THE COMPUTATIONAL DESIGN OF PROTEIN-LIGAND INTERFACES

By

Andrew Morin

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical and Physical Biology

August, 2011

Nashville, Tennessee

Approved:

Professor David W. Piston, Ph.D.

Professor F. Peter Guengerich, Ph.D.

Professor D. Borden Lacy, Ph.D.

Professor Albert Beth, Ph.D.

Professor Lawrence J. Marnett, Ph.D.

Professor Jens Meiler, Ph.D.

ACKNOWLEDGEMENTS

My dissertation research would not have been possible without the support and assistance of my advisor Prof. Jens Meiler and the members of the Meiler lab Kristian Kaufmann, Samuel DeLuca, Gordon Lemmon, Nils Woetzel and Mert Karakas.

I would additionally like to thank Prof. Walter Chazin for his wisdom and guidance, and for his continued advocacy and directorship of the Vanderbilt Center for Structural Biology, without which little of this research would have been possible.

I would also like to express my gratitude to Joel Harp for his assistance and instruction in learning the art and science of X-ray crystallography, as well as Prof. Borden Lacy for her guidance in the field of structural biology.

I furthermore thank the members of my dissertation committee for their continued support in the face of a constantly evolving research project.

Finally, I would like to thank Yoana Dimittrova for all of her support, love and caring, and Mert Karakas and Gulfem Guler for their valued friendship during my years at Vanderbilt.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Å	angstrom
AA	amino acid
ACCRE	Advanced Computing Center for Research and Education
ATP	adenosine triphosphate
BCL	BioChemical Library
BI	backscattering interferometry
C	centigrade
CAPRI	Critical Assessment of PRedicted Interactions
CCD	charged coupled device
CD	circular dichroism
CPD	computational protein design
CPU	central processing unit
D-ala	D-alanine
D-lac	D-lactate
Da	Dalton
dansyl	5-dimethylamino-1-naphthalenesulfonyl
DLS	dynamic light scattering
DNA	deoxyribonucleic acid
Dnase	deoxyribonuclease
ELISA	enzyme-linked-immunosorbent serologic assay
ERT	enzyme replacement therapy
ESI-MS	electrospray ionization mass spectrometry
FA	fluorescence anisotropy
GB	gigabyte

HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HSQC	heteronuclear single quantum coherence
IMAC	immobilized metal affinity chromatography
IPTG	isopropyl-D-thiogalactopyranoside
ITC	isothermal titration calorimetry
K	kelvin
kcal	kilocalorie
K_d	disassociation constant
kDa	kilo Dalton
L	liter
LB	lysogeny broth
LPDB	Ligand Protein Database
MALDI-MS	matrix-assisted laser desorption/ionization mass spectrometry
MCM	Monte-Carlo/Metropolis
mg	milligram
MHz	megahertz
mm	millimeter
MRE	multidrug- (or methicillin) resistant Enterococci
MRSA	multidrug- (or methicillin) resistant Staphylococcus aureus
NADH	nicotinamide adenine dinucleotide, reduced form
nm	nanometer
NMR	nuclear magnetic resonance
PCR	polymerase chain reaction
PDB	Protein Databank
PK/PD	pharmacokinetic /pharmacodynamic
PNAS	Proceedings of the National Academy of Sciences
r.e.u.	ROSETTA energy units

RMSD	root mean square deviation
RNA	ribonucleic acid
SA	solvent accessible
SAXS	small-angle x-ray scattering
SCOP	structural classification of proteins
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEC	size-exclusion chromatography
SER-CAT	Southeast Regional Collaborative Access Team
TPR	tetratricopeptide repeat
UV	ultraviolet
VDW	Van der Waals
VRE	vancomycin resistant Enterococci
VRSA	vancomycin resistant Staphylococcus aureus
WT	wild-type
XML	extensible mark-up language
μM	micro molar

CHAPTER I*

INTRODUCTION AND BACKGROUND

Overview

Interaction between protein and ligand is a fundamental mechanism in biology. Receptor-ligand interaction is the basis of modern pharmacology and many basic cellular processes rely on functional interfaces between proteins and small ligands. Endogenous and exogenous small-molecule and peptide ligands play numerous critical roles in the biology of both normal and disease states, and in the battle between host and pathogen. Knowledge and understanding of these interactions is critical in medicine and the ability to manipulate them is essential in therapeutic development. Although entire disciplines and industries have been created to develop and design small ligands for therapeutic use, substantially less progress has been made from the receptor side, in understanding and designing the protein interfaces to bind small ligands.

The goal of my dissertation research was to develop a general and repeatable method for designing and re-designing protein interfaces to small-molecules and peptide ligands using *in silico*, rational design techniques. Establishing a successful

* Sections of Chapter I have been excerpted from Morin, A. et al., 2011. Computational design of protein-ligand interfaces: potential in therapeutic development. *Trends in Biotechnology*, 29(4), pp.159-66. and the Dissertation proposal of Morin 2007.

and robust computational method for designing protein-ligand interactions would have broad and significant application in both basic science and the development of protein therapeutics to address disease.

To begin developing these computational methods, the D-amino acid peptide target of the glycopeptide antibiotic vancomycin was chosen as proof-of-concept interface system due to its medical and clinical relevance, and extensive prior study. Three distinct protein scaffolds representing diverse binding modes underwent *in silico* design using the ROSETTA protein design program to attempt to produce a novel interface capable of binding the D-alanine target peptide ligand. The resulting protein models were systematically produced and assayed in the laboratory for binding to their intended target using multiple, complementary, high sensitivity assay techniques. Although low affinity interactions were observed for some ROSETTA designed proteins, no high affinity peptide binding proteins were created.

In an attempt to address the failure to produce high affinity binding proteins using ROSETTA, structural characterization of unsuccessful protein designs and additional computational studies of native protein-ligand interfaces were carried out. Taken as a whole, the results obtained in the course of my dissertation research offer insights into the strengths and weaknesses of computational ligand-interface design methods and the structural and biophysical nature of protein-ligand interfaces.

Importance of protein therapeutic development

Protein therapeutics are an important and successful part of today's medical pharmacologic arsenal. The market for clinical protein therapeutics, some $94 billion in 2010, is expected to grow to half of total prescription drug sales by 2014 (1). As of 2008, over 130 therapeutic proteins had been approved for use in humans and treat more than thirty different diseases (2). Therapeutic proteins offer significant potential advantages over classical small-molecule drugs including high specificity, low cross reactivity and off-target effects, novel therapeutic modes and better patient tolerance (3)(2). For a thorough classification and review of the recent state of protein based drugs, see Leader et al. 2008 (4).

CPD as a tool in drug development

Since its inception, computational protein design (CPD) has played an important role in the successful creation and engineering of protein-based drugs. *Post hoc* CPD methods have proven highly successful as a means to modify and refine therapeutic proteins generated through non-computational methods, thereby increasing their utility, functionality and desirable pharmacologic attributes (5). *Post hoc* computational design methods are those used to supplement or extend primary protein function, add additional functionality, or modify secondary protein properties or attributes and in this capacity have proved highly successful. Reduced immunogenicity (6), increased affinity (7), altered pharmacokinetic /pharmacodynamic (PK/PD) properties (8; 9) and the thermostabilization of

medically important proteins (10) have all been successfully achieved using *post hoc* computational methods.

Though a useful and successful part of the protein therapeutic development process, *post hoc* computational methods are typically developed for a specific protein/target system, and are therefore non-generalizable. This restricts the broader applicability of a given *post hoc* method and limits its usefulness in the development of novel therapeutic proteins and strategies.

De novo **versus** ***post hoc*** **design**

In contrast to *post hoc* design, *de novo* design is considered the gold standard for computational protein design method development. Where *post hoc* methods may potentially take advantage of intrinsic properties and attributes of a protein undergoing design, *de novo* protein design, by definition, requires establishing wholly new functionality in a protein that did not previously possess such function. Thus, successful demonstration of a *de novo* design method is generally thought to both require and reflect a more complete understanding of the fundamental biophysics and physiology of a given protein system or function of interest, and is therefore considered to be a necessary first step toward fulfilling the requirements of repeatability and generalizability necessary for establishing CPD as a primary tool for protein therapeutic development.

While *post hoc* CPD methods have been of primary impact to date in the development of therapeutic proteins, recent years have also witnessed the development of exciting new abilities and successful proof-of-concept experiments in the basic science and understanding of *de novo* CPD. The successful *de novo* computational design of novel enzymes (11)(12)(13), protein-protein interactions (14) and DNA endonuclease specificity (15) each demonstrate the immense potential of computational protein therapeutic design, both in the creation of novel therapies as well as dramatically reduced time-to-market for biologic drugs.

Yet there is one specific and important area of the CPD field where basic progress has lagged behind. Once considered a solved problem, the ability to *de novo* design protein interfaces to peptide and small molecule ligands has remained tantalizingly out of reach. In contrast to other basic protein functions, a generalized computational method for the *de novo* creation of ligand binding has yet to be demonstrated.

Biological role of protein-ligand interfaces

Protein-ligand interfaces are essential in biology and many fundamental cellular processes are accomplished and regulated through the interaction of protein with ligand. Such non-covalent protein-ligand interfaces form the functional basis of classical small-molecule pharmacology (16)(3), are critical to endogenous and exogenous receptor-ligand signaling pathways (17) and mediate protein-protein interactions through binding of the unstructured amino acid loops or terminal tails

of larger proteins (18). For this work, we define a ligand as an unstructured amino acid sequence of 10 residues or fewer, or a small molecule of 1,000 Da or below.

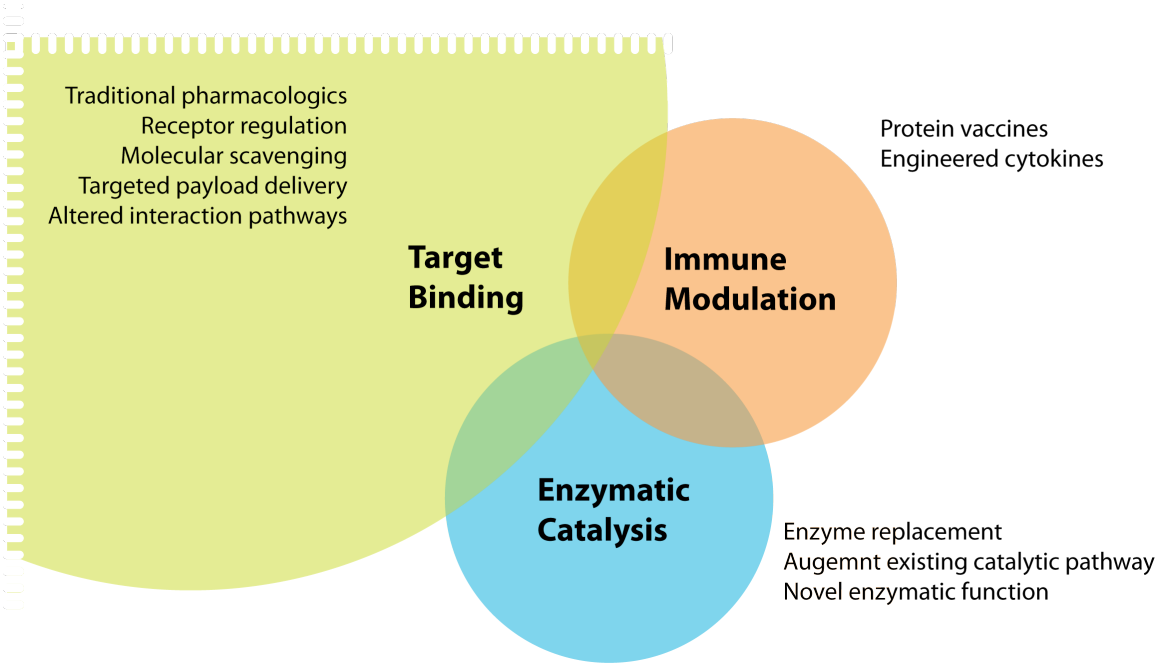


Figure 1.1 Therapeutic functional strategies. Target binding is one of three basic strategies for achieving therapeutic effect, and by far the most commonly employed in drug development. Target binding forms the basis of traditional pharmacology and is the primary mechanism through which direct receptor regulation (antagonistic, allosteric, etc.) may occur. The vast majority of drugs currently available or in development utilizes binding interfaces as their primary mechanism of therapeutic action. The Venn diagram shows the relationship between the three strategies and lists some current and potential functional strategies available through CPD methods.

Functional strategies for protein therapeutic design

Broadly, three basic functional strategies exist for proteins as therapeutic agents: target binding, enzymatic catalysis and immune-modulation (Figure 1.1) (19)(20). Each of these three fundamental strategies has proven successful in protein

therapeutic applications. Protein based conjugate vaccines and cytokines have demonstrated efficacy against pneumococcal and meningococcal bacterial diseases (21) and enzyme replacement therapy (ERT) has provided effective treatment strategies for lysosomal and other enzymatic defect diseases traditional pharmacologic agents could not (22). However, as in classical small-molecule pharmacology, by far the most successful and prevalent functional strategy of the protein based therapies has been target binding, as demonstrated by the domination of the protein therapeutic market by antibody-based drugs (2) which perform their function through tight binding to their targets. Yet, none of the antibody-based therapeutics approved for clinical use have had their binding interfaces designed fully *in silico*, instead relying on conventional laboratory based engineering techniques to generate binding functionality.

This lack of *de novo* ligand binding design capability is unfortunate. Although advancements in other areas of CPD have allowed the development of new types of therapies and therapeutic modes, ligand binding is, and likely will continue to be, the primary mode of therapeutic action for pharmaceutical development into the foreseeable future. Without a reliable and generally applicable method for the *de novo* design of protein interfaces to target ligands, a major avenue of computational protein therapeutic development will remain out of reach.

Not a solved problem

As recently as 2007, the computational design of proteins capable of binding novel peptide and small molecule ligands was considered by some to be a solved problem. Most notably, beginning in 2001 the Hellinga group at Duke University published in several leading peer reviewed journals results of a repeatable computational method for designing proteins capable of binding a variety of ligands, including metals, explosives, biowarfare agents and neurotransmitters (23)(24)(25). Thus, at the time progress in interface design seemed to be keeping pace with other advancements in the CPD field.

Questions arose, however, after researchers were unable to replicate several of the Hellinga results, culminating in a 2009 PNAS article by Schreier et al. titled "Computational design of ligand binding is not a solved problem" (26), and the eventual retraction of key Hellinga group publications (27)(28)(29). With these once accepted achievements in doubt, attention has again focused on demonstrating basic progress in interface design method development. To date, success at *de novo* design of ligand binding interfaces have been generally limited to coordinated binding of metal ions (30), and a broadly applicable and automated computational process for *de novo* design of protein-ligand interfaces has yet to be demonstrated.

Advantages of protein therapeutics over traditional pharmacologics

A critical motivation behind the use of proteins as scaffolds for the rational development of therapeutics versus the design of traditional small molecule pharmacologics lies in the flexibility and manipulability inherent in nature's amino acid building blocks. Using nature's own fundamental machinery to construct designed proteins greatly simplifies and streamlines difficulties associated with their design, production and manufacture (31). The ease with which genetic and amino acid sequences can be manipulated both in the laboratory and *in silico* allows for extremely rapid design and re-design processes. Cloning, expression and large-scale production of designed proteins are achieved via well established molecular biology techniques (20) without the need to develop often complex and laborious chemical synthetic strategies.

Beyond production considerations, the use of protein scaffolds for therapeutic design also offers other engineering and physiologic advantages. The ideal protein therapeutic scaffold would be relatively small (to reduce immunogenicity) (32)(33), single-chain (to facilitate oligomerization or the addition of functional payloads) (20), thermostable, and cysteine-free (to facilitate expression in the reducing environments of the bacterial cytoplasim).

Each of these traits can be conveniently achieved through a combination of careful scaffold selection and rational design. The ability to manipulate cysteine composition also allows for the convenient site-directed coupling of effector compounds (34). The potential fusion to other effector molecules such as toxins

(35) or cytokines (36; 37) expands the therapeutic potential for rationally designed proteins. The design of multivalency or oligovalency (38) is also possible with designed proteins as an effective means of increasing affinity for a target. Additionally, site-directed chemical and post-translational modifications such as PEGylation (39-41) and glycosylation (42) can be readily employed to modulate properties such as serum half-life and tissue penetration (20), endocytic trafficking (31), and immunogenicity (32; 43-45) as well as other pharmacokinetic and metabolic properties (40-42).

Fundamentally, although proteins sometimes suffer from significant drawbacks compared to traditional pharmacologics when used as therapeutic agents – .e.g. immunogenicity, poor gastrointestinal uptake, proteolysis, etc. – the manipulability and functional flexibility inherent in protein/peptide structure conveys considerable advantages in ease and rapidity of design and production.

A large and growing market for protein therapeutics

For many of the above reasons, the market for protein therapeutics, which began in the 1980's, has boomed in the 90's and into this century. Excluding immunoglobulin based therapies, the market for recombinant protein therapeutics has more than doubled in that last five years, and what is today a $51 billion market is projected to continue growing to $87 billion by 2010 (46). Dozens of highly successful protein therapeutics produced by an array of biotechnology companies are currently available for clinical applications, and many dozens more are presently in the

development pipeline. Examples of successful protein therapeutics span a wide range of disease application, including diabetes, anemia, hepatitis, autoimmune diseases and cancer, among others (47). This wide range of successful clinical applications along with the large and rapidly expanding market demonstrates the considerable potential for proteins as therapeutic agents. However, while the market for protein therapies is rapidly growing and the science behind these advances continues to mature, no manufacturer has yet brought to market a protein based antimicrobial therapeutic. This is possibly due to the lack of economic incentives discussed below, but nevertheless represents a critical failure of the biotech and pharmacologic industries to address an urgent public health need.

The need for new antimicrobials and the proof of concept model system

Pharma and biotech are failing to meet the need for new antimicrobials

Modern drug development is a slow and costly endeavor. In 2003 the average time to market for new drugs was estimated to be 15 years, at a cost of $0.8 to $1.7 billion per drug (48; 49). Fewer than 1 in 5000 (0.0002%) of the promising drug candidates that enter pre-clinical testing ultimately receive regulatory approval (50). Due to a convergence of economic and regulatory constraints in the pharmaceutical and biotechnology industries, many urgently needed drugs not fulfilling specific marketing requirements are failing to be developed. Among the categories of therapeutics whose new research and development has been greatly

curtailed or eliminated in recent decades are the antibiotic/antimicrobial compounds (51).

In part, the rate at which antibiotic resistance to a drug arises is a major economic disincentive for pharmaceutical companies and greatly discourages the allocation of substantial resources to the problem (50). In the past 50 years, a total of ten new classes of antibiotics possessing novel modes of action have been discovered, yet just two of those new classes were discovered in the past 30 years (52). In 2002, out of 89 new medicines entering the market, none was an antibiotic (53), and the major industry development programs that remain focused primarily on creating close chemical derivatives of existing antibiotics (51). Because they share a common target and mechanism of action, the useful lifetime of these derivative antibiotics is substantially more limited due to greater susceptibility to enhanced microbial acquisition of multidrug-resistance than are novel classes of antibiotic. Particularly disturbing is that there is no existing antibiotic class for which a bacterial resistance mechanism has not already been documented (52).

The rapidly spreading threat of multidrug microbial resistance

Gram positive microbial pathogens are a major cause of morbidity and mortality around the world. In the U.S., an estimated 19 million hospital patients are at risk for developing gram-positive infection annually, and more than 2 million each year contract an infection from hospital visits (54). Of these hospital acquired (nosocomial) infections, more than 70% do not respond to one or more of the first-

line antibiotics and between 24% and 45% of all gram-positive microbial infections are resistant to multiple classes of antibiotic (53; 55). The microbial pathogens responsible for the majority of resistant infections are the Staphylococcus aureus and Enterococci strains, which together account for greater than 65% of all life-threatening infections (56). Most alarming has been the rapid emergence and spread of multidrug- (or methicillin) resistant strains of S. aureus (MRSA) and Enterococci (MRE) whose prevalence are now widespread and increasing in both hospital and community settings (57; 58). Between 1987 and 1997, reported cases of MRSA in intensive care units approximately doubled, and both MSRA and MRE infections are now epidemic in many hospitals worldwide (59-61). This rapid spread of multidrug-resistant microbes in clinical environments has begun to impose serious limits on treatment options, as few pharmacologic agents remain capable of combating these strains.

The antibiotic of last resort, under threat

Since its introduction in the 1960's, the preferred therapy for treating multidrug resistant microbial infection has been the glycopeptide antibiotic vancomycin. Often referred to as the "antibiotic of last resort", vancomycin's continued utility as an effective treatment for multiply resistant microbial infection is now in doubt due to the recent emergence of additional vancomycin resistance in many of the MRSA and MRE strains (57).

Before 1987 no hospital in the U.S. had reported a case of vancomycin resistant microbial infection. Today, cases of vancomycin-resistant S. aureus and Enterococci, known as vancomycin-resistant S. aureus (VRSA) and vancomycin-resistant Enterococci (VRE), have been reported worldwide and are the third most common cause of healthcare associated infection (57). Equally troubling from a public health perspective has been the recent emergence of reduced vancomycin susceptibility among infections acquired in community settings, such as schools and other public venues (62; 63), demonstrating that resistant microbial pathogens are no longer confined to hospitals.

How vancomycin works

The molecular basis for vancomycin antimicrobial action against Gram-positive bacterial strains is by interfering with proper cell wall biosynthesis. Specifically, vancomycin inhibits peptidoglycan synthesis of the gram-positive bacterial cell wall by binding and sequestering the D-alanyl-D-alanine portion of the cell wall precursor glycopeptide, thereby preventing the peptidoglycan cross-linking necessary for the cells structural integrity, resulting in bacterial lysis and death (64) (Figure 1.2a,b). Studies have shown that binding only a small percentage of this D-ala-D-ala glycopeptide target during cell wall biosynthesis is sufficient to kill gram-positive bacteria (65).

How vancomycin fails

The most common mechanism of acquired resistance to vancomycin observed in pathogenic microbial strains is through the substitution of a D-lactate in place of the D-alanine at the free C-terminus of the bacterial glycopeptide. This single replacement of the C-terminal amino linkage by an ester linkage of the lactate results in loss of an inter-molecular hydrogen bond and introduces a repulsive interaction between oxygen lone pairs, with the latter believed to contribute more to destabilization of the binding interface (66) (Figure 1.2c). This observation suggested that removing the lone pair-lone pair clash might be sufficient to restore vancomycin binding to D-ala-D-lac targets (67; 68). Using this strategy, a vancomycin analogue was synthesized that bound both D-ala-D-ala and D-ala-D-lac peptides with similar affinities and was more effective than vancomycin against VRE (70). Binding, however, was in the millimolar range and therefore not amenable to use as a therapeutic. Moreover, total chemical synthesis of this molecule as a drug is not compatible with large-scale processing.

New approaches to antimicrobial development are urgently needed

Of greatest concern is the discovery that all of the genes necessary for resistance to vancomycin have been found on a single transmissible plasmid, and that cross-species transfer of this plasmid is believed to be responsible for acquisition of bacterial resistance in the wild (71). These resistance genes are collectively known as the VAN system, and numerous variants have been identified to date (72). It is the

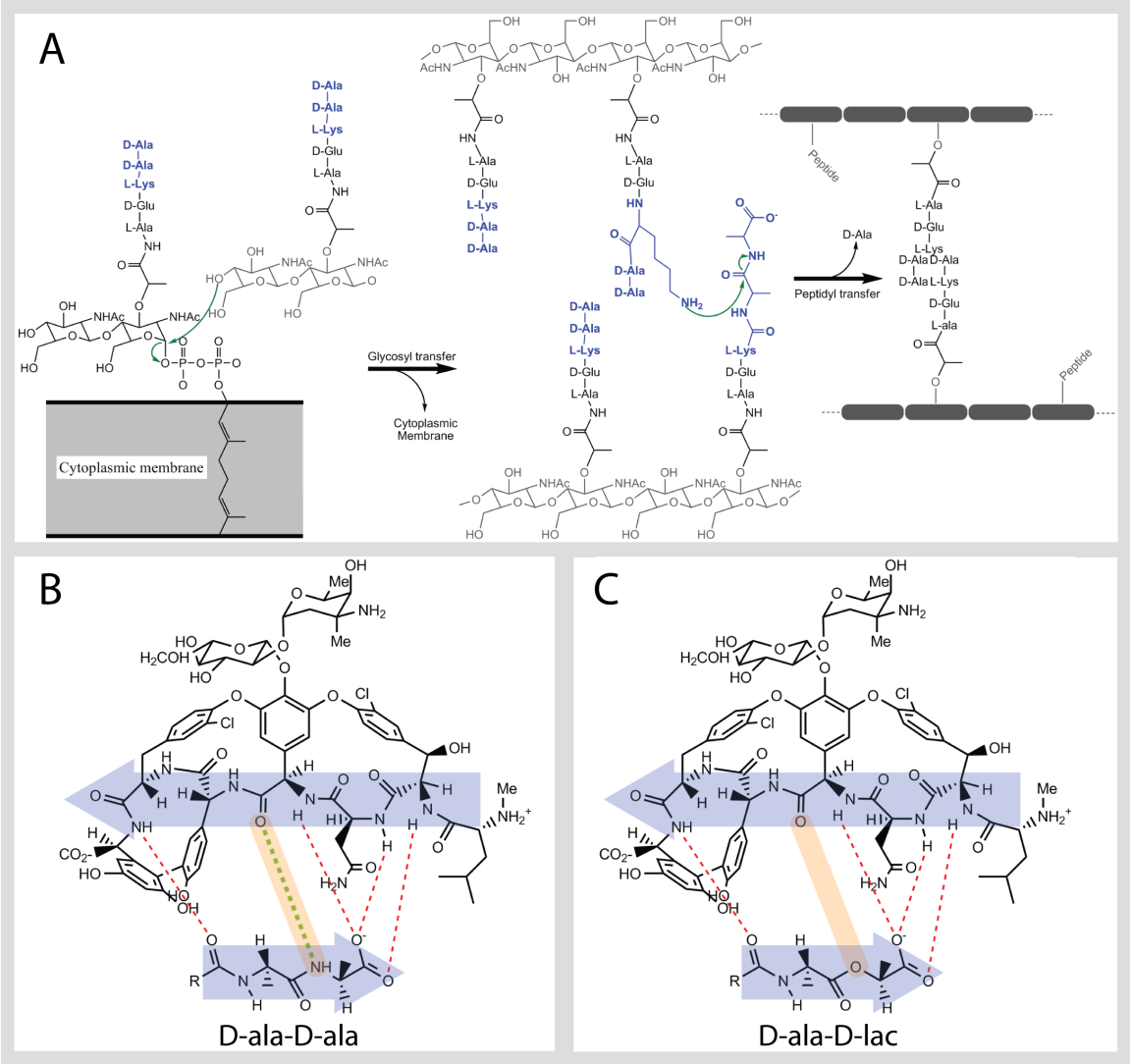


Figure 1.2 Vancomycin Binding Mode and Mechanism. (A) Biosynthesis of gram-positive cell wall via transglycosylation and transpeptidation reactions are inhibited by vancomycin binding the C-terminal -D-ala residues (blue). (B) Five hydrogen bonds stabilize the interaction between vancomycin and the -D-ala-D-ala peptide in a back-bone/β-strand binding mode (lt. blue arrows). (C) Upon substitution of -D-lac at the peptide C-terminus, the loss of a hydrogen bond and resulting lone pair repulsion impart vancomycin resistance. Adapted from(69)

existence of this plasmid and its ability to efficiently confer vancomycin resistance within and across species that constitutes the greatest threat to continued efficacy

of current last-line antimicrobial therapeutics, and portends the greatest need for new classes of therapeutic agents.

This confluence of factors – the rapid rise and spread of MSRA and MRE, the recent emergence of vancomycin resistant microbial strains via a readily transmissible genetic element, and the dismaying lack of industry innovation and investment in discovering new classes of antibiotic, point to an urgent need for new approaches to antimicrobial therapeutic development.

Past ROSETTA protein design studies and successes

The ROSETTA suite of programs encompasses a number of computational functionalities focused on protein prediction and design (73; 74). A central idea behind ROSETTA is to reduce the complexity of the conformational search space by sampling discrete conformations of protein sidechains (rotamers) (75-78). ROSETTA energy functions used for design and scoring of sampled models rely on statistical parameters derived from databases of known protein structures. These "knowledge-based potentials" increase the accuracy of scoring functions for evaluating the designed sequences (79; 80).

In the past decade, ROSETTA has enjoyed tremendous success in application to a wide variety of design problems, including the thermostabilization of an enzyme (81) and design of a novel sequence and topology (80). The latter result was particularly exciting as the designed protein, top7, was soluble, monomeric and exceptionally

stable, and the accuracy of the design was confirmed by the high resolution crystal structure which confirmed the design model to within 1.2Å. The creation of an entirely new fold represents a milestone in the design field as it demonstrates the ability to access regions of protein space not yet observed in nature.

More recently, ROSETTA has been used to reengineer protein complexes (82-84)and to successfully redesign the specificity of a protein-protein interface (85-87). In one study, ROSETTA designed Dnase-inhibitor protein pairs exhibited sub-nanomolar affinities *in vitro*, and a high-resolution crystal structure of the designed complexes confirmed the computational model (88). Moreover, the designed proteins were functional and specific *in vivo*. This study illustrates the potential of computational interface design to create new protein pairs that are both specific and functional in their biological context within living cells.

ROSETTA's ability to predict protein-protein interactions has been demonstrated in the Critical Assessment of PRedicted Interactions (CAPRI). Researchers are given the structures of two proteins and challenged to predict the structure of the complex. There, ROSETTA predictions for targets were strikingly accurate. Not only were the rigid-body orientations of the two partners predicted perfectly, but interface sidechains were also modeled with a high degree of accuracy (89; 90).

Most immediately relevant for this proposal are two ROSETTA developments. The first is ROSETTA's successful application to enhancing the affinity of a protein-peptide complex. Peptide extensions were designed for p53 and dystroglycan-based peptides that bind with increased affinity to the Mdm2 oncoprotein and to

dystrophin (91). This experiment established ROSETTA's applicability to the design of protein-peptide interfaces similar to that attempted in this project. Second, was the implementation and testing of a ROSETTA version that enables, for the first time, design with modified amino acids and other small molecules such as the D-ala-D-lac peptide (92) employed here. This work was completed by my advisor, Jens Meiler, and is the primary enabling computational advancement underlying my dissertation research.

General methods for computational interface design

De novo protein interface design is a specific branch of the larger CPD field. Accordingly, the established computational tools used for interface design are derived from generalized CPD and structure prediction methods. CPD is often described as an inverse-folding problem, with the goal of identifying amino acid sequences compatible with a given three-dimensional protein structure (93). This definition can be extended to interface design, where the goal is to identify a sequence capable of forming a three-dimensional ligand-binding interface. Thus, the focus of interface design is more localized than general CPD, and requires higher accuracy and precision.

Both the generalized protein and interface specific design methods share two general components: a search algorithm to efficiently sample the often vast sequence-conformation space, and a scoring function (also referred to as a fitness function) for discriminating optimal from sub-optimal sequences. For in-depth

reviews of general computational protein design methods, see Lippow & Tidor, 2007 and Alviso et al., 2007 (94)(95).

Sequence/structure search algorithms

Sequence space search algorithms can be classified as either stochastic, or deterministic. Commonly used stochastic search algorithms are Monte Carlo-Metropolis with simulated annealing (Metroplois et al. 1953)(Kirkpactric et al. 1983), fast and accurate side-chain topology and energy refinement (96), genetic algorithms (97), and self-consistent mean-field optimization (98). Stochastic algorithms have the advantage that they will always find a solution to a search query, though the solution is not mathematically guaranteed to be the most optimal. These algorithms can be scaled to take advantage of massively parallel or distributed computing resources. For a review of the commonly used search algorithms see Volgt et al., 2000 and Tian, 2010 (99)(100).

Conversely, deterministic search algorithms such as dead-end elimination will not always be able to arrive at a solution to a given design problem and can be difficult to scale. However, when a deterministic search algorithm is able reach a solution, it can be mathematically proven to be the global minimum-energy conformation for the given input parameters (Desmet et al. 1992).

A common method to further facilitate efficient search of the sequence-conformation space are predefined rotamer libraries. Roatmers are preferred low

energy conformations of each amino acid side chain, derived statistically from the protein data bank (PDB). Rotamer libraries are pre-computed sets of the most common rotamers for each residue type, and can be either backbone-dependent, or backbone-independent (101). By using rotamers as the basis of a sequence-conformation search, the two variables of amino acid identity and conformation can be combined, greatly reducing compute times.

Energy, scoring and fitness functions

Once the search algorithm identifies a specific protein sequence-conformation, a potential energy function is used to evaluate each protein model based on the overall energetics of the system. There are two general approaches to the potential energy functions used in protein design, knowledge-based and physics-based energy potentials. For in-depth reviews of potential energy functions used in protein design, see Boas & Harbury, 2007 and Lippow & Tidor, 2007 (94; 102).

The knowledge-based energy potentials are derived statistically from structures deposited in the PDB, where the 3D coordinates of each protein are converted first into a statistical potential, and then into an energy potential for a given sequence-structure parameter (103). Knowledge based energy potentials typically contain individual terms for van der Waals, electrostatic, hydrogen-bonding, internal entropy, solvation and other energy components.

Knowledge based approaches have the advantage of being able to capture large amounts of empirically derived data into efficient mathematical functions. These functions can then be used to score and evaluate protein sequence-structure models (104).

Physics based energy potentials rely on more complex mathematical models of the basic physical forces that constitute a protein free energy (105). They can be more accurate than knowledge-based methods, but are computationally more expensive. Protein design applications are typically performed using knowledge-based methods due to combinatorial and compute time constraints involved in sampling large sequence-conformation spaces. However, recent attempts to validate physics-based molecular mechanics potential energy functions in protein design have met with modest success in the design of low affinity ligand interfaces (106) and may find further application in the long-term with continued increases in generalized computing power.

General protein design algorithms and protocols

Repetitive, cyclical application of search and scoring algorithms form the basis of a generalized protein design algorithm (Figure 1.3). 3D structure coordinates for starting ligand and protein design scaffold are input, along with an appropriate rotamer library, into the sequence-structure search algorithm. Multiple cycles (often tens to hundreds of thousands or more) of sequence-structure search followed by scoring of the identified protein model are used to evaluate the design search space.

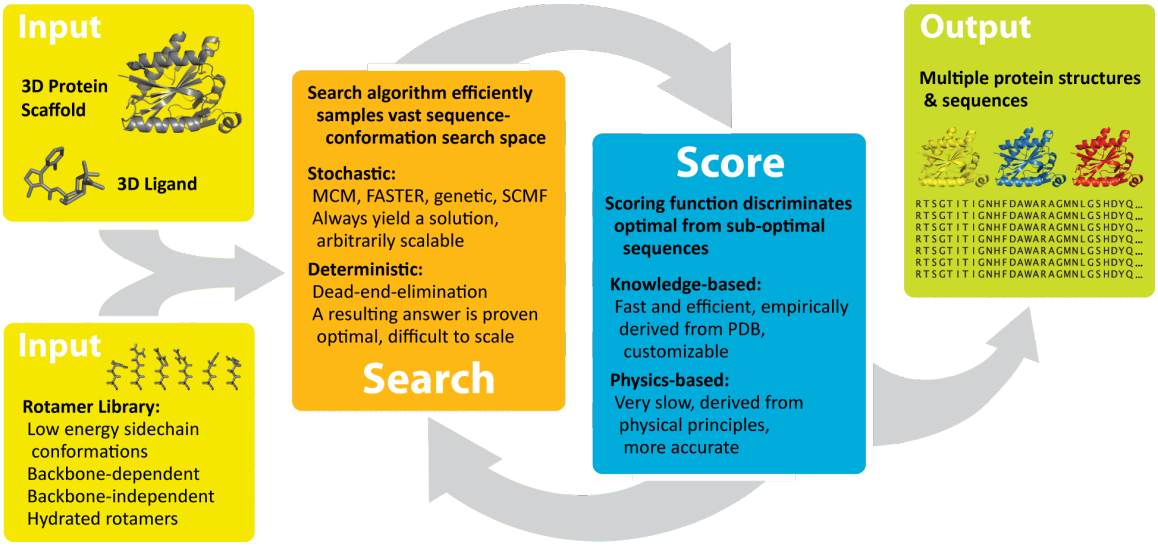


Figure 1.3 General components of an interface design algorithm. Protein-ligand interface design algorithms require the input of a 3-dimentional protein scaffold and ligand on which to perform design. Various rotamer libraries of statistically likely, low-energy conformations of amino acid sidechains or ligands may also be input to reduce search degrees-of-freedom. The design algorithm proceeds through repeated rounds of sequence-conformation search, followed by scoring of each resulting model. If a given sequence-conformation model does not meets predetermined scoring criteria, that model undergoes further sequence-conformation perturbation by the search algorithm. The cycle continues until a sequence-conformation model meets scoring criteria and is output as a sequence and/or 3D protein-ligand model. Typically, multiple models are output for further iterative rounds of design and evaluation.

Models that the scoring function determines meet specified criteria are output as 3D coordinate files. Typically, many thousands of protein models are generated to assure sufficient and unbiased sampling of the sequence-structure search space. These accepted output models might then be further evaluated for other desirable design characteristics not otherwise encompassed by the scoring function such as ligand pose, sequence diversity, etc. Design protocols composed of multiple, iterative rounds of design and model generation can be tailored to a specific design goal.

