

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

---

Student Research Projects, Dissertations, and  
Theses - Chemistry Department

Chemistry, Department of

---

2-2013

## Utilizing NMR Spectroscopy and Molecular Docking as Tools for the Structural Determination and Functional Annotation of Proteins

Jaime Stark

University of Nebraska-Lincoln, [jaime.stark@huskers.unl.edu](mailto:jaime.stark@huskers.unl.edu)

Follow this and additional works at: <https://digitalcommons.unl.edu/chemistrydiss>



Part of the [Analytical Chemistry Commons](#), [Biochemistry Commons](#), [Bioinformatics Commons](#), and the [Structural Biology Commons](#)

---

Stark, Jaime, "Utilizing NMR Spectroscopy and Molecular Docking as Tools for the Structural Determination and Functional Annotation of Proteins" (2013). *Student Research Projects, Dissertations, and Theses - Chemistry Department*. 40.  
<https://digitalcommons.unl.edu/chemistrydiss/40>

This Article is brought to you for free and open access by the Chemistry, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Student Research Projects, Dissertations, and Theses - Chemistry Department by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

UTILIZING NMR SPECTROSCOPY AND MOLECULAR DOCKING AS TOOLS  
FOR THE STRUCTURAL DETERMINATION AND FUNCTIONAL ANNOTATION  
OF PROTEINS

By

Jaime L. Stark

A DISSERTATION

Presented to the Faculty of  
The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Chemistry

Under the Supervision of Professor Robert Powers

Lincoln, Nebraska

February, 2013



UTILIZING NMR SPECTROSCOPY AND MOLECULAR DOCKING AS TOOLS  
FOR THE STRUCTURAL DETERMINATION AND FUNCTIONAL ANNOTATION  
OF PROTEINS

Jaime L. Stark, Ph.D.

University of Nebraska, 2013

Advisor: Robert Powers

With the completion of the Human Genome Project in 2001 and the subsequent explosion of organisms with sequenced genomes, we are now aware of nearly 28 million proteins. Determining the role of each of these proteins is essential to our understanding of biology and the development of medical advances. Unfortunately, the experimental approaches to determine protein function are too slow to investigate every protein. Bioinformatics approaches, such as sequence and structure homology, have helped to annotate the functions of many similar proteins. However, despite these computational approaches, approximately 40% of proteins still have no known function. Alleviating this deficit will require high-throughput methods that combine experimental and computational approaches.

Nuclear magnetic resonance (NMR) ligand affinity screens are an experimental approach that can detect protein-ligand interactions, measure a corresponding dissociation constant, and reliably identify the ligand binding site. Correspondingly, molecular docking is a computational tool that can be used predict the location of the binding site and conformation of a compound when bound to a protein using only the structures of both the protein and the compound. Molecular docking provides an

rapid way to generate protein-ligand costructures and evaluate numerous compounds in a large chemical library. Together, molecular docking and NMR ligand affinity screens provide valuable information for determining the function of a protein.

This dissertation describes the high-throughput application of the Functional Annotation Screening Technology by NMR (FAST-NMR), which combines NMR ligand affinity screens, molecular docking, and bioinformatics to help determine the function of 20 previously uncharacterized proteins. Additionally, new tools were developed to utilize 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC (heteronuclear single quantum coherence) chemical shift perturbations (CSPs) and molecular docking to generate consensus binding sites (CSP-Consensus) and protein-ligand costructures (AutoDockFilter). Virtual screening was also successful utilized to identify a potential natural ligand and propose a function for the YndB protein from *Bacillus subtilis*. Finally, the solution structure of human protein DNAJA1 was determined and its potential role in pancreatic cancer investigated.

**DEDICATED TO:**

My parents: Dan and Holle Stark

**ACKNOWLEDGEMENTS**

Even though my name is the only one at the front of this dissertation, the work presented within was truly a collaboration of many. I would like to thank everyone that has supported me and made this dissertation possible.

My greatest gratitude goes to my research advisor, Dr. Robert Powers. You gave me the freedom to explore and guidance when I got lost. You taught me how to ask the interesting questions and express my ideas. Ultimately, your support and encouragement helped me through the difficult challenges, both in research and personally. You will always be able to count on my help, and I look forward to future collaborations, conference meetings, and hopefully plenty of funding for both of us.

I would like to thank the members of my committee, Dr. David Hage, Dr. Gerard Harbison, Dr. James Takacs, and Dr. Paul Blum, for holding me and my research to a high standard. Our discussions have truly challenged me understand my research better and succeed at UNL.

During my time in graduate school, I have found a love of teaching that helped make me a better communicator. I would like to thank Dr. Eric Malina for helping to develop my confidence standing in front of hundreds of students. I would also like to

thank Dr. David Clevette of Doane College for giving me the opportunity to teach my own class and develop my own voice.

Scientific research is a collaborative effort, and I would also like to extend a warm thank you to the outside collaborators who have been tremendous resources for the research presented here: Dr. Gaetano Montelione and Dr. Rong Xiao from the Northeast Structural Genomics Consortium, and Dr. Pankaj Singh from the University of Nebraska Medical Center.

Surviving the many vagaries of being a graduate student would not have been possible without the help of the members of the Powers Group, past and present. I would like to thank Matt Shortridge, Kelly Mercier, Brad Worley, Steve Halouska, Bo Zhang, Teklab Gebregiworgis, Darrell Marshall, Jonathan Catazaro, Shulei Lei, Nick Sisco, Quin Schultz, and Emily Snell. I have enjoyed sharing research results, class notes, research conference hotel rooms, and long group meetings with all of you. I wish you all the best of luck.

None of this would have been possible without the love and patience of my family. Mom and Dad, from a very early age, you have encouraged my curiosity in the world and did everything you could to give the opportunity to explore it. My success is the direct result of your support. Thank you, and I love you both. To my sister Jennee, despite being your older brother, you often took it upon yourself to look after me. I may not have thought I needed it then, but I did. Thank you, and I love you. To my nephews, Conner and Traden, watching the two of you grow up has me excited for the future. I'm positive you both will succeed with anything you choose to do.

Finally, I would like to thank my girlfriend Jenni. Your encouragement, advice, and companionship these past few years kept me motivated and sane (somewhat), and I will never forget sharing the experience of graduate school with you.

## TABLE OF CONTENTS

<b>CHAPTER 1 .....</b>	<b>1</b>
<b>1.1 STRUCTURAL AND FUNCTIONAL GENOMICS.....</b>	<b>1</b>
1.1.1 Defining protein function .....	1
1.1.2 Protein sequence to function .....	3
1.1.3 Protein structure to function .....	4
1.1.4 Protein active-site to function.....	5
<b>1.2 HIGH-THROUGHPUT SCREENING BY NMR .....</b>	<b>7</b>
1.2.1 Ligand-based NMR screens .....	7
1.2.2 Target-based NMR screens.....	10
<b>1.3 MOLECULAR DOCKING AND VIRTUAL SCREENING .....</b>	<b>13</b>
1.3.1 Docking .....	15
1.3.2 Scoring .....	19
1.3.3 Virtual screening and assessment.....	23
<b>1.4 SUMMARY OF WORK .....</b>	<b>28</b>
<b>1.5 REFERENCES .....</b>	<b>30</b>
<b>CHAPTER 2 THE FUNCTIONAL ANNOTATION OF HYPOTHETICAL PROTEINS FROM THE NORTHEAST STRUCTURAL GENOMICS CONSORTIUM .....</b>	<b>41</b>
<b>2.1 INTRODUCTION .....</b>	<b>41</b>
<b>2.2 MATERIALS AND METHODS .....</b>	<b>44</b>
2.2.1 Hypothetical proteins from the NESG.....	44
2.2.2 Function-based compound library and mixtures .....	46
2.2.3 Ligand-based screen .....	46
2.2.4 Target-based screen.....	47
2.2.5 Rapid generation of protein-ligand costructures .....	47
2.2.6 CPASS .....	48
2.2.7 Other bioinformatics tools .....	48
<b>2.3 RESULTS AND DISCUSSION .....</b>	<b>50</b>
2.3.1 <i>Bacillus subtilis</i> yjcQ (NESG ID: SR346).....	55
2.3.2 <i>Bacillus subtilis</i> ykvR (NESG ID: SR358).....	57

2.3.3 <i>Bacillus subtilis</i> ynzC (NESG ID: SR384) .....	59
2.3.4 <i>Bacillus subtilis</i> yozE (NESG ID: SR391) .....	61
2.3.5 <i>Bacteroides vulgatis</i> BVU_3908 (NESG ID: BvR153) .....	63
2.3.6 <i>Bordetella bronchiseptica</i> BB0938 (NESG ID: BoR11) .....	65
2.3.7 <i>Caulobacter crescentus</i> CC_0527 (NESG ID: CcR55).....	67
2.3.8 <i>Escherichia coli</i> ytfP (NESG ID: ER111) .....	69
2.3.9 <i>Escherichia coli</i> yrbA (NESG ID: ER115) .....	71
2.3.10 <i>Escherichia coli</i> yggU (NESG ID: ER14) .....	74
2.3.11 <i>Escherichia coli</i> yjbR (NESG ID: ER226) .....	77
2.3.12 <i>Escherichia coli</i> ydfO (NESG ID: ER251).....	79
2.3.13 <i>Escherichia coli</i> ygdR (NESG ID: ER382A).....	82
2.3.14 <i>Escherichia coli</i> ykfF (NESG ID: ER397) .....	84
2.3.15 <i>Escherichia coli</i> yeiV (NESG ID: ER541).....	86
2.3.16 <i>Porphyromonas gingivalis</i> PG_0361 (NESG ID: PgR37A) .....	88
2.3.17 <i>Rhodobacter sphaeroides</i> RHOS4_12090 (NESG ID: RhR5).....	90
2.3.18 <i>Salmonella typhimurium</i> STM0327 (NESG ID: StR65) .....	92
2.3.19 <i>Silicibacter pomeroyi</i> SPO1678 (NESG ID: SiR5).....	94
2.3.20 <i>Staphylococcus saprophyticus</i> SSP0609 (NESG ID: SyR11) .....	97
2.4 CONCLUSIONS.....	99
2.5 REFERENCES .....	100
APPENDIX 2A.....	106
CHAPTER 3 RAPID PROTEIN-LIGAND COSTRUCTURES USING CHEMICAL SHIFT PERTURBATIONS AND AUTODOCK .....	116
3.1 INTRODUCTION .....	116
3.2 MATERIALS AND METHODS .....	119
3.2.1 Preparation of the ligand and target protein .....	119
3.2.2 Prediction of ligand binding sites and chemical shift perturbations .....	120
3.2.3 Molecular docking .....	121
3.2.4 Filtering of docked ligand conformations .....	121
3.2.5 Molecular docking using a flexible binding site .....	122
3.2.6 Molecular docking using experimental NMR data.....	123
3.3 RESULTS AND DISCUSSION .....	125
3.3.1 Protein-ligand model systems.....	126

3.3.2 Comparison of blind docking and CSP-guided docking.....	127
3.3.3 Lowest-energy cluster is not necessarily the best conformer .....	129
3.3.4 CSP-guided docking with ADF filtering .....	133
3.3.5 CSP-guided docking with ADF filtering using apoproteins .....	137
3.3.6 Docking with experimental NMR data .....	140
3.4 CONCLUSIONS.....	143
3.5 REFERENCES .....	144
CHAPTER 4 AUTODOCKFILTER 2.0 AND CSP-CONSENSUS.....	150
4.1 INTRODUCTION .....	150
4.2 MATERIAL AND METHODS .....	151
4.2.1 Modifications to AutoDockFilter 2.0 .....	151
4.2.2 Defining the binding site with CSP-Consensus .....	154
4.2.3 Evaluation on protein-ligand systems with experimental CSPs.....	155
4.3 RESULTS AND DISCUSSION .....	158
4.3.1 Modifications to AutoDockFilter 2.0 .....	158
4.3.2 Defining the binding site with CSP-Consensus .....	160
4.3.3 Evaluation on protein-ligand systems with experimental CSPs.....	163
4.4 CONCLUSIONS.....	169
4.5 REFERENCES .....	170
CHAPTER 5 EVALUATION OF FUNCTIONAL SIMILARITY WITH CPASS 2.0.....	173
5.1 INTRODUCTION .....	173
5.2 MATERIAL AND METHODS .....	175
5.2.1 Evaluation of CPASS functional similarity.....	175
5.2.2 Tolerance of active-site variations .....	176
5.2.3 CPASS comparisons of proteins of unknown function .....	181
5.3 RESULTS AND DISCUSSION .....	182
5.3.1 Evaluation of CPASS functional similarity.....	182
5.3.2 Tolerance of active-site variations .....	185
5.3.3 CPASS comparisons of proteins of unknown function .....	191
5.4 CONCLUSIONS.....	193
5.5 REFERENCES .....	195
CHAPTER 6 THE SOLUTION STRUCTURE AND FUNCTION OF YNDB, AN AHSA1 PROTEIN FROM <i>B. SUBTILIS</i> .....	199



<b>6.1 INTRODUCTION .....</b>	<b>199</b>
<b>6.2 MATERIAL AND METHODS .....</b>	<b>203</b>
6.2.1 Solution structure of <i>B. subtilis</i> YndB .....	203
6.2.2 Sequence and structure similarity to YndB .....	204
6.2.3 Virtual screening of a lipid compound library .....	204
6.2.4 NMR titration experiment .....	207
6.2.5 <i>B. subtilis</i> YndB-ligand costructures .....	208
<b>6.3 RESULTS AND DISCUSSION .....</b>	<b>209</b>
6.3.1 Solution structure of <i>B. subtilis</i> YndB .....	209
6.3.2 Sequence and structure similarity to YndB .....	212
6.3.3 Virtual screening of a lipid compound library .....	213
6.3.4 NMR titration experiment .....	214
6.3.5 <i>B. subtilis</i> YndB-ligand costructures .....	218
<b>6.4 CONCLUSIONS .....</b>	<b>222</b>
<b>6.5 REFERENCES .....</b>	<b>226</b>
<b>CHAPTER 7 VIRTUAL SCREENING OF A FUNCTION-BASED COMPOUND LIBRARY .....</b>	<b>231</b>
7.1 INTRODUCTION .....	231
7.2 MATERIALS AND METHODS .....	233
7.3 RESULTS AND DISCUSSION .....	234
7.4 CONCLUSIONS .....	239
7.5 REFERENCES .....	240
<b>CHAPTER 8 HUMAN DNAJA1: A POTENTIAL THERAPEUTIC TARGET FOR PANCREATIC CANCER .....</b>	<b>242</b>
8.1 INTRODUCTION .....	242
8.2 MATERIALS AND METHODS .....	243
8.2.1 Selection of DNAJA1 from a pancreatic cancer ‘omics database .....	243
8.2.2 Effect of DNAJA1 overexpression on pancreatic cell stress modulation .....	245
8.2.3 Solution structure of the DNAJA1 J-domain .....	246
8.2.4 Identification of a ligand binding site on the DNAJA1 J-domain .....	249
8.3 RESULTS AND DISCUSSION .....	250
8.3.1 Selection of DNAJA1 from a pancreatic cancer ‘omics database .....	250
8.3.2 Effect of DNAJA1 overexpression on pancreatic cell stress modulation .....	254

	xi
8.3.3 Solution structure of the DNAJA1 J-domain .....	257
8.3.4 Identification of a ligand binding site on the DNAJA1 J-domain .....	268
8.4 CONCLUSIONS.....	276
8.5 REFERENCES .....	278
CHAPTER 9: SUMMARY AND FUTURE WORK .....	285
9.1 REFERENCES.....	294

"Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry."

- Richard Feynman

"Careful. We don't want to learn anything from this."

- Bill Watterson, "Calvin and Hobbes"

## CHAPTER 1

### INTRODUCTION<sup>†</sup>

#### 1.1 STRUCTURAL AND FUNCTIONAL GENOMICS

The completion of the Human Genome Project<sup>1,2</sup> in 2001 and the technology associated with it has led to an explosion in the number of organisms with sequenced genomes. As of September 2012, the Genomes OnLine Database (GOLD) indicates that 3,738 genomes have been completely sequenced, with another 11,605 genome projects in progress.<sup>3</sup> As a result of these sequencing projects, the DNA sequence of every gene in the genome in a particular organism could be determined. Because each gene carries the blueprint for the production of a protein, knowledge of the sequence of gene also provides knowledge of the amino acid sequence of the resulting protein. While genes are the basic unit of heredity from parent to offspring, proteins are responsible for most of the structural and biochemical functions in an organism, which makes determining the role of each protein essential to our understanding of biology and the development of medical advances.

**1.1.1 Defining protein function.** Protein function can be described in several ways that make it difficult to define and compare to other proteins.<sup>4</sup> The function of a protein could be described in terms of the cellular process with which the protein is involved, the specific enzymatic reaction the protein catalyzes, or a physiological role that the protein plays. In order to address this problem of definition, several projects were developed to provide a classification or ontology to protein function.

---

<sup>†</sup> Chapter 1, Section 2 and 3 were adapted from Stark, J. L. and Powers, R. Application of NMR and molecular docking in structure-based drug discovery. *Top Curr Chem* **326**, 1-34 (2012). Reprinted with permission, copyright 2012 by Springer.

The Enzyme Commission (E.C.) classification<sup>5</sup> is a hierarchical approach to define the function of enzymes, where each protein is assigned a four-numbered code, or E.C. number. Each number, from left to right, represents a more specific classification of the catalyzed reaction. While this approach works well for enzymes, non-enzymatic protein functions are not included. On the other hand, Gene Ontology (GO)<sup>6</sup> annotations create a well-defined vocabulary to represent the functions of a protein. Additionally, these GO terms can be divided into three categories: molecular function (i.e., the specific reaction or binding interaction the protein is involved in), cellular component (i.e., where in the organism the protein performs its function), and biological process (i.e., the higher level process that the protein function helps to accomplish).

There are several experimental methods that can be used to determine the function of a protein: genotype to phenotype studies,<sup>7</sup> monitoring gene expression levels,<sup>8,9</sup> protein-protein interaction assays,<sup>10,11</sup> ChIP-seq and RNA-seq transcription analysis,<sup>12,13</sup> enzymatic assays,<sup>14</sup> and RNA interference.<sup>15,16</sup> While many of these approaches are high-throughput, the results of these experiments often do not provide definitive evidence of the function of a protein.<sup>17-20</sup> Therefore, additional experimentation often requires even more time and resources.

Because of the success of the various genome projects, there are currently 27,661,073 protein sequences in the UniProtKB database (October 2012), which is continually growing.<sup>21,22</sup> Of these, 10,401,675 (37.6%) have no known function and have been assigned as “putative”, “uncharacterized”, or “unknown” proteins. Determining the function of each of these proteins is clearly not feasible using only an experimental approach. Due to these restraints, computational approaches are often utilized to annotate

the function of a protein by comparing it to the sequences/structures of homologous proteins with an experimentally determined function. Thus, only a small number of protein annotations are actually generated from experimental data.

**1.1.2 Protein sequence to function.** With over 27 million protein sequences available, determining sequence similarity between proteins is one of the most common approaches to identifying homology between proteins.<sup>4</sup> Two proteins that have very similar amino acid sequences have a common evolutionary origin and likely the same function. There are several tools available that align amino acid sequences using various alignment algorithms, which include BLAST,<sup>23,24</sup> HMMER,<sup>25</sup> SAM,<sup>26,27</sup> and ClustalW.<sup>28,29</sup> The results of these tools are typically a representation of the aligned sequences and a score that indicates the number of aligned residues as well as a probability that the alignment is correct.

As protein sequences diverge, the assumption is that the proteins are less homologous and it becomes more difficult to infer functional similarity.<sup>4,30</sup> However, sequence similarity does not directly indicate functional similarity since homologous proteins can be orthologs or paralogs. Orthologs are proteins that originated from a common ancestor, while paralogs are the result of gene duplication in the same genome; thus, orthologs tend to have more conserved protein function relative to paralogs.<sup>31</sup> Additionally, there is no definitive cutoff for sequence similarity that will guarantee a similar function. There are several cases of proteins, such as duck eye lens proteins<sup>32</sup> or phosphoglucose isomerase,<sup>33,34</sup> that have the same sequence yet have completely unrelated functions. Because a small number of changes in amino acid sequence can lead to large changes in the function, sequence similarity is not safe to definitively annotate

proteins. However, most of the “annotated” proteins in the UniProtKB have functions based primarily on global sequence similarity to other proteins.<sup>35</sup> This is troubling because only 538,259 (1.9%) of the total protein sequences have actually been manually reviewed and verified.

**1.1.3 Protein structure to function.** Proteins are composed of a string of amino acids with various structural and chemical properties (i.e. charge, polarity, size, etc.). The sequence or arrangement of these amino acids determines the specific three-dimensional structure of a protein. Unfortunately, experimentally determining the structure of a protein using X-ray crystallography or NMR spectroscopy is much more time intensive than determining the sequence.<sup>36,37</sup> The RCSB Protein Data Bank (PDB; [www.rcsb.org](http://www.rcsb.org))<sup>38,39</sup> has 85,848 protein structures deposited, however, this only represents the structures of 28,806 unique proteins. Compared to 27 million protein sequences, structural efforts are understandably far behind.

Structural genomics efforts, such as the Protein Structure Initiative (PSI),<sup>40</sup> have focused structure determination on protein sequences that have little amino acid sequence identity to proteins with known structures. The structure and sequence of that protein can be used to predict the structure of proteins with similar sequences.<sup>41</sup> This approach allows for the more efficient investigation of protein structure space and has resulted in 5,107 new protein structures deposited in the PDB (October 2012), many of which have unique structural architectures or folds.

Ideally, proteins that have the same chemical composition and the same structure should have the same molecular function. Sequence similarity approaches attempt to leverage this relationship by only comparing amino acid sequences, which is a primary

factor in the resulting protein structure. However, even small changes in sequence can have significant changes in the structure of a protein, which may also change the function. Therefore, directly comparing the structures of proteins should provide additional evidence for homology between proteins even in the absence of high sequence similarity.<sup>42,43</sup>

The most common methods/programs to compare the structures of proteins are DALI,<sup>44,45</sup> SSM,<sup>46</sup> FATCAT,<sup>47</sup> CATHEDRAL,<sup>48</sup> and MAMMOTH.<sup>49</sup> These programs typically attempt to find the best match between protein backbone atoms or protein secondary structure elements. The result is typically a structure file of the aligned residues and a score that indicates the root mean square difference (RMSD) between the aligned regions as well as the proportion of the protein that was aligned.

However, having a similar overall structure does not necessarily correlate directly to function. As with sequences, proteins can have the same fold yet perform completely different functions.<sup>50</sup> This is not too surprising since only ~8,000 distinct protein folds are predicted to exist.<sup>51,52</sup> On the other hand, proteins with completely different structures can have the same function.<sup>53</sup>

**1.1.4 Protein active-site to function.** Inferring a protein function from global sequence and structure comparisons has significant drawbacks. The molecular function of a protein primarily depends on the structural arrangement and chemical properties of the amino acids exposed on the surface of the protein and the resulting non-covalent interactions (binding) with other molecules, such as proteins, DNA/RNA, and small molecules. While a protein may experience numerous weak interactions with many different molecules, typically the molecule (often called a ligand) that maximizes the



number of interactions by having complementary chemistry and structure often defines the function of a protein. The region of the protein where the ligand binds is called the binding site or active-site and is often a cavity in the surface of the protein. Because the active-site of a protein is directly related to the function of the protein, the active-site is more evolutionarily stable than the rest of the protein sequence, which explains the difficulties of assigning function from global sequence or structure similarities.<sup>54</sup>

Several computational tools exist that attempt to locate and compare protein active-sites using various approaches: identifying common sequence motifs,<sup>55,56</sup> computational calculation of ligand “hot-spots”,<sup>57</sup> locating structural cavities/clefts,<sup>58-60</sup> or using evolutionary conservation.<sup>61-64</sup> These approaches typically work by using the information of known binding sites, which may be problematic when dealing with proteins that have unique functions that haven’t been characterized previously.

One approach to minimize ambiguity in locating the active-site is to identify a ligand that binds to the protein at the active-site.<sup>65-68</sup> Searching the entirety of chemical space ( $\sim 10^{60}$  compounds)<sup>69</sup> for a ligand that binds to the protein is obviously not feasible. Therefore, using a chemical library that is focused on functionally relevant compounds is important for this approach.<sup>68,70</sup> Despite a focused compound library, a significant number of compounds still need to be evaluated for binding activity. Two high-throughput screening approaches are commonly used in this situation: ligand affinity screens by nuclear magnetic resonance (NMR) spectroscopy and molecular docking/virtual screening. Analysis of the experimentally determined ligand binding site can then be utilized to infer a function based on sequence and structural similarities to other ligand binding sites.<sup>71,72</sup>

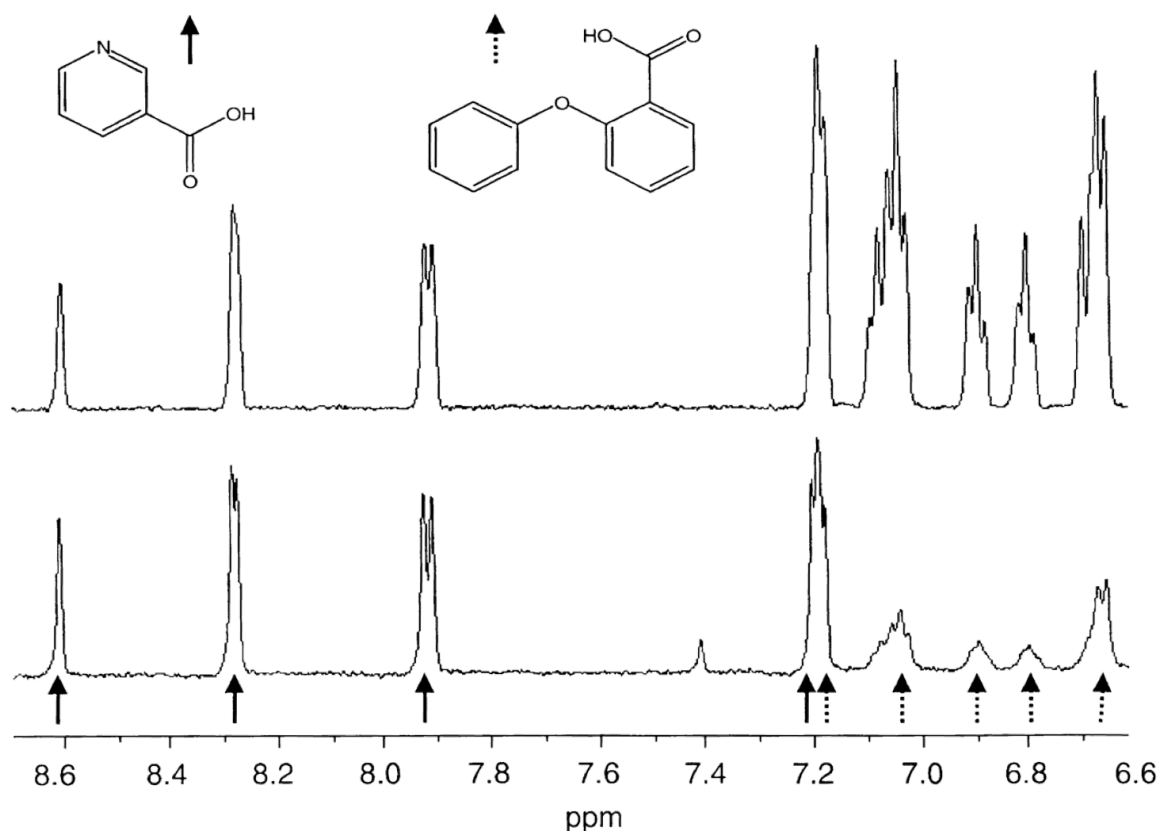
## 1.2 HIGH-THROUGHPUT SCREENING BY NMR

Nuclear magnetic resonance (NMR) spectroscopy is uniquely qualified to assist in the high-throughput functional annotation of proteins.<sup>73</sup> High-throughput screening by NMR is useful for several reasons: (1) it can directly detect the interaction between the ligand and protein using a variety of techniques; (2) samples are typically analyzed under native or near-native conditions; (3) hundreds of samples can be analyzed per day; and (4) information on the binding site and binding affinity can be readily obtained.

In high-throughput screening by NMR, a binding event is detected by the relative differences between the protein or ligand NMR spectrum in the bound and unbound states.<sup>74</sup> However, the specific type of information obtained about the binding process depends on whether a ligand-based or target-based NMR experiment is used.

**1.2.1 Ligand-based NMR screens.** Ligand-based NMR screens typically monitor the NMR spectrum of a ligand under free and bound conditions. Distinguishing between a free ligand and a protein-ligand costructure is generally based on the large molecular weight difference that affects several NMR parameters. Small molecular weight molecules have slow relaxation rates ( $R_2$ ), negative nuclear Overhauser effect (NOE) cross-peaks, and large translational diffusion coefficients ( $D_t$ ). If a protein-ligand binding event occurs, the ligand adopts the properties of the larger molecular-weight protein, increasing  $R_2$ , producing positive NOE cross-peaks, and decreasing  $D_2$ , all of which can be observed by NMR.<sup>75</sup> Most ligand-based NMR screens use one-dimensional (1D)  $^1\text{H}$ -NMR experiments to monitor these changes, which provide significant benefits for a high-throughput screen. 1D NMR experiments are typically fast (2-5 min) and routinely

use mixtures without the need to deconvolute.<sup>76</sup> The deconvolution of mixtures is avoided by ensuring the NMR ligand peaks do not overlap in the NMR spectrum [Figure 1.1]. The application of mixtures allows for hundreds to thousands of compounds to be screened in a single day. Another advantage of ligand-based NMR methods is the minimal amount of protein required ( $<10\ \mu\text{M}$ ) for each experiment. Additionally, isotopically labeled proteins are not needed for the NMR ligand affinity screen and protein molecular weight is not a limiting factor.<sup>77</sup> In fact, higher molecular-weight proteins enhance the observation of a binding event in a ligand-based NMR screen. All of these characteristics make ligand-based NMR screens a routinely used high-throughput screening technique.



**Figure 1.1** An example of the use of a ligand-detect NMR experiment to observe the line broadening (increase  $R_2$ ) that occurs when one compound, in a mixture of two compounds, binds a protein target. The <sup>1</sup>H-NOESY spectra of nicotinic acid (*left structure*) and 2-phenoxybenzoic acid (*right structure*) in a mixture without protein (*top spectrum*) and with the protein, p38 MAP kinase, added (*bottom spectrum*). The *solid* and *dashed arrows* represent the resonances of nicotinic acid and 2-phenoxybenzoic acid, respectively. In this case, the resonances corresponding to 2-phenoxybenzoic acid are broadened, indicating binding of this compound to the protein. (Reprinted with permission, copyright 2001 by Academic Press)<sup>78</sup>

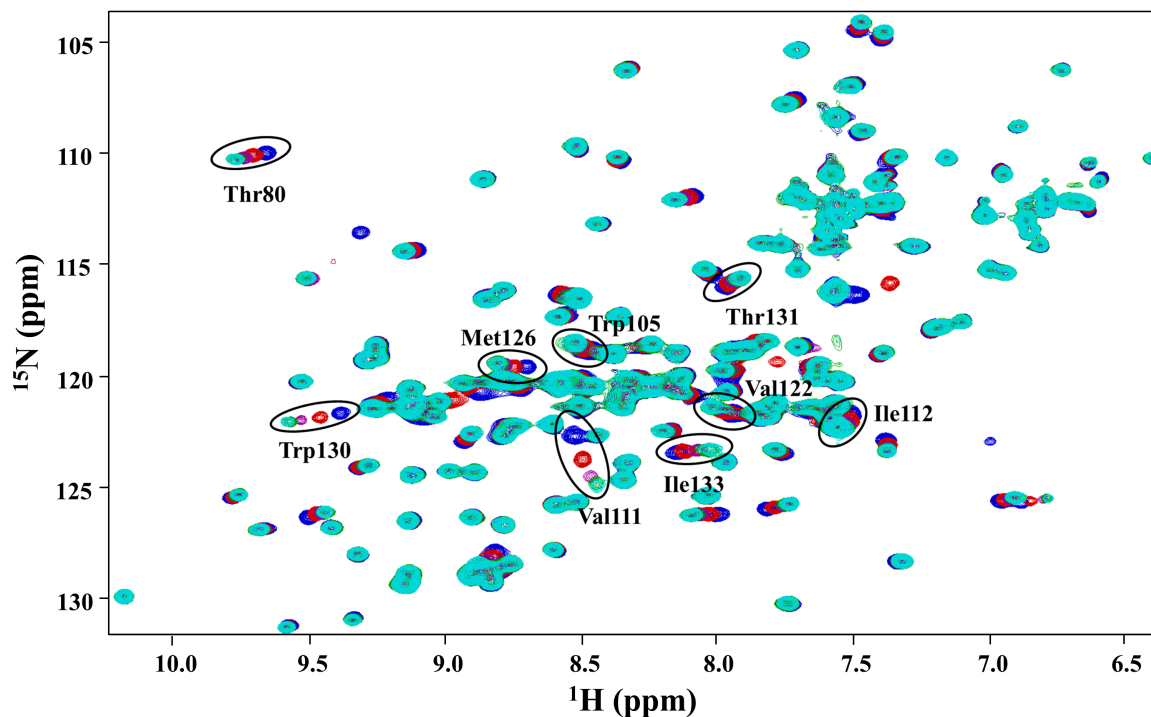
There are several screening techniques created from ligand-based NMR experiments: line broadening,<sup>79</sup> STD NMR,<sup>80</sup> WaterLOGSY,<sup>81</sup> SLAPSTIC,<sup>82</sup> TINS,<sup>83</sup> transferred NOEs,<sup>84</sup> FAXS,<sup>85,86</sup> FABS,<sup>87,88</sup> and diffusion measurements.<sup>89,90</sup> Each of these methods utilizes a specific NMR parameter that indicates ligand-binding, such as a

change in ligand NMR peak width or diffusion, a saturation transfer from the protein or solvent to the ligand, an NOE transfer between the free and bound ligand, a spin-label induced paramagnetic relaxation, or fluorine chemical shift anisotropy. The choice of which method to use typically depends upon the protein target and the compound library being screened. In addition, line broadening and STD, among other techniques, can be used to measure dissociation constants ( $K_D$ ).<sup>91,92</sup> On the other hand, ligand-based NMR screens don't provide any structural information about the protein-ligand complex.

**1.2.2 Target-based NMR screens.** A target-based screen focuses on changes in the protein (or other target) NMR spectrum to identify a binding event. Typically, chemical shift perturbations (CSPs) occur in the protein NMR spectrum upon ligand binding. The complexity and severe peak overlap in a protein 1D  $^1\text{H}$  NMR spectrum makes it impractical to observe subtle CSPs for weak binding ligands. Instead, two-dimensional (2D) heteronuclear NMR<sup>93-95</sup> experiments are typically used for target-based NMR ligand affinity screens.<sup>74</sup> 2D  $^1\text{H}$ - $^{13}\text{C}$ / $^{15}\text{N}$  HSQC/TROSY NMR experiments require a significant increase in experiment time (>10 min) due to the additional dimension and the need to collect a reference spectrum for the ligand-free protein. Also, the protein needs to be  $^{15}\text{N}$  and/or  $^{13}\text{C}$  isotopically labeled. However, 2D  $^1\text{H}$ - $^{13}\text{C}$ / $^{15}\text{N}$  HSQC/TROSY NMR experiments are useful because they provide additional information about the ligand binding site.

A binding ligand often results in the observation of CSPs of the resonances in a 2D  $^1\text{H}$ - $^{15}\text{N}$ - or  $^1\text{H}$ - $^{13}\text{C}$ -HSQC spectrum [Figure 1.2]. These CSPs are usually caused by a change in the chemical environment for residues proximal to the bound ligand or residues undergoing ligand-induced conformational changes. The availability of the protein

structure and the NMR sequence assignments (correlation of an NMR resonance with a specific amino acid residue) allows for the CSPs to be mapped onto a three-dimensional (3D) representation of the protein's surface. A cluster of residues on the protein surface with observed CSPs often identifies the ligand-binding site.

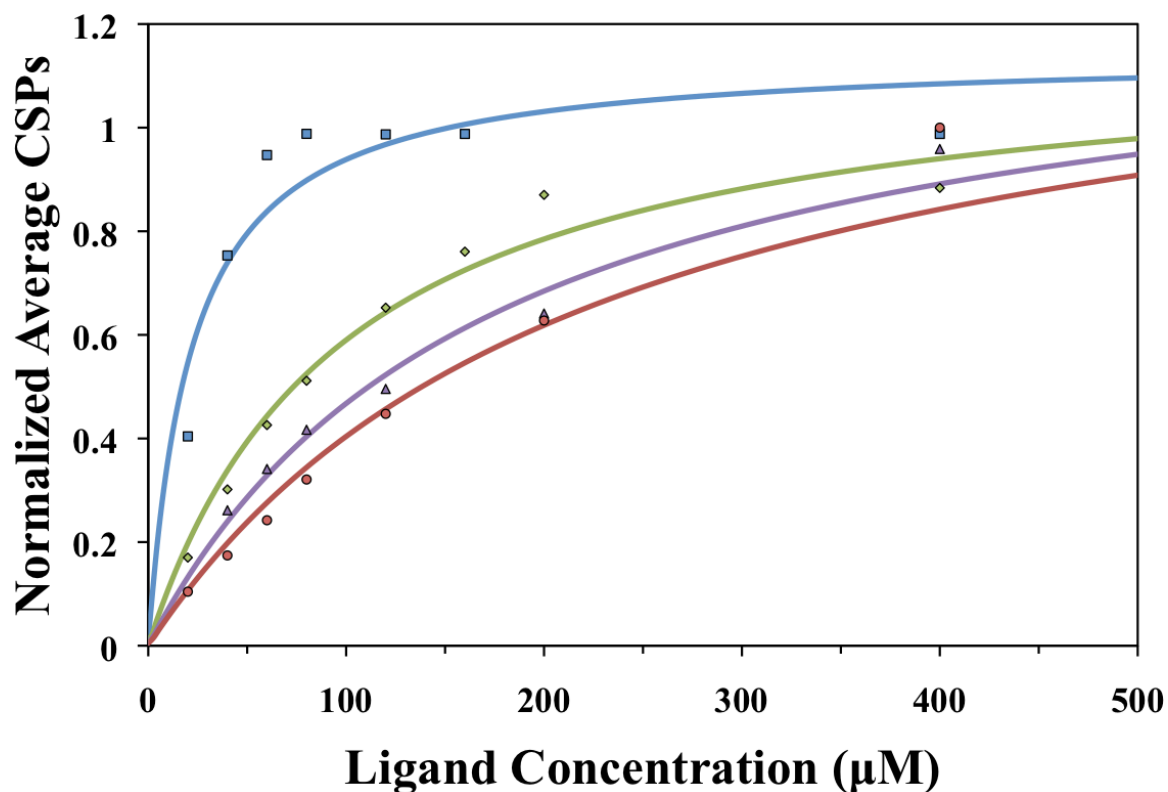


**Figure 1.2** An overlay of the 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectra for the protein YndB titrated with increasing amounts of chalcone. The perturbed residues can be used to identify a consensus binding site. (Reprinted with permission, copyright 2010 by John Wiley and Sons)<sup>68</sup>

The ligand binding affinity ( $K_D$ ) is also routinely determined from CSPs measured from a series of 2D  $^1\text{H}$ ,  $^{13}\text{C}/^{15}\text{N}$ -HSQC/TROSY NMR experiments. The magnitude of the CSPs at varying ligand concentrations is correlated to the  $K_D$  for the protein-ligand costructure using the following equation:<sup>96,97</sup>

$$\text{CSP}_{\text{obs}} = \text{CSP}_{\text{max}} \frac{(K_D + [L] + [P]) - \sqrt{(K_D + [L] + [P])^2 - 4([L][P])}}{2[P]} \quad (1.1)$$

where  $[P]$  is the protein concentration,  $[L]$  is the ligand concentration,  $\text{CSP}_{\text{max}}$  is the maximum CSP observed for a fully bound protein, and  $\text{CSP}_{\text{obs}}$  is the observed CSP at a particular ligand concentration. A least squares fit of Equation 1 to the experimental CSP data is used to calculate a  $K_D$  [Figure 1.3].



**Figure 1.3** NMR titration data for YndB bound to chalcone (*blue*), flavanone (*green*), flavone (*purple*), and flavanol (*orange*). The magnitude of the chemical shift perturbation can be used to calculate the dissociation constants for each compound. (Reprinted with permission, copyright 2010 John Wiley and Sons)<sup>68</sup>

As previously mentioned, because target-based screens require the use of multidimensional NMR experiments, data collection is significantly longer relative to ligand-based NMR screens. Also, target-based screens require higher protein concentrations (>50 μM compared to < 10 μM). This severely limits the utility of target-

based NMR screens for the high-throughput analysis of large compound libraries. Instead, the approach is typically used to validate hits from a high-throughput screen or the analysis of relatively small fragment-based libraries.<sup>98-100</sup> A fragment-based library consists of low molecular-weight compounds (<250-350 Da) that are fragments of known drugs or have drug-like properties.<sup>101</sup> Recent advances like the SOFAST-HMQC experiment<sup>102,103</sup> and the Fast-HSQC experiment<sup>104</sup> have decreased the time and amount of protein necessary for a target-based screen. Nevertheless, the combination of ligand-based and target-based NMR screens are still very resource intensive, requiring a significant amount of time and material. Also, because any high-throughput screen produces a significant amount of negative data (most ligands don't bind or inhibit a protein),<sup>105</sup> a more efficient approach is to screen a library of compounds with a higher probability of binding the protein target. In effect, a virtual or *in silico* screen can be used to enrich a library with likely binders.

### 1.3 MOLECULAR DOCKING AND VIRTUAL SCREENING

Molecular docking is a computational tool that predicts the binding site location and conformation of a compound when bound to a protein using only the knowledge of the structures of the protein and ligand.<sup>106-113</sup> This approach has been found to be fairly successful in redocking compounds into previously solved protein-ligand costructures,<sup>114</sup> where more than 70% of the redocked ligands reside within 2 Å RMSD of the actual ligand pose. During the prediction of protein-ligand costructures, molecular docking programs calculate a binding score that allows for the selection of the best ligand pose. The binding score is typically based on a combination of geometric and energetic



functions (bond lengths, dihedral angles, van der Waals forces, Lennard-Jones and electrostatic interactions, etc.) in conjunction with empirical functions unique to each specific docking program.<sup>115-120</sup>

These binding scores are also routinely used to rank different ligands from a compound library after being docked to a protein target. The virtual or *in silico* screening of a library composed of thousands of theoretical compounds can be accomplished in a day with minimal cost.<sup>121-123</sup> Thus, a virtual screen can significantly accelerate the hit identification and optimization process while reducing the amount of experimental effort. However, a virtual screen does have significant limitations that prevent it from completely replacing experimental high-throughput screening.<sup>108,124-126</sup> These limitations include inaccurate scoring functions, use of rigid proteins, and simplified solvation models. In essence, a virtual screen only increases the likelihood that a predicted ligand actually binds the protein target; experimental verification is still essential.

An accurate prediction of the interactions between two molecules requires an in-depth understanding of the energetics that led to a stable biomolecular complex. Unfortunately, a model that correctly accounts for all the factors involved in a productive protein-ligand interaction is currently unknown. Furthermore, the problem is exponentially more complex than just modeling the specifics of a protein-ligand interaction. A protein contains thousands of atoms that have specific interactions with each other, with the solvent, and with other ions in addition to the bound ligand. Because of this complexity, computational efforts that attempt to model protein-ligand interactions require significant amounts of processing power and time. Many efforts that utilize molecular dynamics and distributed computing<sup>127,128</sup> are generally limited to a detailed

analysis of a single system. These methods are generally not practical for the majority of researchers interested in conducting a virtual screen of a library containing upwards of millions of compounds. To make molecular docking computationally feasible and easily accessible, many simplifications and trade-offs in the process are necessary.

Many computer programs are available to perform or assist with molecular docking, such as: AutoDock,<sup>129</sup> DOCK,<sup>130</sup> FlexX,<sup>131</sup> Glide,<sup>132</sup> HADDOCK,<sup>133</sup> and LUDI.<sup>134,135</sup> Each docking program does have some unique features that make them particularly useful for a given situation or problem. However, nearly all the docking programs consist of two primary components: docking (or searching) and scoring.<sup>106,107</sup> Docking refers to the sampling of the ligand's conformation space and its orientation relative to a receptor. Scoring is used to evaluate and rank the current pose of the ligand.

**1.3.1 Docking.** The docking process requires, at a minimum, two inputs: the three-dimensional structures of the receptor (protein) and the ligand. The most common simplification to the docking process is to keep the structure of the receptor rigid and stationary. Only the ligand is typically allowed to be flexible as it is docked to the protein. Keeping the protein rigid significantly minimizes the complexity of the calculation. Sampling the conformations and orientations of the ligand is done using systematic or stochastic methods.<sup>106,107</sup>

Systematic search methods attempt to sample all of the possible conformations of a ligand by incrementing the torsional angles of each rotatable bond. Unfortunately, this technique is computationally expensive due to the exponential increase in the number of possible conformations ( $N_{\text{conf}}$ ) as the number of rotatable bonds increases:<sup>106</sup>

$$N_{\text{conf}} = \prod_{i=1}^N \prod_{j=1}^{n_{\text{inc}}} \frac{360}{\theta_{i,j}} \quad (1.2)$$

where  $N$  represents the number of rotatable bonds,  $n_{\text{inc}}$  is the number of incremental rotations for each rotatable bond, and  $\theta_{i,j}$  is the size of the incremental rotation for each rotatable bond. As a result, purely brute force systematic approaches are generally not used. Instead, most systematic searches require the use of efficient shortcuts.

Perhaps the most commonly utilized systematic search method is incremental construction, which is used by DOCK,<sup>130</sup> FlexX,<sup>131</sup> E-Novo,<sup>136</sup> LUDI,<sup>134,135</sup> ADAM,<sup>137</sup> and TrixX.<sup>138</sup> In this particular method, the ligand is split into fragments. The most rigid fragments are often used as the core or anchor and are docked first into the receptor binding pocket. The remaining fragments are incrementally added back onto the core fragment, where each addition is systematically rotated to evaluate the most optimal conformation. Thus, incremental construction drastically reduces the number of possible conformations that need to be searched in order to identify the optimal pose.

Another systematic approach uses rigid docking in combination with a predefined library of ligand conformations, which is implemented in OMEGA,<sup>139</sup> FLOG,<sup>140</sup> Glide,<sup>132</sup> and the TrixX Conformer Generator.<sup>141</sup> This technique generates several low energy conformers for a ligand that are clustered by RMSD. A representative conformer from each cluster is then docked into the receptor. The approach is very fast because the docking process keeps the ligand rigid, eliminating the need to spend computation time on searching torsional space. A tradeoff for this increase in speed is a potential loss in accuracy, because the binding potential for all possible conformers may not be explored. However, a major benefit of the technique is the fact that the library of structural

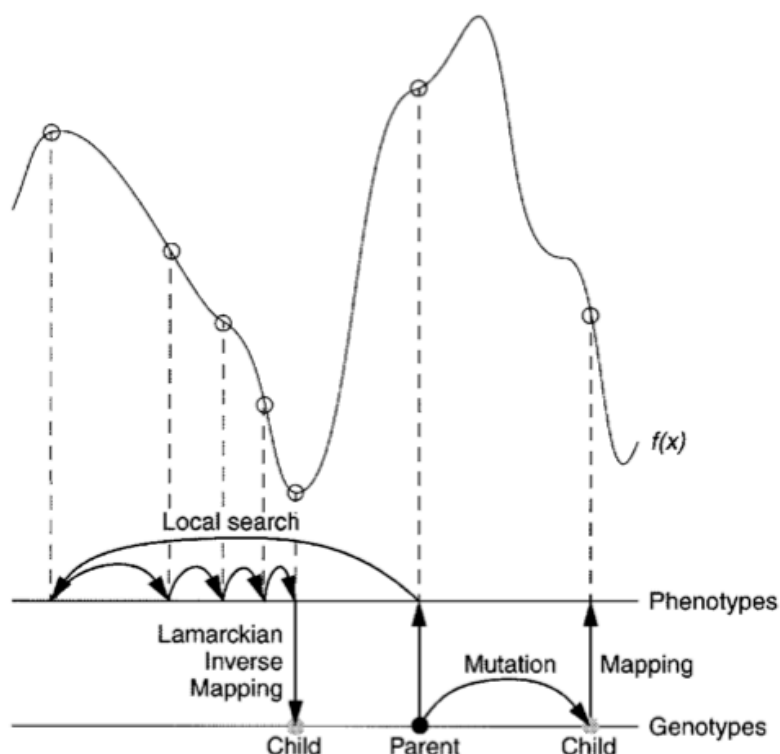
conformers only needs to be generated once. This is a significant savings in time for the pharmaceutical industry, where screening libraries may consist of millions of compounds.

Unlike systematic approaches that attempt to sample all possible ligand conformations, stochastic searches explore conformational space by making random torsional changes to a single ligand or a population of ligands. The structural changes are then evaluated using a probability function. There are three types of stochastic searches: Monte Carlo algorithms,<sup>142</sup> genetic algorithms,<sup>143</sup> and tabu search algorithms.<sup>144</sup> The most basic stochastic method is the Monte Carlo algorithm, which utilizes a Boltzmann probability function to determine whether to accept a particular ligand pose:<sup>145</sup>

$$P \sim \exp \left[ \frac{-(E_1 - E_0)}{K_B T} \right] \quad (1.3)$$

where  $P$  is the probability the conformation is accepted,  $E_0$  and  $E_1$  are the ligand's energy before and after the conformational change,  $K_B$  is the Boltzmann constant, and  $T$  is the absolute temperature. The simple scoring function used by the Monte Carlo algorithms is more effective than molecular dynamics in avoiding local minima and finding the global minimum.<sup>145</sup> Alternatively, genetic algorithms utilize the theory of evolution and natural selection to search ligand conformation space. In this case, the conformations, orientations, and coordinates of a ligand are encoded into variables representing a “genetic code.” A population of ligands with random genetic codes is allowed to evolve using mutations, crossovers, and migrations. The new population is evaluated using a fitness function that eliminates unfavorable ligand poses. Eventually, a final population converges to ligands with the most favorable “genes” or conformations [Figure 1.4]. Tabu searches, like other stochastic methods, randomly modify the conformation and coordinates of a ligand, score the conformer, and then repeat the process for a new

conformation. Tabu searches utilize a tabu list to remember previous ligand states. A pose is immediately rejected if it is close to a prior conformation. The tabu list encourages the search to progress to unexplored regions of conformational space.



**Figure 1.4** An illustration of the genetic algorithm approach, where the states of the ligand (translation, orientation, and conformation relative to the protein) are interpreted as the ligand genotype and the atomic coordinates represent the phenotype. A plot of the change in the fitness function ( $f(x)$ ) as the ligand population is allowed to mutate, crossover, and migrate. The genetic evolution of the ligand effectively samples conformational space where the best conformer is identified by a minimum in the fitness function. (Reprinted with permission, copyright 1998 by John Wiley and Sons)<sup>146</sup>

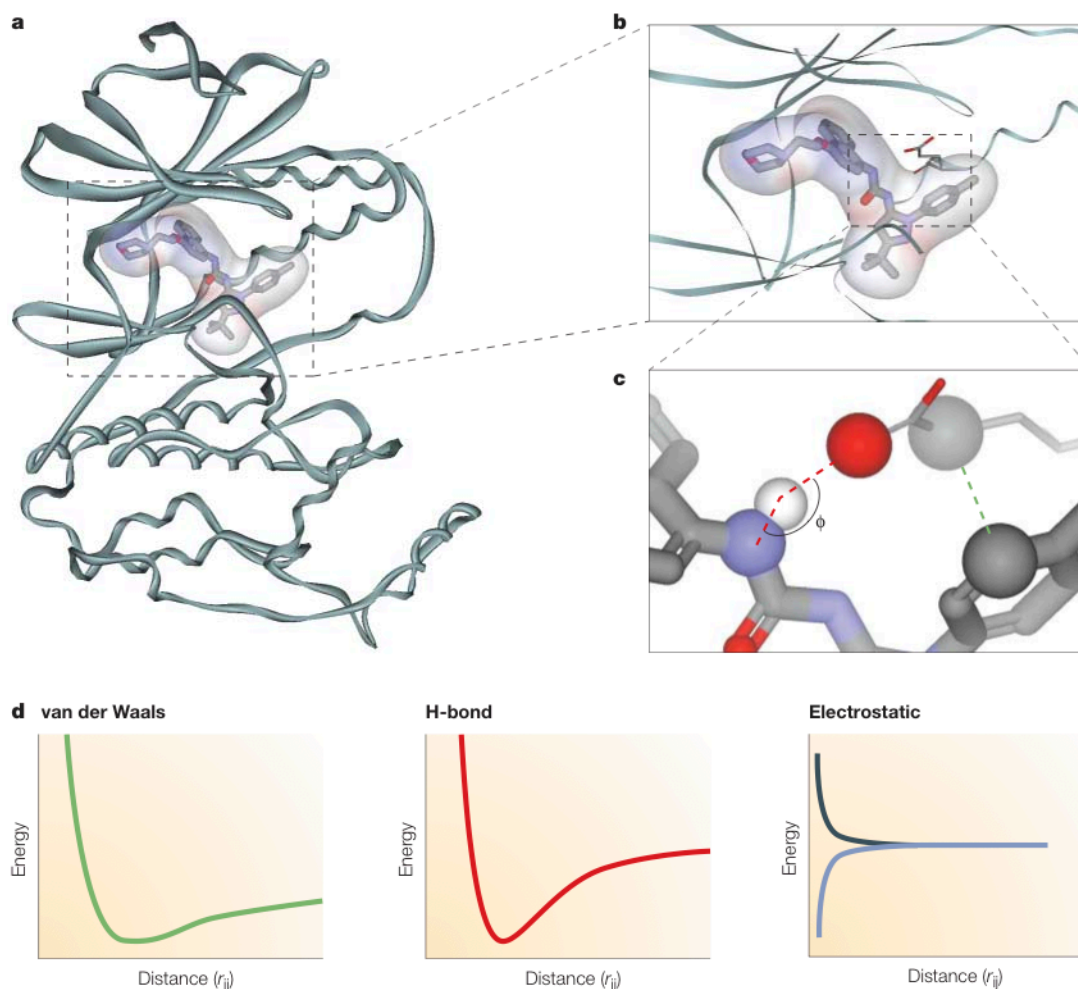
**1.3.2 Scoring.** While docking algorithms are generally efficient at generating the correct ligand pose, it is important for the docking program to actually select the correct ligand conformation from an ensemble of similar conformers. In essence, the scoring function should be able to distinguish between the true or optimal binding conformation and all the other poses. The scoring function is also used to rank the relative binding affinities for each compound in the library. Ideally, the scoring function should be able to

calculate the free energy ( $\Delta G_{\text{binding}}$ ) of the protein-ligand binding interaction, which is directly related to the dissociation constant ( $K_D$ ):<sup>96,106,147</sup>

$$\Delta G_{\text{binding}} = -RT \ln \frac{1}{K_D} \quad (1.4)$$

Unfortunately, accurately calculating the binding free energy is challenging due to the many forces that influence binding. In molecular docking, there are five primary types of scoring functions: force field-based, empirical, knowledge-based, shape-based, and consensus.<sup>148-150</sup>

Force field-based scoring functions<sup>106,107</sup> are used to calculate the free energy of binding by combining the receptor-ligand interaction energy and the change in internal energies of the ligand based on its bound conformation [Figure 1.5] The internal energy of the receptor is usually ignored because the receptor is kept rigid for most docking programs. The protein-ligand binding energies are typically defined by van der Waals forces, hydrogen bonding energies, and electrostatic energy terms. The van der Waals and hydrogen bonding terms often utilize a Lennard-Jones potential function, while the electrostatic terms are described by a coulombic function. Unfortunately, these interaction energies were originally derived from measuring enthalpic interactions in the gas phase.<sup>106</sup> Of course, protein-ligand binding interactions typically occur in an aqueous solution, which introduces additional interactions between the solvent molecules, the receptor, and the ligand. Protein-ligand binding energies are also dependent on the entropic changes that occur upon binding, which include torsional, vibrational, rotational, and translational entropies. Most entropy and solvation-based energy terms can't be calculated using force field-based scoring functions. As a result, force field-based scoring functions are incomplete and inaccurate.



**Figure 1.5** (a) A representation of p38 mitogen-activated protein kinase structure bound to BIRB796 and (b) an expanded view of the binding site. (c) A representation of the hydrogen-bonding (*red*) and electrostatic interactions (*green*) between the atoms of the protein and the atoms of the ligand. (d) A representation of three force-field energy terms (van der Waals, hydrogen-bonding, and electrostatic) as distance between the interacting atom pairs change. (Reprinted with permission, copyright 2004 by the Nature Publishing Group)<sup>106</sup>

Empirical scoring functions<sup>151-153</sup> are similar to force field-based scoring functions because they use a summation of individual energy terms. But empirical scoring functions also attempt to include solvation and entropic terms. This is typically achieved by using experimentally determined binding energies of known ligand-receptor



interactions to train the scoring system using regression analysis. Empirical scoring functions are fast, but the accuracy is completely dependent upon the experimental data set used to train the scoring function. In general, empirical scoring functions are reliable for ligand-receptor complexes that are similar to the training set.

Knowledge-based scoring functions<sup>154-156</sup> are fundamentally different from force field-based and empirical scoring functions. Knowledge-based scoring functions don't attempt to calculate the free energy of binding. Instead, these scoring functions utilize a sum of protein-ligand atom pair interaction potentials to calculate a binding affinity. The atom pair interaction potentials are generated based upon a probability distribution of interatomic distances found in known protein-ligand structures. The probability distributions are then converted into distance-dependent interaction energies. In this manner, knowledge-based scoring functions allow for the modeling of binding interactions that are not well understood. The approach is also simple, which is useful for screening large compound libraries. Unfortunately, knowledge-based scoring functions are designed to reproduce known experimental structures, and the binding score generated has little relevance to an actual binding affinity. This is an issue similar to empirical scoring functions; the accuracy of the scoring function is strongly dependent on the similarity of the protein-ligand costructure to the training data set.

As implied, shape-based scoring functions are based on a shape match between the ligand and the ligand binding site.<sup>157</sup> These scoring functions are typically used as prefilters to eliminate compounds that are unable to fit into the ligand binding site.<sup>68,158</sup> Shape-based scoring functions are fast, but are limited relative to more accurate scoring functions that calculate binding affinities. Shape-based scoring functions typically

generate smooth energy surfaces using Gaussian functions,<sup>158</sup> which are more tolerant to atomic variations and make protein clash interactions “softer.” This essentially helps minimize the effect of small structural variations that may occur during ligand binding.

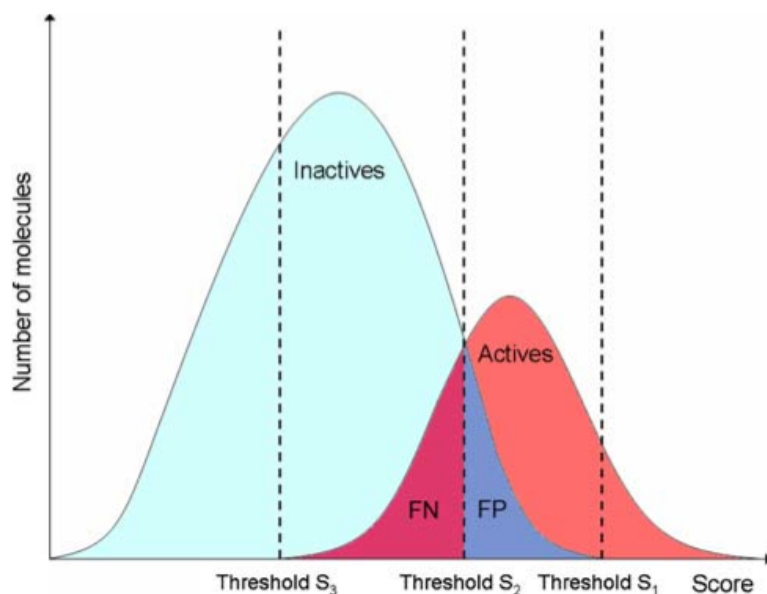
While the above scoring methods are generally useful in describing protein-ligand interactions, the simplifications used in each approach limits the overall accuracy in predicting the correct docked ligand pose.<sup>159,160</sup> The major weakness of most docking programs has been shown to be the scoring function. One approach to compensate for this deficiency is to use a consensus score from a combination of scoring functions to rescore a docked pose. Consensus scoring<sup>107,161</sup> has been shown in several examples to improve docking results compared to a single scoring function. However, like individual scoring functions, the improvement is not consistent and the proper choice of scoring functions to calculate a consensus score is typically based on trial and error.

**1.3.3 Virtual screening and assessment.** Using molecular docking to identify lead candidates is an attractive approach for both functional genomics and the pharmaceutical industry; it allows for the rapid evaluation of millions of chemical compounds while using minimal resources. The process by which molecular docking is used to rank different compounds within a library based on a predicted binding affinity is known as virtual screening.<sup>162,163</sup> A virtual screen requires a balance between optimizing speed and maximizing accuracy. Specifically, the goal of a virtual screen is the rapid and efficient separation of a small subset of active compounds from a relatively large random library of inactive compounds. Unfortunately, determining the effectiveness of a specific virtual screening process is challenging, where independent evaluators routinely generate inconsistent results.<sup>111,164-166</sup>

The ambiguous nature of the results from a virtual screen requires additional methods to evaluate its success. Typically, a virtual screening process is evaluated against a protein target with a set of known binders. Assessing the performance of a virtual screen is primarily based on the accuracy of the predicted ligand pose and binding affinity. The correct binding pose is often evaluated by calculating the RMSD between the docked and experimental ligand structures. The evaluation of binding affinity is typically based on the accurate ranking of known binders instead of the absolute scores because of the known limitations with calculating a binding energy. Other modes of performance assessment involve evaluating enrichment and generating diverse hit lists.

162,166

In a virtual screening protocol, every compound in a library ( $N_{\text{tot}}$ ) is docked to the protein and a corresponding binding score is calculated. The binding score for the ligand's best docked pose is used to rank the ligand relative to the entire library. A virtual screen never results in all the truly active compounds being top ranked.<sup>167</sup> Instead, most virtual screening protocols set a binding score or ranking threshold to identify the predicted active compounds or "hits." In general, top ranked compounds are expected to be enriched with active compounds compared to a random selection [Figure 1.6]



**Figure 1.6** A theoretical distribution of compounds in a virtual screen based upon the docking score. The overlap between active and inactive compounds indicates that the scoring threshold used to identify a hit by virtual screening is critical. (Reprinted with permission, copyright 2008 by Springer)<sup>166</sup>

A high enrichment factor ( $EF > 10$ ) is considered the benchmark of success for a virtual screen.<sup>168</sup> Enrichment is dependent on sensitivity ( $Se$ ) and specificity ( $Sp$ ). Sensitivity represents the true positive rate, which is the ratio of true positives ( $TP$ ) found by the virtual screening vs. the total number of actives ( $A$ ) in the library. The number of actives corresponds to both true positive ( $TP$ ) and false negative ( $FN$ ):<sup>166</sup>

$$Se = \frac{TP}{TP + FN} \quad (1.5)$$

Specificity is the measure of the true negative rate, which represents the ratio of true negatives ( $TN$ ) to the total number of inactive compounds. The number of inactive compounds corresponds to both true negatives ( $TN$ ) and false positives ( $FP$ ):<sup>166</sup>

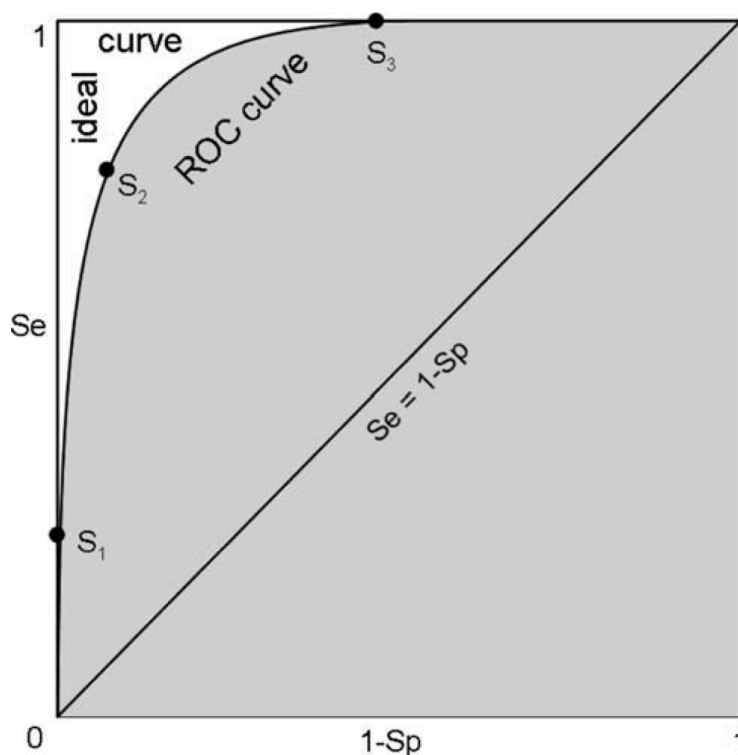
$$Sp = \frac{TN}{TN + FP} \quad (1.6)$$

The enrichment factor is a common method for evaluating the enrichment capabilities of a virtual screen.<sup>166</sup>

$$EF = \frac{\left(\frac{TP}{TP + FP}\right)}{\left(\frac{TP + FN}{N_{\text{tot}}}\right)} \quad (1.7)$$

The enrichment factor is dependent upon the ratio of active compounds to the total number of compounds in the library. As a result, enrichment scores are difficult to compare between virtual screens with different libraries. Also, the enrichment factor does not distinguish between high and low ranking compounds.

Perhaps the more popular approach for evaluating enrichment is to generate a receiver operating characteristic (ROC) curve.<sup>169</sup> The ROC curve is a plot of the true positive rate (*Se*) against the false positive rate (*1-Sp*) at varying thresholds for determining a hit. A ROC curve allows for the evaluation of a virtual screening method without using an arbitrary scoring threshold. Enrichment occurs when the resulting data point at a particular threshold resides above the diagonal (*Se* = *1-Sp*), which corresponds to a random selection of compounds. In a perfect virtual screen where every active compound is identified as a hit and every inactive compound falls below the threshold, the ROC curve approaches the top left corner (*Se* = 1 and *1-Sp* = 0) [Figure 1.7].



**Figure 1.7** A ROC curve is used to evaluate the enrichment of a virtual screen and select a scoring threshold. A ROC curve that approaches  $Se = 1$  and  $1-Sp = 0$  represents perfect enrichment. The area under the ROC curve (AUC) represents the probability that a true active is identified. (Reprinted with permission, copyright 2008 by Springer)<sup>166</sup>

Hit list diversity is also an important consideration for the success of a virtual screen because there is more value in identifying a few unique compounds instead of many compounds all based on the same chemical scaffold. One way that diversity can be determined is by comparing the structural similarities of hits from a virtual screen using a Tanimoto index<sup>170</sup> and then clustering the results. Basically, a Tanimoto index is calculated based on the fraction of similar chemical sub-structures present in two structures. Generally, 1,365 chemical substructures are used to describe a structure. The substructures include individual elements, two-atom substructures, single rings, condensed rings, aromatic rings, other rings, chains, branches, and functional groups:<sup>170</sup>

$$TI = \frac{C}{A + B + C} \quad (1.8)$$

where  $A$  represents the substructural features present in the first structure,  $B$  represents the substructural features present in the second structure, and  $C$  represents the substructural features common to both structures. Identical structures have a  $TI$  score of 1, where completely dissimilar structures have a  $TI$  value of 0.

## 1.4 SUMMARY OF WORK

Combining ligand affinity screens by NMR with molecular docking/virtual screens provides an efficient approach to probing protein-ligand interactions. This dissertation focuses on the application of both NMR and molecular docking/virtual screening towards the structural determination and functional annotation of proteins, which may lead to novel therapeutic targets for disease treatment.

Chapter 2 will illustrate the application of the Functional Annotation Screening Technology by NMR (FAST-NMR) towards the functional annotation of 21 hypothetical proteins from the Northeast Structural Genomics Consortium (NESG). FAST-NMR incorporates high-throughput NMR screens using both ligand-based and target-based approaches, which is followed by the use of molecular docking to generate protein-ligand costructures. These costructures, or more specifically the protein binding sites indicated by the location of the ligand relative to the protein, are then compared, by structure and sequence, to a database of binding sites using the Comparison of Active-site Similarity (CPASS) program. The function of the hypothetical protein is then inferred based upon the similarities of the binding site to those found in the database. In addition, the FAST-

NMR procedure has been updated and streamlined to accommodate the high-throughput analysis of these proteins and will be described here.

Chapter 3 and 4 will describe the development of the program, AutoDockFilter, that utilizes a post-filtering approach for rapidly (~35-45 min) identifying a protein-ligand costructure generated by the molecular docking program AutoDock that best agrees with the experimentally determined chemical shift perturbations obtained from a simple 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR screen. This tool is currently being used to generate the protein-ligand costructures used during the FAST-NMR process. Additionally, a companion program, CSP-Consensus, is described which attempts to remove some of the ambiguity in selecting a consensus binding site from the perturbed residues in a 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR experiment.

The CPASS program has been updated to version 2, which includes surface accessibility, ligand root-mean square difference, C $\beta$  distances, and implementation on the Open Science Grid (OSG; <http://www.opensciencegrid.com>). Chapter 5 will show the evaluation of the enrichment of functionally similar proteins in CPASS and the effect any structural differences in the binding site between apo- and holoproteins may have on the CPASS score. These investigations are then evaluated in the context of utility for the FAST-NMR project.

Chapter 6 illustrates the application of virtual screening to replace the experimental high-throughput screening methodology of FAST-NMR. In this particular case, the protein YndB from *Bacillus subtilis* was determined to have a hydrophobic cavity which indicates a likelihood of binding lipid-like molecules. Unfortunately, lipids are not well-represented in the FAST-NMR compound library due to their hydrophobic



nature. A virtual screen of a lipid library was performed on YndB, which resulted in the identification of three subclasses of flavonoids as being highly enriched: chalcones/hydroxychalcones, flavanones, and flavones/flavonols. Experimental 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR titrations of chalcone, flavanone, flavone, and flavonol all showed binding to YndB that mimicked the virtual screening ranking. Interestingly, these molecules have not been identified among the natural products of *Bacillus* organisms but are important precursors to plant antibiotics, which hints at a stress response function for YndB that is important for the symbiotic relationship between *B. subtilis* and plants.

Because virtual screening worked as a replacement for experimental high-throughput NMR screens as seen in the YndB example, could it also be used to replace or supplement the high-throughput screens used in FAST-NMR? Chapter 7 will compare the results of virtual screening the function-based compound library to the results of the experimental high-throughput screens of the same compounds.

Chapter 8 outlines the development of a pancreatic cancer ‘omics database and the process by which new therapeutic targets for pancreatic cancer might be discovered. From this database, an uncharacterized human protein DNAJA1 (DnaJ-like homolog family A member 1) was identified as being significantly down-regulated in the pancreatic cancer ‘omics studies. The solution structure of the J-domain of DNAJA1 was determined by NMR and then screened using the FAST-NMR process to identify a ligand-defined binding site.

## 1.5 REFERENCES

1. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).

2. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Pagani, I. *et al.* The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40**, D571–9 (2012).
4. Galperin, M. Y. & Koonin, E. V. From complete genome sequence to ‘complete’ understanding? *Trends Biotechnol* **28**, 398–406 (2010).
5. {International Union of Biochemistry and Molecular Biology Nomenclature Committee} & Webb, E. C. *Enzyme Nomenclature 1992*. (Academic Press, 1992).
6. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29 (2000).
7. Nadeau, J. H. *et al.* Sequence interpretation. Functional annotation of mouse genome sequences. *Science* **291**, 1251–1255 (2001).
8. Harrington, C. A., Rosenow, C. & Retief, J. Monitoring gene expression using DNA microarrays. *Curr Opin Microbiol* **3**, 285–291 (2000).
9. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836 (2000).
10. Phizicky, E. M. & Fields, S. Protein-protein interactions: methods for detection and analysis. *Microbiol Rev* **59**, 94–123 (1995).
11. Berggård, T., Linse, S. & James, P. Methods for the detection and analysis of protein-protein interactions. *Proteomics* **7**, 2833–2842 (2007).
12. Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* **36**, 5221–5231 (2008).
13. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).
14. Sun, H., Chattopadhyaya, S., Wang, J. & Yao, S. Q. Recent developments in microarray-based enzyme assays: from functional annotation to substrate/inhibitor fingerprinting. *Anal Bioanal Chem* **386**, 416–426 (2006).
15. Hannon, G. J. RNA interference. *Nature* **418**, 244–251 (2002).
16. Arenz, C. & Schepers, U. RNA interference: from an ancient mechanism to a state of the art therapeutic application? *Naturwissenschaften* **90**, 345–359 (2003).
17. del Val, C. *et al.* High-throughput protein analysis integrating bioinformatics and experimental assays. *Nucleic Acids Res* **32**, 742–748 (2004).
18. Joshi, T., Chen, Y., Becker, J. M., Alexandrov, N. & Xu, D. Genome-scale gene function prediction using multiple sources of high-throughput data in yeast *Saccharomyces cerevisiae*. *OMICS* **8**, 322–333 (2004).
19. Lee, Y.-H. *et al.* Gene knockdown by large circular antisense for high-throughput functional genomics. *Nat Biotechnol* **23**, 591–599 (2005).
20. Oude Elferink, R. One step further towards real high-throughput functional genomics. *Trends Biotechnol* **21**, 146–7– discussion 147–8 (2003).
21. Dimmer, E. C. *et al.* The UniProt-GO Annotation database in 2011. *Nucleic Acids Res* **40**, D565–70 (2012).
22. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **40**, D71–5 (2012).

23. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
24. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).
25. Bateman, A. *et al.* Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res* **27**, 260–262 (1999).
26. Karplus, K. *et al.* Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* **53 Suppl 6**, 491–496 (2003).
27. Karplus, K. *et al.* SAM-T04: what is new in protein-structure prediction for CASP6. *Proteins* **61 Suppl 7**, 135–142 (2005).
28. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680 (1994).
29. Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protocols Bioinformatics* 2.3.1–2.3.22 (2002).
30. Whisstock, J. C. & Lesk, A. M. Prediction of protein function from protein sequence and structure. *Q Rev Biophys* **36**, 307–340 (2003).
31. Theissen, G. Secret life of genes. *Nature* **415**, 741 (2002).
32. Wistow, G. & Piatigorsky, J. Recruitment of enzymes as lens structural proteins. *Science* **236**, 1554–1556 (1987).
33. Jeffery, C. J. Moonlighting proteins. *Trends Biochem Sci* **24**, 8–11 (1999).
34. Jeffery, C. J., Bahnson, B. J., Chien, W., Ringe, D. & Petsko, G. A. Crystal structure of rabbit phosphoglucose isomerase, a glycolytic enzyme that moonlights as neuroleukin, autocrine motility factor, and differentiation mediator. *Biochemistry* **39**, 955–964 (2000).
35. Valencia, A. Automatic annotation of protein function. *Curr Opin Struct Biol* **15**, 267–274 (2005).
36. Scapin, G. Structural Biology and Drug Discovery. *Curr Pharm Des* **12**, 2087–2097 (2006).
37. Powers, R. Applications of NMR to structure-based drug design in structural genomics. *J Struct Funct Genomics* **2**, 113–123 (2002).
38. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
39. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **10**, 980 (2003).
40. Stevens, R. C., Yokoyama, S. & Wilson, I. A. Global efforts in structural genomics. *Science* **294**, 89–92 (2001).
41. Mirkovic, N., Li, Z., Parnassa, A. & Murray, D. Strategies for high-throughput comparative modeling: applications to leverage analysis in structural genomics and protein family organization. *Proteins* **66**, 766–777 (2007).
42. Teichmann, S. A., Murzin, A. G. & Chothia, C. Determination of protein function, evolution and interactions by structural genomics. *Curr Opin Struct Biol* **11**, 354–363 (2001).
43. Shapiro, L. & Harris, T. Finding function through structural genomics. *Curr*

- Opin Biotechnol* **11**, 31–35 (2000).
44. Holm, L. & Sander, C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res* **25**, 231–234 (1997).
  45. Holm, L., Kaariainen, S., Rosenstrom, P. & Schenkel, A. Searching protein structure databases with DaliLite v.3. *Bioinformatics* **24**, 2780–2781 (2008).
  46. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D* **60**, 2256–2268 (2004).
  47. Ye, Y. & Godzik, A. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res* **32**, W582–5 (2004).
  48. Redfern, O. C., Harrison, A., Dallman, T., Pearl, F. M. G. & Orengo, C. A. CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput Biol* **3**, e232 (2007).
  49. Ortiz, A. R., Strauss, C. E. M. & Olmea, O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci* **11**, 2606–2621 (2002).
  50. Rost, B. Enzyme function less conserved than anticipated. *J Mol Biol* **318**, 595–608 (2002).
  51. Liu, X., Fan, K. & Wang, W. The number of protein folds and their distribution over families in nature. *Proteins* **54**, 491–499 (2004).
  52. Neumann, S., Hartmann, H., Martin-Galiano, A. J., Fuchs, A. & Frishman, D. Camps 2.0: exploring the sequence and structure space of prokaryotic, eukaryotic, and viral membrane proteins. *Proteins* **80**, 839–857 (2012).
  53. Galperin, M. Y., Walker, D. R. & Koonin, E. V. Analogous enzymes: independent inventions in enzyme evolution. *Genome Research* **8**, 779–790 (1998).
  54. Gerlt, J. A. & Babbitt, P. C. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem* **70**, 209–246 (2001).
  55. Attwood, T. K. The quest to deduce protein function from sequence: the role of pattern databases. *Int J Biochem Cell Biol* **32**, 139–155 (2000).
  56. Sigrist, C. J. A. *et al.* PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* **38**, D161–6 (2010).
  57. Tuncbag, N., Gursoy, A. & Keskin, O. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics* **25**, 1513–1520 (2009).
  58. Glaser, F., Morris, R. J., Najmanovich, R. J., Laskowski, R. A. & Thornton, J. M. A method for localizing ligand binding pockets in protein structures. *Proteins* **62**, 479–488 (2006).
  59. Davis, I. W., Raha, K., Head, M. S. & Baker, D. Blind docking of pharmaceutically relevant compounds using RosettaLigand. *Protein Sci* **18**, 1998–2002 (2009).
  60. Dundas, J. *et al.* CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res* **34**, W116–8 (2006).

61. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* **38**, W529–33 (2010).
62. Innis, C. A. siteFiNDER|3D: a web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Res* **35**, W489–94 (2007).
63. Yao, H., Mihalek, I. & Lichtarge, O. Rank information: a structure-independent measure of evolutionary trace quality that improves identification of protein functional sites. *Proteins* **65**, 111–123 (2006).
64. Mihalek, I., Res, I. & Lichtarge, O. Evolutionary and structural feedback on selection of sequences for comparative analysis of proteins. *Proteins* **63**, 87–99 (2006).
65. Campbell, S. J., Gold, N. D., Jackson, R. M. & Westhead, D. R. Ligand binding: functional site location, similarity and docking. *Curr Opin Struct Biol* **13**, 389–395 (2003).
66. Mercier, K. A. *et al.* FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* **128**, 15292–15299 (2006).
67. Powers, R., Mercier, K. A. & Copeland, J. C. The application of FAST-NMR for the identification of novel drug discovery targets. *Drug Discov Today* **13**, 172–179 (2008).
68. Stark, J. L. *et al.* Solution structure and function of YndB, an AHSA1 protein from *Bacillus subtilis*. *Proteins* **78**, 3328–3340 (2010).
69. Fink, T. & Reymond, J.-L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J Chem Inf Model* **47**, 342–353 (2007).
70. Mercier, K. A., Germer, K. & Powers, R. Design and characterization of a functional library for NMR screening against novel protein targets. *Comb Chem High Throughput Screen* **9**, 515–534 (2006).
71. Powers, R. *et al.* Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **65**, 124–135 (2006).
72. Powers, R., Copeland, J. & Stark, J. L. Searching the protein structure database for ligand-binding site similarities using CPASS v. 2. *BMC Res Notes* (2011).
73. Powers, R. Advances in Nuclear Magnetic Resonance for Drug Discovery. *Expert Opin Drug Discov* **4**, 1077–1098 (2009).
74. Shuker, S. B., Hajduk, P. J., Meadows, R. P. & Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **274**, 1531–1534 (1996).
75. Lepre, C. A., Moore, J. M. & Peng, J. W. Theory and applications of NMR-based screening in pharmaceutical research. *Chem Rev* **104**, 3641–3676 (2004).
76. Mercier, K. A. & Powers, R. Determining the optimal size of small molecule mixtures for high throughput NMR screening. *J Biomol NMR* **31**, 243–258 (2005).
77. Zartler, E. R. & Mo, H. Practical aspects of NMR-based fragment discovery. *Curr Top Med Chem* **7**, 1592–1599 (2007).
78. Peng, J. W., Lepre, C. A., Fejzo, J., Abdul-Manan, N. & Moore, J. M. Nuclear magnetic resonance-based approaches for lead generation in drug discovery.

