were the fundamental processes that we modeled. For example, one of the first techniques of this type cut four parents sequences into random fragments of 10–50 base-pairs each [187]. Different experimental procedures for performing recombination contain different biases (for

example, on where these cuts occur), but experimental genetic recombination procedures have been getting closer to the ideal in our model: all possible cut locations between codons are equally likely, where the probability of obtaining an allowed sequence is increased because codon boundaries are respected and the gene length and fragment positions are unchanged.

In our analysis of the distribution of mutations in error-prone PCR simulations shown in Figure 5.3, we found that the distributions fit well to Poisson distributions where the expected number of mutations per sequence (at the low mutation rate limit) is the mutation probability times the number of rounds of error-prone PCR times the number of DNA bases in the sequence, divided by two. Becacuse the probability of a DNA mutation being non-silent is 0.760417, the distribution of amino acid mutations at the end of each simulation is also approximately Poisson. Again at the low mutation rate limit, the expected number of amino acid mutations per sequence is the DNA mutation probability times 0.760417 times the number of rounds of error-prone PCR times the number of DNA bases in the sequence, divided by two. In this Poisson analysis, stop codons are considered to code for a "stop" amino acid, which always creates a non-folding sequence in our model.

5.4.2 Sequence-space graph generation

Crystal structures for WW-domain, BPTI, and B-domain were from the Protein Data Bank (PDB codes 1BPI, 1F8A, and 1IGD at 1.09, 1.84, and 1.10 Å resolution, respectively [10, 177–179]). Computational protein sequence and structure search methods were used to determine the allowed sequences for each protein core. These proteins were chosen because they are small but have well-packed cores, making our computational sequence and structure search feasible. The steric constraints in the core restrict the available sequence space, making the allowed sequence spaces small enough to be efficiently searched. Because a wide variety of amino acids would be allowed at surface positions due to reduced steric constraints, we performed our search on the core residues only. These cores were composed of 12, 17, and 11 positions for WW-domain, BPTI and B-domain, respectively (see caption to Figure 5.6), meaning

that full sequence spaces of size 20^{12}, 20^{17}, and 20^{11} were considered. To obtain upper and lower bounds on the possible steric constraints on these core residues, we followed two different procedures for handling the surface residues. In the first procedure, we cut all the surface residues back to alanine to reduce their steric impact as much as possible. In the second procedure, the surface residues were untouched during the design and remained in their crystal structure conformations. The results from both are sufficiently similar that only those from the first procedure are shown.

These sequence spaces were searched using the dead-end elimination and A* algorithms [20, 21], where the pairwise energy function was the sum of van der Waals, solvent-accessible surface area [39], and a Coulombic elecstrostatic term with a dielectric constant of 4 times the distance between each pair of atoms. Each protein backbone remained rigid during the search, and only the cardinal rotamers of each amino acid were considered during the search, using a backbone-independent rotamer library [130, 131, 188]. The lowest-energy 2,999 sequences for the WW-domain, BPTI, and B-domain cores were recorded and were used to define the allowed sequence space for each protein. The calculated energies for this list of sequences spanned 10, 12, and 55 kcal/mol for WW-domain, BPTI, and B-domain, respectively. The wild-type sequence was added to the list to create a set of 3,000 allowed sequences.

We also created a definition of a "functional" sequence to probe how this additional type of restriction influenced our results. We selected random sets of 30, 300, or 1,000 sequences from our lists of 3,000 allowed sequences to test the ability of our error-prone PCR and genetic recombination simulations to locate these functional subsets of sequence space. These definitions mean that the fraction of the foldable space that was functional was 1/100, 1/10, or 1/3.

Sequence-space graphs were created by associating a graph node with each protein sequence computed to be foldable. For every pair of proteins, the mutation distance between them was defined as the smallest number of DNA-base mutations required to mutate from one protein sequence to the other, if any set of codons could be chosen for those protein sequences. If the mutation distance was one, then an edge was drawn between the nodes corresponding to those two proteins. These graphs were

used for sequence selection only and were ignored during the evolution simulations, so the approximations here did not influence those results. Java 1.4.2 was used for all graph creation, and the JUNG (Java Universal Network/Graph) Framework was used [189]. Note that if 25 surface positions were mutated to alanine, and we selected 3,000 sequences for the core, then if truly any amino acids were allowed at the surface, the "true" sequence-space graph for that protein would have 3000×20^{25} nodes. However, this graph would have 20^{25} identical sub-graphs of size 3,000 (representing the core residues) with edges between them representing mutations of the surface residues. Therefore, we represent the allowed sequence space as the graph of the sequences available to the core only.

5.4.3 Measurement of graph properties

A component in a graph is a set of nodes where some path of edges connects every pair of nodes. The size of a component in a graph is the number of nodes that are connected in this way. When no edges connect two components in the graph, those components are called separated or disconnected. The degree of a node is the number of edges that touch that node. A node's clustering coefficient describes the connectivity of that node's neighbors and is defined as the fraction of the possible edges between those neighbors that exist. The shortest path between two nodes in a graph is the number of edges along the shortest connected path of edges between two nodes. The average of the shortest paths is computed after calculating the length of the shortest path between every pair of nodes. Shortest paths and component sizes were calculated using a breadth-first search algorithm [190]. The shortest path between a node and itself is defined to be zero. The diameter is the length of the longest shortest path in the graph.

Chapter 6

General conclusions

In conclusion, computational techniques were applied to the analysis of several molecular systems, providing a better understanding of the factors influencing molecular interaction as well as suggestions for future molecular design. Existing computational tools and a new method for protein analysis were used to study series of changes to molecular systems designed both by nature and in the laboratory. Strategies were also presented on making changes to molecular systems with computational guidance, including a new method for protein engineering.

Electrostatic analysis of a series of small molecule inhibitors for neuraminidase demonstrated that charge optimization is a useful strategy for improving the electrostatic component of the binding free energy and that chemical changes made that correspond to these optimal charges are frequently tighter-binding inhibitors. We also demonstrated the robustness of this approach to the choice of atoms to optimize and provide practical suggestions for using charge optimization in a drug design setting such as adopting chemical changes sequentially. This work suggests that optimizing molecular charges is a useful design goal and suggests that charge optimization will be useful in the design of new inhibitors.