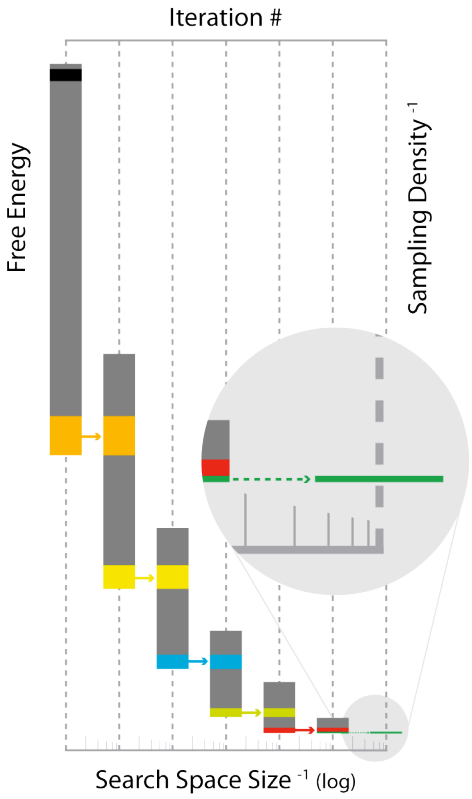


Figure 1.4 Example of an iterative design protocol. Many protocols for protein-ligand interface design use consecutive rounds of computation and enrichment of the best scoring models. The black portion in the upper leftmost bar represents the starting 3D structures of protein with crudely placed ligand. The computed free energy of these starting models is generally high due to clashes of the ligand with protein atoms in the interface intended for design. During the first iteration, the design algorithm identifies and outputs models lower in computed free energy. Of the models generated in the first iteration, a portion possessing the lowest energy scores (orange) are used as starting structures for the second iterative round of design in which the search permutation parameters (e.g. translation, rotation, sequence, conformation, etc.) are narrowed. This narrowing of search parameters serves to decrease the total size of the search space, and increase the sampling density. This iterative process of enrichment and increase in sampling resolution continues until computed energy levels begin to plateau and/or sequences converge – often several rounds or more, depending on the size of the interface under design and degrees of search freedom (e.g. ligand flexibility, multiple ligands, co-factors, etc.). At the end of the iterative design process, a small portion of the protein-ligand interface models possessing the lowest overall free energy are often evaluated manually to assess success.

A typical iterative design protocol for an MCM search function is shown in Figure 1.4. In each iterative round of design, low energy models are selected and carried over to the next design round. At each successive round, the size of the search space decreases through eliminating non-productive areas of the sequence-structure space, causing the sampling density of the remaining search to increase. When decreases in free energy of the protein-ligand complex plateau, a small number of the best scoring models are output as the end product of the design protocol. These proteins may then go on to be expressed and tested experimentally for their predicted function.

The protein design search space

The calculations necessary to perform *in silico* protein design can be vast and difficult to conceptualize. For this reason, it can be helpful to imagine the multiple parameters involved in generating and evaluating a good from a bad protein design – things like protein core stability, satisfied H-bond donors and acceptors, amino acid identity and conformation, van der Waals interaction, etc. – as a multi-dimensional space on which each given permutation of the parameter set is a point. This multi-dimensional parameter space is often referred to as the "search space" for a given protein design computation.

Systematic or brute-force computation of all possible permutations in the search space can quickly grow beyond astronomical scales for even relatively simple systems. For example, varying just a single parameter such as the primary sequence

using only the 20 natural amino acids, in a relatively modest sized protein of 100 amino acids, yields the following search space size: $20^{100} \approx 10^{130}$ which is 50 orders of magnitude larger than the estimated number of atoms in the observed universe. Add to this that the average amino acid contains ~3 rotatable bonds conferring a near-infinite number of possible conformational states, and that amino acid substitutions will often result in changes in backbone phi/psi angles. Further add conformational flexibility, rotational and translational degrees of freedom of a ligand, etc. and we see that an already intractable calculation expands exponentially still. Indeed, in mathematical terms, protein design has proven to be an NP-hard calculation (107), the most difficult category of computational problems to solve. Needless to say, systematic evaluation of even simple protein design problems is infeasible for the foreseeable future.

To address the difficulties of a vast search space, several approximations are typically made. First, the conformational flexibility of amino acid sidechains are represented by a set of discrete, low energy conformations called rotamers (101). The use of rotamer libraries during design thus combines the identity and conformational degrees of freedom into a single pool of sequence-conformation parameters. Second, the protein backbone is often kept fixed during the early rounds of design computations. Third, computational design protocols consisting of multiple, iterative rounds of increasing resolution and complexity are used to exclude large regions of the search space, which cannot result in productive protein designs, while subsequently focusing more intensely on potentially productive regions (See Figure 1.4). However, even with the use of these approximations to

reign in the potential search space, a systematic evaluation remains infeasible. Thus, efficient algorithms are employed to sample the large multi-dimensional search space in a fashion that insures both sufficient density and breadth to identify low energy designs (see Figure 1.3).

These and similar methods have made feasible the computational design of proteins and protein interfaces. Nevertheless, all but the simplest computational protein design efforts are undertaken on modern grid- and supercomputing clusters and can require tens and even hundreds of thousands of CPU-hours per design.

Overview of ROSETTA design methods

The ROSETTA program uses a combination of Monte-Carlo and gradient-based search algorithms together with knowledge-based statistical analysis and rotamer libraries to create protein models. This combination of search methods assures an unbiased sampling of the global conformational energy landscape of the protein system. Sampled structures are scored by environmentally dependent, atomic resolution energy functions derived from first-principle calculations and knowledge-based statistical methods (108). The ROSETTA program code is designed to take full advantage of modern advances in grid-computing architecture such as Vanderbilt's Advanced Computing Center for Research and Education (ACCRE).

In the ROSETTA dock/design mode utilized in my dissertation research, peptide conformers from a pre-computed ligand library are placed into the binding site of

the protein scaffolds to create starting template structure files, one for each ligand conformer scaffold pair. ROSETTA then performs a "random" Monte-Carlo translation and rotation of the peptide ligand, followed by Monte-Carlo substitutions of sidechains in the protein binding site – referred to as "repacking" – using rotamer libraries with simulated annealing. All amino acids in the first and second shell of the peptide binding site are included in the design process. Following each permutation, the free energy of the model is calculated and accepted or rejected using the Metropolis criterion (109). If the energy of the new model is lower, the old model is discarded based on probability criteria and the permutation and scoring process begins again using the new model. If the new model possesses higher energy than the previous model, the new model is discarded based on probability criteria and the process is repeated using the previous best scoring model as a start point for further permutations. ROSETTA repeats this process until the desired number of low energy models for each starting structure is reached. A dock/design computational round is complete upon outputting the desired number of designed models.

At the completion of each computational round, the models are sorted based on lowest overall free energy of the system, and the best scoring models – at some user-designated level of energy cutoff – are carried into the next cycle. A minimum of five cycles of computation are performed. The intent of this multi-round, iterative approach is to achieve sufficient sampling density of the total conformation/energy search space, while progressively increasing the sampling density of a subset of the

search space shown to be enriched for low energy conformational minima of the protein-ligand complex.

After all design rounds are complete, selected lowest energy models undergo gradient based minimization and repacking without design using a “hard-repulsive” scoring function that is more accurate, but too computationally slow and restrictive to use during the design rounds. Further holistic evaluation of these lowest energy models yields a small number of designed structures that can be prioritized for laboratory expression and assay.

CHAPTER II

RESEARCH DESIGN AND EXPERIMENTAL STRATEGY

The experimental proof-of-concept of my dissertation research employed a rational design approach to develop and test computational methods for the *de novo* design of protein interfaces to small ligands. Naturally occurring PDZ, TPR and 1m4w proteins were used as scaffolds to design high affinity binding to the D-alanine-D-alanine and vancomycin resistant D-alanine-D-lactate target peptides.

Vancomycin resistant bacteria replace an amide bond with an ester bond at the C-terminus of the vancomycin target peptide (Figure 1.2e). This loss of a stabilizing hydrogen bond along with the resulting oxygen lone pair repulsion is sufficient to eliminate vancomycin activity. The proposed designed proteins were intended to be capable of counteracting this D-ala to D-lac substitution by providing a compensating hydrogen bond donor while retaining the potential to bind both D-ala and D-lac peptides, thus potentially generating a bi-modal binder equally effective against resistant and non-resistant bacterial strains.

My dissertation sought to implement three specific Aims: Aim 1 focused on the identification of suitable protein scaffolds, and the computational design of the proposed protein-ligand interface using ROSETTA. Aim 2 involved the laboratory

production of the designed proteins and development and validation of assays to measure protein/peptide binding. Aim 3 was to quantify the binding affinities of the designed proteins for their intended peptide targets, and initiate high resolution structural characterization of the designed proteins to assess the accuracy and efficacy of the *in silico* design process.

The logical flow of this research design implemented stepwise filtering and enrichment of results from one research phase to the next (Figure 2.1). At each phase, only the most promising design candidates were carried forward to the next step in the protocol. The computational process generated and evaluated the energy of hundreds of millions protein-peptide permutations, and output the 0.1% (several thousand) lowest energy models from each cycle of computation. At the end of each stage, additional filters were applied to select the most promising candidates for protein production. Such filters seek to remove designs with limited access of the peptide N-terminus to the protein binding site, minimize the total number of mutations, and ensure native-like binding modes and energies.

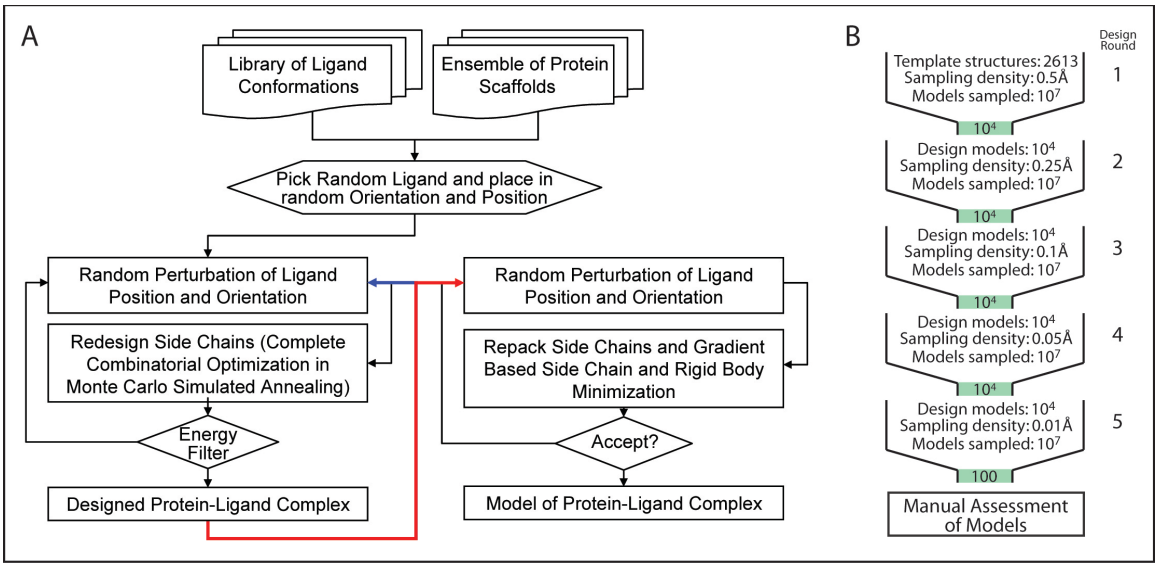


Figure 2.1 Diagram of computational protocols and strategies. (A) Flowchart of ROSETTA computational process showing the multi-step, iterative nature of the ROSETTA design and scoring procedures. Only models that achieve specified minimum energies are accepted and output. (B) Schematic of design protocol. At each cycle, starting structures are used to create a large number of designs, which then undergo filtering before being carried to the next cycle. In each of five cycles, the sampling density is increased by reducing the design perturbation parameters. After the final round of design, the output models are manually assessed to determine the best overall candidate designs.

Subsequent to computations, but before synthesis and expression of the designed proteins had begun, a structure/function alignment of all candidate sequences of a given scaffold design was performed to identify overlapping binding modes. Since many of these sequences shared mutations, by evaluating the binding-sequence space, it was possible to devise a maximally efficient strategy for gene synthesis that minimized redundancy between designs. Following protein expression, proper folding was initially confirmed using CD and NMR spectroscopy and the solution properties of the proteins were determined. Binding affinities and kinetics were quantified using a combination of ITC, fluorescence methods and/or NMR

spectroscopy. The rationale behind using a variety of methods with overlapping sensitivity ranges was the expectation of a wide range of binding constants and thermodynamic properties (Figure 2.2).

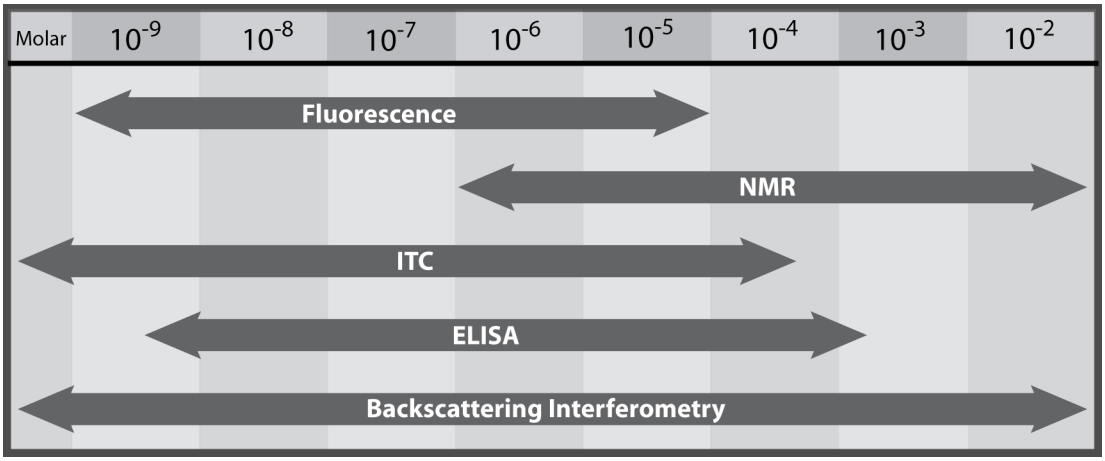


Figure 2.2 Complementary, Overlapping Assays. Representation of the optimal concentration ranges (in M) of the proposed assays for quantifying the binding constants of the designed proteins. Assay complementarity assures that accurate assessment of binding affinities can be obtained even if one or more techniques fail to yield results.

Expressed protein designs of interest were selected for high-resolution structural characterization to assess the accuracy and efficacy of the *in silico* design process at the atomic level. The objective of structural characterization was to assess the causes for lack of observed binding in all of the designed proteins. Defects or noteworthy deviations from prediction were assessed and potential refinements to the computational protocols and ROSETTA program code were evaluated. Due to the unsuccessful design of high affinity binding in any of the produced proteins, preliminary studies into application of the developed methods to other microbial targets were not initiated.

The above research design enabled comprehensive sampling of protein conformational and sequence space *in silico* while permitting experimental characterization of a practical number of proteins at different levels of resolution.

Computational design of high affinity protein binder to peptide targets

Identifying protein scaffolds suitable for design

The first step of the design process was the proper selection of a native protein scaffold on which to begin computations. This was accomplished by searching the PDB for proteins possessing high-resolution crystal structures and a molecular weight of between 10 and 40 kDa. This molecular weight range was supported by studies which have shown the size exclusion limit for the peptidoglycan cell wall of gram positive microbes to be approximately 50 kDa (110; 111). Another highly desirable trait of the scaffolds is thermostability, which was assumed to permit manipulation of binding site residues with decreased risk of destabilizing the overall protein fold.

At the end of the selection process, three protein scaffolds (referred to as PDZ, TPR and 1m4w) were chosen (see Figure 2.3), each possessing a distinct binding mode. All of the chosen scaffolds are stable at temperatures well above physiologic norms, while one of the scaffolds – 1m4w – is derived from a thermophillic organism and possesses a wild-type melting transition above 100°C (112).

Scaffold	PDB I.D.	Fold	Protein Family	Topology	Molecular Function	Biological process	Mass	Species
PDZ	1obx	Mainly beta	PDZ domain	Roll	Protein-protein interaction	Multi-specific protein interaction	~8.4 kDa	Homo sapiens
TPR	1elr	Alpha-alpha superhelix	Tetratricopeptide repeat (TPR)	Horseshoe	Protein-protein interaction	Aerine threonine protein phosphatase 5	~15.5 kDa	Homo sapiens
1m4w	1m4w	Twisted beta-shet	Xylanase/ endoglucacase	Jelly role	Hydrolase activity, polysaccharide binding	Carbohydrate metabolism, phospahte transport	~21 kDa	Nonomuraea flexuosa

Figure 2.3 Biochemical characteristics of chosen protein design scaffolds. The native species and molecular weight are of particular consideration when designing a protein therapeutic due to immunogenic and pharmacodynamic properties.

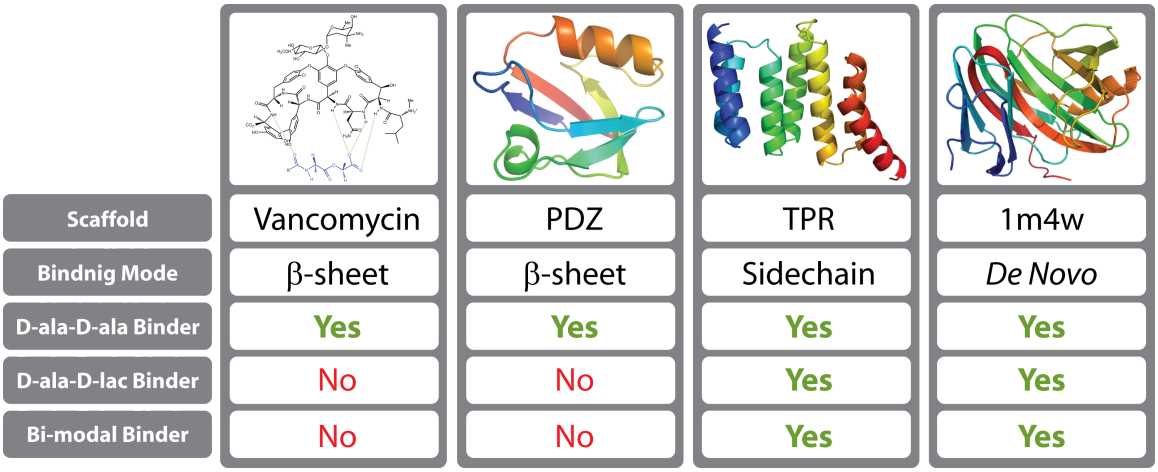


Figure 2.4 Scaffold progression. The logical progression of scaffold choice and design goals for the D-ala-D-ala andD-ala-D-lac vancomycin target peptide, from replication of backbone binding mode (PDZ), to design of a sidecain binding mode (TPR), to *de novo* redesign of a peptide binding interface (1m4w).

Because vancomycin is a glycopeptide and employs a backbone/β-sheet type binding mode, the first design milestone was to replicate this binding interaction using a PDZ domain scaffold. PDZs are mixed α/β domains of ~8 kDa that recognize 4-7 residues at the C-terminus of their peptide targets. Their native mode of binding is primarily by forming highly stable β-sheet type hydrogen bond networks in a manner similar to that of vancomycin (Figure 2.4, 1.2). Although the shared features of this binding mode make it an ideal starting point for designs, analogous lone-pair

clashing of backbone/ester oxygen may prove difficult for these PDZ designs to overcome, and thus make it unlikely for the PDZ designs to be able to efficiently bind D-ala-D-lac peptide targets (Figure 2.5).

The TPR domain is a repeating helix-turn-helix motif of ~16 kDa (Figure 2.3) that, like PDZ domains, natively bind C-terminal residues of protein targets. Unlike the PDZs, the ligand binding mode of the TPRs is exclusively sidechain-mediated. For design applications, sidechains offer a more diverse set of functional groups and increased conformational flexibility. Thus, individual designs created using the TPR scaffolds were anticipated to be capable of binding both D-ala-D-ala and D-ala-D-lac peptides with similar affinities (Figure 2.5).

1m4w is a ~23 kDa thermophilic β-1,4-xylanase which possesses a mainly β-sheet "jelly-roll" topology (Figure 2.3). 1m4w was selected as a design scaffold not only for its high thermostability, but also for the distinct and suitable geometry of its catalytic cleft. This represented the logical next step in the scaffold selection strategy. By designing a peptide binding site *de novo* it was thought possible to create a binding mode that exploits both the high stability of backbone mediated bonds and the flexibility of sidechain mediated bonds (Figure 2.5).

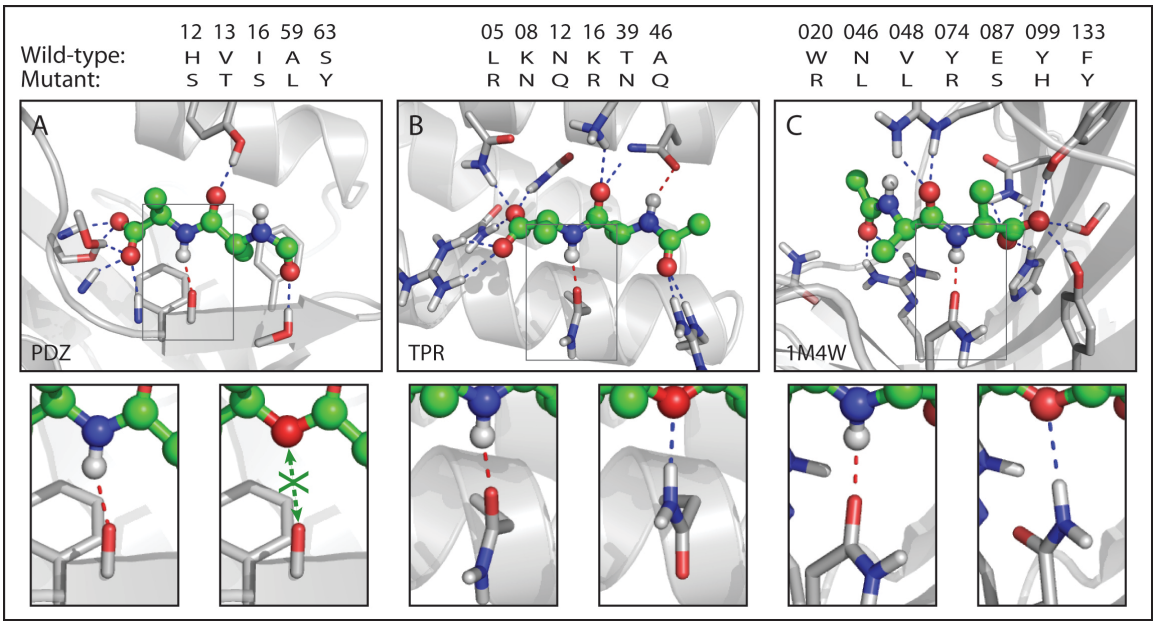


Figure 2.5 Detailed View of Binding Modes of ROSETTA Designed Proteins. (A) Top panel illustrates the backbone mediated hydrogen bond network of a designed PDZ domain. The two bottom panels show close-ups of computed interactions with -D-ala-D-ala (left) and -D-ala-D-lac (right) peptides. The green arrow (lower right panel) denotes a repulsive lone pair interaction identical to that seen for vancomycin, suggesting that the backbone binding mode of PDZ will not allow for -D-ala-D-lac binding. At top, mutations from wild-type are summarized for each design. (B,C) Respective information for TPR (B) and 1m4w (C) designs. Both proteins exhibit a sidechain binding mode. Illustrated in the bottom panels is a single asparagine residue binding both the amide of the -D-ala-D-ala and the ester oxygen of the -D-ala-D-lac peptides through conformational rearrangement, thus acting as a bi-modal binder.

Scaffold progression

The above choice of protein design scaffolds represents a logical, step-wise progression in binding mode, ligand binding functionality and algorithmic complexity involved in the design process. Figure 2.4 shows this progression from vancomycin's mode of binding and inability to bind the D-lac peptide, through the PDZ, TPR and finally *de novo* scaffold of 1m4w with its predicted ability to bind both

D-ala and D-lac peptide equally. This progression also traces the technological complexity of interface design toward progressively more difficult scaffolds, from minimal redesign on PDZ domains which maintain a similar native beta-sheet binding mode, through the design of sidechain binding functionality in the TPR scaffold, to the design novel functionality utilizing both sidechain and backbone binding in the 1m4w scaffold (see Figure 2.5). This step-wise progression was anticipated to permit the incremental testing and validation of ROSETTA's design capabilities at each stage, while also allowing a thorough exploration of the structural and physical properties of the distinct design scaffolds and strategies.

Generation of target peptide conformational library

The peptidoglycan cell wall precursors of all gram positive bacteria share a common sequence, L-lysine-D-alanine-D-alanine, at their C-terminus. Beyond these three residues, the peptidoglycan compositions of different species diverge. It was therefore decided to limit the length of the target peptides to the backbone atoms of the three terminal amino acids plus the methyl sidechains (Figure 2.6). This approach removes the conformational variability of the lysine sidechain, substantially simplifying ligand conformation and protein design calculations while preserving the common gram-positive target sequence.

Because the ROSETTA design mode employed at the time these studies were initiated was unable to impart conformational flexibility to the peptide ligand during the design process, prior to the start of computations, it was necessary to create a

library of target peptide conformations against which to perform the designs. Library creation was accomplished by first calculating allowed φ,ψ angles for the D-ala-D-ala peptide ligand. These computed angles were then used to generate an ensemble of ligand conformations representing systematic permutations around each rotatable bond. For efficiency, the ligand library was parsed to ~2600 peptide conformations for each of the D-ala-D-ala and D-ala-D-lac peptides, which were then used during design.

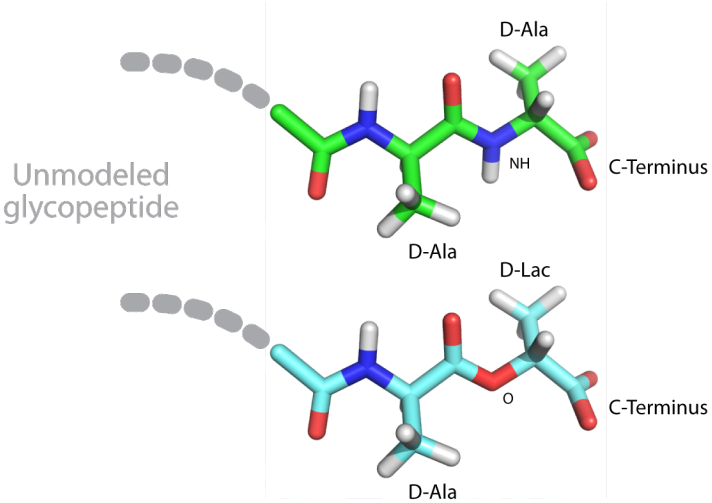


Figure 2.6 Model of D-ala-D-ala and D-ala-D-lac peptide ligands used in computations. D-ala-D-ala (green) and D-ala-D-lac (cyan) ligand models. Only the substitution of a oxygen in place of the C-terminal amide nitrogen distinguishes the two. The remainder of the glycopeptide precursor denoted by grey dashed line, was left unmodeled.

Performance of ROSETTA design computations

Prior to design computations, scaffold structures were subjected to gradient energy minimization to remove atomic level clashes or other defects present from the original crystallographic structure refinement calculations.

Iterative ROSETTA dock/design calculations were performed on ACCRE using a single processor for each staring protein-ligand structure file. (See "Overview of ROSETTA design methods" above.) In the first round of design, 50 low energy models of each of the 2600 starting structures were output for both the D-ala and D-lac ligands (2 x 2600 x 50 = 260,000, Figure 2.1b). Between each computational design round, filtering of output models was performed. Models that occluded egress of the N-terminus of the peptide from the binding pocket were discarded because connection to the remainder of the non-modeled glycopeptide would be impossible. The remaining models were then sorted based on lowest overall free energy of the system, and the best scoring 10,000 were carried into the next iteration of the design commutations. In subsequent rounds of design computation, 100 low energy models are produced for every starting structure for both ligands (2 x 10,000 x 100 = 2,000,000, Figure 2.1b). For all of the designs computations conducted on each protein scaffold, a minimum of five iterative cycles of computation was performed.

After the multiple design rounds were completed, several thousand of the lowest energy models underwent gradient based minimization and repacking using a "hard-repulsive" energy function. One hundred to two hundred of these output lowest energy models were then manually examined and assessed for "desirable" qualitative binding properties that ROSETTA energy functions may not adequately capture.

Structure/function alignments of the binding sequence space for the selected designs were then made and a maximally efficient strategy for expression is devised

that agrees with the gene synthesis strategy (Figure 2.7). Typically, 6 to 12 of the protein designs are chosen for production.

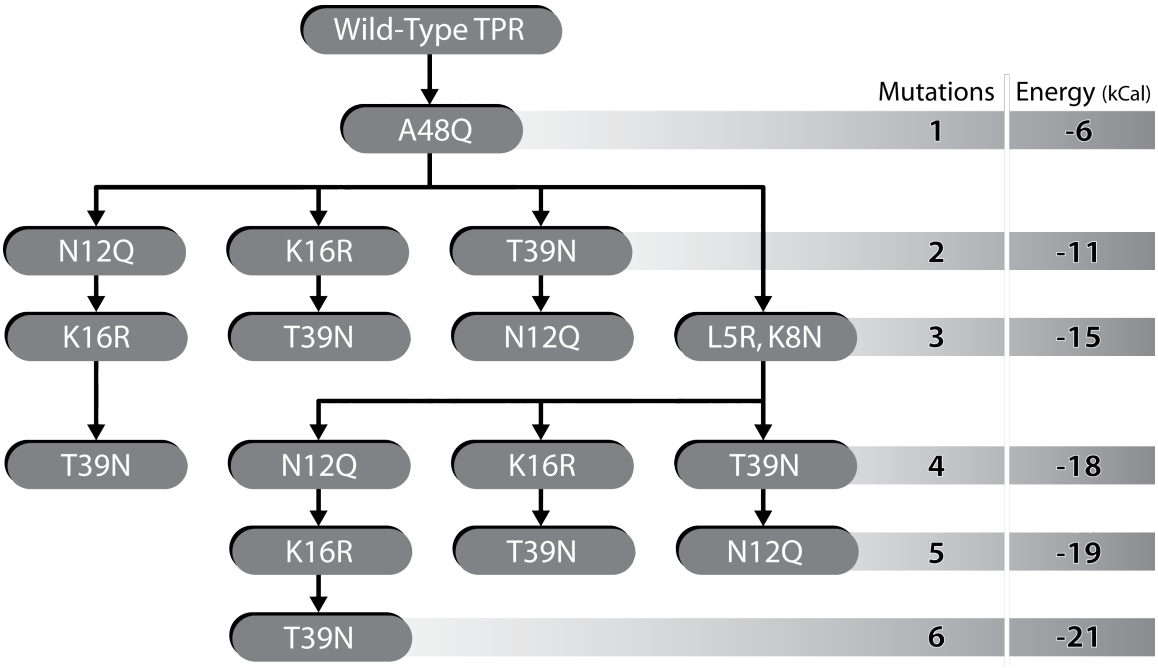


Figure 2.7 Example production strategy for TPR designs. Experimental production strategy for the TPR designs intended to maximize efficiency of protein production and minimize redundancy of sampled binding modes. The four primary target designs highlighted in grey were assembled using recursive PCR, while the remaining designs were produced in a second stage using mutational PCR, starting from the primary set of designed mutants. At right, number of mutations from wild-type for each design, and the ROSETTA predicted energies of binding.

Laboratory production of designed proteins and development of binding assays

Synthesis, expression and purification of designed proteins

Once ROSETTA computations had been completed and an experimental production strategy developed, amino acid sequences were obtained directly from the chosen ROSETTA designed models. Following cloning of the genes, the designed proteins were individually over-expressed and purified. ^{15}N-labeled proteins for NMR studies

were produced in similar fashion using minimal media for which ^{15}N-NH_4Cl was the sole nitrogen source. Pure proteins were then concentrated and assayed for proper folding using CD and NMR spectroscopy and the solution properties of the proteins were determined.

Below is an outline of the general methods used for the production and purification of the ROSETTA designed proteins. All methods and techniques used in the production phase of this proposal followed established laboratory protocols, and whenever possible made use of the previously published experimental conditions for the chosen scaffolds (112-114).

A gene encoding the wild-type scaffold protein was constructed via re-cursive PCR, an efficient method of assembling small to medium size genes (115). Long overlapping oligonucleotides are assembled in a 'one pot' reaction and then amplified to obtain the full-length gene. Mutations for each of the ROSETTA designed sequences were introduced into each PCR assembled scaffold protein gene by simply swapping oligonucleotides. Once assembled, each gene was cloned into a T7-driven E. coli expression vector, which expresses the protein with an N-terminal 3C protease cleavable hexa-histidine tag. Subsequent sequence mutations were achieved using StrataGene QuickChange site-directed mutagenesis method (116). All constructs were confirmed by DNA sequencing.

Following gene synthesis, designed proteins were expressed in E. coli BL21 Star (DE3) cells under standard 37˚C growth conditions in LB medium. Protein over-expression is induced by isopropyl-D-thiogalactopyranoside (IPTG) and cells were

harvested 3-4 hours post-induction. The histidine tagged fusion proteins were purified by immobilized metal affinity chromatography (IMAC). Final purity level of all proteins was greater than 90% as assessed by SDS-PAGE and molecular weights were confirmed by electrospray ionization mass spectrometry (ESI-MS) (See Appendix E). The concentration of the proteins was originally assessed by colorimetric Bradford assay and thereafter by absorbance at 280nm.

Assay for proper folding and solution properties

The secondary structure of each of the purified proteins was determined by far-UV CD spectroscopy and the raw data are converted to mean residue ellipticity. Several designed proteins of each scaffold type underwent 1D 1H-NMR spectroscopy on a 600 MHz Bruker Advance spectrometer equipped with a cryoprobe to confirm that the protein was properly folded. Dynamic light scattering (DLS), SDS-PAGE and analytical Size-Exclusion Chromatography (SEC) were also used to assure proper aggregation state, solubility, high-order structure and the general solution properties of the designed proteins. The DLS measurements were performed with the cooperation of Martin Egli.

Assay of designed proteins to target peptides

Quantify binding of designed proteins using multiple assays

Fluorescence techniques provide a rapid and sensitive means of quantifying binding, and were the primary means of assaying binding in this work. Binding assays using a 5-dimethylamino-1-naphthalenesulfonyl (dansyl-) D-ala-D-ala peptide are well-established for studying vancomycin (117). For these studies, binding titration experiments using dansyl-labeled D-ala-D-ala peptides were performed in solution by adding unlabeled protein to labeled peptide and monitoring both fluorescence emission and anisotropy. The resulting data formed saturation binding curves from which the equilibrium dissociation constants were calculated. Fluorescence experiments were carried out with the assistance of the Beth lab on a T-type polarimeter.

Isothermal titration calorimetry (ITC) was used to measure the change in heat upon complexation of the protein with ligand with no chemical tagging, immobilization or other potentially confounding chemistry required. When applicable, it allowed accurate determination of binding constants, reaction stoichiometry, enthalpy and entropy. However, in the course of these studies, only the TPR based designs were amenable to ITC analysis, as aggregation of the PDZ and negligible evolved heats of complexation of the 1m4w designs prevented their analysis with this method.

NMR chemical shift perturbation assays were used extensively to determine the 2-dimensional ^{1}H-^{15}N HSQC spectra and were collected using ^{15}N -labeled proteins produced and purified as described previously. The uniformly labeled proteins were

concentrated and the target peptide titrated into to the labeled protein solution at specific molar ratios. ^{1}H-^{15}N HSQC spectra were obtained after each titration of the peptide, and the change in chemical shift for each peak was determined and plotted as a saturation binding curve.

Alternative assay methods

Development of a medium throughput antibody-based ELISA screen to rapidly identify binding candidates was briefly pursued, but abandon due to the suitability of other established assay methods and the decision not to produce large numbers of designed proteins.

A Backscattering interferometry (BI) assay was also evaluated in cooperation with Daryl Bornhop's lab as an extremely sensitive, label-free method of detecting protein-ligand interaction. However, following several preliminary experiments, the technology at the time of evaluation was judged to be of insufficient maturity to justify continued testing and validation.

(See Appendix A & B for details of these assays.)

High-resolution structural characterization

Atomic level structural characterization was originally planed on select designed proteins to allow high-resolution confirmation and refinement of the computational

design process through comparison of the predicted *in silico* designed structures to those demonstrated *in vitro*. Though full NMR assignments are available for the PDZ scaffold (118), published crystallization conditions were available for all three - PDZ, TPR, and 1m4w (112-114). The published crystallization conditions for each design scaffold were to be used as a starting point for screening conditions of the mutants. If a protein design demonstrated high affinity binding to its target peptide, crystallization in complex with the ligand was to be attempted. However, in the absence of tight binding, an atomic detail structure of the apo protein would nonetheless be informative. Its structure could be compared with the computational prediction to pinpoint inaccuracies in the computational design protocol and identify the reasons for failure.

Unfortunately, the PDZ scaffold designs exhibited significant aggregation upon purification and were unsuited for both NMR characterization and crystallization trials. Although the TPR based designs possess good solution properties, further examination of the literature revealed that crystallization required the presence and binding of peptide ligand. Thus, initial crystallization screens of the TPR designs were attempted, but abandon upon failure to obtain crystal "hits". The 1m4w designs however, after extensive screening, were crystallized successfully and several high-resolution structures were determined.

CHAPTER III*

COMPUTATIONAL DESIGN OF AN ENDO-1,4-β-XYLANASE LIGAND BINDING SITE

Abstract

The field of computational protein design has experienced important recent success. However, the *de novo* computational design of high-affinity protein/ligand interfaces is still largely an open challenge. Using the ROSETTA program, we attempted the *in silico* design of a high-affinity protein interface to a small peptide ligand. We chose the thermophilic endo-1,4-β-xylanase from *Nonomuraea flexuosa* as the protein scaffold on which to perform our designs. Over the course of the study, twelve proteins derived from this scaffold were produced and assayed for binding to the target ligand. Unfortunately, none of the designed proteins displayed evidence of high-affinity binding. Structural characterization of four designed proteins revealed that although the predicted structure of the protein model was highly accurate, this structural accuracy did not translate into accurate prediction of binding affinity. Crystallographic analyses indicate the lack of binding affinity is possibly due to unaccounted for protein dynamics in the "thumb" region of our design scaffold intrinsic to the family 11 β-xylanase fold. Further computational

* Chapter III is excerpted from Morin, A. et al., 2011. Computational design of an endo-1,4-{beta}-xylanase ligand binding site. *Protein engineering, design & selection*

analysis revealed two specific, single amino acid substitutions responsible for an observed change in backbone conformation, and decreased dynamic stability of the catalytic cleft. These findings offer new insight into the dynamic and structural determinants of the β-xylanase proteins.