- Methods Enzymol* **338**, 202–230 (2001).
79. Hajduk, P. J., Olejniczak, E. T. & Fesik, S. W. One-dimensional relaxation-and diffusion-edited NMR methods for screening compounds that bind to macromolecules. *J Am Chem Soc* **119**, 12257–12261 (1997).
  80. Mayer, M. & Meyer, B. Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angew Chem Int Edit* **38**, 1784–1788 (1999).
  81. Dalvit, C. *et al.* Identification of compounds with binding affinity to proteins via magnetization transfer from bulk water. *J Biomol NMR* **18**, 65–68 (2000).
  82. Jahnke, W., Rudisser, S. & Zurini, M. Spin label enhanced NMR screening. *J Am Chem Soc* **123**, 3149–3150 (2001).
  83. Vanwetswinkel, S. *et al.* TINS, target immobilized NMR screening: an efficient and sensitive method for ligand discovery. *Chem Biol* **12**, 207–216 (2005).
  84. Fejzo, J. *et al.* The SHAPES strategy: an NMR-based approach for lead generation in drug discovery. *Chem Biol* **6**, 755–769 (1999).
  85. Dalvit, C., Flocco, M., Veronesi, M. & Stockman, B. J. Fluorine-NMR competition binding experiments for high-throughput screening of large compound mixtures. *Comb Chem High Throughput Screen* **5**, 605–611 (2002).
  86. Dalvit, C., Fagerness, P. E., Hadden, D. T. A., Sarver, R. W. & Stockman, B. J. Fluorine-NMR experiments for high-throughput screening: theoretical aspects, practical considerations, and range of applicability. *J Am Chem Soc* **125**, 7696–7703 (2003).
  87. Dalvit, C. *et al.* A general NMR method for rapid, efficient, and reliable biochemical screening. *J Am Chem Soc* **125**, 14620–14625 (2003).
  88. Dalvit, C., Ardini, E., Fogliatto, G. P., Mongelli, N. & Veronesi, M. Reliable high-throughput functional screening with 3-FABS. *Drug Discov Today* **9**, 595–602 (2004).
  89. Price, W. S. Pulsed-field gradient nuclear magnetic resonance as a tool for studying translational diffusion: Part 1. Basic theory. *Concepts Mag Res* **9**, 299–336 (1997).
  90. Price, W. S. Pulsed-field gradient nuclear magnetic resonance as a tool for studying translational diffusion: Part 2. Experimental aspects. *Concepts Mag Res* **10**, 197–237 (1998).
  91. Shortridge, M. D., Hage, D. S., Harbison, G. S. & Powers, R. Estimating protein-ligand binding affinity using high-throughput screening by NMR. *J Comb Chem* **10**, 948–958 (2008).
  92. Ji, Z., Yao, Z. & Liu, M. Saturation transfer difference nuclear magnetic resonance study on the specific binding of ligand to protein. *Anal Biochem* **385**, 380–382 (2009).
  93. Muhandiram, D. R., Farrow, N. A., Xu, G.-Y., Smallcombe, S. H. & Kay, L. E. A gradient <sup>13</sup>C NOESY-HSQC experiment for recording NOESY spectra of <sup>13</sup>C-labeled proteins dissolved in H<sub>2</sub>O. *J Magn Reson B* **102**, 317–321 (1993).
  94. Sklenar, V., Piotto, M., Leppik, R. & Saudek, V. Gradient-tailored water suppression for <sup>1</sup>H-<sup>15</sup>N HSQC experiments optimized to retain full sensitivity. *J Magn Reson A* **102**, 241–245 (1993).
  95. Pervushin, K., Riek, R., Wider, G. & Wüthrich, K. Attenuated T<sub>2</sub> relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy

- indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Natl Acad Sci USA* **94**, 12366–12371 (1997).
96. Fielding, L. NMR methods for the determination of protein-ligand dissociation constants. *Prog Nucl Mag Res Spect* **51**, 219–242 (2007).
  97. Morton, C. J. *et al.* Solution structure and peptide binding of the SH3 domain from human Fyn. *Structure* **4**, 705–714 (1996).
  98. Stoll, F. Library Design. *CHIMIA* **57**, 224–228 (2003).
  99. Erlanson, D. A., McDowell, R. S. & O'Brien, T. Fragment-based drug discovery. *J Med Chem* **47**, 3463–3482 (2004).
  100. Siegal, G., Ab, E. & Schultz, J. Integration of fragment screening and library design. *Drug Discov Today* **12**, 1032–1039 (2007).
  101. Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today Technologies* **1**, 337–341 (2004).
  102. Schanda, P., Kupce, E. & Brutscher, B. SOFAST-HMQC experiments for recording two-dimensional heteronuclear correlation spectra of proteins within a few seconds. *J Biomol NMR* **33**, 199–211 (2005).
  103. Schanda, P. & Brutscher, B. Hadamard frequency-encoded SOFAST-HMQC for ultrafast two-dimensional protein NMR. *J Magn Reson* **178**, 334–339 (2006).
  104. Mori, S., Abeygunawardana, C., Johnson, M. O. & van Zijl, P. C. Improved sensitivity of HSQC spectra of exchanging protons at short interscan delays using a new fast HSQC (FHSQC) detection scheme that avoids water saturation. *J Magn Reson B* **108**, 94–98 (1995).
  105. Lahana, R. How many leads from HTS? *Drug Discov Today* **4**, 447–448 (1999).
  106. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **3**, 935–949 (2004).
  107. Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409–443 (2002).
  108. Warren, G. L. *et al.* A critical assessment of docking programs and scoring functions. *J Med Chem* **49**, 5912–5931 (2006).
  109. Kuntz, I. D., Meng, E. C. & Shoichet, B. K. Structure-based molecular design. *Accounts Chem Res* **27**, 117–123 (1994).
  110. Krovat, E. M., Steindl, T. & Langer, T. Recent Advances in Docking and Scoring. *Curr Comput Aided Drug Des* **1**, 93–102 (2005).
  111. Cole, J. C., Murray, C. W., Nissink, J. W. M., Taylor, R. D. & Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins* **60**, 325–332 (2005).
  112. Wandzik, I. Current molecular docking tools and comparisons thereof. *Comm Mathematical Comp Chem* **55**, 271–278 (2006).
  113. Dias, R. & de Azevedo, W. F. Molecular docking algorithms. *Curr Drug Targets* **9**, 1040–1047 (2008).
  114. Hartshorn, M. J. *et al.* Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem* **50**, 726–741 (2007).
  115. Jacobsson, M., Lidén, P., Stjernschantz, E., Boström, H. & Norinder, U. Improving structure-based virtual screening by multivariate analysis of scoring

- data. *J Med Chem* **46**, 5781–5789 (2003).
116. Loving, K., Salam, N. K. & Sherman, W. Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation. *J Comput Aided Mol Des* **23**, 541–554 (2009).
  117. Salaniwal, S., Manas, E. S., Alvarez, J. C. & Unwalla, R. J. Critical evaluation of methods to incorporate entropy loss upon binding in high-throughput docking. *Proteins* **66**, 422–435 (2007).
  118. Vasilyev, V. & Bliznyuk, A. Application of semiempirical quantum chemical methods as a scoring function in docking. *Theor Chem Acc* **112**, (2004).
  119. Wei, D., Zheng, H., Su, N., Deng, M. & Lai, L. Binding energy landscape analysis helps to discriminate true hits from high-scoring decoys in virtual screening. *J Chem Inf Model* **50**, 1855–1864 (2010).
  120. Zawodszky, M. I., Stumpff-Kane, A. W., Lee, D. J. & Feig, M. Scoring confidence index: statistical evaluation of ligand binding mode predictions. *J Comput Aided Mol Des* **23**, 289–299 (2009).
  121. Cerqueira, N. M. F. S. A., Sousa, S. F., Fernandes, P. A. & Ramos, M. J. Virtual screening of compound libraries. *Methods Mol Biol* **572**, 57–70 (2009).
  122. Ripphausen, P., Nisius, B., Peltason, L. & Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J Med Chem* **53**, 8461–8467 (2010).
  123. Sousa, S. F., Cerqueira, N. M. F. S. A., Fernandes, P. A. & Ramos, M. J. Virtual screening in drug design and development. *Comb Chem High Throughput Screen* **13**, 442–453 (2010).
  124. Eckert, H. & Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov Today* **12**, 225–233 (2007).
  125. Merz, K. M. Limits of Free Energy Computation for Protein-Ligand Interactions. *J Chem Theory Comput* **6**, 1018–1027 (2010).
  126. Proschak, E., Rupp, M., Derksen, S. & Schneider, G. Shapelets: possibilities and limitations of shape-based virtual screening. *J Comput Chem* **29**, 108–114 (2008).
  127. Taufer, M., Crowley, M., Price, D. J., Chien, A. A. & Brooks, C. L., III. Study of an accurate and fast protein-ligand docking algorithm based on molecular dynamics. *Concurrency and Computation: Practice and Experience* **17**, 1627–1641 (2005).
  128. Garcia-Sosa, A. T., Sild, S. & Maran, U. Docking and virtual screening using distributed grid technology. *QSAR Comb Sci* **28**, 815–821 (2009).
  129. Morris, G. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* (2009). doi:10.1002/jcc.21256
  130. Ewing, T. J. A., Makino, S., Skillman, A. G. & Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* **15**, 411–428 (2001).
  131. Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **261**, 470–489 (1996).
  132. Friesner, R. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* **47**, 1739–



- 1749 (2004).
133. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **125**, 1731–1737 (2003).
  134. Bohm, H. J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J Comput Aided Mol Des* **6**, 61–78 (1992).
  135. Bohm, H. J. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J Comput Aided Mol Des* **6**, 593–606 (1992).
  136. Pearce, B. C., Langley, D. R., Kang, J., Huang, H. & Kulkarni, A. E-novo: an automated workflow for efficient structure-based lead optimization. *J Chem Inf Model* **49**, 1797–1809 (2009).
  137. Mizutani, M. Y., Tomioka, N. & Itai, A. Rational automatic search method for stable docking models of protein and ligand. *J Mol Biol* **243**, 310–326 (1994).
  138. Schlosser, J. & Rarey, M. Beyond the virtual screening paradigm: structure-based searching for new lead compounds. *J Chem Inf Model* **49**, 800–809 (2009).
  139. Boström, J., Greenwood, J. R. & Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J Mol Graph Model* **21**, 449–462 (2003).
  140. Miller, M. D., Kearsley, S. K., Underwood, D. J. & Sheridan, R. P. FLOG: a system to select ‘quasi-flexible’ ligands complementary to a receptor of known three-dimensional structure. *J Comput Aided Mol Des* **8**, 153–174 (1994).
  141. Griewel, A., Kayser, O., Schlosser, J. & Rarey, M. Conformational sampling for large-scale virtual screening: accuracy versus ensemble size. *J Chem Inf Model* **49**, 2303–2311 (2009).
  142. Hart, T. N. & Read, R. J. A multiple-start Monte Carlo docking method. *Proteins* **13**, 206–222 (1992).
  143. Fuhrmann, J., Rurainski, A., Lenhof, H.-P. & Neumann, D. A new Lamarckian genetic algorithm for flexible ligand-receptor docking. *J Comput Chem* **31**, 1911–1918 (2010).
  144. Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R. & Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **33**, 367–382 (1998).
  145. Burt, S., Hutchins, C. & Zielinski, P. J. A Monte Carlo method for finding important ligand fragments from receptor data. *J Comput Aided Mol Des* **11**, 243–255 (1997).
  146. Morris, G., Goodsell, D., Halliday, R. & Huey, R. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**, 1639–1662 (1998).
  147. Fielding, L. NMR methods for the determination of protein-ligand dissociation constants. *Curr Top Med Chem* **3**, 39–53 (2003).
  148. Huang, S.-Y. & Zou, X. Advances and challenges in protein-ligand docking. *Int J Mol Sci* **11**, 3016–3034 (2010).
  149. Huang, S.-Y., Grinter, S. Z. & Zou, X. Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Phys Chem Chem Phys* **12**, 12899–12908 (2010).
  150. Huang, S.-Y. & Zou, X. in *Annual Reports in Computational Chemistry* **6**, 280–

- 296 (Elsevier, 2010).
151. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. & Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des* **11**, 425–445 (1997).
  152. Tao, P. & Lai, L. Protein ligand docking based on empirical method for binding affinity estimation. *J Comput Aided Mol Des* **15**, 429–446 (2001).
  153. Wang, R., Lai, L. & Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des* **16**, 11–26 (2002).
  154. Muegge, I. & Martin, Y. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem* **42**, 791–804 (1999).
  155. Gohlke, H., Hendlich, M. & Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* **295**, 337–356 (2000).
  156. Velec, H. F. G., Gohlke, H. & Klebe, G. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J Med Chem* **48**, 6296–6303 (2005).
  157. Kortagere, S., Krasowski, M. & Ekins, S. The importance of discerning shape in molecular pharmacology. *Trends Pharmacol Sci* (2009). doi:10.1016/j.tips.2008.12.001
  158. McGann, M., Almond, H., Nicholls, A., Grant, J. & Brown, F. Gaussian docking functions. *Biopolymers* **68**, 76–90 (2003).
  159. Merlitz, H., Herges, T. & Wenzel, W. Fluctuation analysis and accuracy of a large-scale in silico screen. *J Comput Chem* **25**, 1568–1575 (2004).
  160. Tirado-Rives, J. & Jorgensen, W. L. Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. *J Med Chem* **49**, 5880–5884 (2006).
  161. Charifson, P. S., Corkery, J. J., Murcko, M. A. & Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* **42**, 5100–5109 (1999).
  162. Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today* **11**, 580–594 (2006).
  163. Schneider, G. & Böhm, H.-J. Virtual screening and fast automated docking methods. *Drug Discov Today* **7**, 64–70 (2002).
  164. Chen, H., Lyne, P. D., Giordanetto, F., Lovell, T. & Li, J. On evaluating molecular-docking methods for pose prediction and enrichment factors. *J Chem Inf Model* **46**, 401–415 (2006).
  165. Kontoyianni, M., McClellan, L. M. & Sokol, G. S. Evaluation of docking performance: comparative data on docking algorithms. *J Med Chem* **47**, 558–565 (2004).
  166. Kirchmair, J., Markt, P., Distinto, S., Wolber, G. & Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection--what can we learn from earlier mistakes? *J Comput Aided Mol Des* **22**, 213–228 (2008).
  167. Scior, T. *et al.* Recognizing Pitfalls in Virtual Screening: A Critical Review. *J*

- Chem Inf Model* (2012). doi:10.1021/ci200528d
168. Bender, A. & Glen, R. C. A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *J Chem Inf Model* **45**, 1369–1375 (2005).
  169. Truchon, J.-F. & Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the ‘early recognition’ problem. *J Chem Inf Model* **47**, 488–508 (2007).
  170. Scsibraný, H., Karlovits, M., Demuth, W., Müller, F. & Varmuza, K. Clustering and similarity of chemical structures represented by binary substructure descriptors. *Chemometrics and Intelligent Laboratory Systems* **67**, 95–108 (2003).

## CHAPTER 2

### THE FUNCTIONAL ANNOTATION OF HYPOTHETICAL PROTEINS FROM THE NORTHEAST STRUCTURAL GENOMICS CONSORTIUM<sup>‡</sup>

#### 2.1 INTRODUCTION

With over 27 million protein sequences known, determining the functional role of each protein represents an important opportunity to identify new therapeutic targets for drug discovery.<sup>1</sup> The vast majority of these proteins are functionally annotated using sequence and/or structural homology to proteins of known function. However, the success of the various genome projects and structural genomics consortia is adding new protein sequences and structures, ~40% of which are often classified as "putative", "uncharacterized", or "unknown" proteins. Unfortunately, uncharacterized proteins are often "orphaned", because detailed and time intensive biochemical studies are required to functionally annotate a protein. These studies may include analyzing cell phenotypes through knockout libraries, monitoring gene expression levels, or utilizing pull-down assays.<sup>2-5</sup> Nevertheless, "orphaned" proteins provide a unique opportunity to explore functional space and identify new biological targets for drug discovery. Therefore, developing methods to rapidly and accurately assign a function to uncharacterized proteins is of great importance.

Because the interactions of proteins with other biomolecules or small molecules is the basis of a functional definition or classification, identifying the functional ligand, the functional epitope or ligand binding site, and the 3D structure of the protein-ligand costructure are invaluable for a functional annotation. A functional epitope or ligand

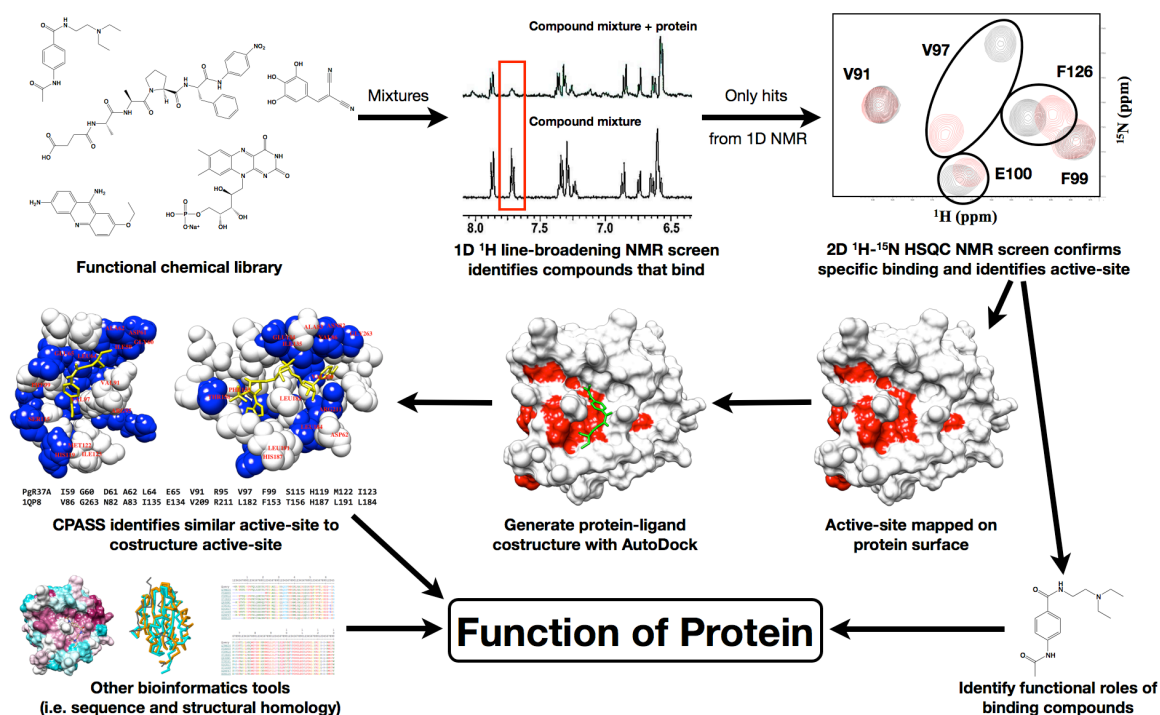
---

<sup>‡</sup> Jennifer Copeland contributed equally to all FAST-NMR work

binding site is evolutionarily conserved relative to the rest of the protein structure in order for the protein to maintain its biological function. Therefore, proteins that share similar binding site structures are expected to be functional homologs and bind a similar set of ligands.<sup>6,7</sup> Correspondingly, numerous *in silico* approaches attempt to infer a function for an uncharacterized protein by predicting ligand binding sites using geometry-based, information-based, and energy-based algorithms.<sup>8-10</sup> Unfortunately, unambiguously identifying the ligand binding site on a protein can be challenging without experimental evidence, especially for proteins with no known function.

Functional Annotation Screening Technology using NMR (FAST-NMR)<sup>6,7</sup> is an experimental approach that combines high throughput screening (HTS) by NMR with molecular docking and bioinformatics analysis in order to assign a function to a protein [Figure 2.1]. In this process, a compound library that contains approximately 460 biologically relevant compounds<sup>11</sup> is screened by NMR using a multistep approach.<sup>12</sup> First, a ligand-based screen using 1D <sup>1</sup>H-NMR line-broadening experiments identifies potential binders. These hits are then verified in a target-based screen using a 2D <sup>1</sup>H, <sup>15</sup>N-HSQC experiment, where the occurrence of chemical shift perturbations (CSPs) allows for the identification of the ligand binding site. Molecular docking is then used to generate a rapid protein-ligand costructure<sup>13</sup> that serves as input for the Comparison of Protein Active-Site Structures (CPASS) program.<sup>14,15</sup> CPASS compares the sequence and structure of this NMR modeled ligand binding site to ~36,000 unique experimental ligand binding sites from the RCSB Protein Databank.<sup>1</sup> Thus, the function of an uncharacterized protein can be inferred due to a similarity in binding sites to a protein with a known function that shares a similar ligand binding site.<sup>16</sup> Additionally, information from the

identity of the ligands shown to bind in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen, such as other proteins known to bind the ligand or metabolic pathways the ligand is involved with, can be used to help characterize the protein. Combining the experimental results of the FAST-NMR screen with available bioinformatics allows for a proposed function to be assigned to the protein.



**Figure 2.1** Flow chart for the application of FAST-NMR.

The FAST-NMR and CPASS approach has been used previously for the successful annotation of two hypothetical proteins, SAV1430 from *S. aureus*<sup>7</sup> and PA1324 from *P. aeruginosa*.<sup>17</sup> It has also been used to identify a structural and functional similarity between the bacterial type III secretion system and eukaryotic apoptosis.<sup>18</sup> In order to demonstrate the high-throughput nature of FAST-NMR, 20 functionally

uncharacterized proteins were screened using the FAST-NMR methodology and subsequently annotated.

## **2.2 MATERIALS AND METHODS**

**2.2.1 Hypothetical proteins from the NESG.** A total of 32 different proteins were received from the Northeast Structure Genomics Consortium (NESG: <http://www.nesg.org>), with each protein consisting of an unlabeled and a labeled ( $^{15}\text{N}$  or  $^{13}\text{C}$ - $^{15}\text{N}$ ) sample. These proteins were selected as representatives for both hypothetical and annotated proteins. Of these 32 proteins, 25 of them were successfully screened by FAST-NMR, where 20 of the proteins were hypothetical (Table 2.1).

**Table 2.1 Uncharacterized proteins screened by FAST-NMR**

<b>Organism</b>	<b>Gene ID</b>	<b>UniProt ID</b>	<b>NESG ID</b>	<b>PDB ID</b>	<b>No. of Amino Acids</b>
<i>Bacillus subtilis</i>	yjcQ	O31639	SR346	2HGC	94
<i>Bacillus subtilis</i>	ykvR	O31683	SR358	2JN9	96
<i>Bacillus subtilis</i>	ynzC	O31818	SR384	3BHP	77
<i>Bacillus subtilis</i>	yoze	O31864	SR391	2FJ6	74
<i>Bacteroides vulgatus</i>	BVU_3908	A6L747	BvR153	2L01	69
<i>Bordetella bronchiseptica</i>	BB0938	Q7WNU7	BoR11	2EXN <sup>19</sup>	128
<i>Caulobacter crescentus</i>	CC_0527	Q9AAR9	CcR55	2JQN	114
<i>Escherichia coli</i>	ytfP	P0AE48	ER111	1XHS <sup>20</sup>	113
<i>Escherichia coli</i>	yrbA	P0A9W6	ER115	1NY8	89
<i>Escherichia coli</i>	yggU	Q8XCU6	ER14	1YH5 <sup>21</sup>	100
<i>Escherichia coli</i>	yjbR	P0AF50	ER226	2FKI <sup>22</sup>	118
<i>Escherichia coli</i>	ydfO	P76156	ER251	2HH8	150
<i>Escherichia coli</i>	ygdR	P65294	ER382A	2JN0	53
<i>Escherichia coli</i>	ykfF	P75677	ER397	2HJJ	79
<i>Escherichia coli</i>	yeiV	P0AFV4	ER541	2K1G <sup>23</sup>	162
<i>Porphyromonas gingivalis</i>	PG_0361	Q7MX54	PgR37A	2KW7	124
<i>Rhodobacter sphaeroides</i>	RHOS4_12090	Q3J357	RhR5	2JRT	92
<i>Salmonella typhimurium</i>	STM0327	Q8ZRJ2	StR65	2JN8	107
<i>Silicibacter pomeroyi</i>	SPO1678	Q5LST8	SiR5	2OA4	93
<i>Staphylococcus saprophyticus</i>	SSP0609	Q49ZM2	SyR11	2K3A <sup>24</sup>	155



**2.2.2 Function-based compound library and mixtures.** The compound library used for the FAST-NMR methodology consists of functional ligands such as metabolites, substrates, inhibitors, and cofactors that have been shown to bind proteins and influence activity.<sup>11</sup> This focused library allows for functional information about the protein to be determined based upon the types of compounds that are shown to bind. The number of compounds in the library fluctuates as some compounds are removed due to cost and stability issues while other new compounds are added. Currently, the compound library contains 460 active compounds. The exact composition of the library can be found in the BioScreen database (<http://bionmr.unl.edu/ligands>). Stock solutions for each compound are stored at -80 °C in either dimethyl sulfoxide-d<sub>6</sub> (DMSO-d<sub>6</sub>; Isotec) or D<sub>2</sub>O (Isotec) at a concentration of 20 mM. In order to minimize the number of NMR samples during the 1D line-broadening screen, the compounds are combined into 117 mixtures that consist of 3-4 compounds each, which has been determined to be the optimal 1D NMR mixture size.<sup>25</sup> The mixtures are created using equal volumes of individual compound stock solutions, leading to a final concentration of 4 mM per ligand in each mixture.

**2.2.3 Ligand-based screen.** The NMR samples for the 1D line-broadening screen were prepared in 10 mM d<sub>19</sub>-bis-Tris (Isotec, St. Louis, MO) buffer at pH 6.5 in 99.99% D<sub>2</sub>O and 11.1 uM TMSP-d<sub>4</sub> (Isotec, St. Louis, MO) to act as a chemical shift reference. Each sample had a 100 µM final concentration for each ligand in the mixture while the protein concentration varied from 10-25 µM. The 1D <sup>1</sup>H NMR spectra were collected on a Bruker 500 MHz Avance spectrometer with a triple-resonance, Z-axis gradient cryoprobe and BACS-120 sample changer using 1D <sup>1</sup>H excitation sculpting pulse sequence to improve water suppression and signal-to-noise. The data was processed using

ACD 1D NMR Processor. The spectra of the ligand mixtures with protein were visually compared to the spectra of the ligand mixtures without protein. A mixture was flagged as a potential binding event if the peak intensity of a protein-ligand sample decreased relative to the ligand-only sample. The potential binding ligand was identified by comparing the broadened peak to reference spectra of ligands known to be in the mixture.

**2.2.4 Target-based screen.** For each compound that showed binding in the 1D line-broadening screen, an NMR sample of that compound in the presence of protein was prepared. The NMR sample consisted of the ligand at 400  $\mu\text{M}$  concentration and the protein ( $^{15}\text{N}$ -labeled or 5%  $^{13}\text{C}$ , 100%  $^{15}\text{N}$ -labeled) at 30  $\mu\text{M}$  concentration in a 10 mM bis-Tris (Sigma-Aldrich) buffer at pH 6.5 and 10%  $\text{D}_2\text{O}$  (Isotec). The 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC experiments were collected on the same 500 MHz spectrometer described above using the WATERGATE and water flip-back pulses for solvent suppression. The data was processed using NMRPipe<sup>26</sup> and visualized in CCPNMR Analysis (<http://www.ccpn.ac.uk>).<sup>27</sup> The resulting spectra were overlaid with the spectrum of the free protein, and protein-ligand spectra that showed significant perturbations of NMR peaks relative to the free protein spectra were designated as binders.

**2.2.5 Rapid generation of protein-ligand costructures.** The ligand that caused the greatest magnitude chemical shift perturbations (CSPs) in the 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC screen was used to define the consensus binding site using CSP-Consensus (described in Chapter 4). The three dimensional structures of both the ligand and the protein were prepared for docking (correcting missing atoms) using UCSF Chimera.<sup>28</sup> AutoDock 4.2.3<sup>29-31</sup> with the AutoDockTools 1.5.4<sup>31,32</sup> (<http://mgltools.scripps.edu>) graphical interface was used to calculate 120 protein-ligand costructures. The AutoDock grid was

set to encompass the consensus binding site identified from CSP-Consensus with a grid spacing of 0.375 Å. The docking was performed using the Lamarckian genetic algorithm with a population of 300 and 2,500,000 energy evaluations. The docked structure that best agreed with the experimental CSPs from the 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC experiment was identified using AutoDockFilter 2.0 (described in Chapter 3 and Chapter 4).<sup>13</sup>

**2.2.6 CPASS.** The protein-ligand costructure generated from AutoDock and AutoDockFilter was submitted to the Comparison of Protein Active Site Similarity (CPASS) program and database (described in more detail in Chapter 5).<sup>14,15</sup> CPASS takes the experimental binding site defined by the protein-ligand costructure and compares its structure and sequence to the active sites of ~36,000 proteins from the Protein Data Bank (PDB). Proteins that have very similar active sites would suggest a similarity in molecular function.

**2.2.7 Other bioinformatics tools.** In addition to the results of CPASS, other bioinformatics approaches are used to provide additional context to the results of the FAST-NMR screen (Table 2.2). The identity of the compounds shown to bind in the target-based screen provides a significant amount of information as well. Knowledge of the type of proteins or metabolic pathways that incorporate the binding compound can allow for an additional line of evidence to evaluate possible functions. Overall, the agreement between multiple approaches toward determining the function of a protein provides greater confidence in the proposed annotation and could be used to guide further experiments.

**Table 2.2 Bioinformatics tools used for protein functional annotation**

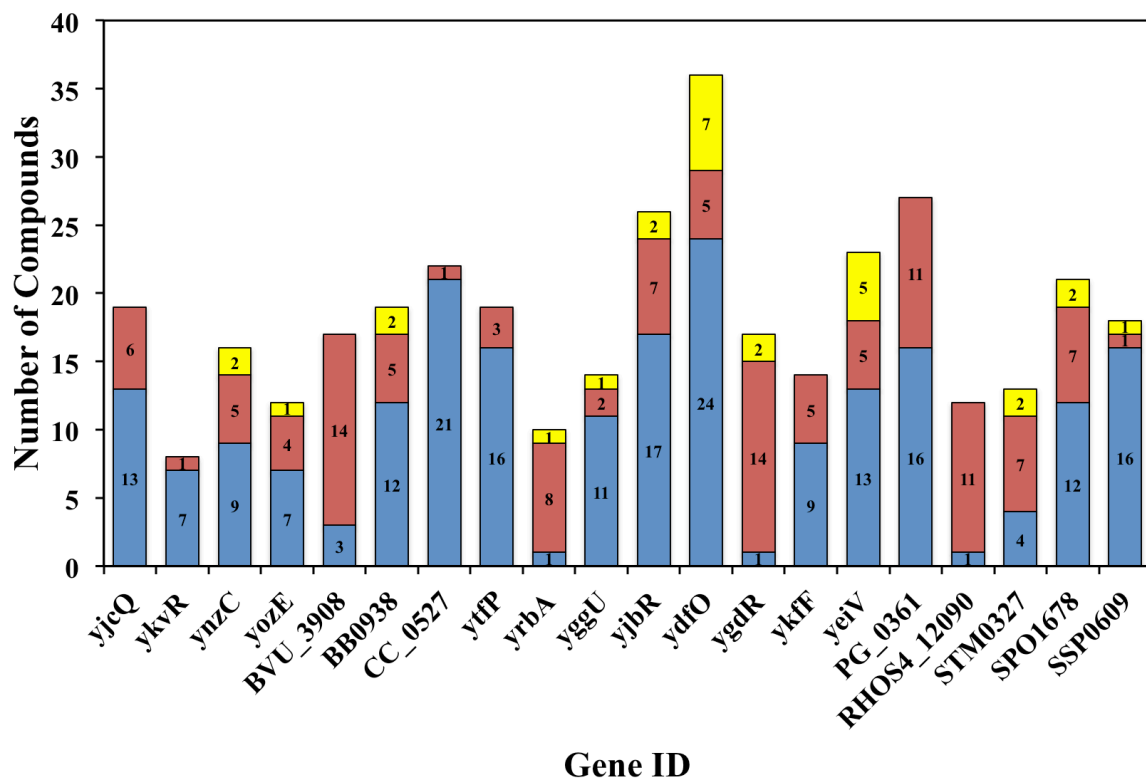
<b>Bioinformatics Tool</b>	<b>Use</b>	<b>Web Address</b>
BioCyc <sup>33</sup>	Locating nearby genes/operons	<a href="http://biocyc.org/">http://biocyc.org/</a>
BLAST <sup>34,35</sup>	Identifying sequence homology	<a href="http://blast.ncbi.nlm.nih.gov">http://blast.ncbi.nlm.nih.gov</a>
CASTp <sup>36</sup>	Comparison of experimental binding sites to binding pockets based on cavity size/shape	<a href="http://sts-fw.bioengr.uic.edu/castp">http://sts-fw.bioengr.uic.edu/castp</a>
CombFunc	Gene ontology function prediction server which includes ConFunc, <sup>37</sup> BLAST, <sup>35</sup> InterPro, <sup>38</sup> Pfam2GO, <sup>39</sup> Phyre2, <sup>40</sup> and 3DLigandSite <sup>41</sup> searches	<a href="http://www.sbg.bio.ic.ac.uk/~mwass/combfunc">http://www.sbg.bio.ic.ac.uk/~mwass/combfunc</a>
ConSurf <sup>42</sup>	Locating evolutionarily conserved residues to compare to experimental binding sites	<a href="http://consurf.tau.ac.il/">http://consurf.tau.ac.il/</a>
CPASS <sup>14,15</sup>	Comparison of experimentally determined binding site to known protein binding sites	<a href="http://cpass.unl.edu/">http://cpass.unl.edu/</a>
Delphi <sup>43</sup>	Predict electrostatic surface charge to compare to experimental binding sites	<a href="http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:DelPhi">http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:DelPhi</a>
ESG <sup>44</sup>	Sequence similarity-based GO annotation prediction	<a href="http://kiharalab.org/web/esg.php">http://kiharalab.org/web/esg.php</a>
IntAct <sup>45</sup>	Database of experimental protein-protein interactions	<a href="http://www.ebi.ac.uk/intact/">http://www.ebi.ac.uk/intact/</a>
InterPro <sup>38</sup>	Predict function based on protein family classification and domain identification	<a href="http://www.ebi.ac.uk/interpro/">http://www.ebi.ac.uk/interpro/</a>
KEGG <sup>46</sup>	Database for identifying pathways which involve the protein or ligand	<a href="http://www.genome.jp/kegg/">http://www.genome.jp/kegg/</a>
MarkUs <sup>47</sup>	Webtool to compare structure and sequence to any known functional annotations	<a href="http://honiglab.c2b2.columbia.edu/MarkUs/cgi-bin/submit.pl">http://honiglab.c2b2.columbia.edu/MarkUs/cgi-bin/submit.pl</a>
PDBeFold <sup>48</sup>	Global structure comparisons	<a href="http://www.ebi.ac.uk/msd-srv/ssm/">http://www.ebi.ac.uk/msd-srv/ssm/</a>
Pfam <sup>49</sup>	Protein family classification based on multiple sequence alignments and hidden Markov models	<a href="http://pfam.sanger.ac.uk/">http://pfam.sanger.ac.uk/</a>
PROFESS <sup>50</sup>	Protein function and evolution analysis	<a href="http://cse.unl.edu/~profess/">http://cse.unl.edu/~profess/</a>
ProFunc <sup>51</sup>	Structure-based function prediction server that includes BLAST searches, structural similarity, structural template comparison, etc.	<a href="http://www.ebi.ac.uk/thornton-srv/databases/profunc/index.html">http://www.ebi.ac.uk/thornton-srv/databases/profunc/index.html</a>
RCSB PDB <sup>52,53</sup>	Protein structure database	<a href="http://www.rcsb.org/">http://www.rcsb.org/</a>
STRING <sup>54</sup>	Database of experimental and	<a href="http://string-db.org/">http://string-db.org/</a>

	predicted protein interactions	
UniProtKB <sup>55</sup>	General reference for information on protein	<a href="http://www.uniprot.org/">http://www.uniprot.org/</a>

## 2.3 RESULTS AND DISCUSSION

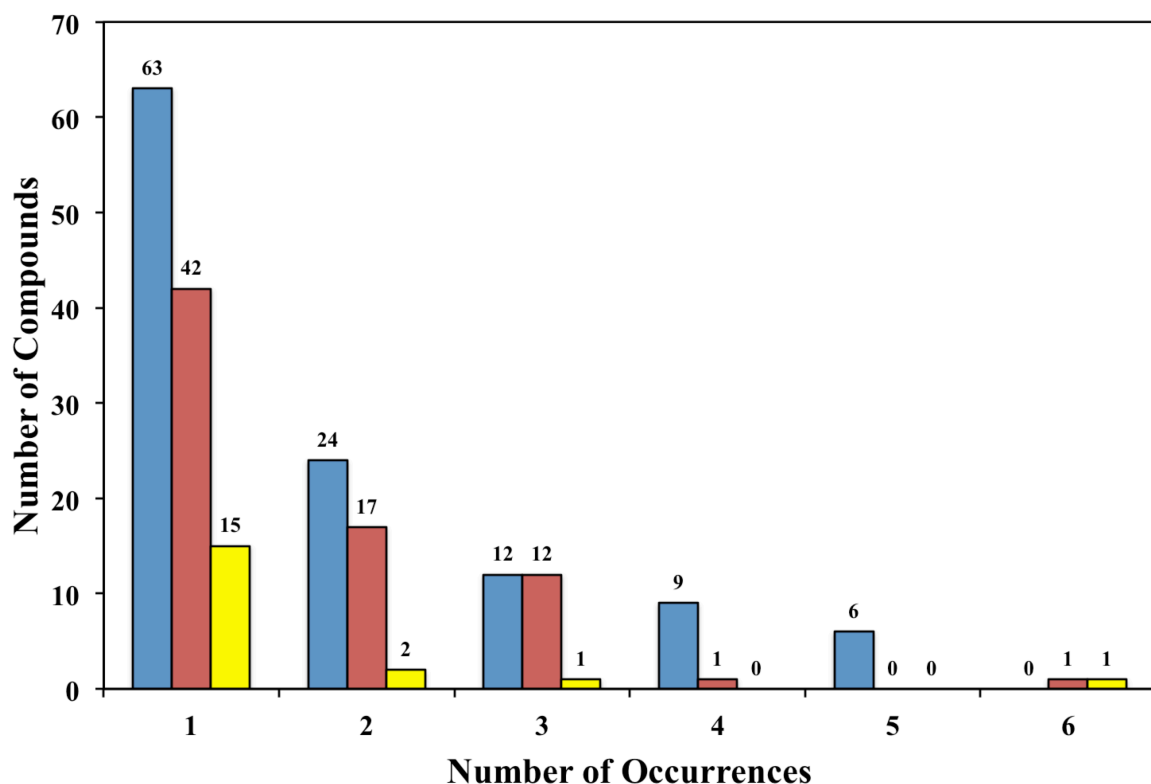
The 1D line-broadening screens for the 20 NESG proteins generated a total of 363 total hits with an average  $18 \pm 7$  hits per protein, which corresponds to a hit rate of approximately 4% for each protein [Figure 2.2]. Many of these 363 hits represent compounds that exhibited line-broadening for multiple proteins. When this duplication is accounted for, the 1D line-broadening screens of all 20 proteins sampled 32.4% of the functional compound library (149 unique hits).

The majority of ligands identified during the 1D  $^1\text{H}$  NMR line-broadening screen were typically specific binders. Approximately 58.7% (213 total) of the binders from the 1D  $^1\text{H}$  NMR line-broadening screen were shown to bind in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screens, which lowers the global hit rate to  $10.7 \pm 6.7$  hits per protein (2.3%) [Figure 2.2], which was still significantly better than seen for traditional high-throughput assays in drug discovery ( $\sim 0.5\%$ ). The 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screens identified a total of 114 unique binders, which represents 24.8% of the total functional compound library. Additionally, 7.7% of the binders from the 1D  $^1\text{H}$  NMR line-broadening screen actually caused precipitation or aggregation of the protein as the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra often exhibit unfolded characteristics or the disappearance of protein peaks.



**Figure 2.2** A histogram demonstrating the number of compounds found to bind during the 1D  $^1\text{H}$  NMR line-broadening screen for each protein. Each bar is further divided based upon the results of the these binders during the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen: binder (blue); non-binder (red); and precipitation/aggregation (yellow).

Compounds routinely identified during the 1D  $^1\text{H}$  NMR line-broadening screen may infer potentially promiscuous compounds (non-specific binders, protein aggregation, functional ambiguity, etc.). Compounds consistently found to show line-broadening in the 1D  $^1\text{H}$  NMR screen are listed in Figure 2.3. For example, suramin underwent line-broadening in 10 different protein screens, which is the most of any compound. Suramin was confirmed as binder in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen for four cases, but caused precipitation/aggregation for the remaining six proteins. Similarly, identifying



**Figure 2.3** A histogram depicting how often a particular compound occurs as a 2D binder (blue), 2D non-binder (red), 2D precipitation/aggregation (yellow).

compounds that consistently yield positive results in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen may indicate the compound is a promiscuous, non-specific binder. Ebselen may be an example of a non-specific binder, since it was identified in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen for 6 different proteins. Nevertheless, there is still value in investigating promiscuous binders, since a compound that binds to a large number of proteins may not represent a functionally relevant interaction. In the FAST-NMR screens of the NESG proteins, 6 compounds (histamine, S-(-)-carbidopa, tyrphostin 25, (-)-riboflavin, nifedipine, Bay 11-7082) appeared as specific binders in 5 different proteins. However, care should be taken before classifying such compounds as promiscuous binders. Some compounds, like histamine or riboflavin, are utilized in many biological pathways and, correspondingly,

are expected to have many binding interactions. So, the biological function(s) associated with the ligand also needs to be considered before classification as a promiscuous binder. As the number of FAST-NMR screens increases, the reliability of classifying particular compounds as promiscuous will improve.

All 20 proteins obtained from NESG received a proposed functional annotation [Figure 2.3] based upon multiple factors: the identity of binders from the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen, a similarity to a functionally characterized protein's active site using CPASS, and the results of common bioinformatics tools based on sequence and structure similarities. While all the data generated from the FAST-NMR screen does not guarantee finding a definitive and correct function for the protein, it did provide information to efficiently guide future experiments. FAST-NMR or similar approaches are, thus, necessary given the large number of functionally uncharacterized proteins that require some means of justifying initiating a detailed investigation.

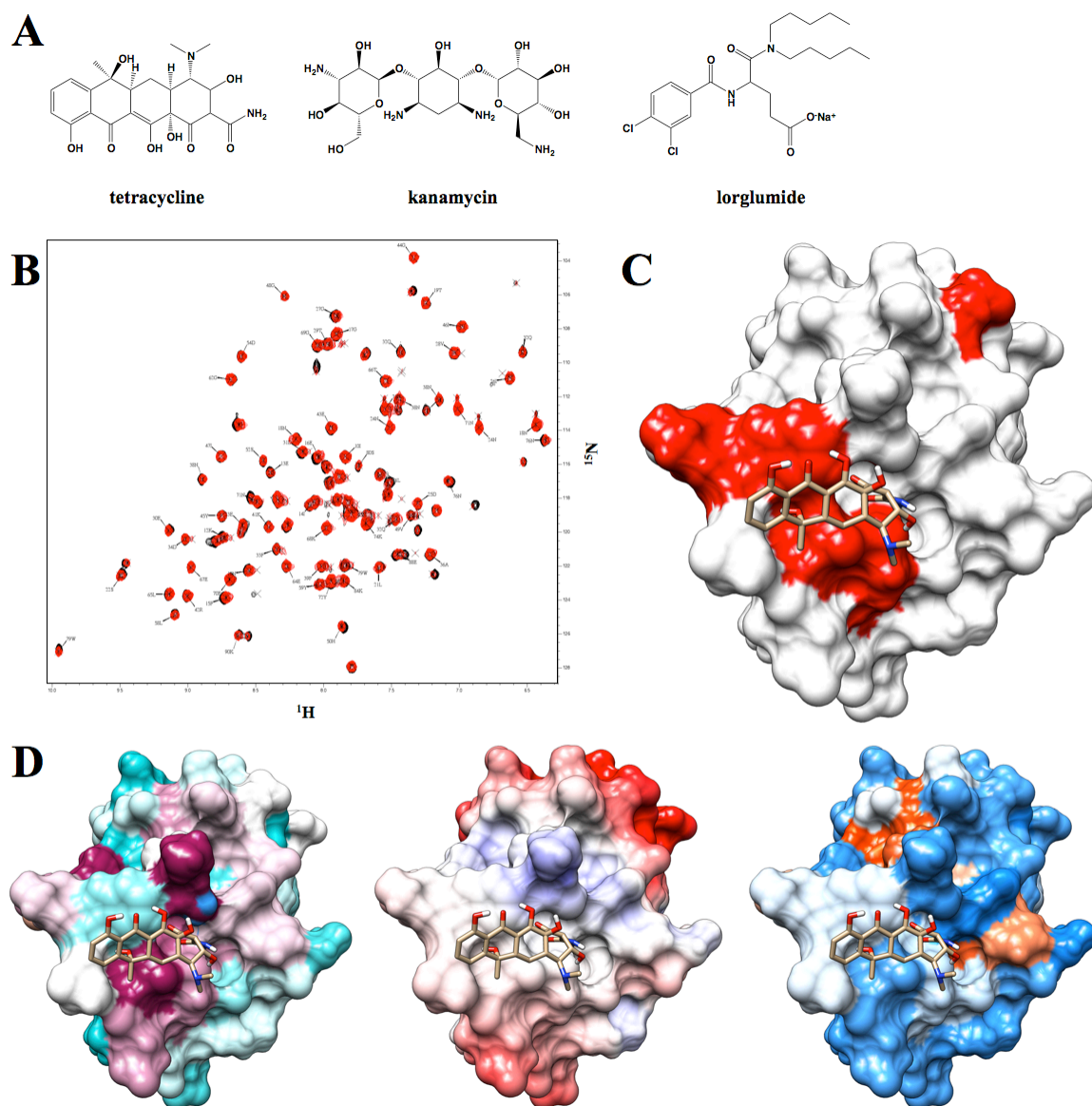


**Table 2.3 Proposed FAST-NMR functional annotations for the 20 NESG proteins.**

<b>Gene ID</b>	<b>NESG ID</b>	<b>UniProt Annotation</b>	<b>Proposed FAST-NMR Annotation</b>
yjcQ	SR346	Uncharacterized protein yjcQ	MarR-like transcriptional regulator involved in chemotaxis
ykvR	SR358	Uncharacterized protein ykvR	Oxidoreductase involved in catechol response
ynzC	SR384	UPF0291 protein ynzC	Activator of SOS-like response to pyrogallols
yoze	SR391	UPF0346 protein yoze	DNA/RNA transferase related to sporulation and/or DNA packing
BVU_3908	BvR153	Putative uncharacterized protein	Z-DNA-binding protease/helicase
BB0938	BoR11	Putative uncharacterized protein	MOSC-like metal-sulfur cluster biosynthesis
CC_0527	CcR55	Putative uncharacterized protein	Dihydropyrimidase-like protein involved in pyrimidine metabolism
ytfP	ER111	Gamma-glutamylcyclotransferase family protein ytfP	BtrG-like aminotransferase
yrbA	ER115	Uncharacterized protein yrbA	BolA-like transcriptional regulator involved in cell wall formation/morphology during stress response
yggU	ER14	UPF0235 protein yggU	Dehydrogenase related to amino acid biosynthesis under diverse environments
yjbR	ER226	Uncharacterized protein yjbR	Transcriptional regulator involved in enterotoxin production
ydfO	ER251	Uncharacterized protein ydfO	Qin prophage protein that responds to stress such as $\beta$ -lactam antibiotics
ygdR	ER382A	Uncharacterized protein ygdR	Membrane-associated oligopeptide transporter

ykfF	ER397	UPF0401 protein ykfF	CP4-6 prophage protein involved in general stress response
yeiV	ER541	Probable endopeptidase Spr	Endopeptidase acting on phenylalanine and/or tyrosine cleavage sites
PG_0361	PgR37A	Conserved domain protein	Methanol dehydrogenase
RHOS4_12090	RhR5	Putative uncharacterized protein	DNA-binding transcriptional repressor involved in signal transduction
STM0327	StR65	Putative cytoplasmic protein	Permease transport and/or signaling
SPO1678	SiR5	Putative uncharacterized protein	DNA-binding transcriptional repressor involved in signal transduction
SSP0609	SyR11	Putative secretory antigen	N-acetylmuramoyl-L-alanine amidase-like protein involved in peptidoglycan hydrolysis/cell lysis

**2.3.1 *Bacillus subtilis* yjcQ (NESG ID: SR346).** The 1D  $^1\text{H}$  NMR line-broadening screen of SR346 identified a total of 19 compounds. Of these 19 compounds, 13 (68.4%) were confirmed as binders in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen, but only three compounds showed significant chemical shift perturbations (CSPs) [Figure 2.4A]. The tightest binder, tetracycline, produced 5 significant perturbations and several smaller perturbations [Figure 2.4B], which present a well-defined consensus binding site [Figure 2.4C]. The binding site does have a few well-conserved residues but no significant hydrophobic or electrostatic characteristics [Figure 2.4D].

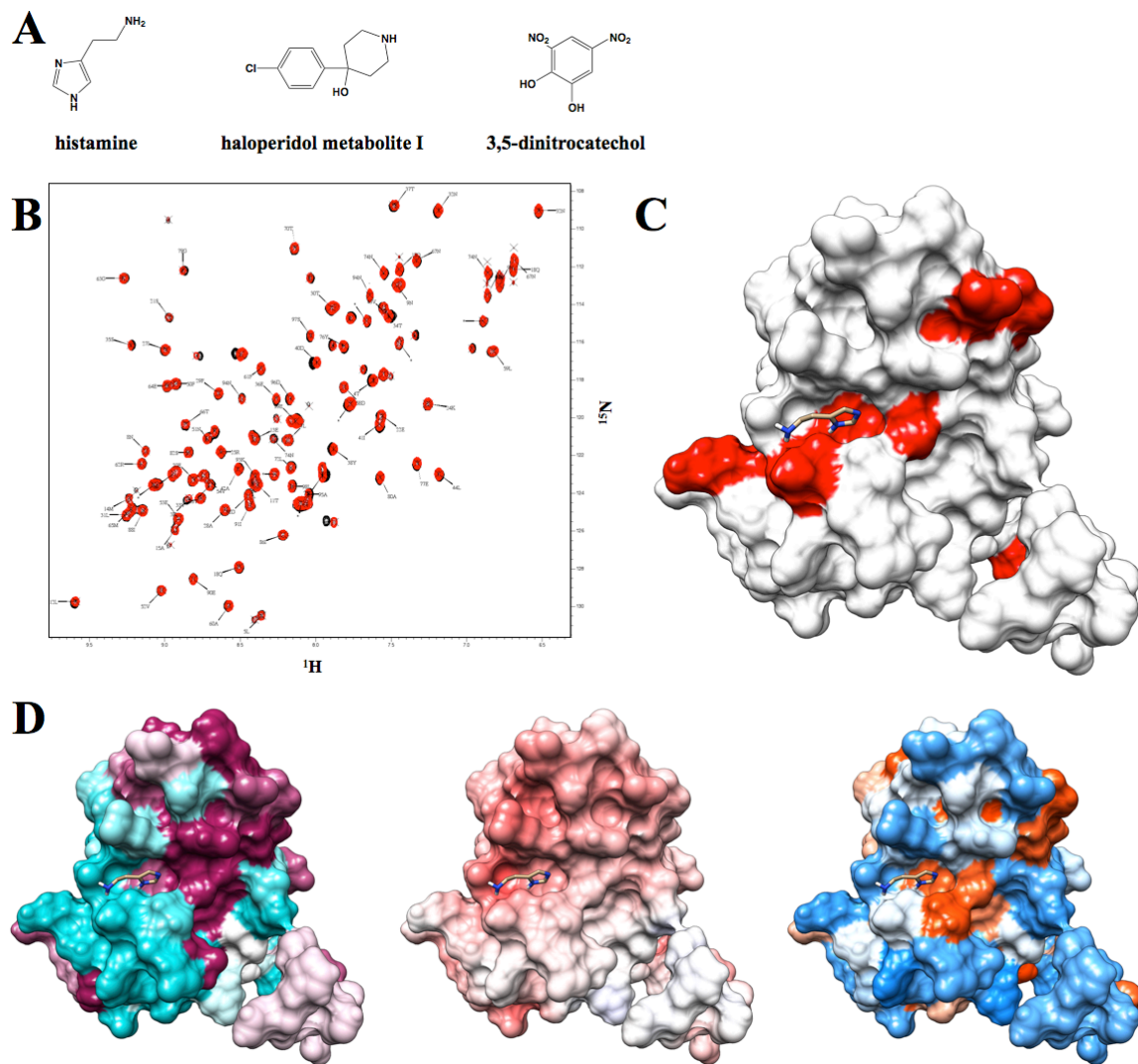


**Figure 2.4** (A) Chemical structures of three compounds shown to bind SR346 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free SR346 (black) and SR346 bound with tetracycline (red). (C) The SR346-tetracycline costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the SR346-tetracycline costructure: (left) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (center) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (right) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes SR346 as an "uncharacterized protein yjcQ." Based on the results of the FAST-NMR approach, the function of SR346 appears to be similar to a MarR-like transcriptional regulator that is likely involved in chemotaxis. The yjcQ gene is regulated by SigD, which is important for the transcription of flagellin and motility genes as well as chemotaxis.<sup>56,57</sup> Most sequence and structural analysis of SR346 identifies it as a winged-helix DNA-binding protein with significant similarities to the MarR-like transcriptional regulators. MarR transcriptional regulators are important proteins for regulating multiple antibiotic resistance or organisms.<sup>58</sup> This is intriguing as the two compounds that showed the greatest CSPs are both antibiotics. As a transcriptional regulator, the protein should bind DNA, yet the experimental ligand binding site does not occur in the predicted DNA binding site ( $\alpha$ 3-helix).<sup>59</sup> This may indicate ligand binding regulates (negatively or positively) DNA binding, and is a possible allosteric binding site.<sup>60</sup> Unfortunately, CPASS was not able to identify many similar ligand binding sites. The highest similarity score was only 35.69%, which matched an immunoglobulin E protein. This may represent a general similarity in chemotaxis and an immunological response, where both sense an external molecule/antigen and initiate a large scale response by the organism.

**2.3.2 *Bacillus subtilis* ykvR (NESG ID: SR358).** The 1D <sup>1</sup>H NMR line-broadening screen of SR358 identified a total of 8 compounds. Seven (87.5%) of these compounds were confirmed as binders in the 2D <sup>1</sup>H,<sup>15</sup>N-HSQC screen, while only three compounds showed significant CSPs [Figure 2.5A]. The tightest binder, histamine, showed 6 significant CSPs [Figure 2.5B], which formed a consensus binding site [Figure

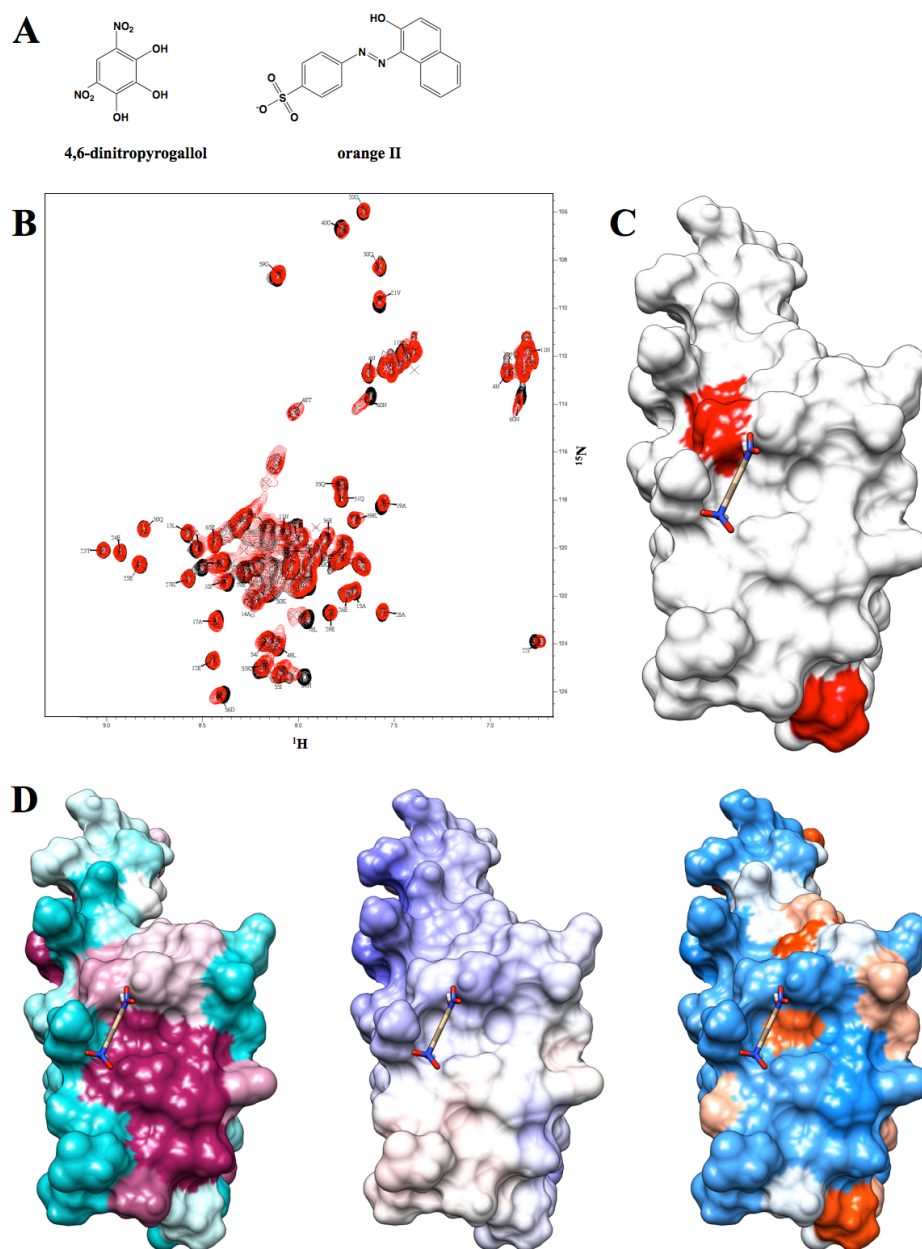
2.5C]. This binding site is near highly conserved residues and has a slightly negative charge to complement the positively charged histamine [Figure 2.5D]



**Figure 2.5** (A) Chemical structures of three compounds shown to bind SR358 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free SR358 (black) and SR358 bound with histamine (red). (C) The SR358-histamine costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the SR358-histamine costructure: (*left*) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (*center*) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (*right*) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes SR358 as an "uncharacterized protein ykvR." Based on the results of the FAST-NMR approach, the function of SR358 may be an oxidoreductase related to the response of *B. subtilis* to catechols, which may include sporulation. The ykvR gene is located near a sporulation-related gene as well as a predicted AdoMet-dependent methyltransferase. ESG strongly predicts an oxidoreductase function, such as dehydrogenases or oxidases, for this protein (100% probability). CPASS does identify a dihydrodiol dehydrogenase as a hit at 41.18%. Dihydrodiol dehydrogenases convert aromatic hydrocarbons into catechols,<sup>61</sup> which is interesting since 3,5-dinitrocatechol was identified as a significant binder. Additionally, in *B. subtilis*, catechols are toxic and can induce a stress response, but it can be catabolized by a catechol-2,3-dioxygenase.<sup>62</sup>

**2.3.3 *Bacillus subtilis* ynzC (NESG ID: SR384).** The 1D <sup>1</sup>H NMR line-broadening screen of SR384 identified a total of 16 compounds that showed line-broadening. Nine (56.3%) of these compounds were confirmed as binders in the 2D <sup>1</sup>H, <sup>15</sup>N-HSQC screen while only two compounds showed significant perturbations [Figure 2.6A]. The tightest binder, 4,6-dinitropyrogallol, showed 9 significant perturbations [Figure 2.6B]. Unfortunately, only two of these perturbations could be reliably assigned. These two residues did not form a consensus binding site [Figure 2.6C]. While both regions were investigated, one perturbed residues bordered upon a well-conserved region of the protein [Figure 2.6D].

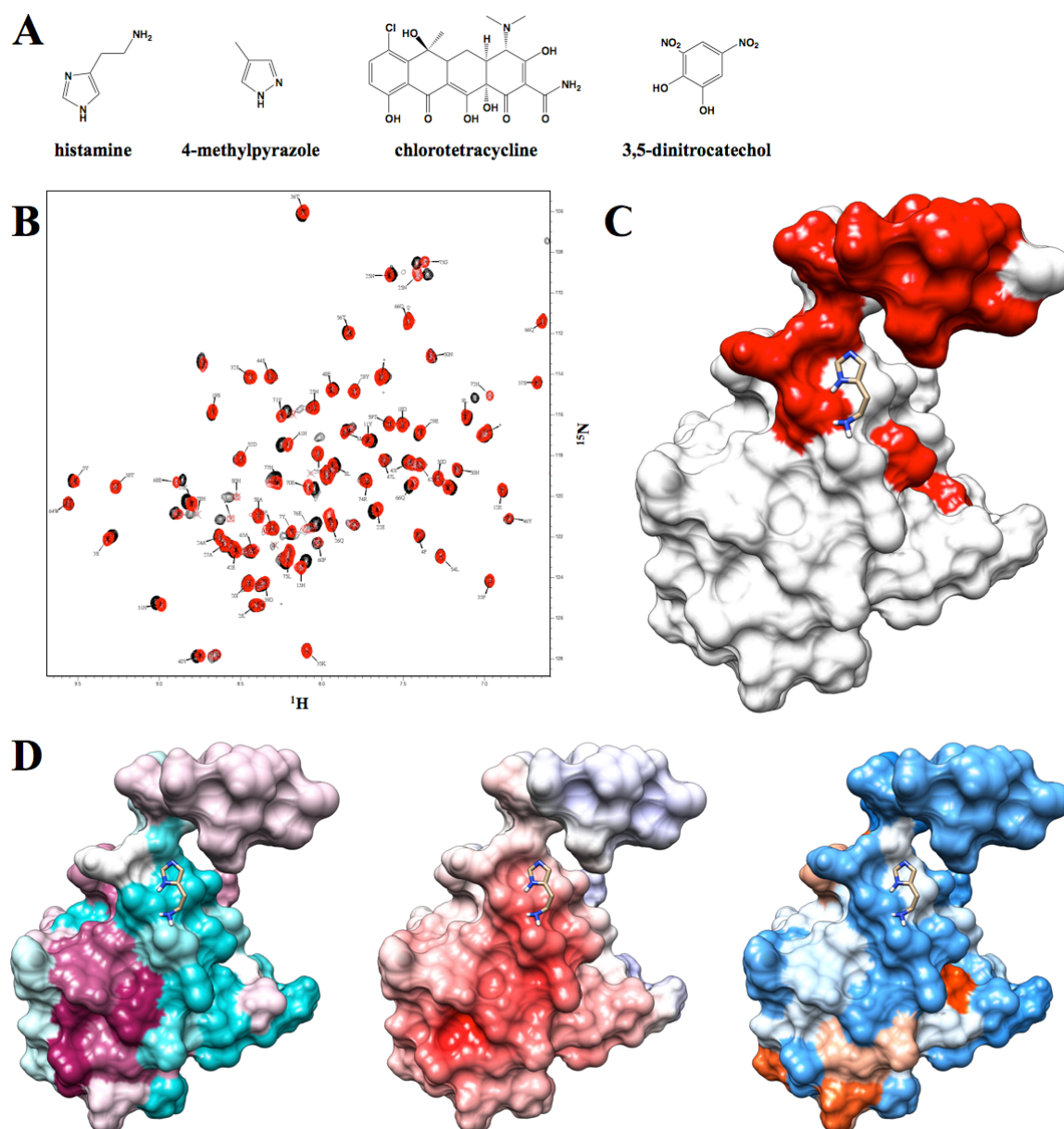


**Figure 2.6** (A) Chemical structures of two compounds shown to bind SR384 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free SR384 (black) and SR384 bound with 4,6-dinitropyrogallol (red). (C) The SR384-4,6-dinitropyrogallol costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the SR384-4,6-dinitropyrogallol costructure: (left) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (center) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (right) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes SR384 as a "UPF0291 protein ynzC." Based on the results of the FAST-NMR approach, SR384 may be involved in detecting pyrogallol-like compounds, which inhibits quorum sensing, which activates an SOS-like response. The genes nearby ynzC are related to cell division and the SOS response, which is responsible for DNA repair.<sup>63,64</sup> BLAST sequence similarity shows that the sequence is well conserved among Gram-positive organisms and the 2D binder 4,6-dinitropyrogallol and other pyrogallol compounds have been implicated in the inhibition of bacterial quorum sensing, which is involved in biofilm formation, bacterial virulence, and drug resistance.<sup>65</sup> However, CPASS does not appear to identify any proteins directly related to cell division, SOS response, or bacterial pathogenicity. The top hit for CPASS (46.66%) is a protein related to fatty acid biosynthesis, which does not appear to be related to the proposed function. Since only 2 CSPs could be used to define the binding site, the proposed binding site may not be accurate, yet it does coincide with the highly conserved region of the protein.

**2.3.4 *Bacillus subtilis* yozE (NESG ID: SR391).** The 1D <sup>1</sup>H NMR line-broadening screen of SR391 identified a total of 13 compounds that showed line-broadening. Eight (61.5%) of these compounds were confirmed as binders in the 2D <sup>1</sup>H,<sup>15</sup>N-HSQC screen most of which showed significant perturbations [Figure 2.7A]. Histamine showed the most significant perturbations [Figure 2.7B], which generates a fairly consistent binding region [Figure 2.7C]. The experimental binding site does not appear to be well conserved, but does have a slightly negative electrostatic surface [Figure 2.7D].

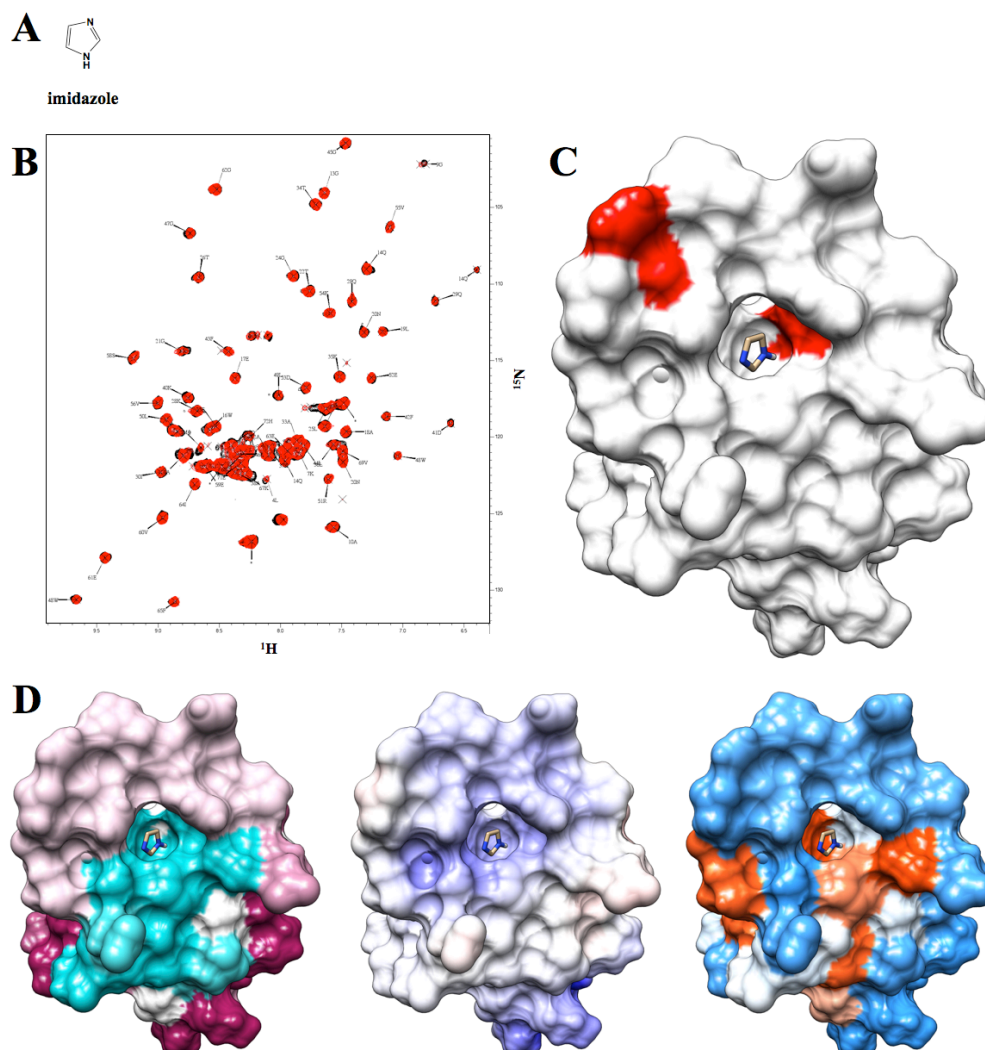




**Figure 2.7** (A) Chemical structures of four compounds shown to bind SR391 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free SR391 (black) and SR391 bound with histamine (red). (C) The SR391-histamine costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the SR391-histamine costructure: (left) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (center) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (right) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes SR391 as a "UPF0346 protein yozE." Based on the results of the FAST-NMR approach, SR391 is likely a DNA/RNA transferase involved with bacterial communication regarding sporulation and/or DNA packing. The yozE gene has a similar phylogenetic profile to a sporulation regulatory gene involved in communication between bacteria<sup>66</sup> and to a protein that promotes RNA polymerase assembly by catalyzing the reaction of a nucleoside triphosphate with RNA. Two of the weaker binders are nucleoside phosphates (UTP and AMP). A structural similarity search identifies the N-terminal fragment of Rad51 from humans, which is similar to the RecA protein from *E. coli*, but includes the additional N-terminal domain. There is some prediction of DNA binding activity for this N-terminal region in humans.<sup>67</sup> The results of CPASS identify a type II DNA topoisomerase, which regulates super-twisting of DNA, with a very similar binding site (42.94%). Binding to DNA/RNA appears to be a common theme and appears to support the proposed FAST-NMR functional annotation.

**2.3.5 *Bacteroides vulgatis* BVU\_3908 (NESG ID: BvR153).** The 1D <sup>1</sup>H NMR line-broadening screen of BvR153 identified a total of 17 compounds that showed line-broadening. Only 3 (17.6%) of these compounds were confirmed as binders in the 2D <sup>1</sup>H, <sup>15</sup>N-HSQC screen, which only showed a small number of significant perturbations [Figure 2.8A]. Imidazole showed the most significant perturbations [Figure 2.8B]. While some of these CSPs do agree with imidazole binding in a small cavity on the protein surface, the small number of CSPs made it difficult to make a definitive assignment [Figure 2.8C]. Nevertheless, this pocket appears to be a likely ligand binding site, has a slight positive charge and is primarily hydrophobic, but is not well conserved [Figure 2.8D].

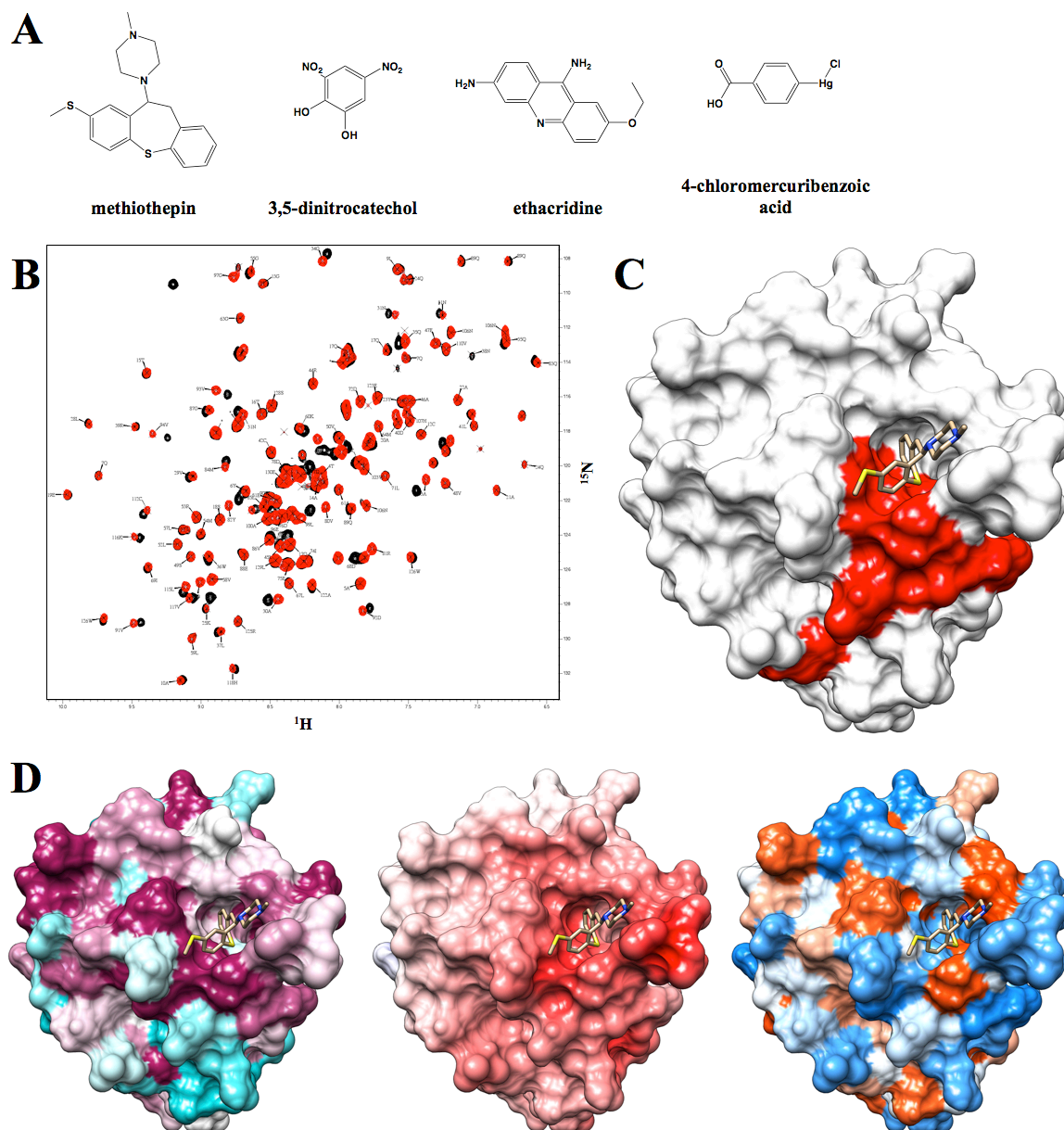


**Figure 2.8** (A) Chemical structures of a compound shown to bind BvR153 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free BvR153 (black) and BvR153 bound with imidazole (red). (C) The BvR153-imidazole costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the BvR153-imidazole costructure: (*left*) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (*center*) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (*right*) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes BvR153 as a "putative uncharacterized protein." Based on the results of the FAST-NMR approach, BvR153 is a Z-alpha protease/helicase, which

is involved in the suppression of the mammalian interferon response pathway. The positive surface charge on this protein suggests DNA binding. BvR153 does appear to have significant structural similarity to several winged helix proteins that bind DNA according to PDBeFold. The best results all feature the Z-alpha domain of adenosine deaminases (1.95 Å RMSD), which bind Z-DNA and convert adenosine to inosine.<sup>68</sup> The binding compound imidazole represents a substructure of both adenosine and inosine. Viruses often utilize the Z-alpha domains to regulate interferon response pathway of their host.<sup>69</sup> CPASS identifies a NS3 protease/helicase bound to an indoline-based inhibitor at 46.02% similarity. NS3 protease/helicase is an essential protein in the life cycle of the hepatitis C virus, and is often the target of drug discovery efforts.<sup>70</sup> The *Bacteriodes vulgatis* organism, from which BvR153 originates, is a gut microbe but is also responsible for infections, and may utilize BvR153 to suppress the interferon response pathway in a similar manner to hepatitis C NS3 protease/helicase.

**2.3.6 *Bordetella bronchiseptica* BB0938 (NESG ID: BoR11).** The 1D <sup>1</sup>H NMR line-broadening screen of BoR11 identified a total of 19 compounds that showed line-broadening. Twelve (63.2%) of these compounds were confirmed as binders in the 2D <sup>1</sup>H, <sup>15</sup>N-HSQC screen, which had many significant perturbations [Figure 2.9A]. The most significant perturbations occurred upon binding with methiothepin [Figure 2.9B], and the perturbations form a well-defined consensus binding site [Figure 2.9C]. This region is very well conserved and exhibits a negatively charged surface [Figure 2.9D].

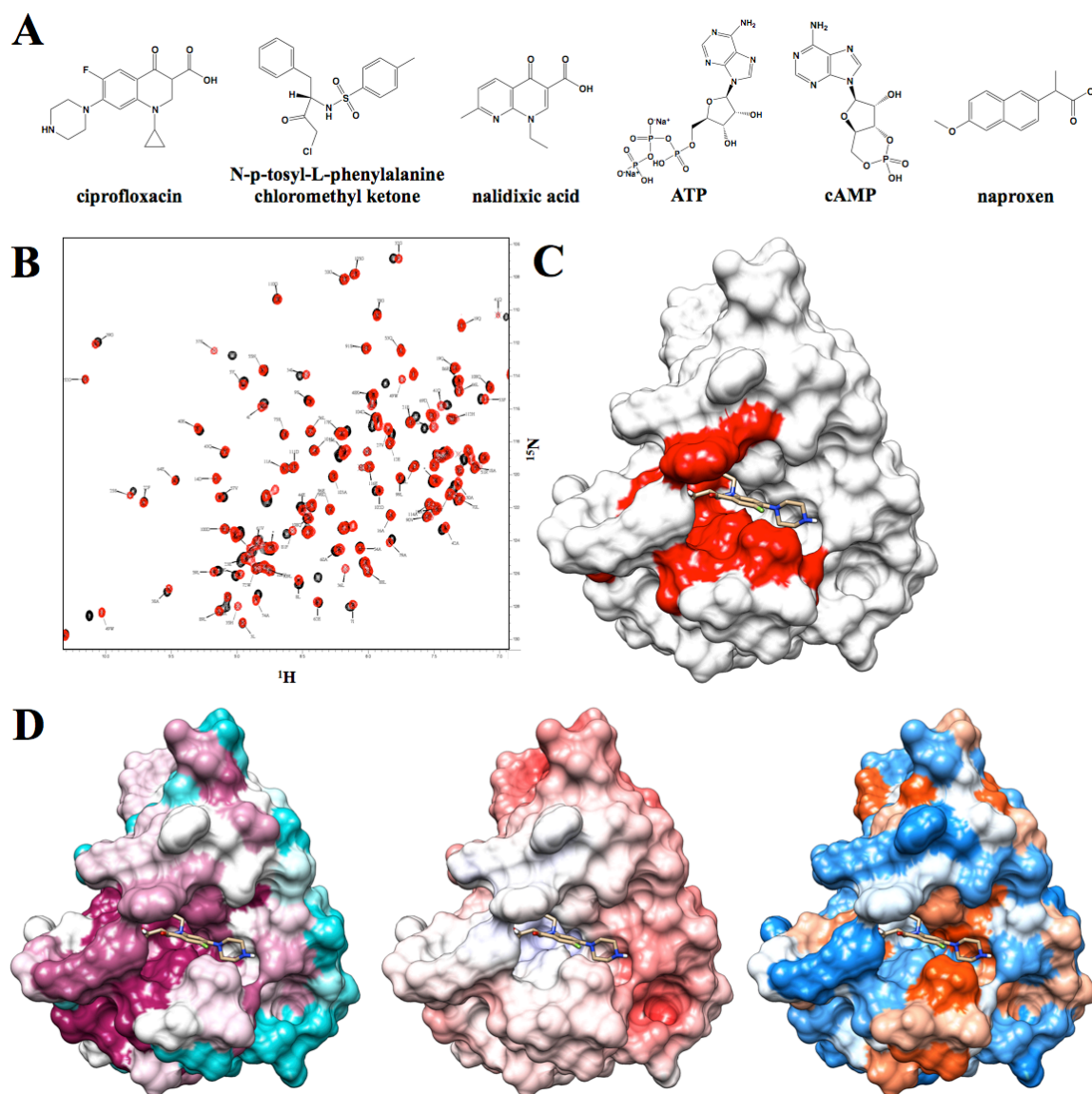


**Figure 2.9** (A) Chemical structures of four compounds shown to bind BoR11 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free BoR11 (black) and BoR11 bound with methiothepin (red). (C) The BoR11-methiothepin costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the BoR11-methiothepin costructure: (*left*) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (*center*) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (*right*) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes BoR11 as a "putative uncharacterized protein." Based on the results of the FAST-NMR approach, the function of BoR11 is likely involved in metal-sulfur cluster biosynthesis similar to MOSC. InterPro and Pfam suggests that BoR11 has a MOSC-like domain, which is important for metal-sulfur cluster biosynthesis.<sup>71</sup> There does not appear to be any structures in the PDB similar to BoR11, but it does adopt a  $\beta$ -barrel type fold, which is commonly found in membrane proteins, transport proteins, as well as MOSC proteins. The tightest binder, methiothepin, contains a couple of sulfur groups while the binder ethacridine bears structural similarity to the molybdenum cofactor present in MOSC proteins. The results of CPASS indicate a strong preference for sulfur-based chemistry with nitrite reductase (34.68%), arylsulfatase (33.08%), and methyltransferases (32.63%) as some of the top hits.