Computational protein modeling played a role in each part of this thesis and was the method most frequently used for molecular analysis. Sequence and structural analysis of the PDZ3–CRIPT protein–protein complex suggested an interaction between these two synaptic proteins before the evolution of the animal synapse. A new method for the analysis of protein sequence alignments was developed and used to locate co-evolving residues and residues important for binding specificity in this protein–protein complex. A structural and electrostatic analysis of this binding interface also pointed to other residues important for binding. Computational protein

design techniques were used to create variants of the enzyme horseradish peroxidase and small molecule docking and binding energy calculations were used to analyze new active mutants of this enzyme. The results of this work suggest that minimizing each mobile side chain as a high-resolution step of the protein design process could be important. Computational protein design techniques were also used in conjunction with evolutionary simulations to evaluate the efficiency of the sequence space search by directed evolution experiments. This work suggested new methods for computationally guiding evolutionary protein engineering projects.

In the past, computational design of individual small molecules, stable proteins and tight-binding protein complexes have met with considerable success [110–112]. However, that these design tools are not widely used is a testament to the difficulty of most design problems and the inconsistent performance of these techniques. This thesis addresses this problem by considering series of solutions to design problems. This both improves the chance of locating a successful design and gives insight into the factors determining a successful design when more than one solution seems possible. Most directly, this was done in Chapter 4 where computational protein design techniques were used to design a library of horseradish peroxidase mutants computed to be foldable. Because foldability is a prerequisite for enzyme functionality we reasoned that a larger fraction of active enzymes would be found a computationally-designed library of foldable sequences than would be found in the entire sequence space available to the enzyme active site. Although we were not able to answer this question experimentally, we did find a new active enzyme in the computationally-restricted library. This enzyme would not have been found if we had focused on only the few protein folds computationally predicted to be most stable. In Chapter 5 we propose a method for computationally guiding directed evolution experiments toward the feasible portions of a protein's sequence space. Directed evolution already takes the approach of searching many possible design solutions, screening upwards of 10^7 sequences, and our methods are predicted to increase the fraction of stable proteins searched by such a technique. In Chapter 2 the charge optimization suggestions for the neuraminidase inhibitors frequently correspond to more than one possible chemical change. In the

set of known binders more than one chemical change often *is* made, and one of the changes typically results in a tighter-binding molecule than does the other. This suggests that even in the design of new small molecules based on electrostatic optimization, making libraries of mutants consistent with computational suggestions would be prudent.

Evolution is another theme of this thesis as we have both studied evolution and used it as a design technique. In Chapter 3 the evolution of the PDZ3–CRIPT binding interface was studied across a variety of organisms from sponge to human, highlighting that changes in protein sequence do not always lead to changes in function. The evolution between different PDZ domain types points to the ability of protein sequence mutations to alter binding function and to maintain binding specificity between different partner types. Evolution as a design technique entails the accumulation of many small changes in a large pool of possible design solutions and then selecting for improved variants. In Chapter 5 we defined sets of allowed sequences for three small proteins and evaluated the efficiency of different evolutionary engineering procedures at finding those sequences. The experimental variables that improved the efficiency of the search for foldable proteins in our models provide suggestions for guiding experimental evolutionary engineering procedures toward the most feasible portions of the sequence space. This type of design takes advantage of the insights provided by computational protein design techniques without losing the benefits provided by evolution, namely screening many small changes for new function. In Chapter 2 the series of neuraminidase inhibitor changes from weak-binding sialyic acid to tight-binding tamiflu and BCX-1812 can also be viewed as an evolution of sorts. As these molecules were developed by medicinal chemists, small changes were made to each inhibitor without complete knowledge of how each change would impact binding, and tighter-binding molecules were selected along the way. Our charge optimization results suggest that even when using the optimal charges to strategically make chemical changes, these changes should be made sequentially, one functional group at a time. This is because a functional group change is unlikely to correspond exactly to the optimal charges for that group. The new charges and atom locations of the real func-

tional group may change the optimal charges for other functional groups. Overall, nature has proved evolution to be an effective design procedure and this thesis work suggests that it can be exploited and computationally guided as a tool for molecular design.

The work presented in this thesis provides valuable insight into the factors influencing molecular interaction in several systems. The results argue that computational molecular modeling is useful for the analysis of molecular interactions and that it can suggest new ideas for molecular design. Future work could include experimental validation of some of the computational ideas presented here. The synthesis of new neuraminidase inhibitors based on charge optimization and a directed evolution experiment guided by our sequence-space graphs would be useful validation. Additionally, the idea that protein stability is a prerequisite for new function is theoretically sound, and so the design of a combinatorial library of stable proteins from which to select functional mutants is still promising and could still be experimentally validated. Focusing on a small number of design positions would allow each mutant in both the exhaustive and computationally-reduced library to be screened.

Bibliography

[1] M. E. M. Noble, J. A. Endicott, and L. N. Johnson. Protein kinase inhibitors: insights into drug design from structure. *Science*, 303(5665):1800–1805, 2004.

[2] J. J. Onuffer and R. Horuk. Chemokines, chemokine receptors and small-molecule antagonists: recent developments. *Trends Pharma. Sci.*, 23(10):459–467, 2002.

[3] S. Dev. Ancient-modern concordance in ayurvedic plants: Some examples. *Env. Health Perspectives*, 107(10):783–789, 1999.

[4] P. Dioscorides. *De Materia Medica (Arabic copy)*. Baghdad, c. 1334.

[5] A. L. Demain and R. P. Elander. The beta-lactam antibiotics: past, present, and future. *Antonie Van Leeuwenhoek*, 75(1-2):5–19, 1999.

[6] A. G. Millgate, B. J. Pogson, I. W. Wilson, T. M. Kutchan, M. H. Zenk, W. L. Gerlach, A. J. Fist, and P. J. Larkin. Analgesia: morphine-pathway block in top1 poppies. *Nature*, 431(7007):413–414, 2004.

