Winter 11-29-2010

Nuclear Magnetic Resonance Affinity Screening Methods for Functional Annotation of Proteins and Drug Discovery

Matthew D. Shortridge PhD
department of chemistry, mds8575@huskers.unl.edu

NUCLEAR MAGNETIC RESONANCE AFFINITY SCREENING METHODS FOR FUNCTIONAL ANNOTATION OF PROTEINS AND DRUG DISCOVERY

By:

Matthew D. Shortridge

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

December, 2010

NUCLEAR MAGNETIC RESONANCE AFFINITY SCREENING METHODS FOR FUNCTIONAL ANNOTATION OF PROTEINS AND DRUG DISCOVERY

Matthew D. Shortridge, Ph.D.

University of Nebraska, 2010

Advisor: Robert Powers

With nearly 1,350 complete genome sequences available our understanding of biology at the molecular level has never been more complete. A consequence of these sequencing projects was the discovery of large functionally unannotated segments of each genome. The genes (and proteins they encode) found in these unannotated regions are considered "hypothetical proteins". Current estimates suggest between 12%-50% of the known gene sequences are functionally unannotated. Incomplete functional annotation of the various genomes significantly limits our understanding of biology. Pragmatically, identifying the functions of these proteins could lead to new therapeutics; making functional annotation of paramount importance.

This dissertation describes the development of new methods for protein functional annotation independent of homology transfer. The hypothesis is proteins with similar function have significantly similar active sites. Nuclear magnetic resonance ligand affinity screening was employed to identify and define protein active sites. The methods developed were tested on a series of functionally diverse, annotated proteins including, serum albumins (*H. sapiens*, *B. taurus*), α and β amylases (*B. licheniformis*, *A. oryzae*, *B. amyloliquefaciens H. vulgare, I. batatas*), primase C-terminal domain (*S. aureus*), nuclease (*S. aureus*) and the type three secretion system protein PrgI (*S. typhirium*).

Functional annotation using protein active sites require a high-resolution three-dimensional structure of the protein. In addition to method development, this dissertation describes the NMR solution structure of *Staphylococcus aureus* primase carboxy-terminal domain (CTD). The primase CTD is essential for bacterial DNA replication and distinctly different from eukaryotes. With the rapid rise in antibiotic resistance, the primase CTD of *S. aureus* is an attractive antibiotic target. The methods used for functional annotation were used to screen *S. aureus* primase CTD to identify the compound acycloguanosine as a binding ligand to primase CTD.

Dedicated to:
Orville E. Miller

Acknowledgements:

The successful completion of this long academic journey required focus, enthusiasm and above all else an extreme amount of support from colleagues, family and friends. Without their helpful encouragement I would not be writing this dissertation. I would first like to extend my greatest gratitude to my research advisor Dr. Robert Powers. Your belief in me and constant encouragement lead to my success at Nebraska. With your guidance and patience I was able to take an interest in NMR and turn it into an understanding of NMR. Additionally, you have provided me with skills beyond the technical that will continually assist me on the even larger academic journey I now begin.

My interest in science stretches beyond the halls of Hamilton and started at an early age. This is because of the excellent science education I was lucky to receive early in life. Thank you to all my teachers and professors at all levels of my education that pushed me to excel. I would specifically like to thank my committee members, Dr. Mark Griep, Dr. Liangcheng Du, Dr. James Takacs, and Dr. Greg Somerville for giving me the freedom to explore and the focus to finish.

While working as a teaching assistant in the instrumentation facility I was fortunate to work with two very talented spectroscopists. A warm and special thank you goes to Dr. Joe Dumais for teaching me much more than how to shim a magnet. Your guidance, mentorship and friendship were pivotal to my success both academically and personally at Nebraska. An equally warm thank you goes to Sara Basagia. Sara, our time spent filling magnets with helium and occasionally our bellies with beer (after the

fills of course) will be some of the happiest memories I leave with. I will truly miss you both.

I was lucky to work with a number of talented colleagues and collaborators here at Nebraska and throughout the scientific community. In the Powers lab I would like to thank Andy, Bo, Jamie, Jenni, Kate, Kelly, Lisa, Mark, Mike, Paxton, Steve, and Visu, each of you helped make our days analyzing spectra, going to conferences, and studying for classes much more fun and rewarding. I was able to work with two very talented undergraduate students, Andy Kichner and Michael Bokemper; both contributed greatly to my research and this dissertation. Mike good luck in medical school and when you get done with that I have a gel for you to run. Thank you to Dave Nelson and Chris Frey for helping me with all my protein expression and purification questions, your experience and insight was greatly beneficial and appreciated. I would also like to extend a warm thank you to the talented collaborators I was lucky to work with, Dr. Peter Revesz and Thomas Triplet both were pivotal in the development of the PROFESS database and the structure comparison work described in chapter 6.

To all my family and friends who have supported me during this long and rigorous journey I am in much debt to you all. Mom and Dad, thank you and I love you both. You have always believed in me and have helped keep me on tract so that I could reach my goals. Every day I strive to model my life after your high level of honesty, integrity, unconditional love and determination. To my sisters Megan and Mindy, I love you both and wish you luck and fortune in the upcoming years. I know you both will go far.

To my grandparents, you have been the constant and solid foundation in my life and I love you all. The positive environment the four of you provided me while I was young has contributed greatly to my current success. Specifically, my love of nature and science is a direct consequence of my time spent outdoors with my grandparents, Orville, and JoAnn. Together they instilled in me a respect of the natural world that cannot be learned in any textbook. Additionally, I must attribute most of my mechanical abilities to my grandparents Harley and Rose. The summers spent working in their shop provided me with the troubleshooting skills and hard work ethic I now use every day.

To my new family Ron, Roxanne, and Veronica, you have taken me in and treated me as your own from day one. I love you and look forward to building new memories with you. I will even start rooting for the USC Trojans on football Saturdays, at least when they are not playing the Huskers!

Lastly, to my wife Ray, I love you with all my heart. Meeting you has been the luckiest and best thing to ever happen to me. Being able to share this journey with you and our two boys, Watson and Beaker, is truly a gift.

TABLE OF CONTENTS:

"Evolution is an obstacle course not a freeway…." S.J Gould

CHAPTER 1:

GENERAL INTRODUCTION

1.1 General introduction to functional genomics. Protein science has a long history inevitably intertwined with the advancements in chemistry, biology and physics. The term "protein" was initially used by Jöns Jakob Berzelius and Gerhardus Johannes Mulder who performed the first elemental analysis of a protein in 1839.[1] Surprisingly, all proteins Berzelius and Mulder studied contained the general empirical formula; $C_{400}H_{620}N_{100}O_{120.}$[1]

Nearly a century after Mulder's work, Jensen *et. al*. discovered the first amino acids in a protein.[2] This discovery eventually lead to the first complete amino acid sequence of a protein elucidated by F. Sanger in 1955.[3, 4] Sanger followed up his work on protein sequencing with developing techniques for DNA sequencing[5, 6] and successfully completed the first entire sequenced genome in 1977.[7] Twenty-two years later, *Haemophilus influenzae* became the first living organism to have its entire genome sequenced.[8] The following 6 years uncovered the complete genome sequences for *Escherichia coli*[9], *Drosophila melanogaster*[10], and in 2001 *Homo sapiens*.[11]

Since the first published genome in 1977[7] there has been an explosion in the number of complete genome sequences (figure 1.1). As of August 2010, a total of 1350 genomes have been completed and published representing all branches in the tree of life with nearly 6500 additional sequencing projects currently in progress.[12] In addition to individual species sequencing efforts, the technological advances in genome sequencing and relative low cost have help push the development of metagenomics. Metagenomics is the sequencing of samples collected directly from their environment. This has led to

the complete sequencing of the human gut "microbiome"[13] and the identification of various soil[14] and ocean[15, 16] microbes that could not be cultured in a laboratory setting.

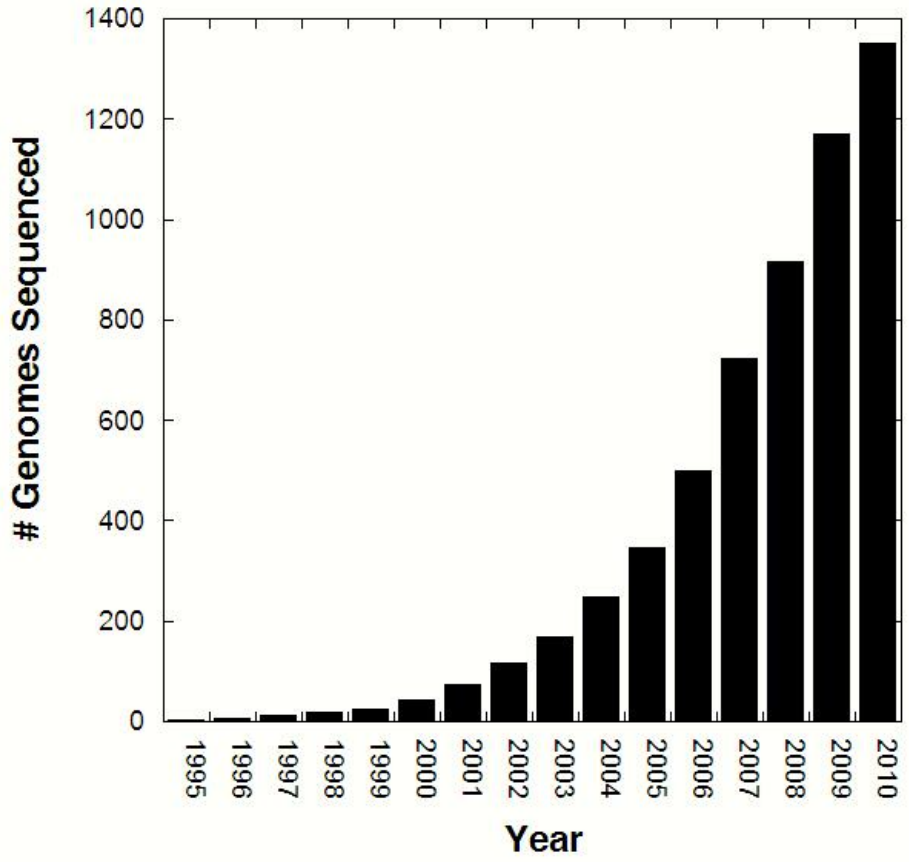


Figure 1.1 The rapid increase in sequenced genomes. Since the first genome of a living organism was sequenced in 1995 there has been a dramatic increase in the total number of completed genomes. The data was collected from the current status of the GOLD database[12] (August 2010) which listed a total of 1350 completed genomes.

A consequence of these sequencing projects was the discovery of large functionally unannotated segments of each genome. The genes (and proteins they encode) found in these unannotated regions are considered "hypothetical proteins". The term hypothetical protein is synonymous with novel gene product, unknown protein, non-characterized protein or putative uncharacterized gene product. Current estimates suggest

the percent of unannotated proteins found in all sequenced genomes is between 12%-50%.[17-19] For example, an estimated 50% of the genes in the *Escherichia coli* genome have not been experimentally annotated.[20, 21]

The large number of hypothetical proteins initially suggested these proteins were adaptations to specific environmental niches and therefore species specific.[22] Considering the large degree of biodiversity this seemed like a reasonable assumption.[22, 23] However, most hypothetical proteins are not species specific, but rather found in a range of phylogenetic distributions generating families of "conserved hypothetical proteins".[24] For example the *E. coli* hypothetical protein *yrdC* is a member of a hypothetical protein family. Homologous sequences to *yrdC* are also found in *Bacillus subtilis*, yeast, and humans.[24] Proteins such as *yrdC* are annotated as conserved hypothetical proteins because no member in the family is completely functionally annotated.[24, 25]

The most accurate and manually edited source for indentifying conserved hypothetical protein families, the Cluster of Orthologous Groups database (COG),[26] reports 2143 uncharacterized, putative or predicted orthologous families in bacteria.[25, 27] The large number of hypothetical and conserved hypothetical proteins significantly limits our understanding of biology. From a pragmatic viewpoint, identifying the functions of these proteins could lead to new therapeutics; making functional annotation of these proteins of paramount importance.

1.2 Introduction to protein functional annotation. The most basic level of functional annotation involves associating experimental evidence for a particular biochemical, biological process, or interaction to a specific gene. A number of experimental methods exist to annotate protein function. These include various

enzymatic assays,[28, 29] protein-protein interaction hybrid assays,[30, 31] knockout studies,[32, 33] gene silencing methods using antisense oligodeoxynucleotides,[34] ribozymes[35] or RNA interference [36-38] and recently metabolomic data.[39-41]

While powerful and direct, often a single biochemical method cannot fully annotate a gene. For example, with knockout and gene silencing studies a function is inferred from the change in observed phenotype between the wild-type and knockout organsim.[42] Knockout studies of essential genes are relatively straightforward with the appropriate control experiments because if the gene is no longer active the cell dies.[43] However, these studies only prove the knockout gene is essential for survival. These studies do not suggest a molecular function. For knockout studies of non-essential genes the issue becomes even more problematic. If multiple different genes carry out a particular function the knockout of gene may give no change in phenotype. Often this happens when a redundant gene compensates for the knockout.[44]

Functional annotations from enzymatic assays generally describe the substrate used in the study or reaction mechanism. For example, the general function ascribed to the enzyme responsible for catalyzing the oxidation of ethanol to acetaldehyde using nicotinamide adenine dinucleotide (NAD+) as a coenzyme, is alcohol dehydrogenase.[45] In humans, the alcohol dehydrogenase family consists of 7 unique genes each bind a range of alcohol substrates.[45] The problem becomes, if a gene has multiple *in vitro* functions, which one is the "correct" *in vivo* function? For alcohol dehydrogenase this problem is even larger with multiple genes binding a range of substrates.

The enzyme classification (EC) scheme attempts to standardize functional annotation from experimental methods.[46] The enzyme classification scheme annotates

proteins based on 6 broad functional classes (oxidoreductases, transferases, hydrolases, lyases isomerases and ligases). The functional annotation of the enzyme is further refined based on substrate and reaction chemistry. For the alcohol dehydrogenase example, all 7 genes in human are classified with the EC number of E.C 1.1.1.1 with each number designating a specific level of functional annotation (scheme 1.1)

Scheme 1.1. Example of enzyme classification (EC) nomenclature.

EC 1.1.1.1 Alcohol dehydrogenase

E.C 1.-.-.-	Oxidoreductases
E.C 1.1.-.-	Acting on the CH-OH group of donors
E.C 1.1.1.-	With NAD(+) or NADP(+) as acceptor
E.C 1.1.1.1	Alcohol dehydrogenase

The EC method provides a concise method to annotate experimental functions down to specific reaction chemistry. However, the problem becomes, what level of enzyme activity (K_m, V_{max} etc…) is needed to assign an EC number? Additionally, for *Escherichia coli*, only 30% of the genome encodes for enzymes, the remaining 70% encodes for transport proteins, response regulators, structural proteins, and other non-enzyme functions.[47]

The sheer number of unannotated proteins significantly limits complete biochemical analysis of every gene within an organism. A search of the NCBI protein sequence database for the term “hypothetical” retrieves nearly 1.5 million hits (August 2010). Correspondingly, nearly 2730 unique structures deposited in the Protein Data Bank (PDB) are annotated as hypothetical (August 2010). The large number of unannotated proteins makes pure experimental work impractical and supports the necessity of bioinformatics and hybrid bioinformatic/experimental methods.[21]

Since the early stages of protein and gene sequencing, it was shown sequences directly relate to the evolution of a protein and in some instances the organism.[48-51] This triggered the development of many sequence comparison methods attempting to accurately measure sequence relatedness.[52-59] Today, multiple sequence alignments (MSA) are routinely used to identify sequence similarity, build phylogenetic relationships, and to measure evolutionary clocks.[60-65] In addition to sequence based approaches, the three dimensional structure of a protein is related to molecular and organism evolution. A number of reports have shown a protein structure can also generate structure based phylogenetic trees[66-70] and protein domain complexity scales with organism complexity.[71] Mapping the evolutionary relationship between proteins is fundamental to current automatic functional annotation methods.

Current bioinformatic methods for functional annotation rely on gene and protein sequence, structure or hybrid sequence/structure similarity searches to automatically annotate protein function.[72] These methods use the evolutionary conservation of a protein to infer a generalized function; 'inheritance through homology'.[73-75] Homology is a hypothesis of the evolutionary relatedness between two or more proteins based on relative sequence or structure similarities.[76, 77] The degree sequence similarity needed to infer homology is still being debated. However, for highly similar sequences ($\geq$ 70%) this method is effective at annotating function.[78]

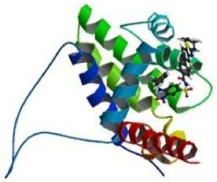
Functional annotation using homology transfer is the standard method of automated functional prediction. Many databases exist for automated functional prediction including, PFAM,[79] Gene Ontology,[80-82] UniProt/RefSeq/Swiss-Prot,[83] ProFunc,[84] and STRING.[85] Similarly, the COG/KOG[26] and eggNOG[17] databases often

get used for functional prediction because they contain large sets of orthologous genes. These databases and others have been reviewed in depth previously.[86] Each database uses different methods of protein representation, different algorithms for comparison and different scoring functions, in the majority of cases the result is a generalized functional annotation.

These automated functional annotation tools are necessary for managing the large volume of sequence data and remain the most popular.[87] However, these methods often lead to spurious annotations because homology does not necessarily imply conservation of function.[88] Additionally, these methods are often error prone and based on a small set of experimentally annotated proteins.[87, 89-92] The maximum reported error rate for automatic functional annotation is 63% for all unannotated genes.[87, 91] For enzymes approximately 30% of current automatic functional annotations are incorrect.[90] Differences in protein active site structure leading to different ligand specificities and enzyme efficiencies are suspected to be a major source of errors in automatic functional annotations.[78, 90, 93]

In addition to the problems stated above, many of the automatic function prediction methods have reached an apparent maximum effectiveness.[94] Essentially, (i) proteins with known function become overly populated in the databases so no new information is reported, or (ii) hypothetical proteins only match other hypothetical proteins. Figure 1.2 shows a structure based similarity search of the protein Bcl-xL, which only retrieves other Bcl-xL proteins. Alternatively, a search of a hypothetical protein YtfP from *E. coli* only retrieves other proteins of unknown function. Similar results are obtained using sequence similarity.

A



PDB	Z	%	Protein
1ysn-A	30.2	100	APOPTOSIS REGULATOR BCL-X;
2o1y-A	29.7	100	APOPTOSIS REGULATOR BCL-X;
2o2m-A	23.4	98	APOPTOSIS REGULATOR BCL-X;
2o2n-A	22.3	98	APOPTOSIS REGULATOR BCL-X;
1r2i-A	21	99	APOPTOSIS REGULATOR BCL-X;
1r2g-A	20.9	98	APOPTOSIS REGULATOR BCL-X;
1af3	20.9	98	APOPTOSIS REGULATOR BCL-X;
1af3-A	20.9	98	APOPTOSIS REGULATOR BCL-X;
1r2h-A	20.8	99	APOPTOSIS REGULATOR BCL-X;
1maz	20.7	99	BCL-XL;
1maz-A	20.7	99	BCL-XL;
1r2d-A	20.6	99	APOPTOSIS REGULATOR BCL-X;
3cva-X	20.5	98	APOPTOSIS REGULATOR BCL-X;
1r2e-A	20.4	99	APOPTOSIS REGULATOR BCL-X;
2yxj-B	20.2	100	APOPTOSIS REGULATOR BCL-X;
2yxj-A	20.2	100	APOPTOSIS REGULATOR BCL-X;
1pq0-A	20.1	97	APOPTOSIS REGULATOR BCL-X;
2bzw-A	19.1	94	APOPTOSIS REGULATOR BCL-X;
1pq1-A	18.9	95	APOPTOSIS REGULATOR BCL-X;

B

PDB	Z	%id	Protein
1xhs-A	25.5	100	HYPOTHETICAL UPF0131 PROTEIN YTFP;
1v30-A	12.3	31	HYPOTHETICAL UPF0131 PROTEIN PH0828;
2qik-A	10.2	23	UPF0131 PROTEIN YKQA;
1vkb-A	9.4	20	HYPOTHETICAL PROTEIN;
2g0q-A	9.2	23	AT5G39720.1 PROTEIN;
2kl2-A	9	20	AIG2-LIKE DOMAIN-CONTAINING PROTEIN 1;
2i5t-A	8.9	17	PROTEIN C7ORF24;
2pn7-B	8.7	17	HUMAN GAMMA-GLUTAMYL CYCLOTRANSFERASE;
2i5t-B	8.6	17	PROTEIN C7ORF24;
2rbh-A	8.6	16	GAMMA-GLUTAMYL CYCLOTRANSFERASE;
2q53-B	8.6	17	UNCHARACTERIZED PROTEIN C7ORF24;
2rbh-B	8.6	17	GAMMA-GLUTAMYL CYCLOTRANSFERASE;
2pn7-A	8.6	17	HUMAN GAMMA-GLUTAMYL CYCLOTRANSFERASE;
3cry-A	8.6	17	GAMMA-GLUTAMYL CYCLOTRANSFERASE;
3cry-B	8.5	17	GAMMA-GLUTAMYL CYCLOTRANSFERASE;
2q53-A	8.5	17	UNCHARACTERIZED PROTEIN C7ORF24;
2jqv-A	8.2	22	AIG2 PROTEIN-LIKE;
1s5j-A	2.5	6	DNA POLYMERASE I;
1nrk-A	2	16	YGFZ PROTEIN;

Figure 1.2 Structure based similarity searching to predict protein function. (A) The anti-apoptosis protein Bcl-xL (1YSN) was compared to the Dali FSSP database[95-98] to identify potential new functions. The only significant hits (Z>2.0) were redundantly solved protein structures of the same sequence or Bcl-2 homologs. (B) The structure of the hypothetical protein YtfP (1XHS) was compared to the same database with the most significant hits having no known function or a range of predicted functions. This example highlights two common problems with current structure based database searches: (i) a protein with known function is overly populated so no new information is reported, or (ii) hypothetical proteins only match other hypothetical proteins.

1.3 Annotation of function using ligand binding. A major source of error in automatic function prediction is differences in active site structure and ligand specificity. Could using active site information increase functional annotation? Proteins interact with biological molecules including other proteins, DNA, RNA or small molecule ligands. Therefore, the active site of a protein must be intrinsically linked to the function of the protein.[99]

Active site similarity tools for functional prediction and annotation are a rapidly growing trend.[100-106] Using ligands to probe protein function is an evolutionary independent method to predict protein function. This should reduce the error rate of traditional homology based methods because active site annotations are not limited to correct ortholog detection.[107] Additionally, traditional homology based methods do not account for post-translational modifications or the occurrence of gene sharing. Both of which have dramatic biological significance.[108-110]

A corollary to function prediction using ligand binding is using similar functions to predict off-target side effects of drugs.[105, 106, 111, 112] Recent observations of potential drug leads binding a range of protein targets with similar function further support the idea of using ligand binding to predict protein function. Attempts have been made to relate ligand binding to sequence or structure similarity with minimal success.[113] To date only a handful of studies have attempted to relate ligand binding with protein function.[106, 114-117]

The work reported in this dissertation uses this most basic definition of protein function to establish a uniform method for identifying functional similarity. The hypothesis is proteins with similar function will bind to a set of similar biologically relevant small molecules at a specific active site. The hypothesis is supported by reports

showing functional regions of a protein are more stable relative to the remainder of the protein sequence undergoing random drift.[118, 119] The correlation between ligand binding sites, ligand structure and protein function has also been demonstrated by a network of ligand binding-site.[120] A variety of computational methods have attempted to exploit the stability of functional regions by identifying ligand binding sites as a method to predict function.[121, 122] Unfortunately, the combined requirements of predicting the ligand, the binding site, and a similarity to an annotated proteins leads to a high level of ambiguity.

This dissertation will discuss the development of high-throughput screening methods to detect ligand binding and discovery protein active sites. There is an inherent similarity between the methods used to detect ligand binding for functional annotation and drug discovery. In this dissertation the high-throughput NMR screening method to detect ligand binding were originally developed to identify binding ligands and protein active sites for attractive drug targets.

Drug discovery is a uniquely complex problem in science and medicine.[123, 124] This is further complicated by the fact that each disease is distinct and requires its own efficient strategy to successfully develop safe therapeutics.[125] A central theme in drug discovery research is attempting to identify highly specific ligands (nM-pM K_D) that bind a biological target. Therefore, the methods used to detect ligand binding in drug discovery research are also amenable to identifying binding ligands for functional annotation. In this dissertation, the techniques developed for high-throughput NMR screening were used to identify binding ligand to a number of functionally diverse proteins including, serum albumins (*H. sapiens*, *B. taurus*), α and β amylases (*B. licheniformis, A. oryzae, B. amyloliquefaciens H. vulgare, I. batatas*), primase C-terminal

domain (*S. aureus*), nuclease (*S. aureus*) and the type three secretion system protein PrgI (*S. typhirium*).

1.4 General principles of high-throughput nuclear magnetic resonance screening. From target selection to pre-clinical trials, nuclear magnetic resonance (NMR) has established itself as an invaluable tool for the chemist working in the drug discovery industry.[126, 127] The flexibility provided by NMR comes from various molecular probes that include chemical shifts, relaxation parameters (T_1, T_2), spatial information (nuclear Overhauser effects, NOE), and diffusion rates. Each parameter is uniquely sensitive to the local chemical and physical environment of a sample and provides structural information at atomic resolution for both small (<1000 Da) and large (> 1000 Da) biological molecules. Additionally, in recent years NMR has proven more valuable to the drug discovery process than simply a tool for structural studies of biological molecules.[126] This is most apparent with the increase use of NMR as a critical component for high-throughput screening (HTS). The current methods for NMR affinity screening methods are listed in table 1.1.

NMR affinity screening methods complement structural biology efforts by validating chemical leads prior to initiating a structure-based drug design program.[128-133] Target focused screening techniques, such as SAR by NMR,[134] RAMPED-UP NMR,[135] STD-NMR,[136] and NMR-SOLVE[137] were developed to identify ligands that bind a therapeutic target in a biologically relevant manner (table 1.1). This is often done by observing chemical shift changes in two-dimensional 2D ^{1}H-^{15}N HSQC spectra of the protein in the presence and absence of a small molecule ligand. Target based NMR screening methods provide invaluable information about the nature of a ligand binding

site.[134, 138-143] However, these methods often require high concentrations (≥ 100 μM) of expensive ^{15}N isotope enriched protein and demand large amounts of data collection time. Therefore target based screening methods are often better suited for secondary follow-up screens.[115, 144, 145]

Unlike the target focused methods, ligand focused techniques detect binding events by identifying changes in the free ^{1}H ligand spectrum upon the addition of a protein. Many ligand focused methods have been developed that exploit various NMR molecular probes including saturation transfer differences,[136, 146] line-broadening changes,[147-150] diffusion rate changes,[151] ^{19}F NMR,[149, 152] spin labels,[150] and transfer NOEs[153] (table 1.1). The ligand focused methods are relatively quick (1-5 min), do not require ^{15}N enriched samples, and are sensitive at much lower protein concentrations (≤ 5 μM). Therefore, these methods have rapidly become invaluable to high-throughput screening with a high success rate of identifying potential inhibitors.[154-159]

This dissertation will focus on using and developing screening methods for high-throughput NMR screening to functionally annotate proteins with no known function. The two central methods used are the 1D line broadening experiment and the 2D ^{1}H-^{15}N HSQC experiment. These methods will be used in tandem for each screen to detect and confirm ligand binding. The tiered approach to NMR screening maximizes the total number of hits identified while reducing the overall sample and data collection requirements.[144]

Table 1.1 Various NMR screening methods and the NMR parameter used to detect ligand binding.

Screening Technique	NMR Parameter used to Detect Ligand Binding	Labeled Protein?	Limited by Protein MW?	Ref.
MS/NMR	Retention on size-exclusion column & chemical shift changes	Yes	Yes	145
Multi-Step NMR	Line-broadening change (T_2) & chemical shift changes	Yes	Yes	144
RAMPED-UP NMR	Chemical shift changes, screening multiple proteins	Yes	Yes	135
SAR by NMR	Chemical shift changes	Yes	Yes	134
SMILI-NMR	In-cell chemical shift changes	Yes	Yes	160
STINT-NMR	In-cell chemical shift changes	Yes	Yes	161, 162
SLAPSTIC	Line-broadening change (T_2) due to protein spin label	Yes	No	150
AIDA-NMR	Line-broadening change (T_2) due to protein-protein complex formation, labeled protein or Trp reporter in ligand binding site	Yes/No	Yes	147, 148
TINS	Line-broadening change (T_2) due to binding to an immobilized protein target	No	Yes	163
3-FABS	Chemical shift changes, requires fluorinated ligands	No	No	152
Affinity NMR	Change in translational diffusion	No	No	151
FAXS	Line-broadening change (T_2) due to ligand competition, requires fluorinated ligands	No	No	149
INPHARMA	Transfer nuclear Overhauser effect (NOE)	No	No	164
NOE pumping	Transfer nuclear Overhauser effect (NOE)	No	No	153
SALMON	Saturation transfer difference from solvent	No	No	165
STD NMR	Saturation transfer difference from protein	No	No	136
WaterLOGSY	Saturation transfer difference from solvent	No	No	146

1.5 Summary of work. The challenges of functional annotation described above are a product of having only a limited collection of bioinformatic tools based on a small set of experimentally characterized proteins. This dissertation focuses on the development and implementation of new experimental approaches to extend functional annotation of unknown proteins. First, in chapter 2 I will discuss the development of a technique to measure relative dissociation constants (K_D) from an NMR high-throughput ligand affinity screen. The method is used to qualitatively select the best binding ligand(s) that will be used to probe the active sites of the various targets and identify a biological function.

Chapters 3 and 4 will discuss the implementation and optimization of the Functional Annotation Screening Technology by NMR (FAST-NMR).[166] The FAST-NMR method is a tiered approach to high-throughput NMR affinity screening to identify binding ligands and proteins active sites. The protein active sites are compared to a database of active sites using the Comparison of Active Site Similarity (CPASS) tool.[103] Functional similarity is inferred through similarities in protein active sites.

In chapter 3 I show the utility of the FAST-NMR method by establishing a functional similarity between the type III secretion system (T3SS) protein PrgI from *S. typhirium* and the human apoptosis regulating protein Bcl-xL. This relationship would not have been identified with current methods because sequence and structure similarity are below the limit of acceptable homology. In chapter 4, I validate the FAST-NMR method by expressing, purifying and screening the *S. aureus* nuclease protein. I show the FAST-NMR method correctly identifies the best binding ligand thyamdine-5'-triphosphate and active site of the protein. Additionally, the FAST-NMR binding site

correctly identified a nucleotide (thymidine-3',5'-diphosphate) bound nuclease structure (1TR5) as the best match in the CPASS search. I will also discuss the implementation to two new pulse sequences and improvements to automated data collection. These improvements to the screening technology dramatically increase throughput and flexibility of FAST-NMR.

The FAST-NMR method was initially developed as a tool for functional annotation. However, the generalized tiered approach to NMR screening used by FAST-NMR is also valuable to drug discovery. In chapter 5 I will discuss the structure, dynamics and high-throughput screening of the DnaG primase C-terminal domain from *Staphylococcus aureus*. The C-terminal domain of primase specifically interacts with the DnaC helicase to initiate primer synthesis and is therefore an attractive drug target for antibiotic development. Using the FAST-NMR screening methods I show acycloguanosine binds to the C-terminal domain of primase at the important helicase interaction site. This result was used to identify a set of structurally similar compounds for further antibiotic development.

A surprising result from the *S. aureus* primase CTD structure was the observation of a potential phylogenetic dependence on protein structure similarity. In chapter 6 I expand on this observation by completing a thorough analysis of functionally identical protein structures and report a maximum sequence and structure similarity between the two bacterial phyla, *Firmicutes* and *Proteobacteria*. Additionally, the results from chapter 6 were used to show a constant rate of structural drift during protein evolution.

Finally, in chapter 7 I discuss a new technique for functional annotation that evolved from the FAST-NMR methodology, but is independent of sequence, structure or

evolutionary information. The method involves the development of a robust scoring system to measure ligand binding profile similarities. A ligand binding profile is defined as a set of ligands that bind a protein from a high-throughput ligand affinity screen using a standardized chemical library. Functional annotation is inferred by clustering unknown proteins with previously annotated proteins that share similar ligand binding profiles. The method was tested on two sets of control proteins, 2 serum albumins (*H. sapiens, B. taurus*) and 5 amylases (*B. licheniformis, A. oryzae, B. amyloliquefaciens H. vulgare, I. batatas*).

1.6 REFERENCES

1. Teich, M.; Needham, D. M., *A Documentary History of Biochemistry 1770-1940* Fairleigh Dickinson University Press 1992.
2. Jensen, H.; Evans, E. A., Studies on Crystalline Insulin: The Nature of the free amino groups in insulin and the isolation of phenylalanine and proline from crystalline insulin. *Journal of Biological Chemistry* **1934,** 108, (1), 1-9.
3. Ryle, A. P.; Sanger, F.; Smith, L. F.; Kitai, R., The disulphide bonds of insulin. *Biochem J* **1955,** 60, (4), 541-56.
4. Sanger, F.; Thompson, E. O.; Kitai, R., The amide groups of insulin. *Biochem J* **1955,** 59, (3), 509-18.
5. Sanger, F.; Donelson, J. E.; Coulson, A. R.; Kossel, H.; Fischer, D., Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. *Proc Natl Acad Sci U S A* **1973,** 70, (4), 1209-13.

6. Sanger, F.; Donelson, J. E.; Coulson, A. R.; Kossel, H.; Fischer, D., Determination of a nucleotide sequence in bacteriophage f1 DNA by primed synthesis with DNA polymerase. *J Mol Biol* **1974,** 90, (2), 315-33.

7. Sanger, F.; Coulson, A. R.; Friedmann, T.; Air, G. M.; Barrell, B. G.; Brown, N. L.; Fiddes, J. C.; Hutchison, C. A., 3rd; Slocombe, P. M.; Smith, M., The nucleotide sequence of bacteriophage phiX174. *J Mol Biol* **1978,** 125, (2), 225-46.

8. Fleischmann, R. D.; Adams, M. D.; White, O.; Clayton, R. A.; Kirkness, E. F.; Kerlavage, A. R.; Bult, C. J.; Tomb, J. F.; Dougherty, B. A.; Merrick, J. M.; et al., Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **1995,** 269, (5223), 496-512.

9. Blattner, F. R.; Plunkett, G., 3rd; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A.; Rose, D. J.; Mau, B.; Shao, Y., The complete genome sequence of Escherichia coli K-12. *Science* **1997,** 277, (5331), 1453-62.

10. Adams, M. D.; Celniker, S. E.; Holt, R. A.; Evans, C. A.; Gocayne, J. D.; Amanatides, P. G.; Scherer, S. E.; Li, P. W.; Hoskins, R. A.; Galle, R. F.; George, R. A.; Lewis, S. E.; Richards, S.; Ashburner, M.; Henderson, S. N.; Sutton, G. G.; Wortman, J. R.; Yandell, M. D.; Zhang, Q.; Chen, L. X.; Brandon, R. C.; Rogers, Y. H.; Blazej, R. G.; Champe, M.; Pfeiffer, B. D.; Wan, K. H.; Doyle, C.; Baxter, E. G.; Helt, G.; Nelson, C. R.; Gabor, G. L.; Abril, J. F.; Agbayani, A.; An, H. J.; Andrews-Pfannkoch, C.; Baldwin, D.; Ballew, R. M.; Basu, A.; Baxendale, J.; Bayraktaroglu, L.; Beasley, E. M.; Beeson, K. Y.; Benos, P. V.; Berman, B. P.;

Bhandari, D.; Bolshakov, S.; Borkova, D.; Botchan, M. R.; Bouck, J.; Brokstein, P.; Brottier, P.; Burtis, K. C.; Busam, D. A.; Butler, H.; Cadieu, E.; Center, A.; Chandra, I.; Cherry, J. M.; Cawley, S.; Dahlke, C.; Davenport, L. B.; Davies, P.; de Pablos, B.; Delcher, A.; Deng, Z.; Mays, A. D.; Dew, I.; Dietz, S. M.; Dodson, K.; Doup, L. E.; Downes, M.; Dugan-Rocha, S.; Dunkov, B. C.; Dunn, P.; Durbin, K. J.; Evangelista, C. C.; Ferraz, C.; Ferriera, S.; Fleischmann, W.; Fosler, C.; Gabrielian, A. E.; Garg, N. S.; Gelbart, W. M.; Glasser, K.; Glodek, A.; Gong, F.; Gorrell, J. H.; Gu, Z.; Guan, P.; Harris, M.; Harris, N. L.; Harvey, D.; Heiman, T. J.; Hernandez, J. R.; Houck, J.; Hostin, D.; Houston, K. A.; Howland, T. J.; Wei, M. H.; Ibegwam, C.; Jalali, M.; Kalush, F.; Karpen, G. H.; Ke, Z.; Kennison, J. A.; Ketchum, K. A.; Kimmel, B. E.; Kodira, C. D.; Kraft, C.; Kravitz, S.; Kulp, D.; Lai, Z.; Lasko, P.; Lei, Y.; Levitsky, A. A.; Li, J.; Li, Z.; Liang, Y.; Lin, X.; Liu, X.; Mattei, B.; McIntosh, T. C.; McLeod, M. P.; McPherson, D.; Merkulov, G.; Milshina, N. V.; Mobarry, C.; Morris, J.; Moshrefi, A.; Mount, S. M.; Moy, M.; Murphy, B.; Murphy, L.; Muzny, D. M.; Nelson, D. L.; Nelson, D. R.; Nelson, K. A.; Nixon, K.; Nusskern, D. R.; Pacleb, J. M.; Palazzolo, M.; Pittman, G. S.; Pan, S.; Pollard, J.; Puri, V.; Reese, M. G.; Reinert, K.; Remington, K.; Saunders, R. D.; Scheeler, F.; Shen, H.; Shue, B. C.; Siden-Kiamos, I.; Simpson, M.; Skupski, M. P.; Smith, T.; Spier, E.; Spradling, A. C.; Stapleton, M.; Strong, R.; Sun, E.; Svirskas, R.; Tector, C.; Turner, R.; Venter, E.; Wang, A. H.; Wang, X.; Wang, Z. Y.; Wassarman, D. A.; Weinstock, G. M.; Weissenbach, J.; Williams, S. M.; WoodageT; Worley, K. C.; Wu, D.; Yang, S.; Yao, Q. A.; Ye, J.; Yeh, R. F.; Zaveri, J. S.; Zhan, M.; Zhang, G.; Zhao,

Q.; Zheng, L.; Zheng, X. H.; Zhong, F. N.; Zhong, W.; Zhou, X.; Zhu, S.; Zhu, X.; Smith, H. O.; Gibbs, R. A.; Myers, E. W.; Rubin, G. M.; Venter, J. C., The genome sequence of Drosophila melanogaster. *Science* **2000,** 287, (5461), 2185-95.

11. Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J. C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, E. R.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J. F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R. A.; Muzny, D. M.; Scherer, S. E.; Bouck, J. B.; Sodergren, E. J.; Worley, K. C.; Rives, C. M.; Gorrell, J. H.; Metzker, M. L.; Naylor, S. L.; Kucherlapati, R.

S.; Nelson, D. L.; Weinstock, G. M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D. R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H. M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R. W.; Federspiel, N. A.; Abola, A. P.; Proctor, M. J.; Myers, R. M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D. R.; Olson, M. V.; Kaul, R.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G. A.; Athanasiou, M.; Schultz, R.; Roe, B. A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W. R.; de la Bastide, M.; Dedhia, N.; Blocker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J. A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D. G.; Burge, C. B.; Cerutti, L.; Chen, H. C.; Church, D.; Clamp, M.; Copley, R. R.; Doerks, T.; Eddy, S. R.; Eichler, E. E.; Furey, T. S.; Galagan, J.; Gilbert, J. G.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L. S.; Jones, T. A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W. J.; Kitts, P.; Koonin, E. V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T. M.; McLysaght, A.; Mikkelsen, T.; Moran, J. V.; Mulder, N.; Pollara, V. J.; Ponting, C. P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A. F.; Stupka, E.; Szustakowski, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y. I.; Wolfe, K. H.; Yang, S. P.; Yeh, R. F.; Collins, F.; Guyer, M. S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K. A.; Patrinos, A.;

Morgan, M. J.; de Jong, P.; Catanese, J. J.; Osoegawa, K.; Shizuya, H.; Choi, S.; Chen, Y. J., Initial sequencing and analysis of the human genome. *Nature* **2001,** 409, (6822), 860-921.

12. Liolios, K.; Chen, I. M.; Mavromatis, K.; Tavernarakis, N.; Hugenholtz, P.; Markowitz, V. M.; Kyrpides, N. C., The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **2009,** 38, (Database issue), D346-54.
13. Ley, R. E.; Lozupone, C. A.; Hamady, M.; Knight, R.; Gordon, J. I., Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* **2008,** 6, (10), 776-88.
14. Daniel, R., The metagenomics of soil. *Nat Rev Microbiol* **2005,** 3, (6), 470-8.
15. Nealson, K. H.; Venter, J. C., Metagenomics and the global ocean survey: what's in it for us, and why should we care? *Isme J* **2007,** 1, (3), 185-7.
16. Bohannon, J., Metagenomics. Ocean study yields a tidal wave of microbial DNA. *Science* **2007,** 315, (5818), 1486-7.
17. Muller, J.; Szklarczyk, D.; Julien, P.; Letunic, I.; Roth, A.; Kuhn, M.; Powell, S.; von Mering, C.; Doerks, T.; Jensen, L. J.; Bork, P., eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res* **2010,** 38, (Database issue), D190-5.
18. Jensen, L. J.; Julien, P.; Kuhn, M.; von Mering, C.; Muller, J.; Doerks, T.; Bork, P., eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* **2008,** 36, (Database issue), D250-4.

19. Sivashankari, S.; Shanmughavel, P., Functional annotation of hypothetical proteins - A review. *Bioinformation* **2006,** 1, (8), 335-8.

20. Galperin, M. Y.; Koonin, E. V., From complete genome sequence to 'complete' understanding? *Trends Biotechnol* **2010,** 28, (8), 398-406.

21. Kolker, E.; Makarova, K. S.; Shabalina, S.; Picone, A. F.; Purvine, S.; Holzman, T.; Cherny, T.; Armbruster, D.; Munson, R. S., Jr.; Kolesov, G.; Frishman, D.; Galperin, M. Y., Identification and functional analysis of 'hypothetical' genes expressed in Haemophilus influenzae. *Nucleic Acids Res* **2004,** 32, (8), 2353-61.

22. Daubin, V.; Ochman, H., Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. *Genome Res* **2004,** 14, (6), 1036-42.

23. Siew, N.; Fischer, D., Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins* **2003,** 53, (2), 241-51.

24. Galperin, M. Y., Conserved 'hypothetical' proteins: new hints and new puzzles. *Comp Funct Genomics* **2001,** 2, (1), 14-8.

25. Galperin, M. Y.; Koonin, E. V., 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res* **2004,** 32, (18), 5452-63.

26. Tatusov, R. L.; Natale, D. A.; Garkavtsev, I. V.; Tatusova, T. A.; Shankavaram, U. T.; Rao, B. S.; Kiryutin, B.; Galperin, M. Y.; Fedorova, N. D.; Koonin, E. V., The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **2001,** 29, (1), 22-8.

27. Tatusov, R. L.; Fedorova, N. D.; Jackson, J. D.; Jacobs, A. R.; Kiryutin, B.; Koonin, E. V.; Krylov, D. M.; Mazumder, R.; Mekhedov, S. L.; Nikolskaya, A. N.; Rao, B. S.; Smirnov, S.; Sverdlov, A. V.; Vasudevan, S.; Wolf, Y. I.; Yin, J.

J.; Natale, D. A., The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **2003,** 4, 41.

28. Benz, E. W., Jr.; Reinberg, D.; Vicuna, R.; Hurwitz, J., Initiation of DNA replication by the dnaG protein. *J Biol Chem* **1980,** 255, (3), 1096-106.

29. Sims, J.; Benz, E. W., Jr., Initiation of DNA replication by the Escherichia coli dnaG protein: evidence that tertiary structure is involved. *Proc Natl Acad Sci U S A* **1980,** 77, (2), 900-4.

30. Arifuzzaman, M.; Maeda, M.; Itoh, A.; Nishikata, K.; Takita, C.; Saito, R.; Ara, T.; Nakahigashi, K.; Huang, H. C.; Hirai, A.; Tsuzuki, K.; Nakamura, S.; Altaf-Ul-Amin, M.; Oshima, T.; Baba, T.; Yamamoto, N.; Kawamura, T.; Ioka-Nakamichi, T.; Kitagawa, M.; Tomita, M.; Kanaya, S.; Wada, C.; Mori, H., Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res* **2006,** 16, (5), 686-91.

31. Schwikowski, B.; Uetz, P.; Fields, S., A network of protein-protein interactions in yeast. *Nat Biotechnol* **2000,** 18, (12), 1257-61.

32. Capecchi, M. R., Altering the genome by homologous recombination. *Science* **1989,** 244, (4910), 1288-92.

33. Capecchi, M. R., The new mouse genetics: altering the genome by gene targeting. *Trends Genet* **1989,** 5, (3), 70-6.

34. Liao, L.; Li, Z., Correlation between gene silencing activity and structural features of antisense oligodeoxynucleotides and target RNA. *In Silico Biol* **2007,** 7, (4-5), 527-34.

35. Welch, P. J.; Marcusson, E. G.; Li, Q. X.; Beger, C.; Kruger, M.; Zhou, C.; Leavitt, M.; Wong-Staal, F.; Barber, J. R., Identification and validation of a gene involved in anchorage-independent cell growth control using a library of randomized hairpin ribozymes. *Genomics* **2000,** 66, (3), 274-83.

36. Fire, A. Z., Gene silencing by double-stranded RNA. *Cell Death Differ* **2007,** 14, (12), 1998-2012.

37. Conte, D., Jr.; Mello, C. C., RNA interference in Caenorhabditis elegans. *Curr Protoc Mol Biol* **2003,** Chapter 26, Unit 26 3.

38. Mello, C. C.; Conte, D., Jr., Revealing the world of RNA interference. *Nature* **2004,** 431, (7006), 338-42.

39. Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E., Metabonomics: a platform for studying drug toxicity and gene function. *Nat Rev Drug Discov* **2002,** 1, (2), 153-61.

40. Wu, L.; van Winden, W. A.; van Gulik, W. M.; Heijnen, J. J., Application of metabolome data in functional genomics: a conceptual strategy. *Metab Eng* **2005,** 7, (4), 302-10.

41. Raamsdonk, L. M.; Teusink, B.; Broadhurst, D.; Zhang, N.; Hayes, A.; Walsh, M. C.; Berden, J. A.; Brindle, K. M.; Kell, D. B.; Rowland, J. J.; Westerhoff, H. V.; van Dam, K.; Oliver, S. G., A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat Biotechnol* **2001,** 19, (1), 45-50.

42. Consortium, T. I. M. K., A Mouse for All Reasons. *Cell* **2007,** 129, (1), 9-13.

43. Knuth, K.; Niesalla, H.; Hueck, C. J.; Fuchs, T. M., Large-scale identification of essential Salmonella genes by trapping lethal insertions. *Mol Microbiol* **2004,** 51, (6), 1729-44.

44. Deutscher, D.; Meilijson, I.; Schuster, S.; Ruppin, E., Can single knockouts accurately single out gene functions? *BMC Systems Biology* **2008,** 2, (1), 50.

45. Wagner, F. W.; Burger, A. R.; Vallee, B. L., Kinetic properties of human liver alcohol dehydrogenase: oxidation of alcohols by class I isoenzymes. *Biochemistry* **1983,** 22, (8), 1857-63.

46. *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the No (Enzyme Nomenclature)* 1ed.; Academic Press: San Diego, 1992; p 862.

47. Riley, M.; Serres, M. H., Interim report on genomics of Escherichia coli. *Annu Rev Microbiol* **2000,** 54, 341-411.

48. Fitch, W. M.; Margoliash, E., Construction of phylogenetic trees. *Science* **1967,** 155, (760), 279-84.

49. Fitch, W. M.; Margoliash, E., The construction of phylogenetic trees. II. How well do they reflect past history? *Brookhaven Symp Biol* **1968,** 21, (1), 217-42.

50. Prager, E. M.; Wilson, A. C., Construction of phylogenetic trees for proteins and nucleic acids: empirical evaluation of alternative matrix methods. *J Mol Evol* **1978,** 11, (2), 129-42.

51. Woese, C. R., Bacterial evolution. *Microbiol Rev* **1987,** 51, (2), 221-71.

52. Smith, T. F.; Waterman, M. S., Identification of common molecular subsequences. *J Mol Biol* **1981,** 147, (1), 195-7.

53. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *J Mol Biol* **1990,** 215, (3), 403-10.

54. Karlin, S.; Altschul, S. F., Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A* **1990,** 87, (6), 2264-8.

55. Needleman, S. B.; Wunsch, C. D., A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **1970,** 48, (3), 443-53.

56. Fitch, W. M., An improved method of testing for evolutionary homology. *J Mol Biol* **1966,** 16, (1), 9-16.

57. Dayhoff, M. O., Computer analysis of protein evolution. *Sci Am* **1969,** 221, (1), 86-95.

58. Churchill, G. A., Stochastic models for heterogeneous DNA sequences. *Bull Math Biol* **1989,** 51, (1), 79-94.

59. Pearson, W. R.; Lipman, D. J., Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **1988,** 85, (8), 2444-8.

60. Feng, J.-a., Improving pairwise sequence alignment between distantly related proteins. *Methods Mol. Biol. (Totowa, NJ, U. S.)* **2007,** 395, (Comparative Genomics, Volume 1), 255-268.

61. Chang, G. S.; Hong, Y.; Dae Ko, K.; Bhardwaj, G.; Holmes, E. C.; Patterson, R. L.; van Rossum, D. B., Phylogenetic profiles reveal evolutionary relationships within the "twilight zone" of sequence similarity. *Proc Natl Acad Sci USA* **2008,** 105, (36), 13474-13479, S13474/1-S13474/14.

62. Thompson, J. D.; Gibson, T. J.; Higgins, D. G., Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* **2002,** Chapter 2, Unit 2 3.

63. Konstantinidis, K. T.; Tiedje, J. M., Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **2005,** 187, (18), 6258-64.

64. Doolittle, R. F.; Feng, D. F.; Cho, G., Determining divergence times with protein clocks. *Biol Bull* **1999,** 196, (3), 356-7; discussion 357-8.

65. Doolittle, R. F.; Feng, D. F.; Tsang, S.; Cho, G.; Little, E., Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* **1996,** 271, (5248), 470-7.

66. Agarwal, G.; Rajavel, M.; Gopal, B.; Srinivasan, N., Structure-based phylogeny as a diagnostic for functional characterization of proteins with a cupin fold. *PLoS One* **2009,** 4, (5), e5736.

67. Ribas de Pouplana, L.; Brown, J. R.; Schimmel, P., Structure-based phylogeny of class IIa tRNA synthetases in relation to an unusual biochemistry. *J Mol Evol* **2001,** 53, (4-5), 261-8.

68. Breitling, R.; Laubner, D.; Adamski, J., Structure-based phylogenetic analysis of short-chain alcohol dehydrogenases and reclassification of the 17beta-hydroxysteroid dehydrogenase family. *Mol Biol Evol* **2001,** 18, (12), 2154-61.

69. Johnson, M. S.; Sali, A.; Blundell, T. L., Phylogenetic relationships from three-dimensional protein structures. *Methods Enzymol* **1990,** 183, 670-90.

70. Johnson, M. S.; Sutcliffe, M. J.; Blundell, T. L., Molecular anatomy: phyletic relationships derived from three-dimensional structures of proteins. *J Mol Evol* **1990,** 30, (1), 43-59.

71. Yang, S.; Bourne, P. E., The evolutionary history of protein domains viewed by species phylogeny. *PLoS One* **2009,** 4, (12), e8378.

72. Rentzsch, R.; Orengo, C. A., Protein function prediction--the power of multiplicity. *Trends Biotechnol* **2009,** 27, (4), 210-9.

73. Lee, D.; Redfern, O.; Orengo, C., Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* **2007,** 8, (12), 995-1005.

74. Watson, J. D.; Laskowski, R. A.; Thornton, J. M., Predicting protein function from sequence and structural data. *Curr Opin Struct Biol* **2005,** 15, (3), 275-84.

75. Goonesekere, N. C.; Shipely, K.; O'Connor, K., The challenge of annotating protein sequences: The tale of eight domains of unknown function in Pfam. *Comput Biol Chem* **2010,** 34, (3), 210-214.

76. Fitch, W. M., Homology a personal view on some of the problems. *Trends Genet* **2000,** 16, (5), 227-31.

77. Reeck, G. R.; de Haen, C.; Teller, D. C.; Doolittle, R. F.; Fitch, W. M.; Dickerson, R. E.; Chambon, P.; McLachlan, A. D.; Margoliash, E.; Jukes, T. H.; et al., "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell* **1987,** 50, (5), 667.

78. Rost, B., Enzyme function less conserved than anticipated. *J Mol Biol* **2002,** 318, (2), 595-608.

79. Sonnhammer, E. L.; Eddy, S. R.; Durbin, R., Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **1997,** 28, (3), 405-420.

80. Blake, J. A.; Harris, M. A., The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis. *Curr Protoc Bioinformatics* **2008,** Chapter 7, Unit 7 2.

81. Camon, E.; Barrell, D.; Brooksbank, C.; Magrane, M.; Apweiler, R., The Gene Ontology Annotation (GOA) Project--Application of GO in SWISS-PROT, TrEMBL and InterPro. *Comp Funct Genomics* **2003,** 4, (1), 71-4.

82. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput Biol* **2009,** 5, (7), e1000431.

83. The UniProt Consortium, The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* **2009,** 37, (Database issue), D169-74.

84. Laskowski, R. A.; Watson, J. D.; Thornton, J. M., ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* **2005,** 33, (Web Server issue), W89-93.

85. von Mering, C.; Huynen, M.; Jaeggi, D.; Schmidt, S.; Bork, P.; Snel, B., STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* **2003,** 31, (1), 258-61.

86. Ashurst, J. L.; Collins, J. E., Gene annotation: prediction and testing. *Annu Rev Genomics Hum Genet* **2003,** 4, 69-88.

87. Schnoes, A. M.; Brown, S. D.; Dodevski, I.; Babbitt, P. C., Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* **2009,** 5, (12), e1000605.

88. Dessailly, B. H.; Redfern, O. C.; Cuff, A.; Orengo, C. A., Exploiting structural classifications for function prediction: towards a domain grammar for protein function. *Curr Opin Struct Biol* **2009,** 19, (3), 349-56.

89. Green, M. L.; Karp, P. D., Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Res* **2005,** 33, (13), 4035-9.

90. Devos, D.; Valencia, A., Intrinsic errors in genome annotation. *Trends Genet* **2001,** 17, (8), 429-31.

91. Brenner, S. E., Errors in genome annotation. *Trends Genet* **1999,** 15, (4), 132-3.

92. Andorf, C.; Dobbs, D.; Honavar, V., Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach. *BMC Bioinformatics* **2007,** 8, 284.

93. Tian, W.; Skolnick, J., How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* **2003,** 333, (4), 863-82.

94. Bork, P., Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Res* **2000,** 10, (4), 398-400.

95. Holm, L.; Kaariainen, S.; Rosenstrom, P.; Schenkel, A., Searching protein structure databases with DaliLite v.3. *Bioinformatics* **2008,** 24, (23), 2780-1.

96. Holm, L.; Sander, C., The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res* **1996,** 24, (1), 206-9.

97. Holm, L.; Sander, C., Mapping the protein universe. *Science* **1996,** 273, (5275), 595-603.

98. Holm, L.; Sander, C., Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res* **1997,** 25, (1), 231-4.

99. Thirlway, J.; Soultanas, P., In the Bacillus stearothermophilus DnaB-DnaG complex, the activities of the two proteins are modulated by distinct but overlapping networks of residues. *J Bacteriol* **2006,** 188, (4), 1534-9.

100. Gold, N. D.; Deville, K.; Jackson, R. M., New opportunities for protease ligand binding site comparisons using SitesBase. *Biochem Soc Trans* **2007,** 35, (Pt 3), 561-5.

101. Kinoshita, K.; Nakamura, H., eF-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics* **2004,** 20, (8), 1329-30.

102. Kuhn, D.; Weskamp, N.; Schmitt, S.; Hullermeier, E.; Klebe, G., From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J Mol Biol* **2006,** 359, (4), 1023-44.

103. Powers, R.; Copeland, J. C.; Germer, K.; Mercier, K. A.; Ramanathan, V.; Revesz, P., Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **2006,** 65, (1), 124-35.

104. Hoffmann, B.; Zaslavskiy, M.; Vert, J. P.; Stoven, V., A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinformatics* **2010,** 11, 99.

105. Xie, L.; Bourne, P. E., Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc Natl Acad Sci U S A* **2008,** 105, (14), 5441-6.

106. Xie, L.; Li, J.; Bourne, P. E., Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput Biol* **2009,** 5, (5), e1000387.

107. Theissen, G., Secret life of genes. *Nature* **2002,** 415, (6873), 741.

108. McLoughlin, S. Y.; Copley, S. D., A compromise required by gene sharing enables survival: Implications for evolution of new enzyme activities. *Proc Natl Acad Sci U S A* **2008,** 105, (36), 13497-502.

109. Andersson, J. O., Convergent evolution: gene sharing by eukaryotic plant pathogens. *Curr Biol* **2006,** 16, (18), R804-6.

110. Ribet, D.; Cossart, P., Post-translational modifications in host cells during bacterial infection. *FEBS Lett* **2010,** 584, (13), 2748-58.

111. Hert, J.; Keiser, M. J.; Irwin, J. J.; Oprea, T. I.; Shoichet, B. K., Quantifying the relationships among drug classes. *J Chem Inf Model* **2008,** 48, (4), 755-65.

112. Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K., Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* **2007,** 25, (2), 197-206.

113. Najmanovich, R. J.; Allali-Hassani, A.; Morris, R. J.; Dombrovsky, L.; Pan, P. W.; Vedadi, M.; Plotnikov, A. N.; Edwards, A.; Arrowsmith, C.; Thornton, J. M., Analysis of binding site similarity, small-molecule similarity and experimental

binding profiles in the human cytosolic sulfotransferase family. *Bioinformatics* **2007,** 23, (2), e104-9.

114. Shortridge, M. D.; Bokemper, M.; Copeland, J. C.; Stark, J.; Powers, R., A Seqeunce and Strucutre Independent Method to Predict Protien Function. *J Am Chem Soc* **2010,** Submitted.

115. Mercier, K. A.; Baran, M.; Ramanathan, V.; Revesz, P.; Xiao, R.; Montelione, G. T.; Powers, R., FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* **2006,** 128, (47), 15292-9.

116. Mercier, K. A.; Cort, J. R.; Kennedy, M. A.; Lockert, E. E.; Ni, S.; Shortridge, M. D.; Powers, R., Structure and function of Pseudomonas aeruginosa protein PA1324 (21-170). *Protein Sci* **2009,** 18, (3), 606-18.

117. 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* **2009,** 4, (10), e7442.

118. Gerlt, J. A.; Babbitt, P. C., Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem* **2001,** 70, 209-46.

119. Mirny, L. A.; Shakhnovich, E. I., Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol* **1999,** 291, (1), 177-96.

120. Park, K.; Kim, D., Binding similarity network of ligand. *Proteins* **2008,** 71, (2), 960-71.

121. Campbell, S. J.; Gold, N. D.; Jackson, R. M.; Westhead, D. R., Ligand binding: functional site location, similarity and docking. *Curr Opin Struct Biol* **2003,** 13, (3), 389-95.

122. Powers, R.; Mercier, K. A.; Copeland, J. C., The application of FAST-NMR for the identification of novel drug discovery targets. *Drug Discov Today* **2008,** 13, (3-4), 172-9.

123. Sams-Dodd, F., Drug discovery: selecting the optimal approach. *Drug Discovery Today* **2006,** 11, (9 & 10), 465-472.

124. Terstappen, G. C.; Schluepen, C.; Raggiaschi, R.; Gaviraghi, G., Target deconvolution strategies in drug discovery. *Nat. Rev. Drug Discovery* **2007,** 6, (11), 891-903.

125. Karlberg, J. P. E., Trends in disease focus of drug development. *Nat. Rev. Drug Discovery* **2008,** 7, (8), 639-640.

126. Pellecchia, M.; Bertini, I.; Cowburn, D.; Dalvit, C.; Giralt, E.; Jahnke, W.; James, T. L.; Homans, S. W.; Kessler, H.; Luchinat, C.; Meyer, B.; Oschkinat, H.; Peng, J.; Schwalbe, H.; Siegal, G., Perspectives on NMR in drug discovery: a technique comes of age. *Nat Rev Drug Discov* **2008,** 7, (9), 738-45.

127. Powers, R., Advances in nuclear magnetic resonance for drug discovery. *Expert Opin. Drug Discovery* **2009,** 4, (10), 1077-1098.

128. Fejzo, J.; Lepre, C. A.; Peng, J. W.; Bemis, G. W.; Ajay; Murcko, M. A.; Moore, J. M., The SHAPES strategy: and NMR-based approach for lead generation in drug discovery. *Chem. Biol.* **1999,** 6, 755-769.

129. 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* **1997,** 119, 12257-12261.

130. Shuker, S. B.; Hajduk, P. J.; Meadow, R. P.; Feisk, S. W., Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science* **1996,** 274, (5292), 1531-1534.

131. Wagner, G.; Hyberts, S. G.; Havel, T. F., NMR Structure Determination in Solution: A Critique and Comparison with X-Ray Crystallography. *Annu. Rev. Biophys. Biomol. Struct* **1992,** 21, 167-198.

132. Vanwetswinkel, S.; Heetebrij, R. J.; Duynhoven, J. v.; Hollander, J. G.; Filippov, D. V.; Hajduk, P.; Siegal, G., TINS, Target Immobilized NMR Screening. *Chem. Biol.* **2005,** 12, 207-216.

133. Peng, J. W.; Moore, J.; Abdul-Manan, N., NMR experiments for lead generation in drug discovery. *Prog. Nucl. Magn. Reson. Spectrosc.* **2004,** 44, 225-256.

134. Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W., Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **1996,** 274, (5292), 1531-4.

135. Zartler, E. R.; Hanson, J.; Jones, B. E.; Kline, A. D.; Martin, G.; Mo, H.; Shapiro, M. J.; Wang, R.; Wu, H.; Yan, J., RAMPED-UP NMR: multiplexed NMR-based screening for drug discovery. *J Am Chem Soc* **2003,** 125, (36), 10941-6.

136. Mayer, M.; Meyer, B., Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angew. Chem., Int. Ed.* **1999,** 38, (12), 1784-1788.

137. Sem, D. S.; Bertolaet, B.; Baker, B.; Chang, E.; Costache, A. D.; Coutts, S.; Dong, Q.; Hansen, M.; Hong, V.; Huang, X.; Jack, R. M.; Kho, R.; Lang, H.; Ma,

C.-T.; Meininger, D.; Pellecchia, M.; Pierre, F.; Villar, H.; Yu, L., Systems-Based Design of Bi-Ligand Inhibitors of Oxidoreductases Filling the Chemical Proteomic Toolbox. *Chemistry & Biology* **2004,** 11, (2), 185-194.

138. Gonzalez-Ruiz, D.; Gohlke, H., Steering protein-ligand docking with quantitative NMR chemical shift perturbations. *J Chem Inf Model* **2009,** 49, (10), 2260-71.

139. Lugovskoy, A. A.; Degterev, A. I.; Fhamy, A. F.; Zhou, P.; Gross, J. D.; Yuan, J.; Wagner, G., A Novel Approach for Characterizing Protein Ligand Complexes: Molecular Basis for Specificity of Small-Molecule Bcl-2 Inhibitors. **2002,** 124, (7), 1234-1240.

140. McCoy, M. A.; Wyss, D. F., Alignment of weakly interacting molecules to protein surfaces using simulations of chemical shift perturbations. *J Biomol NMR* **2000,** 18, (3), 189-98.

141. 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* **2002,** 124, (39), 11758-63.

142. 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. **2000,** 122, (6), 1241-1242.

143. Stark, J.; Powers, R., Rapid protein-ligand costructures using chemical shift perturbations. *J Am Chem Soc* **2008,** 130, (2), 535-45.

144. 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* **2009,** 12, (3), 285-95.

145. Moy, F. J.; Haraki, K.; Mobilio, D.; Walker, G.; Powers, R.; Tabei, K.; Tong, H.; Siegel, M. M., MS/NMR: a structure-based approach for discovering protein ligands and for drug design by coupling size exclusion chromatography, mass spectrometry, and nuclear magnetic resonance spectroscopy. *Anal Chem* **2001,** 73, (3), 571-81.

146. Dalvit, C.; Pevarello, P.; Tato, M.; Veronesi, M.; Vulpetti, A.; Sundstrom, M., Identification of compounds with binding affinity to proteins via magnetization transfer from bulk water. *J. Biomol. NMR* **2000,** 18, (1), 65-68.

147. Bista, M.; Kowalska, K.; Janczyk, W.; Doemling, A.; Holak, T. A., Robust NMR Screening for Lead Compounds Using Tryptophan-Containing Proteins. *J Am Chem Soc* **2009,** 131, (22), 7500-7501.

148. D'Silva, L.; Ozdowy, P.; Krajewski, M.; Rothweiler, U.; Singh, M.; Holak, T. A., Monitoring the Effects of Antagonists on Protein-Protein Interactions with NMR Spectroscopy. *J Am Chem Soc* **2005,** 127, (38), 13220-13226.

149. 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* **2003,** 125, (25), 7696-7703.

150. Jahnke, W.; Ruedisser, S.; Zurini, M., Spin label enhanced NMR screening. *J Am Chem Soc* **2001,** 123, (13), 3149-3150.

151. Lin, M.; Shapiro, M. J.; Wareing, J. R., Diffusion-Edited NMR-Affinity NMR for Direct Observation of Molecular Interactions. *J Am Chem Soc* **1997,** 119, (22), 5249-5250.

152. Dalvit, C.; Ardini, E.; Fogliatto, G. P.; Mongelli, N.; Veronesi, M., Reliable high-throughput functional screening with 3-FABS. *Drug Discovery Today* **2004,** 9, (14), 595-602.

153. Chen, A.; Shapiro, M. J., NOE Pumping: A Novel NMR Technique for Identification of Compounds with Binding Affinity to Macromolecules. *J Am Chem Soc* **1998,** 120, (39), 10258-10259.

154. Ekonomiuk, D.; Su, X. C.; Ozawa, K.; Bodenreider, C.; Lim, S. P.; Yin, Z.; Keller, T. H.; Beer, D.; Patel, V.; Otting, G.; Caflisch, A.; Huang, D., Discovery of a non-peptidic inhibitor of west nile virus NS3 protease by high-throughput docking. *PLoS Negl Trop Dis* **2009,** 3, (1), e356.

155. Grandy, D.; Shan, J.; Zhang, X.; Rao, S.; Akunuru, S.; Li, H.; Zhang, Y.; Alpatov, I.; Zhang, X. A.; Lang, R. A.; Shi, D. L.; Zheng, J. J., Discovery and characterization of a small molecule inhibitor of the PDZ domain of dishevelled. *J Biol Chem* **2009,** 284, (24), 16256-63.

156. Liu, J.; Begley, D.; Mitchell, D. D.; Verlinde, C. L.; Varani, G.; Fan, E., Multivalent drug design and inhibition of cholera toxin by specific and transient protein-ligand interactions. *Chem Biol Drug Des* **2008,** 71, (5), 408-19.

157. Rothweiler, U.; Czarna, A.; Krajewski, M.; Ciombor, J.; Kalinski, C.; Khazak, V.; Ross, G.; Skobeleva, N.; Weber, L.; Holak, T. A., Isoquinolin-1-one inhibitors of the MDM2-p53 interaction. *ChemMedChem* **2008,** 3, (7), 1118-28.

158. Shi, Z.; Tabassum, S.; Jiang, W.; Zhang, J.; Mathur, S.; Wu, J.; Shi, Y., Identification of a potent inhibitor of human dual-specific phosphatase, VHR,

from computer-aided and NMR-based screening to cellular effects. *Chembiochem* **2007,** 8, (17), 2092-9.

159. Tsao, D. H.; Sutherland, A. G.; Jennings, L. D.; Li, Y.; Rush, T. S., 3rd; Alvarez, J. C.; Ding, W.; Dushin, E. G.; Dushin, R. G.; Haney, S. A.; Kenny, C. H.; Malakian, A. K.; Nilakantan, R.; Mosyak, L., Discovery of novel inhibitors of the ZipA/FtsZ complex by NMR fragment screening coupled with structure-based design. *Bioorg Med Chem* **2006,** 14, (23), 7953-61.

160. Xie, J.; Thapa, R.; Reverdatto, S.; Burz, D. S.; Shekhtman, A., Screening of small molecule interactor library by using in-cell NMR spectroscopy (SMILI-NMR). *J Med Chem* **2009,** 52, (11), 3516-22.

161. Burz, D. S.; Dutta, K.; Cowburn, D.; Shekhtman, A., In-cell NMR for protein-protein interactions (STINT-NMR). *Nat. Protoc.* **2006,** 1, (1), 146-152.

162. Burz, D. S.; Dutta, K.; Cowburn, D.; Shekhtman, A., Mapping structural interactions using in-cell NMR spectroscopy (STINT-NMR). *Nat. Methods* **2006,** 3, (2), 91-93.

163. Vanwetswinkel, S.; Heetebrij, R. J.; van Duynhoven, J.; Hollander, J. G.; Filippov, D. V.; Hajduk, P. J.; Siegal, G., TINS, Target Immobilized NMR Screening: An Efficient and Sensitive Method for Ligand Discovery. *Chem. Biol.* **2005,** 12, (2), 207-216.

164. Sanchez-Pedregal, V. M.; Reese, M.; Meiler, J.; Blommers, M. J. J.; Griesinger, C.; Carlomagno, T., The INPHARMA method: Protein-mediated interligand NOEs for pharmacophore mapping. *Angew. Chem., Int. Ed.* **2005,** 44, (27), 4172-4175.

165. Ludwig, C.; Michiels, P. J. A.; Wu, X.; Kavanagh, K. L.; Pilka, E.; Jansson, A.; Oppermann, U.; Guenther, U. L., SALMON: Solvent Accessibility, Ligand binding, and Mapping of ligand Orientation by NMR Spectroscopy. *J. Med. Chem.* **2008,** 51, (1), 1-3.

166. Mercier, K. A.; Baran, M.; Ramanathan, V.; Revesz, P.; Xiao, R.; Montelione, G. T.; Powers, R., FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* **2006,** 128, (47), 15292-9.

CHAPTER 2:

ESTIMATING PROTEIN-LIGAND BINDING AFFINITY USING HIGH-THROUGHPUT SCREENING BY NMR

2.1 INTRODUCTION

In chapter 1 the general principles for using NMR as a high-throughput screening tool were discussed in the context of functional annotation and drug discovery. For a high-throughput screen to be successful a hit must efficiently alter the biological activity of a protein or other biological target molecule and disrupt its normal function.[1] This is generally accomplished by a small molecule changing the dynamics of a protein [2] or interfering with a critical protein-protein interaction.[3] For a small molecule hit identified from a high-throughput screen to become a viable drug candidate it must elicit these effects while simultaneously demonstrating *in vivo* efficacy in the absence of toxic side-effects. Thus, an important component of the drug discovery process is the verification that a small molecule ligand actually binds the protein target in a selective and biologically relevant fashion.

Selectively is measured by binding affinity which is governed by the equilibrium parameters of a binding interaction. The equilibrium state of a binding interaction is described by the concentration of the free ligand $[L]_F$, free receptor $[P]_F$, and the receptor-ligand complex [PL]. For single-site binding, the relative ratios of these concentrations are governed by the kinetic on (k_{on}) and off (k_{off}) rates between the free and bound forms as described in eq 2.1.

$$[L]_F + [P]_F \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} [PL] \qquad [2.1]$$

The strength of a ligand's binding affinity is quantified by the dissociation constant (K_D), or simply the ratio of k_{off} and k_{on} rates.

$$K_D = \frac{k_{off}}{k_{on}} = \frac{[L]_F[P]_F}{[PL]} \qquad [2.2]$$

The selectivity of a ligand to a particular target biological molecule is inversely proportional to the strength of the K_D. Highly selective compounds will have a K_D in the pM-nM range while weaker ligands will exhibit binding in the μM-mM range. Often a HTS will rely on the identification of μM-mM binding ligands coupled with structural information to develop high affinity ligands through combinatorial approaches.

Similar to traditional measurements,[4] NMR methods rely on the collection of multiple data points to accurately determine a K_D for a protein-ligand interaction. This approach is usually impractical in a high–throughput mode that requires a rapid method for characterizing and ranking binding affinities. Examples of high-throughput K_D measurements using 1D NMR experiments have been described that use ^{19}F-containing compounds[5, 6] or the displacement of known low-affinity inhibitors. [7, 8] Unfortunately, these approaches are typically limited in practice because known low-affinity inhibitors or a large library of "drug-like" and structurally diverse ^{19}F-containing compounds are not available for a wide range of protein targets. To increase the utility and throughput of NMR affinity screening a rapid and universal method to determine binding affinity was still needed.

This chapter discusses a new NMR screening method that can determine the relative ranking of binding affinities using a variation of traditional 1D ^{1}H NMR line-broadening experiments.[9, 10] This approach correlates the ratio of NMR peak intensities

for free and bound ligands to the fraction of bound ligand in a protein-ligand complex. This method is illustrated by using human serum albumin (HSA) as a model protein. HSA is also an important secondary target for efficacy screening and a well-established system for monitoring protein-ligand interactions.[11]

2.2 THEORY

2.2.1 Single point K_D measurements. Binding interactions between a protein (MW > 5000 Da) and a low molecular weight ligand (MW < 500 Da) can be examined by using the decrease in NMR peak intensity that occurs upon the addition of a protein to a solution with constant ligand concentration. NMR line-broadening experiments follow an opposite protocol from typical experiments that measure K_D values, where variable protein concentrations are added to solutions that contain a constant ligand concentration. Thus, a different form for the standard Langmuir binding isotherm (eq 2.2) was required.

Rearrangement of eq 2.2 produces the following binding isotherm, in which f_B represents the "fractional occupancy", or the fraction of bound ligand.

$$f_B = \frac{[PL]}{[L]_T} = \frac{1}{1 + \frac{K_D}{[P]_F}} \qquad [2.3]$$

It is assumed in many types of binding studies the total ligand concentration $[L]_T$ is approximately equal to the free ligand concentration; however, this assumption is not applicable to the NMR line-broadening experiments used in this study because $[L]_T$ is not necessarily in excess of the maximum complex concentration [PL]. Also, a direct measurement of the free protein concentration is not possible for the method described in this report. Therefore, eq 2.3 was derived to describe this situation in terms of the total

protein concentration $[P]_T$ and total ligand concentration $[L]_T$ that are known to be present in the system (see appendix 2A and 2B for definition of variables and equation derivation respectively).

$$f_B = \frac{[PL]}{[L]_T} = \frac{1}{1 + \frac{2K_D}{([P]_T - [L]_T - K_D) + \sqrt{([P]_T - [L]_T + K_D)^2 + 4K_D[L]_T}}} \quad [2.4]$$

Equation 2.4 can be simplified to approximate the fractional occupancy in terms of the total ligand concentration $[L]_T$ and total protein concentration $[P]_T$ by using a Taylor series expansion and the assumption that $[L]_T > [P]_T$.

$$f_B = \frac{[PL]}{[L]_T} \approx \frac{[P]_T}{([L]_T + K_D)} \quad [2.5]$$

The fractional occupancy for a protein-ligand complex can be measured using a ratio of NMR peak intensities ($1-I_B/I_F$), where I_B is the sum of ligand NMR peak intensities in the presence of the protein and I_F is the sum of NMR peak intensities for the free ligand. Therefore, B_{expt} (the NMR peak intensity ratio) represents an easily measurable response of ligand binding that can be described in terms of the fraction of bound ligand (f_B) and the NMR linewidth for the free (ν_F) and bound (ν_B) states (see Appendix B for derivation).

$$B_{expt} = 1 - \frac{I_B}{I_F} = 1 - \frac{1}{1 + f_B(\frac{\nu_B}{\nu_F} - 1)} \quad [2.6]$$

Combining eq 2.5 and eq 2.6 leads to a new binding isotherm for this system, as shown below,

$$B_{expt} = 1 - \frac{I_B}{I_F} = 1 - \frac{1}{1 + \frac{c[P]_T}{[L]_T + K_D}} \qquad \text{where } c = \frac{\nu_B}{\nu_F} - 1 \qquad [2.7]$$

The unit-less NMR linewidth ratio constant (c), as defined in eq 2.7, accounts for the proportional change in ligand linewidth upon binding of a ligand to a protein. Once a ligand is bound, the free ligand linewidth (ν_F) of a ligand resonance adopts the linewidth of the protein (ν_B) and the increase in linewidth produces a corresponding decrease in peak intensity measured by the ratio of NMR peak intensity (B).

The dissociation equilibrium constant for a protein-ligand complex that is calculated using eq 2.7 is based on relative changes in NMR peak intensity by fitting the given binding isotherm to a complete protein titration curve. This is impractical in the context of an NMR high-throughput screen where only a single titration point is measured. However, eq 2.7 can be rearranged to solve for K_D to yield an estimate for K_D that is based on $[P]_T$, $[L]_T$, c and B_{single}, where B_{single} is the fractional occupancy at a single protein concentration. The resulting expression is shown in eq 2.8.

$$K_D = \left[\left(\frac{c[P]_T}{B_{single}} - c[P]_T \right) - [L]_T \right] \qquad [2.8]$$

For proteins such as HSA that possess multiple non-specific binding sites, the decrease in ligand signal at a relatively high protein concentration will be an average of specific and non-specific binding. To correct for this effect, the non-specific binding term $n[P]_T$ that corresponds to a linear increase in fraction bound with the addition of protein is simply added to eq 2.7, as shown in eq 2.9.

$$B = 1 - \frac{I_B}{I_F} = 1 - \frac{1}{1 + \frac{c[P]_T}{[L]_T + K_D}} + n[P]_T \qquad [2.9]$$

2.3 EXPERIMENTAL

2.3.1 Materials. The HSA (essentially fatty acid free, ≥ 96% pure), choline bromide (~ 99% pure), clofibrate, furosemide, phenol red, phenylbutazone, phenytoin (~ 99% pure), sodium salicylate, tolbutamide, uridine 5'-monophosphate (98-100% pure) and warfarin (> 98% pure) were purchased from Sigma (St. Louis, MO). The bromophenol blue (ACS reagent grade, 95% pure), bromocresol green (ACS reagent grade, 95% pure), and ibuprofen were from Sigma-Aldrich (Milwaukee, WI). The dimethyl sulfoxide-d_6 (99.9% D), deuterium oxide (99.9 atom% D) and naproxen (98% pure) were obtained from Aldrich (Milwaukee, WI). The 3-(trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt (98% D) was purchased from Cambridge Isotope (Andover, MA). The potassium phosphate dibasic salt (anhydrous, 99.1% pure) and monobasic salt (crystal, 99.8% pure) were purchased from Mallinckrodt (Phillipsburg, NJ).

2.3.2 Apparatus. All NMR spectra were collected on a Bruker 500 MHz Avance spectrometer (Billerica, MA) equipped with a triple-resonance, Z-axis gradient cryoprobe and using a Bruker BACS-120 sample changer and IconNMR software for automated data collection. Spectra were collected at 298 K using 512 transients, a sweep-width of 6009 Hz, 16 K data points and a relaxation delay of 2.0 s. The residual H_2O resonance signal was suppressed with presaturation during the recycle delay and a composite pulse train prior to the 90^o excitation pulse. The total experiment time, including sample changing for each spectrum, was approximately 33 min.

2.3.3 Sample Preparation. All small-molecule ligands that were used in this study were selected based on their previously reported K_D values for HSA and their good solubility in an aqueous solution.[11] The small-molecule ligand samples were individually prepared in 10 mL stock solutions that contained 20 μM ligand, 1% (v/v) dimethyl sulfoxide-d_6 (DMSO-d_6), 10 μM 3-(trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt (TSP) and pH 7.0 (uncorrected) 50 mM potassium phosphate buffer prepared in deuterium oxide.

A series of ten HSA stock solutions were prepared in deuterium oxide by making serial dilutions from a 200 μM master solution of HSA in deuterium oxide. The final concentrations of HSA in these stock solutions ranged from 0 μM to 200 μM and were prepared so that a 10 μL addition of the HSA stock solution to 490 μL of a free ligand solution resulted in final concentrations of 0 μM, 0.1 μM, 0.2 μM, 0.4 μM, 0.6 μM, 0.8 μM, 1 μM, 2 μM, 3 μM, and 4 μM HSA, respectively. These mixtures were prepared individually for each ligand in 1.5 mL microcentrifuge tubes and then transferred to NMR tubes. The sample for each titration that contained 0 μM HSA was used as the reference for calculating the free ligand intensities (I_F) and free ligand linewidths (ν_F). All binding studies performed with these solutions were conducted at 25ºC.

2.3.4 1D ^{1}H NMR binding curves. Spectra were processed with the ACD/1D NMR manager (Advanced Chemistry Development, Inc., Toronto, Ontario). A linear prediction algorithm was applied to the FID in the forward direction and the resulting FID was Fourier transformed. The NMR spectrum was phase-adjusted and baseline-corrected. The residual water signal was removed for spectrum clarity by the solvent removal function in ACD. This function zeros' the spectrum baseline at the residual

water signal. All ligand resonance peaks were visually selected and peak positions were measured relative to a TSP reference set to 0.0 ppm. Peak intensities were measured relative to the DMSO-d_6 peak at 2.69 ppm that was normalized to an intensity of 1.00. The DMSO-d_6 peak was completely recovered during the 1D ^{1}H NMR experiment using a 2.0 s recycle delay. This is >3x the T_1 for DMSO in D_2O at 298K (0.3-0.5 s) and is acceptable for complete relaxation.[12, 13] Individual peak intensities in the aromatic region for each ligand were summed to obtain the free (I_F) and bound (I_B) intensities at each titration point. The peak-intensity ratios were plotted versus total protein concentration and fit to eq 2.9 using the program KaleidaGraph version 3.52 for Windows (Synergy Software., Reading, PA) to estimate the K_D value for each protein-ligand complex. The average NMR linewidth ratio (c) for each ligand was estimated by using eq 2.7, where ν_B was taken to be approximately 94.2 Hz using a previously measured correlation time for HSA of 41 ns.[14] The value for ν_F was calculated as described in the next section. The fit of each binding curve was constrained so that $K_D \geq 0$ in these studies.

2.3.5 Measuring a free ligand NMR linewidth (ν_F). To measure the free ligand linewidth (ν_F) for use in eq 2.7, the NMR spectrum for each free ligand (i.e., as obtained in a solution containing no HSA) was processed as described above to avoid any distortion in linewidth resulting from processing. NMR peak linewidths were measured using the ACD/1D NMR manager peak fitting routine. The average peak linewidth was used to report ν_F for each ligand and to calculate the NMR linewidth ratio.

2.3.6 Simulated high-throughput screening by NMR. To simulate the outcome of an NMR high-throughput screening assay, a single protein concentration $[P]_T$ from the full titration curve was used. On average, the 0.2 μM HSA titration point yielded a large

response for all 12 ligands without reaching saturation. The static total ligand concentration $[L]_T$ was 20 μM. A simulated response curve was generated by fitting a range of K_D values to a range of ideal B_{single} values calculated using eq 2.8. The measured B_{single} value for each ligand at the 0.2 μM HSA titration point was used to calculate a single-point binding constant from eq 2.8 and compared to the simulated response curve. This simulated experiment used both the individual c values calculated for each ligand from the full titration experiment and an average c value calculated from the 12 NMR titration curves. The single-point dissociation equilibrium constant for each ligand was calculated using this average c value.

2.4 RESULTS AND DISCUSSION

2.4.1 Measuring K_D from 1D ^{1}H NMR line-broadening experiments. The development of NMR-based screening assays that monitor changes in chemical-shifts or linewidth as a means to identify or verify initial chemical leads has evolved to become an increasingly important component of drug discovery efforts in the biotechnology and pharmaceutical industry.[15, 16] Nevertheless, the direct measurement of a binding affinity from a high-throughput NMR screen is generally lacking.[5, 6, 8, 17, 18] A decrease in the intensity of a ligand's NMR signal in the presence of a protein is commonly used in NMR-based screens to monitor the formation of a protein-ligand complex. 1D ^{1}H NMR spectra of small-molecules (MW ≤ 500 Da) usually have extremely sharp peaks due to slow dipole-dipole relaxation (T_2).[19] Binding to a high molecular weight agent like a protein induces peak broadening and a corresponding decrease in the ligand's NMR signal intensity because the bound ligand now experiences the shorter relaxation time of

the protein. This effect is illustrated in figure 2.1 using binding by the protein HSA to the drugs phenytoin and naproxen as examples.

The observed increase in ligand linewidth in such an experiment will depend on a number of factors that include the dissociation equilibrium constant for the protein-ligand interaction, K_D. In general, the observed change in the ligand's linewidth (ν_{obs}) for the fast exchange limit will follow the result shown below.

$$\nu_{obs} = \nu_F + f_B(\nu_B - \nu_F) \qquad \text{where} \quad f_B \approx \frac{[P]_T}{[L]_T + K_D} \qquad [2.10]$$

In eq 2.10, f_B is the fraction of the bound protein-ligand complex, ν_F is the free ligand NMR linewidth, and ν_B is the linewidth for the bound state of the ligand (see the appendix B for an explanation regarding the above expression for f_B). Eq 2.10 shows that an increase in the observed ligand linewidth will be related to the free and bound ligand linewidths and the value of K_D for the protein-ligand complex. If it is assumed the linewidth of the protein-ligand complex is significantly larger than that for the free ligand, the ratio of the ligand linewidth in the presence and absence of the protein should represent the remaining free ligand concentration, as indicated by eq 2.7.

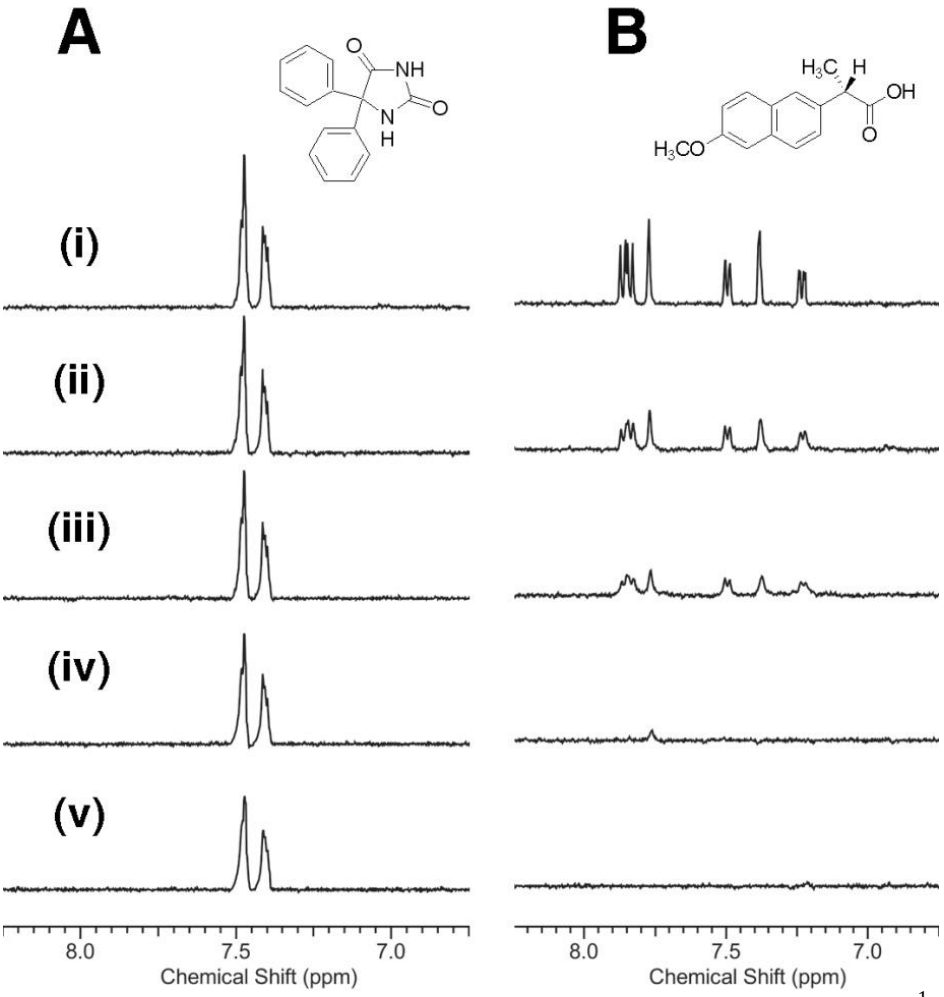


Figure 2.1. Relative line broadening of in response to protein binding. 1D ^{1}H NMR spectra for titration of 20 μM of the drugs phenytoin (A) and naproxen (B) with increasing concentrations of HSA. The concentrations of HSA were as follows: (i) 0 μM, (ii) 0.4 μM, (iii) 1 μM, (iv) 2 μM, and (v) 4 μM. As the protein concentration increases, the intensity of the ligand NMR signal decreases due to the bound ligand adopting the shorter relaxation time of the protein. The decrease in the ratio of NMR signal intensity ($\frac{I_B}{I_F} - 1$) is proportional to the degree of binding such that tighter binding ligands (B) will relax more quickly than weaker binding ligands (A).

This relationship assumes there is a lack of any significant contribution of chemical or dynamic exchange to the observed change in linewidth. This is a reasonable assumption in the context of a high-throughput NMR screen against a single protein target. First, initial chemical leads tend to be weak binders in the fast exchange regime, where the linewidth change of the ligand will be dominated by the linewidth of the protein. Second, biologically relevant binders will interact with the same or similar binding sites on the protein. Under these circumstances, the ligand may experience a relatively constant contribution of chemical and dynamical line-broadening. Thus, the minimal contribution of linewidth from exchange processes should not affect the relative ranking of the ligand binding affinities that are obtained when using such an experimental approach.

The validity of this method for high-throughput screening by NMR was examined by using twelve ligands with previously determined binding affinities to HSA.[11, 20-23] These ligands were used to examine the relationship between the estimated values for K_D and the relative ratios of the NMR Peak intensity. Samples containing 20 μM of any given ligand were titrated with solutions that contained 0 to 4 μM of HSA to develop full binding curves for each of the twelve ligands. As a control, two suspected non-binding ligands (i.e., choline bromide and uridine-5'-monophosphate) were also screened in the presence of HSA with no observable decrease in signal (data not shown). The K_D values that were obtained by this method (see table 2.1) were experimentally determined by directly fitting the resulting binding curve of each ligand to eq 2.9. These fits gave a sum of residuals squared that ranged between 0.977 and 0.998 over the ten concentrations of HSA that were tested. Figure 2.2 shows the results that were obtained for three of the

tested ligands, which have previously reported dissociation equilibrium constants that ranged from 0.7 to 36.8 μM. These figures and the corresponding fits illustrate the ability of this approach to be used with ligands that have weak-to-moderate strength binding to proteins such as HSA.

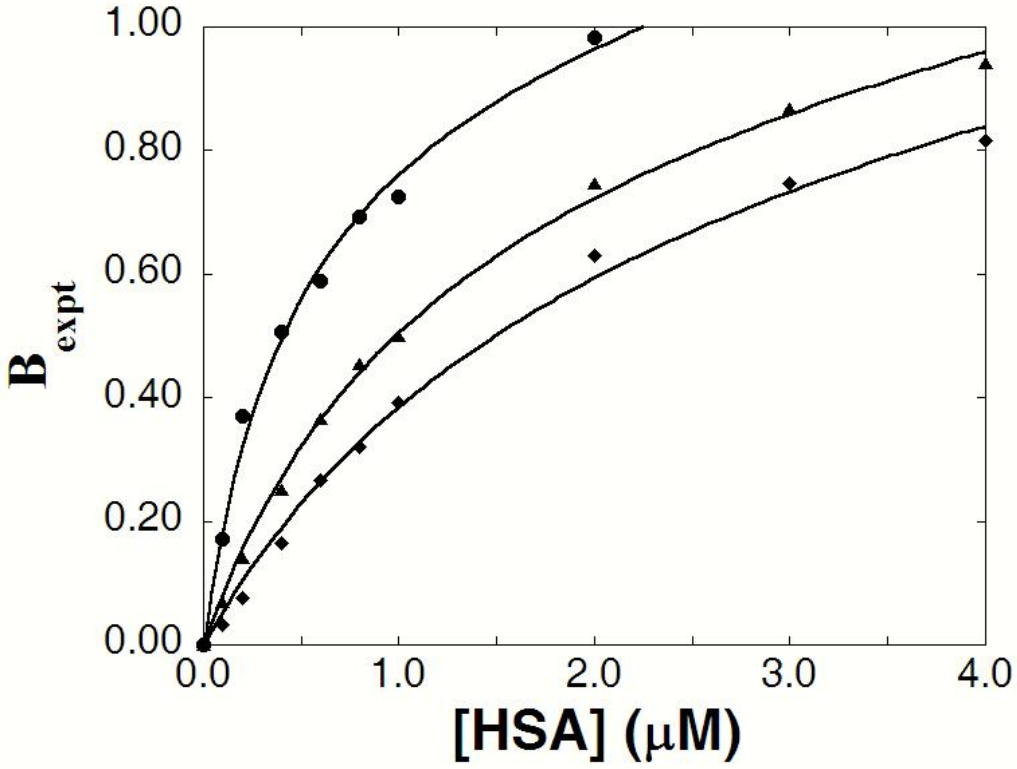


Figure 2.2. NMR ligand binding titration. Experimental fractional occupancy (B_{expt}) for naproxen (●), tolbutamide (▲), and phenol red (◆) versus the total concentration of HSA. The best-fit lines were obtained using eq 2.9. The r^2 for these best-fit lines are given in the text and the K_D values that were obtained from these lines are provided in table 2.1.

2.4.2 Co-variance of K_D and the NMR linewidth ratio (c). Ideally, the dissociation equilibrium constant (K_D) and the NMR linewidth ratio (c) could be simultaneously derived by fitting eq 2.7 to the experimental NMR binding curves. Unfortunately, K_D and c are completely covariant. This requires an approximation for c in

order to calculate K_D from the NMR binding curves. The linewidth of a protein (ν_P) may provide a lower estimate of ν_B if it is assumed that ν_B is dominated by the protein linewidth (ν_P). Estimations of ν_P can be made from the correlation time (τ_c) of the protein by using the intramolecular dipole-dipole relaxation rate constant (T_2^{-1}). [24]

$$T_2^{-1} = \frac{3}{20} b^2 \{3J(0) + 5J(\omega_0) + 2J(2\omega_0)\} \qquad [2.11]$$

Where

$$b = -\frac{\mu_o}{4\pi}\frac{h\gamma^2}{r^3}, \; J(\omega) = \frac{\tau_c}{1+\omega^2\tau_c^2} \text{ and } \omega_0 = -\gamma B_0 \qquad [2.12]$$

In these equations, $J(\omega)$ is the normalized spectral density function, μ_o is the vacuum permeability, $\gamma$is the magnetogyric ratio, ω is frequency (rad s^{-1}), $\hbar$ is Plank's constant, B_0 is the static magnetic field strength and r is the hydrodynamic radius of the protein. In addition, the Stokes-Einstein equation can be used to relate τ_c to the molecular weight (MW) for a globular protein,[25]

$$\tau_c = \frac{4\pi\rho\eta r^3}{3kT} \quad \text{with, } \tau_c \approx \rho * \frac{MW}{2400}(ns) \qquad [2.13]$$

where T is the temperature, k is the Boltzmann constant, η is the viscosity of the solvent, r is the radius and ρ is the shape constant.

The reliability of eq 2.13 to approximate a protein correlation time from its molecular weight is illustrated from a comparison between 27 experimental τ_c values [26, 27] and correlation times predicted using eq 2.13 (figure 2.3A). A linear best-fit was obtained with an R^2 of 0.81 in this case. For a high-throughput screen, ν_p can be estimated from the molecular weight of a protein by using this approximation for τ_c with a shape constant

of 1.32 combined with eq 2.11 and 2.12. The shape constant was determined by optimizing a linear fit between the experimental and predicted τ_c values shown in figure 2.3A by varying ρ. The result is an approximate correlation between ν_P and MW_P, as shown in eq 2.14.

$$\nu_p = 1.26 \bullet MW_p \qquad [2.14]$$

This dependency of linewidth on the size and shape of a protein is plotted in figure 2.3B. For HSA (MW, 66 kDa), the correlation time (41 ns) has previously been measured using time-resolved fluorescence.[14] This correlation time was used to calculate the value used for ν_P, which was 94.2 Hz.

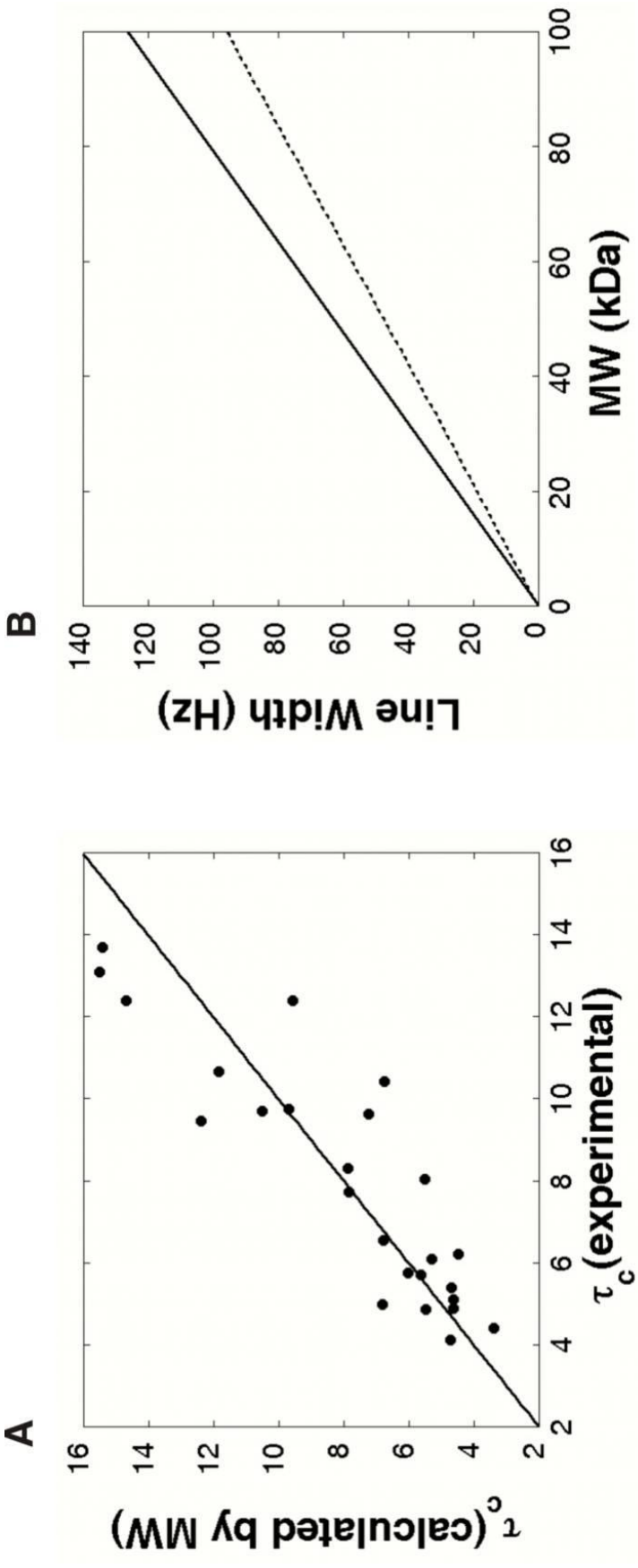

A
τ_c(calculated by MW)
τ_c(experimental)
2
4
6
8
10
12
14
16
B
Line Width (Hz)
MW (kDa)
0
20
40
60
80
100
120
140

Figure 2.3. Approximation of protein linewidth based on molecular weight. (A) Comparison of 27 experimental protein correlation times determined using NMR dynamics data with correlation times predicted from protein MW using eq 2.13 and a shape constant of 1.32. A best-fit line is shown with a slope of 1 and an R^2 of 0.81. (B) A plot of linewidth versus protein molecular weight based on eq 2.13 for spherical proteins with ρ of 1 (solid line) and elliptical proteins with ρ of 1.32 (dashed).

The free ligand linewidth (ν_F) can be measured directly from the NMR spectra of the free ligand using an average ligand linewidth. Average ν_F values measured from the free ligand NMR spectra are reported in table 2.1. However, for large and diverse chemical libraries it may not be feasible to measure an accurate linewidth for each compound. Alternatively, ν_F is generally between 1 and 2 Hz for many small-molecules (MW, 500 < Da), which provides a reasonable estimate for ν_F to calculate an average value for c.

2.4.3 Sensitivity of K_D and NMR linewidth Ratio (c). A closer examination of eq 2.7 indicates the NMR linewidth ratio (c) acts as a scaling factor in the calculation of K_D, with a larger c value resulting in a proportionally larger K_D value. Unfortunately, small variations or errors in the measurement of ν_F will result in proportionally larger variations in both c and K_D. In the context of high-throughput screening by NMR, an incorrect estimate of c will result in a systematic underestimation or overestimation of K_D. However, the relative ranking of the ligand binding affinities will be maintained. In addition, a lower limit to c is inherently defined by eq 2.7.

2.4.4 Comparison of estimated K_D values with literature values. Table 2.1 shows the dissociation equilibrium constants that were measured for twelve ligands known to bind HSA by using the 1D ^{1}H NMR line-broadening method that is described herein. Previously reported K_D values from the literature are also listed for these twelve ligands.[20-22, 28-52]

Table 2.1. Comparison of K_D values determined by NMR and reported in the literature under similar conditions

Ligand	Literature K_D (μM)	Line width (Hz)	c	Measured K_D (μM)
Ibuprofen	0.3[50] 0.33[45] 0.37[49] 0.5[45] 0.52[34] 1.0[47] 1.25[30] 1.26[40] 1.74[41] 1.89[33] 2.08[30] 2.8[35] 4.76[41] 5.56 [35] 5.68[38] 8.33[33] 7.17[29] 18.2[50] 23.81[53] 25.64[53]	2.3 ± 0.2	41.5	0.5 ± 1.0
Naproxen	0.83[42] 1.25[47] 7.09[39] 10.6[46] 23.7[44]	1.8 ± 0.6	51.3	0.7 ± 1.2
Clofibrate	1.32[47]	1.7 ± 0.1	54.3	1.7 ± 3.4
Bromophenol Blue	0.67[43]	2.5 ± 0.4	37.8	3.0 ± 2.3
Furosimide	5.26[36] 52.63[28]	1.5 ± 0.8	57.6	3.4 ± 3.0
Warfarin	1.61[39] 2.17[37] 2.27[37] 2.94[36] 3.03[37] 3.4[31] 3.7[46] 3.85[32] 4.76[37] 5.3[31] 6.8[31]	2.3 ± 0.9	41.7	4.0 ± 2.8
Phenylbutazone	0.67[36] 1.9[31] 5.43[32] 8.4[31] 11[31] 15.13[34]	3.7 ± 0.6	25.2	6.5 ± 2.9
Salicylate	5.26[36] 15.15[37] 32.15[37] 35.71[37] 141[46]	1.4 ± 0.8	63.6	7.2 ± 2.9
Bromocresol Green	0.63[48] 1.43[43]	2.7 ± 0.3	35.1	7.4 ± 2.1
Tolbutamide	25[36] 31.25[39]	2.7 ± 0.4	34.9	10.2 ± 1.2
Phenol Red	35.7[43]	1.6 ± 0.5	58.7	36.8 ± 6.5
Phenytoin	50[20] 58.8[20] 62.5[20] 71.43[53] 96.15[20] 111[20] 1342.3[20] 153.85[20] 211[20] 244[20] 568.2[20]	2.0 ± 0.6	46.8	131.6 ± 12.5

In general, there is good agreement between the K_D values that were estimated by NMR and those values reported in the literature. Variations in temperature, pH or buffer conditions may partly explain the range of K_D values observed in the literature. There may have also been differences in the fatty acid content of the HSA preparations, which can affect the reported K_D values. Thus, 1D ^{1}H NMR line-broadening measurements appear to provide reliable preliminary estimates for binding affinities as part of a high-throughput screening assay.

One limitation of the model that was used for this analysis is the assumption of only a single site interaction between the ligand and protein. There are many cases for which multisite binding or other effects (e.g., allosteric interactions) are present that give rise to more complex binding models.[11, 19] Multisite binding also contributes to the relatively large range of K_D values reported in the literature for HSA ligands. In these situations, the K_D values listed in table 2.1 (for both the NMR and literature results) should be regarded as weighted averages and as measures of the global affinity for a particular ligand with HSA. This averaging effect may be more pronounced for the NMR method than for other techniques because of the practical limit in ligand concentration that could be used to provide a measurable signal. There is also a practical limit to the number of concentrations and data points that could be sampled to give a binding curve. This effect may explain why the NMR-derived K_D values tend to be lower than the literature values, because the use of higher concentrations for the NMR studies would give a higher weight and likelihood to the detection of weaker interactions between the ligand and protein.

A number of other practical limitations also need to be considered in the use of NMR for these binding studies. For instance, the NMR resonances that are specifically involved with protein binding have been shown to exhibit the most dramatic changes in linewidth.[9, 10] Therefore, there are inherent errors caused by summing all peak intensity and selectively excluding ligand peaks due to an overlap with buffer and protein resonances. In addition, errors in the measurement of peak intensity might arise at lower ligand concentrations due to the difficulty of accurately identifying and selecting peaks under these conditions. The result could be either a low or high estimate for K_D, depending on the disparity in linewidth changes and on which peaks are excluded. Using overlapping peaks would introduce an alternative error because the observed intensity is the sum of multiple peaks that cannot be easily de-convoluted. Also, the analysis of hundreds to thousands of NMR spectra in a high-throughput screening assay precludes a manual inspection to selectively determine which peaks to include or exclude.

2.4.5 Estimating K_D based on single-point 1D ^{1}H NMR line-broadening Measurements. Since NMR-based screens are a common component of the drug discovery process in the pharmaceutical industry, single-point estimates of ligand binding affinities could be an extremely valuable tool to initially rank and prioritize chemical leads. During the iterative drug optimization process, it is typical to focus on a small set (i.e., 3-5 compounds) of structurally distinct chemical classes that are amenable to synthetic modification and that exhibit drug-like characteristics.[54] For this work, an NMR screen could be used to verify the presence of a specific and biologically-relevant interaction involving a protein target and to rank the relative binding affinity of the screened ligands to simplify the selection of promising lead compounds. This approach

was illustrated in this study by simulating NMR high-throughput screening results for the twelve compounds that were used in the previous binding study.

First, using an average c value of 45.7 ±11.6 and an HSA concentration of 0.2 μM, single point K_D values were calculated for a range of B_{single} values using eq 2.8. The results of this calculation are shown in figure 2.4. Superimposed on the single point curve in figure 2.4A are the K_D values reported in table 2.1 plotted versus the experimental B values at 0.2 μM HSA. Superimposed on the single point curve in figure 2.4B are the K_D values from table 2.1, where the corresponding c values were used to determine a best-fit to eq 2.9. This represents the typical protocol that would be used in a high-throughput screen and shows that an average value of c is acceptable for use when individual estimates of c may not be practical. A comparison of figure 2.4B with the theoretical curve based on eq 2.8 indicates the single-point method can provide a reasonable approximation for K_D.

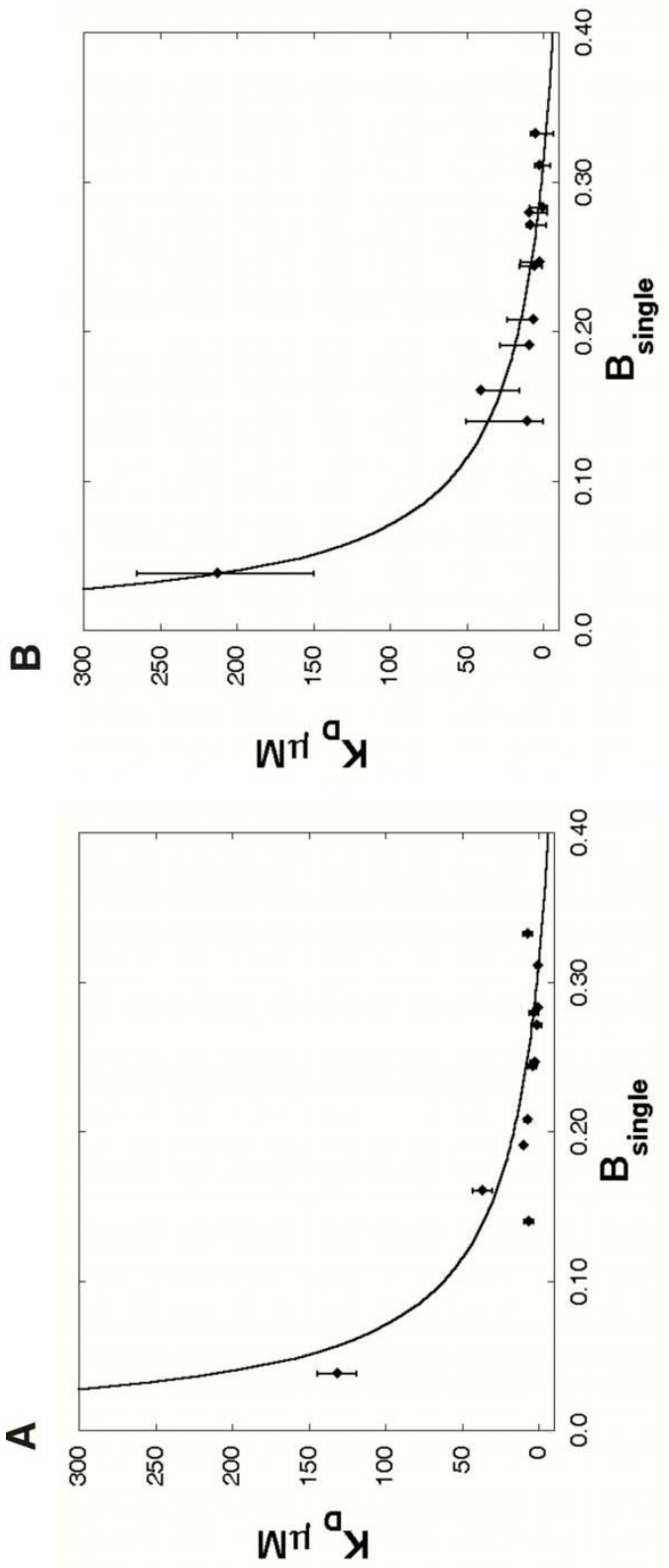

A
B
K_D μM
B_{single}
0
50
100
150
200
250
300
0.0
0.10
0.20
0.30
0.40

Figure 2.4. Use of NMR in a single-point binding analysis for several small-molecule ligands with known interactions with the protein HSA. The curves in (A) and (B) represents the ideal single-point K_D values calculated from eq 2.8 with 0.2 μM HSA and an average c value of 45.7 ± 11.6. (A) The K_D values and errors reported in table 2.1 are superimposed on the ideal fit. The K_D values are based on the best-fit to eq 2.9 using the c values determined for each individual compound. (B) The K_D for each compound was re-calculated based on the best-fit to eq 2.9 using the c values from table 2.1. The error bars in B represent the range of K_D values measured from the range of c values with the error in the free ligand linewidth, ν_F, propagated.

For the twelve compounds that were considered in figure 2.4B, all compounds gave single-point estimates that agreed within a range of one standard deviation over the range of binding affinities and concentrations that were tested. All twelve compounds had experimental and single-point estimates for K_D that agreed within two standard deviations. A higher deviation was observed in figure 2.4A for ligands with higher K_D values. This occurs because of differences between the individual c values and the average c values. Also, eq 2.9 is more sensitive to small changes in c at these high K_D values. This occurs because, at high K_D values, vanishingly small differences in NMR intensities correspond to large differences in K_D. In other words, this method is reaching a practical limit of detection since K_D rapidly approaches infinity as NMR peak intensity changes approach zero.

The relative ranking of the K_D values were also the same for results that were obtained by the single-point calculations or the full titration method. These results indicate the single-point method can, at least in cases such as these, provide a preliminary estimate of K_D values and binding affinities that can be used in the context of a high-throughput screening assay. At a minimum, the relative changes in linewidth provide a rapid and efficient mechanism to prioritize NMR screening leads for further evaluation. However, it is still recommended that a more robust approach for measuring binding affinities for promising leads follow the NMR ligand affinity screen. This precaution follows, in part, from the fact that the accuracy of the K_D values that are measured from the single-point ^{1}H NMR line-broadening experiments will be strongly dependent on having a reasonable estimate for the value of NMR linewidth ratio (c) in such a study.

2.5 REFERENCES

1. Anderson, A. C., The Process of Structure-Based Drug Design. *Chem. Biol.* **2003,** 10, (9), 787-797.
2. Lee, G. M.; Craik, C. S., Trapping Moving Targets with Small Molecules. *Science (Washington, DC, U. S.)* **2009,** 324, (5924), 213-215.
3. Keskin, O.; Gursoy, A.; Ma, B.; Nussinov, R., Towards drugs targeting multiple proteins in a systems biology approach. *Curr. Top. Med. Chem. (Sharjah, United Arab Emirates)* **2007,** 7, (10), 943-951.
4. Copeland, R. A., *Enzymes: A practical Introduction to Structure, Mechanism, Data Analysis*. Second Edition ed.; Wiley-VCH: New York, 2000; p 397.
5. Dalvit, C.; Ardini, E.; Flocco, M.; Fogliatto, G. P.; Mongelli, N.; Veronesi, M., A General NMR Method for Rapid, Efficient, and Reliable Biochemical Screening. *J Am Chem Soc* **2004,** 125, (47), 14620-14625.
6. Tengel, T.; Fex, T.; Emtenäs, H.; Almqvist, F.; Sethson, I.; Kihlberg, J., Use of 19F NMR spectroscopy to screen chemical libraries for ligands that bind to proteins. *Org. Biomol. Chem.* **2004,** 5, 725-731.
7. Dalvit, C.; Flocco, M.; Stockman, B. J.; Veronesi, M., Competition Binding Experiments for Rapidly Ranking Lead Molecules for their Binding Affinity to Human Serum Albumin. *Comb. Chem. High Throughput Screening* **2002,** 5, 645-650.
8. Jahnke, W.; Floersheim, P.; Ostermeier, C.; Zhang, X.; Hemmig, R.; Hurth, K.; Uzunov, D. P., NMR-Reporter Screening for the Detection of HIGH-Affinity Ligands. *Angew. Chem. Int. Ed.* **2002,** 41, (18), 3420-3423.

9. Fischer, J. J.; Jardetzky, O., Nuclear Magnetic Relaxation Study of Intermolecular Complexes. The Mechanism of Penicillin Binding to Serum Albumin. *J Am Chem Soc* **1965,** 87, (14), 3237-3244.

10. Sarrazin, M.; Sari, J. C.; Bourdeaux-Pontier, M.; Briand, C., NMR Study of the Interactions between Flurazepam and Human Serum Albumin. *Molecular Pharmacology* **1972,** 15, 71-77.