Introduction

The ability to rationally design proteins through computational methods has long been a goal of biotechnology and pharmaceutical researchers. The development of widely applicable, repeatable and accurate rational protein design methods is expected to enable the development of protein based therapeutics for human medical applications and improved enzymatic processes essential in industry and manufacturing. The market for clinical protein therapeutics, some $94 billion in 2010, is expected to grow to half of total prescription drug sales by 2014 (1), and industrial use of engineered proteins will soon reach over $5 billion per year (119).

Computational protein design has experienced important success in recent years, with significant achievements in the design of novel enzymes (11; 12; 120), biocatalysts (121)(122), antivirals (123)(124), protein-protein interfaces (125)(126)(84), diagnostics (127)(128), and novel protein folds (80). However, a particular aspect of computational protein design that has proved more difficult is the design of protein-ligand interfaces, particularly the design of proteins capable of tightly binding small-molecules and peptides (129)(130).

The goal of the current study was to develop and experimentally validate computational tools and protocols for designing high-affinity protein-ligand interfaces using the ROSETTA protein design program {http://www.ROSETTAcommons.org/}. The protein design functionality of the ROSETTA program has demonstrated prior success at designing enzymes (12)(11)(131), altering the specificity of protein-protein interactions (84)(83)(87), creating novel protein folds never before seen in nature (80), and predicting protein-peptide specificity (91). Here we set out to expand the application of ROSETTA to the design of a *de novo*, high-affinity interface to a small peptide ligand.

The target ligand system we chose for our proof-of-concept study was the D-alanine-D-alanine C-terminal dipeptide of the peptidoglycan precursor from *Staphylococcus aureus.* These terminal D-ala-D-ala peptides are critical to *S. aureus* cell wall biosynthesis and are the primary target for the glycopeptide antibiotic vancomycin, an antibiotic of last resort for treating multiple resistant gram-positive infection (132). Vancomycin acts by binding and sequestering the D-ala terminus of the peptidoglycan precursor (Figure 3.1a) preventing its incorporation into the bacterial cell wall (Figure 3.1c). This compromises the integrity of the bacterial cell wall, rendering it vulnerable to lysis due to normal osmotic pressure changes (133). Some bacteria acquire resistance to vancomycin by replacing this C-terminal dipeptide with a D-alanine-D-lactate moiety (D-ala-D-lac) (134).

We attempted to use ROSETTA to perform the *de novo* re-design of the family 11 endo-1,4-β-xylanase from *Nonomuraea flexuosa* (PDB ID 1m4w) to replicate the

binding and sequestration mode of action of the vancomycin antibiotic. This protein was chosen due to its available 2.1Å-resolution 3D coordinates, thermostability, expression and production characteristics, molecular mass and the geometry and size of its enzymatic cleft (112). We were encouraged that previously successful ROSETTA enzyme design work had been performed using this protein, proving its feasibility as a scaffold for computational design (12)(11).

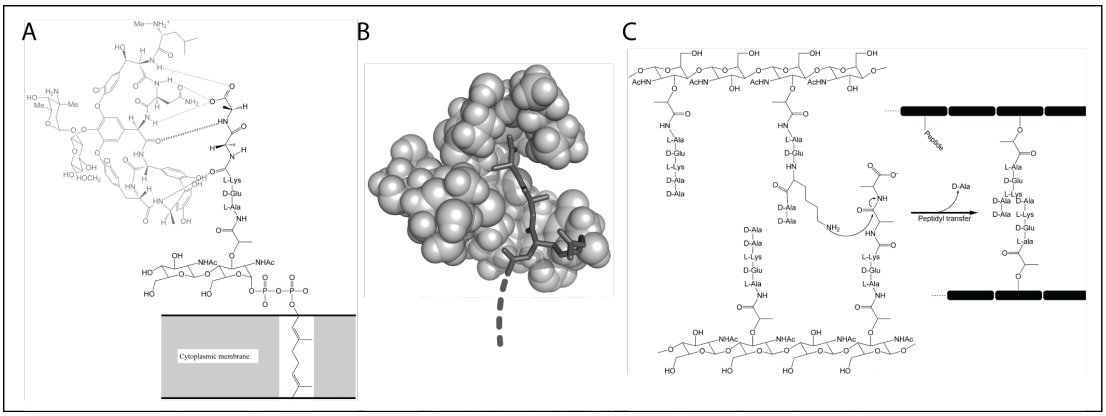


Figure 3.1 The D-ala-D-ala peptidoglycan and vancomycin's mode of action. (A) Vancomycin (light grey) forms 5 critical hydrogen bonds to terminal D-ala-D-ala residues of the S. aureus peptidoglycan precursor anchored in the cytoplasmic membrane. (B) Space filling model showing how vancomycin binds and sequesters the terminal D-ala peptides, thus preventing the peptidyl transfer cross-linking (C) of glycopeptide chains essential for cell wall biosynthesis.

In the course of ROSETTA computations, the scaffold protein's enzymatic cleft is mutated *in silico* to form an interface capable of binding to the target D-ala-D-ala or D-ala-D-lac dipeptides (Figure 2.6). Following computations, the *in silico* designed protein sequences were produced in the laboratory and assayed for binding to the target dipeptide using multiple, complementary methods.

Unfortunately, none of the designed proteins demonstrated high-affinity binding to their target ligands (K_d<100μM). Subsequent structure determination of four of the ROSETTA designed proteins revealed conformational changes in the protein backbone and altered protein dynamics as significant contributing factors to the lack of observed ligand binding affinity. The results presented here can additionally be utilized as a benchmark case for the further development of computational design algorithms.

Materials and methods

Selection of thermostable scaffold protein

To identify protein scaffolds suitable for ROSETTA design, a search of the PDB was conducted for proteins with high-resolution crystallographic structures (<2.5Å), possessing no structurally important metal atoms, having a molecular weight below 50 kDa and a binding surface or pocket of the appropriate geometry to accommodate a dipeptide ligand. Preference was give to thermostable proteins under the assumption that their robustness would allow more extensive design mutations without destabilizing the overall protein fold.

The PDB file of the selected scaffold was prepared for ROSETTA design by the removal of all redundant protein chains and non-proteinacious molecules, including crystallographic water and reagent molecules. All ligand atoms were removed, and any “anisou” or alternate atom positions or sidechain rotamers were discarded,

retaining only the 3D coordinates and identities of protein main-chain and sidechain atoms.

Ligand model and generation of ligand ensemble

The D-ala-D-ala dipeptide ligand moiety consists of 25 atoms – the 12 heavy atoms and 12 hydrogen atoms of the D-ala-D-ala terminus of the target glycopeptide, plus the carbonyl carbon of the preceding lysine residue comprising the peptido linkage. A D-ala-D-lac ligand representing a resistant form of the *S. aureus* glycopeptide was generated by substituting the C-terminal amide nitrogen of the D-ala-D-ala ensemble with oxygen (Figure 2.6). To account for potential conformational flexibility of the dipeptide, an ensemble of conformers was created using the MOE (Molecular Operating Environment) software. The ensemble was populated by systematically rotating the backbone phi/psi angles of the target peptide in 10° increments, then removing all conformers not possessing "allowed" beta-sheet Ramachandran angles for D-amino acids. Each conformer was then output as an individual .pdb file. Design calculations were performed with a representative conformer ensemble of 225 D-ala-D-ala and 225 D-ala-D-lac dipeptide structures.

ROSETTA computations

De novo computational design and ligand docking of the chosen scaffold with the target ligand ensemble was performed using the ROSETTALIGAND module of ROSETTA

version 2.3 (92). RosettaLigand utilizes a monte carlo/metropolis (MCM) simulated annealing search algorithm to dock the ligand molecule with three translational and two rotational degrees of freedom. Simultaneously, RosettaLigand designs the protein scaffold by varying the identities of the amino acids comprising the binding interface (Figure 2.1a). The knowledge-based energy function combines Van der Waals (VDW) attractive and repulsive interactions, hydrogen bonding energy, a desolvation penalty and pair-wise electrostatics (135), as well as sidechain rotamer probabilities derived from the PDB (136).

All peptide conformations were placed manually into the ligand binding site. In an iterative protocol, RosettaLigand simultaneously optimizes ligand position and protein sequence. During computations, ligand position and orientation are randomly perturbed before all interface residues are redesigned to optimize protein ligand interactions. This "dock-design" protocol is repeated five times in an iterative fashion. Following each round of dock-design, 10,000 of the 100,000 models generated were selected based on predicted ligand binding energy normalized by the number of mutations from wild-type, degree of ligand burial, ligand hydrogen-bond donor/acceptor saturation, and egress of the N-terminal extension of the glycopeptide ligand. These best scoring 10,000 models were then used as starting models in the following round of dock-design computations (Figure 2.1b). At each successive round, perturbation of the initial ligand position and orientation was narrowed, leading to increased conformational search density from round-to-round. While the first round allowed for complete ligand reorientation and movement of up to 5Å, the final round limited movement to 5° and 0.5Å. The protocol uses a softened

repulsive VDW scoring potential to smooth the energy landscape. After five dock-design iterations, predicted ligand binding energies plateaued and the amino acid sequences of designed proteins converged. In a final step, 10,000 models were energy minimized using hard-repulsive VDW scoring potentials to discriminate the best protein sequences based on predicted ligand binding energy. This process allowed for minimal ligand movement and optimization of sidechain conformations.

Selection of designed mutant proteins for expression

The resulting protein designs were clustered according to binding pose and sequence and the top scoring models of each sequence group were ranked according to predicted binding energy. Interestingly, the best scoring models shared the same principal binding mode and a subset of mutations. Models were then analyzed at atomic detail on a residue-by-residue basis, examining for hydrogen bonding geometries, hydrophobic packing, burial of polar groups, and binding pocket access/occlusion. Additional filtering of the models for each of the 1m4w scaffold designs was performed to accommodate egress of the N-terminal extension of the glycopeptide target. The best nine models for each target ligand were chosen for experimental evaluation of predicted ligand binding (Table 3.1). Later, three additional point mutants of the design 1m4w_6 were created (see below).

Maximally efficient gene synthesis strategy

A hierarchical strategy for gene construction of the nine mutant proteins was devised to minimize mutational primers and reaction steps (Figure 3.2a). Genes were assembled using recursive PCR (137) from *E. coli* codon-optimized oligonucleotides designed using the Gene2Oligo web server (http://berry.engin.umich.edu/gene2oligo/) (138). Once assembled, the genes were cloned into a T7-driven pET29b expression vector. Point mutations were introduced using Quickchange™ (Stratagene). All constructs were confirmed by DNA sequencing.

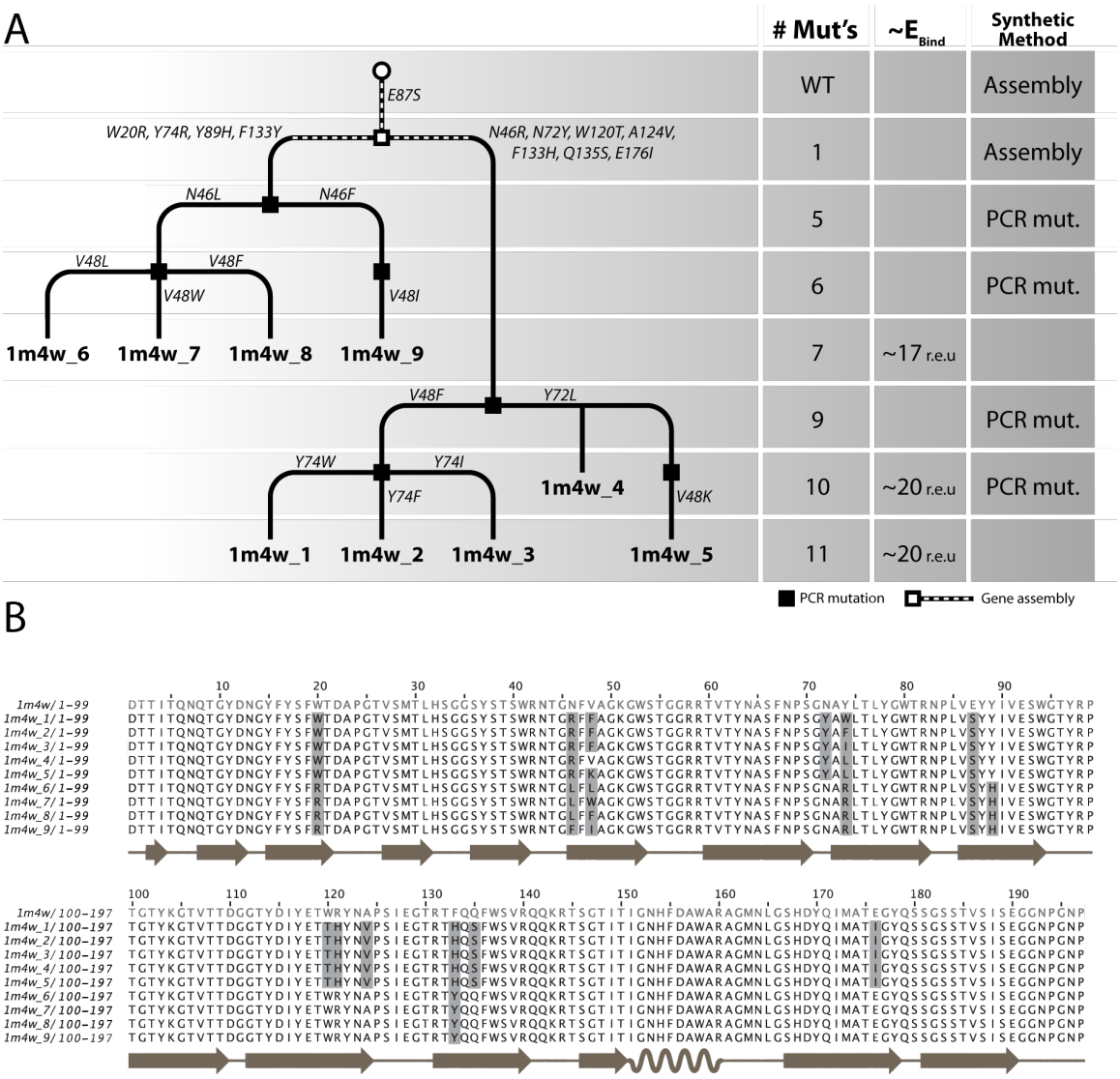


Figure 3.2 Experimental protein synthetic strategy and sequence alignment of 1m4w designs. (A) Chart of sequence mutations from wild-type, synthetic methods and predicted binding properties. Diagram shows at each step in the gene synthetic process by which method mutations were introduced. White-filled circle and box indicate synthesis by gene assembly, beginning with the wild-type sequence at top. Solid boxes indicate mutagenesis by PCR. Italicized text indicates which residues were mutated at each stage. Bold text at line termini denote completed ROSETTA designed proteins. Table at right of diagram shows number of mutations at each synthetic step, approximate predicted energy of binding (in r.e.u.) and synthetic method used. (B) Sequence alignment of wild-type 1m4w (top, grey type) and the nine ROSETTA designed proteins designated 1m4w_1 through 1m4w_9. Mutations from wild-type are indicated by grey boxes and secondary structure is below.

Expression and purification of designed proteins

Proteins were expressed in e. coli BL21(DE3) pLysS cells (Stratagene). Cells were grown in LB media supplemented with kanamycin at 37°C until an OD(600nm) of 0.4-0.6 was reached. The cells were then transferred to 16°C. After 30 minutes, the samples were induced with IPTG to a final concentration of 150μM and grown approximately 14 hours . Cells were then harvested by centrifugation.

Cells were lysed using French-press in 25mM HEPES, 100mM NaCl, 5mM imidazole, 5% glycerol v/v, pH 7.6-7.8 buffer containing protease inhibitor cocktail (Roche). A single step IMAC purification protocol using TALON™ cobalt-affinity resin (Clontech) was sufficient to obtain greater than 95% purity as assessed by SDS-PAGE. Following purification, proteins were immediately dialyzed into a buffer containing 25mM HEPES, 100mM NaCl and 5% glycerol v/v at pH 7.6-7.8.

Molecular weights were confirmed by MALDI-MS on a PerSeptive Biosystems Voyager-DE STR instrument. Protein aggregation state and solution properties were assessed by dynamic light scattering using a DynaPro ProteinSolutions molecular sizing instrument (Wyatt Technology Corporation). Proper protein folding was confirmed by circular dichroism (CD) using a Jasco J-810 Spectropolarimeter and 1D-NMR on a Bruker Avance 600-MHz spectrometer.

^{15}N-labeled proteins for NMR were obtained by expression in M9 minimal media with $^{15}NH_4Cl$ as the sole nitrogen source. For X-ray diffraction and NMR structural characterization, proteins were purified by IMAC as described above followed by size-exclusion chromatography using a HiLoad 16/60 Superdex 75 gel filtration

column (GE Healthcare). This additional purification step gave >99% purity as assessed by SDS-PAGE.

Peptides for protein-ligand binding studies

Peptides were purchased from Genscript. N-terminally acylated L-lys-D-ala-D-ala tripeptide or L-lys-D-ala-D-lac were used in ITC and NMR titrations. Three dansylated peptiedes were used for fluorescence studies: (Dansyl)-L-lys-D-ala-D-ala peptide with the dansyl label covalently linked to the N-terminal nitrogen; L-lys-(Dansyl)-D-ala-D-ala peptide with dansyl label attached to the lysine ε-amino group, and (Dansyl)-AEEAE-L-lys-D-ala-D-ala with a pentapeptide linker that separates the target peptide from the dansyl group.

All assays were carried out in 100mM NaCl, 25mM HEPES, 5% glycerol v/v aqueous buffer at pH 7.7 unless otherwise noted. Protein concentrations were measured at 280nm using a Shimadzu UV-mini 1240 spectrophotometer and calculated extinction coefficients (ExPASy ProtParam server {http://www.expasy.ch/tools/protparam.html}) See Appendix D.

Fluorescence anisotropy

Fluorescence anisotropy (FA) titrations were carried out at 25°C using a T-format PTI Quantamaster 2000-7SE spectrofluorometer equipped with excitation and emission polarizers. The fluorescence emission intensities parallel and

perpendicular to the vertically polarized excitation light were analyzed to determine the steady state anisotropy values for each point in the titration. During the titrations, the concentration of dansyl labeled peptide ligand was held constant while increasing concentrations of protein were added. Dansylated samples were excited at 340nm and the fluorescence emission signal was monitored at 520nm with both excitation and emission slit widths set to 1mm.

NMR chemical-shift perturbation assay

NMR experiments were performed using a Bruker Avance 600-MHz spectrometer equipped with a cryoprobe. ^{1}H-^{15}N HSQC spectra were acquired with ^{15}N-labeled proteins at 200-600 μM in 25 mM HEPES, pH 7.6-7.8, 100 mM NaCl and 2.5% glycerol v/v H_2O/ 10% D_2O. A series of ^{15}N-^{1}H HSQC spectra were acquired of protein titrated with 0, 1, 5, and 10 molar equivalents of peptide at 298K. Data were processed using Topspin 2.0b (Bruker) and analyzed with Sparky (http://www.cgl.ucsf.edu/home/sparky/).

Isothermal titration calorimetry

Isothermal titration calorimetry (ITC) experiments were performed at 30°C using a MicroCal VP-ITC instrument. Unlabeled peptide was titrated into the cell containing 0.6-1.1mM protein in 100mM NaCl, 25mM HEPES, 5% glycerol v/v, pH 7.6-7.8

buffer. Ligand concentrations were 15-20 times the molar concentration of the protein.

Crystallization of proteins derived from model 1m4w_6

Crystallization screens of designed 1m4w_6 as well as three derivative point mutants (Table 3.1) were built from Hampton research HR2-130 Crystal Screen HT reagents using a Thermo Fisher Scientific MaxCell™ crystallization workstation incorporating a MicroLab Starlet™ (Hamilton Corporation, Reno, NV) liquid handling robot and a Mosquito™ nanoliter drop setting robot (TTP LabTech, Oxford, UK). All screening was performed using 96-well MRC plates (Hampton Research) and experiments were visualized and recorded using a Thermo Fisher Scientific Rhombix™ Tablestore automated imaging system. Protein was concentrated to 10mg/mL in 100mM NaCl, 25mM HEPES, 5% glycerol v/v, pH 7.8 buffer. Initial hits from the robotic screen were optimized in 24-well sitting-drop plates using individual Hampton Research Optimize reagents.

Diffraction data collection and processing.

Complete data sets were acquired in-house using a Bruker Microstar rotating-anode X-ray generator and a Bruker Proteum PT135 CCD area detector. Crystals were maintained at 100K using a Bruker Kryo-Flex cryostat. Data collection sweeps were optimized using Cosmo (Bruker AXS, 2008) software and data integrated and scaled

using SADABS (Bruker AXS, 2008) and XPREP (Bruker AXS, 2008) in the PROTEUM2 (Bruker AXS, 2008) package. The cryoprotectant used was the crystallization buffer supplemented with 30% ethylene glycol v/v.

Additional X-ray diffraction data were collected at Southeast Regional Collaborative Access Team (SER-CAT), beamline 22-ID, Advanced Photon Source, Argonne National Laboratory using a MAR165 CCD area detector. A total of 360 frames with a 0.5° oscillation angle were collected at 100 K using a wavelength of 1.00Å and a crystal-to-detector distance of 150 mm.

Data processing and structure refinement

Diffraction data were phased by molecular replacement with the program MOLREP (139), using the 1m4w coordinates obtained from the PDB or ROSETTA designed models. Molecular replacement phases were then used to initiate automated model building with the program, Arp/wArp (140). Model refinement was performed using REFMAC5 (141) with iterated manual fitting using COOT (142). All data analysis and refinement were performed using the CCP4 package (Collaborative Computational Project, Number 4. 1994) and ccp4i gui (143).

Results

Scaffold selection

We began by attempting of identify a suitable protein scaffold for our *de novo* protein-peptide interface design effort. 1m4w is a thermophilic endo-1,4-β-xylanase (EC 3.2.1.8) from *Nonomuraea flexuosa* with a crystal structure determined at 2.10Å resolution (112). Its β-jelly-roll topology of two twisted beta-sheets forms a large cleft where enzymatic endoxylanase activity occurs, typical to family 11 xylanases. The protein does not naturally interact with peptide ligands, instead binding large polysaccharides on its outer surface, while residues inside the cleft catalyze the glycosidic cleavage of xylanose subunits. The overall molecular weight of approximately 22 kDa, the size and geometry of its enzymatic cleft and the lack of native ligand binding function were all well suited to a *de novo* redesign strategy. Additionally, the thermostable nature of 1m4w was expected to allow a more extensive redesign of residues in the binding cleft without significant destabilization of the protein backbone.

ROSETTALIGAND computations

The ROSETTALIGAND module of the ROSETTA suite of programs was used to accommodate the non-standard nature of the D-ala and D-lac ligands during design of the protein-peptide interface. The goal of ROSETTALIGAND dock-design computation is to identify the smallest set of mutations to the native scaffold protein

sequence, which also provides the highest affinity binding to the target dipeptide ligands. The best scoring nine sequences possessing binding energies of at least -1.5 ROSETTA energy units (r.e.u.) per amino acid mutation from wild-type were selected for laboratory expression and assay (Table 3.1; Figure 3.2a). Each of the nine proteins is 197 amino acids in length and displays a unique combination of between seven and eleven mutations. All of the mutations are located in the catalytic cleft on the inside of the concave jelly-roll protein fold, in one of three regions that directly interact with the ligand. These regions are referred to as the "thumb", "palm" or "finger" (see Figure 3.3a)(112). The nine selected protein designs were labeled sequentially as 1m4w_1, through 1m4w_9 (Table 3.1).

	WT	1	2	3	4	5	6	7	8	9	v48	w20	w20v48		
AA position														SS-type	Region
20	W	W	W	W	W	W	R	R	R	R	R	W	W		
46	N	R	R	R	R	R	L	L	L	F	L	L	L	Strand	
48	V	F	F	F	V	K	L	W	F	I	V	L	V		Finger
72	N	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	Loop	
74	Y	W	F	I	L	L	R	R	R	R	R	R	R		
87	E	S	S	S	S	S	S	S	S	S	S	S	S		Palm
89	Y	Y	Y	Y	Y	Y	H	H	H	H	H	H	H		
120	W	T	T	T	T	T	W	W	W	W	W	W	W		
121	R	H	H	H	H	H	R	R	R	R	R	R	R	Strand	
124	A	V	V	V	V	V	A	A	A	A	A	A	A		Thumb
133	F	H	H	H	H	H	Y	Y	Y	Y	Y	Y	Y		
135	Q	S	S	S	S	S	Q	Q	Q	Q	Q	Q	Q		
176	E	I	I	I	I	I	E	E	E	E	E	E	E		Finger
# Mutations	0	11	11	11	10	11	7	7	7	7	6	6	5		
Ligand	--	lac	lac	lac	lac	lac	ala	ala	ala	ala	ala	ala	ala		
E_{bind} (r.e.u)	--	19.9	19.9	19.9	20.2	19.8	17.6	17.4	17.2	17.1	12.9	13.3	15.4		
Affinity* (kcal/mol)	--	-7.4	-7.4	-7.4	-7.6	-7.4	-6.2	-6.0	-5.9	-5.9	-3.5	-3.8	-4.9		
PDB ID	1M4W						3MF6				3MF9	3MFC	3MFA		

Table 3.1 Sequence characteristics of the 1m4w protein designs. Amino acid identities at given sequence positions for wild-type 1m4w plus twelve designed mutants. Designation of each 1m4w_"X" protein at top. Grey type denotes mutated amino acids. Secondary structure and protein region of mutations shown at far right. Number of mutations from wild-type, ligand target (D-ala-D-ala or D-ala-D-lac), computed ROSETTALIGAND energy of binding in ROSETTA energy units (r.e.u), ROSETTALIGAND predicted affinity (in kcal/mol from the method of Meiler & Baker 2006) and PDB IDs for the deposited structures are at bottom.

Figure 3.3 Backbone opening of the binding pocket and prediction of interface rotamer conformations between 1m4w_6 predicted model (light grey) and X-ray structure (dark grey). (A) Cartoon representation of the model and X-ray structure showing the 1.25Å shift in the backbone configuration of the "thumb" region. (B) Detailed comparison of the residues comprising the ligand interface. Most of the residue sidechains are super-imposable, while several are out of position due to the altered backbone conformation. Only two sidechain rotamers assume substantially different conformations from prediction. (C) Residues identified as directly responsible for binding pocket opening. W20-P125 (shown with VDW spheres) form a hydrophobic interaction between "thumb" and "fingers" at the top of the binding pocket, while V48 lies lower in the "palm" of the protein.

During the design process, many of the residues in the catalytic site of the 1m4w enzyme were altered in favor of the new peptide binding function, thus eliminating the proteins native catalytic functionality. The wide and deep catalytic cleft of the protein was transformed by the design process into a tightly fitting binding pocket, closely contacting the target D-ala-D-ala or D-ala-D-lac dipeptide ligands on all sides except the N-termini, thus allowing for egress of the un-modeled remainder of the glycopeptide (Figure 3.4a and Figure 2.6). Predicted binding energies for the initial

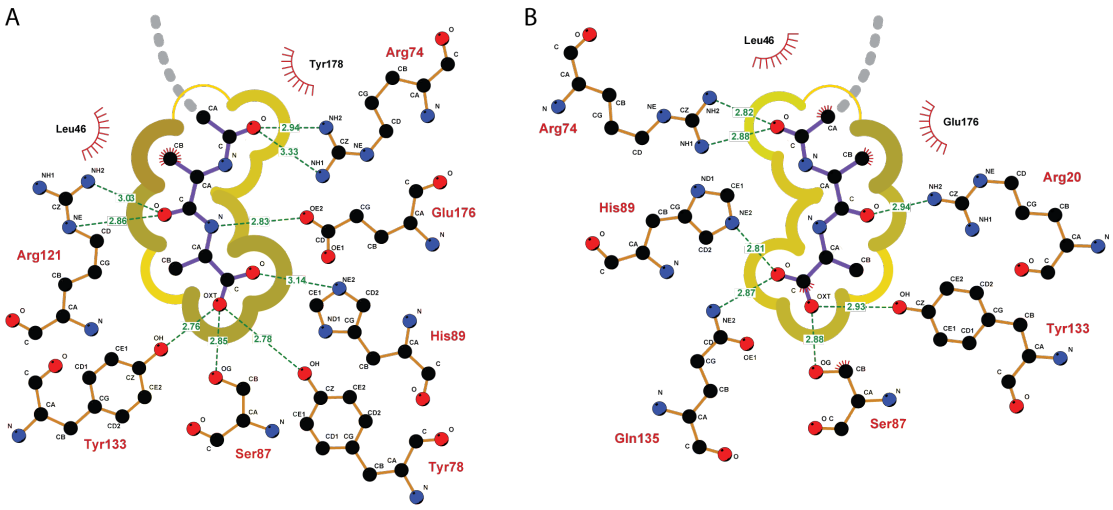


Figure 3.4 Detailed schematic of ligand interface. (A) ROSETTALIGAND predicted interface of 1m4w_6 showing individual residues and H-bonds involved in binding, and the degree of solvent accessibility to the ligand. Darker yellow, thicker lines indicate low exposed surface area; lighter, thinner lines indicate more solvent exposure. Grey dashed line denotes the path of the unmodeled portion of glycopeptide ligand. (B) Detail of the X-ray determined 1m4w_6 apo interface with ligand re-docked. Note the decrease in number of H-bonds and increase in degree of solvent exposure. Solvent accessibility was computed with NACCESS (Hbbard & Thornton, 1992) using a probe radius of 1.4Å and visualized with LigPlot (144).

nine ROSETTALIGAND protein designs ranged from -17 to -20 r.e.u. (Table 3.1). Previous studies by Meiler & Baker found that ROSETTA energy units correspond to experimentally determined binding energies with a correlation of 0.63 (92). Using the Meiler & Baker method, the ROSETTA energies for the initial nine chosen designs correspond to a predicted free energy of binding of -5.82 to -7.50 ± 1.9 kcal/mol and a K_d of 54 ± 34μM to 3 ± 2μM, respectively. Additionally, good hydrophobic packing of both ligand methyl groups and strong binding of the carboxyl terminus were common features in each of the nine protein designs.

Expression characteristics and solution properties of designed proteins

Expression of the ROSETTALIGAND designed proteins proceeded as outline in the Methods section. All of the 1m4w designed proteins expressed well, yielding between 7 and 12 mg/L induction. All 1m4w proteins were found to express greater than 50% soluble, with most greater than 75% soluble. Dynamic light scattering (DLS) and size-exclusion chromatography of each of the expressed proteins indicated that the 1m4w designs existed in solution as homogeneous, monomeric species.

Far-UV CD spectra of the 1m4w designed proteins indicated secondary structure composition similar or identical to wild type (Figure 3.5a). NMR results confirmed that all of the 1m4w proteins were well folded and stable (See Appendix F). Additionally, the 1m4w designed proteins exhibited a high degree of stability and resistance to proteolysis. Samples left at room temperature for several weeks following purification showed no signs of degradation.

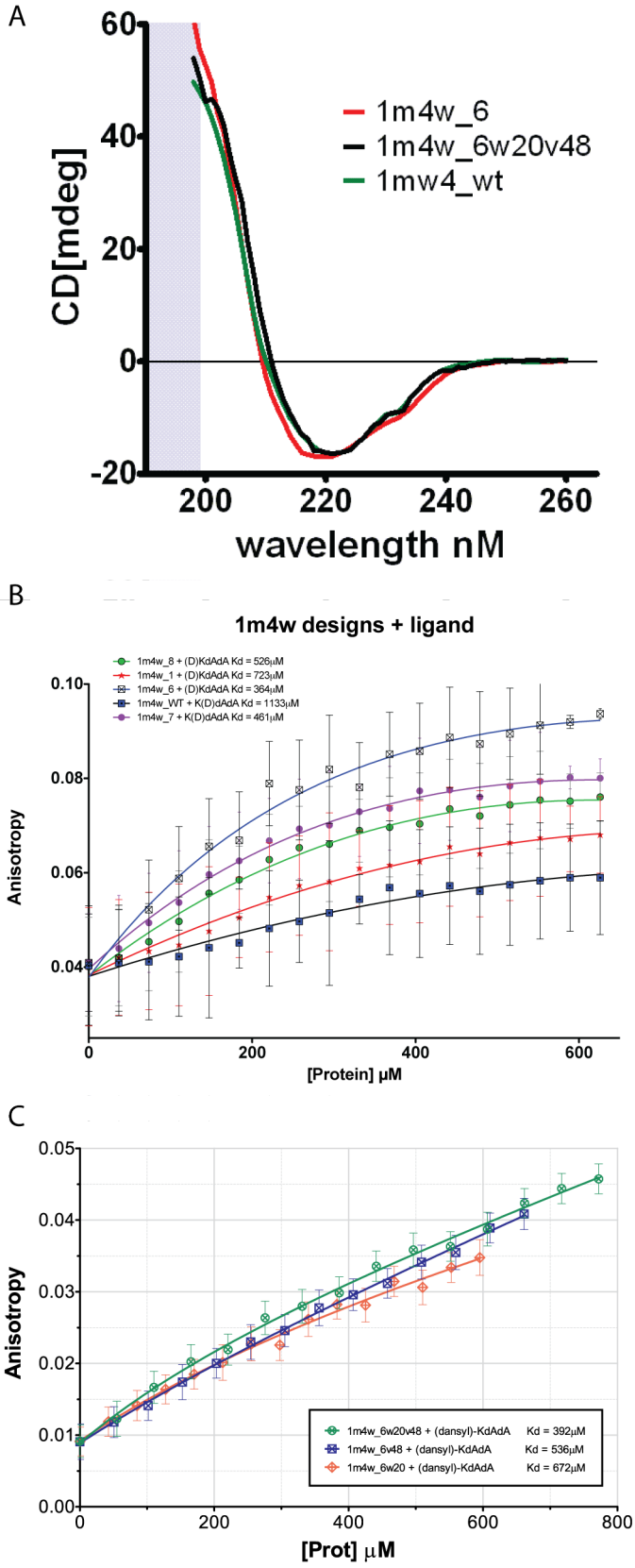


Figure 3.5 CD and binding assay plots for representative designed 1m4w proteins. (A) CD spectra for the wild-type 1m4w, designed 1m4w_6 and re-designed 1m4w_6w20v48 proteins demonstrating similar tertiary structure composition. (B) FA binding assay plots for several of the designed mutants titrated with danslyated KdAdA peptide. (C) FA plots for re-designed proteins titrated with danslyated EKdAdA peptide. Non-linear regression curves in B & C were calculated by GrpahPad Prism software using a one site binding (hyperbolic) curve fitting equation. Concentration of the dansylated ligand was held constant at 10uM while protein concentration was diluted from the maximum values as seen in the plot.

Assay of predicted binding affinity of designed proteins

Following computational design and expression of the chosen interface designs, biophysical binding assays were performed to validate the predicted binding affinities. Unfortunately, none of the designed proteins tested in this study yielded evidence of specific, high affinity binding to their target peptide. We thus conclude that the ROSETTALIGAND interface designs were not successful.

Using fluorescence anisotropy, several of the 1m4w designs indicated low to moderate affinity binding, with K_d values between 367μM to 449μM (Figure 3.5b). Non-specific, background binding affinities for the 1m4w designs during FA measurements were observed to be at or above 850μM. These negative results for high-affinity binding were later confirmed by ITC and NMR spectroscopy.

Structure determination of 1m4w_6

To determine a cause for the lack of observed binding among the designed proteins, a high-resolution X-ray diffraction structure of 1m4w6 was determined. After numerous rounds of refinement an optimal crystallization buffer contained 0.1 M NaCl, 1.125 M ammonium sulfate, 0.1 M Bis-Tris pH 5.5, 3% Jeffamine M600 pH 7.0 and grown at 20°C produced diffracting, single, rod shaped crystals of up to 150μM x 450μM (Figure 3.6). The final conditions differed significantly from that of the wild type 1m4w structure (112). Data sets were collected for 1m4w_6 crystals in the apo

form to a resolution of 1.28Å. Refinement statistics for the structure of the 1m4w_6 designed mutant are listed in Table A.1.