[7] R. L. Mueller and S. Scheidt. History of drugs for thrombotic disease. Discovery, development, and directions for the future. *Circulation*, 89:432–449, 1994.

[8] D. DiMaio and D. Nathans. Directed mutagenesis. *Annu. Rev. Genet.*, 15:265–294, 1981.

[9] F. H. Arnold. Design by directed evolution. *Acc. Chem. Res.*, 31(3):125–131, 1998.

[10] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nuc. Acids Res.*, 28(1):235–242, 2000.

[11] E. DeClercq. Antiviral drugs in current clinical use. *J. Clinical Virology*, 30(2):115–133, 2004.

[12] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.

[13] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.

[14] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M Merz, D. M. Ferguson, D. C. Pellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995.

[15] W. L. Jorgensen and J. Tiradorives. The OPLS potential functions for proteins — energy minimizations for crystals of cyclic-peptides and crambin. *J. Am. Chem. Soc.*, 110(6):1657–1666, 1988.

[16] W. F. Van Gunsteren and H. J. C. Berendsen. Computer simulation of molecular dynamics — methodology, applications, and perspectives in chemistry. *Angew. Chem. Int. Ed.*, 29(9):992–1023, 1990.

[17] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force-field for molecular mechanical simulation of nucleic-acids and proteins. *J. Am. Chem. Soc.*, 106:765–784, 1984.

[18] M. K. Gilson, K. A. Sharp, and B. Honig. Calculating the electrostatic potential of molecules in solution: Method and error assessment. *J. Comp. Chem.*, 9:327–335, 1987.

[19] K. A. Sharp and B. Honig. Electrostatic interactions in macromolecules: Theory and applications. *Annu. Rev. Biophys. Chem.*, 19:301–332, 1990.

[20] J. Desmet, M. Demaeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356(6369):539–542, 1992.

[21] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, 33(2):227–239, 1998.

[22] L.P. Lee and B. Tidor. Optimization of electrostatic binding free energy. *J. Chem. Phys.*, 106(21):8681–8690, 1997.

[23] E. Kangas and B. Tidor. Optimizing electrostatic affinity in ligand-receptor binding: Theory, computation, and ligand properties. *J. Chem. Phys.*, 109(17):7522–7545, 1998.

[24] A.M. Davis and S.J. Teague. Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. *Angew. Chem. Int. Ed.*, 38(6):736–749, 1999.

[25] I.D. Kuntz B.K. Shoichet, A.R. Leach. Ligand solvation in molecular docking. *Proteins*, 34(1):4–16, 1999.

[26] M. K. Gilson and B. Honig. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. *Proteins*, 4(1):7–18, 1988.

[27] K. A. Sharp and B. Honig. Calculating total electrostatic energies with the nonlinear Poisson–Boltzmann equation. *J. Phys. Chem.*, 94(19):7684–7692, 1990.

[28] P. Ferrara, H. Gohlke, D.J. Price, G. Klebeand, and C.L. Brooks 3rd. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.*, 47(12):3032–3047, 2004.

[29] L. P. Lee and B. Tidor. Optimization of binding electrostatics: Charge complementarity in the barnase-barstar protein complex. *Protein Science*, 10(2):362–377, 2001.

[30] E. Kangas and B. Tidor. Electrostatic complementarity at ligand binding sites: Application to chorismate mutase. *J. Phys. Chem.*, 105(4):880–888, 2001.

[31] P. A. Sims, C. F. Wong, and J. A. McCammon. Charge optimization of the interface between protein kinases and their ligands. *J. Comp. Chem.*, 25(11):1416–1429, 2004.

[32] J. N. Varghese, W. G. Laver, and P. M. Colman. Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9-Å-resolution. *Nature*, 303(5912):35–40, 1983.

[33] N. R. Taylor, A. Cleasby, O. Singh, T. Skarzynski, A. J. Wonacott, P. W. Smith, S. L. Sollis, P. D. Howes, P. C. Cherry, R. Bethell, P. Colman, and J. Varghese. Dihydropyrancarboxamides related to zanamivir: a new series of inhibitors of influenza virus sialidases. 2. Crystallographic and molecular modeling study of complexes of 4-amino-4H-pyran-6-carboxamides and sialidase from influenza virus types A and B. *J. Med. Chem.*, 41(6):798–807, 1998.

[34] B.J. Smith, P. M. Colman, M. von Itzstein, B. Danylec, and J. N. Varghese. Analysis of inhibitor binding in influenza virus neuraminidase. *Protein Science*, 10(4):689–696, 2001.

[35] B.J. Smith, J. L. McKimm-Breshkin, M. McDonald, R. T. Fernley, J. N. Varghese, and P. M. Colman. Structural studies of the resistance of influenza virus neuraminidase to inhibitors. *J. Med. Chem.*, 45(11):2207–2212, 2002.

[36] J. N. Varghese, P. M. Colman, A. van Donkelaar, T. J. Blickand A. Sahasrabudhe, and J. L. McKimm-Breschkin. Structural evidence for a second sialic acid binding site in avian influenza virus neuraminidases. *Proc. Natl. Acad. Sci. U. S. A.*, 94(22):11808–11812, 1997.

[37] J.N. Varghese, V.C. Epa, and P.M. Colman. Three-dimensional structure of the complex of 4-guanidino-Neu5Ac2en and influenza virus neuraminidase. *Protein Science*, 4(6):1081–1087, 1995.

[38] J. N. Varghese, P. W. Smith, S. L. Sollis, T. J. Blick, A. Sahasrabudhe, J. L. McKimm-Breschkin, and P. M. Colman. Drug design against a shifting target: a structural basis for resistance to inhibitors in a variant of influenza virus neuraminidase. *Structure*, 6(6):735–746, 1998.

[39] D. Sitkoff, K. A. Sharp, and B. Honig. Accurate calculation of hydration free-energies using macroscopic solvent models. *J. Phys. Chem.*, 98(7):1978–1988, 1994.