**2.3.7 *Caulobacter crescentus* CC\_0527 (NESG ID: CcR55).** The 1D  $^1\text{H}$  NMR line-broadening screen of CcR55 identified a total of 22 compounds that showed line-broadening. Nearly all of the compounds, 21 (95.5%), were confirmed as binders in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen [Figure 2.10A]. Fifteen of these compounds produce significant CSPs, but ciprofloxacin was selected to generate a costructure [Figure 2.10B]. The CSPs create a well-defined consensus binding site [Figure 2.10C], which is well conserved [Figure 2.10D].



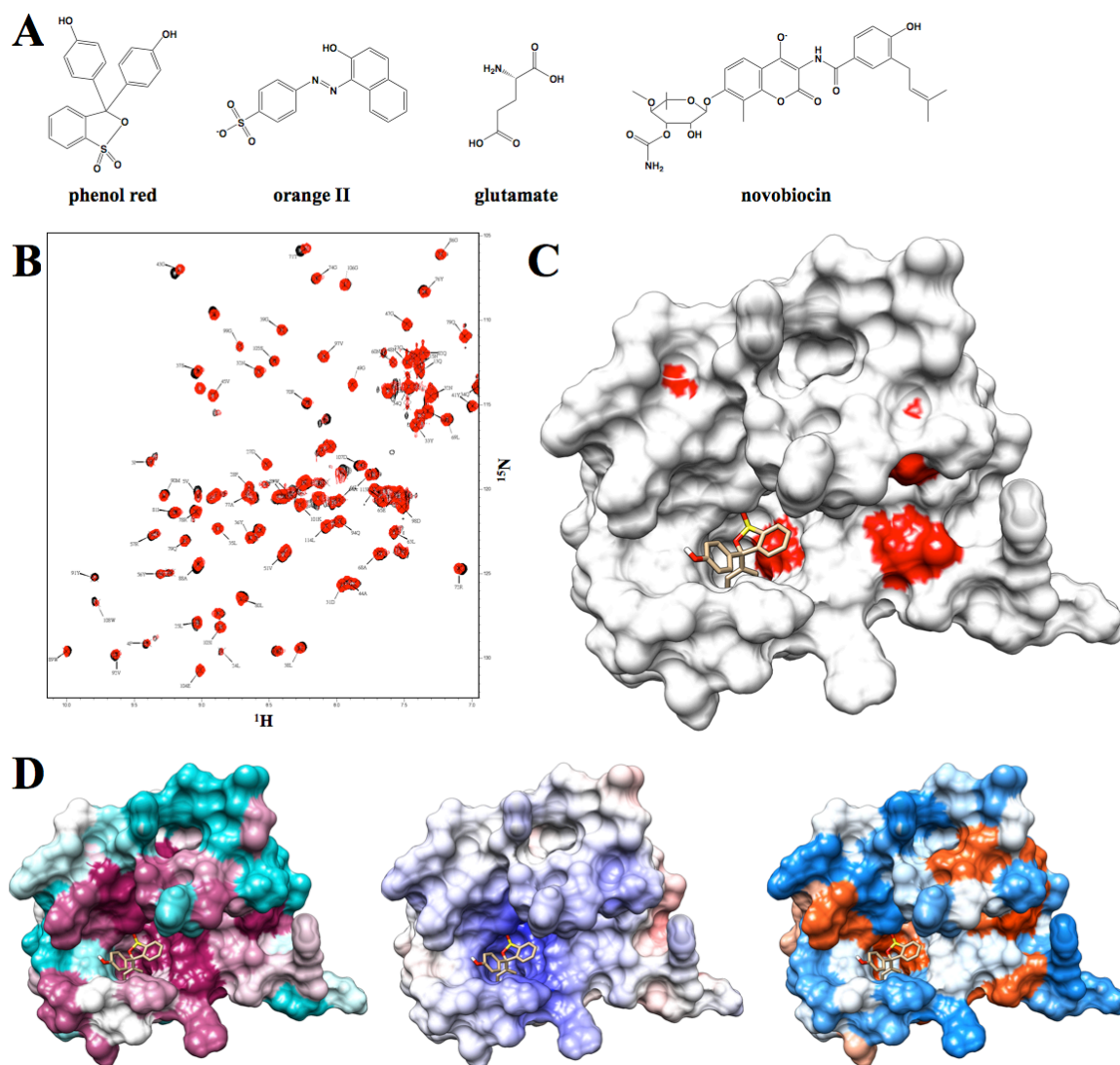


**Figure 2.10** (A) Chemical structures of six compounds shown to bind CcR55 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free CcR55 (black) and CcR55 bound with ciprofloxacin (red). (C) The CcR55-ciprofloxacin costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the CcR55-ciprofloxacin costructure: (left) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (center) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (right) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes CcR55 as a "putative uncharacterized protein." Based on the results of the FAST-NMR approach, CcR55 is likely involved in pyrimidine metabolism and acts similarly to dihydropyrimidase. The CC\_0527 gene is found near a dihydroorotate dehydrogenase, which is involved in pyrimidine metabolism. Sequence similarity searches identify only uncharacterized proteins. A PDBeFold structural similarity search identified a modest similarity (2.9 Å RMSD) to cholera enterotoxin, where both proteins exhibit the ADP-ribosylation fold. While ADP is not found in the compound library, ATP and cyclic AMP were both found as binders in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The top CPASS hit is dihydropyrimidase bound to lysine Nz-carboxylic acid (27.79%), which is involved in pyrimidine metabolism and is similar to the dihydroorotate dehydrogenase.

**2.3.8 *Escherichia coli* ytfP (NESG ID: ER111).** The 1D  $^1\text{H}$  NMR line-broadening screen of ER111 identified a total of 18 compounds that showed line-broadening. Of these compounds, 15 (83.3%) were confirmed as binders in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen, which had many significant perturbations [Figure 2.11A]. The most significant perturbations occurred upon binding with glutamate, however phenol red was used for costructure generation since the residues with significant CSPs could be more reliably assigned [Figure 2.11B]. The perturbed residues identify a pocket for binding [Figure 2.11C], which is highly conserved, positively charged, and fairly hydrophobic [Figure 2.11D].





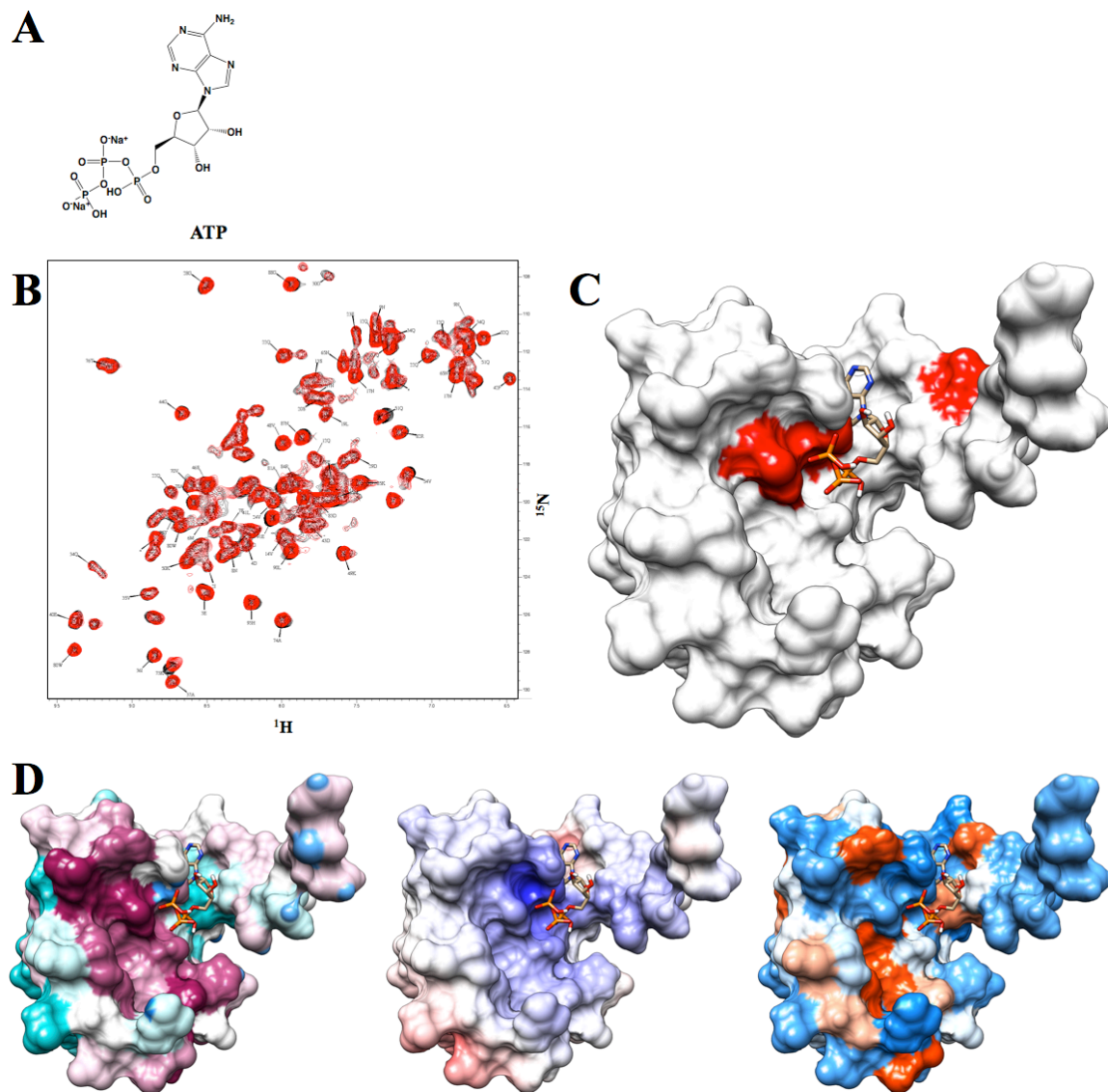
**Figure 2.11** (A) Chemical structures of four compounds shown to bind ER111 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free ER111 (black) and ER111 bound with phenol red (red). (C) The ER111-phenol red costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the ER111-phenol red costructure: (left) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (center) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (right) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes ER111 as a "gamma-glutamylcyclotransferase family protein ytfP." Based on the results of the FAST-NMR approach, ER111 is likely a BtrG-

like aminotransferase. The *ytfP* gene is located near several potentially related genes: predicted outer membrane protein and surface antigen; glycogen synthase; methionine sulfoxide reductase A; quinolinate synthase; tRNA sulfurtransferase; glucose-1-phosphate adenylyltransferase; and ChpB-ChpS toxin-antitoxin system. However, most sequence and structure similarity searches However, InterPro identifies ER111 as having a AIG2-like domain which includes BtrG, a protein used during butirosin antibiotic production,<sup>72</sup> and gamma-glutamyl cyclotransferases, proteins involved in glutathione production. BLAST sequence similarity and PDBeFold structural similarity also identifies high similarity to AIG2-like proteins. Both of these proteins function by cleaving a protecting glutamate, consistent with our observation that ER111 binds glutamate. BtrG cleaves off a protecting glutamyl group from an antibiotic, such as an aminocoumarin like novobiocin, which was also identified as a strong binder in the FAST-NMR screen. ER111 does not have the conserved Glu residue that serves as a proton acceptor in enzymes with gamma-glutamyl cyclotransferase activity,<sup>73</sup> which indicates that ER111 is likely not a gamma-glutamyl cyclotransferase. The binding site generated from the costructure is large (26 amino acids), which causes most CPASS results to fall under a 30% similarity. Nevertheless, a human gamma-glutamyl cyclotransferase bound to pyroglutamic acid was identified by CPASS, but with a very low 20.57% similarity.

**2.3.9 *Escherichia coli* yrbA (NESG ID: ER115).** The 1D <sup>1</sup>H NMR line-broadening screen of ER115 identified a total of 10 compounds that showed line-broadening. Only one compound (10%) was confirmed as a binder in the 2D <sup>1</sup>H, <sup>15</sup>N-HSQC screen [Figure 2.12A]. The only binding compound, ATP, only exhibit a few

significant CSPs [Figure 2.12B]. These CSPs form a consensus binding site [Figure 2.12C], which is not well conserved but is positively charged [Figure 2.12D].

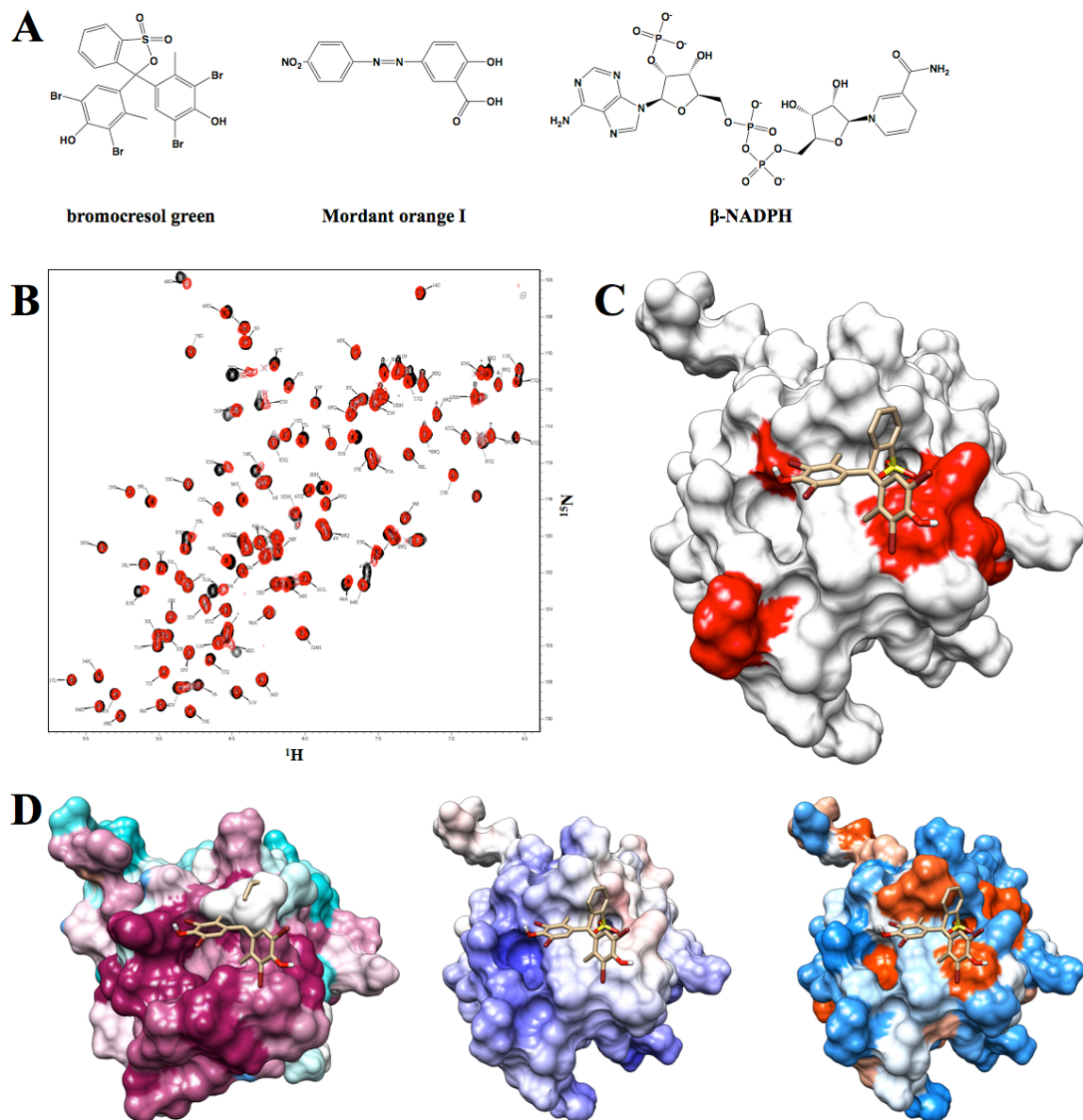


**Figure 2.12** (A) Chemical structures of a compound shown to bind ER115 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free ER115 (black) and ER115 bound with ATP (red). (C) The ER115-ATP costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the ER115-ATP costructure: (*left*) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (*center*) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (*right*) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes ER115 as an "uncharacterized protein." Based on the results of the FAST-NMR approach, ER115 is likely a BolA-like transcriptional regulator involved in cell wall formation/morphology during stress response. A BLAST sequence similarity search identifies numerous proteins that are considered DNA-binding transcriptional regulators or members of the BolA family. InterPro and Pfam also identifies ER115 as a BolA domain. PDBeFold suggests that ER115 is also structurally similar to the BolA proteins (2.76 Å RMSD). BolA proteins has distinct effects on the morphology of the peptidoglycan cell wall, are involved in stress response, and induces the transcription of penicillin binding proteins and  $\beta$ -lactamases.<sup>74,75</sup> This is consistent with the observed binding to ATP in a positively charged surface. STRING has also identified the BolA protein and signal peptidase I to have similar phylogenetic profiles, while UDP-N-acetylglucosamine 1-carboxyvinyltransferase<sup>76</sup> and a predicted ABC-type organic solvent transporter<sup>77</sup> are found to coexpress with ER115. All of these proteins are related to the cell wall. Additionally, the binding site has a positive region that may indicate DNA-binding as a transcriptional regulator. All of this supports the role that ER115 likely plays in cell wall formation/morphology during stress response. CPASS was unable to find any related binding sites in the database.

**2.3.10 *Escherichia coli* yggU (NESG ID: ER14).** The 1D <sup>1</sup>H NMR line-broadening screen of ER14 identified a total of 14 compounds that showed line-broadening. Of these compounds, 11 (78.6%) were confirmed as binders in the 2D <sup>1</sup>H, <sup>15</sup>N-HSQC screen [Figure 2.13A]. Bromocresol green produced the greatest number of perturbations [Figure 2.13B]. The consensus binding site identified from the

significant perturbations forms around a small pocket [Figure 2.13C], which is highly conserved [Figure 2.13D].

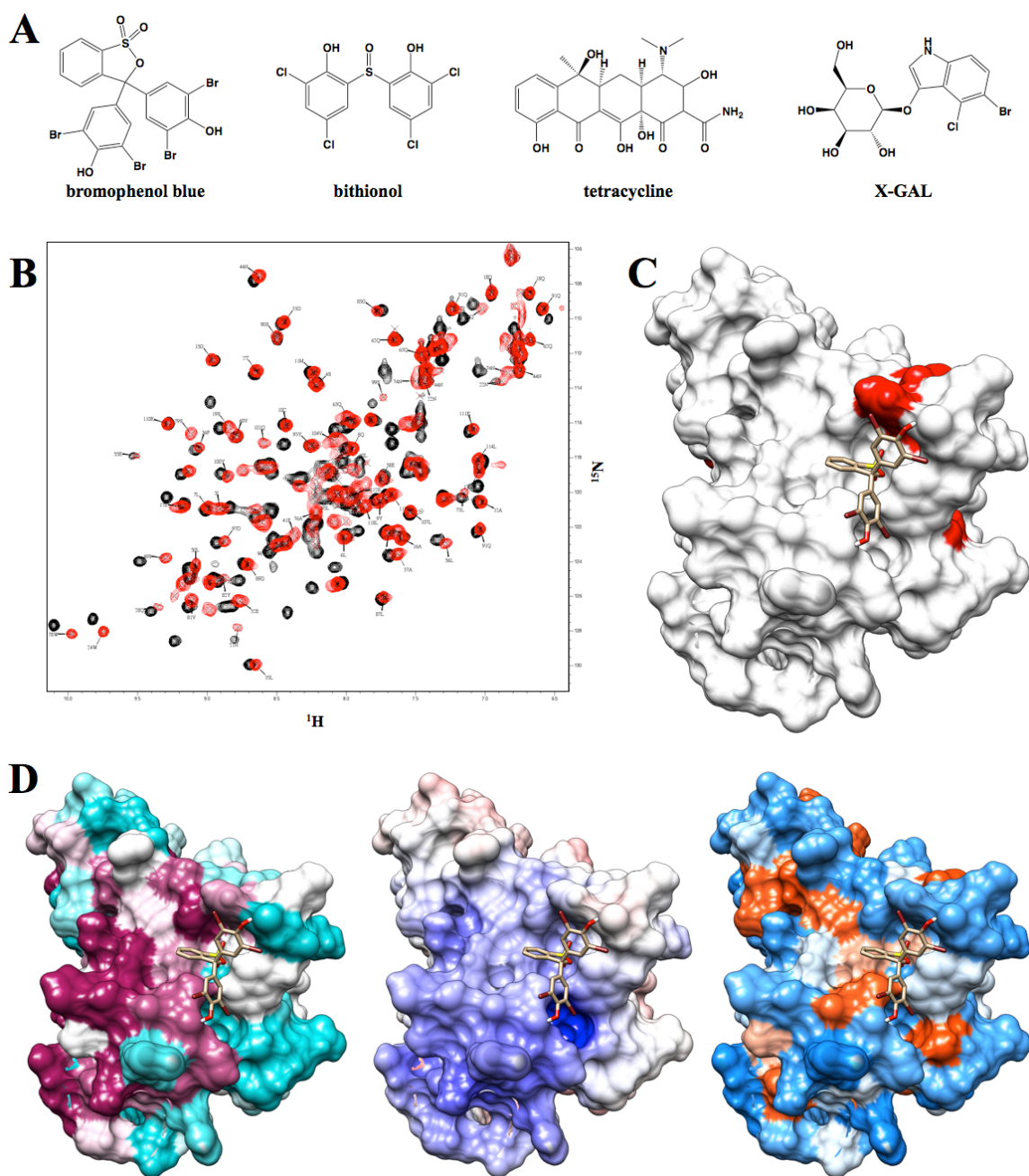


**Figure 2.13** (A) Chemical structures of three compounds shown to bind ER14 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free ER14 (black) and ER14 bound with bromocresol green (red). (C) The ER14-bromocresol green costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the ER14-bromocresol green costructure: (left) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (center) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (right) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes ER14 as a "UPF0235 protein yggU." Based on the results of the FAST-NMR approach, ER14 may be a dehydrogenase probably related to amino acid biosynthesis under diverse environments. A BLAST sequence similarity search primarily finds uncharacterized proteins, except there is 69% sequence identity to an osmotic shock response integral membrane protein. The ER14 protein is structurally similar to some ribosomal proteins (2.47 Å RMSD) according to PDBeFold. CPASS identifies a formaldehyde dehydrogenase (methane metabolism/microbial metabolism in diverse environments) bound to NAD at 38.55% similarity, and quinate/shikimate hydrogenase (phenylalanine, tryptophan, and tyrosine biosynthesis) bound to NAD at 37.10% similarity. Additionally, of the several genes near yggU, one is pyrroline-5-carboxylate reductase, an NAD-dependent enzyme involved in proline biosynthesis.

**2.3.11 *Escherichia coli* yjbR (NESG ID: ER226).** The 1D  $^1\text{H}$  NMR line-broadening screen of ER226 identified a total of 26 compounds that showed line-broadening. Of these compounds, 17 (65.4%) were confirmed as binders in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen [Figure 2.14A]. Both bromophenol blue and bithionol produced many significant perturbations, but bromophenol blue was used for further analysis [Figure 2.14B]. Only some of the perturbed residues could be assigned, leaving a small binding pocket with perturbed residues on the edge [Figure 2.14C]. The binding site is not particularly well conserved and is only slightly positively charged [Figure 2.14D].



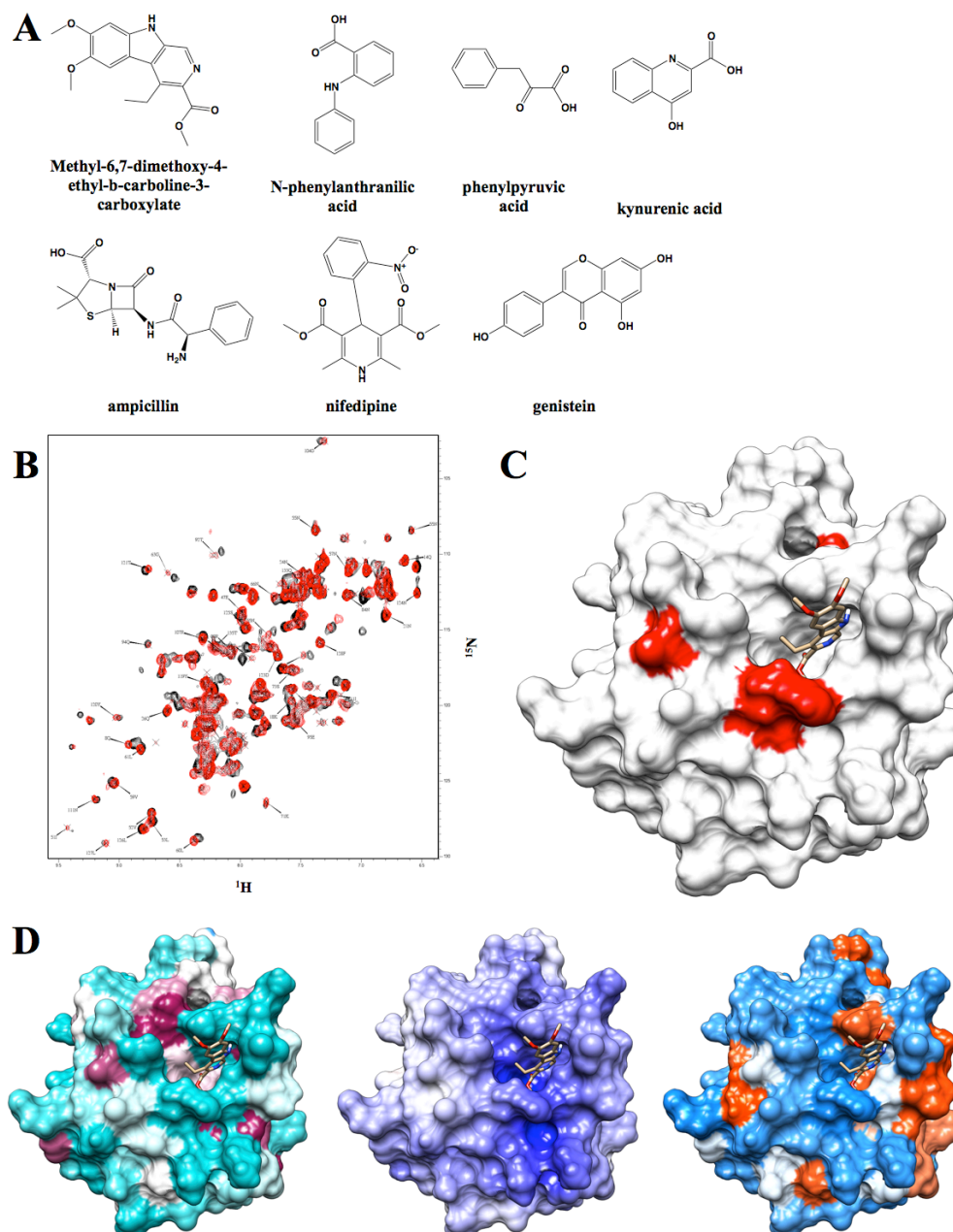


**Figure 2.14** (A) Chemical structures of four compounds shown to bind ER226 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free ER226 (black) and ER226 bound with bromophenol blue (red). (C) The ER226-bromophenol blue costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the ER226-bromophenol blue costructure: (*left*) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (*center*) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (*right*) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes ER226 as an "uncharacterized protein yjbR." Based on the results of the FAST-NMR approach, ER226 is likely a transcriptional regulator possibly related to the production of enterotoxin synthesis. STRING identifies some genes with similar phylogenetic profiles to yjbR, which include a bifunctional DNA-binding transcriptional repressor/NMN adenylyltransferase (NAD biosynthesis) and an acetolactate synthase II (branched chain amino acid synthesis). The structure of ER226<sup>22</sup> indicates “double wing” DNA-binding motif with structural similarity to MotCF and TATA-binding proteins.<sup>78,79</sup> Unfortunately, while the DNA-binding capability is clear, the role of the DNA-binding is less clear. BLAST identifies some sequence similarity to methylated-DNA-(protein)-cysteine S-methyltransferases (34%), which is involved in alkylated DNA repair. PDBeFold indicates that ER226 has some structural similarity a type III secretion chaperones (3.48 Å RMSD), which are often used to inject virulence proteins into the cells of their host.<sup>80</sup> CPASS identifies cholera toxin B bound to galactoside-based inhibitors as the top hits (45.78%). Cholera toxin is related to the heat-labile enterotoxin found in *Escherichia coli*, and a galactose-based compound, X-GAL, was identified as a binder. This hints at a possible virulence role for ER226.

**2.3.12 *Escherichia coli* ydfO (NESG ID: ER251).** The 1D <sup>1</sup>H NMR line-broadening screen of ER251 identified a total of 36 compounds that showed line-broadening. Of these compounds, 24 (66.7%) were confirmed as binders in the 2D <sup>1</sup>H, <sup>15</sup>N-HSQC screen [Figure 2.15A]. Fifteen of these compounds showed a significant number of perturbations that made selecting the tightest binder difficult. Ultimately, methyl-6,7-dimethoxy-4-ethyl-b-carboline-3-carboxylate (DMCM) was chosen to

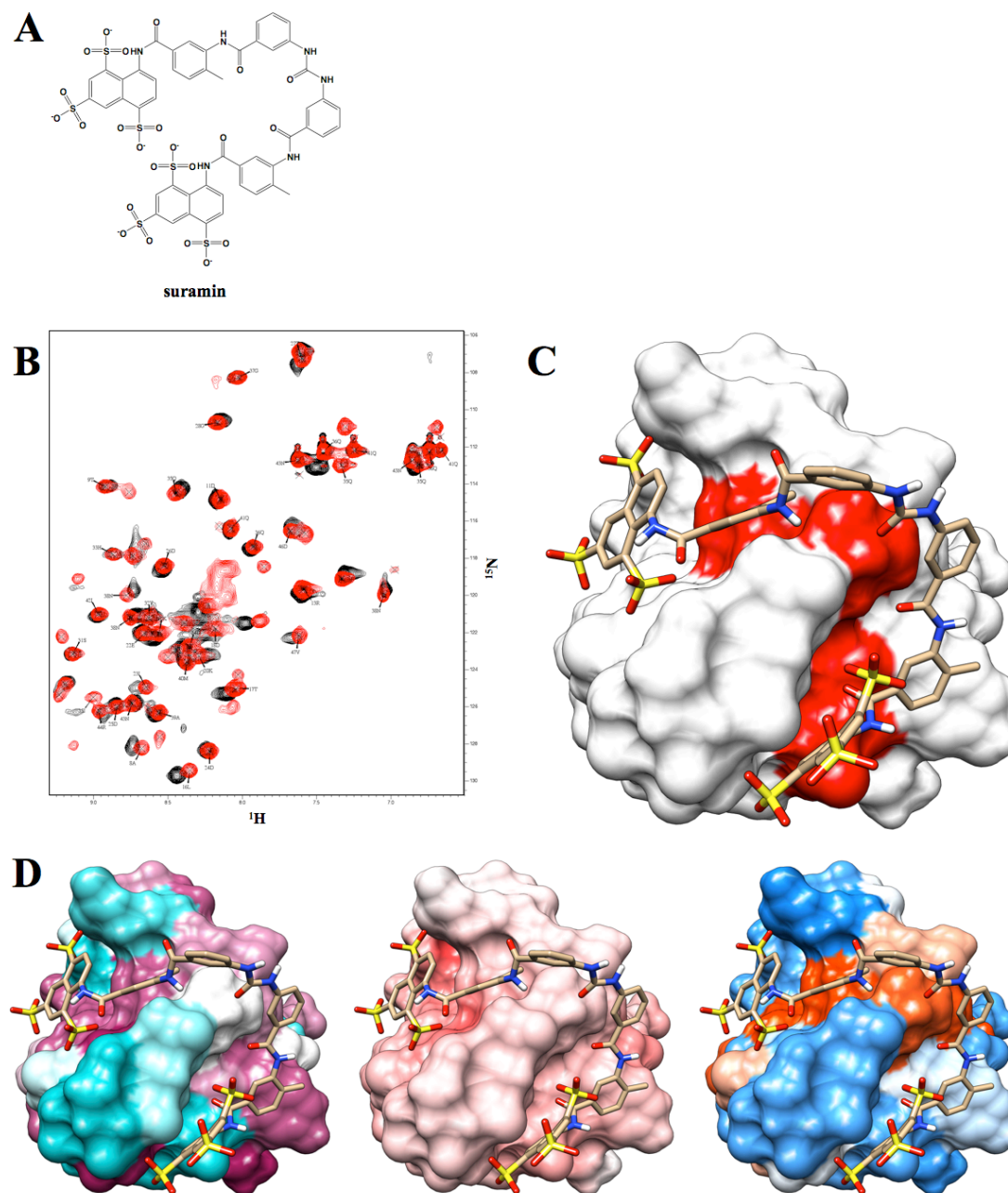
represent the tightest binder [Figure 2.15B]. Despite the many perturbations, the perturbed residues were very difficult to assign reliably. The most significant perturbations that could be assigned appear near a potential binding pocket [Figure 2.15C]. This binding pocket has some nearby highly conserved residues and is also positively charged [Figure 2.15D].



**Figure 2.15** (A) Chemical structures of seven compounds shown to bind ER251 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free ER251 (black) and ER251 bound with DMCM (red). (C) The ER251-DMCM costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the ER251-DMCM costructure: (*left*) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (*center*) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (*right*) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes ER251 as an "uncharacterized protein ydfO." Based on the results of the FAST-NMR approach, ER251 is related to a Qin prophage protein that promotes a bacterial response to environmental stress such as  $\beta$ -lactam antibiotics. STRING indicates that the ydfO gene has a similar phylogenetic profile as the Qin prophage and DLP12 prophage. Pfam predicts that ydfO belongs to a bacteriophage DE3 family that only occurs in *E. coli* and *Salmonella*. The gene ydfO has also been experimentally-determined to be related to the Qin prophage.<sup>81</sup> It has also been experimentally shown to be related to improving survivability of *E. coli* while under peroxide stress.<sup>81</sup> Additionally, deletion of the entire Qin prophage gene, which includes ydfO, significantly lowers survivability in the presence of ampicillin,<sup>81</sup> which is a significant binder to ER251. CPASS was unable to produce any significant results to support or refute the proposed annotation.

**2.3.13 *Escherichia coli* ygdR (NESG ID: ER382A).** The 1D <sup>1</sup>H NMR line-broadening screen of ER251 identified a total of 17 compounds that showed line-broadening. Only one of these compounds (5.9%) could be confirmed as a binder in the 2D <sup>1</sup>H, <sup>15</sup>N-HSQC screen [Figure 2.16A]. Suramin was the only binder but it did produce numerous perturbations [Figure 2.16B]. The location of the CSPs on the protein surface forms a groove, where suramin appears to readily bind [Figure 2.16C]. This binding groove has some nearby conserved residues, a slight negative charge, and hydrophobic characteristics [Figure 2.16D].

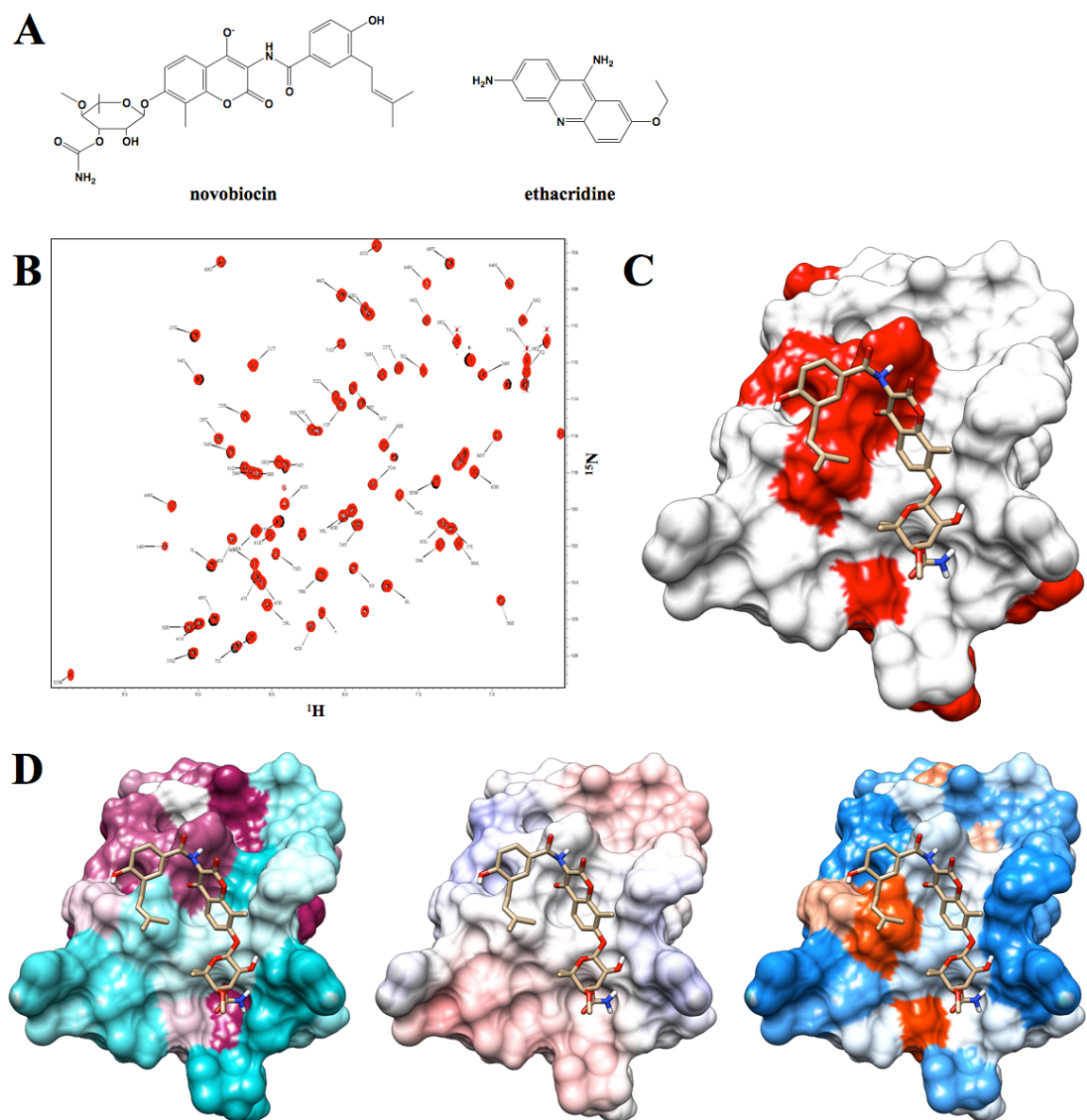


**Figure 2.16** (A) Chemical structures of a compound shown to bind ER382A in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free ER382A (black) and ER382A bound with suramin (red). (C) The ER382A-suramin costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the ER382A-suramin costructure: (left) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (center) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (right) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes ER382A as an "uncharacterized lipoprotein." Based on the results of the FAST-NMR approach, ER382A is likely membrane-associated oligopeptide transporter. The only genes near *ygdR* are a predicted NADP-dependent aldo-keto reductase and a predicted inner membrane protein. However, PDBeFold indicates that ER382A is structurally similar (1.95 Å RMSD) to the C-terminal domain of the biotin holoenzyme synthetase, which has been shown to biotinylate a 23-residue peptide.<sup>82</sup> While the best binder suramin is a heparin mimic and not a peptide, it does have some characteristics similar to a peptide, such as the larger size and amide groups. Additionally, experimental evidence shows that the entire ER382 protein transports di- and tripeptides, and likely belongs to the proton-dependent oligopeptide transporter (POT) family.<sup>83</sup> The electrostatics and hydrophobicity of the ER382A binding site likely provides the specificity for the types of peptides that are transported. CPASS did not find any similar binding sites to the experimental binding site, likely due to the very large size of the binding site.

**2.3.14 *Escherichia coli* ykfF (NESG ID: ER397).** The 1D <sup>1</sup>H NMR line-broadening screen of ER397 identified a total of 14 compounds that showed line-broadening. Nine of these compounds (64.3%) were confirmed as binders in the 2D <sup>1</sup>H, <sup>15</sup>N-HSQC screen [Figure 2.17A]. The perturbations were small in magnitude, where novobiocin represented the tightest binder [Figure 2.17B]. The majority of CSPs formed a consensus binding site on the protein surface [Figure 2.17C], where some residues within this binding site were identified as being highly conserved by ConSurf. [Figure 2.17D].



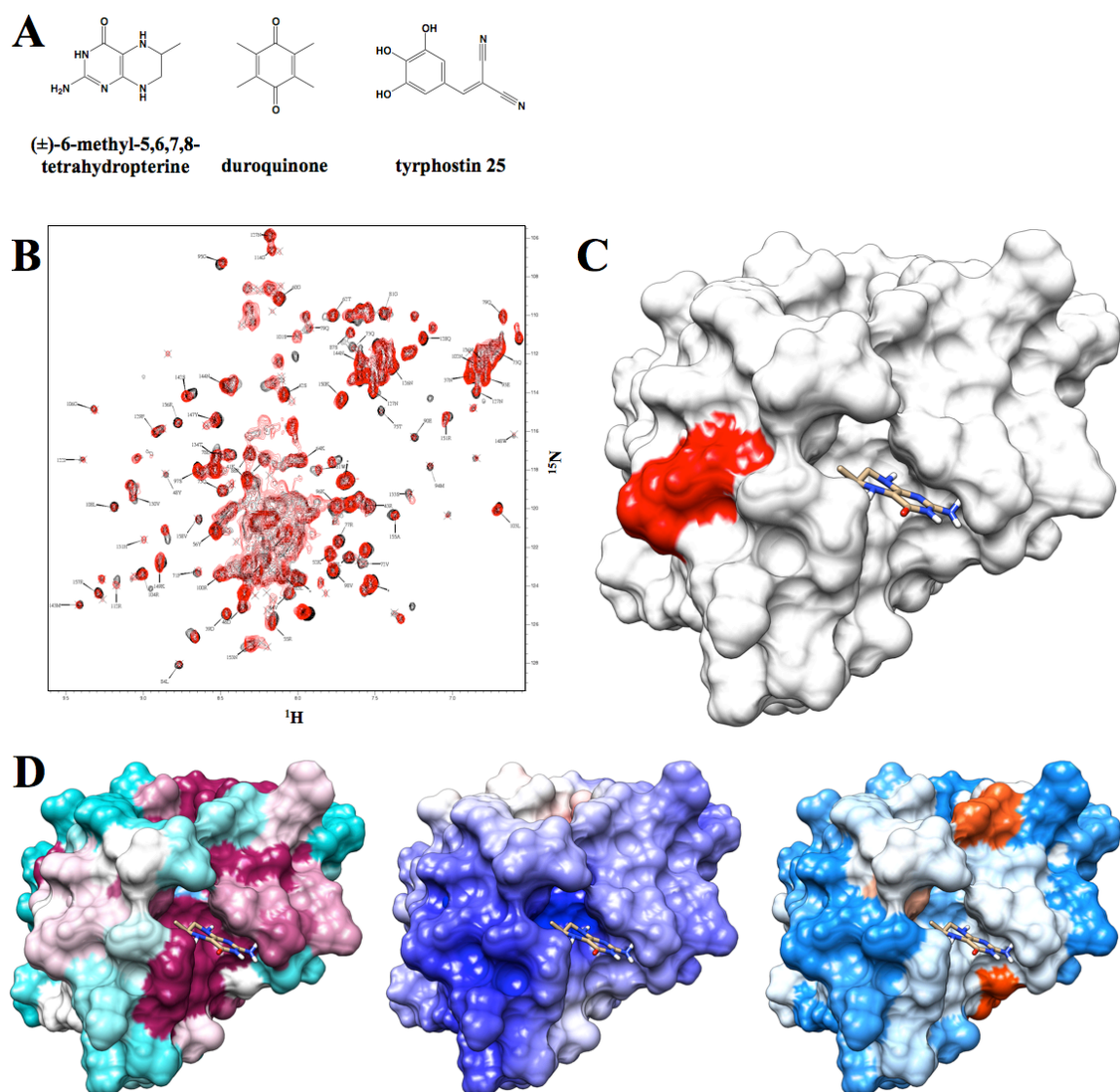


**Figure 2.17** (A) Chemical structures of two compounds shown to bind ER397 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free ER397 (black) and ER397 bound with novobiocin (red). (C) The ER397-novobiocin costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the ER397-novobiocin costructure: (left) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (center) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (right) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.



UniProtKB characterizes ER397 as a "UPF0401 protein ykfF." Based on the results of the FAST-NMR approach, ER397 is likely a CP4-6 prophage involved in survivability of the cell and antibiotic resistance. The gene for ykfF is found in between CP4-6 prophage genes indicating that ykfF is likely a CP4-6 prophage protein, but the function is unknown. The entire CP4-6 prophage has been shown to increase survivability of the cell in the presence of quinolone and  $\beta$ -lactam antibiotics.<sup>81</sup> Novobiocin (aminocoumarin) and amoxicillin ( $\beta$ -lactam) were both shown to be binders, which supports the relationship to antibiotics. From PDBeFold, ER397 is structurally similar to an integrin cassette protein (1.82 Å RMSD) and a glyoxalase/bleomycin resistance protein (2.33 Å RMSD), but BLAST sequence similarity did not identify any similarity between ER397 and functionally characterized proteins. CPASS was unable to identify any similar binding sites.

**2.3.15 *Escherichia coli* yeiV (NESG ID: ER541).** The 1D  $^1\text{H}$  NMR line-broadening screen of ER541 identified a total of 23 compounds that showed line-broadening. Of these compounds, 13 (56.5%) were confirmed as binders in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen [Figure 2.18A]. The compound ( $\pm$ )-6-methyl-5,6,7,8-tetrahydropterine (6-MPH4) represents the best binder with several significant perturbations [Figure 2.18B]. The majority of the perturbations were not able to be assigned, however the perturbations occur near a binding pocket that can fit the 6-MPH4 molecule [Figure 2.18C]. This binding pocket is highly conserved and positively charged [Figure 2.18D].

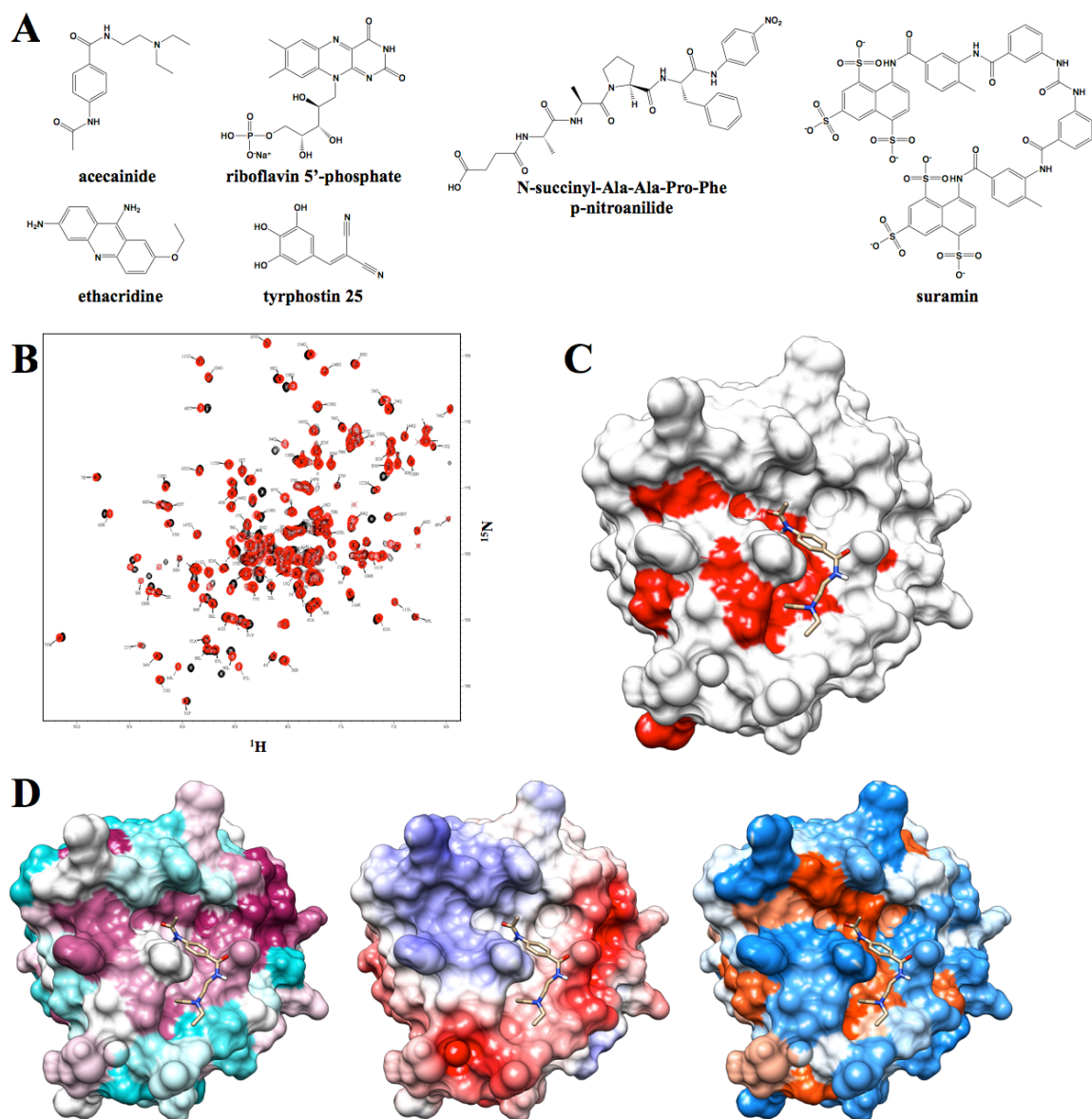


**Figure 2.18** (A) Chemical structures of three compounds shown to bind ER541 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free ER541 (black) and ER541 bound with 6-MPH4 (red). (C) The ER541-6-MPH4 costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the ER541-6-MPH4 costructure: (*left*) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (*center*) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (*right*) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes ER541 as a "probable endopeptidase Spr." Based on the results of the FAST-NMR approach, ER541 is most likely an endopeptidase involved in

cell wall hydrolysis. The *yeiV* gene is located near a putative lipoprotein and is found to coexpress with the same lipoprotein and a DNA-binding transcriptional dual regulator. According to PDBeFold, there is a high structural similarity (1.87 Å RMSD) between ER541 and the NlpC/P60 protein domain families, which belong to the C40 peptidases. These peptidases typically hydrolyze specific peptide linkages in bacterial cell walls, and are likely involved in cell wall hydrolysis during cell growth, division, and lysis.<sup>84</sup> The active site is predicted to include a Cys-His-His catalytic triad, which is present in the experimental binding site for FAST-NMR.<sup>23</sup> CPASS also finds a NLP/P60 family protein bound to cysteinesulfonic acid (34.80%). Given this information, ER541 is likely an endopeptidase. However, the specific peptide linkage that is cleaved is unknown. Among the binders, 6-MPH4 is a cofactor of phenylalanine and tyrosine hydroxylase while tyrphostin 25 is a tyrosine kinase inhibitor. Additionally, there are other amino-acid based binders such as cyclo(His-Pro) and N-p-tosyl-L-phenylalanine chloromethyl ketone. This may suggest that the cleavage site might involve a phenylalanine or tyrosine.

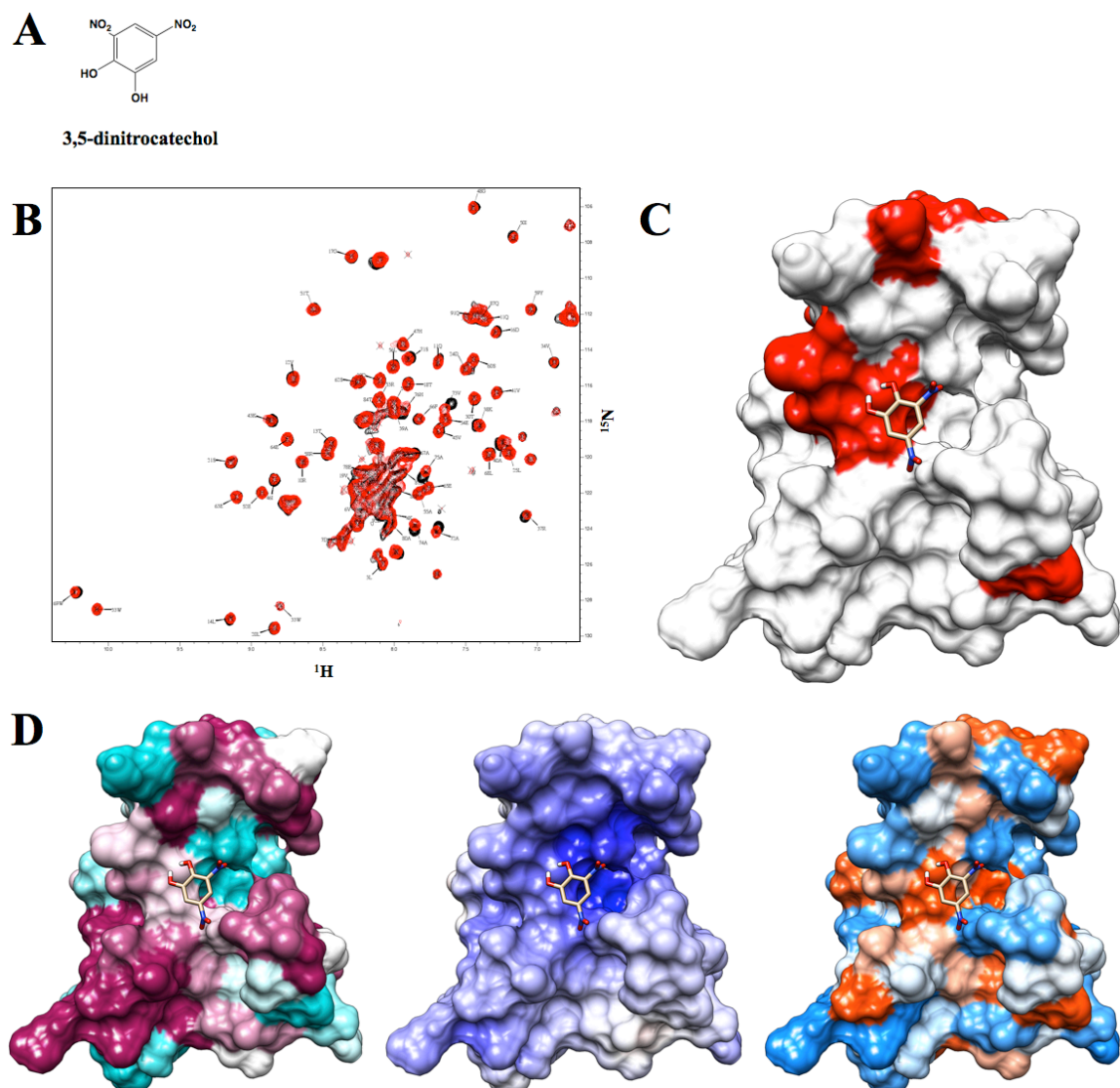
**2.3.16 *Porphyromonas gingivalis* PG\_0361 (NESG ID: PgR37A).** The 1D <sup>1</sup>H NMR line-broadening screen of PgR37A identified a total of 27 compounds that showed line-broadening. Of these compounds, 16 (59.3%) were confirmed as binders in the 2D <sup>1</sup>H, <sup>15</sup>N-HSQC screen [Figure 2.19A]. Seven of these compounds produce significant perturbations with acecainide being selected to generate a costructure [Figure 2.19B]. The perturbations map to a groove on the protein in which acecainide fits [Figure 2.19C]. This is highly conserved and slightly hydrophobic [Figure 2.19D].



**Figure 2.19** (A) Chemical structures of six compounds shown to bind PgR37A in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free PgR37A (black) and PgR37A bound with acecainide (red). (C) The PgR37A-acecainide costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the PgR37A-acecainide costructure: (*left*) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (*center*) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (*right*) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes PgR37A as a "conserved domain protein." Based on the results of the FAST-NMR approach, PgR37A may be an alcohol dehydrogenase, possibly a methanol dehydrogenase. A BLAST sequence similarity search identified mostly uncharacterized proteins, except for the beta-propeller domain from methanol dehydrogenase,<sup>85</sup> which has a similar domain structure as PgR37A. A 44.1% sequence identity was observed between PgR37A and the beta-propeller domain. The Pfam and InterPro also suggests PgR37A has a TPM domain, which is predicted to be involved in the photosystem II (PSII) repair cycle. PDBeFold indicates some structural similarity to alanyl-tRNA synthetases (3.14 Å RMSD) and glycerol dehydratase (4.16 Å RMSD). The CPASS similarity search found a formate dehydrogenase bound to NADPH as a promising match (27.24%) to the PgR37A-acecainide complex. Despite the low CPASS similarity, the binding site share very similar residues. One of the binders identified from the FAST-NMR screen is the strongly oxidizing riboflavin 5'-phosphate, which is capable of catalyzing a dehydrogenase reaction. The above evidence appears to support a dehydrogenase activity.

**2.3.17 *Rhodobacter sphaeroides* RHOS4\_12090 (NESG ID: RhR5).** The 1D <sup>1</sup>H NMR line-broadening screen of RhR5 identified a total of 12 compounds that showed line-broadening. However, only one of these compounds (8.3%) was confirmed as a binder in the 2D <sup>1</sup>H, <sup>15</sup>N-HSQC screen [Figure 2.20A]. This compound, 3,5-dinitrocatechol, did exhibit 7 significant perturbations [Figure 2.20B]. The majority of the perturbations form a consensus binding site [Figure 2.20C], which has a few conserved residues, a positive charge, and is hydrophobic [Figure 2.20D].



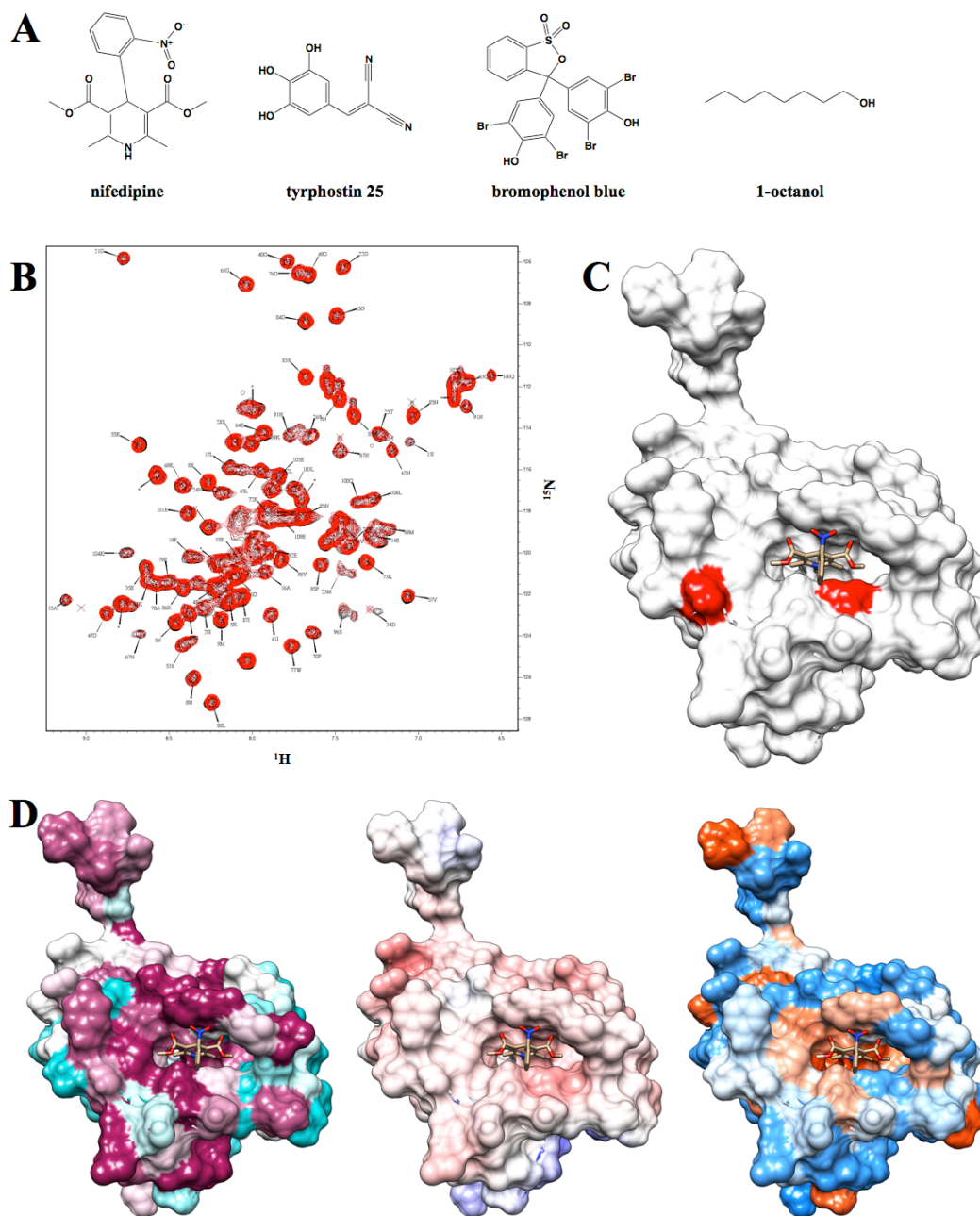
**Figure 2.20** (A) Chemical structures of one compound shown to bind RhR5 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free RhR5 (black) and RhR5 bound with 3,5-dinitrocatechol (red). (C) The RhR5-3,5-dinitrocatechol costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the RhR5-3,5-dinitrocatechol costructure: (*left*) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (*center*) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (*right*) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes RhR5 as a "putative uncharacterized protein." Based on the results of the FAST-NMR approach, RhR5 is likely a DNA-binding transcriptional

repressor involved in signal transduction. STRING identifies three proteins that have similar phylogenetic profiles: *ctrA* two component transcriptional regulator (signal transduction during cell cycle regulation); 2'-deoxycytidine 5'-triphosphate deaminase (deoxypyrimidine metabolism); and histidyl-tRNA synthetase (histidine transfer RNA biosynthesis). A BLAST sequence similarity search only identified uncharacterized proteins, but both Pfam and InterPro predict RhR5 to have a winged-helix-turn-helix transcription repressor DNA-binding domain, which is also supported by the presence of a positively charged surface. RhR5 also has significant structural similarity (1.5 Å - 2.5 Å RMSD) to several helix-turn-helix DNA-binding proteins. CPASS finds several high scoring hits (48%), where the top 5 hits all bind to a sugar phosphate, thus hinting at the DNA-binding properties.

**2.3.18 *Salmonella typhimurium* STM0327 (NESG ID: StR65).** The 1D  $^1\text{H}$  NMR line-broadening screen of StR65 identified a total of 13 compounds that showed line-broadening. Four of these compounds (30.8%) were confirmed as binders in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen [Figure 2.21A]. However, the perturbations for all of these compounds are very small [Figure 2.21B]. One of the two perturbations borders a cavity on the protein [Figure 2.21C], which is well conserved [Figure 2.21D].



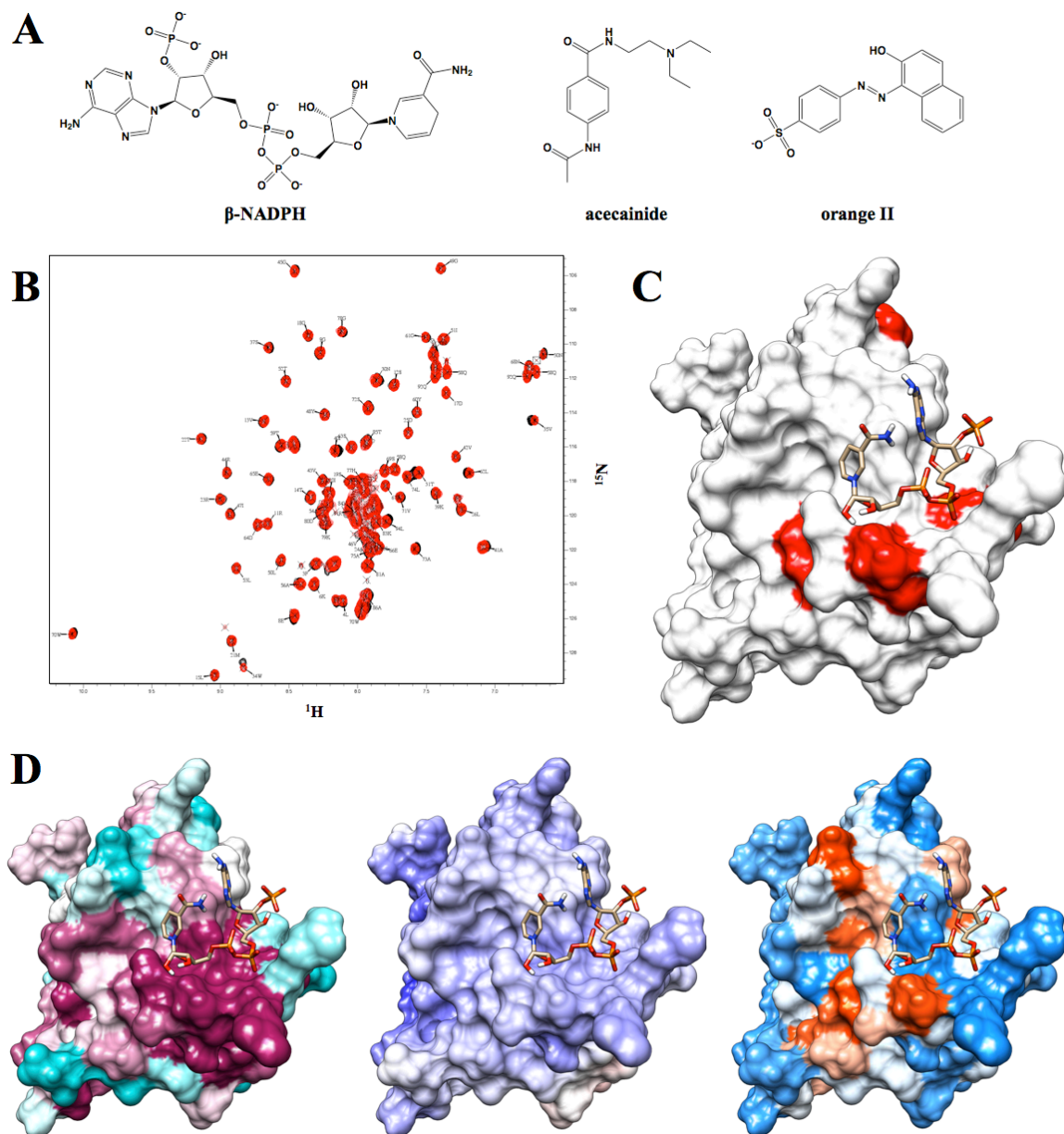


**Figure 2.21** (A) Chemical structures of four compounds shown to bind StR65 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free StR65 (black) and StR65 bound with nifedipine (red). (C) The StR65-nifedipine costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the StR65-nifedipine costructure: (*left*) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (*center*) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (*right*) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.



UniProtKB characterizes StR65 as a "putative cytoplasmic protein." Based on the results of the FAST-NMR approach, StR65 is probably related to permease transport or signaling. The only nearby gene to STM0327 is a putative permease. A BLAST sequence similarity search found only uncharacterized proteins, while a PDBeFold structure similarity search identified an oxygen detoxification protein (2.23 Å RMSD), a circadian clock protein (3.00 Å RMSD), and a regulatory protein involved in recombination (3.77 Å RMSD). CPASS identifies numerous proteins with the top hit being DNA repair and telomere maintenance protein bound to N-dimethyl-lysine (37.96%). However, the annotation of StR65 is very difficult due to the lack of supporting information and the weak binding of the compounds.

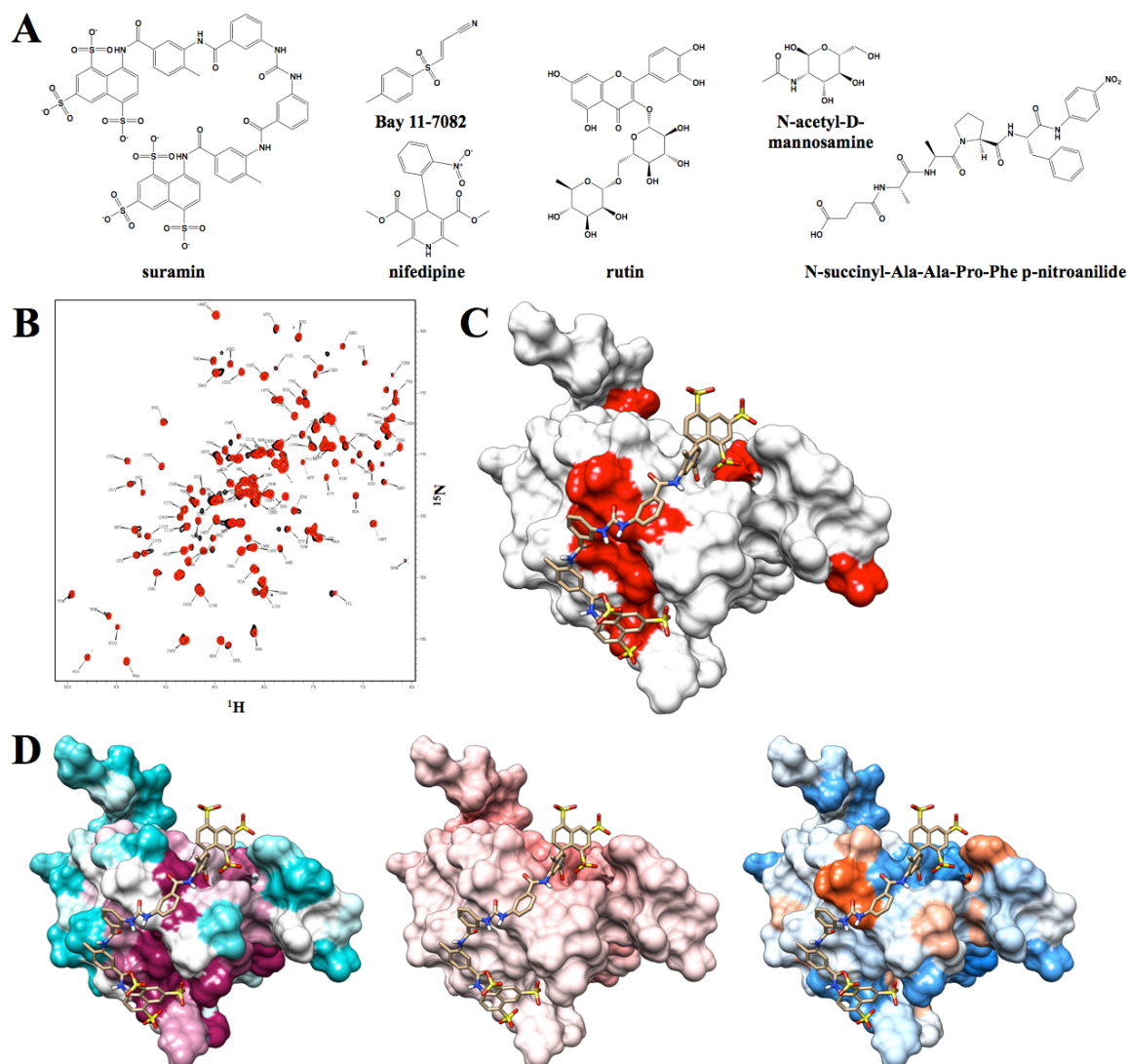
**2.3.19 *Silicibacter pomeroyi* SPO1678 (NESG ID: SiR5).** The 1D  $^1\text{H}$  NMR line-broadening screen of SiR5 identified a total of 22 compounds that showed line-broadening. Of these compounds, 13 (59.1%) were confirmed as binders in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen [Figure 2.22A].  $\beta$ -NADPH produces the greatest number of perturbations [Figure 2.22B], which when mapped to the protein surface highlight a binding groove [Figure 2.22C]. This groove is well conserved and has a slight positive charge [Figure 2.22D].



**Figure 2.22** (A) Chemical structures of three compounds shown to bind SiR5 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free SiR5 (black) and SiR5 bound with  $\beta$ -NADPH (red). (C) The SiR5- $\beta$ -NADPH costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the SiR5- $\beta$ -NADPH costructure: (left) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (center) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (right) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes SiR5 as a "putative uncharacterized protein." Based on the results of the FAST-NMR approach, SiR5 is probably a DNA-binding transcriptional regulator involved in signal transduction. The only nearby gene to SPO1678 is a two component transcriptional regulator, which is involved in signal transduction during . Three additional proteins have similar phylogenetic profiles: ctrA two component transcriptional regulator (signal transduction during cell cycle regulation); 2'-deoxycytidine 5'-triphosphate deaminase (deoxypyrimidine metabolism); histidyl-tRNA synthetase (histidine transfer RNA biosynthesis); and a flagellar protein. Interestingly, with the exception of the flagellar protein, the phylogenetic profiles for SiR5 are very similar to the results previously observed for the *Rhodobacter sphaeroides* RhR5 protein (section 2.3.17). A BLAST search demonstrated that SiR5 has 52% identity to RhR5 and 40% identity to transposases. InterPro and Pfam predict that SiR5 has a winged-helix-turn-helix transcription repressor DNA-binding domain, which is supported by the presence of the slightly positive charged surface. PDBeFold indicates that SiR5 also has significant structural similarity to the helix-turn-helix proteins, including RhR5 (2.29 Å RMSD). The CPASS similarity search does not find many high scoring hits. Like RhR5, SiR5 is likely a DNA-binding transcriptional repressor. However, the FAST-NMR screening results for SiR5 and RhR5 are significantly different in the types of ligands bound and the location of the binding site on the protein, despite the many similarities in the proteins. The binding sites of both proteins are both along the edge of a highly conserved region of the protein where the DNA would likely bind. Despite the differences, both RhR5 and SiR5 have enough similarities to suggest a similar function.

**2.3.20 *Staphylococcus saprophyticus* SSP0609 (NESG ID: SyR11).** The 1D  $^1\text{H}$  NMR line-broadening screen of SyR11 identified a total of 19 compounds that showed line-broadening. Of these compounds, 17 (89.5%) were confirmed as binders in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen [Figure 2.23A]. Suramin produced the greatest perturbations [Figure 2.23B]. The perturbed residues extend across the surface of the protein [Figure 2.23C], and represents residues that are well conserved [Figure 2.23D].



**Figure 2.23** (A) Chemical structures of six compounds shown to bind SyR11 in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. The entire list of binding ligands can be found in Appendix 2A. (B) An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free SyR11 (black) and SyR11 bound with suramin (red). (C) The SyR11-suramin costructure generated by AutoDock where residues with significant CSPs are colored red. (D) Surface representations of the SyR11-suramin costructure: (left) ConSurf residue conservation where highly conserved residues are magenta and poorly conserved residues are cyan; (center) Delphi electrostatics surface where the positively-charged surface is blue and negatively-charged surface is red; and (right) UCSF Chimera hydrophobicity surface where the hydrophilic surface is blue and the hydrophobic surface is orange.

UniProtKB characterizes SyR11 as a "putative secretory antigen." Based on the results of the FAST-NMR approach, SyR11 is likely involved in cell lysis by

metabolizing peptidoglycan linkages in the cell wall. STRING identifies a methicillin resistance protein as having a similar phylogenetic profile, which is also a protein involved in the formation of the peptidoglycan cell wall.<sup>86</sup> A BLAST search only finds a sequence similarity (70% identity) between SyR11 and secretory antigen SsaA in other *Staphylococcus* organisms. SsaA belong to the CHAP domain family, which are N-acetylmuramoyl-L-alanine amidases that function mainly in cell wall metabolism by hydrolyzing the link between amino acids and N-acetylmuramoyl (sugar group).<sup>87</sup> PDBeFold identifies a couple of structures that are similar to SyR11: a putative staphyloxanthin biosynthesis protein (2.79 Å RMSD) and a bifunctional glutathionylspermidine sythetase/amidase (3.13 Å RMSD). Both of these protein matches exhibit the CHAP domain. Additionally, the structure of N-acetylmuramoyl-L-alanine compound is very similar to N-acetyl-D-mannosamine, which was shown to bind in the FAST-NMR screen. CPASS does not find any similar binding sites above 30% due to the large size of the binding site, but there is a 1,4- $\beta$ -N-acetylmuramidase lysozyme bound with sucrose found as the second best hit (22.98%). This is interesting as 1,4- $\beta$ -N-acetylmuramidase lysozyme catalyzes the hydrolysis of N-acetylmuramic acid and N-acetyl-D-glucosamine residues in the peptidoglycan cell wall.

## 2.4 CONCLUSIONS

Determining the function of proteins for which sequence and structural homology have failed is a daunting task that normally would require a significant amount of time and resources to solve using standard approaches like gene knockouts, pull-down assays, or monitoring of expression levels. The FAST-NMR methodology provides a high-

throughput, tiered approach that can potentially annotate a protein over the span of a few days. The speed of this approach is of particular interest given the large number of "orphaned" proteins for which little is known. Even though the FAST-NMR approach does not necessarily provide definitive proof of a protein's function, it can provide a logical initial assumption that can be used to intelligently guide future studies by identifying the protein's binding site and the types of compounds that bind the protein.

The 20 NESG proteins screened in this study required the preparation of approximately 3,000 NMR samples and nearly a month of cumulative NMR experiment time. Each protein, by itself, could realistically go from receipt of a protein sample to a proposed function in approximately a week. However, much of that time can be reduced by automating the sample preparation and NMR data analysis. This truly makes FAST-NMR rapid enough to address the ever-growing number of "orphan" proteins.

## 2.5 REFERENCES

1. Galperin, M. Y. & Koonin, E. V. From complete genome sequence to 'complete' understanding? *Trends Biotechnol* **28**, 398–406 (2010).
2. Tucker, C. L. High-throughput cell-based assays in yeast. *Drug Discov Today* **7**, S125–30 (2002).
3. Lee, Y.-H. *et al.* Gene knockdown by large circular antisense for high-throughput functional genomics. *Nat Biotechnol* **23**, 591–599 (2005).
4. Joshi, T., Chen, Y., Becker, J. M., Alexandrov, N. & Xu, D. Genome-scale gene function prediction using multiple sources of high-throughput data in yeast *Saccharomyces cerevisiae*. *OMICS* **8**, 322–333 (2004).
5. del Val, C. *et al.* High-throughput protein analysis integrating bioinformatics and experimental assays. *Nucleic Acids Res* **32**, 742–748 (2004).
6. Powers, R., Mercier, K. A. & Copeland, J. C. The application of FAST-NMR for the identification of novel drug discovery targets. *Drug Discov Today* **13**, 172–179 (2008).
7. Mercier, K. A. *et al.* FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* **128**, 15292–15299 (2006).
8. Laurie, A. T. & Jackson, R. M. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Curr*

- Protein Pept Sci* **7**, 395–406 (2006).
9. Blundell, T. L. *et al.* Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. *Philos Trans R Soc Lond, B, Biol Sci* **361**, 413–423 (2006).
  10. Vajda, S. & Guarnieri, F. Characterization of protein-ligand interaction sites using experimental and computational methods. *Curr Opin Drug Discov Devel* **9**, 354–362 (2006).
  11. Mercier, K. A., Germer, K. & Powers, R. Design and characterization of a functional library for NMR screening against novel protein targets. *Comb Chem High Throughput Screen* **9**, 515–534 (2006).
  12. Mercier, K. A., Shortridge, M. D. & Powers, R. A multi-step NMR screen for the identification and evaluation of chemical leads for drug discovery. *Comb Chem High Throughput Screen* **12**, 285–295 (2009).
  13. Stark, J. L. & Powers, R. Rapid protein-ligand costructures using chemical shift perturbations. *J Am Chem Soc* **130**, 535–545 (2008).
  14. Powers, R. *et al.* Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **65**, 124–135 (2006).
  15. Powers, R., Copeland, J. & Stark, J. L. Searching the protein structure database for ligand-binding site similarities using CPASS v. 2. *BMC Res Notes* (2011).
  16. Park, K. & Kim, D. Binding similarity network of ligand. *Proteins* **71**, 960–971 (2008).
  17. Mercier, K. A. *et al.* Structure and function of *Pseudomonas aeruginosa* protein PA1324 (21-170). *Protein Sci* **18**, 606–618 (2009).
  18. Shortridge, M. D. & Powers, R. Structural and functional similarity between the bacterial type III secretion system needle protein PrgI and the eukaryotic apoptosis Bcl-2 proteins. *PLoS ONE* **4**, e7442 (2009).
  19. Rossi, P. *et al.* 1H, 13C, and 15N resonance assignments for the protein coded by gene locus BB0938 of *Bordetella bronchiseptica*. *J Biomol NMR* **33**, 197 (2005).
  20. Aramini, J. M. *et al.* Solution NMR structure of *Escherichia coli* ytfP expands the structural coverage of the UPF0131 protein domain family. *Proteins* **68**, 789–795 (2007).
  21. Aramini, J. M. *et al.* Resonance assignments for the hypothetical protein yggU from *Escherichia coli*. *J Biomol NMR* **27**, 285–286 (2003).
  22. Singarapu, K. K. *et al.* NMR structure of protein yjbR from *Escherichia coli* reveals ‘double-wing’ DNA binding motif. *Proteins* **67**, 501–504 (2007).
  23. Aramini, J. M. *et al.* Solution NMR structure of the NlpC/P60 domain of lipoprotein Spr from *Escherichia coli*: structural evidence for a novel cysteine peptidase catalytic triad. *Biochemistry* **47**, 9715–9717 (2008).
  24. Rossi, P. *et al.* Structural elucidation of the Cys-His-Glu-Asn proteolytic relay in the secreted CHAP domain enzyme from the human pathogen *Staphylococcus saprophyticus*. *Proteins* **74**, 515–519 (2009).
  25. Mercier, K. A. & Powers, R. Determining the optimal size of small molecule mixtures for high throughput NMR screening. *J Biomol NMR* **31**, 243–258 (2005).
  26. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6**, 277–293 (1995).
  27. Vranken, W. F. *et al.* The CCPN data model for NMR spectroscopy: development



- of a software pipeline. *Proteins* **59**, 687–696 (2005).
28. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612 (2004).
  29. Morris, G., Goodsell, D., Halliday, R. & Huey, R. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**, 1639–1662 (1998).
  30. Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* **28**, 1145–1152 (2007).
  31. Morris, G. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* (2009). doi:10.1002/jcc.21256
  32. Sanner, M. F. Python: a programming language for software integration and development. *J Mol Graph Model* **17**, 57–61 (1999).
  33. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* **40**, D742–53 (2012).
  34. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
  35. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).
  36. Dundas, J. *et al.* CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res* **34**, W116–8 (2006).
  37. Wass, M. N. & Sternberg, M. J. E. ConFunc--functional annotation in the twilight zone. *Bioinformatics* **24**, 798–806 (2008).
  38. Hunter, S. *et al.* InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res* **40**, D306–D312 (2012).
  39. Forslund, K. & Sonnhammer, E. L. L. Predicting protein function from domain content. *Bioinformatics* **24**, 1681–1687 (2008).
  40. Kelley, L. A. & Sternberg, M. J. E. Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protocols* **4**, 363–371 (2009).
  41. Wass, M. N., Kelley, L. A. & Sternberg, M. J. E. 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res* **38**, W469–73 (2010).
  42. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* **38**, W529–33 (2010).
  43. Honig, B. & Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **268**, 1144–1149 (1995).
  44. Chitale, M., Hawkins, T., Park, C. & Kihara, D. ESG: extended similarity group method for automated protein function prediction. *Bioinformatics* **25**, 1739–1745 (2009).
  45. Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res* **40**, D841–6 (2012).
  46. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, D109–14 (2012).