11. Peters, T., *All About Albumin : Biochemistry, Genetics, and Medical Applications*. Academic Press Inc.: London, 1995.

12. Cavanagh, J.; Fairbrother, W. J.; III, A. G. P.; Skelton, N. J., *Protein NMR Spectroscopy: Princples and Practice*. Academic Press: San Diego, 1996; p 587.

13. Packer, K. J.; Thomlinson, D. J., Nuclear Spin Relaxation and Self-Diffusion in the binary System, Dimethylsulphoxide (DMSO)+Water. *Transactions of the Faraday Society* **1971,** 67, 1302-1314.

14. Helms, M. K.; Petersen, C. E.; Bhagavan, N. V.; Jameson, D. M., Time-resolved fluorescence studies on site-directed mutants of human serum albumin. *FEBS Lett* **1997,** 408, (1), 67-70.

15. Zartler, E. R.; Yan, J.; Mo, H.; Kline, A. D.; Sharpiro, M. J., 1D NMR Methods in Ligand-Receptor Interactions. *Curr. Top. Med. Chem.* **2003,** 3, 25-37.

16. Coles, M.; Heller, M.; Kessler, H., NMR-based screening technologies. *Drug Discovery Today* **2003,** 8, (17), 803-810.

17. Siriwardena, A. H.; Tian, F.; Noble, S.; Prestegard, J. H., A straightforward NMR-Spectroscopy-Based Method for Rapid Library Screening. *Angew. Chem. Int. Ed.* **2002,** 41, (18), 3454-3457.

18. Dalvit, C.; Flocco, M.; Knapp, S.; Mostardini, M.; Perego, R.; Stockman, B. J.; Veronesi, M.; Varasi, M., High-Throughput NMR-Based Screening with Competition Binding Experiments. *J Am Chem Soc* **2002,** 124, 7702-7709.

19. 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* **1997,** 119, 12257-12261.

20. Chen, J.; Ohnmacht, C.; Hage, D. S., Studies of phenytoin binding to human serum albumin by high-performance affinity chromatography. *J Chromatogr B Analyt Technol Biomed Life Sci* **2004,** 809, (1), 137-45.

21. Kurtzhals, P.; Havelund, S.; Joanassen, I.; Markussen, J., Effect of Fatty Acids and Selected Drugs on the Albumin Binding of a Long-Acting, Acylated Insulin Analogue. *J. Pharm. Sci.* **1997,** 86, (12), 1365-1368.

22. Sengupta, A.; Hage, D. S., Characterization of minor site probes for human serum albumin by high-performance affinity chromatography. *Anal Chem* **1999,** 71, (17), 3821-7.

23. Kragh-Hansen, U.; Chuang, V. T. G.; Otagiri, M., Practical Aspects of the Ligand-Binding and Enzymatic Properties of Human Serum Albumin. *Bio. Pharm. Bull.* **2002,** 25, (6), 695-704.

24. Levitt, M. H., *Spin Dynamics: Basics of Nuclear Magnetic Resonance*. John Wiley & Sons, LTD: Chichester, UK, 2001; p 520-537.

25. Cantor, C. R.; Schimmel, P. R., *Biophysical Chemistry Part II: Techniques for the study of biological structure and function*. W. H. Freeman and Co.: San Francisco, 1980; p 461.

26. Bernado, P.; Garcia de la Torre, J.; Pons, M., Interpretation of 15N NMR relaxation data of globular proteins using hydrodynamic calculations with HYDRONMR. *J Biomol NMR* **2002,** 23, (2), 139-50.

27. Garcia de la Torre, J.; Huertas, M. L.; Carrasco, B., HYDRONMR: prediction of NMR relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. *J Magn Reson* **2000,** 147, (1), 138-46.

28. Ascenzi, P.; Bocedi, A.; Notari, S.; Menegatti, E.; Fasano, M., Heme impairs allosterically drug binding to human serum albumin Sudlow's site I. *Biochem Biophys Res Commun* **2005,** 334, (2), 481-6.

29. Ascoli, G.; Bertucci, C.; Salvadori, P., Stereospecific and competitive binding of drugs to human serum albumin: a difference circular dichroism approach. *J Pharm Sci* **1995,** 84, (6), 737-41.

30. Cheruvallath, V. K.; Riley, C. M.; Narayanan, S. R.; Lindenbaum, S.; Perrin, J. H., A quantitative circular dichroic investigation of the binding of the enantiomers of ibuprofen and naproxen to human serum albumin. *J Pharm Biomed Anal* **1997,** 15, (11), 1719-24.

31. Epps, D. E.; Raub, T. J.; Kezdy, F. J., A general, wide-rage spectrofluorometric method for measuring the site-specific affinities of drugs toward human serum albumin. *Anal Biochem* **1995,** 227, (2), 342-50.

32. Hage, D. S., High-performance affinity chromatography: a powerful tool for studying serum protein binding. *J Chromatogr B Analyt Technol Biomed Life Sci* **2002,** 768, (1), 3-30.

33. Hage, D. S.; Noctor, T. A.; Wainer, I. W., Characterization of the protein binding of chiral drugs by high-performance affinity chromatography. Interactions of R- and S-ibuprofen with human serum albumin. *J Chromatogr A* **1995,** 693, (1), 23-32.

34. Hollosy, F.; Valko, K.; Hersey, A.; Nunhuck, S.; Keri, G.; Bevan, C., Estimation of volume of distribution in humans from high-throughput HPLC-based measurements of human serum albumin binding and immobilized artificial membrane partitioning. *J Med Chem* **2006,** 49, (24), 6958-71.

35. Itoh, T.; Saura, Y.; Tsuda, Y.; Yamada, H., Stereoselectivity and enantiomer-enantiomer interactions in the binding of ibuprofen to human serum albumin. *Chirality* **1997,** 9, (7), 643-9.

36. Kragh-Hansen, U.; Chuang, V. T.; Otagiri, M., Practical aspects of the ligand binding and enzymatic properties of human serum albumin. *Biol Pharm Bull* **2002,** 25, (6), 695-704.

37. Lagercrantz, C.; Larsson, T., Comparative studies of the binding of some ligands to human serum albumin non-covalently attached to immobilized Cibacron Blue, or covalently immobilized on Sepharose, by column affinity chromatography. *Biochem J* **1983,** 213, (2), 387-90.

38. Lockwood, G. F.; Albert, K. S.; Szpunar, G. J.; Wagner, J. G., Pharmacokinetics of ibuprofen in man--III: Plasma protein binding. *J Pharmacokinet Biopharm* **1983,** 11, (5), 469-82.

39. Margarita Valero, B. E., Rafael Pelaez, Licesio J. Rodriguez, Naproxen: Hydroxypropyl-β-Cyclodextrin:Polyvinylpyrrolidone Ternary Complex

Formation. *Journal of Inclusion Phenomena and Macrocyclic Chemistry* **2004,** 48, 157-163.

40. Montero, M. T. E., J.; Valls, O., Binding of nonsteroidal anti-inflammatory drugs to human serum albumin. *International Journal of Pharmaceutics* **1990,** 62, (1), 21-25.

41. Paliwal, J. K.; Smith, D. E.; Cox, S. R.; Berardi, R. R.; Dunn-Kucharski, V. A.; Elta, G. H., Stereoselective, competitive, and nonlinear plasma protein binding of ibuprofen enantiomers as determined in vivo in healthy subjects. *J Pharmacokinet Biopharm* **1993,** 21, (2), 145-61.

42. Patonay, G.; Salon, J.; Sowell, J.; Strekowski, L., Noncovalent labeling of biomolecules with red and near- infrared dyes. *Molecules* **2004,** 9, (3), 40-9.

43. Peters, T., *All About Albumin: Biochemistry, Genetics, and Medical Applications*. San Diego, 1996.

44. Pirnau, A. a. B., M, Investigation of the Interaction Between Naproxen and Human Serum Albumin. *Romanian Journal of Biophysics* **2008,** 18, (1), 49-55.

45. Rahman, M. H.; Maruyama, T.; Okada, T.; Imai, T.; Otagiri, M., Study of interaction of carprofen and its enantiomers with human serum albumin--II. Stereoselective site-to-site displacement of carprofen by ibuprofen. *Biochem Pharmacol* **1993,** 46, (10), 1733-40.

46. Rich, R. L.; Day, Y. S.; Morton, T. A.; Myszka, D. G., High-resolution and high-throughput protocols for measuring drug/human serum albumin interactions using BIACORE. *Anal Biochem* **2001,** 296, (2), 197-207.

47. Sowell, J.; Mason, J. C.; Strekowski, L.; Patonay, G., Binding constant determination of drugs toward subdomain IIIA of human serum albumin by near-infrared dye-displacement capillary electrophoresis. *Electrophoresis* **2001,** 22, (12), 2512-7.

48. Trivedi, V. D.; Saxena, I.; Siddiqui, M. U.; Qasim, M. A., Interaction of bromocresol green with different serum albumins studied by fluorescence quenching. *Biochem Mol Biol Int* **1997,** 43, (1), 1-8.

49. Whitlam, J. B.; Crooks, M. J.; Brown, K. F.; Pedersen, P. V., Binding of nonsteroidal anti-inflammatory agents to proteins--I. Ibuprofen-serum albumin interaction. *Biochem Pharmacol* **1979,** 28, (5), 675-8.

50. Yamasaki, K.; Rahman, M. H.; Tsutsumi, Y.; Maruyama, T.; Ahmed, S.; Kragh-Hansen, U.; Otagiri, M., Circular dichroism simulation shows a site-II-to-site-I displacement of human serum albumin-bound diclofenac by ibuprofen. *AAPS PharmSciTech* **2000,** 1, (2), E12.

51. Chen, J.; Hage David, S., Quantitative analysis of allosteric drug-protein binding by biointeraction chromatography. *Nat Biotechnol* **2004,** 22, (11), 1445-8.

52. Trivedi, V. D., On the role of lysine residues in the bromophenol blue-albumin interaction. *Italian Journal of Biochemistry* **1997,** 46, (2), 67-73.

53. Chen, J. O., Corey; Hage, David, Quantitative analysis of allosteric drug-protein binding by biointeraction chromatography. *Nature biotechnology* **2004,** 22, (11).

54. Hajduk, P.; Betz, S. F.; Mack, J.; Ruan, X.; Towne, D. L.; Lerner, C. G.; Beutel, B. A.; Fesik, S. W., A strategy for high-throughput assay development using leads

derived from nuclear magnetic resonance-based screening. *J Biomol Screen* **2002,** 7, (5), 429-432.

Appendix A: Glossary of terms found in chapter 2

L	small-molecule ligand
$[L]_T$	total ligand concentration
$[L]_F$	free ligand concentration
P	protein target
$[P]_T$	total protein concentration
$[P]_F$	free protein concentration
[PL]	protein-ligand complex concentration
I_B	NMR peak height of bound ligand
I_F	NMR peak height of free ligand
K_D	dissociation equilibrium constant for a protein-ligand complex
c	NMR linewidth ratio constant
B	NMR signal response dependent on fraction of bound ligand
B_{single}	NMR signal response dependent on fraction of bound ligand at a single
ν_F	linewidth of the free ligand
ν_B	linewidth of the bound protein-ligand complex
ν_P	linewidth of the protein
ν_{obs}	observed linewidth change upon addition of protein or ligand
f_B	fraction bound complex in solution
f_F	fraction of free ligand in solution
T_2^{-1}	dipole-dipole relaxation constant
τ_c	correlation time

$J(\omega)$	normalized density function of T_2^{-1}
B_o	static magnetic field strength
ω_o	Larmor frequency
MW_P	molecular weight of a protein target

Appendix B: Derivation of equations for rapid K_D method found in chapter 2

The binding of a protein (P) with a single small ligand (L) can be represented by the following reaction.

$$[PL] \Leftrightarrow [P] + [L] \qquad [B1]$$

The dissociation equilibrium constant for this system is described by the expression in eq B2, where the concentrations $[P]_F$, $[L]_F$ and [PL] represent the concentration of free protein, free ligand, and protein-ligand complex, respectively.

$$K_D = \frac{[P]_F[L]_F}{[PL]} \qquad [B2]$$

Based on mass balance, eq B3 can be used to express $[L]_F$ and [PL] in terms of the total ligand concentration and other concentrations in this system.

$$[P]_T - [P]_F = [PL] = [L]_T - [L]_F \Rightarrow [L]_F = [L]_T - [P]_T + [P]_F \qquad [B3]$$

Substitution of these relationships into eq B2 gives eq B4.

$$K_D = \frac{[P]_F\ [L]_T\text{-}[P]_T\text{+}[P]_F}{[P]_T\text{-}[P]_F} \qquad [B4]$$

Eq B4 can now be rearranged into the following form,

$$\begin{aligned} &K_D\ [P]_T\text{-}[P]_F = [P]_F\ [L]_T\text{-}[P]_T\text{+}[P]_F \\ &\Rightarrow [P]_F^2 + [L]_T\text{-}[P]_T\text{+}K_D\ [P]_F\text{-}K_D[P]_T = 0 \end{aligned} \qquad [B5]$$

which makes it possible to solve for $[P]_F$ by using the quadratic formula, as indicated in eq B6, where only the positive root has any meaning in a real protein-ligand system.

$$[P]_F = \frac{-\ [L]_T\text{-}[P]_T\text{+}K_D \pm \sqrt{[L]_T\text{-}[P]_T\text{+}K_D\ ^2\text{+}4K_D[P]_T}}{2} \qquad [B6]$$

The bound fraction of ligand f_B is next defined as given in eq B7.

$$f_B = \frac{[PL]}{[PL]+[L]_F} = \frac{1}{1+\frac{K_D}{[P]_F}} \qquad [B7]$$

If we substitute the positive root of eq B6 into eq B7, the result is eq B8.

$$f_B = \frac{1}{1+\frac{2K_D}{-[L]_T-[P]_T+K_D+\sqrt{[L]_T-[P]_T+K_D{}^2+4K_D[P]_T}}} = \frac{1}{1+\left(\frac{2K_D}{[L]_T-[P]_T+K_D}\right)\left(\sqrt{1+\frac{4K_D[P]_T}{[L]_T-[P]_T+K_D{}^2}}-1\right)^{-1}} \qquad [B8]$$

A further simplification of eq B8 can be accomplished by expanding the square root as a power series where $x = \frac{4K_D[P]_T}{[L]_T-[P]_T+K_D{}^2}$ about $x = 0$. This approach is valid as long as the ligand is in considerable excess relative to the protein. The power series that is used here is shown below.

$$\sqrt{1+x} = 1+\frac{x}{2}-\frac{x^2}{8}+\ldots \qquad [B9]$$

If eq B9 is truncated at the second term, this allows the square root term in eq B8 to be written in the approximate form that is given in eq B10.

$$\sqrt{1+\frac{4K_D[P]_T}{[L]_T-[P]_T+K_D{}^2}} \approx 1+\frac{2K_D[P]_T}{[L]_T-[P]_T+K_D{}^2} \qquad [B10]$$

The overall result of this simplification is that eq B8 converts to the expression shown below, there the fraction of bound ligand f_B is now described in terms of only K_D, the total ligand concentration and the total protein concentration.

$$f_B \approx \frac{1}{1+\frac{[L]_T-[P]_T+K_D}{[P]_T}} = \frac{[P]_T}{[L]_T+K_D} \qquad [B11]$$

If it is assumed the observed free and bound NMR linewidths are represented by ν_F and ν_B, respectively, and that exchange occurs between free and bound states, the general solution to the NMR lineshape is bilorentzian. In the slow limit, the spectrum is obviously just a sum of the spectra of free and bound species, weighted by their relative abundances. If exchange rates become comparable to the inverse linewidths, then a conventional solution of the pair of coupled linear differential equations, including auto and cross relaxation terms but neglecting any chemical-shift difference between the states, gives a time domain (free induction decay):

$$f(t) = c_+ e_+ + c_- e_- \qquad [B12.a]$$

with

$$e_\pm = \exp\left[\left(\Theta \pm \sqrt{\Delta}\right)t\right] \qquad [B12.b]$$

$$c_\pm = c_2 \pm \frac{c_1}{\sqrt{\Delta}} \qquad [B12.c]$$

$$c_1 = \tfrac{1}{4}\left[\left(\bar{\bar{K}}_{11}+2\bar{\bar{K}}_{21}-\bar{\bar{K}}_{22}\right)M_L(0)-\left(\bar{\bar{K}}_{11}-2\bar{\bar{K}}_{12}-\bar{\bar{K}}_{22}\right)M_{PL}(0)\right] \qquad [B12.d]$$

$$c_2 = \tfrac{1}{2}\left(M_L(0)+M_{PL}(0)\right) \qquad [B12.e]$$

$$\Delta = \left(\frac{\bar{\bar{K}}_{11}-\bar{\bar{K}}_{22}}{2}\right)^2 + \bar{\bar{K}}_{12}\bar{\bar{K}}_{21} \qquad [B12.f]$$

$$K_{11} = -\frac{1}{T_{2,f}} - k_1[P] \qquad [B12.g]$$

$$K_{22} = -\frac{1}{T_{2,b}} - k_{-1} \qquad \text{[B12.h]}$$

$$K_{12} = k_{-1} \qquad \text{[B12.i]}$$

$$K_{21} = k_1[\mathrm{P}] \qquad \text{[B12.j]}$$

where M_L and M_{PL} are the magnetization of the free and bound species, respectively. In the fast exchange limit, the solution is still formally biexponential, but the coefficient c_- goes to zero, and the free induction decay signal, normalized to unity at zero time, becomes

$$f(t) = \exp\left(-\frac{1}{T_{2,f}}\left\{\frac{[\mathrm{L}]}{([\mathrm{L}]+[\mathrm{PL}])}\right\} - \frac{1}{T_{2,b}}\left\{\frac{[\mathrm{PL}]}{([\mathrm{L}]+[\mathrm{PL}])}\right\}\right) = \exp\left(-\frac{f_f}{T_{2,f}} - \frac{f_b}{T_{2,b}}\right) \qquad \text{[B13]}$$

Fourier transforming, the fast exchange NMR signal height can be written as shown in eq B14:

$$\mathrm{I_B} = \frac{\mathrm{I_F}\upsilon_\mathrm{F}}{\mathrm{f_F}\upsilon_\mathrm{F} + \mathrm{f_B}\upsilon_\mathrm{B}} \qquad \text{[B14]}$$

where I_F is the height of the ligand signal in the absence of protein and I_B is the observed peak height of the bound complex. This is exactly the same as the height of the free ligand signal in extreme slow exchange! Rearranging eq B14 explains the observed decrease in NMR peak signal for a free small-molecule ligand upon its binding to a protein. The relative ratio of NMR peak height ($\frac{\mathrm{I_B}}{\mathrm{I_F}}$) is now in terms of the fraction of free ligand (f_F) and the fraction of bound ligand (f_B) and is dependent on the observed increase in NMR linewidth upon the binding of a ligand to a protein.

$$1-\frac{I_B}{I_F}=1-\frac{\upsilon_F}{f_F\upsilon_F+f_B\upsilon_B}=1-\frac{1}{f_F+\ f_B\,\upsilon_B/\upsilon_F}=1-\frac{1}{1-f_B+\ f_B\,\upsilon_B/\upsilon_F}$$

$$=1-\frac{1}{1+f_B\left(\frac{\nu_B}{\nu_F}-1\right)} \qquad \text{[B15]}$$

Inserting B11 into B15 provides a measure of the dissociation equilibrium constant for the protein-ligand complex by relating the fraction of bound ligand to the observed change in NMR peak height.

$$B=1-\frac{I_B}{I_F}=1-\frac{1}{1+\frac{[P]_T}{[L]_T+K_D}(\frac{\nu_B}{\nu_F}-1)}=1-\frac{1}{1+\frac{c[P]_T}{[L]_T+K_D}} \qquad \text{where} \quad c=\frac{\upsilon_B}{\upsilon_F}-1 \qquad \text{[B16]}$$

The NMR linewidth ratio, c, is then measured by using the free ligand NMR spectrum and by assuming the linewidth of the bound complex approximates the linewidth of the protein.

CHAPTER 3:

STRUCTURAL AND FUNCTIONAL SIMILARITY BETWEEN THE BACTERIAL TYPE III SECRETION SYSTEM NEEDLE PROTEIN PRGI AND THE EUKARYOTIC APOPTOSIS BCL-2 PROTEINS

3.1 INTRODUCTION

The previous chapter discussed the development of a high-throughput screening methodology to measure and rank relative binding affinities. One of the primary reasons for developing such a method was for the use in the Functional Annotation Screening Technology by NMR (FAST-NMR).[1] The FAST-NMR method is a multi-step approach to high-throughput screening using nuclear magnetic resonance (NMR). A target protein is screened with a library of biologically functional compounds to identify which compounds bind to the target protein, known as a "hit". The first step in the FAST-NMR approach is a 1D ^{1}H line broadening experiment, similar to the experiments described in chapter 2.

The 1D ^{1}H line broadening experiment is a ligand focused experiment in which the response of the free ligand is compared to a sample with the target protein added. For FAST-NMR, the 1D ^{1}H line broadening step is used to identify potential hits as an initial screen. The method developed in chapter 2 is then used to prioritize which ligand-protein interactions are further studied using a secondary target focused 2D ^{1}H-^{15}N HSQC screen based on relative binding affinity. The 2D ^{1}H-^{15}N HSQC monitors the changes in the protein spectrum upon addition of the binding ligands. FAST-NMR also utilizes the Comparison of Protein Active Site Structures (CPASS) software and database to identify similar sequence and structure characteristics between experimentally identified ligand binding sites for proteins of known and unknown function.[2]

Functional regions of a protein are more stable relative to the remainder of the protein sequence undergoing random drift.[3, 4] The correlation between ligand binding sites, ligand structure and protein function has also been demonstrated by a network of ligand binding-site similarity described by Park & Kim.[5] A variety of computational methods have attempted to exploit the stability of functional regions by identifying ligand binding sites as a method to predict function.[6, 7] Unfortunately, the combined requirements of predicting the ligand, the binding site, and a similarity to an annotated proteins leads to a high level of ambiguity. The FAST-NMR approach attempts to experimentally identify ligand binding sites to annotate proteins of unknown function.[7-9] Applying the FAST-NMR method to previously annotated systems also enables experimental ligand binding site data to identify functional relationships that otherwise would not be recognized based solely on global sequence and structure similarity.

The type three secretion system (T3SS) is composed of 20-25 different proteins, which are assembled in a highly choreographed mechanism similar to the assembly of flagella.[10-12] In *Salmonella typhimurium*, the needle complex is responsible for puncturing a host's cell membrane to allow effector proteins (SipB, SipC, SipD) from *S. typhimurium* to enter the host.[13] Many of these effectors can activate bacterial induced apoptosis of a hosts' cell by interacting with capsase-1[14] in a mechanism similar to apoptosis in eukaryotic cells.[15] The needle complex is a large homomultimer composed of ~120 repeated copies of the monomeric protein PrgI, a small helical protein of 83 amino acids.[16] The monomeric form of PrgI is a helix-turn-helix motif with two symmetrically charged surfaces and a conserved loop region, PxxP domain, which are important for needle assembly.[16-18] The charged surfaces of PrgI responsible for needle

assembly also provide a potential binding site for small molecule ligands. This makes PrgI an attractive drug target to disrupt the formation of the needle complex and prevent infection by *S. typhimurium*. However, to date there has been no reported ligands that bind to either region of this protein-protein interaction site.

The PrgI needle complex protein from *S. typhimurium* T3SS was screened in our FAST-NMR assay, which resulted in the identification of a functional similarity between the ligand binding sites of PrgI and the anti-apoptosis protein Bcl-xL. Additionally, Dali[19] and T-Coffee[20] analysis found regions of structure and sequence similarity between the two proteins consistent with the FAST-NMR results. The predicted active-site similarity between PrgI and Bcl-xL was also used to experimentally verify that chelerythrine,[21] a ligand known to inhibit Bcl-xL and induce apoptosis, also binds PrgI. These results provide experimental evidence that suggest a functional relationship between the bacterial type III secretion systems and apoptosis. This is consistent with a general conservation in function between PrgI and the Bcl-2 family of proteins that includes Bcl-xL; both form membrane pores through oligomerization using a conserved helix-turn-helix motif to release effectors to stimulate cell death.

3.2 EXPERIMENTAL

3.2.1 FAST-NMR screen of PrgI. The *Salmonella typhimurium* type three secretion protein (T3SS) PrgI was screened with a functional library [22] using the FAST-NMR assay.[7, 8] Unlabeled and ^{15}N labeled monomeric PrgI was graciously provided by Dr. Roberto DeGuzman (University of Kansas) along with the assigned 2D ^{1}H-^{15}N HSQC spectrum. Sample preparation and experimental parameters for the NMR screen

were executed in the same manner as described previously [8]. Briefly, each ligand mixture was screened at 100 μM/ligand concentration with 25 μM protein in a 99.99% D_2O buffered solution of 20 mM d_{19}-bis-Tris at pH 7.0 with 5% DMSO-d_6 to maintain ligand solubility and 11.1 μM 3-(trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt as a chemical shift reference. 1D ^{1}H NMR spectra for each sample was collected using a presaturation pulse sequence with 64 real transients, 8 dummy transients with 8 k data points, a sweep width of 11.0 ppm and a recycle delay of 2.0 s. Data was Fourier transformed, auto-phase and baseline corrected. Each 1D ^{1}H NMR spectrum were compared to the corresponding free ligand mixture reference spectrum and visually analyzed to identify binding ligands. A binding event was identified by the decrease in ligand intensity of the nuclease-mixture relative to the free ligand mixture. Total data collection time including sample changing was approximately 10 min/spectrum. All 1D ^{1}H NMR spectra were collected on a Bruker 500 MHz Avance spectrometer (Billercia, MA) equipped with a triple resonance, Z-axis gradient cryoprobe and using a Bruker BACS-120 sample changer and IconNMR software for automated data collection. All spectra were collected at 298 K.

All 2D ^{1}H-^{15}N HSQC spectra were collected at 298K using the same instrumentation with the standard 2D ^{1}H-^{15}N HSQC (*hsequetf3gp*) pulse sequence implemented in Bruker TopSpin 1.3 with optimized sample specific 90^o pulse lengths. A total of 16 real scans and 128 dummy scans were collected with 2 k data points with a sweep width of 9.5 ppm in the ^{1}H dimension and 128 data points with a sweep width of 28.0 ppm in the ^{15}N dimension. A ligand free 2D ^{1}H-^{15}N HSQC spectrum was collected using the same buffer conditions with 95% H_2O/5% D_2O to ensure the protein was

properly folded prior to addition of ligands. Total experiment time was approximately 1.5 hrs/spectrum.

A total of 113 1D ^{1}H NMR line-broadening spectra were collected to identify 5 binding ligands from the functional chemical library of 437 compounds. Measurement of binding dissociation constants were completed as described in chapter 2 and as described previously.[23] Secondary 2D ^{1}H-^{15}N HSQC NMR experiments were collected only for the 5 compounds identified as binders in the line-broadening experiments. Chemical shift perturbations (CSPs) (eq 3.1) from the 2D ^{1}H-^{15}N HSQC NMR experiments were used to identify the PrgI ligand binding site, where only residues with a CSP greater than one standard deviation from the mean were used

$$CSP = \left[\frac{(\Delta NH)^2 + (\frac{\Delta 15N}{5})^2}{2}\right]^{1/2} \quad [3.1]$$

where *ΔNH* is the difference between free and bound ^{1}H amide chemical shifts (ppm) and *Δ15N* is the difference between free and bound ^{15}N chemical shifts (ppm).

A rapid approach to determine a ligand binding orientation was employed to determine a PrgI co-structure in the same manner as described previously.[24] The CSPs minimize the search space by using a significantly reduced AutoDock 3D grid. AutoDock 4.0 was used to generate 100 docked PrgI-ligand co-structures using the Lamarckian search algorithm with a population size of 300 and 500,000 energy evaluations.[25] The AutoDockFilter (ADF) program then uses an NMR energy function based on the magnitude of CSPs to select the best ligand conformation.

$$E_{NMR} = k\sum_{i=1}^{n}(\Delta_{Dist})^2 \quad \Delta_{Dist} = \begin{cases} d_{CSP} & d_{CSP} < d_s \\ 0 & d_s \leq d_{CSP} \end{cases} \quad [3.2]$$

where ADF calculates a pseudo-distance (d_{CSP}) based on the magnitude of the NH CSP, which 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. Comparison of these CSP-directed and selected ligand-docked structures with experimental x-ray and NMR structures has yielded an overall average rmsd of 1.17 ± 0.74 Å.[24]

A co-structure of the lipid derivative didecyldimethylammonium bromide (DDAB) bound to PrgI was uploaded to the CPASS database (http://cpass.unl.edu) to identify proteins with similar ligand binding sites by maximizing an rmsd weighted BLOSUM62[26, 27] scoring function ($\boldsymbol{S_{ab}}$).

$$S_{ab} = \sum_{i,j=1}^{i=n,j=m} \frac{d_{min}}{d_i} \left(e^{-\Delta rmsd_{i,j}}\right)^2 p_{i,j} \quad \Delta rmsd_{i,j} = \begin{cases} rmsd_{i,j} - 1 & rmsd_{i,j} > 1\text{Å} \\ 0 & rmsd_{i,j} \leq 1\text{Å} \end{cases} \qquad [3.3]$$

where active site ***a*** contains ***n*** residues and is compared to active site ***b*** from the CPASS database which contains ***m*** residues, $\boldsymbol{p_{i,j}}$ is the BLOSUM62 probability for amino-acid replacement for residue ***i*** from active site ***a*** with residue ***j*** from active site ***b***, $\boldsymbol{\Delta_{i,j}}$ is a corrected root-mean square difference in the Cα coordinate positions between residues ***i*** and ***j***, and $\boldsymbol{d_{min}/d_i}$ is the ratio of the shortest distance to the ligand among all amino-acids in the active site compared to the current amino-acid's shortest distance to the ligand. $\boldsymbol{S_{ab}}$ is only summed over the optimal alignment for residue ***i*** from active site ***a*** with residue ***j*** from active site ***b***. It is not summed over all possible combinations of ***i*** and ***j***. If the number of residues are not identical between active sites ***a*** and ***b*** (n ≠ m), then the additional residues will not have a corresponding match. Each residue can only be used once in the alignment. If active site ***a*** contains unmatched residues, then no contribution is made to $\boldsymbol{S_{ab}}$ which effectively reduces the maximal possible score that can be achieved

for active site ***a***. At the time of this study (May 2008), there were ~35,000 protein-ligand structures in the CPASS database. CPASS was run on a 16 node Beowulf Linux cluster, requires approximately 40 sec for each pair-wise comparison and took ~24 hrs to complete a full search against the entire database.

3.2.2 Structure similarity searching Native protein structures for PrgI (PDB ID: 2JOW) [16] and Bcl-xL (PDB ID: 1YSN)[28] were uploaded to the DaliLite[29] web server (http://www.ebi.ac.uk/DaliLite/) to identify regions of structure homology between the two proteins. To identify structure similarity and possible homology with other proteins within the PDB, the structures were also uploaded to the full Dali[19] web server (http://www.ebi.ac.uk/dali/). A truncated version of the Bcl-xL structure was generated by identifying the amino acids within regions of structure similarity and removing these residues from the native PDB file. The truncated PDB file was searched for regions of similarity using the DaliLite web server (http://www.ebi.ac.uk/Tools/dalilite/index.html).

3.2.3 Sequence similarity searching using BLAST and T-Coffee. Sequences from the T3SS and apoptosis regulation were downloaded from the NCBI server (http://www.ncbi.nlm.nih.gov/) and included PrgI (gi|16766179), InvJ (gi|16766198) and InvG (gi|474941) from *S. typhimurium*, and Bcl-xL, (gi|510901), Bak1 (gi|82571458), Bid (gi|4557361) and Bax (gi|231632) from *Homo sapiens* respectively. A full BLAST search was completed using these sequences associated with both systems as queries.[30] All BLAST sequence searches used default settings. In addition, the sequences and structures for Bcl-xL (PDB-ID: 1YSN), *S. typhimurium* PrgI (PDB-ID: 2JOW), *B. pseudomallei* BsaL (PDB-ID: 2GOU) and *S. flexneri* MxiH (PDB-ID: 2CA5) obtained from the PDB were uploaded to the T-Coffee[20] web server (http://www.tcoffee.org/) to

obtain a multiple sequence alignment using the EXPRESSO(3DCoffee) software.[31] Only the sequence region of the Bcl-xL structure that contained the pore-forming domain and yielded the highest alignment score was used for the multiple sequence alignment.

3.2.4 Secondary binding site similarity between Bcl-xL and PrgI. To further support a structural and functional similarity between Bcl-xL and PrgI, the BindingDB[32] (http://www.bindingdb.org/) was searched for commercially available compounds to test for binding to PrgI. The free 2D ^{1}H-^{15}N HSQC spectrum was collected using 100 μM ^{15}N labeled PrgI in 20 mM bis-Tris buffer with 100 mM sodium chloride at pH 7.0. A second PrgI sample was prepared in the same manner as above with the addition of 500 μM chelerythrine to generate the bound 2D ^{1}H-^{15}N HSQC spectrum. Chemical shift perturbations and a PrgI-chelerythrine docked co-structure were determined as described previously[24] and was compared to the Bcl-xL-chelerythrine model [33].

3.3 RESULTS

3.3.1 Results from the FAST-NMR screen. The needle complex protein, PrgI, from *S. typhimurium* is an attractive antibacterial target because the protein is exposed to the cell surface and blocking this target could prevent injection of virulence factors into the host.[34] The interaction of PrgI with the host membrane stimulates the delivery of effectors from the bacteria into the host cytosol to induce cell death. Recently an NMR structure was determined for a monomeric form of PrgI,[16] which enabled the screening of PrgI using the FAST-NMR assay.[8] FAST-NMR combines NMR ligand affinity screening[35] using a fragment-based functional library[22] with structural biology and bioinformatics[2] to rapidly determine protein-ligand complexes[24] and infer functional

relationship between proteins based on similarities in functional epitopes. Also, the resulting protein-ligand co-structure provides a valuable starting point for structure-based drug design.

FAST-NMR applies a tiered approach to screening[35] to minimize resources and increase throughput (figure 3.1). First, PrgI was screened with the functional chemical library using 1D ^{1}H NMR line-broadening experiments. Five compounds (L-carnitine inner salt, didecyldimethylammonium bromide, 1-methylimidazole, methiothepin mesylate salt, sucrose) were found to bind PrgI by showing a significant decrease in ^{1}H peak intensity upon addition of 25 μM of PrgI. This was determined by comparing normalized ^{1}H ligand peak intensities between the free and bound NMR spectra (figure 3.1A). However, the secondary 2D ^{1}H-^{15}N HSQC experiments identified the lipid derivative didecyldimethylammonium bromide (DDAB) as the only specific PrgI binder (figure 3.1B) based on the observation of a significant number of chemical shift changes in the spectrum. The remaining four compounds elicited no change in chemical shifts in the PrgI 2D ^{1}H-^{15}N HSQC spectrum, which suggest the compounds bound non-specifically to PrgI. PrgI was found to bind DDAB with a K_D of 553 μM as calculated by a 1D ^{1}H line broadening method of chapter 2.[23] Finding a lipid derivative that specifically binds to PrgI is consistent with the protein's function; sensing new host cells and signaling secretion through an interaction with the host membrane.[36]

Chemical shift perturbations (CSPs) in the 2D ^{1}H-^{15}N HSQC experiments between free PrgI and the complex identified the PrgI residues that bind DDAB. Mapping these CSPs onto the PrgI surface identified the DDAB binding site as corresponding to residues at the bifurcation point of the two helices (figure 3.1C). Specifically, residues

S6, L9, S13, K15, and D17 of helix 1 and N59, V65, K66, V67, F68, K69, D70, D72, A73 and L76 of helix 2 showed significant CSPs in the presence of DDAB as calculated by eq 3.1. This ligand binding site has been shown to be important for the formation of the T3SS needle complex in which PrgI forms a repeating coiled-coils structure.[11] According to recent alanine scanning and structural studies, the surface residues in the region between the bifurcation point of the two helices and the conserved loop region, PxxP domain, are important for needle assembly.[16-18] These residues bind to the backside of the bifurcation point of the two helices in a stacked N-terminus to C-terminus manner.[16-18]

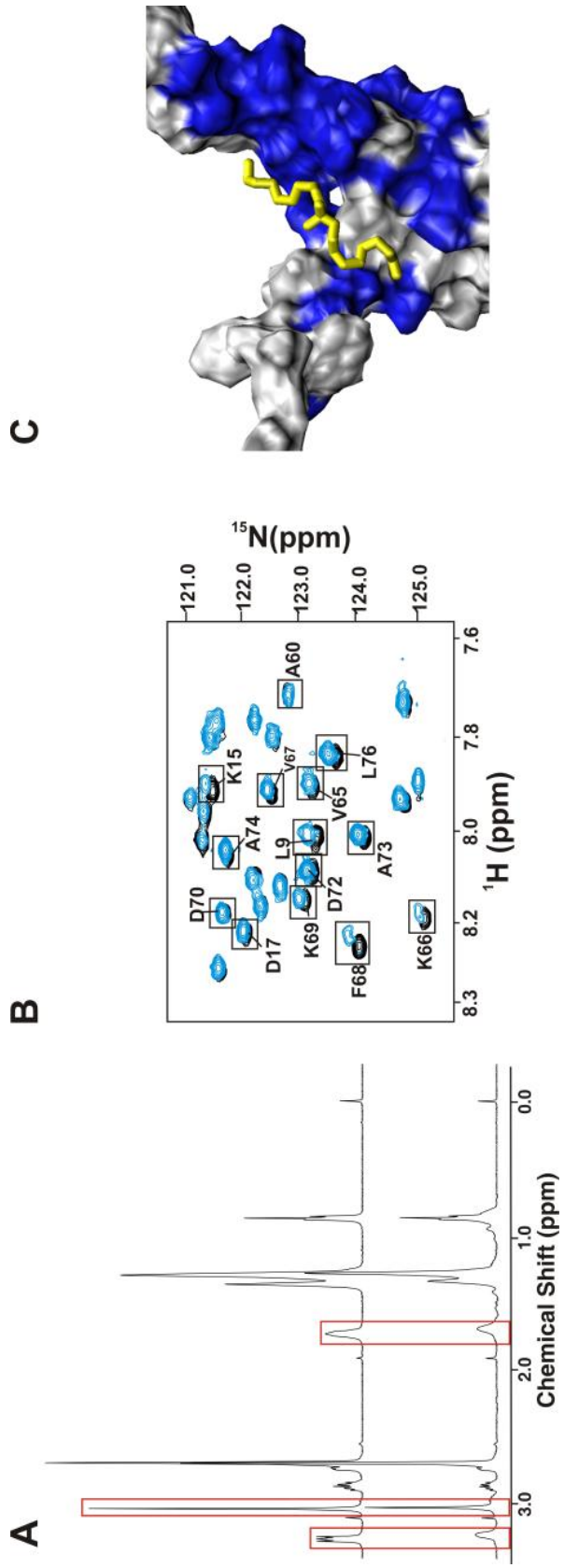

A
Chemical Shift (ppm)
3.0
2.0
1.0
0.0
B
^{15}N(ppm)
121.0
122.0
123.0
124.0
125.0
^{1}H (ppm)
8.3
8.2
8.0
7.8
7.6
A60
K15
V67
L76
V65
A74
L9
A73
D72
D70
K69
D17
F68
K66
C

Figure 3.1: Identification of PrgI Binding Ligands. (A) DDAB NMR spectra in the absence (*top*) and presence (*bottom*) of PrgI illustrating changes in NMR intensities (boxed) upon binding PrgI. Both free and bound 1D ^{1}H NMR spectra were normalized to a constant DMSO signal intensity. (B) Expanded view of the superimposed 2D ^{1}H-^{15}N HSQC spectra of the free and DDAB bound PrgI NMR samples. Residues that incur a chemical shift perturbation are boxed. (C) Expanded view of PrgI surface rendered in VMD[37] where residues that incur a chemical shift change are colored blue and DDAB is colored yellow. Co-structure based on NMR determined ligand binding site using AutoDock and our AutoDockFilter program.

The PrgI residues exhibiting significant CSPs upon binding DDAB were used to guide and filter a molecular docking simulation based on our method to rapidly determine protein-ligand co-structures.[24] AutoDock 4.0[25] was used to calculate 100 docked structures within a 3D grid defined by the CSPs. Our AutoDock Filter program (ADF) selected the best conformer based on consistency with the magnitude of chemical shift changes.[24] The ligand is expected to be closest to the protein residues that incurred the largest CSPs. The best PrgI-DDAB docked structure is shown in figure 3.1C, where DDAB adopts an extended conformation that straddles both helices of PrgI.

3.3.2 Analysis of CPASS and structure similarity results. Comparison of Protein Active Site Structures (CPASS) analysis of the PrgI-DDAB complex identified a human Bcl-2 protein family member (the anti-apoptosis regulating protein Bcl-xL (PDB-ID:1YSN) complexed to an acyl-sulfonamide-based inhibitor (ABT-737))[28] as the top hit based on a ligand binding-site CPASS similarity score of 37.7%. The CPASS alignment is shown in figure 3.2A and is based on maximizing the spatial orientation of similar residue types between the two ligand binding sites. All other proteins with a CPASS similarity > 30% were also evaluated, but Bcl-xL was the only protein that gave a reliable CPASS score and showed some level of structure or sequence similarity to PrgI. It is important to note the CPASS identified similarity between PrgI and Bcl-xL was fundamentally dependent on the existence of a Bcl-xL-ligand complex in the PDB. Ligand complexes for other members of the Bcl-2 protein family (Bax, Bid) currently do not exist.

While DDAB and ABT-737 are distinctly different ligands, the compounds share strong similarities in their mode of protein interactions. ABT-737 binds Bcl-xL edge-on

in an elongated conformation where a minimal number of atoms contact the hydrophobic binding cleft of Bcl-xL. In this manner, DDAB mimics this edge contact interaction of ABT-737 with the similar hydrophobic binding cleft in PrgI. Also, ABT-737 binds in a protein-protein binding interface similar to DDAB, where inhibiting protein interactions is the drugs mechanism of action in cancer cells.[28] Thus, the PrgI and Bcl-xL ligand binding-sites are functionally similar.

A pairwise structure alignment using DaliLite[29] yielded a non-significant Z-score of 1.4 and only 6% sequence identity between PrgI (PDB ID:2JOW) and Bcl-xL (PDB ID:1YSN). Nevertheless, the helix-turn-helix structure of PrgI (residues S13-V65) overlaps the buried helix-turn-helix motif (N136-I182) in Bcl-xL that corresponds to helices α5 (residues W137-D156) and α6 (residues L162-D176) (figure 3.2B). A focused pairwise comparison between the full PrgI protein and the α5 and α6 helices of Bcl-xL gave a low but significant Z-score of 3.3 with an root-mean-square-difference (rmsd) of 3.1Å. The sequence identity also increases from 6% to 9% between the full and focused pairwise alignments, respectively.

While there is an overlap between the DaliLite alignment of PrgI with Bcl-xL and the protein ligand binding sites identified by CPASS, these sites are not identical. This arises because the CPASS similarity is not confined by the primary sequence of the two proteins, but simply captures the spatial orientation of conserved residues around a ligand binding site. This is illustrated by the non-sequential sequence alignment of the PrgI and Bcl-xL ligand binding sites in figure 3.2. The exclusion of the sequence connectivity as a constraint to determine an alignment illustrates the advantage of CPASS in identifying a functional relationship over global sequence and structure alignments.[29, 30]

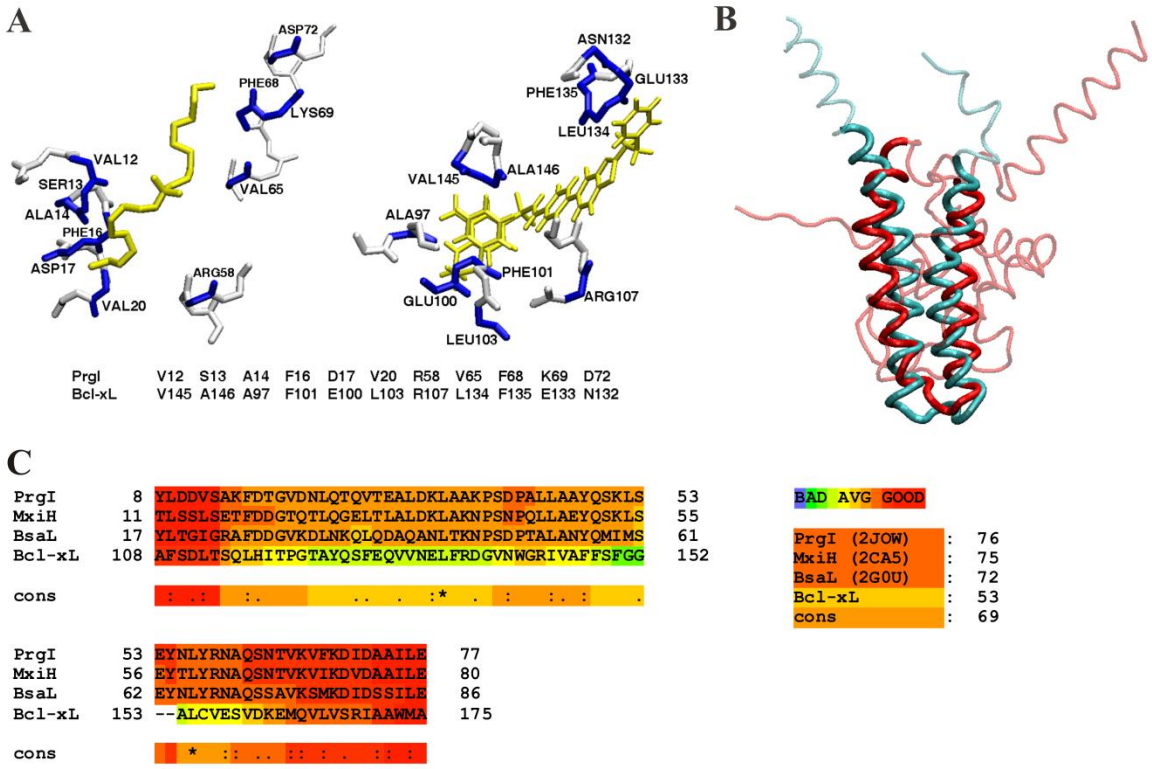


Figure 3.2. Active Site Similarity between PrgI and Bcl-xL. (A) CPASS alignment of the *S. typhimurium* PrgI active-site complexed to DDAB with the active-site of human Bcl-2 protein (Bcl-xL) complexed with acyl-sulfonamide-based inhibitor. The residues aligned by CPASS are labeled and colored blue in the structures. The active site sequence alignment is also shown below the structures. The ligands are colored yellow. (B) Overlay of the human Bcl-2 protein (red) with *S. typhimurium* PrgI (turquoise) based on a DaliLite alignment. (C) Multiple-sequence alignment of the three known T3SS structures of *S. typhimurium* PrgI, *B. pseudomallei* BsaL, and *S. flexneri* MxiH with the human Bcl-2 protein (Bcl-xL). The reliability of the each amino acid alignment is color-coded from blue (poor) to red (good) using the CORE index.[38] The consensus alignment received a score of 69, where a perfect alignment receives a score of 100.

3.3.3 Sequence similarity results. A BLAST[30] homology search using the PrgI and Bcl-xL sequences did not yield any significant information relating PrgI to Bcl-xL. The Bcl-xL sequence only identified homology to other Bcl-2 proteins. Similarly, the PrgI sequence was only aligned to other T3SS needle proteins. This is consistent with a ClustlW2[39] sequence alignment between PrgI and Bcl-xL that resulted in a low 14.3% sequence similarity, which falls below the twilight zone of sequence similarity.[40] Also, focused BLAST searches did not provide any new information. Searching microbial genomes using the Bcl-2 sequences or searching the human genome with T3SS sequences did not identify any sequence alignments with significant E-values. Thus, global sequence alignments did not readily result in identifying any relationship between T3SS and apoptosis proteins. This highlights the power of active site similarity searches to identify potentially new functional similarities in proteins.

Hidden Markov model (HMM) methods[41] provide an alternative and more robust approach to identify homology between distantly related proteins with low sequence similarity relative to traditional BLAST searches. The T-Coffee web server (http://www.tcoffee.org/) provides a consensus sequence alignment (M-Coffee) using multiple HMM protocols.[20] A reliable alignment of conserved residues (figure 3.2C) was obtained between the known T3SS structures of PrgI (PDB ID: 2JOW), BsaL (PDB ID: 2G0U) from *Burkholderia pseudomallei*, and MxiH (PDB ID: 2CA5) from *Shigella flexneri* with the human Bcl-xL (PDB ID: 1YSN) protein. The multiple-sequence alignment was obtained using EXPRESSO(3DCoffee) [31] that combines structural information with a HMM sequence alignment method. The reliability of the per residue alignment is color-coded using the color index,[38] where the majority of residues where in

the average to good range. The alignment of Bcl-xL with the three T3SS structures received a score of 53, where a score of 100 results from a perfect alignment. For comparison, the alignment of the three known human T3SS proteins resulted in a range of scores from 72 to 76. Conversely, scores that range from the 20 to the 30 indicate poor or insignificant alignments. Thus, PrgI aligns preferentially to the other T3SS proteins, but its alignment to the pore forming helices in Bcl-xL is significant and reliable. Importantly, the sequence alignment of PrgI with Bcl-xL encompasses the same residues involved in the ligand binding sites identified by CPASS and the structural similarity identified by DaliLite.

3.3.4 Identification of a second PrgI ligand binding site. The identification of a compound that binds similarly to both PrgI and Bcl-xL would further establish a functional relationship between these two proteins. BindingDB[32] was used to identify potential inhibitors of PrgI based on the CPASS predicted active site similarity with Bcl-xL. A total of 71 ligands were reported to bind Bcl-xL. A majority of the compounds were piperazine derivatives and were not readily available. Two compounds, chelerythrine and sanquinarine were identified as having affinity to Bcl-xL and were both available from commercial suppliers. Chelerythrine was selected over sanquinarine based on previous NMR screening and docking studies that suggested chelerythrine binds between α4, α5 and α6 of Bcl-xL.[33] This region of Bcl-xL was predicted to overlap with PrgI based on the pairwise Dali alignment (figure 3.2B). Conversely, sanquinarine bound the BH3 binding cleft of Bcl-xL and thus was not selected for this secondary binding analysis.[33]

A comparison between the free and chelerythrine bound PrgI 2D ^{1}H-^{15}N HSQC spectra (figure 3.3A) identified a chelerythrine binding site on PrgI (figure 3.3B). The PrgI residues that exhibited chemical shift changes upon binding chelerythrine include residues A14, K15 in helix 1 and residues Y57, N59, A60, V65, K66, V67, F68, and D72 in helix 2. The AutoDock/ADF docked structure of PrgI with chelerythrine suggests PrgI residues K15 and Y57 are the most important residues for chelerythrine binding based on a close contact with the ligand (figure 3.3B). Many of the residues that show significant CSPs for PrgI bound to chelerythrine overlap with the DDAB residues, however, the chelerythrine binding site is on the opposite face of PrgI (figure 3.4). This indicates there are two ligand binding sites on PrgI that is consistent with the two known protein-protein interaction sites for PrgI self-oligermization. The chelerythrine AutoDock docking energy decreased significantly compared to DDAB, -0.43 to -5.29 kcal/mol, respectively

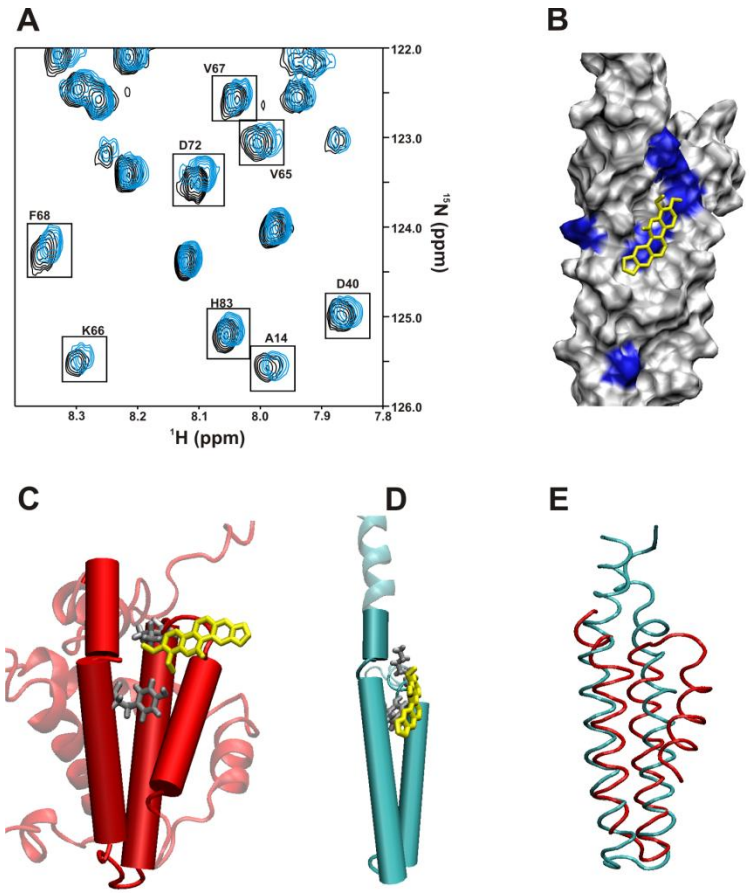


Figure 3.3. Verification the Bcl-xL inhibitor chelerythrine also binds PrgI. (A). Expanded overlay of the 2D ^{1}H-^{15}N HSQC spectra for free PrgI (black) and PrgI bound to chelerythrine (blue). CSPs greater than one standard deviation are boxed. (B) An AutoDock/ADF docked structure of PrgI complexed with chelerythrine based on the observed CSPs from (A). (C) The Bcl-xL region shown to bind chelerythrine is highlighted while the reaming protein structure is transparent. Chelerythrine is colored yellow and is drawn with licorice bonds. Side-chains for Y173 and V135 are shown as licorice bonds and colored grey. (D) A ribbon diagram of the AutoDock/ADF docked PrgI-chelerythrine co-structure. The PrgI-chelerythrine binding region that overlaps with Bcl-xL is highlighted. Chelerythrine is colored yellow and is drawn with licorice bonds. Side-chains for Y57 and K15 are shown as licorice bonds and colored grey. (E) An expanded view of the overlay of Bcl-xL (red) with PrgI (blue) illustrating the structural similarity of the chelerythrine binding sites.

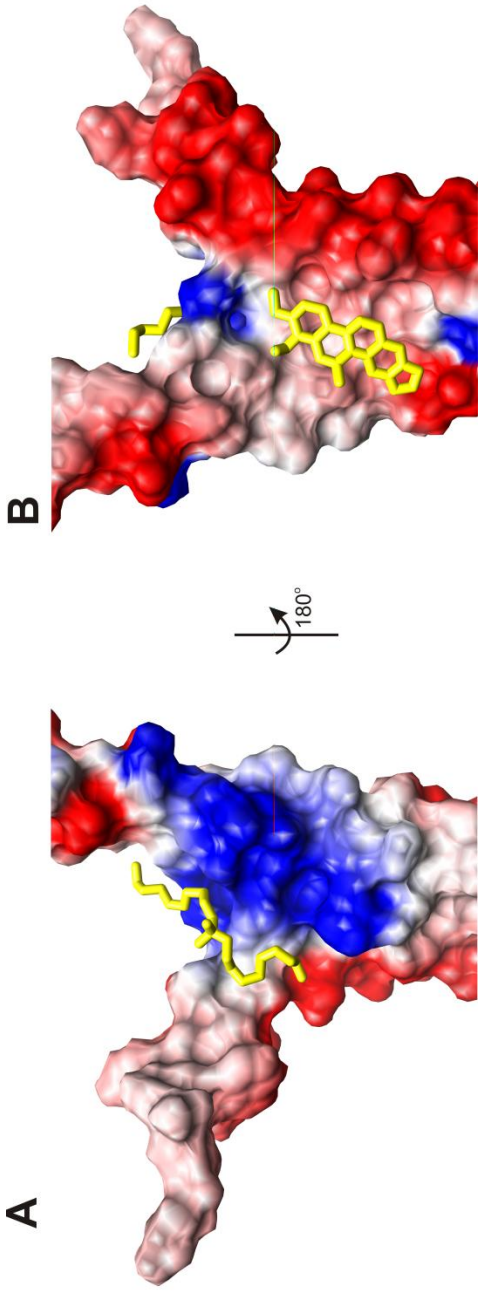

B
180°
A

Figure 3.4. The two PrgI ligand binding sites identified using FAST-NMR. The two PrgI ligand binding sites are highlighted on an electrostatic potential surface (blue positive charge, red negative charge) calculated with the DelPhiController implemented in Chimera [42]. The didecyldimethylammonium bromide binding site (A) is found in a region responsible for needle formation while the chelerythrine binding site (B) is found on the opposite face of PrgI.

The binding site of chelerythrine on PrgI is nearly identical to the binding site of chelerythrine to Bcl-xL (figure 3.3C and 3.3D). In Bcl-xL the chelerythrine binding site is described as being located in the BH groove of helix α4, α5 and α6, which is composed of residues F131, R132, V135, Y173 and H177 (figure 3.3C).[33] Pairwise structure analysis between PrgI and Bcl-xL shows that Y173 of Bcl-xL and Y57 on PrgI are overlapping residues and K15 from PrgI is proximal to V135 from Bcl-xL (figure 3.3E). The primary difference between the two proteins is the lack of α-helix 4 in PrgI, where helix 4 of Bcl-xL appears to act as a 'cap' encasing the ligand and effecting its relative binding orientation. Chelerythrine binds flat in the PrgI binding site, while the compound points into the corresponding Bcl-xL binding site partially overlaying helix α4. Again, both of these structures are docked models based on NMR CSPs and require a high-resolution x-ray or NMR structure to confirm the conformation of the chelerythrine binding site. It is paramount to note that this similarity in chelerythrine binding between the two proteins would have not been discovered if it was not for the identification of the initial conserved ligand binding site between PrgI and Bcl-xL using the FAST-NMR method in combination with the CPASS database.

