

January 2010

Examination of Molecular Recognition in Protein-Ligand Interactions

Yat Tang

Washington University in St. Louis

Follow this and additional works at: <https://openscholarship.wustl.edu/etd>

Recommended Citation

Tang, Yat, "Examination of Molecular Recognition in Protein-Ligand Interactions" (2010). *All Theses and Dissertations (ETDs)*. 341.
<https://openscholarship.wustl.edu/etd/341>

This Dissertation is brought to you for free and open access by Washington University Open Scholarship. It has been accepted for inclusion in All Theses and Dissertations (ETDs) by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS
School of Arts and Sciences
Division of Biology and Biomedical Sciences
Computational Biology Program
Department of Biochemistry and Molecular Biophysics

Thesis Examination Committee:
Garland R. Marshall, Chair
Rajendra S. Apte
James J. Havranek
Kevin D. Moeller
Jay W. Ponder
Gary D. Stormo

EXAMINATION OF MOLECULAR RECOGNITION IN PROTEIN-LIGAND
INTERACTIONS: PROSPECTIVE STUDY AND METHODS DEVELOPMENT

by

Yat T. Tang

A dissertation presented to the Graduate School of Arts and Sciences
of Washington University in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2010
Saint Louis, Missouri

copyright by

Yat T. Tang

2010

ABSTRACT OF THE DISSERTATION

Examination of Molecular Recognition in Protein-Ligand Interactions: Prospective
Study and Methods Development

by

Yat T. Tang

Doctor of Philosophy in Computational Biology

Washington University in St. Louis, 2010

Research Advisor: Professor Garland R. Marshall

This dissertation is a compilation of two main projects that were investigated during my thesis research. The first project was a prospective study which identified and characterized drug-like inhibitors of a prototype of bacterial two-component signal transduction response regulator using computational and experimental methods. The second project was the development and validation of a scoring function, PHOENIX, derived using high-resolution structures and calorimetry measurements to predict binding affinities of protein-ligand interactions.

Collectively, my thesis research aimed to better understand the underlying driving forces and principles which govern molecular recognition and molecular design. A prospective study coupled computational predictions with experimental validation resulted in the discovery of first-in-class inhibitors targeting a signal transduction module important for bacterial virulence. Development and validation of the PHOENIX scoring function for binding affinity prediction derived using high-resolution structures

and calorimetry measurements should guide future molecular recognition studies and endeavors in computer-aided molecular design.

To request for an electronic copy of this dissertation, please email the author (yattang at gmail dot com).

Acknowledgments

I am very grateful to have the privilege to pursue scientific research, and to be a part of a stimulating and collaborative environment like the Center for Computational Biology and Washington University.

First and foremost, I would like to express my gratitude to Garland. Thank you for your patience, and for providing me the freedom and a relaxed, yet challenging environment to pursue a broad scope of projects. I've learned from the tremendous breadth and depth of your 'scar tissues' the past 5 years, spanning lessons in academia, science, and life to mature into an independent scientist and a better person. Also, I thank you for your unwavering support and words of encouragement. I will continue to strive to develop into the man and trailblazer that you are.

Thank you to past and present members of the Marshall Research Lab, Center for Computational Biology, and colleagues and collaborators who have been instrumental in making this achievement possible. Malcolm for tremendous computational support and helpful technical advice; Christy for serving as a catalyst for both computational and experimental aspects of this work, and also for providing career guidance; JW, Dan, and Chris for maintaining computational resources; Sage, Gregory, Xiang, Eric, Jon, Hata for scientific discussions; Yaniv and Abby for teaching me peptide and synthetic chemistry; Matt W. for significant help with the formatting of this dissertation; Ann, Rong and other Stock Lab members for scientific advice and contributions; Eduardo for providing lab resources and advice; Groisman Lab members (Wonsik, Tammy, Chris, Deb, Pablo, Mike, Ephraim, Kerry, Eunjin, Tricia, Guy, Varsha, Vero, Alex, and Liesl) for technical advice, assistance, and providing a stimulating and enjoyable environment; Havranek Lab members (Adam and Chi) for technical assistance; Jeff for always being open to provide resources and assistance; and Kaveri for help with performing biological assays. My deepest apologies to those whom I have inadvertently left off.

To past and present thesis committee members: Gary for providing valuable insight, Jim for significant help with protein purification, assay development, and scientific insights, Nathan for rigorous questioning and chairing the meetings, Kevin for providing advice and critical reading of this dissertation, Jay for insightful comments, Raj for willingness to provide experimental resources.

Gratitude is expressed to the PhRMA Foundation, WUSM Vision Sciences, and WUSM Siteman Cancer Center for partial financial support of this accomplishment.

To my parents, William and Joyce, for their unwavering support. My brothers Terry and Barry for holding down the family fort while I embarked on this academic adventure. Nancy, for her never-ending support, patience, and understanding.

Yat T. Tang

Washington University in Saint Louis
August 2010

This dissertation is dedicated to my parents
William and Joyce.

Contents

Abstract	ii
Acknowledgments	iv
List of Tables	x
List of Figures	xi
1 Introduction	1
1.1 Molecular Recognition	2
1.1.1 Driving Forces: Enthalpy and Entropy	3
1.1.2 Entropic Enhancements Through Ligand Preorganization	6
1.1.3 Enthalpy-Entropy Compensation	9
1.1.4 Challenges in Entropy Estimation	10
1.2 Computational Methods for Molecular Design	12
1.2.1 Computational Methods for Molecular Design	12
1.2.2 Structure-Based Drug Design	13
1.3 Molecular Docking and Virtual Screening	14
1.3.1 Motivation	14
1.3.2 Methodology	15
1.3.3 Applications	15
1.3.4 Limitations and Practical Considerations	16
1.4 Binding Affinity Predictions	18
1.4.1 Computational Methods for Affinity Predictions	19
1.4.2 Empirical Scoring Functions	19
1.4.3 Applications	20
1.4.4 Limitations and Practical Considerations	21
1.5 Experimental Data for Methods Development and Validation	21
1.5.1 Biomolecular Structures Derived by X-ray Crystallography	22
1.5.2 Thermodynamic Measurements by Isothermal Titration Calorimetry	23
1.5.3 Protein-Ligand Complexes for Scoring Function Development	24
1.5.4 Inherent Inaccuracies in Experimental Data	25
1.6 Synopsis of Thesis	25
1.6.1 Summary of Challenges	26
1.6.2 Discoveries and Insights Presented	27

2	Background	28
2.1	Empirical Scoring Functions	29
2.1.1	Training Set	29
2.1.2	Descriptors	30
2.1.3	Linear Regression Method	31
2.2	Descriptors to Estimate Entropy Changes	32
2.2.1	Rotamers and Hydrophobicity	32
2.2.2	Shape- and Volume-Based Descriptors	33
2.3	Protein-Protein Interactions	34
2.3.1	Hot Spots	35
2.3.2	Challenges in Targeting PPI	36
2.3.3	Examples of Inhibitors	37
2.4	Bacterial Two-Component Signal Transduction Systems	41
2.4.1	Mechanism	41
2.4.2	OmpR/PhoB Structural Subfamily	42
2.4.3	<i>Salmonella enterica</i> PhoP Response Regulator	43
2.4.4	Drug Discovery	47
3	Discovery of PhoP Response Regulator Inhibitors	48
3.1	Introduction	48
3.2	Methods and Materials	52
3.2.1	Overview of Computational Strategy	52
3.2.2	Compound Library and Crystal Structure	54
3.2.3	Structure-Based Virtual Screening	54
3.2.4	Overview of Experimental Strategy	57
3.2.5	Electrophoretic Mobility-Shift Assays (EMSA)	58
3.2.6	Similarity Search	60
3.2.7	Native Polyacrylamide Gel Electrophoresis (PAGE)	60
3.2.8	Size-Exclusion Chromatography (SEC)	60
3.2.9	Förster resonance energy transfer (FRET) analyses	61
3.2.10	<i>S. enterica</i> PhoP Expression and Purification	62
3.3	Results	63
3.3.1	Computational Predictions	63
3.3.2	Experimental Validation and Characterization	63
3.4	Discussion	67
3.4.1	Selective Inhibition of <i>S. enterica</i> PhoP	67
3.4.2	Potential Modes of Action	67
3.4.3	Challenges of Structure-Based Design	68
3.4.4	Discovery of First-In-Class PhoP Inhibitors	69
3.5	Conclusions and Future Directions	69
3.5.1	Summary	70
3.5.2	Future Directions	70
3.6	Acknowledgements	70

4	PHOENIX: Scoring Function to Predict Binding Affinities	72
4.1	Introduction	72
4.2	Materials and Methods	75
4.2.1	Training Set	75
4.2.2	Structure Preparation	76
4.2.3	External Test Sets	80
4.2.4	Descriptors Set	81
4.2.5	Function Parameterization	83
4.3	Results	83
4.3.1	Regression Analysis	83
4.3.2	Internal Cross-Validation	87
4.3.3	Testing on External Data Sets	89
4.3.4	PHOENIX Scoring Function	100
4.4	Discussion	103
4.5	Conclusion	106
4.6	Acknowledgements	107
5	Molecular Recognition and Structure-Based Drug Design	108
5.1	Summary of Results and Discussion	108
5.1.1	PhoP Response Regulator Inhibitors	108
5.1.2	PHOENIX Scoring Function for Binding Affinity Predictions	109
5.2	Future Directions	110
5.2.1	Further Characterization and Design of PhoP Inhibitors	110
5.2.2	Advances to Achieve Accurate Binding Affinity Predictions	110
5.3	Conclusion	111
5.3.1	Targeting Protein-Protein Interactions for Drug Discovery	111
5.3.2	Binding Affinity Prediction	111
5.3.3	Future of Computer-Aided Molecular Design	112
	References	113
	Curriculum Vitae	123

List of Tables

4.1	Coefficients and intercepts derived from partial least squares regression for the descriptor set used in the final PHOENIX scoring function. . .	78
4.2	Descriptor set used in the final PHOENIX scoring function and its relative contributions.	79
4.3	Partial least squares (PLS) regression statistics of the change in enthalpy (ΔH).	84
4.4	Partial least squares (PLS) regression statistics of the change in entropy ($T\Delta S$).	84
4.5	Partial least squares (PLS) regression statistics of the change in binding free energy (ΔG).	85
4.6	Correlation statistics of the PHOENIX scoring function using different training sets on the PDBbind v2002 database.	90
4.7	Correlation statistics of the PHOENIX scoring function compared to other commonly used scoring functions on the PDBbind v2002 set. . .	91
4.8	Correlation statistics of the PHOENIX scoring function compared to X-Score scoring functions on high- ($0 \leq 2 \text{ \AA}$) and low-resolution ($2 \leq 2.5 \text{ \AA}$) complexes of the PDBbind 2002 set.	92
4.9	Correlation statistics of the PHOENIX scoring function compared to other commonly used scoring functions on HIV-1 protease complexes of the PDBbind 2002 set.	94
4.10	Correlation statistics of the PHOENIX scoring function compared to other commonly used scoring functions on trypsin complexes of the PDBbind 2002 set.	95
4.11	Correlation statistics of the PHOENIX scoring function compared to other commonly used scoring functions on carbonic anhydrase II complexes of the PDBbind 2002 set.	96
4.12	Assessment of the ability of the PHOENIX scoring function to classify complexes into three binding affinity groups of the PDBbind 2002 set. . .	97
4.13	Correlation statistics of the PHOENIX scoring function on the PDBbind 2004 refined set.	98
4.14	Correlation statistics of the PHOENIX scoring function on the PDBbind 2009 refined set.	99
4.15	Correlation statistics of the PHOENIX scoring function on the PDBbind 2007 core set.	101
4.16	Success rates for correctly ranking the low-, medium-, and high-affinity ligands in the PDBbind 2007 core set.	102

List of Figures

1.1	Forces underlying molecular recognition of protein-ligand interactions.	3
1.2	Thermodynamic cycle of protein-ligand interactions.	5
1.3	Diagram of ligand preorganization to enhance affinity.	7
1.4	Hydrocarbon-stapled BH3 peptide for affinity and property enhancement.	8
1.5	Scatter plot of change of enthalpy versus change of entropy.	9
1.6	Enthalpic optimization of HMG-coA reductase and HIV-1 protease inhibitors to achieve best-in-class compounds.	10
1.7	Effect of water molecule on thermodynamic parameters.	17
2.1	Examples of how a protein interacts with its natural protein partner and with a small molecule.	38
2.2	Examples of small-molecule inhibitors of protein-protein interactions.	39
2.3	Schematic diagram of a typical two-component signal transduction (TCST) system.	42
2.4	Dimer structure of the PhoB receiver domain, a member of the OmpR/PhoB subfamily.	43
2.5	Sequence alignment of the $\alpha 4$ - $\beta 5$ - $\alpha 5$ region of <i>E. coli</i> response regulators.	44
2.6	Schematic diagram of the <i>Salmonella enterica</i> PhoQ/PhoP two-component signal transduction system.	45
2.7	Schematic diagram of the <i>Salmonella enterica</i> PhoQ/PhoP two-component signal transduction system.	46
3.1	Schematic diagram of the <i>Salmonella enterica</i> PhoQ/PhoP two-component signal transduction system.	50
3.2	Schematic diagram of the <i>Salmonella enterica</i> PhoQ/PhoP two-component signal transduction system.	51
3.3	Schematic diagram of the computational workflow to predict for PhoP response regulator inhibitors.	53
3.4	Schematic diagram of the experimental workflow to test for PhoP response regulator inhibitors.	53
3.5	Docking grids for the PhoP interface used for virtual screening. . . .	56
3.6	Chemical structures of 8 PhoP inhibitors.	64
3.7	EMSA dose-response curves for PhoP inhibitors.	65
3.8	Native PAGE results of 8 PhoP inhibitors.	65
3.9	FRET results of 8 PhoP inhibitors.	66

4.1	Distribution histogram of the change in binding free energy (ΔG) of complexes in the final PHOENIX training set.	76
4.2	Distribution histogram of the change in enthalpy (ΔH) of complexes in the final PHOENIX training set.	77
4.3	Distribution histogram of the change in entropy ($T\Delta S$) of complexes in the final PHOENIX training set.	77
4.4	Distribution histogram of the molecular weight (Da) of complexes in the final PHOENIX training set.	80
4.5	Calculated versus experimental values for change in enthalpy (ΔH) of the final PHOENIX training set.	85
4.6	Calculated versus experimental values for change in entropy ($T\Delta S$) of the final PHOENIX training set.	86
4.7	Calculated versus experimental values for change in binding free energy (ΔG) of the final PHOENIX training set.	86
4.8	Calculated versus experimental values for change in enthalpy (ΔH) of the internal cross-validation test set.	88
4.9	Calculated versus experimental values for change in entropy ($T\Delta S$) of the internal cross-validation test set.	88
4.10	Calculated versus experimental values for change in binding free energy (ΔG) of the internal cross-validation test set.	89
4.11	Calculated versus experimental binding affinities ($-\log K_d$) of the PDB-bind 2009 refined set.	99

Chapter 1

Introduction

Protein-ligand interactions form the molecular basis of many ubiquitous and essential biological functions. Molecular recognition of protein-ligand interactions governs the affinity and specificity of complex formation and determines their biological functions. Gaining a better qualitative and quantitative understanding of the physical forces underlying protein-ligand interactions provides a rational guide to therapeutic design. Understanding molecular recognition in protein-ligand interactions is, therefore, of enormous scientific and practical importance.

Computer-aided molecular design (CAMD) utilizes the principles of molecular recognition by taking advantage of increasing computational power to develop and apply various theoretical models for molecular discovery and design. Through the years, CAMD have demonstrated to be effective and instrumental in guiding drug discovery and molecular recognition. Due to the resources, time, and other practical limitations required for performing experimental methods, computational tools that are fast and sacrifice some accuracy are valuable for molecular discovery and design.

Overview

This introductory chapter is organized as follows: the first section is an introduction to molecular recognition in protein-ligand interactions, the second section introduces computer-aided molecular design, the third section introduces molecular docking and virtual screening, the fourth section introduces methods for binding affinity prediction, the fifth section presents the resources available for the development and testing of computational tools, the sixth section provides a synopsis of this thesis.

This thesis focuses on the examination of molecular recognition in protein-ligand interactions through the development, validation, and application of molecular modeling tools for lead discovery. A prospective study which resulted in the discovery of first-in-class drug-like compounds inhibiting a prototype of bacterial two-component signal transduction response regulators provided insight into the current limitations of computational methods for design studies targeting protein-protein interactions. Development and validation of the PHOENIX scoring function demonstrated that models derived using high-resolution structures and calorimetry measurements may be a key advance towards more accurate binding affinity predictions. These investigations provide a deeper understanding into the underlying principles governing molecular recognition in protein-ligand interactions. A better understanding of molecular recognition will guide the development of more accurate and applicable computational methods for molecular design.

1.1 Molecular Recognition

Molecular recognition is a critical phenomenon observed in a variety of biological systems (e.g., protein-ligand, protein-protein, protein-DNA interactions) that generates both specificity and duration of action. Molecular recognition is defined as the specific interaction between two or more molecules through non-covalent bonding, such as electrostatics, hydrogen-bonding, and van der Waals forces (see Fig: 1.1).

Molecular recognition determines and regulates essential cellular interactions such as the binding of hormones to receptors for signaling, protein-protein interactions to detect and respond to various external stimuli as part of a signaling cascade, and small-molecules reversibly binding and inhibiting the function of an enzyme or receptor for therapeutic purposes. Understanding the principles of molecular recognition, specifically the underlying physical forces which govern affinity and specificity, is therefore of great fundamental significance in biomedical research.

This section will primarily discuss methods for estimating entropic changes in protein-ligand interactions, since current molecular design methods implement a limited, at best, representation of the entropic changes associated with complex formation. Lack of accurate entropic descriptions may be the main culprit for inaccuracies in current

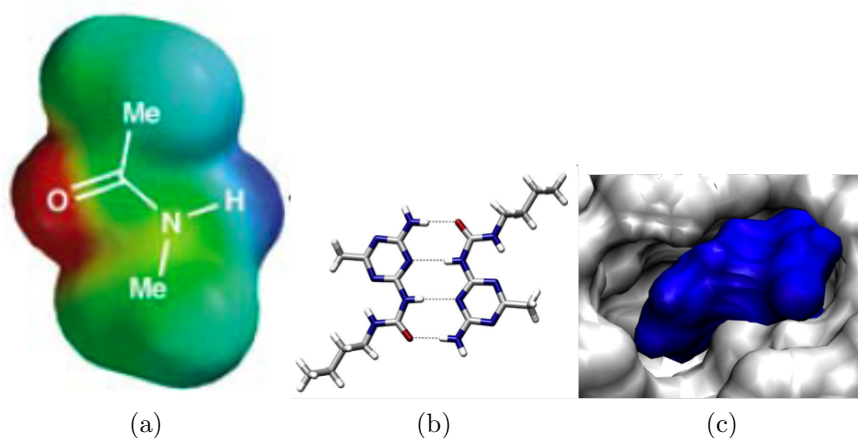


Figure 1.1: (a) Molecular electrostatic potential surfaces plotted on the van der Waals' surface of the molecule calculated by using AM1 of N-methyl acetamide. Positive regions are shown in blue, negative regions are shown in red, and green is neutral. Figure taken from Hunter. (Hunter, 2004) (b) Hydrogen-bonding between 2 molecules (shown in stick representation, oxygen in red, nitrogen in blue hydrogen in white, carbon in gray, respectively). Hydrogen-bonds are formed when the geometries (distance and angle) are ideal. (c) van der Waals interactions between ligand (blue surface) and protein binding site (gray surface).

computer-aided, molecular design methodology. Methods to better capture entropic changes should lead to more accurate methods for molecular design.

1.1.1 Driving Forces: Enthalpy and Entropy

Affinity and specificity of protein-ligand interactions is determined by the change in binding free energy of the complex compared with other potential targets. Change in binding free energy (ΔG°) is composed of two independent thermodynamics forces: change in enthalpy (ΔH°) and change in entropy (ΔS°) written as,

$$\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ \quad (1.1)$$

In protein-ligand interactions, the change in enthalpy is primarily composed of van der Waals interactions, electrostatics, and hydrogen bonding. Binding entropy is a result of the changes in the disorder, or the degrees of freedom of the system. Change

in entropy of the system (protein, ligand, and solvent) can be largely attributed to the solvation and desolvation energies, hydrophobic features, and conformational and configurational changes of both ligand and receptor.

Change in enthalpy (ΔH) of protein-ligand interactions is derived primarily from steric complementarity, electrostatics, and hydrogen-bonding. Forces contributing to steric complementarity include van der Waals interactions, and hydrophobic and hydrophilic surface complementarity. Electrostatic interactions arise from the distance pairing of complementary positively- and negatively-charged groups. Hydrogen-bonding arise from the geometric constraints of hydrogen-bond donors and acceptors.

Change in entropy (ΔS) in protein-ligand interactions is derived mostly from solvation and desolvation, and the degrees of freedom of both the ligand and protein during complex formation. Energy is expended (Fig: 1.2) during the process in which the ligand is transferred from the hydrophilic environment of the solvent to the predominantly hydrophobic environment of the binding site.

Water molecules solvating the binding site cavity are typically entropically unfavorable due to the conformational constraints of the binding site surface, which hinders its ability to form hydrogen bonds. Displacement of bound waters from the binding site to the external environment leads to favorable entropic changes, a primary reason why ligands are designed to be hydrophobic to complement a hydrophobic binding site (and also favorable energetics for getting out of solvent). Ligands designed in this manner generate favorable entropy upon complex formation, and thus enhances its binding affinity.

Change in entropy of the ligand (ΔS_{ligand}) consists of conformational entropy (from accessible rotamers), configurational entropy (translational and rotational), and vibrational entropy (minimal contributions and computationally expensive to model) Chang (2007) as follows,

$$\Delta S_{ligand} = \Delta S_{conformation} + \Delta S_{configurational} + \Delta S_{vibrational}. \quad (1.2)$$

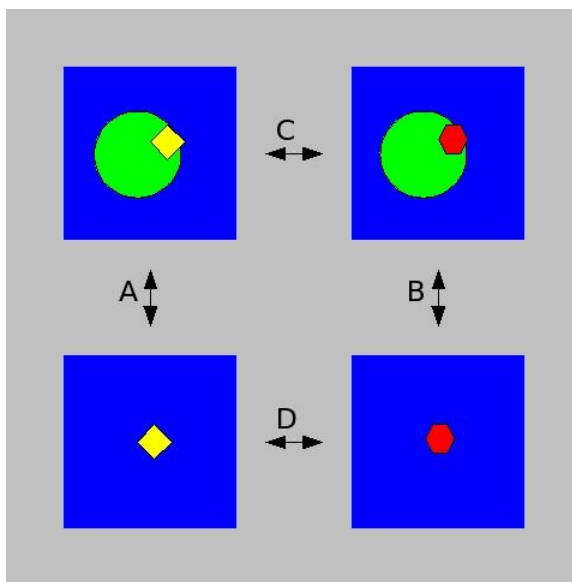


Figure 1.2: Thermodynamic cycle of protein-ligand interactions. Solvation energy changes for 4 states are estimated to derive the change in solvation energy in protein-ligand interactions. (A) Change in solvation energy of the unbound conformation ligand (yellow diamond) and bound protein-ligand complex (protein in green circle). (B) Change in solvation energy of the bound conformation ligand (red hexagon) and bound protein-ligand complex. (C) Change in solvation energy of the unbound conformation ligand-protein complex and bound conformation ligand-protein complex. (D) Change in solvation energy of the unbound conformation ligand and bound conformation ligand.

Change in entropy of the protein ($\Delta S_{protein}$) is also composed of the same components as the ones for the ligand. However, due to necessary simplification for model development, change in entropy of the protein ($\Delta S_{protein}$) is typically not considered, based on the assumption that the change in entropy of the protein remains constant when bound to different ligands (which may not necessarily be true due to entropy-entropy compensation, see Section 1.1.4).

Change in enthalpy (ΔH) and change in entropy (ΔS) are intimately related in their contributions to change in binding free energy. Based on experimental and theoretical studies, enhancing entropy may lead to less favorable enthalpy, and vice versa, therefore canceling out the improved thermodynamic force and resulting in a minimal change in binding free energy. This observation, known as enthalpy-entropy compensation, will be discussed in a greater detail in Section 1.1.3.

To improve ligand binding affinity, a common and effective molecular design strategy to enhance the change in entropy, known as preorganization, is used to result in more favorable overall change in entropy. Preorganization will be discussed in the following section.

1.1.2 Entropic Enhancements Through Ligand Preorganization

In molecular design, a common strategy to optimize ligand affinity is by preorganization to lessen the entropic penalty for protein-ligand complex formation. Preorganization aims to prestabilize the bound conformation (e.g., reduce the number of rotamers of the ligand) in order to decrease the entropic penalty upon complex formation, resulting in an enhanced binding affinity.

A number of studies have demonstrated that rigidifying (decreasing the number of rotamers to minimize flexibility) a ligand or peptide enhances its binding affinity (0.7-1.6 kcal/mol for each restricted rotor) (Gerhard et al., 1993; Searle and Williams, 1992), at times by several orders of magnitude (Sawyer, presumably by decreasing the entropic penalty upon ligand binding. (Fig: 1.3) (DeLorbe et al., 2009)

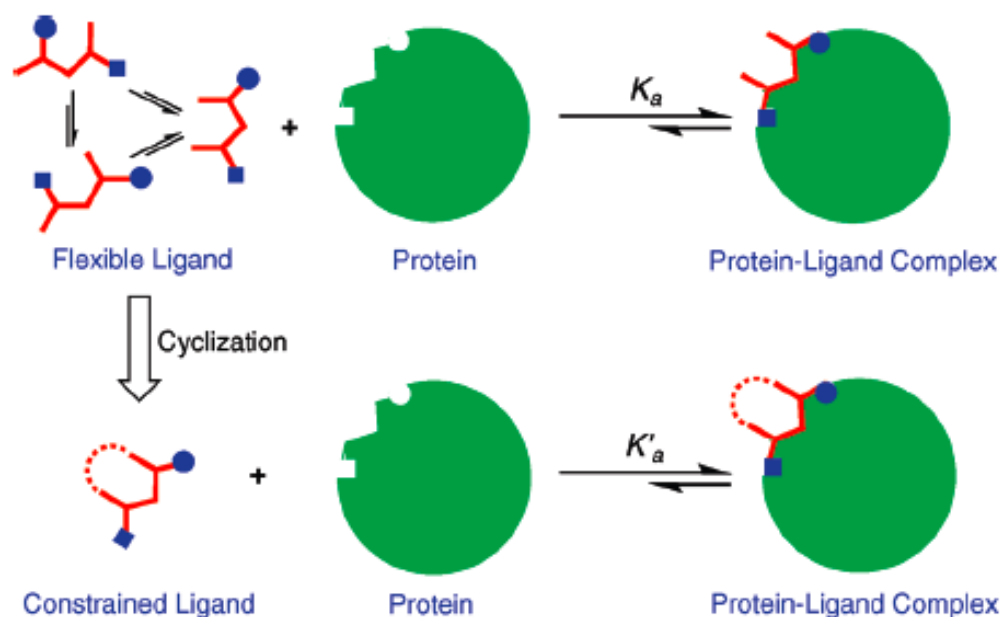


Figure 1.3: Diagram of the energetic effects associated with ligand preorganization. Cyclization of a flexible ligand limits the degrees of freedom and reduces the number of conformational isomers in solution. The likelihood that of the to be in the biologically active conformation in enhanced, resulting in a more favorable entropy of binding, assuming that the flexible and rigid ligands are interacting with the solvent and protein in the same manner. Figure taken from DeLorbe et al. (2009).

Examples of preorganization strategies to enhance affinity include replacing rotatable bonds with ring structures to decrease flexibility, 'cyclizing' of peptides, 1.3 and 'hydrocarbon-stapling' of peptides 1.4. (Walensky et al., 2004)

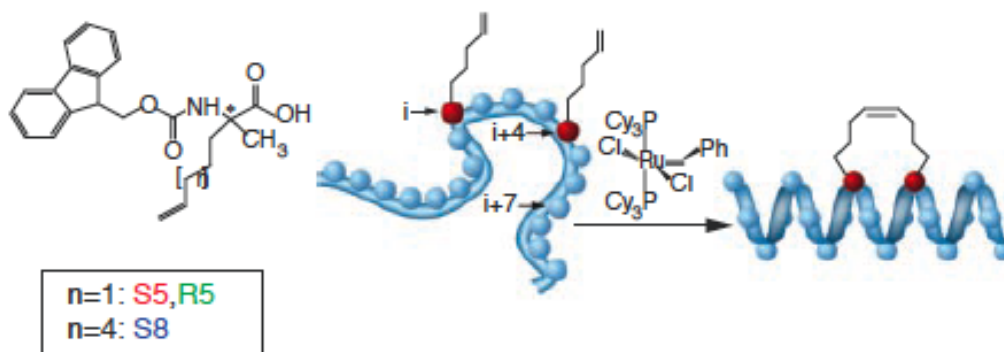


Figure 1.4: Enhanced helicity, protease resistance, and serum stability of hydrocarbon-stapled peptide representing the BH3 helix. Figure taken from Walensky et al. (2004).

Recent calorimetric studies have suggested a paradox to this accepted notion. (Martin, 2007) What presumably was entropy optimization by ligand preorganization have actually resulted in enthalpy optimization. DeLorbe et al. (2009) have shown that cyclizing a series of pseudopeptides to enhance entropy upon binding to Grb2 SH2 domain resulted in enhancement of enthalpy instead. (DeLorbe et al., 2009) This may be due to the fact that more direct polar contacts resulted from the ligand preorganization, which explains the enhanced relative enthalpy. Further calorimetric studies on different series of ligands and targets will need to be performed to assess if this study is an exception, or if the observed binding energetics (favorable binding enthalpy and unfavorable binding entropy) occurs more commonly than expected.

Preorganization does not necessarily enhance changes in relative configurational entropy (translational and rotational entropy) which has recently been demonstrated to contribute a significant portion to relative binding entropy in theoretical studies on HIV-1 protease inhibitors. (Chang, 2007)

1.1.3 Enthalpy-Entropy Compensation

A phenomenon known as enthalpy-entropy compensation (EEC) hinders the effects of optimizing each individual thermodynamic force for enhancing binding free energies. EEC was first observed as a linear relationship between enthalpy and entropy changes in biomolecular interactions. (Lumry and Rajender, 1970; Pimentel and McClennan, 1971; Linert and Jameson, 1989; Weber, 1995; Crunwald and Steel, 1995) (see Fig. 1.5)

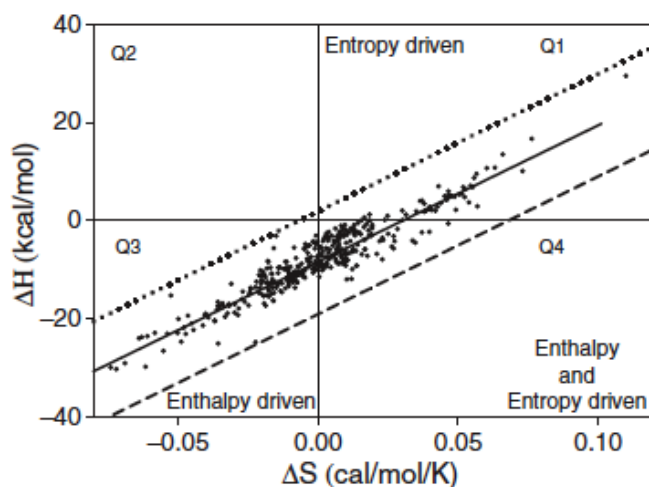


Figure 1.5: Enthalpy-entropy compensation in protein-ligand interactions. The dotted and dashed lines correspond to the upper and lower limit of the dissociation constant. The solid line corresponds to a linear fit of the data. Figure taken from Li et al. (2008)

Searle and Williams discussed EEC in agonists versus antagonists; binding of agonists are mostly enthalpic driven, binding of antagonists are mostly entropic driven. (Searle and Williams, 1992)

With the increasing use of ITC in molecular design studies, relative binding enthalpy and entropy can now be directly measured to elucidate the thermodynamic contributions to binding free energies. (Ladbury et al., 2010) Using ITC to elucidate thermodynamic contributions of HIV-1 protease and HMG-CoA reductase inhibitors, the Freire group observed that first-in-class inhibitors were entropy optimized. Enthalpy optimization was observed in enhancing binding affinity from first-in-class to

best-in-class. (See Fig: 1.6) These calorimetric studies illustrate the importance of understanding independent entropic and enthalpic contributions for optimizing binding affinity in molecular design.