47. Fischer, M. *et al.* MarkUs: a server to navigate sequence-structure-function space. *Nucleic Acids Res* **39**, W357–61 (2011).
48. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D* **60**, 2256–2268 (2004).
49. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res* **40**, D290–D301 (2012).
50. Triplet, T. *et al.* PROFESS: a PROtein function, evolution, structure and sequence database. *Database* **2010**, baq011 (2010).
51. Laskowski, R. A., Watson, J. D. & Thornton, J. M. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* **33**, W89–93 (2005).
52. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
53. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **10**, 980 (2003).
54. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**, D808–15 (2013).
55. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **40**, D71–5 (2012).
56. Serizawa, M. *et al.* Systematic analysis of SigD-regulated genes in *Bacillus subtilis* by DNA microarray and Northern blotting analyses. *Gene* **329**, 125–136 (2004).
57. Chen, Y. F. & Helmann, J. D. The *Bacillus subtilis* flagellar regulatory protein sigma D: overproduction, domain analysis and DNA-binding properties. *J Mol Biol* **249**, 743–753 (1995).
58. Cohen, S. P., Hächler, H. & Levy, S. B. Genetic and functional analysis of the multiple antibiotic resistance (mar) locus in *Escherichia coli*. *J Bacteriol* **175**, 1484–1492 (1993).
59. Hong, M., Fuangthong, M., Helmann, J. D. & Brennan, R. G. Structure of an OhrR-ohrA operator complex reveals the DNA binding mechanism of the MarR family. *Mol Cell* **20**, 131–141 (2005).
60. Wilkinson, S. P. & Grove, A. Ligand-responsive transcriptional regulation by members of the MarR family of winged helix proteins. *Curr Issues Mol Biol* **8**, 51–62 (2006).
61. Carbone, V., Hara, A. & El-Kabbani, O. Structural and functional features of dimeric dihydrodiol dehydrogenase. *Cell Mol Life Sci* **65**, 1464–1474 (2008).
62. Tam, L. T. *et al.* Differential gene expression in response to phenol and catechol reveals different metabolic activities for the degradation of aromatic compounds in *Bacillus subtilis*. *Environ Microbiol* **8**, 1408–1427 (2006).
63. Kawai, Y., Moriya, S. & Ogasawara, N. Identification of a protein, YneA, responsible for cell division suppression during the SOS response in *Bacillus subtilis*. *Mol Microbiol* **47**, 1113–1122 (2003).
64. Aramini, J. M. *et al.* Solution NMR structure of the SOS response protein YnzC from *Bacillus subtilis*. *Proteins* **72**, 526–530 (2008).
65. Ni, N., Choudhary, G., Li, M. & Wang, B. Pyrogallol and its analogs can antagonize bacterial quorum sensing in *Vibrio harveyi*. *Bioorg Med Chem Lett* **18**,

- 1567–1572 (2008).
66. Branda, S. S. *et al.* Genes involved in formation of structured multicellular communities by *Bacillus subtilis*. *J Bacteriol* **186**, 3970–3979 (2004).
  67. Aihara, H., Ito, Y., Kurumizaka, H., Yokoyama, S. & Shibata, T. The N-terminal domain of the human Rad51 protein binds DNA: structure and a DNA binding surface as revealed by NMR. *J Mol Biol* **290**, 495–504 (1999).
  68. Schade, M. *et al.* The solution structure of the Zalpha domain of the human RNA editing enzyme ADAR1 reveals a prepositioned binding surface for Z-DNA. *Proc Natl Acad Sci USA* **96**, 12465–12470 (1999).
  69. Athanasiadis, A. Zalpha-domains: at the intersection between RNA editing and innate immunity. *Semin. Cell Dev. Biol.* **23**, 275–280 (2012).
  70. Ontoria, J. M. *et al.* The design and enzyme-bound crystal structure of indoline based peptidomimetic inhibitors of hepatitis C virus NS3 protease. *J Med Chem* **47**, 6443–6446 (2004).
  71. Anantharaman, V. & Aravind, L. MOSC domains: ancient, predicted sulfur-carrier domains, present in diverse metal-sulfur cluster biosynthesis proteins including Molybdenum cofactor sulfurases. *FEMS Microbiol Lett* **207**, 55–61 (2002).
  72. Llewellyn, N. M., Li, Y. & Spencer, J. B. Biosynthesis of butirosin: transfer and deprotection of the unique amino acid side chain. *Chem Biol* **14**, 379–386 (2007).
  73. Oakley, A. J., Coggan, M. & Board, P. G. Identification and characterization of gamma-glutamylamine cyclotransferase, an enzyme responsible for gamma-glutamyl-epsilon-lysine catabolism. *J Biol Chem* **285**, 9642–9648 (2010).
  74. Santos, J. M., Freire, P., Vicente, M. & Arraiano, C. M. The stationary-phase morphogene *bolA* from *Escherichia coli* is induced by stress during early stages of growth. *Mol Microbiol* **32**, 789–798 (1999).
  75. Santos, J. M., Lobo, M., Matos, A. P. A., De Pedro, M. A. & Arraiano, C. M. The gene *bolA* regulates *dacA* (PBP5), *dacC* (PBP6) and *ampC* (AmpC), promoting normal morphology in *Escherichia coli*. *Mol Microbiol* **45**, 1729–1740 (2002).
  76. Skarzynski, T., Kim, D. H., Lees, W. J., Walsh, C. T. & Duncan, K. Stereochemical course of enzymatic enolpyruvyl transfer and catalytic conformation of the active site revealed by the crystal structure of the fluorinated analogue of the reaction tetrahedral intermediate bound to the active site of the C115A mutant of MurA. *Biochemistry* **37**, 2572–2577 (1998).
  77. Malinverni, J. C. & Silhavy, T. J. An ABC transport system that maintains lipid asymmetry in the gram-negative outer membrane. *Proc Natl Acad Sci USA* **106**, 8009–8014 (2009).
  78. Finnin, M. S., Hoffman, D. W. & White, S. W. The DNA-binding domain of the MotA transcription factor from bacteriophage T4 shows structural similarity to the TATA-binding protein. *Proc Natl Acad Sci USA* **91**, 10972–10976 (1994).
  79. Li, N., Sickmier, E. A., Zhang, R., Joachimiak, A. & White, S. W. The MotA transcription factor from bacteriophage T4 contains a novel DNA-binding domain: the ‘double wing’ motif. *Mol Microbiol* **43**, 1079–1088 (2002).
  80. Page, A.-L. & Parsot, C. Chaperones of the type III secretion pathway: jacks of all trades. *Mol Microbiol* **46**, 1–11 (2002).
  81. Wang, X. *et al.* Cryptic prophages help bacteria cope with adverse environments. *Nat Comm* **1**, 147 (2010).

82. Beckett, D., Kovaleva, E. & Schatz, P. J. A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. *Protein Sci* **8**, 921–929 (1999).
83. Weitz, D. *et al.* Functional and structural characterization of a prokaryotic peptide transporter with features similar to mammalian PEPT1. *J Biol Chem* **282**, 2832–2839 (2007).
84. Anantharaman, V. & Aravind, L. Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes. *Genome Biol* **4**, R11 (2003).
85. Anthony, C. The quinoprotein dehydrogenases for methanol and glucose. *Archives of Biochemistry and Biophysics* **428**, 2–9 (2004).
86. Benson, T. E. *et al.* X-ray crystal structure of *Staphylococcus aureus* FemA. *Structure* **10**, 1107–1115 (2002).
87. Bateman, A. & Rawlings, N. D. The CHAP domain: a large family of amidases including GSP amidase and peptidoglycan hydrolases. *Trends Biochem Sci* **28**, 234–237 (2003).

**APPENDIX 2A** A list of the results of the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen for each protein of the compound hits found during the 1D line-broadening screen. The compounds are categorized as 2D binders if any peaks were perturbed in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC. The compounds are categorized as 2D non-binders if no peaks are perturbed. If the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra do not represent a well-folded protein, the compound is classified as 2D precipitation/aggregation.

<b>Gene ID (NESG ID)</b>	<b>2D Binders</b>	<b>2D Non-Binders</b>	<b>2D Precipitation / Aggregation</b>
<b>yjcQ (SR346)</b>	tetracycline	pyridoxin (vitamin B6)	
	kanamycin	pepstatin A	
	lorglumide	L-histadine	
	(±)-epinephrine	aquacobalamin	
	penicillin G	citrate	
	allopurinol	hydralazine	
	L-2-aminoadipic acid		
	berenil		
	lidocaine		
	Bay 11-7085		
	bepiridil		
	histamine		
	adenine (vitamin B4)		
<b>ykvR (SR358)</b>	histamine	citrate	
	haloperidol metabolite I		
	3,5- dinitrocatechol		
	L-2-aminoadipic acid		
	allopurinol		
	adenine (vitamin B4)		
	dansylglycin		
<b>ynzC (SR384)</b>	4,6- dinitropyrogallol	penicillin G	suramin
	orange II	adrenochrome	bromocresol green

	triethylenetetramine trientine	tyrphostin 1	
	(-)-riboflavin	idazoxan	
	4-amino-2-nitrophenol	acetoacetate	
	tyrphostin 25		
	4-methylpyrazole		
	2-amino-phenol		
	imidazole		
<b>yozE (SR391)</b>	histamine	citrate	bromophenol blue
	4-methylpyrazole	2-amino-4-methylphenol	
	chlorotetracycline	berenil	
	3,5-dinitrocatechol	clofibrate	
	UTP		
	N-succinyl-Ala-Ala-Pro-Phe p-nitroanilide		
	AMP		
<b>BVU_3908 (BvR153)</b>	imidazole	6,7-dimethyl-5,6,7,8-tetrahydropterine	
	S-(-)-carbidopa	(±)-epinephrine	
	(-)-riboflavin	(2S,3S)-trans-3-phenyl-2-oxiranylmethyl-4-nitrophenyl carbonate	
		dimethyl-2-oxoglutarate	
		2-aminofluorene	
		nalidixic acid	
		Mordant orange I	
		lumicolchicine	
		PTH-tryptophan	
		flavanone	
		tyrphostin 1	
		2-deoxyguanosine-5-monophosphate	
		nitrendipine	
		Bay 11-7082	

<b>BB0938 (BoR11)</b>	methiothepin	$\beta$ -NADPH	suramin
	3,5-dinitrocatechol	6-amino-3-methylpurine	bisbenzimidazole H 33258
	ethacridine	N-tert-butyltrimethylsilyl-N-methyltrifluoroacetamide	
	4-chloromercuribenzoic acid	adrenochrome	
	(-)-riboflavin	ebselen	
	acecainide		
	idazoxan		
	S-(-)-carbidopa		
	L-valine		
	2-pyridineacetic acid		
	tyrphostin 1		
	nifedipine		
<b>CC_0527 (CcR55)</b>	ciprofloxacin	trans-chalcone	
	N-p-tosyl-L-phenylalanine chloromethyl ketone		
	nalidixic acid		
	bromocresol green		
	ATP		
	( $\pm$ )-6-methyl-5,6,7,8-tetrahydropterine		
	Mordant orange I		
	suramin		
	duroquinone		
	naproxen		
	phenylbutazone		
	N-		

	phenylanthranilic acid		
	8-methoxypsoralen		
	cAMP		
	chelerythrine		
	6-amino-3-methylpurine		
	picotamide		
	captopril		
	penicillin G		
	idazoxan		
	ebselen		
<b>ytfP (ER111)</b>	phenol red	aminophylline	
	orange II	berenil	
	L-glutamate	ciprofloxacin	
	novobiocin		
	lidocaine		
	rifampicin		
	progesterone		
	buspirone		
	kaempferol		
	L-2-aminoadipic acid		
	eserine (physostigmine)		
	N-acetyl-L-tryptophan 3,5-bis(trifluoromethyl) benzyl ester		
	tetracycline		
	amoxicillin		
	adenine (vitamin B4)		
	ethacridine		
<b>yrbA (ER115)</b>	ATP	apigenin	Bay 11-7085
		lumicolchicine	



		clofibrate	
		adenine (vitamin B4)	
		mycophenolic acid	
		ebselen	
		nifedipine	
		2-cyclohexen-1-one	
<b>yggU (ER14)</b>	bromocresol green	chelerythrine	suramin
	Mordant orange I	ciprofloxacin	
	$\beta$ -NADPH		
	ATP		
	colchicine		
	8-methoxypsoralen		
	imidazole		
	penicillin G		
	( $\pm$ )-6-methyl-5,6,7,8-tetrahydropterine		
	safrole		
	bisbenzimidazole H 33258		
<b>yjbR (ER226)</b>	bromophenol blue	2,6-diisopropylphenol	doxycycline
	bithionol	nalidixic acid	suramin
	tetracycline	adenine (vitamin B4)	
	X-GAL	1-octanol	
	CTP	( $\pm$ )-camphor	
	bepiridil	1-methylhistamine	
	thiamine pyrophosphate	cAMP	
	histamine		
	flutamide		
	lumicolchicine		
	chlorotetracycline		
	cephalexin		
	( $\pm$ )-verapamil		
	3-(1-naphthyl)-D-alanine		

	homovanillic acid		
	PTH-tryptophan		
	menadione (vitamin K3)		
<b>ydfO (ER251)</b>	DMCM	$\beta$ -NADPH	indomethacin
	N-phenylanthranilic acid	cyclo(His-Pro)	sulindac sulfide
	phenylpyruvic acid	L-tyrosine	meclofenamic acid
	kynurenic acid	6,7-dimethyl-5,6,7,8-tetrahydropterine	bromocresol green
	ampicillin	acetoacetate	diclofenac
	nifedipine		adrenochrome
	acecainide		orange II
	daidzen		
	progesterone		
	Mordant orange I		
	Bay 11-7082		
	S-(4-nitrobenzyl)-6-thioinosine		
	4,6-dinitropyrogallol		
	resveratrol		
	genistein		
	doxycycline		
	( $\pm$ )-camphor		
	trans-chalcone		
	menadione (vitamin K3)		
	S-(-)-carbidopa		
	cromolyn		
	tyrphostin 25		
	( $\pm$ )-6-methyl-5,6,7,8-tetrahydropterine		
	(-)-riboflavin		

<b>ygdR (ER382A)</b>	suramin	daphnetin	L-histadine
		CTP	tyrphostin 25
		1-phenyl-1-cyclopropanecarboxylic acid	
		histamine	
		(-)-cotinine	
		2-deoxyguanosine-5-monophosphate	
		1-methylimidazole	
		timolol	
		mecamylamine	
		AMP	
		Bay 11-7085	
		nifedipine	
		bepiridil	
		aquacobalamin	
<b>ykfF (ER397)</b>	novobiocin	nalidixic acid	
	ethacridine	L-histadine	
	allopurinol	L-2-aminoadipic acid	
	lidocaine	tyrphostin 1	
	3-indoleacetic acid	cyclo(His-Pro)	
	clindamycin		
	amoxicillin		
	lorglumide		
	bepiridil		
<b>yeiV (ER541)</b>	(±)-6-methyl-5,6,7,8-tetrahydropterine	acetoacetate	coenzyme A
	duroquinone	imidazole	β-NADPH
	tyrphostin 25	(2S,3S)-trans-3-phenyl-2-oxiranylmethyl-4-nitrophenyl carbonate	Mordant orange I
	hydralazine	ebselen	orange II
	menadione	6-amino-3-methylpurine	sulindac sulfide

	(vitamin K3)		
	ampicillin		
	S-(-)-carbidopa		
	adrenochrome		
	safrole		
	cyclo(His-Pro)		
	N-p-tosyl-L-phenylalanine chloromethyl ketone		
	Bay 11-7082		
	dimethyl 2-oxoglutarate		
<b>PG_0361 (PgR37A)</b>	acecainide	4-methylpyrazole	
	bromocresol green	Bay 11-7085	
	riboflavin 5'-phosphate	serotonin	
	ethacridine	adenine (vitamin B4)	
	tyrphostin 25	(±)-6-methyl-5,6,7,8-tetrahydropterine	
	N-succinyl-Ala-Ala-Pro-Phe p-nitroanilide	(±)-epinephrine	
	ATP	nalidixic acid	
	histamine	S-(-)-carbidopa	
	methiothepin	ebselen	
	1-methylimidazole	nifedipine	
	AMP	flavanone	
	2'-deoxyadenosine 5'-monophosphate		
	phenylpyruvic acid		
	(-)-riboflavin		
	Bay 11-7082		
<b>RHOS4_1209</b>	3,5-	penicillin G	

<b>0 (RhR5)</b>	dinitrocatechol		
		1-methylhistamine	
		berenil	
		bepiridil	
		adenine (vitamin B4)	
		6-amino-3-methylpurine	
		hydralazine	
		lorglumide	
		allopurinol	
		lidocaine	
		cyclo(His-Pro)	
<b>STM0327 (StR65)</b>	nifedipine	pepstatin A	suramin
	1-octanol	L-glutamate	ebselen
	tyrphostin 25	2,6-diisopropylphenol	
	bromophenol blue	aquacobalamin	
		clofibrate	
		acridine	
		duroquinone	
<b>SPO1678 (SiR5)</b>	$\beta$ -NADPH	riboflavin-5'-phosphate	sulindac sulfide
	acecainide	ebselen	suramin
	orange II	6,7-dimethyl-5,6,7,8-tetrahydropterine	
	nifedipine	N-p-tosyl-L-phenylalanine chloromethyl ketone	
	Bay 11-7082	PTH-tryptophan	
	2-aminophenol	adrenochrome	
	hydralazine	penicillin G	
	menadione (vitamin K3)		
	duroquinone		
	4-amino-2-nitrophenol		
	S-(-)-carbidopa		



## CHAPTER 3

### RAPID PROTEIN-LIGAND COSTRUCTURES USING CHEMICAL SHIFT PERTURBATIONS AND AUTODOCK<sup>§</sup>

#### 3.1 INTRODUCTION

Structure-based drug design utilizes the known three-dimensional structures of biologically relevant proteins to develop drug candidates in a rational yet relatively rapid manner.<sup>1,2</sup> However, this process requires an in-depth understanding of the molecular processes that govern the interaction between a target protein and a potential drug. Knowledge of the precise location and orientation, or pose, of the drug molecule when bound to the protein-ligand complex can be experimentally determined using X-ray crystallography or NMR, but these techniques require a significant amount of time, usually on the order of weeks to months.<sup>3,4</sup> A number of NMR approaches have been described to shorten this time frame that includes NOE based protein-ligand models,<sup>5,6</sup> differential chemical shift perturbations between two or more bound ligands,<sup>7</sup> SOS-NMR,<sup>8</sup> and NMR-DOC.<sup>9</sup> These approaches still suffer from significant experimental drawbacks that limit their practical use to routine determination of a large number of protein-ligand costructures. In order to facilitate the high-throughput screening of thousands of compounds, the application of molecular docking simulations for filtering and evaluating drug candidates is a common alternative.<sup>10</sup>

Molecular docking is the process of predicting the structure of a protein-ligand complex using only the structures of the individual components. Most molecular docking

---

<sup>§</sup> Chapter was adapted from Stark, J. and Powers, R. Rapid protein-ligand costructures using chemical shift perturbations. *J Am Chem Soc* **130**, 535-545 (2008). Reprinted with permission, copyright 2012 by the American Chemical Society.

software applications have two key parts: (1) a search algorithm that samples different locations and conformations of the ligand with respect to the protein and (2) a scoring method to evaluate the results of the search algorithm.<sup>11</sup> For molecular docking to be useful in drug discovery, these key parts should be both fast and accurate. These two requirements are often in opposition to each other, requiring necessary compromises that commonly end in ambiguous results or failure.<sup>12-16</sup> There are numerous molecular docking software applications that utilize different search and scoring algorithms, where AutoDock is currently the most cited of these applications<sup>17</sup> and has been demonstrated to outperform other docking tools in a virtual screen of a compound library.<sup>15</sup>

In AutoDock 4,<sup>18,19</sup> the protein is represented as a three-dimensional grid which is searched with a Lamarckian genetic algorithm that explores the different translational, rotational, and torsional degrees of freedom of the ligand relative to the grid. An estimated free energy of binding is used to evaluate the docked ligand conformations and comprises several terms that include dispersion/repulsion, directional hydrogen bonding, electrostatics, desolvation, and conformational energy. As a result of the searching algorithm, the accuracy of an AutoDock calculation is often dependent on the number of torsional degrees of freedom in the ligand and the size of the grid that represents the protein or the binding site.<sup>16,20</sup> The accuracy can be improved by increasing both the population size and the number of energy evaluations for the Lamarckian genetic algorithm.<sup>16,21</sup> Unfortunately, these modifications often lead to a drastic increase in computational time (tens of hours) that significantly reduces the throughput required for iterative structure-based drug design. Furthermore, increasing these parameters does not



guarantee that the lowest-energy conformer predicted by AutoDock will result in a correct protein-ligand model.

Prior knowledge of the ligand binding site would potentially improve the accuracy of the docking calculations by minimizing the grid volume that must be searched as well as limiting the possible conformations of the ligand that have energetically favorable interactions with the protein.<sup>16</sup> One rapid method of locating the binding site is by identifying the amino acid residues that experience chemical shift perturbations (CSPs) in a 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR titration experiment due to their proximity to the bound ligand.<sup>7,22,23</sup>

Chemical shift perturbations (CSPs) can also be used to filter the docking results by selecting a pose consistent with the observed chemical shift changes.<sup>24</sup> The protein-protein docking program HADDOCK<sup>25</sup> uses CSPs and mutagenesis to create ambiguous interaction restraints, which define an upper boundary for the distance one residue can be from any atom of the bound molecule. These restraints are combined with a complete set of structural restraints that define the protein-free conformation in a simulated annealing protocol using CNS<sup>26</sup> to calculate a costructure. A similar approach can be used to provide criteria to select the best ligand conformation(s) generated from an AutoDock calculation.

Our approach for rapidly determining an accurate ligand binding orientation utilizes CSPs from a 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR experiment to both guide and filter an AutoDock costructure calculation. By using CSPs to define the likely ligand binding site, the AutoDock 3D grid is reduced to a volume encompassing only the binding site, thus decreasing the search space sampled by the ligand. Furthermore, an NMR energy

function based on the magnitude of CSPs is shown to be an effective filtering tool to select the best ligand conformation.

### 3.2 MATERIALS AND METHODS

**3.2.1 Preparation of the ligand and target protein.** The analysis of the reliability of using CSPs to guide and filter molecular docking was demonstrated with the X-ray structures for 19 distinct protein-ligand complexes (Table 3.1) present in the Protein Data Bank (<http://www.pdb.org>).<sup>27,28</sup> The ligands were removed from the protein-ligand complex and saved as a separate coordinate file. All solvent molecules and ions were also removed with the exception of ions deemed to be biologically relevant to ligand binding. Any missing heavy atoms for the amino acid residues were added using Swiss-PDBView (<http://www.expasy.org/spdbv>).<sup>29</sup> All hydrogens were added to the protein and ligand using standard protonation states at a neutral pH.

Docking was also performed using the corresponding unbound structures for each of the 19 protein-ligand complexes. This permitted a comparison of the docking performance using both bound and unbound protein structures. Protein files were prepared in the same manner as above. In addition, the backbone coordinates of the apoprotein were aligned with the bound protein structure prior to the AutoDock calculation (Table 3.1). The ligand conformation in the original X-ray structure of the complex was then used to measure a root mean square deviation (RMSD) between the docked ligand conformers calculated using the bound and apoprotein structures.

**Table 3.1** RMSD comparison between the ligand-bound and unbound proteins.

PDB ID bound/unbound	RMSD(Å)			
	full protein backbone	binding site backbone	binding site all atom	resolution
1A6W/1A6U	0.33	0.29	0.77	2.00/2.10
1ACJ/1QIF <sup>30,31</sup>	0.38	0.30	0.63	2.80/2.10
1BLH/1DJB <sup>32,33</sup>	0.25	0.20	1.39	2.30/2.10
1BYB/1BYA <sup>34</sup>	0.29	2.48	2.36	1.90/2.20
1C83/1SUG <sup>35,36</sup>	0.23	0.19	0.77	1.83/1.95
1IVD/1NNA <sup>37,38</sup>	1.04	0.54	0.82	1.90/2.50
1LPC/1LP8 <sup>39</sup>	0.14	0.27	0.58	1.70/1.65
MRG/1AHC <sup>40,41</sup>	0.27	0.17	1.15	1.80/2.00
MTW/2TGA <sup>42,43</sup>	0.34	0.93	1.10	1.90/1.80
QPE/3LCK <sup>44,45</sup>	0.25	0.31	0.40	2.00/1.70
RBP/1BRQ <sup>46,47</sup>	0.59	0.73	1.55	2.00/2.50
SNC/1STN <sup>48,49</sup>	0.67	0.85	2.09	1.65/1.70
1STP/2RTA <sup>50,51</sup>	0.77	0.40	1.11	2.60/1.39
2CTC/2CTB <sup>52</sup>	0.17	0.38	1.72	1.40/1.50
1H4N/2CBA <sup>53,54</sup>	0.21	0.17	0.26	1.90/1.54
1PK4/1KRN <sup>55,56</sup>	0.50	0.25	1.10	2.25/1.67
2SIM/2SIL <sup>57</sup>	0.14	0.16	0.23	1.60/1.60
3PTB/2PTN <sup>43,58</sup>	0.11	0.16	0.31	1.70/1.55
5CPA/5CPA <sup>59,60</sup>	0.36	0.52	1.68	2.00/1.54

### 3.2.2 Prediction of ligand binding sites and chemical shift perturbations.

The NMR-predicted binding site for each protein complex was determined by identifying all of the amino acid residues within 6.0 Å of any atom in the ligand using RasMol 2.7.3.1.<sup>61</sup> These residues were anticipated to incur a chemical shift perturbation when the protein is titrated with the ligand. The coordinates of the residues that composed this binding site were then saved as a separate structure file that was used to define the grid size for the

guided docking. Chemical shift perturbations were then estimated using a simple linear relationship based on the distance between the amide nitrogen for each residue in the binding site to the nearest ligand atom.

**3.2.3 Molecular docking.** AutoDock 4.01<sup>18,19</sup> with the AutoDockTools 1.4.5 (<http://mgltools.scripps.edu>) graphical interface was used to simulate 120 different binding conformations for each protein-ligand pair. In the analysis where the docking was not guided by the NMR-predicted binding site (blind-docking), grid maps were generated with 0.547 spacing and set to an appropriate size that encompasses the entire protein. The CSP-guided docking analysis also used the 0.547 Å spacing, but the grid map size was set to encompass those amino acid residues that were determined to be within 6.0 Å of the ligand. The docking calculations were performed using the Lamarckian genetic algorithm default settings with a population size of 300 and 500,000 energy evaluations. The AutoDock calculations took, on average,  $37 \pm 32$  min per protein-ligand pair to complete on an Intel Xeon 3.06 GHz dual processor Linux workstation. The calculation time increased proportionally with the number of rotatable bonds in the ligand.

**3.2.4 Filtering of docked ligand conformations.** The resulting 120 docked ligand conformations were filtered using our AutoDockFilter (ADF) program, which utilized the magnitude of the chemical shift perturbations to select the best conformers instead of relying on the ambiguity inherent in choosing the best cluster based solely on the AutoDock empirical binding energy.

ADF calculates a pseudodistance ( $d_{\text{CSP}}$ ) based on the magnitude of the NH chemical shift perturbations for each residue in a  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR experiment. We

assumed linear relationships are present between the magnitudes of the CSPs and the distances to the nearest ligand atom. Also, the shortest possible CSP pseudodistance allowed is 3 Å. This minimizes any bias to large chemical shift changes that may result from multiple factors in addition to proximity of the ligand. This pseudodistance is then compared to the shortest distance ( $d_s$ ) between any atom in the residue that incurred an NH CSP and any atom in each docked ligand conformer. A violation energy is attributed to the conformer only when the shortest distance in the docked protein-ligand costructure is larger than the pseudodistance predicted from CSPs. Thus, the pseudodistance based on CSP only represents an upper distance boundary. The violation energy is summed for each separate CSP to generate an overall NMR energy ( $E_{\text{NMR}}$ ):

$$E_{\text{NMR}} = k \sum_{i=1}^n (\Delta_{\text{Dist}})^2 \quad (3.1)$$

where

$$\Delta_{\text{Dist}} = \begin{cases} d_{\text{CSP}} - d_s & d_{\text{CSP}} < d_s \\ 0 & d_s \leq d_{\text{CSP}} \end{cases} \quad (3.2)$$

An RMSD is then calculated by ADF between each docked ligand conformation relative to the ligand with the lowest NMR energy. The structures are then clustered based on this RMSD value using the *k*-means method.<sup>62</sup> If a particular docked conformation has an NMR energy that is beyond two standard deviations from the average, it is excluded from the cluster. Only the best cluster from ADF was used for further analysis. The graphical representations of the proteins and ligands in this paper were prepared using VMD Molecular Graphics Viewer.<sup>63</sup>

**3.2.5 Molecular docking using a flexible binding site.** A new feature in AutoDock 4.01 allows for rotatable bonds in the side chain of any selected residue in the

receptor protein. A flexible binding site was used in the docking of tacrine to the free acetylcholinesterase structure (PDB-ID: 1QIF). This was done to account for the observation that Phe330 had flipped into the active site of acetylcholinesterase, partially blocking access to tacrine. Seven residues within the previously defined binding site were allowed to have flexible sidechains: Trp84, Tyr121, Phe330, Tyr334, Trp432, His440, and Tyr442. The amino acids were chosen based on their proximity to Phe330 and tacrine in the X-ray crystal structure of the complex (PDB-ID: 1ACJ). Not all of the residues in the binding site were defined as flexible due to a limitation in the number of allowable rotatable bonds. The AutoDock calculation with flexible side chains took approximately 4.5 h to complete using the identical docking parameters and computer hardware as the previous calculations.

A further analysis of CSPs to guide and filter an AutoDock molecular docking calculation was performed using published  $^1\text{H}$  and  $^{15}\text{N}$  chemical shift data obtained from the solution structure of staphylococcal nuclease in both the unbound<sup>64</sup> and thymidine 3',5'-bisphosphate complexed forms.<sup>65,66</sup> The magnitude of the CSPs were calculated using a common weighting approach:

$$CSP = \sqrt{\frac{\left(\frac{\delta_N}{5}\right)^2 + \delta_H^2}{2}} \quad (3.3)$$

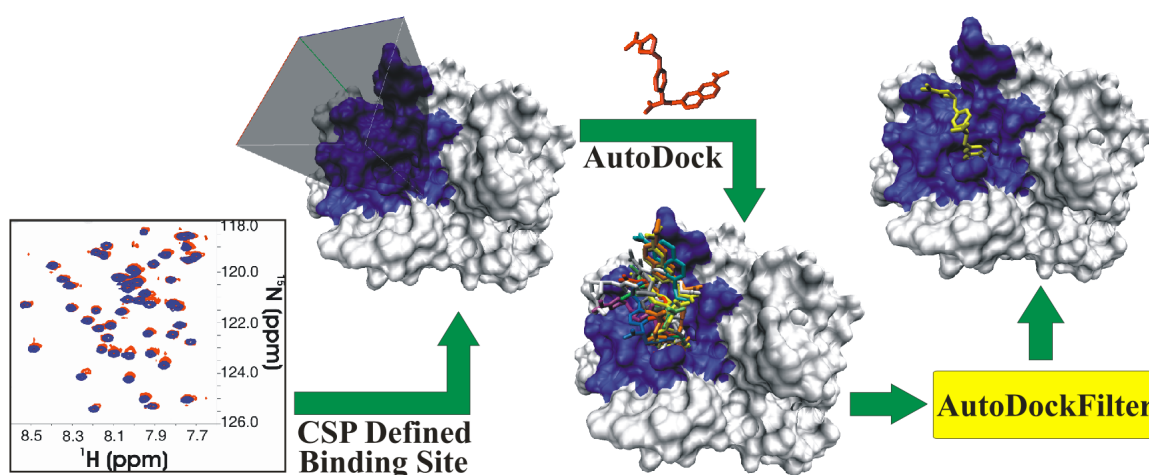
where  $\delta_N$  and  $\delta_H$  represent respectively the changes in  $^{15}\text{N}$  and  $^1\text{H}$  chemical shifts upon ligand binding.<sup>67</sup>

The binding site was determined by first selecting residues with CSPs greater than one standard deviation from the mean and mapping these residues onto the *Staphylococcus aureus* molecular surface. These residues corresponded to Ile18, Asp19,

Phe34, Leu37, Leu38, Val39, Lys84, Ala60, Lys110, Tyr113, Val114, and Tyr115. This was further filtered by visually selecting only those residues that clustered together on the protein's molecular surface, consistent with a consensus binding site. It was important to select the residues predicted to interact with the ligand in the consensus binding site. Some general factors that were considered include the presence of a contiguous surface of residues with CSPs (residues separated by  $<5 \text{ \AA}$  from nearest neighbors), residues clustered about a central point (encircling a binding pocket), surface accessibility, proximity to a surface feature (presence of intervening residues), relative distance to the main cluster of residues, and the relative magnitude of the observed CSPs. Another consideration was the number of residues that form a cluster, where a larger cluster size ( $\geq 4$ ) increases the likelihood that a ligand binding site has been correctly identified.

Residues Leu37, Leu38, Val39, Tyr113, Val114, and Tyr115 form the main contiguous CSP surface along one edge of a pocket on the *S. aureus* molecular surface. Residues Ile18, Asp19, and Lys84 exhibit some of the largest CSPs and were proximal to the same pocket as the main CSP cluster of six residues. In effect, residues Ile18, Asp19, Leu37, Leu38, Val39, Lys84, Tyr113, Val114, and Tyr115 encircle this binding pocket. Residue Ala60 was excluded because it is  $> 10 \text{ \AA}$  from this main cluster of residues and is on the opposite face of the protein. Residue Phe34 was excluded because it is not surface exposed and is part of the hydrophobic core of the protein. Residue Lys110 was excluded because it is outside the ring of residues encircling the binding pocket (i.e., residues Val39 and Tyr113 separate Lys110 from the binding pocket). Lys110 would not be expected to interact directly with thymidine 3',5'-bisphosphate. It also had the second-lowest CSP among the 12 residues initially selected.

Docking calculations were performed using the X-ray structures of the unbound (PDB-ID: 1EY0) and the thymidine 3',5'-bisphosphate complexed (PDB-ID: 1SNC) staphylococcal nuclease protein structure. The ligand coordinates were removed from the protein-ligand complex and stored as a separate file for docking. The AutoDock grid was positioned and sized to cover the residues with experimental CSPs. The grid was also large enough to include the entire thymidine 3',5'-bisphosphate molecule. Docking calculations were performed using the same parameters as before. Filtering of results using ADF was performed using the experimental CSPs.



**Figure 3.1** Flow diagram illustrating the overall process of generating a rapid protein-ligand costructure using CSP data to guide and filter the molecular docking results from AutoDock.

### 3.3 RESULTS AND DISCUSSION

The overall methodology for the rapid determination of a protein-ligand costructure has three steps [Figure 3.1]: (i) identification of the binding site by mapping CSPs from a  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR experiment, (ii) guiding the AutoDock calculations



using the identified binding site, and (iii) using the relative magnitude of the CSPs to filter the resulting ligand conformers from AutoDock. The ability of the protocol to accurately predict a protein-ligand costructure based on CSPs was demonstrated using multiple model systems and an experimental data set for *S. aureus* nuclease complexed to thymidine 3',5'-bisphosphate.

**3.3.1 Protein-ligand model systems.** Due to the scarcity of available chemical shift and structural data for complexed and unbound forms of multiple protein-ligand systems, the methodology was primarily demonstrated using empirically predicted chemical shift perturbations based on existing X-ray structures in the PDB. A total of 19 distinct pairs of protein structures with a variety of biological activity were identified and used for the docking simulation (Table 3.1).

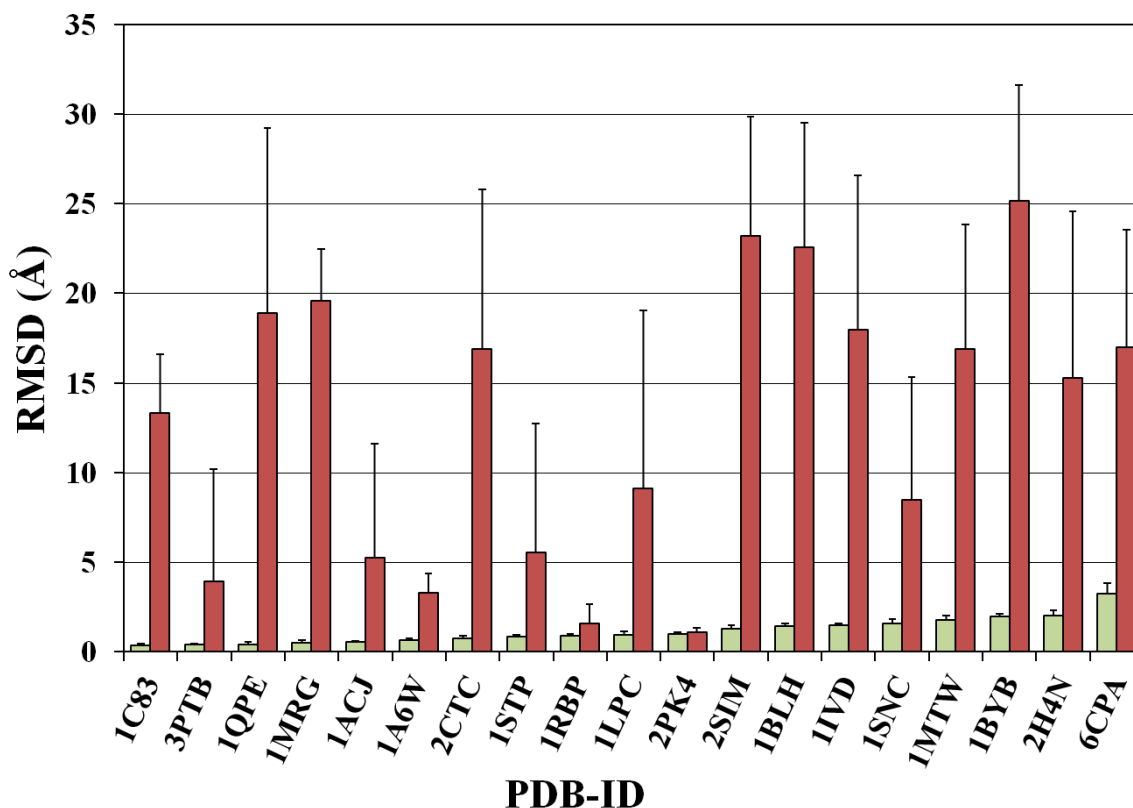
NMR chemical shift changes were routinely and widely used to map protein-ligand interactions based on the generally accepted protocol that residues proximal to the bound ligand will experience higher CSPs compared to residues distal to the ligand binding site.<sup>68,69</sup> Thus, CSPs were estimated by assuming a simple linear relationship between the magnitude of the CSPs and the distance from the amino acids' NH atoms and the nearest ligand atom. This was clearly a simple approximation since other factors besides proximity to the bound ligand contribute to CSPs. These factors include hydrogen bonding, electrostatics and ring current effects.<sup>70</sup> Unfortunately, a robust approach to predict ligand-induced chemical shift changes in a protein-ligand system using *ab initio* methods is not available because of the complexity of the system (e.g., number of atoms).

The absolute magnitude of the predicted CSPs is not critical since the CSPs are only used as an upper bound constraint to filter the poses predicted by AutoDock.

Additionally, a 3 Å distance cutoff is used to avoid any bias or distortion from unusually large CSPs. Effectively, the relative magnitude of the CSPs determines the conformer(s) selected by the AutoDockFilter (ADF) program. Again, this is based on the generally accepted premise that residues that incur the largest relative CSPs are predicted to be closer to the docked ligand. ADF simply identifies the conformer that maximizes an interaction with residues with the largest CSPs. This is also similar to the protocol implemented by HADDOCK<sup>25</sup> where CSPs and mutagenesis are used to create ambiguous interaction restraints, which define upper boundaries for the distances residues may be from any atom of the bound molecules in a protein-protein docking calculation.

**3.3.2 Comparison of blind docking and CSP-guided docking.** Blind docking is commonly used to generate protein-ligand complexes when a binding site is undetermined.<sup>16,21</sup> The approach requires scanning the entire protein surface, where the scoring function is used to both identify the binding site and select the best conformer. It may be extremely challenging to identify the correct ligand binding site when the binding energies between multiple distinct binding sites are within the error of the calculations. The AutoDock binding energies are estimated to have an error of 2.2 kcal/mol.<sup>71</sup> A comparison of results obtained between CSP-guided docking and blind docking demonstrated the expected advantages of guided docking to a known ligand binding site. On average, blind docking generated  $63 \pm 37$  distinct clusters in AutoDock using a 2.0 Å RMSD tolerance for each cluster. This large number of clusters generated a corresponding large average RMSD of  $15.40 \pm 6.40$  Å relative to the original X-ray structure.

The CSP-guided docking calculations yielded a relatively tight clustering of conformers ( $16 \pm 23$  distinct clusters) where the average RMSD was  $2.31 \pm 1.15$  Å. This was a dramatic and expected improvement relative to blind docking. A comparison using all the conformers within 2.2 kcal/mol of the lowest-energy conformer was made between the CSP-guided and blind dockings. This comparison demonstrates that the average RMSD of the CSP-guided conformers was significantly better relative to the blind docking conformers [Figure 3.2]. In addition, molecular dockings for ligands with greater than five rotatable bonds typically showed better results with the CSP-guided docking. Flexible ligands added a significant amount of complexity to the AutoDock calculation that was simplified by applying a smaller search space.

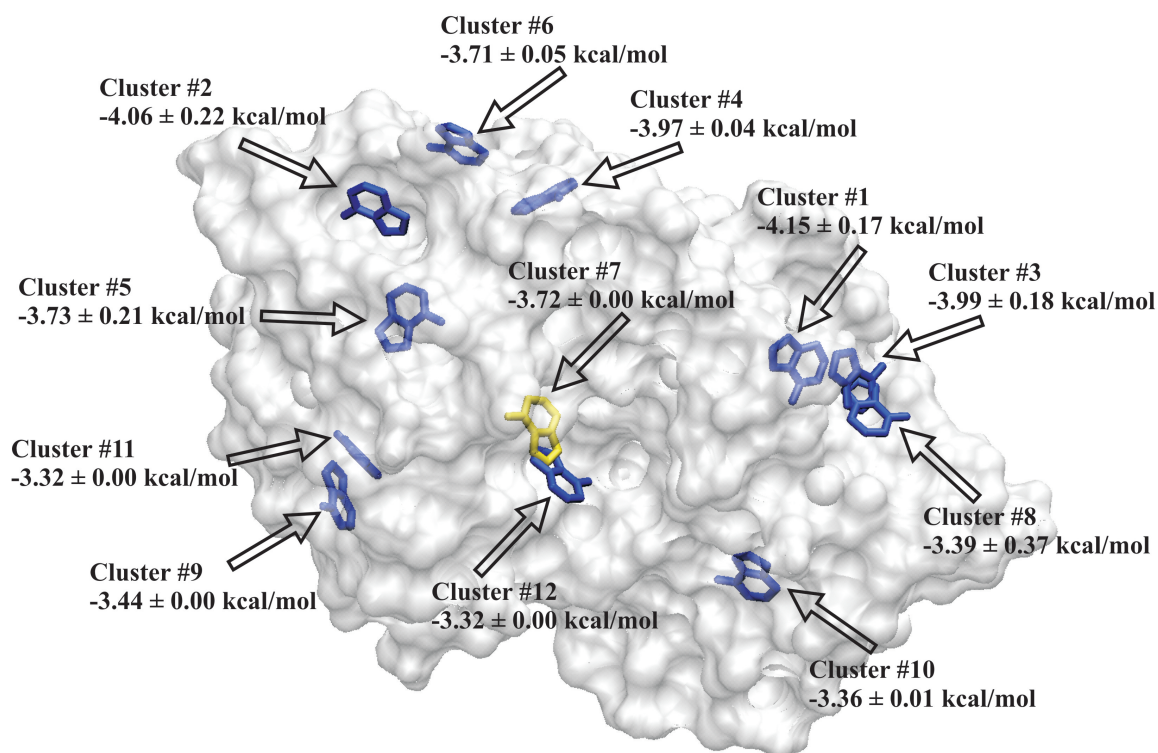


**Figure 3.2** Comparison of RMSD values of the docked ligand conformers relative to the original ligand conformation in the protein-ligand X-ray structure. AutoDock calculations used either CSP-guided docking without ADF filtering (green) or blind docking (red). Conformers within 2.2 kcal/mol of the lowest energy conformer were selected.

Focusing a docking calculation into a smaller volume of the protein minimizes the search space for the ligand and allows the docking resources to be spent orienting the ligand into an energetically favorable position and conformation instead of finding the binding site. Using the binding site determined by CSPs to focus an AutoDock grid is expected to eliminate some of the inherent ambiguity in identifying the correct ligand pose since the uncertainty regarding the correct ligand binding site has been removed.

**3.3.3 Lowest-energy cluster is not necessarily the best conformer.** Using a blind approach, AutoDock calculated a large number of clusters that are clearly outside of the known ligand binding site observed in the original X-ray crystal structure [Figure 3.3].

For AutoDock, selecting the lowest-energy cluster and the most populated cluster were the two most common methods for identifying the most accurate conformers. In our analysis, the lowest-energy cluster represented the best docked conformers in 10 out of the 19 docking calculations performed (53% accuracy). The best docked cluster of conformers was only selected in 6 instances (32% accuracy) when the most populated cluster was used, and 5 of those were also the lowest-energy cluster. This caused a significant ambiguity in evaluating the accuracy of any particular protein-ligand costructure based solely on the AutoDock binding energy. However, it should be noted that AutoDock does generate at least one conformer out of the 120 conformers that is within 2.0 Å of the actual binding pose in 14 out of the 19 protein-ligand blind docking calculations. Thus, an accurate conformer is often generated by an AutoDock calculation, but the binding energy is not a reliable mechanism to identify the best conformer.



**Figure 3.3** Surface representation of the ribosome-inactivating protein (PDB ID: 1MRG) with the lowest-energy adenosine conformer for each cluster calculated by AutoDock superimposed on the protein structure. The conformer with the lowest RMSD relative to the X-ray structure is colored yellow. Each cluster is labeled with the cluster ranking, the average binding energies, and standard deviation of the binding energies.

One of the main reasons the lowest-energy cluster may not contain the most accurately docked conformer appears to arise from the low sensitivity of the AutoDock binding energy to identify distinct binding sites and ligand conformations. This is especially true when the relative differences in the binding energies between the clusters calculated by AutoDock are taken into consideration. A representative conformer for each cluster calculated by AutoDock for adenosine docked to the ribosome-inactivating protein (RIP, PDB-ID: 1MRG) is illustrated in [Figure 3.3]. A difference of only 0.83 kcal/mol is observed between the lowest- and highest-energy clusters. This energy difference is significantly smaller than the 2.2 kcal/mol estimate for the error in the

AutoDock binding energy. Thus, the lowest-energy cluster is not appreciably different from the remaining clusters in the docking of adenosine to RIP. Specifically, an energy difference of only 0.43 kcal/mol is observed between the cluster with the lowest RMSD relative to those of the X-ray structure and the lowest-energy cluster. If the error in the AutoDock binding energy follows a normal distribution, this would partially explain why the correct conformer is not always present in the lowest-energy cluster.

An additional reason the best conformer is not present in the lowest-energy cluster may be attributed to the protocol AutoDock uses to define members of a cluster. AutoDock selects the conformer with the lowest binding energy to represent the first conformer of the first cluster. Conformers that are within the RMSD tolerance of the lowest-energy conformer (2.00 Å for this study) are placed in the first cluster regardless of binding energy. The remaining conformer with the lowest binding energy starts the next cluster that will include all of the remaining conformers within the RMSD tolerance. This continues until all of the conformers are placed in a cluster. Thus, the best conformer can easily be excluded from the first cluster, despite similar binding energies.

Even though the conformers in any particular cluster are within 2.0 Å of each other, the binding energies can vary enough that some conformers in the lowest-energy cluster actually have a higher binding energy than conformers in other clusters. This again suggests that simply using the binding energy to select for the best pose is ambiguous and unreliable for any specific protein-ligand costructure.

On the basis of our results, choosing the lowest-energy clusters in a blind docking will result in identifying an incorrect binding site and a wrong conformer 47% of the time. Interestingly, the lowest-energy cluster represents the best docked pose in 14 of the

19 docking calculations with CSP-guided docking. Again, this is a significant improvement relative to blind docking; however, an ambiguity in identifying the correct conformer still remains. A wrong conformer is still identified 26% of the time. This ambiguity can be remedied by also using the CSPs to further filter the docking results to select conformers that agree with the NMR experimental data.

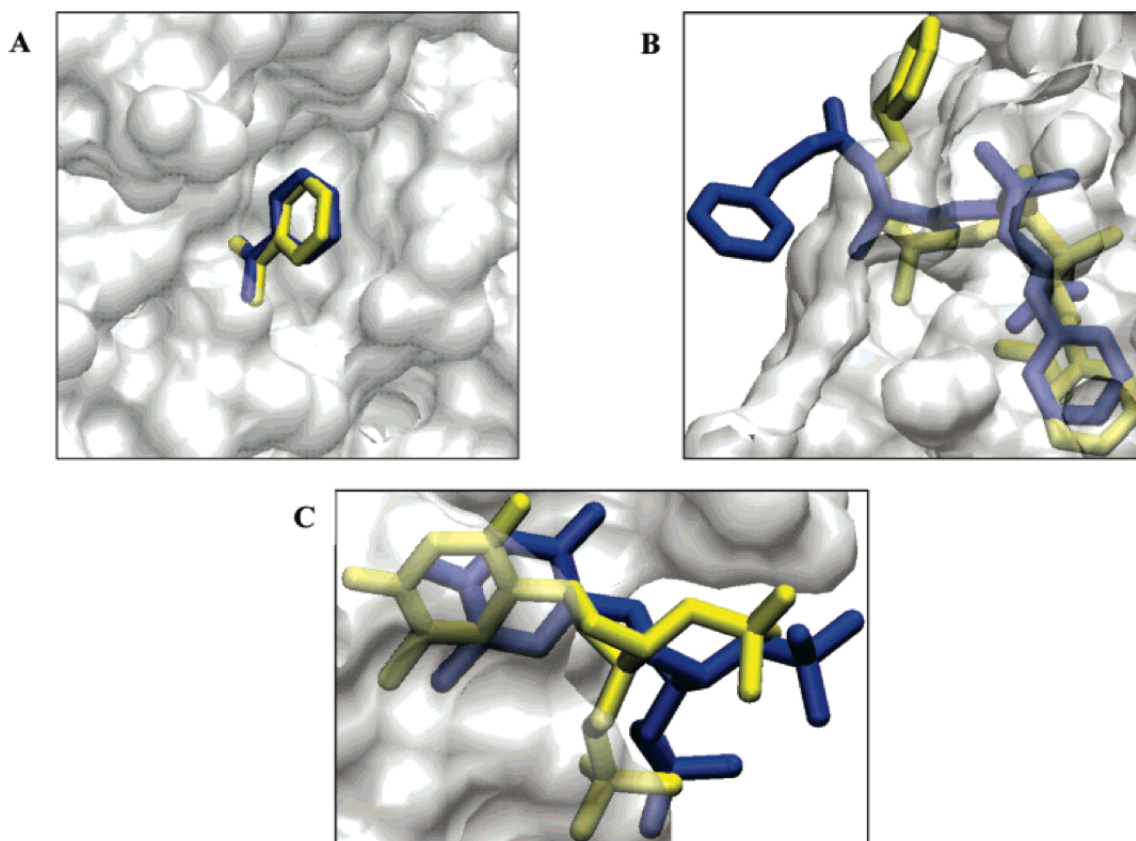
**3.3.4 CSP-guided docking with ADF filtering.** Our AutoDockFilter (ADF) program identifies the “best” conformer from an AutoDock calculation based on a consistency with the experimental CSPs. ADF replaces the ambiguous AutoDock binding energy with an NMR violation energy [Equations 3.1 and 3.2]. The ADF-selected conformer has a minimal distance between each protein residue with a CSP and the docked ligand. Filtering the CSP-guided docking with ADF consistently resulted in the identification of a cluster of conformers with a high similarity to the original X-ray costructure [Figure 3.4A]. This is a significant improvement in accuracy relative to the blind docking and the CSP-guided docking with ADF. It completely eliminates the ambiguity encountered by relying on the AutoDock binding energy.

Obtaining an  $\text{RMSD} < 3 \text{ \AA}$  from an experimental protein-ligand structure was generally considered a good result in a docking calculation.<sup>72</sup> The conformers selected from the CSP-guided docking with ADF filtering had RMSD averages of  $1.17 \pm 0.74 \text{ \AA}$ . These results are also significantly better than the conformers selected by CSP-guided docking without ADF filtering ( $2.02 \pm 1.05 \text{ \AA}$ ). The improvement is even more pronounced when compared to blind docking ( $12.91 \pm 7.81 \text{ \AA}$ ). [Figure 3.5] illustrates the improvement obtained using the CSP-guided docking with ADF filtering for each of the

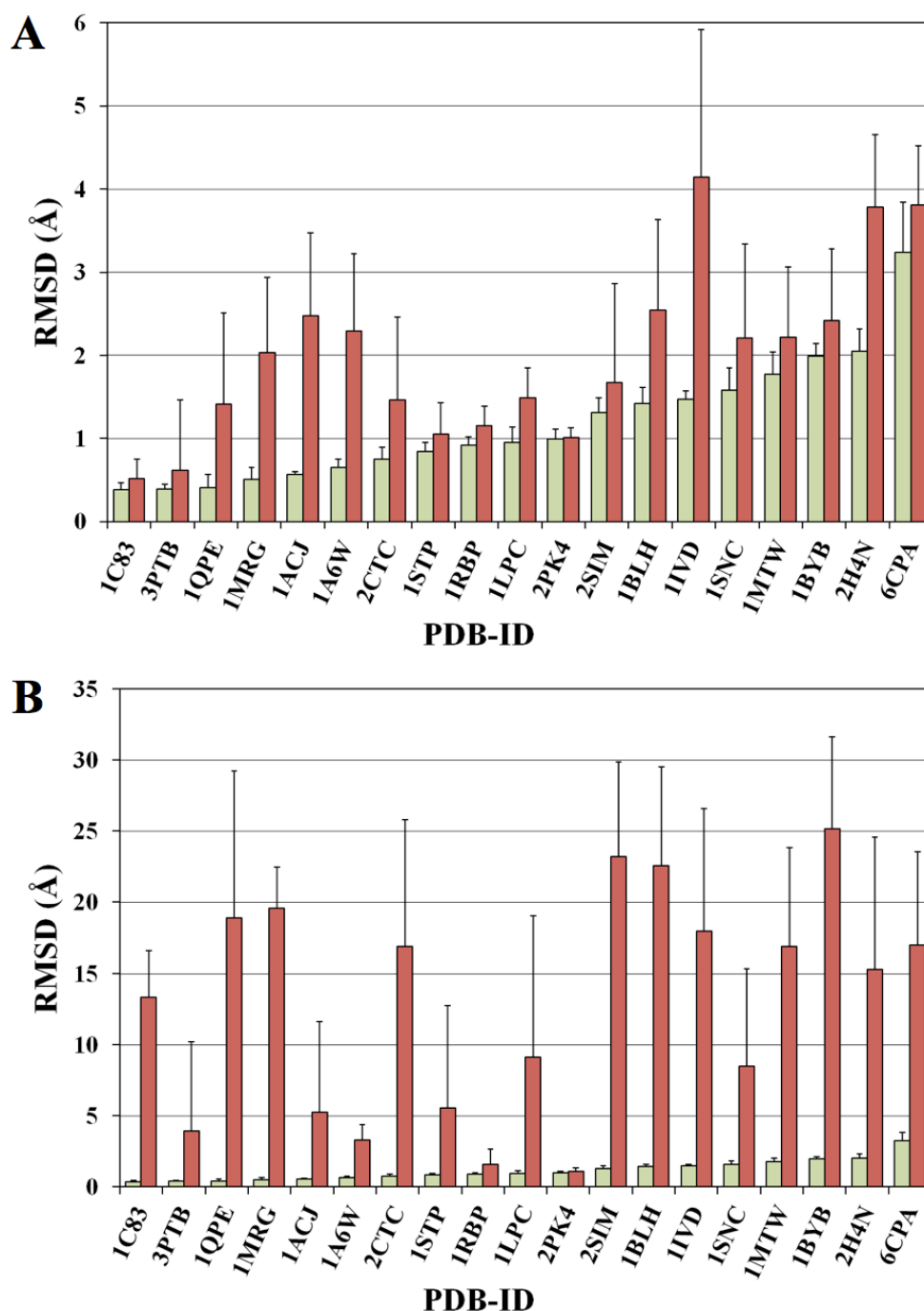


19 protein-ligand structures. Thus, the CSP-guided docking with ADF filtering represents a significant improvement in rapidly obtaining accurate protein-ligand structures.

Only one ligand docking exhibited an average RMSD of over 2.00 Å when guided and filtered using CSPs. The relatively large RMSD average and deviation for phosphonate docked to carboxypeptidase A (PDB-ID: 6CPA) occurs because the ligand is partially solvent exposed [Figure 3.4B]. In solution, this solvent exposed region of the phosphonate is probably ill-defined and adopts multiple conformations similar to the results seen with AutoDock. Therefore, the higher RMSD difference observed between the X-ray structure and the docked conformer is irrelevant because the X-ray structure simply represents one of many equivalent conformations.



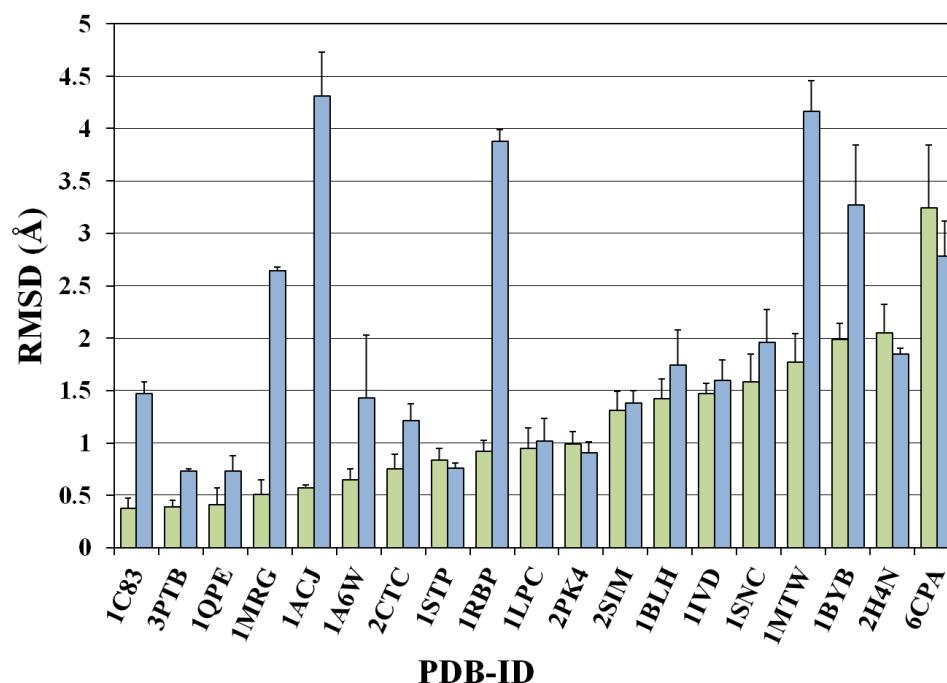
**Figure 3.4** Superposition of the CSP-guided docking with ADF filtering conformers (blue) with the original X-ray structures (yellow) for (A) benzamidine complexed with trypsin (PDB-ID: 3PTB), (B) a phosphonate complexed with carboxypeptidase (PDB-ID: 6CPA), and (C) *S. aureus* nuclease complexed to thymidine 3',5'-bisphosphate (PDB-ID: 1SNC). The nuclease-thymidine 3',5'-bisphosphate docked model is based on experimental NMR chemical shift data.



**Figure 3.5** Comparison of RMSD values of the docked ligand conformers relative to the original ligand conformation in the protein-ligand X-ray structure. AutoDock calculations used the CSP-guided docking with ADF filtering (green) and (A) the CSP-guided docking without ADF filtering (red) or (B) blind docking (red). Conformers within 2.2 kcal/mol of the lowest-energy conformer were selected for the CSP-guided docking without ADF filtering and blind docking.

**3.3.5 CSP-guided docking with ADF filtering using apoproteins.** A ligand binding to a protein structure may result in significant changes in the protein structure. This is illustrated by the backbone or active-site RMSD differences observed between the 19 apoprotein structures and the corresponding protein-ligand complexes (Table 3.1). A deviation as large as 2.48 Å was observed for the ligand binding site of  $\beta$ -amylase (PDB-ID: 1BYB, 1BYA) when it binds glucose. Therefore, the accuracy of the conformers obtained using the CSP-guided docking with ADF filtering protocol was further evaluated using unbound structures.

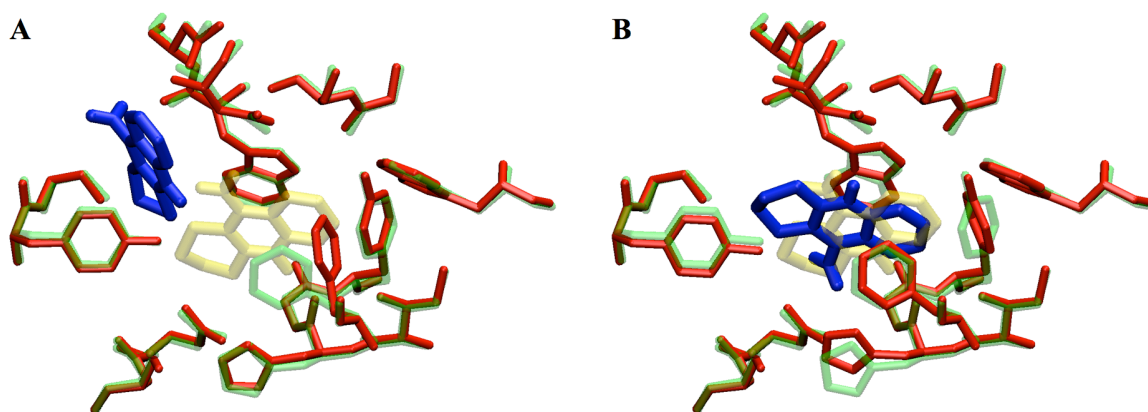
The AutoDock calculation and analysis was repeated using the 19 apoprotein structures following the identical procedure applied to the bound protein structures. The CSP-guided docking and ADF filtering with the apoproteins yielded results similar to those from the docking with the bound protein structures [Figure 3.6]. However, there were four examples where docking of the ligand to the unbound protein structure resulted in a  $\geq 2.00$  Å increase in the RMSD relative to the corresponding protein-ligand X-ray structure. The observed RMSD binding site differences between the bound and free protein structures did not solely explain these docking results (Table 3.1). In fact, the protein X-ray structures with the largest binding site changes upon ligand-binding did not necessarily yield significantly different docking results.



**Figure 3.6** Comparison of RMSD values of the best docked ligand conformers relative to the original ligand conformation in the protein-ligand X-ray structure. The AutoDock calculation was performed by docking the ligand to either the bound protein structure (green) or the free protein structure (blue). The calculations used CSP-guided docking followed by ADF filtering to select the best conformers.

The worst docking results were seen with the free acetylcholinesterase structure (PDB-ID: 1QIF) that incurred a modest all atom deviation of 0.63 Å in the tacrine binding site between the bound and free X-ray structures. The best AutoDock-calculated tacrine conformer using the free acetylcholinesterase structure had a 3.91 Å RMSD from the original acetylcholinesterase-tacrine X-ray structure (PDB-ID: 1ACJ). This compared poorly to an average RMSD of 0.51 Å obtained for the tacrine conformers docked to the bound form of acetylcholinesterase. This large deviation in the docked tacrine conformers was due to the side chain of Phe330 flipping into the free acetylcholinesterase binding site and essentially blocking tacrine from binding deep into the ligand pocket [Figure 3.7A]. The side-chain flipping of Phe330 is a known “gatekeeper” mechanism of ligand

binding to acetylcholinesterase.<sup>73</sup> This steric hindrance due to side-chain dynamics appears to be a common source of docking error that was encountered using apoproteins.



**Figure 3.7** Comparison of tacrine docking to the free acetylcholinesterase structure (PDB-ID: 1QIF) using (A) static and (B) flexible acetylcholinesterase binding site residues. An overlay of the binding site residues (green) and tacrine (yellow) from the X-ray acetylcholinesterase-tacrine structure (PDB-ID: 1ACJ) with the binding site residues (red) for the free acetylcholinesterase structure used for the AutoDock calculation with the resulting best docked tacrine conformation (blue). The RMSD's between the docked and X-ray structure of tacrine using the static and flexible binding sites are 3.91 and 1.78 Å, respectively.

Importantly, the magnitude of the NMR violation energy provides a valuable measure of the inherent accuracy of the AutoDock results and an efficient means to identify these incorrectly docked structures due to protein mobility. A comparison of the NMR violation energy with the RMSD between the docked and X-ray ligand structures is shown in [Figure 3.8]. An NMR violation energy > 1500 is correlated with the poorly docked ligand conformers obtained with the apostructures. Therefore, obtaining a large NMR violation energy would call into question the reliability of the docked structure. This would imply that further analysis or a detailed dynamic simulation would be required in order to obtain an accurate docked protein-ligand costructure. Of course, the NMR violation threshold of 1500 is based on our simulated docking using empirically

determined CSPs. This threshold may change for protein-ligand costructures that are calculated using experimentally determined CSPs.

Docking a ligand into a static protein structure is a common simplification to improve performance, but it is an assumption that can cause inaccuracies as observed above. AutoDock 4 attempts to alleviate the static receptor problem by incorporating side-chain flexibility in addition to the existing ligand flexibility. However, adding this additional flexibility significantly increased the AutoDock calculation time. Different ligand conformers needed to be evaluated against the various amino acid side-chain orientations in the binding site. AutoDock calculations using a static receptor required, on average,  $37 \pm 32$  min. Conversely, an AutoDock calculation that docked tacrine into the free acetylcholinesterase structure allowing 7 amino acid side chains within the binding site to be flexible required 4.5 hours to complete. Despite the dramatic increase in calculation time, there was a significant improvement in the accuracy of the docked tacrine structure [Figure 3.7B]. The RMSD of the docked tacrine structure relative to the original X-ray structure dropped from 3.91 to 1.78 Å [Figure 3.8].

**3.3.6 Docking with experimental NMR data.** While the analysis using empirically predicted chemical shift perturbations appeared to support the reliability of using CSP-guided docking and ADF filtering to rapidly obtain accurate protein-ligand costructures, a full evaluation of the methodology required an analysis with experimental chemical shift perturbation data.  $^1\text{H}$  and  $^{15}\text{N}$  chemical shift data for the free solution structure of staphylococcal nuclease<sup>64</sup> and the complex with thymidine 3',5'-bisphosphate were readily available.<sup>65,66</sup> Similarly, the X-ray structures of the unbound

(PDB-ID: 1EY0) and the complexed (PDB-ID: 1SNC) forms of staphylococcal nuclease were accessible through the PDB.

Experimental CSPs may arrive from either a direct interaction with the bound ligand or indirectly through a protein conformational change that may occur distal from the ligand binding site. Thus, correct utilization of the CSPs to guide and filter an AutoDock calculation requires removing CSPs not resulting from a direct ligand interaction. In practice, CSPs greater than one standard deviation from the average CSP are mapped onto the protein surface to visually identify a consensus ligand binding site. This subset of CSPs is then used to guide the AutoDock grid for the docking calculation which is followed by ADF to select the best conformers based on consistency with these CSPs.

The experimental staphylococcal nuclease CSPs were used to guide and filter the docking of thymidine 3',5'-bisphosphate to the bound conformation of the protein [Figure 3.4C]. The best conformers using experimental CSPs had an average RMSD of  $1.63 \pm 0.35$  Å relative to the original nuclease-thymidine 3',5'-bisphosphate X-ray structure with a corresponding average NMR violation energy of  $681 \pm 333$ . These results compared well to the thymidine 3',5'-bisphosphate conformers obtained with the empirical CSPs, where an average RMSD of  $1.58 \pm 0.27$  Å and an average NMR violation energy of  $788 \pm 409$  were obtained.

A similar comparison was observed for the docking of thymidine 3',5'-bisphosphate to the unbound nuclease structure (PDB-ID: 1EY0). The experimental CSPs generated conformers with an average RMSD of  $1.96 \pm 0.32$  Å and an average NMR violation energy of  $667 \pm 449$  where the empirically predicted CSPs resulted in



conformers with an average RMSD of  $1.96 \pm 0.31$  Å and an average NMR violation energy of  $914 \pm 327$ . In effect, the experimental and empirical CSPs yielded essentially identical models of a nuclease-thymidine 3',5'-bisphosphate costructure [Figure 3.8]. This suggested that the use of empirically predicted CSPs to evaluate the reliability of CSP-guided docking and ADF filtering method is a reasonable approach. Also, the results with the experimental staphylococcal nuclease data clearly indicated that the CSP-guided docking and ADF filtering method works equally well with experimental CSPs. It also suggests that an NMR violation threshold of 1500 may be applicable to identifying poorly docked structures using experimental CSPs since the NMR violation energies calculated for the nuclease-thymidine 3',5'-bisphosphate costructure using experimental and empirical CSPs were similar. These results also demonstrate that the high accuracy obtained with the method is not simply an artifact of the empirical CSPs.

### 3.4 CONCLUSIONS

Combining experimental NMR chemical shift perturbations (CSPs) with AutoDock ligand docking calculations provides an efficient approach to rapidly obtain accurate ( $1.17 \pm 0.74$  Å) protein-ligand models. The CSPs are first used to guide an AutoDock calculation by defining the size and position of the AutoDock grid. The CSPs

are then used in combination with our AutoDockFilter (ADF) program to select the best conformer cluster consistent with the CSPs using an empirical NMR violation energy. ADF assumes a linear relationship between the magnitude of CSPs and the distance between the ligand and the protein residues that incurred the CSP. The NMR violation energy correlates with the accuracy of the docked structures obtained using the empirical CSPs and the *S. aureus* nuclease experimental CSPs, where an observed energy  $> 1500$  implies an unreliably docked structure. The poor docking generally occurred within apoprotein structures that required a conformational change to accommodate the bound ligand. Additional protein dynamics such as side-chain flexibility are required to improve the accuracy of these docked ligands.

The docking method described typically requires  $37 \pm 32$  min per protein-ligand complex. Since a 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR experiment can be collected in  $\leq 15$  min, reliable and accurate protein-ligand costructures can be rapidly obtained in less than an hour. Thus, an efficient initial approach to structure-based drug discovery can be achieved by combining high-throughput NMR screening with CSP-guided and ADF-filtered AutoDock calculations. Of course, X-ray or NMR structures will still be required as the project matures since further refinement of the chemical leads mandates a higher-quality costructure than the 1-2 Å accuracy obtained by molecular docking.

### 3.5 REFERENCES

1. Congreve, M., Murray, C. W. & Blundell, T. L. Structural biology and drug discovery. *Drug Discov Today* **10**, 895–907 (2005).
2. Orry, A. J. W., Abagyan, R. A. & Cavasotto, C. N. Structure-based development of target-specific compound libraries. *Drug Discov Today* **11**, 261–266 (2006).
3. Scapin, G. Structural Biology and Drug Discovery. *Curr Pharm Des* **12**, 2087–2097 (2006).

4. Powers, R. Applications of NMR to structure-based drug design in structural genomics. *J Struct Funct Genomics* **2**, 113–123 (2002).
5. Moy, F. J. *et al.* NMR solution structure of the catalytic fragment of human fibroblast collagenase complexed with a sulfonamide derivative of a hydroxamic acid compound. *Biochemistry* **38**, 7085–7096 (1999).
6. Chen, J. M. *et al.* Structure-based design of a novel, potent, and selective inhibitor for MMP-13 utilizing NMR spectroscopy and computer-aided molecular design. *J Am Chem Soc* **122**, 9648–9654 (2000).
7. Medek, A., Hajduk, P. J., Mack, J. & Fesik, S. W. The use of differential chemical shifts for determining the binding site location and orientation of protein-bound ligands. *J Am Chem Soc* **122**, 1241–1242 (2000).
8. Chen, A. & Shapiro, M. J. NOE Pumping: A Novel NMR Technique for Identification of Compounds with Binding Affinity to Macromolecules. *J Am Chem Soc* **120**, 10258–10259 (1998).
9. Pellecchia, M. *et al.* NMR-based structural characterization of large protein-ligand interactions. *J Biomol NMR* **22**, 165–173 (2002).
10. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **3**, 935–949 (2004).
11. Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409–443 (2002).
12. Verkhivker, G. M. *et al.* Deciphering common failures in molecular docking of ligand-protein complexes. *J Comput Aided Mol Des* **14**, 731–751 (2000).
13. Watson, J. D., Laskowski, R. A. & Thornton, J. M. Predicting protein function from sequence and structural data. *Curr Opin Struct Biol* **15**, 275–284 (2005).
14. Arakaki, A. K., Zhang, Y. & Skolnick, J. Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics* **20**, 1087–1096 (2004).
15. Park, H., Lee, J. & Lee, S. Critical assessment of the automated AutoDock as a new docking tool for virtual screening. *Proteins* **65**, 549–554 (2006).
16. Hetényi, C. & van der Spoel, D. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci* **11**, 1729–1737 (2002).
17. Sousa, S. F., Fernandes, P. A. & Ramos, M. J. Protein-ligand docking: current status and future challenges. *Proteins* **65**, 15–26 (2006).
18. Morris, G., Goodsell, D., Halliday, R. & Huey, R. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**, 1639–1662 (1998).
19. Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* **28**, 1145–1152 (2007).
20. Bursulaya, B. D., Totrov, M., Abagyan, R. & Brooks, C. L. Comparative study of several algorithms for flexible ligand docking. *J Comput Aided Mol Des* **17**, 755–763 (2003).
21. Hetényi, C. & van der Spoel, D. Blind docking of drug-sized compounds to proteins with up to a thousand residues. *FEBS Lett* **580**, 1447–1450 (2006).

22. Shuker, S. B., Hajduk, P. J., Meadows, R. P. & Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **274**, 1531–1534 (1996).
23. McCoy, M. A. & Wyss, D. F. Spatial localization of ligand binding sites from electron current density surfaces calculated from NMR chemical shift perturbations. *J Am Chem Soc* **124**, 11758–11763 (2002).
24. Schieborr, U. *et al.* How much NMR data is required to determine a protein-ligand complex structure? *ChemBioChem* **6**, 1891–1898 (2005).
25. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **125**, 1731–1737 (2003).
26. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D* **54**, 905–921 (1998).
27. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **10**, 980 (2003).
28. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
29. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723 (1997).
30. Harel, M. *et al.* Quaternary ligand binding to aromatic residues in the active-site gorge of acetylcholinesterase. *Proc Natl Acad Sci USA* **90**, 9031–9035 (1993).
31. Weik, M. *et al.* Specific chemical and structural damage to proteins produced by synchrotron radiation. *Proc Natl Acad Sci USA* **97**, 623–628 (2000).
32. Chen, C. C., Rahil, J., Pratt, R. F. & Herzberg, O. Structure of a phosphonate-inhibited beta-lactamase. An analog of the tetrahedral transition state/intermediate of beta-lactam hydrolysis. *J Mol Biol* **234**, 165–178 (1993).
33. Chen, C. C. *et al.* Structure and kinetics of the beta-lactamase mutants S70A and K73H from *Staphylococcus aureus* PC1. *Biochemistry* **35**, 12251–12258 (1996).
34. Mikami, B., Degano, M., Hehre, E. J. & Sacchettini, J. C. Crystal structures of soybean beta-amylase reacted with beta-maltose and maltal: active site components and their apparent roles in catalysis. *Biochemistry* **33**, 7779–7787 (1994).
35. Andersen, H. S. *et al.* 2-(oxalylamino)-benzoic acid is a general, competitive inhibitor of protein-tyrosine phosphatases. *J Biol Chem* **275**, 7101–7108 (2000).
36. Pedersen, A. K., Peters, G., G. Ü. H., Möller, K. B., Iversen, L. F. & Kastrop, J. S. Water-molecule network and active-site flexibility of apo protein tyrosine phosphatase 1B. *Acta Crystallogr D* **60**, 1527–1534 (2004).
37. Jedrzejas, M. J. *et al.* Structures of aromatic inhibitors of influenza virus neuraminidase. *Biochemistry* **34**, 3144–3151 (1995).
38. Bossart-Whitaker, P. *et al.* Three-dimensional structure of influenza A N9 neuraminidase and its complex with the inhibitor 2-deoxy 2,3-dehydro-N-acetyl neuraminic acid. *J Mol Biol* **232**, 1069–1083 (1993).
39. Kurinov, I. V., Rajamohan, F. & Uckun, F. M. High resolution X-ray structure and potent anti-HIV activity of recombinant dianthin antiviral protein. *Arzneimittelforschung* **54**, 692–702 (2004).
40. Huang, Q., Liu, S., Tang, Y., Jin, S. & Wang, Y. Studies on crystal structures,

- active-centre geometry and depurinating mechanism of two ribosome-inactivating proteins. *Biochem J* **309** ( Pt 1), 285–298 (1995).
41. Ren, J., Wang, Y., Dong, Y. & Stuart, D. I. The N-glycosidase mechanism of ribosome-inactivating proteins implied by crystal structures of alpha-momorcharin. *Structure* **2**, 7–16 (1994).
  42. Stubbs, M. T., Huber, R. & Bode, W. Crystal structures of factor Xa specific inhibitors in complex with trypsin: structural grounds for inhibition of factor Xa and selectivity against thrombin. *FEBS Lett* **375**, 103–107 (1995).
  43. Walter, J. *et al.* On the disordered activation domain in trypsinogen: chemical labelling and low-temperature crystallography. *Acta Crystallogr B-Stru* **38**, 1462–1472 (1982).
  44. Zhu, X. *et al.* Structural analysis of the lymphocyte-specific kinase Lck in complex with non-selective and Src family selective kinase inhibitors. *Structure* **7**, 651–661 (1999).
  45. Yamaguchi, H. & Hendrickson, W. A. Structural basis for activation of human lymphocyte kinase Lck upon tyrosine phosphorylation. *Nature* **384**, 484–489 (1996).
  46. Cowan, S. W., Newcomer, M. E. & Jones, T. A. Crystallographic refinement of human serum retinol binding protein at 2A resolution. *Proteins* **8**, 44–61 (1990).
  47. Zanotti, G., Ottonello, S., Berni, R. & Monaco, H. L. Crystal structure of the trigonal form of human plasma retinol-binding protein at 2.5 Å resolution. *J Mol Biol* **230**, 613–624 (1993).
  48. Loll, P. J. & Lattman, E. E. The crystal structure of the ternary complex of staphylococcal nuclease, Ca<sup>2+</sup>, and the inhibitor pdTp, refined at 1.65 Å. *Proteins* **5**, 183–201 (1989).
  49. Hynes, T. R. & Fox, R. O. The crystal structure of staphylococcal nuclease refined at 1.7 Å resolution. *Proteins* **10**, 92–105 (1991).
  50. Weber, P. C., Ohlendorf, D. H., Wendoloski, J. J. & Salemme, F. R. Structural origins of high-affinity biotin binding to streptavidin. *Science* **243**, 85–88 (1989).
  51. Katz, B. A. Binding of biotin to streptavidin stabilizes intersubunit salt bridges between Asp61 and His87 at low pH. *J Mol Biol* **274**, 776–800 (1997).
  52. Teplyakov, A., Wilson, K. S., Orioli, P. & Mangani, S. High-resolution structure of the complex between carboxypeptidase A and L-phenyl lactate. *Acta Crystallogr D* **49**, 534–540 (1993).
  53. Lesburg, C. A., Huang, C., Christianson, D. W. & Fierke, C. A. Histidine --> carboxamide ligand substitutions in the zinc binding site of carbonic anhydrase II alter metal coordination geometry but retain catalytic activity. *Biochemistry* **36**, 15780–15791 (1997).
  54. Håkansson, K., Carlsson, M., Svensson, L. A. & LILJAS, A. Structure of native and apo carbonic anhydrase II and structure of some of its anion-ligand complexes. *J Mol Biol* **227**, 1192–1204 (1992).
  55. Wu, T. P., Padmanabhan, K., Tulinsky, A. & Mulichak, A. M. The refined structure of the epsilon-aminocaproic acid complex of human plasminogen kringle 4. *Biochemistry* **30**, 10589–10594 (1991).
  56. Stec, B., Yamano, A., Whitlow, M. & Teeter, M. M. Structure of human plasminogen kringle 4 at 1.68 Å and 277 K. A possible structural role of disordered

- residues. *Acta Crystallogr D* **53**, 169–178 (1997).
57. Crennell, S. J. *et al.* The Structures of Salmonella typhimurium LT2 Neuraminidase and its Complexes with Three Inhibitors at High Resolution. *J Mol Biol* **259**, 264–280 (1996).
  58. Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Crystallogr B-Stru* **39**, 480–490 (1983).
  59. Kim, H. & Lipscomb, W. N. Crystal structure of the complex of carboxypeptidase A with a strongly bound phosphonate in a new crystalline form: comparison with structures of other complexes. *Biochemistry* **29**, 5546–5555 (1990).
  60. Rees, D. C., Lewis, M. & Lipscomb, W. N. Refined crystal structure of carboxypeptidase A at 1.54 Å resolution. *J Mol Biol* **168**, 367–387 (1983).
  61. Sayle, R. A. & Milner-White, E. J. RASMOL: biomolecular graphics for all. *Trends Biochem Sci* **20**, 374 (1995).
  62. Kanungo, T. *et al.* A local search approximation algorithm for k-means clustering. in *SoCG* 10–18 (ACM Press, 2002). doi:10.1145/513400.513402
  63. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* **14**, 33–8–27–8 (1996).
  64. Wang, J. F., Hinck, A. P., Loh, S. N., LeMaster, D. M. & Markley, J. L. Solution studies of staphylococcal nuclease H124L. 2. <sup>1</sup>H, <sup>13</sup>C, and <sup>15</sup>N chemical shift assignments for the unligated enzyme and analysis of chemical shift changes that accompany formation of the nuclease-thymidine 3',5'-bisphosphate-calcium ternary complex. *Biochemistry* **31**, 921–936 (1992).
  65. Wang, J. F., LeMaster, D. M. & Markley, J. L. Two-dimensional NMR studies of staphylococcal nuclease. 1. Sequence-specific assignments of hydrogen-1 signals and solution structure of the nuclease H124L-thymidine 3',5'-bisphosphate-Ca<sup>2+</sup> ternary complex. *Biochemistry* **29**, 88–101 (1990).
  66. Wang, J. F., Hinck, A. P., Loh, S. N. & Markley, J. L. Two-dimensional NMR studies of staphylococcal nuclease. 2. Sequence-specific assignments of carbon-13 and nitrogen-15 signals from the nuclease H124L-thymidine 3',5'-bisphosphate-Ca<sup>2+</sup> ternary complex. *Biochemistry* **29**, 102–113 (1990).
  67. Garrett, D. S., Seok, Y.-J., Peterkofsky, A., Clore, G. M. & Gronenborn, A. M. Identification by NMR of the binding surface for the histidine-containing phosphocarrier protein HPr on the N-terminal domain of enzyme I of the Escherichia coli phosphotransferase system. *Biochemistry* **36**, 4393–4398 (1997).
  68. Carlomagno, T. Ligand-target interactions: what can we learn from NMR? *Annu Rev Biophys Biomol Struct* **34**, 245–266 (2005).
  69. Lepre, C. A., Moore, J. M. & Peng, J. W. Theory and applications of NMR-based screening in pharmaceutical research. *Chem Rev* **104**, 3641–3676 (2004).
  70. Oldfield, E. Chemical shifts in amino acids, peptides, and proteins: from quantum chemistry to drug design. *Annu Rev Phys Chem* **53**, 349–378 (2002).
  71. Rosenfeld, R. J. *et al.* Automated docking of ligands to an artificial active site: augmenting crystallographic analysis with computer modeling. *J Comput Aided Mol Des* **17**, 525–536 (2003).
  72. Wolf, A., Zimmermann, M. & Hofmann-Apitius, M. Alternative to consensus scoring--a new approach toward the qualitative combination of docking

- algorithms. *J Chem Inf Model* **47**, 1036–1044 (2007).
73. Kryger, G., Silman, I. & Sussman, J. L. Structure of acetylcholinesterase complexed with E2020 (Aricept): implications for the design of new anti-Alzheimer drugs. *Structure* **7**, 297–307 (1999).