3. 4 DISCUSSION

3.4.1 Ligand binding similarity of the Bcl-2 family of proteins with PrgI. A structural and functional similarity between PrgI, a type three secretion system protein, and Bcl-xL, a member of the Bcl-2 family of proteins involved in eukaryotic apoptosis, was identified from a FAST-NMR ligand affinity screen in combination with a bioinformatic analysis. This association is fundamentally based on the similarity in ligand

binding sites depicted in figure 3.2A, where the conserved helix-turn-helix motif simply provides secondary support of a PrgI and Bcl-2 functional link. While similar active sites provide a measure of functional similarity, inferring homology based solely on the observation of a similar helix-turn-helix motif is questionable. The helix-turn-helix is a common motif and without a global sequence similarity, an evolutionary lineage based solely on active site similarity cannot be readily established. However, identifying similar ligand binding sites between the two proteins does provide support the proteins share a common function and are expected to bind similar ligands.

The initial identification of the conserved DDAB ligand binding site between Bcl-xL and PrgI was used to predict, test and confirm that chelerythrine binds PrgI in a similar manner to Bcl-xL. This further supports the structural and functional similarity between PrgI and Bcl-xL, but also demonstrates the utility of active site similarity as a predictive tool for ligand binding. Chelerythrine was only tested for PrgI binding because of the proposed active site similarity with Bcl-xL. Thus, these studies have identified the first known ligands to bind PrgI (DDAB and chelerythrine). Both ligand binding sites are associated with the functionally important PrgI self-oligomerization sites. Therefore, compounds based on either the DDAB or chelerythrine scaffold may disrupt PrgI oligomerization. These compounds may serve as valuable chemical leads to develop novel antibiotics. Additionally, since the ligands bind in separate locations on the PrgI surface (figure 3.4), the compounds present two distinct approaches for developing drugs targeting PrgI. Unfortunately, because chelerythrine also binds Bcl-xL it is reasonable to expect that an antibiotic designed using chelerythrine as a scaffold may produce undesirable off-target side effects. This issue may be minimized or eliminated by simply

improving the PrgI binding affinity for chelerythrine derivatives. This illustrates another important feature of the FAST-NMR protocol; active site similarity is a useful tool to predict potential side effects due to off target inhibition in addition to predicting potential drug leads. While computational methods for predicting potential drug toxicity[43] are useful because of their speed, validation requires experimental methods such as the FAST-NMR approach.

3.4.2 Functional similarity of the Bcl-2 family of proteins with PrgI. The Bcl-2 family of proteins are essential for eukaryotic apoptosis; where Bcl-xL is responsible for repressing cell death activity.[15] The *in vivo* binding partners of Bcl-xL include the pro-apoptosis proteins Bax, Bak and Bid. It has been shown that expression levels of repressor (Bcl-xL) and pro-apoptosis proteins (Bax, Bak and Bid) are reciprocal in nature suggesting precise regulation of eukaryotic apoptosis.[44] A combination of mutational and structure work has shown the BH3 binding domain of Bcl-xL is critical for binding interactions with other Bcl-2 proteins and apoptosis regulation.[44]

The structure of Bcl-xL very closely resembles the structures of Bax, Bid, Bcl-2, and other members of the Bcl-2 family of proteins, which all resemble pore-forming domains of bacterial toxins.[45-47] Bcl-2, Bcl-xL, Bax and the truncated active form of Bid (tBid) have all been shown to form pores in liposomes, but a similar cellular function has only been observed for Bax.[44, 48, 49] In healthy cells, Bax is a monomer in the cytosol. Many different apoptotic signals result in the transfer of Bax to the outer mitochondrial membrane where an interaction with Bid and the lipid membrane induces Bax to form a supramolecular opening in the outer mitochondrial membrane.[50, 51] This pore structure causes the release of pro-apoptotic factors from the mitochondria into the cytoplasm to

induce cell death[52] and contains ~22 copies of Bax with a diameter of ~20 nm. The interaction of Bcl-xL with Bax prevents Bax induced cell death,[53] where drugs that disrupt Bcl-xL interacting with Bcl-2 proteins are a promising form of cancer therapy.[54] Bcl-xL has been described as a dominant-negative version of Bax.[55]

PrgI comprises the T3SS needle structure, which is formed by a PrgI homomultimer composed of ~ 120 copies of the protein.[10-12] This needle structure senses and punctures host membranes forming a pore to transfer proteins to induce cell death in a mechanism similar to eukaryotic apoptosis.[13-15] A general conservation in function between PrgI and the Bcl-2 protein family is thus maintained and readily apparent; both form membrane pores via a helix-turn-helix motif through oligomerization to release effectors to stimulate cell death. Additionally, PrgI requires PrgJ for oligomerization into the needle[11] while Bax requires Bid to induce pore formation.[51] Thus, a protein interaction with other members of the Bcl-2 family is required to either promote (Bid) or inhibit (Bcl-xL) Bax oligomerization. It is also interesting that PrgI was found to bind to a lipid analog and lipids have been found to play a role in Bax oligomerization.[51]

Importantly, the experimentally observed ligand binding sites for both PrgI and Bcl-xL are functionally equivalent and within the conserved helix-turn-helix motif. Both sites correspond to functionally critical protein-protein interaction sites required for oligomerization and pore formation. The DDAB binding site on PrgI overlaps with key residues involved in PrgI oligomerization and needle assembly. Similarly, ABT-737 is an inhibitor of apoptosis and functions by inhibiting Bcl-xL protein interactions.[56] Thus, the similarity in the ligand binding sites helps establish a functional link between the two proteins.

3.4.3 Structural similarity of the Bcl-2 family of proteins with PrgI. The Bax pore-forming domain is conserved in Bcl-xL, Bcl-2 and Bid[45, 57] and corresponds to the helix-turn-helix motif (helices α5 and α6) that was identified by CPASS to be similar to PrgI (figure 3.2A). Also, a comparison of the Bcl-xL and PrgI structure by Dalilite resulted in the alignment of the PrgI structure with this conserved Bcl-2 helix-turn-helix motif (figure 3.2B). Additionally, a multiple sequence alignment indicated a reliable similarity between T3SS needle-forming proteins and the Bcl-2 pore-forming region (figure 3.2C). Thus, the PrgI structure can be viewed as a minimalistic version of the Bcl-2 structure, and corresponds to the functionally essential and conserved core pore-forming domain.

Gene duplication along with insertion and/or deletions of sub-structures into variable genetic regions are known methods for the evolution of protein function.[58, 59] These processes may explain the evolution of the Bcl-2 family of proteins from a smaller PrgI-like ancestor. Since the PrgI structure overlaps with residues N136 to I182, this may suggest N- and C-terminal insertions generated a Bcl-2 protein from a PrgI-like ancestor. This is consistent with the hypothesis proposed by Aouacheria *et al.*,[60] where the ancestral toxic pore forming domain (helices α5 and α6) required developing a means to prevent inappropriate apoptosis and to regulate cell death.

Presumably, a main function of the N- and C-terminal inserts into a PrgI-like ancestor would be to stabilize the monomer form of Bax until an apoptotic signal occurs. In effect, the insertions would provide a stronger control over the pore formation process. This is consistent with what has been experimentally observed, both the N- terminus and C-terminus residues of Bax are essential to maintain the monomer form of Bax in the

cytosol.[47, 61, 62] Deletion of the first 20 amino acids from the N-terminus results in Bax being localized to the mitochondria.[61, 62] Similarly, the Bax structure indicates the C-terminal hydrophobic helix α9 is bent in a hydrophobic groove, but contains some critical solvent exposed polar residues that are necessary to maintain solubility.[47] In fact, a model for the translocation of Bax from the cytosol to the mitochondria requires a conformational change in Bax that opens up helix α9 and exposes the pore forming region composed of helices α5 and α6.[47, 63] Deletions of 21 residues from the C-terminus, which includes part of helix α6, prevents oligomerization.[64]

While Bax oligomerizes to form a circular pore structure containing ~22 copies, this oligomerization process does not extend to form layers like the PrgI needle structure. The conformational change in Bax results in the globular domain remaining in the cytosol and sterically prevents oligomerization perpendicular to the membrane.[65] Thus, the structural insert that maintains a monomer Bax in the cytosol also prevents an unnecessary linear extension of the Bax oligomer out of the mitochondria membrane. Conversely, regulating PrgI oligomerization is not necessary since the assembly of the T3SS system is not detrimental to the cell. Therefore, a minimal pore-forming structure is all that is necessary for the T3SS system. The length of the PrgI needle is controlled by the proper assembly of the inner rod (PrgJ) that requires the InvJ protein.[66] The deletion of InvJ results in long non-functional needles.

3.4.4 An evolutionary relationship between T3SS and eukaryotic apoptosis?

Based on the observed similarity in the structure and function between PrgI and the Bcl-2 protein family it is tempting to hypothesize the proteins share a common ancestor. The structural comparison of PrgI with the Bcl-2 family of proteins discussed above suggests

a possible evolutionary path. A common ancestral protein has been suggested for the Bcl-2 protein family, where pore formation using helices α5 and α6 is the ancestral proteins predicted primary function.[60] Similarly, T3SS are also predicted to evolve from a single gene[67] that is a simple but versatile export system.[68] Again, the helix-turn-helix is a common and ancient motif[69] demonstrating both its diverse utility and evolutionary stability. Thus, it is plausible that a simple and ancient PrgI-like protein could be an evolutionary precursor to both the Bcl-2 protein family and PrgI. It also appears unlikely that PrgI and the Bcl-2 protein family would evolve through a convergent process since the helix-turn-helix is such a simple and ancient motif[69] and essential to the function of both proteins. Evolving a readily available helix-turn-helix protein into either PrgI or the Bcl-2 protein family seems like a simpler path than the conversion of a uniquely distinct fold to incorporate a core helix-turn-helix motif. Also, the evolution of proteins from simple structural components has been previously proposed[70] and is consistent with other general evolutionary trends where complex systems evolve from simpler systems.[71]

By analogy, the sharing of a common ancestor by PrgI and the Bcl-2 family of proteins would imply an evolutionary relationship between the T3SS and eukaryotic apoptosis systems. T3SS is a prime example of a vestigial system and an important illustration of the stepwise evolution of the flagella machinery.[72, 73] Therefore, it is reasonable to expect that other systems will be identified that share an evolutionary relationship with T3SS. T3SS is also an ancient system and clearly predates the origin of the mitochondria from prokaryote endosymbiosis.[74, 75] α-proteobacteria,[74] which are close relatives of the mitochondria, are known to contain T3SS.[68, 76, 77] Could an obsolete T3SS system contribute valuable components to the eukaryotic apoptosis system after

endosymbiosis? An evolutionary link has already been observed between a mitochondrial and T3SS protein.[78, 79] Furthermore, a detailed analysis of the origin of apoptotic proteins suggests a pivotal role for bacterial proteins in the evolution of eukaryotic apoptosis.[80]

3.5 REFERENCES

1. Mercier, K. A.; Baran, M.; Ramanathan, V.; Revesz, P.; Xiao, R.; Montelione, G. T.; Powers, R., FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* **2006,** 128, (47), 15292-9.
2. Powers, R.; Copeland, J. C.; Germer, K.; Mercier, K. A.; Ramanathan, V.; Revesz, P., Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **2006,** 65, (1), 124-35.
3. Gerlt, J. A.; Babbitt, P. C., Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem* **2001,** 70, 209-46.
4. Mirny, L. A.; Shakhnovich, E. I., Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol* **1999,** 291, (1), 177-96.
5. Park, K.; Kim, D., Binding similarity network of ligand. *Proteins* **2008,** 71, (2), 960-71.
6. Campbell, S. J.; Gold, N. D.; Jackson, R. M.; Westhead, D. R., Ligand binding: functional site location, similarity and docking. *Curr Opin Struct Biol* **2003,** 13, (3), 389-95.

7. Powers, R.; Mercier, K. A.; Copeland, J. C., The application of FAST-NMR for the identification of novel drug discovery targets. *Drug Discov Today* **2008,** 13, (3-4), 172-9.

8. Mercier, K. A.; Baran, M.; Ramanathan, V.; Revesz, P.; Xiao, R.; Montelione, G. T.; Powers, R., FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* **2006,** 128, (47), 15292-9.

9. Powers, R., Functional genomics and NMR spectroscopy. *Comb Chem High Throughput Screen* **2007,** 10, (8), 676-97.

10. Cornelis, G. R.; Van Gijsegem, F., Assembly and function of type III secretory systems. *Annu Rev Microbiol* **2000,** 54, 735-74.

11. Kimbrough, T. G.; Miller, S. I., Assembly of the type III secretion needle complex of Salmonella typhimurium. *Microbes Infect* **2002,** 4, (1), 75-82.

12. Macnab, R. M., How bacteria assemble flagella. *Annu Rev Microbiol* **2003,** 57, 77-100.

13. Galan, J. E., Salmonella interactions with host cells: type III secretion at work. *Annu Rev Cell Dev Biol* **2001,** 17, 53-86.

14. Grassme, H.; Jendrossek, V.; Gulbins, E., Molecular mechanisms of bacteria induced apoptosis. *Apoptosis* **2001,** 6, (6), 441-5.

15. Yan, N.; Shi, Y., Mechanisms of apoptosis through structural biology. *Annu Rev Cell Dev Biol* **2005,** 21, 35-56.

16. Wang, Y.; Ouellette, A. N.; Egan, C. W.; Rathinavelan, T.; Im, W.; De Guzman, R. N., Differences in the electrostatic surfaces of the type III secretion needle proteins PrgI, BsaL, and MxiH. *J Mol Biol* **2007,** 371, (5), 1304-14.

17. Deane, J. E.; Roversi, P.; Cordes, F. S.; Johnson, S.; Kenjale, R.; Daniell, S.; Booy, F.; Picking, W. D.; Picking, W. L.; Blocker, A. J.; Lea, S. M., Molecular model of a type III secretion system needle: Implications for host-cell sensing. *Proc Natl Acad Sci U S A* **2006,** 103, (33), 12529-33.

18. Torruellas, J.; Jackson, M. W.; Pennock, J. W.; Plano, G. V., The Yersinia pestis type III secretion needle plays a role in the regulation of Yop secretion. *Mol Microbiol* **2005,** 57, (6), 1719-33.

19. Dietmann, S.; Park, J.; Notredame, C.; Heger, A.; Lappe, M.; Holm, L., A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res* **2001,** 29, (1), 55-7.

20. Poirot, O.; O'Toole, E.; Notredame, C., Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Res* **2003,** 31, (13), 3503-6.

21. Chan, S. L.; Lee, M. C.; Tan, K. O.; Yang, L. K.; Lee, A. S.; Flotow, H.; Fu, N. Y.; Butler, M. S.; Soejarto, D. D.; Buss, A. D.; Yu, V. C., Identification of chelerythrine as an inhibitor of BclXL function. *J Biol Chem* **2003,** 278, (23), 20453-6.

22. 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* **2006,** 9, (7), 515-34.

23. 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* **2008,** 10, (6), 948-58.

24. Stark, J.; Powers, R., Rapid protein-ligand costructures using chemical shift perturbations. *J Am Chem Soc* **2008,** 130, (2), 535-45.

25. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **1998,** 19, (14), 1639-1662.

26. Henikoff, S.; Henikoff, J. G., Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **1992,** 89, (22), 10915-9.

27. Henikoff, S.; Henikoff, J. G., Performance evaluation of amino acid substitution matrices. *Proteins* **1993,** 17, (1), 49-61.

28. Oltersdorf, T.; Elmore, S. W.; Shoemaker, A. R.; Armstrong, R. C.; Augeri, D. J.; Belli, B. A.; Bruncko, M.; Deckwerth, T. L.; Dinges, J.; Hajduk, P. J.; Joseph, M. K.; Kitada, S.; Korsmeyer, S. J.; Kunzer, A. R.; Letai, A.; Li, C.; Mitten, M. J.; Nettesheim, D. G.; Ng, S.; Nimmer, P. M.; O'Connor, J. M.; Oleksijew, A.; Petros, A. M.; Reed, J. C.; Shen, W.; Tahir, S. K.; Thompson, C. B.; Tomaselli, K. J.; Wang, B.; Wendt, M. D.; Zhang, H.; Fesik, S. W.; Rosenberg, S. H., An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature* **2005,** 435, (7042), 677-81.

29. Holm, L.; Park, J., DaliLite workbench for protein structure comparison. *Bioinformatics* **2000,** 16, (6), 566-7.

30. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997,** 25, (17), 3389-402.

31. Armougom, F.; Moretti, S.; Poirot, O.; Audic, S.; Dumas, P.; Schaeli, B.; Keduas, V.; Notredame, C., Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. **2006,** 34, (Web Server), W604-W608.

32. 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* **2007,** 35, (Database issue), D198-201.

33. Zhang, Y. H.; Bhunia, A.; Wan, K. F.; Lee, M. C.; Chan, S. L.; Yu, V. C.; Mok, Y. K., Chelerythrine and sanguinarine dock at distinct sites on BclXL that are not the classic BH3 binding cleft. *J Mol Biol* **2006,** 364, (3), 536-49.

34. Muller, S.; Feldman, M. F.; Cornelis, G. R., The Type III secretion system of Gram-negative bacteria: a potential therapeutic target? *Expert Opin Ther Targets* **2001,** 5, (3), 327-339.

35. 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* **2009,** 12, (3), 285-95.

36. Kubori, T.; Sukhan, A.; Aizawa, S. I.; Galan, J. E., Molecular characterization and assembly of the needle complex of the Salmonella typhimurium type III protein secretion system. *Proc Natl Acad Sci U S A* **2000,** 97, (18), 10225-30.

37. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *J Mol Graph* **1996,** 14, (1), 33-8, 27-8.

38. Notredame, C.; Abergel, C., Using multiple alignment methods to assess the quality of genomic data analysis. In *Bioinformatics and genomes: current*

perspectives, Andrade, M. A., Ed. Horizon Scientific Press: Wymondham (United Kingdom), 2003.

39. Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G., Clustal W and Clustal X version 2.0. *Bioinformatics* **2007,** 23, (21), 2947-8.

40. Rost, B., Twilight zone of protein sequence alignments. *Protein Eng* **1999,** 12, (2), 85-94.

41. Bystroff, C.; Krogh, A., Hidden Markov Models for prediction of protein features. *Methods Mol Biol* **2008,** 413, 173-98.

42. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **2004,** 25, (13), 1605-12.

43. Xie, L.; Li, J.; Bourne, P. E., Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput Biol* **2009,** 5, (5), e1000387.

44. Chao, D. T.; Korsmeyer, S. J., BCL-2 family: regulators of cell death. *Annu Rev Immunol* **1998,** 16, 395-419.

45. Chou, J. J.; Li, H.; Salvesen, G. S.; Yuan, J.; Wagner, G., Solution structure of BID, an intracellular amplifier of apoptotic signaling. *Cell* **1999,** 96, (5), 615-24.

46. McDonnell, J. M.; Fushman, D.; Milliman, C. L.; Korsmeyer, S. J.; Cowburn, D., Solution structure of the proapoptotic molecule BID: a structural basis for apoptotic agonists and antagonists. *Cell* **1999,** 96, (5), 625-34.

47. Suzuki, M.; Youle, R. J.; Tjandra, N., Structure of Bax: coregulation of dimer formation and intracellular localization. *Cell* **2000,** 103, (4), 645-54.

48. Vander Heiden, M. G.; Thompson, C. B., Bcl-2 proteins: regulators of apoptosis or of mitochondrial homeostasis? *Nat Cell Biol* **1999,** 1, (8), E209-16.

49. Yan, L.; Miao, Q.; Sun, Y.; Yang, F., tBid forms a pore in the liposome membrane. *FEBS Lett* **2003,** 555, (3), 545-50.

50. Annis, M. G.; Soucie, E. L.; Dlugosz, P. J.; Cruz-Aguado, J. A.; Penn, L. Z.; Leber, B.; Andrews, D. W., Bax forms multispanning monomers that oligomerize to permeabilize membranes during apoptosis. *Embo J* **2005,** 24, (12), 2096-103.

51. Kuwana, T.; Mackey, M. R.; Perkins, G.; Ellisman, M. H.; Latterich, M.; Schneiter, R.; Green, D. R.; Newmeyer, D. D., Bid, Bax, and lipids cooperate to form supramolecular openings in the outer mitochondrial membrane. *Cell* **2002,** 111, (3), 331-42.

52. Breckenridge, D. G.; Xue, D., Regulation of mitochondrial membrane permeabilization by BCL-2 family proteins and caspases. *Curr Opin Cell Biol* **2004,** 16, (6), 647-52.

53. Fletcher, J. I.; Meusburger, S.; Hawkins, C. J.; Riglar, D. T.; Lee, E. F.; Fairlie, W. D.; Huang, D. C. S.; Adams, J. M., Apoptosis is triggered when prosurvival Bcl-2 proteins cannot restrain Bax. *Proc Natl Acad Sci USA* **2008,** 105, (47), 18081-18087, S18081/1-S18081/7.

54. Kang, M. H.; Reynolds, C. P., Bcl-2 Inhibitors: Targeting Mitochondrial Apoptotic Pathways in Cancer Therapy. *Clin. Cancer Res.* **2009,** 15, (4), 1126-1132.

55. Billen, L. P.; Kokoski, C. L.; Lovell, J. F.; Leber, B.; Andrews, D. W., Bcl-XL inhibits membrane permeabilization by competing with Bax. *PLoS Biol.* **2008,** 6, (6), 1268-1280.

56. Wendt, M. D., Discovery of ABT-263, a Bcl-family protein inhibitor: observations on targeting a large protein-protein interaction. *Expert Opin. Drug Discovery* **2008,** 3, (9), 1123-1143.

57. Muchmore, S. W.; Sattler, M.; Liang, H.; Meadows, R. P.; Harlan, J. E.; Yoon, H. S.; Nettesheim, D.; Chang, B. S.; Thompson, C. B.; Wong, S. L.; Ng, S. L.; Fesik, S. W., X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death. *Nature* **1996,** 381, (6580), 335-41.

58. Jiang, H.; Blouin, C., Insertions and the emergence of novel protein structure: a structure-based phylogenetic study of insertions. *BMC Bioinformatics* **2007,** 8, 444.

59. Morett, E.; Bork, P., Evolution of new protein function: recombinational enhancer Fis originated by horizontal gene transfer from the transcriptional regulator NtrC. *FEBS Lett* **1998,** 433, (1-2), 108-12.

60. Aouacheria, A.; Brunet, F.; Gouy, M., Phylogenomics of life-or-death switches in multicellular animals: Bcl-2, BH3-Only, and BNip families of apoptotic regulators. *Mol Biol Evol* **2005,** 22, (12), 2395-416.

61. Cartron, P. F.; Oliver, L.; Martin, S.; Moreau, C.; LeCabellec, M. T.; Jezequel, P.; Meflah, K.; Vallette, F. M., The expression of a new variant of the pro-apoptotic molecule Bax, Baxpsi, is correlated with an increased survival of glioblastoma multiforme patients. *Hum Mol Genet* **2002,** 11, (6), 675-87.

62. Goping, I. S.; Gross, A.; Lavoie, J. N.; Nguyen, M.; Jemmerson, R.; Roth, K.; Korsmeyer, S. J.; Shore, G. C., Regulated targeting of BAX to mitochondria. *J Cell Biol* **1998,** 143, (1), 207-15.

63. Desagher, S.; Osen-Sand, A.; Nichols, A.; Eskes, R.; Montessuit, S.; Lauper, S.; Maundrell, K.; Antonsson, B.; Martinou, J. C., Bid-induced conformational change of Bax is responsible for mitochondrial cytochrome c release during apoptosis. *J Cell Biol* **1999,** 144, (5), 891-901.

64. Er, E.; Lalier, L.; Cartron, P. F.; Oliver, L.; Vallette, F. M., Control of Bax homodimerization by its carboxyl terminus. *J Biol Chem* **2007,** 282, (34), 24938-47.

65. Lalier, L.; Cartron, P. F.; Juin, P.; Nedelkina, S.; Manon, S.; Bechinger, B.; Vallette, F. M., Bax activation and mitochondrial insertion during apoptosis. *Apoptosis* **2007,** 12, (5), 887-96.

66. Marlovits, T. C.; Kubori, T.; Lara-Tejero, M.; Thomas, D.; Unger, V. M.; Galan, J. E., Assembly of the inner rod determines needle length in the type III secretion injectisome. *Nature* **2006,** 441, (7093), 637-40.

67. Liu, R.; Ochman, H., Stepwise formation of the bacterial flagellar system. *Proc Natl Acad Sci U S A* **2007,** 104, (17), 7116-21.

68. Gophna, U.; Ron, E. Z.; Graur, D., Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events. *Gene* **2003,** 312, 151-163.

69. Rosinski, J. A.; Atchley, W. R., Molecular evolution of helix-turn-helix proteins. *J. Mol. Evol.* **1999,** 49, (3), 301-309.

70. Kannan, N.; Neuwald, A. F., Did Protein Kinase Regulatory Mechanisms Evolve Through Elaboration of a Simple Structural Component? *J. Mol. Biol.* **2005,** 351, (5), 956-972.

71. Shapiro, R., Small molecule interactions were central to the origin of life. *Q Rev Biol* **2006,** 81, (2), 105-25.

72. McCann, H. C.; Guttman, D. S., Evolution of the type III secretion system and its effectors in plant-microbe interactions. *New Phytol* **2008,** 177, (1), 33-47.

73. Gophna, U.; Ron, E. Z.; Graur, D., Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events. *Gene* **2003,** 312, 151-63.

74. Gray, M. W.; Burger, G.; Lang, B. F., The origin and early evolution of mitochondria. *GenomeBiology [online computer file]* **2001,** 2, (6), No pp given.

75. Koonin, E. V.; Aravind, L., Origin and evolution of eukaryotic apoptosis: the bacterial connection. *Cell death and differentiation* **2002,** 9, (4), 394-404.

76. Gauthier, A.; Thomas, N. A.; Finlay, B. B., Bacterial Injection Machines. *Journal of Biological Chemistry* **2003,** 278, (28), 25273-25276.

77. Batut, J.; Andersson, S. G. E.; O'Callaghan, D., The evolution of chronic infection strategies in the a-proteobacteria. *Nature Reviews Microbiology* **2004,** 2, (12), 933-945.

78. Pallen Mark, J.; Bailey Christopher, M.; Beatson Scott, A., Evolutionary links between FliH/YscL-like proteins from bacterial type III secretion systems and second-stalk components of the FoF1 and vacuolar ATPases. *Protein Sci* **2006,** 15, (4), 935-41.

79. Mulkidjanian, A. Y.; Makarova, K. S.; Galperin, M. Y.; Koonin, E. V., Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nature Reviews Microbiology* **2007,** 5, (11), 892-899.

80. Koonin, E. V.; Aravind, L., Origin and evolution of eukaryotic apoptosis: the bacterial connection. *Cell Death Differ* **2002,** 9, (4), 394-404.

CHAPTER 4:

OPTIMIZATION AND VALIDATION OF THE FAST-NMR METHOD

4.1 INTRODUCTION

The predicted functional similarity between PrgI and the Bcl-2 family of proteins described in chapter 3 illustrates the enhanced benefit of combining experimental data with bioinformatics. The Functional Annotation Screening Technology by NMR (FAST-NMR) is an initial step in achieving high-throughput functional analysis of proteins, independently of global sequence or structure homology transfer. FAST-NMR uses a tiered approach to NMR screening to identify protein active sites.[1, 2] First, each protein is screened with a library of approximately 437 compounds distributed across 113 mixtures.[3, 4] Binding is detected using the 1D ^{1}H NMR line broadening methods discussed in chapters 2 and 3.[2] The compounds that show the tightest binding are passed to the second tier screening step (2D ^{1}H-^{15}N HSQC) to identify the protein active site. The experimentally identified active site is compared to a database of known protein active sites using the CPASS database and software.[5] Finally, protein function is inferred by identifying similar active sites in the CPASS database.

The tiered approach to screening, along with screening in mixtures, reduces the total amount of time and sample requirements needed to identify a protein active site.[1, 3, 4] However, the reduction in data collection time is relative to screening a protein with individual compounds. A significant bottleneck in the process remains the large data collection time for screening all 113 mixtures and relatively large sample requirements needed in the 2D ^{1}H-^{15}N screening step.

NMR is a relatively insensitive technique that uses signal averaging to increase signal-to-noise. Data collection time is directly proportional to the total number of scans needed and the recycle delay between each scan. To maximize signal-to-noise and suppress residual solvent signal; the initial FAST-NMR 1D ^{1}H screening step required 64-128 scans with a recycle delay of 1.0-2.0 s (chapter 3). The total time to collect an NMR spectrum for a mixture and move to the next sample is approximately 10-14 min (2-6 min data collection, 8 min sample change and set up). This correlates to approximately 19-26 hrs of total 1D ^{1}H experiment time for each protein screen.

The tiered approach to NMR screening saves experimental time and protein sample by prioritizing which ligands from the 1D screen get passed to the 2D ^{1}H-^{15}N HSQC conformation screens.[1] However, the average number of hits for the 4 proteins (PA1324, SAV1430, PrgI, and *S. aureus* primase CTD) screened with the initial method was 16.75 ± 10.75 ligands with a range between 5-30 ligands. Using the tiered approach method still requires large sample concentrations and experimental time. For the 30 ligands identified that bound SAV1430 the total amount of ^{15}N labeled protein was approximately 30 mgs and nearly 80 hrs of data collection (2.5 hrs/HSQC with 8min/sample change).[2] For PrgI, the protein with the lowest number of hits, the total time for the 2D ^{1}H-^{15}N HSQC screen was approximately 11.5 hrs (1.5 hrs/HSQC with 8min/sample changing).[6] Obviously, spending between 30-100 hrs of total experiment time and the large protein requirement significantly limits the throughput of the FAST-NMR method.

In this chapter I will discuss the optimization of the FAST-NMR method by implementing two new pulse sequences and making significant updates to the automated

NMR data collection. The improvements made to the screening method decreased the total experimental time for the 1D ^{1}H NMR screening step by approximately 8 hrs per protein. Additionally, the improvements made to the 2D ^{1}H-^{15}N HSQC screening step provides a greater flexibility in data collection by reducing the total amount of sample needed or reducing to over experimental time.

I evaluated the improvements to FAST-NMR screening using *Staphylococcus aureus* nuclease, a well-established model protein for NMR screening with a number of previously solved free and ligand-bound NMR structures.[7-9] I demonstrate the improved FAST-NMR screening method can correctly identify the previously reported nuclease ligand binding site in a high-throughput manner. Additionally, the binding site found by FAST-NMR was used by CPASS to correctly identified the reference nuclease structure from the CPASS ligand binding site database.[5]

4.2 EXPERIMENTAL

4.2.1 Materials. The bromocresol green (ACS reagent grade, 95% pure) was purchased from Sigma-Aldrich (Milwaukee, WI). The dimethyl sulfoxide-d_6 (99.9% D), 2,2-Bis(hydroxymethyl)-2,2′,2″-nitrilotriethanol-d_{19} (98% D), naproxen (98% pure) and deuterium oxide (99.9% D) was obtained from Aldrich (Milwaukee, WI). The 3-(trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt (98% D) was purchased from Cambridge Isotope (Andover, MA). The potassium phosphate dibasic salt (anhydrous, 99.1% pure) and monobasic salt (crystal, 99.8% pure) were purchased from Mallinckrodt (Phillipsburg, NJ). *E. coli* cells containing the pET28a(+) plasmid with nuclease sequence and kanamycin resistance gene was obtained from Dr. Greg Somerville's lab

(see appendix A for nuclease sequence). The plasmid isolation kit, Quickclean 5M miniprep, was purchased from GenScript (Piscataway, NJ). All competent cell lines were purchased from Stratagene (La Jolla, CA). All unlabeled growth media components including tryptone, yeast extract, agar, sodium chloride, and IPTG were purchased from Aldrich. Cobalt affinity resin was purchased from ClonTech (Mountain View, CA).

4.2.2 Apparatus. Two different pulse sequences with improved solvent suppression were implemented to decrease sample requirements and data collection time in the FAST-NMR method. All NMR spectra were collected on a Bruker 500 MHz Avance spectrometer (Billerica, MA) equipped with a triple-resonance, Z-axis gradient Cryoprobe. All samples were tuned, matched and shimmed to optimize the observed signal. All sample volumes were at a constant 600 μL volume in a 178 mm long x 5 mm OD NMR tube rated for 500-700 MHz (NE-UL5-7 New Era Enterprise, Vineland, NJ) to minimize shimming between samples. All samples were collected at 298K. 1D ^{1}H data was processed using ACD labs v. 12.0 while 2D ^{1}H-^{15}N HSQC data was processed using NMRPIPE[10] and visualized using PIPP[11] and CCCPNMR.[12]

4.2.3 Optimization of automated data collection. As described in chapter 2 and 3, the FAST-NMR method utilizes the Bruker BACS-120 sample changer and IconNMR software for automated data collection. To increase throughput, the automatic receiver gain adjustment was turned off and each sample was collected at a constant receiver gain. Additionally, an automatic shimming routine using a single iteration of the Bruker gradient shimming to optimize Z1 and Z2 field axes was developed to minimize the time needed to shim a sample while providing adequate line shape.

4.2.4 Implementation of the 1D ^{1}H excitation sculpting pulse sequence. The 1D ^{1}H excitation sculpting pulse sequence[13] came as a standard and compiled pulse sequence in the Bruker pulse sequence library (*zgesgp*). All ^{1}H 90^o pulse lengths were optimized by finding a 360^o spectral null at a constant power level of -4.3 dB. The optimized ^{1}H 90^o pulse length was used to calculate all ^{1}H pulses used in the sequence. A total of 64 real transients and 8 dummy transients at 8k data points were collected with a recycle delay of 1.0 s. Total experiment time was approximately 1.25 min.

The excitation sculpting pulse sequence was compared to the presaturation pulse sequence to examine differences in spectral quality, signal to noise and ability to measure a single point binding constant (chapter 2). The presaturation sequence was executed in the same manner as the excitation sculpting sequence with a recycle delay of 2.0 s to maximize solvent suppression. Total experiment time was approximately 2.5 min.

A free ligand solution was prepared in a 5 mL stock containing 50 μM naproxen, 5% (v/v) deuterated dimethyl sulfoxide-d6 (DMSO-d6), 11.1 μM 3-(trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt (TSP) and 50 mM potassium phosphate buffer pH 7.0 (uncorrected) in 99.98% deuterium oxide. Five replicate samples were made from the 5 mL stock solution and transferred to individual NMR tubes. These 5 samples were used for calculating the average free ligand intensities (I_F) and average free ligand linewidths (ν_F). Data for each sample was collected using the excitation sculpting sequence and presaturation sequence.

A bound ligand solution was prepared in a 5 mL stock solution containing 50 μM naproxen, 5 μM human serum albumin (HSA), 5% (v/v) deuterated dimethyl sulfoxide-d6 (DMSO-d6), 11.1 μM 3-(trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt (TSP)

and 50 mM potassium phosphate buffer pH 7.0 (uncorrected) in 99.98% deuterium oxide. Five replicate samples were made from this stock solution and **t**ransferred to individual NMR tubes. These 5 samples were used to calculate the average bound ligand intensities (I_B). Data for each sample was collected using the excitation sculpting sequence and presaturation sequence.

4.2.5 Implementation of the 2D ^{1}H -^{15}N HSQC with WATERGATE and water flip-back for solvent suppression. The 2D ^{1}H-^{15}N HSQC utilizing WATERGATE and water flip-back pulses came as a standard and compiled pulse sequence in the Bruker pulse sequence library (*hsqcfpf3gpphwg*). All ^{1}H 90^{o} pulse lengths were optimized in the same manner as describe above in section 4.2.2.1. Additionally, all ^{13}C and ^{15}N pulse powers were optimized using the Bruker *dec90* and *dec90F3* pulse sequences, respectively. A 100 μM ^{15}N labeled *S. aureus* nuclease sample and 5 μM ^{15}N labeled *S. aureus* nuclease sample were used to test the pulse sequence. Both samples were prepared in a 95% H_2O/5%D_2O buffered solution of 50 mM KPO_4 (pH 7.0) with 300 mM NaCl.

4.2.6 Expression of unlabeled and ^{15}N labeled *S. aureus* nuclease. The pET28a(+) plasmid with the recombinant nuclease sequence and kanamycin resistance gene was extracted from the stock *E. coli* cells using the method outlined in the Genscript Quickclean 5M miniprep kit (appendix 4B). The plasmid was transformed into Bl21-DE3-pLySs and Bl21-DE3-codon+ competent *E. coli* cells using the method described in the Stratagene manual. Transformed cells were grown at 37 ^{o}C for 12 hrs on LB agar plates containing 50 mg/L kanamycin. Only the Bl21-DE3-codon+ cells produced any colonies.

Three different isolated colonies were selected from the agar plate, individually inoculated into three different centrifuge tubes containing 25 mL of LB broth and left to grow for 12 hrs in an incubated shaker at 37 ^{o}C. A 1 mL sample from each 12 hr growth was inoculated into three different 25 mL cultures of LB broth and left to grow until an O.D of 0.6 at 600nm was reached. Protein expression was induced by the addition of 500 mM IPTG. Induced cells were grown for an additional three hours and expression was checked by running a 15% PAGE gel of the whole cells (figure 4.5A). The colony that gave the best expression was used to make a 25 mL 40% glycerol stock suspension stored in 1 mL aliquots for future expressions as previously described.[14]

A 1 mL glycerol stock sample was thawed and used to make a LB agar streak plate. Unlabeled *S. aureus* nuclease was expressed by isolation of a single colony from the streak, growing strain BL21(DE3)codon+/pET28a(+) in 25 mL LB broth containing kanamycin at 50 mg/L at 37 ^{o}C for 12 hrs. A 5 mL sample of the 12 hr growth was inoculated into 1 L cultures (2 L total) of LB broth containing kanamycin at 50 mg/L at 37 ^{o}C until an absorbance of 0.67 at 600 nm was reached (~4 hrs). Protein expression was induced by the addition of 500 mM IPTG to each culture and shaken for an additional 3.5 hrs at 37 ^{o}C (appendix 4C). Cells were harvested by centrifugation at 10,000 G and stored frozen at -80 ^{o}C.

A 1 mL glycerol stock sample was thawed and used to make a LB agar streak plate. ^{15}N-labeled *S. aureus* nuclease was expressed by isolation of a single colony from the streak, growing strain BL21(DE3)codon+/pET28a(+) in 25 mL M9 minimal media broth (2 mL 1M MgSO4, 100 uL 1M $CaCl_2$, 10 mL 100x Basal Medium Eagle Vitamin Solution (Gibco), 1.0 g ^{15}N-NH_4Cl, 4 g d-glucose, 200 mL of 5xM9 salts) containing

kanamycin at 50 mg/L at 37 ^{o}C for 12 hrs. A 5 mL sample of the 12 hr growth was inoculated into 1 L cultures (2 L total) of M9 minimal media broth until an absorbance of 0.79 at 600 nm was reached (6.75 hr) (appendix 4C). Protein expression was induced by the addition of 500 mM IPTG to each culture and shaken for an additional 3 hrs at 37 ^{o}C. Cells were harvested by centrifugation at 10,000 G and stored frozen at -80 ^{o}C.

4.2.7 Purification of unlabeled and ^{15}N-labeled *S. aureus* nuclease. Both unlabeled and ^{15}N-labeled *S. aureus* nuclease expressions were treated the same for purification. Cells were thawed, re-suspend in equilibration/wash buffer (50 mM sodium phosphate pH 7.0 and 300 mM NaCl) in 25 mL aliquots and sonicated on ice 3 times at 45 s intervals. The lysate was centrifuged for 30 min at 10,000 g and incubated with 10 mL Talon Cobalt affinity resin for 30 min at 4 ^{o}C. The protein bound resin was washed by passing 4 column bed volumes of equilibration/wash buffer through the resin bed. Nuclease was eluted with 5 column bed volumes of elution buffer (50 mM sodium phosphate pH 7.0, 300 mM NaCl and 150 mM imidazole) and stored at 4 ^{o}C.

4.2.8 FAST-NMR screening of *S. aureus* nuclease. A FAST-NMR screen, AutoDock ligand bound co-structures and CPASS analysis of the *S. aureus* nuclease ligand binding site was completed using the methods described in chapter 3 with additional modifications. Specifically, the FAST-NMR ligand affinity screen of nuclease utilized the pulse sequences and experimental parameters described above. The increase in solvent suppression efficiency required using less protein sample per screen due to aliphatic protein resonance overlap with the ligand signals. The total protein concentration was reduced from 25 μM in the PrgI screen to 5 μM in the nuclease screen.

A 10 mL volume of stock nuclease was buffer exchanged with equilibration/wash buffer to remove residual imidazole. Each buffer exchange involved centrifuging the nuclease sample at 5,000 G for 5 min to a volume of ~1 mL using a 15 mL Amicon Ultra-15, 10,000 MW cutoff centrifugal filter unit (Millipore, Billerica, MA). After each centrifugation, 10 mL of equilibration/wash buffer was added to the Amicon Ultra-15 and the process was repeated 5 times. After the final buffer exchange the nuclease sample was concentrated to 5 mL. The final concentration of the sample was approximately 1 mM nuclease in a buffered solution of 50 mM sodium phosphate pH 7.0 and 300 mM NaCl.

1D ^{1}H NMR ligand affinity screening was completed in a similar manner described in chapter 3. Briefly, 5 μM nuclease was added to each ligand mixture (100 μM/ligand) in a 99.99% D_2O buffered solution of 20 mM d_{19}-bis-Tris at pH 7.0 with 5% DMSO-d_6 to maintain ligand solubility and 11.1 μM 3-(trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt as a chemical shift reference. A total of 113 mixture samples were prepared. 1D ^{1}H NMR spectra for each sample was collected using the excitation sculpting sequence with 64 real scans, 8 dummy scans with 8 k data points, a sweep width of 12.0 ppm and a recycle delay of 1.0 s. Data was Fourier transformed, auto-phase and baseline corrected. The 1D ^{1}H NMR spectra were compared to free ligand mixture reference spectra and visually analyzed to identify binding ligands. A binding event was identified by the decrease in ligand intensity of the nuclease-mixture relative to the free ligand mixture. Total data collection time including sample changing was approximately 6 min/spectrum

All 2D ^{1}H-^{15}N HSQC affinity screens were completed by the addition of 500 μM ligand to a 100 μM ^{15}N labeled nuclease sample in a 95% H_2O/5%D_2O buffered solution of 20 mM bis-Tris at pH 7.0 with 5% DMSO-d_6 to maintain ligand solubility. 2D ^{1}H-^{15}N HSQC spectra were collected using the WATERGATE/flip-back pulse sequence described in section 4.2.5 with 8 real scans, 128 dummy scans, 1 k data points in the ^{1}H dimension and 128 data points in the ^{15}N dimension. The sweep width of the spectrum was 12.0 ppm in the ^{1}H dimension and 30.0 ppm in the ^{15}N dimension. A recycle delay for the pulse sequence was set to 1.0 s. Total data collection time was approximately 20 min/spectrum. Spectra were processed using the same parameters as described in section 4.2.4.

4.3 RESULTS AND DISCUSSION

As described in chapters 2 and 3, NMR affinity screening generally involves collecting an NMR spectrum at a low analyte concentration in an aqueous buffer. This poses a significant challenge when developing a high-throughput NMR screening methods. The relative concentration of residual protons in 99.99% D_2O is 1100 mM compared to 20-100 μM for the free ligand. The ~10-50 fold intensity difference between solvent and analyte peaks decreases the limit of detection (figure 4.1A). A number of solvent suppression techniques exist to selectively irradiate the solvent peak and increase the detection limit.

The initial pulse program used for the 1D ^{1}H NMR affinity screening in the FAST-NMR method was a presaturation sequence with a composite pulse train prior to the 90^{o} pulse (figure 4.1B). The presaturation pulse is a low power pulse implemented

during the recycle delay at the frequency of the solvent signal. As the presaturation pulse length is increased there is an increase in solvent suppression. To maximize signal, the recycle delay is set at 1-5 times larger than the T_1 relaxation rate for the analyte. A recycle delay of 2.0 s (see chapter 2 and 3) is used for the presaturation method.

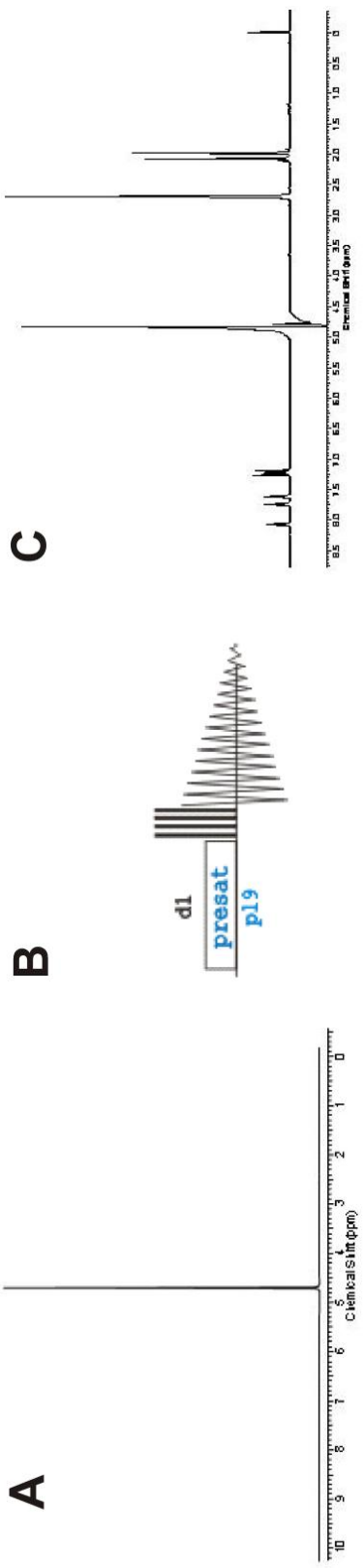

A
B
C
d1
presat
p19
ChemicalShift(ppm)

Figure 4.1 Solvent suppression for low concentration experiments. (A) A 100 μM sample of bromocresol green in an aqueous buffer prepared with 99.99% D2O. The residual protons from the water (~1100 mM) squelched the bromocresol green signals giving one strong peak in the center of the spectrum. (B) The presaturation with composite pulse water suppression technique is used to selectively suppress the solvent signal. Quality of solvent suppression is dependent on the power (p19) of the presaturation pulse (presat) and the pulse length, which is the same as the recycle delay (d1). A composite pulse (thin vertical black bars) is applied for analyte excitation to decrease the effect of inhomogeneities in the applied B_1 field. (C) A 100 mM sample of bromocresol green after solvent suppression. The experiment was collected using the pulse program in (B) with a recycle delay of 2.0 s and 64 scans. The time to collect one spectrum is approximately 2.25 min. The residual solvent peak can be removed for clarity during processing, but has a baselinewidth of 117 Hz.

4.3.1 Optimization of automated data collection. The largest limiting factor for high-throughput ligand binding studies using the FAST-NMR method is the nearly 8 min required for sample changing and experimental set up. This correlates to approximately 15 hrs of "dead time" between data collection. A large portion of the time was tied to the receiver gain adjustment and shimming routine (~3 min). The previously described method for FAST-NMR (chapter 3) required samples to be prepared in 500 μL volumes to reduce sample requirements. However, this inadvertently required a longer shimming routine because the sample was not uniformly covering the receiver coil. By preparing samples at a larger 600 μL volume that extends beyond the receiver coil, a shorter gradient shimming routine was implemented while maintaining good line shape and linewidth. Using a simple gradient shim routine saved nearly 2 min between samples. Additionally, setting the receiver gain to a constant value based on the first sample and removing the automatic receiver gain adjustment saved nearly 1 min of sample set up time. A total time savings of ~3 min per sample was seen by making small adjustments to the automatic data collection protocols reducing the total time between samples to ~5 min. For the FAST-NMR library of 113 mixtures this correlates to a savings of ~5.5 hrs, reducing the total time for sample changing and set up during a FAST-NMR screen to 9.4 hrs.

4.3.2 Improving 1D ^{1}H NMR screening efficiency. To further increase throughput of the FAST-NMR screen requires using a 1D ^{1}H pulse sequence that will give comparable or better results with a shorter recycle delay. The excitation sculpting pulse sequence for solvent suppression developed by Hwang *et. al,*[13] uses gradient pulses

to selectively irradiate the water. This removes the dependency on the recycle delay as found in the presaturation sequence.

I compared the standard FAST-NMR presaturation pulse sequence with the excitation sculpting sequence to determine if there were any improvements in the ligand binding analysis. Specifically, I was looking for improvement in water suppression, increases in signal to noise and overall spectral quality. Spectral quality was determined by the amount of post-processing editing required. I was also looking for differences in measured single point binding dissociations constants (K_D) as described in chapter 2.

Five replicate samples were made at two different human serum albumin (HSA) concentrations (0 μM and 5 μM) containing 50 μM naproxen, 5% (v/v) deuterated dimethyl sulfoxide-d6 (DMSO-d6), 11.1 μM 3-(trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt (TSP) and 50 mM potassium phosphate buffer pH 7.0 (uncorrected) in 99.98% deuterium oxide. All samples at 0 μM HSA were used for calculating the average free ligand intensities (I_F) and average free ligand linewidths (ν_F). All samples at 5 μM HSA were used for calculating the average bound ligand intensities (I_B). A 1D ^{1}H NMR spectrum using the presaturation sequence and the excitation sculpting sequence were collected sequentially for each sample. All samples were collected under the same conditions at a constant receiver gain of 32.

The excitation sculpting method efficiently suppressed the solvent signal such that no residual solvent signal remained (figure 4.2C&D). The resulting baseline was flat and did not require any baseline corrections. The presaturation sequence did not completely remove the residual solvent signal and required post-processing editing of the residual

solvent signal. Additionally, a baseline correction was needed because of a rolling edge near 10.0 ppm (figure 4.2A&B).

The signal to noise ratio for both sequences was comparable at the constant receiver gain set to 32 with the excitation sculpting sequence S/N = 74.9 and the S/N for the presaturation was 68.8 compared to the methyl peak in naproxen. The presaturation sequence is limited to a low receiver gain because the residual water signal is still large relative to the analyte concentration. However, the improved water suppression of the excitation sculpting method allows a larger receiver gain (1 k). This improves the S/N to 431.6 relative to the methyl peak in naproxen. This was a 6 fold improvement in S/N compared to the initial presaturation pulse sequence.

The naproxen average linewidth for the presaturation sequence was 3.52 ± 0.5 Hz, where the average linewidth for the excitation sculpting sequence was to 2.52 ± 0.1 Hz. The differences are due to removing the residual water signal. For example, without removing the residual water signal in the presaturation sequence the reference TSP peak has a half width of 7.79 Hz calculated by the peak fitting routine in ACD labs. Once the water peak is removed the reference peak linewidth drops to 1.91 Hz. The difference is caused by the automatic peak fitting routine in ACD misreading the true baseline of the spectrum, this is due to negative data points in the residual solvent signal (figure 4.2 A) and the baseline roll at the edge of the spectrum (figure 4.2 B). The excitation sculpting sequence does not have these issues and therefore the average calculated linewidth is smaller.

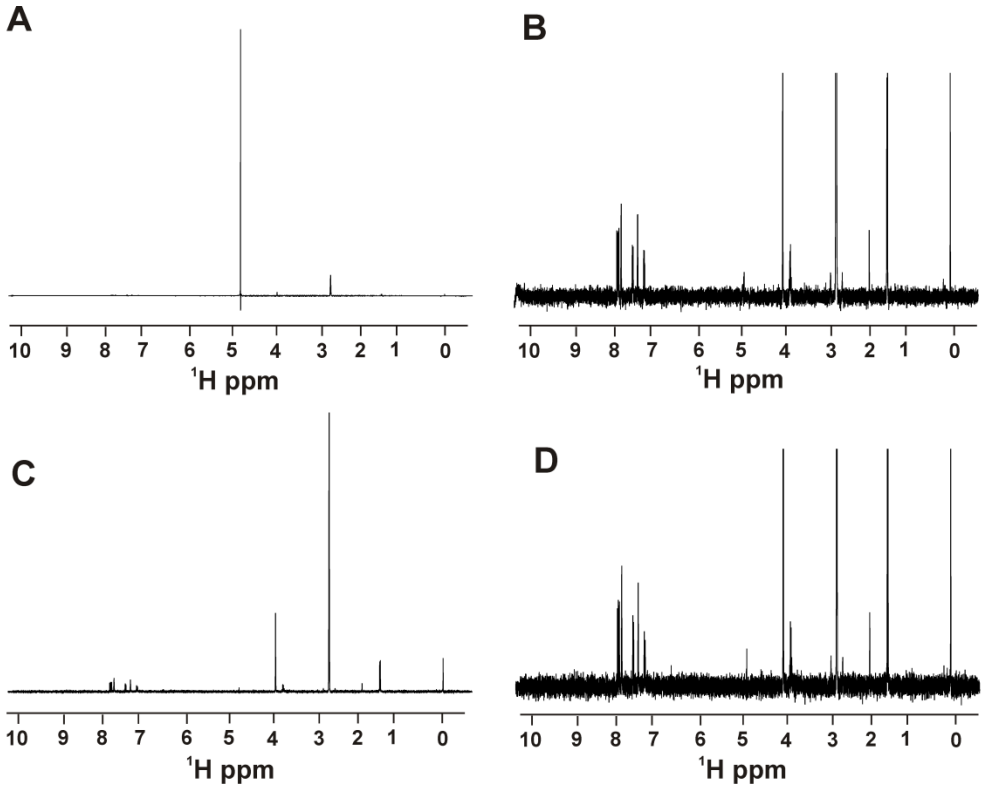


Figure 4.2. Comparison between two water suppression techniques. (A) The presaturation pulse sequence of a 50 μM naproxen sample in a 99.99% D_2O buffer. The pulse sequence does not completely suppress the residual water signal at 4.69 ppm. (B) An expanded view of the presaturation spectrum. In addition to not sufficiently suppressing the solvent signal the presaturation sequence generates a baseline roll at the edge of the spectrum around 10 ppm. These issues distort the accurate measurement of the free ligand linewidth and introduce significant error in the K_D measurement. (C) The excitation sculpting sequence efficiently suppresses the solvent signal so no post processing editing is required. (D) Additionally the excitation sculpting method does not introduce baseline roll in the spectrum. The total time to collect a spectrum using the excitation sculpting sequence is approximately 1.25 min compared to 2.5 min for the presaturation method.

The difference in measured average linewidth between the two pulse programs has a dramatic effect on the accuracy in measuring single point binding constants (K_D). The peak height for each spectrum was summed and then averaged to calculate B_{expt} (eq 2.7) and the single point K_D (eq 2. 8) (see chapter 2 for method description). The measured K_D for naproxen biding to HSA was 0.36 μM using the excitation sculpting method and -43.7 μM using the presaturation pulse. The non-sense K_D value from the presaturation pulse was caused by the over estimation of the free ligand linewidth. In chapter 2, the average ligand linewidth using the presaturation pulse sequence was 1.8 Hz and the single point K_D was 0.7 ± 1.2. The data for chapter 2 was collected under analytical conditions with a large number of scans (512) and long experiment time of 33 min per sample. This is not amenable to high-throughput screening. The problem with accurately measuring a free ligand linewidth under high-throughput conditions severely limits the utility of the presaturation pulse sequence.

The results from the excitation sculpting sequence show a significant improvement over the presaturation pulse for the FAST-NMR screening. Solvent suppression using this sequence is not dependent on the recycle delay reducing the total time needed to collect a single spectrum by approximately 1.25 min. Additionally, the results from an excitation sculpting sequence do not need post processing solvent filtering which dramatically improves the single point K_D method for high-throughput NMR ligand affinity screens.

4.3.3 Improving 2D ^{1}H-^{15}N HSQC NMR screening efficiency. The standard 2D ^{1}H-^{15}N HSQC spectrum correlates the amide proton to the amide nitrogen giving a single peak for each amino acid (figure 4.3). The current method using the standard 2D

^{1}H-^{15}N HSQC requires between 1.5-2.5 hrs per spectrum and is therefore a significant portion of the screening time for FAST-NMR. In addition to the data collection time, the method requires a minimum of 100 μM ^{15}N labeled protein per sample equaling between 1-30 mgs of protein depending on the number of hits from the 1D ^{1}H NMR screen.

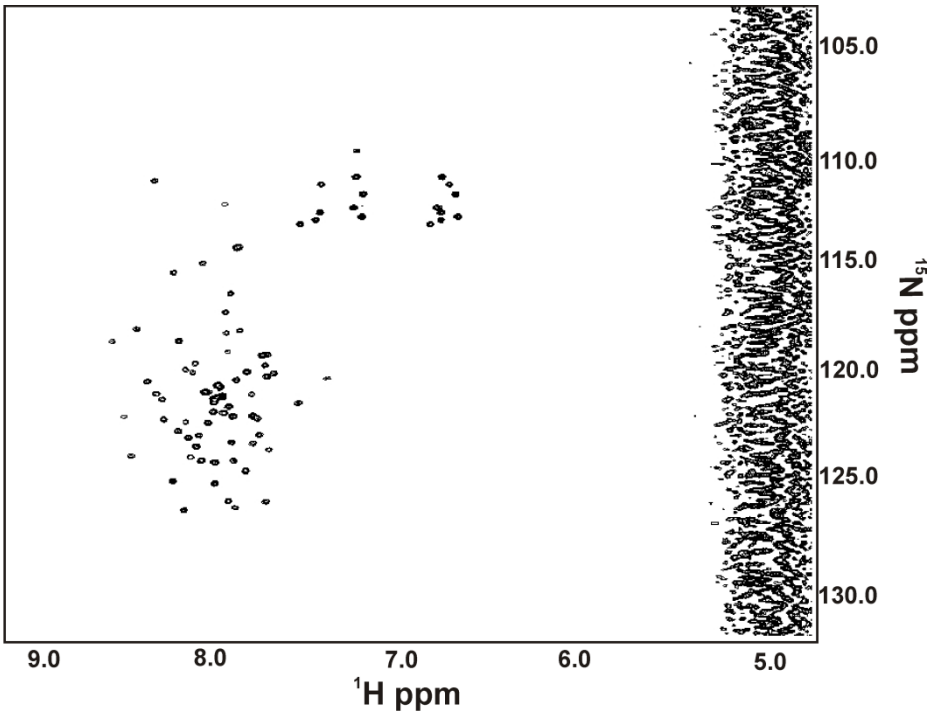


Figure 4.3 A standard 2D ^{1}H-^{15}N HSQC Spectrum. The standard 2D ^{1}H-^{15}N HSQC correlates each backbone amide proton with its corresponding backbone amide nitrogen. Samples are collected in 95% H_2O/5%D_2O buffers with a large residual solvent streak (5.0 ppm). The relative ratio of analyte to solvent signal reduces the overall signal to noise requiring larger concentrations of analyte and longer data collection times (1.5-2.5 hrs). The sample was 100 μM PrgI in 95%H2O/5%D2O buffered solution of 20 mM bis-Tris pH 7.0).