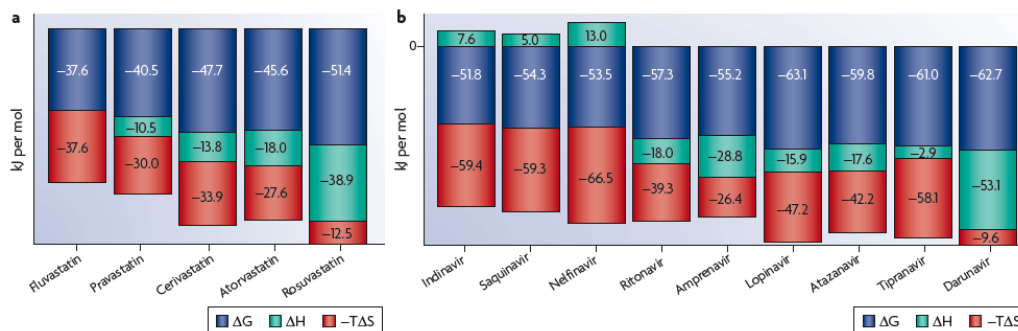


Figure 1.6: (A) Thermodynamic profile of the binding of a series of statins to 3-hydroxy-3-methylglutaryl coenzyme A (HMG-coA) reductase. the sum of the change in enthalpy (ΔH ; green) and the change in entropy (ΔS) multiplied by the absolute temperature (T ; red) gives the change in free energy (ΔG ; blue). Figure taken from Ladbury et al. (2010).

1.1.4 Challenges in Entropy Estimation

To rigorously and accurately estimate the change in entropy, the partition function from statistical mechanics needs to be calculated in order to obtain the statistical properties of the system in thermodynamic equilibrium. Due to the computational expense needed to calculate the partition function, which includes the incorporation of multiple parameters (e.g., temperature, volume of the system, number of constituent particles) and the sampling of various microstates of the system to derive a canonical ensemble (statistical ensemble representing probability distribution of microstates), heuristic methods that are less physically accurate but much faster are instead used to estimate entropic changes.

Physical phenomena contributing to entropic changes that are challenging, and energetically critical, to accurately model by heuristic methods are solvation/desolvation, multiple binding states, entropy-entropy compensation, and configurational entropy.

Solvation and desolvation have long been known to play a critical role in entropic contributions to binding free energy. (Chandler, 2005) A number of theoretical methods utilizing molecular dynamics simulations and implicit solvent representation (e.g., MM-GBSA, MM-PBSA, etc.) (Massova and Kollman, 2000) have been developed and applied in molecular studies to estimate solvation energies involved in protein-ligand interactions. Recent theoretical studies using an explicit solvent representation have also suggested the importance of binding-site water molecules in entropic contributions. These more rigorous studies employed the inhomogeneous solvation theory (Lazaridis, 1998) to estimate the hydration thermodynamics with relatively short (10 ns) molecular dynamics simulations. Abel et al. used explicit solvent simulations to more accurately describe the physical interactions of binding site waters, and subsequently developed descriptors to qualitatively predict affinities of ligands binding to factor Xa with higher accuracy. (Abel et al., 2008) Other studies using coupling inhomogeneous solvation theory and molecular dynamics simulations provided atomic insight into the roles of water molecules and their contributions to entropy and binding free energy. (Young et al., 2007; Huang et al., 2009; Young et al., 2010)

Accurate estimation of the change in entropy upon complex formation requires the incorporation of multiple binding states. In docking, multiple potential ligand binding poses are generated in order to predict the native binding pose. Ruvinsky demonstrated the use of the multiple ligand binding poses from docking experiments to heuristically estimate the partition function. (Ruvinsky and Kozintsev, 2005) This method to estimate the ensemble of complex microstates demonstrated enrichment in binding free energy estimation compared with using a single binding mode, in AutoDock as well as other docking programs. (Ruvinsky, 2007) However, the caveat in this simplified method is that the multiple potential conformations of the binding site is not accounted for, which certainly will lead to inaccuracies in change of entropy ($T\Delta S$) estimates.

Binding site flexibility of either protein or receptor is largely ignored in docking, due to the added computational expense to accurately model the dynamics of a larger molecule. However, this oversimplification used in docking ignores the phenomenon known as entropy-entropy compensation. In entropy-entropy compensation, the binding of the ligand results in unfavorable entropy, leading to entropy compensation of the protein or receptor, typically by increasing its flexibility (conformational dynamics) as

a means to balance the change of entropy of the system. Experimental studies using NMR relaxation have demonstrated this entropy-entropy compensation in calmodulin upon protein domain binding. (Marlow et al., 2010) Due to the use of a rigid receptor in docking, entropic changes arising from entropy-entropy compensation is not described, and therefore contributes to the inaccuracies in entropy estimates.

Configuration entropy has long been known to contribute significantly to change in entropy, but have not been evaluated quantitatively. The Gilson group used molecular dynamics simulations and Mining Minima calculations to estimate the magnitude of configurational entropys contribution to the overall change of entropy in the system for HIV-1 protease inhibitors. (Chang, 2007) Their results suggested that configuration entropy (translational and rotational entropy) was actually a larger component than conformation entropy (ligand and protein rotamers). These studies illustrate the need to design conformationally restricted ligands to enhance affinity, and also the need to better capture changes in configuration entropy in CAMD methods.

1.2 Computational Methods for Molecular Design

Numerous computational and theoretical methods have been developed for molecular design. Methods have focused on the pragmatic issue of balancing computational speed versus accuracy and applicability. As scientific knowledge and computational power have progressed significantly over the last few decades, the computational methods developed have been increasingly fast, applicable, and accurate. Computational methods for molecular design are now commonly employed in both industry and academic settings, and have been recognized to be a valuable, sometimes essential, tool for molecular recognition and ligand design.

1.2.1 Computational Methods for Molecular Design

Computer-aided molecular design (CAMD) is used to describe the application of computational chemistry and molecular modeling methods to molecular design, mainly to expedite and rationalize each of the different stages of the process. As computational

power continues to increase, CAMD continues to progress in terms of the development and application of more physically accurate models to examine more complex biomolecular systems in a greater detail.

One of the most encountered challenges in drug discovery is the identification of potential lead compounds for a novel therapeutic target. For a novel target where there are no known putative ligands or inhibitors that can be used for searching similar chemical analogs, a standard experimental method for lead discovery is to perform a high-throughput screening (HTS). (Bleicher et al., 2003) While obtaining biological activity data for millions of compounds may seem promising in theory, HTS studies are often not practical or as fruitful in practice as one might conceive. Not only are the screens costly and require significant resources (e.g., robotics setup, compounds, solutions, biological target of interest), it also requires a robust biological assay, which may not be readily available for a novel system in the early stages of the study. Moreover, HTS are also prone to identifying false positives, which may lead to significant time and resources spent on characterizing and optimizing these false leads. (McGovern et al., 2002) Due to these requirements and limitations, the use of computational methods for lead discovery is a much more practical early-stage strategy, and is now widely employed to expedite and complement the lead discovery process.

1.2.2 Structure-Based Drug Design

In the case where an X-ray crystal structure of the molecular design target has been solved, structure-based virtual screening can be used to screen large libraries (e.g., ZINC: over 8,000,000 commercially available compounds) (Irwin and Shoichet, 2005) and prioritize compounds for experimental testing. The workflow of virtual screening can be categorized as a two-step process: docking and scoring. In docking, a conformational search is performed on the ligand to identify a set of low-energy conformations that complement the protein binding site (in the case of rigid protein docking). With the predicted binding poses, scoring is used to assess the relative binding free energy of the complexes as a post-processing step to filter out energetically-unfavorable poses based on physical and chemical properties (e.g., steric complementarity, electrostatics, potential for hydrogen bonding). Molecular docking will be presented in a greater

detail in Section 1.3, and binding affinity predictions will be presented in a greater detail in Section 1.4 of this introduction.

1.3 Molecular Docking and Virtual Screening

Molecular docking, known simply as docking, is a computational method which aims to predict the conformation of a ligand bound to a binding site, given the structure of the protein (typically in the form of an X-ray crystal structure). Conformation of the protein is typically kept rigid, although recent docking programs have incorporated binding site dynamics using flexible side chains and protein backbones (e.g., AutoDock (Morris et al., 1998), RosettaLigand (Davis and Baker, 2009), etc.). Conformational search is performed on the ligand to identify a set of low-energy conformations which complement (in terms of shape and electrostatics) the binding site. Virtual screening is the application of molecular docking to predict the binding modes of compounds in a library, and prioritize these compounds based on an estimated binding affinity to identify potential leads. (Walters et al., 1998) Virtual screening has proven to be useful in the early stages of lead discovery. (Shoichet, 2004)

1.3.1 Motivation

Molecular docking serves to rapidly and accurately identify potential low-energy conformations of protein-ligand complexes. Experimental methods to resolve bound complexes at an atomic scale, such as X-ray crystallography or nuclear magnetic resonance techniques (e.g., SAR by NMR (Shuker et al., 1996)) are both labor- and time- intensive, sometimes limiting their applicability in molecular design. Under circumstances where time and resources are limited, virtual docking is a valuable method for providing rapid and relatively accurate structural information for rationalization and visualization.

1.3.2 Methodology

In recent years, significant progress has been made on the development of docking programs. (Ewing and Kuntz, 1997) A typical docking program is composed of three key components: representation of the binding site, the conformation search algorithm for generation of binding poses, and a scoring function used for affinity predictions. The docking programs developed vary in each of these components to some degree, which leads to different performances depending on the protein target and compound library of interest (e.g., AutoDock (Morris et al., 1998), DOCK (Ewing et al., 2001), GOLD (Jones et al., 1997), etc.). Because of this, there is no single best-performing docking method for CAMD; the best-performing method depends heavily on the system of interest and individual parameters set by the user.

One of the more widely used docking programs, and also available for free to academia, is AutoDock (Morris et al., 1998). In AutoDock, the binding site targeted by the user is represented based on a grid method. The binding site is defined by a cubic grid, with varying degrees of resolution (from 0.200 to 0.375 Å). AutoDock encodes the structural information of the ligand as genes, and applies a Lamarckian genetic algorithm to perform the conformational searches. The AutoDock scoring function is based on a semi-empirical method using force field-based terms as descriptors and experimental binding data to derive an equation to estimate the binding free energy of the complex. (Huey et al., 2007) AutoDock has proven to be useful in the hands of a large number of investigators for a variety of structural targets and compound libraries. In addition, a number of preparation and analysis tools have been developed for use with AutoDock, making the use of this docking program more user-friendly than other academically-available ones. (Morris et al., 2009) Because of its proven applicability and robustness in our own as well as in other research groups, AutoDock was selected as the docking method for virtual screening studies in this thesis.

1.3.3 Applications

Molecular docking has been applied to a variety of CAMD applications, most notably for binding pose prediction and lead compound identification. AutoDock has been shown to be useful in retrospective binding pose prediction based on the criteria of

enrichment factor and receiver operator curve (ROC). As discussed in the previous section, AutoDock has been a valuable virtual screening tool for lead discovery as demonstrated in a number of studies. (Goodsell et al., 1996; Osterberg et al., 2002; Park et al., 2006)

1.3.4 Limitations and Practical Considerations

A number of limiting assumptions and oversimplifications are used in most application of docking methods (e.g., rigid binding site, lack of accurate solvent representation, stochastic methods used in conformational searches, limitations in affinity predictions, etc.). Due to these often necessary simplifications, inherent inaccuracies in their implementations have led to a high percentage of false positives in many retrospective and prospective virtual screening studies.

Rigid representation of the protein binding site (no backbone or side-chain flexibility) limits the utility of docking programs to identify novel ligands (ones that are not structural similar to the ligand in the complex used for docking) which may actually bind to the protein by an induced fit mechanism. Cozzini et al. have recently reviewed the different methods to incorporate protein flexibility to achieve better accuracy in SDDD. (Cozzini et al., 2008) The rigid protein assumption is also a crude oversimplification, since the potential structural changes which may lead to significant energetic changes important in protein-ligand binding are completely ignored. In addition, entropy-entropy compensation (discussed in Section 1.1.4) cannot be accounted for using a rigid binding site.

In docking, solvation and desolvation are typically represented implicitly. In protein-ligand interactions where waters play an important role in complex formation (e.g., aldose reductase), docking methods may fail to predict the correct pose simply due to the lack of a correct explicit representation of bound water. To achieve more accurate docking predictions in these cases, explicit waters in the binding site need to be included. Solvation and desolvation effects will need to be represented in a more physically accurate manner to achieve accurate binding affinity estimations. (See Figure 1.7)

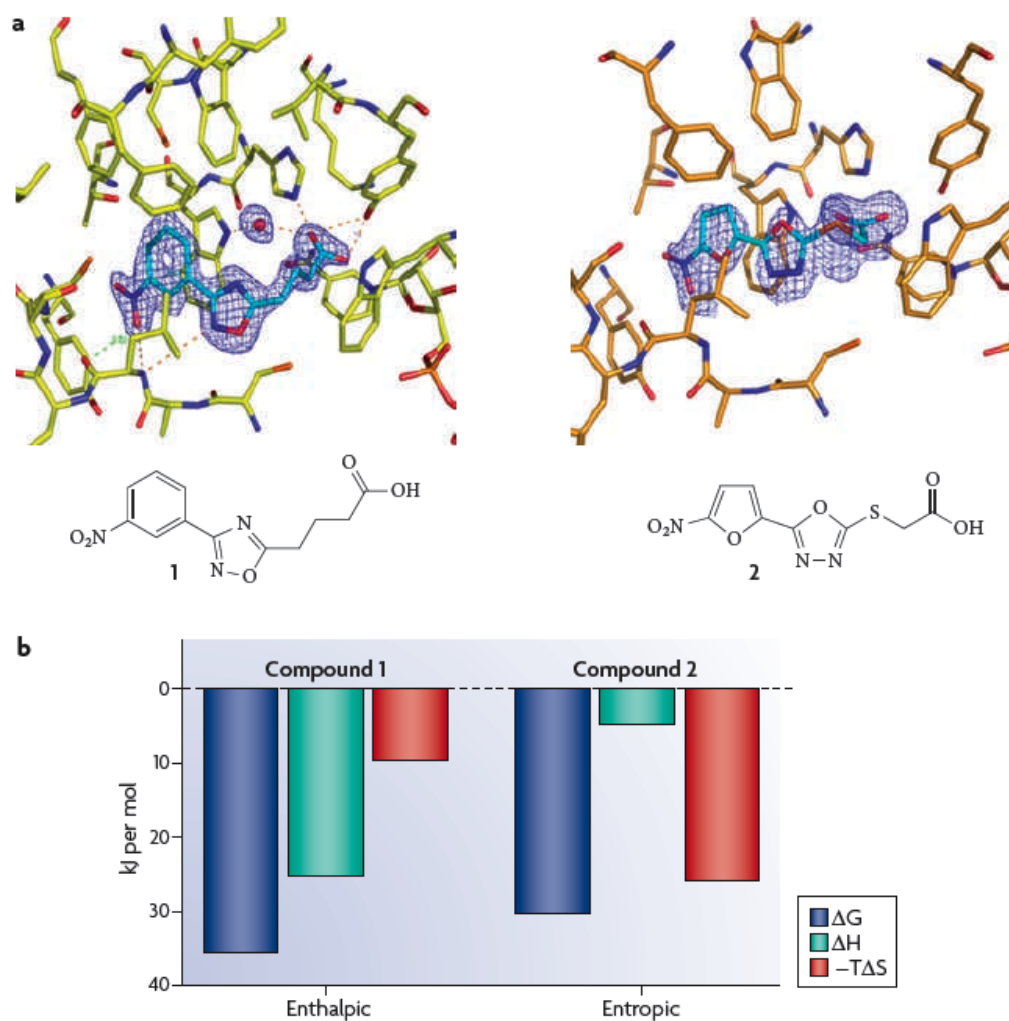


Figure 1.7: (A) Two similar ligands shown differ significantly in their enthalpic and entropic contributions in binding to aldose reductase. The figure on the left contains a water molecule, while the figure on the right does not. (B) Inclusion of a water molecule in the complex on the left results in less favorable entropy (red) by more favorable enthalpy (green), due to additional hydrogen bonds. Binding free energy is shown in blue. Figure taken from Ladbury et al. (2010).

Another limitation in docking, in particular with AutoDock (Morris et al., 1998) and most other programs available to academia, is in the stochastic algorithms used to perform conformations searches. Due to the random nature of these methods, results may not necessarily be replicated even when using the same structures and parameters. To circumvent this limitation, docking tools based on systematic searches (e.g., SKATE (Feng and Marshall, 2010)) can be used to correctly identify the native binding pose among those sterically allowed, given that the scoring method used can be invoked to correctly prioritize the affinities of different poses.

It is generally appreciated that affinity prediction is one of the most difficult yet critical parts to CAMD. Current scoring methods used with docking methods are of limited accuracy and applicability. Binding affinity predictions will be discussed in Section 1.4.

1.4 Binding Affinity Predictions

Predicting binding affinity is one of the most critical and challenging components to computer-aided structure-based drug design. (Ajay and Murcko, 1995; Gohlke and Klebe, 2002) Methods for predicting binding affinity are essential in multiple applications, including molecular docking to identify a native binding mode, virtual screening of compound libraries to identify lead compounds, and lead optimization for enhancing binding affinity and target specificity. (Kitchen et al., 2004; Lyne, 2002; Shoichet, 2004) Despite significant advances in first-principle methods for predicting binding affinity (Beveridge and Dicapua, 1989; Massova and Kollman, 2000; Hansson et al., 1998; Wang et al., 2001; Jiao et al., 2008), empirical scoring functions that are fast and relatively accurate are still widely used in drug discovery. (Bohm and Stahl, 2002) For virtual screening studies where libraries up to millions of compounds are screened against a target of interest, a scoring function is needed to rapidly assess multiple binding modes of each of the multiple conformers generated for each potential ligand. This is also the case for in silico lead optimization where a large number of structural analogs are computationally constructed and assessed. In addition to advantages in speed, empirical scoring functions should not require careful validation

for each individual therapeutic target, but be portable once validated, making them more suitable for general use in high-throughput applications.

1.4.1 Computational Methods for Affinity Predictions

This section provides an overview of methods for estimating and experimentally measuring binding affinity of protein-ligand interactions. First-principles methods such as free energy perturbation (FEP) (Kollman, 1993), linear interaction energy (LIE) (Hansson et al., 1998), and MM-PBSA/GBSA (Massova and Kollman, 2000) have all been demonstrated to accurately estimate and rank binding affinities for series of analogs in a number of different systems. Despite the successes achieved by these more rigorous physics-based methods, their speed (requires molecular dynamics simulations) and need for calibration on novel systems limit their applicability in high-throughput screening. In addition, only modest structural perturbations to the ligand are tolerated to achieve accurate relative affinities. Scoring functions derived in a more simplified (less physically accurate) manner are still widely used in drug discovery due to advantages in speed and overall applicability.

Scoring functions can be broadly categorized into the following groups: force field-based, knowledge-based (heuristic), and empirical. Force field-based methods use terms derived from molecular mechanics calculations as a means to capture the physicochemical forces governing protein-ligand interactions. (Kuntz, 1992) Knowledge-based methods use experimentally determined structures and derive a statistical potential based on observed data for predictions. (Gohlke et al., 2000) Empirical scoring functions will be discussed in a greater detail in the Section 1.4.2.

1.4.2 Empirical Scoring Functions

Empirical scoring functions aim to represent the atomic interactions of protein-ligand complexes by the use of simple descriptors to adequately capture the physicochemical forces governing protein-ligand complex formation. The underlying assumption in scoring functions is that the physical and chemical interactions of protein-ligand interactions can be quantitatively captured using a set of descriptors. Each descriptor

is weighted by a coefficient, derived by a linear regression method through training on experimental data from binding assays, resulting in an equation for calculating binding affinities,

$$\Delta G_{complex} = a_1 X_1 + a_2 X_2 + \cdots + k \quad (1.3)$$

where a_i = i^{th} coefficient of i^{th} descriptor,

X_i = value of i^{th} descriptor,

k = constant derived from linear regression.

Over the last 20 years, a number of scoring functions have been developed, with some notable ones being SCORE1 (Bohm, 1994), SCORE2 (Bohm, 1998), ChemScore (Eldridge et al., 1997), X-Score (Wang et al., 2002), Lig-Score (Krammer et al., 2005), PLP (Gehlhaar et al., 1995; Verkhivker et al., 2000), DrugScore (Gohlke et al., 2000), and SFCscore (Sottriffer et al., 2008). These scoring functions differ by their choice and implementation of descriptors to capture the physicochemical interactions, the size and diversity of the training set, and the regression method used to derive the predictive equations. A number of reviews on scoring functions and assessments of their performance and applicability have been published. (Halperin et al., 2002; Wang et al., 2003; Stahl and Rarey, 2001; Wang et al., 2004; Ferrara et al., 2004; Warren et al., 2006; Cheng et al., 2009)

1.4.3 Applications

Empirical scoring functions are widely used in molecular design studies such as virtual screening for lead identification and lead optimization. A number of recent scoring functions, notably SFCscore and X-Score, have provided accurate and robust affinity estimates on complexes in PDBbind, a database compiled of high-quality protein-ligand complexes.

Any particular scoring function will inevitably provide limited accuracy for certain classes of targets or compounds, depending on the implementation and training data used. To circumvent this problem, the consensus scoring strategy is used to take

advantage of the strengths of each scoring function in order to correctly prioritize binding affinities. Consensus scoring, a strategy in which multiple scoring functions are used to derive an average affinity estimate, have been demonstrated to provide enrichment compared to the use of a single scoring function (Wang et al., 2003; Cheng et al., 2009).

1.4.4 Limitations and Practical Considerations

Despite the proven utility of empirical scoring functions in CAMD, there remains significant room for improvement in both accuracy and applicability. Accurate affinity predictions remain elusive for novel targets that are not represented in the scoring function training set. Correct ranking of binding poses according to relative affinities is still a challenge. False positives and false negatives are common in most, if not all, virtual screening studies.

Although accuracy and applicability are the primary objectives in scoring function development, computing power is also a concern. Scoring function will still need to be fast enough to assess over tens of millions of poses (considering a library of 1,000,000 compounds with 30-50 poses per compound) that is typical in a lead discovery study. However, with the continuing availability of computational power, speed has become less of a concern in the development of scoring functions.

1.5 Experimental Data for Methods Development and Validation

Development of computer-aided molecular design methods requires the use of experimental data to validate and assess their predictability and robustness. With the increasing amount of biomolecular structures and experimental data in the form of X-ray crystal structures and affinity measurements from biological assays, a wealth of information exists for the development and validation of CAMD methods.

The Protein Data Bank (PDB) serves as the central depository of biomolecular structures. The SCORPIO (Olsson et al., 2008) and PDBcal (Li et al., 2008) databases have recently been compiled with protein-ligand complexes with X-ray crystal structures and thermodynamics parameters measured directly by isothermal titration calorimetry (ITC). (Freire, 2004) PDBbind is the premier resource for high-quality protein-ligand complexes available from the PDB. With the increasing availability of structural and biological/biophysical data, CAMD methods may now be rigorously tested for accuracy and overall target applicability.

1.5.1 Biomolecular Structures Derived by X-ray Crystallography

Protein Data Bank (PDB) is the most comprehensive depository of biomolecular structures (containing over 66,000 structures as of August 2010) solved by experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR). PDB serves as a valuable resource of protein-ligand complexes for the development and validation of computational methodologies. With the increasing number of complex structures, a sufficient number of structures with high nominal resolution (2 Å) are available for model development. In addition, diversity of protein families and ligand structures are now available for the compiling more diverse training sets for methods development.

Recent studies have highlighted the need to use high-quality protein-ligand complexes for CAMD methods development and validation. In addition to nominal resolution, 2 other metrics to better describe the quality of X-ray crystal structures are the free R value (R_{free}) (Brunger and Rice, 1997) and diffraction-component precision index (DPI) (Cruickshank, 1999). R_{free} is a measure of the degree to which an atomic model predicts a subset of the observed diffraction data that has been omitted from the refinement. R_{free} is defined by the equation:

$$R_{free} = \frac{\sum |F_{obs} - F_{calc}|}{\sum |F_{obs}|} \quad (1.4)$$

where F_{obs} = the observed reflection amplitudes,
 F_{calc} = the reflection amplitudes calculated from the model.

DPI is measure of the quality of the structural model derived from the diffraction data. DPI is defined by the equation:

$$\sigma(x, B_{avg}) = 1.28 N_{atoms}^{1/2} V_a^{1/3} N_{obs}^{-5/6} R_{free}, \quad (1.5)$$

where N_{atoms} = number of atoms in the unit cell,
 V_a = volume of unit cell,
 N_{obs} = number of crystallographic observations.

Collectively, these 3 metrics should provide a better quantitative measure of structural quality, and can be used as criteria to exclude low-quality structures to minimize structural noise which may contribute to inaccuracies when used for training and testing CAMD methods.

1.5.2 Thermodynamic Measurements by Isothermal Titration Calorimetry

Isothermal titration calorimetry (ITC) is a biophysical technique used in molecular design studies to determine the thermodynamic parameters contributing to the change in binding free energy of the system. (Freire, 2004) ITC directly measures the heat energy associated with a molecular interaction between 2 or more molecules to provide quantitative measurements of binding affinity (K_a), enthalpy changes (ΔH), and binding stoichiometry (n), from which the binding free energy and change in entropy can be derived by the equation,

$$\Delta G = -RT \ln K_a = \Delta H - T \Delta S \quad (1.6)$$

where R = gas constant,
 T = temperature.

ITC is a valuable tool because the experiments do not require chemical modifications, labeling, immobilization, or constraints on system size, providing measurements in a more realistic environment. However, ITC is limited to studying mostly high-affinity ligands, since ligands that are weak binders will require an intractable concentration of protein for conducting these experiments. ITC also requires significant time and labor to perform the experiments, and is a low-throughput method compared to other biophysical techniques for affinity measurements (e.g., surface plasmon resonance). Even with these drawbacks, ITC remains the gold standard experimental technique to determine thermodynamic parameters.

With the increasing use of isothermal titration calorimetry in molecular recognition studies, databases such as SCORPIO (Olsson et al., 2008) and PDBcal (Li et al., 2008) have been compiled for use in the development of methods for affinity predictions. SCORPIO contains 254 protein-ligand complexes with X-ray crystal structures and thermodynamic parameters from ITC, while PDBcal contains over 400 protein-ligand complexes. Taken together, these 2 databases provide both valuable structural and thermodynamic information for the development and validation of more accurate and applicable CAMD methods.

1.5.3 Protein-Ligand Complexes for Scoring Function Development

PDBbind is one of the most comprehensive databases of bound protein-ligand complexes with X-ray crystallographic structure and experimentally measured binding affinity data (K_d , K_i , IC_{50}). The 2002 ($n = 800$) and 2004 ($n = 1091$) versions of the refined set has been used in a number of comparison studies. The recent 2009 version contains over 1600 complexes for scoring function evaluation purposes. PDBbind has become the gold-standard benchmark set for evaluation of binding affinity predictions.

PDBbind has been used to assess the accuracy and reliability of the top-performing empirical scoring functions (e.g., X-Score, SFCscore, ChemScore, etc.). In addition, a core set of 195 complexes in the 2007 version has been used to assess the performance of scoring functions on non-redundant structures based on docking power, scoring power, and ranking power. (Cheng et al., 2009)

1.5.4 Inherent Inaccuracies in Experimental Data

Structures of protein-ligand complexes solved by X-ray crystallography and experimentally measured binding affinity data are typically used to develop and validate CAMD methods. In crystallography, the conditions used to induce crystallization (e.g., temperature, salt concentrations, buffer conditions) do not necessarily reflect physiological conditions. In addition, current scoring functions have used lower nominal resolution (>2 Å) structures for model training and testing, which may lead to atomic-level structural inaccuracies for CAMD methods development.

Experimentally measured binding affinity data used for CAMD methods development comes from assays performed by different groups, and are not necessarily performed under the same conditions (e.g., temperature, salt concentrations, buffer conditions), leading to inaccuracies in measurements. A combination of K_d , K_i , and IC_{50} are used to derive the binding affinity, which may lead to discrepancies when used for direct affinity comparisons.

In CAMD methods development, the use of high-resolutions structures (≤ 2 Å) and thermodynamics parameters determined by ITC may alleviate the potential sources of experimental inconsistencies.

1.6 Synopsis of Thesis

The thesis is a compilation of 2 main projects. The first project aimed to identify drug-like inhibitors targeting bacterial two-component signal transduction response regulators as a strategy to inhibit virulence. The second project involves the development of an empirical scoring function derived using high-resolution structures and

thermodynamic parameters measured by ITC to predict binding affinities of protein-ligand interactions.

1.6.1 Summary of Challenges

The first project includes 2 significant challenges: identification of potential antibiotics with novel modes of action, and targeting protein-protein interactions using drug-like compounds. Due to the increasing resistance of bacteria to current clinical antibiotics, there is an urgent need to develop therapeutics with novel modes of action. Inhibiting virulence should lead to less resistance generation since less selective pressure is presented to bacteria. To date, antibiotics development has proven to be a significant challenge (Payne et al., 2006), especially ones with novel modes of action. Targeting protein-protein interactions is the next frontier in therapeutics development. Because protein-protein interactions are ubiquitous in signal transduction and essential for cellular function, targeting this class of interactions provide a new avenue for drug discovery. However, the relatively flat and solvent-exposed surfaces typically present in protein-protein interfaces are challenging binding sites to target. Only few examples of successful cases have been demonstrated.

In the second project, an empirical function is developed to accurately predict binding affinity and thermodynamic parameters of protein-ligand interactions. Scoring functions for affinity predictions have been relatively accurate, at best. Current scoring functions lack an accurate representation of entropic changes, which should lead to inaccuracy in binding affinity predictions. In addition, scoring functions also use lower-resolution crystal structures for model training and testing, which may lead to inaccuracies in estimating affinities. To my knowledge, no empirical scoring functions are available that can predict the enthalpic and entropic changes contributing to binding free energy changes. Estimating the thermodynamic parameters to binding free energy will provide further insight into lead discovery and optimization.

1.6.2 Discoveries and Insights Presented

This thesis consists of a prospective study on using CAMD methods for lead discovery, and also the development of a scoring function for binding affinity predictions.

1. Discovery of First-In-Class Response Regulator Inhibitors

Eight compounds have been identified to disrupt formation of the PhoP-DNA complex necessary for regulating gene expression. These compounds do not effect dimerization, as initially targeted, but may potentially be binding at the N-terminal domain to act in an allosteric manner, or at the C-terminal DNA-binding domain, both of which represent novel modes of action.

2. Accurate Affinity Predictions Using High-Resolution Structures and ITC Measurements

In the development of the PHOENIX empirical scoring function, a set of 112 complexes (X-ray crystal structures with resolution $\leq 2\text{\AA}$ and thermodynamic parameters from ITC) was used to derive a scoring function for affinity prediction. PHOENIX was rigorously tested using a variety of complex sets, and has proven to perform comparably to other top-performing scoring functions, and better in a few cases.

Taken together, these advances in molecular design will provide addition insights and a deeper understanding to guide the development of CAMD methods.

Chapter 2

Background

This chapter aims to provide background information on several key concepts studied in this dissertation: empirical scoring functions for affinity predictions, descriptors to estimate entropy changes, drug design targeting protein-protein interactions, and bacterial two-component signal transduction. Empirical scoring functions and descriptors to estimate entropy changes are the fundamental concepts for the second project of this thesis: the development of the PHOENIX scoring function for affinity predictions derived using high-resolution structures and calorimetry measurements. Drug design to target protein-protein interactions and bacterial two-component signal transduction are important background information for understanding the motivation and significance of the first project of this thesis: discovery of drug-like compounds targeting the PhoP response regulator to inhibit bacterial virulence.

The first section will present the objectives of an empirical scoring function and its main components. The second section will present heuristic descriptors to capture entropic contributions to binding free energy in protein-ligand interactions. The third section will present the background and significance of targeting protein-protein interactions for drug design. The fourth section will provide the background on bacterial two-component signal transduction and its significance as a therapeutic target.

2.1 Empirical Scoring Functions

Empirical scoring functions aim to represent the atomic interactions of protein-ligand complexes by the use of simple quantitative descriptors to capture the physicochemical forces governing protein-ligand complex formation. The underlying assumption in scoring functions is that the physical and chemical interactions of protein-ligand interactions can be quantified using a set of such descriptors. Each descriptor is weighted by a coefficient, derived by a linear regression method through training on experimental data from binding assays, resulting in an equation to estimate binding affinities. Within the last 20 years, a number of scoring functions have been developed, with some notable ones being SCORE1 (Bohm, 1994), SCORE2 (Bohm, 1998), ChemScore (Eldridge et al., 1997), X-Score (Wang et al., 2002), Lig-Score (Krammer et al., 2005), GOLD (Jones et al., 1997), PLP (Gohlke et al., 2000), and SFCscore (Sotriffer et al., 2008). These scoring functions differ by their choice and implementation of descriptors to capture the physicochemical interactions, the size and diversity of the training set, and the regression method used to derive the predictive equations.