## CHAPTER 4

### AUTODOCKFILTER 2.0 AND CSP-CONSENSUS

#### 4.1 INTRODUCTION

In Chapter 3, the theoretical principles behind the AutoDockFilter (ADF) program were described. This program was shown to be feasible as a technique to rapidly generate protein-ligand costructures using chemical shift perturbations (CSPs) from a 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC experiment to guide the docking process and then filter the resulting docked poses that best agree with those same CSPs.<sup>1</sup> The ADF program proved to be effective in removing the ambiguity that results from the 2.2 kcal/mol estimate for the error in the AutoDock binding energy.<sup>1,2</sup>

Unfortunately, the original ADF program was tested under ideal conditions on protein-ligand systems where the CSPs were calculated, with only one example, being tested on experimental CSPs. As is often the case when developing software, the application of the original ADF program during the FAST-NMR screens reported in Chapter 2 revealed some flaws that required addressing: (i) the use of absolute CSP magnitudes to calculate the pseudodistance, (ii) the difficulty of defining the consensus binding site, and (iii) the uneven number of CSPs within a proposed binding site.

In order to address these problems, the ADF program was updated and underwent significant modifications. The new version of the program, AutoDockFilter 2.0 (ADF 2.0), was written in Python (<http://www.python.org>). ADF 2.0 was designed to increase the usability and flexibility of the program. This included the incorporation of modifiable

parameters, the automatic output of the protein-ligand co-structure coordinate files, and easy implementation as a web-based program.

As a companion to the ADF 2.0 program, another program, CSP-Consensus (CSPC), was developed to minimize the guesswork involved with defining a consensus binding site from CSPs. This program uses hierarchical clustering of the distances between perturbed residues in order to generate a dendrogram. The program can then use a distance cutoff based on the size of the ligand to define the consensus binding site. Importantly, the program generates a grid map file that encompasses the size and position of the consensus binding site for easy use in AutoDock.

## 4.2 MATERIAL AND METHODS

**4.2.1 Modifications to AutoDockFilter 2.0.** The primary modification to the AutoDockFilter program involved how the pseudodistances are calculated. The original ADF program generated specific pseudodistances based on the absolute magnitude of the CSPs. Unfortunately, this caused problems when dealing with ligand binding events that had smaller CSPs than average because of weak binding interactions. The presence of small CSPs would result in long pseudodistances when absolute magnitude of the CSPs was employed. These long pseudodistances were unlikely to be violated resulting in NMR violation energies equal to zero for all of the docked poses. This effectively rendered the filtering of the AutoDock poses useless. As discussed in Chapter 1 [Equation 1.1], the magnitude of a chemical shift perturbation is related to the dissociation constant ( $K_D$ ) of the ligand and the concentration of the ligand.<sup>3</sup> If a 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC titration of protein-ligand pair that exhibited large CSPs were repeated with a

smaller concentration of ligand, the absolute magnitude of the CSPs would decrease despite the fact that the ligand would be located in the same binding site with identical intermolecular distances. Therefore, using the absolute magnitude to determine the pseudodistance would only be useful under very specific and defined circumstances. Instead, a relative CSP was used to calculate a pseudodistance by comparing the CSP for each amino acid relative to each other CSP in the dataset.

In ADF 2.0, the CSP magnitudes are converted to a pseudodistance ( $d_{\text{CSP}}$ ) as defined by a linear relationship between an upper and lower boundary (see eqn. 4.1). By default, ADF 2.0 uses 1.5 standard deviations from the median CSP magnitude to define the upper boundary ( $\text{CSP}_{\text{U}}$ ), while the median CSP magnitude is used to define the lower boundary ( $\text{CSP}_{\text{L}}$ ). The median is used instead of the mean to minimize the effect of extreme CSPs on defining the lower boundary. These boundaries are then converted to pseudodistances where the upper CSP boundary ( $\text{CSP}_{\text{U}}$ ) is set at a minimum pseudodistance of 3 Å ( $d_{\text{min}}$ ) and the lower CSP boundary ( $\text{CSP}_{\text{L}}$ ) is set at a longer pseudodistance of 6.5 Å ( $d_{\text{long}}$ ). These boundaries are used to calculate the pseudodistance ( $d_{\text{CSP}}$ ) for a specific CSP [Equation 4.1]:

$$d_{\text{CSP}} = \frac{(\text{CSP} - 2\text{CSP}_{\text{U}} - \text{CSP}_{\text{L}})}{(\text{CSP}_{\text{U}} - \text{CSP}_{\text{L}})/(d_{\text{long}} - d_{\text{min}})} \quad (4.1)$$

As in the original ADF, 3 Å is the default minimum pseudodistance that is assigned, while there is no limit to the maximum pseudodistance. ADF 2.0 has built in scalable settings so that these boundaries and distances can be set to different values during execution. CSPs that exceed upper CSP boundary are considered extreme perturbations and are treated differently than in the original ADF program.

The original ADF program set these extreme CSPs to the minimum pseudodistance of 3 Å. An evaluation of the amino acid residues with extreme CSPs showed that the majority of these residues were not significantly closer to the ligand than the other residues with relatively normal CSPs. This was not too surprising as extreme CSPs were likely the result of the residue experiencing other significant factors that influence the magnitude of the CSP, such as electrostatics, and ring current effects.<sup>4</sup> Therefore, the filtering process should not bias the results towards these residues with a small pseudodistance. In order to address this issue in ADF 2.0, extreme CSPs can now be set to a specific distance value (the default is 5 Å).

The new ADF 2.0 program also modifies the way the results are clustered by using a rank and file method similar to the clustering system used in the AutoDock program. The ligand poses are sorted from the lowest NMR violation energy to the highest. The first ligand pose is automatically assigned to the first cluster. The next pose is also assigned to the first cluster if its structural root mean squared deviation (RMSD) is less than a defined RMSD tolerance (default is 2 Å); otherwise, the pose becomes the first member of a second cluster. The process is continued for all remaining docked poses by comparing its structural RMSD to each cluster, and then either assigning the pose to the cluster with the lowest RMSD or forming a new cluster if the RMSD is not lower than the RMSD tolerance.

In order to evaluate the beneficial impact of the new pseudodistance calculation and clustering process, the AutoDock results and calculated CSPs for the protein-ligand systems described in Chapter 3 were repeated using the new ADF 2.0 program. The results obtained from ADF 1.0 and ADF 2.0 were then compared.

In addition to the fundamental changes to the ADF program, several beneficial tools were added. ADF 2.0 now has the option to modify most of the parameters used in the filtering calculation. Additionally, ADF 2.0 automatically outputs the protein-ligand costructure of the lowest energy cluster. Additional options include outputting the structures of other clusters as well as the costructure with the lowest AutoDock binding energy. ADF 2.0 has also been made available as a web-based client (courtesy of Brad Worley). Both the source code and web-based client can be found at <http://bionmr.unl.edu>.

**4.2.2 Defining the binding site with CSP-Consensus.** CSPs are commonly used as markers for identifying residues near to the bound ligand.<sup>5,6</sup> Ideally, every residue within 6 Å of that ligand would exhibit significant chemical shift changes, and the binding site would be clearly identified. In practice, this is rarely observed. Chemical shift changes occur due to a change in the chemical environment of the residue. These changes may result from the proximity of a bound ligand, the formation or breaking of hydrogen bonds, perturbations in electrostatics, presence or absence of ring current effects, or an overall changes in the protein structure resulting from a binding event.<sup>4</sup> Therefore, not all significant CSPs observed are the result of proximity to the ligand nor do all residues near a bound ligand exhibit significant CSPs. This ultimately makes defining the binding site more difficult than just selecting the most perturbed residues.

The program CSP-Consensus (CSPC) was developed in order to address this difficulty by selecting the residues with significant CSPs and identifying the residues that form a consensus binding site using the size of the ligand and hierarchical clustering

analysis. CSPC attempts to minimize the subjective nature of visually defining a binding site by mapping the perturbed residues to the surface of a protein.

CSPC requires the input of the residues with significant CSPs as well as the protein and ligand structures. From the list of significantly perturbed residues, a distance matrix is created from the amide-to-amide distance between each residue. This distance matrix is then subjected to a centroid/UPGMC hierarchical clustering algorithm<sup>7</sup>, which starts by grouping the two residues with the smallest pairwise amide-amide distance. A node is placed at the centroid distance between these two points. This new node is then treated as a single cluster for the next comparison, which looks for the shortest distance from another residue or cluster to this centroid distance. This process continues until all the residues have been clustered. The final clusters are defined by setting a maximum cluster distance threshold, which is based on the size of the ligand. By default, this threshold is set to one-half of the longest distance between any atom in the ligand plus 6 Å. This distance represents the possibility of clustering two perturbed residues that occur on opposite sides of the ligand. If no ligand file is input, the distance threshold is set to a default of 15 Å. The CSPC program outputs several files that include: (i) a results table, (ii) a dendrogram of the pairwise clustering, (iii) a CSP file to be used in ADF 2.0, and (iv) an AutoDock grid file with the size and location of the grid set around the clustered residues.

**4.2.3 Evaluation on protein-ligand systems with experimental CSPs.** The evaluation of the ability of ADF 2.0 to guide and filter AutoDock results was performed using 8 protein-ligand complexes (Table 4.1) in the RCSB Protein Data Bank<sup>8,9</sup> (PDB; <http://www.rcsb.org>) for which the chemical shifts of both the apo- and holo- forms of

the protein were available in the Biological Magnetic Resonance Data Bank (BMRB; <http://www.bmrb.wisc.edu>). The ligands were removed from the protein-ligand complex and saved as a separate coordinate file. Any missing heavy atoms for the amino acid residues were added using UCSF Chimera<sup>10</sup> (<http://www.cgl.ucsf.edu/chimera>). All hydrogen atoms were added to the protein and ligand using standard protonation states at pH 7.0.

**Table 4.1 RMSD comparison between the ligand-bound and unbound proteins**

PDB ID bound/unbound	RMSD (Å)			BMRB Entry bound/unbound
	full protein backbone	binding site backbone	binding site all atom	
1AKE/4AKE <sup>11,12</sup>	8.257	3.063	3.541	5746/5720
1EII/1B4M <sup>13,14</sup>	3.158	2.723	3.342	4682/4681
1JKN/1F3Y <sup>15,16</sup>	4.163	1.883	2.676	5054/4448
1JOK/1JOO <sup>17</sup>	5.092	0.902	2.725	494/495
1JW5/1JW4 <sup>18</sup>	0.296	0.259	1.112	4987/4986
1X6K/1SVJ <sup>19,20</sup>	1.755	0.661	1.724	6030/6029
2B8G/2B8F	0.870	1.526	3.278	6599/6600
2H3I/2H3F <sup>21</sup>	6.507	1.823	2.286	5960/7250

Ligands were also docked into the apo-structures for each of the 8 protein-ligand complexes in order to evaluate the utility of this approach in the likely scenario that a holoprotein structure is unavailable. The backbone coordinates of the apoprotein were aligned with the holoprotein structure prior to the AutoDock calculation. The ligand conformation in the original holoprotein structure was then used to measure an RMSD between the docked ligand conformers calculated using both the apo- and holoprotein structures.

The NMR-predicted binding site for each protein complex was determined using the published  $^1\text{H}$  and  $^{15}\text{N}$  chemical shift data obtained from the solution structure of each protein in both the unbound and bound forms. The magnitude of the CSPs were calculated using a common weighting approach:

$$CSP = \sqrt{\frac{\left(\frac{\delta_N}{5}\right)^2 + \delta_H^2}{2}} \quad (4.2)$$

where  $\delta_N$  and  $\delta_H$  represent the changes in  $^{15}\text{N}$  and  $^1\text{H}$  chemical shifts, respectively, upon ligand binding.<sup>22</sup>

In order to evaluate the capabilities of ADF 2.0 to identify the docking conformers that most agree with the experimental CSPs, the binding sites of the proteins were first defined by determining all residues with an amide group within 6 Å of the ligand in the holoprotein structure. The CSPs for each of these residues is then used by ADF 2.0 to filter the molecular docking results.

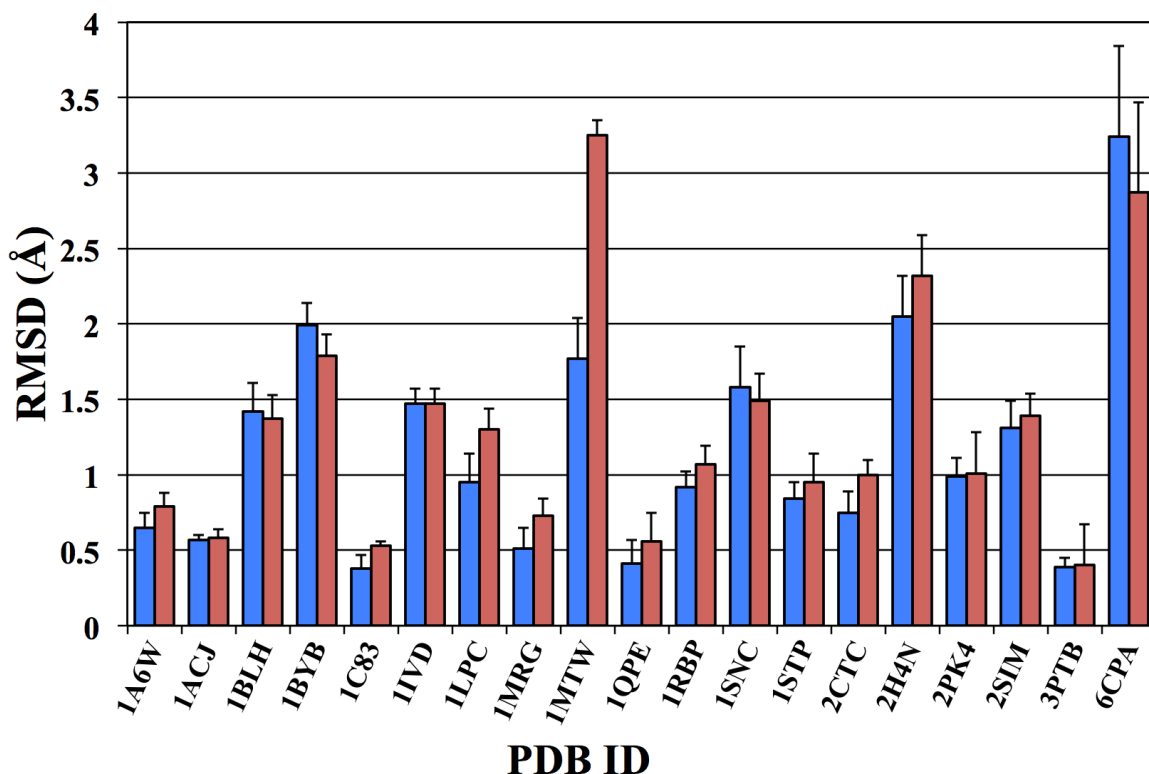
Unfortunately, identifying the binding site without prior knowledge would represent the most likely case for programs like AutoDock and ADF 2.0. Therefore, an evaluation of the ADF guiding and filtering process was performed using only knowledge of the chemical shifts for both the apo- and holoproteins. A preliminary binding site is determined by selecting residues with CSPs greater than one standard deviation from the mean. The residues that compose this preliminary binding site are then input into the new companion program, CSP-Consensus (described in section 4.2.2), to select a consensus binding site. The consensus binding site and the associated CSPs are then used to guide and filter the docking process.



AutoDock 4.2.3<sup>23-25</sup> with AutoDockTools 1.5.4<sup>25,26</sup> (<http://mgltools.scripps.edu>) graphical interface was used to simulate 120 different binding conformations for each protein-ligand pair. The grid maps were generated using 0.375 Å spacing and set to an appropriate size that encompasses all of the perturbed residues as well the size of the ligand. The docking calculations were performed using the Lamarckian genetic algorithm default settings with a population size of 300 and 2,500,000 energy evaluations. The results of each docked calculation were then input into the ADF 2.0 program using the default parameters.

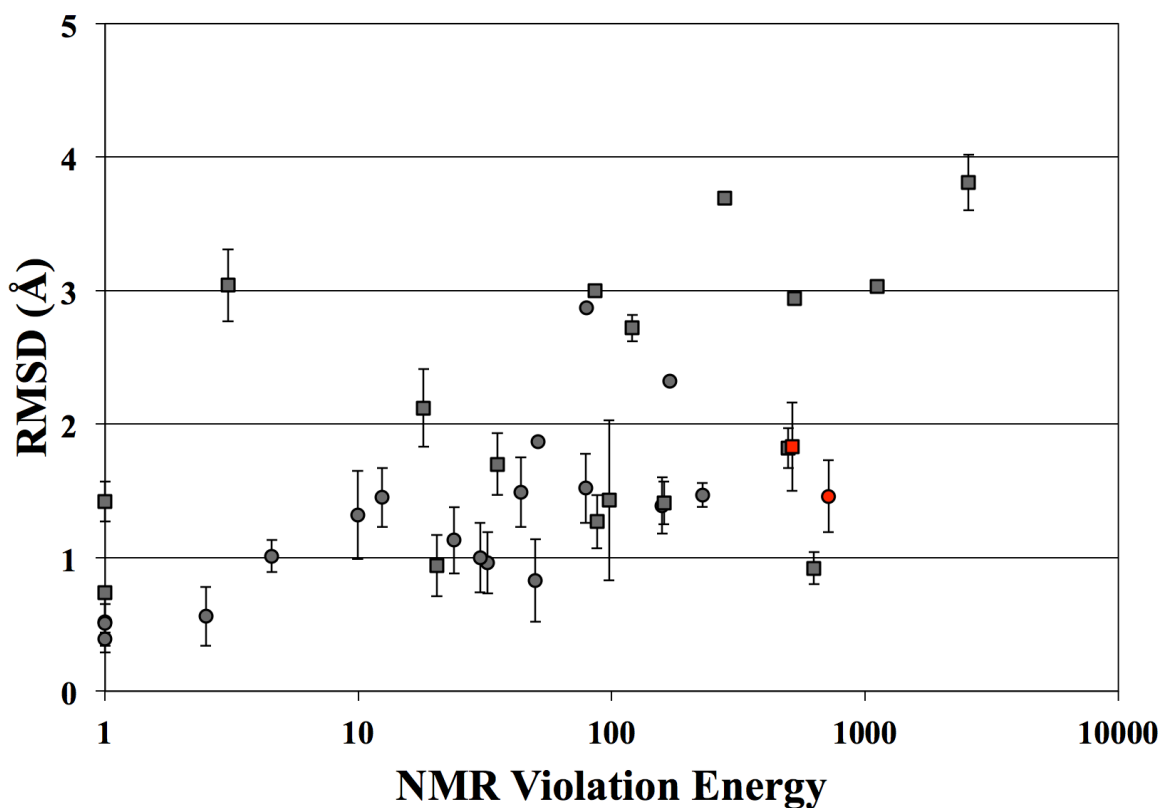
## 4.3 RESULTS AND DISCUSSION

**4.3.1 Modifications to AutoDockFilter 2.0.** The new ADF 2.0 software incorporates a significant modification to the calculation of the pseudodistances from CSP magnitudes. Instead of using absolute CSP magnitudes, relative CSP magnitudes are used to account for the effects of weaker binders. A comparison between the results obtained with ADF 1.0 and ADF 2.0 using the original 19 protein-ligand systems described in Chapter 3 is shown in Figure 4.1. As expected, ADF 2.0 performs essentially equivalent to the original ADF program since the original dataset did not have an inherent scaling problem. The absolute CSP magnitudes from the original 19 protein-ligand dataset were equivalent in magnitude to the CSPs used to define the linear relationship between CSPs and intermolecular distances. This illustrates that the new pseudodistance calculations used in ADF 2.0 results in the same relative NMR violation energy for a given protein-ligand system.



**Figure 4.1** Comparison of the RMSD values of the docked ligand conformers relative to the original ligand conformation in the protein-ligand X-ray structure. AutoDock calculations used the CSP-guided docking with filtering using the original ADF program (green) and the updated ADF 2.0 (red). ADF 2.0 does appear to mimic the success of the original program when dealing with calculated CSPs.

However, the biggest effect of changing to a relative CSP magnitude in ADF 2.0 is the comparison of NMR energies between protein-ligand systems. In Chapter 3, an obvious correlation existed between the NMR violation energy and the RMSD of the ADF-selected docked poses from the actual ligand pose found in the holoprotein structure. An NMR violation energy above 1,500 indicated an inaccurate result due to either poor docking results or structural changes between the free and bound versions of the protein. The use of relative CSP magnitudes in ADF 2.0 does appear to follow this same general correlation with regards to the calculated data [Figure 4.2]. But a clear distinction between a correct and incorrect model is not as clearly apparent.

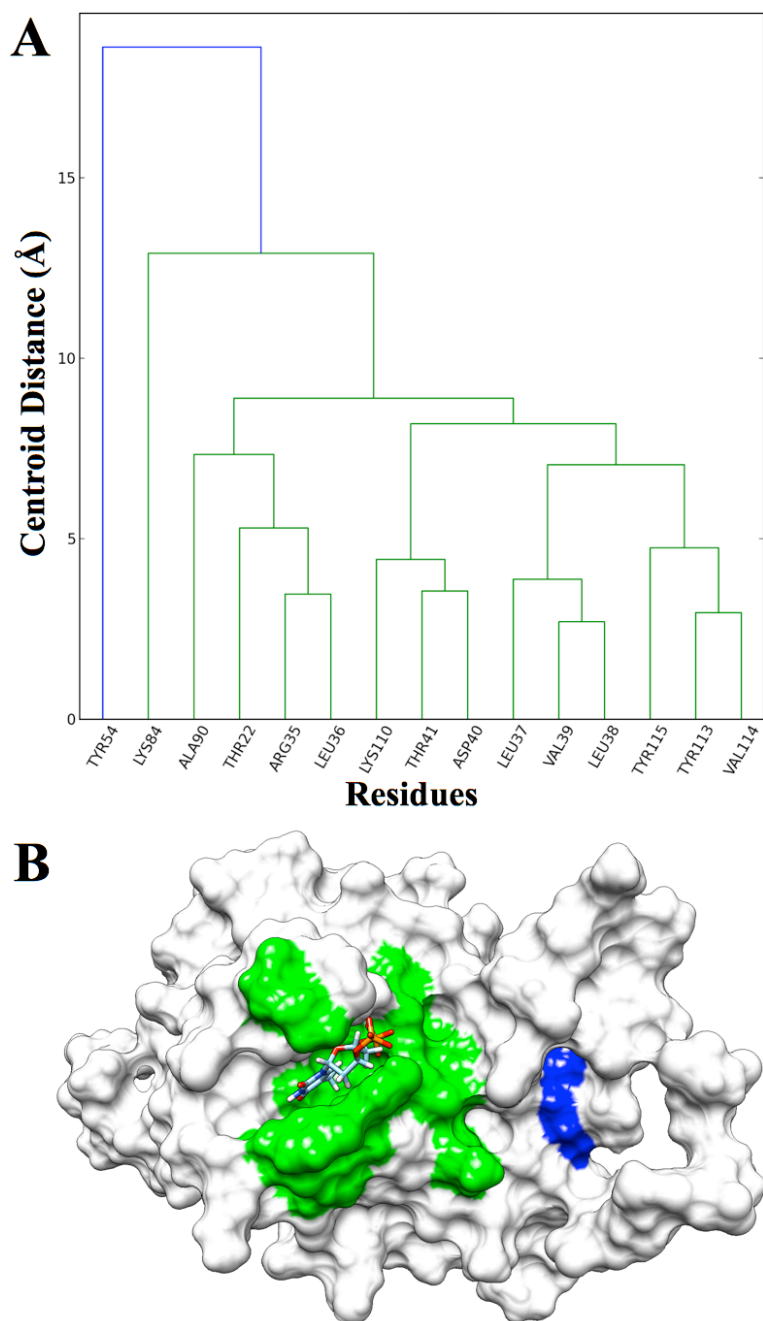


**Figure 4.2** Comparison of the NMR violation energy (logarithmic scale) against the corresponding RMSD for the best-docked conformers using calculated CSPs and the bound/free (circle/square) forms of the protein structure. The RMSD is relative to the ligand's conformation in the original X-ray structure. The red data points correspond to the docking results using the experimental CSPs for staphylococcal nuclease. Error bars represent variability within the best cluster selected by ADF 2.0.

**4.3.2 Defining the binding site with CSP-Consensus.** The CSPC program attempts to objectively evaluate whether residues exhibiting significant CSPs should be clustered together to define a consensus binding site using the size of the ligand and distance between each perturbed residue. The CSPC program outputs a dendrogram where the distance between each cluster represents the pairwise distance between the

centroids. A threshold distance that is based on the size of the ligand is then used to set the cutoff for the final cluster size.

The clustering technique used by CSPC is effective at clustering together perturbed residues that would visually appear to define a consensus binding site [Figure 4.3A,B]. In the case of thymidine-3',5'-bisphosphate (THP) binding with staphylococcal nuclease, 14 of the 15 residues that exhibited significant CSPs would be clustered together based on amide-amide distances. The excluded residue (Tyr 54) is 18.63 Å from the centroid of the other 14 residues. This would suggest that THP, with a length of 10.6 Å, is unlikely to cause a perturbation in Tyr54 due to proximity while also causing a similar perturbation to the 14 other residues.

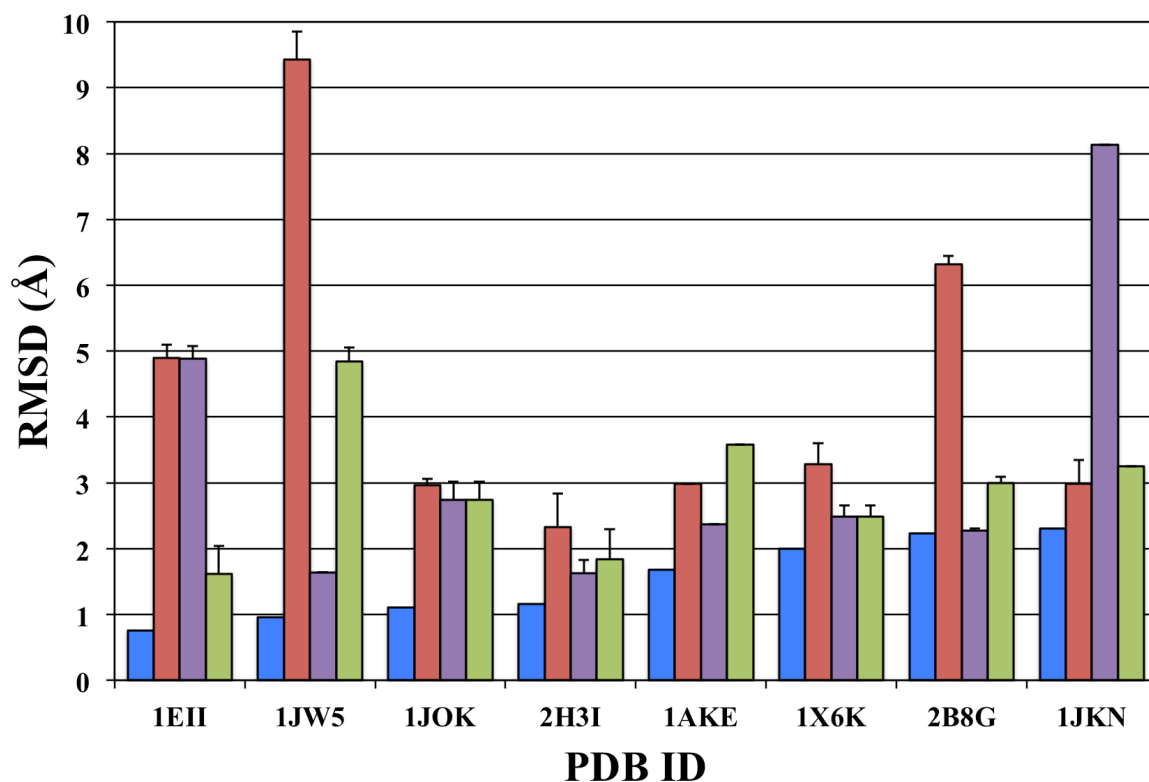


**Figure 4.3** (A) A dendrogram of amino acid residues in staphylococcal nuclease (SNase) that show a significant chemical shift perturbation upon binding of thymidine-3',5'-bisphosphate (THP). The residues are clustered using the centroid/UPGMC hierarchical clustering algorithm. Amino acids are placed in the same cluster if the node connecting two or more residues/clusters falls under a specified distance threshold that is defined by the size of the ligand causing the perturbation (green). In this case, Tyr54 (blue) is above the distance 11 Å distance threshold and is placed in its own cluster. (B) A representation of THP bound to SNase with the residues identified in the CSP-Consensus dendrogram are mapped onto the surface of the protein. Tyr54 (blue) is clearly separated from the consensus of perturbed residues and is excluded.

While these 14 residues are clustered together to form a consensus binding site to be used for guiding and filtering AutoDock, only 8 of these residues are within 6 Å of the correct ligand pose. While those 6 “extra” residues would make the AutoDock grid size bigger, the real concern involves the bias these “distant” residues may have on the filtering process. However, when dealing with protein-ligand systems where the holoprotein structure is not known, there is no foolproof way to exclude these 6 residues. It would be possible to combine information of predicted consensus binding site residues from CSPC and residues in predicted binding cavities from programs like CASTp<sup>27</sup> or ConSurf<sup>28</sup> to eliminate these residues.

**4.3.3 Evaluation on protein-ligand systems with experimental CSPs.** Because the development of the original ADF program, the availability of both apo- and holoprotein solution structures has allowed for the further evaluation of ADF using experimentally determined CSPs. The results of ADF 2.0 filtering on 8 distinct protein-ligand systems with experimental data were evaluated using two approaches. In the first approach, the CSPs used to define the binding site were selected based on which residues are known to be within 6 Å of the correct pose. This approach includes residues that would not be considered to have significant CSPs, and is intended to illustrate the capabilities of ADF 2.0 to filter the AutoDock results with the ideal binding site. The second evaluation explored the capabilities of ADF 2.0 filtering under experimental conditions where the binding site is determined based on only those residues that exhibit significant perturbations and form a consensus cluster in CSPC. Both of these approaches were then compared to the ligand pose AutoDock suggests is best based on binding

energy (first AutoDock cluster) as well as the best pose generated by AutoDock that agrees most with the correct ligand pose [Figure 4.4].



**Figure 4.4** Comparison of the RMSD values of the docked ligand conformers on the bound protein relative to the original ligand conformation in the protein-ligand NMR structure. The best possible docked pose calculated by AutoDock (blue) is compared to the docked poses with the best AutoDock binding energy (red), the best-docked poses selected by ADF 2.0 using the CSPs of the known binding site (purple), and the best-docked poses selected by ADF 2.0 using experimentally determined binding sites (green). Error bars represent variability within the best cluster selected by either AutoDock or ADF 2.0.

As indicated in Chapter 3, AutoDock is pretty effective at generating a docked pose that is close to the correct ligand pose ( $1.52 \pm 0.61$  Å RMSD). This is especially remarkable because only three of the protein-ligand systems (1eii, 1jok, 2b8g) have ligands with less than 10 torsional degrees of freedom, and one system (1ake) has a

ligand with 22 torsional degrees of freedom. However, when AutoDock ranks these docked poses by binding energy, the best-docked pose (first AutoDock cluster) is often not selected ( $4.40 \pm 2.42$  Å RMSD). This is likely due to the inherent error of approximately 2.2 kcal/mol when AutoDock calculates a binding energy.<sup>2</sup>

ADF 2.0 improves upon the results of AutoDock in almost every case when the best ADF 2.0 pose (first ADF 2.0 cluster) using the ideal binding site is selected ( $3.53 \pm 1.71$  Å RMSD). However, the best ADF-predicted costructures for 1eii and 1jkn still had poor RMSD values. A visual inspection of these costructures show that the poor RMSD values are the result of a flipped ligand pose being selected. This is not unexpected as ADF only evaluates the minimum distance from any atom in the ligand to the amide of the residues; it does not specify that a particular ligand atom must be near the residue. This leads to flipped ligands that have small NMR violation energies. While this is not ideal for predicting ligand poses for a drug discovery effort, it does not hinder the ability of ADF 2.0 to predict the binding site that could then be compared to other binding sites for functional annotation using FAST-NMR and CPASS as described in Chapter 2. If these ligand poses were flipped back, the average RMSD would significantly improve to an RMSD of  $2.25 \pm 0.60$  Å.

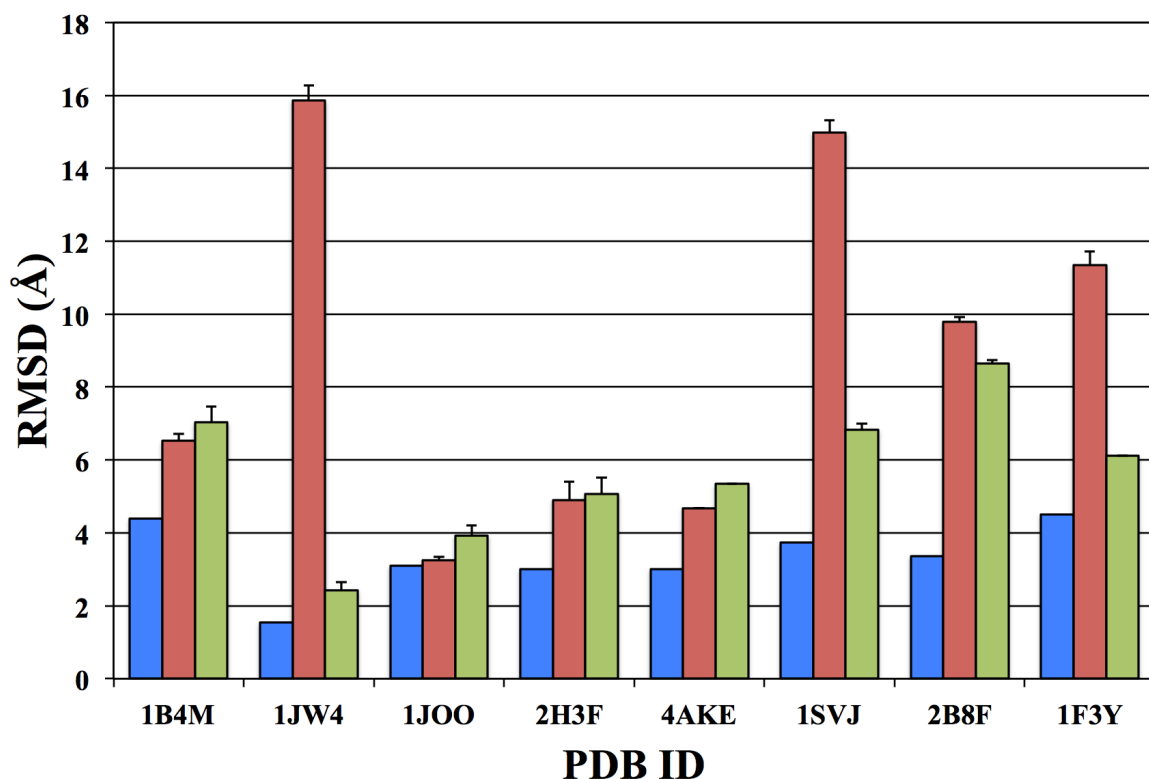
ADF 2.0 still improves upon the results of AutoDock when using ligand-defined binding sites from experimental CSPs ( $2.92 \pm 1.02$  Å RMSD). As previously described, experimentally-defined binding sites may include residues that are not within 6 Å of the correct ligand pose. These residues at the edge of the predicted ligand binding site may bias the filtering of the AutoDock poses. This is exactly the case with the best ADF pose for 1jw5, where the extra residues from the experimentally-defined binding site had



greater perturbations compared to the residues of the ideal binding site. This effectively shifts the selection of the best ADF pose closer to the extra residues with greater CSPs. Despite this observation, the filtering results for ADF 2.0 still significantly improved upon the best-docked structures predicted by AutoDock without ADF.

The new ADF 2.0 has also been shown to improve upon the AutoDock results when using CSPs to dock a ligand into an apoprotein structure. [Figure 4.5]. The best-docked pose generated by AutoDock using an apoprotein structure has a higher RMSD relative to the correct docked pose from the holoprotein structure ( $3.33 \pm 0.93$  Å RMSD). This difference is likely a result of errors associated from aligning the bound and unbound forms of the structure as well as any structural differences in the binding site. Again, the best-docked poses predicted by AutoDock using only binding energies fails to correlate with the correct ligand pose in the holoprotein structure ( $8.92 \pm 4.84$  Å RMSD). While some of this failure is due to structural differences between the bound and unbound forms, four of the predicted poses have an RMSD of upwards of 16 Å. The ADF 2.0 predicted poses are essentially equivalent to or significantly better than the AutoDock predicted poses ( $5.68 \pm 1.94$  Å RMSD), yet these results are consistently worse than the results seen using the bound form of the protein. The structural differences between the apo- and holo- structures can affect several factors such as: (i) correctly defining the consensus binding site, (ii) side chain residues sterically hindering access into the binding site, (iii) distances between perturbed residues may be further apart in the apostructure, increasing the NMR violation energy for the correct pose. These factors appear to have a significant effect on the results of the docking and filtering. Thus, using an apostructure is more likely to lead to a lower accuracy costructure due to the rigid nature of the protein

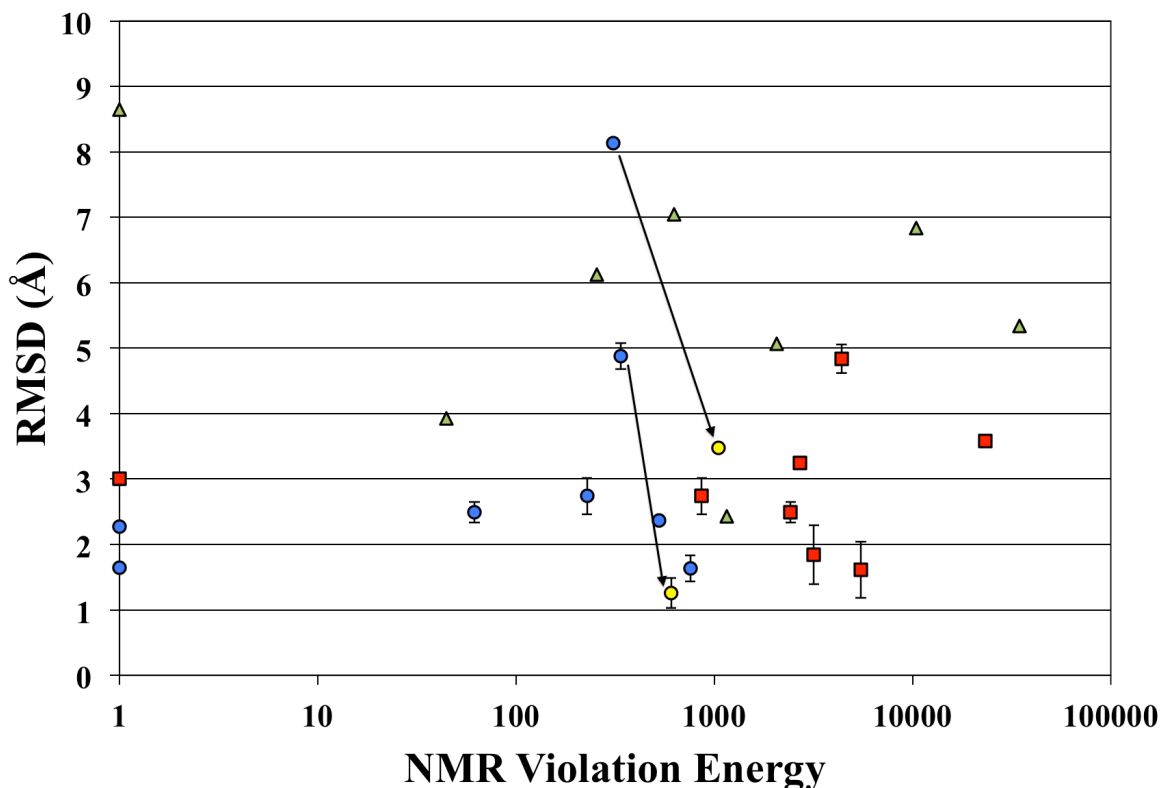
during docking. Unfortunately, this issue can only really be addressed by implementing fully flexible proteins into the docking process or utilizing protein structure ensembles, which comes with its own share of difficulties.<sup>1,29-32</sup>



**Figure 4.5** Comparison of the RMSD values of the docked ligand conformers on the unbound protein relative to the original ligand conformation in the protein-ligand NMR structure. The best possible docked pose calculated by AutoDock (blue) is compared to the docked poses with the best AutoDock binding energy (red), and the best-docked poses selected by ADF 2.0 using experimentally determined binding sites (green). Error bars represent variability within the best cluster selected by either AutoDock or ADF 2.0.

The NMR violation energies when evaluated with the experimental data indicate that there is no real trend between an absolute NMR violation energy and RMSD from correct pose when comparing different protein-ligand systems [Figure 4.6]. Essentially, the NMR violation energy has little value outside the particular protein-ligand system being investigated. This is not too surprising since experimental binding sites will include

some extraneous residues and may be missing other relevant residues. Additionally, CSPs are not solely affected by proximity, so the relative perturbation of a residue in one protein-ligand system does not necessarily have the same pseudodistance in another protein-ligand system despite a similar relative perturbation. This effectively eliminates the utility of using the NMR violation energy as a direct evaluation of the accuracy of the predicted costructure. However, there may still be value in the NMR violation energy with respect to the identity of residues with significantly violated pseudodistances. The per residue violation energy may indicate residues that should be excluded from the binding site definition or residues that should be defined as conformationally flexible in the AutoDock simulation.



**Figure 4.6** Comparison of the NMR violation energy (logarithmic scale) against the corresponding RMSD for the best-docked conformers using the experimental CSPs of the known binding site (blue circle) and the experimentally predicted binding site using the holoprotein structure (red square) and apoprotein structure (yellow triangle). The RMSD is relative to the ligand's conformation in the original NMR structure. The yellow points and arrows represent the corrected RMSD and NMR violation energy for the docked poses that are flipped. Error bars represent variability within the best cluster selected by ADF 2.0.

#### 4.4 CONCLUSIONS

AutoDockFilter has previously demonstrated the benefits of using chemical shift perturbations from a 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR experiment to guide and filter the results of docking calculations from AutoDock. AutoDockFilter 2.0 builds upon the success of the original program by implementing the use of relative CSP magnitudes to calculate the pseudodistance while adding features to expand the usability and flexibility of the

program. AutoDockFilter 2.0 was able to identify accurate protein-ligand costructures using CSPs from the ideal binding site ( $2.25 \pm 0.60$  Å RMSD) and the experimentally determined binding site ( $2.92 \pm 1.02$  Å RMSD). This is a definite improvement over the best protein-ligand costructures predicted by AutoDock alone ( $4.40 \pm 2.42$  Å RMSD). Additionally, a companion program, CSP-Consensus, was developed in order to address the difficulties of determining a consensus binding site. CSP-Consensus calculates amide-amide distances between perturbed residues and uses hierarchical clustering and the size of the ligand to identify the residues that best represent a consensus binding site.

#### 4.5 REFERENCES

1. Stark, J. L. & Powers, R. Rapid protein-ligand costructures using chemical shift perturbations. *J Am Chem Soc* **130**, 535–545 (2008).
2. Rosenfeld, R. J. *et al.* Automated docking of ligands to an artificial active site: augmenting crystallographic analysis with computer modeling. *J Comput Aided Mol Des* **17**, 525–536 (2003).
3. Fielding, L. NMR methods for the determination of protein-ligand dissociation constants. *Prog Nucl Mag Res Spect* **51**, 219–242 (2007).
4. Oldfield, E. Chemical shifts in amino acids, peptides, and proteins: from quantum chemistry to drug design. *Annu Rev Phys Chem* **53**, 349–378 (2002).
5. Carlomagno, T. Ligand-target interactions: what can we learn from NMR? *Annu Rev Biophys Biomol Struct* **34**, 245–266 (2005).
6. Lepre, C. A., Moore, J. M. & Peng, J. W. Theory and applications of NMR-based screening in pharmaceutical research. *Chem Rev* **104**, 3641–3676 (2004).
7. Jain, A. N. Bias, reporting, and sharing: computational evaluations of docking methods. *J Comput Aided Mol Des* **22**, 201–212 (2008).
8. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
9. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **10**, 980 (2003).
10. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612 (2004).
11. Müller, C. W. & Schulz, G. E. Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap5A refined at 1.9 Å resolution. A model for a catalytic transition state. *J Mol Biol* **224**, 159–177 (1992).
12. Müller, C. W., Schlauderer, G. J., Reinstein, J. & Schulz, G. E. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding.

- Structure* **4**, 147–156 (1996).
13. Lu, J. *et al.* Binding of retinol induces changes in rat cellular retinol-binding protein II conformation and backbone dynamics. *J Mol Biol* **300**, 619–632 (2000).
  14. Lu, J. *et al.* The structure and dynamics of rat apo-cellular retinol-binding protein II in solution: comparison with the X-ray structure. *J Mol Biol* **286**, 1179–1195 (1999).
  15. Fletcher, J. I., Swarbrick, J. D., Maksel, D., Gayler, K. R. & Gooley, P. R. The structure of Ap(4)A hydrolase complexed with ATP-MgF(x) reveals the basis of substrate binding. *Structure* **10**, 205–213 (2002).
  16. Swarbrick, J. D. *et al.* The three-dimensional structure of the nudix enzyme diadenosine tetraphosphate hydrolase from *Lupinus angustifolius* L. *J Mol Biol* **302**, 1165–1177 (2000).
  17. Wang, J. *et al.* Solution structures of staphylococcal nuclease from multidimensional, multinuclear NMR: nuclease-H124L and its ternary complex with Ca<sup>2+</sup> and thymidine-3',5'-bisphosphate. *J Biomol NMR* **10**, 143–164 (1997).
  18. Duan, X. & Quioco, F. A. Structural evidence for a dominant role of nonpolar interactions in the binding of a transport/chemosensory receptor to its highly polar ligands. *Biochemistry* **41**, 706–712 (2002).
  19. Haupt, M. *et al.* The holo-form of the nucleotide binding domain of the KdpFABC complex from *Escherichia coli* reveals a new binding mode. *J Biol Chem* **281**, 9641–9649 (2006).
  20. Breitenlechner, C. B. *et al.* Structure-based optimization of novel azepane derivatives as PKB inhibitors. *J Med Chem* **47**, 1375–1390 (2004).
  21. Saad, J. S. *et al.* Structural basis for targeting HIV-1 Gag proteins to the plasma membrane for virus assembly. *Proc Natl Acad Sci USA* **103**, 11364–11369 (2006).
  22. Garrett, D. S., Seok, Y.-J., Peterkofsky, A., Clore, G. M. & Gronenborn, A. M. Identification by NMR of the binding surface for the histidine-containing phosphocarrier protein HPr on the N-terminal domain of enzyme I of the *Escherichia coli* phosphotransferase system. *Biochemistry* **36**, 4393–4398 (1997).
  23. Morris, G., Goodsell, D., Halliday, R. & Huey, R. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**, 1639–1662 (1998).
  24. Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* **28**, 1145–1152 (2007).
  25. Morris, G. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* (2009). doi:10.1002/jcc.21256
  26. Sanner, M. F. Python: a programming language for software integration and development. *J Mol Graph Model* **17**, 57–61 (1999).
  27. Dundas, J. *et al.* CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res* **34**, W116–8 (2006).
  28. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* **38**, W529–33 (2010).
  29. Huang, S.-Y. & Zou, X. Ensemble docking of multiple protein structures:

- considering protein structural variations in molecular docking. *Proteins* **66**, 399–421 (2007).
30. Erickson, J. A., Jalaie, M., Robertson, D. H., Lewis, R. A. & Vieth, M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J Med Chem* **47**, 45–55 (2004).
  31. Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A. & Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem* **49**, 534–553 (2006).
  32. B-Rao, C., Subramanian, J. & Sharma, S. Managing protein flexibility in docking and its applications. *Drug Discov Today* (2009). doi:10.1016/j.drudis.2009.01.003

## CHAPTER 5

### EVALUATION OF FUNCTIONAL SIMILARITY WITH CPASS 2.0

#### 5.1 INTRODUCTION

Determining the structural and biochemical functions of proteins is essential to our understanding of biology and the potential application for medical advances. The success of the various genome projects<sup>1,2</sup> is rapidly increasing the number of known protein sequences (~27,000,000) in the UniProtKB database.<sup>3,4</sup> Obviously, an experimental determination of a function for each of these proteins is not feasible, thus, computational approaches are often used to annotate the function based on global sequence or structural homology to proteins with an experimentally determined function.<sup>5</sup> Despite this approach, nearly 40% of these proteins are still functionally unannotated.

The molecular function of a protein is often defined by the structural arrangement and chemical properties of the amino acids that interact with a functionally relevant molecule. This binding site, or active site, tends to be more evolutionarily stable than the rest of the protein. Correspondingly, proteins with a similar function would likely also have similar active sites.<sup>6</sup> There are several computational tools that attempt to predict the location of a protein active site by finding common sequence motifs,<sup>7,8</sup> ligand “hotspots”,<sup>9</sup> structural cavities/clefts,<sup>10-12</sup> evolutionarily conserved residues,<sup>13-16</sup> similar molecular surfaces,<sup>17</sup> or structural motifs.<sup>18-20</sup> However, the predictive nature of these approaches often leads to the identification of ambiguous binding sites, which makes functional characterization difficult.



The Comparison of Protein Active Site Structures (CPASS; <http://cpass.unl.edu>)<sup>21,22</sup> program is used in conjunction with the FAST-NMR<sup>23,24</sup> methodology (described in Chapter 2) to compare the sequence and structure of an experimentally-determined binding site to a database of binding sites in order to elucidate a functional relationship. Because CPASS defines a binding site as every amino acid residue within 6 Å of a bound ligand, a comprehensive database of binding-sites can be created from the protein-ligand costructures found in the RCSB Protein Data Bank<sup>25,26</sup> (PDB; <http://www.rcsb.org>). Therefore, a CPASS calculation performs an exhaustive search, where a comparison occurs between the query binding site and every unique binding site from the PDB.

The CPASS program and database has been recently updated to version 2, which adds significant enhancements that reduce the time to run a calculation to ~1 hour, improves the user interface, and increases the size of the CPASS database to ~36,000 distinct binding sites.<sup>22</sup> Additionally, the CPASS v2.0 similarity function has been modified to include three new features that enriches the selection of similar active sites: (1) a comparison of the solvent accessible surface area (SASA) of the aligned residues; (2) the root mean square different (RMSD) between the two bound ligands in the active site; and (3) the addition of the C $\beta$  position to the distance calculation between aligned residues.<sup>22</sup> This chapter focuses on the enhanced capabilities of CPASS v2.0 to identify similar protein active sites; and the potential effect of experimental variability on the reliability of the CPASS similarity score. Additionally, CPASS v2.0 was used during the functional annotation of 21 proteins of unknown function (described in Chapter 2), and the overall trends are reported here.

## 5.2 MATERIAL AND METHODS

**5.2.1 Evaluation of CPASS functional similarity.** Evaluating the reliability of CPASS to identify a functional homolog requires a quantitative approach to define functional similarity, which is difficult to determine.<sup>27</sup> Gene Ontology (GO) terms<sup>28</sup> are currently the most common approach to functional annotation, but they are often incomplete, generic, and often dependent on global sequence homology. Conversely, the Enzyme Commission (E.C.) classification,<sup>29</sup> provides a well-defined, hierarchical approach to define functional similarity between enzymes due to its focus on enzyme activity. Thus, E.C. was chosen to evaluate the ability of CPASS to identify functional homologs.

The capabilities of CPASS v2.0 were evaluated using two different proteins with enzymatic activity: aspartate transaminase (PDB: 1yaa; E.C. 2.6.1.1);<sup>30</sup> and glutamine-tRNA ligase (PDB: 1gtr; E.C. 6.1.1.18).<sup>31</sup> The ligand binding site from each protein structure was compared against the entire CPASS database of ~36,000 ligand-defined binding sites using the default CPASS v2.0 parameters, which includes the new additions to the similarity function of ligand RMSD, solvent accessible surface area, and the C $\beta$  position within the distance calculation.

Three different methods were used to define what constitutes as a functionally similar active site (true positive). The first method only considered proteins with the same E.C. classification (i.e., all four E.C. numbers are identical) as true positives. The second method used a broader definition of E.C. classification (i.e., only the first three E.C. numbers are identical). The third method used a very broad definition of functional

homology by defining all active sites that bind the same ligand as the query protein as being functionally similar. Receiver operating characteristics (ROC) curves (described in Chapter 1) were generated using the three different definitions of a true positive. The true positive rates were plotted against false positive rates over the full range of CPASS similarity scores. In addition, distribution curves were used to plot the fraction of negatives and positives at each CPASS similarity score using a bin size of 10. The fraction simply corresponds to the number of positives or negatives per bin relative to the total number of positives or negatives.

**5.2.2 Tolerance of active-site variations.** CPASS was designed to compare ligand binding sites. Ideally, most ligand binding sites would be identified from an experimentally determined protein-ligand costructure, but that approach is not always available. In some cases, the location of the ligand binding site may be inferred from other sources, such as site-directed mutagenesis, NMR chemical shift perturbations, bioinformatics, or computer modeling. However, these approaches may introduce variations in the location of the binding site that would negatively affect the CPASS comparison.

In many cases, models of protein-ligand costructures are generated from NMR or X-ray structures of the unbound form of the protein. However, many proteins undergo significant structural changes upon the binding of a ligand, which may lead to significant changes in the binding site of the protein. Since the CPASS database is generated from experimentally determined protein-ligand costructures from the PDB, CPASS comparisons to ligand-defined binding sites using the unbound protein structure could hinder the ability of CPASS to identify a similar binding site. Therefore, comparing a

ligand-defined binding site using the unbound protein structure may hinder the ability of CPASS to identify a similar binding site.

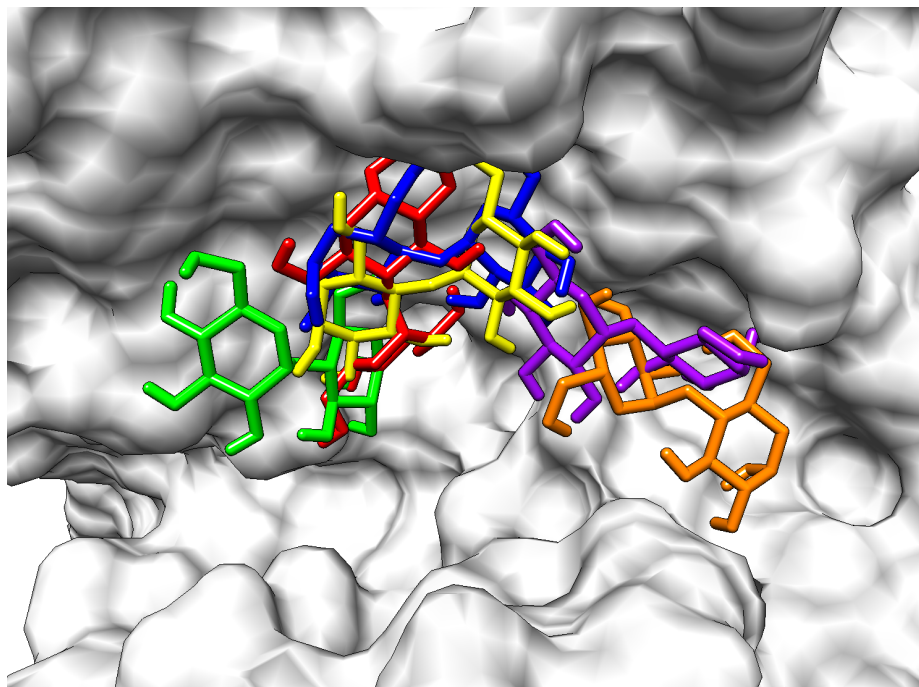
To evaluate the effect that structural changes in the bound and unbound forms of the protein may have on the CPASS comparison, 10 proteins with known bound and unbound structures (Table 5.1) were selected from the PDB for comparison in CPASS. The bound and unbound forms of the protein were structurally aligned in UCSF Chimera<sup>32</sup> (<http://www.cgl.ucsf.edu/chimera>). After the alignment, the coordinates of the ligand in the bound protein structure were added to the unbound protein structure. CPASS calculations with default parameters were performed on both the bound structure and unbound structure with the added ligand coordinates.

**Table 5.1 RMSD and residue comparison between the ligand-bound and unbound proteins**

PDB ID bound/unbound	RMSD(Å)		number of binding site residues bound/unbound (matched)
	binding site backbone	binding site all atom	
1AKE/4AKE <sup>33,34</sup>	3.063	3.541	67/42 (27)
1EII/1B4M <sup>35,36</sup>	2.723	3.342	45/42 (33)
1JKN/1F3Y <sup>37,38</sup>	1.883	2.676	31/19 (18)
1JOK/1JOO <sup>39</sup>	0.902	2.725	19/20 (19)
1JW5/1JW4 <sup>40</sup>	0.259	1.112	21/21 (21)
1MX8/1MX7	1.565	2.323	39/37 (35)
1X6K/1SVJ <sup>41,42</sup>	0.661	1.724	35/32 (30)
2B8G/2B8F	1.526	3.278	10/10 (10)
2H3I/2H3F <sup>43</sup>	1.823	2.286	31/33 (30)
2K0X/2K0Y	0.920	2.362	11/13 (11)

In addition to the structural differences between the bound and unbound forms of the protein, determining the exact location, or pose, of the ligand with respect to the protein can introduce additional variability when compared to an experimentally determined protein-ligand costructure. Because CPASS defines the ligand binding site for a protein as any residue within 6 Å of the ligand, the location of the ligand is important. Tools like NMR chemical shift perturbations and molecular docking can be combined to improve the accuracy of the ligand pose (as seen in Chapter 3 and Chapter 4); however, trying to predict the exact ligand pose is still very challenging.

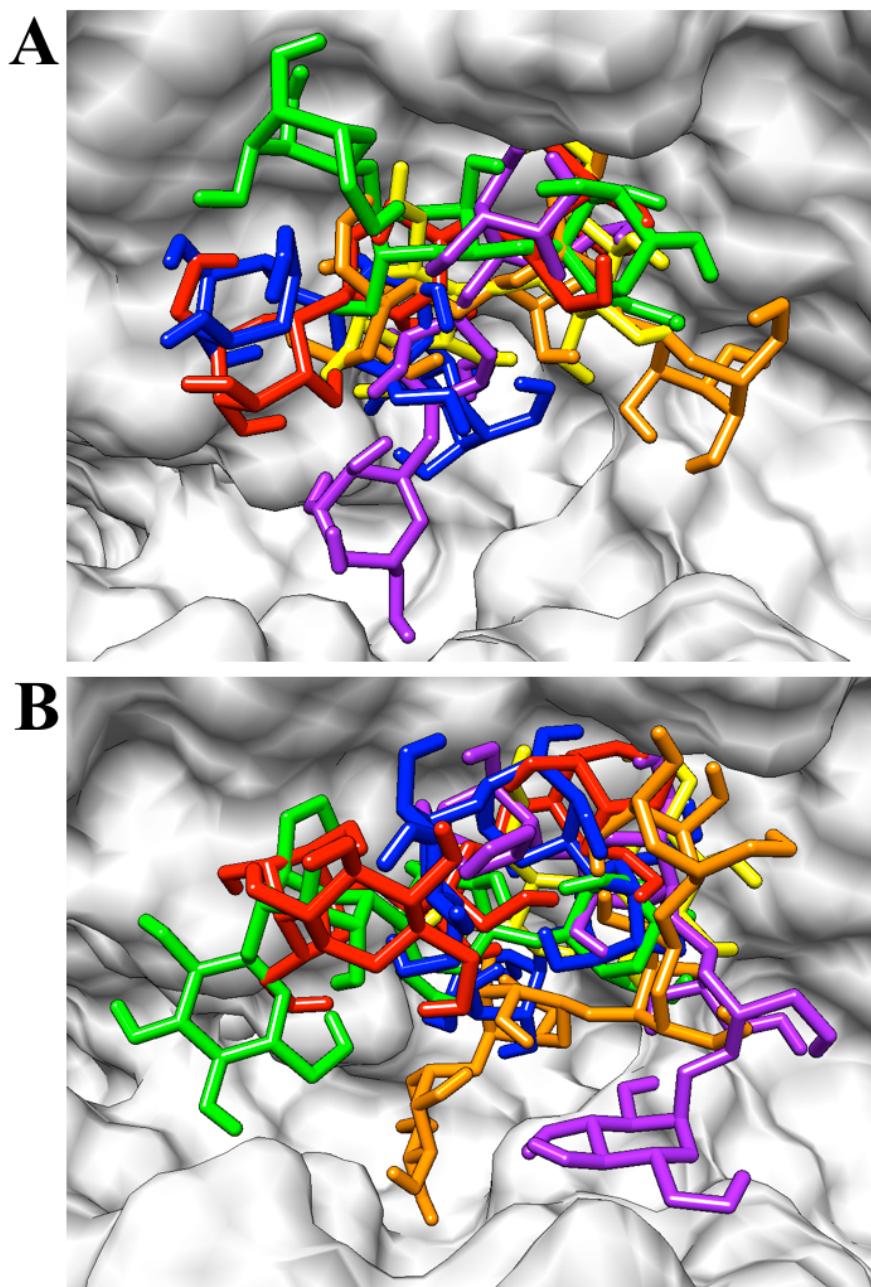
The effect of the variability of a ligand's pose was evaluated by redocking maltose into the bound form of maltodextrin-binding protein (PDB ID: 1ANF).<sup>40</sup> AutoDock 4.2.3<sup>44-46</sup> with the AutoDockTools 1.5.4<sup>46,47</sup> (<http://mgltools.scripps.edu>) graphical interface was used to simulate 100 different ligand poses for maltose docked with the maltodextrin-binding protein. The docking calculation was done using the Lamarckian genetic algorithm with default parameters. From the results of this calculation, 5 poses were selected to represent increasing variation from the correct pose found in the bound protein structure [Figure 5.1]. The coordinates of these 5 poses were then added to the bound structure file. The maltose-maltodextrin-binding protein docked model was then submitted to CPASS using default parameters.



**Figure 5.1** An illustration of docked maltose in the binding pocket of maltodextrin-binding protein (PDB: 1JW5). The experimental ligand conformation (yellow) is compared to 5 docked ligand poses of decreasing accuracy to the correct structure: 1.50 Å RMSD (blue); 2.02 Å RMSD (red); 3.27 Å RMSD (green); 4.81 Å RMSD (purple); and 6.77 Å RMSD (orange).

In many cases where the protein has an unknown function, the identity of the molecule(s) that interact with that protein is also unknown. High-throughput screens of a compound library are often used to help identify molecules that show specific binding to the protein in question. Depending upon the composition of the compound library, compounds that are shown to be binders are probably not the natural ligand for the protein *in vivo*. However, these compounds often have similar physicochemical features to the natural ligand, but some differences in the binding pose of the ligand would be expected. CPASS does not directly use the ligand in the database search, but it does define the size of the ligand binding site. Thus, the identity of the ligand may affect the CPASS outcome.

In order to test the effect of ligand size on CPASS performance, two larger saccharides similar to maltose, 3-glucose amylose and 4-glucose amylose, were also used to define the ligand binding site for a CPASS search. 3-glucose amylose and 4-glucose amylose were docked to the bound form of maltodextrin-binding protein<sup>40</sup> using AutoDock and the same docking parameters described above. From each of the docking calculations, 5 unique poses were selected that show significant overlap with the true binding pose of maltose, but also exhibited an extension in the size of the ligand binding site [Figure 5.2A,B]. These unique poses were then submitted to CPASS for evaluation.



**Figure 5.2** An illustration of the experimental pose of maltose (yellow) and 5 docked poses of (A) 3-glucose amylose and (B) 4-glucose amylose in the binding pocket of maltodextrin-binding protein (PDB: 1JW5). The docked poses of each sugar were selected based on a similarity to the original maltose pose and sampling of unique space around the maltose binding site.

**5.2.3 CPASS comparisons of proteins of unknown function.** To date, most of the evaluations of CPASS used well-characterized proteins of known function, where a



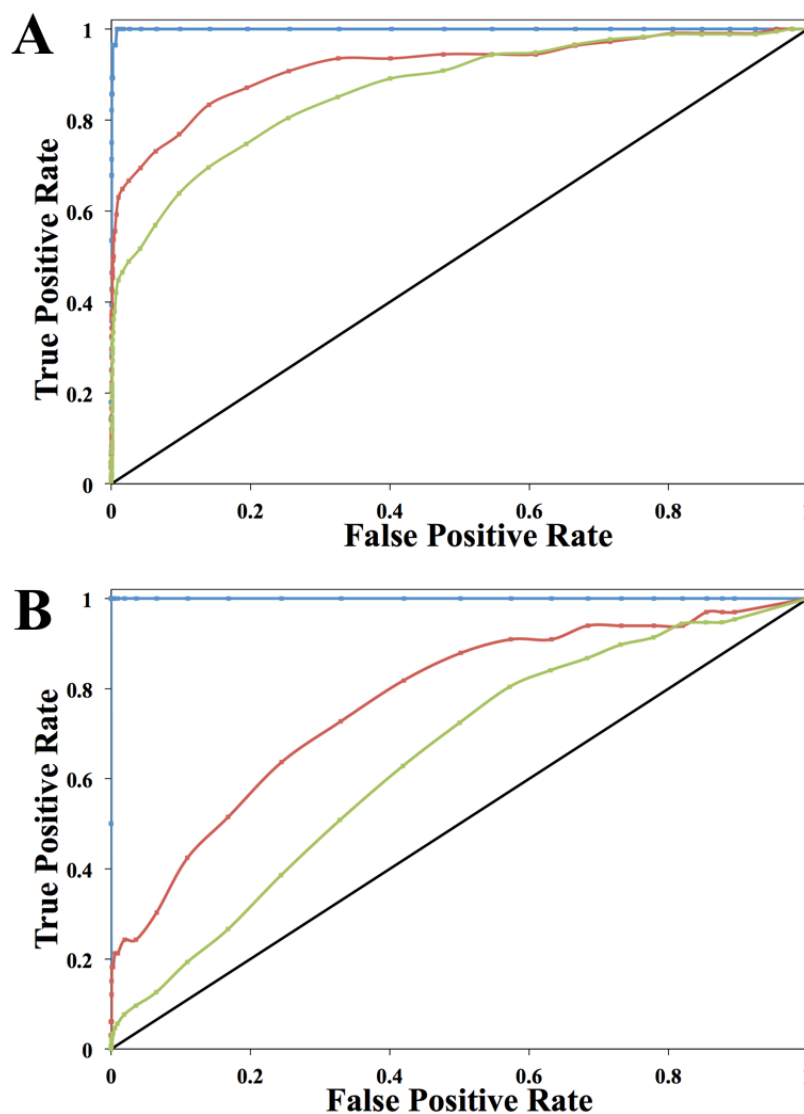
correct result was known and could be found by CPASS. Additionally, CPASS has previously contributed towards the functional annotation of a few other proteins with unknown functions: *Staphylococcus aureus* protein SAV1430,<sup>23</sup> *Pseudomonas aeruginosa* protein PA1324,<sup>48</sup> *Pyrococcus horikoshii* protein PH1320,<sup>24</sup> *Homo sapiens* protein Q13206,<sup>24</sup> and *Salmonella typhimurium* protein PrgI.<sup>49</sup> However, evaluating the performance of CPASS under experimental conditions similar to FAST-NMR is necessary for the continued improvement of the program.

Twenty-one proteins of unknown function from the Northeast Structural Genomics Consortium (NESG; <http://www.nesg.org>) were screened using the FAST-NMR approach (see Chapter 2), which resulted in the identification of compounds shown to bind to each protein and the generation of protein-ligand costructures using a guided and filtered molecular docking approach (see Chapter 3 and Chapter 4). The experimentally determined, ligand-defined binding sites from these costructures were then compared to other ligand-defined binding sites from the PDB using CPASS v2.0.

## 5.3 RESULTS AND DISCUSSION

**5.3.1 Evaluation of CPASS functional similarity.** Receiver operating characteristics (ROC) curves for the CPASS calculations of aspartate transaminase [Figure 5.3A] and glutamine-tRNA ligase [Figure 5.3B] illustrate the overall ability of CPASS to identify true positives relative to false positives. The straight-line in the graph indicates the expected results if the CPASS predictions were completely random. Curves that approach the upper-left of the graph indicate better performance of the method to select true positives from the database based on CPASS similarity scores. As apparent in

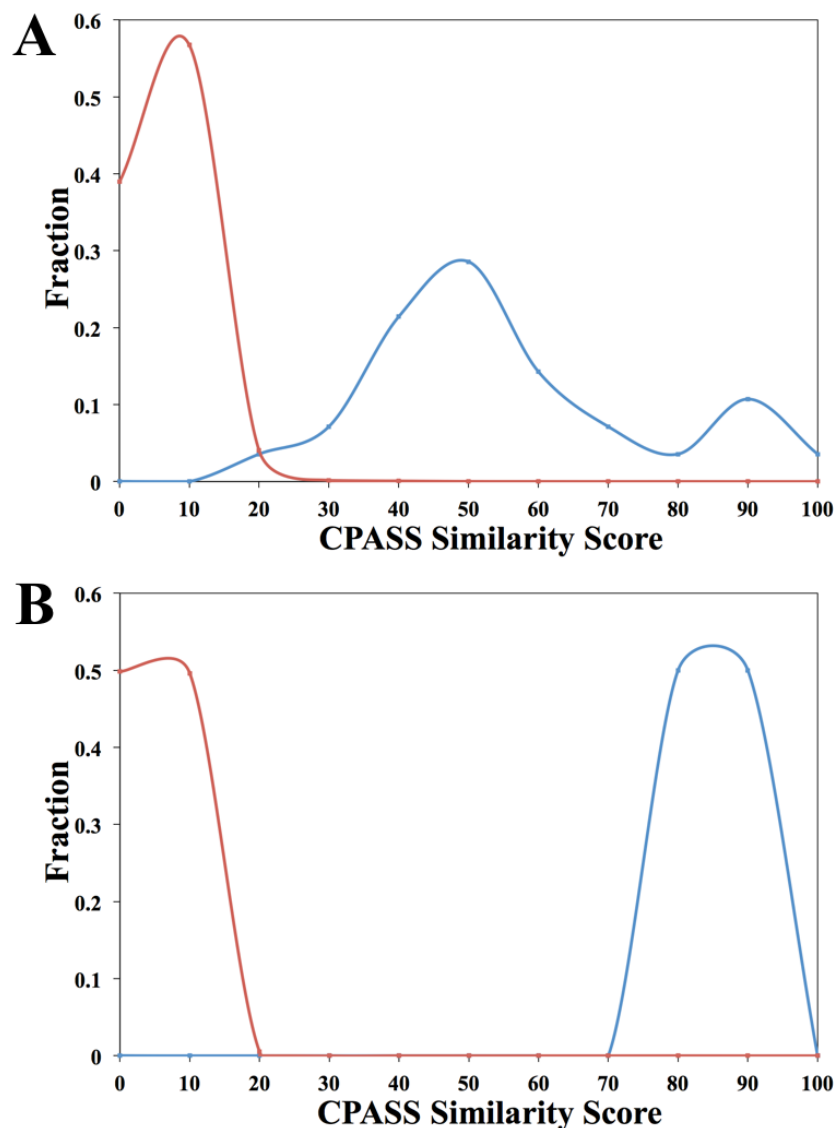
[Figure 5.3A,B], the enrichment in the ROC curves and the corresponding improvement in CPASS performance follow the increasing strictness in the functional classification of true positives. The ROC curves, where true positives are based on identical E.C. numbers, is essentially ideal. The ROC curve based on a broader E.C. similarity (only the first three numbers are identical) performs worse than the ROC curves of the stricter E.C. classification. However, this is still an improvement over the ROC curves that are generated when true positives are based only on proteins binding the same ligand. These results are not surprising since a stricter classification of function minimizes the number of proteins that are incorrectly characterized as true positives. It is important to note that in all cases the number of true positives based on identical E.C. numbers for both proteins is extremely small (2 – 28) relative to the size of the CPASS database (~36,000), even for proteins that are well represented in the PDB. This occurs because the CPASS database has been deliberately filtered to remove identical or highly similar ligand-binding sites.



**Figure 5.3** ROC curves showing the true positive rate relative the false positive rate of CPASS calculations for (A) aspartate transaminase (E.C. 2.6.1.1) and (B) glutamine-tRNA ligases (E.C. 6.1.1.18). True positives are defined based on three levels of functional homology between the query protein and the CPASS database: true positives have exact same E.C. classification (blue); true positives have first three E.C. numbers being identical (red); and true positives have same binding ligand (green).

The high CPASS performance is also illustrated by the distribution of the true positives and true negatives as a function of CPASS scores for both aspartate transaminase [Figure 5.4A] and glutamine-tRNA ligase [Figure 5.4B], which indicate that true negatives peak at a CPASS score of ~10%. True positives have a range of

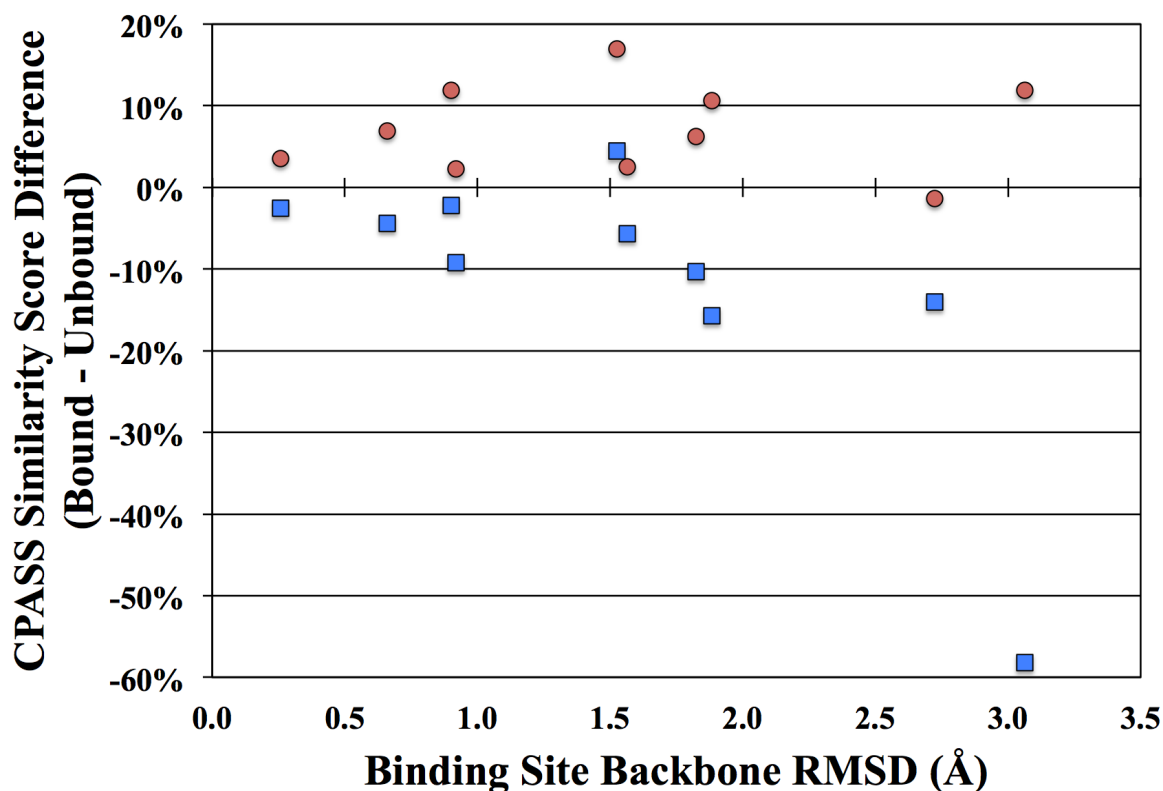
CPASS scores, but a threshold of ~20 – 30% is expected to identify the majority of functionally homologous proteins, while excluding essentially all of the true negatives.



**Figure 5.4** Distribution curves showing the fraction of negatives (red) and positives (blue) as a function of CPASS similarity scores (bin size of 10%) for (A) aspartate transaminase (E.C. 2.6.1.1) and (B) glutamine-tRNA ligases (E.C. 6.1.1.18).