To increase the efficiency and versatility of the 2D ^{1}H-^{15}N HSQC screening step in FAST-NMR, a solvent suppressed 2D ^{1}H-^{15}N HSQC pulse sequence was implemented. The sequence uses the WATERGATE and water flip back method for solvent suppression.[15] Suppressing the residual water in 2D ^{1}H-^{15}N HSQC increases the flexibility for NMR screening. If sample is a limiting factor, the pulse sequence can detect protein concentrations as low as 5 μM with an extended acquisition time. If sample concentration is not a limiting factor, the pulse program can be used to collect an NMR spectrum in approximately 20 min at 100 μM protein concentration (figure 4.4).

A 5 μM sample of ^{15}N labeled *S. aureus* nuclease was prepared in a 95% H_2O/5%D_2O buffer with 50 mM KPO_4 and 300 mM NaCl. Data was collected using the WATERGATE[15]/water flip back 2D ^{1}H-^{15}N HSQC[16] pulse sequence with 400 real scans, 128 dummy scans, 1 k data points in the ^{1}H dimension and 128 data points in the ^{15}N dimension. The sweep width of the spectrum was 17.0 ppm in the ^{1}H dimension and 30.0 ppm in the ^{15}N dimension. A recycle delay for the pulse sequence was set to 1.0 s. The total experiment time was 13 hrs. A 100 μM sample of ^{15}N labeled *S. aureus* nuclease was prepared using the same buffer conditions. An NMR spectrum was collected similar to the 5 μM sample, but with only 8 real scans. The total time to collect a 2D ^{1}H-^{15}N HSQC at 100 μM protein concentration was approximately 20 min. There was no difference in peak position between the two experiments. No differences were observed compared to a standard 2D ^{1}H-^{15}N HSQC spectrum of free nuclease at 1.2 mM protein concentration (figure 4.5C). The WATERGATE/water flip back 2D ^{1}H-^{15}N HSQC experiments sufficiently suppressed the residual solvent peak such that no post-processing editing was required.

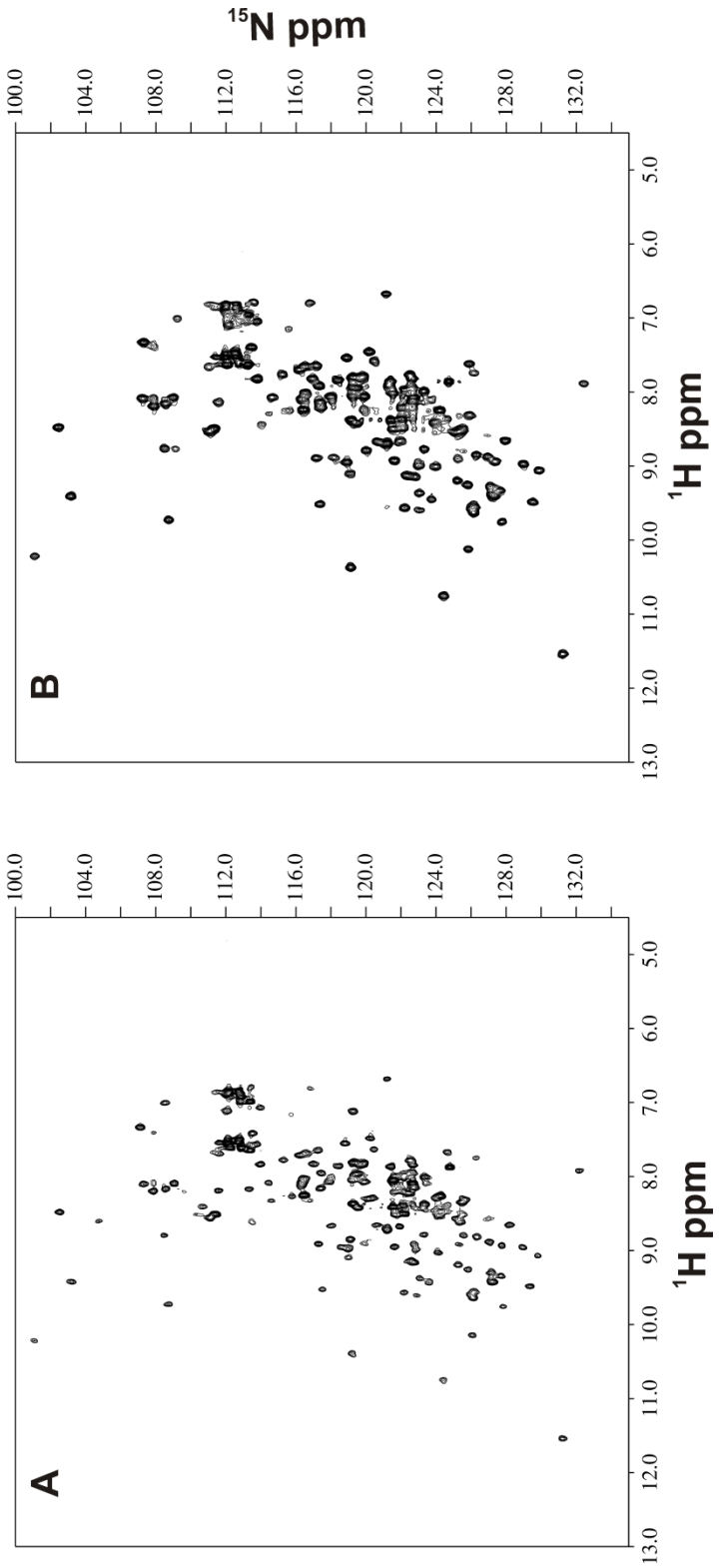

A
B
^{15}N ppm
^{1}H ppm
100.0
104.0
108.0
112.0
116.0
120.0
124.0
128.0
132.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0

Figure 4.4 Concentration study of the WATERGATE/water flip back 2D ^{1}H-^{15}N HSQC. (A) A 5 μM sample of ^{15}N labeled *S. aureus* nuclease was prepared in the a 95% H_2O/5%D_2O buffer with 50 mM KPO_4 and 300 mM NaCl. NMR spectrum was collected with 400 scans and experiment time was approximately 13 hrs. (B) A 100 μM sample of ^{15}N labeled *S. aureus* nuclease under the same conditions collected with 8 scans. Total experiment time was approximately 20 min.

4.3.4 FAST-NMR screen of *S. aureus* nuclease. The 19 kDa protein *Staphylococcus aureus* nuclease is a well-studied NMR model system and was used to test the FAST-NMR optimization and validate its functional annotation.[7-9] The goal of the experiment was to identify nuclease binding ligands using the FAST-NMR screening methods, to identify the active site of the protein, and to complete a successful CPASS analysis. The hypothesis was that we would find the same binding site as previously reported for the nuclease- thymidine-3',5'-diphosphate ligand bound co-structure.[7-9] Furthermore, we would identify a preferential similarity between this nuclease's ligand binding site and other nuclease ligand binding sites.

Unlabeled and uniformly ^{15}N labeled nuclease was expressed and purified as described in the experimental sections 4.2.6 and 4.2.7. Expression was checked by comparing induced and non-induced samples of three growth cultures (figure 4.5A). A total of 75.6 mg/L unlabeled and 23.4 mg/L ^{15}N labeled purified nuclease was obtained from 2 L growths. All concentrations were measured by maximum UV absorbance at 280 nm with a molar extinction coefficient, ε, of 17,420 M^{-1} cm^{-1}. A 2D -^{1}H-^{15}N HSQC was collected on the purified sample of ^{15}N labeled nuclease and compared to the reported 2D ^{1}H-^{15}N HSQC spectrum[7-9] and associated assignments to check for proper folding of the protein (figure 4.5C). The spectrum for the expressed nuclease was comparable to the reference spectrum[7-9] with differences most likely accounted for by differences in buffer, temperature, spectral resolution and the slightly longer sequence of the expressed nuclease (9 additional N-terminal amino acids, see appendix 4A for comparison between expressed nuclease and reference sequence).

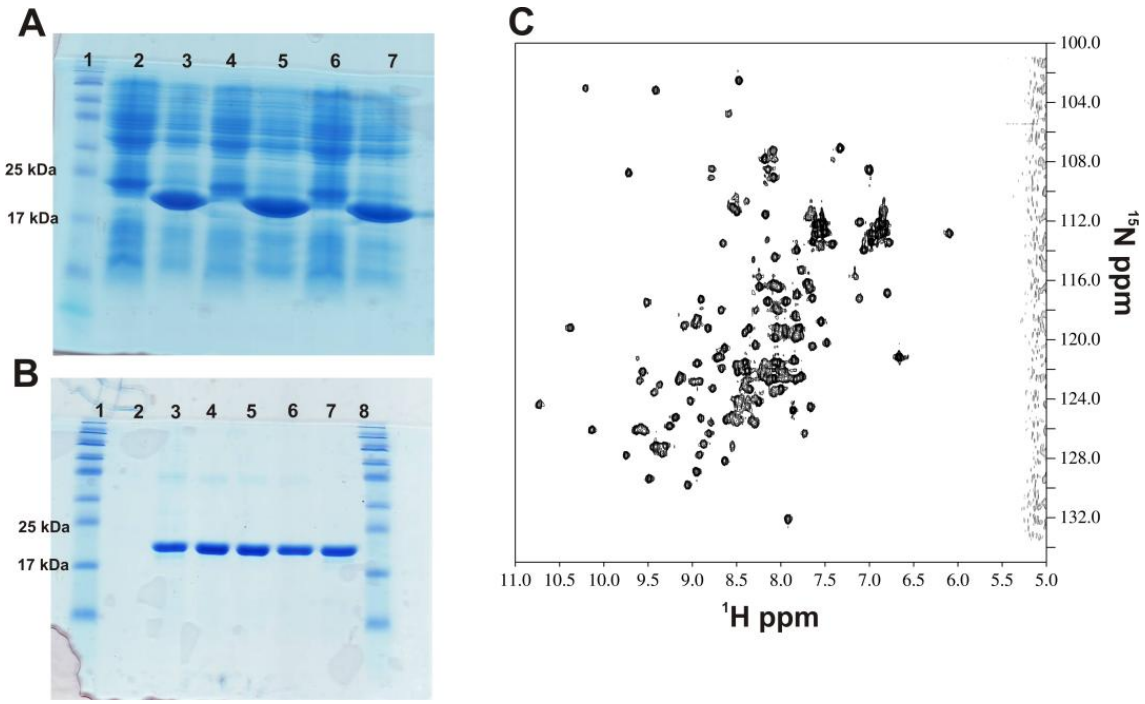


Figure 4.5 Expression and purification of *S. aureus* nuclease. (A) Induced cultures (lane 3, 5 and 7) of unlabeled nuclease from 3 randomly selected colonies of *E. coli* BL21-DE3-codon+(nuc) were compared to non-induced cultures (lane 2, 4, 6). A dark band was identified in the induced cultures at approximately 19 kDa (MW lane 1). (B) Purification of the culture media with a his-tag resin gave 5 isolated bands (lane 3-7) at the same molecular weight as in (A). (C) The expression and purification was repeated under minimal media conditions for expression of ^{15}N labeled nuclease. A 2D ^{1}H-^{15}N HSQC was collected on a sample from the purified stock solution (1.2 mM). The protein spectrum was dispersed indicating a folded protein.

Unlabeled nuclease was screened against the compound library as described in the experimental section 4.2.8. A total of 18 ligands were identified in the 1D ^{1}H NMR screening. Table 4.1 reports the list of all nuclease-binding ligands found in our chemical library. All nucleotides in the chemical library bound nuclease. Binding of thymidine-5'-triphosphate was indicated by relative changes in peak height between the free and bound spectrum and the appearance of enzymatic turnover of the ligand (figure 6A). Two new peaks at 7.32 ppm and 8.3 ppm are visible when the 5 μM nuclease is added to the sample.

Table 4.1 Ligands identified to bind nuclease from a high-throughput NMR screen.

Binding ligand
Adenosine-5'-triphosphate
Guanosine-5'-triphosphate
Uracil-5'-triphosphate
Cytosine-5'-triphosphate
Thymadine-5'-triphosphate
3'-5'-cyclic guanosine monophosphate
3'-5'-cyclic adenosine monophosphate
Suramin
Mitoxantrone dihydrochloride
Phosphocholine
4-Hydroxy-3-methoxyphenylglycol
Aquocobalamin
L-leucine
Bepridil dihydrocholoride
Ciprofloxacin
Diminazene
Lumicolchince
Acebutolol hydrochloride

Confirmation of ligand binding was completed by monitoring changes in ^{15}N and ^{1}H chemical shifts upon addition of 500 μM thymidine-5'-triphosphate ligand to a 100 μM ^{15}N labeled nuclease sample. Binding site residues between nuclease and thymidine-5'-triphosphate were identified in a similar manner as described in chapter 3. A total of 17 residues were identified to have greater than 1 standard deviation from the average chemical shift difference upon addition of the ligand. Of these 17 residues, 6 were identified in the side chain amide region, 6 residues were unambiguously identified and the remaining 5 were either not assigned in the reference spectrum or could not be unambiguously identified. Four of the unambiguously identified residues, F34, R35, K84, and Y113, were residues found in the active site of the reference structure (figure 4.6B&C). The reference binding pocket is composed of 8 amino acids F34, R35, and L36, T82, D83, K84, Y115, V114, and Y113. Thymidine-5'-triphophate was the only ligand titrated with ^{15}N labeled nuclease because thymidine-3',5'-diphosphate is the bound ligand for the reference nuclease structure (PDB 1JOK), which was not in the FAST-NMR chemical library at the time of screening.

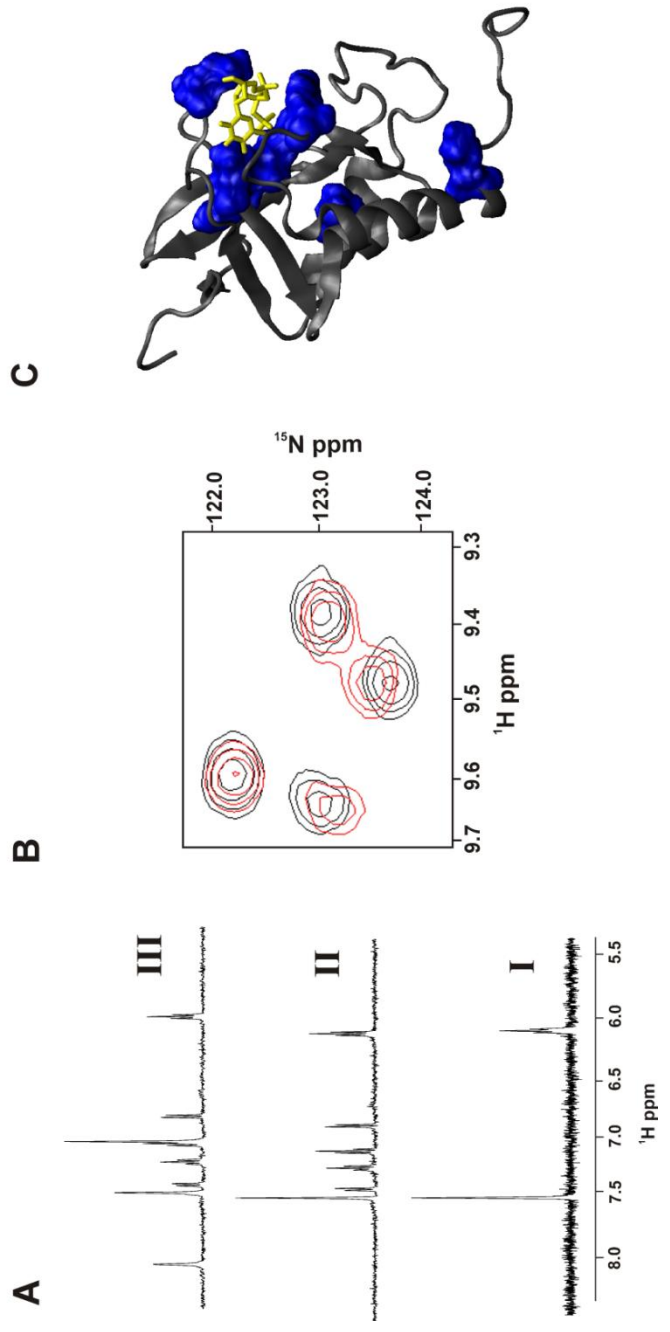
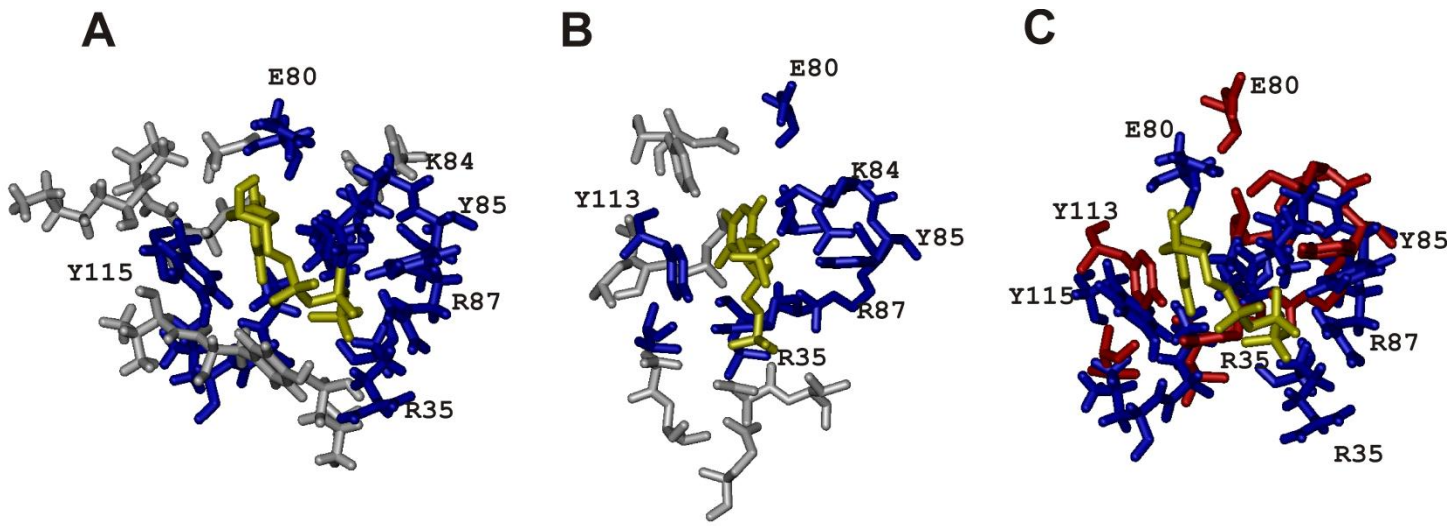

A
III
II
I
8.0
7.5
7.0
6.5
6.0
5.5
^{1}H ppm
B
^{15}N ppm
122.0
123.0
124.0
9.3
9.4
9.5
9.6
9.7
^{1}H ppm
C

Figure 4.6 FAST-NMR screen of *S. aureus* nuclease. (A) Unlabeled nuclease was screened with the FAST-NMR compound library as described in chapter 2 using the new pulse sequences as described in section 4.2.8. 18 ligands were found to bind nuclease with thymidine-5'-triphosphate (AI single, AII mixture free) showing possible enzymatic turnover (AIII bound) in addition to a decrease in signal. Two new NMR resonance not found in the free mixture (AII) are observed in the complex (AIII). The assignment of these peaks is not clear, but most likely correspond to the formation of thymidine-5'-diphosphate from thymidine-5'-triphosphate. The non-binding compounds in the mixture include biotin and acetylsalicylic acid. (B) 17 peaks significantly changed in a 2D ^{1}H-^{15}N HSQC spectrum upon the addition of 500 μM thymidine-5'-triphophate to a 100 μM sample of ^{15}N labeled nuclease. For clarity, an example of the relative change upon ligand binding for two residues (F34 and R35) is shown (black free nuclease, red ligand bound nuclease). (C) The residues identified in the 2D ^{1}H-^{15}N HSQC spectrum are highlighted on the protein structure (1JOK) and used to generate a ligand bound co-structure. Structure images were generated with VMD[17]

A ligand bound co-structure of nuclease with thymidine-5'-triphosphate was generated in the same manner as described in detail in chapter 3. The residues identified in the 2D ^{1}H-^{15}N HSQC spectrum binding study that overlapped with the known binding site were used to define the grid search space for AutoDock. The Autodock Filter program[18] was run to select the best conformation. Finally, the ligand bound co-structure was uploaded to CPASS.[5]

The best hit for the nuclease-thymidine-5'-triphosphate docked co-structure was a *Staphylococcus* nuclease protein (PDB 1TR5) bound to thymidine-3,-5'-diphosphate. The active site similarity score was 47.47% with an average rmsd of 0.69 ± 0.3 Å for the overlapping active site residues. CPASS did not find the structure used to generate the ligand bound structure (PDB 1JOK) because the program filters out proteins with ≥ 95% sequence similarity and/or ligand binding sites with ≥ 80% sequence similarity. However, recent updates to the CPASS database and software now allow for pairwise active site comparisons. The pairwise comparison between the docked nuclease co-structure and the experimental co-structure bound to thymidine-3',5'-diphosphate had a pairwise active site similarity score of 48.1% with and an average rmsd of 0.76 ± 0.3 Å for the overlapping active site residues.




```
1JOK: R35 L36 L37 G79 E80 T82 N83 K84 Y85 R87 L89 Y113 V114 Y115 K116 P117 N118
1TR5: R35  -  L36  -  E80  -  N83 K84 Y85 R87 L80  -    -   Y113  -    -    -
```

Figure 4.7 CPASS analysis of *S. aureus* nuclease. (A) The ligand bound co-structure for nuclease complexed with thymidine-5'-triphosphate (yellow) was uploaded to the CPASS database. (B) The best match was the *Staphylococcus* nuclease protein (PDB 1TR5) bound to thymidine-3,-5'-diphosphate (yellow). (C) An overlay of the two active sites (1JOK blue, 1TR5 red) gave an overall rmsd of 0.69± 0.3 Å and a CPASS similarity score of 47.47%. The sequence alignment of the two ligand binding sites is shown below the figures, where the aligned residues are colored blue in A and B. Structure images were generated with VMD.[17]

4.4 REFERENCES

1. 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* **2009,** 12, (3), 285-95.
2. Mercier, K. A.; Baran, M.; Ramanathan, V.; Revesz, P.; Xiao, R.; Montelione, G. T.; Powers, R., FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* **2006,** 128, (47), 15292-9.
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* **2006,** 9, (7), 515-34.
4. Mercier, K. A.; Powers, R., Determining the optimal size of small molecule mixtures for high-throughput NMR screening. *J. Biomol. NMR* **2005,** 31, (3), 243-258.
5. Powers, R.; Copeland, J. C.; Germer, K.; Mercier, K. A.; Ramanathan, V.; Revesz, P., Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **2006,** 65, (1), 124-35.
6. 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* **2008,** 10, (6), 948-58.
7. Wang, J. F.; Hinck, A. P.; Loh, S. N.; Markley, J. L., Two-dimensional NMR studies of staphylococcal nuclease: evidence for conformational heterogeneity from hydrogen-1, carbon-13, and nitrogen-15 spin system assignments of the

aromatic amino acids in the nuclease H124L-thymidine 3',5'-bisphosphate-Ca2+ ternary complex. *Biochemistry* **1990,** 29, (17), 4242-53.

8. 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-Ca2+ ternary complex. *Biochemistry* **1990,** 29, (1), 102-13.
9. 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-Ca2+ ternary complex. *Biochemistry* **1990,** 29, (1), 88-101.
10. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **1995,** 6, (3), 277-93.
11. Garrett, D. S.; Powers, R.; Groenenborn, A. M.; Clore, G. M., A common sense approach to peak picking in two-, three-, and four-dimensional spectra using automatic computer analysis of contour diagrams. *Journal of Magnetic Resonance (1969-1992)* **1991,** 95, (1), 214-20.
12. Fogh, R.; Ionides, J.; Ulrich, E.; Boucher, W.; Vranken, W.; Linge, J. P.; Habeck, M.; Rieping, W.; Bhat, T. N.; Westbrook, J.; Henrick, K.; Gilliland, G.; Berman, H.; Thornton, J.; Nilges, M.; Markley, J.; Laue, E., The CCPN project: an interim report on a data model for the NMR community. *Nat Struct Biol* **2002,** 9, (6), 416-8.

13. Hwang, T.-L.; Shaka, A., Water Suppression That Works. Excitation Sculpting Using Arbitrary Waveforms and Pulsed-Field Gradients. *J Mag Res, Series A* **1995,** 112, (2), 275-279.

14. Maniatis, T.; Fritsch, E. F.; Sambrook, J., *Molecular Cloning. A laboratory manual*. Cold Spring Harbor Laboratory: Cold Spring Harbor, NY, 1982.

15. Piotto, M.; Saudek, V.; Sklenar, V., Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J Biomol NMR* **1992,** 2, (6), 661-5.

16. Andersson, P.; Gsell, B.; Wipf, B.; Senn, H.; Otting, G., HMQC and HSQC experiments with water flip-back optimized for large proteins. *J Biomol NMR* **1998,** 11, (3), 279-88.

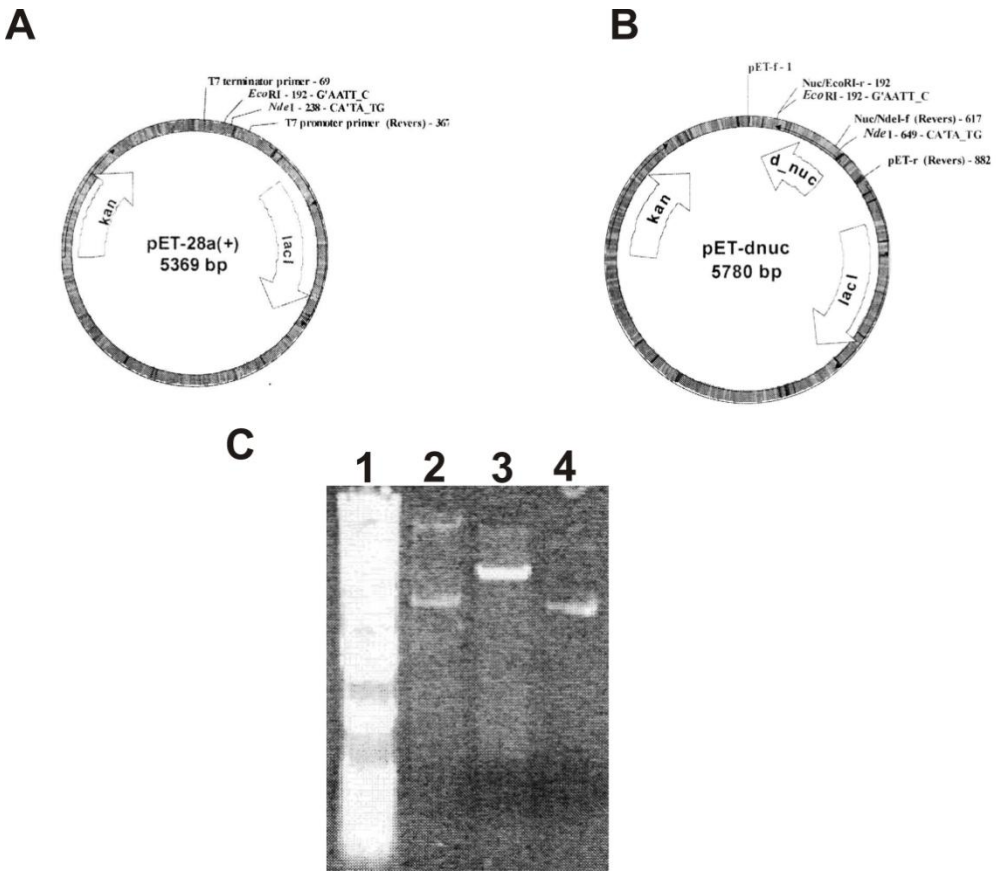
17. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *J Mol Graph* **1996,** 14, (1), 33-8, 27-8.

18. Stark, J.; Powers, R., Rapid protein-ligand costructures using chemical shift perturbations. *J Am Chem Soc* **2008,** 130, (2), 535-45.

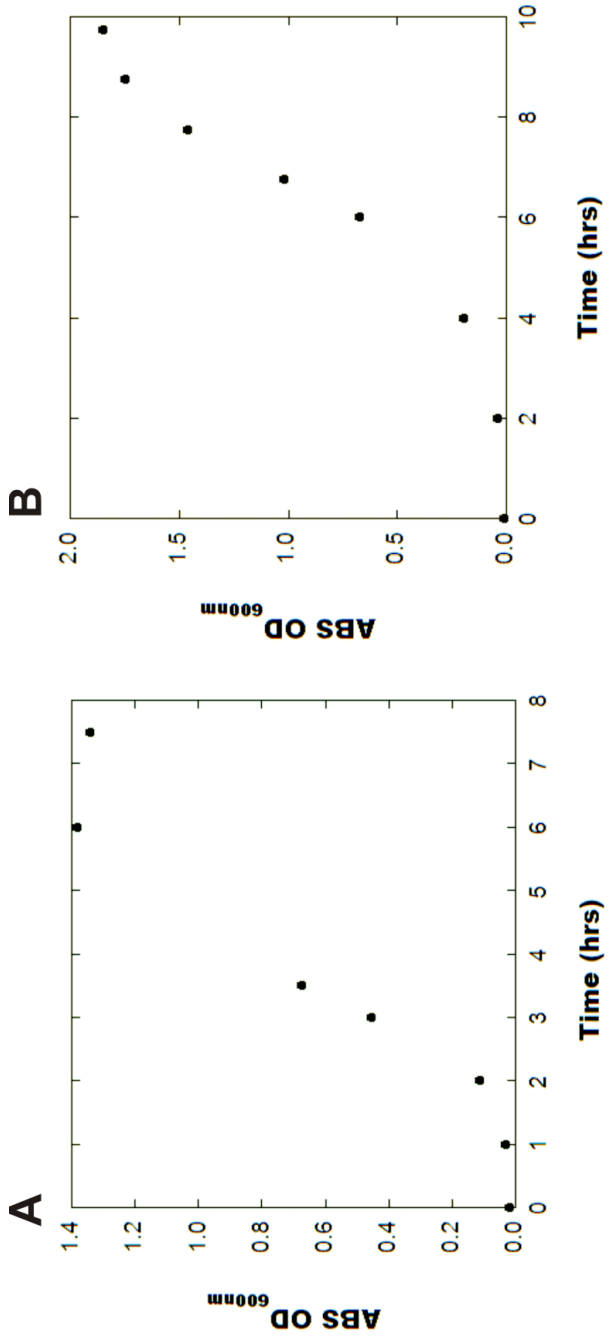
Appendix 4A. Sequence of *S. aureus* nuclease. A ClustalW sequence alignment with the sequence of the expressed nuclease (dNuclease) is shown with the reference nuclease sequence (refNuclea). The reference nuclease sequence was reported from the PDB ID 1JOK[7-9]

```
dNuclease       MGSSHHHHHHSSGLVPRGSHMATSTKKLHKEPATLIKAIDGDTVKLMYKGQPMTFRLLLV 60
refNuclea       ----------------------ATSTKKLHKEPATLIKAIDGDTVKLMYKGQPMTFRLLLV 39
                                       ************************************

dNuclease       DTPETKHPKKGVEKYGPEASAFTKKMVENAKKIEVEFDKGQRTDKYGRGLAYIYADGKMY 120
refNuclea       DTPETKHPKKGVEKYGPEASAFTKKMVENAKKIEVEFDKGQRTDKYGRGLAYIYADGKMV 99
                ************************************************************

dNuclease       NEALVRQGLAKVAYVYKPNNTHEQLLRKSEAQAKKEKLNIWSEDNADSGQ 170
refNuclea       NEALVRQGLAKVAYVYKPNNTHEQLLRKSEAQAKKEKLNIWSEDNADSGQ 149
                ***************************************************
```

Appendix 4B. Comparison of the standard pET-28a(+) and nuclease inserted pET-28a(+)–nuc plasmids used for the nuclease expression. (A) Standard pET-28a(+) plasmid (B) nuclease inserted plasmid. (C) 1% agarose gel of the isolated pET-28a(+)-nuc plasmid (lane2), digested plasmid (lane 3) and control pET-28a(+) plasmid (lane 4).



Appendix 4C. Growth curves for nuclease expression. (A) Expression of unlabeled nuclease with IPTG induction at 3.5hrs OD_{600nm} 0.67. (B) Expression of ^{15}N labeled nuclease with IPTG induction at 6.75 hrs OD_{600nm} 0.79. The difference in growth rates was caused by the difference in growth media.



CHAPTER 5:
THE STRUCTURE, DYNAMICS AND LIGAND SCREENING OF THE PRIMASE C-TERMINAL DOMAIN (CTD) FROM *STAPHYLOCOCCUS AUREUS*

5.1 INTRODUCTION

Bacterial primase (DnaG) is a conserved and essential enzyme responsible for the synthesis of Okazaki fragments during DNA replication.[1] The protein is composed of three domains; N-terminal domain responsible for DNA binding, the catalytic core responsible for synthesis of Okazaki fragments, and the C-terminal domain (CTD) responsible for the interaction between primase and bacterial helicase (DnaB).[2, 3] Full length primase is conserved among all organisms and exhibits relatively large sequence similarity.[1, 4, 5] However, the sequence conservation is limited to the N-terminus and catalytic core.[1] The C-terminal domain (primase CTD) is highly variable; even among similar species.[1, 4, 5] The functional consequence of the low sequence conservation of the C-terminal domain is still unclear, but it could play a role in regulating species-specific DNA replication.[6]

The solution structures of primase CTD shows significant variability between *Geobacillus stearothermophilus* (PDB 1Z8S) and *Escherichia coli* (PDB 2HAJ).[4, 7] Generally, the primase CTD structure is composed of two sub-domains, an N-terminal six-helix bundle (sub-domain C1) that is essential for DnaB activity and correct primer length and a helical hairpin (sub-domain C2) that mediates binding to DnaB.[4, 8] The two solution structures share significant structure similarity at the N-terminal bundle (C1 sub-domain) but show a sharp difference in the corresponding C2 sub-domain.[4, 7, 8] In *E. coli* primase CTD is composed of 7 helices with the two sub-domains connected through a long ridged helix 6.[7] In *G. stearothermophilus* the helix linking the two sub-domains is

kinked at a Pro556 residue forming two distinct helices (helix 6 and 7).[4] The recent structure of the DnaG-DnaB complex shows that both the C1 and C2 sub-domains are important for binding to helicase.[9]

The discrepancy in the DnaG CTD structures has yet to be fully resolved. However, it has been shown that primer synthesis is only carried out when primase CTD and helicase N-terminal domain (NTD) interact.[6] The differences in sequence and structure of the primase CTD between the two organisms suggest a species-specific method of replication regulation.[6] The *S. aureus* sequence (see appendix 5A) for primase CTD is more similar to the *G. stearothermophilus* sequence with 20% sequence identity and 58% sequence similarity. However the sequence similarity between *S. aureus* sequence and the *E. coli* sequence (57% similarity and only 10% sequence identity) is comparable to the sequence similarity between *S. aureus* and *G. stearothermophilus*. The comparable sequence similarities make direct homology modeling challenging because either structure is a possible model for *S. aureus*. But, by comparing the sequence similarities of the loop region between helix 6 and 7, the proline residue (Pro556) that forms the kink in the linking helix in *G. stearothermophilus* is replaced with a glycine in *S. aureus* (appendix 5A). Glycine has the second largest propensity (second to proline) to be found in a loop region.[10, 11] Conversely, the *E. coli* sequence contains a methionine in the corresponding position consistent with a rigid helix 6. This single amino acid substitution between the three proteins suggests the *S. aureus* primase structure is more likely to be similar to the *G. stearothermophilus* structure.

The rapid rise in community acquired antibiotic resistance, particularly to *S. aureus,* requires the rapid identification of new antibiotic targets and potential drugs.[12]

The primase-helicase interaction is an attractive antibiotic target because it is functionally conserved in bacteria, essential for DNA replication and the bacterial DnaG-DnaB interaction is distinctly different from that of eukaryotes.[1, 6] Additionally, the high degree of sequence variability and differences in structure suggest a possible means to tailor antibiotic development to a specific organism.

As described in chapter 1, the 1D ^{1}H and 2D ^{1}H-^{15}N HSQC screening methods used for FAST-NMR was originally developed for high-throughput drug discovery. However, to specifically find an active site for structure based drug discovery, the complete backbone resonance assignments and a high-resolution, three-dimensional (3D) structure are required. In this chapter, I will discuss the NMR determination of the solution structure for *S. aureus* primase CTD. I will examine a potential phylum dependency on the two sub-domain structures using sequence and structure similarities. I will also report protein dynamics for the conformation of a loop region between the two sub-domains. Finally, I will discuss the discovery of a potential lead compound that binds to the C2 sub-domain of primase CTD.

5.2 EXPERIMENTAL

5.3.1 Materials. For the DnaG primase CTD structure determination, NMR dynamics analysis of the structure, and the NMR ligand affinity screens, purified and uniformly ^{13}C, ^{15}N labeled [U-^{13}C, ^{15}N] DnaG primase CTD and ^{5}N labeled [U-^{15}N] DnaG primase CTD was purchased from Nature Technologies (Lincoln, NE) (see figure 5.1A for gel). The dimethyl sulfoxide-d_6 (99.9% D) and deuterium oxide (99.9% D) were obtained from Aldrich (Milwaukee, WI). The 3-(trimethylsilyl)propionic-2,2,3,3-d_4

acid sodium salt (98% D) was purchased from Cambridge Isotope (Andover, MA). The potassium phosphate dibasic salt (anhydrous, 99.1% pure) and monobasic salt (crystal, 99.8% pure) were purchased from Mallinckrodt (Phillipsburg, NJ). All compounds used for screening were obtained as described in chapters 3, 4 and elsewhere.[13] Briefly, the compound library is composed of 437 known biologically active compounds distributed across 113 mixtures with 3-4 compounds in each mixture.

5.3.2 Apparatus. All NMR experiments used for the protein backbone assignments of DnaG primase CTD were collected at 298 K on a five channel 600 MHz Bruker Avance spectrometer equipped with a 5 mm TXI probe. NMR experiments used for the protein side chain resonances and distance constraints were collected at the Rocky Mountain Regional 900 MHz NMR Facility on a four channel 900 MHz Varian INOVA spectrometer equipped with a 5 mm HCN probe. Assignments of the backbone and side chain resonances were obtained from the following spectra: 2D ^{1}H-^{15}N-HSQC, 2D ^{1}H-^{13}C-HSQC, HNCO, HNCA, CBCACONH, CBCANH, HNHA, HBHACONH, CCCONH, HCCCONH and H(CCH)-COSY (collected on 900MHz).[14] Distance constraints were obtained from 3D ^{15}N-edited NOESY and 3D ^{13}C-editied NOESY (collected at 900 MHz).[14]

Hydrogen bond constraints were determined using the (CLEANEX-PM)-FHSQC experiment.[15] A total of 2048 data points were collected in the ^{1}H dimension and 128 data points were collected in the ^{15}N dimension. The spectrum was collected with 16 transients and a sweep width of 8012.82 Hz in the ^{1}H dimension and 1613.424 Hz in the ^{15}N dimension. The mixing time was set to 100 ms with a CLEANEX spinlock power of 2 KHz.

All NMR experiments for protein dynamics analysis were collected on a Bruker 500 MHz Avance spectrometer (Billercia, MA) equipped with a triple resonance, Z-axis gradient Cryoprobe. Experiments used for dynamics study have been described previously[16-18] and included a 2D $^{1}H^{-15}NN$ HSQC experiment (*hsqct1etf3gpsi*) designed to measure T_1 relaxation rates with delay times of 0.0 ms, 5.39 ms, 53.92 ms, 134.80 ms, 269.60 ms, 404.40 ms, 539.20 ms, 674.00 ms and 1078.40 ms, a 2D ^{1}H-^{15}N HSQC experiment (*hsqct2etf3gpsi*) designed to measure T_2 relaxation rates with delay times of 0.0 ms, 17.6 ms, 35.2 ms, 52.8 ms, 70.4 8 ms, 105.6 ms, 123.2 ms, 140.8 ms, 158.4 ms, 176.0 ms, and a 2D ^{1}H-^{15}N HSQC experiment (*hsqcnoef3gpsi*) designed to measure NOE enhancements.

The relaxation rates (T_1,T_2) for each DnaG primase CTD amino acid was calculated by fitting the intensity of each peak to the intensity decay curve (appendix 5B) (eq 5. 1) where I_t is the intensity of each peak at the delay time *t*, I_0 is the initial steady state intensity

$$I_t = I_0 \exp\left(-\frac{t}{T_{1,2}}\right) \quad [5.1]$$

The NOE values were determined by the ratio of peak intensity between the saturated (I_{sat}) and unsaturated (I_{unsat}) spectra

$$NOE = I_{sat}/I_{unsat} \quad [5.2]$$

All T_1, T_2 and NOE data measurements were used to calculate an overall correlation time (τ_r), order parameters (S^2), internal motion (τ_e) or chemical exchange (R_{ex}) using the Lipari-Szabo model free method[19] implemented by FAST-MODEL FREE.[20]

All NMR experiments used for the ligand binding screen were collected on a

Bruker 500 MHz Avance spectrometer (Billercia, MA) equipped with a triple resonance, Z-axis gradient cryoprobe and using a Bruker BACS-120 sample changer and IconNMR software for automated data collection. All 1D ^{1}H NMR spectra were collected at 298K using the pulse sequence described in chapter 3. To increase throughput, only 64 transients were signal averaged for each spectrum with 8k data points. All 2D ^{1}H-^{15}N HSQC spectra were collected at 298K using the standard pulse sequence implemented in Bruker TopSpin 1.3 with optimized sample specific 90^{o} pulse lengths.

All multidimensional experiments were processed using NMRpipe,[21] analyzed using PIPP[22] or CCPNMR.[23] All 1D ^{1}H NMR spectra were processed with the ACD/1D NMR manager v. 12.0 (Advanced Chemistry Development, Inc., Toronto, Ontario). All ligand protein docking studies were completed as described in chapter 3 and 4.

5.3.3 Sample preparation. For NMR backbone assignment experiments, uniformly ^{13}C, ^{15}N labeled [U-^{13}C, ^{15}N] DnaG primase CTD was concentrated to 1.2 mM in a 95% H_2O/5% D_2O buffered solution of 100 mM NaCl, 25 mM KPO_4 pH 6.64 (uncorrected) using an Amicon ultra centricon (MW cutoff 10 000 Da). 50 mM arginine and 50 mM glutamine was added to the NMR sample for long term stability. For side chain experiments uniformly ^{13}C, ^{15}N labeled [U-^{13}C, ^{15}N] DnaG primase CTD was concentrated to 1.4 mM in the same buffer conditions used for the NMR backbone assignment experiments.

NMR dynamics data was collected using a uniformly ^{15}N labeled [U-^{15}N] sample of DnaG primase CTD concentrated to 1.2 mM in a 95% H_2O/5% D_2O buffered solution of 100 mM NaCl, 25 mM KPO_4 pH 6.64 (uncorrected) using an Amicon ultra centricon

(MW cutoff 10 000 Da). 50 mM arginine and 50 mM glutamine was added to the NMR sample for long term stability.

Sample preparation and experimental parameters for the NMR ligand affinity screen were executed in the same manner as described previously[24] and in chapter 3. Briefly, each ligand mixture (113 total) was screened using 1D ^{1}H NMR at 100 μM ligand concentration with 25 μM protein in a 99.99% D_2O buffered solution of 20 mM d_{19}-bis-Tris at pH 7.0 (uncorrected) with 2% DMSO-d_6 to maintain ligand solubility and 11.1 μM 3-(trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt as a chemical shift reference. 1D ^{1}H NMR spectra for each sample was collected using a pre-saturation pulse sequence with 64 real transients, 8 dummy transients with 8 K data points, a sweep width of 11.0 ppm and a recycle delay of 2.0 s. Data was Fourier transformed, auto-phase and baseline corrected. Each 1D ^{1}H NMR spectrum were compared to the corresponding free ligand mixture reference spectrum and visually analyzed to identify binding ligands. A binding event was identified by the decrease in ligand intensity of the nuclease-mixture relative to the free ligand mixture. Total data collection time including sample changing was approximately 10 min/spectrum

Additionally, a ligand free 2D ^{1}H-^{15}N HSQC spectrum was collected using the same buffer conditions with 95% H_2O/5% D_2O to ensure the protein was properly folded prior to addition of each ligand.

5.3.4 NMR Structure calculations and refinement. NOE assignments were obtained by using 3D ^{15}N-edited NOESY and 3D ^{13}C-edited NOESY experiments. NOE intensities were sorted visually into four classes: strong (1.8–2.5), medium (1.8–3.0), weak (1.8–4.0), very weak (3.0–5.0). Upper limits for distances involving methyl protons

and nonstereospecifically assigned methylene protons were corrected appropriately for center averaging. Initial NOE assignment was completed by the program Autostructure[25] which identified 1055 intra-residue, 173 sequential, 312 medium range ($1 \geq 5$) and 73 long range (5 >) NOEs. Due to significant peak overlap, even at a high magnetic field (900 mHz), manual refinement was needed to complete NOE assignment. All torsion angle constraints were obtained by chemical shift analysis using the TALOS[26] software program, and measured coupling constants from an HNHA experiment.[27]

Hydrogen bond constraints were determined using the (CLEANEX-PM)-FHSQC experiment.[15] The (CLEANEX-PM)-FHSQC spectrum was compared with the 2D ^{1}H-^{15}N HSQC spectrum, where amides with missing peaks were assigned hydrogen bond constraints. These residues were selected because the (CLEANEX-PM)-FHSQC spectrum identifies amide residues with fast water exchange rates. The hydrogen bond distance constraints were set at 2.8 Å between the carboxyl oxygen and the amide nitrogen, and 1.8 Å between the carboxyl oxygen and the amide proton. Carboxyl groups within 2.5 Å of the slowly exchanging amide groups were selected to be involved in a hydrogen bond.

The structures were refined using the hybrid distance geometry dynamical-simulated annealing method[28] with minor modifications[29] using the program XPLOR-NIH[30] adapted to incorporate pseudopotentials for $^{3}J(HN\text{-}H\alpha)$ coupling constants,[31] secondary $^{13}C\alpha/^{13}C\beta$ chemical shift constraints,[32] and a conformational database potential.[33-35] A total of 1000 structures were calculated. The 20 lowest energy structures were then subjected to further energy minimization with CNS using explicit water solvation that included Lennard-Jones and electrostatic potentials using a modification of

the procedure and forcefield of Nilges.[36, 37] An average DnaG primase CTD structure was calculated from these 20 structures.

The target function that is minimized during restrained minimization and simulated annealing comprises quadratic harmonic terms for covalent geometry, ^{3}J(HN-Hα) coupling constants, and secondary ^{13}Cα/^{13}Cβ chemical shift constraints, square-well quadratic potentials for the experimental distance and torsion angle constraints, and a quadratic van der Waals term for nonbonded contacts. The force constant for the conformational database was kept relatively low (0.5–1.0 kcal/mol) throughout the simulation to allow the experimental distance and torsion angle constraints to predominately influence the resulting structures. The force constant for the NOE and dihedral constraints were 30 times and 10 times stronger than the force constants used for the conformational database.[38] All peptide bonds were constrained to be planar and trans. There were no hydrogen-bonding, electrostatic, or 6–12 Lennard-Jones empirical potential energy terms in the target function.

5.3 RESULTS AND DISCUSSION

5.3.1 NMR assignments and secondary structure prediction of primase C-terminal domain from *Staphylococcus aureus*. The backbone resonance assignments were completed using the NMR experiments described above (^{1}H-^{15}N HSQC, HNCO, HNCA, HNCOCA, CBCACONH, CBCANH, HNHA, HBHACONH and the ^{1}H-^{15}N HSQC edited NOESY) and manually analyzed using PIPP[22] and CCPNMR.[23] The backbone resonance assignment was 85% complete with 139 amino acids of the 163 unambiguously assigned in the ^{1}H-^{15}N HSQC (figure 5.1). Unassigned residues in the

^{1}H-^{15}N HSQC include M1-H13, D19, E28, H37, L38, M9, T58, R94, E95, E101, P109, and Y110. The majority of the unassigned residues correlate to the engineered N-terminal sequence (MGHNHNHNHNHNGGDDDD) for purification, residues M1-H13 correlated with the N-terminal his-tag, and residue D19 is part of an engineered proteolytic cleavage site. Excluding the purification tag the backbone assignments were 94% complete. The ten amino acids found in the primase sequence that were not assigned were primarily found in unstructured loop regions, turns between two helices or at the edge of a helix. Residues H37-M39 were in a turn region between helix 1 and 2, residue T58 was in an unstructured loop region between helix 2 and helix 3 and residues E101, P109 and Y110 were in an unstructured loop region between helix 5 and helix 6. Residues R94 and E95 are the second and third residues of helix 5. An exhaustive analysis of the NMR data set was unable to yield an assignment for these residues, suggesting the end of the helix may undergo partial unfolding and exchange broadening.

Aliphatic side chain carbon chemical shift assignments were completed using the CCCONH experiment correlating the preceding (i-1) residue to the following (i) backbone amide chemical shift. Aliphatic side chain proton chemical shifts were completed with the HCCH-COSY and HCCCONH experiments. Aromatic side chain assignments were completed using the 3D ^{13}C-edited NOESY experiment. The statistics for resonance assignment include, 139/163 HN, 139/201 N, 139/163 Cα, 134/168 Hα, 128/141 Cβ, 148/181 Hβ, 85/92 Cγ, 89/160 Hγ, 49/64 Cδ, 49/64 Hδ, 19/31 Cε 19/31 Hε, 4/7 C_{ζ} 4/7 H_{ζ} and 132/143 CO**.** All assignments will be uploaded to the BMRB.[39]

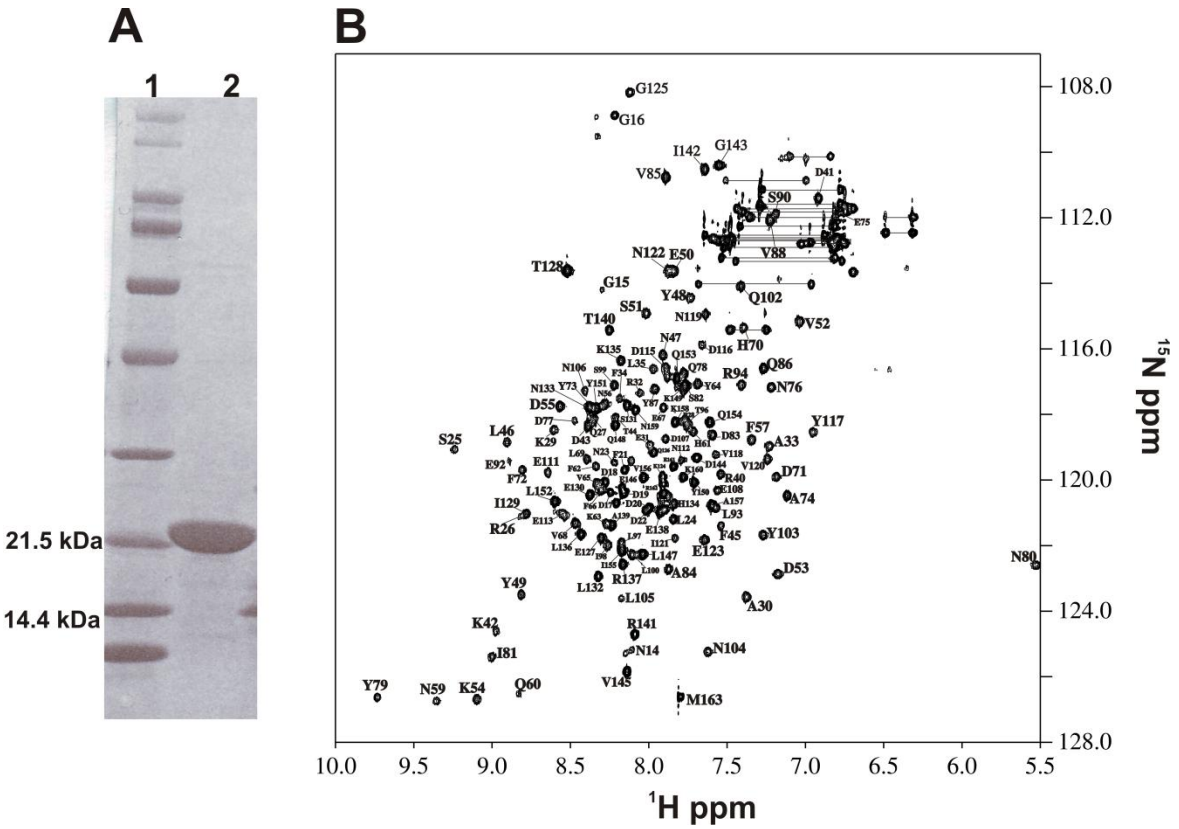


Figure 5.1 Assigned 2D ^{1}H-^{15}N HSQC spectrum of *S. aureus* primase CTD. (A) Purification of *S. aureus* primase CTD, lane 1 MW marker, lane 2 shows the expressed and purified ^{13}C/^{15}N labeled *S. aureus* primase CTD used for all studies in this work. (B) Complete backbone ^{1}H and ^{15}N assignments of the DnaG primase CTD from *S. aureus*. The spectrum was fully assigned with the exception of one peak at ^{1}H 7.90 ppm and ^{15}N 120.4 ppm. The peak is large and broad relative to other peaks in the spectrum and is likely the remaining unassigned his tag residues.

Secondary structure prediction using the difference in backbone $\Delta^{13}C\alpha/^{13}C\beta$ carbon chemical shifts between the assigned residues and random coil chemical shifts predict an all α-helical protein with 8 helices (figure 5.2A). Helical structures in primase CTD include Helix 1 Arg26-Lys36, Helix 2 Asp41-Glu50, Helix 3 Gln60-Glu75, Helix 4 Ile81-Tyr87, Helix 5 Asn91-Gln102, Helix 6 Try110-Lys124, Helix 7 Ile129-Arg141, Helix 8 Glu146-Glu161. The C1 sub-domain of primase CTD includes Helix 1-6 and the C2 sub-domain includes Helix 7-8 (figure 5.3B). This is consistent with the *S. aureus* DnaG primase CTD homology modeling predicted from the *Geobacillus stearothermophilus* structure[4] and the secondary structure prediction server NetSurfP (figure 5.2B).[40]

Of particular interest are the residues between the predicted helices 6 and 7 (residues K124-T128). This region is significantly different in *G. stearothermophilus* primase CTD compared to *E. coli* primase CTD solution structures (PDB 1Z8S and 2HAJ, respectively).[7] In *G. stearothermophilus*, this region is a loop forming two distinct sub-domains (C1, C2) of primase CTD. In *E. coli*, the region is a long and rigid α helix. For *S. aureus*, the experimental secondary structure $^{13}C\alpha$ and $^{13}C\beta$ chemical shift differences suggest that region is similar to the *G. stearothermophilus* structure with an extend loop region starting at residue Gly125 (figure 5.2). The $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts for residues in this region are near random coil chemical shift values.