2.1.1 Training Set

To develop empirical scoring functions, a large and diverse training set of protein-ligand complexes, typically from X-ray crystallographic structures, are used. Compilation of the training set aims to adequately represent the diversity of protein-ligand interactions by including complexes from various protein families (e.g., proteases, metalloenzymes, etc.) and ligands with different structural and chemical properties (e.g., drug-like compounds, sugars, peptides, peptidomimetics, etc.) in order to capture the different types of physical interactions involved in molecular recognition.

Protein-ligand complexes used are of high structural quality with binding affinities measured from biological assays. For example, in the training set used to develop X-Score, a set of 200 protein-ligand complexes with a nominal resolution of 2.5 Å or less with experimentally measured K_d or K_i values were used. In the development of SFCscore (Sotriffer et al., 2008), a larger training set of 290 complexes (some obtained from corporate databases) were used to achieve accurate affinity predictions. As the

availability of high-quality crystal structures with binding affinity data continues to increase, larger sets of protein-ligand complexes may be included in training sets to develop more accurate and versatile empirical scoring functions.

2.1.2 Descriptors

Empirical scoring functions use simple descriptors to quantitatively capture the physicochemical forces of molecular recognition important in protein-ligand interactions. Degree of descriptor complexity may range from relatively simple ones such as the number of ligand heavy atoms as a simple estimate of van der Waals contacts with the binding site, to more complex ones such as molecular dynamics simulations using a polarizable potential to estimate the electrostatics contributions to binding free energy. (Jiao et al., 2008)

Typical terms used in empirical scoring functions include descriptors to quantify the size and potential conformations of the ligand (e.g., molecular weight, number of rotatable bonds), surface complementarity, and hydrophobic and hydrophilic properties. Implementation of descriptors can sometimes vary amongst scoring functions, resulting in different estimations of physicochemical properties. Descriptors sets also vary in size and diversity depending on the implementation and training set used. A descriptor set should contain a minimal number of descriptors to explain the variance of the training set in order to develop the simplest models, which allow for a more direct interpretation of the rationale and physical forces captured by the scoring function.

Due to the relatively large number of descriptors and a relatively small number of protein-ligand complexes for used in developing empirical scoring functions, the issues with collinearity amongst descriptors often occur and may limit prediction accuracy. (Wold et al., 1984) Methods such as removing descriptors with greater than 90-95% correlation, or removing ones that have low correlation to binding affinity may help to identify a more informative set of descriptors by removal of ones contributing to noise. Feature selection methods are sometimes used to identify a subset of descriptors explaining to the most variance in the model.

2.1.3 Linear Regression Method

A linear regression method is typically used to assign coefficients to each descriptor and derive equations to fit the experimentally measured binding affinities. Partial least squares of latent variables (PLS) regression is a statistical method commonly used in empirical scoring function development. (Geladi and Kowalski, 1986) PLS finds a linear regression model by projecting the predicted variables (X) and the observable variables (Y) to a new space. A PLS model will try to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space. PLS is suited for use in instances where the number of variables is more than the number of observations, and when there is multicollinearity among X values, often the case with physicochemical descriptors.

Leave-one-out cross validation (LOO-CV) is performed to identify the appropriate number of PLS components (to explain variance while avoiding overfitting) to use for equation modeling. LOO-CV statistics such as q^2 in the following equation,

$$q^2 = 1 - \frac{\sum_i (X_{i,pred} - X_{i,exp})^2}{\sum_i (X_{i,exp} - X_{i,mean})^2} \quad (2.1)$$

and cross-validation standard errors (S_{PRESS}) in the following equation,

$$S_{PRESS} = \sqrt{\frac{\sum_i (X_{i,pred} - X_{i,exp})^2}{(N - k - 1)}} \quad (2.2)$$

where N = number of complexes,

k = number of PLS components,

are used to assess the predictive ability of the equation. Internal cross-validation, in which a training set is separated into a smaller training set and a test set, is performed to better assess overfitting and model robustness (e.g., not hyper-sensitive to a particular complex or set of complexes used in the training set). Statistics such as r_{pred}^2

$$r_{pred}^2 = 1 - \frac{\sum_i (X_{i,pred} - X_{i,exp})^2}{\sum_i (X_{i,exp} - X_{i,mean})^2} \quad (2.3)$$

and standard error (SE_{pred}) in the following equation,

$$SE_{pred} = \sqrt{\frac{\sum_i (X_{i,pred} - X_{i,exp})^2}{(N - 1)}} \quad (2.4)$$

where N = number of complexes,
are used to assess the predictive ability of the equation.

2.2 Descriptors to Estimate Entropy Changes

Entropy changes in protein-ligand interactions have a limited representation in existing scoring functions. Most of the commonly used scoring functions include descriptors such as number of ligand rotamers and partition coefficient to estimate the conformational entropy and desolvation effects of ligand binding. Recent theoretical studies have illustrated that certain components of entropic changes in protein-ligand interactions are not adequately captured by scoring functions: desolvation entropy and configurational entropy.

The first section will present descriptors traditionally used to capture entropy changes, primarily the change in conformational entropy upon ligand binding. The second section will introduce shape- and volume-based descriptors to better capture desolvation effects and configurational entropy.

2.2.1 Rotamers and Hydrophobicity

One common descriptor used to simply capture the change in conformational entropy in protein-ligand interactions is the number of rotatable bonds. Searle and Williams have estimated that each rotatable bond contributes 1.2-1.6 kcal/mol to change of binding free energy, assuming complete loss of rotational freedom. (Searle and Williams, 1992) In most of the commonly used scoring functions, the number of rotatable bonds on

the ligand is used to estimate change of entropy upon ligand binding. To estimate motions from binding site residues to the change of entropy, number of rotatable bonds of binding-site residues has been included in scoring functions. However, such descriptors did correlate with the change in binding free energy as used in previous studies (e.g., X-Score) (Wang et al., 2002).

Descriptors to quantify hydrophobicity of the ligand and binding site are also commonly used to capture entropy changes in solvation and desolvation effects. The octanol/water partition coefficient (logP) is commonly used to estimate the lipophilicity of the ligand based on the properties of its atoms. (Wang et al., 1997) This simple descriptor has demonstrated to be an effective indicator of ligand lipophilicity to adequately capture entropic changes, in addition to being an important solubility predictor and pharmacological property.

Descriptors to quantify the complementary hydrophobic surface area of protein-ligand complexes are also commonly used to capture solvation and desolvation entropy changes from the transfer of the ligand in solvent to the binding site, and also the displacement of water molecules in the binding site.

2.2.2 Shape- and Volume-Based Descriptors

As discussed earlier, current scoring functions have used a limited representation of thermodynamically important phenomena such as solvation effects and configuration entropy, which may explain their inaccuracies in affinity predictions. As a heuristic method to quantify solvation and desolvation effects, shape- and volume-based descriptors can be used to capture these effects. Volume descriptors capture configurational entropy (translational and rotational entropy) by characterizing the steric complementarity of the ligand bound to its binding site.

Recent advances in algorithms to quantify molecular shape and volume have provided new tools for calculating descriptors for scoring function development. A number of programs are now available for the identification and characterization of protein binding sites and cavities (e.g., VOIDOO, LIGSITE, POCKET, POCKET-FINDER, CAST, PASS, APROPOS, SURFNET, Q-SITEFINDER, POCKETPICKER, etc.).

(Kleywegt and Jones, 1994; Hendlich et al., 1997; Levitt, 1992; An et al., 2005; Liang et al., 1998; Brady and Stouten, 2000; Peters et al., 1996; Laskowski, 1995; Laurie and Jackson, 2005; Weisel et al., 2007) Nicholls et al. (2010) has reviewed the application and performance of current methods to estimate molecular shape in computer-aided design.

For practical reasons, programs that are accurate, freely available, and user-friendly are utilized to calculate descriptors for scoring function development. A couple of programs recently developed, FPOCKET and VICE, fit these criteria. FPOCKET used an algorithm based on Voronoi tessellation and alpha spheres to characterize the shape and volume of the ligand and binding site. (Guillox et al., 2009) VICE (Vectorial Identification of Cavity Extents) used an algorithm based on vector representations to characterize the shape and depth of protein binding sites. (Tripathi and Kellogg, 2009) FPOCKET was used to calculate shape and volume descriptors included in the development of the PHOENIX scoring function. Descriptors calculated by VICE will be evaluated in later versions of PHOENIX.

2.3 Protein-Protein Interactions

Protein-protein interactions (PPI) are ubiquitous and essential in establishing the affinity and selectivity of biomolecular interactions for biological function. The ability of a protein to bind specifically to its intended partner and form stable complexes to carry out a certain function is fundamental to biology. PPI plays an important role in the function of metabolic networks and signal transduction. In signal transduction, proteins are typically activated (e.g., phosphorylation) and interact to form homodimers, or complexes with other partners to process a signal. These common interactions serve as attractive targets for signal modulation and inhibition. With the discovery of PPI hot spots responsible for the interactions at the interfaces, much effort have been devoted to targeting these critical regions with small molecules and peptidomimetics as a novel strategy to modulate biological function.

The first section will provide background on hot spots of PPI, the second section will illustrate the challenges of targeting PPI, and the third section will present recent successes of PPI inhibition using small molecules.

2.3.1 Hot Spots

Protein-protein interaction interfaces were traditionally accepted as having free energy of binding evenly distributed amongst residues across interfaces. However, recent experimental studies and data analyses have led to a paradigm shift of this concept. Pioneering studies by Clackson and Wells (Clackson and Wells, 1995) on a hormone-receptor interface by alanine-scanning mutagenesis and X-ray crystallography demonstrated that only a subset of residues were contributing significantly to the binding free energy of protein-protein complexes. Jones and Thornton examined the structures available in the Protein Data Bank (PDB) to gain a better general understanding of the structural characteristics of various classes (e.g., homodimer, heterodimer, enzyme-inhibitor, antibody-protein) of PPI, and observed a similar trend in addition to other properties. (Jones and Thornton, 1996) This small subset of residues, known as hot spots, contribute to a significant amount to binding free energy for PPI, and determines a proteins affinity and specificity to various partners as well as its biological function.

Fundamental properties which govern PPI are the following: size and shape, surface complementarity, (Fernandez and Scheraga, 2003) residue interface propensities, hydrophobicity and hydrogen-bonding. Contact surfaces of PPI interfaces are generally large (1500-3000 \AA^2 , compared to 300-1000 \AA^2 for protein-ligand interfaces) and hydrophobic with a high percentage of buried nonpolar residues. These interactions occur through van der Waals contacts and are energetically favorable due to the preference of the interfaces to transfer from a polar to a nonpolar environment. PPI interactions consist of close packing of residues, with some in patches that protrude to serve as recognition motifs. The number of patches range from 1 to 15 residues within a distance of 200-400 \AA to result in displacement of water molecules at the interface and a gain in change in entropy. The enthalpic contributions from van der Waals contacts in addition to the entropic contributions from desolvation make PPI energetically favorable. Electrostatics forces from the charged residues at the interface contribute to complex formation and are important in determining the duration of interactions. (Sheinerman et al., 2000)

Hot spots are functionally characterized as residues where alanine mutations cause a significant increase (≥ 2.0 kcal/mol) in the binding free energy. (Thorn and Bogan,

2001) Thorn and Bogan have compiled a database to classify PPI hot spots using experimental data from alanine-scanning mutagenesis studies. Analyses performed by Moreira et al. (2007) determined that on average 9.5% of the interfacial residues are hot spots. PPI hot spots are functionally and structurally adaptive, and play important roles in binding to its cognate partner, as well as to other proteins. Hot spots are typically located near the center of PPI interfaces and are critical for complex formation. They have also been observed to overlap with structurally conserved residues. Identification and characterization of PPI hot spots should provide a better understanding of its potential binding partners and its biological functions. This knowledge can be used to target hot spots for molecular design of PPI inhibitors in therapeutics development.

2.3.2 Challenges in Targeting PPI

PPI interfaces have long been recognized as challenging structural targets for molecular design. In general, PPI interfaces are hydrophobic, flat surfaces decorated with critical charge groups to determine binding specificity. While hydrophobic and charged binding sites are typically encountered in molecular design for enzymes and receptors, the flatness of the binding site makes it challenging for small molecules to bind with sufficient affinity and selectivity. Enzymes and receptors typically possess grooves and pockets at protein surfaces to serve as binding sites. Also, PPI interfaces do not have a starting natural ligand or substrate to serve as a lead molecule for ligand-based searches and affinity and property optimization often known when targeting proteases and receptors.

X-ray crystal structures available of PPI interfaces are typically solved with its cognate protein partner, rather than with a bound ligand or substrate that is common with enzymes. For structure-based design, crystal structure bound to another protein, typically as a homodimer or heterodimer, may not necessarily reveal the potential regions for small molecules to bind. Ligands may bind at PPI interfaces by an induced-fit mechanism, which can be challenging to accommodate when using a crystal structure with a bound protein partner. A number of PPI studies have demonstrated that the interfaces are adaptable, and binding sites may not necessarily be observed from the free protein or the PPI complex. Plasticity of PPI interfaces comes from the

dynamics of the side chains and loops in the region. Molecular dynamics simulations have suggested the occurrence of transient pockets suitable as small molecule binding sites. (Brown and Hajduk, 2006; Eyrisch and Helms, 2007) The use of static crystal structures for molecular design may neglect potential small molecule binding sites. (See Fig: 2.1.)

Solvation effects upon ligand binding, as well as electrostatics contributions to binding free energy are not well understood in PPI. Computational tools designed to capture solvation and electrostatics effects for traditional medicinal chemistry targets may not necessarily be suitable for use in PPI interfaces since the binding sites are located in different environments (e.g., deep hydrophobic cavities for enzymes; shallow and solvent-exposed regions for PPI surfaces). More detailed experimental and theoretical studies will be necessary to achieve a better understanding the roles of solvation and electrostatics to binding free energy in PPI.

The increasing availability of experimental data from PPI studies (e.g., alanine-scanning mutagenesis to identify hot spots; free, protein-protein, and protein-ligand crystal structures; binding affinity and thermodynamic parameters; biophysical studies to assess dynamics for transient binding pockets and potential for induced-fit binding) will further improve understanding of the underlying physical principles of molecular recognition in PPI. To summarize, a better understanding of PPI interface dynamics and energetics is necessary to achieve accurate molecular design. Additional complexity in computational analyses of ligand binding may be essential for accurate predictions.

2.3.3 Examples of Inhibitors

Despite the challenges faced when targeting PPI interfaces with small molecules, a number of successes have been reported in the past decade. These notable successes in the discovery and design of PPI inhibitors have been reviewed by Wells and McClendon, in particular, examples of discontinuous PPI for which small molecules that compete directly with one of the protein partners have been discovered. (See Fig: 2.2) (Wells and McClendon, 2007)

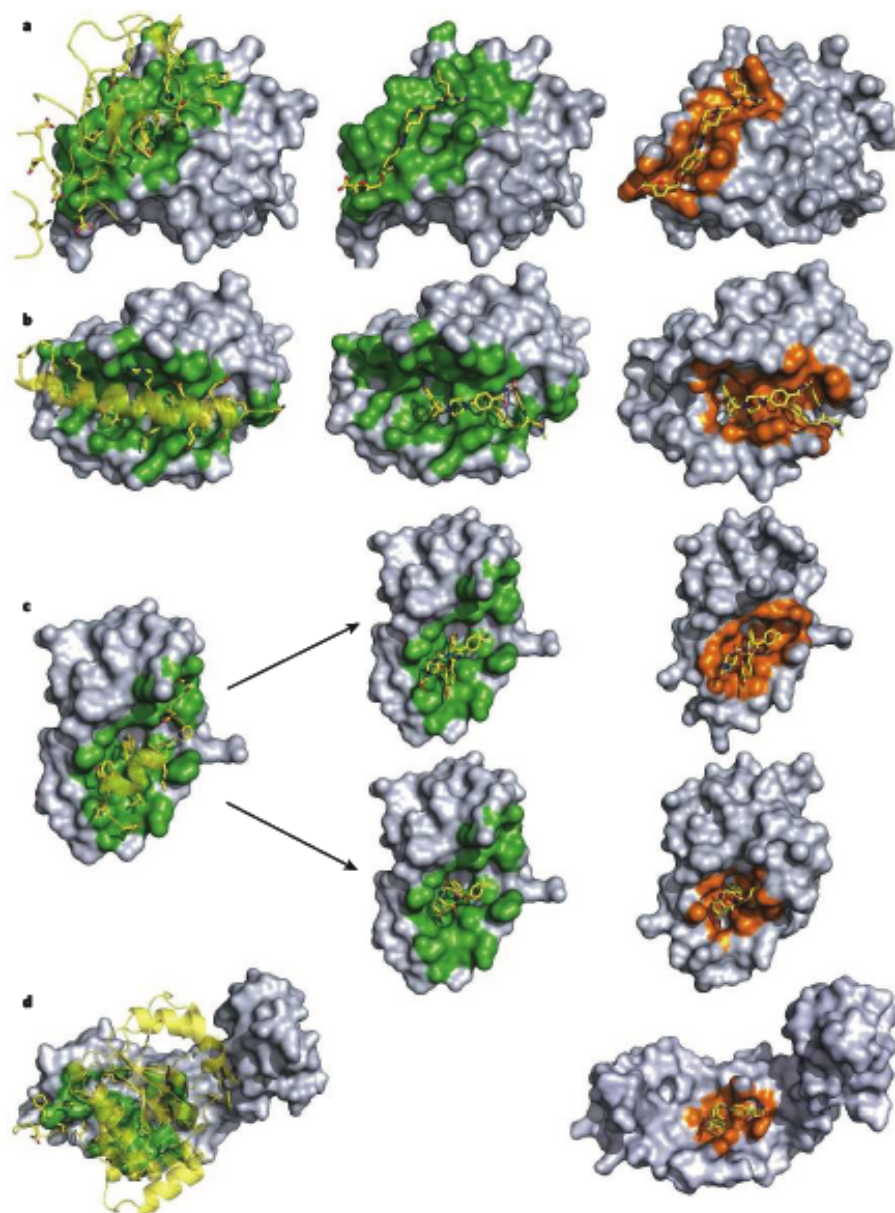


Figure 2.1: The structures of protein-protein or protein-peptide complexes are shown on the left. The target protein (gray), the binding protein or peptide (yellow), and selected side chains (shown in sticks: carbon in yellow, oxygen in red, nitrogen in blue). The contact surface of the target protein (green) is within 4.5 Å from the binding molecule. The structures of the protein-small-molecule complexes are shown on the right, with the contact surface shown in orange. Examples are (a) IL-2, (b) Bcl-X_L, (c) HDM2, (d) HPV-18 E2. Figure taken from Wells and McClendon (2007).

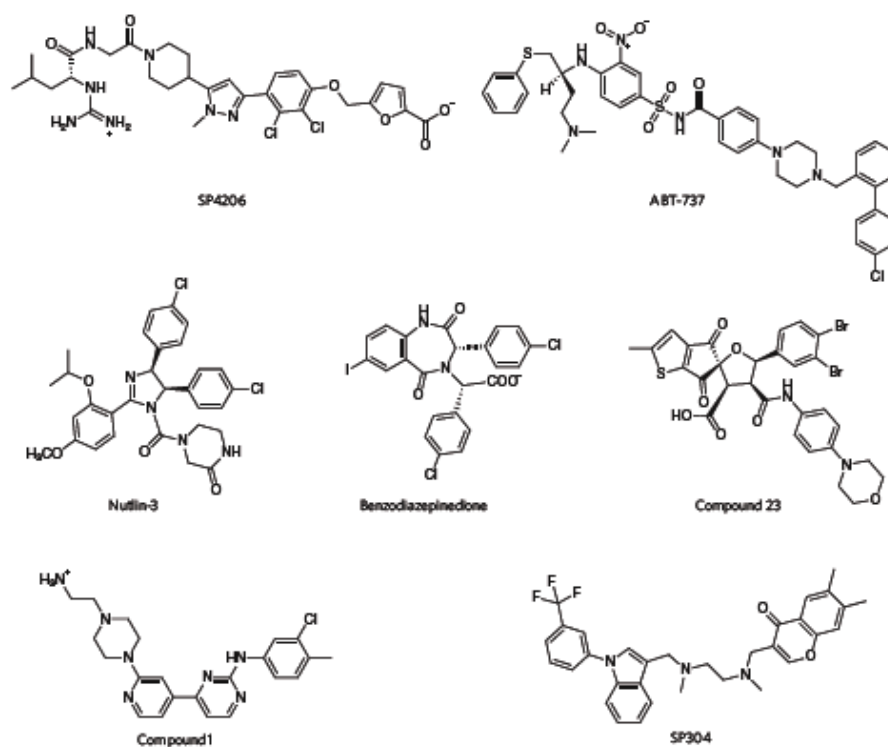


Figure 2.2: Six examples of small-molecule inhibitors of protein-protein interactions that have been discovered and reviewed in Wells and McClendon (2007). These compounds represent different areas of chemical spaces compared with each other.

SP4206 binds to IL-2. ABT-737 binds to Bcl-XL. Nutlin-3 and the benzodiazepinedione shown above bind to HDM2. Compound 23 binds to HPV E2. Compound 1 binds to ZipA. And SP304 binds to TNF. Figure taken from Wells and McClendon (2007).

These examples have available crystal structures of both the PPI complex as well as the protein-ligand complex, and serve as model systems for future structure-based design of PPI inhibitors.

Six recent examples of targets for PPI inhibitors with crystal structure information are as follows: the cytokine interleukin-2 (IL-2), members of B-cell lymphoma 2 (Bcl-XL), human protein double minute 2 (HDM2), human papilloma virus transcription factor E2 (HPV E2), bacterial membrane protein ZipA, and cytokine tumor-necrosis factor (TNF). Initial hits were identified by high-throughput screening were in the micromolar range, and optimized into nanomolar inhibitors by medicinal chemistry efforts. One thing to note from the screening studies was the high percentage of false positives. This may be due to the chemical space representation of the compound libraries, which typically contain compounds targeting traditional drug discovery targets (e.g., proteases, G-protein-coupled receptors). As the number of PPI inhibitors increase, retrospective studies can be performed to identify the appropriate chemical space in which to search.

In addition to discovery through screening efforts, rational, structure-based design studies have also been demonstrated to be effective in targeting PPI interfaces. In particular, the use of alpha-helical mimetic scaffolds has proven to be effective in a number of targets. Walensky et al. used a hydrocarbon stapling strategy to design stabilized alpha-helix of BCL-2 domains which was shown to be helical, protease-resistant, and cell-permeable to bind to BCL-2 pockets with enhanced affinity. (?) Subsequent theoretical studies have provided insight into the underlying forces responsible for its α -helical propensity. In another notable series of molecular design studies, the Hamilton group used various non-peptidic, small-molecule scaffolds as α -helix mimetics and demonstrated their effectiveness with a variety of PPI targets.

Recent examples of successes from both screening studies and rational design demonstrated the feasibility of targeting PPI interfaces for modulation and inhibition of biological functions. Due to its prevalence and importance in biology, PPI serves as the next frontier of structural targets for the development of therapeutics with novel modes of action.

2.4 Bacterial Two-Component Signal Transduction Systems

Two-component signal transduction (TCST) is the predominant signaling scheme used in bacteria to sense and respond to environmental changes in order to survive and thrive. Bacteria typically contain over a dozen TCST systems that are involved in regulating physiological functions such as metabolism and motility, as well as more specialized functions such as virulence and development. Since their discovery approximately 20 years ago, TCST systems have been extensively studied across a number of different bacterial species. Due to the prevalence and importance of TCST in bacteria, gaining a better understanding of these signaling systems may be useful for the development of antibiotics to target these conserved and ubiquitous modular signaling schemes.

This section provides background on the mechanism of TCST, characteristics of the largest structural family of TCST response regulators (OmpR/PhoB), the PhoP response regulator in *Salmonella enterica*, and the attractiveness of targeting response regulators for therapeutics development.

2.4.1 Mechanism

Two-component signal transduction (TCST) is the predominant signaling scheme in bacteria to sense and respond to environmental changes for survival and proliferation. (Stock et al., 2000; Gao et al., 2007; Gao and Stock, 2009; Stock and Guhaniyogi, 2006; Gao and Stock, 2010) TCST regulatory systems are modular in terms of their arrangement of domains within their proteins within various pathways. In general, TCST regulatory systems are comprised of a transmembrane sensor histidine kinase (HK) and an intracellular receiver response regulator (RR) with conserved sequence, structural, and biochemical properties, allowing them to readily adapt to various modes of intracellular signaling. These signaling systems typically couple environmental stimuli to an adaptive response, participating in fundamental processes such as regulating metabolism, as well as more specialized functions such as controlling virulence for the pathogens host.

A TCST system in its simplest form consists of 2 conserved components: a histidine kinase (HK) and a response regulator (RR). The HK is the input component to detect external stimuli to generate a response through the signaling pathway. The RR is the output component regulated by the HK to produce a response to external changes. In the signaling scheme, the HK autophosphorylates at a histidine residue, resulting in a high-energy phosphoryl group that is subsequently transferred to an aspartic acid residue of the RR. Phosphorylation induces a conformational change of the RR to activate its function as a transcription factor to respond to the external stimuli transduced by the HK by gene regulation. (See Fig: 2.3.)

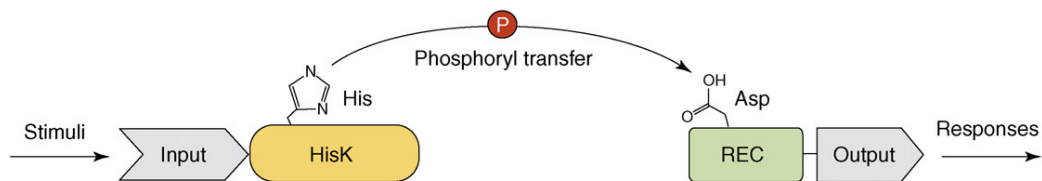


Figure 2.3: Schematic diagram of a typical two-component signal transduction (TCST) system. The conserved histidine-containing kinase domain (HisK) of the histidine kinase is shown in yellow and the conserved aspartic-acid-containing receiver domain (REC) of the response regulator is in green. The variable domains that confer specificity of input (the sensing domain) and output (effector domain) to each TCST system are shown in gray. Figure taken from Gao et al. (2007).

2.4.2 OmpR/PhoB Structural Subfamily

PhoP is a member of the OmpR/PhoB structural family, the largest structural family making up approximately 30% of all TCST response regulators. OmpR/PhoB family members are typically composed of 2 domains connected by a flexible linker: an N-terminal receiver domain and a C-terminal DNA-binding domain. Response regulators of the OmpR/PhoB family are characterized by a conserved $\alpha 4$ - $\beta 5$ - $\alpha 5$ structural motif of the receiver domain, a plastic surface important for dimerization. OmpR/PhoB subfamily members are characterized with a conserved and functionally important $\alpha 4$ - $\beta 5$ - $\alpha 5$ structural motif at the dimer interface. (See Fig: 2.4 and Fig: 2.5.)

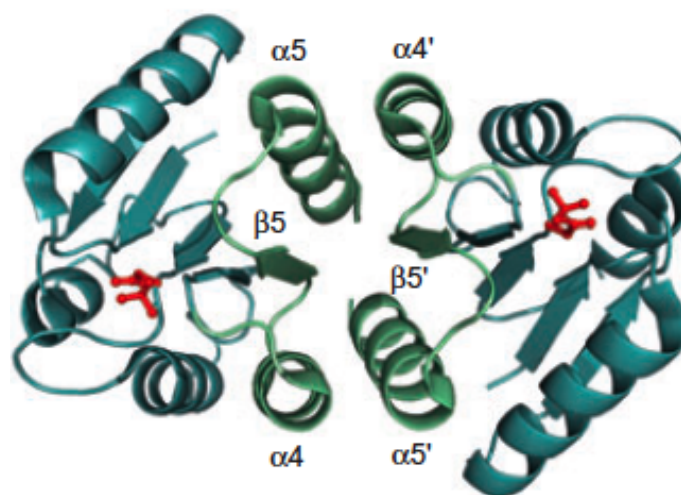


Figure 2.4: Dimer structure of the PhoB receiver domain (PDB ID: 1ZES). The receiver domain has a conserved $\alpha 4$ - $\beta 5$ - $\alpha 5$ (teal blue) and OmpR/PhoB subfamily RRs appear to share a conserved dimeric structure once phosphorylated. The non-covalent phosphoryl analogue beryllofluoride (BeF_3^-) coordinates to the conserved aspartate residue (red), allosterically perturbing the $\alpha 4$ - $\beta 5$ - $\alpha 5$ surface (green) and promoting dimerization. Figure taken from Gao et al. (2008).

The N-terminal $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface share a common set of hydrophobic and charged residues involved in van der Waals contact and salt-bridges important for homodimerization and function, and are conserved amongst other response regulators across different bacteria species. (Gao et al., 2007)

2.4.3 *Salmonella enterica* PhoP Response Regulator

The PhoQ/PhoP two-component regulator system is a major regulator of virulence in *Salmonella enterica* serovar Typhimurium, and also in a number of other gram-negative bacterial pathogens (e.g., *Shigella flexneri*, *Yersinia pestis*, *Neisseria meningitidis*). (Groisman, 2001; Vescovi et al., 1996; Moss et al., 2000; Oyston et al., 2000; Jamet et al., 2009; Johnson et al., 2001) PhoQ/PhoP in *S. enterica* is activated by low extracellular Mg^{2+} levels, acidic pH, and antimicrobial peptides (typical of human gut conditions during infection) to control various physiological and virulence functions. (Vescovi et al., 1996; Choi et al., 2009; Kato and Groisman, 2008; Shi et al., 2004)

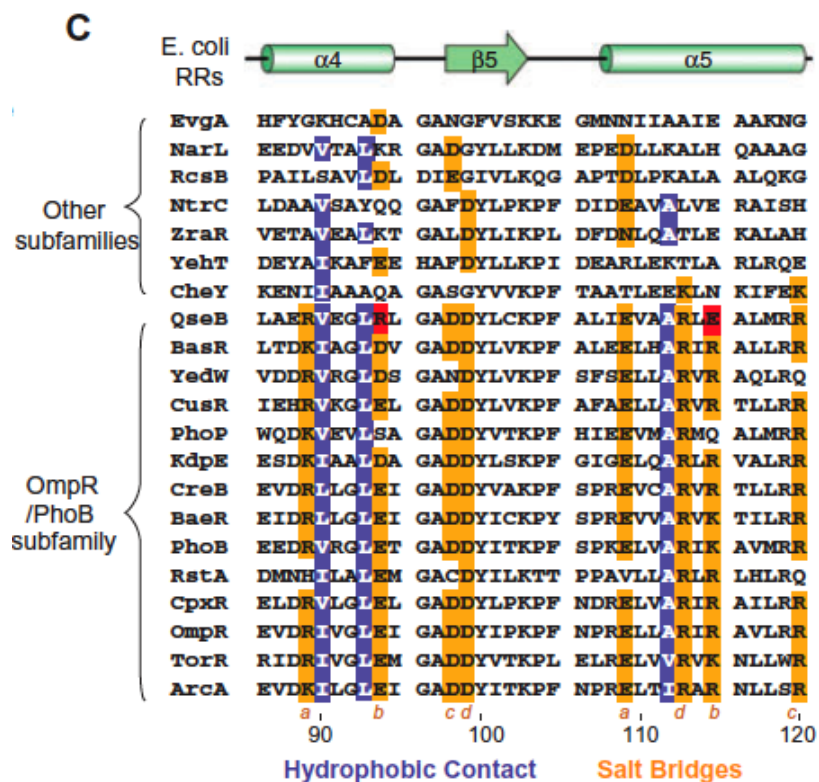


Figure 2.5: Sequence alignment of the $\alpha 4$ - $\beta 5$ - $\alpha 5$ region of *E. coli* response regulators. Response regulators include all *E. coli* OmpR/PhoB subfamily members and some representatives from other subfamilies. Among the highly conserved residues within the OmpR/PhoB subfamily are the highlighted residues that are involved in intermolecular interactions: hydrophobic contacts (blue); charged residues for salt bridge formation (orange). The pairing of charged residues is labelled by four pairs of letters a, b, c and d. A red highlight represents a pair of residues that are not conserved but could still complement each other with reversed charges. All these highlighted residues are not well conserved in response regulators from other families. Figure taken from Gao et al. (2008).