**5.3.2 Tolerance of active-site variations.** Ten proteins with known bound and unbound structures were evaluated with CPASS. The RMSD between the bound and

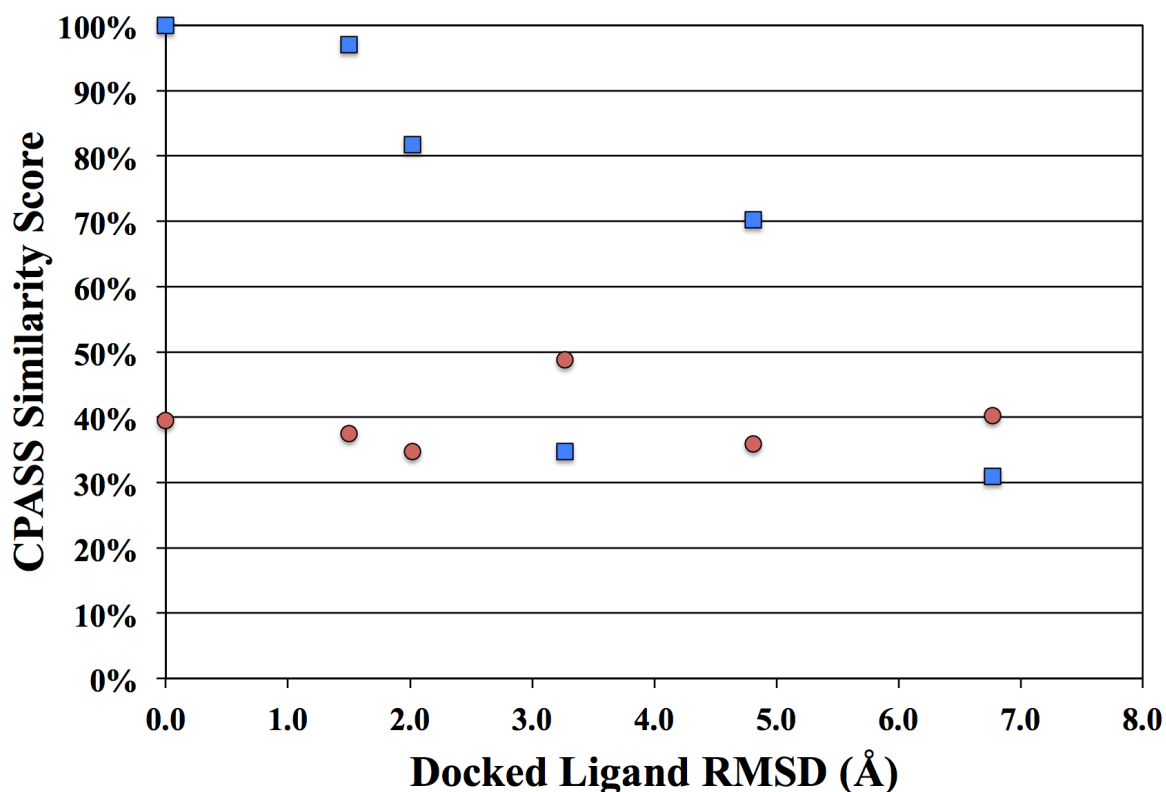
unbound backbone amino acid residues ranged from a small difference (0.259 Å) up to significant difference (3.063 Å). Superimposing the ligand coordinates of the bound protein structure into the unbound protein structure allowed for a comparison of ligand-defined binding sites using CPASS. The CPASS similarity scores for the first functional match (true positive) and the first functional non-match (false positive) were compared between the bound and unbound binding sites. The difference in CPASS similarity score for both was plotted against the active site backbone RMSD [Figure 5.5]. It is not surprising to see that as the structural differences between the binding sites of the bound and unbound proteins grows, the CPASS similarity score for the first functional match decreases. As the RMSD between the binding sites reaches  $\sim 2.0$  Å, the CPASS similarity score typically drops by at least 10%. For an RMSD greater than 2.0 Å, the CPASS score can change even more drastically, where a decrease in the CPASS similarity score of almost 60% occurred for an RMSD of  $\sim 3.0$  Å. While the first functional non-match typically had improved CPASS similarity scores for the unbound binding site, the CPASS scores did not increase with increasing RMSDs. In other words, functional non-matches (false positives) are unlikely to have a significantly high CPASS similarity score when using the unbound form of the protein. Instead, the CPASS similarity scores for the true positives decrease. Thus, the gap between the similarity scores of a true positive and a false positive will diminish with an increase in the structural difference between the bound and unbound forms. It is actually possible for the true positive to score much lower than the false positive.



**Figure 5.5** Comparison of the difference in CPASS similarity scores for the first functionally similar result (blue square) and the first functionally dissimilar result (red circle) using the original protein-ligand costructure and the unbound costructure with the ligand superimposed.

The location of the bound ligand can also have an effect on the CPASS similarity score.

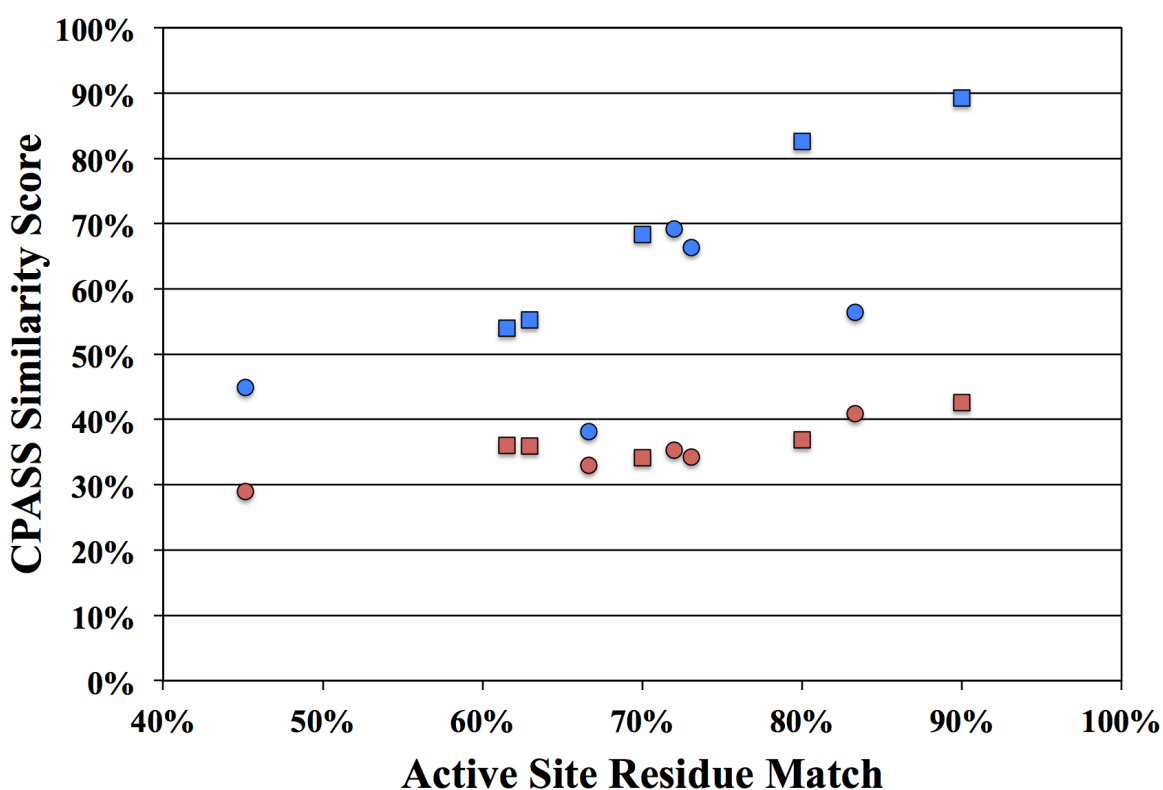
In the case of maltodextrin-binding protein, the maltose ligand was redocked into the bound protein structure. The comparison between the RMSDs of the correct ligand pose and 5 docked ligand poses to the CPASS similarity scores of the first functional match (true positive) and first functional non-match (false positive) illustrates the importance of the ligand conformation [Figure 5.6]. Identifying a correct functional match declines as the ligand conformation gets further from the correct ligand pose. Once again, the effect on the CPASS scores for functional non-matches is not significantly influenced.



**Figure 5.6** Comparison of CPASS similarity scores for the first functionally similar result (blue square) and the first functionally dissimilar result (red circle) against the RMSD of the docked maltose from the correct maltose pose in maltodextrin-binding protein (PDB: 1JW5).

Similar to ligand location, ligand size also influences the ligand-defined binding site, and, correspondingly, the CPASS similarity score. Docking a 3-sugar and 4-sugar amylose into the binding site of the maltodextrin-binding protein, allows for the evaluation of using larger ligands in a CPASS search. The ligand-defined binding site will likely include additional residues, which increases the number of residues to match. If these additional residues are not functionally relevant and are unmatched in a functional homolog, it will decrease the CPASS similarity score. The CPASS similarity scores for the first functional match (true positive) and functional non-match (false positive) was compared to the percentage of residues in the larger binding sites that

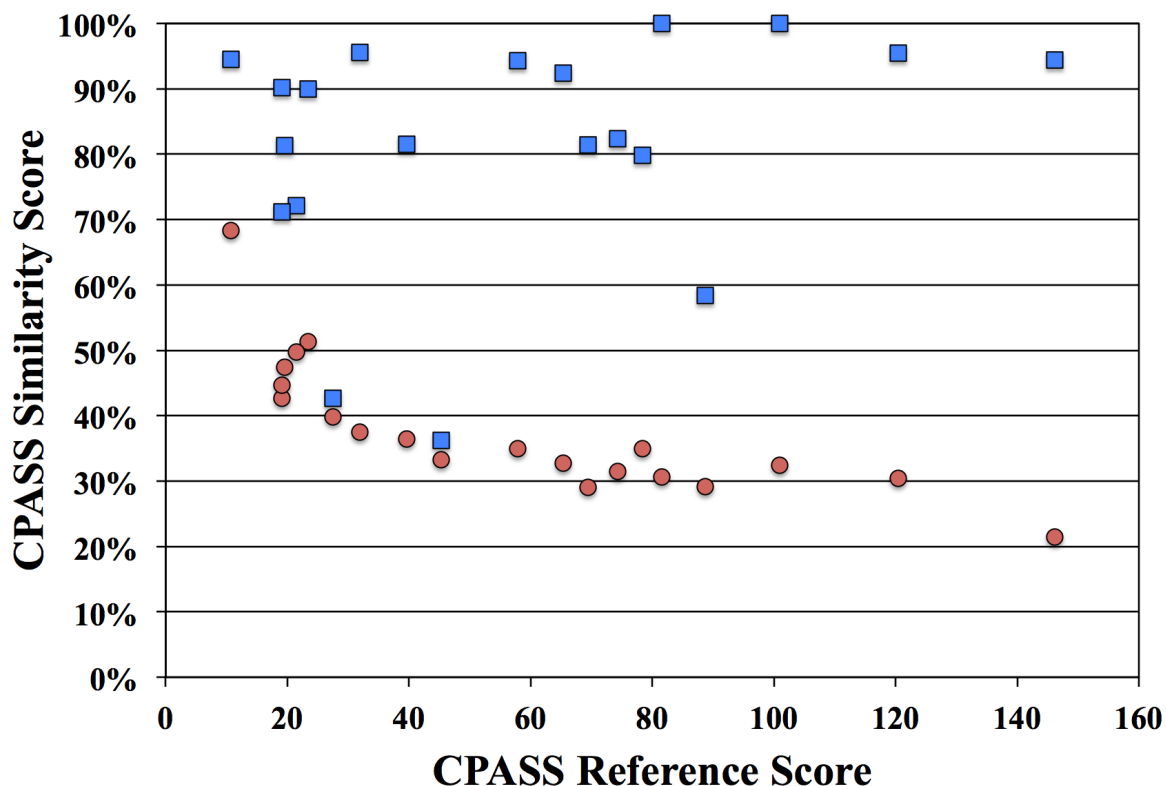
match the original protein-ligand costructure binding site [Figure 5.7]. As expected, the CPASS similarity score decreased with the increasing, irrelevant size of the ligand binding site. It should be noted that the CPASS similarity scores for the functional non-matches is also reduced. Unfortunately, the decrease in the CPASS similarity score for functional matches is more pronounced and leads to a smaller gap in the similarity scores between true positives and false positives.



**Figure 5.7** CPASS similarity scores for the first functionally similar result (blue) and the first functionally dissimilar result (red) as a function of the percent similarity in the ligand-binding site query. The active site residue match corresponds to the percentage of the residues in the 3-glucose amylose (square) and 4-glucose amylose (circle) defined ligand-binding site that are the same as the ligand-defined binding site from the original maltose-bound protein (PDB: 1JW5). Five unique poses of the docked 3-glucose amylose and 4-glucose amylose are shown.



All three evaluations illustrate that any change to the ligand-defined binding site has an effect on the CPASS similarity score. The degree of the effect is primarily dependent upon the size of the query's binding site. An increase in the number of residues in the binding site leads to a higher CPASS reference score, the sum of the CPASS similarity function scores for each residue in the query's binding site when compared to itself. Comparing the CPASS reference score for the binding sites of the 10 bound and unbound proteins with the resulting CPASS similarity score shows that smaller binding sites have a smaller separation between true positives and false positives [Figure 5.8]. The CPASS similarity score for functionally similar binding sites is not dependent on the size of the ligand binding site, as long as the ligand-binding site is accurately defined. Interestingly, the size of the binding site does exhibit a trend in CPASS similarity scores when evaluating functionally dissimilar binding sites. Smaller binding sites tend to have higher CPASS similarity scores for functionally dissimilar proteins. Simply, the smaller number of residues in the query binding site have a higher probability of finding a match to a functionally unrelated protein with a large binding site. A few residues may serendipitously share a similar spatial arrangement.

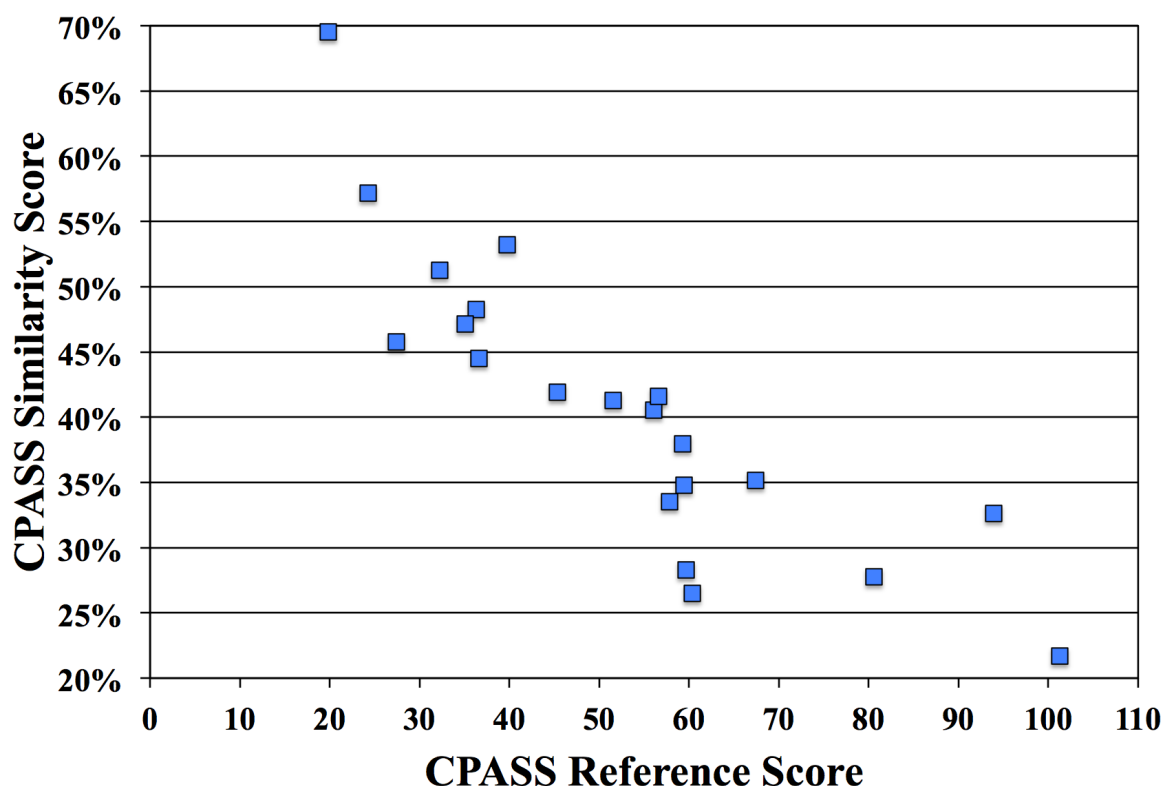


**Figure 5.8** Comparison between the ligand-binding site size dependent CPASS reference score and the CPASS similarity score. The first functionally similar (blue square) and first functionally dissimilar (red circle) results are shown.

**5.3.3 CPASS comparisons of proteins of unknown function.** The true biological function of the proteins investigated with FAST-NMR in Chapter 2 is unknown. Thus, a direct evaluation of the CPASS performance is not possible. However, an exploration of the general trends seen in the CPASS results for proteins of unknown function screened by FAST-NMR does provide valuable insights to further improve CPASS.

For the 21 proteins of unknown function successfully screened by FAST-NMR, the CPASS similarity score averaged  $41.0\% \pm 11.4\%$ . The highest score was 69.54% and the lowest score was 21.71%. A low CPASS score is probably indicative of a lack of a functional homolog in the CPASS database. The CPASS similarity scores follow the

previously described size dependent trend [Figure 5.9]. The highest similarity score has the smallest binding site (7 residues), while the lowest similarity score has the largest binding site (35 residues). This trend does not directly indicate that the top hit in a CPASS search is not a functional homolog; however, none of the proposed functions for the proteins described in chapter 2 matched the top CPASS hit.



**Figure 5.9** Comparison between the CPASS reference score and the highest CPASS similarity score for the 21 functionally unannotated proteins screened by FAST-NMR.

As mentioned previously, there are several factors from the FAST-NMR screening result that may affect the CPASS similarity score. There may be significant structural differences between the bound and unbound forms of the protein. The FAST-NMR methodology uses the protein's unbound structure in combination with molecular

docking to generate a protein-ligand costructure. There is no reliable method to correct for structural variations due to ligand binding except by experimentally determining the protein-ligand costructure using X-ray crystallography or NMR. Additionally, the location of the ligand is determined using a combination chemical shift perturbations and molecular docking. While this approach is certainly more accurate in determining the ligand pose than either method alone, the average RMSD from the correct ligand pose is still approximately 3.5 Å (see Chapter 4). As previously demonstrated, this conformational error may significantly lower the CPASS similarity score for a functional match [Figure 5.6]. Finally, the size of ligands shown to bind in a FAST-NMR screen often varies significantly, and are likely different from the natural ligand. Again, this has been shown to negatively impact the CPASS similarity score [Figure 5.7]. This last issue might be addressed in FAST-NMR by performing a detailed CPASS analysis using known binders of varying sizes.

## 5.4 CONCLUSIONS

CPASS 2.0 has been shown to be very effective in selecting for proteins with high functional similarity. In general, a CPASS similarity score of 30% or greater indicates a greater likelihood of functional homology. However, the variability in defining a binding site using predictive or screening approaches can significantly influence CPASS similarity calculations, where differences in bound and unbound structures, errors in docking a ligand to the structure, and variations in ligand size all appear to decrease the CPASS similarity score for a well-characterized protein. Additionally, an evaluation of the CPASS results from a FAST-NMR screen of 21 proteins of unknown function

highlights the impact of these issues. Specifically, these experimental problems hinder the ability of CPASS to reliably rank functional homologs as the top hit..

While experimental and computational variability will always be an issue, one approach to minimize their effect would be to prioritize the matching process. As mentioned previously, the reason protein structural differences and ligand size/location affect the CPASS similarity score is due to the way CPASS defines the ligand binding site. Every residue within 6 Å of the ligand is defined by CPASS as the binding site. While residues at the edge of the 6 Å cutoff are scaled to minimize the impact of small structural variations, the majority of the residues in the binding site are essentially equivalent in terms of importance to the CPASS scoring function. Unfortunately, not every residue located within this binding site is necessary for the molecular function of the protein, and thus would not necessarily be conserved. Implementing a weighting function for each residue in the query based on predicted or known importance may help prioritize binding sites in the database that have similar sequence and structure between these important residues.

Unfortunately, the greatest problem facing the functional annotation of unknown proteins with CPASS is the size of functional space represented by the database. The number of protein structures represented in the PDB is still significantly smaller than the number of known protein sequences. Therefore, the query protein may represent the first member of a functional class of proteins present in the PDB. Structural genomics is attempting to address this problem by prioritizing experimental structure determination efforts. This is leading to significant increase in the number of unique protein structures. But, unfortunately, the majority of structures deposited in the PDB lack a biologically

relevant ligand. Since the CPASS database is generated from proteins with bound ligands in the PDB, a functional class that does not have a representative structure with a bound ligand would not appear in the results from a CPASS query. This is especially a concern for proteins that binds another biomolecule (protein, DNA, RNA) instead of a small molecular weight compound.. These problems may be addressed by expanding the size of the CPASS database to potentially include predicted binding sites for unique proteins with no bound ligands. The inclusion of protein-protein or protein-DNA binding sites would also expand the searchable functional space. However, introducing either of these approaches to expand the database does introduce its own challenges.

Despite these issues, CPASS still provides valuable information based on the partial similarities that do exist, especially when combined with other bioinformatics approaches and experimental data. Additionally, CPASS can also be used to help understand the evolutionary relationships between proteins based on the changes that occur in the binding site.

## 5.5 REFERENCES

1. Bernal, A., Ear, U. & Kyrpides, N. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res* **29**, 126–127 (2001).
2. Pagani, I. *et al.* The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40**, D571–9 (2012).
3. Dimmer, E. C. *et al.* The UniProt-GO Annotation database in 2011. *Nucleic Acids Res* **40**, D565–70 (2012).
4. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **40**, D71–5 (2012).
5. Lee, D., Redfern, O. & Orengo, C. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* **8**, 995–1005 (2007).
6. Gerlt, J. A. & Babbitt, P. C. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem* **70**, 209–246 (2001).

7. Attwood, T. K. The quest to deduce protein function from sequence: the role of pattern databases. *Int J Biochem Cell Biol* **32**, 139–155 (2000).
8. Sigrist, C. J. A. *et al.* PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* **38**, D161–6 (2010).
9. Tuncbag, N., Gursoy, A. & Keskin, O. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics* **25**, 1513–1520 (2009).
10. Glaser, F., Morris, R. J., Najmanovich, R. J., Laskowski, R. A. & Thornton, J. M. A method for localizing ligand binding pockets in protein structures. *Proteins* **62**, 479–488 (2006).
11. Davis, I. W., Raha, K., Head, M. S. & Baker, D. Blind docking of pharmaceutically relevant compounds using RosettaLigand. *Protein Sci* **18**, 1998–2002 (2009).
12. Dundas, J. *et al.* CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res* **34**, W116–8 (2006).
13. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* **38**, W529–33 (2010).
14. Innis, C. A. siteFiNDER|3D: a web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Res* **35**, W489–94 (2007).
15. Yao, H., Mihalek, I. & Lichtarge, O. Rank information: a structure-independent measure of evolutionary trace quality that improves identification of protein functional sites. *Proteins* **65**, 111–123 (2006).
16. Mihalek, I., Res, I. & Lichtarge, O. Evolutionary and structural feedback on selection of sequences for comparative analysis of proteins. *Proteins* **63**, 87–99 (2006).
17. Kinoshita, K. & Nakamura, H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci* **12**, 1589–1595 (2003).
18. Stark, A., Sunyaev, S. & Russell, R. B. A model for statistical significance of local similarities in structure. *J Mol Biol* **326**, 1307–1316 (2003).
19. Laskowski, R. A., Watson, J. D. & Thornton, J. M. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* **33**, W89–93 (2005).
20. Watson, J. D., Laskowski, R. A. & Thornton, J. M. Predicting protein function from sequence and structural data. *Curr Opin Struct Biol* **15**, 275–284 (2005).
21. Powers, R. *et al.* Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **65**, 124–135 (2006).
22. Powers, R., Copeland, J. & Stark, J. L. Searching the protein structure database for ligand-binding site similarities using CPASS v. 2. *BMC Res Notes* (2011).
23. Mercier, K. A. *et al.* FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* **128**, 15292–15299 (2006).
24. Powers, R., Mercier, K. A. & Copeland, J. C. The application of FAST-NMR for the identification of novel drug discovery targets. *Drug Discov Today* **13**, 172–179 (2008).

25. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
26. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **10**, 980 (2003).
27. Friedberg, I. Automated protein function prediction--the genomic challenge. *Brief Bioinformatics* **7**, 225–242 (2006).
28. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29 (2000).
29. {International Union of Biochemistry and Molecular Biology Nomenclature Committee} & Webb, E. C. *Enzyme Nomenclature 1992*. (Academic Press, 1992).
30. Jeffery, C. J., Barry, T., Doonan, S., Petsko, G. A. & Ringe, D. Crystal structure of *Saccharomyces cerevisiae* cytosolic aspartate aminotransferase. *Protein Sci* **7**, 1380–1387 (1998).
31. Rould, M. A., Perona, J. J. & Steitz, T. A. Structural basis of anticodon loop recognition by glutamyl-tRNA synthetase. *Nature* **352**, 213–218 (1991).
32. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612 (2004).
33. Müller, C. W. & Schulz, G. E. Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap5A refined at 1.9 Å resolution. A model for a catalytic transition state. *J Mol Biol* **224**, 159–177 (1992).
34. Müller, C. W., Schlauderer, G. J., Reinstein, J. & Schulz, G. E. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* **4**, 147–156 (1996).
35. Lu, J. *et al.* Binding of retinol induces changes in rat cellular retinol-binding protein II conformation and backbone dynamics. *J Mol Biol* **300**, 619–632 (2000).
36. Lu, J. *et al.* The structure and dynamics of rat apo-cellular retinol-binding protein II in solution: comparison with the X-ray structure. *J Mol Biol* **286**, 1179–1195 (1999).
37. Fletcher, J. I., Swarbrick, J. D., Maksel, D., Gayler, K. R. & Gooley, P. R. The structure of Ap(4)A hydrolase complexed with ATP-MgF(x) reveals the basis of substrate binding. *Structure* **10**, 205–213 (2002).
38. Swarbrick, J. D. *et al.* The three-dimensional structure of the nudix enzyme diadenosine tetraphosphate hydrolase from *Lupinus angustifolius* L. *J Mol Biol* **302**, 1165–1177 (2000).
39. Wang, J. *et al.* Solution structures of staphylococcal nuclease from multidimensional, multinuclear NMR: nuclease-H124L and its ternary complex with Ca<sup>2+</sup> and thymidine-3',5'-bisphosphate. *J Biomol NMR* **10**, 143–164 (1997).
40. Duan, X. & Quijcho, F. A. Structural evidence for a dominant role of nonpolar interactions in the binding of a transport/chemosensory receptor to its highly polar ligands. *Biochemistry* **41**, 706–712 (2002).
41. Haupt, M. *et al.* The holo-form of the nucleotide binding domain of the KdpFABC complex from *Escherichia coli* reveals a new binding mode. *J Biol Chem* **281**, 9641–9649 (2006).
42. Breitenlechner, C. B. *et al.* Structure-based optimization of novel azepane derivatives as PKB inhibitors. *J Med Chem* **47**, 1375–1390 (2004).
43. Saad, J. S. *et al.* Structural basis for targeting HIV-1 Gag proteins to the plasma



- membrane for virus assembly. *Proc Natl Acad Sci USA* **103**, 11364–11369 (2006).
44. Morris, G., Goodsell, D., Halliday, R. & Huey, R. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**, 1639–1662 (1998).
  45. Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* **28**, 1145–1152 (2007).
  46. Morris, G. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* (2009). doi:10.1002/jcc.21256
  47. Sanner, M. F. Python: a programming language for software integration and development. *J Mol Graph Model* **17**, 57–61 (1999).
  48. Mercier, K. A. *et al.* Structure and function of *Pseudomonas aeruginosa* protein PA1324 (21-170). *Protein Sci* **18**, 606–618 (2009).
  49. Shortridge, M. D. & Powers, R. Structural and functional similarity between the bacterial type III secretion system needle protein PrgI and the eukaryotic apoptosis Bcl-2 proteins. *PLoS ONE* **4**, e7442 (2009).

## CHAPTER 6

### THE SOLUTION STRUCTURE AND FUNCTION OF YNDB, AN AHSA1 PROTEIN FROM *B. SUBTILIS* \*\*

#### 6.1 INTRODUCTION

The Bet v 1 protein from birch is a major allergen with high sequence similarity to the plant PR-10 pathogenesis-related proteins, which are involved in the response of plants toward microbial infection.<sup>1</sup> As the Bet v 1 proteins structure was solved,<sup>2</sup> numerous other proteins from among eukaryotes, archaea, and bacteria have been identified as having the same characteristic fold.<sup>3</sup> The Bet v 1-like superfamily of proteins now contains ~10,135 sequences and consists of 13 unique families. The four largest families in the Bet v 1-like superfamily are the polyketide cyclases (3,475 sequences), the ring hydroxylase  $\alpha$ -chain (2,022 sequences), the activator of Hsp90 ATPase homolog 1-like protein (AHSA1) family (1,762 sequences), and the StAR-related lipid transfer (START) family (1,026 sequences). The sequence similarity among the different Bet v 1-like families tends to be relatively low (0-38%), but all contain the same helix-grip fold that forms a hydrophobic cavity in between the long C-terminal  $\alpha$ -helix and the antiparallel  $\beta$ -sheet.<sup>3</sup> This hydrophobic cavity has been shown to preferentially bind to lipids, sterols, polyketide antibiotics, and other hydrophobic molecules.<sup>3</sup>

Although the Bet v 1-like superfamily members share a similar fold, the biological functions vary across the different families. The ring hydroxylases degrade

---

\*\* Chapter 5 was adapted from Stark, J. L., et al. Solution structure and function of YndB, an AHSA1 protein from *Bacillus subtilis*. *Proteins* **78**(16), 3328-3340 (2010). Reprinted with permission, copyright 2010 by John Wiley and Sons.

polycyclic aromatic hydrocarbons into non-aromatic *cis*-diols,<sup>4</sup> the START family appears to be involved in steroidogenesis,<sup>5,6</sup> whereas the polyketide cyclase family is involved with the biosynthesis of polyketide-based antibiotics and pigments.<sup>7</sup> Members of the AHSA1 family are named after the human activator of Hsp90 ATPase protein (Aha1). Although the proteins of this family have similar structures, the functions for most of the AHSA1 family members, except for its namesake, are ambiguous and are currently classified by UniProtKB<sup>8</sup> as either a general stress protein or a conserved putative protein of unknown function. The eukaryotic protein Aha1 is proposed to interact with the middle domain of heat shock protein 90, which stimulates its ATPase activity.<sup>9,10</sup> The domain organization of many homologous eukaryotic proteins in the AHSA1 family also suggests a function that is similar to Aha1. Conversely, homologous prokaryotic proteins have a much more diverse domain organization suggesting a wide range of possible functions.<sup>3</sup>

Of the 80 total structures solved for 59 members of the Bet v 1-like superfamily, 32 have ligands bound. The types of ligands that have been experimentally determined to bind Bet v 1-like proteins include membrane lipids, plant hormones, secondary metabolites, polycyclic aromatic hydrocarbons, and DNA/RNA.<sup>3</sup> There are 12 total proteins in the AHSA1 family with known structures. The only protein in the AHSA1 family with a solved structure of its protein-ligand complex is the self-sacrificing resistance protein CalC from *Micromonospora echinosporato*,<sup>11</sup> where CalC is shown bound to calicheamicin  $\gamma$ 1,<sup>12</sup> a potent antitumor antibiotic compound. Both Pfam<sup>13</sup> and SCOP<sup>14</sup> databases classify CalC as belonging to the AHSA1 family due to its 43-55% sequence similarity to other uncharacterized bacterial members of AHSA1. However,

CalC contains a break in the C-terminal helix that is uncharacteristic of most Bet v 1-like proteins and would likely indicate a new CalC-like family within the Bet v 1-like superfamily. This leaves only the human Aha1 with a proposed function within the AHSA1 family.

The *Bacillus subtilis* YndB protein is a protein of unknown biological function targeted for structural analysis by the Northeast Structural Genomics Consortium (NESG; <http://www.nesg.org>; NESG target: SR211). We previously reported the near complete nuclear magnetic resonance (NMR) assignments for *B. subtilis* YndB,<sup>15</sup> where the protein was originally identified as being a member of the START domain<sup>15,16</sup> due to the similar helix-grip fold found in the structure of two homologous proteins and based on CATH comparisons.<sup>17</sup> The NMR structures reported for *Bacillus cereus* protein BC4709 (PDB ID: 1xn6) and *Bacillus halodurans* protein BH1534 (PDB ID: 1xn5) led to their START domain classification.<sup>16</sup> These two proteins are 64% and 57% homologous to YndB, respectively, inferring a similar annotation for YndB. However, the SCOP and Pfam databases have suggested that YndB, BC4709, and BH1534 belong to the AHSA1 family. Sequence similarity searches with YndB only identify proteins annotated as either AHSA1 or proteins of an unknown function. The primary difference between START domain and AHSA1 structures is that START domain proteins typically contain two additional N-terminal  $\beta$ -strands and an  $\alpha$ -helix, which also makes the proteins larger. The structure of BC4709 and BH1534 do not have these additional structural components further supporting their AHSA1 classification.

Assigning a function to an uncharacterized protein like YndB can be a daunting task that involves obtaining a high-resolution structure<sup>18</sup> combined with detailed studies

that may include generating knockout libraries to analyze cell phenotypes, monitoring gene expression levels, or performing pull-down assays, all of which require in-depth bioinformatics analyses.<sup>19-23</sup> As the biological function of a protein is, by definition, derived from its interactions with other biomolecules or small molecules, identifying interacting partners is an alternative route to obtaining a functional annotation. One such technique, FAST-NMR,<sup>24,25</sup> utilizes a small biologically focused compound library combined with NMR high-throughput screening (HTS), rapid protein-ligand costructures using AutoDock<sup>26</sup> and chemical shift perturbations (CSPs),<sup>27</sup> and a comparison of protein active site structures<sup>28,29</sup> to assist the functional annotation of proteins. However, the utility of FAST-NMR relies on structural analogs being found within the diverse functional chemical library. In the case of YndB, the known Bet v 1-like superfamily ligands combined with the expected hydrophobic cavity for YndB already suggests the protein is likely to bind lipid-like molecules. This eliminates the need for screening a diverse array of compounds found in the FAST-NMR compound library and instead requires an extensive screen against a focused lipid-like library. Because of the large number of biologically relevant lipid-like compounds<sup>30</sup> and the corresponding limited commercial availability, an HTS assay is not practical or cost effective. Instead, an *in silico* screen<sup>31,32</sup> provides an attractive alternative method to identify specific classes of compounds that may interact with YndB and to focus follow-up *in vitro* efforts.

To better understand the general biological role of AHSA1 proteins, the structure and putative biological functions of the *B. subtilis* YndB protein was determined using NMR spectroscopy and the *in silico* ligand-binding screen. The three-dimensional solution structure of YndB (PDB ID: 2kte) is described herein and is consistent with

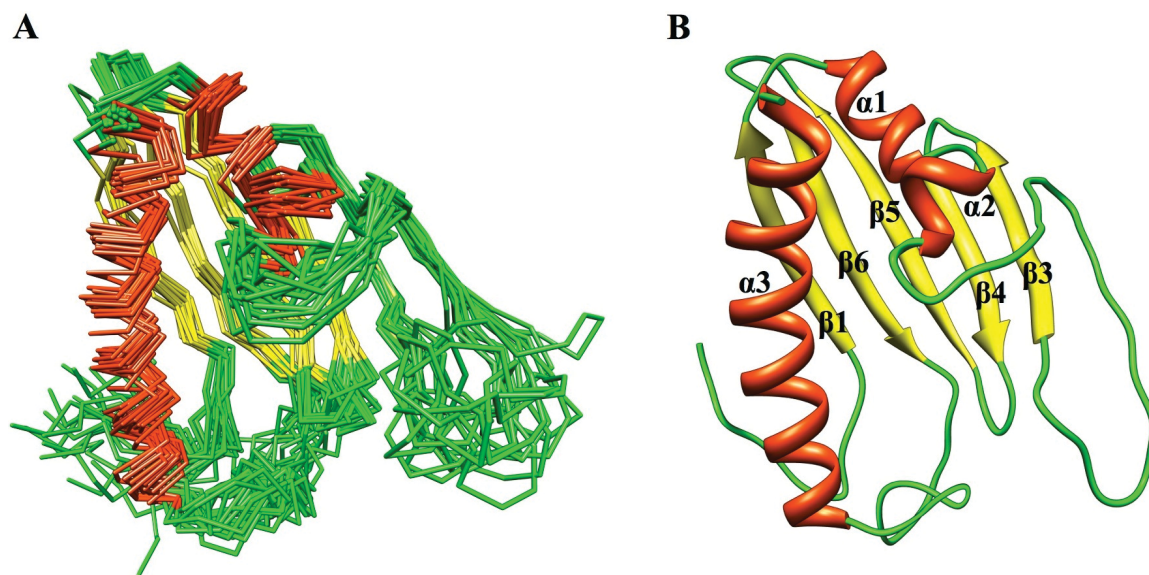
other AHSA1 proteins.<sup>††</sup> As most Bet v 1-like and AHSA1 proteins contain a hydrophobic ligand-binding pocket, the *in silico* screen of a ~18,500 lipid compound library<sup>30</sup> was performed to identify a particular class of lipids that preferentially bind YndB and to provide insight into its biological function. The *B. subtilis* YndB protein was shown to experimentally bind *trans*-chalcone, a member of an important class of antibiotics and an important plant metabolite produced by chalcone synthase (CHS).<sup>33</sup> Three other compounds similar in structure to chalcones (flavanone, flavone, and flavonol) and part of the same metabolic pathway were also shown to bind YndB, albeit weaker binders than *trans*-chalcone. These chalcone-like molecules are often found as precursors to flavonoids that play a key role in plant-microbe signaling and defense, where *Bacillus* strains have been shown to have a beneficial impact on plant health by protecting against fungal and bacterial pathogens.<sup>34</sup> This suggests *B. subtilis* YndB may respond to a plant infection signal and induce a stress response.

## 6.2 MATERIAL AND METHODS

**6.2.1 Solution structure of *B. subtilis* YndB.** The three dimensional structure of Yndb is report as an 18 structure ensemble in the PDB (PDB ID: 2kte). [Figure 6.1A] The structure of the YndB protein was visualized and evaluated using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (<http://www.cgl.ucsf.edu/chimera>).<sup>35</sup> Putative binding sites of YndB and homologous proteins BC4709 and BH1534 were investigated and compared using CASTp,<sup>36</sup> which attempts to identify protein ligand binding sites and active sites by defining the molecular surface and determining surface accessible pockets.

---

<sup>††</sup> Dr. Kelly Mercier was responsible for determining the solution structure of YndB.



**Figure 6.1** The NMR solution structure of *B. subtilis* protein YndB: (A) a backbone trace of the 18 lowest energy conformation models and (B) a ribbon diagram where the  $\alpha$ -helices are colored red, the  $\beta$ -strands are colored yellow, and the loops are colored green.

**6.2.2 Sequence and structure similarity to YndB.** To identify homologous proteins and elucidate a possible function, multiple similarity comparisons were performed. The pair-wise sequence alignment of YndB to protein sequences in a nonredundant database was performed using BLASTP<sup>37-39</sup> and the default BLOSUM62 scoring matrix. DaliLite v.3<sup>40</sup> was used to perform the structural similarity comparisons of YndB (model #10) with proteins from the RCSB PDB. ClustalW<sup>41</sup> was used to align the sequences of YndB and the two homologous proteins, BC4709 and BH1534, for a detailed analysis of conserved amino acid residues that make up functionally relevant components of each protein. The ClustalW sequence alignments used the default settings.

**6.2.3 Virtual screening of a lipid compound library.** The *in silico* screen of YndB against a lipid library was performed to identify classes of lipid molecules that are favored to bind the protein. The lipid library used in this study was obtained from the

Nature Lipidomics Gateway ([www.lipidmaps.org](http://www.lipidmaps.org)), which contains two-dimensional structures of 21,824 lipid molecules (as of January 2010) found in mammalian species.<sup>30,42,43</sup> Clearly, the lipid library is not exhaustive and many lipid molecules found in nonmammalian organisms are not represented, but the goal of the virtual screening effort is to identify a structural homolog to the natural ligand or to identify a particular class of lipid that preferentially binds YndB. Eight major categories of lipids are represented in the Nature Lipidomics Gateway library: fatty acyls (3,476 structures), glycerolipids (3,012 structures), glycerophospholipids (1,958 structures), sphingolipids (3,376 structures), sterol lipids (2,125 structures), prenol lipids (1,156 structures), saccharolipids (13 structures), and polyketides (6,708 structures). The eight major categories are further divided into a total of 538 distinct subclasses of lipid compounds. The two-dimensional structure files provided by the Nature Lipidomics Gateway were converted into three-dimensional conformers using the program OMEGA 2.3.2<sup>44</sup> (OpenEye Scientific Software, Sante Fe, NM). OMEGA generates a database of multiple three-dimensional conformers for each ligand in the compound library using fragment assembly, ring conformation enumeration, and torsion driving. In this study, OMEGA was used to generate a maximum of 600 unique ( $> 0.5$  Å rmsd) conformers for each lipid molecule for a total searchable database consisting of  $\sim 10,000,000$  conformers. OMEGA failed to generate conformers for 3,306 of the lipid structure files, leaving a chemical library of 18,518 compounds for the *in silico* screen. Most of these failures occurred during the processing of the sphingolipid category of lipids (3,196 out of 3,376 failed) due largely to the large size and number of branches/rotatable bonds of the molecules in this category.



The docking program FRED 2.2.5<sup>45,46</sup> (OpenEye Scientific Software, Sante Fe, NM) was used for the virtual screen of YndB against the lipid library. FRED is a rigid docking program, which uses the multiple conformers of each ligand created in OMEGA and generates 100 docked poses within the defined binding site by rotating and translating the rigid molecule to optimize shape complementarity. The poses of each ligand conformer are then ranked using the built-in consensus scoring method, where only the top scoring pose is kept. As the conformers are rigid during this docking process, FRED has been shown to be very fast as compared with other docking programs that allow for ligand flexibility.<sup>47</sup> This speed is necessary in order to screen the large lipid-like library in a reasonable amount of time. Although some accuracy may be lost due to rigid docking and a lack of a biologically relevant conformation for the ligand, FRED was primarily used to rapidly filter out compounds that could not fit into the YndB ligand-binding pocket. Before initiation of the docking, model 10 from the YndB PDB file (PDB ID: 2kte) was prepared using FRED Receptor 2.2.5 (OpenEye Scientific Software, Sante Fe, NM), where a high-quality shape potential grid of 3403 Å<sup>3</sup> was generated that encompassed the proposed binding cavity. Model 10 was selected as the target receptor for the virtual screen as it had the lowest violation energies during the solution structure calculations. The lipid library compounds were ranked using the default Chemgauss3 scoring function that includes descriptors of shape and molecular chemical properties. Chemgauss3 incorporates steric and hydrogen bond interactions as well as protein and ligand desolvation parameters that are smoothed using a Gaussian function. The relative enrichment for each lipid class within the top 1000, the top 500, the top 200, the top 100, and the top 50 ranked compounds were calculated according to the following equation:

$$\%RE = \frac{\%Ab_{Lib} - \%Ab_{FRED}}{\%Ab_{Lib}} \times 100\% \quad (6.1)$$

where %RE is percent relative enrichment, %Ab<sub>Lib</sub> is the percent abundance of a lipid class in the Nature Lipidomics Gateway library, and %Ab<sub>FRED</sub> is the percent abundance of a lipid class observed in either the top 1000, 500, 200, 100, or 50 ranked compounds by FRED.

**6.2.4 NMR titration experiment.** Based on the results of the virtual screen, three classes of lipid molecules were identified as possible binders: flavones/flavonols, flavanones, and chalcones/hydroxychalcones. Experimental validation of these possible binders was performed using chemical shift perturbations (CSPs) in 2D <sup>1</sup>H-<sup>15</sup>N HSQC NMR spectra collected on a Bruker 500 MHz Avance spectrometer equipped with a triple-resonance, Z-axis gradient cryoprobe. The 2D <sup>1</sup>H-<sup>15</sup>N HSQC NMR experiment was collected at 298 K with 32 scans, 1024 data points and a spectral width of 36 ppm in the indirect <sup>15</sup>N-dimension. The ligands selected to represent each of the three potential binding lipid classes were *trans*-chalcone, flavanone, flavone, and flavonol (Sigma-Aldrich, St. Louis, MO). Flavone and flavonol belong to the same class of lipids, but there was interest in how the binding would be affected with the addition of a polar functional group. The fatty acyl oleic acid (Sigma-Aldrich, St. Louis, MO) was also selected as a negative control. These compounds were selected based on availability, a simple scaffold that clearly represented the lipid class, and cost. Each compound was dissolved in “100%” deuterated DMSO-*d*<sub>6</sub> (Sigma-Aldrich, St. Louis, MO) before titration. The titration analysis was performed with an 80 μM <sup>15</sup>N-labeled YndB sample (20 mM MES buffer, pH 6.5 with 10% D<sub>2</sub>O, 0.02% NaN<sub>3</sub>, 10 mM DTT, 5 mM CaCl<sub>2</sub>,

100 mM NaCl, and 50  $\mu$ M 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS)) and increasing concentrations (ranging from 0 to 600  $\mu$ M) of each ligand. The NMR data were processed with NMRpipe<sup>48</sup> and the spectra viewed using NMRViewJ.<sup>49</sup> Kaleidagraph 3.5 (Synergy Software) was used to fit the NMR data to the following equation.<sup>50,51</sup>

$$\text{CSP}_{\text{obs}} = \text{CSP}_{\text{max}} \frac{(K_D + [L] + [P]) - \sqrt{(K_D + [L] + [P])^2 - 4([L][P])}}{2[P]} \quad (6.2)$$

where  $\text{CSP}_{\text{obs}}$  is the 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC CSPs,  $[P]$  is the protein concentration,  $[L]$  is the ligand concentration, and  $K_D$  is the dissociation constant.

**6.2.5 *B. subtilis* YndB-ligand costructures.** Co-structures of YndB bound to each compound used in the NMR titration experiment (*trans*-chalcone, flavanone, flavone, and flavonol) were generated to analyze the ligand-binding pocket. Although a definitive identification of the binding pose of these ligands to YndB would require extensive NMR experiments and data analysis similar to the original effort to solve the apo-YndB structure, molecular docking can provide a rapid and reliable NMR-based model to examine the details of the binding interactions.

AutoDock 4.01<sup>26,52</sup> with the AutoDockTools 1.5.2 (<http://mgltools.scripps.edu>) graphical interface was used to simulate 100 different binding poses for each YndB-ligand complex. AutoDock was used instead of FRED due to the accuracy gained from flexible ligand docking and because it is one of the most highly cited docking programs available.<sup>53</sup> The grid map was generated with 0.375 Å spacing with *xyz* grid point dimensions of 50 x 58 x 48, which is of sufficient size to encompass the proposed binding pocket previously identified in CASTp and FRED. The docking calculations

were performed using the Lamarckian genetic algorithm default settings with a population size of 300 and 5,000,000 energy evaluations.

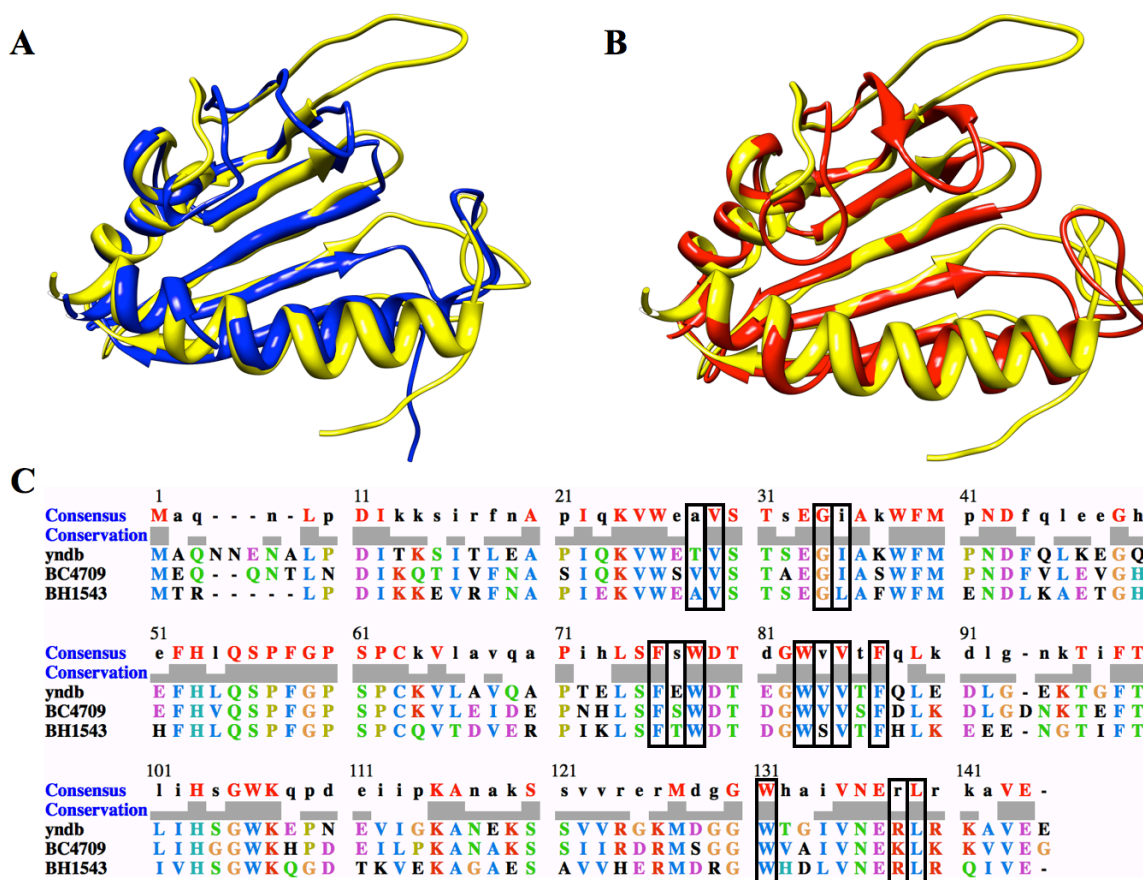
## 6.3 RESULTS AND DISCUSSION

**6.3.1 Solution structure of *B. subtilis* YndB.** The observed secondary structure and fold for *B. subtilis* YndB are characteristic of the helix-grip fold found in the Bet v 1-like superfamily. The helix-grip fold consists of a  $\beta$ -sheet with two small and one long  $\alpha$ -helix. The  $\beta$ -sheet is comprised of five strands instead of the normal six. The missing short strand, which is normally  $\beta 2$ , forms sheet like interactions in only two of the 18 structures in the ensemble and appears to protect the edge of  $\beta 3$ ; unprotected edges can be adventitious interaction sites for aggregation.<sup>54</sup> However, the strands are annotated 1, 3, 4, 5, and 6 to facilitate comparisons with other family members: residues 12-18 ( $\beta 1$ ), 63-69 ( $\beta 3$ ), 73-78 ( $\beta 4$ ), 83-91 ( $\beta 5$ ), and 96-104 ( $\beta 6$ ). The three  $\alpha$ -helices are comprised of residues 22-28 ( $\alpha 1$ ), 33-36 ( $\alpha 2$ ), and 120-143 ( $\alpha 3$ ) [Figure 6.1B]. There is significant variability in the loop regions of the protein corresponding to residues 37-63 and 105-120. These loops appear to be important for the structure of the hydrophobic ligand-binding cavity [Figure 6.1B].

Like other proteins with a helix-grip fold, YndB has an exposed hydrophobic core, likely used in the binding of lipid-like molecules. Analysis with CASTp<sup>Å</sup> shows that the volume of this putative binding cavity is 790 Å<sup>3</sup>. The core of YndB consists primarily of aromatic side chains. One element of the YndB binding pocket is the long  $\alpha 3$ -helix. This helix is anchored to the  $\beta$ -sheet by residues W130, V134, and L138, which show NOE interactions to the sheet residues S15, T17, and L18. Helix  $\alpha 3$  has previously

been identified as being crucial to the function of the structurally related START domains.<sup>55,56</sup> In both of these earlier studies, the removal of part of the  $\alpha 3$ -helix eliminated ligand binding. Based on the NMR solution structure for YndB, removing the C-terminal residues would result in  $\alpha 3$  no longer being associated with the  $\beta$ -sheet, and therefore, the protein would probably not be folded properly. Hence, we surmise that the previous results were likely due to protein instability and not the activity of specific residues to ligand binding.

The Bet v 1-like superfamily classification for YndB and the reliability of the NMR structure is further supported by the structural similarities to two homologous proteins, BC4709 and BH1534 [Figure 6.2A,B]. The YndB protein exhibits a backbone rmsd of 1.1 Å and 1.2 Å to BC4709 and BH1534, respectively, when only secondary structural elements are included in the alignment. The main difference among the structures lies in the loop regions, where there appears to be a significant difference in the loop conformation of residues 37-63 and 105-120 for YndB. This difference affects the size of the hydrophobic cavity for YndB, where BC4709 and BH1534 have much smaller volumes (199 Å<sup>3</sup> and 106 Å<sup>3</sup>, respectively) relative to YndB.



**Figure 6.2** An overlay of the NMR solution structure of *B. subtilis* protein YndB (yellow), with (A) *Bacillus cereus* protein BC4709 (blue) (PDB ID: 1xn6) and (B) *Bacillus halodurans* protein BH1543 (red) (PDB ID: 1xn5). (C) The multiple sequence alignment from ClustalW of YndB, BC4709 and BH1543 with the 14 active site residues (< 5 Å from bound ligand) indicated in black rectangles.

As expected from the high-sequence identity, the sequence compositions of the ligand binding sites are also similar [Figure 6.2C]. Nine of the 14 residues that line the cavity are identical and predominantly hydrophobic (V29, G34, F76, W78, W83, V85, F87, W130, and L138) and two others show high similarity. Although, none of these residues exist within the loop regions, the large loop from residues 37-63 is very well conserved with 16 residues being identical among the three proteins, once again indicating the importance of these loops for ligand binding. The loop regions corresponding to residues 37-63 and 15-120 are predicted to be conformationally flexible

based on the lack of NMR assignments.<sup>15</sup> Of the 43 amino acids that comprise the two loop regions, a total of 21 residues are unassigned. Correspondingly, these loop regions have a limited number of structural constraints resulting in the observed conformational variability in the ensemble of calculated structures [Figure 6.1A]. Although the lack of NMR assignments and NOEs suggests conformational flexibility, these observations are not sufficient to define the loops as dynamic and requires further experimental evidence for verification.<sup>57</sup> Nevertheless, it is anticipated that a bound substrate would restrict the loop conformation near the YndB ligand binding site.

**6.3.2 Sequence and structure similarity to YndB.** The BLASTP search of YndB against a non-redundant protein sequence database identified 58 proteins with an *E*-value of  $1.0 \times 10^{-21}$  or lower that included BC4709 and BH1534. All of these proteins belong to the Gram-positive organisms of the order Bacillales and have sequence identities  $> 39\%$  and sequence similarities  $> 58\%$ . There is a clear division in sequence similarity between the 58 Bacillales proteins and other Gram-positive bacteria proteins homologous to YndB. The sequence similarity score drops significantly from *E*-values of  $10^{-21}$  for Bacillales proteins to *E*-values of  $\geq 10^{-9}$  for other Gram-positive proteins. This sequence distinction may indicate a function specific to the Bacillales organism.

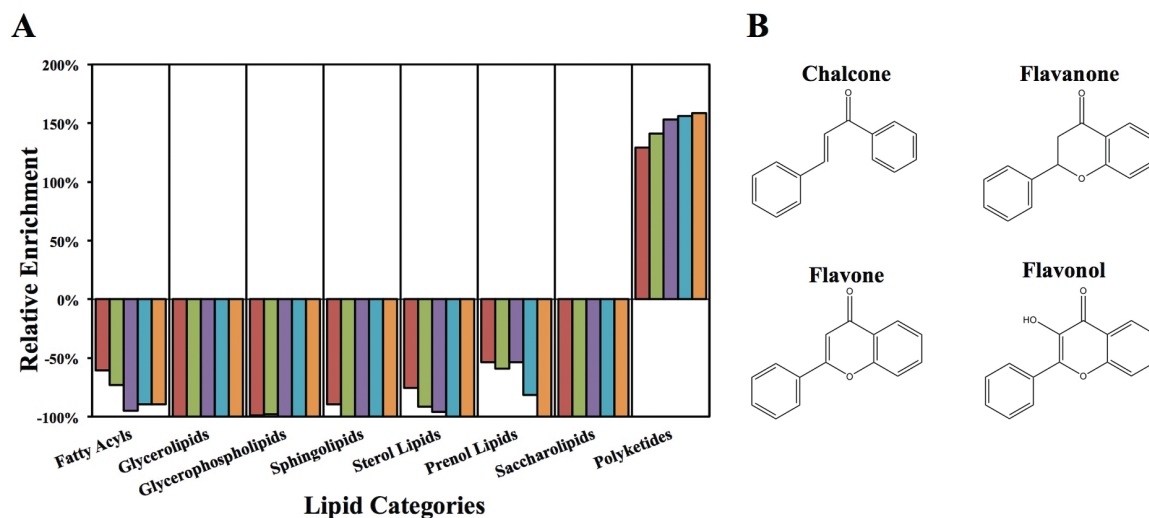
The structural similarity search using DaliLite identified 590 proteins with a *Z*-score over 2.0. Once again, the two proteins with the greatest structural similarity are BC4709 and BH1534 with *Z*-scores of 14.0 and 14.2, respectively. The top 100 proteins with the highest structural similarities have *Z*-scores ranging from 9.0 to 14.2 with sequence identities  $< 25\%$ , except for BC4709 and BH1534. All of the proteins identified

in this range are either uncharacterized or members of the Bet v 1-like superfamily, which includes Bet v 1-like proteins in plants.

**6.3.3 Virtual screening of a lipid compound library.** The *in silico* screen of YndB with the entire Nature Lipidomics Gateway lipid library took ~44 hours with the computation dispersed across 16 nodes of a Linux Beowulf cluster. Of the 18,518 structures in the library, FRED successfully docked 17,475 compounds to YndB. The relative enrichment [Equation 6.2] of each lipid class from the FRED docking is plotted in [Figure 6.3A]. Only one lipid category, the polyketides, had a positive relative enrichment among the top 1000 docked lipid molecules. Polyketides represent 86.8% of the molecules in the top 1000, whereas they only make up 37.9% of the entire compound library for a relative enrichment of 129%. The polyketide representation increases significantly as the cutoff for the FRED scoring energy for molecules accepted in the top rankings is decreased. If only the top 50 docked compounds are considered, 98.0% of these compounds are polyketides, with only one hit being a member of the fatty acyl category. All of the polyketides identified as hits belong to the flavonoid class of lipids. Within the flavonoids, three subclasses emerge as favorable hits from the virtual screen: chalcones/hydroxychalcones, flavanones, and flavones/flavonols.

The chalcone/hydroxychalcone subclass turns out to be the most significant hit as 44.9% of the flavonoids in the top 50 hits were chalcones, whereas they only make up 9.4% of the library of flavonoids. The remaining flavonoids in the top 50 hits belong to the flavanone (28.6%) and flavone/flavonol (14.3%) subclasses. The molecules from these three subclasses all have very similar chemical structures, which consist of at least two benzene rings and contain only a few rotatable bonds [Figure 6.3B].



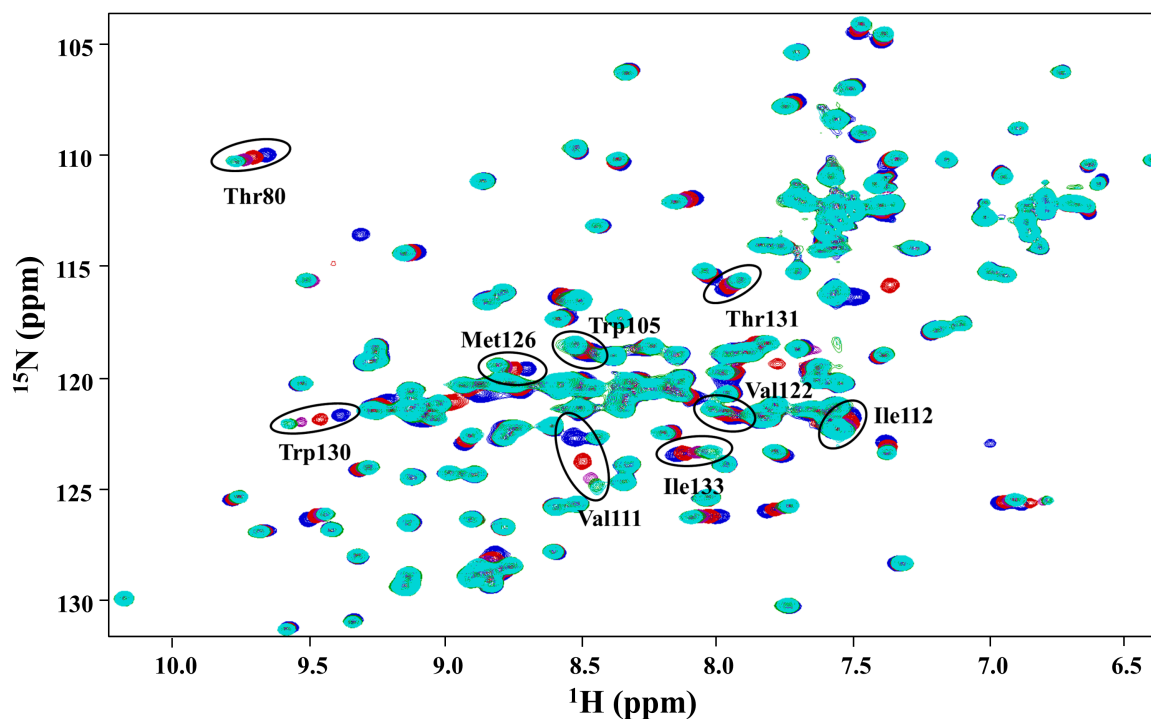


**Figure 6.3** Summary of the FRED *in silico* screening results of the Nature Lipidomics Gateway lipid library and *B. subtilis* protein YndB. Approximately 18,500 lipid structures corresponding to ~10,000,000 conformers were docked into YndB. The compounds were ranked using the FRED Chemgauss3 scoring function. **(A)** A plot of the relative enrichment [Equation 6.1] for each of the eight major lipid categories within the top 1,000 hits (red), the top 500 hits (green), the top 200 hits (purple), the top 100 hits (cyan), and the top 50 hits (orange). Only the polyketides were positively enriched in the virtual screen relative to their representation in the original lipid library. **(B)** The chemical structures of the four flavonoid compounds chosen to represent the three most enriched subclasses of lipids (chalcones/hydroxychalcones, flavanones, and flavones/flavonols) identified from the *in silico* screen.

**6.3.4 NMR titration experiment.** Although virtual screening appears to be a useful tool for identifying particular classes of lipids that have structural and chemical properties amenable to binding to YndB, these results require validation by experimental methods. NMR is routinely used to evaluate protein-ligand interactions, to measure dissociation constants ( $K_D$ ) and to identify ligand-binding sites through the observation of CSPs.<sup>27,58,59</sup> CSPs were calculated by comparing the average  $^1\text{H}$  and  $^{15}\text{N}$  resonance changes between ligand-free and ligand-bound YndB 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR spectra. The advantage of this approach is the speed and minimal amount of protein and ligand required. Unfortunately, access to the specific lipid compounds predicted to bind YndB is

very limited due to low commercial availability and/or high cost. Based upon the *in silico* screening of YndB with a lipid library, chalcones/hydroxychalcones, flavanones, and flavones/flavonols were identified to be the most likely to bind YndB. Therefore, representative molecules were sought for each class containing the basic structural scaffold that would likely have characteristic binding properties. A member of the fatty acyl category of lipids was also sought for use as a negative control.

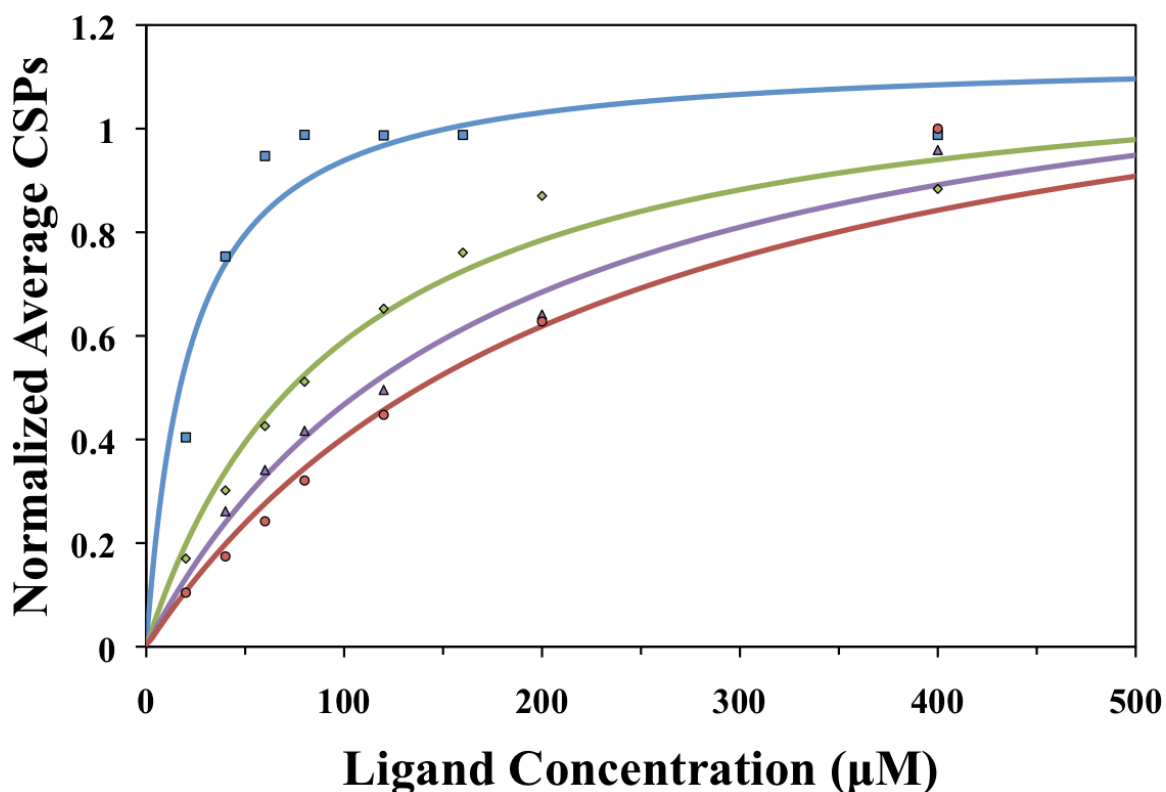
For the chalcone/hydroxychalcone subclass of lipids, *trans*-chalcone was selected to represent the basic structural scaffold for this class. The titration of YndB with *trans*-chalcone resulted in significant CSPs [Figure 6.4]. Nine YndB residues with the most significant CSPs (greater than two standard deviations from the mean) were identified: Thr80, Trp105, Val111, Ile112, Val122, Met126, Trp130, Thr131, and Ile133. These residues line the opening of the proposed binding pocket [Figure 6.6] identified by CASTp. Six more residues with significant perturbations (greater than one standard deviation from the mean) were also identified: Glu110, Val121, Arg123, Asp127, Gly128, and Asn135. These amino acids reside in the long  $\alpha$ 3-helix and contribute to a portion of the ligand binding pocket. Their perturbation may indicate a structural change in  $\alpha$ 3-helix upon binding. The remaining loop residues that define the binding pocket are unassigned in apo-YndB.



**Figure 6.4** Overlay of the 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectra of *B. subtilis* protein YndB titrated with chalcone, where the chalcone concentration was increased from 0  $\mu\text{M}$  (blue) to 160  $\mu\text{M}$  (cyan). The significant CSPs of the nine assigned amino acid residues used to determine the dissociation constant ( $K_D$ ) are highlighted with a black oval and labeled accordingly. Not all of the perturbed peaks were assigned; these residues are likely from the loop regions.

The normalized CSPs for each of the nine amino acid residues were plotted as a function of protein-ligand concentration ratios and fit to a binding isotherm [Equation 6.2] to determine a dissociation constant. *trans*-Chalcone binds tightly to YndB with a  $K_D$  of  $\leq 1$   $\mu\text{M}$  and a stoichiometry of 1:1. The binding stoichiometry is based on the observation that a two-site model does not fit the data as evidenced by the fact that the CSPs reaches a maximum at  $\sim 1:1$  protein-chalcone concentration ratio [Figure 6.5]. Calculating an exact  $K_D$  for *trans*-chalcone was not possible given the YndB concentration (80  $\mu\text{M}$ ) used for the 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC titration experiments, and significantly lowering the YndB concentrations was not feasible. Superimposed on the

*trans*-chalcone NMR titration data [Figure 6.5] is a theoretical curve for a  $K_D$  of 1  $\mu\text{M}$ , implying an upper-limit for the *trans*-chalcone dissociation constant.

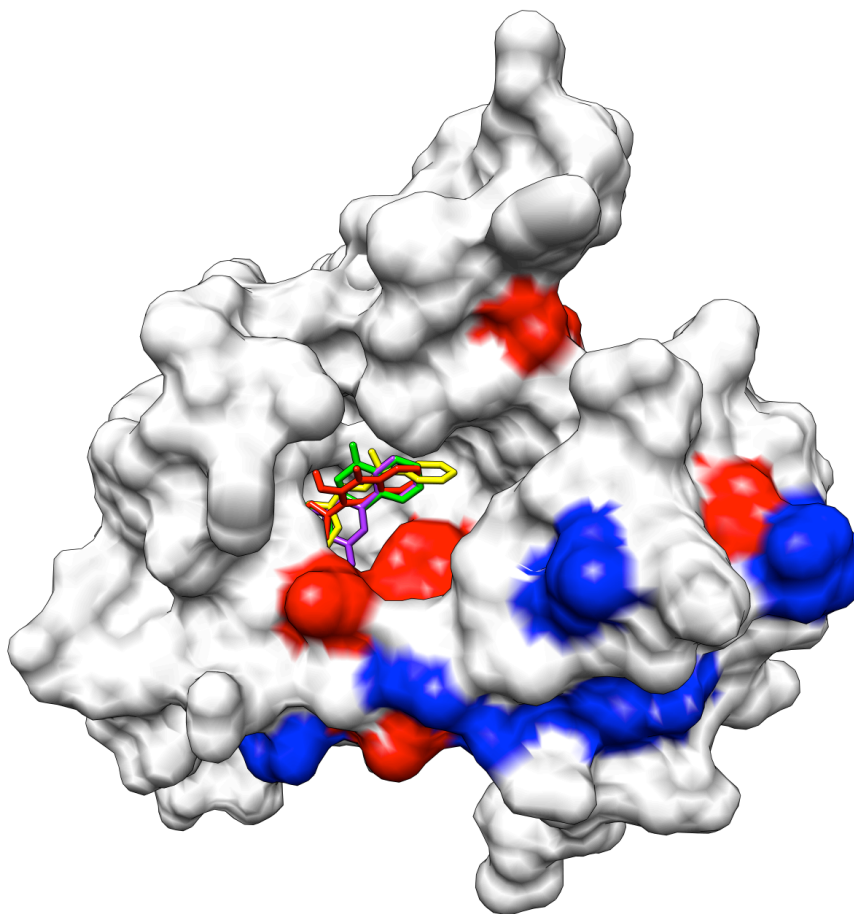


**Figure 6.5** NMR titration data for *trans*-chalcone (blue), flavanone (green), flavone (purple), and flavonol (orange). The normalized CSPs for the nine most perturbed residues are plotted versus the protein-ligand concentration ratios. The titration curves were fit to a binding isotherm [Equation 6.2] using Kaleidagraph 3.5 (Synergy Software). The best-fit curves are shown as a solid line. The theoretical curve displayed for *trans*-chalcone corresponds to a  $K_D$  of 1  $\mu\text{M}$  and represents the upper-limit for the  $K_D$ . The measured  $K_D$  values are  $\leq 1 \mu\text{M}$  (*trans*-chalcone),  $32 \pm 3 \mu\text{M}$  (flavanone),  $62 \pm 9 \mu\text{M}$  (flavone), and  $86 \pm 16 \mu\text{M}$  (flavonol).

Representing the flavanone and flavone/flavonol subclasses, flavanone, flavone, and flavonol all showed CSPs of the same residues found to be perturbed in the *trans*-chalcone titration, indicating that all the molecules bind in a similar manner. However, flavanone, flavone, and flavonol bound YndB significantly weaker than *trans*-chalcone

with dissociation constants of  $32 \pm 3 \mu\text{M}$ ,  $62 \pm 9 \mu\text{M}$ , and  $86 \pm 16 \mu\text{M}$ , respectively [Figure 6.5]. The range of the dissociation constants mirrors the representation of each subclass in the virtual screen, where chalcones were the most abundantly ranked compounds, followed by flavanones and then the flavones/flavonols. Titration of oleic acid to YndB, which was used to represent the fatty acyl category of the lipids, showed no significant CSPs and therefore no evidence of binding (data not shown).

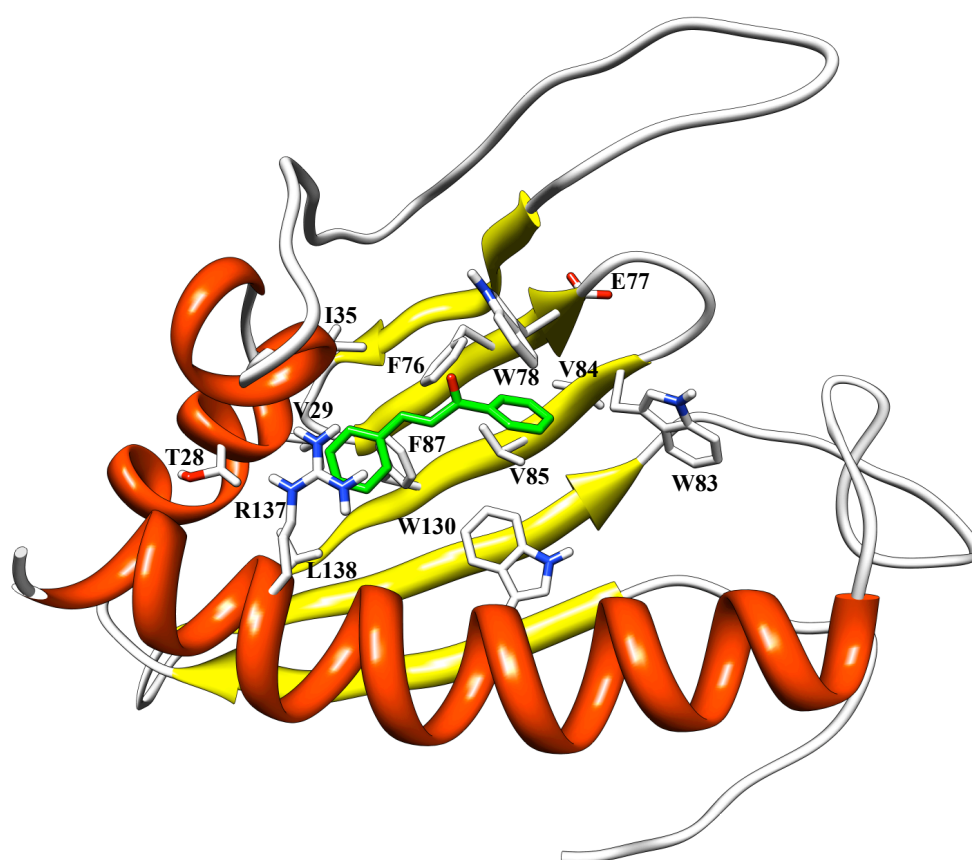
**6.3.5 *B. subtilis* YndB-ligand costructures.** Using AutoDock, the three-dimensional structures of *trans*-chalcone, flavanone, flavone, and flavonol were each docked into the YndB binding pocket identified by CASTp and supported by NMR CSPs. The docking of each compound did not result in much variation between the poses for each of the ligands. Out of 100 docked poses for each ligand, at least 80 were within a 2.0 Å rmsd of each other. A comparison of the most energetically favorable poses for each ligand shows that the compounds essentially bind with the same orientation [Figure 6.6]. The docked structures are also consistent with the 1:1 binding stoichiometry predicted by the NMR titration experiments and CSPs. Binding two or more compounds in the YndB binding pocket is sterically prohibitive and the NMR CSPs do not identify a secondary ligand binding site.



**Figure 6.6** A representation of the *B. subtilis* YndB protein surface using the NMR solution structure, where amino acid residues that exhibited NMR CSPs caused by the titration of *trans*-chalcone, flavanone, flavone, and flavonol are colored red ( $\geq 2$  standard deviations from the mean) and blue ( $\geq 1$  standard deviations from the mean). The residues with the largest CSPs can be found near the entrance to the ligand binding cavity, whereas the remaining residues are associated with helix  $\alpha 3$ . Shown within the ligand binding cavity are the docked conformations of the four ligands experimentally determined to bind YndB: chalcone (yellow), flavanone (green), flavone (purple), and flavonol (red).

The free energy of binding predicted by AutoDock was essentially identical for each compound, averaging  $-7.2$  kcal/mol which correlates to a dissociation constant of  $\sim 5$   $\mu$ M. With the exception of *trans*-chalcone, this is a stronger binding affinity than observed for the three compounds in the NMR titration experiments. This is not surprising since predicting the actual free energy of binding using AutoDock has an

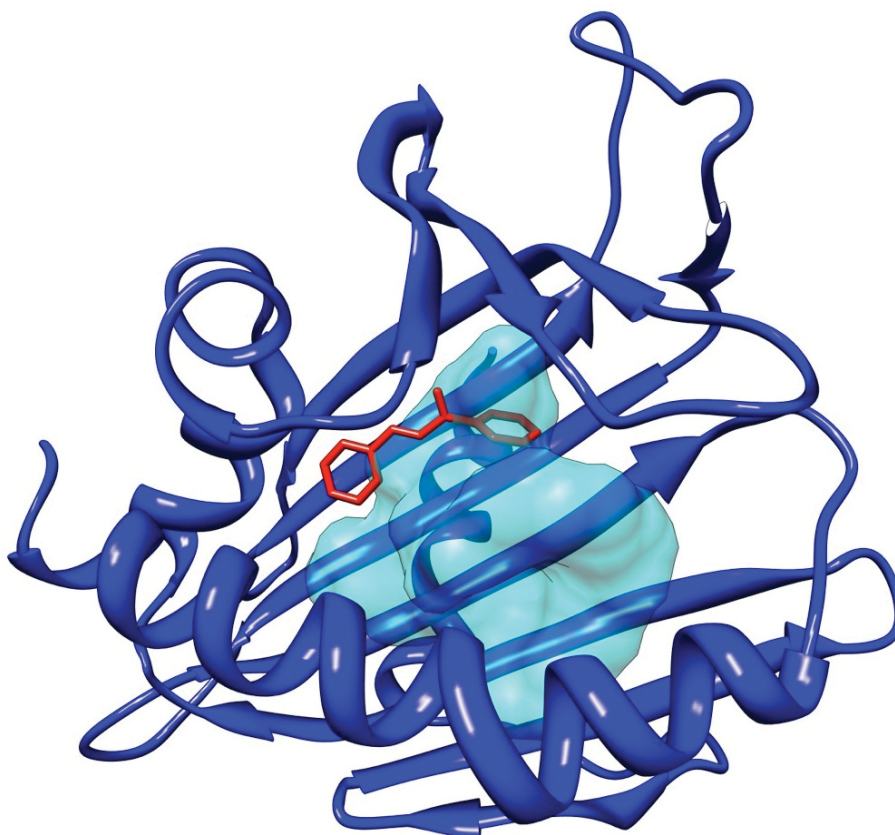
estimated error of 2.2 kcal/mol.<sup>60</sup> In the YndB-chalcone modeled structure, there are 14 residues that reside within 5 Å of the docked *trans*-chalcone, where five of these residues are aromatic [Figure 6.7]. These aromatic residues presumably have a strong influence on ligand binding and selectivity, consistent with the hydrophobic and aromatic nature of *trans*-chalcone and the other flavonoids. The binding of *trans*-chalcone to YndB does not appear to involve any hydrogen bonding interactions.



**Figure 6.7** The NMR solution structure of YndB docked with *trans*-chalcone (green). The sidechains for the 14 amino acid residues with 5 Å of the ligand are shown and labeled. Five aromatic sidechains surround the *trans*-chalcone molecule and form a hydrophobic pocket.

Most of the difficulty in generating an accurate protein-ligand co-structure for YndB stems from the suspected flexibility of the two loop regions that define the hydrophobic cavity. Any variation in the orientation of the loop sidechains directly results in changes in the binding site conformation that may be required to accommodate a ligand. This effect can be seen in some of the YndB structures found in the NMR ensemble. The YndB-chalcone model and the relative *trans*-chalcone orientation does correlate well with the binding site for the human phosphatidylcholine transfer protein (PC-TP) complexed with dilinoleoylphosphatidylcholine (PDB ID: 1ln1),<sup>61</sup> a protein structure in the related START domain family [Figure 6.8]. Although the sequence identity between YndB and PC-TP is low (5%), both proteins have structural similarities (3.6 Å rmsd) with binding pockets located in the same region of the protein. However, the binding pocket of YndB is significantly smaller than the pocket found in PC-TP due to the tighter packing of the  $\beta$ -sheet with the loop regions and the long  $\alpha$ 3-helix. Nevertheless, the overlay of the YndB-chalcone model with the PC-TP complex indicates that chalcone and the other flavonoids bind within a pocket similar to the large dilinoleoylphosphatidylcholine-binding pocket.