[40] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

[41] C. I. Bayly, P. Geplak, W. D. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for determining atom-centered charges: the RESP model. *J. Phys. Chem.*, 97(40):10269–10280, 1993.

[42] W. D. Cornell, P. Cieplak, C. I. Bayly, and P. A. Kollman. Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *J. Am. Chem. Soc.*, 115(32):9620–10280, 1993.

[43] A. Nicholls and B. Honig. A rapid finite-difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. *J. Comp. Chem.*, 12(4):435–445, 1991.

[44] M. E. Davis, J. D. Madura, B. A. Luty, and J. A. McCammon. Electrostatics and diffusion of molecules in solutions — simulations with the University-of-Houston Brownian Dynamics program. *Comput. Phys. Comm.*, 62(2-3):187–197, 1991.

[45] M. Holst and F. Saied. Multigrid solution of the nonlinear Poisson–Boltzmann equation and calculation of titration curves. *J. Comp. Chem.*, 98(7):1978–1988, 1994.

[46] H. Oberoi and N. M. Allewell. Multigrid solution of the Poisson–Boltzmann equation. *Biophys. J.*, 65(1):48–55, 1993.

[47] J. O'M. Bockris and A. K. N. Reddy. *Modern Electrochemistry.* Plenum Press, New York, 1973.

[48] L. T. Chong, S. E. Dempster, Z. S. Hendsch, L. P. Lee, and B. Tidor. Computation of electrostatic complements to proteins: A case of charge stabilized binding. *Protein Science*, 7(1):206–210, 1998.

[49] E. Kangas and B. Tidor. Charge optimization leads to favorable electrostatic binding. *Phys. Rev. E*, 59(5):5958–5961, 1999.

[50] E. Kangas and B. Tidor. Electrostatic specificity in molecular ligand design. *J. Chem. Phys.*, 112(20):9120–9131, 2000.

[51] R. J. Vanderbei. LOQO: An interior point code for quadratic programming. *Optimization Methods and Software*, 11:451–484, 1999.

[52] C. A. Sarkar, K. Lowenhaupt, T. Horan, T. C. Boone, B. Tidor, and D. A. Lauffenburger. Rational cytokine design for increased lifetime and enhanced potency using pH-activated "histidine switching". *Nat. Biotechnol.*, 20(9):908–913, 2002.

[53] J. S. Ahn, M. L. Radhakrishnan, M. Mapelli, S. Choi, B. Tidor, G. D. Cuny, A. Musacchio, L. A. Yeh, and K. S. Kosik. Defining CDK5 ligand chemical space with small molecule inhibitors of tau phosphorylation. *Chem. Biol.*, 12:811–823, 2005.

[54] T. Sulea and E. O. Purisima. Optimizing ligand charges for maximum binding affinity. A solvated interaction energy approach. *J. Phys. Chem. B.*, 105:889–899, 2001.

[55] T. Sulea and E. O. Purisima. Profiling charge complementarity and selectivity for binding at the protein surface. *Biophys. J.*, 84:2883–2896, 2003.

[56] D. F. Green and B. Tidor. *Escherichia coli* glutaminyl-tRNA synthetase is electrostatically optimized for binding of its cognate substrates. *J. Mol. Biol.*, 10:435–452, 2004.

[57] D. F. Green and B. Tidor. Design of improved protein inhibitors of HIV-1 cell entry: Optimization of electrostatic interactions at the binding interface. *Proteins*, 60:644–657, 2005.

[58] V. Stoll, K. D. Stewart, C. J. Maring, S. Muchmore, V. Giranda, Y. Y. Gu, G. Wang, Y. Chen, M. Sun, C. Zhao, A. L. Kennedy, D. L. Madigan, Y. Xu, A. Saldivar, W. Kati, G. Laver, T. Sowin, H. L. Sham, J. Greer, and D. Kempf. Influenza neuraminidase inhibitors: Structure-based design of a novel inhibitor series. *Biochemistry*, 42(3):718–727, 2003.

[59] Y. S. Babu, P. Chand, S. Bantia, P Kotian, A. Dahghani, Y. El-Kattan, T. Lin, T. L. Hutchison, A. J. Elliott, C. D. Parker, S. L. Ananth, L. L. Horn, G. W. Laver, and J. A. Montgomery. BCX-1812 (RWJ-270201): Discovery of a novel, highly potent, orally active, and selective influenza neuraminidase inhibitor through structure-based drug design. *J. Med. Chem.*, 43(19):3482–3486, 2000.

[60] M. Fornabaio, P. Cozzini, A. Mozzarelli, D. J. Abraham, and G. E. Kellog. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 2. Computational titration and pH effects in molecular models of neuraminidase-inhibitor complexes. *J. Med. Chem.*, 46:4487–4500, 2003.

[61] K. M. Masukawa, P. A. Kollman, and I. D. Kuntz. Investigation of neuraminidase-substrate recognition using molecular dynamics and free energy calculations. *J. Med. Chem.*, 46:5628–5637, 2003.

[62] C.U. Kim, W. Lew, M.A. Williams, H. Liu, L. Zhang, S. Swaminathan, N. Bischofberger, M.S. Chen, D.B. Mendel, C.Y. Tai, W.G. Laver, , and R.C. Stevens. Influenza neuraminidase inhibitors posessing a novel hydrophobic interaction in the enzyme active site: Design, synthesis, and structural analysis of carbocyclic sialic acid analogues with potent anti-influenza activity. *J. Am. Chem. Soc.*, 119(4):681–690, 1997.

[63] N. A. Campbell, J. B. Reece, and L. G. Mitchell. *Biology, Fifth Edition.* Addison Wesley Longman, New York, 1999.

[64] T. Cavalier-Smith, E. Chao, N. Boury-Esnault, and J. Vacelet. Sponge phylogeny, animal monophyly, and the origin of the nervous system: 18S rRNA evidence. *Can. J. Zool.*, 74(11):2031–2045, 1996.