In the signaling cascade (Figure 1), the PhoQ histidine kinase is activated by low extracellular magnesium levels and is autophosphorylated at a histidine residue. PhoQ subsequently transfers the phosphate group from the conserved histidine of PhoQ to the conserved aspartate on the PhoP response regulator. Phosphorylation of PhoP presumably induces a conformational change to mediate homodimerization for DNA binding. The PhoP homodimer functions as a transcription factor by recognizing and binding to phoP boxes in promoters of PhoP-regulated genes. Through this mechanism, PhoP regulates expression of approximately 3% of the *Salmonella* genes in response to low magnesium levels to control physiological and virulence functions.

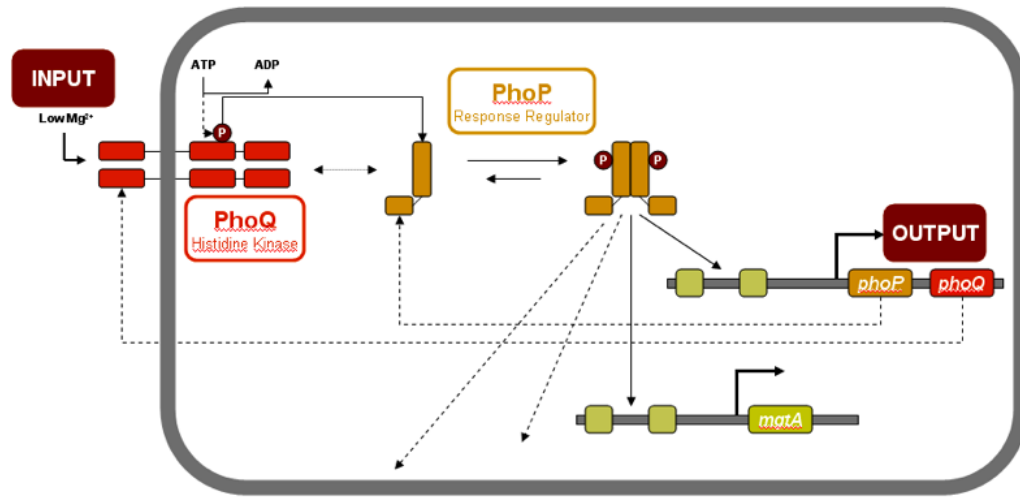


Figure 2.6: Schematic diagram of the *Salmonella enterica* PhoQ/PhoP two-component signal transduction system. PhoQ is the sensor histidine kinase, and PhoP is the response regulator. PhoQ senses low extracellular magnesium levels, leading to autophosphorylation at a conserved histidine residue. PhoQ transfers to phosphate group to a conserved aspartate residue on PhoP. Phosphorylation of PhoP mediates activation by causing a conformational change, allowing PhoP to homodimerize. PhoP recognizes phoP boxes at its DNA promoters (e.g., phoP, phoQ, mgtA) and function as a transcription factor to regulate virulence gene expression (including positive feedback loop).

It was hypothesized that bacterial virulence could be inhibited by disruption of the PhoQ/PhoP signaling pathway, specifically by targeting the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface of the PhoP response regulator. Preventing this essential protein-protein interaction (PPI)

would inhibit formation of the PhoP-DNA complex and its function as a transcription factor to regulate gene expression. 3.2 PhoP was chosen in this study as a prototype of the OmpR/PhoB family of response regulators to probe for PPI hot spots at the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface (Figure 2), in efforts to identify and characterize potential binding sites.

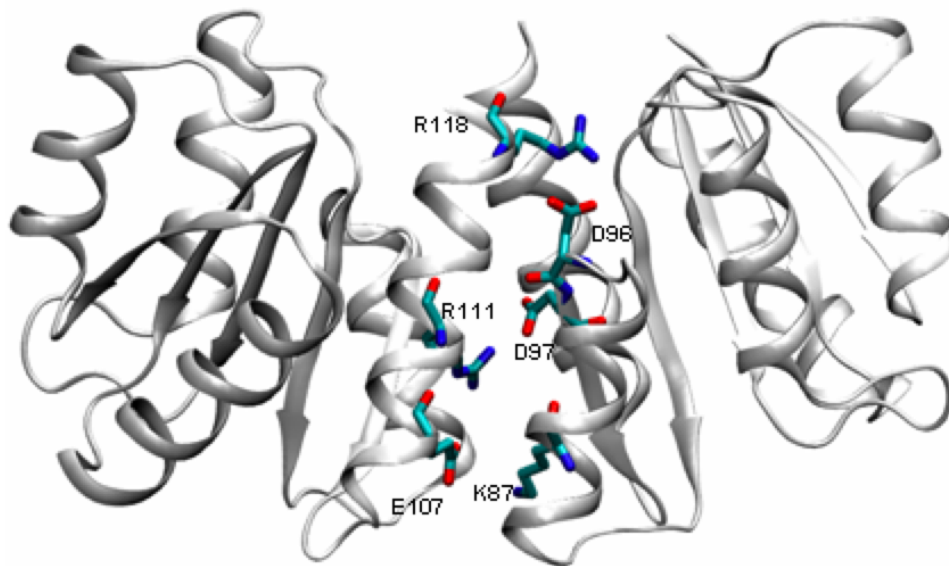


Figure 2.7: Critical salt-bridges at the PhoP $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface important for PhoP homodimerization and function. Residues important for dimerization (site-directed mutagenesis studies, Stock et al., unpublished) are shown in capped sticks and labeled by their residue name and number. Mutation of one of these residues decreases its ability to homodimerize. The PhoP homodimer (PDB ID: 2PKX) (Bachhawat and Stock, 2007) (cartoon) binds to phoP boxes in promoters of PhoP-regulated genes to modulate virulence gene expression.

The *S. enterica* PhoQ/PhoP signaling pathway is one of the better characterized bacterial TCST systems demonstrated to be important for virulence regulation. High-resolution X-ray crystal structures of both inactivated (PDB ID: 2PKX) and activated (PDB ID: 2PL1) *E. coli* PhoP were available with an interface highly similar in sequence to the one in *S. enterica* (differing only by 1 residue at the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface). (Bachhawat and Stock, 2007) For these reasons, *S. enterica* PhoP was an attractive target for investigation via structure-based drug design to test the effects of response regulator inhibition and its potential for virulence regulation.

2.4.4 Drug Discovery

A promising strategy towards antibacterial development is to target TCST regulatory systems to disrupt the expression of genes important for virulence. Targeting bacterial signal-transduction systems has only recently been demonstrated to be an effective potential strategy for antibiotics development. Rasko et al. (2008) targeted the sensor TCST component, the QseC histidine kinase, by the prevention of autophosphorylation, which led to disruption of the signaling cascade important for virulence regulation. Hung et al. (2005); Shakhnovich et al. (2007) demonstrated the feasibility of a small-molecule for homodimer inhibition and virulence gene regulation when they discovered virstatin to target the *Vibrio cholerae* ToxT. These studies demonstrated the feasibility of drug-like molecules targeting TCST-regulated gene expression important for virulence as a potential strategy for antibiotics development.

Chapter 3

Discovery of PhoP Response Regulator Inhibitors

3.1 Introduction

Infectious diseases have evolved resistance to most clinical antibiotics. Antibiotic resistance occurs at low levels in natural populations, but can become prevalent within a few years of the clinical adoption of an antibiotic. Antimicrobial therapeutics currently in clinical use either inhibit bacterial growth or induce death. One promising strategy is to combat virulence *per se* without inhibiting growth or inducing death, so less selective pressure will cause the bacteria to generate resistance. With the emergence of bacterial strains resistant to multiple antibiotics, there is an urgent need for the development of antibiotics with different modes of action less subjective to development of resistance.

Two-component signal transduction (TCST) is the predominant signaling scheme in bacteria to sense and respond to environmental changes for survival and proliferation. (Stock et al., 2000; Gao et al., 2007; Gao and Stock, 2009; Stock and Guhaniyogi, 2006; Gao and Stock, 2010) TCST regulatory systems are modular in terms of their arrangement of domains within their proteins within various pathways. In general, TCST regulatory systems are comprised of a transmembrane sensor histidine kinase (HK) and an intracellular receiver response regulator (RR) with conserved sequence, structural, and biochemical properties, allowing them to readily adapt to various modes of intracellular signaling. These signaling systems typically couple environmental stimuli to an adaptive response, participating in fundamental processes such

as regulating metabolism, as well as more specialized functions such as controlling virulence for the pathogens host.

The PhoQ/PhoP two-component regulator system is a major regulator of virulence in *Salmonella enterica* serovar Typhimurium, and also in a number of other gram-negative bacterial pathogens (e.g., *Shigella flexneri*, *Yersinia pestis*, *Neisseria meningitidis*). (Groisman, 2001; Vescovi et al., 1996; Moss et al., 2000; Oyston et al., 2000; Jamet et al., 2009; Johnson et al., 2001) PhoQ/PhoP in *S. enterica* is activated by low extracellular Mg^{2+} levels, acidic pH, and antimicrobial peptides (typical of human gut conditions during infection) to control various physiological and virulence functions. (Vescovi et al., 1996; Choi et al., 2009; Kato and Groisman, 2008; Shi et al., 2004) In the signaling cascade (Fig: 3.1), the PhoQ histidine kinase is activated by low extracellular magnesium levels and is autophosphorylated at a histidine residue. PhoQ subsequently transfers the phosphate group from the conserved histidine of PhoQ to the conserved aspartate on the PhoP response regulator. Phosphorylation of PhoP presumably induces a conformational change to mediate homodimerization for DNA binding. The PhoP homodimer functions as a transcription factor by recognizing and binding to phoP boxes in promoters of PhoP-regulated genes. Through this mechanism, PhoP regulates expression of approximately 3% of the *Salmonella* genes in response to low magnesium levels to control physiological and virulence functions. The *S. enterica* PhoQ/PhoP signaling pathway is one of the better characterized bacterial TCST systems demonstrated to be important for virulence regulation.

PhoP is a member of the OmpR/PhoB structural family, the largest structural family making up approximately 30% of all TCST response regulators. OmpR/PhoB family members are typically composed of 2 domains connected by a flexible linker: an N-terminal receiver domain and a C-terminal DNA-binding domain. Response regulators of the OmpR/PhoB family are characterized by a conserved $\alpha 4$ - $\beta 5$ - $\alpha 5$ structural motif of the receiver domain, a plastic surface important for dimerization. Due to the conservation and importance of the $\alpha 4$ - $\beta 5$ - $\alpha 5$ structural motif for function amongst OmpR/PhoB family members, it was hypothesized that bacterial virulence could be inhibited by disruption of the PhoQ/PhoP signaling pathway, specifically by targeting the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface of the PhoP response regulator. Preventing this essential protein-protein interaction (PPI) would inhibit formation of the PhoP-DNA complex

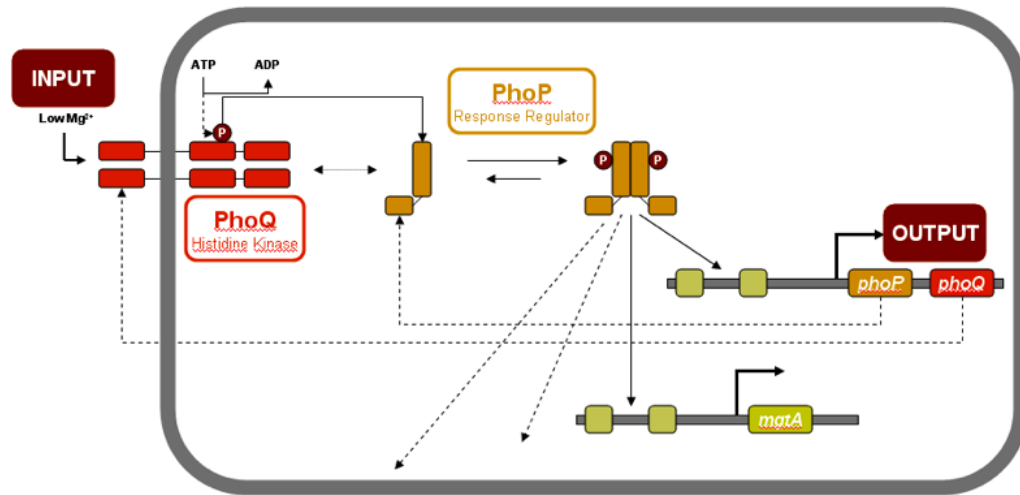


Figure 3.1: Schematic diagram of the *Salmonella enterica* PhoQ/PhoP two-component signal transduction system. PhoQ is the sensor histidine kinase, and PhoP is the response regulator. PhoQ senses low extracellular magnesium levels, leading to autophosphorylation at a conserved histidine residue. PhoQ transfers to phosphate group to a conserved aspartate residue on PhoP. Phosphorylation of PhoP mediates activation by causing a conformational change, allowing PhoP to homodimerize. PhoP recognizes phoP boxes at its DNA promoters (e.g., *phoP*, *phoQ*, *mgtA*) and function as a transcription factor to regulate virulence gene expression (including positive feedback loop).

and its function as a transcription factor to regulate gene expression. PhoP was chosen in this study as a prototype of the OmpR/PhoB family of response regulators to probe for PPI hot spots at the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface (Fig: 3.2), in efforts to identify and characterize potential binding sites. The response regulator N-terminal $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface share a common set of hydrophobic and charged residues involved in van der Waals contact and salt-bridges important for homodimerization and function, and are conserved amongst other response regulators across different bacteria species. (Gao et al., 2007) High-resolution X-ray crystal structures of both inactivated (PDB ID: 2PKX) and activated (PDB ID: 2PL1) *E. coli* PhoP were available with an interface highly similar in sequence to the one in *S. enterica* (differing only by 1 residue at the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface). (Bachhawat and Stock, 2007) For these reasons, *S. enterica* PhoP was an attractive target for investigation via structure-based drug design to test the effects of response regulator inhibition and its potential for virulence regulation.

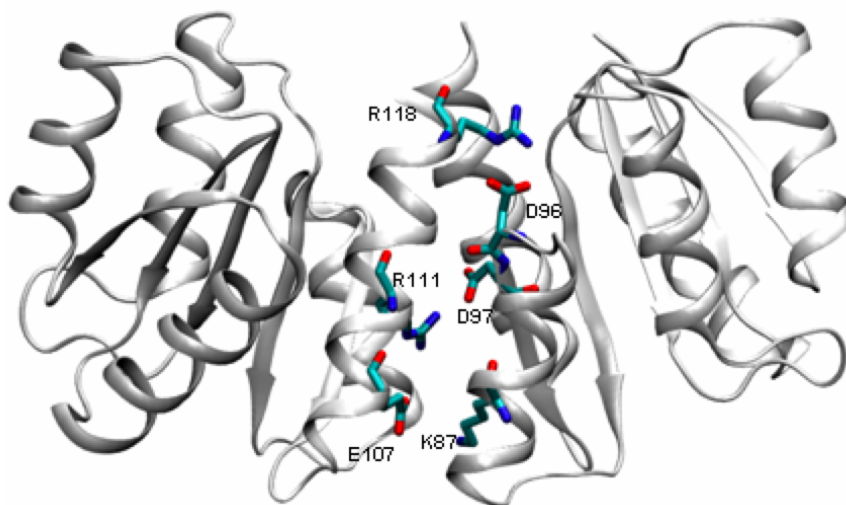


Figure 3.2: Critical salt-bridges at the PhoP $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface important for PhoP homodimerization and function. Residues important for dimerization (site-directed mutagenesis studies, Stock et al., unpublished) are shown in capped sticks and labeled by their residue name and number. Mutation of one of these residues decreases its ability to homodimerize. The PhoP homodimer (PDB ID: 2PKX) (Bachhawat and Stock, 2007) (cartoon) binds to phoP boxes in promoters of PhoP-regulated genes to modulate virulence gene expression.

Targeting bacterial signal-transduction systems has only recently been demonstrated to be an effective potential strategy for antibiotics development. Rasko et al. (2008)

targeted the sensor TCST component, the QseC histidine kinase, by the prevention of autophosphorylation, which led to disruption of the signaling cascade important for virulence regulation. Hung et al. (2005); Shakhnovich et al. (2007) demonstrated the feasibility of a small-molecule for homodimer inhibition and virulence gene regulation when they discovered virstatin to target the *Vibrio cholerae* ToxT. These studies demonstrated the feasibility of drug-like molecules targeting TCST-regulated gene expression important for virulence as a potential strategy for antibiotics development.

A prototype of the predominant class of bacterial signal transduction important for bacterial virulence is investigated as a proof-of-concept study towards this new strategy for antibiotics development. TCST systems predominate in control of bacterial expression and are completely absent in humans, making them an attractive class of targets for the development of new antibiotics with novel modes of action. To our knowledge, there are currently no known inhibitors of TCST response regulators. Drug-like compounds targeting PhoP, specifically the functionally important $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface, should selectively disrupt its function as a transcription factor and inhibit the expression of critical virulence genes. In this study, a hybrid approach coupling computational (Fig: 3.3) and experimental (Fig: 3.4) methods was used to predict, validate, and characterize drug-like inhibitors of the *S. enterica* PhoP response regulator.

3.2 Methods and Materials

3.2.1 Overview of Computational Strategy

3.3 illustrates an overview of the computational strategy to prioritize compounds targeting the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface of PhoP. Structure-based virtual screening was used to screen a drug-like version of the National Cancer Institute (NCI) Diversity library ($n = 1420$) with the crystal structure of the activated PhoP (PDB ID: 2PL1) (?). The docking procedure (labeled Docking) was performed using AutoDock to predict binding poses. A consensus scoring method (labeled Scoring) consisting of the predicted binding affinities from AutoDock (Morris et al., 1998), X-Score (Wang et al., 2002), and CScore (SYBYL) was used to better assess the binding affinities of the

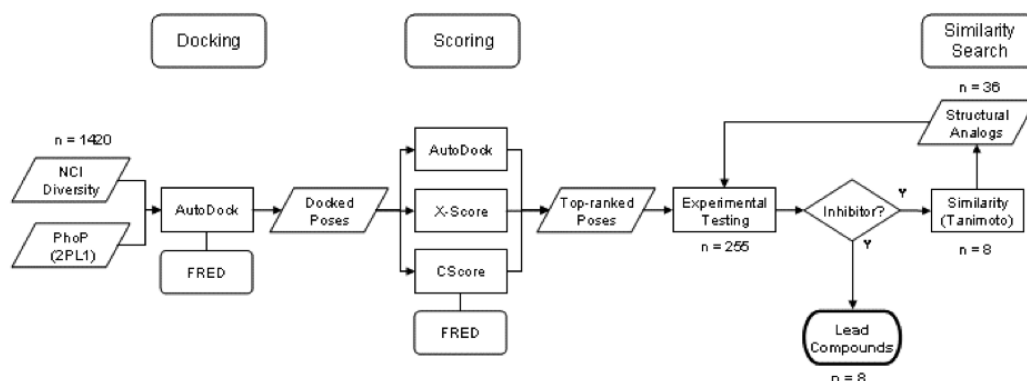


Figure 3.3: Schematic diagram of the computational workflow to predict for PhoP response regulator inhibitors. In the computational method, a drug-like version of the NCI Diversity library ($n = 1420$) was screened for inhibitors of activated PhoP (PDB ID: 2PL1). Docking was performed using AutoDock 4.0.1 and FRED (OpenEye). The predicted binding poses were rescored using X-Score and CScore.

A consensus scoring scheme composed of the predict affinities from AutoDock, X-Score, and CScore was used to rank the compounds. The top-ranked compounds were tested experimentally. Similarity search (Tanimoto Index) was performed on 8 compounds with inhibition activity to search for structural analogs.

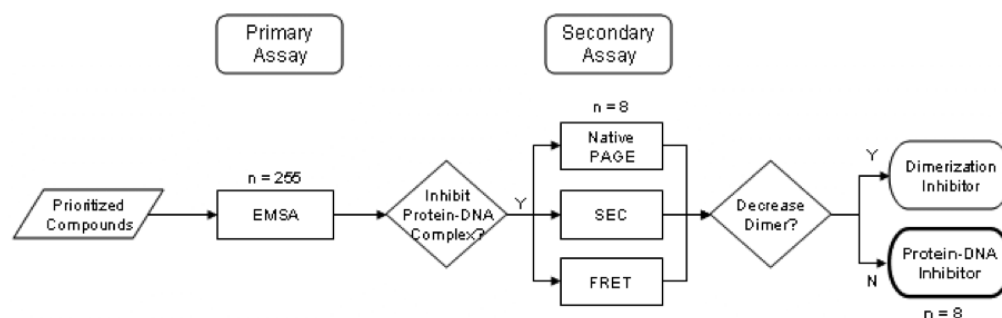


Figure 3.4: Schematic diagram of the experimental workflow to test for PhoP response regulator inhibitors. In the experimental strategy, the set of prioritized compounds were initially tested using EMSA. Compounds displaying inhibition activity by EMSA were further characterized for effects on dimerization using native PAGE, SEC, and FRET. Eight compounds inhibited formation of the protein-DNA complex necessary for virulence gene regulation.

resulting poses. The top-ranked compounds were tested experimentally to assess biological activity and characterize the mode of action. As an alternative computational method to identify potential PhoP inhibitors, FRED (OpenEye) was used for both docking and scoring (consensus method using PLP, ChemScore, OEChemScore). The top-ranked compounds from FRED ($n = 40$) were also tested experimentally.

From the experimental results, a similarity search based on the Tanimoto index (labeled Similarity Search) was performed to identify structural analogs ($n = 36$). These analogs were also experimentally tested. A total of 255 compounds were tested experimentally, and verified 8 compounds that inhibited formation of the PhoP-DNA complex necessary for gene regulation.

3.2.2 Compound Library and Crystal Structure

A drug-like version of the National Cancer Institute (NCI) Diversity I library composed of 1420 compounds (derived from a larger version of the NCI library of 140,000 compounds filtered based on criteria characteristic of drug-like compounds derived from Lipinskis Rule of 5), was screened. Compounds in the NCI Diversity library have purity 90% or better as characterized by liquid chromatography-mass spectrometry (LC-MS). The compound library was downloaded in SD format and converted to mol2 format using OpenBabel. Inactivated and activated *E. coli* PhoP (PDB ID: 2PKX and 2PL1, respectively) were downloaded from the Protein Data Bank. (Bachhawat and Stock, 2007)

3.2.3 Structure-Based Virtual Screening

AutoDock

AutoDock 4.0.1 was used to predict binding poses for compounds in the NCI Diversity library for experimental testing. (Morris et al., 1998; Huey et al., 2007) AutoDock was used to convert the ligand structures from the mol2 format to the AutoDock pdbqt format, with explicit hydrogen atoms and calculated Gasteiger charges. AutoDock-Tools (ADT) was used to prepare the protein structures. (Morris et al., 2009) Polar

hydrogens were added to the protein target, and Gasteiger charges were assigned. The structure files were saved in the pdbqt format.

Docking was performed to identify low-energy conformations (binding poses) of the compounds to sterically and chemically complement the binding site. The protein search area (grid spacing of 0.375 Å with dimensions of 40 × 40 × 40 Å) was centered at the α 4- β 5- α 5 motif where residues critical for dimerization in PhoP (e.g., Arg 111 and Arg 118 of activated PhoP) and other OmpR/PhoB family members are located, as shown in 3.5. (30) Lamarckian genetic algorithm was used to perform the ligand conformational searches to result in 30 binding poses for each ligand. The default parameters were used; to increase the sampling and accuracy, the following parameters were modified: `ga_pop_size` = 200; `ga_num_evals` = 5,000,000; `ga_run` = 30.

A drug-like version of the NCI Diversity library (1420 compounds) (?) filtered based on drug-like and lead-like features was screened using AutoDock 4.0.1. Virtual screening was performed on the *E. coli* activated PhoP (PDB ID: 2PL1), since the sequence at the dimer interface was highly similar to the *S. enterica* PhoP (*S. enterica* PhoP contains SER 93, while *E. coli* PhoP has ALA 93). AutoDock was used to generate docked poses of the NCI Diversity compounds at the PhoP dimer interface.

Fast Rigid Exhaustive Docking (FRED)

FRED (OpenEye) was used as an alternative computational tool for both docking and scoring. OMEGA2 (OpenEye) was used to generate bioactive ligand conformations according to the parameters from the study by Bostrom et al. Default parameters were used for FRED docking searches. Binding site was marked by using a docked ligand from the NCI Diversity library, and specifying the regions within 5 Å of it. The default FRED scoring function was used to predict binding affinities to rank ligands for experimental testing.

Rescoring of Predicted Poses

Due to the known limitations of scoring functions to accurately predict binding affinities, (Stahl and Rarey, 2001; Kitchen et al., 2004; Ferrara et al., 2004) especially in

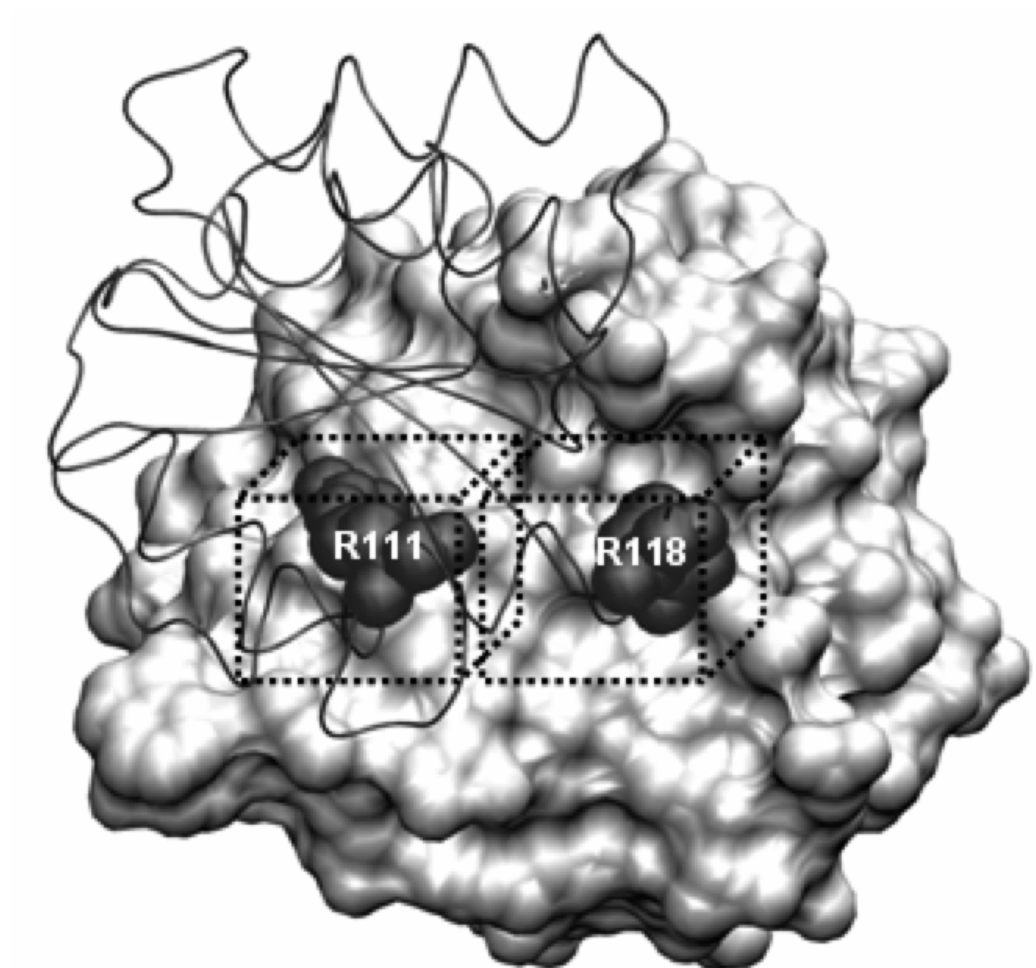


Figure 3.5: To prioritize compounds in NCI Diversity ($n = 1420$) to identify PhoP dimerization inhibitors, docking searches were constrained at the interface regions where salt-bridges critical for dimerization are located (AutoDock grid boxes used illustrated by hashed lines). Residues critical for dimerization (R111 and R118) are labeled and shown in CPK, in addition to PhoP (PDB ID: 2PKX) chain A (ribbon) and chain B (surface).

the case where the compounds are docked to a novel target with a relatively flat binding site typical of protein dimer interfaces, a consensus scoring scheme was used to prioritize and identify tight-binding ligands. (Halperin et al., 2002; Stahl and Rarey, 2001; Wang et al., 2004; Ferrara et al., 2004; Warren et al., 2006) Consensus scoring has proven to be effective in increasing the enrichment rate in a number of studies. (Wang et al., 2003; Cheng et al., 2009) Ligand binding poses predicted by AutoDock were rescored using CScore in SYBYL 7.3, which is composed of 4 separate scoring functions (D-score, ChemScore, PMF-score, G-score), and X-Score. (Wang et al., 2002; Eldridge et al., 1997) The consensus score for CScore resulted in integer scores between 1 and 4, with 4 being the highest. Docked poses were then sorted based on the consensus CScore value, and subsequently by the ChemScore binding affinities. The consensus score estimated by X-Score with default parameters, composed of 3 different scoring functions (HPScore, HMScore, HSScore), was used. The top 15% of the docked poses ranked by each of the scoring functions (the AutoDock binding free energy, the CScore consensus method, and the X-Score consensus method) were selected, and the docked poses found in all 3 scoring methods (6390 poses representing 179 compounds) were selected for experimental testing.

This computational protocol of docking and consensus scoring has been demonstrated to be effective in our research group for identifying lead compounds from the NCI Diversity library for various PPI studies. (Taylor et al., 2008) Since there are no known inhibitors of PhoP or any TCST response regulators with which to validate a computational protocol (there are only a small number of structural examples of protein-protein interaction inhibitors available as reviewed by Wells (Wells and McClendon, 2007), and no examples to our knowledge with a binding site that is structurally related to PhoP), this protocol was initially used for virtual screening.

3.2.4 Overview of Experimental Strategy

Figure 3.4 illustrates an overview of the experimental strategy to assess biological activity and characterize the mode of action. of the predicted compounds. As a primary assay, electrophoretic mobility-shift assays (EMSA) were used to assess whether the compounds ($n = 255$) can inhibit the formation of the *S. enterica* PhoP-DNA complex necessary for gene regulation. To test whether the compounds active in EMSA ($n = 8$)

affected PhoP dimerization, secondary assays used included native polyacrylamide gel electrophoresis (PAGE), size exclusion chromatography, and Förster resonance energy transfer (FRET)-based assay. Native PAGE was used to detect qualitative effects on *S. enterica* PhoP dimerization. SEC was used to better characterize the compounds ability to inhibit *S. enterica* PhoP dimerization by measuring the monomer-dimer profile. FRET was used as a higher-throughput method to quantify the effects of dimerization on *E. coli* OmpR/PhoB members to assess compound selectivity at the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface. Experimental results from the primary and secondary assays suggest the compounds inhibit formation of the PhoP-DNA complex, not by dimer inhibition, but in an allosteric manner to prevent DNA binding or by blocking the C-terminal DNA-binding domain. Potential mode of actions will be presented in a greater detail in Section 3.4.2.

3.2.5 Electrophoretic Mobility-Shift Assays (EMSA)

As a primary assay, an electrophoretic mobility-shift assay, also known as a gel-shift assay, was used to test the ability of the compounds to inhibit formation of the PhoP dimer-DNA complex necessary for gene expression. In the gel-shift assays, the protein is first activated by phosphorylation to induce the monomer-dimer equilibrium. Radiolabeled DNA is incubated with the protein, and the compound of interest is subsequently added to the mixture. After an incubation period, the sample is loaded and electrophoresed on a gel to observe formation of the protein-DNA complex. Compounds which inhibit protein-DNA complex formation will lead to a darker band further down in the gel representing unbound DNA, while compounds which do not effect complex formation will result in a single band higher up on the gel (bound complex). Band intensities can be quantified and fitted to a sigmoidal curve to estimate IC_{50} values in dose-response experiments.

PhoP was activated by in vitro phosphorylation with acetyl phosphate to induce formation of monomer-dimer equilibrium. Compounds dissolved in DMSO (99.9%) were then incubated with the monomer-dimer mix, which did not exceed 1% volume of DMSO in the final solution concentration. Radioactive-labeled DNA containing phoP boxes (GGTTTxxxxTGTTTA) were subsequently incubated with the mix to allow

formation of the PhoP-DNA complex. Samples were loaded on 4-20% TBE gels (Invitrogen) and electrophoresed. Gels were dried and analyzed using a phosphorimager (FujiFilm BAS-5000). Compounds that did not effect DNA-binding displayed a single band upstream which represented the bound PhoP-DNA complex. Compounds that inhibited formation of the PhoP-DNA complex displayed a shift, resulting in a band of the PhoP-DNA complex and/or a downstream band which represented unbound DNA.