**Figure 6.8** A structural alignment of human PC-TP complexed with dilinoleoylphosphatidylcholine (PDB ID: 1ln1) with the *B. subtilis* YndB-chalcone NMR-based model. Only the *trans*-chalcone is shown from the YndB-chalcone structure. The structural alignment indicates the location of the docked *trans*-chalcone (red) relative to the PC-TP protein structure (blue) and a transparent molecular surface (cyan) representation of the bound dilinoleoylphosphatidylcholine. The binding pockets of the two proteins are in the same region, indicating a reasonable docking of *trans*-chalcone to YndB.

## 6.4 CONCLUSIONS

The NMR structure for *B. subtilis* protein YndB indicates that the protein adopts a helix-grip fold and is clearly a member of the Bet v 1-like superfamily. The YndB structure contains an apparent hydrophobic cavity between the long C-terminal  $\alpha$ -helix and the antiparallel  $\beta$ -sheet. Like other members of the Bet v 1-like superfamily, the cavity suggests YndB binds to lipids, sterols, polyketide antibiotics, or other hydrophobic molecules as part of its biological function. The YndB protein was originally assigned as

a START domain protein based on the high-sequence similarity to *B. cereus* BC4709 and *B. halodurans* BH1534, which were assigned to START domains based on common structural features.<sup>15,16</sup> Instead, SCOP and Pfam databases have suggested that YndB belongs to the closely related AHSA1 subfamily. Also, YndB, BC4709, and BH1534 do not have the additional N-terminal  $\beta$ -strands and the additional  $\alpha$ -helix that are characteristics of a START domain structure.<sup>3</sup> Likewise, a BLASTP sequence alignment search indicates YndB is more appropriately assigned as a member of AHSA1. The BLASTP search identified 58 proteins from organisms belonging to the Gram positive Bacillales order that are homologous to YndB with sequence identities > 39%. The functions for prokaryotic AHSA1 family members are typically classified as either a general stress protein or a conserved putative protein of unknown function. Likewise, the Dali search identified a large number of structural homologs to Bet v 1-like proteins, AHSA1 family members and an abundance of hypothetical proteins or proteins of unknown function.

To further explore the potential functional annotation of YndB, the *in silico* screen against a ~18,500 lipid-like chemical library was conducted. The best binders identified from the *in silico* screen were from the three general lipid classes of flavones/flavonols, flavanones, and chalcones/hydroxychalcones. Representative compounds from all three classes were screened by NMR, where *trans*-chalcone, flavanone, flavone, and flavonol were all shown to bind in the YndB hydrophobic cavity with  $K_D$  values of  $\leq 1$ , 32, 63, and 86  $\mu\text{M}$ , respectively. The fact that all four molecules chosen from the *in silico* screen were shown to bind YndB is rather remarkable and indicative of the inherent value of our approach. The typical hit rate for a high-throughput

screen is generally low (0.1-0.5%),<sup>62</sup> where the *in silico* screens may result in improved hit rates of up to 35-90%.<sup>31,63,64</sup> A model for the YndB-chalcone complex was shown to be consistent with the binding of dilinoleoylphosphatidylcholine to human PC-TP, a related START domain protein.

The binding of chalcone and flavanone to a *B. subtilis* protein is an intriguing observation because these molecules are primarily found in plants as precursors to flavonoid molecules used for antimicrobial defense, flower pigmentation, absorption of harmful UV radiation, and signaling between plants and beneficial microbes.<sup>65-69</sup> A number of structural homologs to YndB identified by Dali were Bet v1-like proteins in plants. In plants, chalcones are often synthesized from a cinnamoyl-CoA molecule followed by malonyl-CoA additions. The conversion of the resulting molecule into chalcone is catalyzed by the protein chalcone synthase (CHS).<sup>33</sup> Thus, chalcone is a key substrate for antibiotics or other flavonoid-based compounds. After the synthesis of chalcone, the chalcone isomerase (CHI) enzyme converts chalcone to flavanone, which is another compound identified to bind YndB. The other two binding compounds, flavone, and flavonol, are products of the various flavonoid synthesis pathways that initiate with the chalcone scaffold.<sup>33</sup> Bacteria do not possess CHS or CHI proteins and thus do not produce chalcone or flavanone. Some bacteria, including *B. subtilis*, do have proteins that appear to be homologous to the CHS proteins found in plants. These homologous proteins, known as type III polyketide synthases, appear to be pervasive in bacteria, indicating a possible mechanism for antimicrobial biosynthesis from chalcones, thus supporting the similarities of YndB to polyketide cyclases.<sup>7,70</sup> Likewise, homologs of CHI have also been identified in bacteria.<sup>71</sup> However, no Type III polyketide synthase

has been identified that is known to synthesize chalcones, and flavonoids have not been identified among the natural products of *Bacillus*. It seems unlikely that *B. subtilis* is producing chalcone-based antibiotics. These observations support the possibility of an exchange of genes between plants and bacteria, where these proteins are evolved for a unique bacterial function.<sup>72</sup>

The chalcone-binding property of YndB may be related to stress response as originally indicated based on the relationship to the eukaryotic Aha1 protein. In addition, Bet v1-like proteins in plants, which were shown to be structural homologs of YndB, are also primarily related to a stress response caused by a pathogen infection. An evaluation of the genes found near YndB in *B. subtilis* supports the stress response explanation. Although there are numerous uncharacterized membrane proteins identified in this cluster, one of these genes codes for the membrane bound protein, amino acid permease, which is involved in spore germination. The gene for protein BH1534 from *B. halodurans* also contains sporulation factors upstream and nucleotide metabolism downstream along with numerous putative membrane proteins. Likewise, the BC4709 gene for *B. cereus* has numerous membrane proteins in its cluster but also includes an ArsR transcriptional regulator and multidrug resistance proteins. The genes for all three of these proteins exist within regions containing stress response factors. Many other homologous proteins contain similar gene arrangements. These consistent gene arrangements hint at a likely stress-response mechanism, which is also supported by the similarities of these proteins to the eukaryotic Aha1 protein. Aha1 interacts with Hsp90, whereas prokaryotes have a homologous version of Hsp90 called HtpG, which has been

shown to be induced under high heat stress conditions. It should be noted, however, there is currently no direct correlation between HtpG and YndB or spore formation.<sup>73</sup>

*B. subtilis* is a plant growth promoting rhizobacterium, which is often found on the surface of plant roots and provides protection against pathogens through biofilm formation.<sup>74,75</sup> As chalcones are a key precursor to many antibiotics used by plants, it seems reasonable that *B. subtilis* has developed a mechanism of response toward chalcone-based compounds. Therefore, we hypothesize that the YndB protein, along with the homologous proteins BC4709 and BH1534, initiates a stress response-pathway when exposed to chalcone or chalcone-like compounds during the plant's response to pathogen infection. Potentially, this stress response may either induce processes to help control plant pathogens<sup>76</sup> and/or lead toward spore formation to protect *B. subtilis* from the impending release of antibiotics from the plant.<sup>77</sup> As flavonoids are routinely used as signaling molecules between plants and microbes during pathogen infections,<sup>78</sup> it is reasonable to consider chalcone binding as part of the symbiotic relationship between *B. subtilis* and plants.

## 6.5 REFERENCES

1. van Loon, L. C., Rep, M. & Pieterse, C. M. J. Significance of inducible defense-related proteins in infected plants. *Annu Rev Phytopathol* **44**, 135–162 (2006).
2. Gajhede, M. *et al.* X-ray and NMR structure of Bet v 1, the origin of birch pollen allergy. *Nat Struct Biol* **3**, 1040–1045 (1996).
3. Radauer, C., Lackner, P. & Breiteneder, H. The Bet v 1 fold: an ancient, versatile scaffold for binding of large, hydrophobic ligands. *BMC Evol Biol* **8**, 286 (2008).
4. Habe, H. & Omori, T. Genetics of polycyclic aromatic hydrocarbon metabolism in diverse aerobic bacteria. *Biosci Biotechnol Biochem* **67**, 225–243 (2003).
5. Ponting, C. P. & Aravind, L. START: a lipid-binding domain in StAR, HD-ZIP and signalling proteins. *Trends Biochem Sci* **24**, 130–132 (1999).
6. Stocco, D. M. StAR protein and the regulation of steroid hormone biosynthesis. *Annu Rev Physiol* **63**, 193–213 (2001).

7. Ames, B. D. *et al.* Crystal structure and functional analysis of tetracenomyacin ARO/CYC: implications for cyclization specificity of aromatic polyketides. *Proc Natl Acad Sci USA* **105**, 5349–5354 (2008).
8. Apweiler, R., Bairoch, A. & Wu, C. H. Protein sequence databases. *Curr Opin Chem Biol* **8**, 76–80 (2004).
9. Lotz, G. P., Lin, H., Harst, A. & Obermann, W. M. J. Aha1 binds to the middle domain of Hsp90, contributes to client protein activation, and stimulates the ATPase activity of the molecular chaperone. *J Biol Chem* **278**, 17228–17235 (2003).
10. Panaretou, B. *et al.* Activation of the ATPase activity of hsp90 by the stress-regulated cochaperone aha1. *Mol Cell* **10**, 1307–1318 (2002).
11. Singh, S. *et al.* Structural insight into the self-sacrifice mechanism of enediyne resistance. *ACS Chem Biol* **1**, 451–460 (2006).
12. Halcomb, R. L. Organic synthesis and cell biology: partners in controlling gene expression. *Proc Natl Acad Sci USA* **91**, 9197–9199 (1994).
13. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res* **40**, D290–D301 (2012).
14. Andreeva, A. *et al.* Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* **36**, D419–25 (2008).
15. Mercier, K. A. *et al.* (1)H, (13)C, and (15)N NMR assignments for the *Bacillus subtilis* yndB START domain. *Biomol NMR Assign* **3**, 191–194 (2009).
16. Liu, G. *et al.* NMR data collection and analysis protocol for high-throughput protein structure determination. *Proc Natl Acad Sci USA* **102**, 10487–10492 (2005).
17. Greene, L. H. *et al.* The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* **35**, D291–7 (2007).
18. Montelione, G. T. *et al.* Unique opportunities for NMR methods in structural genomics. *J Struct Funct Genomics* **10**, 101–106 (2009).
19. del Val, C. *et al.* High-throughput protein analysis integrating bioinformatics and experimental assays. *Nucleic Acids Res* **32**, 742–748 (2004).
20. Joshi, T., Chen, Y., Becker, J. M., Alexandrov, N. & Xu, D. Genome-scale gene function prediction using multiple sources of high-throughput data in yeast *Saccharomyces cerevisiae*. *OMICS* **8**, 322–333 (2004).
21. Lee, Y.-H. *et al.* Gene knockdown by large circular antisense for high-throughput functional genomics. *Nat Biotechnol* **23**, 591–599 (2005).
22. Tucker, C. L. High-throughput cell-based assays in yeast. *Drug Discov Today* **7**, S125–30 (2002).
23. Oude Elferink, R. One step further towards real high-throughput functional genomics. *Trends Biotechnol* **21**, 146–7– discussion 147–8 (2003).
24. Mercier, K. A. *et al.* FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* **128**, 15292–15299 (2006).
25. Powers, R., Mercier, K. A. & Copeland, J. C. The application of FAST-NMR for the identification of novel drug discovery targets. *Drug Discov Today* **13**, 172–179 (2008).
26. Morris, G., Goodsell, D., Halliday, R. & Huey, R. Automated docking using a

- Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**, 1639–1662 (1998).
27. Stark, J. L. & Powers, R. Rapid protein-ligand costructures using chemical shift perturbations. *J Am Chem Soc* **130**, 535–545 (2008).
  28. Powers, R. *et al.* Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **65**, 124–135 (2006).
  29. Powers, R., Copeland, J. & Stark, J. L. Searching the protein structure database for ligand-binding site similarities using CPASS v. 2. *BMC Res Notes* (2011).
  30. Fahy, E., Sud, M., Cotter, D. & Subramaniam, S. LIPID MAPS online tools for lipid research. *Nucleic Acids Res* **35**, W606–12 (2007).
  31. Ghosh, S., Nie, A., An, J. & Huang, Z. Structure-based virtual screening of chemical libraries for drug discovery. *Curr Opin Chem Biol* **10**, 194–202 (2006).
  32. Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today* **11**, 580–594 (2006).
  33. Austin, M. B. & Noel, J. P. The chalcone synthase superfamily of type III polyketide synthases. *Nat Prod Rep* **20**, 79–110 (2003).
  34. Choudhary, D. K. & Johri, B. N. Interactions of *Bacillus* spp. and plants--with special reference to induced systemic resistance (ISR). *Microbiol Res* **164**, 493–513 (2009).
  35. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612 (2004).
  36. Dundas, J. *et al.* CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res* **34**, W116–8 (2006).
  37. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
  38. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).
  39. Gish, W. & States, D. J. Identification of protein coding regions by database similarity search. *Nat Genet* **3**, 266–272 (1993).
  40. Holm, L., Kaariainen, S., Rosenstrom, P. & Schenkel, A. Searching protein structure databases with DaliLite v.3. *Bioinformatics* **24**, 2780–2781 (2008).
  41. Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protocols Bioinformatics* 2.3.1–2.3.22 (2002).
  42. Fahy, E. *et al.* A comprehensive classification system for lipids. *J Lipid Res* **46**, 839–861 (2005).
  43. Fahy, E. *et al.* Update of the LIPID MAPS comprehensive classification system for lipids. *J Lipid Res* **50 Suppl**, S9–14 (2009).
  44. Boström, J., Greenwood, J. R. & Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J Mol Graph Model* **21**, 449–462 (2003).
  45. McGann, M., Almond, H., Nicholls, A., Grant, J. & Brown, F. Gaussian docking functions. *Biopolymers* **68**, 76–90 (2003).
  46. McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J Chem Inf Model* **51**, 578–596 (2011).
  47. Kellenberger, E., Rodrigo, J., Muller, P. & Rognan, D. Comparative evaluation of

- eight docking tools for docking and virtual screening accuracy. *Proteins* **57**, 225–242 (2004).
48. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6**, 277–293 (1995).
  49. Johnson, B. A. Using NMRView to visualize and analyze the NMR spectra of macromolecules. *Methods Mol Biol* **278**, 313–352 (2004).
  50. Morton, C. J. *et al.* Solution structure and peptide binding of the SH3 domain from human Fyn. *Structure* **4**, 705–714 (1996).
  51. Fielding, L. NMR methods for the determination of protein-ligand dissociation constants. *Curr Top Med Chem* **3**, 39–53 (2003).
  52. Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* **28**, 1145–1152 (2007).
  53. Sousa, S. F., Fernandes, P. A. & Ramos, M. J. Protein-ligand docking: current status and future challenges. *Proteins* **65**, 15–26 (2006).
  54. Richardson, J. S. & Richardson, D. C. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci USA* **99**, 2754–2759 (2002).
  55. Arakane, F. *et al.* Steroidogenic acute regulatory protein (StAR) retains activity in the absence of its mitochondrial import sequence: implications for the mechanism of StAR action. *Proc Natl Acad Sci USA* **93**, 13731–13736 (1996).
  56. Feng, L., Chan, W., Roderick, S. & Cohen, D. High-level expression and mutagenesis of recombinant human phosphatidylcholine transfer protein using a synthetic gene: evidence for a C-terminal membrane binding domain. *Biochemistry* **39**, 15399–15409 (2000).
  57. Markwick, P. R. L., Malliavin, T. & Nilges, M. Structural biology by NMR: structure, dynamics, and interactions. *PLoS Comput Biol* **4**, e1000168 (2008).
  58. Mercier, K. A. *et al.* Structure and function of *Pseudomonas aeruginosa* protein PA1324 (21-170). *Protein Sci* **18**, 606–618 (2009).
  59. Shortridge, M. D., Hage, D. S., Harbison, G. S. & Powers, R. Estimating protein-ligand binding affinity using high-throughput screening by NMR. *J Comb Chem* **10**, 948–958 (2008).
  60. Rosenfeld, R. J. *et al.* Automated docking of ligands to an artificial active site: augmenting crystallographic analysis with computer modeling. *J Comput Aided Mol Des* **17**, 525–536 (2003).
  61. Roderick, S. L. *et al.* Structure of human phosphatidylcholine transfer protein in complex with its ligand. *Nat Struct Biol* **9**, 507–511 (2002).
  62. Dove, A. Drug screening--beyond the bottleneck. *Nat Biotechnol* **17**, 859–863 (1999).
  63. Doman, T. N. *et al.* Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J Med Chem* **45**, 2213–2221 (2002).
  64. Konstantinou-Kirtay, C., Mitchell, J. B. O. & Lumley, J. A. Scoring functions and enrichment: a case study on Hsp90. *BMC Bioinformatics* **8**, 27 (2007).
  65. Forkmann, G. & Martens, S. Metabolic engineering and applications of flavonoids. *Curr Opin Biotechnol* **12**, 155–160 (2001).
  66. Steinkellner, S. *et al.* Flavonoids and strigolactones in root exudates as signals in



- symbiotic and pathogenic plant-fungus interactions. *Molecules* **12**, 1290–1306 (2007).
67. Long, S. R. Rhizobium-legume nodulation: life together in the underground. *Cell* **56**, 203–214 (1989).
  68. Winkel-Shirley, B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol* **126**, 485–493 (2001).
  69. Perret, X., Staehelin, C. & Broughton, W. J. Molecular basis of symbiotic promiscuity. *Microbiol Mol Biol Rev* **64**, 180–201 (2000).
  70. Shen, B. & Hutchinson, C. R. Deciphering the mechanism for the assembly of aromatic polyketides by a bacterial polyketide synthase. *Proc Natl Acad Sci USA* **93**, 6600–6604 (1996).
  71. Gensheimer, M. & Mushegian, A. Chalcone isomerase family and fold: no longer unique to plants. *Protein Sci* **13**, 540–544 (2004).
  72. Bode, H. B. & Müller, R. Possibility of bacterial recruitment of plant genes associated with the biosynthesis of secondary metabolites. *Plant Physiol* **132**, 1153–1161 (2003).
  73. Versteeg, S., Escher, A., Wende, A., Wiegert, T. & Schumann, W. Regulation of the *Bacillus subtilis* heat shock gene *htpG* is under positive control. *J Bacteriol* **185**, 466–474 (2003).
  74. Rudrappa, T., Quinn, W. J., Stanley-Wall, N. R. & Bais, H. P. A degradation product of the salicylic acid pathway triggers oxidative stress resulting in down-regulation of *Bacillus subtilis* biofilm formation on *Arabidopsis thaliana* roots. *Planta* **226**, 283–297 (2007).
  75. Rudrappa, T. & Bais, H. P. *Arabidopsis thaliana* Root Surface Chemistry Regulates in Planta Biofilm Formation of *Bacillus subtilis*. *Plant Signal Behav* **2**, 349–350 (2007).
  76. Weller, D. M. & Thomashow, L. S. in *Molecular Ecology of Rhizosphere Microorganisms: Biotechnology and the Release of GMOs* (O'Gara, F., Dowling, D. N. & Boesten, B.) 1–18 (Wiley-VCH Verlag GmbH, 1994). doi:10.1002/9783527615810.ch1
  77. Piggot, P. J. & Hilbert, D. W. Sporulation of *Bacillus subtilis*. *Curr Opin Microbiol* **7**, 579–586 (2004).
  78. Mabood, F., Jung, W. J. & Smith, D. L. in *Soil Biology: Molecular Mechanisms of Plant and Microbe Coexistence* **15**, 291–318 (Springer, 2008).

## CHAPTER 7

### VIRTUAL SCREENING OF A FUNCTION-BASED COMPOUND LIBRARY

#### 7.1 INTRODUCTION

The vast majority of initial leads in drug discovery are identified from high-throughput screens (HTS).<sup>1-3</sup> Pharmaceutical companies have invested heavily in developing and maintaining large chemical libraries (>1,000,000 compounds), which are screened using automated, biological assays intended to monitor a specific response or biological effect.<sup>3</sup> Unfortunately, HTS is extremely inefficient due to the high cost of developing, maintaining, and screening such large libraries of compounds. Furthermore, the random search for an effective drug in the vastness of chemical space ( $\sim 10^{60}$  compounds)<sup>4</sup> is extremely challenging. Thus, HTS hit rates are typically very low, where <0.5% of compounds exhibit any inhibitor activity in an assay.<sup>5</sup> Correspondingly, HTS assays are highly inefficient since most of the screening effort is spent on the analysis of negative data. HTS assays, by nature, are mechanistic “black boxes,” and a response does not provide any information on the mechanism of inhibition. This often leads to numerous false positives from undesirable interactions<sup>6-8</sup> that may lead the drug discovery project astray. Improving the efficiency of drug discovery requires the implementation of advanced techniques that better guide the selection of lead candidates without sacrificing speed.

Ideally, an entirely *in silico* approach to screening a large compound library would significantly improve efficiency and reduce costs.<sup>9,10</sup> However, several assessments of virtual screens have concluded that, without prior in-depth analysis of the

protein's ligand binding site, only a marginal improvement in finding successful leads is observed relative to standard HTS.<sup>11</sup> The inherent errors of a docking program to calculate the strength of interactions for each unique protein-ligand system makes it difficult to reliably rank the compound library from best binder to worst.

Fragment-based drug discovery is an alternative method that focuses on a bottom-up approach to finding a good binder. Low molecular weight compounds (< 250 amu) are used to find subpockets within the overall active site.<sup>12,13</sup> Compounds that are found in nearby subpockets can be chemically linked to produce a theoretically stronger inhibitor or binder. This approach allows for a smaller library to be screened.<sup>12,13</sup> However, experimental screens of fragment-based libraries are difficult due to the low affinity of the compounds.<sup>14,15</sup> Unfortunately, using virtual screening to prioritize a fragment-based compound libraries is challenging since small compounds tend to be more promiscuous binders and any error in the calculation of an interaction has a greater effect on the overall rankings of the docked compounds.<sup>16</sup> While fragment-based virtual screens are more efficient, they are not often used to prioritize a fragment-based compound library because of these challenges.

In drug discovery, the large compound libraries and fragment-based compound libraries are intended to explore structural diversity in order to find a novel drug. Similarly, the FAST-NMR approach<sup>17,18</sup> to the functional annotation of uncharacterized proteins uses a compound library of approximately 420 compounds that explores functional diversity by only including compounds that have been shown to have biological activity.<sup>19</sup> This function-based compound library has a greater likelihood of

including compounds that are biologically relevant to the molecular function of the protein.<sup>19</sup>

In FAST-NMR, the function-based compound library is initially screened against the protein of interest using a 1D <sup>1</sup>H line-broadening NMR screen. The compounds that show line broadening are then followed by a 2D <sup>1</sup>H, <sup>15</sup>N-HSQC NMR screen to validate specific binding and locate the binding site. The 1D line-broadening screens require a significant amount of resources for preparation, execution, and analysis. Could virtual screening be used to prioritize the compounds in the function-based library instead of performing a 1D line-broadening screen?

Two proteins, SAV1430 and PA1324, have been previously screened using the FAST-NMR approach. The results of the 1D line-broadening screens for these proteins are compared to the results of the virtual screens using the same function-based compound library. The ability of virtual screening to prioritize compounds shown to experimentally bind these proteins is then evaluated.

## 7.2 MATERIALS AND METHODS

Virtual screens using the function-based compound library were performed on two proteins previously screened by FAST-NMR: *Staphylococcus aureus* protein SAV1430 (PDB ID: 1PQX) and *Pseudomonas aeruginosa* protein PA1324 (PDB ID: 1XPN).<sup>17,20</sup> The function-based compound library used for the virtual screens consists of the same 420 compounds that were experimentally screened by FAST-NMR. The two-dimensional structure for each compound in the library was converted into a three-dimensional structure using MM2 energy minimization in Chem3D (CambridgeSoft;

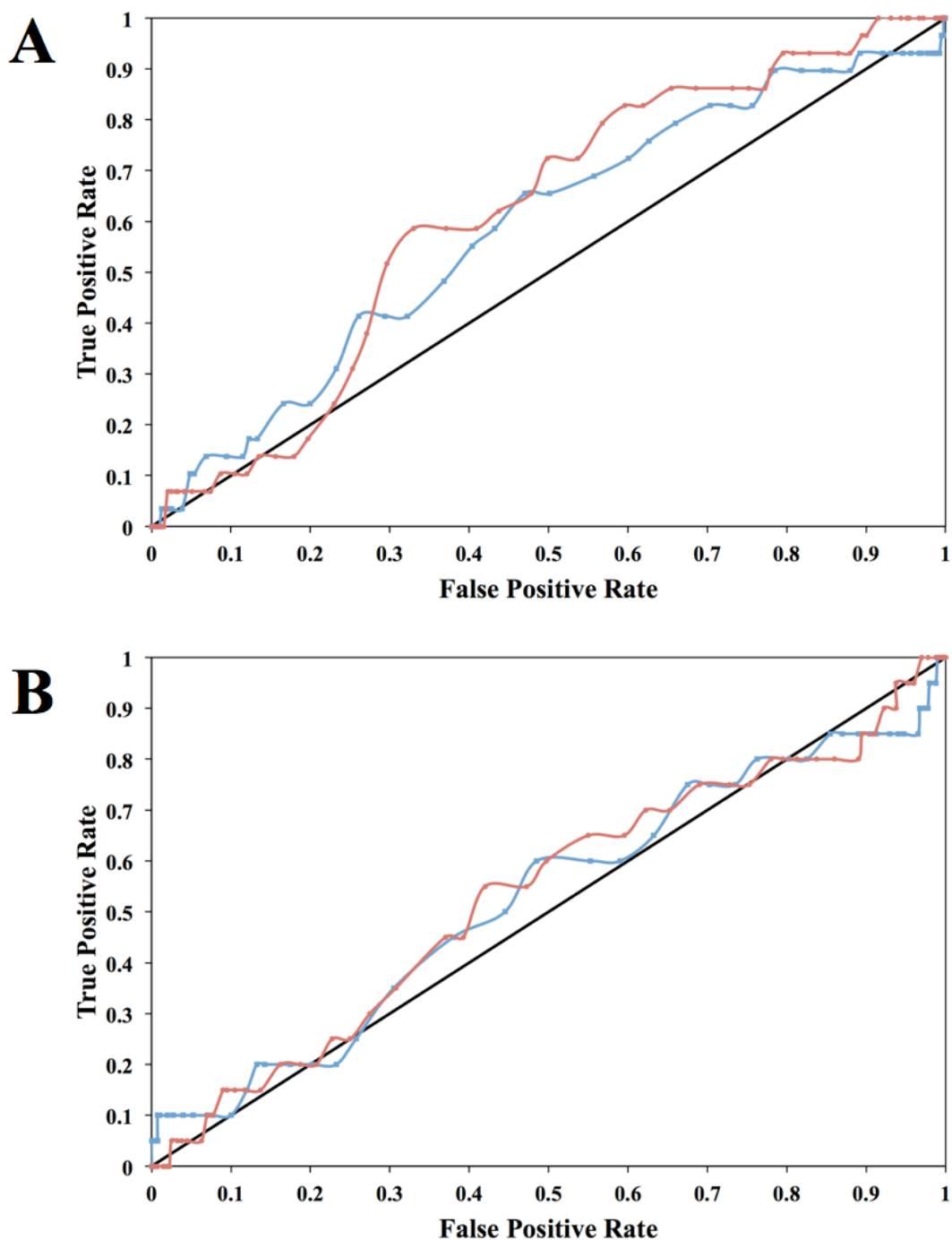
<http://www.cambridgesoft.com>). All hydrogens were added to the protein and compounds using standard protonation states at a neutral pH.

Docking was performed using AutoDock 4.01<sup>21-23</sup> with the AutoDockTools 1.4.5 (<http://mgltools.scripps.edu>)<sup>23,24</sup> graphical interface, where 10 different binding conformations were simulated for each compound binding with the protein. Each protein was docked to the compound library twice, once with the grid encompassing the entire protein (blind docking) and again with the grid encompassing the experimental binding site identified during the FAST-NMR screens (guided docking). Grid maps were generated with 0.447 Å spacing in both cases with enough size to accommodate the largest compounds. The docking calculations were performed using the Lamarckian genetic algorithm default settings with a population size of 300 and 5,000,000 energy evaluations. The calculations were performed on an Intel Xeon 3.06 GHz dual processor Linux workstation and required approximately 10 days to complete.

The docked conformer with the lowest binding energy for each protein-ligand pair was selected and then compared to the lowest energy conformers of the other protein-ligand pairs. Receiver operating characteristics (ROC) curves (described in Chapter 1) were generated using the compounds identified as binders during the 1D line-broadening screen in FAST-NMR as true positives. From these prior experiments, 21 binders were identified for SAV1430<sup>17</sup> and 20 were identified for PA1324.<sup>20</sup> The true positive rates were plotted against false positive rates over the full range of AutoDock binding energies for both the blind and guided docking.

### 7.3 RESULTS AND DISCUSSION

The ideal goal of virtual screens is to enrich the compound library such that selecting a small fraction of highly ranked compounds leads to a greater likelihood of including true binders in the experimental screen. In the blind virtual screens of SAV1430 and PA1324, neither screen resulted in any significant enrichment of true binders [Figure 7.1A,B]. In fact, the true binders had a wide range of binding energies. Thus selecting the lowest energy protein-ligand complexes provided no benefit in selecting true binders. Of the 20 docked compounds with the lowest binding energies, SAV1430 had two true binders while PA1324 had one. None of these true binders for were the best binder. In fact, the two true binders for SAV1430 weren't even docked into the experimentally determined binding site; indicating that the high ranking was erroneous. The best binder identified in FAST-NMR for both SAV1430 (O-phospho-L-tyrosine) and PA1324 (suramin) ranked in the bottom 50% of the virtual screen (Table 7.1). Additionally, neither of these best binders was properly docked into the experimentally-determined binding sites of their respective proteins.



**Figure 7.1** ROC curves showing the true positive rate relative to the false positive rate of virtual screens for (A) SAV1430 and (B) PA1324. True positives are defined as the compounds that were shown to experimentally bind the proteins in the FAST-NMR 1D line-broadening screens. The red ROC curve indicates the results based on a blind virtual screen where the docking grid encompasses the entire protein. The blue ROC curve indicates the results for a guided virtual screen where the docking grid is focused on the experimental binding site.

**Table 7.1 AutoDock binding energies<sup>a</sup> for virtual screens of compound library**

	SAV1430		PA1324	
	Blind	Guided	Blind	Guided
<b>Range</b>	-9.02 to -1.69	-7.21 to 0.25	-9.54 to -0.67	-7.35 to -1.41
<b>Average</b>	-5.51	-4.54	-5.23	-4.82
<b>Best experimental binder<sup>b</sup></b>	-5.52 (214 <sup>th</sup> )	-5.35 (92 <sup>nd</sup> )	-4.40 (331 <sup>st</sup> )	-7.35 (1 <sup>st</sup> )

<sup>a</sup> Binding energies in kcal/mol

<sup>b</sup> Binding energies and rank among docked compounds in library for best experimental binder identified in FAST-NMR screens.

The results of the blind virtual screen are not surprising. Blind docking indicates no prior knowledge of the binding site, thus the docking process must spend time searching for an energetically favorable binding site as well as orienting the ligand into a more favorable conformation and pose.<sup>25,26</sup> Despite using a greater number of energy evaluations (5,000,000) than the default, these additional resources still could not enrich the results of the virtual screens. Additionally, most of the lowest energy binders in these virtual screens did not dock into the experimental binding site identified by FAST-NMR.

The results of the blind screen indicates a virtual screen is unlikely to replace the 1D line-broadening screen of FAST-NMR as a means to prioritize the ligands for screening by 2D <sup>1</sup>H, <sup>15</sup>N-HSQC. But does the use of a focused function-based compound library provide any benefit towards virtual screening enrichment? When the virtual screen was run again with the grid set to encompass only the experimentally-determined binding site from FAST-NMR, the resulting enrichment did not significantly improve [Figure 7.1A,B]. Once again, selecting the lowest energy docked compounds did not improve the chances of identifying a true binder. Of the 20 docked compounds with the



lowest binding energies, only two of the SAV1430 and PA1324 true binders were identified in the virtual screen. Neither of these compounds was identified in the blind virtual screen.

The best binder for SAV1430 (O-phospho-L-tyrosine) had very similar binding energies in both the blind virtual screen (-5.52 kcal/mol) and the guided virtual screen (-5.35 kcal/mol) despite docking to completely different regions on the protein. Remarkably, the guided virtual screen did identify the best binder for PA1324 (suramin) as having the lowest binding energy. However, this is likely due to the size of suramin (1,291 amu), which would undergo a significant number of interactions compared to most molecules in the compound library. This is apparent from a comparison of the binding energies, where the energy terms defined by van der Waals forces, hydrogen bonding, and desolvation had the lowest values (-9.35 kcal/mol in blind screen; -7.80 kcal/mol in guided screen) for suramin compared to all the other compounds. On the other hand, the positive torsional energy (4.39 kcal/mol) for suramin offset these very low energies, which is why suramin was not identified as the lowest energy docked compound in the blind virtual screen. The docking energetics of suramin in the guided virtual screen was more favorable since the binding pocket contains a very positive electrostatic surface that interacts with several negative charges on suramin. This electrostatic interaction produced a low electrostatic energy (-3.14 kcal/mol) that contributed significantly to the total binding energy. This interaction was also previously determined as an important binding component during the functional annotation.<sup>20</sup> In the blind screen, the suramin molecule was unable to find this positive electrostatic surface.

## 7.4 CONCLUSIONS

In FAST-NMR, a typical 1D line-broadening screen of the compound library with a protein target identifies approximately 20 compounds that would be validated by a 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen. For virtual screens to be a viable replacement, selecting the 20 compounds with the lowest binding energies to be validated by  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screens should include a significant number of actual binders. Unfortunately, virtual screens of SAV1430 and PA1324 showed that no enrichment of actual binders occurred. Thus, replacing the experimental 1D line-broadening screen of FAST-NMR with a purely computational virtual screen of the same function-based compound library is not beneficial. The composition of the compound library does not appear to improve the enrichment process.

Virtual screening does have some benefits for FAST-NMR beyond just replacing the 1D line-broadening screen. The guided virtual screens of SAV1430 and PA1324 show that while the compound library was not enriched, it is still possible to identify tight binders. However, this is typically dependent upon the protein system being explored. PA1324 has a positive electrostatic surface in the binding pocket that interacts strongly with the negative charges of suramin, which made the interaction easier to score. In Chapter 6, the virtual screen of YndB was successful because the focused lipid library and well-defined hydrophobic pocket of the protein made steric hindrance and hydrophobic interactions the primary driving forces.<sup>27</sup>

Should virtual screening be used for FAST-NMR? Because the resource cost for virtual screening is relatively low, there is some merit in utilizing it to supplement a FAST-NMR assay, because it may provide information on favorable binding regions and

predict dominant protein-ligand interactions. Virtual screening can also help investigate compounds that are not present in the function-based compound library with minimal additional effort. In the case of YndB (Chapter 6),<sup>27</sup> virtual screening was used because experimentally screening a lipid library by NMR was not feasible, and there was significant prior evidence to suggest the location of the binding site the identity of likely ligands.

## 7.5 REFERENCES

1. Kenny, B. A., Bushfield, M., Parry-Smith, D. J., Fogarty, S. & Treherne, J. M. The application of high-throughput screening to novel lead discovery. *Prog Drug Res* **51**, 245–269 (1998).
2. Davis, A. M., Keeling, D. J., Steele, J., Tomkinson, N. P. & Tinker, A. C. Components of successful lead generation. *Curr Top Med Chem* **5**, 421–439 (2005).
3. Sams-Dodd, F. Drug discovery: selecting the optimal approach. *Drug Discov Today* **11**, 465–472 (2006).
4. Fink, T. & Reymond, J.-L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J Chem Inf Model* **47**, 342–353 (2007).
5. Lahana, R. How many leads from HTS? *Drug Discov Today* **4**, 447–448 (1999).
6. Goode, D. R., Totten, R. K., Heeres, J. T. & Hergenrother, P. J. Identification of promiscuous small molecule activators in high-throughput enzyme activation screens. *J Med Chem* **51**, 2346–2349 (2008).
7. McGovern, S. L., Caselli, E., Grigorieff, N. & Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J Med Chem* **45**, 1712–1722 (2002).
8. McGovern, S. L., Helfand, B. T., Feng, B. & Shoichet, B. K. A specific mechanism of nonspecific inhibition. *J Med Chem* **46**, 4265–4272 (2003).
9. Foloppe, N. *et al.* Identification of chemically diverse Chk1 inhibitors by receptor-based virtual screening. *Bioorg Med Chem* **14**, 4792–4802 (2006).
10. Richardson, C. M. *et al.* Discovery of a potent CDK2 inhibitor with a novel binding mode, using virtual screening and initial, structure-guided lead scoping. *Bioorg Med Chem Lett* **17**, 3880–3885 (2007).
11. Warren, G. L. *et al.* A critical assessment of docking programs and scoring functions. *J Med Chem* **49**, 5912–5931 (2006).
12. Carr, R. A. E., Congreve, M., Murray, C. W. & Rees, D. C. Fragment-based lead

- discovery: leads by design. *Drug Discov Today* **10**, 987–992 (2005).
13. Congreve, M., Chessari, G., Tisi, D. & Woodhead, A. J. Recent developments in fragment-based drug discovery. *J Med Chem* **51**, 3661–3680 (2008).
  14. Pellecchia, M. *et al.* Perspectives on NMR in drug discovery: a technique comes of age. *Nat Rev Drug Discov* **7**, 738–745 (2008).
  15. Sun, C., Petros, A. M. & Hajduk, P. J. Fragment-based lead discovery: challenges and opportunities. *J Comput Aided Mol Des* **25**, 607–610 (2011).
  16. Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today* **11**, 580–594 (2006).
  17. Mercier, K. A. *et al.* FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* **128**, 15292–15299 (2006).
  18. Powers, R., Mercier, K. A. & Copeland, J. C. The application of FAST-NMR for the identification of novel drug discovery targets. *Drug Discov Today* **13**, 172–179 (2008).
  19. Mercier, K. A., Germer, K. & Powers, R. Design and characterization of a functional library for NMR screening against novel protein targets. *Comb Chem High Throughput Screen* **9**, 515–534 (2006).
  20. Mercier, K. A. *et al.* Structure and function of *Pseudomonas aeruginosa* protein PA1324 (21-170). *Protein Sci* **18**, 606–618 (2009).
  21. Morris, G., Goodsell, D., Halliday, R. & Huey, R. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**, 1639–1662 (1998).
  22. Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* **28**, 1145–1152 (2007).
  23. Morris, G. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* (2009).doi:10.1002/jcc.21256
  24. Sanner, M. F. Python: a programming language for software integration and development. *J Mol Graph Model* **17**, 57–61 (1999).
  25. Hetényi, C. & van der Spoel, D. Blind docking of drug-sized compounds to proteins with up to a thousand residues. *FEBS Lett* **580**, 1447–1450 (2006).
  26. Stark, J. L. & Powers, R. Rapid protein-ligand costructures using chemical shift perturbations. *J Am Chem Soc* **130**, 535–545 (2008).
  27. Stark, J. L. *et al.* Solution structure and function of YndB, an AHSA1 protein from *Bacillus subtilis*. *Proteins* **78**, 3328–3340 (2010).

## CHAPTER 8

### HUMAN DNAJA1: A POTENTIAL THERAPEUTIC TARGET FOR PANCREATIC CANCER

#### 8.1 INTRODUCTION

Despite the decline of cancer-related mortality in the past decade, effective approaches to early diagnosis and treatment of pancreatic cancer remain elusive. Although it accounts for only 3% (43,000 new cases every year) of all cancers, pancreatic cancer is the fourth leading cause of cancer death in the United States (37,000 deaths annually) and has the highest mortality rate of any cancer.<sup>1,2</sup> Those with an operable early-stage of the disease have a 5-year survival rate of about 20%.<sup>1,3</sup> Unfortunately, 80% of all pancreatic cancer diagnoses indicate an advanced stage of the disease that is beyond the point of surgery.<sup>2-4</sup> Inoperable forms of pancreatic cancer have a five-year survival rate of only 3%. The difficulty in detecting or diagnosing pancreatic cancer has several causes: the early stages of pancreatic cancer don't typically exhibit symptoms; the symptoms that do occur are often similar to other illnesses; and the location of the pancreas behind other organs can hinder detection.<sup>2</sup>

Most patients with advanced pancreatic cancer are treated with chemotherapy based on gemcitabine, which is a cytotoxic nucleoside drug that primarily inhibits DNA synthesis.<sup>5</sup> However, this treatment is only mildly effective for patients with an advanced stage of pancreatic cancer and only provides a 5.91 month increase in the median survival rate.<sup>6</sup> Also, gemcitabine-resistant forms of pancreatic cancer or acquired resistance during treatment are common problems.<sup>7</sup> Correspondingly, there have been numerous attempts to combine gemcitabine with other cytotoxic agents, such as 5-fluorouracil or capecitabine, however these approaches have been mostly unsuccessful.<sup>8</sup> It is apparent

that a cytotoxic approach to treating pancreatic cancer is not an effective therapy. Therefore, identifying novel, but druggable, protein targets for the treatment of pancreatic cancer and improving the quality of life for patients is an essential need.

This chapter reports the development of a pancreatic cancer ‘omics database (Borg) designed to identify potentially interesting protein targets for drug discovery. The human protein DnaJ homolog subfamily A member 1 (DNAJA1) was selected as a potentially interesting therapeutic target for the treatment of pancreatic cancer. The potential importance of DNAJA1 to pancreatic cancer is demonstrated with stress-response cell-based assays using cell-lines overexpressing DNAJA1. Additionally, the structure of the J-domain of DNAJA1 (A1-JD) was determined by NMR spectroscopy followed by the identification of potential binding sites using a ligand-based NMR screen.

## **8.2 MATERIALS AND METHODS**

**8.2.1 Selection of DNAJA1 from a pancreatic cancer ‘omics database.** The scientific literature was searched to identify proteins potentially associated with pancreatic cancer. Five separate proteomic studies identified a total of 844 unique proteins that were differentially expressed in various pancreatic cancer cell lines.<sup>9-13</sup> Additionally, three separate genomics studies of mutation frequency or gene expression in 71 pancreatic cancer cell lines identified 4,492 genes that are significantly modulated.<sup>13-15</sup> The resulting 5,336 proteins/genes were combined into a pancreatic cancer ‘omics database and evaluated with a simplistic, unsupervised method (Borg) to

bridge the gap between the large number of proteins/genes and human protein annotations.

In Borg, each protein or gene identified in the database was manually assigned a reviewed UniProtKB<sup>16,17</sup> accession number in order to facilitate a uniform starting point for bioinformatics analysis. The UniProtKB accession numbers were used to cross-reference the RCSB Protein Data Bank (PDB)<sup>18,19</sup> to determine if an experimental structure exists for each protein. Additionally, homologous structures were identified with a BlastP<sup>20,21</sup> sequence search against the RCSB PDB using an E-value of less than  $1 \times 10^{-10}$ . These protein sequences were also used to search the PSI: Knowledgebase<sup>22</sup> to identify similar structures (structure count with E-value  $< 0.001$ ) and targets (target count with E-value  $< 0.001$ ).

Prioritization of potential therapeutic targets proceeded by generating a functional network of the proteins identified in the 'omics database using protein function annotations from GO,<sup>23</sup> OMA,<sup>24</sup> BindingDB,<sup>25</sup> DIP,<sup>26</sup> eggNOG,<sup>27</sup> Ensembl,<sup>28</sup> KEGG,<sup>29,30</sup> PFAM,<sup>31</sup> STRING,<sup>32</sup> and RCSB PDB<sup>18,19</sup> databases.<sup>‡‡</sup> Each protein was given a binary score ("1" if the protein has the annotation or "0" if it does not) for each of the 7,795 possible functional annotations identified from these databases. The dimension space of this dataset was reduced to three dimensions using principal component analysis (PCA), and then clustered into functionally distinct groups using Gaussian mixture models. Proteins in each cluster without experimental structures were then prioritized by assigning scores based on their annotation count, their neighboring nodes in STRING<sup>32</sup>

---

<sup>‡‡</sup> The functional clustering of the pancreatic cancer 'omics database was developed and performed by Brad Worley using the in-house Borg software.

and DIP<sup>26</sup> interaction networks, and the annotation counts of those neighboring nodes.

The DIP ( $W_{D,i}$ ) and STRING ( $W_{S,i}$ ) scores were calculated as follows [Equation 8.1]:

$$W_{D,i} = \frac{1}{l_i} \sum_{j:P_j \in D_i} l_j \quad \text{and} \quad W_{S,i} = \frac{1}{l_i} \sum_{j:P_j \in S_i} l_j \quad (1)$$

where the sum of the normalized lengths ( $l$ ) from the set of all DIP/STRING interaction partners  $j$  with protein  $i$ , divided by the normalized length of protein  $i$ . Normalized length of a protein is the number of annotations divided by the maximum number of annotations in the database.

**8.2.2 Effect of DNAJA1 overexpression on pancreatic cell stress modulation.**<sup>§§</sup> MiaPaCa2 cells were obtained from the American Type Culture Collection (Rockville, MD). Cells were cultured as previously described.<sup>33</sup> Briefly, MiaPaCa2 cells were maintained in Dulbecco's Modified Eagle's Medium (Life Technologies, Inc.) supplemented with 10% heat-inactivated fetal bovine serum (FBS), nonessential amino acids, sodium pyruvate, and penicillin/streptomycin in 37°C incubator with 5% CO<sub>2</sub>. To stably express Full length His-tagged DNAJA1 construct, retroviral transductions were done essentially as described previously.<sup>34</sup>

Cell lysates were prepared by scraping cells (80-90% confluent) into lysis buffer [50 mM Tris-HCl (pH 7.5), 0.15 M NaCl, 1% Triton x-100 (v/v), 1% sodium deoxycholate (w/v), and 0.1% SDS (w/v) 5 mM EDTA and 1 mM phenylmethylsulphonyl fluoride]. Lysates were incubated, on ice, for 30 min. and centrifuged at 4°C for 15 minutes at 13,000 rpm to remove cell debris. Supernatants were transferred to fresh tubes and protein content was determined using the Bradford protein

---

<sup>§§</sup> The evaluation of the effect of DNAJA1 overexpression on pancreatic cell stress modulation was performed by the lab of Dr. Pankaj Singh at the University of Nebraska Medical Center.



assay reagent (Bio-Rad) with various concentrations of bovine serum albumin as standards. Cell lysates were stored at -20 or -80°C. Cell lysate proteins were resolved on 10% or 12% Tris-Glycine denaturing polyacrylamide gels in a 1x SDS-PAGE buffer (1g/liter SDS, 3g/liter Tris base, and 14.4g/liter glycine). Western blotting was performed as previously described.<sup>35,36</sup>

Cell survival was evaluated by MTT assay as described elsewhere.<sup>37</sup> Cells were plated in triplicate in 96-well plates at 5,000 cells per well and incubated at 37°C. 12 hours later, cells were treated with anisomycin or solvent control and incubated for 24 hours. At the end point, the culture medium was removed and 20 µl of MTT solution (5mg/ml, Sigma) was added per well, followed by a 2 hour incubation. MTT was removed and 200 µl of DMSO were added to each well to dissolve formazan. Formazan optical density was determined by utilizing a microplate reader at a wavelength of 540nm.

**8.2.3 Solution structure of the DNAJA1 J-domain.** The full human DNAJA1 protein (397 amino acids) has been targeted by the Northeast Structural Genomics Consortium (NESG; <http://www.nesg.org>) for structural elucidation as HR3099 (UniProt ID: P31689) [Figure 8.1]. The J-domain of DNAJA1 (DNAJA1-JD; NESG ID: HR3099K) was selected for structural determination by NMR due to its high DIP and STRING scores, small size (67 amino acids) and proposed importance for binding to DnaK (Hsp70), an important heat shock protein involved in stress response and cancer.<sup>38-</sup>

42

The NESG provided uniformly labeled <sup>13</sup>C, <sup>15</sup>N-enriched DNAJA1-JD (77 amino acids with 10 non-native residues MGHHHHHHSH at the N-terminus for purification).

The protein construct containing the sequence for DNAJA1-JD was transformed into BL21 (DE3) + Magic cells. The soluble fraction of the lysed cells was collected and purified with a Ni-NTA affinity column (Qiagen) and gel filtration column (HiLoad 26/60 Superdex 75 pg, Amersham Biosciences) chromatography. The NMR protein sample was stored in a sealed Shigemi tube with 20 mM 2-(4-morpholino)ethanesulfonic acid (MES; Sigma-Aldrich) buffer, pH 6.5 (uncorrected) with 10% D<sub>2</sub>O (Isotec), 0.02% NaN<sub>3</sub> (Sigma-Aldrich), 10 mM dithiothreitol (DTT, Sigma-Aldrich), 5 mM CaCl<sub>2</sub> (Sigma-Aldrich), 100 mM NaCl (Sigma-Aldrich), and 50  $\mu$ M 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS; Sigma-Aldrich).

MVKETTYDVLGVKPNATQEELKKAYRKLALKYHPDKNPNEGEKFKQIS  
QAYEVLSDAKKRELYDKGGEQAIKEGGAGGGFGSPMDIFDMFFGGGGRM  
QRERRGKNVVHQLSVTLEDLYNGATRKLALQKNVICDKCEGRGGKKGAV  
ECCPNCRGTMQIRIHQIGPGMVQOIQSVCMECQGHGERISPKDRCKSC  
NGRKIVREKKILEVHIDKGMKDGQKITFHGEGDQEPGLEPGDIIIVLDQ  
KDHAVFTRRGEDLFMCMDIQLVEALCGFQKPISTLDNRTIVITSHPGQI  
VKHGDIKCVLNEGMPYIYRRPYEKGRLIIEFKVNFPENGFLSPDKLSLLE  
KLLPERKEVEETDEMDQVELVDFDPNQERRRHYNGEAYEDDEHHPRGGV  
QCQTS

**Figure 8.1** Protein sequence of DNAJA1 (UniProtKB ID: P31689). The red residues indicate the 67 amino acids of the J-domain (A1-JD).

NMR experiments used for the protein backbone and sidechain assignments of DNAJA1-JD were collected at 298 K on a 600 MHz Bruker Avance spectrometer equipped with a 5 mm TXI probe. The backbone and sidechain assignments were completed using the standard and manual triple-resonance approach<sup>43-45</sup> using the following NMR experiments: two-dimensional (2D)-<sup>1</sup>H,<sup>15</sup>N-HSQC; 2D-<sup>1</sup>H,<sup>13</sup>C-HSQC;

and three-dimensional (3D) HNCO; HN(CA)CO; HNCA; HN(CO)CA; CBCA(CO)NH; CBCANH; HNHA; HBHA(CO)NH; CC(CO)NH; HCC(CO)NH; H(CC)H-COSY; and H(CC)H-TOCSY experiments.  $^{15}\text{N}$ -edited NOESY-HSQC and  $^{13}\text{C}$ -edited NOESY-HSQC experiments were collected on a 500 MHz Bruker Avance spectrometer equipped with a triple-resonance, Z-axis gradient cryoprobe to identify nuclear Overhauser effects (NOEs). Amide hydrogen exchange rates were also evaluated on the 500 MHz Bruker Avance spectrometer using the (CLEANEX-PM)-FHSQC experiment.

The NMR experimental data was processed using NMRPipe<sup>46</sup> and evaluated in CCPNMR Analysis (<http://www.ccpn.ac.uk>).<sup>47</sup> An initial homology model of DNAJA1-JD was generated from the protein backbone resonances using CS-ROSETTA on the WeNMR GRID-enabled web portal ([www.enmr.eu/webportal](http://www.enmr.eu/webportal)).<sup>48-50</sup>

The initial model of DNAJA1-JD was refined with XPLOR-NIH 2.31<sup>51</sup> using the following experimental restraints: NOE distance restraints (1,070 restraints), H-bond distance and angle restraints (50 restraints),  $^3J_{\text{NH}\alpha}$  coupling constant restraints (38 restraints),  $^{13}\text{C}\alpha/^{13}\text{C}\beta$  chemical shift restraints (127 restraints), and dihedral angle restraints (116 restraints) predicted from TALOS+.<sup>52</sup> All peptide bonds were constrained to be planar and *trans*. A total of 400 structures were calculated with the 20 lowest energy structures being subjected to explicit water refinement based on the RECOORD protocols.<sup>53</sup> An average DNAJA1-JD structure was calculated based on the average atom coordinates of the 20 water-refined structures and subsequently minimized using the same explicit water refinement above.

The resulting structures were evaluated using the PSVS software suite,<sup>54</sup> which includes Verify3D,<sup>55</sup> ProsaII,<sup>56</sup> PROCHECK,<sup>57</sup> and Molprobit.<sup>58</sup> The three-dimensional

structures of the proteins are represented here using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (<http://www.cgl.ucsf.edu/chimera>). ClustalW<sup>59</sup> was used (with default settings) to align the sequences of DNAJA1-JD with the J-domains of four homologous proteins: *E. coli* DnaJ J-domain (PDB ID: 1XBL),<sup>60</sup> *H. sapiens* Hsp40 (HDJ-1) J-domain (PDB ID: 1HDJ),<sup>61</sup> *H. sapiens* HSJ1a (PDB ID: 2LGW),<sup>62</sup> and *H. sapiens* DnaJ subfamily C member 12 (PDB ID: 2CTQ). Electrostatic surface potentials of the protein were calculated using Delphi.<sup>63</sup> The identification of evolutionarily conserved residues using both sequence and structure was performed using the ConSurf server with default settings.<sup>64</sup>

**8.2.4 Identification of a ligand binding site on the DNAJA1 J-domain.** To experimentally determine potential small molecule binding sites, a high-throughput NMR ligand affinity screen using the FAST-NMR compound library (described in Chapter 2 and elsewhere)<sup>65,66</sup> was performed. The 1D line broadening ligand-based screen and 2D <sup>1</sup>H, <sup>15</sup>N-HSQC protein-based screen follows the same procedure outlined in Chapter 2 using 10  $\mu$ M and 30  $\mu$ M DNAJA1-JD (5% <sup>13</sup>C, 100% <sup>15</sup>N-labeled; provided by the NESG), respectively. The concentrations for the compounds were 100  $\mu$ M in the 1D line-broadening screen and 400  $\mu$ M in the 2D <sup>1</sup>H, <sup>15</sup>N-HSQC screen.

The NMR spectra for the ligand affinity screens were collected on a 500 MHz Bruker Avance spectrometer equipped with a triple-resonance, Z-axis gradient cryoprobe with a Bruker BACS-120 sample changer. All 1D <sup>1</sup>H NMR spectra were processed with the ACD/NMR Processor (ACD/Labs) and 2D <sup>1</sup>H, <sup>15</sup>N-HSQC spectra were processed with NMRPipe<sup>46</sup> and visualized in CCPNMR Analysis.<sup>47</sup>

The chemical shift perturbations (CSPs) between the resonances of the free protein and ligand-bound protein 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra were used to define the consensus binding site using CSP-Consensus (described in Chapter 4). AutoDock 4.2.3<sup>67-69</sup> with AutoDockTools 1.5.4<sup>69,70</sup> (<http://mgltools.scripps.edu>) graphical interface was used to calculate 120 protein-ligand costructures, which were filtered using AutoDockFilter 2.0 (described in Chapter 3 and Chapter 4)<sup>71</sup> to identify the costructures that best agree with the experimental CSPs.

## 8.3 RESULTS AND DISCUSSION

**8.3.1 Selection of DNAJA1 from a pancreatic cancer ‘omics database.** The goal of the pancreatic cancer ‘omics database is to focus the search for potentially interesting therapeutic targets. The ‘omics database contains 5,336 proteins/genes, which represents approximately 26% of the human genome. Clearly, not all of the proteins identified in these proteomics and genomics studies are related to pancreatic cancer or therapeutically important, thus additional efforts are required to prioritize the most relevant proteins.

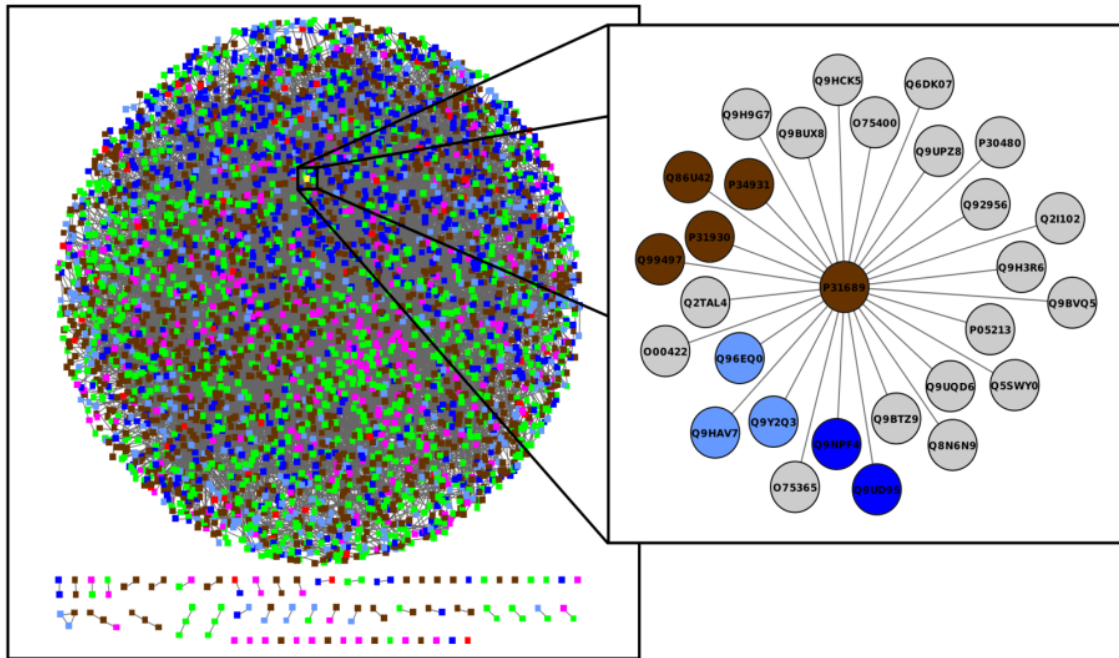
Within the ‘omics database, approximately 26% (1,194 proteins) have no known function and have been assigned as “putative”, “uncharacterized”, or “unknown” proteins. These uncharacterized proteins are potential therapeutic targets for pancreatic cancer, however determining the function of a protein is challenging. One approach to infer the function of these proteins is to leverage sequence and structural homology to proteins with known function. However, only 14% of the functionally uncharacterized proteins have a known structure, which hinders functional annotation by structural

comparisons. Using sequence similarity of the uncharacterized proteins to find homologous proteins with a structure, the number of proteins with an experimental or homologous structure rises to 34%. Unfortunately, these structures, while important, do not provide enough information to functionally annotate the proteins, determine its potential role in pancreatic cancer, and develop a therapy based on that role.

Instead of focusing on the uncharacterized proteins alone, another approach to prioritize the database would be to leverage the annotation and interaction information from several databases to develop a functional network. Functional pathways that are well represented in the 'omics database hint at the importance of the proteins in that pathway towards pancreatic cancer. Additionally, uncharacterized proteins that have an association with many highly annotated proteins would make for potentially interesting protein targets.

Based on the annotations from several databases (7,795 annotations), a functional network was generated that identified six broad functional classifications [Figure 8.2]: (1) DNA binding, transcription regulation, and transcription factors; (2) transmembrane signaling and transport activity; (3) chromatin, histone modification, and transcription regulation; (4) stress response and signaling; (5) translation and biosynthetic processes; and (6) mitosis and cytokinesis. Proteins within these clusters were then evaluated based on DIP ( $W_{D,i}$ ) and STRING ( $W_{S,i}$ ) scores, which identify those proteins that are associated with highly annotated functional pathways. Other factors, such as small size (< 200 amino acids) and likely solubility, were evaluated for amenability to NMR structure determination.

Based on the above criteria, the human protein DnaJ-homolog A member 1 (DNAJA1) was identified as having 25 annotations and DIP ( $W_{D,i}$ ) and STRING ( $W_{S,i}$ ) scores of 4.4 and 11.0, respectively. This indicates that DNAJA1 is associated with other well-annotated proteins in the 'omics database and is involved in a pathway that may be important in pancreatic cancer. In the pancreatic cancer 'omics database, DNAJA1 was shown to be down-regulated five-fold in pancreatic cancer cells relative to normal healthy cells and cells undergoing pancreatitis.<sup>12</sup>



**Figure 8.2** (Left) Functional network of the 5,336 proteins within the 'omics database. Proteins are color-coded based on functional clustering. Proteins involved in small isolated networks are shown below the primary network. (Right) Expanded view of the network for DNAJA1 (UniProtKB: P31689). Only the nearest neighbors are shown. Brown nodes indicate a general functional classification of mitosis and cytokinesis. Dark blue nodes indicate a general functional classification of chromatin, histone modification, and transcription regulation. Light blue nodes indicate a general functional classification of translation and biosynthetic processes. Grey nodes are proteins not present in the pancreatic cancer 'omics database, but have been shown to be associated with DNAJA1.

The protein DNAJA1 belongs to the family of proteins known as DnaJ proteins. The DnaJ proteins, also known as heat shock protein 40 (Hsp40 or Hsc40), are proteins originally identified in *E. coli* that act as co-chaperones to the molecular chaperone DnaK (Hsp70), which is responsible for several cellular processes such as rescuing misfolded proteins, folding polypeptide chains, transport of polypeptides through membranes, assembly and disassembly of protein complexes, and control of regulatory proteins.<sup>72-74</sup> DnaJ primarily facilitates the hydrolysis of ATP from DnaK which is necessary for the chaperone activity of DnaK.<sup>74-76</sup>

There are over 41 members of the DnaJ family encoded in the human genome, where little is known about their specific biological functions.<sup>73</sup> However, the specific DnaJ protein that binds DnaK appears to determine the activity of the complex.<sup>74</sup> Each specific DnaJ protein can be classified into three subfamilies (A, B, C), with subfamily A closely resembling the DnaJ protein of *E. coli*.<sup>77</sup> DnaJ subfamily A proteins typically consist of three distinct domains: (1) a highly conserved J-domain region of approximately 70 amino acids found near the N-terminus, which mediates the interaction with DnaK;<sup>76,78</sup> (2) A G/F-rich region acting as a flexible linker; and (3) a cysteine-rich region containing 4 motifs resembling a zinc-finger domain. DnaJ subfamily B proteins do not typically have the cysteine-rich region, while DnaJ subfamily C proteins only have the J-domain.<sup>74</sup>

In general, J-domain proteins modulate protein assembly, disassembly, and translocation.<sup>79</sup> DnaJ subfamily A member 1 (DNAJA1) human protein has been shown to associate on its own with unfolded polypeptide chains and prevent their aggregation.<sup>80</sup> It has also been shown to regulate androgen receptor signaling and spermatogenesis in

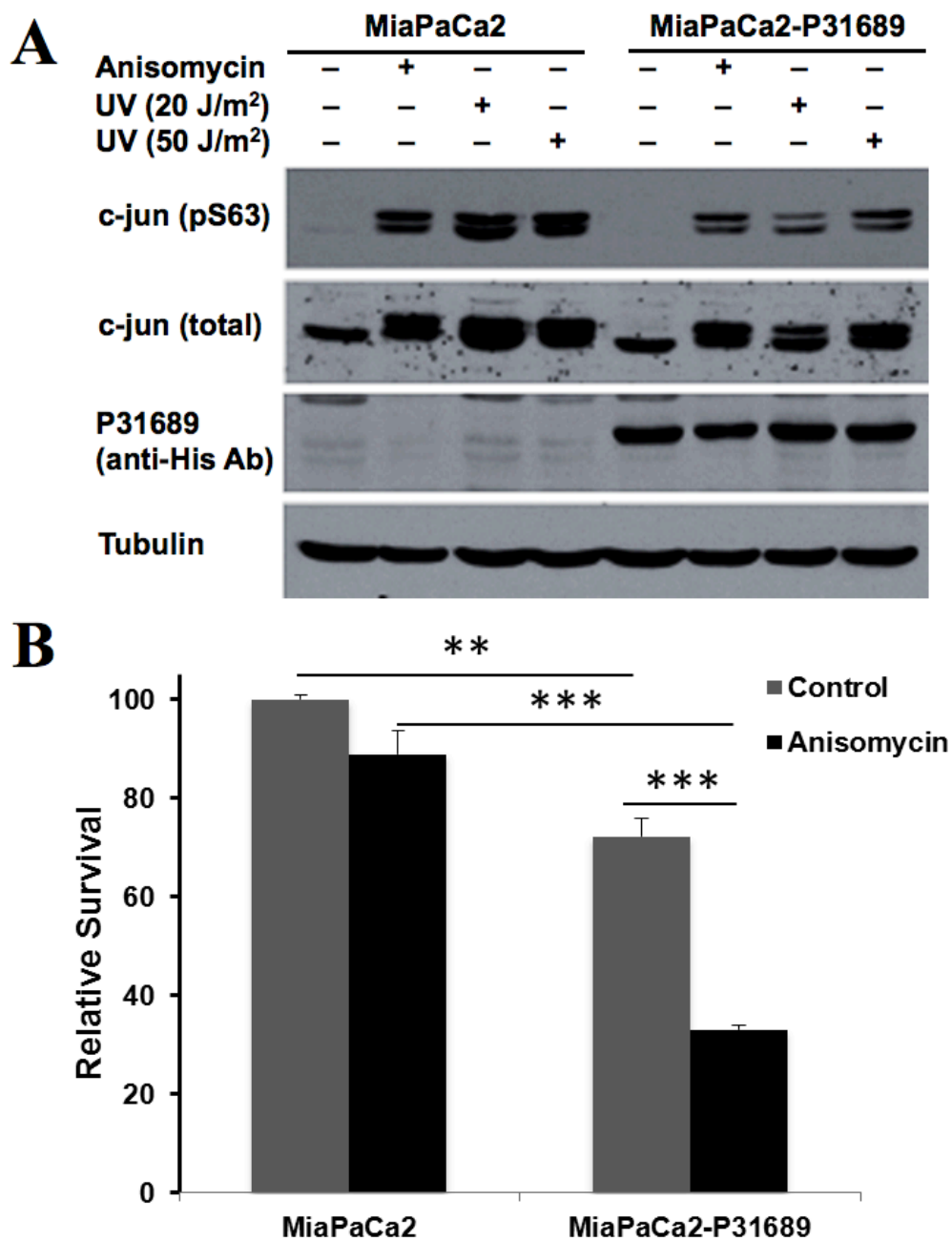


mice.<sup>77</sup> DNAJA1 has been reported to contribute to the resistance of glioblastomas to radiotherapy<sup>81</sup> and has also been targeted as a biomarker for pancreatic cancer in order to evaluate the effects of FPTase inhibitors.<sup>82,83</sup> Additionally, DNAJA1 appears to be involved in importing proteins into the mitochondria.<sup>84,85</sup> Of course, the mitochondrial pathway to apoptosis protects against cancer and requires importing apoptotic factors into the mitochondrial membrane.<sup>86-89</sup> However, there have been no studies of whether the conserved J-domain of DnaJ alone has any role in cancer biology independent of DnaK.<sup>73</sup> Additionally, DNAJA1 is an interesting target based on its association with DnaK, which is expressed abundantly in various tumors and may even promote tumorigenesis by inhibiting cell death.<sup>40-42,90-92</sup>

### **8.3.2 Effect of DNAJA1 overexpression on pancreatic cell stress modulation.**

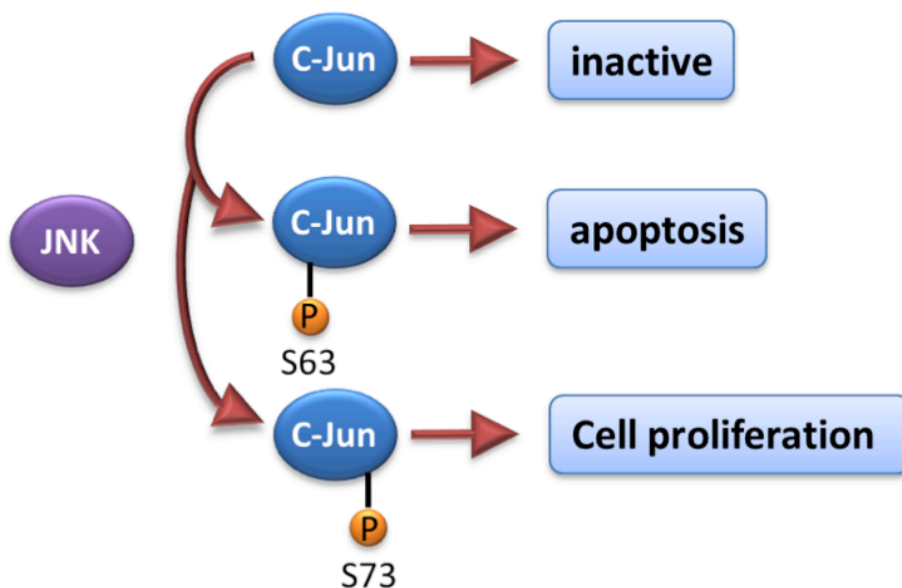
The pancreatic cancer 'omics database and functional networking approaches identified DNAJA1 as a potentially interesting target for investigation. However, it is vital importance that experimental evidence is used to verify that DNAJA1 is a relevant target in pancreatic cancer. Cell-based functional assays were performed in order to identify any cancer related properties. MiaPaCa2 cells stably expressing DNAJA1 or vector control were subjected to a 1-hour incubation with anisomycin (370nM), a protein synthesis inhibitor, or UV treatment at 20 J/m<sup>2</sup> or at 50 J/m<sup>2</sup> doses to mimic stress. Activation of stress-induced JNK pathway was measured by evaluating the downstream phosphorylation of c-jun at 20 min post-treatment. The results indicate that overexpression of DNAJA1 diminishes anisomycin and UV-induced c-jun phosphorylation at S63 [Figure 8.3A]. Cell survival was also evaluated by performing an MTT assay 24 hours after treatment with anisomycin. The results indicate that DNAJA1

expression decreased cell survival under conditions of anisomycin treatment [Figure 8.3B]. Expression of the DNAJA1 protein in MiaPaCa2 pancreatic cancer cell lines resulted in the cells being more susceptible to stress-induced (UV-induced DNA damage or anisomycin-induced inhibition of protein synthesis) apoptosis [Figure 8.3]. This effectively suppresses the phosphorylation-mediated activation of the oncogenic transcription factor, c-Jun, which is often found overexpressed and hyperphosphorylated in cancer<sup>93-95</sup> and has an essential role in pancreatic cancer.<sup>96,97</sup> c-Jun is a member of the JNK signaling pathway, and its transcriptional activity and expression is primarily regulated by the phosphorylation of two serines, Ser63 and Ser73.<sup>98-101</sup> c-Jun regulates a range of cellular processes including apoptosis, tumorigenesis, and cell proliferation, which includes protecting cells from induced cell death.<sup>100,101</sup> This makes c-Jun both a positive and negative regulator of cell death [Figure 8.4A]. However, in several cancers, c-Jun has been shown to inhibit apoptosis, leading to the uncontrolled growth typical of cancer.<sup>102-104</sup>



**Figure 8.3** (A) The expression of His-tagged DNAJA1 in MiaPaCa2 pancreatic cancer cells suppresses the activation of c-Jun in response to anisomycin (370 nM) or UV treatment. 20 min-post treatment cells were subjected to lysis and the levels of phosphor c-jun (S63) and total c-jun were evaluated by western blotting. Expression levels of exogenously expressed DNAJA1 were evaluated by immunoblotting with anti-His antibody, while immunoblotting with anti-tubulin antibody was performed as a loading control (B) The expression of DNAJA1 in MiaPaCa2 pancreatic cancer cells decreases the survival in response to anisomycin treatment-induced stress. The cell survival was measured by MTT assay 24 hour post-treatment. (\*\* indicates  $p < 0.01$ ; \*\*\* indicates  $p < 0.001$ )

Interestingly, heat-shock proteins have also been shown to be regulators of apoptosis, where DnaK (Hsp70) suppresses JNK activity.<sup>105-107</sup> Thus our preliminary results suggest DNAJA1 stimulates the DnaK suppression of a JNK-induced anti-apoptotic signaling pathway by forming a complex with DnaK [Figure 8.4B]. This hypothesis is consistent with the five-fold down-regulation of DNAJA1 in pancreatic cancer cells, the resulting suppression of c-Jun phosphorylation, and the corresponding susceptibility to stress-induced apoptosis by expressing DNAJA1. Obviously, protection from stress-induced apoptosis by down-regulating DNAJA1 would be beneficial since cancer cells, by definition, exist in a stressful environment.

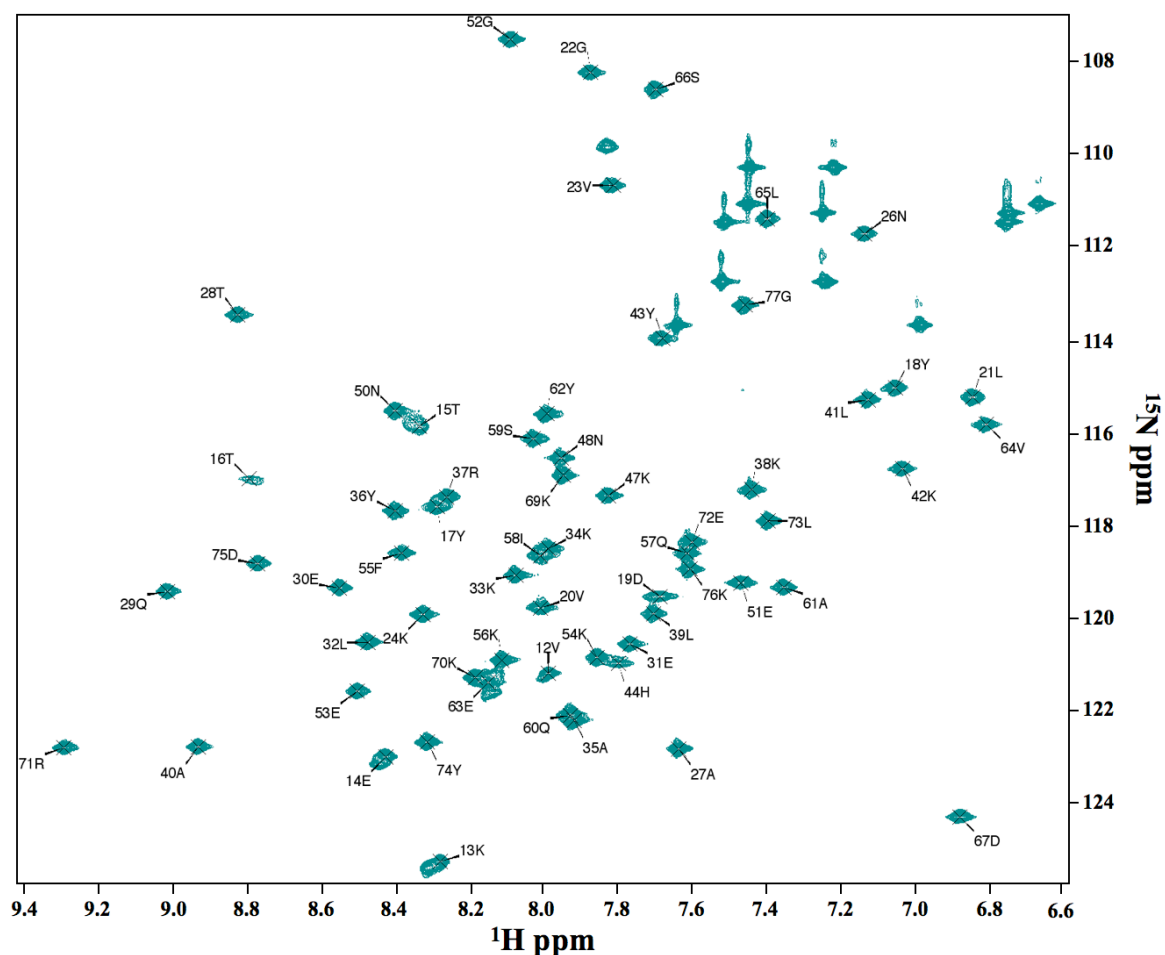


**Figure 8.4** An illustration of the role of c-Jun in the JNK pathway under different phosphorylation states.

**8.3.3 Solution structure of the DNAJA1 J-domain.** The backbone resonance assignments were completed using the 2D and 3D NMR experiments described above

(2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC, HNCO, HN(CA)CO, HNCA, HN(CO)CA, CBCANH, CBCA(CO)NH, and HNHA). This resulted in 85.7% of the 77 amino acids in the protein unambiguously assigned in the  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC. When excluding the non-native 10-residue N-terminal His-tag used for purification and the prolines for which no amide exists, the assignment improves to 98.4% (63/64). The one amino acid for which the amide could not be assigned was Met11, which is the first residue following the His-tag. The side chain assignments were completed using a combination of the CC(CO)NH, HCC(CO)NH, H(CC)H-COSY, and H(CC)-TOCSY 3D NMR experiments. The backbone and sidechain assignments, not including 10-residue His-tag, were nearly complete with 63/67 N, 63/64 HN, 67/67 C $\alpha$ , 70/70 H $\alpha$ , 63/64 C $\beta$ , 113/115 C $\beta$ , 43/62 C $\gamma$ , 66/74 H $\gamma$ , 26/48 C $\delta$ , 40/54 H $\delta$ , 11/21 C $\epsilon$ , 14/34 H $\epsilon$ , 0/9 C $\zeta$ , and 0/1 H $\zeta$  atoms assigned.

Using the backbone resonance assignments, a homology structure of DNAJA1-JD was generated using CS-ROSETTA, which utilizes chemical shifts to select protein fragments from the PDB followed by Monte Carlo assembly and relaxation by Rosetta. This tool has been shown to be effective in predicting protein structures for small proteins up to 16 kDa. The homology model generated from CS-ROSETTA exhibited the same secondary structure as most DnaJ proteins and agreed with the secondary structure predicted from TALOS. All available backbone and side chain chemical shift assignments will be deposited into the Biological Magnetic Resonance Data Bank (BMRB; <http://www.brm.b.wisc.edu>).



**Figure 8.5** Complete backbone of  $^1\text{H}$  and  $^{15}\text{N}$  assignments of human DNAJA1-JD in a 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectrum.

The solution structure of DNAJA1-JD was calculated using 1,120 distance restraints, 116 dihedral restraints, 127  $\text{Ca}/\text{C}\beta$  carbon chemical shift restraints, and 38  $^3\text{J}_{\text{NH}\alpha}$  coupling constant restraints. The restraints used during the structure calculation are reported in [Table 8.1]. XPLOR-NIH was used to calculate 400 structures, and the 20 lowest energy structures were selected for further refinement in water using the RECOORD protocol implemented in XPLOR-NIH. The resulting ensemble and average structures [Figure 8.6] agreed well with the NMR data, where the experimental restraints had low rms deviations [Table 8.1]. The water-refined average structure had no NOE

violations greater than 0.5 Å or dihedral violations greater than 5°. The water-refined ensemble of 20 structures had a backbone RMSD of  $0.709 \pm 0.119$  Å to the unrefined average coordinates. This result improves to  $0.399 \pm 0.086$  Å when only the residues involved in the more stable secondary structure elements are evaluated, which indicates the consistency of the structure calculation using the experimental restraints [Table 8.2]. In addition, comparing the original CS-ROSETTA homology model to the final water-refined average structure showed a backbone RMSD of 1.803 Å (full protein) and 0.889 Å (secondary structure).

The PSVS software suite was used to verify the quality of the ensemble and average structures [Table 8.3]. Few unreasonable atom clashes were identified by the Molprobity module, and overall has very good Z-score (-1.59) compared to the average Z-score for NMR structures in the RCSB PDB (-10.74).<sup>54</sup> This is actually comparable to the Z-scores of medium resolution protein structures found in the RCSB PDB (-1.39).<sup>54</sup> An evaluation of the probable dihedral angles expected for each residue using PROCHECK also shows impressive Z-scores (1.65 for  $\Phi/\psi$  dihedrals and 1.48 for all dihedrals). Additionally, the 98.5% of the residues, for both the ensemble and average structures, fell within the most favored regions of Ramachandran space, where only one residue, Val12, existed in just the allowed region. Overall, the results indicate that the final ensemble and average models are good structures with little to no unreasonable structural features. The coordinates of the water-refined ensemble and the water-refined average structure will be deposited in the RCSB PDB.

**Table 8.1 Structure calculation statistics<sup>a</sup>**

	$\langle \text{SA} \rangle$	$(\overline{\text{SA}})_r$
RMSD distance restraints (experimental) (Å)		
All (1120)	$0.083 \pm 0.001$	0.082
Inter-residue sequential ( $ i-j  = 1$ ) (269)	$0.082 \pm 0.003$	0.086
Inter-residue short range ( $1 <  i-j  < 5$ ) (221)	$0.078 \pm 0.005$	0.073
Inter-residue long range ( $ i-j  \geq 5$ ) (80)	$0.108 \pm 0.008$	0.098
Intra-residue (500)	$0.084 \pm 0.002$	0.085
H-bonds (50)	$0.038 \pm 0.008$	0.029
RMSD Dihedral angle restraints (°) (116)	$0.027 \pm 0.044$	0.00
RMSD C $\alpha$ and C $\beta$ shifts restraints (ppm) (127)	$0.873 \pm 0.036$	0.907
RMSD $^3J_{\text{NH}\alpha}$ restraints (Hz) (38)	$0.536 \pm 0.038$	0.540
RMSD (covalent geometry)		
Bonds (Å)	$0.007 \pm 0.000$	0.007
Angles (°)	$0.649 \pm 0.015$	0.647
Impropers (°)	$0.794 \pm 0.037$	0.778
Energy (kcal/mol)		
Total	$-2620.08 \pm 70.18$	-2834.43
Bond	$37.44 \pm 2.35$	36.16
Angle	$94.07 \pm 4.69$	95.65
Dihedral	$0.02 \pm 0.04$	0.00
Impropers	$33.44 \pm 2.77$	31.56
van der Waals	$-262.40 \pm 10.76$	-273.46
NOE	$230.11 \pm 7.09$	226.01
$^3J_{\text{NH}\alpha}$	$10.99 \pm 1.60$	11.08
C $\alpha$ and C $\beta$ shifts	$49.16 \pm 4.30$	52.92

<sup>a</sup>  $\langle \text{SA} \rangle$  represents the final 20 water refined simulated annealing structures.  $(\overline{\text{SA}})_r$  represents the water refined average structure of all 20 water-refined structures.