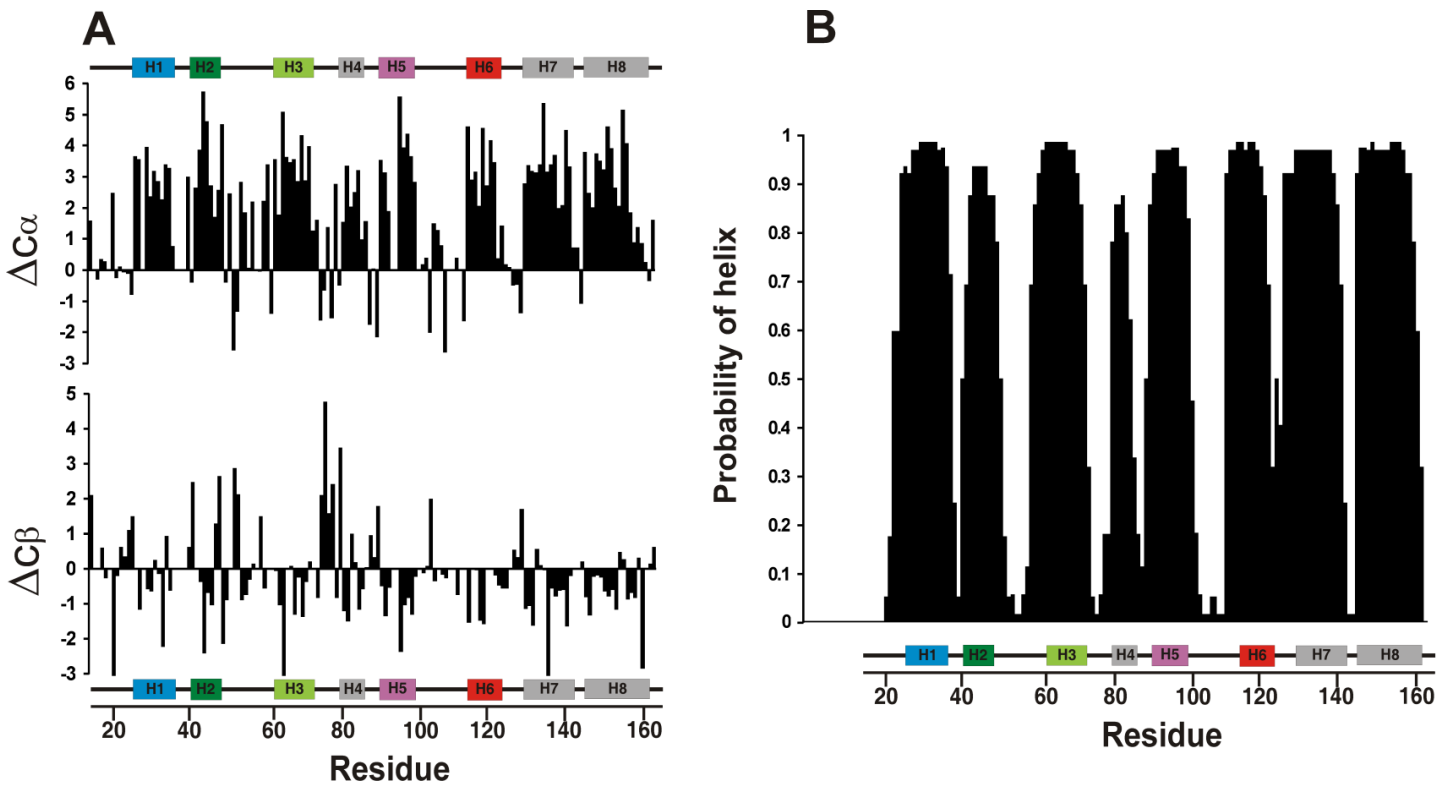

A
H1
H2
H3
H4
H5
H6
H7
H8
Δcα
Δcβ
Residue
B
Probability of helix
Residue

Figure 5.2. Secondary structure prediction for *S. aureus* primase CTD based on $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts. (A) secondary structures in *S. aureus* primase CTD are predicted based on differences in measured $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts compared to random coil chemical shift values. For $\Delta^{13}C\alpha$, positive regions represent α helical structure. For $\Delta^{13}C\beta$, negative values indicate α helix. The secondary structures are overlaid onto the results showing regions of α helix. (B) The predicted secondary structures in *S. aureus* primase CTD measured by NetSurfP.[40] Positive regions are the probability of the sequence stretch adopting α helix secondary structure. Both experimental and predicted secondary structure analysis suggest a loop region between helix 6 and 7 starting at residue G125.

5.3.2 Structure calculation and analysis of primase C-terminal domain (CTD) from *Staphylococcus aureus*. The solution structure of *S. aureus* primase CTD was calculated using 1823 distance restraints, 280 dihedral restraints measured by TALOS[26], 256 $^{13}C\alpha/^{13}C\beta$ carbon chemical shift restraints and 82 $^{3}J_{NH\alpha}$ coupling constant restraints. A complete list of the restraints used for the structure calculation is described in table 5.1. A total of 1000 structures were calculated from 10 individual sets of 100 structures using XPLOR-NIH[30] scripts described previously.[41] The lowest energy structures from each set were consolidated to generate a set of 20 low energy structures which were further refined in a water bath using the RECOORD scripts[37] implemented with CNS.[42, 43]

The resulting *S. aureus* primase CTD structures are consistent with the NMR data as evident by the relatively low rms deviations from experimental distance, dihedral, $^{13}C\alpha/^{13}C\beta$ chemical shift and ^{3}J(HN-Hα) coupling constant constraints (figure 5.3A). Also there are no distance violations > 0.5 Å or dihedral angle violations > 5°. The average root-mean square deviation (RMSD) of the 20 lowest energy structures about the mean coordinate positions is 0.97 ± 0.16 Å for all backbone atoms and 1.73 ± 0.39 Å for all heavy atoms with aligned residues 26-35, 40-49, 60-75, 82-85, 92-100 and 112-124. The final restrained minimized average structure of *S. aureus* primase CTD has an RMSD about the mean coordinate positions of 0.19 Å for all backbone atoms and 0.49 Å for all heavy atoms.

Table 5.1: Structural Statistics and Atomic rms Differences[a]

A. Structural Statistics	<SA>	$(\overline{SA})_r$
rms deviations from experimental distance restraints (Å)		
all (1823)	0.046 ± 0.008	0.079
interresidue sequential (|i-j| = 1) (460)	0.038 ± 0.008	0.096
interresidue short range (1 <|i-j| # 5) (446)	0.058 ± 0.010	0.074
interresidue long-range (|i-j| > 5) (140)	0.074±0.016	0.180
intraresidue (663)	0.003±0.008	0.004
H-bonds (114)[b]	0.072±0.030	0.040
rms deviation from exptl dihedral restraints (deg) (280)[c,d]	1.644±0.754	0.611
rms deviation from exptl Cα restraints (ppm) (130)	1.12 ± 0.05	1.08
rms deviation from exptl Cβ restraints (ppm) (126)	1.06 ± 0.02	1.02
rms deviation from $^3J_{NH\alpha}$ restraints (Hz) (82)	0.83 ± 0.06	1.02
F_{NOE} (kcal mol^{-1})[d]	212 ± 84.6	602.59
F_{tor} (kcal mol^{-1})[d]	36 ± 45	6.36
F_{repel} (kcal mol^{-1})[e]	65.63 ± 24	26.38
$F_{L\text{-}J}$ (kcal mol^{-1})[f]	-553.91 ± 29	-1212.10
deviations from idealized covalent geometry		
bonds (Å) (2684)	0.003±0.0	0.002
angles (deg) (04795)	0.504 ± 0.045	0.035
impropers (deg) (1468)[g]	0.441±0.069	0.344
PROCHECK[h]		
Overall G-Factor	-0.13 ± 0.03	-0.19
% Residues in most favorable region of Ramachandran plot	80.2 ± 3.1	85.5
H-bond energy	0.41 ± 0.41	0.45
Number of bad contacts/100 residues	25 ± 5.3	0.0

B. Atomic rms Differences (Å)

	C1 Domain (residues 26-124)		secondary structure[i]	
	backbone atoms	all atoms	backbone atoms	all atoms
<SA> vs SA	1.2 ± 0.1.7	2.00 ± 0.17	0.70 ± 0.44	2.4 ± 2.40
<SA> vs (SA)r	1.37 ± 0.19	2.32 ± 0.25	0.58 ± 0.22	1.52 ± 0.6
(SA)r vs SA	1.02	1.67	0.52 ± 0.33	1.21 ± 0.58

[a]The notation of the structures is as follows: <SA> are the final 20 simulated annealing structures and $(\overline{SA})_r$ is the restrained minimized mean structure obtained by restrained minimization of the mean structure $\overline{SA}$. The number of terms for the various restraints is given in parentheses. [b]For backbone NH-CO hydrogen bond there are two restraints: r_{NH-O} = 1.5-2.3 Å and r_{N-O} = 2.5 - 3.3 Å. All hydrogen bonds involve slowly exchanging NH protons inferred from calculated structures and CLEANX fast-exchange experiment.[15] [c]The torsion angle restraints comprise 140 ϕ and 140 ψ. [d]The values of the square-well NOE (F_{NOE}) and torsion angle (F_{tor}) potentials (cf. eqs 2 and 3 in [44]) are calculated with force constants of 50 kcal mol $^{-1}$ Å^{-2} and 200 kcal mol^{-1} rad^{-2}, respectively. [e]The value of the quadratic van der Waals repulsion term (F_{rep}) (cf. eq 5 in [45]) is calculated with a force constant of 4 kcal mol $^{-1}$ Å^{-4} with the hard-sphere van der Waals radius set to 0.8 times the standard values used in the CHARMM [46] empirical energy function. [28, 46, 47] [f]E_{L-J} is the Lennard-Jones-van der Waals energy calculated with the CHARMM empirical energy function and is *not* included in the target function for simulated annealing or restrained minimization. [g]The improper torsion restraints serve to maintain planarity and chirality. [h]These were calculated using the PROCHECK program. [i]The residues in the regular secondary structure are: 26–35 (α_1), 40-49(α_2), 60-75(α_3), 82-85(α_4) 92-100(α_5), 112-124 (α_6), 128-143(α_7) and 146-162(α_8) rmsd values were measured by aligning each secondary structure element individually and calculating an average and standard deviation.

The quality of the *S. aureus* primase CTD NMR structure was analyzed using PROCHECK. The results for the average minimized structure (figure 5.3B) show that *S. aureus* primase CTD has an overall G-Factor of -0.13 ± 0.03 with no bad contacts, which are all consistent with a good quality structure. Also, all non-glycine dihedral angles lie within the expected region of the Ramachandran plot, where 85.5% of the backbone dihedral residues lie within the most favorable region with 100% of the residues falling in the allowed region. The PROCHECK analysis of the average minimized structure was completed with the removal of the N-terminal his-tag. The consistency of the dihedral angles further illustrates the quality of the structure. The 20 lowest energy structures and the restrained-minimized average structure will be deposited into the PDB.[48]

The *S. aureus* primase CTD structure is composed of 8 helices. Helical structures in primase CTD include Helix 1 Arg26-Lys36, Helix 2 Asp41-Glu50, Helix 3 Gln60-Glu75, Helix 4 Ile81-Tyr87, Helix 5 Asn91-Gln102, Helix 6 Try110-Lys124, Helix 7 Ile128-Arg141, Helix 8 Glu146-Glu161. The C1 sub-domain of primase CTD includes Helix 1-6 and the C2 sub-domain includes Helix 7-8 (figure 5.3B). Conformation of the loop region between the two sub-domains was established by the lack of sequential NH-NH NOEs in the 2D ^{1}H-^{15}N HSQC edited NOESY and the presence of exchange peaks for each residue (G125, Q126 and E127) in the CLEANX experiment.[15] The results of the CLEANX experiment suggest these residues are undergoing exchange with the solvent and therefore not protected by hydrogen bonding; indicative of a loop structure.

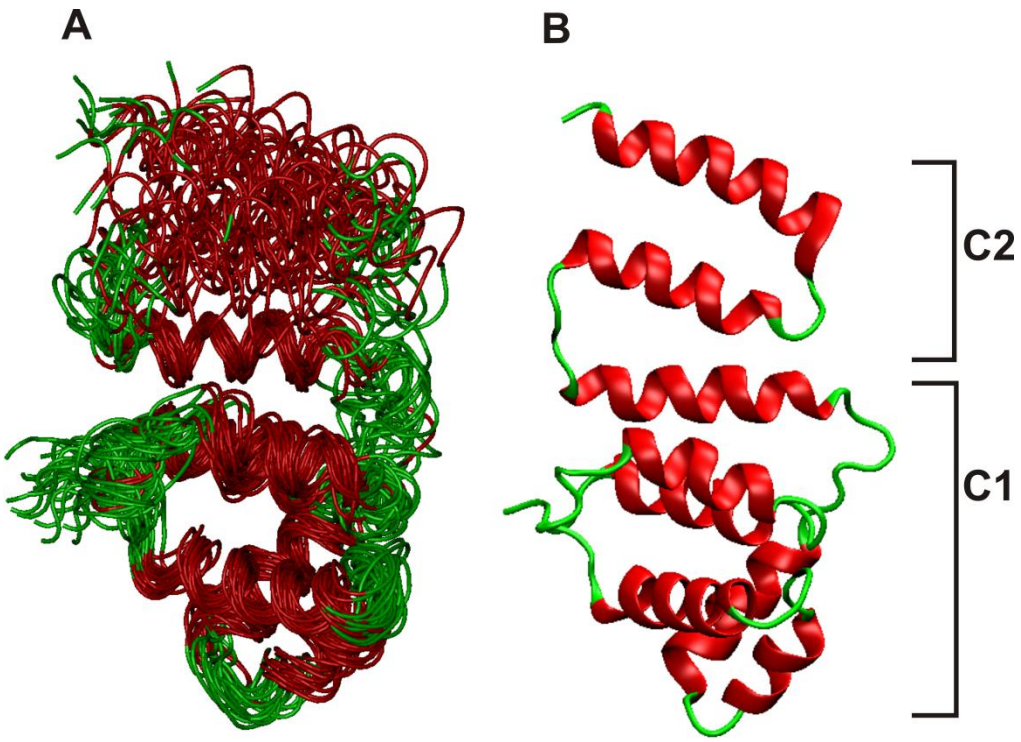


Figure 5.3 Comparison of the ensemble overlay and average minimized structure. (A) An overlay of the backbone trace of the 20 low energy, water refined structures aligned with residues 26-35, 40-49, 60-75, 82-85, 92-100 and 112-124 from the N-terminal C1 sub-domain. (B) A ribbon diagram of the average water refined structure 20. The two sub-domains are labeled C1 and C2. The C1 sub-domain is composed of helices 1-6 and the C2 sub-domain is composed of helices 7-8. Both structures are colored according to the secondary structure: red, α-helix; green, loop both images were generated with VMD.[49]

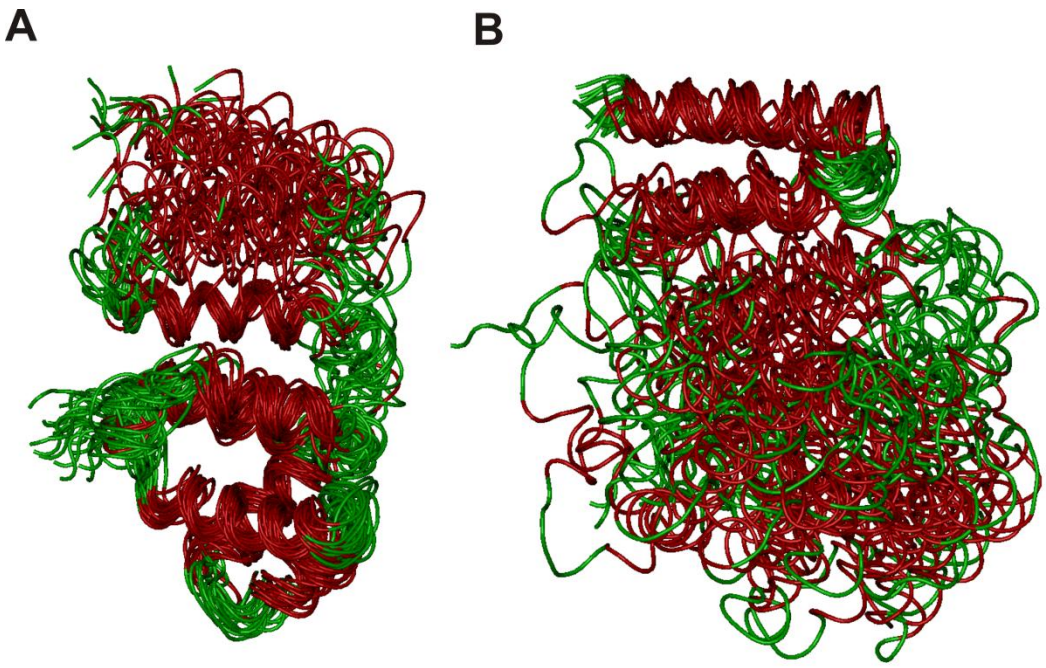


Figure 5.4. Ensemble overlay aligned to either sub-domain C1 or C2. (A) An overlay of the backbone trace of the 20 low energy structures aligned with residues 26-35, 40-49, 60-75, 82-85, 92-100 and 112-124 from the N-terminal C1 sub-domain. (B) An overlay of the backbone trace of the 20 low energy structures aligned with residues 128-141 and 146-161from the C-terminal C2 sub-domain. Both structures are colored according to the secondary structure: red, α-helix; green, loop both images were generated with VMD.[49]

The overall resolution of the protein structure was lower than what is generally possible with current NMR techniques. This lower resolution is an indication of the severe peak overlap in the NMR spectra, even at 900 MHz resolution. Figure 5.5A is the resulting 2D ^{1}H-^{13}C HSQC slice of the aliphatic 3D ^{13}C-edited NOESY experiment collected at the Rocky Mountain Regional 900 MHz NMR Facility. The spectrum was folded to increase resolution with the blue peaks representing the proton-carbon peaks for the Hβ-Cβ, Hγ-Cγ, Hδ-Cδ, Hε-Cε side chain resonances. The orange peaks correspond to the proton-carbon peaks for the Hα-Cα backbone resonances (for absolute Cα chemical shift add 35.804 ppm to each ^{13}C resonance). As an example of the severe peak overlap, the resolved Hα-Cα peaks in figure 5.5A only represent about half of the 163 possible assignments. The remaining peaks are buried in the broad and significantly intense region between 4.5 ppm ^{1}H and 15.0 ppm ^{13}C.

The severe peak overlap is also seen in the 2D ^{1}H-^{1}H slice of the aliphatic 3D ^{13}C-edited NOESY experiment (figure 5.5B). This is particularly problematic in the region 1.0-2.0 ^{1}H and 1.0-2.0 ^{1}H corresponding to the Hγ and Hδ side chain resonances of lysine, leucine, isoleucine and Hγ of valine. These 4 amino acids compose nearly 25% of the total amino acid composition of *S. aureus* primase CTD.

The severe peak overlap significantly complicated the complete side chain assignments with only 56% of Hγ assigned and 76% Hδ assigned. The corresponding number of long range (>5) NOEs was lower than anticipated (only 140) for a protein of 19.6 kDa; the main cause of the lower structure resolution. Peak overlap was also caused by degenerate chemical shifts due to an all α helical protein and by broader peaks due to protein dynamics (see section 5.3.5)

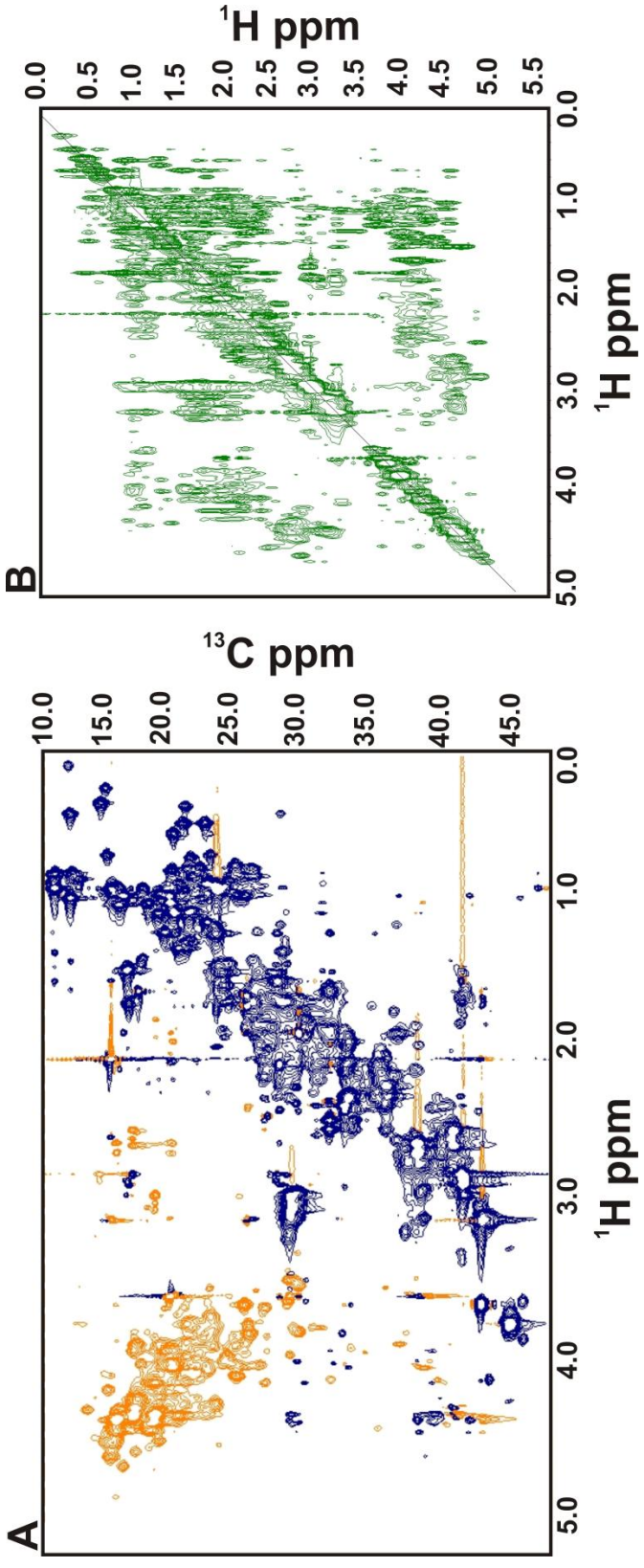

B
^{1}H ppm
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
^{1}H ppm
0.0
1.0
2.0
3.0
4.0
5.0
A
^{13}C ppm
10.0
15.0
20.0
25.0
30.0
35.0
40.0
45.0
^{1}H ppm
0.0
1.0
2.0
3.0
4.0
5.0

Figure 5.5. 3D ^{1}H-^{13}C HSQC edited NOESY of *S. aureus* primase CTD at 900 MHz. (A) The 2D ^{1}H-^{13}C plane of the 3D ^{13}C-edited NOESY spectrum of *S. aureus* primase CTD shows significant peak overlap, specifically in the Hα-Cα region (orange, note spectrum is folded add 35.804 ppm to all orange peaks for absolute chemical shift). A number of broad and intense peaks at 15.0 ppm ^{13}C and ~4.5 ppm ^{1}H show severe degeneracy in chemical shifts. (B) The 2D ^{1}H-^{1}H plane of the 3D ^{13}C-edited NOESY spectrum showing significant peak overlap in the Hγ and Hδ regions (1.5 ppm ^{1}H and 1.5 ppm ^{1}H)

5.3.3 Comparison between the three bacterial DnaG primase CTD structures. In *E. coli*, the primase CTD is composed of 7 helices with a long helix 6 connecting the C-terminus helix to the N-terminal bundle (figure 5.6 C).[7] The *G. stearothermophilus* structure is composed of 8 helices with the long helix 6 of *E. coli* broken into two helices forming two sub-domains (C1, C2) (figure 5.6B). A flexible loop region between helix 6 and 7 separates the two sub-domains in *G. stearothermophilus*.[4] The structure of *S. aureus* primase CTD is also composed of 8 helices with two sub-domains (C1, C2) separated by a flexible loop region between helix 6 and 7 (figure 5.6A). Figure 5.6 shows a side-by-side comparison for all three bacterial primase CTD structures. The residues found in the loop region between helix 6 and 7 are highlighted on figure 5.6.

A pairwise Dali[50] structure based alignment of the three primase CTD structures shows the *S. aureus* structure is similar to the *G. stearothermophilus* structure with a loop region separating the two sub-domains. The Z-scores for the pairwise structure similarities of the three structures are *S. aureus*-*G. stearothermophilus* 8.0, *S. aureus* –*E. coli* 6.5 and *G. stearothermophilus* - *E. coli* 5.3. The structure overlays are found in figure 5.7. Structure similarity between the three proteins is limited the N-terminal (C1) sub-domain. *E. coli* and *G. stearothermophilus* have the same overall fold with a backbone rmsd of 3.2 Å observed for the alignment of the first 6 helices that form an N-terminal helical bundle (C1).[4, 7] The same comparison for *S. aureus* to *E. coli* gives a backbone rmsd of 3.4 Å and the comparison between *S. aureus* and *G. stearothermophilus* gives a backbone rmsd of 2.8 Å.

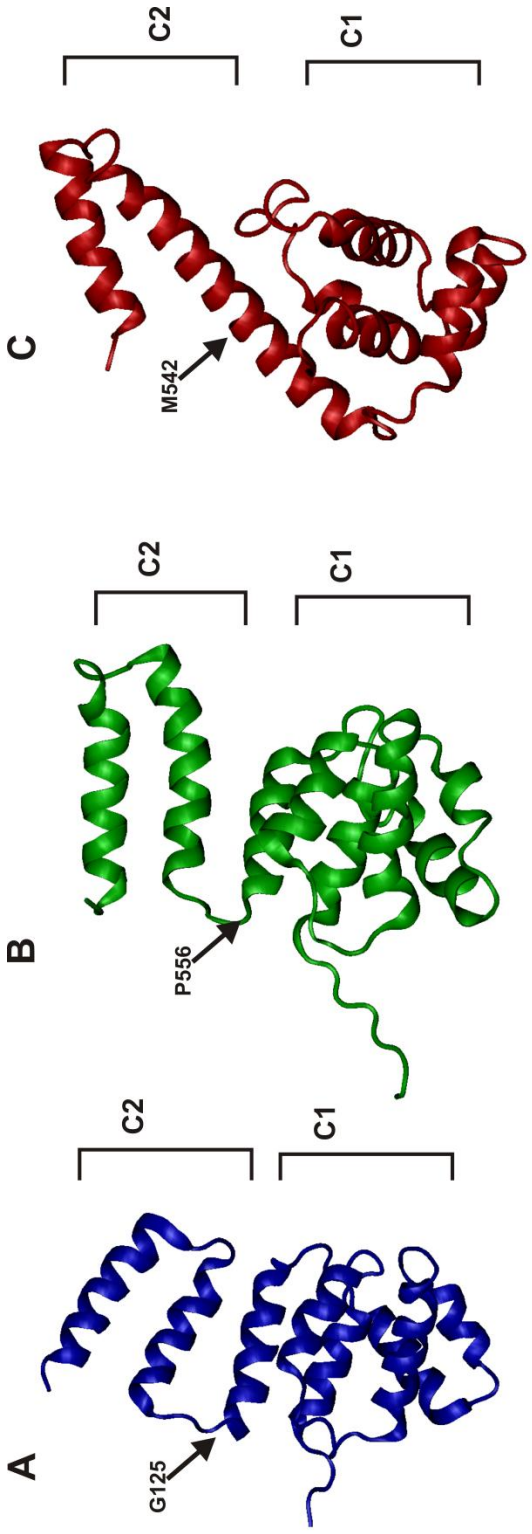

A
G125
C2
C1
B
P556
C2
C1
C
M542
C2
C1

Figure 5.6. Three bacterial primase CTD structures. The three bacterial primase CTD structures are reported showing the two different sub-domains and the residue responsible for the flexible linker. (A) Solution structure of *S. aureus* primase CTD, (B) solution structure of *G. stearothermophilus* primase CTD, and (C) solution structure of *E. coli* primase CTD. In both A and B, the two sub-domains are separated by a loop region linker. In *E. coli*, the loop region forms a ridged, continuous helix with a methionine residue in the structurally similar site to *S. aureus* and *G. stearothermophilus*.

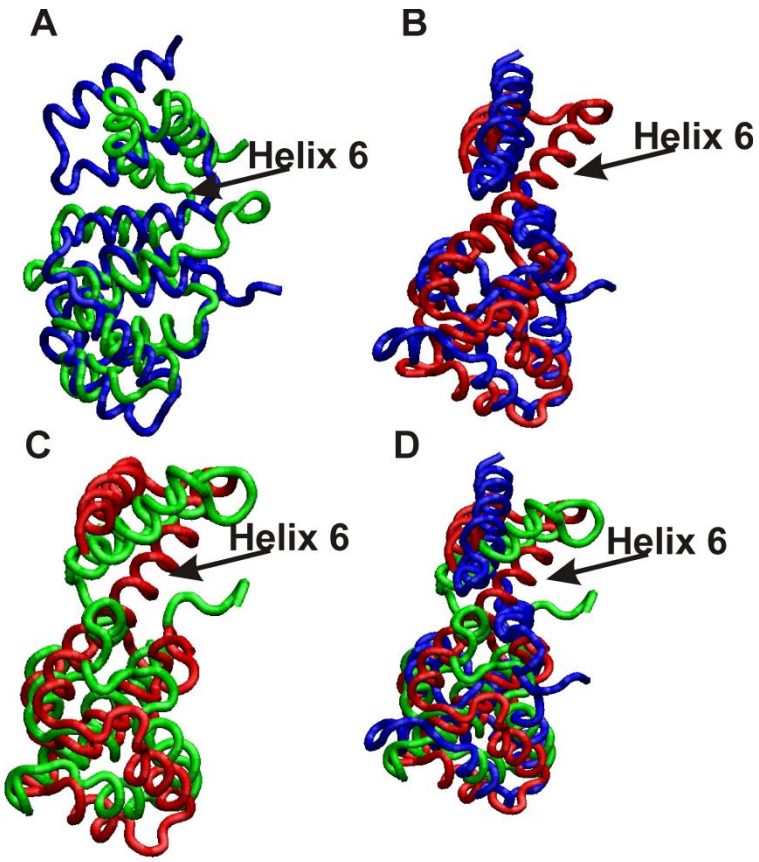


Figure 5.7. Structure similarities between the three primase CTD structures. (A). Comparison between *S. aureus* (blue) and *G. sterarothermophilus* (green) primase CTD gave a Z-score of 8.0and sequence identity of 20%. (B) Comparison between *S. aureus* (blue) and *E. coli* (red) primase CTD gave a Z-score of 6.5 and a sequence identity of 10%. (C) Comparison between *G. sterarothermophilus* (green) and *E. coli* (red) primase CTD gave a Z-score of 5.3 and sequence identity of 14%. (D) Multiple structure alignment of all three structures shows the conservation in the N-terminal bundle. The *E. coli* structure has an extend helix 6, which is broken into two helices in the two *Firmicutes* structures. The long helix 6 of *E. coli* is highlighted to show the primary difference in the structures.

As described previously, the only known structure similarity for the two previously solved primase CTD structures is the N-terminus of the replicateive helicases.[4, 7] The *S. aureus* primase CTD structure is also similar to the N-terminal domain of the replicative helicases. As with *E. coli* and *G. stearothermophilus*, the similarity is limited to the N-terminal helical bundle (C1). A comparison of *S. aureus* primase CTD with the Dali[51, 52] database identified the N-terminal domain of the *G. stearothermophilus* helicase as having the highest similarity (Z-score of 8.4) to *S. aureus* primase CTD. The remaining significant hits included the 3 previously solved primase CTD structures (1Z8S, 1T3W, and 2HAJ) from *G. stearothermophilus* and *E. coli* respectively. Additionally, the N-terminal domains of DnaB helicase from *E. coli* (1B79), *H. pylori* (3GXV) *T. aquaticus* (2Q6T) and *Bacillus phage spp1* (3BGW) were identified as structurally similar to *S. aureus* primase CTD.

5.3.4 Phylum dependency of the helix 6 structure. Similar to *G. stearothermophilus*, the *S. aureus* primase CTD structure also has a loop between helix 6 and 7. Examining the 3 non-redundant structures currently solved for the C-terminal domain of primase suggests a phylum dependency on the helix 6 loop structure. The difference in this helix is the primary reason the two *Firmicutes* structures are more similar to each other than to the *Proteobacteria* primase CTD structure. In *G. stearothermophilus*, the loop between helix 6 and 7 is composed of the amino acids Asn554, Arg555, and Pro556. Conversely, in the *S. aureus* structure, the loop is composed of the amino acids Gly125, Gln126, and Glu127. A multiple sequence alignment suggests the helix breaking proline appears to be limited to bacillus organisms.[7] Correspondingly, the glycine that forms the loop region between helix 6 and

7 in *S. aureus* appears to be limited to other S*taphylococcus* organisms. Thus, the loop that forms the C1 and C2 sub-domains in the *G. stearothermophilus* and *S. aureus* primase CTD structures appears to be phylum dependent and the sequence appears to be species dependent.

To follow this hypothesis further, the secondary structure of primase CTD for 6 different organisms was completed using NetSurfP.[40] The NetSurfP[40] accurately predicted the secondary structure for *S. aureus* primase CTD (figure 5.2B), supporting its reliable for accurately predicting secondary structures. In all three *Firmicutes* sequences (*G. stearothermophilus*, *S. aureus* and *B. anthracis*), a loop is predicted between helix 6 and 7 that forms two independent sub-domains (figure 5.8). Interestingly, the residues that form the loop are not highly conserved (figure 5.9). In all three *Proteobacteria* sequences (*E. coli, Y. pestis and P. aeruginosa*), the loop is not present and a long ridged helix 6 remains (figure 5.8). Again, the residues that make up the ridged portion of helix 6 and are structurally aligned with the loop region in the *Firmicutes* are not highly conserved (figure 5.10). My hypothesis is that primase CTD regulates binding to helicase in a phylum dependent manner based on structure. Secondly, primase CTD binding to DnaB helicase is sequentially regulated in a species-specific manner.

The helicase interaction with the CTD of primase is essential for primer synthesis during DNA replication.[2, 53, 54] It has been previously shown that *S. aureus* helicase will only stimulate primer synthesis when incubated with the cognate primase[6] suggesting a species-specific interaction. The observed difference in *Firmicutes* and *Proteobacteria* primase CTD structures reported here could explain the observed species-specific results of primer synthesis.

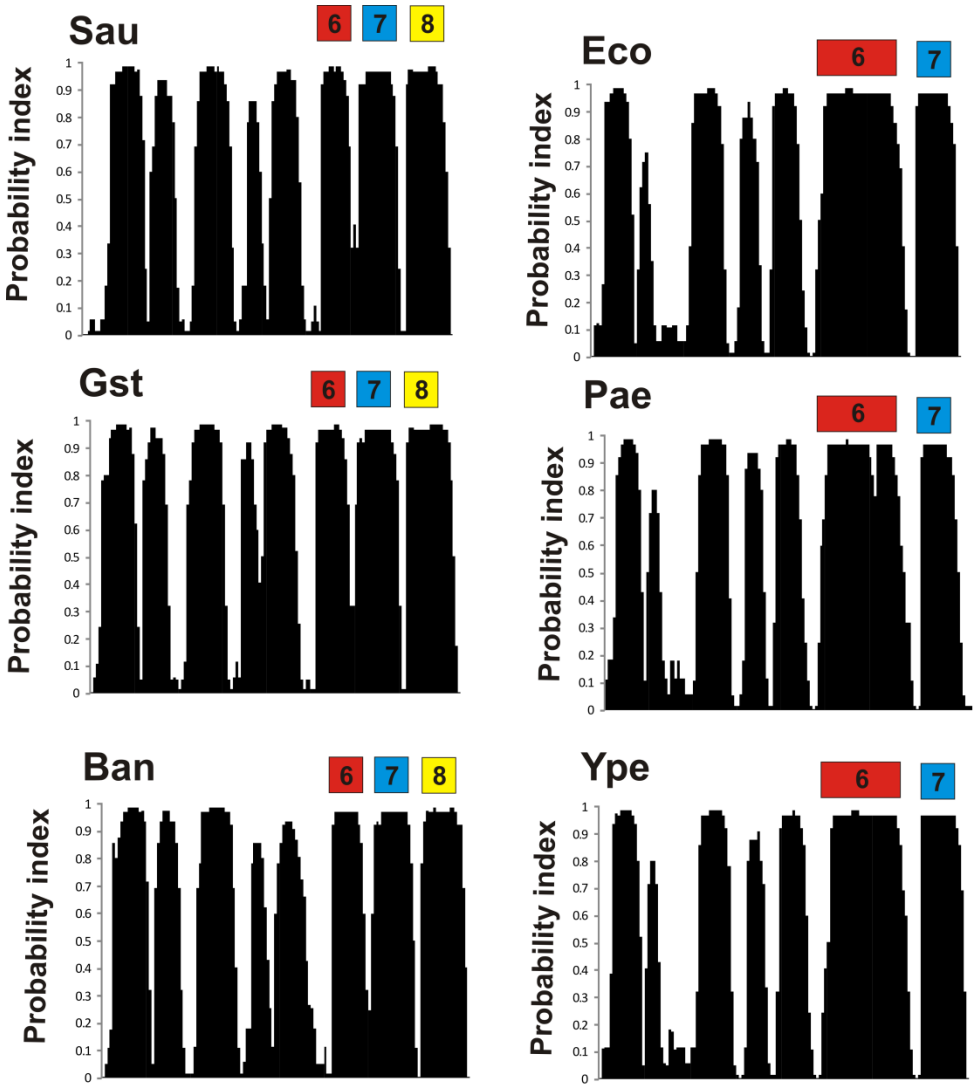


Figure 5.8 Secondary structure prediction of 6 primase CTD sequences. The secondary structures for 6 primase CTD domains were predicted using NetSurfP.[40] The probability index ranges from 0-1 with 0 indicating a loop and 1 indicating a helix. Three *Firmicutes* sequences (*S. aureus Sau*, *G. stearothermophilus Gst*, and *B. anthracis Ban*) all predict 8 helices with a loop region between helix 6 and 7 based on lower probablilty indices forming two sub-domains (C1 and C2). Three *Proteobacteria* sequences (*E. coli Eco*, *P. aeruginosa Pae* and *Y. pestis Ype*) all show 7 helices with a ridged helix 6. The Sau, Eco and Gst structures have all been solved confirming secondary structure prediction.

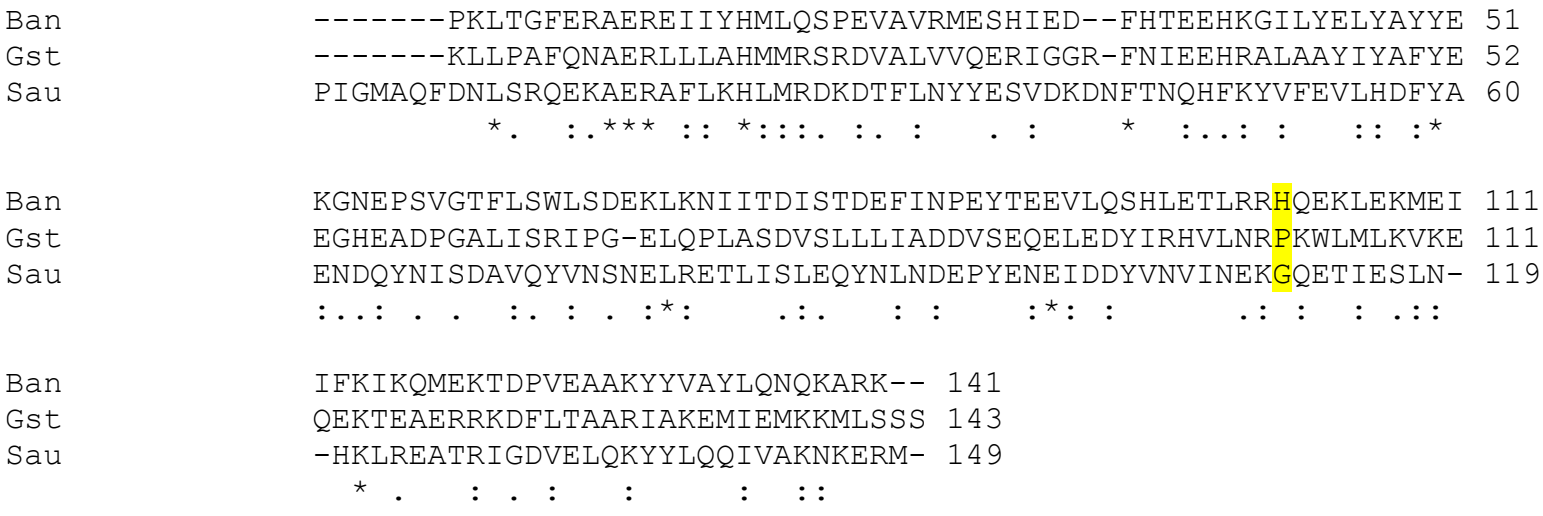


Figure 5.9. Multiple sequence alignment of 3 *Firmicutes* primase CTD sequences. A multiple sequence alignment was completed using ClustlW for 3 *Firmicutes* sequences (Ban, *B. anthracis*, Gst, *G. stearothermophilus*, Sau, *S. aureus*). The residue found in the loop region predicted by NetSurfP is highlighted yellow. In all 3 sequences a loop is predicted between helix 6 and 7. However, the amino acid that forms the loop is not conserved. This suggests a structural and sequence method to regulate primer synthesis through interaction of DnaB helicase.

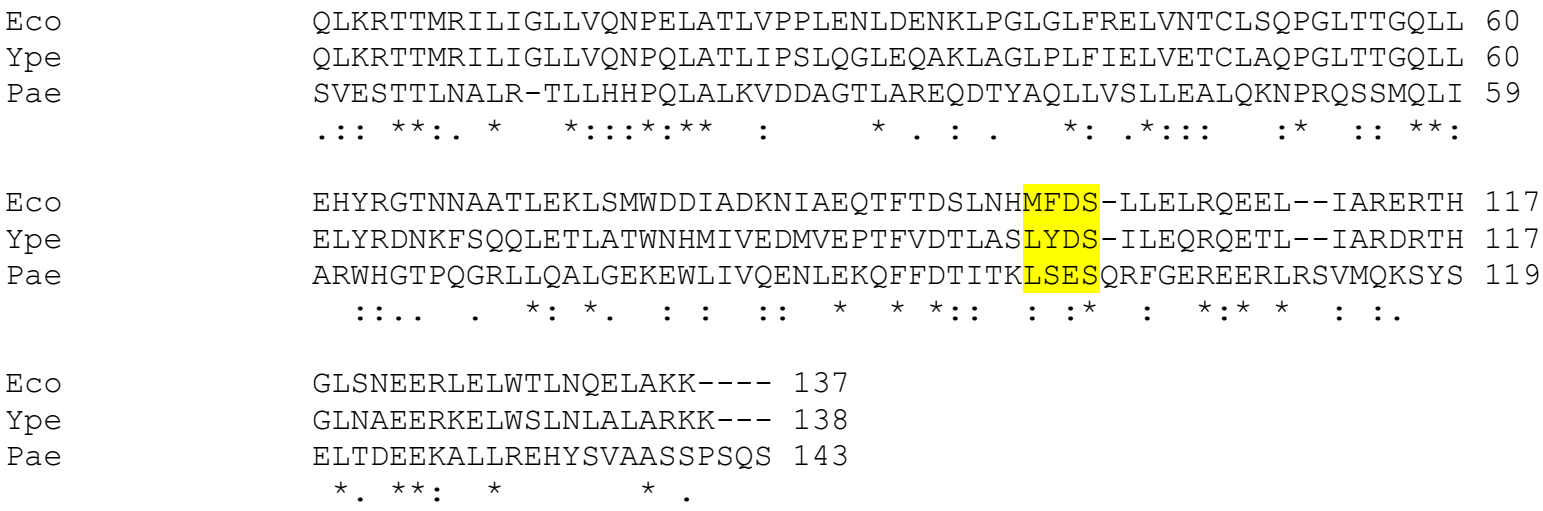


Figure 5.10. Multiple sequence alignment of 3 *Proteobacteria* primase CTD sequences. A multiple sequence alignment was completed using ClustlW for 3 *Proteobacteria* sequences (Eco, *E. coli*, Ype, *Y. pestis*, Pae, *P. aeruginosa*). The residues found in helix 6 that correspond to the residues of the loop region predicted by NetSurfP[40] are highlighted yellow. In all 3 sequences helix 6 is predicted to be ridged. However, the amino acids that form the helix are not highly conserved. This suggests a structural and sequence method to regulate primer synthesis through interaction of DnaB helicase.

5.3.5 Dynamics of the primase C-terminal domain from *S. aureus*. The flexibility of the C2 sub-domain in the *G. stearothermophilus* structure is thought to play an important role in the structural differences with the *E. coli* structure. To determine if the same flexibility is seen in the *S. aureus* structure, the dynamics of the protein was measured using NMR relaxation parameters T_1, T_2 and the relative ratio of NOE enhancement. All T_1, T_2 and NOE values were measured on a per residue basis by eq 5.1 and 5.2 respectively and imported into the program FASTModel Free[20] to measure the Lapri-Szabo order parameters[19] (figure 5.11 D), which show relative local motions in the structure compared to the complete structure.

Generally, S^2 values are near 1.0 for well folded and ridged structures with S^2 values below 0.8 indicative of local motion within a structure. For *S. aureus* primase CTD, the overall model free analysis was very noisy with an average S^2 of 0.83 ± 0.13 for all residues except the his-tag. The large amount of noise in the S^2 data makes identifying significant local motions within the structure challenging based on order parameters alone. The C1 sub-domain (residues 21-124) order-parameters were nosier than the C2 domain (residues 128-163) suggesting more flexibility within the C1 sub-domain. The increased flexibility apparently played a significant role in the lower resolution of the *S. aureus* primase CTD structure due to exchange broadening and a lack of NOE build up, reducing the total number of long range NOEs.

The raw relaxation data provides further support regarding the overall dynamics of the structure. The average T_1 relaxation rate excluding the flexible his-tag residues was 657.7 ± 115.5 ms (figure 5.11A) and the average T_2 relaxation rate was 64.2 ±17.8 ms (Figure 5.8B). The large standard deviations of the relaxation measurements appear

to be caused by a difference in relaxation rates between the two sub-domains of the protein. Individually each sub-domain has an average T_1 of 716 ± 93.5 ms for C1 and 545.2 ± 53.4 ms for C2. The average T_2 for each sub-domain was 55.5 ±7.3 ms for C1 and 78.0 ±12.6 ms for C2. Residue 146 was excluded in these measurements because of the increased local motion of the residues in the loop region between helix 7 and 8.

The difference in the average relaxation times for the two sub-domains and the overall noise associated with the T_1 data suggest the structure is undergoing significant motions. Each sub-domain of the primase CTD is stable and structured as indicated by the average relative ratios of peak intensities between a NOE enhanced and non-enhanced spectra (figure 5.11C). For the C1 sub-domain the average ratio excluding loop regions was 0.96 ±0.13 and the C2 sub-domain was 0.95 ±0.17.

The loop region between helix 6 and 7 (specifically G125) appears to be a pivot point for a change in average relaxation rates. The change in relaxation rates, the lack of distance restraints and the in ability to simultaneously overlay the two sub-domains suggest the two sub-domains act independently of each other on a larger time scale than the model free analysis. The residues of sub-domain C1 fit model 3, which includes both S^2 , a generalized order parameter that reflects the amplitude of internal motions and R_{ex}, which accounts for chemical exchange in T_2 measurements. Proteins that fit model three generally have internal motions on the ms timescale. The observation that sub-domain C1 has a significant R_{ex} contribution accounts for the noise in the S^2 plot (figure 5.11D) and the reduced resolution of the structure. R_{ex} contributions are plotted in figure 5.12. The residues of domain C2 generally fit model 1, which only contributes S^2 order parameter.

If the two sub-domains exhibit independent motion relative to each other, this would result in a separate total correlation times for each sub-domain and contribute to the overall noise observed in the model free analysis. The measured correlation time (τ_m) from the NMR relaxation data for the full protein is 9.8 ns. This is slightly larger than the predicted 8.2 ns based on molecular weight of 19.6 kDa where $\tau_m \approx MW/24000$.[55] The predicted correlation time for the protein based on HYDRONMR[56, 57] is 15.8 ns. HYDRONMR uses the structure of the protein to back calculate the relaxation parameters and predict a correlation time. The predicted correlation time is much larger than predicted based on molecular weight. As described in chapter 2, the molecular weight approximation is for spherical, globular proteins. Having both measured and predicted correlation times larger than the approximation value further suggest internal motion between the two sub-domains (C1, C2). Each sub-domain has a predicted correlation time using HYDRONMR[56, 57] of 10.0 ns for C1 and 3.9 ns for C2. Both predictions are longer than the predicted correlation times based on the molecular weight approximation, 5.5 ns and 1.9 ns for C1 and C2 respectively.

The dynamic nature of the primase CTD structure could play a role in helicase binding. It was shown the loop region of *G. stearothermophilus* becomes more extended upon binding DnaB helicase N-terminal domain.[9] This could also be true for *S. aureus* primase CTD, but further analysis will be needed to confirm this hypothesis. The difference in dynamics between bound and free *G. stearothermophilus* coupled with the phylum specific dependency on the loop region further suggest the primase C-terminal domain is involved in species-specific regulation of DNA replication.

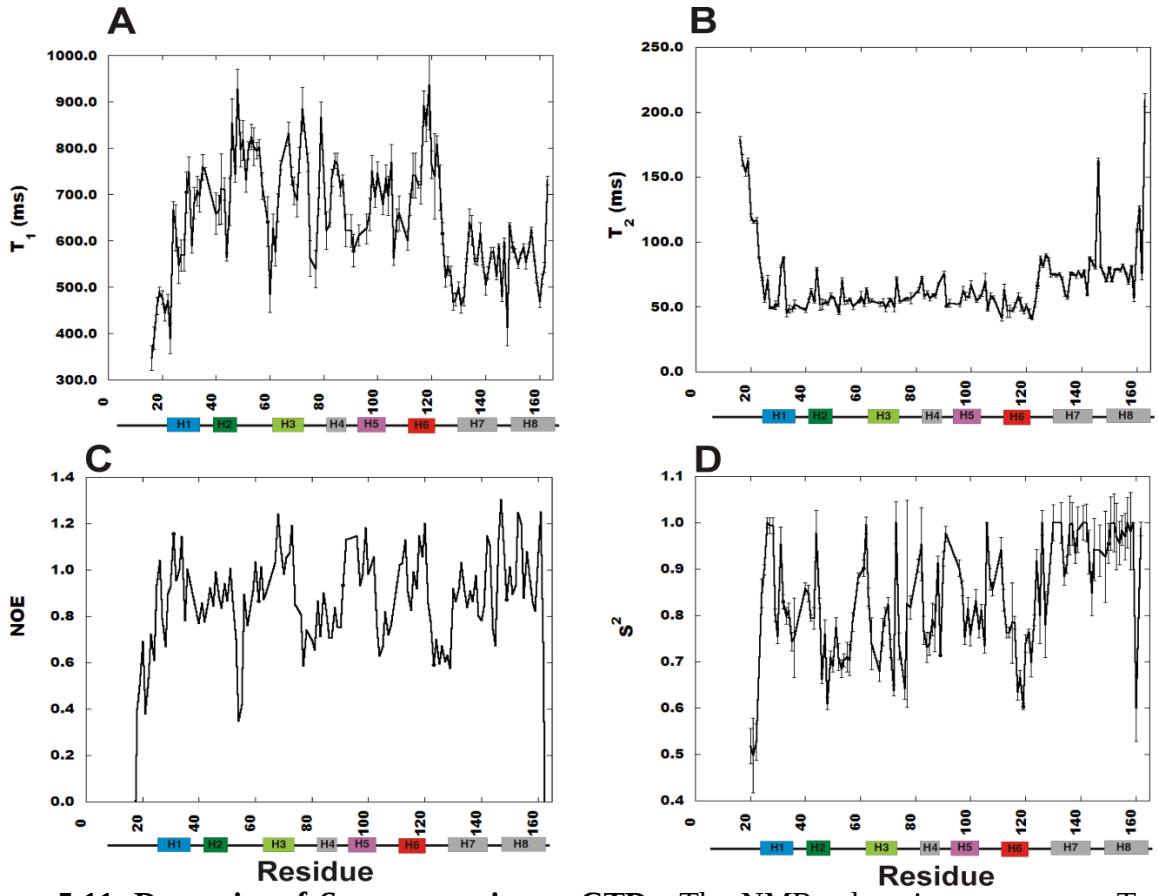


Figure 5.11. Dynamics of *S. aureus* primase CTD. The NMR relaxation parameters T_1 (A) and T_2 (B), NOE enhancements (C) and S^2 order parameters (D) are plotted per residue. The graphs show the relative flexibility between the two sub-domains of *S. aureus* primase CTD. The C2 sub-domain (residues 129-163) has different relaxation rates relative to the N-terminal bundle suggesting dynamic motion between the two sub-domains on a longer time scale than standard Lapri-Szabo Modelfree[19] measurements (> ps-ns).

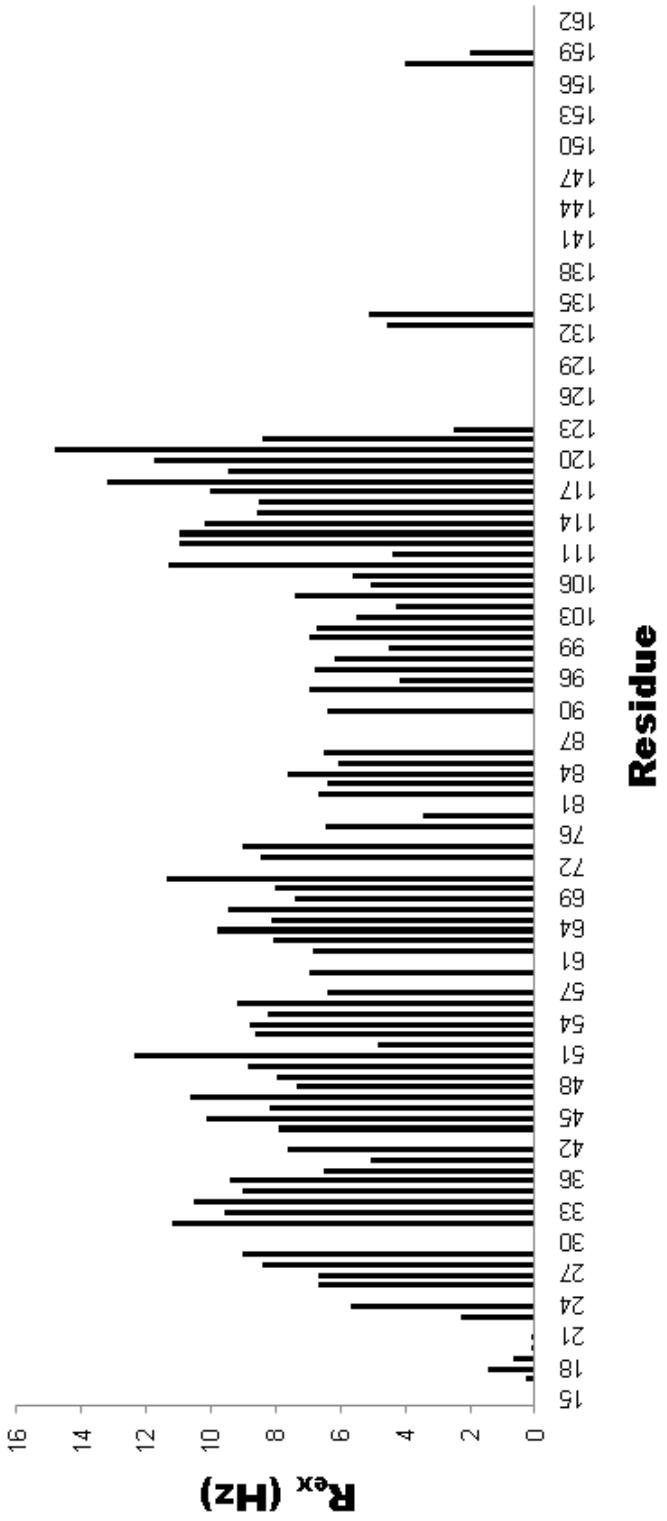

R_{ex} (Hz)
0
2
4
6
8
10
12
14
16
Residue
15
18
21
24
27
30
33
36
42
45
48
51
54
57
61
64
69
72
76
81
84
87
90
96
99
103
106
111
114
117
120
123
126
129
132
135
138
141
144
147
150
153
156
159
162

Figure 5.12. **Contribution of R_{ex} to dynamics of *S. aureus* primase CTD.** The R_{ex} term in model free analysis contributes to chemical exchange due to ms timescale motions of the protein. The majority of the C1 sub-domain has a large value for R_{ex} indicating large degree of flexibility. This increase in chemical exchange caused an increase in overall linewidth leading to large peak overlap, which can account for the lower resolution of the structure.

5.3.6 Identification of binding ligands to *S. aureus* primase CTD. A high-throughput NMR ligand affinity screen of the *S. aureus* primase CTD was completed to identify potential inhibitors of the DnaG - DnaB interaction. A total of 12 compounds (table 5.2) were shown to bind *S. aureus* primase CTD using the 1D ^{1}H NMR screening methods described in this dissertation (see chapters 3 & 4). Two of the ligands, acycloguanosine and mitoxantrone dihydrochloride were previously identified as inhibitors of the DnaG - DnaB interaction in herpes simplex virus.[58, 59] Additionally, the compound myricetin was shown to inhibit the bacterial helicases with an IC_{50} of 10 μM.[60] These three compounds were further analyzed for their binding with primase CTD from *S. aureus*.

A 2D ^{1}H-^{15}N HSQC spectrum was collected for ligand free primase CTD and a bound primase CTD-ligand complex for acycloguanosine, mitoxantrone, and myricetin (figure 5.13). The buffer used for the ligand affinity screen was different from the structural work, but did not significantly change the 2D ^{1}H-^{15}N HSQC spectrum depicted in Figure 5.9A. All three compounds showed primase CTD binding based on chemical shift perturbations in the 2D ^{1}H-^{15}N HSQC spectrum with the addition of the compounds. Acycloguanosine showed the most promising specific interaction based on the magnitude and clustering of chemical shift changes (figure 5.13C). Conversely, myricetin showed a mix of specific and non-specific interactions (figure 5.13B) and mitoxantrone dihydrochloride induced the formation of large molecular weight aggregates (figure 5.13D).

Table 5.2 Ligands identified to bind *S. aureus* primase CTD from a high-throughput NMR ligand affinity screen.