[65] C. Borchiellini, M. Manuel, E. Alivon, N. Boury-Esnault, J. Vacelet, and Y. Le Parco. Sponge paraphyly and the origin of Metazoa. *J. Evol. Biol.*, 14(1):171–179, 2001.

[66] K. J. Peterson and N. J. Butterfield. Origin of the Eumetazoa: Testing ecological predictions of molecular clocks against the Proterozoic fossil record. *Proc. Natl. Acad. Sci. U. S. A.*, 102(27):9547–9552, 2005.

[67] H. Husi, M. A. Ward, J. S. Choudhary, W. P. Blackstock, and S. G. N. Grant. Proteomic analysis of NMDA receptor-adhesion protein signaling complexes. *Nat. Neurosci.*, 3(7):661–669, 2000.

[68] J. Peng, M. J. Kim, D. Cheng, D. M. Duong, S. P. Gygi, and M. Sheng. Semi-quantitative proteomic analysis of rat forebrain postsynaptic density fractions by mass spectrometry. *J. Biol. Chem.*, 279(20):21003–21011, 2004.

[69] O. Sakarya, K. A. Armstrong, M. Adamska, M. Adamski, I. Wang, B. Tidor, B. M. Degnan, T. H. Oakley, and K. S. Kosik. A post-synaptic scaffold at the origin of the animal kingdom.

[70] M. Niethammer, E. Kim, and M. Sheng. Interaction between the C terminus of NMDA receptor subunits and multiple members of the PSD-95 family of membrane-associated guanylate kinases. *J. Neuroscience*, 16(7):2157–2163, 1996.

[71] M. Niethammer, J. G. Valtschanoff, T. M. Kapoor, D. W. Allison, R. J. Weinberg, A. Craig, and M. Sheng. CRIPT, a novel postsynaptic protein that binds to the third PDZ domain of PSD-95/SAP90. *Neuron*, 20:693–707, 1998.

[72] E. Kim, S. J. DeMarco, S. M. Marfatia, A. H. Chishti, M. Sheng, and E. E. Strehler. Plasma membrane CA2+ ATPase isoform 4b binds to membrane-associated guanylate kinase (MAGUK) proteins via their PDZ (PSD-95/Dlg/ZO-1) domains. *J. Biol. Chem.*, 273(3):1591–1595, 1998.

[73] W. Zhang, L. Vazquez, M. Apperson, and M. B. Kennedy. Citron binds to PSD-95 at glutamatergic synapses on inhibitory neurons in the hippocampus. *J. Neuroscience*, 19(1):96–108, 1999.

[74] M. Niethammer, E. Kim, and M. Sheng. Interaction between the C terminus of NMDA receptor subunits and multiple members of the PSD-95 family of membrane-associated guanylate kinases. *J. Neuroscience*, 16(7):2157–2163, 1996.

[75] D. A. Doyle, A. Lee, J. Lewis, E. Kim, M. Sheng, and R. MacKinnon. Crystal structures of a complexed and peptide-free membrane protein-binding domain: Molecular basis of peptide recognition by PDZ. *Cell*, 85(7):1067–1076, 1996.

[76] I. Bezprozvanny and A. Maximov. Classification of PDZ domains. *FEBS Letters*, 509:457–462, 2001.

[77] Y. Hata, H. Nakanishi, and Y. Takai. Synaptic PDZ domain-containing proteins. *Neurosci. Res.*, 32:1–7, 1998.

[78] Z. Songyang, A. S. Fanning, C. Fu, J. Xu, S. M. Marfatia, A. H. Chishti, A. Crompton, A. C. Chan, J. M. Amderson, and L. C. Cantly. Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science*, 275(5296):73–77, 1997.

[79] D. Cowburn. Peptide recognition by PTB and PDZ domains. *Curr. Opin. Struct. Biol.*, 7:835–838, 1997.

[80] B. Z. Harris, F. W. Lau, N. Fujii, R. K. Guy, and W. A. Lim. Role of electrostatic interactions in PDZ domain ligand recognition. *Biochemistry*, 42(10):2797–2805, 2003.

[81] A. S. Fanning and J. M. Anderson. Protein–protein interactions: Pdz domain networks. *Current Biology*, 6(11):1385–1388, 1996.

[82] L. W. Lo, E. I. Shakhnovich, and L. A. Mirny. Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. *Proc. Natl. Acad. Sci. U. S. A.*, 100(8):4463–4468, 2003.

[83] W. R. Atchley, K. R. Wollenberg, W. M. Fitch, W. Terhalle, and A. W. Dress. Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. *Mol. Biol. Evol.*, 17(1):164–178, 2000.

[84] S. W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, 1999.

[85] A. A. Fodor and R. W. Aldrich. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, 56(2):211–221, 2004.

[86] A. C. Wallace, R. A. Laskowski, and J. M. Thornton. LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.*, 8(2):127–134, 1995.

[87] I. K. McDonald and J. M. Thornton. Satisfying hydrogen-bonding potential in proteins. *J. Mol. Biol.*, 238(6):777–793, 1994.

[88] A. P. Turnbull, C. Phillips, G. Bunkoczi, J. Debreczeni, E. Ugochukwu, A. C. W. Pike, F. Gorrec, C. Umeano, J. Elkins, G. Berridge, P. Savitsky, O. Gileadi, F. von Delft, J. Weigelt, A. Edwards, C. Arrowsmith, M. Sundstrom, and D. Doyle. Crystal structure of the 1st PDZ domain of Human DLG3.

[89] I. von Ossowski, W. Oksanena, L. von Ossowski, C. Cai, M. Sundberg, A. Goldman, and K. Keinanen. Crystal structure of the second PDZ domain of SAP97 in complex with a GluR-A C-terminal peptide. *FEBS Journal*, 273(22):5219–5229, 2006.

[90] E. V. Koonin. Orthologs, patalogs, and evolutionary genomics. *Annu. Rev. Genetics*, 39:309–338, 2005.

[91] E. L. L. Sonnhammer and E. V. Koonin. Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, 18(12):619–620, 2002.

[92] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nuc. Acids Res.*, 30(14):3059–3066, 2002.