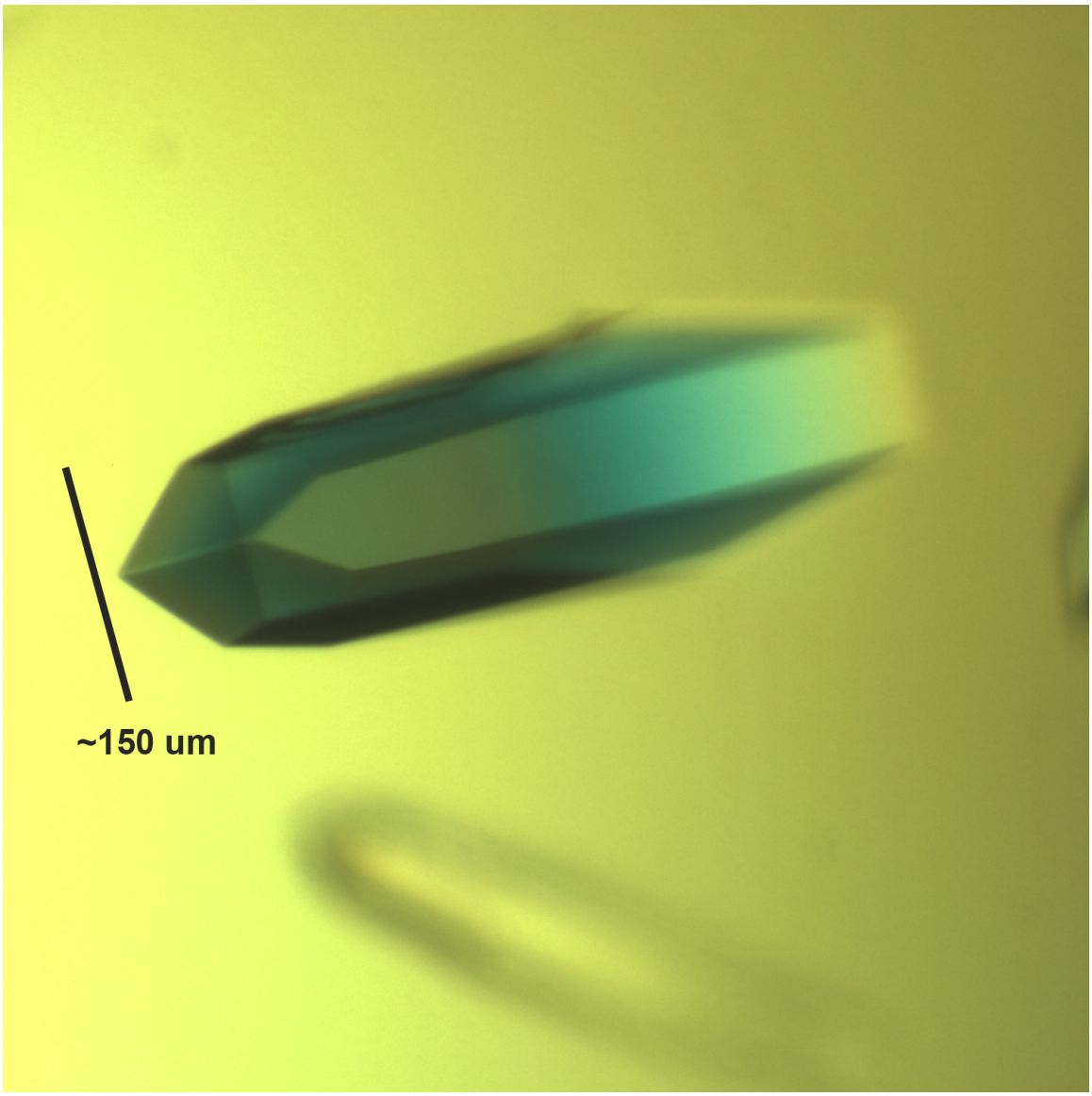


Figure 3.6 Morphology of the 1m4w_6 crystals. This crystal was grown in sitting drop, 24-well plate in buffer containing 0.1 M NaCl, 1.125 M ammonium sulfate, 0.1 M Bis-Tris pH 5.5, 3% Jeffamine M600 pH 7.0 and grown at 20°C. Dimensions of the crystal were ~150µm by 450µm. All 1m4w_6 derivatives had similar crystal morphologies.

Structural analysis of 1m4w_6

Using the newly obtained high-resolution 3D structure of the designed 1m4w_6 protein, a comparative structural analysis was performed. The most identifiable difference between the 1m4w_6 experimental structure (PDB IDs 3mf6) and ROSELLALIGAND predicted 1m4w_6 model is an expansion of the binding pocket. This expansion occurs through a 1.25Å outward movement of the protein "thumb" region when compared to the original 1m4w structure (Figure 3.3a). Moreover, the solvent accessible (SA) surface area of the pocket increases 2.5 times, while normalized SA volume expands by a factor of 2.3 (Figure 3.7c). Although flexibility of residue sidechains within the pocket partially compensate for this "opening" relative to prediction, a significant enlargement of the binding pocket is observed. The all atom RMSD for the whole protein is 0.61A, but rises to 0.96 Å within the binding pocket (Figure 3.3b). Notably, interface residues that contribute most to RMSD are also those possessing the highest crystallographic B-factors. The expansion of the binding pocket disrupts interactions observed in the computational model. When the ligand is re-docked into the crystallographic structure, only eight of eleven predicted hydrogen bond interactions are able to assume correct bonding geometry, while the ratio of ligand surface area in VDW contact with protein decreased from 0.79 to 0.63 (Figure 3.4b). Thus, we hypothesized that the lack of observed ligand binding affinity was due to the expansion of the binding pocket and resulting disruption of predicted binding contacts.

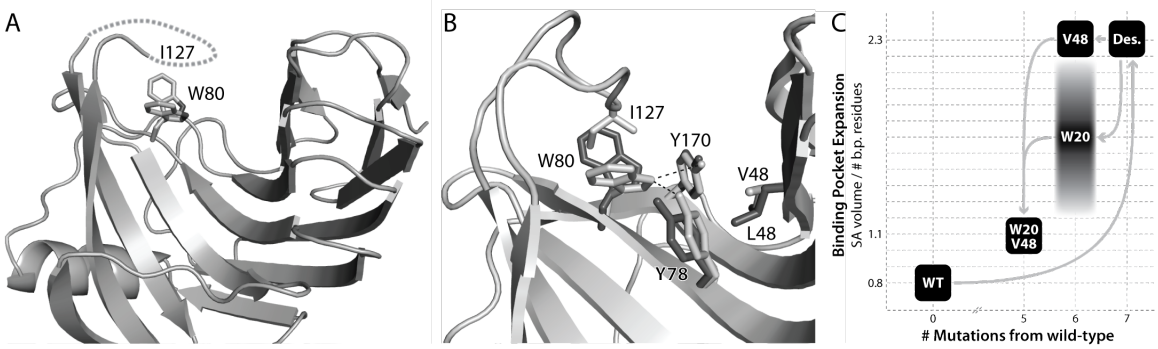


Figure 3.7 Structural determinants of β–xylanase "thumb" destabilization. (A) Loss of resolvable electron density in the "thumb" region is caused by an alternate confirmation of W80 in the protein "palm". (B) A "domino" effect of altered sidechain packing results from the substitution of wild-type (light grey) V to designed (dark grey) L at position 48. This added steric bulk pushes Y78 out of H-bonding position, which then allows W80 to adopt an alternative conformation that clashes with I127 and disrupts the hydrophobic packing of the two, thus destabilizing the "thumb" loop. (C) Chart showing the relative degree of binding pocket expansion for each sequence substitution. Wild-type (WT) and W20V48 proteins display a closed conformation, while the designed (Des) and V48 substitutions result in an "open" conformation. The W20 mutant, due to "thumb" destabilization, dynamically inhabits a range of conformations between "open" and "closed".

ROSETTA analysis of 1m4w_6

To investigate the hypothesis that binding pocket enlargement is responsible for the lack of detected binding affinity, a detailed analysis of residue-level energy contributions to binding affinity was performed comparing the 1m4w_6 experimental and predicted structures. In comparing the two structures, ROSETTALIGAND calculations showed a modest but clear loss of binding affinity as pocket backbone opening increased, as indicated by several of the contributing energy terms (Table 3.2).

1m4w_6 Rotetta model

Residue	atr	rep	sol	hb	cou	
20					-0.12	
22					-0.17	
45					-0.03	
46	-0.50	0.00	0.22		0.22	
47	-0.03		0.04		-0.07	
48	-0.28		0.32		-0.01	
72	-0.13		0.39		-0.04	
74	-1.36		2.42	-1.30	-0.60	
78	-1.29		1.74	-0.86	-1.02	
80	-0.56		0.61		0.02	
87	-0.73		1.20	-1.47	-0.75	
88					0.08	
89	-0.78		0.90	-0.26	-0.80	
94	-0.41				0.00	
97					0.00	
99					0.00	
121	-2.39		4.00	-1.91	0.81	
122	-0.04				0.00	
124	-0.08		0.01		0.04	
125	-0.10		0.15		0.03	
126	-0.23				-0.08	
131					-0.04	
133	-0.91		1.57	-0.93	-0.53	
135	-0.83		1.41	-0.06	-0.62	
137	0.00		0.01		0.03	
176	-1.92		3.20	-1.18	-1.55	
177	-0.09		0.09		-0.20	
178	-2.54	0.02	1.75		0.04	
179					0.03	
	-15.22	0.02	20.04	-7.99	-5.35	-8.49
	-0.72	0.01	1.11	-1.00	-0.18	-0.16
	0.80	0.40	0.60	2.00	0.25	Rosetta Weights
	-12.17	0.01	12.02	-15.97	-1.34	-17.45

1m4w_6 x-ray structure re-docked with ligand

Residue	atr	rep	sol	hb	cou	
20					-0.01	
22					-0.16	
46	-0.64	0.99	0.38		0.18	
47	-0.02		0.04		-0.10	
48	-0.16		0.20		-0.01	
72	-0.03		0.16		-0.01	
74	-1.53		2.91	-1.62	-0.50	
78	-1.04		1.26	-0.23	-0.99	
80	-0.27		0.37		0.00	
87	-0.28		0.52	-0.08	-0.68	
88					0.06	
89	-0.91	0.42	1.45	-0.44	-0.50	
94					0.00	
97					0.00	
99					0.00	
121	-2.34	1.10	4.18	-1.82	1.35	
124					0.01	
125	-0.03		0.04		0.03	
126					-0.05	
133	-0.18		0.40		0.14	
135	-0.52		0.97		-0.48	
137	-0.03		0.07		0.02	
176	-1.82		2.93	-1.11	-1.26	
177	-0.06		0.08		-0.16	
178	-0.47		0.30		0.00	
	-10.34	2.51	16.25	-5.29	-3.11	0.03
	-0.61	0.84	0.96	-0.88	-0.12	0.04
	0.80	0.40	0.60	2.00	0.25	Rosetta Weights
	-8.27	1.01	9.75	-10.59	-0.78	-8.88

Table 3.2 Decompositions of the Rosetta binding energy values (in r.e.u) for each of three 1m4w derived proteins. Binding energies for each model are decomposed into five energy terms: attractive, repulsive, solvation, hydrogen-bonding and coulombic. The protein residues to which each term applies is listed in the leftmost column of each table. The sum of each energy term column (shown in grey at bottom) is multiplied by the 'Rosetta Weights' term, then each terms sum is added to yield the total Rosetta 'lig_sum' energy of ligand binding. The left table shows the individual energies terms of residues which participate in binding of the target ligand for the Rosetta predicted 1m4w_6 model.

For example, the total number of residues involved in the hydrogen bonding network between ligand and protein decreased from 8 to 6, while the number of total hydrogen bonds dropped from 11 to 8. Correspondingly, the total hydrogen bond energy worsened from –8.1 to –5.3 r.e.u. while Van der Waals packing was significantly reduced from -14.5 to -10.3 r.e.u. Similarly, solvation and electrostatic interaction energies worsened as pocket expansion increased. A weighted composite ROSETTA binding energy score for the protein-ligand system decreased from -17.2 to -12.9 r.e.u. From this analysis, we concluded that ROSETTALIGAND can

discriminate between the binding energies of a wild-type backbone configuration and that of an enlarged binding pocket, and that this energy differential could potentially explain the lack of experimentally observed ligand binding.

Additional analysis of pair-wise ROSETTA energies revealed a potentially significant contributor to the backbone opening of the "thumb" region: a Trp 20 to Arg mutation that disrupts an interaction with Pro 125 in wild-type 1m4w. This hydrophobic interaction in the wild-type protein appears to stabilize the "thumb" loop in a "closed" configuration and help keep the binding pocket laterally compact (Figure 3.3c). Additionally, the mutation of Val to the sterically bulkier Leu in position 48 of 1m4w_6 further acts as a "wedge" to "prop-open" the binding pocket in the "palm" region at a position of mechanical advantage (Figure 3.3c), causing added strain within the interface. Evolutionary evidence for the crucial function of these residues can be seen from a sequence alignment of 1m4w with its nearest 250 homologues. In all 250, the Trp 20, Pro125 and Val 48 residues are either strictly or highly conserved.

Structure guided redesign of 1m4w_6

Using the information gleaned from ROSETTALIGAND computational analysis, a structure guided redesign of the 1m4w_6 protein was performed to test the hypotheses that a) the observed lack of binding affinity was due primarily to the unintended expansion of the binding pocket and resulting disruption of the predicted binding interactions, and b) that either or both of two identified

mutations (Arg20 and Leu48) from wild-type were largely responsible for the opening of the "thumb" region and expansion of the binding pocket.

Three separate mutants were made starting from the 1m4w_6 sequence, by reverting Arg 20, Leu 48 and a double reversion of both residues to the wild-type amino acid identities (Table 3.1). These newly designed proteins were used to identify the individual and cumulative contributions by each mutation to the backbone conformational change seen in the 1m4w_6 design. Reverting these mutations, it was hoped, would restore the binding pocket to the predicted (wild-type) geometry thus conferring the originally predicted ligand binding affinity.

Following sight directed mutagenesis and expression of the revertant mutants (see Methods) ligand binding assays for each of the three 1m4w_6 derived proteins were performed using FA and ITC. None of the redesigned 1m4w_6 derived mutants displayed observable binding affinities above those obtained from the original 1m4w_6 design. Using FA, the 1m4w_6w20, 1m4w_v48 and the 1m4w_6w20v48 displayed 672μM, 536μM, and 392μM, respectively (Figure 3.5c).

To understand the lack of binding affinity among the three 1m4w_6 derived revertant mutants, structure determination through X-ray crystallography was again performed. Using close grids screens around successful 1m4w_6 crystallization conditions, high quality, diffracting crystals were obtained for the 1m4w_6v48, 1m4w_6w20 and 1m4w_6w20v48 constructs (PDB IDs 3mf9, 3mfc and 3mfa, respectively). Multiple single crystals formed in several buffers centered around wells containing 0.1 M NaCl, 1.25 M ammonium sulfate, 0.1 M Bis-Tris pH

5.5, 3.5% Jeffamine M600 w/v pH 7.0 at 20°C. Complete data sets down to 1.6-1.7Å were obtained for the three protein constructs using the in-house defractometer (see Appendix). The data sets for all three proteins were phased by molecular replacement using MOLREP and models built using the Apr/warp software suite (see Methods, Chapter II). Attempts to obtain liganded co-crystals were unsuccessful. All protein structures obtained were in the apo configuration.

Structural analysis of 1m4w_6 redesigned proteins

High-resolution structures of the redesigned 1m4w_6 derived revertant mutants revealed the relative contributions of the respective mutations to backbone conformation and binding pocket opening. In agreement with ROSETTALIGAND prediction and part "b" of our hypothesis, the double revertant mutant 1m4w_6w20v48 possessed a native-like "closed" conformation, while the backbone of the 1m4w_6v48 mutant displayed an "open" configuration largely unchanged from 1m4w_6 (Figure 3.7c). The backbone RMSD of 1m4w_6w20v48 was 0.38Å from wild-type, while 1m4w_6v48 was similar to the 1m4w_6 crystallographic structure. Unexpectedly, the "thumb" region of the 1m4w_6w20 mutant was not resolvable due to lack of electron density, indicating a high degree of mobility (Figure 3.7a).

Discussion

The intent of this study was to explore computational methods for designing *de novo* high affinity protein-peptide interfaces. The protein designs described above did not achieve our goal of high affinity binding to their target peptide. Nonetheless, four high-resolution structures of endo-1,4-beta-xylanase derived proteins yielded important insights into the structural dynamics of family 11 xylanase proteins.

Experimental design

The following paragraph summarizes our hypothesis and describes the layout of the work performed: Our hypothesis at the outset of this study was that RosettaLigand was capable of *de novo* design of a high-affinity protein-peptide interface to a non-standard dipeptide ligand. Experimental testing of our original nine protein-peptide interface designs yielded negative results for high affinity ligand binding, thus failing to prove this hypothesis. Subsequent structure determination and detailed analysis of one of the designs, 1m4w_6, led to our second order hypothesis that backbone opening and expansion of the designed ligand binding pocket, caused by specific mutations, resulted in the disruption of predicted binding contacts and consequent lack of ligand affinity. It was hoped that by reverting these specific residues to wild-type, the ligand binding pocket would "re-close", thus allowing the predicted ligand biding interactions to form and bind the target dipeptide with high-affinity.

Testing the second order hypothesis by expression and assay of three redesigned proteins yielded similar negative results for ligand binding. Structure determination and analysis of the three proteins yielded further important insights. Our hypothesis was incorrect in predicting that "re-closing" of the binding pocket would result in high affinity ligand binding. While an expanded, "open" geometry of the binding pocket may contribute to a lack of high affinity binding, a closed geometry, as seen in the structure of the double revertant mutant 1m4w_6w20v48, is not sufficient to confer high affinity ligand binding.

However, part of the second order hypothesis was shown to be true. The two specific residues identified by a detailed ROSETTA energy analysis comparing the predicted and experimentally determined structures of 1m4w_6 were indeed responsible for the binding pocket expansion, and reverting these residues to wild-type restored the predicted geometry of the binding pocket. We speculate that changes in the configurational dynamics of the protein as seen in crystallographic B-factors may be partly responsible for the lack of high-affinity ligand binding. However, confirmation of this hypothesis remains outside the scope of our experimental data. An equally likely contributor to failure may be shortcomings in the ROSETTA energy function, in particular its solvation energy function or treatment of water molecules.

ROSETTALIGAND can accurately predict both the fine and large-scale structure of designed proteins and protein-ligand interfaces

Figure 3.3b compares the position of each sidechain atom for residues that comprise the binding pocket between predicted and experimentally attained 1m4w_6 structures. We see that even with the "opening" of the binding pocket due to expansion of the "thumb" region backbone, the majority of sidechains assume their predicted conformations. Furthermore, even with this "thumb" region backbone shift, the RMSD of all the sidechain atoms in the unliganded 1m4w_6 binding pocket is 0.96Å. This level of accuracy improves still further when the "thumb" region backbone re-adopts the native "closed" conformation, as in the structure if 1m4w_6w20v48, where the residues comprising the unliganded binding pocket attain an RMSD of 0.63Å.

As described in the Results section, we tested ROSETTALIGAND's ability to predict the backbone changes observed in the mutant proteins. This test was omitted in the original design protocol. The original protocol intentionally prevented the protein backbone from adapting in response to mutations introduced during design. The decision to use a fixed-backbone protocol initially was made to increase speed of the calculations and was based on the erroneous assumption that a thermophilic protein scaffold such as 1m4w would be unlikely to experience significant conformational change from the mutation of a small number of residues in the enzymatic cleft. When subsequently using protocols able to accommodate backbone flexibility, ROSETTALIGAND is quantitatively able to predict the shift in backbone configuration when the destabilizing Trp20 and Val48 mutations are alternately

included or removed. If the respective mutations for the "open" 1m4w_6 and "re-closed" 1m4w_6w20v48 are substituted onto the others backbone coordinates, flexible-backbone relaxation protocols in ROSETTALIGAND can accurately recover the backbone conformation observed in the experimental structures and account for binding pocket expansion (Figure 3.8). When the Trp20 and Val48 mutations are introduced onto a native "closed" backbone configuration, the pocket expands to that seen in the 1m4w_6 structure (Figure 3.8c). When the mutations are removed, the backbone "re-closes" to the native 1m4w configuration (Figure 3.8b). Had we adopted a flexible-backbone protocol during our initial design calculations, it is likely that "opening" of the 1m4w_6 design would have been predicted accurately.

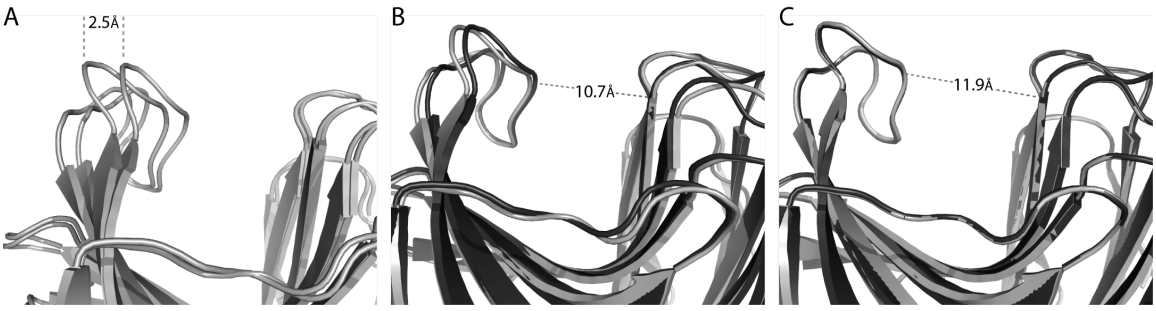


Figure 3.8 ROSETTA flexible backbone protocols can recapitulate backbone conformational shift. (A) 2.5Å magnitude shift in backbone conformation between the "closed" and "opened" confirmations of the 1m4w wild-type and designed protein, respectively. (B) When the W20 and V48 sequence positions are substituted onto an "open" backbone conformation (light grey), ROSETTALIGAND, using flexible backbone protocols, recovers the "closed" configuration (dark grey). (C) Likewise, substituting R20 and L48 onto a "closed" backbone will result in a "re-open" conformation.

We thus conclude that ROSETTALIGAND is able to predict the structure of the 1m4w designs to near atomic resolution of both the binding interface and protein as a

whole, and that the modeling of backbone conformational changes is important when designing protein-peptide interfaces.

Accurate structure prediction of the designed proteins did not translate into binding affinity

Although ROSETTALIGAND can accurately predict large-scale changes in backbone configuration observed in the designed protein structures, the computational protocols employed in this study are significantly limited at addressing complex protein dynamics and potential entropic factors of ligand binding. ROSETTA scoring and binding energy calculations are performed using a single, static, atomic representation of protein and ligand. Although recent advances in flexible backbone and relaxation functionality within ROSETTA have expanded its ability to address structural fluctuation during design (145), the ability to fully predict the effects of dynamics at a protein-ligand interface remains limited.

Analysis of the crystallographic data from all four of the determined 1m4w mutants when compared to wild-type 1m4w indicate a significant increase in both the mobility of the loop forming segments of the proteins "thumb" region and an overall increase in the crystallographic temperature factors (B-factors) of the protein backbone comprising the ligand binding pocket. It is interesting to note that even after the reversion mutations of the 1m4w_6w20v48 protein allowed the "re-closing" of the ligand binding pocket to wild-type dimensions, the global B-factors of the protein, and more significantly those of the "thumb" and "finger" regions which

comprise the two sides of the binding cleft remain elevated an average of more than 60% (Figure 3.9). These elevated B-factors suggest a fundamental alteration in the dynamics of the protein as a whole (146) that could significantly impact the energetics of ligand binding.

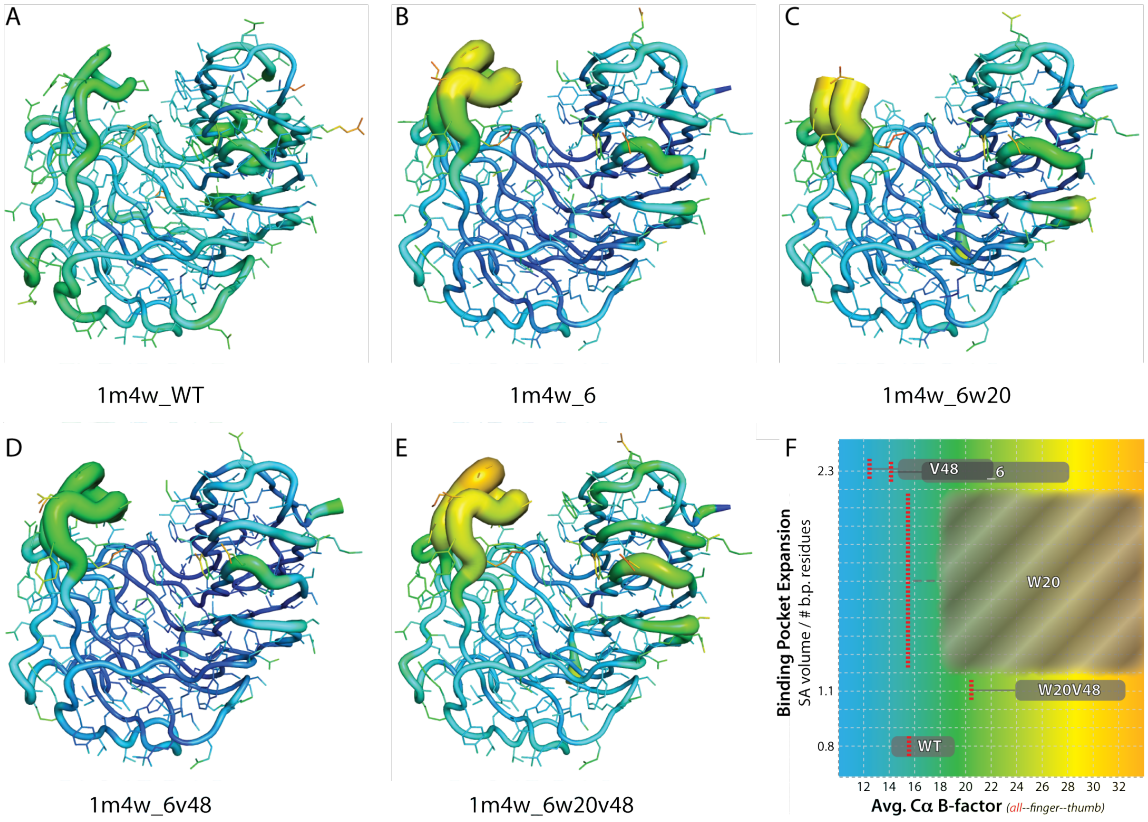


Figure 3.9 Visualization of crystallographic B-factors for wild-type and four 1m4w mutant proteins. Panels A-E: backbone and residue sidechains colored and sized by B-factor values for wild-type 1m4w and X-ray determined structures. Red/thick = higher B-factor, blue/thin = lower B-factor. Panel F displays the average B-factor values (x-axis) as a function of binding pocket volume (y-axis) for each protein (WT = 1m4w; _6 = 1m4w_6; V48 = 1m4w_6v48; W20 = 1m4w_6w20; W20V48 = 1m4w6w20v48). Note that while the average B-factor value for the entire protein (all) decreases for some of the designs, the “thumb” and “finger” B-factors are increased for all designed structures. This suggests a fundamental shift in the overall dynamics of the protein. Also note that the binding pocket volume for 1m4w_6w20 (panel C) is shown as a value range in Panel F due to lack of electron density in the “thumb” region. The binding pocket volume of 1m4w_6v48 and 1m4w_6 are equal. B-factor values for the whole protein (red, dashed line), “finger” region (left extent of grey box) and “thumb” region (right extent of grey box).

Increased dynamic mobility of the "thumb" region specifically can be observed in all four designed structures when compared to wild-type (Figure 3.9). These B-factors are 1.5 to 2.0 fold higher than in the wild-type 1m4w. In the case of the 1m4w_6w20 mutant, the lack of electron density in the "thumb" loop is indicative of increased mobility. This "thumb" region contributes approximately 40% of the ligand interface surface area and 5 of 11 of predicted hydrogen bonds to the ligand. Thus, this observed change in dynamics in the 1m4w "thumb" region is hypothesized to be a contributing factor to the lack of observed ligand binding.

Beyond the implications of altered proteins dynamics, standard ROSETTALIGAND design protocols rely on a bulk, non-explicit solvation term (147) to represent water molecules in and around the binding interface. Entropic factors of binding-pocket desolvation are not well addressed by an implicit solvation term (148). Examination of the four X-ray structures reveal 9 to 11 ordered water molecules within the binding pocket. Due to the increased importance of predicting individual atomic interactions in the design of high-affinity interfaces, the explicit modeling of water molecules is desirable for successful design of protein-ligand interfaces (149)(150). Although recent extensions to ROSETTA now allow explicit interfacial waters to be modeled, this functionality did not exist at the time this study commenced.

Ligand and scaffold selection are important determinants of design success

A dipeptide ligand composed of small, non-polar amino acids is a difficult target for a proof-of-concept experiment and was intended to push the boundaries of ROSETTALIGAND technology. This, however, may have been overly ambitious. A larger, more apolar ligand possessing greater VDW surface area and opportunity for charge-charge interactions would be preferred in future work. Also, it remains an open question as to whether the selection of a "D" peptide target ligand, while theoretically equivalent to "L" amino acids from a chemical and computational standpoint, may have negatively contributed to the difficulty in achieving high affinity binding (151)(152).

More important to the potential success of protein-ligand interface design are the dynamics and conformational stability of a design scaffold protein. As found here, even highly stable, thermophilic proteins with melting temperatures well above 100°C (153) potentially possess dynamic modes that can negatively impact high affinity interface design due to increased entropic penalties for ligand binding. The dynamics of the endo-1,4-beta-xylanase fold, as noted in recent work by Vieira et al., indicate that the 1m4w "thumb" is inherently mobile in solution at elevated *in situ* temperatures (154). Evidence for intensified "thumb" and binding site dynamics can be seen in the crystallographic B-factors of each of the four designed protein structures. The relatively small number of mutations (in the case of 1m4w_6w20, only six) necessary to cause significant destabilizing dynamics was unanticipated for a thermophilic protein. This dynamic propensity is an undesirable trait in a protein

scaffold when attempting to design a well-defined, stable, high-affinity interface. Meticulous and deliberate care is advisable when choosing a *de novo* design scaffold, and particular attention should be given to protein dynamic modes. In this respect, scaffolds that have been extensively classified by NMR, SAXS, molecular dynamic simulations or other methods which yield information on protein dynamics are preferred.

The high-resolution structures of ROSETTALIGAND interface designs reveal critical structural and dynamic determinants of β-xylanase proteins

The most notable feature of the 1m4w_6 designed protein when compared to the wild-type 1m4w protein scaffold is the radial expansion of the binding pocket defined by the "thumb", "palm" and "finger" regions (Figure 3.3a). A similar degree of expansion is also observed in the 1m4w_6v48 derivative of 1m4w_6, where Leu at position 48 has been reverted to wild-type Val. These two designs share a common mutation of Trp to Arg at position 20, which disrupts a critical hydrophobic contact between "finger" (W20) and "thumb" (P125), resulting in expansion of the binding pocket (Figure 3.3c).

Necessary but not sufficient for closure of the binding pocket of 1m4w_6 and its derivatives is the restoration of the hydrophobic contact between residues Trp20 and Pro125. This interaction is crucial to maintaining a closed geometry under crystallization conditions. At higher, *in situ* temperatures near 100°C where this enzyme has evolved to function (112), this interaction may be important in

regulating the dynamics and enzyme kinetics of the 1m4w protein. That this Trp-Pro interaction is highly conserved across multiple species indicates it is likely a key structural, dynamic and kinetic determinant common to family 11 xylanases.

While the hydrophobic Trp20-Pro125 interaction is necessary, it is not sufficient to allow stable closing of the binding pocket. The destabilization and consequent lack of electron density observed in the crystal structure of 1m4w_6w20 results from a clash of an alternative configuration of Trp80 in the "palm" with Ile127 in the loop which forms the "thumb" (Figure 3.7a). This clash is in turn due to the altered packing of Tyr78, which is directly caused by the added steric bulk of the Ile48 mutation in the "fingers". It is this "domino" effect leading from I48 > Y78 > W80 > I127 ("fingers" to "palm" to "thumb") that breaks the contact between Trp20 and Pro125, thereby resulting in added mobility of the "thumb" loop (Figure 3.7b). Thus, although reversion of position 20 to the wild-type Trp is necessary for binding pocket closing, it is not in and of itself sufficient. The designed Leu at position 48 must also be reverted to wild-type Val to result in a "closed" pocket configuration (Figure 3.7c).

It is intriguing that the effects of a single, conservative substitution at a spatially distal amino acid position can have such a pronounced effect on the stability of a thermophilic protein at relatively low temperature – i.e. that the additional bulk of a single carbon atom is transmitted from one side of the protein to the other, through three (bulky) amino acid sidechains, to destabilize a large tertiary structural element at well below physiologic temperature. This suggests that the amino acid

sequence of the 1m4w protein, even in the protein core (palm region), is finely tuned to accommodate this dynamic mobility. This further suggests that the increased dynamic mobility of the "thumb" region due to mutations introduced during design, mimics the effect of increased temperature. These mutations might therefore be thought of as having enabled high-temperature, native-like dynamics at low temperatures.

The continuing challenge of de novo protein-peptide interface design

While the lack of success experienced in the course of this particular study may or may not be attributable to factors such as unfortunate scaffolds selection, unanticipated protein dynamics or the lack of explicitly modeled interfacial waters, it is important to note that progress in the field of *de novo* ligand interface design as a whole has lagged significantly behind other areas of *de novo* protein design. Though not long ago considered by some to be a solved problem, retractions in several key papers in the last several years (29) have led to the conclusion that the design of high affinity protein-ligand interfaces is one of the fundamental areas of basic protein function to remain an open problem (26).

ROSETTA has proven adept at such challenging tasks as design of novel protein folds (80), altered recognition and cleavage specificity of a DNA endonuclease (155) and even the design of enzymes with catalytic modes not found in nature (11; 12; 131). Protein-protein interfaces have been re-designed for altered and multi-specificity (86; 156), while ROSETTA and other techniques have successfully re-designed

protein-peptide interfaces for altered specificity and increased affinity (157; 158)(91).

What is it that makes *de novo* design of protein-ligand interfaces so difficult, and why would *de novo* interface design be significantly more challenging than the re-design of a protein-peptide interface, or the design of a novel enzyme? While a completely satisfactory answer to these questions has yet to be established, one contributing factor could be protein dynamics. The requirement to design and manipulate dynamics may set a higher bar for the *de novo* design of ligand binding. Unfortunately, protein dynamics is also one of the most difficult and least tractable problems for current protein design programs.

De novo protein design by definition entails establishing entirely new functionality in a protein that did not previously possess such function. It requires an ability to recreate and manipulate all properties of a protein necessary for a given function. Conversely, non- *de novo* design, where basic protein functionality is retained but altered – as when re-designing the ligand binding specificity or increasing affinity – relies on conserved intrinsic properties of the protein important to its function. Such conserved intrinsic properties could include protein dynamic modes conducive to ligand binding. Similarly, re-design of protein-protein specificity may benefit from conserved functionality and dynamics, as well as having the added advantage of a larger interface surface area and number of potential interactions to offset small errors in the design algorithms. Such small errors may have a larger impact in ligand

interface design where each of a small number of interactions must be optimal for tight interaction.

Yet surely the creation of novel catalytic function in the *de novo* design of enzymes (11; 12; 131) requires no less precision and accuracy than the design of ligand binding. What has allowed these efforts to succeed where interface design has yet to? A partial answer may lie in the nature of enzyme function. In this case, the precise geometry of the catalytic mechanism is critical, and facilitating this geometry can be thought of as "binding" the chemical transition state. However, the timescale on which transition state "binding" occurs is extremely short, on the order of 10^{-12} seconds, when compared to high affinity ligand binding interactions which must be maintained for seconds or longer (159). Furthermore, recent studies suggest that the chemical step in enzyme catalysis is insensitive to global protein dynamics, which instead affect only enzyme kinetics (160; 161). In this light, it is notable that all of the successful enzyme designs cited above were performed using a naturally occurring enzyme as a design scaffold (some even used 1m4w) and that all of these designed enzymes possess relatively poor kinetic properties, even after undergoing multiple rounds of directed evolution to address the lack of kinetic efficiency (11; 12; 131). The implication of these observations match the findings of this study, which found that ROSETTA was capable of designing interfaces with a high degree of structural/geometric accuracy – as would be needed to stabilize a catalytic transition state intermediate – but lacked the ability to account for or design protein dynamic modes necessary for binding or efficient kinetics. While these speculations

are far from conclusive with the small amount of evidence presented here, it is an intriguing line of thought that may warrant further attention in future studies.

Conclusion

Our attempts at using the ROSETTALIGAND program to design *in silico* a high-affinity protein-peptide interface to a bacterial dipeptide target were unsuccessful. Twelve proteins using 1m4w as a design scaffold were assayed for binding to their intended target. No high-affinity binding was detected for any of these twelve designs.

We have proposed several potential contributors to this apparent lack of success, including overambitious target peptide selection and the lack of explicitly modeled interfacial water molecules. However, extensive evidence indicates that possibly the most significant negative contributor to the study outcome may be the unappreciated nature and extent of dynamics inherent to the design scaffold protein.

We have shown that ROSETTALIGAND is able to predict the structure of a designed interface to near-atomic resolution, and of large-scale protein conformational changes due to mutations introduced during the design process. However, accurate structure prediction did not translate into successful design of high affinity ligand binding. We therefore conclude that the computational design of proteins that tightly bind small molecules remains possibly a greater challenge than the design of enzymes. While computational enzyme design requires accurate structural

prediction of catalytic residues, no tight substrate binding is needed for success, and is less sensitive to the pervasive effects of protein dynamics.

In addition to the lessons and caveats learned above concerning protein design applications, we have also gained new information regarding structural and functional determinants of family 11 endo-1,4-beta-xylanase proteins. Specifically, the four high-resolution X-ray structures complement prior reports of the catalytic dynamics of the “thumb” region of in family 11 xylanases, as well as reveal new insights into individual amino acids involved in the structural and functional dynamics of the beta-xylanase protein fold. These xylanase structures may also serve as benchmark systems for future computational design protocols that model protein-peptide or protein-small molecule interfaces.

CHAPTER IV

ROSETTA SEQUENCE CHARACTERIZATION AND RECAPITULATION OF PROTEIN INTERFACES TO SMALL-MOLECULE AND PEPTIDE LIGANDS

Introduction

The prediction and design of protein-peptide and protein-small-molecule interfaces is an important but relatively unproven capability of ROSETTA3. While significant investigation into designing protein-protein interfaces (82), altering protein-protein interaction specificity (85), engineering catalysis (11; 12), predicting small-molecule binding affinity (162) and designing ligand affinity (91) has and is being addressed, the design of the protein interface to peptides and small-molecules constitutes something of a gap in ROSETTA research. As an initial step into investigating ROSETTA3's proficiency at designing protein-peptide and protein-small-molecule interfaces, an extensive sequence recovery benchmark was performed on a diverse and representative set of liganded protein holostructures derived from the Ligand Protein Database (LPDB)(163). A statistical examination was also made of the structural and sequence-specific properties of wild-type peptide and small-molecule binding interfaces. These wild-type interface propensities were then compared to ROSETTA designed interfaces.