ImageJ was used to quantitate the EMSA band intensities (for both bands representing bound and unbound DNA). KaleidaGraph was used to perform the curve fitting using a sigmoidal function to derive the IC_{50} values from dose response experiments.

The phop DNA fragments for electrophoretic mobility-shift assays were amplified by PCR using primers 312 and 369, and genomic DNA of wild-type *S. enterica* as template. The DNA fragments were isolated by running an electrophoretic gel and purified using QIAquick columns (Qiagen). To radiolabel the DNA fragments, 100 ng of phop DNA was used with T4 polynucleotide kinase and γ - ^{32}P ATP and incubated at 37 C° for 1.5 h. Unincorporated DNA was removed using G-50 microcolumns (Amersham). 20,000 CPM of labeled probe (12 fmol), 200 ng poly(dI-dC)-poly(dI-dC) (Amersham), and phosphorylated PhoP-His₆ were mixed with binding buffer (50 mM Tris-HCl pH 8.0, 50 mM KCl, 50 μ g/ml BSA) in a total volume of 20 μ l and incubated for 20 min at room temperature.

For electrophoretic mobility-shift assays, *S. enterica* PhoP-His₆ was phosphorylated with acetyl phosphate. PhoP (0.6-1.2 nmol) was incubated in 50 μ l of phosphorylation buffer (50 mM Tris-HCl pH 7.5, 50 mM KCl, 10 mM MgCl₂) containing 10 mM acetyl phosphate (Sigma-Aldrich) for 2.5 h at room temperature. Excess acetyl phosphate was removed from phosphorylated PhoP-His₆ using a Micro Bio-Spin 6 Chromatography Column (Bio-Rad) equilibrated with Tris buffer. Phosphorylated PhoP-His₆ were kept at 4 C° and used within 24 h.

3.2.6 Similarity Search

To identify similar compounds of the experimentally confirmed inhibitors, a similarity search (Tanimoto index) using the NCI website (<http://129.43.27.140/ncidb2/>) was performed to search the larger NCI library of 140,000 compounds. Using the experimentally confirmed compounds as the query, 36 compounds from the full NCI library were found, ordered, and experimentally tested by EMSA for inhibition activity.

3.2.7 Native Polyacrylamide Gel Electrophoresis (PAGE)

To test the effect of compounds on PhoP dimerization, a native polyacrylamide gel electrophoresis (PAGE) assay was used. The protein is first activated by phosphorylation to induce formation of the monomer-dimer equilibrium (same method as described in EMSA). The compound of interest is incubated with the equilibrium mixture. Samples are loaded and electrophoresed on a native PAGE gel to separate the monomer and dimer states. Proteins can be visualized by gel staining. Compounds with no effects on dimerization should display 2 bands: one less intense and more downstream band representing the monomer, and a darker and more upstream band representing the dimer. Compounds with inhibition effects on dimerization will display a more intense band representing the monomer, and a less intense band representing the dimer.

S. enterica PhoP-His₆ was phosphorylated in the same manner as by EMSA. Compounds were incubated with the active protein for 30 min, and loading buffer was added after. Samples were loaded and ran on a native gel (6% Tris-Glycine; Invitrogen) at 4 C°. Gels were then stained with coomassie blue to visualize the bands representing the protein.

3.2.8 Size-Exclusion Chromatography (SEC)

Size-exclusion chromatography was used to better characterize the monomer-dimer distribution. Protein is activated by phosphorylation (similar manner to EMSA and native PAGE) to induce the monomer-dimer equilibrium. Compound of interest

is incubated with the equilibrium mix, and injected to and separated by the size-exclusion column. Relative absorbance (A_{280}) is detected from the elutions to quantify the amount of protein present. Compounds which inhibit dimerization will display a larger peak in the earlier elution representing the monomer, and a smaller peak in a later elution representing the dimer. Compounds which do not effect dimerization will not display any changes in the monomer-dimer absorbance profile. (Data not shown)

Purified PhoP-His₆ was phosphorylated using 50 mM ammonium phosphoramidate (synthesized by the method of Sheridan et al.(48)) and 20 mM MgCl₂ for 30 min at room temperature. Compounds were subsequently incubated with PhoP for 30 min. Samples of inactivated and activated PhoP (100 μ l) were individually applied to a Superdex 75 column (GE Healthcare) equilibrated with elution buffer (50 mM Tris/HCl, pH 8.0, 150 mM KCl). Proteins were eluted in the same buffer at a flow rate of 0.5 ml/min. Protein concentration was assessed by measuring the OD₂₈₀. Fractions were collected and analyzed using native PAGE. A molecular weight standards kit (Sigma-Aldrich) was applied to the column for calibration.

3.2.9 Förster resonance energy transfer (FRET) analyses

FRET analyses were used to detect and quantify effects on dimerization on *E. coli* PhoP. In FRET, a cyan fluorescent protein-fused PhoP (CFP-PhoP) and a yellow fluorescent protein-fused PhoP (YFP-PhoP) are activated by phosphorylation. Phosphorylation induces dimerization between CFP-PhoP and YFP-PhoP, and brings the CFP and YFP in proximity where an energy exchange occurs. PhoP dimerization is characterized by a decrease of cyan emission and an increase of yellow emission. The rate of FRET increase depends on the rates of phosphorylation and dimerization. Compounds which inhibit dimerization will lead to a decrease in the FRET signal, while compounds which do not effect dimerization will not change the FRET signal.

FRET analyses of fluorescent protein (FP) fused PhoP proteins, CFP-PhoP and YFP-PhoP, were performed as described in (Gao et al., 2008) Phosphorylation of FP-PhoP was initiated by addition of MgSO₄ and phosphoramidate to the mixture of CFP-PhoP, YFP-PhoP and indicated compounds. The final concentrations were

0.6 μ M CFP-PhoP, 2.5 μ M YFP-PhoP, 20 mM phosphoramidate, 5 mM MgSO_4 , 100 μ M compounds and 1% (v/v) DMSO. Fluorescence was followed at 475 nm and 530 nm with excitation at 430 nm. The ratio of 475 nm and 530 nm emissions was defined as the FRET ratio to monitor the interaction between CFP-PhoP and YFP-PhoP.

3.2.10 *S. enterica* PhoP Expression and Purification

S. enterica PhoP response regulator with a C-terminal His₆-tag were overexpressed in *E. coli* strain BL21(DE3) transformed with the pT-7-7 plasmid. Cells were grown in Luria-Bertani medium with ampicillin (100 mg/liter) and incubated until mid-logarithmic phase at 37 C°. Overexpression was induced with 1 mM isopropyl- β -D-thiogalactopyranoside (IPTG) and incubated overnight at 25 C°.

For purification, cells were harvested by centrifugation and then washed and resuspended in PBS, and stored overnight in -80 C°. Cells were then thawed at 4 C°, suspended in lysis buffer (50 mM NaH_2PO_4 , pH 8.0, 300 mM NaCl, 10 mM imidazole) and lysed by sonication. Cells were then centrifuged and the cell lysate was collected as the supernatant. Cell lysate was applied to a Ni^{2+} column (GE Healthcare). Unbound proteins were removed with elution with Buffer A (50 mM Tris-HCl pH 8.0, 500 mM NaCl, 10 mM imidazole). Bound proteins were eluted with a 0 to 100% gradient of Buffer A to Buffer B (50 mM Tris-HCl pH 8.0, 100 mM NaCl, 500 mM imidazole). Fractions were collected and analyzed by SDS/PAGE. Proteins were concentrated using Amicon Ultra-15 filters (Millipore) and stored in storage buffer (20 mM Tris pH 7.8, 100 mM KCl, 20% glycerol) at -80 C°. Protein concentrations were determined by measuring the OD₂₈₀ (NanoDrop spectrophotometer). Purity was assessed by SDS/PAGE.

3.3 Results

3.3.1 Computational Predictions

Computational strategy identified inhibitors which potentially bind at the PhoP $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface. A total of 259 compounds were tested experimentally. From the results obtained by AutoDock and the consensus scoring scheme, the following number of compounds include: 119 compounds that used a search grid centered at Arg 111, 60 compounds that used a search grid centered at Arg 118. (Fig: 3.5) From the results obtained by docking using FRED (search area centered at Arg 111), 40 compounds were tested.

3.3.2 Experimental Validation and Characterization

Eight compounds inhibited formation of the PhoP-DNA complex. A total of 255 compounds were tested by EMSA: 119 compounds from the Arg 111 binding site screen using AutoDock, 40 compounds from the Arg 111 binding site screen using FRED, 60 compounds from the Arg 118 binding site screen, and 40 compounds identified from a Tanimoto similarity search for analogs of the initial hits within the entire NCI library. Eight compounds (NSC9608, NSC35489, NSC45576, NSC48630, NSC65238, NSC88915, NSC118806, NSC168197) (See Fig: 3.6) displayed inhibition activity via disruption of the PhoP-DNA complex. The 8 compounds inhibited PhoP-DNA complex formation in a dose-dependent manner (See Fig: 3.7), with 6 of the 8 compounds displaying an IC_{50} in the micromolar range (3.6 μM to 285 μM).

Analogues (36 in total) of the 8 inhibitors based on Tanimoto similarity (90% similar) from the NCI library (140,000 compounds) did not display inhibition activity by EMSA. Inactivity of these analogues may be due to the sensitive preference of the PhoP binding site at the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface.

Compounds did not effect PhoP dimerization. The 8 compounds did not display any effects on band intensity for the band representing the PhoP monomer and dimer in the native PAGE assay (See Fig: 3.8). These results suggested that the

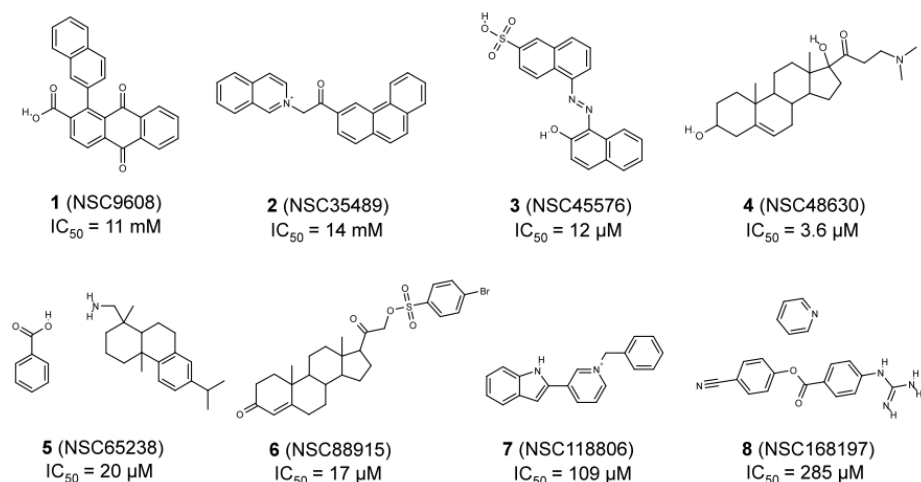


Figure 3.6: Chemical structures of the 8 compounds (1-8) that inhibited formation of the PhoP-DNA complex by electrophoretic mobility-shift assays (EMSA) with their NSC number and their estimated IC_{50} values

compounds do not act as dimerization inhibitors. However, band intensity changes may not have been observed due to the inability of the gel to distinctly separate the phosphorylated monomer and dimer species.

The 8 compounds also did not display any changes on the monomer-dimer profile of PhoP as compared with a control (DMSO) using size-exclusion chromatography, which performs a better separation and characterization of the oligomeric states than native PAGE. (data not shown) One caveat of using SEC as a binding assay is that a compound exhibiting a rapid off-rate cannot be detected with SEC due to time (30 min.) needed to perform the separations.

Results from the FRET assays using *E. coli* PhoP also did not suggest any degree of dimer inhibition (See Fig:3.9). However, with the use of CFP-PhoP and YFP-PhoP, one possibility is that the compounds may bind to CFP or YFP instead of PhoP, resulting in undetected inhibition effects. Some compounds (e.g., 1, 2) were inherently colored and affected the detection of FRET signal, and therefore could not be properly tested for dimer inhibition. Compounds 1, 3 and 8 displayed a smaller FRET ratio changes as compared with the positive control sample (e.g., CY +DMSO, CY PhoP). However, compounds 1 and 3 interfere with the FP fluorescence while compound 8 alters the fluorescence ratio of FP-PhoP pairs. Therefore the FRET

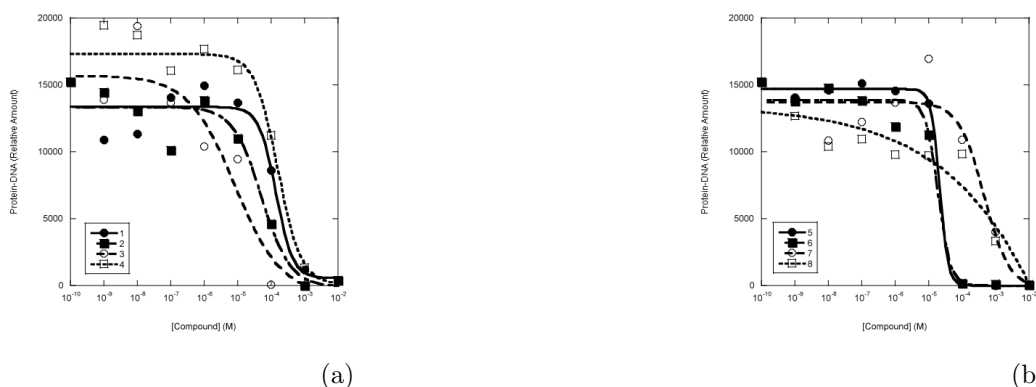


Figure 3.7: Dose response curves (fitted by sigmoidal function, KaleidaGraph (?)) from electrophoretic mobility-shift assays in presence of compounds 1-4 (a) and 5-8 (b). The points plotted represent the relative amount of PhoP-DNA complex formed in EMSA experiments.

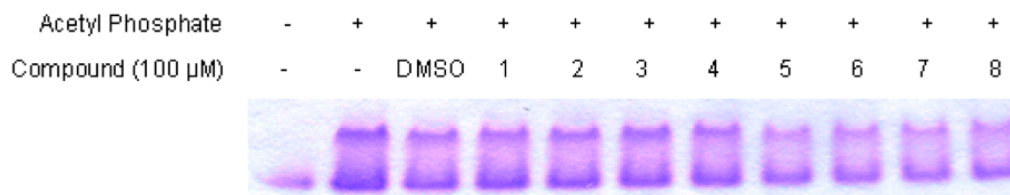


Figure 3.8: Native PAGE results of *S. enterica* PhoP in presence of 8 inhibitor compounds (1-8). PhoP dimerization was induced by phosphorylation via acetyl phosphate. An unphosphorylated sample without compound (first lane from left) was used as a negative control. Samples with buffer only (second lane from left) and with 1% (v/v) DMSO (third lane from left) were used as positive controls.

method is not sufficient to assess the inhibition of compounds 1, 3 and 8. FRET also cannot detect compounds which bind to the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface to inhibit DNA binding in an allosteric manner, since only the N-terminal of the PhoP response regulator bound to CFP or YFP was used.

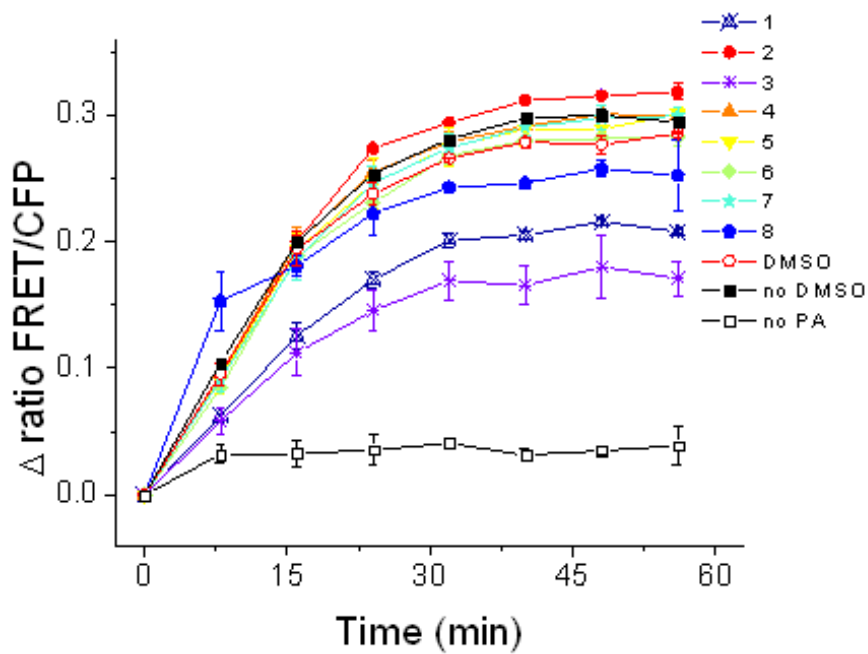


Figure 3.9: Time-dependent FRET of *E. coli* FP-PhoP pairs in the presence of inhibitor compounds (1-8). Inactive PhoP (no PA) was used as a negative control, while activated PhoP (no DMSO) and sample with DMSO (DMSO) were used as positive controls.

3.4 Discussion

3.4.1 Selective Inhibition of *S. enterica* PhoP

The 8 compounds did not display any effects on the signal in the FRET assays with *E. coli* PhoP, which suggested no inhibition of dimerization (See Figure 3.9). The compounds may bind to the *E. coli* PhoP $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface, but have no effects on dimerization. However, since the *E. coli* PhoP linker and C-terminal was not included in the FRET analyses, there remains a possibility that the compounds may bind to *E. coli* PhoP to inhibit DNA binding. These results obtained so far suggested that the compounds might selectively bind to the *S. enterica* PhoP $\alpha 4$ - $\beta 5$ - $\alpha 5$ dimer interface to inhibit formation of the protein-DNA complex, perhaps by an allosteric mechanism.

3.4.2 Potential Modes of Action

Experimental results from a series of in vitro and biophysical assays (EMSA, Native PAGE, SEC, FRET) suggested that the compounds tested disrupt PhoP-DNA complex formation, but not via direct homodimer inhibition. Instead, one potential mode of action is to bind to the C-terminal DNA-binding domain, a more direct mechanism to disrupt formation of the PhoP-DNA complex. Another possible mode of action is by binding to the N-terminal regulatory receiver domain or the linker region to act in an allosteric manner and prevent conformational changes necessary for DNA binding. The former may prove to be a more effective strategy for selective inhibition of response regulators, due to the conserved $\alpha 4$ - $\beta 5$ - $\alpha 5$ structural motif of the dimer interface.

Virtual screening may have helped to identify compounds binding the interface. The inhibitors did not block dimerization, but did prevent formation of PhoP-DNA complex necessary for gene regulation, perhaps by binding at the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface and acting by an allosteric mechanism.

3.4.3 Challenges of Structure-Based Design

In the pursuit of discovering PPI inhibitors, the original objective was to use small-molecules as probes to identify and characterize potential binding sites at the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface of *S. enterica* PhoP to block homodimerization. Through the discovery of dimerization inhibitors presumably binding at the sites where the critical residues are located, it was hoped that the findings would lead to a better understanding of the physical properties underlying molecular recognition of protein-ligand interactions at PPI. While the original intent of this study was to target the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface to block dimerization, the experimental results suggest an unexpected finding: drug-like compounds may bind at the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface and function in an allosteric manner and cause a conformational change to prevent DNA-binding. Further experimental characterization (e.g., X-ray crystallography, NMR) will be necessary to identify potential binding sites and to elucidate the atomic details of the complexes to determine the mode of action.

Structure-based discovery of protein-protein interaction inhibitors remains a significant challenge as seen by the high percentage of false positives from the computational predictions in this study. Docking and structure-based design methods to incorporate protein flexibility will need to be used to identify and design potential allosteric inhibitors, or ones that bind by an induced-fit mechanism. Another potential limitation is in the scoring functions used for virtual screening, which cannot accurately predict binding affinities, especially for systems that are not present in the scoring function training sets (e.g., binding sites with relatively flat surfaces). Even if the docking modes may be correctly identified, limitations in scoring accuracy may classify potential tight-binding ligands as weak-binders, resulting in a high percentage of false positives. Perhaps the use of first-principles methods for estimating binding affinities may help circumvent this limitation, assuming that the correct ligand conformation was predicted by the docking procedure.

Also, availability of a high-resolution crystal structure of PhoP bound to its DNA promoter should better elucidate the conformation of the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface in the biologically-active form bound to DNA, and provide atomic-level detail of potential binding sites necessary for docking and structure-based drug design. However, inherent inaccuracy of using a static structure for molecular design may lead to

false-negatives using structure-based design methods without incorporation of protein flexibility. For more accurate computational modeling, in particular to target PPI, improvements in both docking and scoring methods are necessary. With the increasing interest in targeting PPI and availability of structural and binding affinity data, development of more accurate and robust SBDD methods to target PPI will become possible.

3.4.4 Discovery of First-In-Class PhoP Inhibitors

In this study, 8 first-in-class inhibitors of the *S. enterica* PhoP TCST response regulator were discovered using a hybrid approach coupling computational and experimental methods for molecular design. Potential mode of action of the compounds was characterized by a series of in vitro and biophysical assays. Compounds may potentially bind at the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface and act as allosteric inhibitors, rather than dimerization inhibitors, to prevent DNA binding. In addition, it is also possible that the compounds may act by binding to the C-terminal DNA-binding domain to directly block DNA binding. Discovery of first-in-class PhoP inhibitors should serve as a proof-of-concept for targeting TCST response regulators as a novel strategy to inhibit bacterial virulence.

3.5 Conclusions and Future Directions

Targeting two-component signal transduction response regulators to modulate virulence gene expression is a promising strategy for antibiotics development. With the increasing resistance of bacterial pathogens to current therapeutics, antibiotics that can prevent virulence instead of inhibiting growth or inducing death may lead to less selective pressure for the generation of resistance.

3.5.1 Summary

In this study, 8 compounds have been discovered by coupling computational and experimental methods to disrupt formation of the *S. enterica* PhoP-DNA complex necessary for gene regulation. Eight compounds inhibited the PhoP-DNA complex formation in a dose-dependent manner by EMSA. Based on the experimental results, the PhoP inhibitors may potentially bind to the plastic $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface and act by an allosteric mechanism to prevent DNA binding. Alternative modes of action include binding to other regions of the N-terminal domain to act in an allosteric manner, or the C-terminal DNA-binding domain to directly inhibit formation of the PhoP-DNA complex. Experimental results obtained from a series of biochemical and biophysical assays suggest a potential of these compounds to inhibit bacterial virulence.

3.5.2 Future Directions

Further elucidation of the mode of action to assess the potential of the 8 compounds as virulence inhibitors is planned. Structural analogs of the 8 compounds can be designed to enhance affinity and characterize structure-activity relationships. Structural studies such as X-ray crystallography and NMR must be performed in order to validate the mode of action and elucidate the protein-ligand interactions in atomic detail.

3.6 Acknowledgements

The authors thank Mr. Malcolm Tobias and Dr. Chris Ho of the Center for Computational Biology for technical support; Dr. Christina Taylor for intellectual and technical contributions, and critical reading of this manuscript. Gratitude is expressed to Dr. Won Sik Seo, Dr. J. Christian Perez, Mrs. Tammy Latifi and other members of the Groisman, Stock and Havranek labs for technical assistance and scientific discussions. Y.T.T. is supported by a PhRMA Predoctoral Fellowship in Informatics. G.R.M. acknowledges financial support from NIH GM068460 and the Department of Defense (HDTRA1-09-CHEM-BIO-BAA). A.M.S. and E.A.G. are investigators of

the Howard Hughes Medical Institute. We also acknowledge the NCI Developmental Therapeutics Program for providing compounds for experimental testing.

Chapter 4

PHOENIX: Scoring Function to Predict Binding Affinities

4.1 Introduction

Predicting binding affinity is one of the most critical and challenging components to computer-aided structure-based drug design. (Ajay and Murcko, 1995; Gohlke and Klebe, 2002) Methods for predicting binding affinity are instrumental in a variety of applications, including molecular docking to identify a native binding mode, virtual screening of compound libraries to identify lead compounds, and lead optimization for enhancing binding affinity and target specificity. (Kitchen et al., 2004; Lyne, 2002; Shoichet, 2004) Despite significant advances in first-principle methods for predicting binding affinity, (Beveridge and Dicapua, 1989; Massova and Kollman, 2000; Hansson et al., 1998; Wang et al., 2001; Jiao et al., 2008) empirical scoring functions that are fast and relatively accurate are still widely used in drug discovery. (Bohm and Stahl, 2002) For virtual screening studies where libraries up to millions of compounds are screened against a target of interest, a scoring function is needed to rapidly assess multiple binding modes of each multiple conformers generated for each compound. This is also the case for in silico lead optimization where a large number of analogs are computationally constructed and assessed. In addition to speed of evaluation for virtual screening, other scoring functions can be accurate at an atomic level for structure-based drug design in characterizing the dominant physical forces in molecular recognition during ligand binding. Moreover, empirical scoring functions should be transferable and not require careful individual validation for each system under

study, making them more suitable for use in new problems with limited experimental data.

Empirical scoring functions aim to represent the atomic interactions of protein-ligand complexes by the use of relatively simple quantitative descriptors to capture the physicochemical forces governing protein-ligand complex formation. The underlying assumption in scoring functions is that the physical and chemical interactions of protein-ligand interactions can be quantitatively captured using a set of descriptors, and the sum of these descriptors will accurately predict binding affinities. In practice, each descriptor is weighted by a coefficient, derived by a linear regression method through training on experimental data from binding assays, resulting in an equation for calculating binding affinities.

Over the last 20 years, a number of scoring functions have been developed, with some notable ones being SCORE1 (Bohm, 1994), SCORE2 (Bohm, 1998), ChemScore (Eldridge et al., 1997), X-Score (Wang et al., 2002), Lig-Score (Krammer et al., 2005), PLP (Gehlhaar et al., 1995; Verkhivker et al., 2000), DrugScore (Gohlke et al., 2000), CScore (?), GOLD (Jones et al., 1997), and SFCscore (Sottriffer et al., 2008). These scoring functions differ by their choice and implementation of descriptors to capture the physicochemical interactions, the size and diversity of the training set, and the regression method used to derive the predictive equations. A number of reviews on scoring functions and assessments of their performance and applicability have been published. (Halperin et al., 2002; Wang et al., 2003; Stahl and Rarey, 2001; Wang et al., 2004; Ferrara et al., 2004; Warren et al., 2006; Cheng et al., 2009)

Empirical scoring functions generally predict either the free energy of binding (ΔG) or the dissociation constant (K_d), both of which can be derived from the other. Recent calorimetric studies have elucidated the compensating enthalpic and entropic changes associated with binding free energy. (Kawasaki et al., 2010; Freire, 2008, 2009) In a review from Ladbury, Klebe, and Freire, (Ladbury et al., 2010) the binding free energies of first-in-class HIV-1 protease and HMG-CoA reductase inhibitors were shown to be due largely from optimizing entropy (ΔS), while improving binding affinity of subsequent analogs was predominantly the result of improving enthalpy (ΔH). Marlow et al. (2010) has experimentally demonstrated that changes in protein

conformational dynamics can serve as an indication of the changes in protein conformation entropy, which may also play an important role in high-affinity protein-ligand complexes. Roy and Laughton (2010) have demonstrated using molecular dynamics simulations the importance of phenomena such as entropy-entropy compensation, dewetting of the protein binding site, and ligand configuration entropy in the form of rotational freedom in contributing to changes in entropy. Because the binding free energy is composed of these compensating thermodynamics forces, the ability to accurately predict enthalpy (ΔH) and entropy ($T\Delta S$) independently should provide additional insight during structure-based drug design studies. Results from these experimental and theoretical studies illustrate the importance of considering both enthalpy and entropy contributions separately and in a greater detail for structure-based drug design studies.

Current empirical scoring functions contain descriptors that mainly take into account the changes of enthalpy (ΔH) in binding, and have used rudimentary methods such as the number of rotamers on a ligand, calculated partition coefficient (XlogP), and complementary hydrophobic surface area estimation to describe changes in entropic forces ($T\Delta S$). The lack of an accurate entropic description of protein-ligand interactions is surely the major reason why scoring function accuracy has been limited; they can predict enthalpic contributions accurately, but fail to predict entropic contributions, resulting in limited accuracy in predicting binding free energy. In the development of PHOENIX, addition terms to describe the shape and volume of both the ligand and protein binding site were included. Volume-based descriptors of the ligand and binding site heuristically capture the rotational and translation entropy contributing to the configurational entropy of the system. Developing entropy models using shape and volume-based descriptors should lead to more accurate binding affinity predictions.

Development of PHOENIX aimed to take advantage of the increasing application of isothermal titration calorimetry (ITC) in medicinal chemistry and the recent availability of databases (PDBcal and SCORPIO) (Li et al., 2008; Olsson et al., 2008) containing both X-ray crystallographic structures of protein-ligand complexes and ITC experimental determination of both enthalpic and entropic contributions to binding free energy. PHOENIX, derived from the VALIDATE scoring function, includes additional shape- and volume-based descriptors to better capture entropic contributions

typically not accounted for in scoring functions. A diverse set of 112 protein-ligand complexes with resolution ≤ 2.0 Å and thermodynamics parameters measured from ITC was used for training. A set of 42 descriptors, including 7 shape and volume descriptors calculated using FPOCKET (Guillox et al., 2009), were used as a heuristic method to capture the physicochemical forces underlying protein-ligand interactions. Partial least squares of latent variables (PLS) was used to assign coefficients for each descriptor, and to independently derive regression equations to calculate ΔH and $T\Delta S$.

4.2 Materials and Methods

4.2.1 Training Set

Information on protein-ligand complexes with crystallographic structures and thermodynamic parameters from isothermal titration calorimetry were obtained from PDBcal and SCORPIO databases. Experimental values of ΔG , ΔH , and $T\Delta S$ were obtained from the database websites, while X-ray crystallographic structures were downloaded from the Protein Data Bank (PDB). Only structures of complexes with a crystallographic resolution ≤ 2.5 Å were used in the initial compilation of the training set. Additional metrics such as free R value (R_{free}) (Brunger and Rice, 1997) (See Equation: 1.4) and diffraction-component precision index (DPI) (Blow, 2002) (See Equation: 1.5) were used to assess structural quality. R_{free} is a measure of the degree to which an atomic model predicts a subset of the observed diffraction data that has been omitted from the refinement process. DPI is a measure of the quality of the structural model derived from the diffraction data. However, due to the scarcity of complexes with a resolution of ≤ 2.5 Å, ITC parameters, and R_{free} values, the resolution (≤ 2.0 Å) was used as the final criteria to obtain the PHOENIX training set of 112 complexes. Nine different subsets of the 162 complexes were evaluated for predictive ability: Set 68, includes structures with resolutions ≤ 2.0 Å, $R_{free} \leq 0.3$, DPI ≤ 0.3 , ligand molecular weight < 1000 daltons, ΔH , $T\Delta S$; Set 82, includes structures with resolutions ≤ 2.0 Å, $R_{free} \leq 0.3$, DPI ≤ 0.3 , ligand molecular weight < 1000 daltons; Set 91, includes structures with resolutions ≤ 2.0 Å, $R_{free} \leq 0.3$, DPI ≤ 0.3 ; Set 105, includes structures with resolutions ≤ 2.0 Å, ligand molecular weight < 1000 daltons;

Set 112, includes structures with resolutions ≤ 2.0 Å; Set 127, includes structures with resolutions ≤ 2.0 Å, 15 complexes with resolution between 2.0 and 2.5 Å also present in PDBbind test set; Set 140, includes structures with resolutions ≤ 2.25 Å; Set 153, includes structures with resolutions ≤ 2.5 Å, ligand molecular weight < 1000 daltons; Set 162, includes structures with resolutions ≤ 2.5 Å. These subsets were selected to evaluate whether the quality of the crystal structures and diversity of the training set impacted the performance of the scoring function. Of the 9 subsets tested, Set 112 ?? (resolutions ≤ 2.0 Å) resulted in the best performing binding free energy (ΔG) model.

Distribution histograms of the thermodynamic parameters and molecular weight of the complexes of the final PHOENIX training set are shown in Figures 4.1 4.2 4.3 4.4. Coefficients from PLS regression is shown in Table 4.1.