[93] A. T. Brunger and M. Karplus. Polar hydrogen positions in proteins — empirical energy placement and neutron-diffraction comparison. *Proteins*, 4(2):148–156, 1988.

[94] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM — a program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4(2):187–217, 1983.

[95] A. Sali and T. L. Blundell. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234(3):779–815, 1993.

[96] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. PROCHECK — a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26:283–291, 1993.

[97] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, Cambridge, 2003.

[98] F. H. Arnold. Combinatorial and computational challenges for biocatalyst design. *Nature*, 409:253–259, 2001.

[99] A. M. Klibanov. Improving enzymes by using them in organic solvents. *Nature*, 409:241–246, 2001.

[100] J. D. Bloom, M. M. Meyer, P. Neinhold, C. R. Otey, D. MacMillan, and F. H. Arnold. Evolving strategies for enzyme engineering. *Curr. Opin. Struct. Biol.*, 15(4):447–452, 2005.

[101] M. T. Reetz. Controlling the enantioselectivity of enzymes by directed evolution: Practical and theoretical ramifications. *Proc. Natl. Acad. Sci. U. S. A.*, 101(16):5716–5722, 2004.

[102] J. Kaur and R. Sharma. Directed evolution: An approach to engineer enzymes. *Crit. Rev. Biotech.*, 26(3):165–199, 2006.

[103] P. T. R. Rajagopalan, S. Lutz, and S. J. Benkovic. Coupling interactions of distal residues enhance dihydrofolate reductase catalysis: mutational effects on hydride transfer rates. *Biochemistry*, 41:12618–12628, 2002.

[104] T. Yano, S. Oue, and H. Kagamiyama. Directed evolution of an aspartate aminotransferase with new substrate specificities. *Proc. Natl. Acad. Sci. U. S. A.*, 20:1251–1255, 1998.

[105] A. Glieder, E. T. Farina, and F. H. Arnold. Laboratory evolution of a soluble, self-sufficient, highly active alkane hydroxylase. *Nat. Biotechnol.*, 20:1135–1139, 2002.

[106] P. K. Agarwal, S. R. Billeter, P. T. R. Rajagopalan, S. J. Benkovic, and S. Hammes-Schiffer. Network of coupled promoting motions in enzyme catalysis. *Proc. Natl. Acad. Sci. U. S. A.*, 99(5):2794–2799, 2002.

[107] K. L. Morley and R. J. Kazlauskas. Improving enzyme properties: when are closer mutations better? *Trends Biotechnol.*, 23(5):231–237, 2005.

[108] D. E. Benson, M. S. Wisz, and H. W. Hellinga. Rational design of nascent metalloenzymes. *Proc. Natl. Acad. Sci. U. S. A.*, 97(12):6292–6297, 2000.

[109] D. E. Benson, A. E. Haddy, and H. W. Hellinga. Converting a maltose receptor into a nascent binuclear copper oxygenase by computational design. *Biochemistry*, 41(9):3262–3269, 2002.

[110] M. A. Dwyer, L. L. Looger, and H. W. Hellinga. Computational design of a biologically active enzyme. *Science*, 304(5679):1967–1971, 2004.

[111] G. Dantas, B. Kuhlman, D. Callendar, M. Wong, and D. Baker. A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.*, 332(2):449–460, 2003.

[112] D. N. Bolon, C. A. Voigt, and S. L. Mayo. De novo design of biocatalysts. *Curr. Opin. Chem. Biol.*, 6(2):125–129, 2002.

[113] R. J. Hayes, J. Bentzien, M. L. Ary, M. Y. Hwang, J. M. Jacinto, J. Vielmetter, A. Kundu, and B. I. Dahiyat. Combining computational and experimental screening for rapid optimization of protein properties. *Proc. Natl. Acad. Sci. U. S. A.*, 99(25):15926–15931, 2002.

[114] M. C. Saraf, G. L. Moore, N. M. Goodey, V. Y. Cao, S. J. Benkovic, and C. D. Maranas. IPRO: an iterative computational protein library redesign and optimization procedure. *Biophys J.*, 90(11):4167–4180, 2006.

[115] M. M. Meyer, L. Hochrein, and F. H. Arnold. Structure-guided SCHEMA recombination of distantly related beta-lactamases. *Protein Eng. Des. Sel.*, 19(12):563–570, 2006.

[116] H. Xiang, L. Luo, K. L. Taylor, and D. Dunaway-Mariano. Interchange of catalytic activity within the 2-enoyl-coenzyme A hydratase/isomerase superfamily based on a common active site template. *Biochemistry*, 38(24):7638–7652, 1999.

[117] E. Quemeneur, M. Moutiez, J. B. Charbonnier, and A. Menez. Engineering cyclophilin into a proline-specific endopeptidase. *Nature*, 391(6664):301–304, 1998.

[118] H. Park, S. Nam, J. K. Lee, C. N. Yoon, B. Mannervik, S. J. Benkovic, and H. Kim. Design and evolution of new catalytic activity with an existing protein scaffold. *Science*, 311(5760):535–538, 2006.

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

[120] S. Hammes-Schiffer. Impact of enzyme motion on activity. *Biochemistry*, 41(45):13335–13343, 2002.

[121] T. H. Rod, J. L. Radkiewicz, and C. L. Brooks. Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proc. Natl. Acad. Sci. U. S. A.*, 100(12):6980–6985, 2003.

[122] N. Ota and D. A. Agard. Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. *J. Mol. Biol.*, 351(2):345–354, 2002.

[123] G. I. Berglund, G. H. Carlsson, A. T. Smith, H. Szoke, A. Henriksen, and J. Hajdu. The catalytic pathway of horseradish peroxidase at high resolution. *Nature*, 417:463–468, 2002.

[124] N. C. Veitch. Horseradish peroxidase: a modern view of a classic enzyme. *Phytochemistry*, 65:249–259, 2004.

[125] W. Adam, M. Lazarus, C. R. Saha-Moller, O. Weichold, U. Hoch, D. Haring, and P. Scherier. Biotransformations with peroxidases. *Adv. in Biochem. Eng. and Biotech.*, 63:73–108, 1999.