A prerequisite to the reliable design of novel, functional protein-ligand interfaces is to demonstrate the ability to accurately predict and recapitulate native-like protein-ligand structure and interactions. This may be accomplished on different levels of precision, utilizing several different objective functions. Amongst the most fundamental objective functions to be found in a protein design context is the recapitulation of protein primary sequence. A central principle of structural biology is that a protein primary amino acid sequence dictates the protein three-dimensional structure, and thus its function. Therefore, the ability to accurately and reliably recapitulate the primary sequence of a protein can act as an effective, if somewhat incomplete, proxy for structure prediction, and can thereby be considered a good first-order test of a computational design algorithm. While the recovery of the primary amino acid sequence does not and cannot contain all of the information of other more rigorous objective functions, such as deltaG of binding, it is nonetheless a practical and effective way of assessing the basic competency of a computational design method. These assessments can then be highly useful in diagnosing and remedying flaws and deficiencies in the method under evaluation.

Such was the motivation for the following sequence recovery experiments within the context of my dissertation work. Having previously established significant flaws in ROSETTA's ability to accurately design high affinity protein-ligand interfaces, I set out to determine whether ROSETTA could recapitulate this basic objective function of "native" ligand binding interfaces using primary sequence as a metric.

Results from the prior beta-xylanase design study revealed that ROSETTA is adept at predicting the atomic level structure of a known protein scaffold given a specific set of sequence modifications. However, because this accuracy did not translate into an accurate prediction of high affinity binding, it was decided to take a step back down the hierarchy of design objective functions and evaluate the prediction of protein primary sequence.

In order to properly evaluate this effort, proteins known to natively bind small-molecule and peptide ligands, whose high-resolution structures and thermodynamic binding energies had been determined, were selected for study. ROSETTA was then tasked with predicting the atomic level structure of a set of protein-ligand interfaces without knowledge of the primary sequence. The success with which ROSETTA recovered the native amino acid sequence was evaluated to determine ROSETTA's proclivities, aptitudes and weaknesses, thus yielding a starting point for improvement of ROSETTA's design function.

Furthermore, these sequence recovery experiments could be performed over a spectrum of protocols, ranging from the idealized, to evaluate the theoretical maximum performance of ROSETTA and its component functions, to the practical, by emulating the protocols and functionality found in an interface design application.

Prior Studies

Significant prior work has been performed in developing computational methods to identify protein-ligand binding interfaces. Identifying where on a protein a ligand binds greatly enhances rational approaches to synthetic and computational drug design and promotes a more thorough understanding of the therapeutic mode of action. Numerous computational approaches have demonstrated some level of success in this regard, including machine learning techniques (164), statistical analysis (165) and hybrid computational and experimental methods (166). However, while these methods each display differing abilities at identifying the location of ligand binding sites, they do not in and of themselves shed light on the fundamental sequence level properties of a ligand interface.

Prior examinations of protein-ligand interfaces at the sequence level have been few, and their findings largely lacking on agreement. Villar & Kauvar conducted an analysis of the sequence characteristics of ligand interfaces on a set of 50 protein-ligand complexes chosen from the PDB, 13 of which were enzymes. They found that Gly, Ser, Arg, His, Trp and Tyr were overrepresented at binding sites, while Pro, Lys, Glu and Ala were underrepresented. They also observed a peaking of hydrophobic amino acids at intermediate distances from the binding site that they postulated corresponded to the locations of the protein core relative to the surface ligand binding sites (167). Different findings were reached by a separate group, which examined a larger set of 756 protein-ligand complexes identified using automated means. Their data saw an overrepresentation of Trp, His, Phe, Met, Tyr & Cys, and an

underrepresentation of Pro, Lys, Gln, Ala, Thr and Gly (168). They also failed to observe an increase in hydrophobic residues at intermediate distances form the binding site. While there is some overlap in their findings, a significant discrepancy remains. This discrepancy may be due to the differing compositions of the respective protein-ligand structure test sets used in the studies. From this, I concluded that creating my own test set composed specifically of hand curated, diverse protein-ligand interfaces to small-molecule and peptide ligands was necessary to assure accurate analysis of primary sequence characteristics in protein-ligand interfaces.

Methods

Creation of a protein-ligand test set

For computational studies involving structure and or sequence prediction, it is vitally important that protein structures used to measure the accuracy of a prediction be highly accurate. Unfortunately, it is often the case that small-molecule and non-proteinaceous ligand structures retrieved from the PDB possess significant errors in bond angle, length or other parameterizations. Because computational energy functions can be highly sensitive to these types of input errors, it is important that correct protein-ligand complex structure files be used when evaluating the properties of a ligand interface. To help address this need, the Ligand Protein Database (LPDB)(163) was created. The LPDB combines experimentally derived binding data with protein-ligand structures that have been minimized in the

MSI CHARMM force field to correct ligand parameterization errors introduced in the course of the normal crystallographic structure refinement process.

I began compilation of a representative ligand interface structure set for use in this study by selecting all proteins from the LPDB which possessed a single ligand of two hundred total atoms or less, no interfacial of structurally relevant water molecules or metal ions and whose protein was composed of a single, continuous polypeptide chain. From the remaining files all water and crystallographic reagents were removed, as were any alternate "aniso" atom statements. The files were then renumbered, and the chain IDs of the protein and ligand were assigned "A" and "X" respectively.

This process yielded a set of 174 protein-ligand complex files we designated the "full" set. The full set was then further culled at a 30% sequence homology cutoff using the PISCES server (169), and SCOP protein class, family, fold classifications to assure diverse and non-overlapping structure, function and sequence identity of each protein-ligand complex in the experimental test set. This final culled set is comprised of 43 protein-ligand complexes and was named the "diverse" set. Roughly half of the complexes in the diverse set possess experimentally derived binding affinity data. Additional copies of all 174 complexes were also created with the ligand entirely removed to evaluate the capabilities and propensities of ROSETTA when no ligand is present.

Preparation of protein-ligand test sets for ROSETTA

ROSETTA design computations that utilize non-proteinaceous ligand molecules require each ligand be parameterized to match defined ROSETTA atom types in a ROSETTA .params file. All ligands in the test set were downloaded in .mol file format, and then used to generate .params files for each ligand using the standard ROSETTA molfile_to_params.py script. This script outputs both a .params file and a .pdb format file of the ligand atoms which then replace the ligand statements in the original protein-ligand complex .pdb file. This process was performed for each of the 174 members of the full structure set.

Relaxation of LPDB structures in the ROSETTA force field

Prior to performing the ROSETTA design runs, each of the 174 protein-ligand complexes was extensively minimized using the ROSETTA fast-relax algorithm. Although all of the structures obtained from the LPDB had previously been minimized using the CHARMM force field, because the recapitulation experiment was to be conducted entirely using the ROSETTA design force field, it was desirable to re-minimize the structure files in ROSETTA prior to computations. Each protein-ligand complex underwent successive rounds of protein sidechain and backbone minimization until the per round decrease in overall energy of the complex plateaued, with consecutive energy drops between rounds of less than 2%. The ligand pose and configuration were not altered, except for the rotation of methyl groups. Although no clashes or large magnitude conflicts were observed between

the two force fields, some of the larger protein complexes exhibited as much as a 10% change in overall energy between the CHARMM and ROSETTA minimum energy states. For many of the structures, achieving this level of minimization required as many as 70 or more rounds of relaxation.

Distance binning of the protein amino acids

In order to evaluate the change in native amino acid composition as well as the ability of ROSETTA to design ligand interfaces at different distances, all of the protein amino acids in each of the 174 members of the full protein-ligand complex test set were divided into bins at set distances from the ligand as follows: Any protein backbone beta carbon – or virtual beta carbon in the case of Gly, defined as a point equidistant between each of the Gly alpha carbon hydrogens – within a prescribed distance of any ligand atom was included in a distance bin (see Figure 4.1). Bins were set at 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 16, 18, 20 and 22 angstroms and infinity, inclusive - i.e, the infinite distance bin includes all of the members of the 22 angstrom and lower bins, etc. - however not all of these bins were used in every study. Lists of the residues in each bin were then used to generate .resfiles which describe to ROSETTA at which amino acid positions to perform a desired function, in this case design.

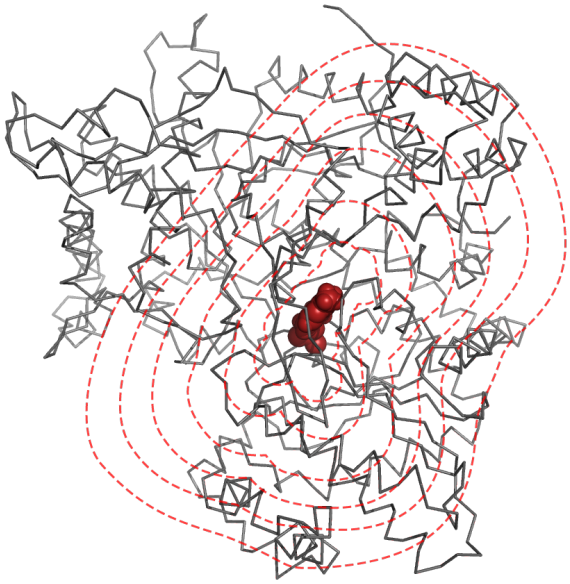


Figure 4.1 Ligand-sidechain distance binning of LPDB protein-ligand complexes. Each protein is divided into eleven bins at 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 angstroms and ∞ based on C-beta to ligand atom distance.

ROSETTA interface design using the XML scripter

Utilizing each of the three previously generated file types (.pdb, .resfile and .params) for all 43 of the protein-ligand complexes in the diverse structure set, a simple design protocol was created using the newly developed XML scripting function of ROSETTA3 (73). For each interface residue in a given distance bin, the XML protocol sampled rotamer combinations of all 20 standard amino acids at each position defined within the bin, without knowledge of the native primary sequence and with minimal perturbation of the ligand pose. In all cases the ending perturbation of the ligand was less than 0.05Å RMSD. (Sample ROSETTA command lines and an example XML script can be found in the Appendix.) The protocol therefore consisted of a

simple repacking and minimization of each designated residue using an extended set of backbone-dependent sidechain rotamers from all 20 amino acids without preference or knowledge of the native sequence.

For each of the 43 complexes at each of the 15 design distance bins, with and without ligand present in the binding site, 100 structures were output. Thus, 129,000 structures were created using ROSETTA for evaluation and analysis (43 x 15 x 2 x 100 = 129,000). Computations were performed on the Vanderbilt Advanced Computing Center for Research and Education (ACCRE).

Analysis of ROSETTA recapitulated interfaces

Evaluation of the output .pdb files was accomplished using the Meiler lab BCL library executable "bcl_app_calculate_sequence_recovery.cpp", which uses a list of output .pdb files, a native .pdb file and the design .resfiles to compute a table of amino acid changes from the native sequence (an example command line for bcl_app_calculate_sequence_recovery.cpp and sample output files can be found in Appendix ##). Interface energies of the output .pdb files was extracted using shell scripting from individual output .pdbs. Both of these data sets were then parsed further and used to generate plots for analysis.

Results and Discussion

Sequence characteristics of native protein-ligand interfaces

From the diverse set of 43 protein-ligand complexes divided into distance bins, an analysis of the sequence level characteristics of ligand interfaces can be performed and compared to results found in past studies (see "Prior Studies", above). When normalized to amino acid frequencies found globally (170), the amino acid composition of residue positions close the ligand interface were seen to be highly skewed. As seen in Figure 4.2, an overrepresentation of Asp, Cys, Gly, Trp and Tyr; and an underrepresentation of Arg, Gln, Glu, Leu, Lys, Pro and Thr is observed, with the frequencies becoming more, but not wholly, normal as we move more distant from the ligand interface. While these observed amino acid frequencies agree in part with the two cited prior studies, they nonetheless diverge significantly for specific amino acids. It is also noted that we saw in increase in the frequency of hydrophobic amino acids at intermediate distances, in agreement with Villar & Kauvar, (1994). Worthy of note is that the four angstrom bin data is not displayed due to the dramatic but trivial propensity towards small amino acids, as would be expected for such a short distance cutoff.

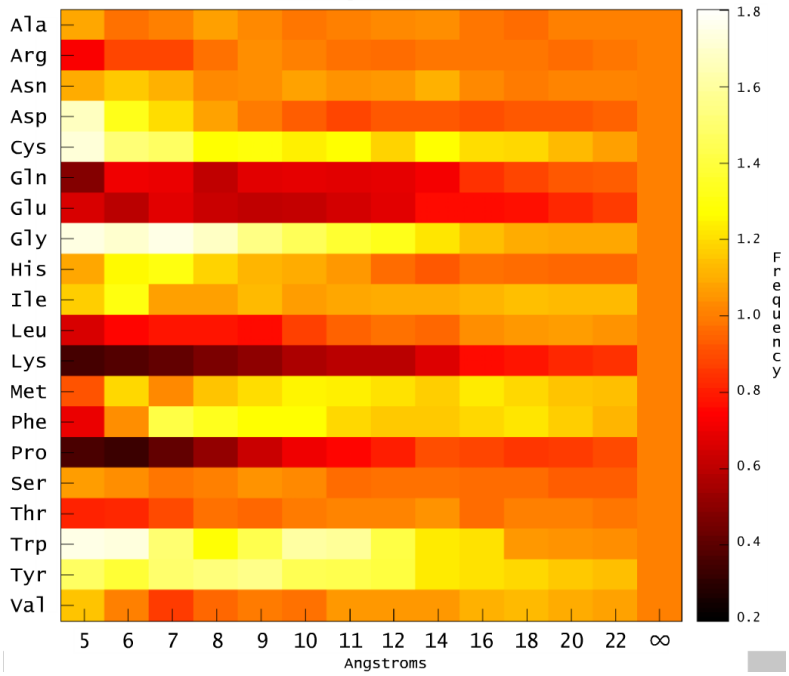


Figure 4.2 Heatmap of normalized amino acid frequency per distance from ligand for 43 diverse proteins. Frequency scale at right denotes ratio of over or under representation of amino acids in test set versus global amino acid frequency.

That three different studies arrive at three different answers to the same question is somewhat disconcerting. The likely explanation is that this analysis is highly dependent on the starting data set of protein complexes used for evaluation. Having compiled a hand-curated and diverse set of highly non-homologous proteins, in contrast to Soga et al., without an overrepresentation of enzymes, as was intentionally done by Villar et al., we feel that our set of proteins yields the most accurate results for the sequence evaluation of native interfaces to small-molecule and peptide ligands.

From these amino acid frequency results of native binding sites, we have gained a point of reference from which to evaluate the ROSETTA recapitulation of ligand

interfaces using this same protein-ligand complex data set. By comparing the ROSETTA results to these data, we now have the ability to identify flaws in the ROSETTA design algorithm.

ROSETTA recapitulation of protein-ligand interfaces

Computational resources:

Using the ROSETTA3 XML scripter and basic design protocol described above, the majority of the 43 protein-ligand complexes completed their 100 output structures with 24-72 hours on individual ACCRE Opteron CPUs, depending on the size of the distance bin, and hence number of residues being designed. Larger proteins generally took longer to complete their runs than did smaller proteins in any given distance bin, however, a few of the 43 complexes computed extremely slowly, irrespective of protein size or bin. Some of these took more than one week to complete 100 structures and required as much as 9GB of ram during design runs. No satisfactory reason for this compute time or memory differential has been discovered.

Percent sequence recovery:

To evaluate the degree of sequence recapitulation achieved by ROSETTA, a percent recovery is calculated by comparing the native amino acid identities to all of the designed amino acids in each distance bin. A plot of average sequence recovery for

each of the 43 complexes in the diverse structure set at each of eleven distance bins is shown in Figure 4.3 both with and without ligand. Here we see several interesting features.

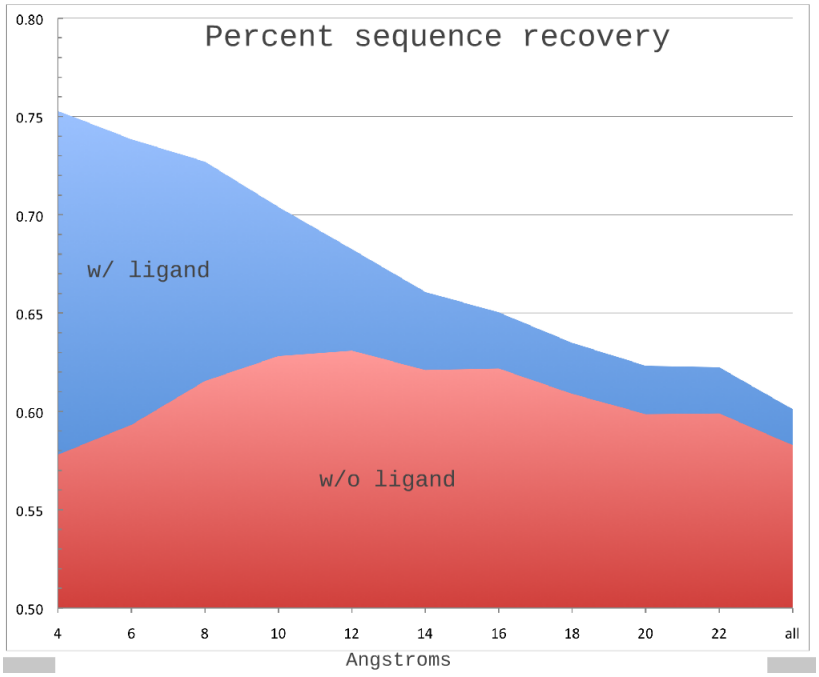


Figure 4.3 Plot of aggregate percent sequence recovery for diverse set of 43 protein-ligand complexes with and without ligand present at each distance bin.

First, we see a clear difference between liganded and non-liganded design runs. In the presence of a ligand, the sequence near the binding site is far more likely to be preserved than in the absence of a ligand. This denotes ROSETTA's clear ability to recognize sequence optimization for ligand interaction of a given interface. Note also how the liganded percent recovery approaches parity with non-liganded as the distance from the ligand grows. Second, the overall rate of sequence recovery is quite good out to around eight angstroms – the extent of direct interaction with the ligand. Third, even without ligand present ROSETTA demonstrates clear aptitude at

recovering primary sequence between 10-16 angstroms, which corresponds generally to the hydrophobic core of the proteins, as has been noted previously. Furthermore, it is worth keeping in mind that some of the proteins included in this experimental set may be promiscuous binders whose sequence is optimized to interact with multiple ligands. In such cases the ideal sequence recovery rate would be below 100%.

This average data can be broken down into percent recovery for each individual amino acid type for each distance bin, both with and without ligand present, as seen in the heatmaps of Figure 4.4. From this data we can clearly see that ROSETTA under predicts Glu, and to a lesser extent Met and Arg at ligand interfaces, while under predicting Lys in general throughout the protein, but especially at the interface. This trend largely reverses itself at larger distances, where we see an under prediction of predominantly charged amino acid types in the non-interfacial regions of the protein. In contrast, the non-liganded designs tend to under predict the same charged residue types throughout the protein. This result is generally encouraging, suggesting that ROSETTA is fairly adept at predicting interfaces at the sequence level, but may imply some need for improvement in ROSETTA's ability to predict charged amino acid interactions.

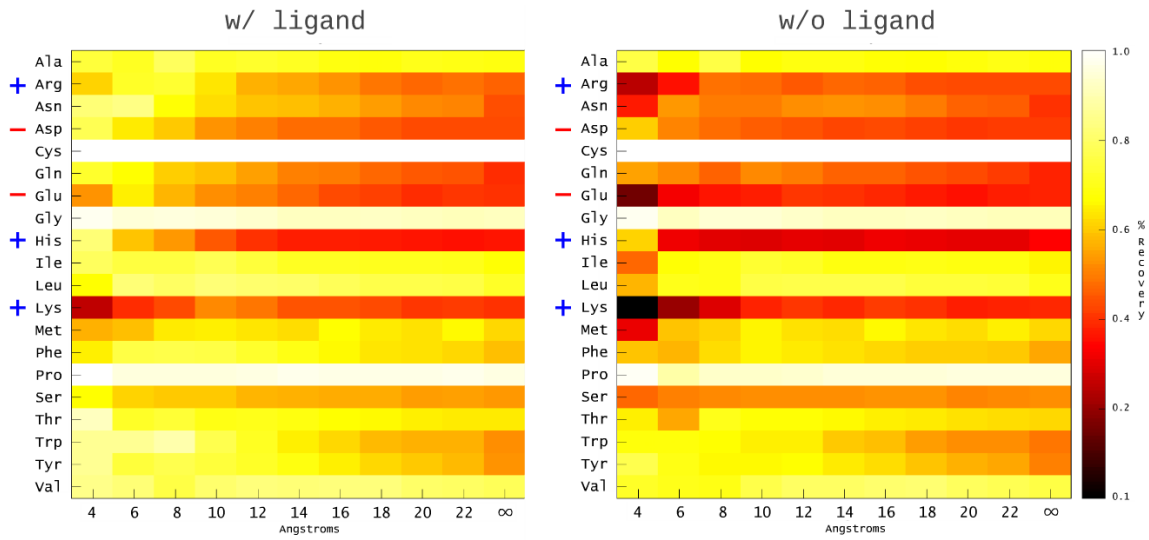


Figure 4.4 Heatmaps of percent sequence recovery by amino acid type for each of eleven distance bins, with and without ligand. Color bar at right denotes degree of sequence recovery. +/- designate charges of specific amino acids.

Amino acid substitution propensity:

We may further break down the sequence recovery data by looking at the likelihood of a specific amino acid change for a given distance bin. In Figure 4.5 we see a substitution table constructed to show the propensity for each amino acid type to change to any other amino acid for the eight angstrom distance bin. On the Y-axis is the native (starting) residue type, and on the X-axis is what those native residues were designed into. Thus we see that out to eight angstroms, all residues are most likely to retain their native identity, demarcated by a dark diagonal from the top left to lower right-hand corners. However, we also see that to a significant degree ROSETTA considers Ser and Ala somewhat interchangeable. Less of a surprise is the propensity for changing Asp to Asn in the 8 angstrom bin.

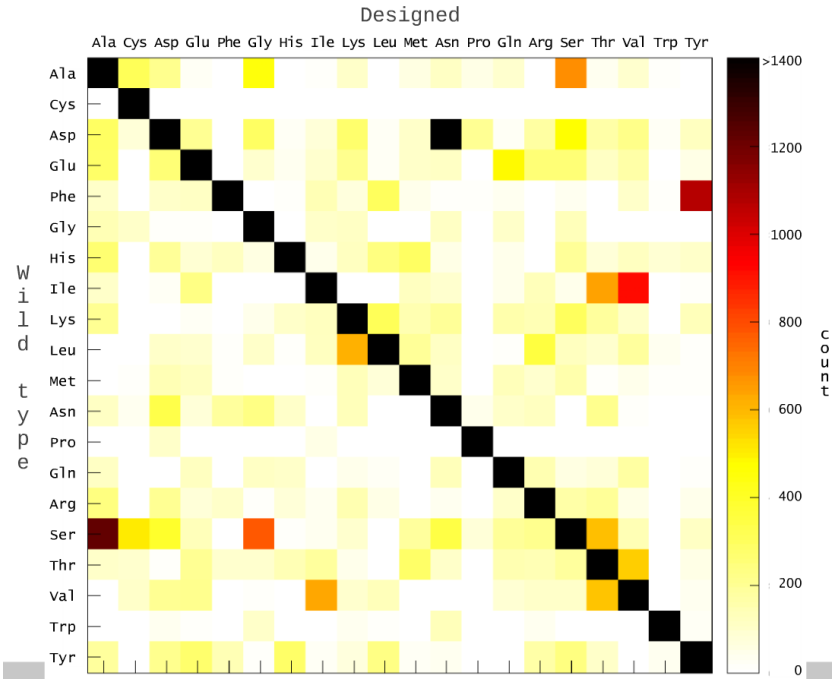


Figure 4.5 Chart of individual amino acid substitutions for the 8 angstrom bin. Y-axis denoted native amino acids while the x-axis denotes design substitutions. Color bar at right denotes total amino acid count.

If we compile all of this type of propensity analysis for each of the eleven distance bins, both with and without ligand present, we can construct an amino acid propensity heatmap (see Figure 4.6). By subtracting the propensities without ligand from the propensities with ligand, we generate a heatmap showing the difference in design propensities between liganded and non-liganded designs of all 43 protein complexes (Figure 4.7). While the 4 angstrom data is difficult to interpret due to the strong size bias at such short bin lengths, we clearly see some irregularities in the ROSETTA predictions. For instance, the large amino acids are highly over predicted and there is an unwarranted preference for Glu over Gln.

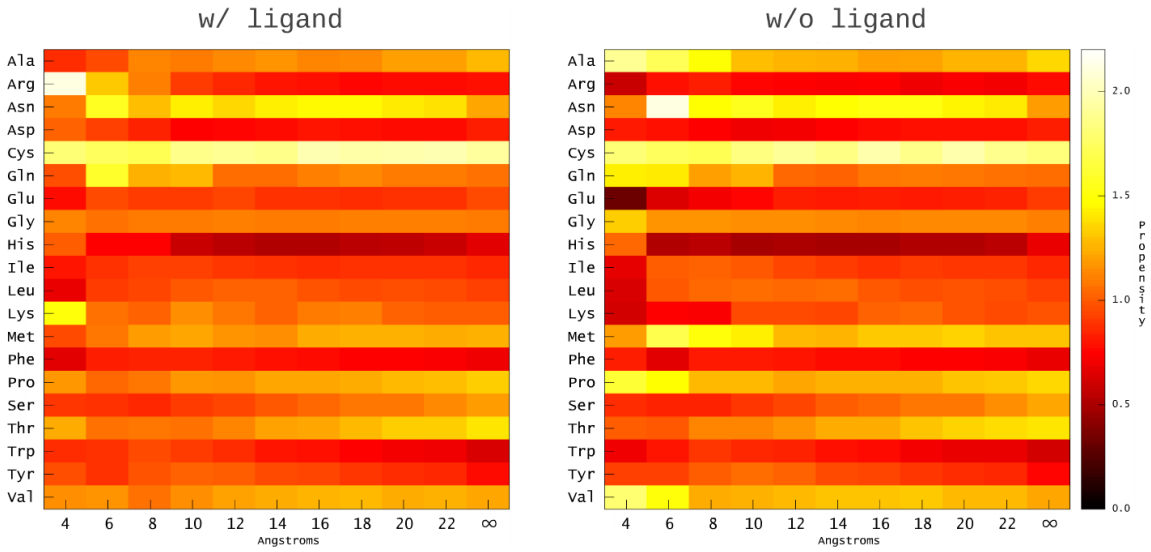


Figure 4.6 Heatmaps of individual amino acid design propensity with and without ligand at each of eleven distance bins. Colorbar at right denotes propensity ratio to global amino acid frequency.

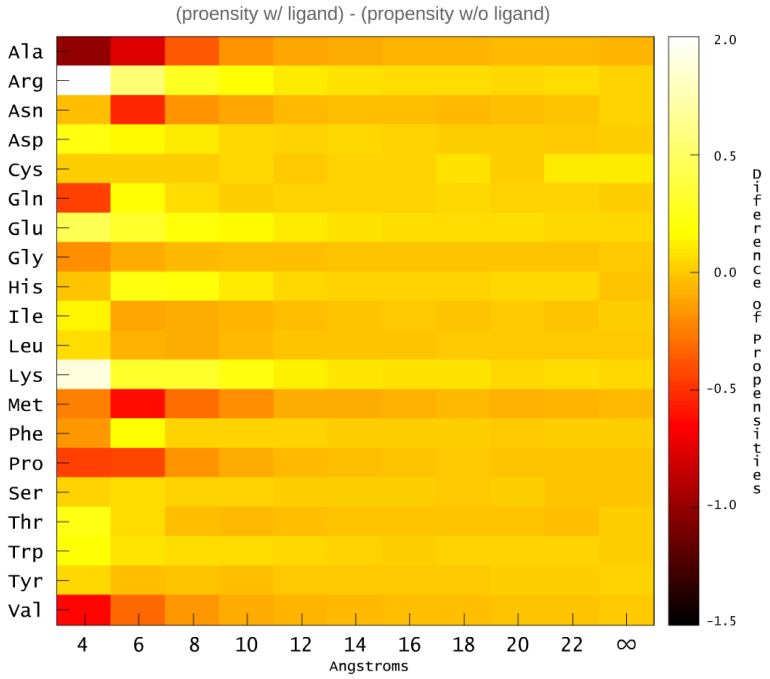


Figure 4.7 Heatmap of difference in amino acid design propensity (propensity w/ ligand – propensity w/o ligand) for each of eleven distance bins. Colorbar at right denotes ratio of propensity difference.

The six angstrom data more consistently indicate ROSETTA's under and over prediction of various amino acid types near the ligand interface. Ala, Asn, Met, Pro and to some extent Val are all under predicted by ROSETTA by between 0.40-0.85 times at the binding interface compared to the protein as a whole, while Arg, Asp, Gln, His, Lys and Phe are over predicted. This type of data can be invaluable in evaluating the strengths and weaknesses of the ROSETTA design energy functions.

Percent recovery as a function of other protein properties:

Beyond the evaluation of percent sequence recovery at designated distance cutoffs, we also examined percent recovery as a function of both binding affinity and ligand size. It seems a reasonable hypothesis that the higher the affinity of the ligand interface, the more optimized the interface sequence should be and therefore the higher the sequence recovery we can expect. Unfortunately this was not the case as seen in the data in Figure 4.8. There is no discernable correlation between experimentally derived deltaG of binding and sequence recovery for our test set. More encouraging findings can be observed in the plot of ligand size versus sequence recovery (Figure 4.9). Here we find that, in general, as the number of heavy atoms in the ligand increases, so too does the lower threshold for sequence recovery.

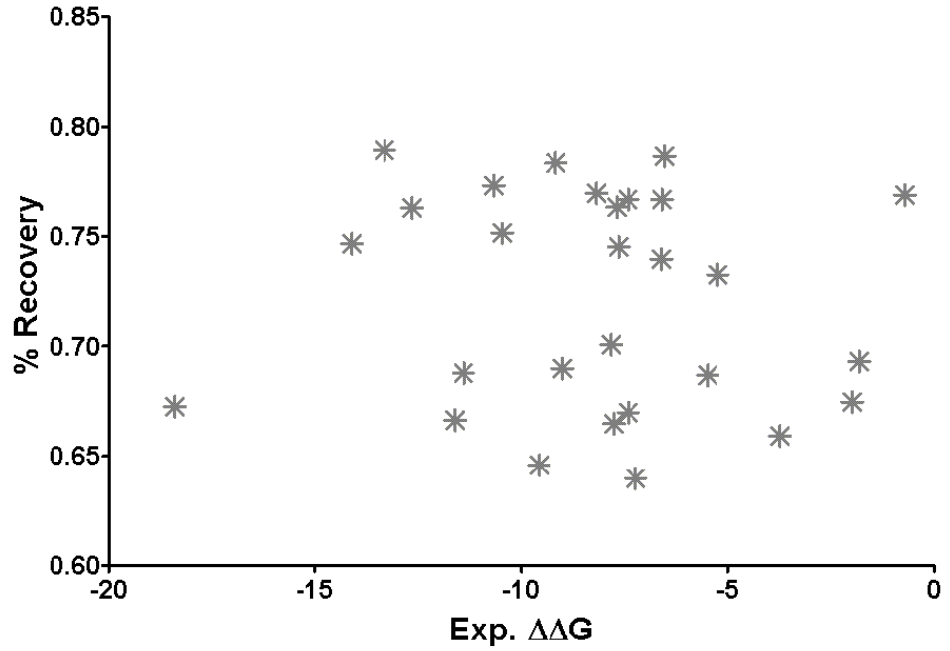


Figure 4.8 Plot of percent amino acid recovery versus experimental delatG of binding for members of the diverse 43 protein-ligand complex test set. X-axis is in kcal/mol.

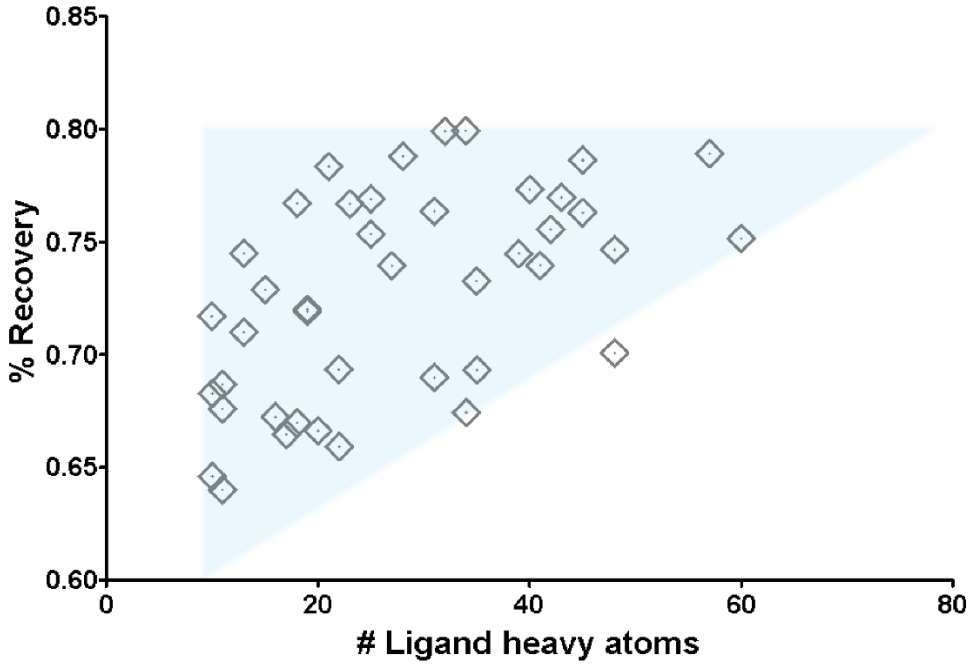


Figure 4.9 Plot of percent amino acid recovery versus size of ligand in number of heavy atoms for members of the diverse 43 protein-ligand complex test set.

Preliminary results from expanded protocols and a newer version of ROSETTA3

With the relative success of the basic sequence recovery experiments, efforts were made to expand the experimental protocols to encompass more structurally probing and realistic functionalities of ROSETTA more closely replicating real world interface design applications and to evaluate recently updated ROSETTA3 dock/design code.

The same set of 43 diverse proteins were narrowed further by removing all complexes containing ligands of over 100 atoms in size, thus biasing the data set towards smaller, more drug-like ligands. The new "small-molecule" protein-ligand set contained a total of 30 proteins taken from the previous diverse set.

Experimental design

In the time since the previous sequence recovery experiments commenced, updated and expanded ROSETTA3 code has been developed which both integrates formerly disparate ROSETTA dock/design functionality and adds new functionality to the XML dock/design scripter. Thus, we set out to both replicate the prior sequence recovery results and add new, expanded XML scripting protocols, which better approximate potential ROSETTA interface design applications.

First, idealized sequence recovery using the native ligand conformation and pose would be replicated as in the prior study. However, this time new XML scripting functionality would allow finer control over ligand perturbations – reducing the

overall translation and rotation from less than 0.05 angstroms to zero angstroms. Percent sequence recovery would be evaluated as in the prior study.

Second, docking runs using new flexible-ligand/fragment functionality would be undertaken where the primary protein sequence would not be altered from the native, while the ligand underwent flexible docking and simultaneous protein sidechain repacking against a rigid protein backbone. RMSD from the native ligand pose versus predicted interface energy and structure recovery of the sidechain rotamers comprising the protein interface would be evaluated.

Third, the first two protocols would be combined into a simultaneous docking/design protocol that would measure sequence recovery concurrent with pose and structure recovery of the ligand and protein interface, respectively. This protocol would most closely resemble a protocol used in applications of the dock/design functionality.

Preparation of flexible-ligand fragment files

In addition to the .pdb, .resfile and .params files, new ligand fragment input files were created which augment the previous .pdb and .params files. These new files contain structural and conformational information about the ligand fragments used during docking and design.

To begin, a set of ligand conformations is generated using “Schrodinger Maestro ConfGen Advanced” ligand conformer generation function. Ligand files in .sdf format

from each of the 30 protein-ligand complexes are used to generate ligand conformations. (see Figure 4.10 for ConfGen parameters). The parameters were set to output between 1-150 conformers for each ligand depending on the number of rotatable bonds and degree of conformational restraints inherent in each ligand. The resulting conformations are saved as a single .pdb file containing all ligand conformations for a given ligand. The standard ROSETTA script "auto_conformers.sh" can then be invoked by supplying both the .sdf file of the ligand and the .pdb file containing all Maestro generated ligand conformers. (see Appendix for sample auto_conformers.sh command line.) The output .pdb and .params files from the "auto_conformers.sh" script are then appended to the ROSETTA3 dock/design XML scripter command line (see Appendix for sample ROSETTA3 command lines and all input files).

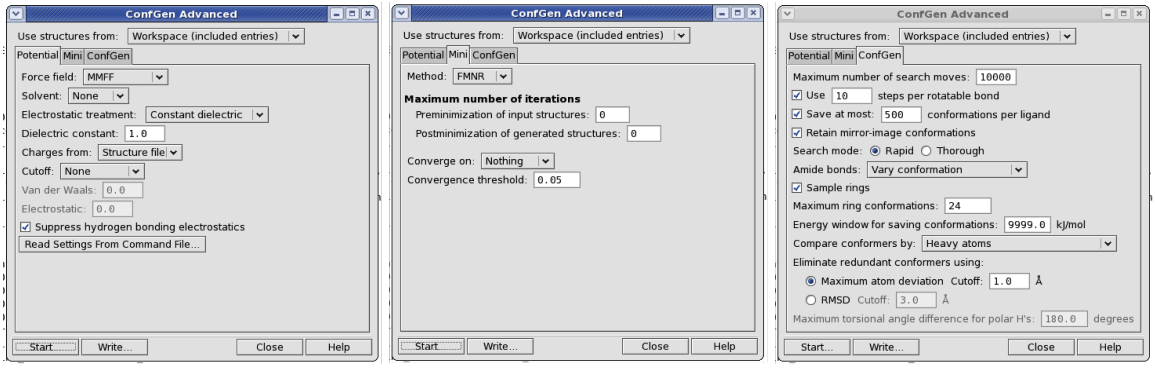


Figure 4.10 Screen captures of the parameter setting of the ConfGen utility in Schrodinger Maestro used to generate ligand conformational ensembles. Each panel shows the parameters for each of 3 different tabs from the configuration window.