Binding ligand
(±)-a-Lipoamide
L-Histidine (His)
Acycloguanosine
Sodium DL-lactate
3-Aminopropionitrile fumarate salt
Sodium creatine phosphate dibasic tetrahydrate
mitoxantrone dihydrochloride
Chelerythrine chloride
5,5-Diphenylhydantoin
1-Methylimidazole
Didecyldimethylammonium bromide
(±)-Propranolol hydrochloride
Myricetin

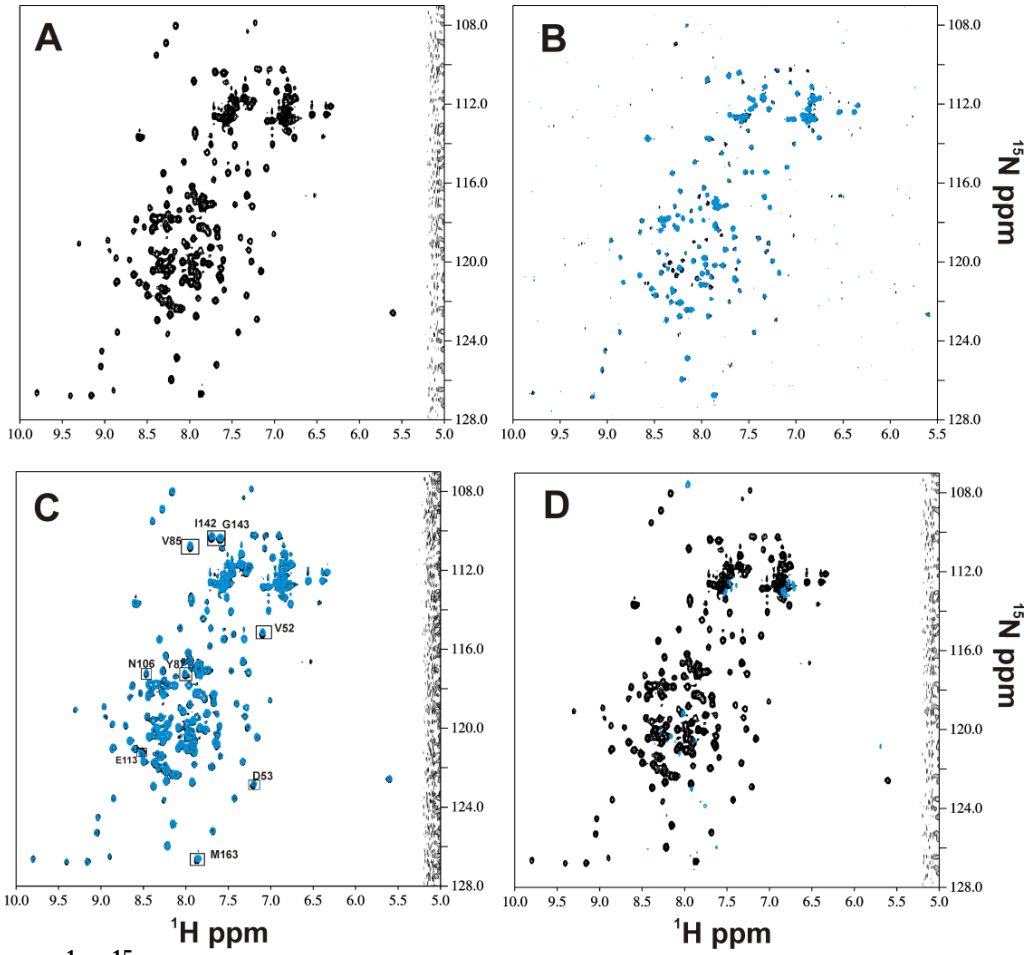


Figure 5.13. 2D ^{1}H-^{15}N HSQC ligand affinity screen for *S. aureus* primase CTD inhibitors. Ligands identified from the 1D ^{1}H NMR line-broadening screen were added to a 100 μM solution of primase CTD to a final concentration of 500 μM (black free primase CTD, blue bound primase CTD). The screening buffer had no effect on the structure or the 2D ^{1}H-^{15}N HSQC spectrum (A). Myricetin showed a mix of specific and non-specific binding to primase CTD indicated by a decrease in peak intensity (B). Acycloguanosine bound specifically to primase CTD (C). Residues corresponding to the acycloguanosine binding site are boxed and labeled. Mitoxantrone dihydrochloride induced large MW aggregates upon binding to primase CTD as indicated by a complete loss of primase CTD signal (D).

The residues in primase CTD showing the largest change upon addition of myricetin (figure 5.13B) were GLY 16, ASP 17, ASP 19, ASP 20, PHE 21, and LEU 24. The corresponding HSQC peaks show a decrease in signal intensity upon binding myricetin, which suggests an exchange broadened non-specific interaction. These residues are primarily found in the his-tag and the extreme N-terminus of the primase CTD structure. In addition to the decrease in intensity of the N-terminus, residues S51, D53, D55, and F57 also showed a significant change in chemical shift as calculated by the weighting equation (eq 5.1 see chapter 3 for discussion).

$$W = \left[\frac{(\Delta NH)^2 + (\frac{\Delta \pm N}{5})^2}{2}\right]^{1/2} \qquad [5.3]$$

The change in chemical shift upon addition of myricetin to primase CTD suggests a specific interaction between the protein and the ligand. If the ligand specifically binds to residues 51, 53, 55, and 57, the non-specific binding of the his-tag can be explained by a transient effect due to the mobility of the his-tag residues and the proximity of the ligand bound to helix 51, 53, 55 and 57.

Acycloguanosine was shown to specifically and significantly interact with primase CTD residues R32, V52, D53, F72, V85, N106, E113, N122, G125, I142, G143, Q154, V156, E161, R161 and M163. These residues exhibited chemical shift changes above 1 standard deviation from the average of all residues (figure 5.13C). Importantly, no decrease in peak intensity was observed, implying a specific interaction.

Mitoxantrone dihydrochloride induced the formation of large molecular weight aggregates upon addition to primase CTD. This is apparent from the complete disappearance of NMR signals in the primase CTD 2D ^{1}H-^{15}N HSQC spectrum, presumably caused by molecular-weight induced peak broadening (figure 5.13D).

5.3.7 Comparison between primase CTD ligand binding site and helicase binding site. The chemical shift differences between free and acycloguanosine bound primase CTD were mapped to the surface of the average water refined structure (figure 5.14A). The largest chemical shift differences were found on the last two helices of the structure. This region has been shown to be required for primase CTD binding to the N-terminal domain of the helicase in *E. coli* and *G. steraothermophilus*.[3, 9, 53, 61] As shown by Bailey *et. al*, [9] both the C1 and C2 sub-domains of primase CTD interact with the N-terminus of helicase. Specifically, the binding of the C2 sub-domain to the N-terminal domain of DnaB helicase is essential for binding the helicase and stimulating primer synthesis, while interaction of the C1 sub-domain with helicase is essential for correct primer synthesis.

Using inference through homology, the binding sites from the *G. stearothermophilus* structure with helicase were color coded onto the *S. aureus* primase CTD structure (figure 5.14B). The residues that undergo the largest chemical shift change upon addition of acycloguanosine are in the same region of the helicase binding sub-domain (C2). There are differences between the two binding sites. Particularly, the acycloguanosine site appears to be on the opposite face of the C2 sub-domain relative to the helicase binding sites. However, the *S. aureus* helicase binding site was only identified by inference through homology with *G. stearothermophilus*. The exact binding site for S. aureus may be different enough to encompass the acycloguanosine binding site. This point highlights the challenges of targeting a large protein-protein interaction sites for drug development.

Conformation of the acycloguanosine binding site and inhibitory activity can be achieved through activity assays showing a decrease in primer synthesis, comparative dynamic studies between free and bound primase CTD in complex with helicase, and the full structure determination of the ligand bound primase CTD structure. These studies are beyond the scope of this work. However, identifying a ligand that appears to bind the C2 sub-domain of *S. aureus* primase CTD suggests a potential mechanism of inhibiting primer-induced helicase activity. Acycloguanosine may be a viable lead compound for a structure-based drug discovery since it may target the essential primase CTD C2 sub-domain mediated DnaG-DnaB interaction. Pending the conformation studies, the results described in this chapter suggest the identification of a new antibiotic drug target; the interaction between primase CTD and helicase N-terminal domain.

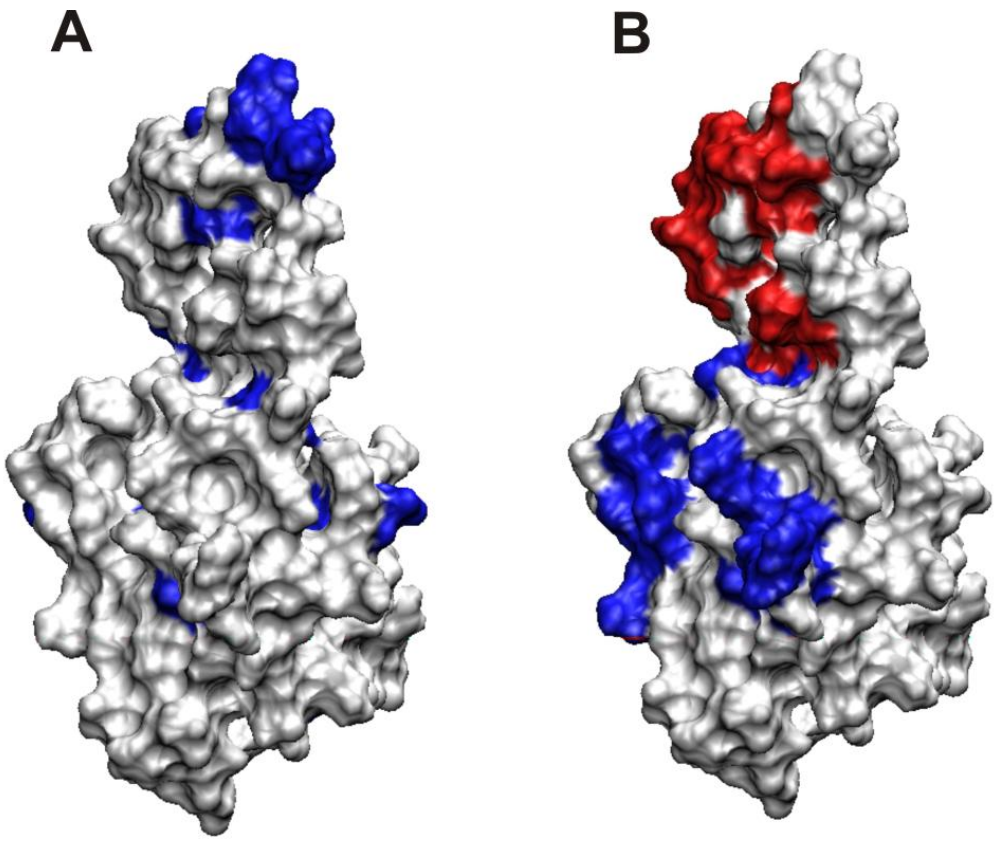


Figure 5.14. Comparison between ligand binding and helicase binding sites. (A) The residues that illustrate the largest chemical shift difference upon addition of acycloguanosine are colored blue on the *S. aureus* primase CTD NMR solution structure. (B) The residues that interact with the N-terminal domain of DnaB helicase are colored based on sub-domain interaction (red C2, blue C1). The helicase interactions are based on homology transfer between the *G. stearothermophilus* primase CTD structure interacting with the N-terminal domain of *G. stearothermophilus* DnaB helicase.[9]

5.4 REFERENCES

1. Frick, D. N.; Richardson, C. C., DNA primases. *Annu Rev Biochem* **2001,** 70, 39-80.
2. Tougu, K.; Marians, K. J., The interaction between helicase and primase sets the replication fork clock. *J Biol Chem* **1996,** 271, (35), 21398-405.
3. Bird, L. E.; Pan, H.; Soultanas, P.; Wigley, D. B., Mapping Protein-Protein Interactions within a Stable Complex of DNA Primase and DnaB Helicase from Bacillus stearothermophilusâ€ *Biochemistry* **1999,** 39, (1), 171-182.
4. Syson, K.; Thirlway, J.; Hounslow, A. M.; Soultanas, P.; Waltho, J. P., Solution structure of the helicase-interaction domain of the primase DnaG: a model for helicase activation. *Structure* **2005,** 13, (4), 609-16.
5. Pan, H.; Wigley, D. B., Structure of the zinc-binding domain of Bacillus stearothermophilus DNA primase. *Structure* **2000,** 8, (3), 231-9.
6. Koepsell, S. A.; Larson, M. A.; Griep, M. A.; Hinrichs, S. H., Staphylococcus aureus helicase but not Escherichia coli helicase stimulates S. aureus primase activity and maintains initiation specificity. *J Bacteriol* **2006,** 188, (13), 4673-80.
7. Su, X. C.; Schaeffer, P. M.; Loscha, K. V.; Gan, P. H.; Dixon, N. E.; Otting, G., Monomeric solution structure of the helicase-binding domain of Escherichia coli DnaG primase. *Febs J* **2006,** 273, (21), 4997-5009.
8. Oakley, A. J.; Loscha, K. V.; Schaeffer, P. M.; Liepinsh, E.; Pintacuda, G.; Wilce, M. C.; Otting, G.; Dixon, N. E., Crystal and solution structures of the helicase-binding domain of Escherichia coli primase. *J Biol Chem* **2005,** 280, (12), 11495-504.

9. Bailey, S.; Eliason, W. K.; Steitz, T. A., Structure of hexameric DnaB helicase and its complex with a domain of DnaG primase. *Science* **2007,** 318, (5849), 459-63.

10. Costantini, S.; Colonna, G.; Facchiano, A. M., Amino acid propensities for secondary structures are influenced by the protein structural class. *Biochem Biophys Res Commun* **2006,** 342, (2), 441-51.

11. Chou, P. Y.; Fasman, G. D., Prediction of protein conformation. *Biochemistry* **1974,** 13, (2), 222-45.

12. Larson, E., Community factors in the development of antibiotic resistance. *Annu Rev Public Health* **2007,** 28, 435-47.

13. Mercier, K. A.; Germer, K.; Powers, R., Design and Characterization of a Functional Library for NMR Screening against Novel Protein Targets. *Combinatorial Chemistry and High Throughput Screening* **2006,** 9, (7), 515-534.

14. Sattler, M.; Schleucher, J.; Griesinger, C., Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999,** 34, (2), 93-158.

15. Hwang, T.-L.; Mori, H.; Shaka, A.; van Zijl, P., Application of Phase-Modulated CLEAN Chemical EXchange Spectroscopy (CLEANEX-PM) to Detect Waterâ^'Protein Proton Exchange and Intermolecular NOEs. **1997,** 119, (26), 6203-6204.

16. Farrow, N. A.; Muhandiram, R.; Singer, A. U.; Pascal, S. M.; Kay, C. M.; Gish, G.; Shoelson, S. E.; Pawson, T.; Forman-Kay, J. D.; Kay, L. E., Backbone

dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by 15N NMR relaxation. *Biochemistry* **1994,** 33, (19), 5984-6003.

17. Kay, L. E.; Torchia, D. A.; Bax, A., Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* **1989,** 28, (23), 8972-9.

18. Mandel, A. M.; Akke, M.; Palmer, A. G., 3rd, Backbone dynamics of Escherichia coli ribonuclease HI: correlations with structure and function in an active enzyme. *J Mol Biol* **1995,** 246, (1), 144-63.

19. Lipari, G.; Szabo, A., Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. **1982,** 104, (17), 4546-4559.

20. Cole, R.; Loria, J. P., FAST-Modelfree: a program for rapid automated analysis of solution NMR spin-relaxation data. *J Biomol NMR* **2003,** 26, (3), 203-13.

21. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **1995,** 6, (3), 277-93.

22. Garrett, D. S.; Powers, R.; Groenenborn, A. M.; Clore, G. M., A common sense approach to peak picking in two-, three-, and four-dimensional spectra using automatic computer analysis of contour diagrams. *Journal of Magnetic Resonance (1969-1992)* **1991,** 95, (1), 214-20.

23. Fogh, R.; Ionides, J.; Ulrich, E.; Boucher, W.; Vranken, W.; Linge, J. P.; Habeck, M.; Rieping, W.; Bhat, T. N.; Westbrook, J.; Henrick, K.; Gilliland, G.; Berman, H.; Thornton, J.; Nilges, M.; Markley, J.; Laue, E., The CCPN project: an interim

report on a data model for the NMR community. *Nat Struct Biol* **2002,** 9, (6), 416-8.

24. Mercier, K. A.; Baran, M.; Ramanathan, V.; Revesz, P.; Xiao, R.; Montelione, G. T.; Powers, R., FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* **2006,** 128, (47), 15292-9.
25. Huang, Y. J.; Tejero, R.; Powers, R.; Montelione, G. T., A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins* **2006,** 62, (3), 587-603.
26. 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* **2009,** 44, (4), 213-23.
27. Vuister, G. W.; Wang, A. C.; Bax, A., Measurement of three-bond nitrogen-carbon J couplings in proteins uniformly enriched in nitrogen-15 and carbon-13. *J Am Chem Soc* **1993,** 115, (12), 5334-5335.
28. Nilges, M.; Clore, G. M.; Gronenborn, A. M., Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Lett* **1988,** 229, (2), 317-24.
29. Clore, G. M.; Appella, E.; Yamada, M.; Matsushima, K.; Gronenborn, A. M., Three-dimensional structure of interleukin 8 in solution. *Biochemistry* **1990,** 29, (7), 1689-96.
30. Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Clore, G. M., The Xplor-NIH NMR molecular structure determination package. *J Magn Reson* **2003,** 160, (1), 65-73.

31. Garrett, D. S.; Kuszewski, J.; Hancock, T. J.; Lodi, P. J.; Vuister, G. W.; Gronenborn, A. M.; Clore, G. M., The impact of direct refinement against three-bond HN-C alpha H coupling constants on protein structure determination by NMR. *J Magn Reson B* **1994,** 104, (1), 99-103.

32. Kuszewski, J.; Qin, J.; Gronenborn, A. M.; Clore, G. M., The impact of direct refinement against 13C alpha and 13C beta chemical shifts on protein structure determination by NMR. *J Magn Reson B* **1995,** 106, (1), 92-6.

33. Kuszewski, J.; Gronenborn, A. M.; Clore, G. M., Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Sci* **1996,** 5, (6), 1067-80.

34. Kuszewski, J.; Gronenborn, A. M.; Clore, G. M., Improvements and extensions in the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids. *J Magn Reson* **1997,** 125, (1), 171-7.

35. Kuszewski, J.; Clore, G. M., Sources of and solutions to problems in the refinement of protein NMR structures against torsion angle potentials of mean force. *J Magn Reson* **2000,** 146, (2), 249-54.

36. Linge, J. P.; Nilges, M., Influence of non-bonded parameters on the quality of NMR structures: a new force field for NMR structure calculation. *J Biomol NMR* **1999,** 13, (1), 51-9.

37. Nederveen, A. J.; Doreleijers, J. F.; Vranken, W.; Miller, Z.; Spronk, C. A.; Nabuurs, S. B.; Guntert, P.; Livny, M.; Markley, J. L.; Nilges, M.; Ulrich, E. L.; Kaptein, R.; Bonvin, A. M., RECOORD: a recalculated coordinate database of

500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins* **2005,** 59, (4), 662-72.

38. Powers, R.; Mirkovic, N.; Goldsmith-Fischman, S.; Acton, T. B.; Chiang, Y.; Huang, Y. J.; Ma, L.; Rajan, P. K.; Cort, J. R.; Kennedy, M. A.; Liu, J.; Rost, B.; Honig, B.; Murray, D.; Montelione, G. T., Solution structure of Archaeglobus fulgidis peptidyl-tRNA hydrolase (Pth2) provides evidence for an extensive conserved family of Pth2 enzymes in archea, bacteria, and eukaryotes. *Protein Sci* **2005,** 14, (11), 2849-61.

39. Markley, J. L.; Ulrich, E. L.; Berman, H. M.; Henrick, K.; Nakamura, H.; Akutsu, H., BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J Biomol NMR* **2008,** 40, (3), 153-5.

40. Petersen, B.; Petersen, T. N.; Andersen, P.; Nielsen, M.; Lundegaard, C., A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol* **2009,** 9, 51.

41. Halouska, S.; Zhou, Y.; Becker, D. F.; Powers, R., Solution structure of the Pseudomonas putida protein PpPutA45 and its DNA complex. *Proteins* **2009,** 75, (1), 12-27.

42. Brunger, A. T., Version 1.2 of the Crystallography and NMR system. *Nat Protoc* **2007,** 2, (11), 2728-33.

43. Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L., Crystallography & NMR system: A

new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* **1998,** 54, (Pt 5), 905-21.

44. Clore, G. M.; Nilges, M.; Sukumaran, D. K.; Bruenger, A. T.; Karplus, M.; Gronenborn, A. M., The three-dimensional structure of a1-purothionin in solution: combined use of nuclear magnetic resonance, distance geometry and restrained molecular dynamics. *EMBO Journal* **1986,** 5, (10), 2729-35.

45. Nilges, M.; Gronenborn, A. M.; Bruenger, A. T.; Clore, G. M., Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Engineering* **1988,** 2, (1), 27-38.

46. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem* **1983,** 4, (2), 187-217.

47. Nilges, M.; Gronenborn, A. M.; Brunger, A. T.; Clore, G. M., Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Eng* **1988,** 2, (1), 27-38.

48. Deshpande, N.; Addess, K. J.; Bluhm, W. F.; Merino-Ott, J. C.; Townsend-Merino, W.; Zhang, Q.; Knezevich, C.; Xie, L.; Chen, L.; Feng, Z.; Green, R. K.; Flippen-Anderson, J. L.; Westbrook, J.; Berman, H. M.; Bourne, P. E., The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res* **2005,** 33, (Database issue), D233-7.

49. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *J Mol Graph* **1996,** 14, (1), 33-8, 27-8.

50. Holm, L.; Park, J., DaliLite workbench for protein structure comparison. *Bioinformatics* **2000,** 16, (6), 566-7.

51. Holm, L.; Sander, C., Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res* **1997,** 25, (1), 231-4.

52. Holm, L.; Sander, C., The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res* **1996,** 24, (1), 206-9.

53. Tougu, K.; Marians, K. J., The extreme C terminus of primase is required for interaction with DnaB at the replication fork. *J Biol Chem* **1996,** 271, (35), 21391-7.

54. Hiasa, H.; Marians, K. J., Initiation of bidirectional replication at the chromosomal origin is directed by the interaction between helicase and primase. *J Biol Chem* **1999,** 274, (38), 27244-8.

55. Cantor, C. R.; Schimmel, P. R., *Biophysical Chemistry Part II: Techniques for the study of biological structure and function*. W. H. Freeman and Co.: San Francisco, 1980; p 461.

56. Garcia de la Torre, J.; Huertas, M. L.; Carrasco, B., HYDRONMR: prediction of NMR relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. *J Magn Reson* **2000,** 147, (1), 138-46.

57. Bernado, P.; Garcia de la Torre, J.; Pons, M., Interpretation of 15N NMR relaxation data of globular proteins using hydrodynamic calculations with HYDRONMR. *J Biomol NMR* **2002,** 23, (2), 139-50.

58. Crute, J. J.; Grygon, C. A.; Hargrave, K. D.; Simoneau, B.; Faucher, A. M.; Bolger, G.; Kibler, P.; Liuzzi, M.; Cordingley, M. G., Herpes simplex virus helicase-primase inhibitors are active in animal models of human disease. *Nat Med* **2002,** 8, (4), 386-91.

59. Kleymann, G.; Fischer, R.; Betz, U. A.; Hendrix, M.; Bender, W.; Schneider, U.; Handke, G.; Eckenberg, P.; Hewlett, G.; Pevzner, V.; Baumeister, J.; Weber, O.; Henninger, K.; Keldenich, J.; Jensen, A.; Kolb, J.; Bach, U.; Popp, A.; Maben, J.; Frappa, I.; Haebich, D.; Lockhoff, O.; Rubsamen-Waigmann, H., New helicase-primase inhibitors as drug candidates for the treatment of herpes simplex disease. *Nat Med* **2002,** 8, (4), 392-8.

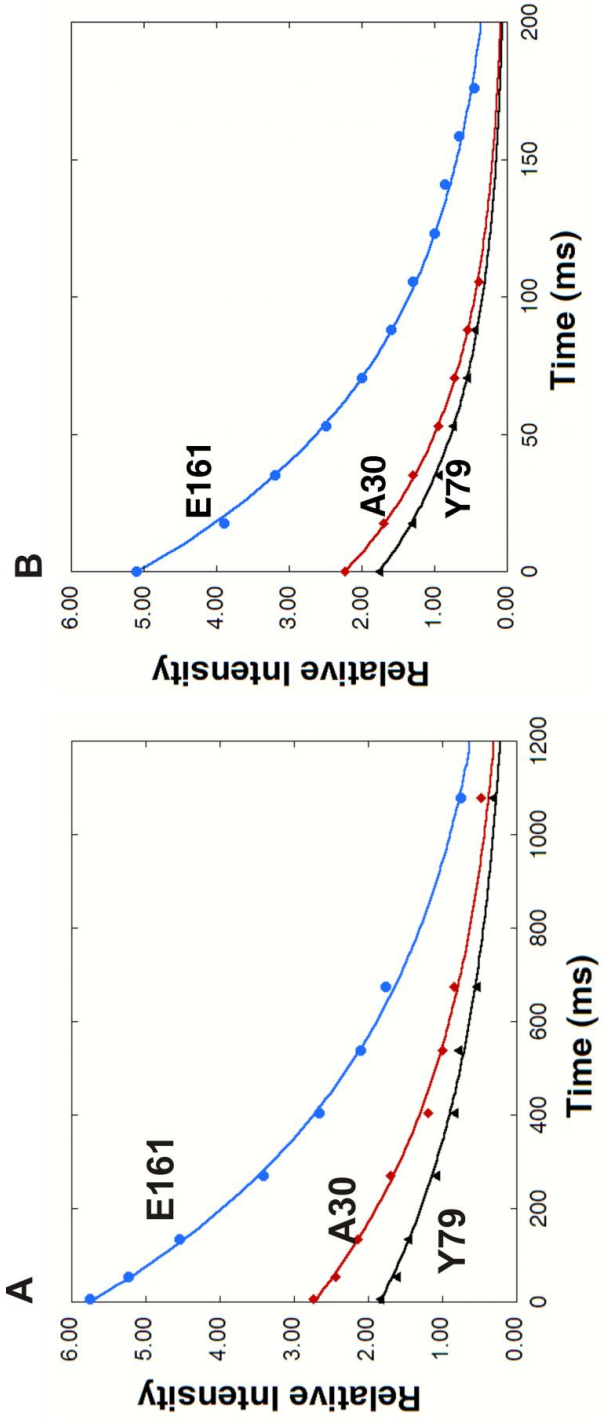
60. Griep, M. A.; Blood, S.; Larson, M. A.; Koepsell, S. A.; Hinrichs, S. H., Myricetin inhibits Escherichia coli DnaB helicase but not primase. *Bioorg Med Chem* **2007,** 15, (22), 7203-8.

61. Thirlway, J.; Soultanas, P., In the Bacillus stearothermophilus DnaB-DnaG complex, the activities of the two proteins are modulated by distinct but overlapping networks of residues. *J Bacteriol* **2006,** 188, (4), 1534-9.

Appendix 5A: Sequence of the *S. aureus* primase C-terminal domain used in these studies. The CTD of DnaG primase is approximately 17.2kDa protein (without his-tag, 19.6kDa with his tag). The glycine residue (G125) that is structurally similar to P543 in the *G. stearothermophilus* structure is highlighted. The sequence has an additional N-terminal his-tag added for purification shown as lower case.

>S.aureus primaseCTD
mghnhnhnhn hnhnggdddd FDNLSRQEKA ERAFLKHLMR
DKDTFLNYYE SVDKDNFTNQ HFKYVFEVLH DFYAENDQYN
ISDAVQYVNS NELRETLISL EQYNLNDEPY ENEIDDYVNV
INEKGQETIE SLNHKLREAT RIGDVELQKY YLQQIVAKNK
ERM

Appendix 5B. Relative relaxation parameters for the *S. aureus* primase CTD structure. Overall, on a per residue basis the T_1 and T_2 values were measured to a high degree of accuracy using eq 5.1 and 5.2 based on the fit quality. (A) T_1 (B) T_2.



CHAPTER 6:

BACTERIAL PROTEIN STRUCTURES REVEAL PHYLUM DEPENDENT DIVERGENCE

6.1 INTRODUCTION

As highlighted in Chapter 5, selecting the best model protein for a biological system can be challenging if limited to sequence and structure information alone. The differences between the primase CTD structures suggest a third constraint, evolution, for selecting a correct model protein. Quantifiable models of protein evolution are useful for developing robust tools to identify suitable drug-binding sites, to predict increases in susceptibility to a human genetic disease, and to study organism niches. Some of the strongest arguments in favor of evolution draw from studies on protein sequence homology.[1] Multiple sequence alignments are routinely used to highlight sequence similarity and variability between organisms and create phylogenetic relationships.[2, 3] Protein evolution is a direct result from changes to the protein's gene sequence, which are selected and modulated by a number of factors including structure.[4, 5]

What is the impact on protein structure as its sequence undergoes genetic drift? Maintaining the correct protein fold is fundamental to preserving its function,[6] but evolving the sequence would also be expected to result in structural changes.[7, 8] The resulting paradox is that sequence determines a protein's structure, but the structure is relatively invariant over a large range of sequences. This paradox is highlighted by the tremendous difference between the number of known protein structures versus protein folds.[9] Even though the Protein Data Bank (PDB)[10] contains 67,529 protein structures as

of August, 2010, there are only 1,110 unique topologies and 1,195 unique folds in the CATH[11] and SCOP[12] structure classification databases, respectively. The significant reduction in the number of protein folds relative to the number of protein sequences implies a strong correlation between structure and function.

While the explicit reason for the reduction in fold space remains unclear, some have suggested that protein fold space may be more appropriately described as a continuum instead of a collection of discreet folds.[13] In this manner, a protein fold should be considered as being plastic, where sequence changes are accommodated by local perturbations in the structure while maintaining the general characteristics of a particular fold.[14-16] Correspondingly, the genetic drift in a protein's sequence may imply a similar gradual divergence in structure instead of a sudden dramatic transition to a new fold. If this perspective is accurate, then a comparative analysis of homologous proteins should identify correlated rates of structure and sequence divergence. Previous studies have examined structure similarities between homologous proteins, but did not evaluate if structure divergence is correlated with phylogeny.[14-16] In this chapter, I expand on this previous work by quantifying a maximum structure/sequence similarity between the two bacterial phyla, *Proteobacteria* and *Firmicutes*. I will also discuss the viability of phylogeny as a suitable constraint for selecting a homology model by showing certain protein folds are more sensitive than others to changes in sequence.

6.2 EXPERIMENTAL

6.2.1 Cluster of Orthologous Groups (COG) assignment of the Protein Data Bank (PDB). Assignment of each bacterial protein in the Protein Data Bank (PDB) to a

COG number in the clusters of orthologous groups[17] database required downloading the complete sequence lists from both databases and running a pairwise Basic Local Alignment Search Tool (BLAST) comparison. The pairwise protein BLAST search was run using the Protein Mapping and Comparison Tool (PROMPT v. 0.9.2)[18] that allowed for large pairwise BLAST searching and reported the best match between the two databases. The BLAST search was run using the BLOSUM62 matrix with a gap penalty of 11, gap extension penalty of 1, a word size of 5, and a BLAST expectation threshold (E-value) of 10^{-9}. This E-value was used to unambiguously match genes in the COG database with proteins in the PDB. All PDB-to-COG matches were reported and stored in the PROFESS (Protein Function, Evolution, Structure, and Sequence) database (http://cse.unl.edu/~profess/).[19]

After matching structures to their representative COG each PDB entry was matched with its source organism and phylum. The data set was then filtered according to the number of unique organisms. Specifically, only those COGs with structures from two or more different source organisms in both *Proteobacteria* and *Firmicutes* were analyzed further.

6.2.2 Pairwise structure comparison. The pairwise structure comparison program DaliLite v. 2.4.2[20] was installed on our 16-node Dual Athlon AMD 2.13 GHz with 1 GB of RAM Beowulf cluster running CentrOS 4.4 Linux with a 2.25 TB RAID array. A C-shell script matches the PDB files from each *Proteobacteria-Proteobacteria* comparison (-/-), *Firmicutes-Firmicutes* comparison (+/+) and *Proteobacteria-Firmicutes* comparison (-/+) and then submits the job to the program DaliLite. Each structural comparison took approximately 2-10 min, depending on the size and relative similarity of

structures. The total time to run all 63,504 comparisons was approximately 7 weeks.

The shell script extracts all structural comparison information reported by DaliLite (comparison files, rmsd, %Sequence ID, Z-score) on a per chain basis. A single PDB file may contain multiple protein chains, where each chain may have a separate COG assignment. All structure information is stored in the PROFESS database, which is parsed to find the largest Z-score for each pairwise structure comparison. The largest Z-score represents the best structure comparison for a pair of proteins and ensures the correct PDB chains were used for the analysis and the correct COG assignments were made. All best matches from each COG were used to calculate the Fractional Structure Similarity score (FSS) described by eq 6.1.

$$FSS = \frac{Z_{AB}}{max(Z_{AA}, Z_{BB})}, \quad [6.1]$$

where Z_{AB} was the Z-score for comparing proteins A and B, Z_{AA} was the Z-score when protein A was compared to itself and Z_{BB} was the Z-score when protein B was compared to itself. Thus, Z_{AA} and Z_{BB} represent the Z-score that can be achieved for perfect similarity.

6.2.3 Manual filtering and data analysis. Manual refinement of the dataset included verification of each PDB assignment to a COG and filtering out redundantly solved structures from the same organism. When multiple structures were reported from the same organism (or organism with synonymous name), the structure that gave the largest Dali Z-score within the COG was kept while remaining structures were discarded from the analysis. This confirmed a single best PDB-to-COG match for each organism. Manual refinement was accomplished by opening all PDB IDs within a COG and checking biological information against the PDB (http://www.rcsb.org/pdb/home), COG

(http://www.ncbi.nlm.nih.gov/COG/) and the NCBI (http://www.ncbi.nlm.nih.gov/) web servers. Consistency in functional and structural assignment within a COG coupled with very low E-values between COG and PDB confirmed the best PDB-to-COG match was made. Additionally, manual refinement was used to verify uniform sample conditions (i.e., the same ligand bound to all proteins within a COG or all proteins correspond to wild-type sequences) for cases of redundantly solved structures.

6.2.4 Structure based phylogenetic trees. In addition to pairwise alignment, all the protein structures from each COG were simultaneously aligned using the multiple structure alignment program MAMMOTH-multi (http://ub.cbm.uam.es/mammoth/mult/).[21] The resulting aligned structures and the structure-based sequence alignment was used with in-house software to calculate an all-versus-all matrix of per-residue Cα distances. Standard boot-strapping techniques were then applied to the all-versus-all matrix of per-residue Cα distances to generate 100 distance-matrix tables. Columns of structure-based sequence alignments with the corresponding Cα distances were randomly selected until the total number of columns in the original sequence alignment was reached. The resulting set of Cα distances were then used to calculate a root mean square deviation (rmsd) between each pair of structures in the matrix. The 100 distance-matrix tables were imported into PHYLIP 3.68[22] to generate a consensus phylogenetic tree and bootstrap confidence levels.

Each set of 100 bootstrapped distance matrices were analyzed by the Fitch-Margoliash method implemented in PHYLIP. Each matrix was jumbled with 100 replicates using 37 as the random number generator seed. This resulted in 10000 unique and random distance matrices for each COG. The best tree was identified with the

program Consense implemented in PHYLIP using the extended majority rule conservation. Since the bootstrapped trees do not show distance relationship, the original distance matrix generated by MAMMOTH-multi was used to generate a distance based phylogenetic tree. Each original distance matrix was jumbled with 100 replicates using 37 as the random number seed. The distance based phylogenetic tree was drawn using the program Drawtree implemented in PHYLIP.

Representative distance based phylogenetic trees are shown in (figure 6.4). Each tree was visually inspected and compared with the DaliLite analysis using the bootstrap values to determine if a tree fit the split, starburst, or split +1. A "split" means the *Firmicutes* and *Proteobacteria* proteins were strongly separated from one another, "Starburst" means there was little to no evidence for a split according to phyla, and "Split +1" means there was strong evidence for a split according to phyla with the exception of one protein

6.2.5 Measuring functional similarity within a COG. Each protein in our dataset was annotated with the corresponding Gene Ontology[23] identification number(s) found in the PDB. By definition, a strong consensus requires each protein to share the same set of GO terms. Instead, a weak consensus set of GO terms was generated for each COG, where only a majority of proteins are required to share the same GO term. A distance was measured between the weak consensus set and the set of GO terms assigned to each individual protein. An average, normalized distance is reported for each COG, where a score of 1 indicates an identical functional classification and a score of 0 indicates a lack of functional similarity. The distance between each protein's GO terms and the consensus GO term set was measured as follows:

$$GO_{sim} = 1 - \left(\frac{\sum(|WC \cup Go_i|) - |WC \cap Go_i|}{|WC \cup Go_i|}\right) \quad [6.2]$$

where GO_{sim} is the normalized GO functional similarity score, WC denotes the weak consensus set of GO terms for the COG, and GO_i denotes the set of GO terms set for each protein in the COG.

6.3 RESULTS

6.3.1 Creating the COG structure families. Current functional annotation tools available in the PDB include the Gene Ontology (GO)[23] and Enzyme Classification (EC).[24] Unfortunately, due to potential for convergence of function, these annotation tools are not useful for the study of homologous structures. To accurately observe phylum dependent structure divergence of proteins, it is important to construct a dataset of functionally similar orthologs. Among the 20 resources for structural classification of proteins, the clusters of orthologous groups (COGs) scheme is the only one that attempts to identify orthology[25] while providing moderate functional information. Therefore, each sequence and structure in the PDB was annotated with one COG number. Additionally, each protein was annotated with GO numbers and the relative functional similarity for each COG was measured (table 6.1).

Table 6.1. COG structure families. [a]COG Structure Families have two or more represented structures from among the *Firmicutes* and two or more from among the *Proteobacteria*. [b]Functonal similarities are measured by overlapping consensus GO terms (eq 6.2) [c]"Split" means the *Firmicutes* and *Proteobacteria* proteins were strongly separated from one another, "Starburst" means there was little to no evidence for a split according to phyla, and "Split +1" means there was strong evidence for a split according to phyla with the exception of one protein. The relative functional similarity of a COG is reported by measuring an average distance between a weak consensus set of Gene Ontology (GO) annotations and the set of all GO annotations for each protein within a COG. Perfect functional similarity is reported as a 1, while no similarity is reported as a 0. See appendix 6B for a list of the PDB files associated with each COG. *No CATH value available for reported structures, CATH values were predicted using a sequence based search in the CATH database where the best match is reported.

COG[a]	COG Function Annotation	COG Function Similarity[b]	Phylogenetic Structure Tree[c]	CATH
28	Thiamine pyrophosphate requiring enzymes	0.59	Split	3.40.50.970
39	Malate/lactate dehydrogenases	0.8	Split	3.40.50.720
394	Protein-tyrosine-phosphatase	0.61	Split	3.40.50.270
604	NADPH:quinone reductase and related Zn-dependent oxidoreductases	0.88	Split	3.40.50.720
605	Superoxide dismutase	0.76	Split	3.20.20.80*
742	N6-adenine-specific methylase	0.73	Split	3.40.50.150*
813	Purine-nucleoside phosphorylase	0.87	Split	3.40.50.1580
1012	NAD-dependent aldehyde dehydrogenases	0.58	Split	3.40.309.10
1075	Predicted acetyltransferases and hydrolases with the alpha/beta hydrolase fold	0.7	Split	3.40.50.1820
1607	Acyl-CoA hydrolase	0.87	Split	3.40.0.1820*
1940	Transcriptional regulator/sugar kinase	0.31	Split	3.30.420.40
2124	Cytochrome P450	0.8	Split	1.10.630.10
2188	Transcriptional regulators	0.89	Split	3.40.1410.10
446	Uncharacterized NAD (FAD) - dependent dehydrogenases	0.85	Split	3.30.390.30
1057	Nicotinic acid mononucleotide adenylyltransferase	0.95	Split	3.40.50.620
242	N-formylmethionyl-tRNA deformylase	0.87	Split +1	3.90.45.10
1052	Lactate dehydrogenase and related dehydrogenases	0.89	Split +1	3.40.50.720
2141	Coenzyme F420-dependent N5,N10-methylene tetrahydromethanopterin	0.76	Split +1	3.20.20.30

	reductase and related flavin-dependent oxidoreductases			
3832	Uncharacterized conserved protein	1	Split +1	3.30.530.20
110	Acetyltransferase (isoleucine patch superfamily)	0.56	Starburst	2.160.10.10
171	NAD synthase	0.85	Starburst	3.40.50.620
251	Putative translation initiation inhibitor, yjgF family	0	Starburst	3.30.1330.40
346	Lactoylglutathione lyase and related lyases	0.11	Starburst	3.10.180.10
366	Glycosidases	0.51	Starburst	2.60.40.1180
454	Histone acetyltransferase HPA2 and related acetyltransferases	0.83	Starburst	3.40.630.30
491	Zn-dependent hydrolases, including glyoxylases	0.5	Starburst	3.60.15.10
500	SAM-dependent methyltransferases	0.59	Starburst	3.40.50.150
526	Thiol-disulfide isomerase and thioredoxins	0.96	Starburst	3.40.30.10
590	Cytosine/adenosine deaminases	0.7	Starburst	3.40.140.10
637	Predicted phosphatase/phosphohexomutase	0.52	Starburst	1.10.164.10
664	cAMP-binding proteins	0.5	Starburst	1.10.10.10
745	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	0.73	Starburst	3.40.50.2300
753	Catalase	0.93	Starburst	3.30.63.10*
778	Nitroreductase	0.64	Starburst	3.40.109.10
784	FOG: CheY-like receiver	0.48	Starburst	3.40.50.2300
796	Glutamate racemase	0.92	Starburst	3.40.50.1860
1028	Dehydrogenases with different	0.84	Starburst	3.40.50.720

	specificities (related to short-chain alcohol dehydrogenases)			
1151	6Fe-6S prismane cluster-containing protein	0.71	Starburst	1.20.1270.30
1309	Transcriptional regulator	0.8	Starburst	1.10.357.10
1396	Predicted transcriptional regulators	0.54	Starburst	1.10.260.40
1404	Subtilisin-like serine proteases	0.6	Starburst	3.40.50.200
1733	Predicted transcriptional regulators	1	Starburst	1.10.510.10*
1846	Transcriptional regulators	0.85	Starburst	1.10.10.10
2159	Predicted metal-dependent hydrolase of the TIM-barrel fold	0.83	Starburst	3.20.20.140*
2367	Beta-lactamase class A	0.93	Starburst	3.40.710.10
2730	Endoglucanase	0.88	Starburst	3.20.20.80
3693	Beta-1,4-xylanase	0.89	Starburst	3.20.20.80
4948	L-alanine-DL-glutamate epimerase and related enzymes of enolase superfamily	0.71	Starburst	3.20.20.120

The development of the PROFESS (**PRO**tein **F**unction, **E**volution **S**equence and **S**tructure) database (http://cse.unl.edu~profess)[19] contains all PDB-to-COG annotations along with other biologically relevant information. This includes associating each structure with its phyla classification, which allowed for the structures from *Firmicutes* and *Proteobacteria* to be easily selected for further analysis.

The most recent COG database was created by finding the genome-specific best-hit for each gene in 66 unicellular genomes (50 bacteria, 13 archaea, and 3 eukaryota). Specifically, the orthologs present in three or more genomes were detected automatically and then multidomain proteins were manually split into component domains to eliminate artifactual lumping. The online COG database contains 192,987 sequences distributed among 4,876 COGs, accounting for 75% of genes in these 66 genomes.

At the time of our COG-to-PDB annotation, the PDB included 45,368 protein structures (August 2008), although many of them were composed of multiple subunits (and therefore associated with an even larger number of sequences). The two best-represented bacterial phyla, which accounts for nearly one-fourth of all structures in the PDB, were selected for annotation. The PDB contains 8,298 *Proteobacteria* protein structures and 3,416 *Firmicutes* structures. The sequences for each of these structures were compared to the COG reference sequences using the Basic Local Alignment Search Tool (BLAST).[26] The initial match between the COG and PDB databases was completed with an expectation cut-off of $1x10^{-9}$ to maximize the likelihood of matching each PDB with its correct COG. The BLAST similarity matching was required for two reasons, first the PDB did not list gene names and secondly to capture structures from organisms that were not present in the COG database. The BLAST comparison matched 82% of the

Firmicutes and *Proteobacteria* sequences to specific COGs, resulting in functional assignments for 2,728 *Firmicutes* structures and 6,881 *Proteobacteria* structures. Of these hits, 27% were 100% identical to the COG reference sequence and 97% matched with greater than 50% sequence identity. To carry out our comparative study, we selected only those COGs that contained a minimum of two *Firmicutes* organisms and two *Proteobacteria* organisms. This requirement gave 281 unique COGs with a total of 3,047 bacterial proteins (1,066 *Firmicutes* and 1,981 *Proteobacteria*).

6.3.2 Pairwise structure similarity. The pairwise structure comparison tool DaliLite[20] was used to perform 63,504 pairwise comparisons between all of the proteins in our dataset. In total, the backbone structure similarity corresponded to 31,542 *Proteobacteria-Proteobacteria* comparisons (-/-), 12,674 *Firmicutes-Firmicutes* comparisons (+/+), and 19,288 *Proteobacteria-Firmicutes* comparisons (-/+). All comparisons were manually filtered within their respective COG to remove all but one redundantly solved structure (the largest contributor to the size of the dataset), multiple or non-functionally relevant conformations (mutant protein, non-native experimental conditions, inhibited ligand complex), and the shorter of two protein structures. The final dataset contained 48 COGs (table 6.1) with a total of 1,713 structural comparisons with 147 *Firmicutes* proteins from 58 unique organisms and 176 *Proteobacteria* proteins from 84 unique organisms (see appendix 6A for complete list of proteins used in this study).

After manual analysis the resulting dataset was predominantly populated with very low E-values further supporting correct annotation of a structure to the correct COG. The distribution of E-values is reported in figure 6.1. The histogram shows only 3 PDB-to-COG matches with the minimum E-value cutoff, with the majority of the PDB-to-

COG matches falling below $1.0E^{-40}$. The median E-value for each COG is reported in appendix 6B with a range between $4E^{-16}$ and 0. Where an E-value of 0 indicates all structures within a COG were perfectly matched.

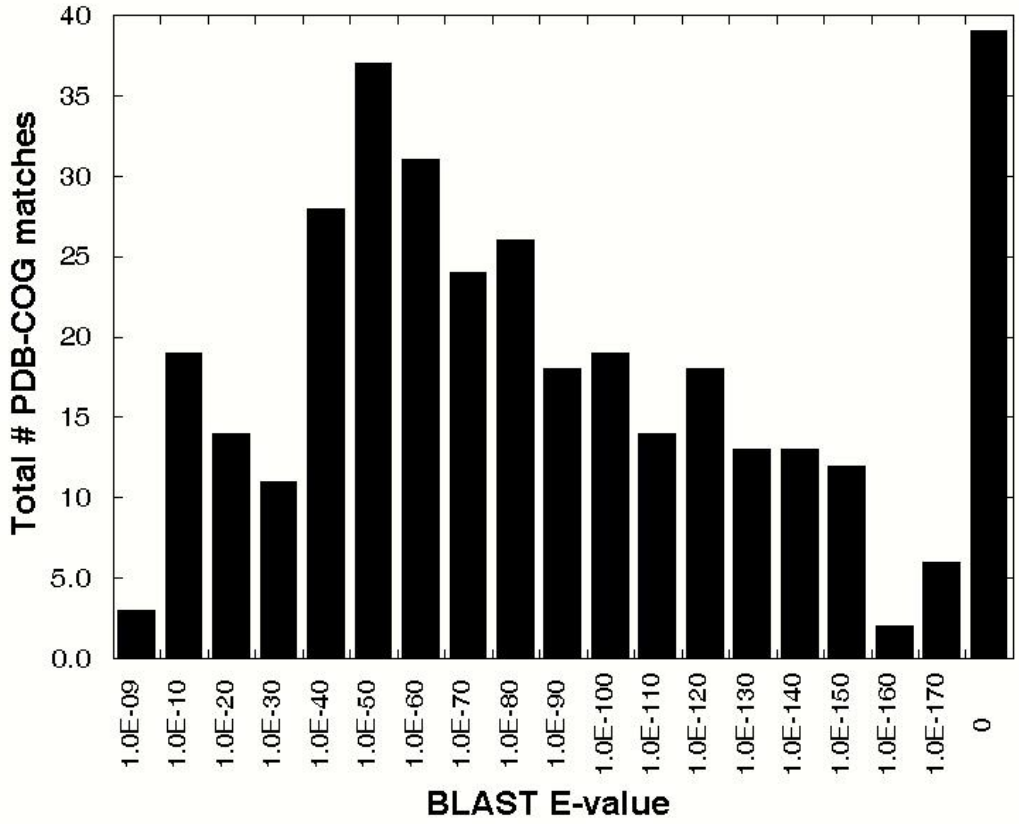


Figure 6.1. Distribution of E-values within manually filtered dataset. A set of E-values at each division consisted of the total number of PDB-to-COG matches between the upper and lower bounds. An E-value of $1x10^{-9}$ approximately relates to a standard significance P-value of $1x10^{-9}$.

The resulting Dali Z-scores from the pairwise structure comparisons were plotted against sequence identity (figure 6.2) to reveal a saturating relationship as the percent identity rose to 100%. The lowest observed Z-score was 5.7 with a corresponding 16% sequence identity. This Z-score was still above the minimum cutoff of 2.0 (dashed line) for matches that were two standard deviations above a random match. This lowest Z-score came from the comparison of two *Firmicutes* proteins in COG0346 (lactoylglutathione lyase and related lyases): 2QH0 (*Clostridium acetobutylicum*); and 2QQZ (*Bacillus anthracis*). The average Z-score for all comparisons was 27 ± 13, indicating that all structural comparisons were significant.

Since Z-scores increase as a function of the protein length, we normalized this effect by calculating the Fractional Structure Similarity (FSS) score as described in eq 6.1. The pairwise FSS scores plotted against sequence identity (figure 6.3) resulted in a hyperbolic curve. All FSS values fell below an upper-limit at each percent identity. In fact, 20% sequence identity yielded a maximal FSS of 60%. This FSS limit was observed when all of the data were used (figure 6.3A), when only the pairwise comparisons within either phyla were used (figure 6.3B and C), or when only the pairwise comparisons between the two phyla were used (figure 6.3D). The pairwise comparison plot between the two phyla (figure 6.3D) showed an abrupt cutoff at 61% sequence identity and a 0.84 FSS score. This abrupt cutoff was not an artifact created by culling the dataset, since a similar plot prior to the manual filtering also demonstrated the same effect (appendix 6C).

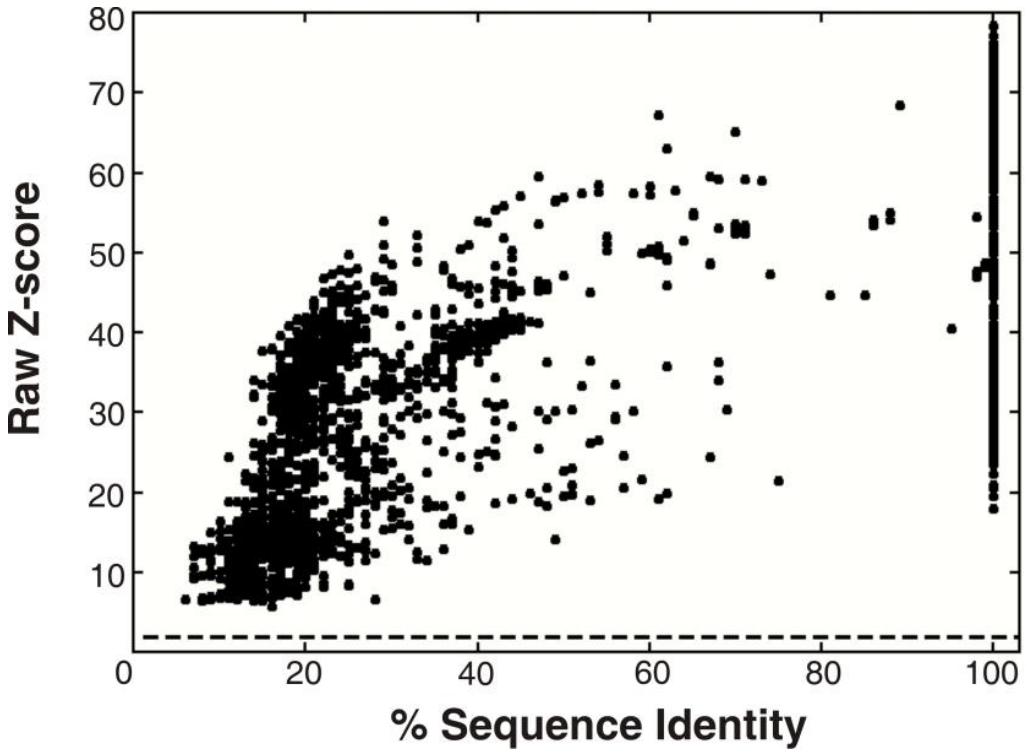


Figure 6.2. The relationship between structure similarity and sequence identity for 48 COGs. Structure similarity is given as the raw Z-score, which increases as the protein length increases. The comparisons were for all proteins against all proteins, and include the comparison for each protein against itself. The dashed line identifies a Dali Z-score of 2, which is the minimal limit for inferring structural similarity.

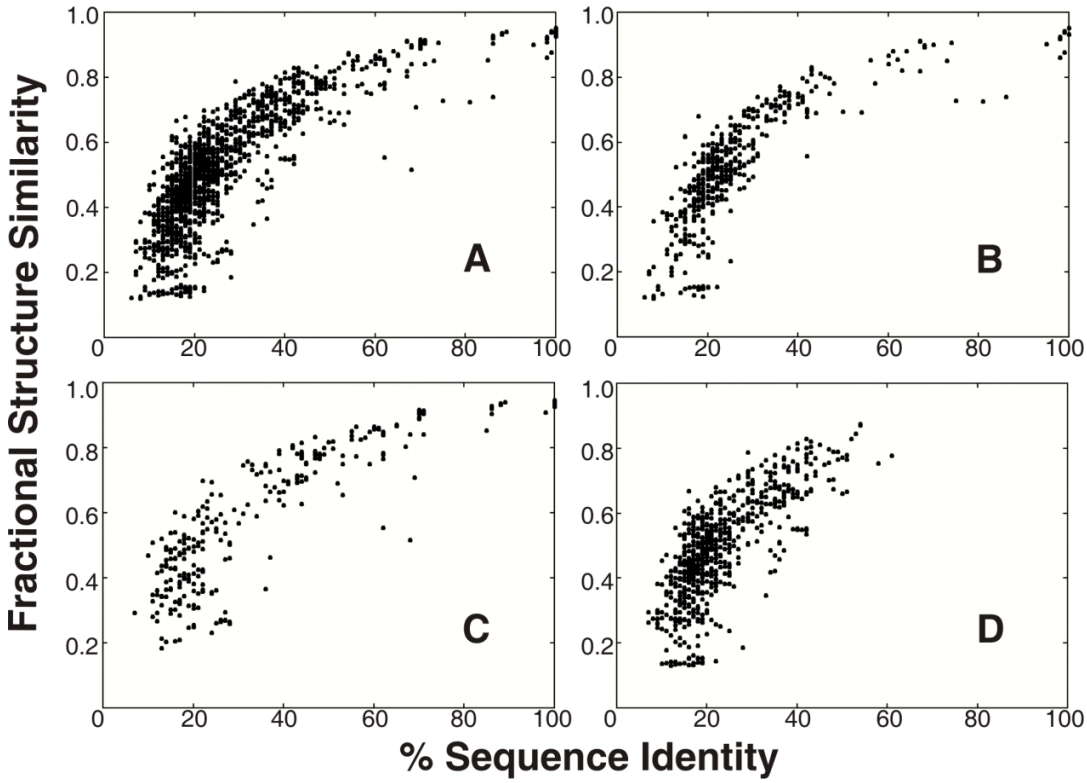


Figure 6.3. The fractional structure similarity (FSS) and sequence identity for 48 COGs. FSS was calculated using eq 6.1 to normalize the Dali Z-scores for their different sizes. The FSS values were plotted against sequence identity for (A) all the pairwise comparisons, (B) only *Proteobacteria-Proteobacteria* comparisons, (C) only *Firmicutes-Firmicutes* comparisons and (D) only *Proteobacteria-Firmicutes* comparisons.

The protein structures in COG0028 (thiamine pyrophosphate requiring enzymes) provides a useful example of the structural divergence that occurred after the *Firmicutes* and *Proteobacteria* phyla split. The overall fold is conserved between the phyla while there are discrete structural elements that are unique to each phylum. The two *Firmicutes* structures (figure 6.4A) yield a Z score of 59.6 and an FSS of 0.83, indicating very high structural conservation. The structure comparison between the 4 representative *Proteobacteria* structures (figure 6.4B) yield an average Z-score of 37.7 ± 1.6 and an average FSS of 0.58 ± 0.03. Again, the structures share a similar fold despite the slightly lower scores.

Comparison of the structures between the *Firmicutes* and *Proteobacteria* (figure 6.4C and D, respectively) phyla yield a lower Z-score of 34.8 ± 1.2 and a lower FSS of 0.49 ± 0.02 then the comparisons within each phylum. This suggests a divergence in structural details while conserving the overall fold. A detailed analysis reveals localized differences between the structures from the two phyla (see red highlights in figure 6.4C and D). In the *Firmicutes* representative structure, there is a continuous helix compared to helical breaks and loop insertions in the *Proteobacteria* structure. This is similar to the C-terminal domain of primase, where a long continuous helix found in the *E. coli* structure is broken by a loop region in *G. stearothermophilus*[27-30] and *S. aureus* (chapter 5).