[126] A. Aharoni, L. Gaidukov, O. Khersonsky, S. M. Gould, C. Roodveldt, and D. S. Tawfik. The 'evolvability' of promiscuous protein functions. *Nat. Genet.*, 37(1):73–76, 2005.

[127] R. A. Chica, N. Doucet, and J. N. Pelletier. Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Curr. Opin. Biotechnol.*, 16(4):378–384, 2005.

[128] A. Bateman, L. Coin, R. Durbin, R. D¿ Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The Pfam protein families database. *Nuc. Acids Res.*, 32:D138–D141, 2004.

[129] S. Hayward and N. Go. Collective variable description of native protein dynamics. *Annu. Rev. Phys. Chem.*, 46:223–250, 1995.

[130] R. L. Dunbrack. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.*, 12(4):431–440, 2002.

[131] R. L. Dunbrack and F. E. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.*, 6(8):1661–81, 1997.

[132] S. M. Lippow. *Computational analysis, design, and experimental validation of antibody binding affinity improvements beyond in vivo maturation.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, U.S.A., 2007.

[133] M. D. Altman. *Computational ligand design and analysis in protein complexes using inverse methods, combinatorial search, and accurate solvation modeling.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, U.S.A., 2006.

[134] W. Sherman, T. Day, M. P. Jacobson, R. A. Friesner, and R. Farid. Novel procedures for modeling ligand/receptor induced fit effects. *J. Med. Chem.*, 49(2):534–553, 2006.

[135] D. W. Colby, B. A. Kellogg, C. P. Graff, Y. A. Yeung, J. S. Swers, and K. D. Wittrup. Engineering antibody affinity by yeast surface display. *Methods Enzymol.*, 388:348–358, 2004.

[136] H. W. T. van Vlijmen and M. Karplus. Analysis of calculated normal modes of a set of native and partially unfolded proteins. *J. Phys. Chem. B.*, 103:3009–3021, 1999.

[137] V. Zoete, O. Michielin, and M. Karplus. Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *J. Mol. Biol.*, 315:21–52, 2002.

[138] Z. W. Cao, X. Chen, and Y. Z. Chen. Correlation between normal modes in the 20-200 cm^{-1} frequency range and localized torsion motions related to certain collective motions in proteins. *J. Mol. Graph. and Model.*, 21:309–319, 2003.

[139] A. M. Levin and G. A. Weiss. Optimizing the affinity and specificity of proteins with molecular display. *Mol. Biosyst.*, 2(1):49–57, 2006.

[140] D. Baker. Prediction and design of macromolecular structures and interactions. *Philos. T. Roy. Soc. B.*, 361(1467):459–463, 2006.

[141] G. L. Butterfoss and B. Kuhlman. Computer-based design of nodel protein structures. *Annu. Rev. Biophys. Biomol. Struct.*, 35:49–65, 2006.

[142] C. L. Vizcarra and S. L. Mayo. Electrostatics in computational protein design. *Curr. Opin. Chem. Biol.*, 9(6):622–626, 2005.

[143] H. Li, R. Helling, C. Tang, and N. Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273(5275):666–669, 1996.

[144] P. Koehl and M. Levitt. Protein topology and stability define the space of allowed sequences. *Proc. Natl. Acad. Sci. U. S. A.*, 99(3):1280–1285, 2002.

[145] L. D. Bogarad and M. W. Deem. A hierarchical approach to protein molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 96(6):2591–2595, 1999.

[146] G. L. Moore and C. D. Maranas. Modeling DNA mutation and recombination for directed evolution experiments. *J. Theor. Biol.*, 205(3):483–503, 2000.

[147] N. Maheshri and D. V. Schaffer. Computational and experimental analysis of DNA shuffling. *Proc. Natl. Acad. Sci. U. S. A.*, 100(6):3071–3076, 2003.

[148] M. Olsen, B. Iverson, and G. Georgiou. High-throughput screening of enzyme libraries. *Curr. Opin. Biotechnol.*, 11(4):331–337, 2000.

[149] P. Amstutz, P. Forrer, C. Zahnd, and A. Pluckthun. In vitro display technologies: novel developments and applications. *Curr. Opin. Biotechnol.*, 12(4):400–405, 2001.

[150] E. T. Boder and K. D. Wittrup. Yeast surface display for directed evolution of protein expression, affinity, and stability. *Methods Enzymol.*, 328:430–444, 2000.

[151] P. S. Daugherty, G. Chen, M. J. Olsen, B. L. Iverson, and G. Georgiou. Antibody affinity maturation using bacterial surface display. *Protein Eng.*, 11(9):825–832, 1998.

[152] P. Samuelson, E. Gunneriusson, P. A. Nygren, and S. Stahl. Display of proteins on bacteria. *J. Biotechnol.*, 96(2):129–154, 2002.

[153] G. P. Smith and V. A. Petrenko. Phage display. *Chem. Rev.*, 97(2):391–410, 1997.

[154] B. Seed. Purification of genomic sequences from bacteriophage libraries by recombination and selection *in vivo*. *Nuc. Acids Res.*, 11(8):2427–2445, 1983.

[155] R. W. Roberts and J. W. Szostak. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 94(23):12297–12302, 1997.

[156] J. Hanes and A. Pluckthun. In vitro selection and evolution of functional proteins by using ribosome display. *Proc. Natl. Acad. Sci. U. S. A.*, 94(10):4937–4942, 1997.

[157] D. S. Tawfik and A. D. Griffiths. Man-made cell-like compartments for molecular evolution. *Nat. Biotechnol.*, 16(7):652–656, 1998.

[158] W. P. Stemmer. Rapid evolution of a protein in vitro by DNA shuffling. *Nature*, 370(6488):389–391, 1994.

[159] D. Wahler and J. L. Reymond. Novel methods for biocatalyst screening. *Curr. Opin. Chem. Biol.*, 5(2):152–158, 2001.