Unexpected results and debate

Early results from the first design protocol yielded unexpectedly high sequence recovery results of near 100%. Closer inspection of the output structures revealed that the large majority of the 30 protein-ligand complexes recovered 100% of the correct sequence, while a handful of the protein complexes experienced 60-90% sequence recovery, thus bring down the overall average to approximately 97%. This high sequence recovery was judged unrealistic and was thought must be due to a bug in the ROSETTA code where native rotameric information from the input .pdb file was being passed inappropriately to the design algorithm. After much effort however, it was determined that in fact no sequence or rotamer information was being passed to the docking algorithm. Instead, if the ligand translation and rotation perturbations were held to zero, the majority of the complexes would exhibit perfect sequence recovery. This is a startling result that far exceeds the percent recovery of prior work both from my own previous study and anecdotal studies of other members of the ROSETTA community. Furthermore, a structural superimposition of the output and input structures revealed that there was near zero RMSD between the native relaxed complex and the designed complexes.

After much discussion and debate, a consensus was reached that the cause of the abnormally high sequence recovery results is a combination of the unprecedentedly rigorous relaxation protocol these protein-ligand complex were subjected to prior to design, and a coincident improvement in the accuracy of the ROSETTA energy functions relating to the backbone-dependent rotamer library used in both

relaxation and design of the complexes. In effect, the complexes were relaxed so extensively that all of the backbone phi/psi angles and sidechain conformations within the protein came to take on a value exactly or near exactly represented in the backbone-dependent rotamer library. Thus, even when all native sidechain information is removed from a given amino acid position, the remaining backbone geometry is sufficient to favor a single rotamer, and therefore amino acid identity, above all others in the library. It can alternatively be stated that use of the backbone-dependent rotamer libraries in ROSETTA imparts a strong "memory" to the backbone configuration such that, following extensive minimizations, backbone geometry alone contains sufficient information to recover the correct specific rotamer.

Further experiments were conducted allowing the ligand to rotate up to 15 degrees during design and separate experiments using the non-relaxed .pdb files of the 30 protein-ligand complexes. Consistent with the "backbone memory" hypothesis, perturbing the interface by allowing the ligand to rotate produced an average sequence recovery of 86%, while using the non-relaxed .pdb complex files resulted in an average sequence recovery of 27%.

There remains debate over whether these results constitutes a significant scientific finding or are merely an insubstantial artifact and flaw in the ROSETTA knowledge-based energy functions dealing with backbone-dependent rotamer libraries.

Further experiments involving the 30 diverse LPDB complexes

Once a satisfactory explanation and understanding has been found for the unusually high sequence recovery results, experiments examining the remaining two protocols encompassing docking and combined dock/design may be resumed. These should yield important information on the efficacy and utility of other ROSETTA functionality.

However, these protocols can be expanded still further to address interesting questions and capabilities in that have yet to be examined. For example, adding backbone flexibility to the flexible-ligand dock/design protocol would go even further towards replicating a "real world" interface design application. Also, further examination into the promiscuous binding of multiple ligands by single proteins might help elucidate sequence and structure mechanisms of multi-specific binding. Using proteins from the LPDB, for example HIV protease, possessing multiple deposited high-resolution protein-ligand holostructures as well as experimental binding data could constitute a significant and important project. Design of protein-ligand interface using individual ligands might be predicted to result in divergent sequences, while the use of a simultaneous ensemble of all known ligands might be predicted to recapitulate the native sequence more faithfully, thus confirming the native protein sequence to be optimized for binding multiple ligands. Subsequent laboratory expression and assay of the protein sequences optimized for binding individual ligands and compared the affinities of wild-type multi-ligand binding proteins would make for a significant investigation.

CHAPTER V*

DISCUSSION AND LESSONS LEARNED

Summary of research

The ROSETTA protein design program was use to design three complementary protein scaffolds to bind D-peptide ligands. The chosen scaffold proteins each represented a distinct ligand-binding mode, as well as degree of functional flexibility and computational complexity. The model ligand interface system was the D-ala and D-lac peptide targets of the glycopeptide antibiotic vancomycin.

Multiple, iterative rounds of ROSETTA design computation were performed on each scaffold, sampling hundreds of millions of model ligand interfaces *in silico*. Between 5 and 12 of the best scoring PDZ, TPR and 1m4w designed proteins were produced in the laboratory and assayed for binding to their target ligand using multiple assays. The PDZ domains were found to be largely insoluble and unamenable to assay, and were thus excluded from the remainder of the study. Both the TPR and 1m4w designs demonstrated no detectable, to low affinity binding to their intended ligands. Due to the nature and goal of the study, this was considered a failure of the computational design process.

* Portions of Chapter V have been excerpted from Morin, A. et al., 2011. Computational design of protein-ligand interfaces: potential in therapeutic development. *Trends in biotechnology*. Additional material from the Dissertation proposal of Morin 2007.

To help understand the lack of successful interface design, structural characterization was pursued of both the TPR and 1m4w designs. Four high-resolution X-ray structures were obtained of distinct 1m4w mutants. Examination of the determined structures revealed that although ROSETTA had accurately predicted the fine-scale structure of the protein-ligand interfaces, this accuracy did not translate into high-affinity binding.

Due to these experimental results, my dissertation research was shifted to exclusively computation, toward a more detailed focus on examining ROSETTA's abilities and deficiencies in predicting native protein-ligand interfaces. A diverse set of native small-ligand binding protein structures were culled from the LPDB. This protein set was analyzed to assess the properties of native protein-ligand interfaces at the sequence level. ROSETTA docking and design protocols replicating the prior interface design work were devised carried out, and results compared the native set to evaluate ROSETTA's sequence-recovery objective function in the design of protein-ligand interfaces. Although interesting findings regarding ROSETTA's propensities in recapitulating ligand interfaces and the strong backbone-dependence of rotamer libraries have been encountered, a full evaluation of these and further necessary experiments await completion at a later date.

Interface design capabilities are just beginning to reflect modern ligand binding paradigms

The first accepted theory for the physical basis of ligand binding, offered in 1894 by Emil Fischer, is known as the lock-and-key principle. It posits that specificity and affinity are the result of preexisting, rigid shape and chemical complementarity between the ligand and receptor. As knowledge of protein dynamics and kinetics continued to grow, D. E. Koshland, Jr. extended the lock-and-key paradigm in 1958 with a new theory of ligand binding known as the induced-fit model. This model proposes that both ligand and receptor are flexible and can mutually induce shape and chemical complementarity. The induced-fit theory was able to help explain many newly recognized phenomenon such as cooperativity, regulation and aspects of specificity. As our understanding of ligand binding continued to evolve, the conformational-selection model of ligand binding was proposed, first in 1965 by Monod et al. (171) relating to protein allostery, then was generalized in 2000 by Kumar et al. (172) to ligand-receptor interactions. The conformational-selection model proposes that receptors exist in an ensemble of conformations and that a ligand binding will select energetically preferred conformations, thereby altering ensemble conformation equilibrium.

Since the introduction of these models, and as our understanding of ligand binding continues to mature, it has been generally accepted that all three are in fact correct, and that one or all of these models may apply to a given protein-ligand system. Indeed, a hybrid fourth model, combining the induced fit and conformational selection paradigms has been recognized where pre-binding interactions of a ligand

with a favored conformer of an ensemble induces conformational changes in both the ligand and protein, which then form the low energy interface (173). Although each of these models of ligand-protein interaction has shown to be involved in ligand binding, lively scientific debate continues as to the relative importance and contributions of each model (174-177). Evidence has even shown that a given protein-ligand system may process from one binding paradigm to another depending on conditions, such as concentration, temperature, etc. (178). For the computational design of protein-ligand interfaces to be successful, it must be able to account for the complex physical nature of ligand-receptor interactions as described by these overlapping and interrelated binding paradigms.

Traditionally, protein design has relied on methods approximating the lock-and-key model of ligand interfaces. Protein backbones were held fixed, while only residue sidechains were allowed to change conformation. In some cases, small phi/psi angle adjustments were allowed on the protein backbone during gradient minimization of the protein-ligand complex to accommodate slight changes in protein conformation. However, the magnitude of these phi/psi changes are too small to sample significant backbone conformational space, and thus these methods were only able to model rigid protein-ligand interfaces. For example, Reina et al. was able to redesign the specificity of class I PDZ protein PSD-95 to bind naturally incompatible class II ligands by performing correlated mutations on both protein and ligand using the backbone atom positions found in the liganded co-crystal structure (Figure 5.1) (179). However, they were unable to achieve a specificity switch from class I to class III PDZ ligands using fixed backbone design due to greater structural diversity in the

class III protein and ligand conformations than could be modeled using a rigid backbone protocol.

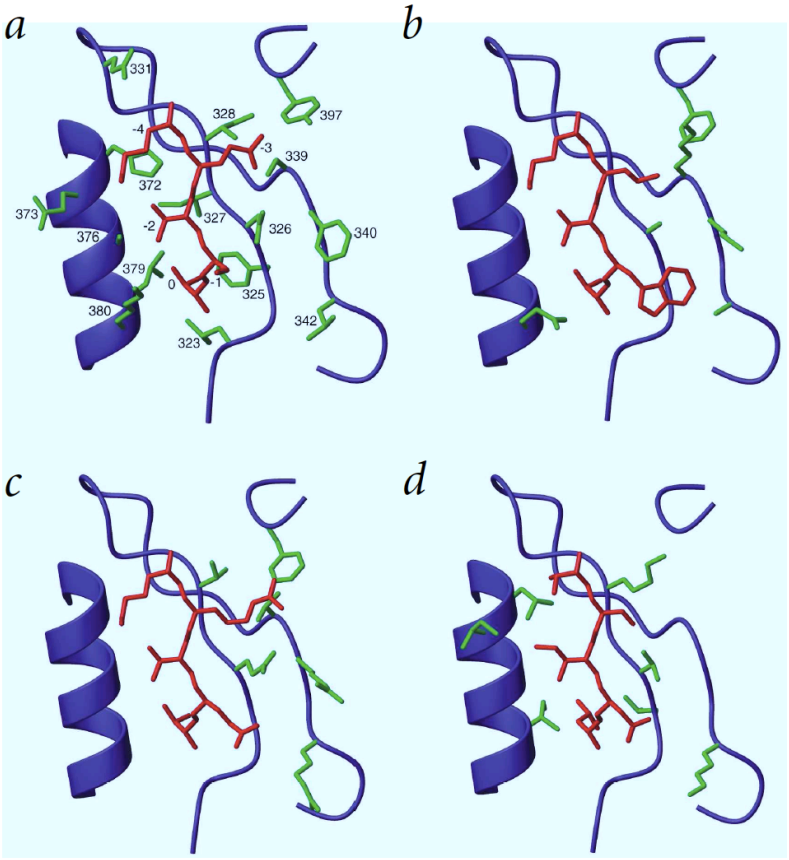


Figure 5.1 Ribbon diagrams of the PDZ–ligand complexes. a,PSD-95 PDZ3 (PDZ-wt) and its natural ligand (KQTSV). b, PDZ-hyd–hyd peptide (KITWV). c, PDZ-pol–pol peptide (KRTWV). d, PDZ-Eg5–Eg5 peptide (TSINL). The residues of the ligand (red), as well as those selected for mutagenesis (green), are numbered in (a). Only the ligand (red) and the mutations suggested by Perla (green) are shown in (b–d). e, Alignment of the target sequences discussed in this work, including the Class I and Class II consensus, Eg5, hyd, pol and the two sequences known to bind the wild type domain, CRIPT (referred to in this work as wt peptide) and NL.

With increases in computational power, new methodologies were introduced attempting to approximate the conformational selection binding model through the use of pre-calculated ensembles of protein backbone and ligand conformations.

Ensembles were often derived from experimental NMR data or calculated using molecular dynamics programs or other computational methods. While these ensemble methods are able to approximate large conformational changes in the protein and ligand, they nonetheless rely on a rigid backbone during ligand docking and sidechain repacking, and are thus unable to account for smaller scale conformational changes which occur during design of the binding interface (180).

More recently, the introduction of backbone flexibility and kinematic loop sampling methodology has allowed computational design of protein-ligand interfaces to match the induced fit binding paradigm more closely (181)(182). Medium-scale sampling of "backrub" phi/psi angle changes in the protein backbone allows efficient approximation of protein conformational changes (183; 184), while kinematic loop modeling (185; 186) allows extra sampling of flexible loops which often explore more conformational space than the protein core.

Ligands can also be modeled flexibly during design calculations. Peptide ligands composed of standard amino acids can be modeled using techniques identical to the protein backbone. However, small-molecule or non-standard amino acid ligands must be specially parameterized to assure that low energy configurations are preferentially sampled. A recent technique for modeling ligand flexibility is to generate fragment based ligand rotamers using the Cambridge Structural Database (CSD) of small molecules (145). Statistical potentials are generated similar to those for the protein design energy functions and used to populate a predefined ligand rotamer library. The computational search algorithm then adds the ligand rotamers

to the amino acid rotamers when sampling the multi-dimensional search space during design (162). Another approach to modeling ligand flexibility was used by Sood et al. They combined loop-modeling and protein design techniques to achieve modest binding affinity increases in the design of N- and C-terminal extensions to natural peptide ligands bound to proteins (91).

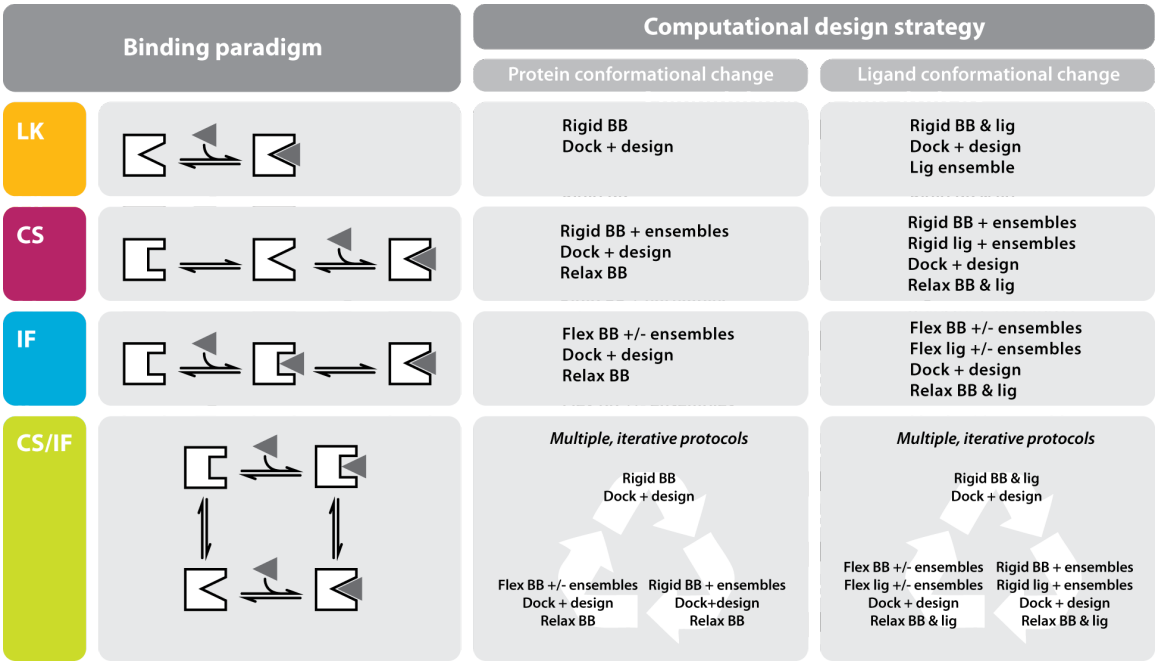


Figure 5.2 Ligand binding paradigms and their corresponding computational design strategies. The four binding paradigms outlined in the text are listed in the colored badges: LK= lock-and-key (orange); CS= conformational selection (purple); IF= induced fit (blue); CS/IF= hybrid conformational selection/induced fit (green). Immediately to the right are schematic representations of the corresponding binding modes. The two rightmost columns are examples of possible computational design strategies for each binding mode in cases of protein conformational change alone (left) and protein plus ligand conformational change (right). Prior interface design methods were limited to addressing the LK binding mode where no protein conformational change occurs, with some limited ability to explore CS binding modes with the use of structural ensembles and gradient relaxation of protein and ligand. Recent design techniques have allowed the potential exploration of IF binding modes through the combination of structural ensembles and protein/ligand flexibility functionality. Current work is investigating potential methods for applications to the hybrid CS/IF modes and requires iteration through multiple combined protocols.

Using methods such as these the computational design of ligand-protein interfaces has begun to simulate the three fundamental physical models of protein-ligand interaction. By combining the above techniques, the hybrid conformational selection/induced fit binding paradigm may also be approximated (Figure 5.2). However, while these techniques can approximate the conformational change observed in the induced fit and conformational selection ligand binding paradigms, they are unable to model the dynamics of the protein or ligand separately or in complex. (See below for further discussion of dynamics in interface design.)

Localized water molecules in the ligand interface are critical to successful design

A critical component to consider in the design of ligand binding is water at the interface (187). Water molecules have significant effects on the change in free energy of ligand binding and contribute significantly to both its entropic and enthalpic energy components (148). The effects of solvation on polar and hydrophobic sidechains and atoms in both the protein and ligand, the displacement of coordinated water molecules and the formation and breaking of bridging hydrogen bonds can each, individually, be sufficient to overwhelm the energetics of ligand binding (188). Studies by Teyra & Pisabarro on a representative set of protein-ligand complexes from the PDB have shown that ~40% of interface forming residues interact through bridging water molecules and that ~15% interact solely by means of water-mediated hydrogen bonds (189). Furthermore, even well

classified ligand interfaces such as the recognition of proline-rich sequences by SH3 domains that have traditionally been regarded as being hydrophobically driven, in fact form a dual-mode interface complemented by a network of bridging water-mediated hydrogen bonds (Figure 5.3) (190). It is therefore critical that the effects of water at the protein-ligand interface be included in the evaluation of interface designs.

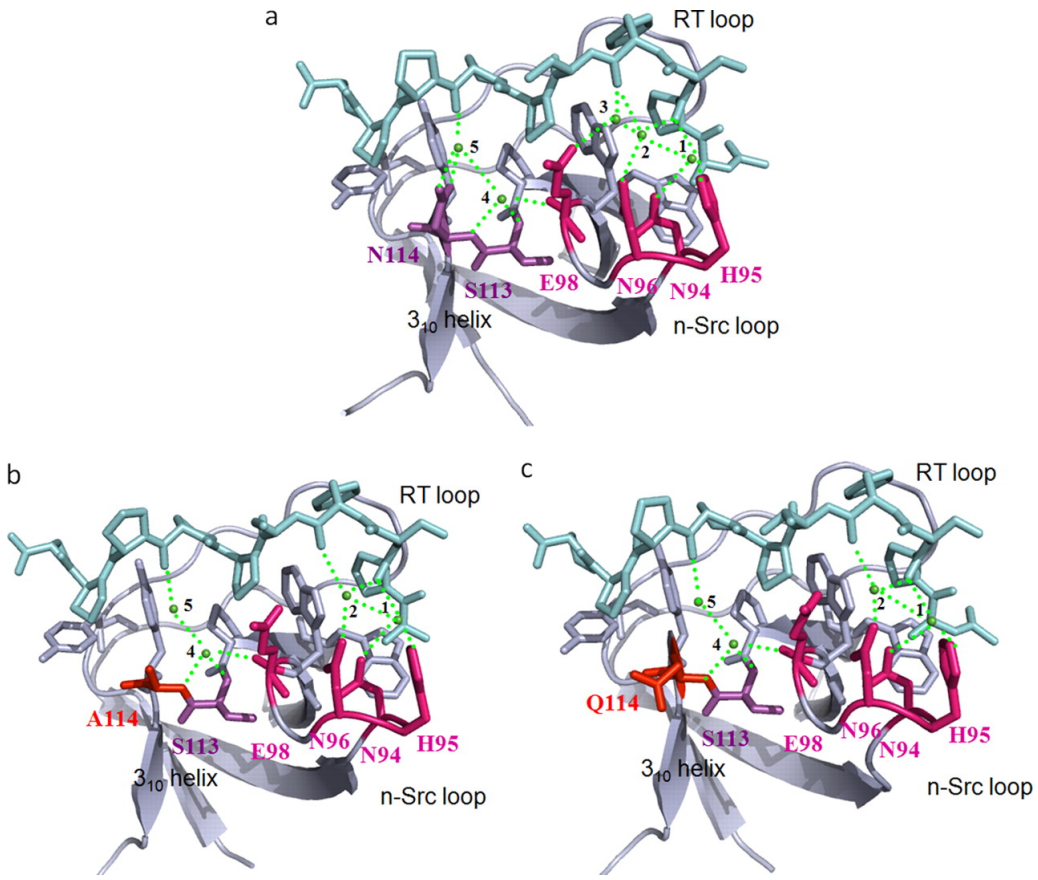


Figure 5.3 Water at the binding interface. Water molecules at the Abl-SH3/p41 binding interface for WT (a), N114A (b), and N114Q (c). The structure of the Abl-SH3domainisshownin a gray schematic. Residues defining the canonical binding site for polyproline recognition are shown as gray sticks. The structure of the p41 peptide is shown as cyan sticks. Fully buried water molecules at the binding interface are shown as green spheres (sites occupied by water molecules are labeled from 1 to 5). Peripheral water-coordinating residues in the 310 and n-Src regions are shown as purple and dark pink sticks, respectively. Water-mediated hydrogen bonds are depicted as dotted green lines. (190)

The most common method for modeling the effects of water at the protein-ligand interface during design is the use of an implicit solvation model that typically includes solvent accessibility and Gaussian-shaped solvent exclusion terms to compute bulk solvent effects. These models have the advantage of being quick to compute and are generally measured over an intact protein-ligand system (94). However, in the design of ligand interfaces, where the steric clash of a single coordinated water molecule or the electrostatic repulsion of one unaccounted for electron pair can potentially disrupt ligand binding, the average accuracy of fast implicit solvation calculations may not always be sufficient (191). In many cases, the explicit modeling of water molecules in the interface will be necessary.

A partial solution to the need for explicit solvation of the protein-ligand interface is the inclusion of solvated rotamers (192). Solvated rotamer libraries are constructed in a fashion similar to normal roatmer libraries (see above), but include the most common positions for coordinated water atoms as observed in the PDB, along with residue sidechain atoms (Figure 5.4). These solvated rotamer libraries can then be included with the normal rotamer libraries during sequence-conformation sampling of the residues comprising the interface. Solvated rotamers are limited by the fixed orientation of the water in relation to the coordinating sidechain atoms, and the concerted coordination of a single water by more than one residue cannot be accommodated. Additionally, expansion of the rotamer libraries dictates a corresponding expansion of the design search space and required computational resources (94).

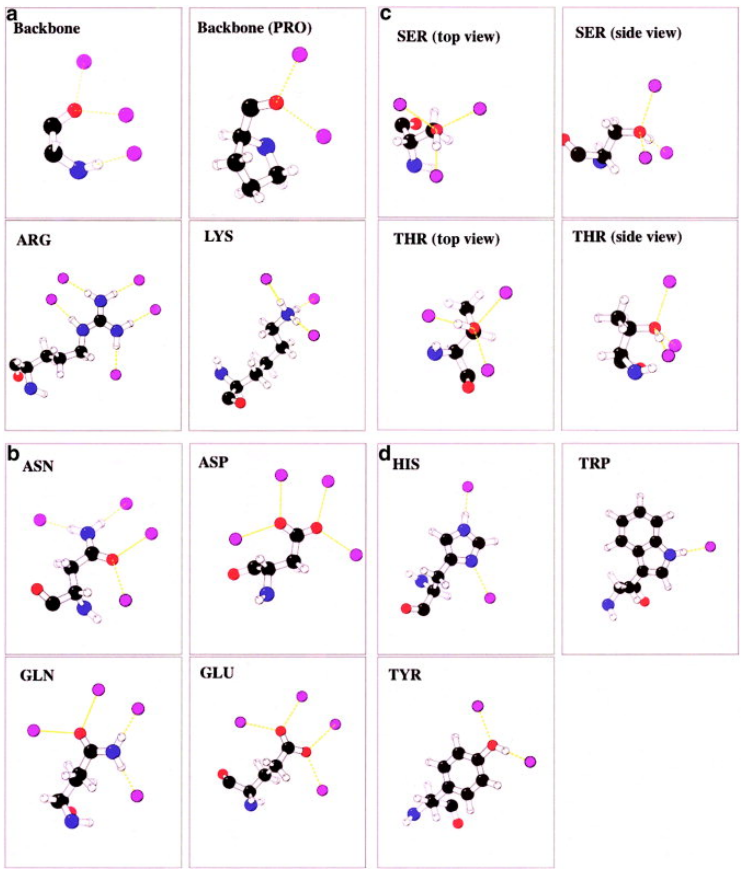


Figure 5.4 Water placement in solvated rotamers. Protein atoms are colored using the CPK convention (nitrogen, blue; oxygen, red; carbon, black; hydrogen, white). Oxygen atoms in water molecules are colored in purple. Hydrogen bonds are indicated by yellow dashed lines. For serine and threonine, two views are shown. These figures were prepared with Molscript. (94).

More recent progress in computational protein design methods has enabled the explicit modeling of limited numbers of water molecules as independent ligands within the interface (Lemmon et al., in preparation). Using these methods, each water is allowed compete degrees of freedom within a defied interface area and can potentially more accurately predict direct hydrogen-bonding, electrostatic and hydrophobic effects of individual water molecules. A disadvantage of these methods is the significantly increased computational complexity accompanying the addition

of each explicit water. This currently limits the practical use of the technique to the inclusion of only a handful of explicit water molecules. Future increases in raw computational power and optimization of search and scoring algorithms should soon allow the modeling of sufficient numbers of water to hydrate most small to medium sized ligand binding interfaces during design.

Expanding functionality and applications of interface design

While the addition of limited explicit modeling of water in the ligand binding interface denotes a significant expansion in interface design capability, other design functionality has also recently been developed which may potentially expand the rage of medically relevant targets and applications to which interface design methods may applied.

Ashworth et al. has demonstrated the ability to redesign the homing specificity and catalytic cleavage functions of a DNA endonuclease (Figure 5.5) (193-195). The inclusion of DNA and RNA as designable ligand interface targets holds great promise for therapeutic applications.

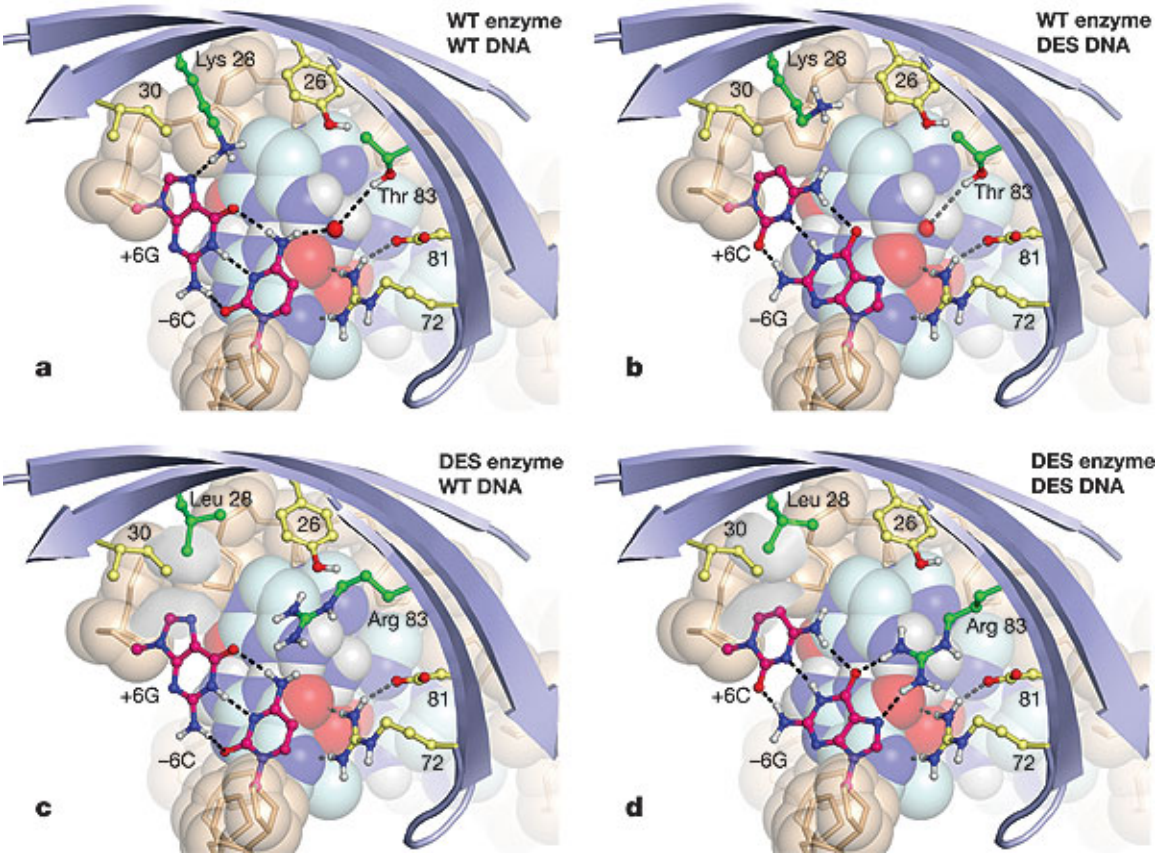


Figure 5.5 Comparison of the predicted interactions in cognate and non-cognate binding complexes, illustrating the designed specificity switch. a, Wild-type I-MsoI, -6CG (wild type). A water molecule present in the original structure16 is shown. b, Wild-type I-MsoI, -6GC. c, I-MsoI-K28L/T83R, -6CG. d, I-MsoI-K28L/T83R, -6GC. In parts c and d, the van der Waals surfaces of Leu 28 and +6C are shown in grey. Figures were generated using the molecular graphics program PyMOL (Delano Scientific). WT, wild type; DES, designed; blue strands, protein backbone; beige spheres and sticks, DNA backbone; other spheres, constant nucleotides; dashed lines, hydrogen bonds. (193-195).

Negative and multi-state design strategies have also proven successful at creating specificity and multivalency in protein-protein interfaces, and may be similarly useful at ligand interface applications. Humphris & Kortemme computationally designed multivalent protein-protein interfaces (156) using multi-state design protocols, whereas Bolon et al. found that inclusion of negative design strategies

was necessary for establishing specificity at a dimer interface (196). The integration of similar strategies into ligand interface design protocols could enable analogous functionality.

Computational protocols comparable to those that allow multiple explicit waters to be modeled in the interface can also allow the simultaneous modeling and design of multiple ligand types as well. Yosef et al. achieved a 900-fold increase in binding specificity when re-designed calmodulin interfaces to large alpha-helical ligands in the presence of Ca2+ co-factors necessary for interface formation (Figure 5.6)(197). The ability to perform interface design while including biologically important molecules such as inorganic cofactors like metal ions and clusters, organic cofactors such as NADH or ATP, or other combinations of small-molecules and ligand could greatly expand the range of targets computational methods can address.

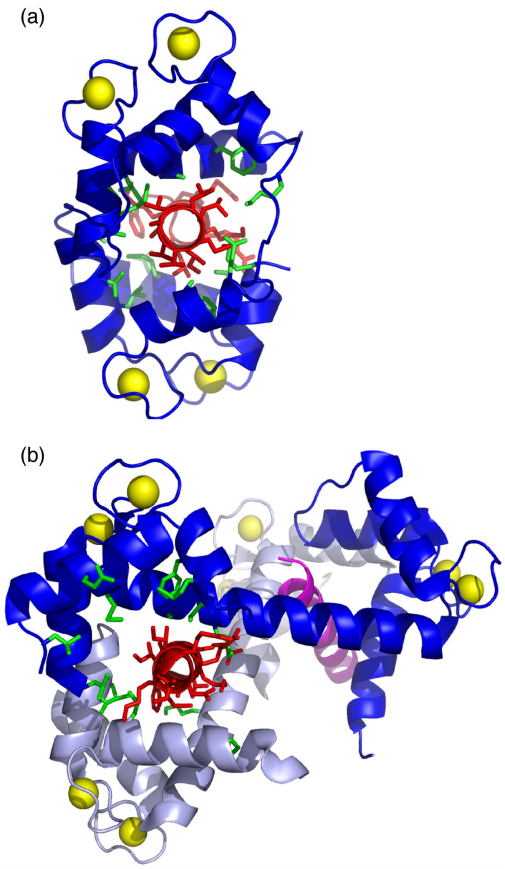


Figure 5.6 X-ray structures of CaM in complex with the two targets. (a) CaM–CaMKIIp complex (PDB code: 1CM1). CaM is shown in blue, and CaMKIIp is shown in red. CaM residues included in the optimization of the CaM–CaMKIIp binding interface are shown in green. Calcium ions are indicated as yellow spheres. (b) CaM fused to CaNp showing two CaM molecules (blue and light blue) and two CaNp molecules (red and purple) (PDB code: 2F2O). The same CaM residues as in (a) are shown in green. (Yosef et al. 2009).

Is *de novo* enzyme design easier?

The most succinct answer to that provocative question is: certainly not. The work performed in 2008 in the laboratory of David Baker at the University of Washington on the *de novo* design of novel enzymes (11)(12) was both innovative and a powerful demonstration of the potential of computational protein design. Very little, if any, of the work that went into these groundbreaking studies could be called easy.

Yet the question remains: why has progress in the computational design of the seemingly more basic functionality of ligand binding lagged behind accomplishments in the *in silico* creation of enzymes not previously observed in nature? After all, binding a ligand would seem to operate through less complex and sensitive mechanisms than enzymatic catalysis. Although the answer to that question is undoubtedly a complex one, clues to understanding the different outcomes in each of the sub-fields lay in both the fundamental physical processes of each protein function, and the limitations of the computational tools used to model them.

Proteins are inherently dynamic. At the most basic level, all protein functions depend on dynamic processes (198). Continued advancement in researchers understanding of the ways in which protein dynamics effects the formation of protein-ligand interfaces have led to development of more a sophisticated understanding of the mechanisms of ligand binding. Similarly, an ongoing and lively debate amongst scientists in the biophysics, structural and computational biology fields regarding the roles of protein dynamics in enzyme catalysis may help to identify at least one reason ligand interface design may be a larger challenge than enzyme design.

One side in this debate, relying on newly developed computational and biophysical methods, argues that the chemical step in catalysis is insensitive to protein dynamics and is instead almost entirely a function of active site geometry. This group further conclude that the major role of protein dynamics in enzyme function

is in determining the kinetics of the enzyme (160; 161; 199). Their opponents in the debate argue for the central importance of dynamics in both enzyme catalysis and kinetics to enzyme function (200-202).

While we await the resolution of this debate, results from the computational protein design field may offer some intriguing insights. Indeed, if the first group in the debate is correct, the chemical step in catalysis is less sensitive to protein dynamics and is strictly a function of geometry (structure), this may offer some explanation as to why enzyme design succeeded where interface design has yet to. The computational tools commonly used for protein design, like those employed in the Baker lab enzyme design studies, have proven to be quite adept at accurately predicting protein structure and conformation down to the atomic level (203), but are largely unable to model protein dynamics on the timescales involved in many functional processes – such as ligand binding or catalytic active-site motion. If dynamics subordinate to structure (geometry) when designing enzymes, but not when creating ligand-interfaces, this could be one reason for the lack of success in computational interface design. Enzyme design is playing to the strengths of the computational methods, whereas interface design is dependent on its weaknesses. Evidence for this possibility might be seen in the results of the Baker enzyme design efforts, where although a high proportion of the tested designed enzymes displayed catalytic activity, none of the enzymes, even after subsequent rounds of directed evolution, were shown to possess more than a low kinetic efficiency.

Thus, it is quite possible that although ligand-interface design seeks to reproduce a more basic function than catalysis, it may in the end have the more difficult task, requiring the further development of new and more advanced computational methods.

Elephant in the room: the dynamic nature of proteins

Although many new capabilities have recently been developed to aid in the computational design of protein-ligand interfaces, there remain significant outstanding questions in the field that have yet to be effectively addressed, and which may require carful consideration when performing interface design calculations. Chief among these is the role of protein dynamics in ligand binding.

Proteins are intrinsically dynamic macromolecules possessing a range of vibrational modes. These dynamic modes can operate on both fast (femtosecond to microsecond) and slow (millisecond to second) time-scales and with motional amplitudes ranging from sub bond-length vibrations to concerted motions of large protein domains. Recent advances in our understanding of protein dynamics in biological processes has lead to the extension of the classic structure-function paradigm to include dynamics as crucial to a complete description of protein function (198). Indeed, an April 2009 special issue of the journal Science (Science, April 10, 2009) was dedicated to the rapidly evolving understanding of dynamics in protein biology, cellular function and drug development.

As the role of dynamics in protein function has become better appreciated, the role of dynamics in protein-ligand binding has also become more clear, and more clearly important. Time-dependent intra- and inter-protein motions directly affect every element of a protein-ligand interface, from the energetics of ligand interaction (204) to the conformational populations available for binding (205). While significant debate remains on the details of how dynamics affect the formation of protein-ligand interfaces (206), it is nonetheless clear that dynamics is a major, and potentially overriding factor in protein-ligand binding. It is therefore a cause for concern that current protein design methods are not able to fully model and account for protein dynamics.