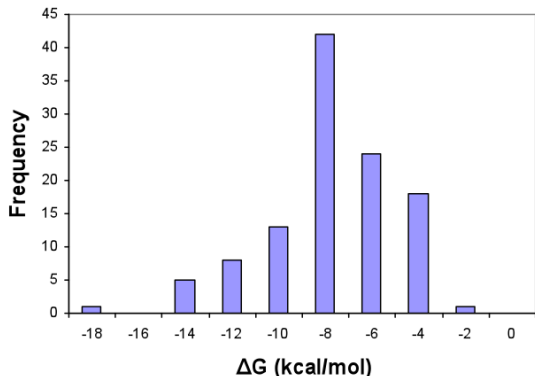


Figure 4.1: Distribution histogram of the change in binding free energy (ΔG) (mean = -8.73 kcal/mol, std. dev. = 2.73 kcal/mol)

4.2.2 Structure Preparation

Protein-ligand complexes downloaded from the PDB were prepared as follows. Protein structure was extracted from the complex using SYBYL 7.3 (Tripos). Water molecules present in the complex were kept as part of the protein structure for an explicit solvent representation. In cases where multiple chains or subunits were present, the chain or subunit that was most complete was selected, which was chain A in most

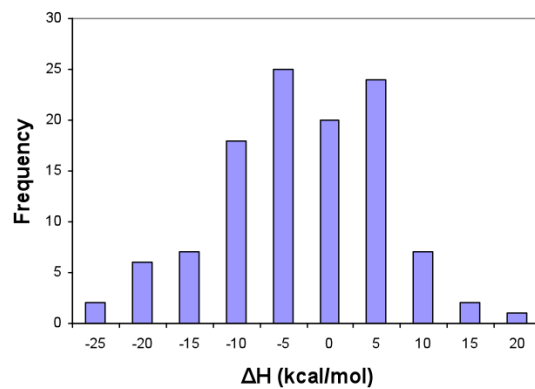


Figure 4.2: Distribution histogram of the change in enthalpy (ΔH)
(mean = -5.40 kcal/mol, std. dev. = 8.96 kcal/mol)

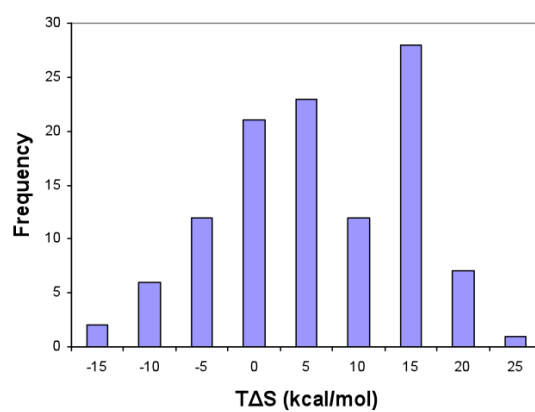


Figure 4.3: Distribution histogram of the change in entropy ($T\Delta S$)
(mean = 3.31 kcal/mol, std. dev. = 8.92 kcal/mol)

Descriptor	ΔH	$T\Delta S$	ΔG
INTERCEPT	-7.064	-1.619	-5.445
Electrostatic Interaction Energy	-0.006	-0.005	-0.001
Steric Interaction Energy	0.008	0.004	0.004
Steric Fit	-0.064	-0.085	0.021
Rotatable Bonds	0.002	0.001	0.001
Ligand Strain Energy	0.023	0.008	0.015
Hydrophobic/Hydrophobic Contact Surface Area 1	0.004	0.007	-0.003
Hydrophilic/Hydrophilic Contact Surface Area 1 (Opposite Charge)	0.009	0.011	-0.002
Hydrophobic/Hydrophilic Contact Surface Area 1	0.009	0.012	-0.003
Hydrophilic/Hydrophilic Contact Surface Area 1 (Same Charge)	0.006	0.011	-0.005
Hydrophobic/Hydrophobic Contact Surface Area 2	0.001	0.001	0
Hydrophilic/Hydrophilic Contact Surface Area 2 (Opposite Charge)	0.006	0.006	0
Hydrophobic/Hydrophilic Contact Surface Area 2	0.003	0.004	-0.001
Hydrophilic/Hydrophilic Contact Surface Area 2 (Same Charge)	-0.002	0	-0.002
Ligand Total Hydrophobic Surface Area	0	0.001	-0.001
Ligand Total Hydrophilic Surface Area	-0.003	-0.004	0.001
Flexibility Index(Rot Bonds/ Non Term Bonds)	3.549	3.424	0.125
Ligand Buried Hydrophobic Surface Area	0	0.001	-0.001
Ligand Buried Hydrophilic Surface Area	-0.006	-0.006	0
Ligand Exposed Hydrophobic Surface Area	-0.001	-0.001	0
Ligand Exposed Hydrophilic Surface Area	-0.003	-0.004	0.001
Receptor Buried Hydrophobic Surface Area	-0.001	0.001	-0.002
Receptor Buried Hydrophilic Surface Area	-0.002	0.002	-0.004
Receptor Exposed Hydrophobic Surface Area	0	0	0
Receptor Exposed Hydrophilic Surface Area	0	0	0
Normalized Ligand Buried Hydrophobic Surface Area	3.487	3.695	-0.208
Normalized Ligand Buried Hydrophilic Surface Area	-2.814	-2.906	0.092
Normalized Ligand Exposed Hydrophobic Surface Area	-4.324	-3.101	-1.223
Normalized Ligand Exposed Hydrophilic Surface Area	-2.993	-5.661	2.668
Total Ligand/Receptor Hydrogen Bonds	0.044	0.032	0.012
Ligand Total Donor/Acceptor Count	-0.059	-0.072	0.013
Ligand Total Hydrogen Bond Atoms	0.007	-0.008	0.015
Ligand Total Buried Donor/Acceptor Count	-0.114	-0.129	0.015
Receptor Total Donor/Acceptor Count	0.053	0.045	0.008
Receptor Total Buried Donor/Acceptor Count	0.079	0.082	-0.003
Partition Coefficient	0.024	0.115	-0.091
Ligand Volume	-0.001	-0.001	0
Pocket Volume	-0.001	-0.001	0
Number of Alpha Spheres	0.004	0.009	-0.005
Proportion of Apolar Alpha Spheres	-4.253	-4.039	-0.214
Mean Local Hydrophobic Density	-0.159	-0.137	-0.022
Polarity Score	0.124	0.114	0.01
Alpha Sphere Density	0.012	-0.019	0.031

Table 4.1: Coefficients and intercepts derived from partial least squares regression for the descriptor set ($n = 42$) used in the final PHOENIX scoring function for change in enthalpy (ΔH) and change in entropy ($T\Delta S$) equations.

Descriptor	ΔH	$T\Delta S$	ΔG
Mean Local Hydrophobic Density	0.08	0.063	0.072
Flexibility Index (Rot Bonds/ Non Term Bonds)	0.066	0.059	0.063
Receptor Total Buried Donor/Acceptor Count	0.063	0.061	0.062
Pocket Volume	0.052	0.051	0.052
Electrostatic Interaction Energy	0.054	0.043	0.049
Hydrophobic/Hydrophilic Contact Surface Area 2	0.043	0.049	0.046
Proportion of Apolar Alpha Spheres	0.049	0.043	0.046
Hydrophobic/Hydrophilic Contact Surface Area 1	0.04	0.049	0.045
Polarity Score	0.048	0.04	0.044
Normalized Ligand Buried Hydrophobic Surface Area	0.035	0.034	0.035
Ligand Total Donor/Acceptor Count	0.031	0.035	0.033
Ligand Buried Hydrophilic Surface Area	0.034	0.03	0.032
Ligand Total Buried Donor/Acceptor Count	0.031	0.032	0.032
Ligand Total Hydrophilic Surface Area	0.029	0.028	0.029
Ligand Strain Energy	0.04	0.014	0.027
Steric Interaction Energy	0.035	0.016	0.026
Normalized Ligand Buried Hydrophilic Surface Area	0.026	0.025	0.026
Hydrophilic/Hydrophilic Contact Surface Area 2 (Opposite Charge)	0.024	0.023	0.024
Hydrophilic/Hydrophilic Contact Surface Area 1 (Opposite Charge)	0.02	0.024	0.022
Hydrophobic/Hydrophobic Contact Surface Area 1	0.015	0.025	0.02
Hydrophilic/Hydrophilic Contact Surface Area 1 (Same Charge)	0.014	0.024	0.019
Ligand Volume	0.021	0.015	0.018
Receptor Exposed Hydrophobic Surface Area	0.013	0.02	0.017
Normalized Ligand Exposed Hydrophilic Surface Area	0.012	0.021	0.017
Normalized Ligand Exposed Hydrophobic Surface Area	0.019	0.013	0.016
Receptor Exposed Hydrophilic Surface Area	0.009	0.02	0.015
Ligand Exposed Hydrophilic Surface Area	0.011	0.015	0.013
Receptor Total Donor/Acceptor Count	0.015	0.011	0.013
Partition Coefficient	0.005	0.021	0.013
Hydrophobic/Hydrophobic Contact Surface Area 2	0.007	0.016	0.012
Total Ligand/Receptor Hydrogen Bonds	0.013	0.009	0.011
Receptor Buried Hydrophilic Surface Area	0.008	0.011	0.01
Ligand Buried Hydrophobic Surface Area	0.004	0.013	0.009
Number of Alpha Spheres	0.005	0.012	0.009
Receptor Buried Hydrophobic Surface Area	0.006	0.009	0.008
Steric Fit	0.006	0.007	0.007
Ligand Exposed Hydrophobic Surface Area	0.008	0.004	0.006
Ligand Total Hydrophobic Surface Area	0.001	0.008	0.005
Hydrophilic/Hydrophilic Contact Surface Area 2 (Same Charge)	0.007	0.001	0.004
Ligand Total Hydrogen Bond Atoms	0.002	0.002	0.002
Alpha Sphere Density	0.001	0.002	0.002
Rotatable Bonds	0.001	0.001	0.001

Table 4.2: Descriptor set ($n = 42$) used in the final PHOENIX scoring function and its relative contribution to change in enthalpy (ΔH), change in entropy ($T\Delta S$), and change in binding free energy (ΔG) calculations. Descriptors are sorted by relative fraction to change in binding free energy (ΔG) in descending order.

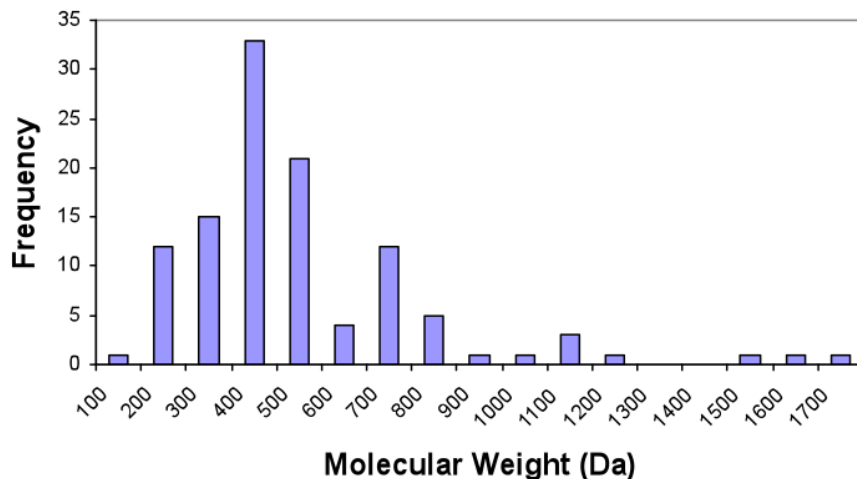


Figure 4.4: Distribution histogram of the molecular weight (Da)
(mean = 455.25 Da, std. dev. = 273.11 Da.)

cases. Missing side chains and neutral terminal groups were added by the Biopolymer Structure Preparation function. Hydrogens were added to both the protein and water using the Biopolymer dictionary. The ligand was extracted from the complex and atom types were assessed and reassigned, if necessary. Hydrogens were added to all atoms. The resulting protein and ligand structures were saved in mol2 format.

4.2.3 External Test Sets

External validation sets include three versions of the PDBbind refined set (2002, 2004, and 2009) (Wang et al., 2004) and the 2007 PDBbind core set downloaded from the PDBbind site. Previous scoring function development studies by Wang, Lu, Fang, and Wang (Wang et al., 2004) and Sotriffer, Sanschagrin, Matter, and Klebe (Sotriffer et al., 2008) used both the 2002 and 2004 versions as benchmark sets, thus were assessed in this study for comparison purposes. The 2002 version contains 800 complexes, the 2004 version contains 1091 complexes, and the 2009 version contains 1741 protein-ligand complexes with resolution ≤ 2.5 Å. The 2007 core set, which consists of 195 complexes with non-redundant protein families and diversity of ligand structures and binding affinities, was also used to assess the general

applicability of PHOENIX. A number of docking and scoring assessments have used this core set as a diverse set benchmark. (Cheng et al., 2009) The protein and ligand structures were downloaded from the PDBbind database. Structures of the proteins were prepared using the same procedure as the training set. The ligands did not require any preparation and were used as is. For the 2004 and 2009 sets, 1071 out of 1091 were used in the 2004 set, while 1612 of 1741 were used in the 2009 set.

4.2.4 Descriptors Set

A set of 42 descriptors were used to derive the PHOENIX scoring function, as listed in Table 1. Of that set, the first 34 of the descriptors listed were calculated using the VALIDATE scoring function. (Head et al., 1996) The calculated partition coefficient, XlogP, was computed based on the Wang, Fu, and Lai study (Wang et al., 1997) using FILTER (OpenEye). FPOCKET (Guillox et al., 2009), a cavity detection program based on Voronoi tessellation and alpha spheres, was used to obtain 7 volume-based descriptors to describe the ligand and protein binding site.

VALIDATE parameters were determined by using both molecular mechanics a heuristics approach in combination with parameters derived from molecular mechanics. Parameters derived from molecular mechanics include electrostatic interaction energy (EIE), steric interaction energy (SIE), and ligand strain energy (LSE). EIE accounts for the electrostatic interactions that contribute to the specificity of protein-ligand interactions, and was calculated using the MacroModel program. Charges for the protein and ligand were derived from the OPLS-AA force field. Nonbonded electrostatic interaction energy was calculated using the explicit sum of the Coulombic potentials. SIE was computed from the explicit sum of the Lennard-Jones potentials, where the required parameters were derived from the OPLS-AA force field. LSE was calculated based on the difference between the energy of the ligand in the binding site and the energy of the ligand by itself.

Descriptors derived from heuristics for both the ligand and protein include steric fit, number of rotatable bonds, total number of ligand/protein hydrogen bonds, total

donor/acceptor count, total hydrogen-bond atoms, and number of buried hydrogen-bond atoms. Steric fit (SF) was used to describe the close packing interactions between the protein and ligand. In order to quantitate surface complementarity between protein and ligand, descriptors were used to capture lipophilic complementarity (nonpolar/nonpolar), hydrophilic complementarity (polar/polar, opposite charge), lipophilic/hydrophilic complementarity (polar/nonpolar), and hydrophilic noncomplementarity (polar/polar, like charge). Two separate methods were used. The first method used an absolute surface area between the protein and ligand similar to the method used by Bohm. The second method was based on a pairwise sum estimate, similar to the approach by Kellogg et al. For a detailed description of the implementation and underlying theory of the 34 VALIDATE descriptors, refer to the original study by Head et al. (Head et al., 1996)

As a heuristic method to capture entropic contributions, volume descriptors were used to represent the amount of water molecules displaced from the protein binding site, as well as the desolvation process of ligand going from unbound to bound state. FPOCKET (Guillox et al., 2009), a cavity detection program based on Voronoi tessellation and alpha spheres, was used to obtain 7 volume-based descriptors to describe the ligand and protein binding site (ligand volume, pocket volume, number of alpha spheres, proportion of apolar alpha spheres, mean local hydrophobic density, polarity score, alpha sphere density).

Feature selection strategies such as excluding descriptors with a correlation coefficient ≥ 0.95 of another descriptor, or excluding descriptors that displayed minimal correlation to the thermodynamics parameters (≤ 0.01 , ≤ 0.05) were assessed to identify a set of descriptors leading to the best performance. In addition, attempts were made to separate ΔH and $T\Delta S$ descriptors by deriving simpler models using subsets ($n = 20-30$) of the final descriptors set ($n = 42$) which contribute qualitatively to each thermodynamic force, to test if more accurate predictions could be achieved. After excluding the descriptors with high correlation and descriptors with low correlation to ΔH and $T\Delta S$ as well as separating descriptors for each thermodynamic force, the models resulted in less accurate predictions when assessing the 2002 version of PDBbind; therefore all 42 descriptors were used in as the PHOENIX scoring function.

4.2.5 Function Parameterization

The weight coefficients for each descriptor and equation for predicting ΔH and $T\Delta S$ were derived by using PLS in SYBYL 7.3. All 42 descriptors were used as input parameters. To derive the regression equations, leave-one-out cross validation was initially performed to identify the optimal number of components to use for the PLS model. The PLS model was subsequently constructed using the number of components with the highest q^2 and least error to calculate the constant and coefficients for each descriptor. Regression statistics such as r^2 , standard error, and F-value were used to assess the predictive ability of the models. The fraction of relative contribution of each descriptor to change in enthalpy (ΔH), change in entropy ($T\Delta S$), and change in binding free energy (ΔG) is listed in Table 2, and the coefficients and intercepts derived from partial least squares regression for the final PHOENIX scoring function ($n = 112$) are listed in Table 4.1.

4.3 Results

4.3.1 Regression Analysis

Regression and leave-one-out cross validation statistics of the different training sets used for PHOENIX along with statistics for change of enthalpy (ΔH) (Table: 4.3), change in entropy ($T\Delta S$) (Table: 4.4), and change in binding free energy (ΔG) (Table: 4.5) are shown. Although the training set of 68 complexes resulted in the best regression statistics and standard errors, equations derived using the training set of 112 complexes were used since its performance on the external test sets were better than ones using the other test sets. Selecting the training set with the best regression or cross-validations statistics to use for external predictions can lead to using a model that may simply be overfitted to the training set. Regression analysis on the different training sets demonstrated that good fits were obtained using partial least squares on the set of 42 descriptors. Experimental versus predicted values of change of enthalpy (ΔH) (Fig: 4.5), change of entropy ($T\Delta S$) (Fig: 4.6), and change of binding free energy (ΔG) (Fig: 4.7) are shown.

No. of Complexes	No. of Components	r^2	s (kcal/mol)	F-value
68	3	0.645	4.03	38.69
82	4	0.566	6.1	25.09
105	2	0.442	6.79	40.44
112	3	0.497	6.44	35.6
127	3	0.503	6.24	41.44
140	3	0.447	6.37	36.64
153	4	0.48	6.22	34.15
162	4	0.466	6.26	34.3

Table 4.3: Partial least squares (PLS) regression statistics of the change in enthalpy (ΔH). Values presented include the number of complexes used in the training set (No. of Complexes), the number of components used to derive the PLS model (No. of Components), the correlation coefficient (r^2), the standard error (s), and F-value (F-value).

No. of Complexes	No. of Components	r^2	s (kcal/mol)	F-value
68	3	0.735	3.81	59.2
82	5	0.722	4.92	39.45
105	2	0.55	6.06	62.26
112	3	0.605	5.69	55.17
127	3	0.606	5.63	63.01
140	2	0.478	6.29	62.7
153	3	0.512	6.2	52.1
162	3	0.534	6.03	60.29

Table 4.4: Partial least squares (PLS) regression statistics of the change in entropy ($T\Delta S$). Values presented include the number of complexes used in the training set (No. of Complexes), the number of components used to derive the PLS model (No. of Components), the correlation coefficient (r^2), the standard error (s), and F-value (F-value).

No. of Complexes	r^2	s (kcal/mol)
68	0.61	1.19
82	-7.00	5.54
105	0.43	1.55
112	0.55	1.34
127	0.44	1.44
140	-0.89	2.62
153	-0.13	2.01
162	0.30	1.56

Table 4.5: Partial least squares (PLS) regression statistics of the change in binding free energy (ΔG). Values presented include the number of complexes used in the training set (No. of Complexes), the correlation coefficient (r^2), and the standard error (s).

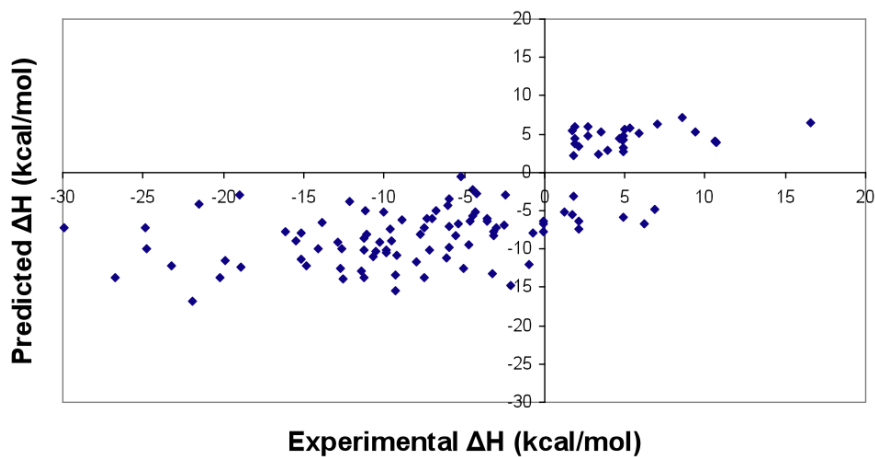


Figure 4.5: Scatter plot from regression analysis of the final PHOENIX training set ($n = 112$). Calculated versus experimental values for change in enthalpy (ΔH). Regression and leave-one-out cross validation statistics are as follows: $r^2 = 0.50$, $s = 6.44$ kcal/mol, $q^2 = 0.37$, $S_{PRESS} = 7.24$ kcal/mol.

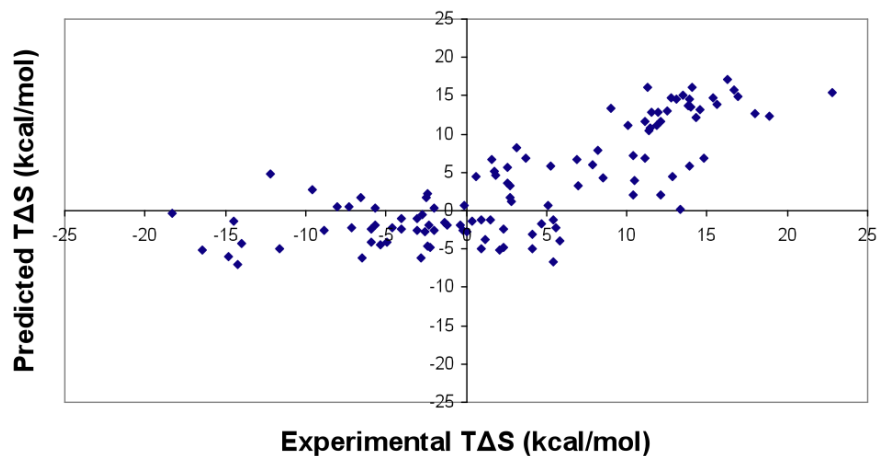


Figure 4.6: Scatter plot from regression analysis of the final PHOENIX training set ($n = 112$). Calculated versus experimental values for change in entropy ($T\Delta S$). Regression and leave-one-out cross validation statistics are as follows: $r^2 = 0.61$, $s = 5.69$ kcal/mol, $q^2 = 0.48$, $S_{PRESS} = 6.50$ kcal/mol.

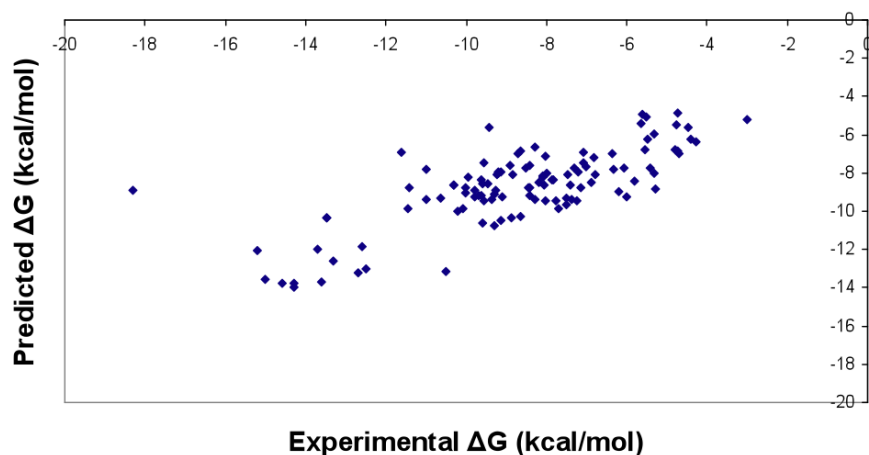


Figure 4.7: Scatter plot from regression analysis of the final PHOENIX training set ($n = 112$). Calculated versus experimental values for change in binding free energy (ΔG). Regression and leave-one-out cross validation statistics are as follows: $r^2 = 0.55$, $s = 1.34$ kcal/mol.

The training set with 112 complexes resulted in models that did not lead to good regression statistics compared with results using the training set of 68 complexes. One possible reason for this is the larger training set contained a wider variety of protein-ligand complexes, especially ones that were difficult to predict, such as streptavidin and biotin complexes. Change in enthalpy and change in entropy values did not vary as much in the set of 68 complexes as the larger training sets, resulting in smaller errors and a better linear fit. However, when validating the model on external test sets such as PDBbind, enthalpy, and entropy, and binding free energy regression equations derived using the set of 112 complexes resulted in better regression statistics, which indicates that diversity in both structural data and thermodynamics data may be necessary to achieve robust predictive ability. When tested using the larger training sets ($n = 127, 140, 153, 162$) which included structures between 2 and 2.5 Å resolution, the performance on the external test sets did not improve. While increasing the size of the training set generally leads to more predictive models, in this case, results from this study suggest that inclusion of lower-resolution structures may actually introduce noise, leading to less predictive binding affinity calculations.

4.3.2 Internal Cross-Validation

Cross-validation studies were performed on the PHOENIX scoring function trained with 112 complexes. The set of 112 complexes was divided into a set of 82 complexes for training, and a set of 30 complexes for testing. PLS was used to derive regression equations, which resulted in the following regression and leave-one-out cross validation statistics: $r^2 = 0.43$, $s = 7.27$ kcal/mol (2 components), $q^2 = 0.34$, $S_{PRESS} = 7.83$ kcal/mol for change of enthalpy (ΔH); $r^2 = 0.56$, $s = 6.37$ kcal/mol (2 components), $q^2 = 0.48$, $S_{PRESS} = 6.89$ kcal/mol for change of entropy ($T\Delta S$). These equations were used to calculate the thermodynamics contributions in the test set. Figures ?? displays the experimental versus predicted values for ΔH , $T\Delta S$, and ΔG , respectively. Predicted statistics for the test set of 30 complexes were as follows: ΔH , $r^2 = 0.25$, $s = 6.32$ kcal/mol; $T\Delta S$, $r^2 = 0.31$, $s = 6.01$ kcal/mol; ΔG , $r^2 = 0.52$, $s = 1.53$ kcal/mol. While the ΔH and $T\Delta S$ calculations resulted in sizable errors, calculating their difference to obtain binding free energy led to a standard error within a reasonable accuracy range.

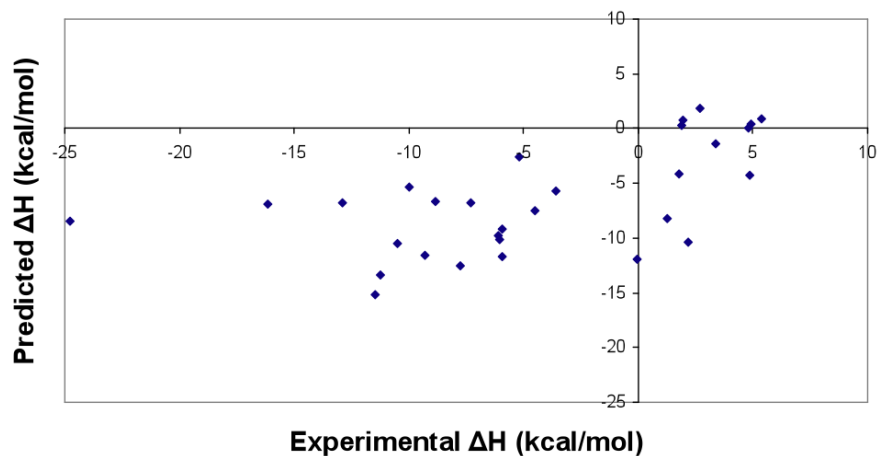


Figure 4.8: Scatter plots from leave-one-out cross validation analyses of the final PHOENIX training set ($n = 112$), separated into a training set of 82 complexes and a test set of 30 complexes. Calculated versus experimental values for change in enthalpy (ΔH). Regression statistics are as follows: $r^2 = 0.25$, $s = 6.32$ kcal/mol.

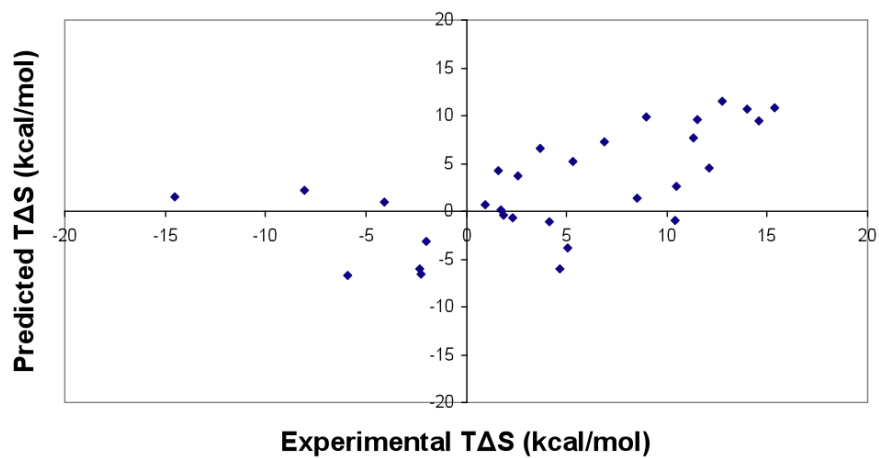


Figure 4.9: Scatter plots from leave-one-out cross validation analyses of the final PHOENIX training set ($n = 112$), separated into a training set of 82 complexes and a test set of 30 complexes. Calculated versus experimental values for change in entropy ($T\Delta S$). Regression statistics are as follows: $r^2 = 0.31$, $s = 6.01$ kcal/mol.

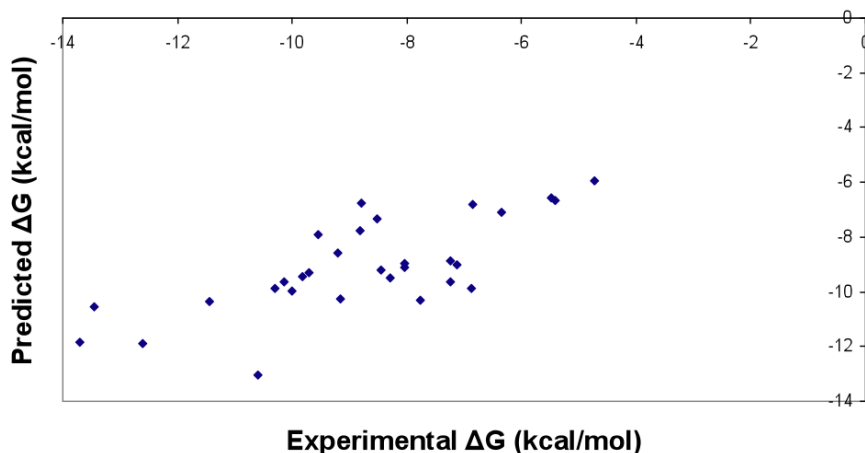


Figure 4.10: Scatter plots from leave-one-out cross validation analyses of the final PHOENIX training set ($n = 112$), separated into a training set of 82 complexes and a test set of 30 complexes. Calculated versus experimental values for change in binding free energy (ΔG). Regression statistics are as follows: $r^2 = 0.52$, $s = 1.53$ kcal/mol.

4.3.3 Testing on External Data Sets

To better assess the performance of PHOENIX on accuracy and applicability of affinity predictions, the scoring function trained with 112 complexes was tested on 4 different versions (2002, 2004, and 2009 refined sets; 2007 core set) of the PDBbind. For the sake of comparison, the assessment was performed in a similar fashion to the scoring function studies of Wang, Lu, Fang, and Wang, (Wang et al., 2004) Sotriffer, Sanschagrın, Matter, and Klebe, (Sotriffer et al., 2008) and Cheng, Li, Li, Liu, and Wang (Cheng et al., 2009). Note the use of a different set of statistical metrics (e.g., Pearson correlation coefficient, Spearman correlation coefficient, etc.) to assess the performance on the external data set for comparison purposes with previous scoring function studies. To assess the performance of PHOENIX in a greater detail, correlation evaluation was performed on protein-ligand complexes categorized based on resolutions, protein families, and binding affinities in the 2002 version. The 2004 and 2009 refined sets were used to assess the performance of PHOENIX on larger and

more diverse data sets. The 2007 core set, consisting of 195 complexes with 65 protein families with 3 ligands of different affinities (low-, medium-, and high-affinity), was used to assess performance on a non-redundant and diverse set of complexes.