[160] R. E. Campbell, O. Tour, A. E. Palmer, P. A. Steinback, G. S. Baird, D. A. Zacharias, and R. Y. Tsien. A monomeric red fluorescent protein. *Proc. Natl. Acad. Sci. U. S. A.*, 99(12):7877–7882, 2002.

[161] B. J. Bevis and B. S. Glick. Rapidly maturing variants of the Discosoma red fluorescent protein (DsRed). *Nat. Biotechnol.*, 20(1):83–87, 2002.

[162] C. A. Voigt, S. L. Mayo, F. H. Arnold, and Z. G. Wang. Computationally focusing the directed evolution of proteins. *J. Cell Biochem. Suppl.*, 37:58–63, 2001.

[163] H. Zhao and F. H. Arnold. Directed evolution converts subtilisin E into a functional equivalent of thermitase. *Protein Eng.*, 12(1):47–53, 1999.

[164] C. Neylon. Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nuc. Acids Res.*, 32(4):1448–1459, 2004.

[165] P. S. Daugherty, G. Chen, B. L. Iverson, and G. Georgiou. Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *Proc. Natl. Acad. Sci. U. S. A.*, 97(5):2029–2034, 2000.

[166] M. Zaccolo and E. Gherardi. The effect of high-frequency random mutagenesis on in vitro protein evolution: a study on TEM-1 beta-lactamase. *J. Mol. Biol.*, 285(2):775–783, 1999.

[167] W. M. Coco, W. E. Levinson, M. J. Crist, H. J. Hektor, A. Darzins, P. T. Pienkos, C. H. Squires, and D. J. Monticello. DNA shuffling method for generating highly recombined genes and evolved enzymes. *Nat. Biotechnol.*, 19(4):354–359, 2001.

[168] H. Zhao, L. Giver, Z. Shao, J. A. Affholter, and F. H. Arnold. Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nat. Biotechnol.*, 16(3):258–261, 1998.

[169] L. A. Rowe, M. L. Geddie, O. B. Alexander, and I. Matsumura. A comparison of directed evolution approaches using the beta-glucuronidase model system. *J. Mol. Biol.*, 332(4):851–860, 2003.

[170] M. Ostermeier, J. H. Shim, and S. J. Benkovic. A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotechnol.*, 17(12):1205–1209, 1999.

[171] D. Zha, A. Eipper, and M. T. Reetz. Assembly of designed oligonucleotides as an efficient method for gene recombination: a new tool in directed evolution. *Chembiochem.*, 4(1):34–39, 2003.

[172] M. A. Mena, T. P. Treynor, S. L. Mayo, and P. S. Daugherty. Blue fluorescent proteins with enhanced brightness and photostability from a structurally targeted library. *Nat. Biotechnol.*, 24(12):1569–1571, 2006.

[173] J. J. Silberg, J. B. Endelman, and F. H. Arnold. SCHEMA-guided protein recombination. *Methods Enzymol.*, 388:35–42, 2004.

[174] M. C. Saraf, A. R. Horswill, S. J. Benkovic, and C. D. Maranas. FamClash: a method for ranking the activity of engineered enzymes. *Proc. Natl. Acad. Sci. U. S. A.*, 101(12):4142–4147, 2004.

[175] C. A. Voigt, S. L. Mayo, F. H. Arnold, and Z. G. Wang. Computational method to reduce the search space for directed protein evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 98(7):3778–3783, 2001.

[176] J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.*, 103(15):5869–5874, 2006.

[177] S. Parkin, B. Rupp, and H. Hope. Structure of bovine pancreatic trypsin inhibitor at 125 K definition of carboxyl-terminal residues Gly57 and Ala58. *Acta. Crystallogr. D. Biol. Crystallogr.*, 52(1):18–29, 1996.

[178] M. A. Verdecia, M. E. Bowman, K. P. Lu, T. Hunter, and J. P. Noel. Structural basis for phosphoserine-proline recognition by group IV WW domains. *Nat. Struct. Biol.*, 7(8):639–643, 2000.

[179] J. P. Derrick and D. B. Wigley. The third IgG-binding domain from streptococcal protein G. an analysis by X-ray crystallography of the structure alone and in a complex with Fab. *J. Mol. Biol.*, 243(5):906–918, 1994.

[180] K. F. Lau and K. A. Dill. Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. U. S. A.*, 87(2):638–642, 1990.

[181] J. D. Bloom, C. O. Wilke, F. H. Arnold, and C. Adami. Stability and the evolvability of function in a model protein. *Biophys. J.*, 86(5):2758–2764, 2004.

[182] M. T. Reetz, J. D. Carballeira, and A. Vogel. Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew. Chem. Int. Ed.*, 45(46):7745–7751, 2006.

[183] A. Wong, S. N. Albright, and M. F. Wolfner. Evidence for structural constraint on ovulin, a rapidly evolving Drosophila melanogaster seminal protein. *Proc. Natl. Acad. Sci. U. S. A.*, 103(49):18644–18649, 2006.

[184] C. M. Hill, W. S. Li, J. B. Thoden, H. M. Holden, and F. M. Raushel. Enhanced degradation of chemical warfare agents through molecular engineering of the phosphotriesterase active site. *J. Am. Chem. Soc.*, 125(30):8990–8991, 2003.

[185] D. F. Green, A. T. Dennis, P. S. Fam, B. Tidor, and A. Jasanoff. Rational design of new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide. *Biochemistry*, 45(41):12547–12559, 2006.

[186] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nuc. Acids Res.*, 25:3389–3402, 1997.

[187] W. P. C. Stemmer. DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 91(22):10747–10751, 1994.

[188] R. L. Dunbrack and M. Karplus. Backbone-dependent rotamer library for proteins - application to side-chain prediction. *J. Mol. Biol.*, 230:543–574, 1993.

[189] J. O'Madadhain, D. Fisher, and S. White. JUNG: Java Universal Network/Graph Framework, http://jung.sourceforge.net.

[190] T. Corman, C. Leiserson, and R. Rivest. *Introduction to Algorithms.* MIT Press, Cambridge, 1990.