Recent interface design work performed in our lab offers a good example of the unexpected and potentially confounding role of protein dynamics in the design of ligand binding (Morin et al., 2010). Attempts to *de novo* design a ligand interface using a thermophilic enzyme failed to produce the computationally predicted high affinity binding in experimental binding assays. Subsequent high resolution X-ray structure determination of four of the designed proteins revealed that although the predicted structures of the ligand interfaces were highly accurate, this structural accuracy did not translate into binding affinity. Crystallographic analysis revealed that in all of the designed proteins, five or fewer of the designed mutations in a putative binding pocket lead to a fundamental change in global dynamics of the protein (Figure 3.9). This alteration of the protein dynamics was sufficient to eliminate high affinity ligand binding. This degree of sensitivity to design mutations of a thermophillic protein's inherent dynamics was unanticipated. An additional

instructional finding from this study was the clear distinction between conformational change and protein dynamics. Although our design algorithms were able to accurately predict an induced conformational change in the designed protein, these algorithms were unable to discern the resulting change in dynamics and consequent effects on ligand affinity. Thus, we see that although protein conformational changes can be effectively modeled, they comprise only a subset of the dynamic information necessary to accurately describe and predict ligand binding.

Similar to protein dynamics, the dynamic properties of ligands are also an important consideration when performing interface design. Not only can ligand dynamic properties influence protein-receptor structure vis-à-vis the induced-fit binding model, but dynamics within the ligand can potentially dominate changes in free energies upon binding used to discriminate successful models during computational design (205), and thus may also require careful consideration in interface design efforts.

Currently, no generalized CPD method is able to effectively compute protein dynamic information. These types of computations remain solely in the domain of the much more computationally expensive molecular and quantum mechanical computation methods, which have not yet become practical for *de novo* design applications (106). Until a more comprehensive modeling of protein dynamics can be achieved using CPD techniques, many functionally important design

considerations relating to protein dynamics will remain outside of the abilities of computational interface design.

Careful scaffold selection will continue to be crucial to successful interface design efforts

The usefulness and functionality of *de novo* computational protein-ligand interface design is still developing, and the lack of protein dynamic information in the design process will continue to necessitate careful consideration of the protein systems chosen to undergo design.

De novo interface design is normally performed on a natural, naïve scaffold protein. *De novo* scaffold proteins are typically selected for favorable characteristics such as molecular weight and stability, laboratory qualities such as expression system, yield and published production protocols, and end-application traits such as percent human sequence content and *in vivo* characteristics (19). Due to the current limitations of *de novo* computational interface design, additional considerations when selecting not only a scaffold protein, but also ligand target will be critical to successful design outcomes.

Due to the inability of current design methods to adequately model dynamic processes, a rigorous vetting of potential interface design scaffolds for dynamics is advisable. Experimental tools such as NMR (207), mass spectrometry (208) and x-ray scattering and diffraction techniques (209-211) in combination with computational tools such as molecular dynamics (212; 213) can help provide insight

into the dynamic properties of potential design scaffolds. Removing from consideration scaffolds possessing large-scale dynamic modes at the proposed ligand interface site, extensive conformational dynamics or other potentially disruptive dynamic processes opaque to computational design methods will help prevent unexpected design difficulties. This requirement raises difficulties of its own however, as the experimental information on a protein scaffolds dynamic properties may not be readily available, and when available, is often non-trivial to interpret. Therefore, it is likely that in the near-term, the best strategy for choosing a design scaffold protein remains to select one possessing the least amount of inherent dynamics. This aim has been greatly aided by the recent establishment of database repositories for protein dynamic information such as Dynameomics (214), DynDom (215) and Molmovdb.org (216).

De novo interface design in drug development

De novo interface design holds great potential for the development of new and novel therapies and modes of therapeutic action. The ability to reliably design ligand binding functionality against any target using a chosen protein scaffold would enable applications in all areas of medicine and greatly expand an already burgeoning protein therapeutic market - a market that his so far achieved success using only *post hoc* computational design techniques.

Though the full potential of *de novo* ligand interface design has not yet been fulfilled, innovations and techniques in related areas of computational protein design, newly

applied and adapted to ligand interfaces, portend a coming renaissance in the rational and rapid computational design protein drugs.

Yet there remain significant challenges to be overcome before reliable and repeatable methods for the *de novo* design of protein-ligand interface can be achieved. Foremost among these challenges is the open question of the role of protein dynamics in ligand binding and how to effectively and efficiently model it during design. Thus far, the modeling of such dynamics remains solely the purview of computationally intensive molecular mechanics simulations; methods that are currently, and for the foreseeable future, too computationally expensive to sample the vast sequence-structure search space on a practical timescale.

Similar difficulties apply to the inclusion of explicit water molecules in the design of ligand interfaces. While the role and importance of individual waters in ligand binding is generally appreciated and understood, the ability for computational design methods to model them remains limited.

In the near-term, it appears the best way to avoid the negative effects of these open and unsolved questions in *de novo* interface design is through the careful selection of design target and scaffold. Choosing to apply the current techniques to targets and proteins which do not possess significant dynamic properties or extensive bridging water-mediated H-bond networks in the interface may offer the best chance for immediate success in interface design.

Though solving these outstanding challenges appears to be a prerequisite to establish *de novo* interface design as a standard tool in drug design and

development, once these outstanding challenges have been addressed, the *de novo* computational design of protein interfaces to target ligands has the potential to radically alter the way in which therapeutic protein drugs are created.

CHAPTER VI

FUTURE DIRECTIONS

The results and findings of my dissertation work indicate a need for further development of Rosetta design functionality to accommodate information on the dynamic nature of proteins (203). Additional improvements to the accuracy of the hydrogen bonding scoring function may also be desirable.

Beyond the efforts to improve Rosetta's hydrogen bonding terms already underway, two potential aims, one medium and one long-term in scope and implementation would help accomplish these improvements. The first and medium term aim would be the identification and validation of design protein scaffold set of known dynamic and other protein design qualities. The second, longer-term aim would be the development and incorporation of knowledge-based protein dynamics scoring function into Rosetta.

While these aims differ in scope and ultimate level of utility and usefulness to the broader protein design community, the fundamental knowledge and information developed over the course of their completion are broadly complimentary. This suggests that an incremental approach, beginning with the first aim, and upon completion, proceeding to the second would be advisable.

Identification and validation of design protein scaffold set

The development of a validated design protein scaffold set possessing known dynamics and other physical, structural and biochemical properties suitable for a given design objective function (e.g. ligand binding, catalysis, protein-protein interaction, etc.) requires the understanding of several phenomena fundamental to protein function (213), as well as the experimental means and techniques used to investigate and classify them. The phenomena include the basic physical, thermodynamic and structural-dynamic basis for protein-ligand, protein-substrate or protein-protein interaction, concordant to the desired design objective function. While the experimental techniques used to elucidate these and other properties of interest span a wide range of physical, chemical and technological disciplines(146).

Additional consideration and evaluation of properties relating to protein production, assay, characterization and application would also require investigation. For example, in addition to obtaining the fullest understanding possible of the dynamic properties of a candidate design scaffold protein, one would also have to classify and select each based on suitability to a specific end-application – such as human therapeutic, industrial process, scientific reagent, etc. – as well as the techniques used to assess and characterize the success of the design process.

Beyond the necessity to develop a phenomenistic understanding of the fundamental processes involved in an objective protein function, a thorough understanding of the experimental techniques and resulting data used in classifying, describing and comprehending protein dynamics would be required. Theoretical and practical

knowledge of NMR (207), x-ray scattering and diffraction (210; 211), fluorescence/photonic and computational molecular and quantum mechanical methods and others (213) would be necessary to understand their significance and consequence in the protein design process. Gaining this knowledge in itself would be a non-trivial process.

Once a thorough understanding these two critical elements – a theoretical basis of dynamic protein function and the experimental techniques used to elucidate them – have been acquired, they could then be applied to identifying proteins suitable for manipulation through protein design.

The process of selecting and parsing candidate protein design scaffold would rely primarily on the databases that currently exist as repositories for the structural, functional and dynamic information – databases for dynamic information (214-216), structural and biochemical information (170; 217), binding and thermodynamic data (163; 218), etc.

For example, one might first cull the PDB for proteins matching the basic biochemical and laboratory properties needed to facilitate production and end-application of the proteins (e.g. molecular weight, origin species, expression strain, number of peptide chains, etc.). This might be expected to yield several hundreds to thousands of matching proteins. These proteins might then be categorized according to native protein function for classification as either de novo or re-design utilization, before being cross-referenced and parsed against other databases containing experimental data on binding thermodynamics or affinity, NMR

experiments, solution dynamics or molecular dynamics calculation information – each of which may reside in separate databases. After the candidate proteins have been parsed and identified, additional experiments to fill in missing information or confirm or resolve conflicting data may be necessary.

At the end of this vetting process, one would expect to have identified and validated no more than a handful of proteins suitable for design applications, possibly no more than 3-5 protein scaffolds initially, with more added over time. However, due to the substantial knowledge and expertise necessary in the execution of this project and the broad and disparate nature of the information involved, this project may very well constitute several years of work of an advanced graduate, or more likely, post-doctoral level. Nevertheless, the benefits and impact of the work would be substantial and could be of great use to the entire protein design field.

Development and incorporation of knowledge-based protein dynamics scoring function into Rosetta

A important, and likely crucial, longer-term goal would be the incorporation of a knowledge-based protein dynamics scoring function into Rosetta protein design. This would endow Rosetta with some ability to predict and approximate the dynamic modes and functions of proteins critical to design applications, whilst maintaining Rosetta's advantage in computational efficiency over the more demanding physics-based computational methods such as molecular dynamics.

There are however two significant hurdles which must be overcome in the course of this endeavor.

The first hurdle applies equally to both this aim and the more modest goal mentioned above – the assembly of a protein design scaffold set. This first significant hurdle is the current lack of a standard set of experimental techniques or methods to comprehensively characterize a proteins dynamics across the applicable time-domains of protein function(198). Indeed, the experimental elucidation of the comprehensive dynamics for even a single protein is often beyond any single technique, and may require the application of several different methods, including molecular mechanical and/or quantum mechanical simulations (213).

This relates directly to the second hurdle to the implementation of knowledge-based protein dynamics scoring function in Rosetta, the general lack of protein dynamic data. It seems likely that a relatively large dataset will be required to construct a knowledge-based potential of this kind. To be useful, this dataset could require extensive dynamic characterization data on hundreds to thousands of proteins. Given that no such database currently exists, and that, as mentioned above, there is also no standard or agreed upon set of experiments for gaining such data, the prospects for developing a knowledge-based dynamics score seem increasingly long-term. However, it is conceivable that even a very “low-resolution” score, one that simply indicated that a given sequence change might lead to a significant change in the overall dynamics of the protein, would be useful in the design of protein function.

Prospects and importance

It is my opinion that the second and most difficult of these aims – the development of a knowledge-based protein dynamics scoring function – is both inevitable and necessary before Rosetta can become a robust, reliable and useful tool for protein functional design. The design of protein-ligand interfaces may rely more heavily on this ability to address the dynamic nature of proteins than other design applications such as enzyme design, but the fundamentally dynamic nature of all proteins indicate that all protein prediction and design methods would benefit from this added functionality. Due to the long-term nature of this ambitious undertaking however, the intermediate step of creating a validated protein design scaffold set may prove highly useful to rational design efforts, and would constitute a significant achievement in its own right.

APPENDIX

APPENDIX A:Development of medium-throughput ELISA assay

Aim: Develop qualitative, medium-throughput antibody-based screen to rapidly identify binding candidates

The motivation for developing an ELISA screen is to enhance the throughput of the experimental design by allowing screening of proteins directly from cell lysates. Once implemented, the screen will take less than one week to perform and can reduce the number of nonproductive designs by more than half.

The designed proteins will be screened for initial binding from whole-cell lysates using an indirect enzyme-linked immunosorbent assay (ELISA) with target peptides immobilized on 96-well plates. The biotinylated D-ala-D-ala target peptide will be added to the plate and bound to the immobilized avidin. The designed histidine tagged proteins will then be added to the wells containing bound peptide, incubated, and washed. Detection of any bound proteins will util-ize an anti-histidine 1º antibody and 2º antibody with an alkaline phosphatase fusion protein for detection by the addi-tion of nitrophenyl phosphate. The signal will be measured at 405 nm using a plate reader. To ensure specificity, con-trol wells without peptide and without designed protein will be included.

Expected Outcome and Interpretation: Development of the ELISA assay will be accomplished using published methods and commercially available reagents, and is

expected to take several months to complete. Caveats and Challenges: A potential problem with the use of the ELISA assay on cell lysates is interference by the high quantity of E. coli proteins. One solution is a rapid purification of the expression products using cobalt affinity resin prior to ELISA screening. This step can be done to quickly remove a high percentage of the contaminating proteins and will increase the chance of identifying even low-binding affinity designs.

APPENDIX B:Testing of Backscattering Interferometry binding assay

Aim: Validate and benchmark Backscattering Interferometry binding assay

Backscattering interferometry (BI) is a new technique currently undergoing development in the laboratory of Daryl Bornhop in the Vanderbilt Department of Chemistry. In its current implementation, BI can be used to detect label-free protein/ligand interaction in solution at femtomolar concentrations in either a substrate immobilized or free-solution mode74. The principal behind BI is a change in the phase of a reflected light beam compared to reference that is de-pendent upon the radius of gyration of the molecular assembly in solution. Thus, binding affinity of ligand to protein can be measured as a function of light phase change, and plotted to obtain standard hyperbolic saturation binding curves75.

The ligand/substrate systems used to validate the efficacy of the BI assay will be the same as those which have al-ready been used to test the standard binding assays listed in Aim 3. Specifically, well established literature Kd values of the antibiotic

vancomycin binding its target -L-lys-D-ala-D-ala peptide ligand9, and a wild-type TPR protein binding its native MEEVD peptide ligand76 will be replicated.

APPENDIX C: Table A.1, Crystallographic statistics for the 1m4w derived proteins

Data Collection	**1m4w_6**	**1m4w_6w20**	**1m4w_6v48**	**1m4w_6w20v48**
Wavelength, Å	1	1.5418	1.5418	1.5418
Resolution (outer shell), Å	55.30-1.28 (1.34-1.28)	38.48-1.69 (1.79-1.69)	49.01-1.70 (1.79-1.70)	55.32-1.63 (1.73-1.63)
Rmerge*, %	7.6 (53.3)	8.6 (40.2)	8.9 (29.6)	4.6 (21.1)
Mean I/sigma(I)	54.89 (3.52)	23.22 (3.63)	28.48 (3.34)	26.44 (3.01)
Completeness, %	99.8 (96.4)	99.7 (97.9)	100.0 (100.0)	88.5 (48.1)
Redundancy	9.70 (5.5)	18.78 (6.77)	21.80 (12.06)	7.53 (1.22)
Unique observations	62177 (4534)	28769 (4289)	28204 (3957)	28568 (2549)
Refinement				
Rcryst/Rfree, % †	18.07/19.37	17.62/21.62	16.40/20.38	18.42/22.63
No. protein atoms	1169	1077	1155	1157
No. solvent waters	386	438	404	366
Bond length rmsd, Å	0.030	0.026	0.028	0.013
Bond angle rmsd, °	2.235	1.952	1.954	1.274
Avg. protein B, Å2	12.476	17.679	15.363	19.194
Ramachandran plot, % ‡				
Most favored	88.3	89.5	89.0	86.3
Allowed	10.5	9.9	9.7	12.4
Generously allowed	1.2	0.6	1.3	1.2
Disallowed	0.0	0.0	0.0	0.0

Outer resolution bin statistics are given in parentheses.
*Rmerge = Shkl(Si|Ihkl,i - <Ihkl>))/Shkl,i<Ihkli>, where Ihkl,i, is the intensity of an individual measurement of the reflection with Miller indices h, k and l, and <Ihkl> is the mean intensity of that reflection.
†Rcryst = S||Fobs, hkl| - |Fcalc, hkl||/|Fobs, hkl|, where |Fobs, hkl| and |Fcalc, hkl| are the observed and calculated structure factor amplitudes. Rfree is equivalent to Rcryst but calculated with reflections (5%) omitted from the refinement process.
‡Calculated with the program PROCHECK

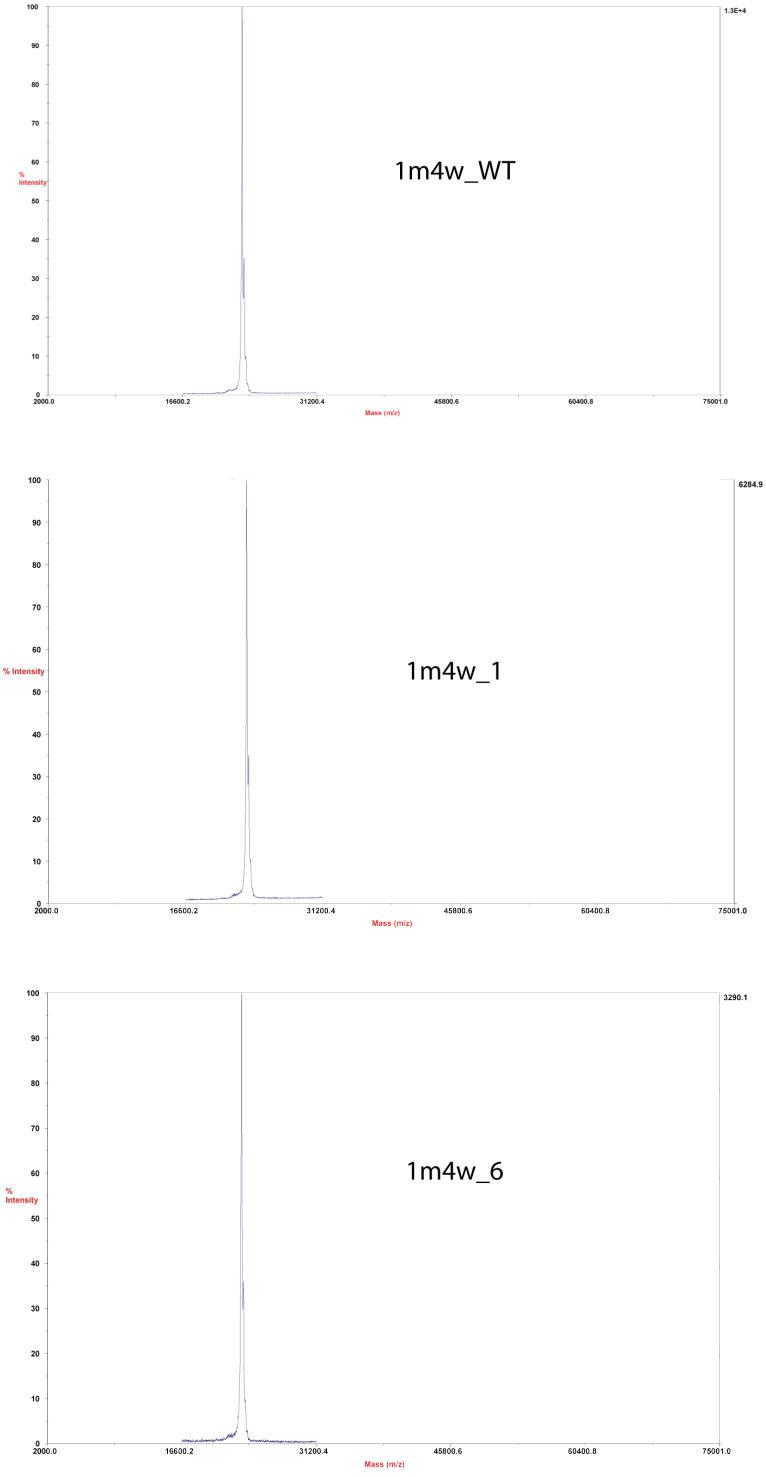
APPENDIX D: Extinction coefficients for 1m4w wild-type and designs

Extinction coefficients are in units of M-1 cm-1, at 280 nm measured in water.

1m4w_WT:	Ext. coefficient 66350
1m4w_1:	Ext. coefficient 67840
1m4w_2:	Ext. coefficient 62340
1m4w_5:	Ext. coefficient 62340
1m4w_6:	Ext. coefficient 60850
1m4w_9:	Ext. coefficient 60850
1m4w_6w20:	Ext. coefficient 66350
1m4w_6v48:	Ext. coefficient 60850
1m4w_6w20v48:	Ext. coefficient 66350

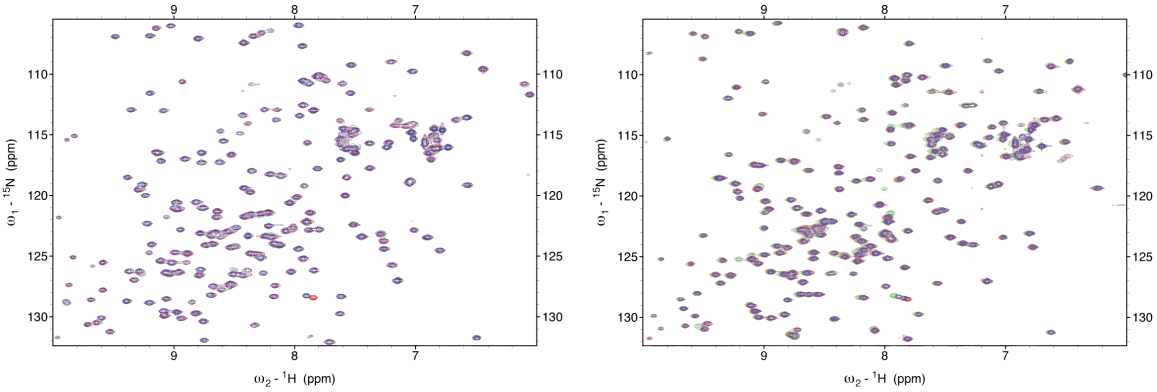
For full biochemical data of the above proteins, see morina archive DVD /1m4w_xylanase_study/1m4w-a/ProtParam_1m4w/

APPENDIX E: Mass spectra of selected 1m4w proteins



Mass spectra of 1m4w wild-type, 1m4w_1 and 1m4w_6 proteins showing relative purity and correct mass. Analysis performed by Vanderbilt Mass Spectrometry Core Facility.

APPENDIX F: NMR Spectra of 1m4w protein titrated with EKAA peptide



Overlaid 2D NMR spectra of 1m4w wild-type (left) and 1m4w_6 designed (right) proteins titrated with increasing concentrations of Glu-Lys-Ala-Ala (EKAA) peptide ligand using a Bruker Avance 600-MHz spectrometer equipped with a cryoprobe. Protein concentrated to 240uM titrated with concentrations of EKAA peptide in 0, 0.1, 0.25, 0,5 and 1 molar ratios. Chemical shift perturbations are consistent with non-specific binding, low affinity binding observed in fluorescence anisotropy assays.

BIBLIOGRAPHY

1. Strohl WR, Knight DM. Discovery and development of biopharmaceuticals: current issues. [Internet]. Current opinion in biotechnology. 2009 ;20(6):668-72.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19896824

2. Aggarwal S. What's fueling the biotech engine--2008. [Internet]. Nature biotechnology. 2009 Nov ;27(11):987-93.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19898448

3. Jones DS, Silverman AP, Cochran JR. Developing therapeutic proteins by engineering ligand-receptor interactions. [Internet]. Trends in biotechnology. 2008 Sep ;26(9):498-505.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18675482

4. Leader B, Baca QJ, Golan DE. Protein therapeutics: a summary and pharmacological classification. [Internet]. Nature reviews. Drug discovery. 2008 Jan ;7(1):21-39.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18097458

5. Lippow SM, Tidor B. Progress in computational protein design. [Internet]. Current opinion in biotechnology. 2007 Aug ;18(4):305-11.Available from: http://www.ncbi.nlm.nih.gov/pubmed/17644370

6. Bryson CJ, Jones TD, Baker MP. Prediction of immunogenicity of therapeutic proteins: validity of computational tools. [Internet]. BioDrugs : clinical immunotherapeutics, biopharmaceuticals and gene therapy. 2010 Jan ;24(1):1-8.[cited 2010 Jul 13] Available from: http://www.ncbi.nlm.nih.gov/pubmed/20055528

7. Lippow SM, Wittrup KD, Tidor B. Computational design of antibody-affinity improvement beyond in vivo maturation. [Internet]. Nature biotechnology. 2007 Oct ;25(10):1171-6.[cited 2010 Jul 13] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2803018&tool=pmcentrez&rendertype=abstract

8. Yu X-Q, Wilson AG. The role of pharmacokinetic and pharmacokinetic/pharmacodynamic modeling in drug discovery and development [Internet]. Future Medicinal Chemistry. 2010 Jun ;2(6):923-928.[cited 2010 Jul 15] Available from: http://www.future-science.com/doi/abs/10.4155/fmc.10.181

9. Hu L, Au JL-S, Wientjes MG. Computational modeling to predict effect of treatment schedule on drug delivery to prostate in humans. [Internet]. Clinical cancer research : an official journal of the American Association for Cancer Research. 2007 Mar ;13(4):1278-87.[cited 2010 Jun 30] Available from: http://www.ncbi.nlm.nih.gov/pubmed/17317840

10. Korkegian A, Black ME, Baker D, Stoddard BL. Computational thermostabilization of an enzyme. [Internet]. Science (New York, N.Y.). 2005 May ;308(5723):857-60.[cited 2010 Jul 13] Available from: http://www.ncbi.nlm.nih.gov/pubmed/15879217

11. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D. Kemp elimination catalysts by computational enzyme design. [Internet]. Nature. 2008 ;453(7192):190-5.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18354394

12. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, Hilvert D, Houk KN, Stoddard BL, Baker D. De novo computational design of retro-aldol enzymes. [Internet]. Science (New York, N.Y.). 2008 ;319(5868):1387-91.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18323453

13. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, St.Clair JL, Gallaher JL, Hilvert D, Gelb MH, Stoddard BL, Houk KN, Michael FE, Baker D. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction [Internet]. Science. 2010 Jul ;329(5989):309-313.[cited 2010 Jul 15] Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1190239

14. Huang P-S, Love JJ, Mayo SL. A de novo designed protein protein interface. [Internet]. Protein science : a publication of the Protein Society. 2007 Dec ;16(12):2770-4.[cited 2010 Sep 3] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2222823&tool=pmcentrez&rendertype=abstract

15. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ, Stoddard BL, Baker D. Computational redesign of endonuclease DNA binding and cleavage specificity. [Internet]. Nature. 2006 Jun ;441(7093):656-9.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16738662

16. Vallance P, Smart TG. The future of pharmacology. [Internet]. British journal of pharmacology. 2006 Jan ;147 Suppl S304-7.[cited 2010 Jul 27] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1760753&tool=pmcentrez&rendertype=abstract

17. Rang HP. The receptor concept: pharmacology's big idea. [Internet]. British journal of pharmacology. 2006 Jan ;147 Suppl S9-16.Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1760743&tool=pmcentrez&rendertype=abstract

18. Keskin O, Gursoy A, Ma B, Nussinov R. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? [Internet]. Chemical reviews. 2008 Apr ;108(4):1225-44.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18355092

19. Skerra A. Alternative non-antibody scaffolds for molecular recognition. [Internet]. Current opinion in biotechnology. 2007 Aug ;18(4):295-304.Available from: http://www.ncbi.nlm.nih.gov/pubmed/17643280

20. Marshall SA, Lazar GA, Chirino AJ, Desjarlais JR. Rational design and engineering of therapeutic proteins [Internet]. Drug discovery today. 2003 ;8(5):212–221.Available from: http://linkinghub.elsevier.com/retrieve/pii/S1359644603026102

21. Moreau V, Fleury C, Piquer D, Nguyen C, Novali N, Villard S, Laune D, Granier C, Molina F. PEPOP: computational design of immunogenic peptides. [Internet]. BMC bioinformatics. 2008 Jan ;971.[cited 2010 Jul 15] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2262870&tool=pmcentrez&rendertype=abstract

22. Brady RO. Enzyme replacement for lysosomal diseases. [Internet]. Annual review of medicine. 2006 Jan ;57283-96.[cited 2010 Jul 27] Available from: http://www.ncbi.nlm.nih.gov/pubmed/16409150

23. Marvin JS, Hellinga HW. Conversion of a maltose receptor into a zinc biosensor by computational design. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2001 Apr ;98(9):4955-60.[cited 2010 Jul 13] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=33145&tool=pmcentrez&rendertype=abstract

24. Looger LL, Dwyer MA, Smith JJ, Hellinga HW. Computational design of receptor and sensor proteins with novel functions. [Internet]. Nature. 2003 May ;423(6936):185-90.[cited 2010 Jul 13] Available from: http://www.ncbi.nlm.nih.gov/pubmed/12736688

25. Allert M, Rizk SS, Looger LL, Hellinga HW. Computational design of receptors for an organophosphate surrogate of the nerve agent soman. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2004 May ;101(21):7907-12.[cited 2010 Jul 13] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=419530&tool=pmcentrez&rendertype=abstract

26. Schreier B, Stumpp C, Wiesner S, Höcker B. Computational design of ligand binding is not a solved problem. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2009 Nov ;106(44):18491-6.[cited 2010 Jul 13] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2773959&tool=pmcentrez&rendertype=abstract

27. Dwyer MA, Looger LL, Hellinga HW. Retraction. [Internet]. Science (New York, N.Y.). 2008 Feb ;319(5863):569.[cited 2010 Jul 10] Available from: http://www.ncbi.nlm.nih.gov/pubmed/18239106

28. Allert M, Dwyer MA, Hellinga HW. Local encoding of computationally designed enzyme activity. [Internet]. Journal of molecular biology. 2007 Feb ;366(3):945-53.[cited 2010 Jul 13] Available from: http://www.ncbi.nlm.nih.gov/pubmed/17196220

29. Check Hayden E. Chemistry: Designer debacle [Internet]. Nature. 2008 May ;453(7193):275-278.[cited 2010 Jul 13] Available from: http://www.nature.com/doifinder/10.1038/453275a

30. Cochran FV, Wu SP, Wang W, Nanda V, Saven JG, Therien MJ, DeGrado WF. Computational de novo design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor. [Internet]. Journal of the American Chemical Society. 2005 Feb ;127(5):1346-7.[cited 2010 Jul 13] Available from: http://www.ncbi.nlm.nih.gov/pubmed/15686346

31. Binz HK, Amstutz P, Plückthun A. Engineering novel binding proteins from nonimmunoglobulin domains. [Internet]. Nature biotechnology. 2005 ;23(10):1257-68.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16211069

32. Schellekens H. Factors influencing the immunogenicity of therapeutic proteins. [Internet]. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association. 2005 ;20 Suppl 6vi3-9.Available from: http://www.ncbi.nlm.nih.gov/pubmed/15958824

33. Chirino A, Ary M, Marshall S. Minimizing the immunogenicity of protein therapeutics. [Internet]. Drug discovery today. 2004 Jan ;9(2):82-90.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/15012932

34. Smith AB, Savinov SN, Manjappara UV, Chaiken IM. Peptide-small molecule hybrids via orthogonal deprotection-chemoselective conjugation to cysteine-anchored scaffolds. A model study. [Internet]. Organic letters. 2002 Nov ;4(23):4041-4.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/12423081

35. Pastan I. Immunotoxins containing Pseudomonas exotoxin A: a short history. [Internet]. Cancer immunology, immunotherapy : CII. 2003 May ;52(5):338-41.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/12700949

36. Helguera G, Penichet ML. Antibody-cytokine fusion proteins for the therapy of cancer. [Internet]. Methods in molecular medicine. 2005 Jan ;109347-74.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/15585931

37. Davis CB, Gillies SD. Immunocytokines: amplification of anti-cancer immunity. [Internet]. Cancer immunology, immunotherapy : CII. 2003 May ;52(5):297-308.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/12700945

38. Willuda J, Kubetzko S, Waibel R, Schubiger PA, Zangemeister-Wittke U, Plückthun A. Tumor targeting of mono-, di-, and tetravalent anti-p185(HER-2) miniantibodies multimerized by self-associating peptides. [Internet]. The Journal of biological chemistry. 2001 Apr ;276(17):14385-92.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/11278961

39. Bailon P, Palleroni A, Schaffer CA, Spence CL, Fung WJ, Porter JE, Ehrlich GK, Pan W, Xu ZX, Modi MW, Farid A, Berthold W, Graves M. Rational design of a potent, long-lasting form of interferon: a 40 kDa branched polyethylene glycol-conjugated interferon alpha-2a for the treatment of hepatitis C. [Internet]. Bioconjugate chemistry. 12(2):195-202.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/11312680

40. Hershfield MS, Chaffee S, Koro-Johnson L, Mary A, Smith AA, Short SA. Use of site-directed mutagenesis to enhance the epitope-shielding effect of covalent modification of proteins with polyethylene glycol. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 1991 Aug ;88(16):7185-9.[cited 2011 Apr 11] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=52258&tool=pmcentrez&rendertype=abstract

41. Harris JM, Martin NE, Modi M. Pegylation: a novel process for modifying pharmacokinetics. [Internet]. Clinical pharmacokinetics. 2001 Jan ;40(7):539-51.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/11510630

42. Macdougall IC, Gray SJ, Elston O, Breen C, Jenkins B, Browne J, Egrie J. Pharmacokinetics of novel erythropoiesis stimulating protein compared with epoetin alfa in dialysis patients. [Internet]. Journal of the American Society of Nephrology : JASN. 1999 Nov ;10(11):2392-5.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/10541299

43. Hermeling S, Crommelin DJA, Schellekens H, Jiskoot W. Structure-immunogenicity relationships of therapeutic proteins [Internet]. Pharmaceutical research. 2004 ;21(6):897–903.Available from: http://www.springerlink.com/index/N740631734543617.pdf

44. Schellekens H. Immunogenicity of therapeutic proteins [Internet]. Nephrology Dialysis Transplantation. 2003 ;18(7):1257-1259.Available from: http://www.ndt.oupjournals.org/cgi/doi/10.1093/ndt/gfg164

45. Kessler M, Goldsmith D, Schellekens H. Immunogenicity of biopharmaceuticals. [Internet]. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association. 2006 ;21 Suppl 5v9-12.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16959792

46. Pavlou AK, Reichert JM. Recombinant protein therapeutics—success rates, market trends and values to 2010 [Internet]. Nature biotechnology. 2004 ;22(12):1513–1519.Available from: http://csdd.tufts.edu/_documents/www/doc_233_7750_826.pdf

47. Gill DS, Damle NK. Biopharmaceutical drug discovery using novel protein scaffolds. [Internet]. Current opinion in biotechnology. 2006 ;17(6):653-8.Available from: http://www.ncbi.nlm.nih.gov/pubmed/17055245

48. Frank RG. New estimates of drug development costs. [Internet]. Journal of health economics. 2003 Mar ;22(2):325-30.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/12606149

49. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. [Internet]. Journal of health economics. 2003 Mar ;22(2):151-85.[cited 2010 Jul 17] Available from: http://www.ncbi.nlm.nih.gov/pubmed/12606142

50. US GAO. New Drug Development: Science, Business, Regulatory, and Intellectual Property Issues Cited as Hampering Drug Development Efforts. Report to Congressional Requesters. 2006 ;

51. Vecchione T. Sluggish antibiotic pipeline driving “superbug” fears. Drug Topics. 2003 ;147:42.(Jun. 2):

52. Barrett JF. Can biotech deliver new antibiotics? [Internet]. Current opinion in microbiology. 2005 ;8(5):498-503.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16125445

53. Peterson LR. Bad bugs, no drugs: no ESCAPE revisited. [Internet]. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2009 Sep ;49(6):992-3.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/19694542

54. Högberg LD, Heddini A, Cars O. The global need for effective antibiotics: challenges and recent advances. [Internet]. Trends in pharmacological sciences. 2010 Nov ;31(11):509-15.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/20843562

55. Neu HC. The crisis in antibiotic resistance. [Internet]. Science (New York, N.Y.). 1992 Aug ;257(5073):1064-73.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/1509257

56. Guillemot D, Gasquet I, Vallet O, David M-F, Laurent C, Mathieu D. Thirty-day mortality of nosocomial systemic bacterial infections according to antibiotic susceptibility in an 800-bed teaching hospital in France. [Internet]. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2005 Jun ;11(6):502-4.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/15882203

57. Rice LB. Antimicrobial resistance in gram-positive bacteria. [Internet]. The American journal of medicine. 2006 ;119(6 Suppl 1):S11-9; discussion S62-70.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16735146

58. Rice LB. Controlling antibiotic resistance in the ICU: different bacteria, different strategies. [Internet]. Cleveland Clinic journal of medicine. 2003 Sep ;70(9):793-800.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/14518574

59. CDC. CDC's campaign to prevent antimicrobial resistance in health-care settings. [Internet]. MMWR. Morbidity and mortality weekly report. 2002 Apr ;51(15):343.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/12004862

60. Kluytmans-Vandenbergh MFQ, Kluytmans JAJW. Community-acquired methicillin-resistant Staphylococcus aureus: current perspectives. [Internet]. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2006 Mar ;12 Suppl 19-15.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/16445719

61. Howard DH, Scott RD. The economic burden of drug resistance. [Internet]. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2005 Aug ;41 Suppl 4S283-6.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/16032567

62. Coque TM, Tomayko JF, Ricke SC, Okhyusen PC, Murray BE. Vancomycin-resistant enterococci from nosocomial, community, and animal sources in the United States. [Internet]. Antimicrobial agents and chemotherapy. 1996 Nov ;40(11):2605-9.[cited 2011 Apr 11] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=163584&tool=pmcentrez&rendertype=abstract

63. Marcinak JF, Frank AL. Epidemiology and treatment of community-associated methicillin-resistant Staphylococcus aureus in children. [Internet]. Expert review of anti-infective therapy. 2006 Feb ;4(1):91-100.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/16441212

64. Loll PJ, Axelsen PH. The structural biology of molecular recognition by vancomycin. [Internet]. Annual review of biophysics and biomolecular structure. 2000 Jan ;29265-89.[cited 2010 Sep 17] Available from: http://www.ncbi.nlm.nih.gov/pubmed/10940250

65. Walsh CT, Fisher SL, Park IS, Prahalad M, Wu Z. Bacterial resistance to vancomycin: five genes and one missing hydrogen bond tell the story. [Internet]. Chemistry & biology. 1996 Jan ;3(1):21-8.Available from: http://www.ncbi.nlm.nih.gov/pubmed/8807824