Correlation evaluation results for the 2002 version of PDBbind compared to scoring functions in the Wang, Lu, Fang, and Wang study (Wang et al., 2004) and SFCscore (Sotriffer et al., 2008) are summarized in Table 4.7. Based on the correlation evaluation of PDBbind 2002, the performance of PHOENIX is comparable to the top-performing scoring functions (e.g., SFCscore and X-Score::HMScore).

Scoring Function	R_p	R_s	SD	ME	a	b
PHOENIX_68	0.499	0.518	2.04	1.62	0.33	4.09
PHOENIX_82	0.41	0.449	6.27	5.94	0.38	8.34
PHOENIX_105	0.473	0.502	2.06	1.6	0.43	3.25
PHOENIX_112	0.524	0.559	1.98	1.56	0.37	4.27
PHOENIX_127	0.517	0.534	2.07	1.65	0.33	4.77
PHOENIX_140	0.424	0.445	2.91	2.41	0.41	2.32
PHOENIX_153	0.333	0.339	2.24	1.81	0.21	4.54
PHOENIX_162	0.492	0.513	2.17	1.72	0.34	3.62

Table 4.6: Correlation evaluation of the PHOENIX scoring function using different training sets on the PDBbind v2002 ($n = 796$) database. Correlation statistics include Pearson correlation coefficient (R_p), Spearman correlation coefficient (R_s), standard deviation (SD), mean error (ME), slope (a) in the linear regression ($y = ax + b$), and intercept (b). The number after the scoring function indicates the total number of complexes used for training (e.g., PHOENIX_68, training set of 68 complexes).

Resolution

To assess the performance of PHOENIX on affinity predictions for low- and high-resolution complexes, the 2002 version of PDBbind was categorized into 2 sets: a high-resolution ($\leq 2 \text{ \AA}$) set of 494 complexes, and a low-resolution ($2 \leq 2.5 \text{ \AA}$) set of 302 complexes. Correlation evaluation results are listed in Table 4.8. PHOENIX affinity predictions on the high-resolution set were comparable to ones obtained from the X-Score functions (HPScore, HMScore, HSScore). PHOENIX affinity predictions on the low-resolution set were inferior to ones obtained from the 3 X-Score functions.

Scoring Function	N	R_p	SD	ME	a	b
PHOENIX	796	0.524	1.98	1.56	0.37	4.27
SFCscore::met	800	0.585	1.8	1.37	0.82	1.23
X-Score::HPScore	800	0.514	1.89	1.47	0.71	2.03
X-Score::HMScore	800	0.566	1.82	1.42	0.92	1.18
X-Score::HSScore	800	0.506	1.9	1.48	0.93	1.24
DrugScore::Pair	800	0.473	1.94	1.51	4.90E-06	4.1
DrugScore::Surf	800	0.463	1.95	1.53	7.20E-05	4.48
DrugScore::Pair/Surf	800	0.476	1.94	1.5	4.70E-06	4.09
Sybyl::D-Score	800	0.322	2.09	1.67	9.70E-03	5
Sybyl::PMF-Score	785	0.147	2.16	1.74	6.43E-03	5.92
Sybyl::G-Score	800	0.443	1.98	1.56	9.13E-03	4.34
Sybyl::ChemScore	797	0.499	1.91	1.5	9.10E-02	3.9
Sybyl::F-Score	732	0.141	2.19	1.77	2.10E-02	6.06
Cerius2::LigScore	717	0.406	2	1.57	0.79	4.63
Cerius2::PLP1	800	0.458	1.96	1.52	2.30E-02	4.09
Cerius2::PLP2	800	0.455	1.96	1.53	2.60E-02	3.93
Cerius2::PMF	795	0.253	2.13	1.71	1.10E-02	5.37
Cerius2::LUDI1	790	0.334	2.08	1.66	2.60E-03	4.88
Cerius2::LUDI2	799	0.379	2.04	1.62	4.20E-03	4.28
Cerius2::LUDI3	800	0.331	2.08	1.67	3.20E-03	4.68
GOLD::GoldScore	694	0.285	2.16	1.72	2.40E-02	5.33
GOLD::GoldScore_opt	772	0.365	2.06	1.63	3.00E-02	4.7
GOLD::ChemScore	741	0.423	2	1.56	8.50E-02	4.65
GOLD::ChemScore_opt	762	0.449	1.96	1.52	8.60E-02	4.41
HINT	800	0.33	2.08	1.65	0.2	6.36

Table 4.7: Correlation evaluation of the PHOENIX scoring function compared to other commonly used scoring functions on the PDBbind v2002 set. Correlation statistics presented are the number of complexes tested (N), Pearson correlation coefficient (R_p), standard deviation (SD), mean error (ME), slope (a) in the linear regression ($y = ax + b$), and intercept (b). Results from the commonly used scoring functions taken from Wang et al. (2004) are presented for comparison purposes.

PHOENIX, as well as the X-Score functions, provided better correlation statistics for the high-resolution set than the low-resolution set. One point to note is that the high-resolution set has 192 more complexes compared to the low-resolution set, yet still achieved better correlation statistics. These results may suggest that scoring functions in general can achieve more accurate predictions using higher-resolution and perhaps higher-quality X-ray crystal structures compared to using low-resolution and low-quality structures.

Scoring Function	N	R_p	SD	ME	R_s	N	R_p	SD	ME	R_s
PHOENIX	494	0.558	1.92	1.52	0.586	302	0.468	2.06	1.6	0.511
X-Score::HPScore	494	0.597	1.95	1.55	0.615	302	0.492	2.13	1.68	0.525
X-Score::HMScore	494	0.575	2.03	1.62	0.589	302	0.48	2.19	1.73	0.525
X-Score::HSScore	494	0.614	1.89	1.49	0.64	302	0.493	2.07	1.62	0.536

Table 4.8: Correlation evaluation of the PHOENIX scoring function compared to X-Score scoring functions on high- ($0 \leq 2 \text{ \AA}$) and low-resolution ($2 \leq 2.5 \text{ \AA}$) complexes of the PDBbind 2002 set. Correlation statistics presented are the number of complexes tested (N), Pearson correlation coefficient (R_p), standard deviation (SD), mean error (ME), Spearman correlation coefficient (R_s).

Protein Families

Three protein families were selected from the 2002 version of PDBbind set to test the performance of PHOENIX on these special cases: HIV-1 protease, trypsin, carbonic anhydrase II. Table 4.9 4.10 4.11 lists the correlation evaluation statistics. The correlation statistics from PHOENIX on the HIV-1 protease set ($R_p = 0.563$, $SD = 1.65$, $ME = 1.35$, $R_s = 0.434$) is better than most of the scoring functions in the Wang, Lu, Fang, and Wang study (Wang et al., 2004) in terms of R_p , and comparable to the top-performing scoring functions (Cerius2::LigScore, $R_p = 0.528$; GOLD::GoldScore_opt, $R_p = 0.555$). This may be due to the inclusion of explicit waters in PHOENIX; water molecules play a critical role in the binding of HIV-1 protease inhibitors. For the trypsin complexes, the correlation statistics from PHOENIX were inferior compared to the other scoring functions tested. Perhaps, the descriptors used in PHOENIX cannot adequately capture the electrostatics involved in the binding of trypsin inhibitors due to the use of monopole electrostatics, which led to larger errors in the affinity calculations. Another potential reason for the poorer performance is that only 2 trypsin

complexes were included in the PHOENIX training set, while other scoring functions included a larger set of trypsin complexes in their training sets. As the availability of crystal structure of complexes with ITC data increases, more trypsin complexes can be included in the training set to improve affinity calculations. For the set of carbonic anhydrase II complexes, the correlation statistics from PHOENIX were comparable to the other scoring functions. SFCscore performed the best, which may primarily be due to the descriptors used to capture interactions with metal atoms present in the binding pocket; metals are involved in critical interactions with the ligand for this class of metalloenzymes. Inferior performance in affinity predictions for the carbonic anhydrase set may be due to the fact that PHOENIX does not contain any descriptors to capture ligand interactions with metal atoms. Again, the use of more sophisticated representation of electrostatic interactions should improve predictability.

Affinities

The 2002 version of PDBbind was categorized into 3 groups: low-affinity ($\text{pK}_d < 5$), medium-affinity ($5 \leq \text{pK}_d \leq 8$), and high-affinity ($\text{pK}_d < 8$). PHOENIX was assessed on its ability to calculate a binding affinity that results in the same group as the experimental binding affinity. Results from this study are listed in Table 4.12. PHOENIX correctly categorized 27% of the low-affinity complexes, 100% of the medium-affinity complexes, and 61% of the high-affinity complexes. PHOENIX performed the best on the medium- and high-affinity complexes compared to the scoring functions from previous studies. The performance on the low-affinity group was the second best (best was SFCscore). This assessment demonstrated that PHOENIX can estimate affinities within a reasonable accuracy range to readily distinguish between a tight-binding ligand from a low-affinity ligand. As minimizing false-positive rates is a significant challenge in computer-aided molecular design, PHOENIX may prove to be advantageous for affinity estimations and relative rankings as well as binding pose prediction, especially when applied to high-resolution structures with high-quality experimental data.

Scoring Function	N	R_p	SD	ME	R_s
PHOENIX	82	0.563	1.65	1.35	0.434
SFSscore::met	74	0.361	na	na	0.312
X-Score::HPScore	82	0.429	1.25	1.01	0.436
X-Score::HMScore	82	0.379	1.28	1.04	0.334
X-Score::HSScore	82	0.4	1.27	1.05	0.322
DrugScore::Pair	82	0.377	1.28	1.04	0.315
DrugScore::Surf	82	0.401	1.27	1.02	0.317
DrugScore::Pair/Surf	82	0.384	1.28	1.04	0.322
Sybyl::D-Score	82	0.342	1.3	1.03	0.305
Sybyl::PMF-Score	82	0.246	1.34	1.09	0.226
Sybyl::G-Score	82	0.35	1.3	1.05	0.335
Sybyl::ChemScore	82	0.376	1.28	1.05	0.35
Sybyl::F-Score	80	0.361	1.31	1.08	0.375
Cerius2::LigScore	81	0.528	1.18	0.99	0.496
Cerius2::PLP1	82	0.458	1.23	1.02	0.395
Cerius2::PLP2	82	0.438	1.25	1.03	0.414
Cerius2::PMF	82	0.411	1.26	1.03	0.342
Cerius2::LUDI1	82	0.208	1.35	1.11	0.123
Cerius2::LUDI2	82	0.274	1.33	1.11	0.181
Cerius2::LUDI3	82	0.248	1.34	1.1	0.174
GOLD::GoldScore	69	0.386	1.25	1	0.391
GOLD::GoldScore_opt	78	0.555	1.13	0.92	0.579
GOLD::ChemScore	78	0.404	1.19	0.98	0.386
GOLD::ChemScore_opt	80	0.429	1.24	1.02	0.393
HINT	82	0.313	1.32	1.04	0.264

Table 4.9: Correlation evaluation of the PHOENIX scoring function compared to other commonly used scoring functions on HIV-1 Protease complexes of the PDBbind 2002 set. Correlation statistics presented are the number of complexes tested (N), Pearson correlation coefficient (R_p), standard deviation (SD), mean error (ME), Spearman correlation coefficient (R_s).

Scoring Function	N	R_p	SD	ME	R_s
PHOENIX	40	0.476	1.9	1.38	0.574
SFSScore::met	40	0.853	na	na	0.848
X-Score::HPScore	45	0.754	1.15	0.88	0.725
X-Score::HMScore	45	0.823	0.99	0.75	0.824
X-Score::HSScore	45	0.753	1.15	0.91	0.766
DrugScore::Pair	45	0.78	1.09	0.82	0.818
DrugScore::Surf	45	0.674	1.29	0.99	0.753
DrugScore::Pair/Surf	45	0.78	1.09	0.82	0.807
Sybyl::D-Score	45	0.617	1.37	0.98	0.736
Sybyl::PMF-Score	37	0.513	1.02	0.86	0.523
Sybyl::G-Score	45	0.58	1.42	1.06	0.728
Sybyl::ChemScore	45	0.761	1.13	0.91	0.749
Sybyl::F-Score	45	0.663	1.31	1.05	0.61
Cerius2::LigScore	40	0.392	1.59	1.27	0.467
Cerius2::PLP1	45	0.729	1.19	0.88	0.785
Cerius2::PLP2	45	0.754	1.15	0.84	0.802
Cerius2::PMF	43	0.775	1.06	0.85	0.74
Cerius2::LUDI1	45	0.67	1.29	1.01	0.698
Cerius2::LUDI2	45	0.696	1.25	0.95	0.725
Cerius2::LUDI3	45	0.679	1.28	1	0.69
GOLD::GoldScore	36	0.029	1.65	1.32	-0.012
GOLD::GoldScore_opt	42	0.59	1.41	1.14	0.673
GOLD::ChemScore	44	0.388	1.61	1.33	0.348
GOLD::ChemScore_opt	44	0.52	1.49	1.21	0.565
HINT	45	0.135	1.73	1.37	0.251

Table 4.10: Correlation evaluation of the PHOENIX scoring function compared to other commonly used scoring functions on trypsin complexes of the PDBbind 2002 set. Correlation statistics presented are the number of complexes tested (N), Pearson correlation coefficient (R_p), standard deviation (SD), mean error (ME), Spearman correlation coefficient (R_s).

Scoring Function	N	R_p	SD	ME	R_s
PHOENIX	39	0.539	3.26	2.97	0.444
SFSScore::met	37	0.717	na	na	0.485
X-Score::HPScore	39	0.544	1.18	0.85	0.547
X-Score::HMScore	39	0.495	1.23	0.95	0.341
X-Score::HSScore	39	0.417	1.28	0.91	0.448
DrugScore::Pair	39	0.622	1.1	0.83	0.501
DrugScore::Surf	39	0.512	1.21	0.97	0.269
DrugScore::Pair/Surf	39	0.623	1.1	0.83	0.495
Sybyl::D-Score	39	0.584	1.14	0.86	0.441
Sybyl::PMF-Score	39	0.655	1.07	0.8	0.652
Sybyl::G-Score	39	0.643	1.08	0.79	0.649
Sybyl::ChemScore	39	0.609	1.12	0.76	0.663
Sybyl::F-Score	35	0.371	1.15	0.87	0.145
Cerius2::LigScore	18	0.154	1.78	1.34	-0.323
Cerius2::PLP1	39	0.718	0.98	0.76	0.606
Cerius2::PLP2	39	0.735	0.96	0.67	0.781
Cerius2::PMF	39	0.604	1.12	0.87	0.603
Cerius2::LUDI1	38	0.065	1.21	0.86	0.335
Cerius2::LUDI2	39	0.47	1.25	0.89	0.519
Cerius2::LUDI3	39	0.433	1.27	0.91	0.554
GOLD::GoldScore	34	0.539	1.25	0.9	0.42
GOLD::GoldScore_opt	37	0.585	1.17	0.86	0.532
GOLD::ChemScore	39	0.498	1.22	0.89	0.307
GOLD::ChemScore_opt	39	0.639	1.08	0.8	0.454
HINT	39	0.599	1.13	0.78	0.689

Table 4.11: Correlation evaluation of the PHOENIX scoring function compared to other commonly used scoring functions on carbonic anhydrase II complexes of the PDBbind 2002 set. Correlation statistics presented are the number of complexes tested (N), Pearson correlation coefficient (R_p), standard deviation (SD), mean error (ME), Spearman correlation coefficient (R_s).

Scoring Function	Low	Medium	High
PHOENIX	52/205=27%	417/417=100%	112/193=61%
SFSscore::met	88/191=46%	309/417=74%	86/192=45%
X-Score::HPScore	33/205=16%	358/402=89%	48/193=25%
X-Score::HMScore	41/205=20%	348/402=87%	65/193=34%
X-Score::HSScore	29/205=14%	350/402=87%	53/193=27%
DrugScore::Pair	24/205=12%	359/402=89%	45/193=23%
DrugScore::Surf	11/205=5%	362/402=90%	45/193=23%
DrugScore::Pair/Surf	24/205=12%	358/402=89%	47/193=24%
Sybyl::D-Score	0/205=0%	384/402=96%	2/193=1%
Sybyl::PMF-Score	0/196=0%	395/396=99%	0/193=0%
Sybyl::G-Score	12/205=6%	359/402=89%	30/193=16%
Sybyl::ChemScore	38/204=19%	349/400=87%	40/193=21%
Sybyl::F-Score	0/182=0%	362/362=100%	0/188=0%
Cerius2::LigScore	11/186=6%	340/366=93%	16/165=10%
Cerius2::PLP1	24/205=12%	364/401=91%	35/193=18%
Cerius2::PLP2	30/205=15%	363/402=90%	32/193=17%
Cerius2::PMF	0/202=0%	390/400=97%	3/193=2%
Cerius2::LUDI1	1/203=0%	379/394=96%	9/193=5%
Cerius2::LUDI2	6/205=3%	378/401=94%	15/193=8%
Cerius2::LUDI3	1/205=0%	387/402=96%	9/193=5%
GOLD::GoldScore	0/178=0%	331/339=98%	4/177=2%
GOLD::GoldScore_opt	3/200=1%	366/385=95%	11/187=6%
GOLD::ChemScore	8/177=5%	345/376=92%	37/188=20%
GOLD::ChemScore_opt	20/187=11%	346/386=90%	38/189=20%
HINT	2/205=1%	388/402=97%	11/193=6%

Table 4.12: Assessment of the ability of the PHOENIX scoring function to classify complexes into three binding affinity groups: low-affinity ($\text{pK}_d < 5.0$), medium-affinity ($5.0 \leq \text{pK}_d \leq 8.0$), and high-affinity ($\text{pK}_d > 8.0$). The number of correctly categorized complexes and total number of complexes in each category, as well as the percentage of the correctly categorized complexes are presented for PHOENIX and the commonly used scoring functions take from the Wang et al. (2004) for comparison purposes.

Recent Versions of PDBbind

Recent versions of the PDBbind dataset (2004 and 2009) were used as external test sets for scoring function validation. The correlation statistics for the 2004 version ($n = 1073$), also used as a test set in the development of SFCscore, are listed in Table 4.13, and the ones for the 2009 version ($n = 1612$) are listed in Table 4.14. Based on the results from the 2004 and 2009 refined sets, PHOENIX demonstrated comparable performance compared with the X-Score functions and SFCscore (the better performing scoring functions). Also, results from the larger and more diverse 2004 and 2009 PDBbind refined sets demonstrated the robustness of PHOENIX in predicting affinities for various types of protein-ligand interactions. See Figure 4.11 for a scatter plot of the calculated versus experimental affinities for the 2009 refined set ($n = 1612$).

Scoring Function	R_p	R_s	SD	ME	a	b
PHOENIX	0.515	0.554	2	1.57	0.35	4.46
X-Score::HPScore	0.557	0.589	2	1.57	0.32	4.24
X-Score::HMScore	0.54	0.572	2.06	1.63	0.29	4.23
X-Score::HSScore	0.561	0.593	1.95	1.53	0.36	4.27

Table 4.13: Correlation evaluation of the PHOENIX scoring function on the PDBbind 2004 ($n = 1073$) refined set. Correlation statistics presented are the number of complexes tested (N), Pearson correlation coefficient (R_p), standard deviation (SD), mean error (ME), slope (a) in the linear regression ($y = ax + b$), and intercept (b). Results from the best-performing scoring functions taken from Wang et al. (2004) are presented for comparison purposes.

Diverse and Non-Redundant Test Set

To further assess the performance of PHOENIX compared with other scoring functions, the PDBbind 2007 core set was used to represent a diverse, yet non-redundant, set of protein-ligand complexes. The 2007 core set includes 65 unique protein family members, each with a low-, medium-, and high-affinity ligand. Binding affinities ranged from 1.40 to 13.96 pK_d , molecular weight from 103 to 974, and number of ligand rotatable bonds from 0 to 32. Performance in the scoring power test similar to the one in Cheng et al. (2009) was used to assess PHOENIX. The statistics from

Scoring Function	R_p	R_s	SD	ME	a	b
PHOENIX	0.575	0.591	1.76	1.41	0.44	3.98
X-Score::HPScore	0.571	0.589	1.78	1.43	0.36	4.05
X-Score::HMScore	0.563	0.581	1.84	1.48	0.34	4.03
X-Score::HSScore	0.565	0.584	1.75	1.42	0.40	4.10

Table 4.14: Correlation evaluation of the PHOENIX scoring function on the PDBbind 2009 ($n = 1612$) refined set. Correlation statistics presented are the number of complexes tested (N), Pearson correlation coefficient (R_p), standard deviation (SD), mean error (ME), slope (a) in the linear regression ($y = ax + b$), and intercept (b). Results from the best-performing scoring functions taken from Wang et al. (2004) are presented for comparison purposes.

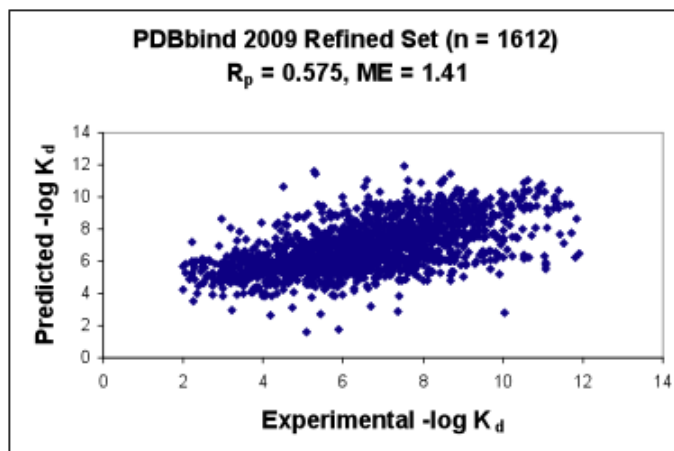


Figure 4.11: Scatter plot of calculated versus experimental binding affinities ($-\log K_d$) of the PDBbind 2009 refined set ($n = 1612$). Regression statistics are as follows:
 $R_p = 0.575$, ME = 1.41.

correlation evaluation on affinity predictions are listed in Table 4.15. To test whether scoring functions provided value over the use of a simple descriptor, the number of heavy atoms was assessed as a scoring method. PHOENIX resulted in the second highest Pearson correlation coefficient, however, the mean error was more than twice as large as the second largest (1.70 compared to 0.71), suggesting that there is still significant room for improvement in the accuracy of affinity predictions. To assess the ranking power of PHOENIX as performed in Cheng et al. (2009), each of the 65 families were assessed to check if the low-, medium-, and high-affinity ligand were ranked in the correct order. Families that were ranked correctly for all 3 complexes were given a score of 1, while a score of 0 is given if there is any deviation from the correct ranking (e.g, low, high, medium; medium, high, low). The success rate in the ranking power study of the 2007 core set is listed in Table 4.16. PHOENIX achieved a success rate of 46.2%, which ranks amongst the best-performing functions, with only 4 other scoring functions with a higher success rate (X-Score::HScore, 58.5%; DS::PLP2, 53.8%; DrugScoreCSD, 52.3%; SYBYL::ChemScore, 47.7%). The performance of PHOENIX in this study demonstrated its utility in structure-based design to correctly rank relative affinities for various types of protein-ligand complexes.

4.3.4 PHOENIX Scoring Function

The final PHOENIX scoring functions (used to predict ΔH , $T\Delta S$, and ΔG ($\Delta H - T\Delta S$)) that resulted in the best performance across multiple versions (2002, 2004, 2009 refined sets; 2007 core set) and subsets (resolutions, protein families, affinities from v2002) of the PDBbind database used a training set of 112 structurally and energetically diverse complexes. A set of 42 descriptors were included in the ΔH , $T\Delta S$, and ΔG ($\Delta H - T\Delta S$) models: 34 derived from molecular mechanics calculations, various surface area terms, hydrogen-bond donors and acceptor count from VALIDATE; 1 to estimate the ligand partition coefficient (XlogP); 7 shape- and volume-based descriptors from FPOCKET to better capture entropic contributions. Partial least squares was used to assign coefficients to each of the terms to derive the master equation to calculate ΔH , $T\Delta S$, and ΔG . While change in enthalpy (ΔH) and change in entropy ($T\Delta S$) predictions were of limited accuracy (standard errors of 6 kcal/mol) individually, the difference between their individual predictions resulted

Scoring Function	N	R_p	SD	ME
PHOENIX	194	0.616	2.16	0.644
X-Score::HMScore	195	0.644	1.83	0.705
DrugScoreCSD	195	0.569	1.96	0.627
SYBYL::ChemScore	195	0.555	1.98	0.585
DS::PLP1	195	0.545	2	0.588
GOLD::ASP	193	0.534	2.02	0.577
SYBYL::G-Score	195	0.492	2.08	0.536
DS::LUDI3	195	0.487	2.09	0.478
DS::LigScore2	193	0.464	2.12	0.507
GlideScore-XP	178	0.457	2.14	0.435
DS::PMF	193	0.445	2.14	0.448
GOLD::ChemScore	178	0.441	2.15	0.452
Number of Heavy Atoms	195	0.431	2.15	0.517
SYBYL::D-Score	195	0.392	2.19	0.447
DS::Jain	189	0.316	2.24	0.346
GOLD::GoldScore	169	0.295	2.29	0.322
SYBYL::PMF-Score	190	0.268	2.29	0.273
SYBYL::F-Score	185	0.216	2.35	0.243

Table 4.15: Correlation evaluation of the PHOENIX scoring function on the PDBbind 2007 core set. Correlation statistics presented are the number of complexes tested (N), Pearson correlation coefficient (R_p), standard deviation (SD), and mean error (ME). The Number of Heavy Atoms was used as a benchmark to assess scoring function enrichment. Results from the commonly used scoring functions taken from Wang et al. (2004) are presented for comparison purposes.

Scoring Function	Success Rate (%)
PHOENIX	46.2
X-Score::HSScore	58.5
DS::PLP2	53.8
DrugScoreCSD	52.3
SYBYL::ChemScore	47.7
SYBYL::D-Score	46.2
SYBYL::G-Score	46.2
GOLD::ASP	43.1
DS::LUDI3	43.1
DS::Jain	41.5
DS::PMF	41.5
SYBYL::PMF-Score	38.5
GOLD::ChemScore	36.9
DS::LigScore2	35.4
GlideScore-XP	33.8
Number of Heavy Atoms	32.3
SYBYL::F-Score	29.2
GOLD::GoldScore	23.1

Table 4.16: Success rates for correctly ranking the low-, medium-, and high-affinity ligands in the PDBbind 2007 core set for the PHOENIX scoring function and 16 other commonly used scoring functions taken from Cheng et al. (2009).

in a relatively accurate change in binding free energy (ΔG) (standard errors of 1.5 kcal/mol). External validation using the 2009 version of the PDBbind refined set ($n = 1612$) (most comprehensive high-quality data set for assessing scoring functions) resulted in a Pearson correlation coefficient (R_p) of 0.575 and a mean error (ME) of 1.41 pK_d, which demonstrated its relative accuracy and robustness in predicting binding affinities.

4.4 Discussion

Predicting binding affinity of protein-ligand interactions remains one of the most critical and challenging problems in computer-aided drug design. The PHOENIX scoring function, derived using a training set of high-resolution structures ($n = 112$) and calorimetry measurements for change of enthalpy (ΔH) and change of entropy ($T\Delta S$) from ITC, has demonstrated an ability to achieve accurate binding affinity predictions across 4 large and diverse sets of protein-ligand complexes (PDBbind 2002, 2004, 2009 refined sets; 2007 core set) using a modest number of descriptors ($n = 42$) to capture key physicochemical interactions. Nine descriptors contributing the most ($>4\%$) to binding free energy (mean local hydrophobic density, flexibility index, receptor total buried donor/acceptor count, pocket volume, electrostatic interaction energy, hydrophobic/hydrophilic contact surface area 2, proportion of apolar alpha spheres, hydrophobic hydrophilic contact surface area 1, polarity score) aimed to capture the key physical forces underlying protein-ligand interactions: enthalpic contributions via van der Waals interactions, hydrogen bonding at the binding site, electrostatics for specificity; entropic contributions via volume and polarity features of binding site and ligand conformational entropy. Overall, the relative contributions from each of the descriptors were fairly distributed (ranging from 0.001 to 0.072), which suggested that each descriptor contains some degree of information for capturing the physics of protein-ligand interactions. Perhaps the use of a larger and more physically-accurate set of descriptors in future studies may help in further capturing the atomic-level details underlying molecular recognition in protein-ligand interactions.

Despite the promising performance in predicting binding affinities, some limitations of PHOENIX have been revealed. The enthalpy (ΔH) and entropy ($T\Delta S$) regression and internal cross-validation results suggest that there is significant room for improvement in deriving these equations. The individual thermodynamics parameters (ΔH and $T\Delta S$) displayed only modest predictive ability with relatively large errors (6-7 kcal/mol). There are several possible reasons for this. Scoring function would benefit from training on a larger and more structurally and thermodynamically diverse set of complexes. More physically-accurate descriptors are needed to more accurately capture and separate enthalpic and entropic contributions due to entropy/enthalpy compensation. Descriptors that can better separate enthalpic (ΔH) and entropic ($T\Delta S$) contributions are needed to derive more accurate independent thermodynamic models. However, developing descriptors to capture primarily enthalpy or entropy is a challenging feat in itself, since any physicochemical interactions that can be experimentally quantitated are correlated and will contain, to some degree, both thermodynamic forces (e.g, flexibility index and total ligand surface area). Inclusion of descriptors to explicitly capture hydrogen bonding interactions may lead to more accurate ΔH predictions. Descriptors to better capture electrostatics interactions such as pi-cation interactions may also help with predicting enthalpy changes. To better quantitate entropic contributions, descriptors to take into account conformational changes of the binding site, such as quantitating the rotamers of the side chains involved in the complex, may provide a measure of entropy changes from the protein upon ligand binding (entropy-entropy compensation). (Trbovic et al., 2009) Classifying water molecules in the binding site according to their energetic preferences as a means to model dewetting will be useful for capturing the entropy change upon ligand binding and displacement of binding-site water molecules. (Abel et al., 2008; Homans, 2007; Young et al., 2007) Inclusion of multiple binding modes to better represent conformational and configurational entropy may help to derive more accurate change in entropy ($T\Delta S$) models as been demonstrated in theoretical studies. (Ruvinsky and Kozintsev, 2005; Ruvinsky, 2007; Stjernschantz and Oostenbrink, 2010; Lee and Seok, 2008; Salaniwal et al., 2007) Moreover, larger sets of high-quality and structurally and thermodynamically diverse protein-ligand complexes will certainly be necessary to achieve more representative statistics for the different protein families and ligand structures.

As presented earlier, the change in enthalpy (ΔH) and change in entropy ($T\Delta S$) calculations were of limited relative accuracy. However, the change in binding free energy (ΔG) calculations was within relative accuracy compared with other commonly used scoring functions. The relative accuracy predicted by the ΔG model (difference of ΔH and $T\Delta S$ model predictions) may have resulted from the cancellation of the overestimated values from independent ΔH and $T\Delta S$ calculations, since the regression coefficient signs are in the same direction for both forces. Overestimates of ΔH and $T\Delta S$ may have been due to the high correlation between the physicochemical descriptors used (e.g., flexibility index to capture conformational entropy is correlated to terms estimating total ligand surface area to capture van der Waals interactions to enthalpy), which were originally intended to be used to estimate ΔG . In other words, a descriptor used to estimate $T\Delta S$ contributions (e.g., flexibility index) may also capture, to some degree, the physical forces underlying ΔH contributions (e.g., ligand total surface area). As an attempt to separate ΔH and $T\Delta S$ descriptors, simpler models using subsets ($n = 20-30$) of the final descriptors set ($n = 42$) that are intuitive to contribute qualitatively to each thermodynamic force were used to test if more accurate predictions can be achieved. However, resulting predictions by these feature selection models were not as accurate as the predictions calculated using models with the full descriptors set. As mentioned before, descriptors that can better distinguish between ΔH and $T\Delta S$ contributions should be developed and included in future development of accurate thermodynamically-based scoring functions.

In developing scoring functions, the inherent inaccuracy of the experimental data, which has been highlighted by a number of scoring function and structure-based design studies, remains the culprit to the limited accuracy in binding affinity predictions. In X-ray crystallography, conditions used to induce crystalization are often in dramatic contrast to physiological conditions under which protein-ligand interactions occur. Another potential source of error is from the thermodynamics measurements by ITC. Experiments conducted with ITC have often been performed under varying temperature and buffer conditions (e.g., salt concentration, pH), that may lead to marked variations in the thermodynamics measurements. Inclusion of such ITC data in the training sets may not necessarily represent the magnitude of thermodynamics forces under physiological conditions. As the use of ITC increases to measure

thermodynamics forces in protein-ligand interactions, diverse structural and thermodynamics data performed under homogenous conditions should become available to help alleviate these limitations.