**Table 8.2 Atomic rms differences<sup>b</sup>**

	Full protein (residues 11-77)		Secondary Structure	
	Backbone atoms	All heavy atoms	Backbone atoms	All heavy atoms
$\langle SA \rangle$ vs $\overline{SA}$	$0.709 \pm 0.119$	$1.417 \pm 0.110$	$0.399 \pm 0.086$	$1.232 \pm 0.134$
$\langle SA \rangle$ vs $(\overline{SA})_r$	$0.888 \pm 0.136$	$1.739 \pm 0.139$	$0.639 \pm 0.155$	$1.605 \pm 0.131$
$(\overline{SA})_r$ vs $\overline{SA}$	0.649	1.217	0.523	1.183

<sup>b</sup>  $\langle SA \rangle$  represents the final 20 water-refined simulated annealing structures.  $\overline{SA}$  represents the average structure of all 20 water-refined structures.  $(\overline{SA})_r$  represents the water-refined average structure of all 20 water-refined structures.

**Table 8.3 Structure evaluation<sup>c</sup>**

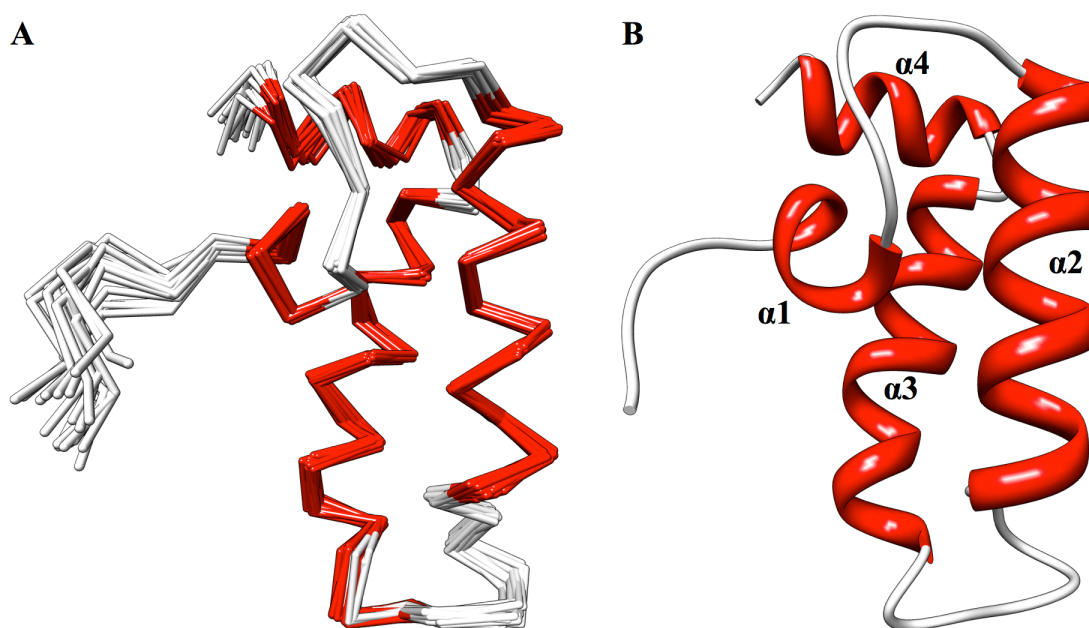
	$\langle SA \rangle$	$(\overline{SA})_r$
PSVS Z-scores <sup>d</sup>		
Verify3D	-0.80	-0.16
ProsaII (-ve)	2.40	2.52
Procheck (phi-psi)	1.65	1.46
Procheck (all)	1.48	1.54
MolProbity clash score	-1.59	-1.56
Ramachandran space <sup>e</sup>		
Most favored regions	98.5 %	98.5 %
Allowed regions	1.5 %	1.5 %
Disallowed regions	0.0 %	0.0 %

<sup>c</sup>  $\langle SA \rangle$  represents the final 20 water-refined simulated annealing structures.  $(\overline{SA})_r$  represents the water-refined average structure of all 20 water-refined structures.

<sup>d</sup> Calculated with PSVS (more positive scores are better)

<sup>e</sup> Calculated with Molprobity module in PSVS

The secondary structure and fold for DNAJA1-JD are characteristic for the other J-domains found in DnaJ homologs in most species. The structure consists of four  $\alpha$ -helices: residues 17-21 ( $\alpha 1$ ); 29-42 ( $\alpha 2$ ); 52-65 ( $\alpha 3$ ); and 68-75 ( $\alpha 4$ ) [Figure 8.6]. The loop between  $\alpha 2$  and  $\alpha 3$  (residues 43-51) contains the highly conserved His-Pro-Asp (HPD) motif (residues 44-46).



**Figure 8.6** (A) An overlay of the backbone trace of the 20 lowest energy, water-refined structures. (B) A ribbon representation of the average structure generated from the average atomic coordinates of the 20 lowest energy, water-refined structures, followed by water refinement of the average structure. Both structures are colored according to secondary structure:  $\alpha$ -helix (red) and loop (white).

When DNAJA1-JD was selected for structural work, there was no example of a DnaJ subfamily A member 1 protein in the RCSB PDB. However, since that time, a solution structure of the J-domain of human DnaJ subfamily A member 1 protein has been deposited in the PDB (PDB ID: 2LO1), which is the same protein (albeit with four more residues) as DNAJA1-JD. A comparison between 2LO1 and DNAJA1-JD show

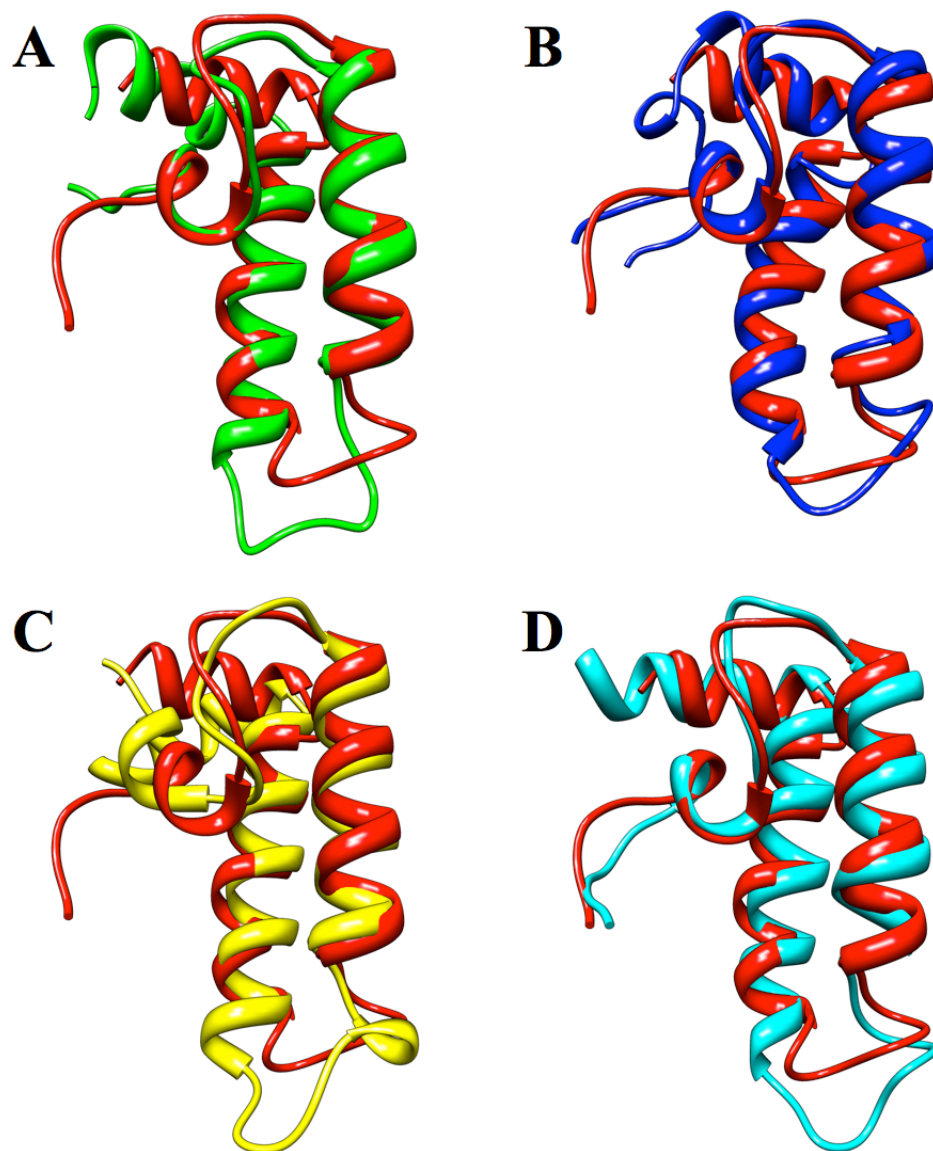
significant agreement between the two structures, with a backbone RMSD of 1.751 Å (all residues) and 1.088 Å (secondary structure). While the two structures are fairly similar, evaluation of the structure of 2LO1 using PSVS indicates some differences from DNAJA1-JD. The Molprobability clash Z-score for 2LO1 (3.32) is significantly better than the score obtained from DNAJA1-JD (-1.56). Also, the Verify3D Z-score for 2LO1 (0.40) is better than DNAJA1-JD (-0.80). However, the Z-score results of ProsaII (1.18), Procheck (phi-psi: -0.24), Procheck (all dihedral: -0.44), and Ramachandran space (6.9% allowed and 1.1% disallowed) for 2LO1 are significantly worse than that seen in DNAJA1-JD. These results indicate that 2LO1 has minimized the steric clashes, probably with a slightly more extended structure [Figure 8.7], at the expense of better stereochemistry and agreement with known protein folds.



**Figure 8.7** Overlay of the ribbon structures for DNAJA1-JD (red) compared to the previously solved structure, 2LO1 (blue).

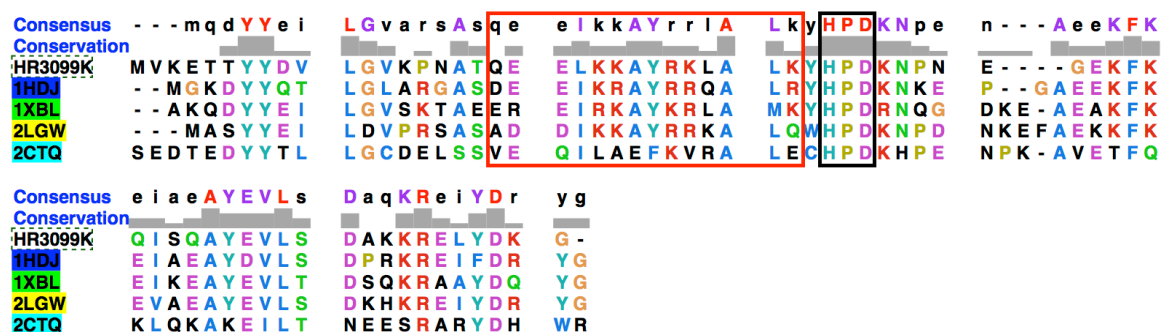
The structures of 28 DnaJ proteins in various organisms have been solved. Most of these proteins (16 structures) are from humans, with the majority belonging to DnaJ subfamily B (6 structures) and DnaJ subfamily C (8 structures). The tertiary structure of DNAJA1-JD was compared to a few representative structures: *E. coli* DnaJ J-domain (PDB: 1XBL); human DnaJ homolog subfamily B member 1 J-domain (PDB: 1HDJ); human DnaJ homolog subfamily B member 2 J-domain (PDB: 2LGW); and human DnaJ homolog subfamily C member 12 J-domain (PDB: 2CTQ). All four proteins have essentially the same tertiary structure as DNAJA1-JD with PDBeFold Z-scores of 5.148 (2.08 Å RMSD), 6.872 (1.65 Å RMSD), 4.541 (2.24 Å RMSD), and 6.904 (1.37 Å RMSD), respectively [Figure 8.8]. The different DnaJ homolog subfamilies do not appear

to have a significant difference in structure especially since the best matched structure to DNAJA1 belongs to DnaJ homolog subfamily C. The true difference in these subfamilies likely stems from differences in the other regions of the DnaJ proteins.

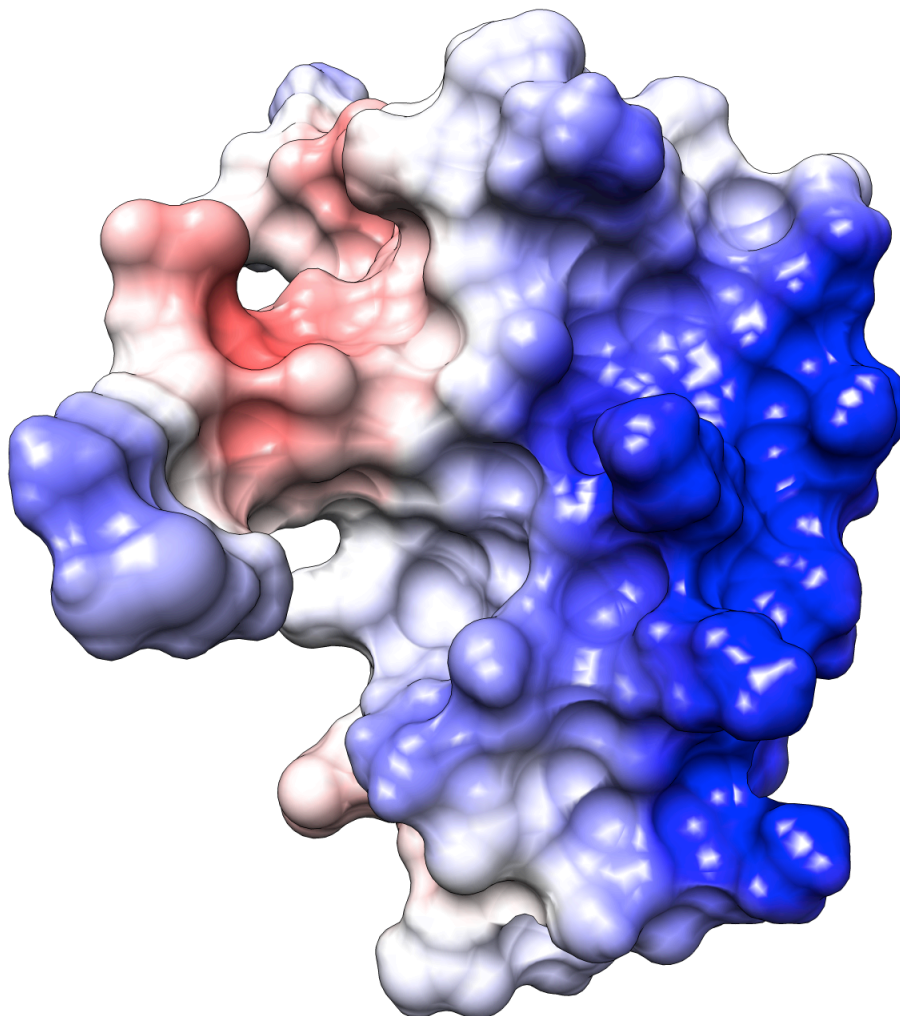


**Figure 8.8** Overlay of the ribbon structure for DNAJA1-JD (red) with (A) *E. coli* DnaJ J-domain (PDB: 1XBL), (B) *H. sapiens* DnaJ homolog subfamily B member 1 J-domain (PDB:1HDJ), (C) *H. sapiens* DnaJ homolog subfamily B member 2 (PDB: 2LGW), and (D) *H. sapiens* DnaJ homolog subfamily C member 12 (PDB: 2CTQ).

The sequence of human DNAJA1-JD is also very well conserved with J-domains from other organisms, with 56 proteins having  $\geq 49\%$  sequence identity. The sequences of the representative proteins 1XBL, 1HDJ, 2LGW, and 2CTQ have sequence identities of 51%, 56%, 47%, and 32%, respectively [Figure 8.9]. As expected, all five proteins have the highly conserved HPD motif present. Additionally, the  $\alpha 2$ -helix is highly conserved for all of the proteins except for 2CTQ. This is interesting as the  $\alpha 2$ -helix tends to be positively charged and represents a possible binding spot for DnaK (Hsp70) [Figure 8.10].



**Figure 8.9** ClustalW comparison of DNAJA1-JD (HR3099K) with 1HDJ (blue), 1XBL (green), 2LGW (yellow), and 2CTQ (cyan). The highly conserved HPD sequence is outlined in a black box. The residues that make up the  $\alpha 2$ -helix are outlined with a red box.

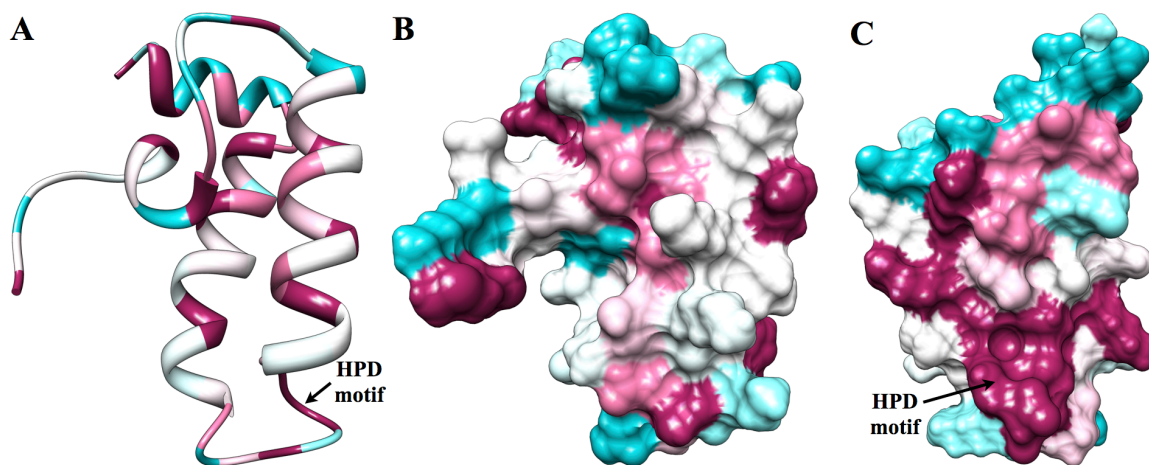


**Figure 8.10** An electrostatic surface representation of DNAJA1-JD (positively charged: blue and negatively charged: red). The large positively charged region corresponds to the residues of the  $\alpha 2$ -helix.

**8.3.4 Identification of a ligand binding site on the DNAJA1 J-domain.** One of the primary functions for DnaJ is to stimulate the ATPase activity of DnaK. If DNAJA1 were to act as a potential therapeutic target, understanding the binding interactions that can be influenced would help in the potential development of drug. The primary function of the J-domain of DnaJ is to bind to the ATPase domain on DnaK. As previously



mentioned, the main feature of the J-domain proteins is a highly conserved HPD motif [Figure 8.11], which may indicate its importance in binding to DnaK.



**Figure 8.11** (A) A ribbon representation of the conserved residues of DNAJA1-JD. (B) A surface representation of the conserved residues of DNAJA1-JD. (C) A 90° rotation of the surface representation of the conserved residues of DNAJA1-JD, which highlights the highly conserved HPD motif. The highly conserved residues (magenta) and poorly conserved residues (cyan) were calculated with Consurf.

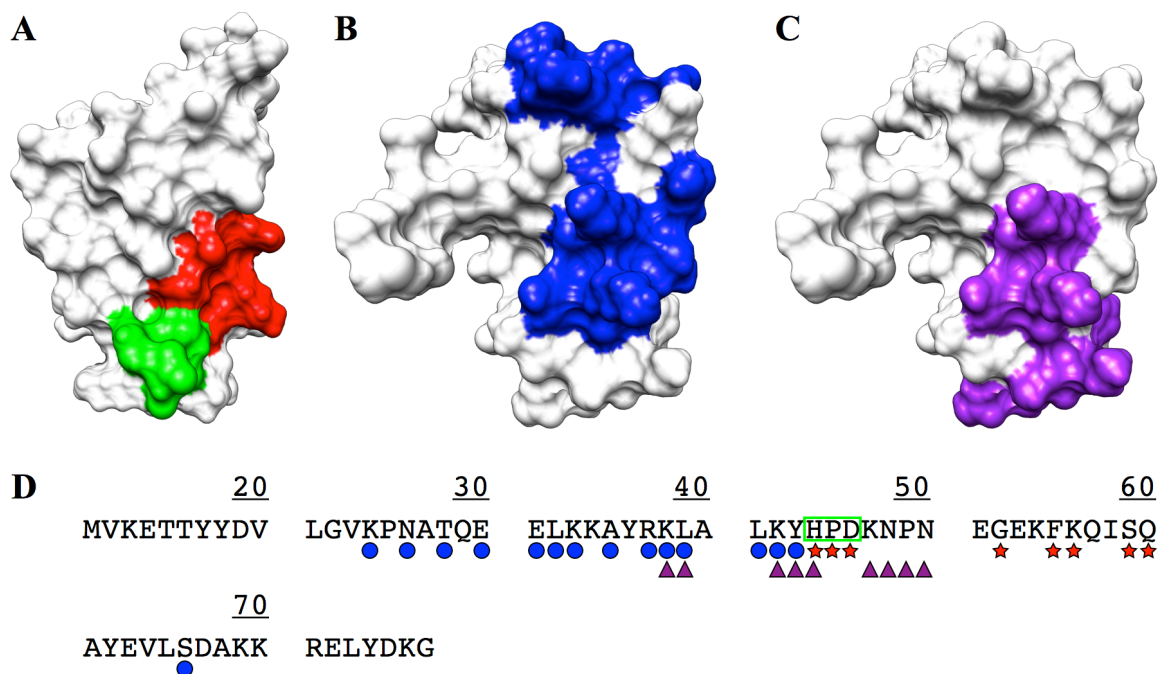
The RCSB PDB contains only one example of a J-domain in complex with DnaK. This example (PDB ID: 2QWN) is a crystal structure of the bovine auxilin (DnaJ homolog subfamily C) J-domain chemically cross-linked with bovine DnaK at the conserved HPD motif.<sup>72,108</sup> A sequence alignment of the residues between the bovine auxilin J-domain and DNAJA1-JD allows for a prediction of the proposed DnaJ-DnaK interaction site [Figure 8.12A]. This proposed binding site includes the highly conserved HPD motif. However, there is some contention as to whether the cross-linked complex is biologically relevant or whether auxilin accurately represents most DnaJ interactions with DnaK.<sup>108-110</sup>



A previous analysis observed the chemical shift perturbations in 2D  $^1\text{H}$ ,  $^{15}\text{N}$  HSQC experiments when the *E. coli* DnaJ J-domain is bound to *E. coli* DnaK.<sup>111</sup> The majority of the perturbed residues in *E. coli* DnaJ J-domain occurred along the  $\alpha$ 2-helix. An alignment of the sequence between *E. coli* DnaJ J-domain and human DNAJA1-JD allows for the mapping of those same perturbed residues [Figure 8.12B], which also indicates that the binding site is along the  $\alpha$ 2-helix, not the HPD motif. The  $\alpha$ 2-helix is intriguing as a binding site because of its positively charged surface [Figure 8.10]. Additionally, the proposed binding site on DnaK has a negatively charged surface, which supports the possibility of the  $\alpha$ 2-helix as the binding site for DnaK. Additionally, mutations of residues in the  $\alpha$ 2-helix inhibit the DnaJ-DnaK interaction.<sup>110,111</sup> This supports the hypothesis that the  $\alpha$ 2-helix on DNAJA1 may represent the likely DnaK binding site.

Another interaction site on DNAJA1-JD may be related to the inhibition of DnaJ activity. A related DnaJ protein, TIM14, is essential for the transport of proteins across the outer membrane of mitochondria by stimulating ATPase activity of mitochondrial Hsp70.<sup>112</sup> Any mutation in the HPD motif of TIM14 effectively inhibits its activity, which indicates the importance of the HPD motif in the function of DnaJ proteins. Additionally, when TIM14 is complexed with TIM16, another J-domain like protein but without the HPD motif, TIM14 activity is inhibited. The location of the TIM14-TIM16 interaction site partially overlaps with the proposed DnaK binding site that includes residues in the  $\alpha$ 2-helix.<sup>113</sup> A sequence alignment of TIM14 with DNAJA1-JD indicates an essentially identical overlap [Figure 8.12C], where 4 of the 15 perturbed residues that

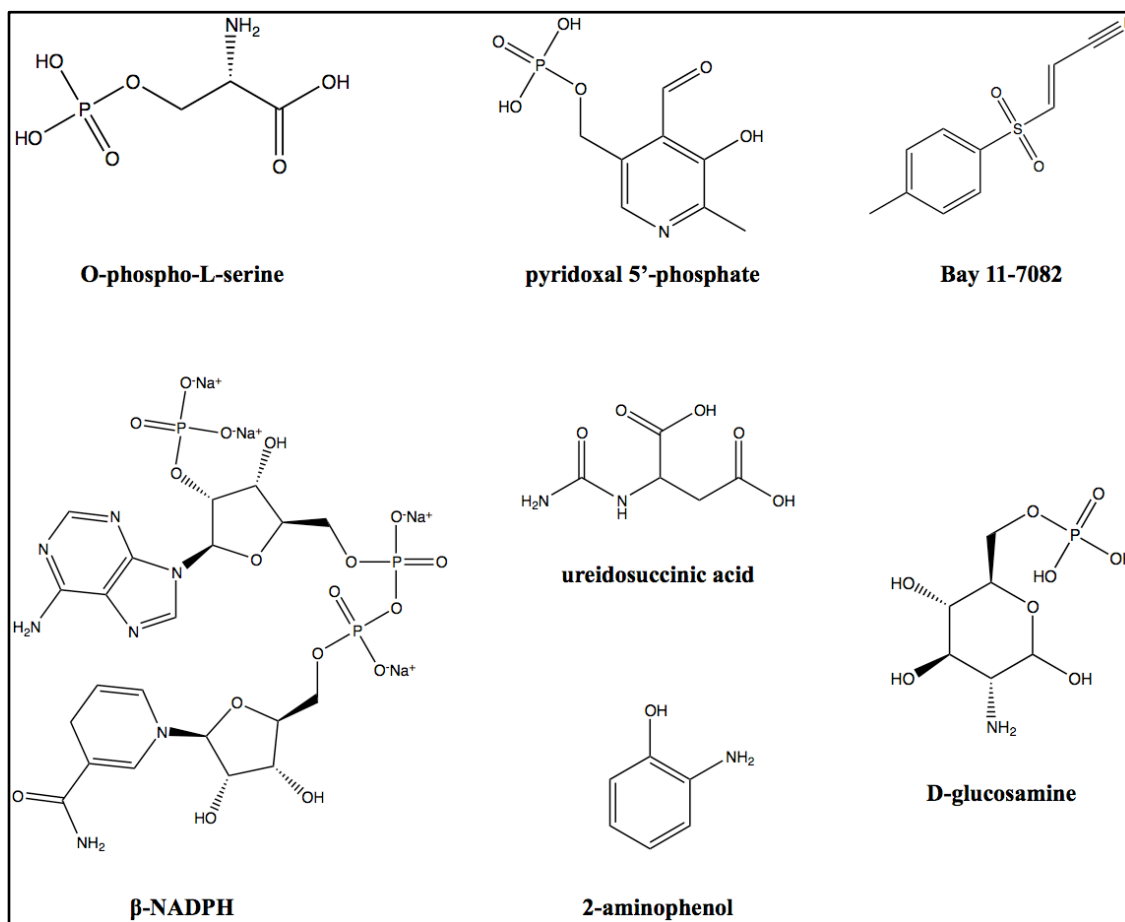
make up the  $\alpha$ 2-helix proposed binding site are also part of the inhibition site [Figure 8.12D].



**Figure 8.12** (A) A surface representation of DNAJA1-JD highlighting another proposed DnaK binding site based upon the bovine auxilin-bovine Hsp70 complex (PDB ID: 2QWN) colored in red and green (conserved HPD motif). (B) A surface representation of DNAJA1-JD (rotated  $\sim 90^\circ$ ) with the proposed DnaK binding site based upon NMR titration data colored in blue. (C) A surface representation of DNAJA1-JD (rotated  $\sim 90^\circ$ ) with the proposed inhibition site based upon the TIM14-TIM16 complex colored in purple. (D) The sequence of DNAJA1-JD with the proposed interaction sites indicated: DnaK binding site from titrations (blue circle); DnaK inhibition site (purple triangles); DnaK binding site from cross-linked auxilin-Hsp70 complex (red stars); highly conserved HPD motif (green box).

A high-throughput NMR ligand affinity screen was performed with the FAST-NMR function-based compound library. The 1D line-broadening screen identified 27 possible binders. A 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen identified only 7 compounds that induced chemical shift perturbations (CSPs) upon binding with DNAJA1-JD [Figure 8.13]. The seven compounds induced CSPs in the same set of residues, inferring a consistent and

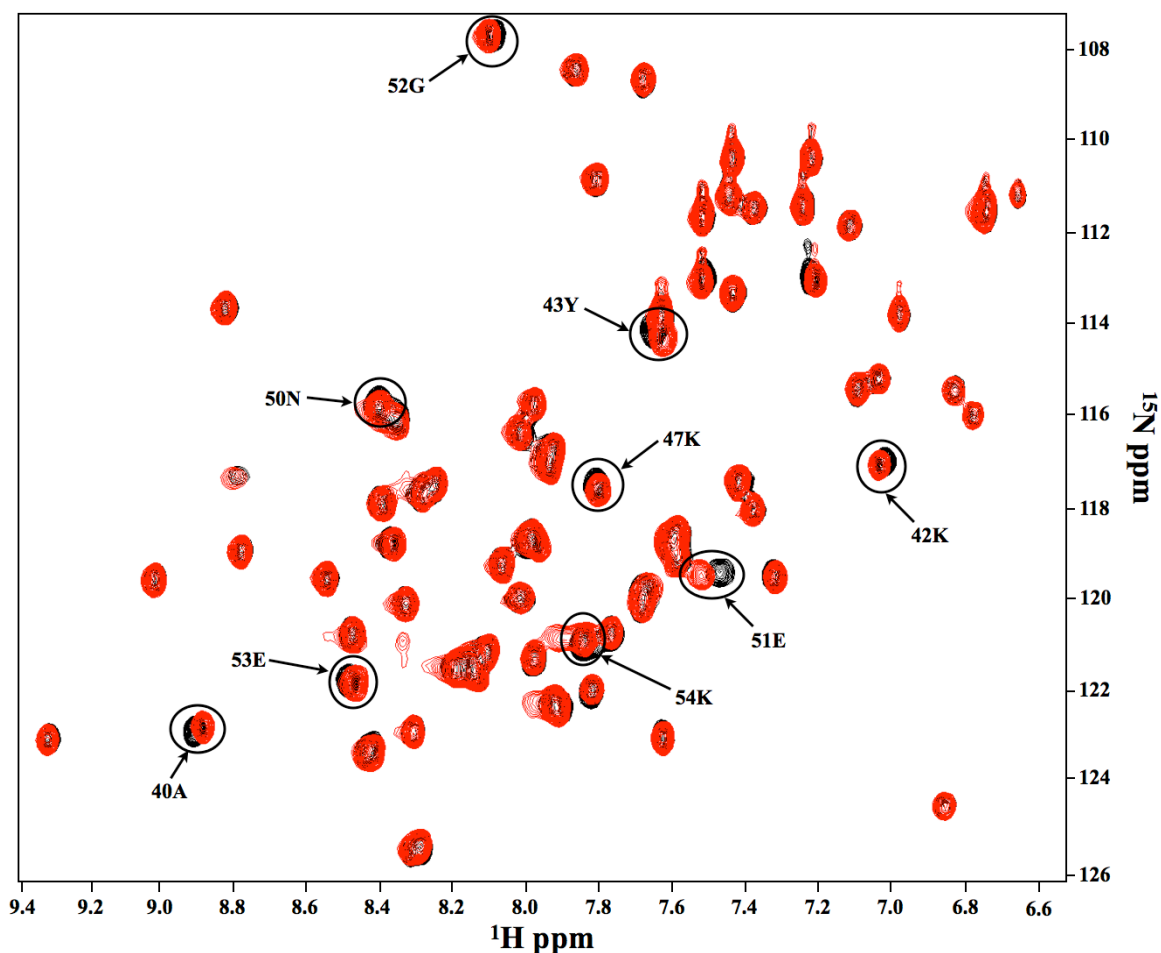
unique ligand binding site. Most of the compounds are small, but a common chemical motif or scaffold is not apparent. However, four of the compounds contain a phosphate group, two contain a carboxylic acid group, and hydroxyl groups are also very common, all these groups are likely to be deprotonated at pH 7.0, leaving negatively charged molecules.



**Figure 8.13** The structures of the compounds that induce significant chemical shift perturbations (CSPs) in a 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of DNAJA1-JD. O-phospho-L-serine showed the greatest number of significant perturbations (9 significant CSPs).

Of the 7 compounds, O-phospho-L-serine had the greatest number of perturbations (9) in the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screen [Figure 8.14]. Using the size of O-

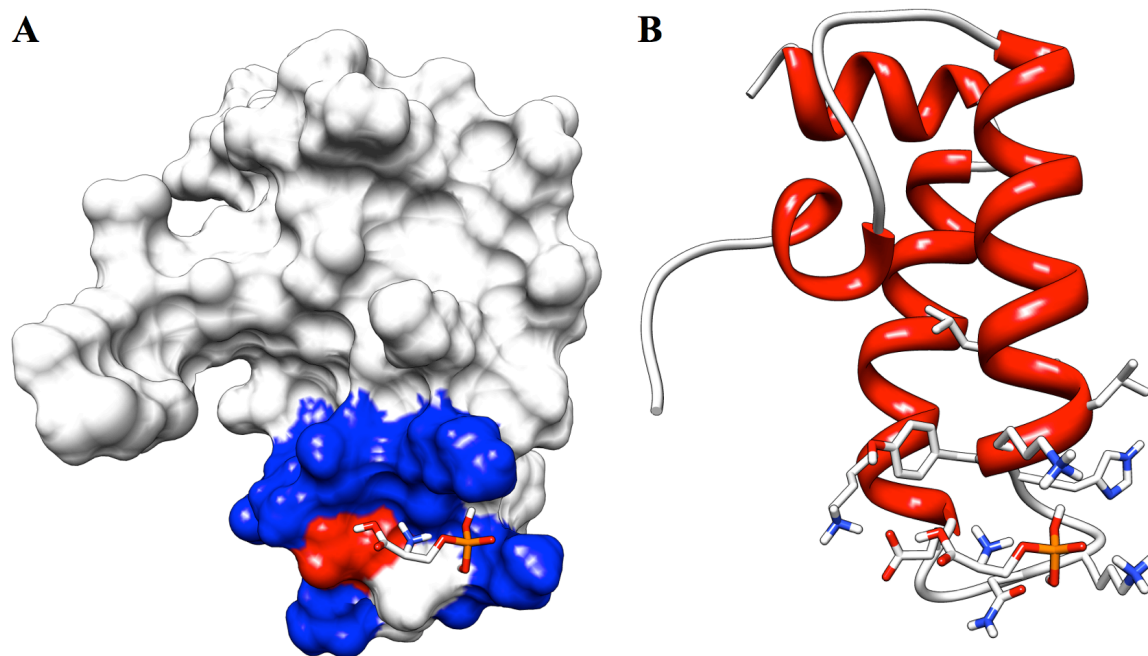
phospho-L-serine and CSP-Consensus, every one of the perturbed residues was determined to be part of the consensus binding site.



**Figure 8.14** An overlay of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC spectra of free DNAJA1-JD (black) and DNAJA1-JD with O-phospho-L-serine (red).

A DNAJA1-JD/O-phospho-L-serine costructure was determined using AutoDock and AutoDockFilter, which identifies the docked pose that best matches the chemical shift perturbation data. The costructure selected by AutoDockFilter [Figure 8.15] had an AutoDock binding energy of -2.66 kcal/mol, which fits with the average AutoDock binding energy of  $-2.54 \pm 0.40$  kcal/mol for all of the docked poses. The costructure has 9

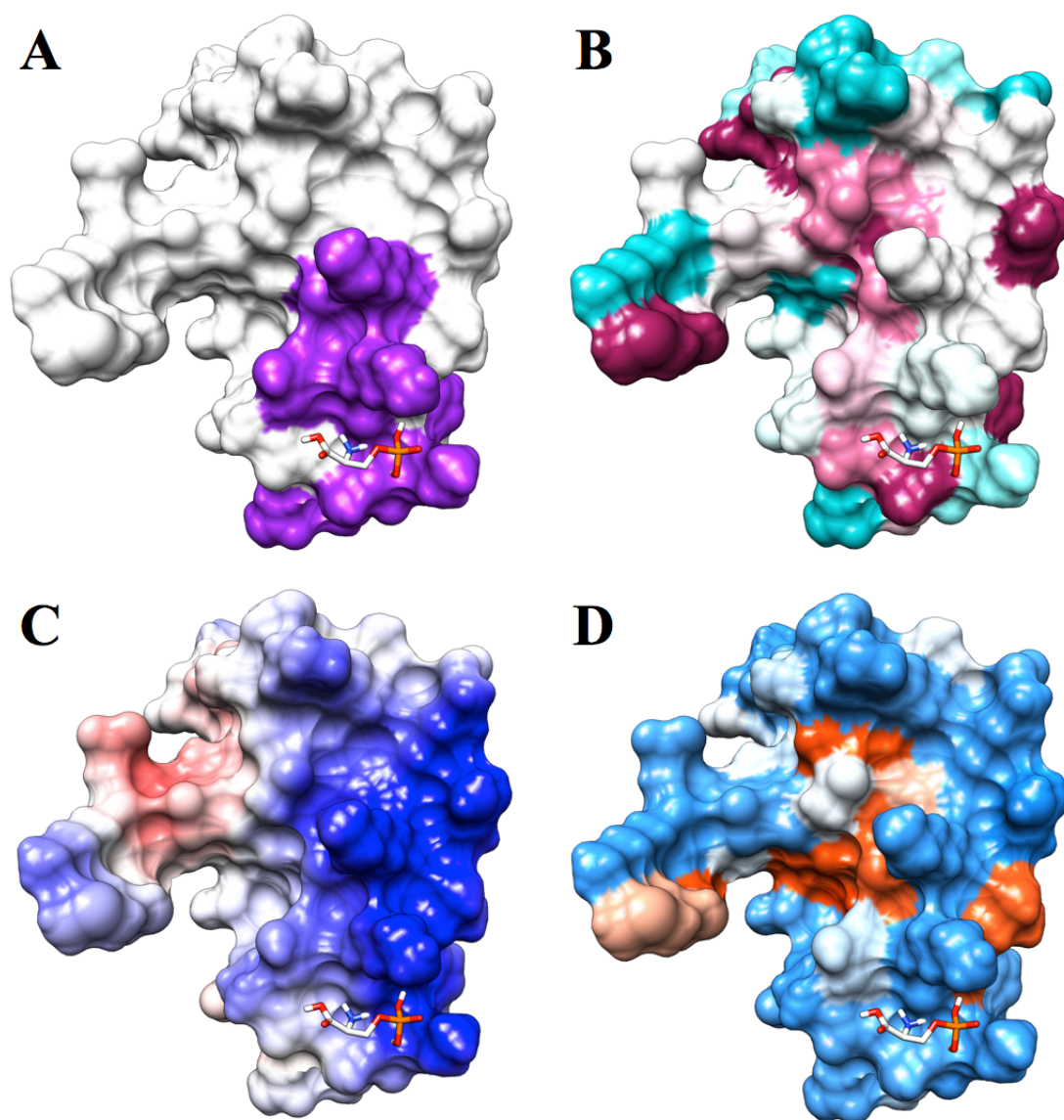
residues within 6 Å of the docked ligand: Leu39, Leu41, Tyr43, His44, Lys47, Asn48, Glu 51, and Lys54 [Figure 8.15B]. Only four of these residues were perturbed in the 2D  $^1\text{H}$ , $^{15}\text{N}$ -HSQC. However, the nearby residues identified in the costructure may be different due to the static nature of the protein during the docking calculation.



**Figure 8.15** (A) A surface representation of DNAJA1-JD with bound O-phospho-L-serine where the residues showing a chemical shift perturbation are colored blue. The one residue, Glu51, which shows the greatest chemical shift perturbation with every binding ligand is colored red. (B) A ribbon representation of DNAJA1-JD bound to O-phospho-L-serine where the sidechains of any residue within 6 Å of the ligand are displayed.

Based on the protein-ligand costructure, the binding site coincides with the predicted inhibition site based on the TIM14-TIM16 complex [Figure 8.16A]. The O-phospho-L-serine binding site is also consistent with the chemical shift perturbations for the other compounds shown to bind from the FAST-NMR assay. Surprisingly, the binding site is not particularly well conserved, evolutionarily [Figure 8.16B]. Since most of the binding ligands have negatively charged groups, it seems likely that the compounds have an electrostatic interaction with the positively charged region on the  $\alpha$ 2-

helix [Figure 8.16C]. Additionally, most of the protein surface, including the binding site, consists of hydrophilic residues [Figure 8.16D] indicating that hydrophobic interactions are unlikely to be energetically favorable. These properties may explain why most of the ligands shown to bind to DNAJA1-JD were small and negatively charged.



**Figure 8.16** A surface representation of DNAJA1-JD bound with O-phospho-L-serine with (A) the proposed inhibition site based on the TIM14-TIM16 interaction (purple), (B) the highly conserved (magenta) and poorly conserved (cyan) residues from Consurf, (C) the positively charged surface (blue) and negatively charged surface (red) from Delphi, and (D) the hydrophilic surface residues (light blue) and hydrophobic surface residues (orange-red).

## 8.4 CONCLUSIONS

Advances in the treatment of pancreatic cancer have been slow despite the urgent need for better therapies. The generation of a pancreatic cancer 'omics database identified

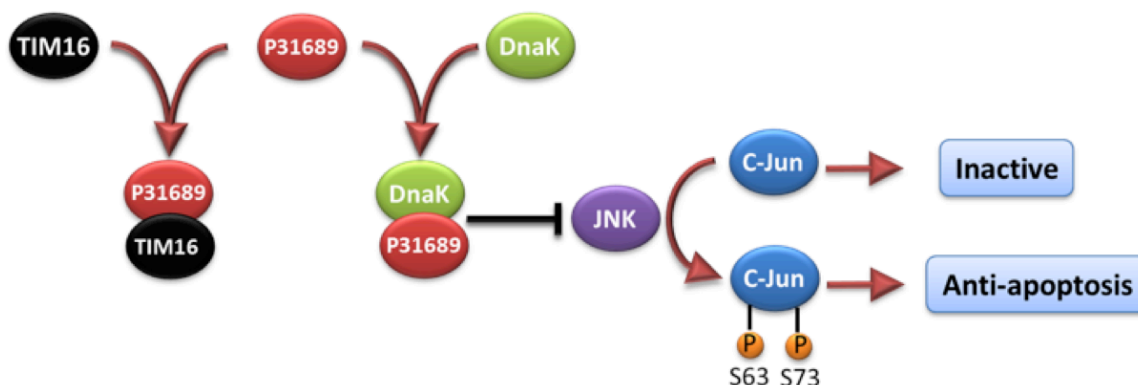
5,336 proteins which underwent significant changes in expression profiles or mutation frequency in pancreatic cancer cells, which represents nearly 25% of the human genome. An 'omics database was designed to find potential therapeutic targets using functional networks based on annotations and interaction information from several other databases. Our database identified the human protein DnaJ homolog subfamily A member 1 (DNAJA1) as a promising therapeutic target. DNAJA1 is significantly down-regulated in pancreatic cancer cells. DNAJA1 is a cochaperone that facilitates the hydrolysis of ATP from DnaK (Hsp70), a chaperone that assists protein folding, prevents aggregation of misfolded proteins, and transports proteins across membranes.

Cell-based functional assays showed that the overexpression of DNAJA1 suppresses the stress response capabilities of the oncogenic transcription factor, c-Jun, which is often overexpressed and hyperphosphorylated in cancer cells. c-Jun is part of the JNK signaling pathway, and its phosphorylation state can promote apoptosis or cell proliferation. DnaK has previously been shown to suppress the JNK pathway, which inhibits the hyperphosphorylated, anti-apoptosis state found in pancreatic cancer cells. The down-regulation of DNAJA1 in pancreatic cancer cells likely lowers the activity of DnaK, which allows for the hyperphosphorylation of c-Jun.

The solution structure of the J-domain of DNAJA1 (DNAJA1-JD) was determined by NMR. The structure has the same features as other homologous J-domains, including the conserved HPD motif in the loop between the  $\alpha 2$  and  $\alpha 3$  helices. A high-throughput ligand affinity screen by NMR identified 7 compounds that bound to the same region of the protein. The strongest binder, O-phospho-L-serine, appeared to bind in a region predicted to inhibit the binding of DnaJ proteins to DnaK.



The structure, bioinformatics analysis, cell-based assays and ligand affinity screen suggest that DNAJA1 has a role in pancreatic cancer. The J-domain of DNAJA1 appears to have overlapping protein-protein interfaces, where one interface activates DnaJ function and the other inhibits it. These interfaces on DNAJA1-JD may be interesting targets for future drug discovery efforts related to pancreatic cancer.



**Figure 8.17** An illustration of the proposed role P31689 (DNAJA1) and DnaK may have on the JNK pathway and c-Jun phosphorylation. The activation of DNAJA1 through the interaction with DnaK appears to suppress the JNK pathway, thus keeping c-Jun in the inactive state. However, inhibiting DNAJA1 binding to DnaK, as TIM16 does, would allow for the hyperphosphorylation of c-Jun which has an anti-apoptosis effect.

## 8.5 REFERENCES

1. Siegel, R., Naishadham, D. & Jemal, A. Cancer statistics, 2012. *CA Cancer J Clin* **62**, 10–29 (2012).
2. Hidalgo, M. Pancreatic cancer. *N. Engl. J. Med.* **362**, 1605–1617 (2010).
3. Li, D., Xie, K., Wolff, R. & Abbruzzese, J. L. Pancreatic cancer. *Lancet* **363**, 1049–1057 (2004).
4. Hidalgo, M. New insights into pancreatic cancer biology. *Ann Oncol* **23 Suppl 10**, x135–8 (2012).
5. Mini, E., Nobili, S., Caciagli, B., Landini, I. & Mazzei, T. Cellular pharmacology of gemcitabine. *Ann Oncol* **17 Suppl 5**, v7–12 (2006).
6. Moore, M. J. *et al.* Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. *J. Clin. Oncol.* **25**, 1960–1966

- (2007).
7. Sheikh, R., Walsh, N., Clynes, M., O'Connor, R. & McDermott, R. Challenges of drug resistance in the management of pancreatic cancer. *Expert Rev Anticancer Ther* **10**, 1647–1661 (2010).
  8. Rivera, F., López-Tarruella, S., Vega-Villegas, M. E. & Salcedo, M. Treatment of advanced pancreatic cancer: from gemcitabine single agent to combinations and targeted therapy. *Cancer Treat. Rev.* **35**, 335–339 (2009).
  9. Yamada, M., Fujii, K., Koyama, K., Hirohashi, S. & Kondo, T. The Proteomic Profile of Pancreatic Cancer Cell Lines Corresponding to Carcinogenesis and Metastasis. *J Proteomics Bioinform* **2**, 001–018 (2009).
  10. Shen, J., Person, M. D., Zhu, J., Abbruzzese, J. L. & Li, D. Protein expression profiles in pancreatic adenocarcinoma compared with normal pancreatic tissue and tissue affected by pancreatitis as detected by two-dimensional gel electrophoresis and mass spectrometry. *Cancer Res* **64**, 9018–9026 (2004).
  11. Chen, R. *et al.* Pancreatic cancer proteome: the proteins that underlie invasion, metastasis, and immunologic escape. *Gastroenterology* **129**, 1187–1197 (2005).
  12. Crnogorac-Jurcevic, T. *et al.* Proteomic analysis of chronic pancreatitis and pancreatic adenocarcinoma. *Gastroenterology* **129**, 1454–1463 (2005).
  13. Grützmann, R. *et al.* Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene* **24**, 5079–5088 (2005).
  14. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008).
  15. Pilarsky, C. *et al.* Activation of Wnt signalling in stroma from pancreatic cancer identified by gene expression profiling. *J. Cell. Mol. Med.* **12**, 2823–2835 (2008).
  16. Dummer, E. C. *et al.* The UniProt-GO Annotation database in 2011. *Nucleic Acids Res* **40**, D565–70 (2012).
  17. UniProt Consortium Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **40**, D71–5 (2012).
  18. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
  19. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **10**, 980 (2003).
  20. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
  21. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).
  22. Gabanyi, M. J. *et al.* The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *J Struct Funct Genomics* **12**, 45–54 (2011).
  23. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29 (2000).
  24. Altenhoff, A. M., Schneider, A., Gonnet, G. H. & Dessimoz, C. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res* **39**, D289–94 (2011).
  25. Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding

- affinities. *Nucleic Acids Res* **35**, D198–201 (2007).
26. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30**, 303–305 (2002).
  27. Powell, S. *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res* **40**, D284–9 (2012).
  28. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res* **40**, D84–90 (2012).
  29. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, D109–14 (2012).
  30. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, D480–4 (2008).
  31. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res* **40**, D290–D301 (2012).
  32. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**, D808–15 (2013).
  33. McDermott, K. M. *et al.* Overexpression of MUC1 reconfigures the binding properties of tumor cells. *Int J Cancer* **94**, 783–791 (2001).
  34. Singh, P. K. *et al.* Platelet-derived growth factor receptor beta-mediated phosphorylation of MUC1 enhances invasiveness in pancreatic adenocarcinoma cells. *Cancer Res* **67**, 5201–5210 (2007).
  35. Wen, Y., Caffrey, T. C., Wheelock, M. J., Johnson, K. R. & Hollingsworth, M. A. Nuclear association of the cytoplasmic tail of MUC1 and beta-catenin. *J Biol Chem* **278**, 38029–38039 (2003).
  36. Singh, P. K. *et al.* Phosphorylation of MUC1 by Met modulates interaction with p53 and MMP1 expression. *J Biol Chem* **283**, 26985–26995 (2008).
  37. Costa, N. R., Paulo, P., Caffrey, T., Hollingsworth, M. A. & Santos-Silva, F. Impact of MUC1 mucin downregulation in the phenotypic characteristics of MKN45 gastric carcinoma cell line. *PLoS ONE* **6**, e26970 (2011).
  38. Brodsky, J. L. & Chiosis, G. Hsp70 molecular chaperones: emerging roles in human disease and identification of small molecule modulators. *Curr Top Med Chem* **6**, 1215–1225 (2006).
  39. Patury, S., Miyata, Y. & Gestwicki, J. E. Pharmacological targeting of the Hsp70 chaperone. *Curr Top Med Chem* **9**, 1337–1351 (2009).
  40. Rohde, M. *et al.* Members of the heat-shock protein 70 family promote cancer cell growth by distinct mechanisms. *Genes Dev* **19**, 570–582 (2005).
  41. Jäättelä, M., Wissing, D., Bauer, P. A. & Li, G. C. Major heat shock protein hsp70 protects tumor cells from tumor necrosis factor cytotoxicity. *EMBO J* **11**, 3507–3512 (1992).
  42. Jäättelä, M., Wissing, D., Kokholm, K., Kallunki, T. & Egeblad, M. Hsp70 exerts its anti-apoptotic function downstream of caspase-3-like proteases. *EMBO J* **17**, 6124–6134 (1998).
  43. Ikura, M., Kay, L. E. & Bax, A. A novel approach for sequential assignment of <sup>1</sup>H, <sup>13</sup>C, and <sup>15</sup>N spectra of proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. *Biochemistry* **29**, 4659–4667 (1990).

44. Kay, L. E., Ikura, M., Tschudin, R. & Bax, A. Three-dimensional triple-resonance NMR Spectroscopy of isotopically enriched proteins. *J Magn Reson* **89**, 496–514 (1990).
45. Bax, A. Triple resonance three-dimensional protein NMR: before it became a black box. *J Magn Reson* **213**, 442–445 (2011).
46. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6**, 277–293 (1995).
47. Vranken, W. F. *et al.* The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* **59**, 687–696 (2005).
48. Bonvin, A. M. J. J., Rosato, A. & Wassenaar, T. A. The eNMR platform for structural biology. *J Struct Funct Genomics* **11**, 1–8 (2010).
49. Shen, Y. *et al.* Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* **105**, 4685–4690 (2008).
50. Shen, Y., Vernon, R., Baker, D. & Bax, A. De novo protein structure generation from incomplete chemical shift assignments. *J Biomol NMR* **43**, 63–78 (2009).
51. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. The Xplor-NIH NMR molecular structure determination package. *J Magn Reson* **160**, 65–73 (2003).
52. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol NMR* **44**, 213–223 (2009).
53. Nederveen, A. J. *et al.* RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins* **59**, 662–672 (2005).
54. Bhattacharya, A., Tejero, R. & Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins* **66**, 778–795 (2007).
55. Eisenberg, D., Luthy, R. & Bowie, J. U. VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol* **277**, 396–404 (1997).
56. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355–362 (1993).
57. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* **26**, 283–291 (1993).
58. Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* **35**, W375–83 (2007).
59. Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protocols Bioinformatics* 2.3.1–2.3.22 (2002).
60. Pellecchia, M., Szyperski, T., Wall, D., Georgopoulos, C. & Wüthrich, K. NMR structure of the J-domain and the Gly/Phe-rich region of the Escherichia coli DnaJ chaperone. *J Mol Biol* **260**, 236–250 (1996).
61. Qian, Y. Q., Patel, D., Hartl, F. U. & McColl, D. J. Nuclear magnetic resonance solution structure of the human Hsp40 (HDJ-1) J-domain. *J Mol Biol* **260**, 224–235 (1996).
62. Gao, X.-C. *et al.* The C-terminal helices of heat shock protein 70 are essential for J-domain binding and ATPase activation. *J Biol Chem* **287**, 6044–6052 (2012).

63. Honig, B. & Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **268**, 1144–1149 (1995).
64. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* **38**, W529–33 (2010).
65. Mercier, K. A., Germer, K. & Powers, R. Design and characterization of a functional library for NMR screening against novel protein targets. *Comb Chem High Throughput Screen* **9**, 515–534 (2006).
66. Mercier, K. A. & Powers, R. Determining the optimal size of small molecule mixtures for high throughput NMR screening. *J Biomol NMR* **31**, 243–258 (2005).
67. Morris, G., Goodsell, D., Halliday, R. & Huey, R. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**, 1639–1662 (1998).
68. Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* **28**, 1145–1152 (2007).
69. Morris, G. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* (2009).doi:10.1002/jcc.21256
70. Sanner, M. F. Python: a programming language for software integration and development. *J Mol Graph Model* **17**, 57–61 (1999).
71. Stark, J. L. & Powers, R. Rapid protein-ligand costructures using chemical shift perturbations. *J Am Chem Soc* **130**, 535–545 (2008).
72. Jiang, J. *et al.* Structural basis of J cochaperone binding and regulation of Hsp70. *Mol Cell* **28**, 422–433 (2007).
73. Mitra, A., Shevde, L. A. & Samant, R. S. Multi-faceted role of HSP40 in cancer. *Clin Exp Metastasis* **26**, 559–567 (2009).
74. Qiu, X.-B., Shao, Y.-M., Miao, S. & Wang, L. The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. *Cell Mol Life Sci* **63**, 2560–2570 (2006).
75. Horne, B. E., Li, T., Genevoux, P., Georgopoulos, C. & Landry, S. J. The Hsp40 J-domain stimulates Hsp70 when tethered by the client to the ATPase domain. *J Biol Chem* **285**, 21679–21688 (2010).
76. Wall, D., Zylicz, M. & Georgopoulos, C. The NH<sub>2</sub>-terminal 108 amino acids of the Escherichia coli DnaJ protein stimulate the ATPase activity of DnaK and are sufficient for lambda replication. *J Biol Chem* **269**, 5446–5451 (1994).
77. Terada, K. *et al.* A type I DnaJ homolog, DjA1, regulates androgen receptor signaling and spermatogenesis. *EMBO J* **24**, 611–622 (2005).
78. Hennessy, F., Nicoll, W. S., Zimmermann, R., Cheetham, M. E. & Blatch, G. L. Not all J domains are created equal: implications for the specificity of Hsp40-Hsp70 interactions. *Protein Sci* **14**, 1697–1709 (2005).
79. Walsh, P., Bursac, D., Law, Y. C., Cyr, D. & Lithgow, T. The J-protein family: modulating protein assembly, disassembly and translocation. *EMBO Rep* **5**, 567–571 (2004).
80. Terada, K. & Oike, Y. Multiple molecules of Hsc70 and a dimer of DjA1 independently bind to an unfolded protein. *J Biol Chem* **285**, 16789–16797

- (2010).
81. Wang, C.-C. *et al.* HDJ-2 as a target for radiosensitization of glioblastoma multiforme cells by the farnesyltransferase inhibitor R115777 and the role of the p53/p21 pathway. *Cancer Res* **66**, 6756–6762 (2006).
  82. Chow, L. Q. M. *et al.* A phase I safety, pharmacological, and biological study of the farnesyl protein transferase inhibitor, lonafarnib (SCH 663366), in combination with cisplatin and gemcitabine in patients with advanced solid tumors. *Cancer Chemother. Pharmacol.* **62**, 631–646 (2008).
  83. Patnaik, A. *et al.* A phase I, pharmacokinetic, and biological study of the farnesyltransferase inhibitor tipifarnib in combination with gemcitabine in patients with advanced malignancies. *Clin Cancer Res* **9**, 4761–4771 (2003).
  84. Kanazawa, M., Terada, K., Kato, S. & Mori, M. HSDJ, a human homolog of DnaJ, is farnesylated and is involved in protein import into mitochondria. *J. Biochem.* **121**, 890–895 (1997).
  85. Terada, K., Kanazawa, M., Bukau, B. & Mori, M. The human DnaJ homologue dj2 facilitates mitochondrial protein import and luciferase refolding. *J. Cell Biol.* **139**, 1089–1095 (1997).
  86. Ott, M., Gogvadze, V., Orrenius, S. & Zhivotovsky, B. Mitochondria, oxidative stress and cell death. *Apoptosis* **12**, 913–922 (2007).
  87. Paschen, S. A., Weber, A. & Häcker, G. Mitochondrial protein import: a matter of death? *Cell Cycle* **6**, 2434–2439 (2007).
  88. Petit, E., Oliver, L. & Vallette, F. M. The mitochondrial outer membrane protein import machinery: a new player in apoptosis? *Front. Biosci.* **14**, 3563–3570 (2009).
  89. Llambi, F. & Green, D. R. Apoptosis and oncogenesis: give and take in the BCL-2 family. *Curr Opin Genet Devel* **21**, 12–20 (2011).
  90. Jäättelä, M. Over-expression of hsp70 confers tumorigenicity to mouse fibrosarcoma cells. *Int J Cancer* **60**, 689–693 (1995).
  91. Gurbuxani, S. *et al.* Selective depletion of inducible HSP70 enhances immunogenicity of rat colon cancer cells. *Oncogene* **20**, 7478–7485 (2001).
  92. Volloch, V. Z. & Sherman, M. Y. Oncogenic potential of Hsp72. *Oncogene* **18**, 3648–3651 (1999).
  93. Wodrich, W. & Volm, M. Overexpression of oncoproteins in non-small cell lung carcinomas of smokers. *Carcinogenesis* **14**, 1121–1124 (1993).
  94. Binétruy, B., Smeal, T. & Karin, M. Ha-Ras augments c-Jun activity and stimulates phosphorylation of its activation domain. *Nature* **351**, 122–127 (1991).
  95. Agarwal, S., Corbley, M. J. & Roberts, T. M. Reconstitution of signal transduction from the membrane to the nucleus in a baculovirus expression system: activation of Raf-1 leads to hypermodification of c-jun and c-fos via multiple pathways. *Oncogene* **11**, 427–438 (1995).
  96. Shin, S. *et al.* Activator protein-1 has an essential role in pancreatic cancer cells and is regulated by a novel Akt-mediated mechanism. *Mol. Cancer Res.* **7**, 745–754 (2009).
  97. Tessari, G. *et al.* The expression of proto-oncogene c-jun in human pancreatic cancer. *Anticancer Res.* **19**, 863–867 (1999).

98. Weston, C. R. & Davis, R. J. The JNK signal transduction pathway. *Curr Opin Cell Biol* **19**, 142–149 (2007).
99. Chen, Y. R. & Tan, T. H. The c-Jun N-terminal kinase pathway and apoptotic signaling (review). *Int. J. Oncol.* **16**, 651–662 (2000).
100. Shaulian, E. AP-1--The Jun proteins: Oncogenes or tumor suppressors in disguise? *Cell. Signal.* **22**, 894–899 (2010).
101. Wisdom, R., Johnson, R. S. & Moore, C. c-Jun regulates cell cycle progression and apoptosis by distinct mechanisms. *EMBO J* **18**, 188–197 (1999).
102. Eferl, R. *et al.* Liver tumor development. c-Jun antagonizes the proapoptotic activity of p53. *Cell* **112**, 181–192 (2003).
103. Mathas, S. *et al.* Aberrantly expressed c-Jun and JunB are a hallmark of Hodgkin lymphoma cells, stimulate proliferation and synergize with NF-kappa B. *EMBO J* **21**, 4104–4113 (2002).
104. Ahmed, S. U. & Milner, J. Basal cancer cell survival involves JNK2 suppression of a novel JNK1/c-Jun/Bcl-3 apoptotic network. *PLoS ONE* **4**, e7305 (2009).
105. Takayama, S., Reed, J. C. & Homma, S. Heat-shock proteins as regulators of apoptosis. *Oncogene* **22**, 9041–9047 (2003).
106. Jolly, C. & Morimoto, R. I. Role of the heat shock response and molecular chaperones in oncogenesis and cell death. *J. Natl. Cancer Inst.* **92**, 1564–1572 (2000).
107. Mosser, D. D. *et al.* The chaperone function of hsp70 is required for protection against stress-induced apoptosis. *Mol. Cell. Biol.* **20**, 7146–7159 (2000).
108. Sousa, R. *et al.* Evaluation of competing J domain:Hsp70 complex models in light of existing mutational and NMR data. *Proc Natl Acad Sci USA* **109**, E734 (2012).
109. Zuiderweg, E. R. P. & Ahmad, A. Reply to Sousa *et al.*: Evaluation of competing J domain:Hsp70 complex models in light of methods used. *Proc Natl Acad Sci USA* **109**, E735–E735 (2012).
110. Ahmad, A. *et al.* Heat shock protein 70 kDa chaperone/DnaJ cochaperone complex employs an unusual dynamic interface. *Proc Natl Acad Sci USA* **108**, 18966–18971 (2011).
111. Greene, M. K., Maskos, K. & Landry, S. J. Role of the J-domain in the cooperation of Hsp40 with Hsp70. *Proc Natl Acad Sci USA* **95**, 6108–6113 (1998).
112. Mokranjac, D., Sichting, M., Neupert, W. & Hell, K. Tim14, a novel key component of the import motor of the TIM23 protein translocase of mitochondria. *EMBO J* **22**, 4945–4956 (2003).
113. Mokranjac, D., Bourenkov, G., Hell, K., Neupert, W. & Groll, M. Structure and function of Tim14 and Tim16, the J and J-like components of the mitochondrial protein import motor. *EMBO J* **25**, 4675–4685 (2006).

## CHAPTER 9

### SUMMARY AND FUTURE WORK

The number of proteins that are considered "uncharacterized" proteins with no known function is constantly growing. Pure biochemical approaches to determining protein function are too slow to address this growth, while using sequence and structure similarities for functional annotation is only reliable for globally similar proteins. Additionally, structural genomics has resulted in a large number of protein with structures for which nothing else is known. These uncharacterized proteins essentially become "orphaned" since it is unknown whether they present an interesting research target. New methodologies are necessary to address this problem.

One alternative approach involves the comparison of the protein active sites instead of entire protein. Because the protein active site is responsible for the interactions of a protein with other biomolecules or small molecules, the properties of this localized functional epitope are typically more evolutionarily conserved.<sup>1,2</sup> This dissertation demonstrates the application of this approach with Functional Annotation Screening Technology by NMR (FAST-NMR), which combines NMR ligand affinity screens and molecular docking in a high-throughput methodology to functionally annotate proteins.<sup>1,2</sup>

Chapter 2 describes the application of FAST-NMR to propose functions for 20 uncharacterized proteins from the Northeast Structural Genomics Consortium (NESG; <http://www.nesg.org>). A library of 460 functionally-relevant compounds<sup>3</sup> was screened in a tiered approach with 1D  $^1\text{H}$  NMR line-broadening screens followed by 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screens to identify binders and a binding site.<sup>4</sup> This was followed by molecular



docking with AutoDock, which generated a protein-ligand costructure.<sup>5,6</sup> The experimental binding site defined by the costructure was then compared to a database of binding sites using the Comparison of Protein Active Site Structures (CPASS) program.<sup>7,8</sup> Additionally, other bioinformatics tools combined with knowledge of the identity of the binding compounds were all used to propose functions for the 20 uncharacterized proteins. The results of this approach are intended to guide future work on these "orphaned" proteins, thus minimizing time and effort. For example, using the weak binders identified during the FAST-NMR screen could guide future investigations of compounds with similar moieties.

The FAST-NMR screen of the 20 NESG proteins represents the first time that the FAST-NMR process was implemented for multiple proteins at the same time. Overall, this resulted in improvements in the FAST-NMR protocol, as well as identifying any bottlenecks. Preparation of nearly 3,000 NMR samples requires a significant amount of time that could be alleviated with automated sample preparation, and should be one focus for improvement in the future.

The other major bottleneck occurs during both the 1D and 2D NMR spectral analysis steps. Analysis of the 1D <sup>1</sup>H line-broadening screen requires overlaying the spectra for the compound mixtures with and without protein. The spectra is then manually evaluated to identify any potentially decrease in peak height (line-broadening). Unfortunately, proteins have many hydrogen atoms that can be observed in the 1D <sup>1</sup>H NMR spectra as well, and these hydrogen peaks can obfuscate the presence of peak line-broadening for the ligands. Future efforts are investigating whether to use a 1D saturation transfer difference (STD)<sup>9</sup> NMR screen instead of a 1D <sup>1</sup>H line-broadening screen to

identify binding ligands. However, 1D STD NMR screens take significantly more NMR time and produce more hits that represent non-specific binders. Another approach currently under investigation would be to use mathematical techniques to subtract the protein peaks from the 1D  $^1\text{H}$  NMR spectra. This would effectively allow for an automated analysis of the spectra, which would increase throughput.

For the 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC screens, the identification of chemical shift perturbations (CSPs) is straightforward. However, the challenge lies in identifying the amino acid resonance being perturbed. Each peak in a 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC experiment correlates to the amide of an amino acid residue. Since the proteins being investigated by FAST-NMR have had their backbones previously assigned by NMR, the resonances for each amino acid residue can be determined using the Biological Magnetic Resonance Data Bank (BMRB; <http://bmr.b.wisc.edu>).<sup>10</sup> However, in some cases during the FAST-NMR analysis, the sample conditions of the protein are different enough to result in amino acid residue resonances that are not the same as those found in the BMRB. This can make it difficult to assign CSPs to amino acid residues, and thus determine a consensus binding site. Additionally, since CSPs, by definition, are the change in the residue resonance peaks of a protein with ligand compared to the protein without ligand, it is often difficult to determine the origin of each perturbed peak in a cluster of perturbed peaks. Both of these issues could be addressed by initially running a series of 2D  $^1\text{H}$ ,  $^{15}\text{N}$ -HSQC experiments that gradually change sample buffer conditions and/or ligand concentrations. This may generate smaller perturbations which could be easier track.

The identity of the compounds that were shown to bind each protein during the FAST-NMR screens was extremely important in proposing a protein function. Currently,

there is no implemented protocol to evaluate the binding compounds as a whole in a systematic manner. There is significant value for functional annotation in identifying common physicochemical features, metabolic pathways, and other binding proteins. While these approaches were certainly used during the FAST-NMR analysis, they also required a significant amount of effort to evaluate individually. The BioScreen Ligands database (<http://bionmr-c1.unl.edu/cgi-bin/ligands/index>) is a great start to providing structural information on the compounds. However, implementing additional data into the database such as KEGG pathways,<sup>11</sup> Tanimoto structural comparisons,<sup>12</sup> the identity of binding proteins from the RCSB PDB,<sup>13</sup> as well as the means to compare these features between a list of ligands would be an invaluable tool.

The FAST-NMR process relies on the generation of protein-ligand costructures to identify an experimentally-determined binding site to be used for comparisons in CPASS. Experimental approaches to determine protein-ligand costructures with NMR spectroscopy or X-ray crystallography can take months.<sup>14,15</sup> Chapters 3 and 4 describe the development of the programs AutoDockFilter (ADF) and CSP-Consensus (CSPC), which utilizes molecular docking and the CSPs from the 2D <sup>1</sup>H, <sup>15</sup>N-HSQC screens to quickly generate a protein-ligand costructure in approximately an hour.<sup>16</sup>

ADF was shown to significantly improve the results of molecular docking with AutoDock by selecting the results that best agree with the experimental CSPs. However, as a filter, ADF still requires AutoDock to produce accurate results. Future versions of ADF could instead act to steer the docking process,<sup>17</sup> which would require adding the ADF violation energy into the AutoDock calculations. Also, ADF does not currently allow for the filtering of AutoDock results if any protein sidechains are allowed to be

flexible during the AutoDock calculation. Addition of this feature would help remove any docking problems that occur due to the use of a protein apostructure.

AutoDock accuracy is improved when the search grid is focused upon the binding site of the ligand.<sup>16,18,19</sup> Normally, CSPs from the 2D <sup>1</sup>H, <sup>15</sup>N-HSQC screens are visually mapped onto the surface of the protein in order to subjectively identify a consensus binding site. CSPC was developed to make determining the consensus binding site more objective by using hierarchical clustering of the amide-amide distances between perturbed residues. When tested on 8 protein-ligand systems with experimental CSP values, it was able to reliably identify the binding site on each protein. Future work for CSPC involves setting up a web-based server for the program and creating a user interface to visually evaluate the clustering results.

After generating the protein-ligand costructure with AutoDock and ADF, the ligand-defined binding site is then submitted to the CPASS program, where it is compared to a database of ligand-defined binding sites generated from the RCSB PDB.<sup>13</sup> The comparison involves the alignment of the binding sites structure and sequence. Recently, CPASS was updated to version 2,<sup>8</sup> which includes additional comparisons that include C $\beta$  position, surface accessibility, ligand alignment, and implementation on the Open Science Grid (OSG). Chapter 5 provided an evaluation of this new version of CPASS by using ROC comparisons to illustrate the enrichment of functionally similar proteins in CPASS.

CPASS 2.0 has been shown to be very effective in selecting for proteins with high functional similarity. In general, a CPASS similarity score of 30% or greater indicates a greater likelihood of functional homology. However, the FAST-NMR screens of the 20

uncharacterized proteins highlighted some potential problems that have made it challenging to use to functionally annotate proteins. However, the variability in defining a binding site using predictive or screening approaches can significantly influence CPASS similarity calculations, where differences in bound and unbound structures, errors in docking a ligand to the structure, and variations in ligand size all appear to decrease the CPASS similarity score for a well-characterized protein. Additionally, an evaluation of the CPASS results from a FAST-NMR screen of 20 proteins of unknown function highlights the impact of these issues. Specifically, these experimental problems hinder the ability of CPASS to reliably rank functional homologs as the top hit.

While experimental and computational variability will always be an issue, one approach to minimize their effect would be to prioritize the matching process. As mentioned previously, the reason protein structural differences and ligand size/location affect the CPASS similarity score is due to the way CPASS defines the ligand binding site. Every residue within 6 Å of the ligand is defined by CPASS as the binding site. While residues at the edge of the 6 Å cutoff are scaled to minimize the impact of small structural variations, the majority of the residues in the binding site are essentially equivalent in terms of importance to the CPASS scoring function. Unfortunately, not every residue located within this binding site is necessary for the molecular function of the protein, and thus would not necessarily be conserved. Implementing a weighting function for each residue in the query based on predicted or known importance may help prioritize binding sites in the database that have similar sequence and structure between these important residues. Two factors could be considered as potential sources of

weighted scoring: distance of amino acid to ligand and/or evolutionary conservation among similar proteins such as that generated by ConSurf.<sup>20</sup>

Additionally, CPASS scoring currently uses only the query active site to generate the reference score by which the similarity scores are generated. However, as illustrated during the FAST-NMR screens, proteins that bind to large compounds like suramin produce large binding sites, and thus large reference scores. Matching the entire binding site is difficult, and often results in a match in substructure but overall match less than 30%. This could be addressed by using the smaller of the two active sites to generate the reference score or by generating a Q-score that implements alignment length into the comparison.

The FAST-NMR screens also illustrated a potential update to CPASS that would be extremely useful to the end user. Since the uncharacterized proteins typically have lower CPASS scores, the relevant hits can often get mixed in with irrelevant hits. One potential approach would be to add KEGG pathways/reactions identifiers, compound name, source organism, and potential GO annotations for each active site in the CPASS database. This could allow for the immediate evaluation of common features found in the results, especially at lower similarity scores where it is more likely to see false positives.

Unfortunately, the greatest problem facing the functional annotation of unknown proteins with CPASS is the size of functional space represented by the database. The number of protein structures represented in the PDB is still significantly smaller than the number of known protein sequences. Therefore, the query protein may represent the first member of a functional class of proteins present in the PDB. Structural genomics is attempting to address this problem by prioritizing experimental structure determination

efforts. This is leading to significant increase in the number of unique protein structures. But, unfortunately, the majority of structures deposited in the PDB lack a biologically relevant ligand. Since the CPASS database is generated from proteins with bound ligands in the PDB, a functional class that does not have a representative structure with a bound ligand would not appear in the results from a CPASS query. This is especially a concern for proteins that binds another biomolecule (protein, DNA, RNA) instead of a small molecular weight compound. These problems may be addressed by expanding the size of the CPASS database to potentially include predicted binding sites for unique proteins with no bound ligands using programs such as ConSurf<sup>20</sup> or CASTp.<sup>21</sup> The inclusion of protein-protein or protein-DNA binding sites would also expand the searchable functional space. However, introducing either of these approaches to expand the database does introduce its own challenges.

Despite these issues, CPASS still provides valuable information based on the partial similarities that do exist, especially when combined with other bioinformatics approaches and experimental data. Additionally, CPASS can also be used to help understand the evolutionary relationships between proteins based on the changes that occur in the binding site.

Recently, the structure of the YndB protein from *Bacillus subtilis* was determined.<sup>22,23</sup> It featured a hydrophobic binding pocket that would likely bind a lipid molecule, which are not presently represented in the function-based compound library. In Chapter 6, the function of the YndB protein was determined using virtual screening of a lipid compound library followed by NMR titrations. Remarkably, the virtual screen identified chalcone molecules as likely binders, which was then verified by 2D <sup>1</sup>H, <sup>15</sup>N-

HSQC titrations. Chalcone was determined to be a tight binder with a  $K_D$  of less than 1  $\mu$ M. However, chalcone is not found as a natural product of *Bacillus* organisms. Instead, it is a precursor to antibiotics produced from plants. This hints at the potential symbiotic role between *Bacillus* and plants that uses chalcone as a signaling molecule to promote sporulation or antibiotic production in *Bacillus*. Current efforts are involved in verifying the effect of chalcone on *Bacillus* using a combination of cell growth analysis, proteomics, and metabolomics.

Because virtual screening worked so well for the YndB protein, could it be used to supplant the 1D  $^1\text{H}$  NMR line-broadening screens in FAST-NMR? The work in Chapter 7 investigated this possibility and found that there are some significant drawbacks to using virtual screening as a complete replacement. Future efforts could explore the use of virtual screen to investigate compounds that are not in the function-based compound library. It could also be used in cases where the binding site and a broad classification of the ligands are already known, as in the case of YndB.

Another alternative approach within FAST-NMR to annotated proteins is to compare ligand binding profiles, which are a set of ligands experimentally shown to bind a specific protein. Proteins with the same function have been demonstrated to bind the same set of small molecules from a function-based chemical library.<sup>24</sup> These ligand binding profiles can be used to identify functionally homologous proteins in a manner similar to sequence similarity techniques such as BLAST<sup>25,26</sup> using only the results of 1D  $^1\text{H}$  NMR line-broadening screen. This approach also has the added benefit of not requiring a protein structure or sequence. Virtual screening could potentially be used in a similar manner to identify a ligand binding profile, but on a large virtual library.



Pancreatic cancer is perhaps the deadliest cancer, yet it is not the most prevalent cancer and most of the current treatments for pancreatic cancer do not improve the prognosis. It's clear that discovery of a novel therapeutic target is important. Chapter 8 discusses a bioinformatics approach (Borg) to prioritize the numerous proteins that have been shown to be involved in pancreatic cancer cells followed by the structural and functional characterization of DNAJA1, a potential therapeutic target for pancreatic cancer. The overexpression of DNAJA1 results in the suppression of an anti-apoptotic mechanism found in pancreatic cancer. The structure of the J-domain of DNAJA1 has identified two potential binding sites, one for activation and one for inhibition. Future work will continue to investigate the relationship between DNAJA1 and pancreatic cancer. This will involve identifying the factors that may inhibit DNAJA1 anti-apoptotic suppression.

## 9.1 REFERENCES

1. Mercier, K. A. *et al.* FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* **128**, 15292–15299 (2006).
2. Powers, R., Mercier, K. A. & Copeland, J. C. The application of FAST-NMR for the identification of novel drug discovery targets. *Drug Discov Today* **13**, 172–179 (2008).
3. Mercier, K. A., Germer, K. & Powers, R. Design and characterization of a functional library for NMR screening against novel protein targets. *Comb Chem High Throughput Screen* **9**, 515–534 (2006).
4. Mercier, K. A., Shortridge, M. D. & Powers, R. A multi-step NMR screen for the identification and evaluation of chemical leads for drug discovery. *Comb Chem High Throughput Screen* **12**, 285–295 (2009).
5. Morris, G. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* (2009).doi:10.1002/jcc.21256
6. Morris, G., Goodsell, D., Halliday, R. & Huey, R. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**, 1639–1662 (1998).
7. Powers, R. *et al.* Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **65**, 124–135 (2006).

8. Powers, R., Copeland, J. & Stark, J. L. Searching the protein structure database for ligand-binding site similarities using CPASS v. 2. *BMC Res Notes* (2011).
9. Wang, Y.-S., Liu, D. & Wyss, D. F. Competition STD NMR for the detection of high-affinity ligands and NMR-based screening. *Magn Reson Chem* **42**, 485–489 (2004).
10. Ulrich, E. L. *et al.* BioMagResBank. *Nucleic Acids Res* **36**, D402–8 (2008).
11. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, D109–14 (2012).
12. Scsibraný, H., Karlovits, M., Demuth, W., Müller, F. & Varmuza, K. Clustering and similarity of chemical structures represented by binary substructure descriptors. *Chemometrics and Intelligent Laboratory Systems* **67**, 95–108 (2003).
13. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **10**, 980 (2003).
14. Scapin, G. Structural Biology and Drug Discovery. *Curr Pharm Des* **12**, 2087–2097 (2006).
15. Powers, R. Applications of NMR to structure-based drug design in structural genomics. *J Struct Funct Genomics* **2**, 113–123 (2002).
16. Stark, J. L. & Powers, R. Rapid protein-ligand costructures using chemical shift perturbations. *J Am Chem Soc* **130**, 535–545 (2008).
17. González-Ruiz, D. & Gohlke, H. Steering protein-ligand docking with quantitative NMR chemical shift perturbations. *J Chem Inf Model* **49**, 2260–2271 (2009).
18. Hetényi, C. & van der Spoel, D. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci* **11**, 1729–1737 (2002).
19. Bursulaya, B. D., Totrov, M., Abagyan, R. & Brooks, C. L. Comparative study of several algorithms for flexible ligand docking. *J Comput Aided Mol Des* **17**, 755–763 (2003).
20. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* **38**, W529–33 (2010).
21. Dundas, J. *et al.* CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res* **34**, W116–8 (2006).
22. Mercier, K. A. *et al.* (1)H, (13)C, and (15)N NMR assignments for the *Bacillus subtilis* yndB START domain. *Biomol NMR Assign* **3**, 191–194 (2009).
23. Stark, J. L. *et al.* Solution structure and function of YndB, an AHSA1 protein from *Bacillus subtilis*. *Proteins* **78**, 3328–3340 (2010).
24. Shortridge, M. D., Bokemper, M., Copeland, J. C., Stark, J. L. & Powers, R. Correlation between Protein Function and Ligand Binding Profiles. *J Proteome Res* **10**, 2538–2545 (2011).
25. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
26. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).