66. Healy VL, Lessard IA, Roper DI, Knox JR, Walsh CT. Vancomycin resistance in enterococci: reprogramming of the D-ala-D-Ala ligases in bacterial peptidoglycan biosynthesis. [Internet]. Chemistry & biology. 2000 May ;7(5):R109-19.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/10801476

67. McComas CC, Crowley BM, Boger DL. Partitioning the loss in vancomycin binding affinity for D-Ala-D-Lac into lost H-bond and repulsive lone pair contributions. [Internet]. Journal of the American Chemical Society. 2003 Aug ;125(31):9314-5.Available from: http://www.ncbi.nlm.nih.gov/pubmed/12889959

68. Ma N, Jia Y, Liu Z, Gonzalez-Zamora E, Bois-Choussy M, Malabarba A, Brunati C, Zhu J. Design and synthesis of macrocycles active against vancomycin-resistant enterococci (VRE): the interplay between d-Ala-d-Lac binding and hydrophobic effect. [Internet]. Bioorganic & medicinal chemistry letters. 2005 ;15(3):743-6.Available from: http://www.ncbi.nlm.nih.gov/pubmed/15664849

69. Walsh TR, Howe RA. The prevalence and mechanisms of vancomycin resistance in Staphylococcus aureus. [Internet]. Annual review of microbiology. 2002 ;56657-75.Available from: http://www.ncbi.nlm.nih.gov/pubmed/12142482

70. Crowley BM, Boger DL. Total synthesis and evaluation of [Psi[CH2NH]Tpg4]vancomycin aglycon: reengineering vancomycin for dual D-Ala-D-Ala and D-Ala-D-Lac binding. [Internet]. Journal of the American Chemical Society. 2006 Mar ;128(9):2885-92.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16506767

71. Appelbaum PC. The emergence of vancomycin-intermediate and vancomycin-resistant Staphylococcus aureus. [Internet]. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2006 Mar ;12 Suppl 116-23.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/16445720

72. Woodford N. Epidemiology of the genetic elements responsible for acquired glycopeptide resistance in enterococci. [Internet]. Microbial drug resistance (Larchmont, N.Y.). 2001 Jan ;7(3):229-36.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/11759084

73. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban Y-EA, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popović Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P. Rosetta3 an object-oriented software suite for the simulation and design of macromolecules. [Internet]. Methods in enzymology. 2011 Jan ;487545-74.[cited 2010 Dec 29] Available from: http://www.ncbi.nlm.nih.gov/pubmed/21187238

74. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D. Progress in modeling of protein structures and interactions. [Internet]. Science (New York, N.Y.). 2005 Oct ;310(5748):638-42.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/16254179

75. Dunbrack RL. Rotamer libraries in the 21st century. [Internet]. Current opinion in structural biology. 2002 Aug ;12(4):431-40.[cited 2010 Aug 22] Available from: http://www.ncbi.nlm.nih.gov/pubmed/12163064

76. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. [Internet]. Protein science : a publication of the Protein Society. 1997 Aug ;6(8):1661-81.[cited 2011 Apr 11] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2143774&tool=pmcentrez&rendertype=abstract

77. Dunbrack RL, Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. [Internet]. Nature structural biology. 1994 May ;1(5):334-40.[cited 2010 Oct 25] Available from: http://www.ncbi.nlm.nih.gov/pubmed/7664040

78. Dunbrack RL, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. [Internet]. Journal of molecular biology. 1993 Mar ;230(2):543-74.[cited 2010 Oct 25] Available from: http://www.ncbi.nlm.nih.gov/pubmed/8464064

79. Poole AM, Ranganathan R. Knowledge-based potentials in protein design. [Internet]. Current opinion in structural biology. 2006 Aug ;16(4):508-13.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/16843652

80. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. [Internet]. Science (New York, N.Y.). 2003 ;302(5649):1364-8.Available from: http://www.ncbi.nlm.nih.gov/pubmed/14631033

81. Korkegian A, Black ME, Baker D, Stoddard BL. Computational thermostabilization of an enzyme. [Internet]. Science (New York, N.Y.). 2005 ;308(5723):857-60.Available from: http://www.ncbi.nlm.nih.gov/pubmed/15879217

82. Kortemme T, Baker D. Computational design of protein-protein interactions. [Internet]. Current opinion in chemical biology. 2004 ;8(1):91-7.Available from: http://www.ncbi.nlm.nih.gov/pubmed/15036162

83. Mandell DJ, Kortemme T. Computer-aided design of functional protein interactions. [Internet]. Nature chemical biology. 2009 ;5(11):797-807.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19841629

84. Sammond DW, Eletr ZM, Purbeck C, Kuhlman B. Computational design of second-site suppressor mutations at protein-protein interfaces. [Internet]. Proteins. 2010 ;78(4):1055-65.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19899154

85. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, Baker D. Computational redesign of protein-protein interaction specificity. [Internet]. Nature structural & molecular biology. 2004 ;11(4):371-9.Available from: http://www.ncbi.nlm.nih.gov/pubmed/15034550

86. Joachimiak LA, Kortemme T, Stoddard BL, Baker D. Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. [Internet]. Journal of molecular biology. 2006 ;361(1):195-208.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16831445

87. Humphris EL, Kortemme T. Design of multi-specificity in protein interfaces. [Internet]. PLoS computational biology. 2007 ;3(8):e164.Available from: http://www.ncbi.nlm.nih.gov/pubmed/17722975

88. Kortemme T, Morozov AV, Baker D. An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein–Protein Complexes [Internet]. Journal of Molecular Biology. 2003 ;326(4):1239-1259.Available from: http://linkinghub.elsevier.com/retrieve/pii/S0022283603000214

89. Schueler-Furman O, Wang C, Baker D. Progress in protein-protein docking: atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility. [Internet]. Proteins. 2005 ;60(2):187-94.Available from: http://www.ncbi.nlm.nih.gov/pubmed/15981249

90. Gray JJ. Protein-protein docking. Nature. 2006 ;(April):

91. Sood VD, Baker D. Recapitulation and design of protein binding peptide structures and sequences. [Internet]. Journal of molecular biology. 2006 ;357(3):917-27.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16473368

92. Meiler J, Baker D. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. [Internet]. Proteins. 2006 ;65(3):538-48.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16972285

93. Pabo C. Molecular technology: Designing proteins and peptides [Internet]. Nature. 1983 Jan ;301(5897):200-200.[cited 2010 Aug 11] Available from: http://www.nature.com/doifinder/10.1038/301200a0

94. Lippow SM, Tidor B. Progress in computational protein design. [Internet]. Current opinion in biotechnology. 2007 Aug ;18(4):305-11.Available from: http://www.ncbi.nlm.nih.gov/pubmed/17644370

95. Alvizo O, Allen BD, Mayo SL. Computational protein design promises to revolutionize protein engineering. [Internet]. BioTechniques. 2007 ;42(1):31, 33, 35 passim.Available from: http://www.ncbi.nlm.nih.gov/pubmed/17269482

96. Allen BD, Mayo SL. Dramatic performance enhancements for the FASTER optimization algorithm. [Internet]. Journal of computational chemistry. 2006 Jul ;27(10):1071-5.[cited 2010 Aug 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/16685715

97. Desjarlais JR, Handel TM. De novo design of the hydrophobic cores of proteins. [Internet]. Protein science : a publication of the Protein Society. 1995 Oct ;4(10):2006-18.[cited 2010 Aug 11] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2142989&tool=pmcentrez&rendertype=abstract

98. Koehl P. Mean-field minimization methods for biological macromolecules [Internet]. Current Opinion in Structural Biology. 1996 Apr ;6(2):222-226.[cited 2010 Aug 11] Available from: http://linkinghub.elsevier.com/retrieve/pii/S0959440X96800789

99. Voigt CA, Gordon DB, Mayo SL. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. [Internet]. Journal of molecular biology. 2000 Jun ;299(3):789-803.[cited 2010 Jul 23] Available from: http://www.ncbi.nlm.nih.gov/pubmed/10835284

100. Tian P. Computational protein design, from single domain soluble proteins to membrane proteins. [Internet]. Chemical Society reviews. 2010 Jun ;39(6):2071-82.[cited 2010 Aug 12] Available from: http://www.ncbi.nlm.nih.gov/pubmed/20407671

101. Dunbrack R. Rotamer Libraries in the 21st Century [Internet]. Current Opinion in Structural Biology. 2002 Aug ;12(4):431-440.Available from: http://linkinghub.elsevier.com/retrieve/pii/S0959440X02003445

102. Boas FE, Harbury PB. Potential energy functions for protein design. [Internet]. Current opinion in structural biology. 2007 Apr ;17(2):199-204.Available from: http://www.ncbi.nlm.nih.gov/pubmed/17387014

103. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. [Internet]. Journal of molecular biology. 1997 Apr ;268(1):209-25.[cited 2010 Aug 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/9149153

104. Dehouck Y, Gilis D, Rooman M. A new generation of statistical potentials for proteins. [Internet]. Biophysical journal. 2006 Jun ;90(11):4010-7.[cited 2010 Jul 18] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1459517&tool=pmcentrez&rendertype=abstract

105. Mackerell AD. Empirical force fields for biological macromolecules: overview and issues. [Internet]. Journal of computational chemistry. 2004 Oct ;25(13):1584-604.[cited 2010 Aug 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/15264253

106. Boas FE, Harbury PB. Design of protein-ligand binding based on the molecular-mechanics energy model. [Internet]. Journal of molecular biology. 2008 Jul ;380(2):415-24.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18514737

107. Pierce NA, Winfree E. Protein design is NP-hard. [Internet]. Protein engineering. 2002 Oct ;15(10):779-82.[cited 2010 Aug 12] Available from: http://www.ncbi.nlm.nih.gov/pubmed/12468711

108. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2000 ;97(19):10383.Available from: http://www.pnas.org/content/97/19/10383.long

109. Metropolis N, Ulam S. The Monte Carlo method. [Internet]. Journal of the American Statistical Association. 1949 Sep ;44(247):335-41.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/18139350

110. Demchick P, Koch a L. The permeability of the wall fabric of Escherichia coli and Bacillus subtilis. [Internet]. Journal of bacteriology. 1996 Feb ;178(3):768-73.Available from: http://www.ncbi.nlm.nih.gov/pubmed/8550511

111. Meroueh SO, Bencze KZ, Hesek D, Lee M, Fisher JF, Stemmler TL, Mobashery S. Three-dimensional structure of the bacterial cell wall peptidoglycan. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2006 ;103(12):4404-9.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16537437

112. Hakulinen N, Turunen O, Janis J, Leisola M, Rouvinen J. Three-dimensional structures of thermophilic beta-1,4-xylanases from Chaetomium thermophilum and Nonomuraea flexuosa. Comparison of twelve xylanases in relation to their thermal stability [Internet]. European Journal of Biochemistry. 2003 ;270(7):1399-1412.Available from: http://www.blackwell-synergy.com/links/doi/10.1046%2Fj.1432-1033.2003.03496.x

113. Kang B, Cooper D, Devedjiev Y, Derewenda U, Derewenda Z. Molecular Roots of Degenerate Specificity in Syntenin's PDZ2 DomainReassessment of the PDZ Recognition Paradigm [Internet]. Structure. 2003 ;11(7):845-853.Available from: http://linkinghub.elsevier.com/retrieve/pii/S0969212603001254

114. Scheufler C, Brinker A, Bourenkov G, Pegoraro S, Moroder L, Bartunik H, Hartl FU, Moarefi I. Structure of TPR Domain – Peptide Complexes: Critical Elements in the Assembly of the Hsp70–Hsp90 Multichaperone Machine. Cell. 2000 ;101199-210.

115. Prodromou C, Pearl LH. Recursive PCR: a novel technique for total gene synthesis. [Internet]. Protein engineering. 1992 Dec ;5(8):827-9.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/1287665

116. Zheng L, Baumann U, Reymond J-L. An efficient one-step site-directed and site-saturation mutagenesis protocol. [Internet]. Nucleic acids research. 2004 Jan ;32(14):e115.[cited 2011 Mar 8] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=514394&tool=pmcentrez&rendertype=abstract

117. Popieniek PH, Pratt RF. A fluorescent ligand for binding studies with glycopeptide antibiotics of the vancomycin class. [Internet]. Analytical biochemistry. 1987 Aug ;165(1):108-13.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/3688425

118. Skelton NJ, Koehler MFT, Zobel K, Wong WL, Yeh S, Pisabarro MT, Yin JP, Lasky LA, Sidhu SS. Origins of PDZ domain ligand specificity. Structure determination and mutagenesis of the Erbin PDZ domain. [Internet]. The Journal of biological chemistry. 2003 Feb ;278(9):7645-54.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/12446668

119. Arora N, Banerjee AK, Mutyala S, Murty US. Comparative characterization of commercially important xylanase enzymes. [Internet]. Bioinformation. 2009 ;3(10):446-53.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19759868

120. Gerlt JA, Babbitt PC. Enzyme (re)design: lessons from natural evolution and computation. [Internet]. Current opinion in chemical biology. 2009 ;13(1):10-8.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19237310

121. Damborsky J, Brezovsky J. Computational tools for designing and engineering biocatalysts. [Internet]. Current opinion in chemical biology. 2009 ;13(1):26-34.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19297237

122. Kaplan J, DeGrado WF. De novo design of catalytic proteins. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2004 Aug ;101(32):11566-70.Available from: http://www.ncbi.nlm.nih.gov/pubmed/15292507

123. Flower DR, McSparron H, Blythe MJ, Zygouri C, Taylor D, Guan P, Wan S, Coveney PV, Walshe V, Borrow P, Doytchinova IA. Computational vaccinology: quantitative approaches. [Internet]. Novartis Foundation symposium. 2003 ;254102-20; discussion 120-5, 216-22, 250-2.Available from: http://www.ncbi.nlm.nih.gov/pubmed/14712934

124. Shi L, Sings HL, Bryan JT, Wang B, Wang Y, Mach H, Kosinski M, Washabaugh MW, Sitrin R, Barr E. GARDASIL: prophylactic human papillomavirus vaccine development--from bench top to bed-side. [Internet]. Clinical pharmacology and therapeutics. 2007 ;81(2):259-64.Available from: http://www.ncbi.nlm.nih.gov/pubmed/17259949

125. Das R, Baker D. Macromolecular modeling with rosetta. [Internet]. Annual review of biochemistry. 2008 ;77(March):363-82.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18410248

126. Karanicolas J, Kuhlman B. Computational design of affinity and specificity at protein-protein interfaces. [Internet]. Current opinion in structural biology. 2009 ;19(4):458-63.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19646858

127. Taillefer R, Edell S, Innes G, Lister-James J. Acute thromboscintigraphy with (99m)Tc-apcitide: results of the phase 3 multicenter clinical trial comparing 99mTc-apcitide scintigraphy with contrast venography for imaging acute DVT. Multicenter Trial Investigators. [Internet]. Journal of nuclear medicine : official publication, Society of Nuclear Medicine. 2000 ;41(7):1214-23.Available from: http://www.ncbi.nlm.nih.gov/pubmed/10914912

128. Sodee DB, Malguria N, Faulhaber P, Resnick MI, Albert J, Bakale G. Multicenter ProstaScint imaging findings in 2154 patients with prostate cancer. The ProstaScint Imaging Centers. [Internet]. Urology. 2000 ;56(6):988-93.Available from: http://www.ncbi.nlm.nih.gov/pubmed/11113745

129. Schreier B, Stumpp C, Wiesner S, Höcker B. Computational design of ligand binding is not a solved problem. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2009 ;106(44):18491-6.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19833875

130. Hayden EC. Key protein-design papers challenged. [Internet]. Nature. 2009 ;461(7266):859.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19829341

131. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, St.Clair JL, Gallaher JL, Hilvert D, Gelb MH, Stoddard BL, Houk KN, Michael FE, Baker D. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction [Internet]. Science. 2010 Jul ;329(5989):309-313.[cited 2010 Jul 15] Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1190239

132. Boneca IG, Chiosis G. Vancomycin resistance: occurrence, mechanisms and strategies to combat it. [Internet]. Expert opinion on therapeutic targets. 2003 ;7(3):311-28.Available from: http://www.ncbi.nlm.nih.gov/pubmed/12783569

133. Loll PJ, Axelsen PH. The structural basis of molecular recognition by vancomycin. Annual Review of Biophysics and Biomolecular Structure. 2000 ;29265-289.

134. Cui L, Iwamoto A, Lian J-qi, Neoh H-min, Maruyama T, Horikawa Y, Hiramatsu K. Novel mechanism of antibiotic resistance originating in vancomycin-intermediate Staphylococcus aureus. [Internet]. Antimicrobial agents and chemotherapy. 2006 ;50(2):428-38.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16436693

135. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2000 Sep ;97(19):10383-8.Available from: http://www.ncbi.nlm.nih.gov/pubmed/10984534

136. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. [Internet]. Protein science : a publication of the Protein Society. 1997 Aug ;6(8):1661-81.Available from: http://www.ncbi.nlm.nih.gov/pubmed/9260279

137. Stemmer WP, Crameri A, Ha KD, Brennan TM, Heyneker HL. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. [Internet]. Gene. 1995 ;164(1):49-53.Available from: http://www.ncbi.nlm.nih.gov/pubmed/7590320

138. Rouillard J-marie, Lee W, Truan G, Gao X, Zhou X, Gulari E. Gene2Oligo: oligonucleotide design for in vitro gene synthesis. [Internet]. Nucleic acids research. 2004 ;32(Web Server issue):W176-80.Available from: http://www.ncbi.nlm.nih.gov/pubmed/15215375

139. Vagin A, Teplyakov A. MOLREP : an Automated Program for Molecular Replacement [Internet]. Journal of Applied Crystallography. 1997 ;30(6):1022-1025.Available from: http://scripts.iucr.org/cgi-bin/paper?S0021889897006766

140. Langer G, Cohen SX, Lamzin VS, Perrakis A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. [Internet]. Nature protocols. 2008 ;3(7):1171-9.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18600222

141. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. [Internet]. Acta crystallographica. Section D, Biological

crystallography. 1997 ;53(Pt 3):240-55.Available from: http://www.ncbi.nlm.nih.gov/pubmed/15299926

142. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. [Internet]. Acta crystallographica. Section D, Biological crystallography. 2004 ;60(Pt 12 Pt 1):2126-32.Available from: http://www.ncbi.nlm.nih.gov/pubmed/15572765

143. Potterton E, Briggs P, Turkenburg M, Dodson E. A graphical user interface to the CCP4 program suite. [Internet]. Acta crystallographica. Section D, Biological crystallography. 2003 ;59(Pt 7):1131-7.Available from: http://www.ncbi.nlm.nih.gov/pubmed/12832755

144. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. [Internet]. Protein engineering. 1995 ;8(2):127-34.Available from: http://www.ncbi.nlm.nih.gov/pubmed/7630882

145. Davis IW, Baker D. RosettaLigand docking with full ligand and receptor flexibility. [Internet]. Journal of molecular biology. 2009 ;385(2):381-92.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19041878

146. Rueda M, Ferrer-Costa C, Meyer T, Pérez A, Camps J, Hospital A, Gelpí JL, Orozco M. A consensus view of protein dynamics. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2007 Jan ;104(3):796-801.Available from: http://www.ncbi.nlm.nih.gov/pubmed/17215349

147. Lazaridis T, Karplus M. Effective energy function for proteins in solution [Internet]. Proteins: Structure, Function, and Genetics. 1999 May ;35(2):133-152.[cited 2010 Jul 28] Available from: http://doi.wiley.com/10.1002/(SICI)1097-0134(19990501)35:2<133::AID-PROT1>3.0.CO;2-N

148. Gilson MK, Zhou H-xiang. Calculation of protein-ligand binding affinities. [Internet]. Annual review of biophysics and biomolecular structure. 2007 ;3621-42.Available from: http://www.ncbi.nlm.nih.gov/pubmed/17201676

149. Thilagavathi R, Mancera RL. Ligand-protein cross-docking with water molecules. [Internet]. Journal of chemical information and modeling. 2010 ;50(3):415-21.Available from: http://www.ncbi.nlm.nih.gov/pubmed/20158272

150. Amadasi A, Spyrakis F, Cozzini P, Abraham DJ, Kellogg GE, Mozzarelli A. Mapping the energetics of water-protein and water-ligand interactions with the “natural” HINT forcefield: predictive tools for characterizing the roles of water in biomolecules. [Internet]. Journal of molecular biology. 2006 ;358(1):289-309.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16497327

151. Yamada R, Kera Y. D-amino acid hydrolysing enzymes. [Internet]. EXS. 1998 ;85145-55.Available from: http://www.ncbi.nlm.nih.gov/pubmed/9949873

152. Sela M, Zisman E. Different roles of D-amino acids in immune phenomena. [Internet]. The FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 1997 ;11(6):449-56.Available from: http://www.ncbi.nlm.nih.gov/pubmed/9194525

153. Sunna A, Gibbs MD, Chin CW, Nelson PJ, Bergquist PL. A gene encoding a novel multidomain beta-1,4-mannanase from Caldibacillus cellulovorans and action of the recombinant enzyme on kraft pulp. [Internet]. Applied and environmental microbiology. 2000 ;66(2):664-70.Available from: http://www.ncbi.nlm.nih.gov/pubmed/10653733

154. Vieira DS, Degrève L, Ward RJ. Characterization of temperature dependent and substrate-binding cleft movements in Bacillus circulans family 11 xylanase: a molecular dynamics investigation. [Internet]. Biochimica et biophysica acta. 2009 ;1790(10):1301-6.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19409448

155. Ashworth J, Taylor GK, Havranek JJ, Quadri SA, Stoddard BL, Baker D. Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. [Internet]. Nucleic acids research. 2010 Apr ;[cited 2010 Jul 23] Available from: http://www.ncbi.nlm.nih.gov/pubmed/20435674

156. Humphris EL, Kortemme T. Design of multi-specificity in protein interfaces. [Internet]. PLoS computational biology. 2007 Aug ;3(8):e164.[cited 2010 Jul 6] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1950952&tool=pmcentrez&rendertype=abstract

157. Jackrel ME, Valverde R, Regan L. Redesign of a protein-peptide interaction: characterization and applications. [Internet]. Protein science : a publication of the Protein Society. 2009 ;18(4):762-74.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19309728

158. Cortajarena AL, Yi F, Regan L. Designed TPR modules as novel anticancer agents. [Internet]. ACS chemical biology. 2008 Mar ;3(3):161-6.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18355005

159. Zhang X, DeChancie J, Gunaydin H, Chowdry AB, Clemente FR, Smith AJT, Handel TM, Houk KN. Quantum mechanical design of enzyme active sites. [Internet]. The Journal of organic chemistry. 2008 Feb ;73(3):889-99.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18179229

160. Kamerlin SCL, Warshel A. At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? [Internet]. Proteins. 2010 May ;78(6):1339-75.[cited 2010 Aug 24] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2841229&tool=pmcentrez&rendertype=abstract

161. Pisliakov AV, Cao J, Kamerlin SCL, Warshel A. Enzyme millisecond conformational dynamics do not catalyze the chemical step. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2009 Oct ;106(41):17359-64.[cited 2010 Aug 24] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2762662&tool=pmcentrez&rendertype=abstract

162. Davis IW, Raha K, Head MS, Baker D. Blind docking of pharmaceutically relevant compounds using RosettaLigand. [Internet]. Protein science : a publication of the Protein Society. 2009 ;18(9):1998-2002.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19554568

163. Roche O, Kiyama R, Brooks CL. Ligand−Protein DataBase: Linking Protein−Ligand Complex Structures to Binding Data [Internet]. Journal of Medicinal Chemistry. 2001 Oct ;44(22):3592-3598.[cited 2011 Apr 8] Available from: http://pubs.acs.org/doi/abs/10.1021/jm000467k

164. Kelley LA, Shrimpton PJ, Muggleton SH, Sternberg MJE. Discovering rules for protein-ligand specificity using support vector inductive logic programming. [Internet]. Protein engineering, design & selection : PEDS. 2009 ;22(9):561-7.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19574295

165. Henrich S, Salo-Ahen OMH, Huang B, Rippmann FF, Cruciani G, Wade RC. Computational approaches to identifying and characterizing protein binding sites for ligand design. [Internet]. Journal of molecular recognition : JMR. 2010 Mar ;23(2):209-19.[cited 2010 Sep 7] Available from: http://www.ncbi.nlm.nih.gov/pubmed/19746440

166. Vajda S, Guarnieri F. Characterization of protein-ligand interaction sites using experimental and computational methods. [Internet]. Current opinion in drug discovery & development. 2006 May ;9(3):354-62.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16729732

167. Villar HO, Kauvar LM. Amino acid preferences at protein binding sites. [Internet]. FEBS letters. 1994 Jul ;349(1):125-30.Available from: http://www.ncbi.nlm.nih.gov/pubmed/8045288

168. Soga S, Shirai H, Kobori M, Hirayama N. Use of amino acid composition to predict ligand-binding sites. [Internet]. Journal of chemical information and modeling. 2007 ;47(2):400-6.Available from: http://www.ncbi.nlm.nih.gov/pubmed/17243757

169. Wang G. PISCES: a protein sequence culling server [Internet]. Bioinformatics. 2003 Aug ;19(12):1589-1591.[cited 2010 Nov 18] Available from: http://www.bioinformatics.oupjournals.org/cgi/doi/10.1093/bioinformatics/btg224

170. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF. Protein identification and analysis tools in the ExPASy server. [Internet]. Methods in molecular biology (Clifton, N.J.). 1999 Jan ;112531-52.[cited 2010 Aug 22] Available from: http://www.ncbi.nlm.nih.gov/pubmed/10027275

171. Monod J, Wyman J, Changeux J-P. On the nature of allosteric transitions: A plausible model [Internet]. Journal of Molecular Biology. 1965 May ;12(1):88-118.[cited 2010 Jul 27] Available from: http://linkinghub.elsevier.com/retrieve/pii/S0022283665802856

172. Kumar S, Ma B, Tsai CJ, Sinha N, Nussinov R. Folding and binding cascades: dynamic landscapes and population shifts. [Internet]. Protein science : a publication of the Protein Society. 2000 Jan ;9(1):10-9.[cited 2010 Aug 11] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2144430&tool=pmcentrez&rendertype=abstract

173. Grünberg R, Leckner J, Nilges M. Complementarity of structure ensembles in protein-protein binding. [Internet]. Structure (London, England : 1993). 2004 Dec ;12(12):2125-36.Available from: http://www.ncbi.nlm.nih.gov/pubmed/15576027

174. Wlodarski T, Zagrovic B. Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2009 Nov ;106(46):19346-51.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19887638

175. Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. [Internet]. Nature chemical biology. 2009 Nov ;5(11):789-96.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19841628

176. Weikl TR, Deuster C von. Selected-fit versus induced-fit protein binding: kinetic differences and mutational analysis. [Internet]. Proteins. 2009 Apr ;75(1):104-10.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18798570

177. Okazaki K-I, Takada S. Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2008 Aug ;105(32):11182-7.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18678900

178. Hammes GG, Chang Y-C, Oas TG. Conformational selection or induced fit: a flux description of reaction mechanism. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2009 Aug ;106(33):13737-41.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19666553

179. Reina J, Lacroix E, Hobson SD, Fernandez-Ballester G, Rybin V, Schwab MS, Serrano L, Gonzalez C. Computer-aided design of a PDZ domain to recognize new target sequences. [Internet]. Nature structural biology. 2002 Aug ;9(8):621-7.Available from: http://www.ncbi.nlm.nih.gov/pubmed/12080331

180. Chaudhury S, Gray JJ. Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. [Internet]. Journal of molecular biology. 2008 Sep ;381(4):1068-87.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18640688

181. Bonvin AMJJ. Flexible protein-protein docking. [Internet]. Current opinion in structural biology. 2006 Apr ;16(2):194-200.[cited 2010 Aug 3] Available from: http://www.ncbi.nlm.nih.gov/pubmed/16488145

182. Mandell DJ, Kortemme T. Backbone flexibility in computational protein design. [Internet]. Current opinion in biotechnology. 2009 Aug ;20(4):420-8.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19709874

183. Smith CA, Kortemme T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. [Internet]. Journal of molecular biology. 2008 Jul ;380(4):742-56.Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2603262&tool=pmcentrez&rendertype=abstract

184. Georgiev I, Keedy D, Richardson JS, Richardson DC, Donald BR. Algorithm for backrub motions in protein design. [Internet]. Bioinformatics (Oxford, England). 2008 Jul ;24(13):i196-204.[cited 2010 Aug 11] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2718647&tool=pmcentrez&rendertype=abstract

185. Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. [Internet]. Nature methods. 2009 Aug ;6(8):551-2.Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2847683&tool=pmcentrez&rendertype=abstract

186. Kolodny R. Inverse Kinematics in Biology: The Protein Loop Closure Problem [Internet]. The International Journal of Robotics Research. 2005 Feb ;24(2-3):151-163.[cited 2010 Aug 11] Available from: http://ijr.sagepub.com/cgi/doi/10.1177/0278364905050352

187. Lu Y, Wang R, Yang C-yie, Wang S. Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes. [Internet]. Journal of chemical information and modeling. 2007 ;47(2):668-75.Available from: http://www.ncbi.nlm.nih.gov/pubmed/17266298

188. Li Z, Lazaridis T. Water at biomolecular binding interfaces. [Internet]. Physical chemistry chemical physics : PCCP. 2007 ;9(5):573-81.Available from: http://www.ncbi.nlm.nih.gov/pubmed/17242738

189. Teyra J, Pisabarro MT. Characterization of interfacial solvent in protein complexes and contribution of wet spots to the interface description. [Internet]. Proteins. 2007 Jun ;67(4):1087-95.[cited 2010 Jul 12] Available from: http://www.ncbi.nlm.nih.gov/pubmed/17397062

190. Palencia A, Camara-Artigas A, Pisabarro MT, Martinez JC, Luque I. Role of interfacial water molecules in proline-rich ligand recognition by the Src homology 3 domain of Abl. [Internet]. The Journal of biological chemistry. 2010 Jan ;285(4):2823-33.[cited 2010 Jun 28] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2807336&tool=pmcentrez&rendertype=abstract

191. Jaramillo A, Wodak SJ. Computational protein design is a challenge for implicit solvation models. [Internet]. Biophysical journal. 2005 Jan ;88(1):156-71.Available from: http://www.ncbi.nlm.nih.gov/pubmed/15377512

192. Jiang L, Kuhlman B, Kortemme T, Baker D. A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. [Internet]. Proteins. 2005 Mar ;58(4):893-904.Available from: http://www.ncbi.nlm.nih.gov/pubmed/15651050

193. Ashworth J, Taylor GK, Havranek JJ, Quadri SA, Stoddard BL, Baker D. Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. [Internet]. Nucleic acids research. 2010 Sep ;38(16):5601-8.[cited 2011 Feb 7] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2938204&tool=pmcentrez&rendertype=abstract

194. Thyme SB, Jarjour J, Takeuchi R, Havranek JJ, Ashworth J, Scharenberg AM, Stoddard BL, Baker D. Exploitation of binding energy for catalysis and design. [Internet]. Nature. 2009 Oct ;461(7268):1300-4.[cited 2010 Jul 27] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2771326&tool=pmcentrez&rendertype=abstract

195. Ashworth J, Baker D. Assessment of the optimization of affinity and specificity at protein-DNA interfaces. [Internet]. Nucleic acids research. 2009 Jun ;37(10):e73.[cited 2010 Jul 6] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2691843&tool=pmcentrez&rendertype=abstract

196. Bolon DN, Grant RA, Baker TA, Sauer RT. Specificity versus stability in computational protein design. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2005 Sep ;102(36):12724-9.[cited 2010 Oct 2] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1200299&tool=pmcentrez&rendertype=abstract

197. Yosef E, Politi R, Choi MH, Shifman JM. Computational design of calmodulin mutants with up to 900-fold increase in binding specificity. [Internet]. Journal of molecular biology. 2009 Feb 6;385(5):1470-80.[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/18845160

198. Henzler-Wildman K, Kern D. Dynamic personalities of proteins. [Internet]. Nature. 2007 Dec ;450(7172):964-72.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18075575

199. Kamerlin SCL, Warshel A. Reply to Karplus: Conformational dynamics have no role in the chemical step [Internet]. Proceedings of the National Academy of Sciences. 2010 Apr ;107(17):E72-E72.[cited 2010 Aug 24] Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1002658107

200. Karplus M. Role of conformation transitions in adenylate kinase. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2010 Apr ;107(17):E71; author reply E72.[cited 2010 Aug 24] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2867848&tool=pmcentrez&rendertype=abstract

201. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. [Internet]. Nature. 2007 Dec ;450(7171):913-6.[cited 2010 Aug 24] Available from: http://www.ncbi.nlm.nih.gov/pubmed/18026087

202. Henzler-Wildman KA, Thai V, Lei M, Ott M, Wolf-Watz M, Fenn T, Pozharski E, Wilson MA, Petsko GA, Karplus M, Hübner CG, Kern D. Intrinsic motions along an enzymatic reaction trajectory. [Internet]. Nature. 2007 Dec ;450(7171):838-44.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18026086

203. Morin A, Kaufmann KW, Fortenberry C, Harp JM, Mizoue LS, Meiler J. Computational design of an endo-1,4-{beta}-xylanase ligand binding site. [Internet]. Protein engineering, design & selection : PEDS. 2011 Feb 24;[cited 2011 Apr 11] Available from: http://www.ncbi.nlm.nih.gov/pubmed/21349882

204. Mobley DL, Dill KA. Binding of small-molecule ligands to proteins: “what you see” is not always “what you get”. [Internet]. Structure (London, England : 1993). 2009 ;17(4):489-98.Available from: http://www.ncbi.nlm.nih.gov/pubmed/19368882

205. Roy J, Laughton CA. Long-Timescale Molecular-Dynamics Simulations of the Major Urinary Protein Provide Atomistic Interpretations of the Unusual Thermodynamics of Ligand Binding [Internet]. Biophysical Journal. 2010 Jul ;99(1):218-226.[cited 2010 Jul 7] Available from: http://linkinghub.elsevier.com/retrieve/pii/S0006349510004212

206. Ming D, Wall ME. Interactions in native binding sites cause a large change in protein dynamics. [Internet]. Journal of molecular biology. 2006 ;358(1):213-23.Available from: http://www.ncbi.nlm.nih.gov/pubmed/16513135

207. Mittermaier A, Kay LE. New tools provide new insights in NMR studies of protein dynamics. [Internet]. Science (New York, N.Y.). 2006 Apr ;312(5771):224-8.[cited 2010 Jul 27] Available from: http://www.ncbi.nlm.nih.gov/pubmed/16614210

208. Wales TE, Engen JR. Hydrogen exchange mass spectrometry for the analysis of protein dynamics. [Internet]. Mass spectrometry reviews. 2006 ;25(1):158-70.[cited 2010 Aug 12] Available from: http://www.ncbi.nlm.nih.gov/pubmed/16208684

209. Collet E. Dynamical structural science. [Internet]. Acta crystallographica. Section A, Foundations of crystallography. 2010 Mar ;66(Pt 2):133-4.[cited 2010 Jul 28] Available from: http://www.ncbi.nlm.nih.gov/pubmed/20164636

210. Cho HS, Dashdorj N, Schotte F, Graber T, Henning R, Anfinrud P. Protein structural dynamics in solution unveiled via 100-ps time-resolved x-ray scattering [Internet]. Proceedings of the National

Academy of Sciences. 2010 Apr ;107(16):7281-7286.Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1002951107

211. Cammarata M, Levantino M, Schotte F, Anfinrud PA, Ewald F, Choi J, Cupane A, Wulff M, Ihee H. Tracking the structural dynamics of proteins in solution using time-resolved wide-angle X-ray scattering. [Internet]. Nature methods. 2008 Oct ;5(10):881-6.[cited 2010 Aug 12] Available from: http://www.ncbi.nlm.nih.gov/pubmed/18806790

212. Liang S, Li L, Hsu W-L, Pilcher MN, Uversky V, Zhou Y, Dunker AK, Meroueh SO. Exploring the molecular design of protein interaction sites with molecular dynamics simulations and free energy calculations. [Internet]. Biochemistry. 2009 Jan ;48(2):399-414.[cited 2010 Aug 12] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2754190&tool=pmcentrez&rendertype=abstract

213. Liu L, Koharudin LMI, Gronenborn AM, Bahar I. A comparative analysis of the equilibrium dynamics of a designed protein inferred from NMR, X-ray, and computations. [Internet]. Proteins. 2009 Dec ;77(4):927-39.[cited 2010 Aug 12] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2767477&tool=pmcentrez&rendertype=abstract

214. Kamp MW van der, Schaeffer RD, Jonsson AL, Scouras AD, Simms AM, Toofanny RD, Benson NC, Anderson PC, Merkley ED, Rysavy S, Bromley D, Beck DAC, Daggett V. Dynameomics: a comprehensive database of protein dynamics. [Internet]. Structure (London, England : 1993). 2010 Mar ;18(4):423-35.Available from: http://www.ncbi.nlm.nih.gov/pubmed/20399180

215. Lee RA, Razaz M, Hayward S. The DynDom database of protein domain motions. [Internet]. Bioinformatics (Oxford, England). 2003 Jul ;19(10):1290-1.[cited 2010 Aug 12] Available from: http://www.ncbi.nlm.nih.gov/pubmed/12835274

216. Gerstein M, Krebs W. A database of macromolecular motions. [Internet]. Nucleic acids research. 1998 Sep ;26(18):4280-90.[cited 2010 Aug 12] Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=147832&tool=pmcentrez&rendertype=abstract

217. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. [Internet]. Nature structural biology. 2003 Dec ;10(12):980.[cited 2011 May 4] Available from: http://www.ncbi.nlm.nih.gov/pubmed/14634627

218. Li L, Dantzer JJ, Nowacki J, O'Callaghan BJ, Meroueh SO. PDBcal: a comprehensive dataset for receptor-ligand interactions with three-dimensional structures and binding thermodynamics from isothermal titration calorimetry. [Internet]. Chemical biology & drug design. 2008 ;71(6):529-32.Available from: http://www.ncbi.nlm.nih.gov/pubmed/18482338