4.5 Conclusion

Towards development of an empirical scoring function to achieve more accurate binding affinity predictions, high-resolutions X-ray crystal structures of protein-ligand complexes and thermodynamic parameters measured by ITC were used to derive models to calculate enthalpic and entropic contributions to binding free energies. Shape- and volume-based descriptors were used as a heuristic method to implicitly capture changes in desolvation entropy and ligand configurational entropy. PHOENIX demonstrated accurate binding affinity predictions comparable to the top-performing scoring functions based on an extensive series of tests on the 4 versions of the PDBbind database. To our knowledge, this is the first empirical scoring function developed using thermodynamics parameters from ITC as a strategy to derive regression equations to calculate binding affinity. Predicting binding affinities is the most critical and also challenging component of structure-based drug design. Often times, a docking program may identify a compound in the native low-energy conformation, but without an accurate scoring function, will be categorized as a non-binder, rendering the docking program of minimal value. Because of the high false-positive and false-negative rates associated with computer-aided drug design methodologies, development of an accurate and reliable scoring function is absolutely necessary for enhancing the performance of these in silico design tools. Development of the PHOENIX scoring function demonstrated the use of high-resolution structural complexes and thermodynamics parameters for model training can be the key advances towards achieving more accurate binding affinity predictions.

4.6 Acknowledgements

The authors thank Mr. Malcolm Tobias and Dr. Chris Ho of the Center for Computational Biology for significant technical assistance. Gratitude is expressed to Dr. Richard Head and Mr. Daniel Lee for technical assistance, Dr. Chris Waller for access to FRED, Prof. Glen Kellogg and Dr. Ashutosh Tripathi for access and support to VICE, Prof. Anatoly Ruvinsky for access to his entropy quantitation program, Dr. Vincent Le Guilloux for support of FPOCKET, Prof. Rino Ragno for scientific discussions, and Dr. Christina Taylor for critical reading of this manuscript. Y.T.T. was supported by a PhRMA Foundation Predoctoral Fellowship in Informatics. In addition, G.R.M. acknowledges financial support from NIH GM068460 and the Department of Defense (HDTRA1-09-CHEM-BIO-BAA).

Chapter 5

Molecular Recognition and Structure-Based Drug Design

5.1 Summary of Results and Discussion

The goal of this dissertation was to examine molecular recognition in protein-ligand interactions. This was achieved in a prospective structure-based drug design study to identify protein-protein interaction inhibitors targeting a bacterial signal transduction pathway important for virulence, and also in the development of a scoring function to predict binding affinities derived using high-resolution crystal structures and thermodynamic parameters determined by isothermal titration calorimetry. The prospective structure-based drug design study was the first of its kind, targeting the conserved interface of the response regulator to discover inhibitors with a new mode of action. The development of the PHOENIX scoring function demonstrated the use of high-quality data (high-resolution crystal structures with thermodynamic parameters determined by ITC) and inclusion of descriptors to better capture entropic contributions might be the key advances to achieve accurate and robust binding affinity predictions.

5.1.1 PhoP Response Regulator Inhibitors

In a prospective structure-based drug design study, 8 drug-like compounds were discovered to inhibit the *S. enterica* PhoP response regulator. A hybrid strategy coupling

computational and experimental methods was used to identify these PPI inhibitors. A computational method consisting of structure-based virtual screening for pose prediction, followed by rescoring using a consensus scoring scheme, prioritized potential compounds for experimental testing. A series of biochemical and biophysical assays were used to validate the biological activity and characterize the potential mode of action. The 8 drug-like compounds discovered represent first-in-class inhibitors of a conserved bacterial signal transduction module important for virulence in a number of gram-negative pathogens. Experimental results suggested that these compounds bind at the functionally important $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface and act in an allosteric manner to inhibit the formation of the PhoP-DNA complex necessary for gene expression. Discovery of these first-in-class inhibitors demonstrated a novel strategy for the development of antibiotics with new modes of action. With the increasing resistance of antibiotics currently in clinical use, this study demonstrated a new approach for combating bacterial virulence. These inhibitors also serve as valuable lead compounds for optimization by medicinal chemistry to improve affinity and pharmacological properties. In addition, such drug-like compounds serve as useful molecular probes to identify and characterize potential small-molecule binding sites at the $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface. These chemical probes can also be used as molecular dials to tune the PhoQ/PhoP signal transduction pathway in a time- and dose-dependent manner to examine the impact of PhoP in regulating virulence.

5.1.2 PHOENIX Scoring Function for Binding Affinity Predictions

The development of the PHOENIX scoring function for binding affinity predictions demonstrated the use of high-resolution crystal structures with thermodynamic parameters determined by ITC and inclusion of descriptors to better capture entropic contributions might be the key advances towards accurate and robust affinity estimations. PHOENIX used a total of 112 complexes and 42 physicochemical descriptors for model training, and partial least squares to derive the regression equations. While the estimations for change of enthalpy (ΔH) and change of entropy ($T\Delta S$) were of marginal accuracy, the results derived for change of binding free energy (ΔG) were

accurate and comparable with the top-performing scoring functions. Affinity predictions by PHOENIX have been demonstrated to be accurate and robust in external tests using multiple versions (2002, 2004, 2007, 2009) and subsets (resolutions, protein families, affinities, diverse and non-redundant complexes) of the PDBbind database. In particular, PHOENIX performs better than existing scoring functions for estimating affinities of high-resolution ($\leq 2 \text{ \AA}$) complexes, demonstrating its value for atomic-level structure-based design studies. PHOENIX is valuable for accurate binding affinity predictions, and should also be useful for binding pose predictions and docking method assessments with high-resolution protein-ligand complexes.

5.2 Future Directions

5.2.1 Further Characterization and Design of PhoP Inhibitors

To determine the mode of action and elucidate protein-ligand interactions at an atomic level, X-ray crystallography and NMR studies must be performed. Cell-based assays and animal model tests will be needed to assess the therapeutic potential of these PhoP inhibitors. Structural analogs of the 8 drug-like compounds can be designed to enhance affinity and pharmacological properties, in addition to characterizing the structure-activity relationships.

5.2.2 Advances to Achieve Accurate Binding Affinity Predictions

More physically accurate descriptors will need to be developed in order to achieve accurate predictions of change of enthalpy (ΔH) and change of entropy ($T\Delta S$). Descriptors that can separate enthalpic and entropic contributions will be instrumental to achieve accurate thermodynamic parameter estimations. Better methods to capture solvation and desolvation contributions (either implicitly and explicitly) and inclusion of multiple binding modes (ligand and protein) to represent the conformational statistics will be necessary to achieve more accurate change of entropy ($T\Delta S$)

estimations. Development of more accurate shape and volume descriptors will be important to estimate the change in configurational (translational and rotational) entropy, illustrated in theoretical studies to be an important component to the overall change of entropy ($T\Delta S$). Better electrostatics models (e.g., polarizable force fields) will need to be used to accurately estimate enthalpic contributions and its individual components (e.g., electrostatics, hydrogen-bonding, van der Waals contacts) to binding free energy.

5.3 Conclusion

5.3.1 Targeting Protein-Protein Interactions for Drug Discovery

Protein-protein interactions are involved in a variety of biological functions such as in metabolism and signal transduction. Due to its prevalence and importance in biology, PPI serve as an attractive structural motif for therapeutics development. Discovery of first-in-class PhoP response regulator inhibitors targeting the functionally important $\alpha 4$ - $\beta 5$ - $\alpha 5$ interface demonstrated this strategy for the design and development of antibiotics with new modes of action. Two-component signal transduction systems are highly prevalent and important in bacterial physiology and virulence. Modulation of these conserved systems by targeting the output response regulator should provide a deeper insight into the biological roles and importance of these gene regulatory elements.

5.3.2 Binding Affinity Prediction

Binding affinity predictions is the most important and challenging component to computer-aided structure-based drug design. Methods for accurate affinity predictions are important and widely employed in computational lead discovery and optimization studies. Advances implemented and assessed in the PHOENIX scoring

function will guide the future development of accurate and robust methods for affinity predictions.

5.3.3 Future of Computer-Aided Molecular Design

Scientific contributions accomplished in this dissertation are a testament to the value of computer-aided molecular design methods to examine molecular recognition and application to drug discovery. Discovery of PPI inhibitors demonstrated a new frontier for antibiotics development. Advances implemented in PHOENIX suggested important advances to implement in order to achieve accurate binding affinity predictions. Taken together, these studies should guide the future development and application of CAMD methods for molecular recognition and structure-based drug design.

References

- R. Abel, T. Young, R. Farid, B. J. Berne, and R. A. Friesner. Role of the active-site solvent in the thermodynamics of factor xa ligand binding. *J. Amer. Chem. Soc.*, 130:2817–2831, 2008.
- Ajay and M. A. Murcko. Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.*, 38:4953, 1995.
- J. An, M. Totrov, and R. Abagyan. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Prot.*, 4:752, 2005.
- P. Bachhawat and A. M. Stock. Crystal structures of the receiver domain of the response regulator phop from escherichia coli in the absence and presence of the phosphoryl analog beryll fluoride. *J. Bacteriol.*, 189:5987, 2007.
- D. L. Beveridge and F. M. Dicapua. Free-energy via molecular simulation - applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Chem.*, 18:431, 1989.
- K. H. Bleicher, H. J. Bohm, K. Muller, and A. I. Alanine. A guide to drug discovery: Hit and lead generation: Beyond high-throughput screening. *Nat. Drug Disc.*, 2: 369–378, 2003.
- D. M. Blow. Rearrangement of crickshank’s formulae for the diffraction-component precision index. *Acta Cryst. Sec. D*, 58:792, 2002.
- H. J. Bohm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput. Aid. Mol. Des.*, 8:243, 1994.
- H. J. Bohm. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3d database search programs. *J. Comput. Aid. Mol. Des.*, 12:309, 1998.
- H. J. Bohm and M. Stahl. The use of scoring functions in drug discovery applications. *Rev. Comput. Chem.*, 18:41, 2002.
- G. P. Brady and P. F. W. Stouten. Fast prediction and visualization of protein binding pockets with pass. *J. Comput. Aid. Mol. Des.*, 14:383, 2000.

- S. P. Brown and P. J. Hajduk. Effects of conformational dynamics on predicted protein druggability. *ChemMedChem*, 1:70, 2006.
- A. Brunger and L. M. Rice. Crystallographic refinement by simulated annealing: methods and applications. *Methods in Enzymology*, 277:243, 1997.
- D. Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, 437:640–647, 2005.
- C. A. Chang. Ligand configurational entropy and protein binding. *Proc. Nat. Acad. Sci.*, 104:1534–1539, 2007.
- T. Cheng, X. Li, Y. Li, Z. Liu, and S. Wang. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.*, 49:1079–1093, 2009.
- E. Choi, E. A. Groisman, and D. Shin. Activated by different signals, the phop/phoQ two-component system differentially regulates metal uptake. *J. Bacteriol.*, 191:7174, 2009.
- T. Clackson and J. A. Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267:383, 1995.
- P. Cozzini, G. E. Kellogg, F. Spyraakis, G. Costantino D. J. Abraham, A. Emerson, F. Fanelli, H. Gohlke, L. A. Kuhn, G. M. Morris, M. Orozco, T. A. Pertinhez, M. Rizzi, and C. A. Sotriffer. Target flexibility: An emerging consideration in drug discovery and design. *J. Med. Chem.*, 51:6237–6255, 2008.
- D. W. J. Cruickshank. Remarks about protein structure precision. *Acta Cryst. Sect. D*, 55:583, 1999.
- E. Crunwald and C. Steel. Solvent reorganization and thermodynamic enthalpy-entropy compensation. *J. Am. Chem. Soc.*, 117:5687–5692, 1995.
- I. Davis and D. Baker. RosettaLigand docking with full ligand and receptor flexibility. *J. Mol. Biol.*, 385:381–392, 2009.
- J. E. DeLorbe, J. H. Clements, M. G. Teresk, A. P. Benfield, H. R. Plake, L. E. Millspaugh, and S. F. Martin. Thermodynamic and structural effects of conformational constraints in protein-ligand interactions: Entropic paradox associated with ligand preorganization. *J. Am. Chem. Soc.*, 131:16758–16770, 2009.
- M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aid. Mol. Des.*, 11:425, 1997.

- T. J. A. Ewing and I. D. Kuntz. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.*, 18:1175–1189, 1997.
- T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. Dock 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.*, 15:411–428, 2001.
- S. Eyrisch and V. Helms. Transient pockets on protein surfaces involved in protein-protein interaction. *J. Med. Chem.*, 50:3457, 2007.
- J. A. Feng and G. R. Marshall. Skate: A docking program that decouples systematic sampling from scoring. *J. Comput. Chem.*, 1:1–2, 2010.
- A. Fernandez and H. A. Scheraga. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc. Nat. Acad. Sci.*, 100:113, 2003.
- P. Ferrara, H. Gohlke, D.J. Price, G. Klebe, and C.L. Brooks. Assessing scoring functions for protein ligand interactions. *J. Med. Chem.*, 47:3032, 2004.
- E. Freire. Isothermal titration calorimetry. *Curr Prot. Cell Biol.*, 2004.
- E. Freire. Do enthalpy and entropy distinguish first in class from best in class? *Drug Discovery Today*, 13:869, 2008.
- E. Freire. A thermodynamic approach to the affinity optimization of drug candidates. *Chem. Biol. Drug Des.*, 74:468, 2009.
- R. Gao and A. M. Stock. Biological insights from structures of two-component proteins. *Annu. Rev. Microbiol.*, 63:133, 2009.
- R. Gao and A. M. Stock. Molecular strategies for phosphorylation-mediated regulation of response regulator activity. *Curr. Opin. Microbiol.*, 13:160, 2010.
- R. Gao, T. R. Mack, and A. M. Stock. Bacterial response regulators: versatile regulatory strategies from common domains. *Trends Biochem. Sci.*, 32:223, 2007.
- R. Gao, Y. Tao, and A. M. Stock. System-level mapping of escherichia coli response regulator dimerization with fret hybrids. *Mol. Microbiol.*, 69:1358, 2008.
- D. K. Gehlhaar, G. M. Verkhivker, P. A. Rejto, C. J. Sherman, D. R. Fogel, L. J. Fogel, and S. T. Freer. Molecular recognition of the inhibitor ag-1343 by hiv-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.*, 2:317, 1995.
- P. Geladi and B. R. Kowalski. Partial least-squares regression: a tutorial. *Anal. Chem. Acta.*, 185:1, 1986.

- U. Gerhard, M. S. Searle, and D. H. Williams. The free energy change of restricting a bond rotation in the binding of peptide analogues to vancomycin group antibiotics. *Bioorg. Med. Chem. Lett.*, 3:803–808, 1993.
- H. Gohlke and G. Klebe. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed.*, 41:2644, 2002.
- H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, 295:337, 2000.
- D. S Goodsell, G. M. Morris, and A. J. Olson. Automated docking of flexible ligands: Applications of autodock. *J. Molec. Recog.*, 9:1–5, 1996.
- E. A. Groisman. The pleiotropic two-component regulatory system phop-phoq. *J. Bacteriol.*, 183:1835, 2001.
- V. Le Guilloux, P. Schmidtke, and P. Tuffery. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinfo.*, 10:1, 2009.
- I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, Bioinformatics*, 47:409, 2002.
- T. Hansson, J. Marelus, and J. Aqvist. Ligand binding affinity prediction by linear interaction energy methods. *J. Comput. Aided Mol. Des.*, 12:27–35, 1998.
- R. D. Head, M. L. Smythe, T. I. Oprea, C. L. Waller, S. M. Green, and G. R. Marshall. Validate: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.*, 118:3959, 1996.
- M. Hendlich, F. Rippmann, and G. Barnickel. Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.*, 15:359, 1997.
- S. Homans. Water, water everywhere - except where it matters? *Drug Discovery Today*, 12:534, 2007.
- L. Huang, R. Abel, R. A. Friesner, and B. J. Berne. Dewetting transitions in protein cavities. *J. Chem. Theory Comput.*, 5:1462–1473, 2009.
- R. Huey, G. M. Morris, A. J. Olson, and D. S Goodsell. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.*, 28:1145–1152, 2007.
- D. T. Hung, E. A. Shakhnovich, E. Pierson, and J. J. Mekalanos. Small-molecule inhibitor of vibrio cholerae virulence and intestinal colonization. *Science*, 310:670, 2005.

- C. A. Hunter. Quantifying intermolecular interactions: Guidelines for the molecular recognition toolbox. *Angew. Chem. Int. Ed.*, 43:5310–5324, 2004.
- J. J. Irwin and B. K. Shoichet. Zinc a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, 45:177–182, 2005.
- A. Jamet, C. Rousseau, J. Monfort, E. Frapy, X. Nassif, and P. Martin. The response regulator phop is important for survival under conditions of macrophage-induced stress and virulence in yersinia pestis. *Microbiol. SGM*, 155:2288, 2009.
- D. Jiao, P. A. Golubkov, T. A. Darden, and P. Ren. Calculation of proteinligand binding free energy by using a polarizable potential. *Proc. Nat. Acad. Sci.*, 105:6290, 2008.
- C. Johnson, J. Newcombe, S. Thorne, H. Borde, L. Eales-Reynolds, and A. Goringe. Generation and characterization of a phop homologue mutant of neisseria meningitidis. *Mol. Microbiol.*, 39:1345, 2001.
- G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267:727–748, 1997.
- S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Science*, 93:13, 1996.
- A. Kato and E. A. Groisman. The phoq/phop regulatory network of salmonella enterica. *Adv. Exp. Med. Biol.*, 631:7, 2008.
- Y. Kawasaki, E. E. Chufan, V. Lafont, K. Hidaka, and Y. Kiso. How much binding affinity can be gained by filling a cavity? *Chem. Biol. Drug Des.*, 75:143, 2010.
- D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Disc.*, 3:935, 2004.
- G. J. Kleywegt and T. A. Jones. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta. Cryst. Sec. D*, 50:178, 1994.
- P. Kollman. Free energy calculations: applications to chemical and biochemical phenomena. *Chem. Rev.*, 93:2395–2417, 1993.
- A. Krammer, P. D. Kirchoff, X. Jiang, C. M. Venkatachalam, and M. Waldman. Ligscore: a novel scoring function for predicting binding affinities. *J. Mol. Graph. Model.*, 23:395, 2005.
- I. D. Kuntz. Structure-based strategies for drug design and discovery. *Science*, 257:1078, 1992.

- J. E. Ladbury, G. Klebe, and E. Freire. Adding calorimetric data to decision making in lead discovery: A hot tip. *Nat. Drug Disc.*, 9:23–27, 2010.
- R. A. Laskowski. Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, 13:323, 1995.
- A. T. R. Laurie and R. M. Jackson. Q-sitefinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, 21:1908, 2005.
- T. Lazaridis. Inhomogeneous fluid approach to solvation thermodynamics. 2. applications to simple fluids. *J. Phys. Chem. B*, 102:3542–3550, 1998.
- J. Lee and C. Seok. A statistical rescoring scheme for protein-ligand docking: Consideration of entropic effect. *Proteins: Structure, Function, Bioinformatics*, 70:1074, 2008.
- M. Levitt. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.*, 226:507, 1992.
- L. Li, J. J. Dantzer, J. Nowacki, B. J. OCallaghan, and S. O. Meroueh. Pdbcal: A comprehensive dataset for receptor ligand interactions with three-dimensional structures and binding thermodynamics from isothermal titration calorimetry. *Chem. Biol. Drug. Des.*, 71:529–532, 2008.
- J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Prot. Sci.*, 7:1884, 1998.
- W. Linert and R. F. Jameson. The isokinetic relationship. *Chem. Soc. Rev.*, 18:477–505, 1989.
- R. Lumry and S. Rajender. Enthalpy-entropy compensation phenomena in water solutions of proteins and small molecules: A ubiquitous property of water. *Biopolymers*, 9:1125–1227, 1970.
- P. D. Lyne. Structure-based virtual screening: an overview. *Drug Discovery Today*, 7:1047, 2002.
- M. S. Marlow, J. Dogan, K. K. Frederick, K. G. Valentine, and A. J. Wand. The role of conformational entropy in molecular recognition by calmodulin. *Nat. Chem. Biol.*, 6:352–358, 2010.
- S. F. Martin. Preorganization in biological systems: Are conformational constraints worth the energy? *Pure Appl. Chem.*, 79:193–200, 2007.
- I. Massova and P. A. Kollman. Combined molecular mechanical and continuum solvent approach (mm-pbsa/gbsa) to predict ligand binding. *Pers. Drug Disc. Des.*, 18:113–135, 2000.

- S. L. McGovern, E. Caselli, N. Grigorieff, and B. K. Shoichet. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.*, 45:1712–1722, 2002.
- I. S. Moreira, P. A. Fernandes, and M. J. Ramos. Hot spots-a review of the protein-protein interface determinant amino-acid residues. *Proteins: Structure, Function, Bioinformatics*, 68:803, 2007.
- G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19:1639–1662, 1998.
- G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.*, 30:2785–2791, 2009.
- J. Moss, P. Fisher, B. Vick, E. A. Groisman, and A. Zychlinsky. The regulatory protein phop controls susceptibility to the host inflammatory response in shigella flexneri. *Cell Microbiol.*, 2:443, 2000.
- A. Nicholls, G. B. McGaughey, R. P. Sheridan, A. C. Good, G. Warren, M. Mathieu, S. W. Muchmore, S. P. Brown, J. A. Grant, J. A. Haigh, N. Nevens, A. N. Jain, and B. Kelley. Molecular shape and medicinal chemistry: a perspective. *J. Med. Chem.*, 2010.
- T. S. G. Olsson, M. A. Williams, W. R. Pitt, and J. E. Ladbury. The thermodynamics of protein-ligand interaction and solvation: insights for ligand design. *J. Mol. Biol.*, 384:1002, 2008.
- F. Osterberg, G. M. Morris, and M. F. Sanner. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in autodock. *Proteins: Structure, Function, and Bioinformatics*, 46:34–40, 2002.
- P. Oyston, N. Dorrell, K. Williams, S. Li, M. Green, and R. Titball. The response regulator phop is important for survival under conditions of macrophage-induced stress and virulence in yersinia pestis. *Infect Immun.*, 68:3419, 2000.
- H. Park, J. Lee, and S. Lee. Critical assessment of the automated autodock as a new docking tool for virtual screening. *Proteins: Structure, Function, and Bioinformatics*, 65:549–554, 2006.
- D. J. Payne, M. N. Gwynn, D. J. Holmes, and D. L. Pompliano. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Disc.*, 6:29, 2006.

- K. P. Peters, J. Fauck, and C. Frommel. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.*, 256:201, 1996.
- C. C. Pimentel and A. L. McClellan. Hydrogen bonding. *Annu. Rev. Phys. Chem.*, 22:347–385, 1971.
- D. A. Rasko, C. G. Moreira, D. R. Li, N. C. Reading, J. M. Ritchie, M. K. Waldor, N. Williams, R. Taussig, S. Wei, M. Roth, D. T. Hughes, J. F. Huntley, M. W. Fina, J. R. Falck, and V. Sperandio. Targeting qsec signaling and virulence for antibiotic development. *Science*, 321:1078, 2008.
- J. Roy and C. Laughton. Long-timescale molecular-dynamics simulations of the major urinary protein provide atomistic interpretations of the unusual thermodynamics of ligand binding. *Biophys. J.*, 99:218, 2010.
- A. Ruvinsky. Role of binding entropy in the refinement of protein-ligand docking predictions: Analysis based on the use of 11 scoring functions. *J Comput. Chem.*, 28:1364–1372, 2007.
- A. Ruvinsky and A. Kozintsev. New and fast statistical-thermodynamic method for computation of protein-ligand binding entropy substantially improves docking accuracy. *J Comput. Chem.*, 26:1089–1095, 2005.
- S. Salaniwal, E. Manas, J. Alvarez, and R. Unwalla. Critical evaluation of methods to incorporate entropy loss upon binding in high-throughput docking. *Proteins: Structure, Function, Bioinformatics*, 66:422, 2007.
- M. S. Searle and D. H. Williams. The cost of conformational order: Entropy changes in molecular associations. *J. Am. Chem. Soc.*, 114:10690–10697, 1992.
- E. A. Shakhnovich, D. T. Hung, E. Pierson, K. Lee, and J. J. Mekalanos. Virstatin inhibits dimerization of the transcriptional activator toxT. *Proc. Nat. Acad. Sci.*, 104:2372, 2007.
- F. B. Sheinerman, R. Norel, and B. Honig. Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.*, 10:153, 2000.
- Y. Shi, M. Cromie, F. Hsu, J. Turk, and E. A. Groisman. Phop-regulated salmonella resistance to the antimicrobial peptides magainin 2 and polymyxin b. *Mol. Microbiol.*, 53:229, 2004.
- B. K. Shoichet. Virtual screening for chemical libraries. *Nature*, 432:862–865, 2004.
- S. B. Shuker, P. J. Hajduk, R. P. Meadows, and S. W. Fesik. Discovering high-affinity ligands for proteins: Sar by nmr. *Science*, 274:1531–1534, 1996.

- C. A. Sotriffer, P. Sanschagrin, H. Matter, and G. Klebe. Sfcscorer: Scoring functions for affinity prediction of proteinligand complexes. *Proteins: Structure, Function, Bioinformatics*, 73:395–419, 2008.
- M. Stahl and M. Rarey. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.*, 44:1035, 2001.
- E. Stjernschantz and C. Oostenbrink. Improved ligand-protein binding affinity predictions using multiple binding modes. *J. Comput. Chem.*, 98:2682, 2010.
- A. M. Stock and P. Guhaniyogi. A new perspective on response regulator activation. *J. Bacteriol.*, 188:7328, 2006.
- A. M. Stock, V. L. Robinson, and P. N. Goudreau. Two-component signal transduction. *Annu. Rev. Biochem.*, 69:183, 2000.
- C. M. Taylor, Y. Barda, O. Kisselev, and G. R. Marshall. Modulating g-protein coupled receptor/g-protein signal transduction by small molecules suggested by virtual screening. *J. Med. Chem.*, 51:5297, 2008.
- K. S. Thorn and A. A. Bogan. Aseddb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17:284, 2001.
- N. Trbovic, J. Cho, R. Abel, R. Friesner, and M. Rance. Protein side-chain dynamics and residual conformational entropy. *J. Am. Chem. Soc.*, 131:615, 2009.
- A. Tripathi and G. E. Kellogg. A novel and efficient tool for locating and characterizing protein cavities and binding sites. *Proteins: Structure, Function, Bioinformatics*, 78:825, 2009.
- G. M. Verkhivker, D. Bouzida, D. K. Gehlhaar, P. A. Rejto, S. T. Freer, and P. W. Rose. Complexity and simplicity of ligandmacromolecule interactions: the energy landscape perspective. *Curr. Opin. Struct. Biol.*, 12:197, 2000.
- E. G. Vescovi, F. Soncini, and E. A. Groisman. Mg²⁺ as an extracellular signal: environmental regulation of salmonella virulence. *Cell*, 84:165, 1996.
- L. D. Walensky, A. L. Kung, I. Escher, T. J. Malia, S. Barbuto, R. D. Wright, G. Wagner, G. L. Verdine, and S. J. Korsmeyer. Activation of apoptosis in vivo by a hydrocarbon-stapled bh3 helix. *Science*, 305:1466–1470, 2004.
- W. P. Walters, M. T. Stahl, and M. A. Murcko. Virtual screeningan overview. *Drug Discovery Today*, 3:160–178, 1998.

- J. M. Wang, P. Morin, W. Wang, and P. A. Kollman. Use of mm-pbsa in reproducing the binding free energies to hiv-1 rt of tibo derivatives and predicting the binding mode to hiv-1 rt of efavirenz by docking and mm-pbsa. *J. Am. Chem. Soc.*, 123: 5221, 2001.
- R. Wang, Y. Fu, and L. Lai. A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comput. Sci.*, 37:615, 1997.
- R. Wang, L. Lai, and S. Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des.*, 16:11–26, 2002.
- R. Wang, Y. Lu, and S. Wang. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.*, 46:2287, 2003.
- R. Wang, Y. Lu, X. Fang, and S. Wang. An extensive test of 14 scoring functions using the pdbind refined set of 800 protein-ligand complexes. *J. Chem. Inf. Comp. Sci.*, 44:2114–2125, 2004.
- G. L. Warren, C. W. Andrews, A. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff, , and M. S. Head. A critical assessment of docking programs and scoring functions. *J. Med. Chem.*, 49:5912, 2006.
- C. Weber. vant hoff revisited: Enthalpy of association of protein subunits. *Phys. Chem.*, 99:1052–1059, 1995.
- M. Weisel, E. Proschak, and G. Schneider. Pocketpicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.*, 1:1, 2007.
- J. A. Wells and C. L McClendon. Reaching for high-hanging fruit in drug discovery at proteinprotein interfaces. *Nature*, 450:1001, 2007.
- S. Wold, A. Ruhe, H. Wold, and W. J. Dunn III. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.*, 5:735, 1984.
- T. Young, R. Abel, H. Kim, B. J. Berne, and R. A. Friesner. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc. Nat. Acad. Sci.*, 104:808–813, 2007.
- T. Young, L. Hua, X. Huang, R. Abel, R. A. Friesner, and B. J. Berne. Dewetting transitions in protein cavities. *Proteins.*, 78:1856–1869, 2010.

Curriculum Vitae

Yat T. Tang
329A N Boyle Ave.
St. Louis, MO 63108
yattang@gmail.com

Born	June 15, 1981 Kowloon, Hong Kong
Degrees	B.S. Cybernetics, June 2003 Bioinformatics Concentration Minor in Mathematics <i>University of California, Los Angeles, CA USA</i> Ph.D. Computational Biology, anticipated August 2010 <i>Washington University, St. Louis, MO USA</i>
Employment	Intern, 5/2008 - 8/2008 <i>Clinical Pharmacokinetics and Pharmacodynamics</i> <i>Merck & Co., Inc, West Point, PA</i> Research Associate, 7/2003 - 5/2004 <i>Dept. of Neurobiology</i> <i>University of California, Los Angeles, CA</i> Research Assistant, 9/1999 - 6/2003 <i>Dept. of Psychiatry and Biobehavioral Sciences</i> <i>University of California, Los Angeles, CA</i>
Professional Societies	American Chemical Society American Peptide Society

Biophysical Society

Select Honors and Awards

PhRMA Foundation Predoctoral Fellowship in Informatics
American Chemical Society Computational Division Chemical Computing Group Award
National Biomedical Computational Resources Summer Institute Travel Award
American Peptide Symposium Travel Award
Washington University Vision Sciences Predoctoral Fellowship
Washington University Siteman Center Cancer Biology Predoctoral Fellowship
National Science Foundation Integrative Graduate Education and Research Traineeship

Publications

Tang, Y. T., Marshall, G. R. Virtual Screening for Lead Discovery. *Drug Design and Discovery: Methods and Protocols*, in press.

Tang, Y. T., Latifi, T., Groisman, E. A., Marshall, G. R. Virtual Screening Targeting the PhoP Response Regulator to Inhibit Bacterial Virulence. *Proceedings of the Twenty-First American Peptide Symposium* 2009.

Cottrell, G. S., Roosterman, D., Marvizon, J. C., Song, B., Wick, E., Pikios, S., Wong, H., Berthelie, C., Tang, Y., Sternini, C., Bunnett, N. W., Grady, E. F. Localization of calcitonin receptor-like receptor and receptor activity modifying protein 1 in enteric neurons, dorsal root ganglia, and the spinal cord of the rat. *Journal of Comparative Neurology* 2005; 490:3; 239-255.

Manuscripts in preparation

Tang, Y. T., Marshall, G. R. PHOENIX: Scoring Function for Affinity Prediction using High-Resolution Crystal Structures and Calorimetry Measurements.

Tang, Y. T., Gao, R., Havranek, J. J., Groisman, E. A., Stock, A.M., Marshall, G. R. Inhibition of Bacterial Virulence: Drug-like Molecules Targeting the Salmonella Enterica PhoP Response Regulator.

Tang, Y. T., Marshall, G. R. Entropic Contributions to Binding Free Energy: Considerations for Affinity and Specificity in Molecular Recognition.

Tang, Y. T., Marshall, G. R. Entropic and Scoring Functions: Strategies Towards Accurate Binding Affinity Predictions.

August 2010

Molecular Recognition in Protein-Ligand Interactions, Tang, Ph.D. 2010