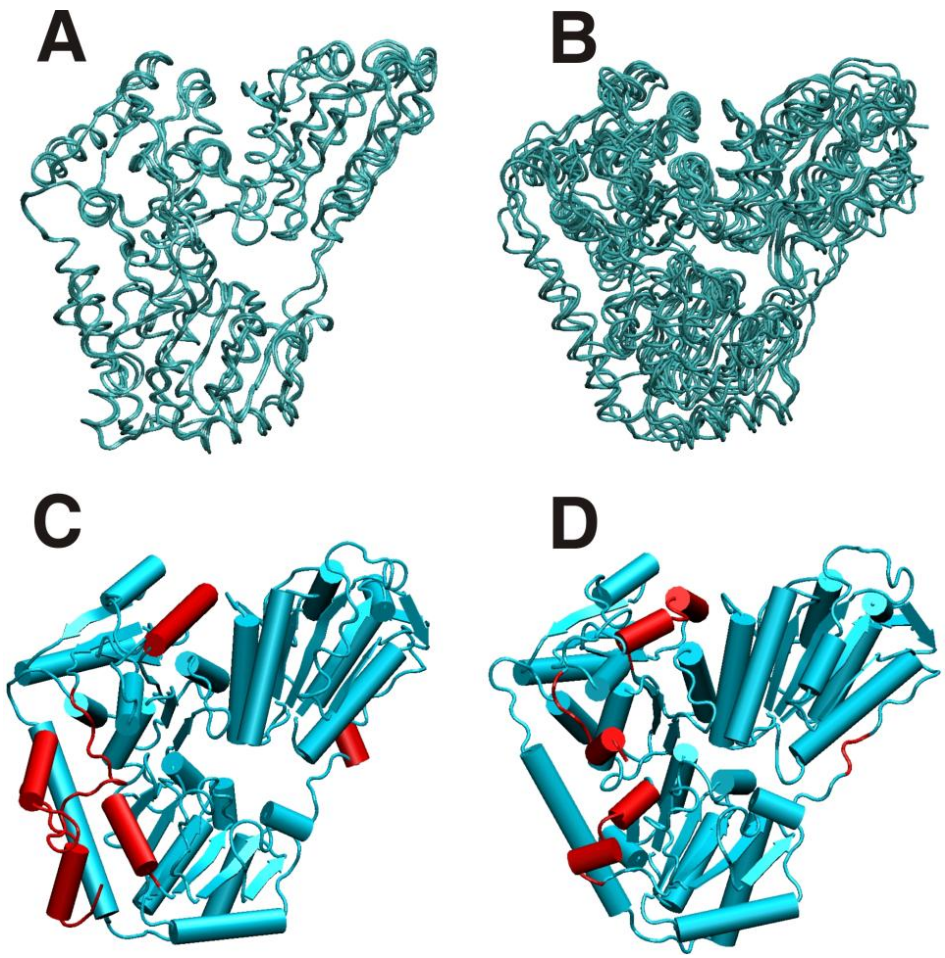


Figure 6.4. Comparison of protein structures for COG0028 between two bacterial phyla. The protein structures for COG0028 thiamine pyrophosphsate requiring enzymes show (A) the two *Firmicutes* structures have highly overlapping structures and (B) the four *Proteobacteria* structures are very similar to each another. See also the phylogenetic structure tree for COG0028 in (figure 6.5). On the other hand, the major structural differences between the *Firmicutes* and *Proteobacteria* are highlighted in red on a representative *Firmicutes* (C) structure from *L. plantarum* (**Lpl**) (PDB ID: 1POW) and the representative *Proteobacteria* structure (D) from *P. fluorescens* (**Pfl**) (PDB ID: 2AG0).

6.3.3 COG structure phylogenies. Structure based phylogenies were created from root-mean square differences (rmsd) in per residue Cα positions for optimally aligned protein structures using MAMMOTH-multi.[21] A separate phylogenetic tree was generated for each COG, where three distinct patterns were observed (table 6.1): 15 exhibited a strong split at the phylum level, 29 exhibited a starburst pattern suggesting little to no evidence for a split according to phyla, and 4 exhibited a strong split at the phylum level but with the exception of a single structure (split +1).

The 15 COG phylogenies with strong phylum-splitting patterns had two branches, one with closely related *Firmicutes* structures and the other with closely related *Proteobacteria* structures. Two examples are COG0028 (Thiamine pyrophosphate requiring enzymes) and COG0446 (Uncharacterized NAD(FAD)-dependent dehydrogenases) (figure 6.5). The structures for both of these COGs are classified in the CATH system as α/β 3-layer sandwiches, but differ in that COG0028 proteins have a Rossmann fold topology (figure 6.4) and COG0046 proteins have a FAD/NAD (P)-binding domain topology.

The 29 COGs with phylogenetic starburst patterns showed no evidence for the separation of structures according to phyla (table 6.1). Two examples were COG0491 (Zn-dependent hydrolases) and COG1309 (Transcriptional regulator) (figure 6.5). The CATH classification for COG0491 *Bacillus cereus* Zinc-dependent beta-lactamase (PDB ID: 1BC2) [31] describes it as an α/β 4-layer sandwich with metallo-beta-lactamase Chain A topology. The large category of beta-lactamases constitutes a collection of enzymes that can be derived from any one of a group of proteins that bind, synthesize, or degrade peptidoglycans. The protein structures assigned to COG0491 gave FSS scores with large

standard deviations, as is evident from the separated clusters within the *Proteobacteria* arm of the phylogenetic tree.

The COG1309 structural family falls into one of two CATH topologies, Arc Repressor Mutant (subunit A) or Tetracycline Repressor (domain 2). Only those structures similar to the Arc Repressor Mutant (subunit A) topology were used for the pairwise comparison, since it was the dominant fold in this COG. The protein structures in the COG1309 structure family gave low FSS scores. However, even with low overall FSS the average absolute Z-score was 13 ± 2 indicating that it has significant overall structure similarity. The high FSS deviations of COG0491 structural family and the low average FSS scores of COG1309 structural family both indicate rapid structural divergence following the phyla split, consistent with the observed starburst phylogenetic patterns.

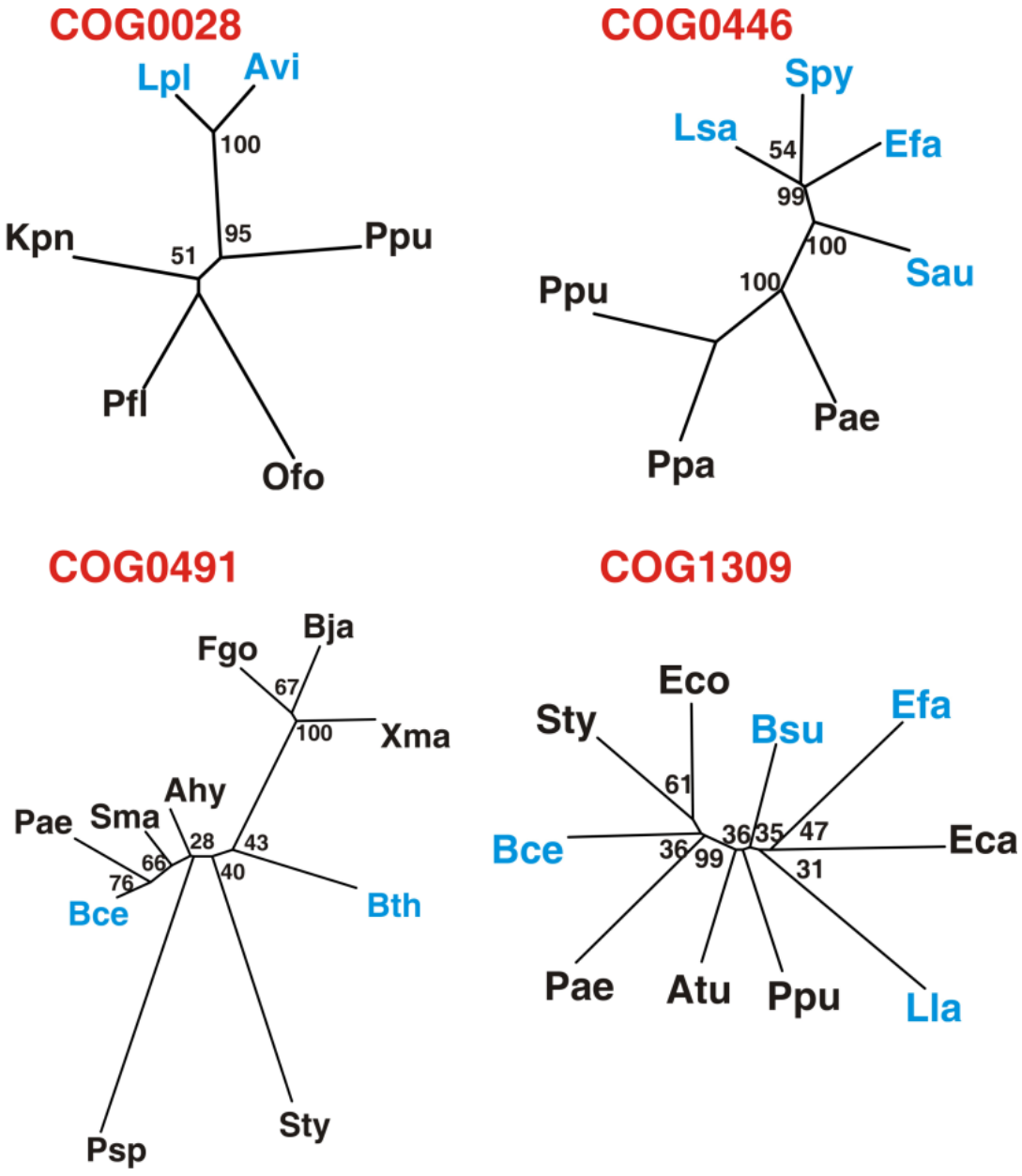

COG0028
Lpl
Avi
100
Kpn
51
95
Ppu
Pfl
Ofo
COG0446
Spy
Lsa
54
Efa
99
100
Sau
Ppu
100
Pae
Ppa
COG0491
Fgo
Bja
67
100
Xma
Ahy
Pae
Sma
28
43
66
40
76
Bth
Bce
Sty
Psp
COG1309
Eco
Sty
Efa
Bsu
61
36
35
47
Bce
36
99
Eca
31
Pae
Atu
Ppu
Lla

Figure 6.5. Protein structure based phylogenetic trees highlighting the split and starburst patterns. The phylogenetic structure trees showed three different patterns: (*top*) strong split according to phyla; *(bottom*) starburst with no clear relationship to a common ancestor; and (figure 6.6) strong splits with the exception of one outlier. The *Firmicutes* protein structures are in blue and the *Proteobacteria* in black. The bootstrap values from 100 bootstrap replicates are indicated on branches and represent how often a branch appeared in the distance matrix. The two examples for the split pattern were from COG0028 (thiamine pyrophosphate requiring enzymes) and COG0446 (uncharacterized NAD(FAD)-dependent dehydrogenases). In the case of a strong split, the central branches were observed more than 95 times out of 100 replicate trials. The two examples for starburst pattern were from COG0491 (Zn-dependent hydrolases) and COG1309 (transcriptional regulator). For starburst patterns, very few branches were observed in more than two-thirds of the 100 replicate trials. The organism abbreviations are: *A. hydrophila* (Ahy) ; *A. tumefaciens* (Atu); *A. viridians* (Avi) ; *B. cereus* (Bce) ; *B. japonicum* (Bja); *B. subtilis* (Bsu) ; *B. thuriagienes* (Bth); *E. carotovora* (Eca); *E. coli* (Eco); *E. faecalis* (Efa); *F. gormanii* (Fgo); *K. pneumonia* (Kpn); *L. lactis* (Lla); *L. sanfranciscens* (Lsa); *L. plantarum* (Lpl); *O. formigens* (Ofo) ; *P. aeruginosa* (Pae); *P. fluorescens* (Pfl); *P. pantotrophus* (Ppa); *P. putida* (Ppu); *P. species* (Psp); *S. aureus* (Sau); *S. marcescens* (Sma); *S. typhimurium* (Sty); and *X. maltophilia* (Xma).

Four COG structure phylogenies showed a strong split pattern with a single outlier (figure 6.6). This result provides further evidence for the observation of phyla split based on structure similarity. The presence of the outlier in a clear split pattern suggests a horizontally transferred gene (table 6.1) or potential paralog. For all four families [COG0242 (N-formylmethionyl-tRNA deformylase) COG1052 (Lactate dehydrogenase and related dehydrogenases), COG2141 (Coenzyme F420-dependent N5,N10-methylene tetrahydromethanopterin reductase and related flavin-dependent oxidoreductases), and COG3832 (Uncharacterized conserved protein)] there was a large and significant average absolute Z-score for all comparisons along with strong BLAST E-values indicating the correct match was made between COG and PDB. For COG0242, the *Bacillus cereus* gene *def* that encodes the N-formylmethionyl-tRNA deformylase protein (PDB ID: 1WS0) has been previously identified as a gene that has undergone horizontal gene transfer.[32]

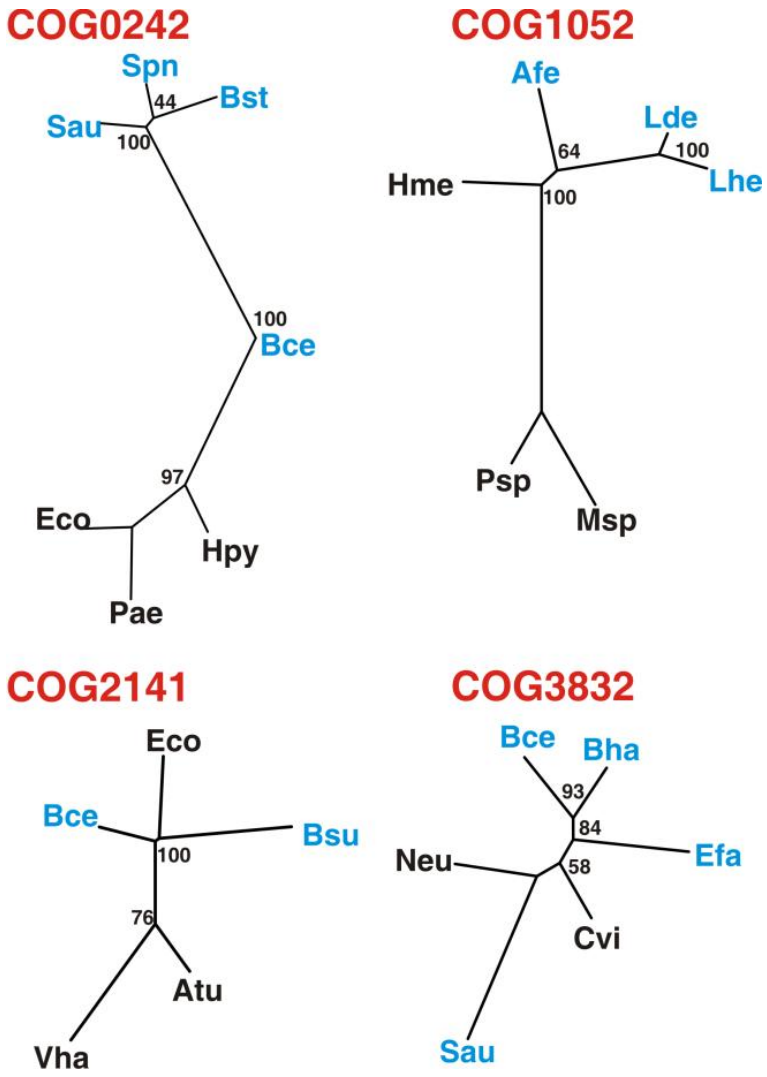

COG0242
Spn
Bst
44
Sau
100
100
Bce
97
Eco
Hpy
Pae
COG1052
Afe
Lde
64
100
Lhe
Hme
100
Psp
Msp
COG2141
Eco
Bce
Bsu
100
76
Atu
Vha
COG3832
Bce
Bha
93
84
Efa
Neu
58
Cvi
Sau

Figure 6.6. **Protein structure based phylogenetic trees highlighting the split +1 pattern.** Protein structure phylogenies of 4 COGs out of 48 had a strong split pattern with the exception of one outlier structure. The phylogenies were very reliable because the central branches were observed in 100 out of 100 replicate trials. When one *Firmicutes* or *Proteobacteria* protein structure clusters on a branch with the other phylum, its structure diverges from its closest relatives while resembling those of the other phyla. The COGs that fit this pattern are from COG0242 (N-formylmethionyl-tRNA deformylase), COG1052 (lactate dehydrogenase and related dehydrogenases), COG2141 (coenzyme F420-dependent N5, N10-methylene tetrahydromethanopterin reductase and related flavin-dependent oxidoreductases), and COG3832 (uncharacterized conserved protein). The organism abbreviations are: *A. fermentans* (Afe); *A. tumefaciens* (Atu); *B. cereus* (Bce); *B. halodurans* (Bha); *B. stearothermophilus* (Bst); *B. subtilis* (Bsu); *C. violaceum* (Cvi); *E. coli* (Eco); *E. faecalis* (Efa); *H. methylovorum* (Hme); *H. pylori* (Hpy); *L. delbrueckii* (Lde); *L. helveticus* (Lhe); *M. species* (Msp); *N. europaea* (Neu); *P. aeruginosa* (Pae) ; *P. species* (Psp), *S. aureus* (Sau); *S. pneumoniae* (Spn); and *V. harveyi* (Vha).

6.3.4 Structure divergence rates across phyla. As a way to quantify the relationship between structure difference and sequence difference, each phylogenetic tree was reduced to a single coordinate by calculating a structure similarity ratio (θ_{FSS}) and a sequence identity ratio (θ_{SeqID}). θ_{FSS} was determined for all 48 COGs by calculating an average FSS score for the *Proteobacteria-Firmicutes* structure comparisons, Avg($FSS_{+/-}$), and dividing by the sum of the average *Proteobacteria-Proteobacteria*, Avg($FSS_{-/-}$), and *Firmicutes-Firmicutes*, Avg($FSS_{+/+}$), comparisons:

$$\theta_{FSS} = \frac{Avg\,(FSS_{+/-})}{Avg\,(FSS_{+/+})/_2 + Avg\,(FSS_{-/-})/_2} \qquad [6.3]$$

Similarly, a sequence identity ratio (θ_{SeqID}) was determined by calculating an average sequence identity for the *Proteobacteria-Firmicutes* structure comparisons, Avg($SeqID_{+/-}$), and dividing by the sum of the average *Proteobacteria-Proteobacteria*, Avg($SeqID_{-/-}$), and *Firmicutes-Firmicutes*, Avg($SeqID_{+/+}$), comparisons:

$$\theta_{SeqID} = \frac{Avg\,(SeqID_{+/-})}{Avg\,(SeqID_{+/+})/_2 + Avg\,(SeqID_{-/-})/_2} \qquad [6.4]$$

In general, most starburst phylogenies (see representative COG0491 and COG1309 in (figure 6.5) had a branch length between members of different phyla that was much shorter than the branch lengths between members within the same phyla. That is, a starburst phylogeny was expected to have θ_{FSS} and θ_{SeqID} values greater than unity. Likewise, most split phylogenies had longer branches between phyla than within each phyla (see representative COG0028 and COG0446 in (figure 6.5) and were expected to yield θ_{FSS} and θ_{SeqID} of less than unity.

When θ_{FSS} and θ_{SeqID} for all 48 COGs were plotted versus one another (figure 6.7), the starburst phylogenies clustered around unity for both structure and sequence

whereas the split phylogenies clustered around 0.85 for structure and 0.70 for sequence. This indicated that split phylogenies occur when the structure differences are significantly less than their sequence differences. In addition, the plot of θ_{FSS} versus θ_{SeqID} conformed to a linear relationship regardless of the shape of the phylogenetic tree indicating that all homologous protein structure differences are constant with respect to homologous protein sequence differences ($\theta_{FSS} = 0.55\theta_{SeqID} + 0.45$; $R^2 = 0.7$). Thus, this curve represents the relative structural drift rate for each COG structural family between the two phyla. The slope indicates that structure branch lengths change approximately half as fast as sequence branch lengths.

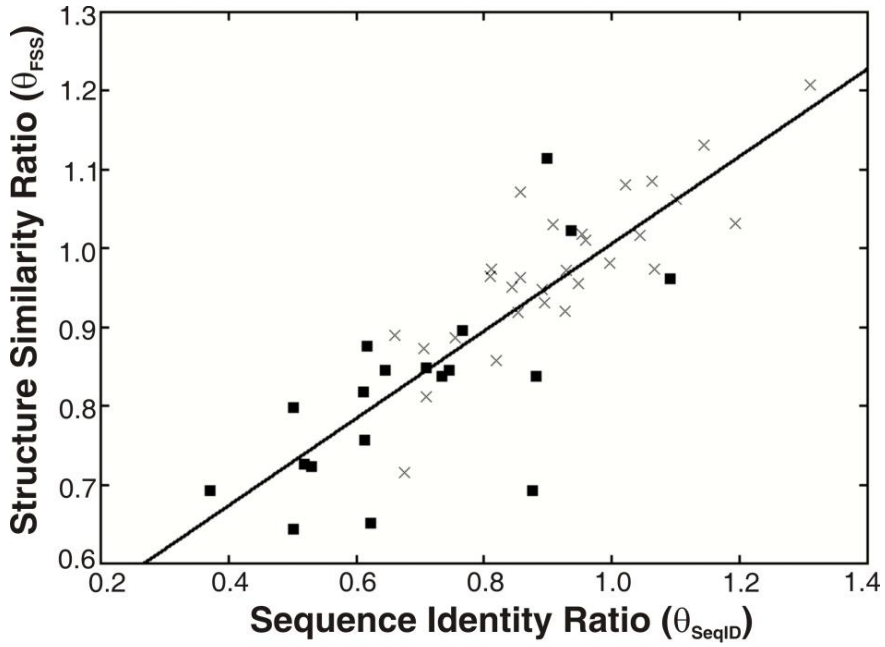


Figure 6.7. Constant rate of structural drift. The relationship between structure and sequence change was constant regardless of the phylogenetic starburst (x) or split (■) pattern. Structure changes measured using a structure similarity ratio (θ_{FSS}), where the average FSS between members of the two phyla (*Firmicutes* versus *Proteobacteria*) was divided by the average FSS between members of the same phyla (see eq 6.3). Sequence change was calculated similarly (see eq 6.4). The best-fit line, $\theta_{FSS}=0.55\theta_{SeqID} + 0.45$, yielded an R^2 of 0.70.

6.3.5 Fold dependency on structure similarity. A plot of FSS vs. sequence identity for the two most populated CATH families in our dataset (figure 6.8) was used to investigate if particular protein architectures are more amenable to structural changes. The largest portion of our data set, 24 of 48 COGs (50%), is represented by CATH 3.40 (α/β, 3-layer ($\alpha\beta\alpha$) sandwich). Within CATH 3.40, 12 of 24 COGs (50%) are represented by the starburst phylogenetic tree pattern. The remaining 12 COGs correspond to 11 splits and 1 split +1 phylogenetic tree patterns.

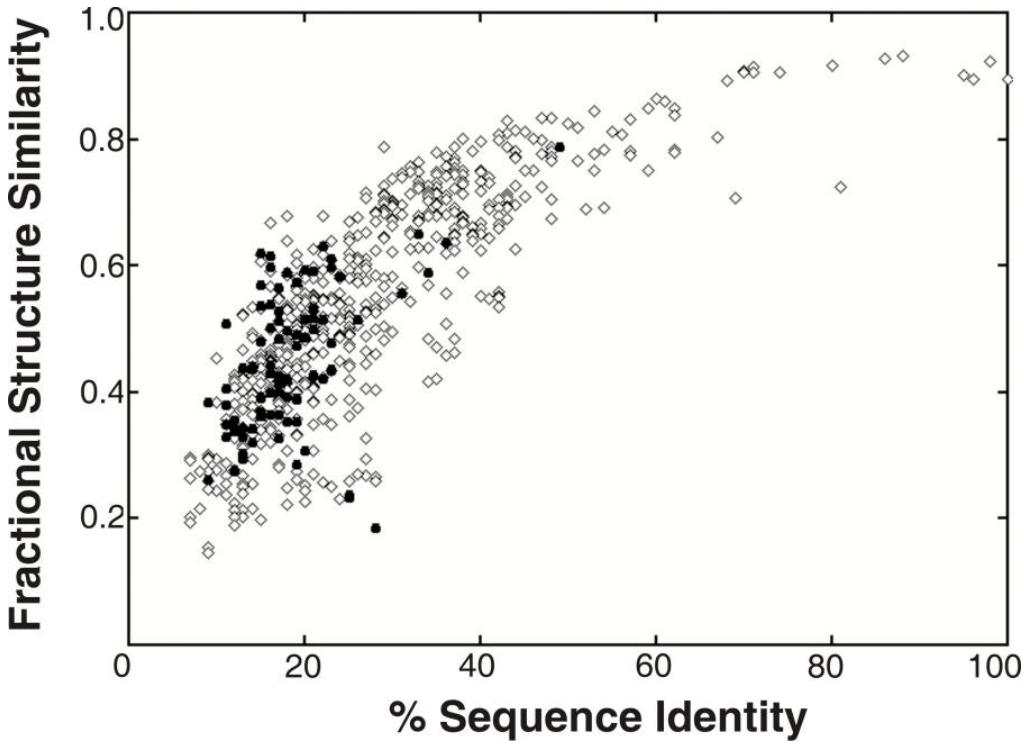


Figure 6.8. Fold dependency on fractional structure similarity (FSS) and sequence comparisons. The FSS between two CATH families, CATH 1.10 (●) CATH 3.40 (◊). CATH 1.10 (mainly α, orthogonal bundle) family is apparently limited to approximately 40% sequence identity and 0.6 FSS while CATH 3.40 (α/β, 3-Layer ($\alpha\beta\alpha$) sandwich) fills in the complete curve. 87.5% of the COGs (7 of 8) represented by CATH 1.10 give a starburst structure similarity tree. Contrastingly, only 50% (12 of 24) of the COGs represented by CATH 3.40 give a starburst structure similarity tree. The remaining 12 COGs formed either split (11 of 12) or split +1 (1 of 12).

The second most populous CATH family is CATH 1.10 (mainly α, orthogonal bundle) with 15% of our COGs belonging to this CATH family. Most (85.7%) of the COGs (6 of 7) in the CATH 1.10 family are represented by the starburst phylogenetic tree pattern with only 1 COG represented by a split pattern. There appears to be a limit in structure similarity at approximately 0.6 FSS and a corresponding sequence identity limit at 40% for CATH 1.10 (figure 6.8, solid circles). This limit is not observed in the CATH 3.40 family (figure 6.8, open diamonds). The sequence and structure similarity limit for CATH 1.10 combined with a larger percentage of COGs assigned to the starburst family suggests that CATH 1.10 is more susceptible to mutations that affect the protein structure. The results suggest a faster evolutionary rate leading to a higher structural divergence relative to other CATH architectures.

6.4 DISCUSSION

There is an inherent challenge in obtaining an accurate functional annotation for a large set of proteins from a relatively small number of experimentally determined functions.[33-37] The available functional information is incomplete, ambiguous and error-prone[38, 39] and requires multiple sources[35] to improve the accuracy in the annotation of a protein. There is also the complicating factor of correctly distinguishing between orthologs and paralogs, where it has been previously noted the COG database does include some paralog members[17, 40]. Thus, the accuracy of this analysis is fundamentally dependent on a reliable functional assignment for each protein structure. Given these challenges, the independent and separate utilization of both COG and GO terms provides a reasonable and robust approach to identify clusters of functionally similar proteins. The

overall high sequence (E-value $\leq 10^{-9}$, sequence identity $\geq 16\%$), structure (Z-score > 5.7) and GO term similarity (0.72 ± 0.21) within each COG supports this conclusion. The lack of identity for the GO term similarity scores should not be interpreted as evidence for functional divergence. GO terms are assigned based on a validated source. So, a missing GO term for a protein is more likely attributed to the fact the protein has not been explicitly tested for the specified activity. Similarly, a protein being assigned a GO term does not provide definitive evidence the function is relevant *in vivo*.[41-44]

The comparison of homologous protein structures with the same function provides quantitative evidence that protein structures diverged following the speciation events that created the modern bacterial phyla of *Firmicutes* and *Proteobacteria*. The abrupt cutoff at 61% sequence identity and 0.84 fractional structure similarity observed between *Firmicutes* and *Proteobacteria* proteins was mirrored by an approximate 60% protein sequence identity between these two phyla observed by 16S rRNA sequence similarity.[45, 46] Thus, this maximum observed sequence identity imparts limits to the maximum possible structure similarity between homologus proteins from these two phyla. This is consistent with prior observations that sequence identity $\leq$ 40-50% sometimes results in significant structural and functional differences.[7, 8, 47] Furthermore, the results imply an inherent allowable structural plasticity that does not perturb function. The random drift after speciation inexorably leads to non-identical structures despite maintenance of function. There are a number of cases where FSS was below 0.20 indicating a significant structural change. Proteins with completely different folds but the same function are extreme examples of the plasticity of the structure-function relationship and include such proteins as peptidyl-tRNA hydrolases (COG1990),[48]

pantothenate kinase (KOG2201).[49] polypeptide release factors [50] and lysyl-tRNA synthetases (COG1190),[51] these proteins are not in our dataset.

Forty percent of the COGs we examined have evolved slowly enough that it was possible to generate phylogenetic trees consistent with this ancient split. The other COGs have either evolved too rapidly or are otherwise subject to few evolutionary constraints to provide evidence for this split. This distinction between the COGs is clearly apparent from the comparison of θ_{FSS} and θ_{SeqID} in (figure 6.7). The linear relationship implies a fixed relative structure drift rate, where structure changes half as fast as sequence across phyla. This correlation in the divergence of protein sequences and protein structures has additional ramifications beyond bacterial evolution. Our analysis implies a continuum of protein folds that adapt to large sequence changes by incurring local structural modifications.[13-16] This continuum of protein folds makes it challenging to apply protein structural classification to identify function, as has been previously noted.[52, 53]

Does the nature of the protein's three-dimensional structure play a role in protein structure divergence? Our analysis demonstrates that some proteins evolve slowly and maintain high sequence identity (>80%) and structure similarity (> 0.80 FSS) while other proteins exhibit rapid evolution rates where sequence identity is ≤ 20% and FSS ≤ 0.40. This implies the underlying architecture of a particular protein may be more or less amenable to amino-acid substitutions in order to maintain functional activity. A specific protein fold may have a higher intrinsic plasticity that enables it to readily accommodate sequence changes through local conformational changes without a detrimental impact on activity. This is exactly what was observed, structural variations were localized to specific regions as illustrated by the comparison of the COG0028 protein structures see

(figure 6.4). This is consistent with the observation of different structure divergence rates within a protein.[54, 55] Regions of the protein that do not impact biological activity are expected to yield a higher divergence rate and incur larger local structural changes.[56, 57] As a result, a fold with a relatively high plasticity would experience an elevated structural diversity between phyla, where the rate of change may closely parallel the mutation rate.[14] Conversely, another fold may be extremely sensitive to amino-acid substitutions, where minor sequence perturbations may result in a decrease in structural integrity and a corresponding loss of activity. As a result, the sequence and structure of this protein class would be relatively conserved. This analysis is consistent with the known range of protein thermodynamic stabilities,[58] and the general observation that most mutations destabilize protein structures.[59]

This chapter illustrates the inherent value in solving structures for functionally identical proteins from multiple organisms. A major challenge in creating our COG-to-PDB dataset was the fundamental requirement to have structures from at least two *Firmicutes* organisms and two *Proteobacteria* organisms. Only 48 (~1%) of the 4,876 COGs meet this stringent requirement. The limited number of multiple homologous structures has partly occurred because structural biology efforts are focused on obtaining single representative structures for each functional class or protein fold[60] and understandably biased toward therapeutically relevant proteins.[61] If we are to achieve a more accurate understanding of the relationship between the evolution of protein fold, protein sequence, and the organisms in which they function, the fields of bioinformatics and structural biology must expand their focus to include efforts to obtain a more diverse set of homologous protein structures.

6.5 REFERENCES

1. Do, C. B.; Katoh, K., Protein multiple sequence alignment. *Methods Mol. Biol. (Totowa, NJ, U. S.)* **2008,** 484, (Functional Proteomics), 379-413.
2. Feng, J.-a., Improving pairwise sequence alignment between distantly related proteins. *Methods Mol. Biol. (Totowa, NJ, U. S.)* **2007,** 395, (Comparative Genomics, Volume 1), 255-268.
3. Chang, G. S.; Hong, Y.; Dae Ko, K.; Bhardwaj, G.; Holmes, E. C.; Patterson, R. L.; van Rossum, D. B., Phylogenetic profiles reveal evolutionary relationships within the "twilight zone" of sequence similarity. *Proc Natl Acad Sci USA* **2008,** 105, (36), 13474-13479, S13474/1-S13474/14.
4. Pal, C.; Papp, B.; Lercher, M. J., An integrated view of protein evolution. *Nat Rev Genet* **2006,** 7, (5), 337-48.
5. Rocha, E. P. C., The quest for the universals of protein evolution. *Trends in Genetics* **2006,** 22, (8), 412-416.
6. Forouhar, F.; Kuzin, A.; Seetharaman, J.; Lee, I.; Zhou, W.; Abashidze, M.; Chen, Y.; Yong, W.; Janjua, H.; Fang, Y.; Wang, D.; Cunningham, K.; Xiao, R.; Acton, T. B.; Pichersky, E.; Klessig, D. F.; Porter, C. W.; Montelione, G. T.; Tong, L., Functional insights from structural genomics. *J Struct Funct Genomics* **2007,** 8, (2-3), 37-44.
7. Chothia, C.; Lesk, A. M., The relation between the divergence of sequence and structure in proteins. *Embo J* **1986,** 5, (4), 823-6.
8. Rost, B., Twilight zone of protein sequence alignments. *Protein Engineering* **1999,** 12, (2), 85-94.

9. Sadreyev, R. I.; Grishin, N. V., Exploring dynamics of protein structure determination and homology-based prediction to estimate the number of superfamilies and folds. *BMC Struct Biol* **2006,** 6, 6.

10. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res* **2000,** 28, (1), 235-42.

11. Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M., CATH--a hierarchic classification of protein domain structures. *Structure* **1997,** 5, (8), 1093-108.

12. Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C., SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **1995,** 247, (4), 536-40.

13. Kolodny, R.; Petrey, D.; Honig, B., Protein structure comparison: Implications for the nature of 'fold space', and structure and function prediction. *Curr. Opin. Struct. Biol.* **2006,** 16, (3), 393-398.

14. Illergard, K.; Ardell, D. H.; Elofsson, A., Structure is three to ten times more conserved than sequence-A study of structural response in protein cores. *Proteins* **2009,** 77, (3), 499-508.

15. Panchenko, A. R.; Wolf, Y. I.; Panchenko, L. A.; Madej, T., Evolutionary plasticity of protein families: coupling between sequence and structure variation. *Proteins* **2005,** 61, (3), 535-44.

16. Williams, S. G.; Lovell, S. C., The effect of sequence evolution on protein structural divergence. *Mol Biol Evol* **2009,** 26, (5), 1055-65.

17. Tatusov, R. L.; Fedorova, N. D.; Jackson, J. D.; Jacobs, A. R.; Kiryutin, B.; Koonin, E. V.; Krylov, D. M.; Mazumder, R.; Mekhedov, S. L.; Nikolskaya, A. N.; Rao, B. S.; Smirnov, S.; Sverdlov, A. V.; Vasudevan, S.; Wolf, Y. I.; Yin, J. J.; Natale, D. A., The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **2003,** 4, 41.

18. Schmidt, T.; Frishman, D., PROMPT: a protein mapping and comparison tool. *BMC Bioinformatics* **2006,** 7, 331.

19. Triplet, T.; Shortridge, M. D.; Griep, M. A.; Stark, J. L.; Powers, R.; Revesz, P., PROFESS: a PROtein function, evolution, structure and sequence database. *Database (Oxford)* 2010, baq011.

20. Holm, L.; Park, J., DaliLite workbench for protein structure comparison. *Bioinformatics* **2000,** 16, (6), 566-7.

21. Lupyan, D.; Leo-Macias, A.; Ortiz, A. R., A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics* **2005,** 21, (15), 3255-63.

22. Felsenstein, J., PHYLIP- Phylogeny Inference Package (Version 3.2). *Cladistics* **1989,** 5, 164-166.

23. Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **2000,** 25, (1), 25-9.

24. Schomburg, I.; Chang, A.; Ebeling, C.; Gremse, M.; Heldt, C.; Huhn, G.; Schomburg, D., BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* **2004,** 32, (Database issue), D431-3.

25. Ouzounis, C. A.; Coulson, R. M.; Enright, A. J.; Kunin, V.; Pereira-Leal, J. B., Classification schemes for protein structure and function. *Nat Rev Genet* **2003,** 4, (7), 508-19.

26. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *J Mol Biol* **1990,** 215, (3), 403-10.

27. Bailey, S.; Eliason, W. K.; Steitz, T. A., Structure of hexameric DnaB helicase and its complex with a domain of DnaG primase. *Science* **2007,** 318, (5849), 459-463.

28. Oakley, A. J.; Loscha, K. V.; Schaeffer, P. M.; Liepinsh, E.; Pintacuda, G.; Wilce, M. C.; Otting, G.; Dixon, N. E., Crystal and solution structures of the helicase-binding domain of *Escherichia coli* primase. *Journal of Biological Chemistry* **2005,** 280, (12), 11495-11504.

29. Su, X. C.; Schaeffer, P. M.; Loscha, K. V.; Gan, P. H.; Dixon, N. E.; Otting, G., Monomeric solution structure of the helicase-binding domain of Escherichia coli DnaG primase. *Febs J* **2006,** 273, (21), 4997-5009.

30. Syson, K.; Thirlway, J.; Hounslow, A. M.; Soultanas, P.; Waltho, J. P., Solution structure of the helicase-interaction domain of the primase DnaG: A model for helicase activation. *Structure* **2005,** 13, (4), 609-616.

31. Fabiane, S. M.; Sohi, M. K.; Wan, T.; Payne, D. J.; Bateson, J. H.; Mitchell, T.; Sutton, B. J., Crystal structure of the zinc-dependent beta-lactamase from Bacillus

cereus at 1.9 A resolution: binuclear active site with features of a mononuclear enzyme. *Biochemistry* **1998,** 37, (36), 12404-11.

32. Garcia-Vallve, S.; Romeu, A.; Palau, J., Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.* **2000,** 10, (11), 1719-1725.

33. Andrade, M. A. In *Automatic genome annotation and the status of sequence databases*, 2003; Horizon Scientific Press: 2003; pp 107-121.

34. Frishman, D., Protein Annotation at Genomic Scale: The Current Status. *Chem. Rev. (Washington, DC, U. S.)* **2007,** 107, (8), 3448-3466.

35. Rentzsch, R.; Orengo, C. A., Protein function prediction - the power of multiplicity. *Trends Biotechnol.* **2009,** 27, (4), 210-219.

36. Valencia, A., Automatic annotation of protein function. *Curr. Opin. Struct. Biol.* **2005,** 15, (3), 267-274.

37. Karp, P. D.; Paley, S.; Zhu, J., Database verification studies of SWISS-PROT and GenBank. *Bioinformatics* **2001,** 17, (6), 526-532.

38. Schnoes, A. M.; Brown, S. D.; Dodevski, I.; Babbitt, P. C., Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS Comput Biol* **2009,** 5, (12), e1000605.

39. Benitez-Paez, A., Considerations to improve functional annotations in biological databases. *OMICS* **2009,** 13, (6), 527-532.

40. Dessimoz, C.; Boeckmann, B.; Roth, A. C. J.; Gonnet, G. H., Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res.* **2006,** 34, (11), 3309-3316.

41. Canevascini, S.; Caderas, D.; Mandel, t.; Fleming, A. J.; Dupuis, I.; Kuhlemeier, C., Tissue-specific expression and promoter analysis of the tobacco Itp1 gene. *Plant Physiol.* **1996,** 112, (2), 513-524.

42. Lindorff-Larsen, K.; Lerche, M. H.; Poulsen, F. M.; Roepstorff, P.; Winther, J. R., Barley lipid transfer protein, LTP1, contains a new type of lipid-like post-translational modification. *J. Biol. Chem.* **2001,** 276, (36), 33547-33553.

43. Otsuka, T.; Takagi, H.; Horiguchi, N.; Toyoda, M.; Sato, K.; Takayama, H.; Mori, M., CCl4-induced acute liver injury in mice is inhibited by hepatocyte growth factor overexpression but stimulated by NK2 overexpression. *FEBS Lett.* **2002,** 532, (3), 391-395.

44. West, G.; Nymalm, Y.; Airenne, T. T.; Kidron, H.; Mattjus, P.; Salminen, T. T., Crystallization and x-ray analysis of bovine glycolipid transfer protein. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004,** D60, (4), 703-705.

45. Konstantinidis, K. T.; Tiedje, J. M., Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **2005,** 187, (18), 6258-64.

46. Konstantinidis, K. T.; Tiedje, J. M., Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* **2005,** 102, (7), 2567-72.

47. Rost, B., Enzyme function less conserved than anticipated. *Journal of Molecular Biology* **2002,** 318, (2), 595-608.

48. Powers, R.; Mirkovic, N.; Goldsmith-Fischman, S.; Acton, T. B.; Chiang, Y.; Huang, Y. J.; Ma, L.; Rajan, P. K.; Cort, J. R.; Kennedy, M. A.; Liu, J.; Rost, B.; Honig, B.; Murray, D.; Montelione, G. T., Solution structure of Archaeglobus fulgidis peptidyl-tRNA hydrolase (Pth2) provides evidence for an extensive

conserved family of Pth2 enzymes in archea, bacteria, and eukaryotes. *Protein Sci* **2005,** 14, (11), 2849-61.

49. Yang, K.; Eyobo, Y.; Brand, L. A.; Martynowski, D.; Tomchick, D.; Strauss, E.; Zhang, H., Crystal structure of a type III pantothenate kinase: insight into the mechanism of an essential coenzyme A biosynthetic enzyme universally distributed in bacteria. *J Bacteriol* **2006,** 188, (15), 5532-40.

50. Kisselev, L., Polypeptide release factors in prokaryotes and eukaryotes: same function, different structure. *Structure* **2002,** 10, (1), 8-9.

51. Ibba, M.; Morgan, S.; Curnow, A. W.; Pridmore, D. R.; Vothknecht, U. C.; Gardner, W.; Lin, W.; Woese, C. R.; Soll, D., A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. *Science* **1997,** 278, (5340), 1119-22.

52. Hadley, C.; Jones, D. T., A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure* **1999,** 7, (9), 1099-112.

53. Pascual-Garcia, A.; Abia, D.; Ortiz, A. R.; Bastolla, U., Cross-over between discrete and continuous protein structure space: Insights into automatic classification and networks of protein structures. *PLoS Comput. Biol.* **2009,** 5, (3), No pp given.

54. Lin, Y.-S.; Hsu, W.-L.; Hwang, J.-K.; Li, W.-H., Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Molecular Biology and Evolution* **2007,** 24, (4), 1005-1011.

55. Chirpich, T. P., Rates of protein evolution. Function of amino acid composition. *Science (Washington, DC, United States)* **1975,** 188, (4192), 1022-3.

56. Chothia, C.; Lesk, A. M., The relation between the divergence of sequence and structure in proteins. *EMBO Journal* **1986,** 5, (4), 823-6.

57. Lesk, A. M.; Chothia, C., How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *Journal of Molecular Biology* **1980,** 136, (3), 225-70.

58. Robertson, A. D.; Murphy, K. P., Protein Structure and the Energetics of Protein Stability. *Chem Rev* **1997,** 97, (5), 1251-1268.

59. Sanchez, I. E.; Tejero, J.; Gomez-Moreno, C.; Medina, M.; Serrano, L., Point mutations in protein globular domains: contributions from function, stability and misfolding. *J Mol Biol* **2006,** 363, (2), 422-32.

60. Chandonia, J. M.; Brenner, S. E., Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. *Proteins* **2005,** 58, (1), 166-79.

61. Mestres, J., Representativity of target families in the Protein Data Bank: impact for family-directed structure-based drug discovery. *Drug Discovery Today* **2005,** 10, (23/24), 1629-1637.

Appendix 6A. A table of all manually curated proteins used in chapter 6 with their associated COG annotation, phylogenetic pattern, phylum classification and source organism.

Split COG	PDB	Phylum	Source
28	2JI7	*Proteobacteria*	*OXALOBACTER FORMIGENES*
28	2AG0	*Proteobacteria*	*PSEUDOMONAS FLUORESCENS*
28	1YNO	*Proteobacteria*	*PSEUDOMONAS PUTIDA*
28	1OZF	*Proteobacteria*	*KLEBSIELLA PNEUMONIAE*
28	1V5E	*Firmicutes*	*AEROCOCCUS VIRIDANS*
28	1POW	*Firmicutes*	*LACTOBACILLUS PLANTARUM*
39	2PWZ	*Proteobacteria*	*ESCHERICHIA COLI*
39	1B8P	*Proteobacteria*	*AQUASPIRILLUM ARCTICUM*
39	1Y6J	*Firmicutes*	*CLOSTRIDIUM THERMOCELLUM*
39	1LDN	*Firmicutes*	*BACILLUS STEAROTHERMOPHILUS*
39	1EZ4	*Firmicutes*	*LACTOBACILLUS PENTOSUS*
394	2GI4	*Proteobacteria*	*CAMPYLOBACTER JEJUNI*
394	2FEK	*Proteobacteria*	*ESCHERICHIA COLI*
394	1LJL	*Firmicutes*	*STAPHYLOCOCCUS AUREUS*
394	1JL3	*Firmicutes*	*BACILLUS SUBTILIS*
446	2V3A	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
446	1Q1R	*Proteobacteria*	*PSEUDOMONAS PUTIDA*
446	1D7Y	*Proteobacteria*	*PARACOCCUS PANTOTROPHUS*
446	2CDU	*Firmicutes*	*LACTOBACILLUS SANFRANCISCENSIS*
446	2BC0	*Firmicutes*	*STREPTOCOCCUS PYOGENES*
446	1YQZ	*Firmicutes*	*STAPHYLOCOCCUS AUREUS*
446	1F8W	*Firmicutes*	*ENTEROCOCCUS FAECALIS*
604	1WLY	*Proteobacteria*	*BURKHOLDERIA SP. WS*
604	1QOR	*Proteobacteria*	*ESCHERICHIA COLI*
604	1XA0	*Firmicutes*	*BACILLUS STEAROTHERMOPHILUS*
604	1TT7	*Firmicutes*	*BACILLUS SUBTILIS*
605	2BKB	*Proteobacteria*	*ESCHERICHIA COLI*
605	1DT0	*Proteobacteria*	*PSEUDOMONAS PUTIDA*
605	1XRE	*Firmicutes*	*BACILLUS ANTHRACIS*
605	1JR9	*Firmicutes*	*BACILLUS HALODENITRIFICANS*

742	2IFT	*Proteobacteria*	*HAEMOPHILUS INFLUENZAE*
742	2FPO	*Proteobacteria*	*ESCHERICHIA COLI*
742	2FHP	*Firmicutes*	*ENTEROCOCCUS FAECALIS*
742	2ESR	*Firmicutes*	*STREPTOCOCCUS PYOGENES*
813	1VHJ	*Proteobacteria*	*VIBRIO CHOLERAE*
813	1ECP	*Proteobacteria*	*ESCHERICHIA COLI*
813	2AC7	*Firmicutes*	*BACILLUS CEREUS G9241*
813	1XE3	*Firmicutes*	*BACILLUS ANTHRACIS*
1012	2HG2	*Proteobacteria*	*ESCHERICHIA COLI*
1012	1EYY	*Proteobacteria*	*VIBRIO HARVEYI*
1012	1T90	*Firmicutes*	*BACILLUS SUBTILIS*
1012	1EUH	*Firmicutes*	*STREPTOCOCCUS MUTANS*
1057	1YUM	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
1057	1K4M	*Proteobacteria*	*ESCHERICHIA COLI*
1057	2H2A	*Firmicutes*	*STAPHYLOCOCCUS AUREUS*
1057	1KAQ	*Firmicutes*	*BACILLUS SUBTILIS*
1075	1TAH	*Proteobacteria*	*BURKHOLDERIA GLUMAE*
1075	1OIL	*Proteobacteria*	*BURKHOLDERIA CEPACIA*
1075	1EX9	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
1075	2HIH	*Firmicutes*	*STAPHYLOCOCCUS HYICUS*
1075	1KU0	*Firmicutes*	*BACILLUS STEAROTHERMOPHILUS*
1607	2GVH	*Proteobacteria*	*AGROBACTERIUM TUMEFACIENS STR. C58*
1607	1YLI	*Proteobacteria*	*HAEMOPHILUS INFLUENZAE*
1607	1Y7U	*Firmicutes*	*BACILLUS CEREUS*
1607	1VPM	*Firmicutes*	*BACILLUS HALODURANS CProteobacteria25*
1940	1Z6R	*Proteobacteria*	*ESCHERICHIA COLI*
1940	1Z05	*Proteobacteria*	*VIBRIO CHOLERAE O1 BIOVAR ELTOR*
1940	2QM1	*Firmicutes*	*ENTEROCOCCUS FAECALIS V583*
1940	2GUP	*Firmicutes*	*STREPTOCOCCUS PNEUMONIAE*
1940	1XC3	*Firmicutes*	*BACILLUS SUBTILIS*
2124	1YRD	*Proteobacteria*	*PSEUDOMONAS PUTIDA*
2124	1T2B	*Proteobacteria*	*CITROBACTER BRAAKII*
2124	1Q5E	*Proteobacteria*	*POLYANGIUM CELLULOSUM*
2124	1IZO	*Firmicutes*	*BACILLUS SUBTILIS*
2124	1FAG	*Firmicutes*	*BACILLUS MEGATERIUM*

2188	2PKH	*Proteobacteria*	*PSEUDOMONAS SYRINGAE PV. TOMATO STR. DC3000*
2188	2FA1	*Proteobacteria*	*ESCHERICHIA COLI*
2188	2OOI	*Firmicutes*	*STAPHYLOCOCCUS AUREUS*
2188	2OGG	*Firmicutes*	*BACILLUS SUBTILIS*

Split with HGT

COG	PDB	Phylum	Source
242	2EW7	*Proteobacteria*	*HELICOBACTER PYLORI*
242	1IX1	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
242	1ICJ	*Proteobacteria*	*ESCHERICHIA COLI*
242	2AI9	*Firmicutes*	*STAPHYLOCOCCUS AUREUS*
242	1WS0	*Firmicutes*	*BACILLUS CEREUS*
242	1LQY	*Firmicutes*	*BACILLUS STEAROTHERMOPHILUS*
242	1LM6	*Firmicutes*	*STREPTOCOCCUS PNEUMONIAE*
1052	2GSD	*Proteobacteria*	*MORAXELLA SP.*
1052	2GO1	*Proteobacteria*	*PSEUDOMONAS SP.*
1052	1GDH	*Proteobacteria*	*HYPHOMICROBIUM METHYLOVORUM*
1052	2DLD	*Firmicutes*	*LACTOBACILLUS HELVETICUS*
1052	1XDW	*Firmicutes*	*ACIDAMINOCOCCUS FERMENTANS*
1052	1J4A	*Firmicutes*	*LACTOBACILLUS DELBRUECKII SUBSP. BULGARICUS*
2141	2I7G	*Proteobacteria*	*AGROBACTERIUM TUMEFACIENS*
2141	1M41	*Proteobacteria*	*ESCHERICHIA COLI*
2141	1BRL	*Proteobacteria*	*VIBRIO HARVEYI*
2141	2B81	*Firmicutes*	*BACILLUS CEREUS*
2141	1TVL	*Firmicutes*	*BACILLUS SUBTILIS*
3832	1Z94	*Proteobacteria*	*CHROMOBACTERIUM VIOLACEUM ATCC 12472*
3832	1XFS	*Proteobacteria*	*NITROSOMONAS EUROPAEA*
3832	2NN5	*Firmicutes*	*ENTEROCOCCUS FAECALIS*
3832	2IL5	*Firmicutes*	*STAPHYLOCOCCUS AUREUS*
3832	1XN6	*Firmicutes*	*BACILLUS CEREUS*
3832	1XN5	*Firmicutes*	*BACILLUS HALODURANS*

Starburst

COG	PDB	Phylum	Source
110	2NPO	*Proteobacteria*	*CAMPYLOBACTER JEJUNI*
110	1KRR	*Proteobacteria*	*ESCHERICHIA COLI*
110	2IC7	*Firmicutes*	*GEOBACILLUS KAUSTOPHILUS*
110	1KK5	*Firmicutes*	*ENTEROCOCCUS FAECIUM*

171	1XNG	*Proteobacteria*	*HELICOBACTER PYLORI*
171	1WXE	*Proteobacteria*	*ESCHERICHIA COLI*
171	2PZB	*Firmicutes*	*BACILLUS ANTHRACIS*
171	1KQP	*Firmicutes*	*BACILLUS SUBTILIS*
251	2IG8	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
251	1QU9	*Proteobacteria*	*ESCHERICHIA COLI*
251	1J7H	*Proteobacteria*	*HAEMOPHILUS INFLUENZAE*
251	2EWC	*Firmicutes*	*STREPTOCOCCUS PYOGENES*
251	1XRG	*Firmicutes*	*CLOSTRIDIUM THERMOCELLUM*
251	1QD9	*Firmicutes*	*BACILLUS SUBTILIS*
346	2PJS	*Proteobacteria*	*AGROBACTERIUM TUMEFACIENS*
346	1R9C	*Proteobacteria*	*RHIZOBIUM LOTI*
346	1NPB	*Proteobacteria*	*SERRATIA MARCESCENS*
346	1MPY	*Proteobacteria*	*PSEUDOMONAS PUTIDA*
346	1LQK	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
346	1LGT	*Proteobacteria*	*BURKHOLDERIA SP.*
346	1KMY	*Proteobacteria*	*BURKHOLDERIA CEPACIA*
346	1F9Z	*Proteobacteria*	*ESCHERICHIA COLI*
346	1EIL	*Proteobacteria*	*PSEUDOMONAS SP.*
346	1ECS	*Proteobacteria*	*KLEBSIELLA PNEUMONIAE*
346	2QQZ	*Firmicutes*	*BACILLUS ANTHRACIS STR. AMES*
346	2QH0	*Firmicutes*	*CLOSTRIDIUM ACETOBUTYLICUM*
346	2P7K	*Firmicutes*	*LISTERIA MONOCYTOGENES*
346	2P25	*Firmicutes*	*ENTEROCOCCUS FAECALIS*
346	2I7R	*Firmicutes*	*STREPTOCOCCUS PNEUMONIAE*
346	1ZSW	*Firmicutes*	*BACILLUS CEREUS*
346	1SS4	*Firmicutes*	*BACILLUS CEREUS*
366	1ZJA	*Proteobacteria*	*PSEUDOMONAS MESOACIDOPHILA*
366	1M53	*Proteobacteria*	*KLEBSIELLA SP. LX3*
366	1G5A	*Proteobacteria*	*NEISSERIA POLYSACCHAREA*
366	1B0I	*Proteobacteria*	*PSEUDOALTEROMONAS HALOPLANKTIS*
366	1WZA	*Firmicutes*	*HALOTHERMOTHRIX ORENII*
366	1W9X	*Firmicutes*	*BACILLUS HALMAPALUS*
366	1UA7	*Firmicutes*	*BACILLUS SUBTILIS*
366	1PAM	*Firmicutes*	*BACILLUS SP.*
366	1OT2	*Firmicutes*	*BACILLUS CIRCULANS*
366	1JI1	*Firmicutes*	*THERMOACTINOMYCES VULGARIS*
366	1HVX	*Firmicutes*	*BACILLUS STEAROTHERMOPHILUS*
366	1E3X	*Firmicutes*	*BACILLUS AMYLOLIQUEFACIENS*
366	1BPL	*Firmicutes*	*BACILLUS LICHENIFORMIS*

454	2Q0Y	*Proteobacteria*	*RALSTONIA EUTROPHA JMP134*
454	2OZH	*Proteobacteria*	*XANTHOMONAS CAMPESTRIS PV. CAMPESTRIS*
454	2GE3	*Proteobacteria*	*AGROBACTERIUM TUMEFACIENS*
454	2FT0	*Proteobacteria*	*ESCHERICHIA COLI*
454	2FIW	*Proteobacteria*	*RHODOPSEUDOMONAS PALUSTRIS CGA009*
454	2EUI	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
454	1S3Z	*Proteobacteria*	*SALMONELLA ENTERITIDIS*
454	1GHE	*Proteobacteria*	*PSEUDOMONAS SYRINGAE PV. TABACI*
454	2PC1	*Firmicutes*	*STREPTOCOCCUS AGALACTIAE 2603V/R*
454	2OH1	*Firmicutes*	*LISTERIA MONOCYTOGENES STR. 4B F2365*
454	2JDC	*Firmicutes*	*BACILLUS LICHENIFORMIS*
454	2ATR	*Firmicutes*	*STREPTOCOCCUS PNEUMONIAE TIGR4*
454	2AJ6	*Firmicutes*	*STAPHYLOCOCCUS AUREUS*
454	1Z4E	*Firmicutes*	*BACILLUS HALODURANS*
454	1Y9K	*Firmicutes*	*BACILLUS CEREUS*
454	1U6M	*Firmicutes*	*ENTEROCOCCUS FAECALIS*
454	1TIQ	*Firmicutes*	*BACILLUS SUBTILIS*
491	2OBW	*Proteobacteria*	*SALMONELLA TYPHIMURIUM*
491	2GMN	*Proteobacteria*	*BRADYRHIZOBIUM JAPONICUM*
491	2FM6	*Proteobacteria*	*XANTHOMONAS MALTOPHILIA*
491	2FHX	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
491	1X8G	*Proteobacteria*	*AEROMONAS HYDROPHILA*
491	1WUO	*Proteobacteria*	*SERRATIA MARCESCENS*
491	1P9E	*Proteobacteria*	*PSEUDOMONAS SP.*
491	1K07	*Proteobacteria*	*FLUORIBACTER GORMANII*
491	2BTN	*Firmicutes*	*BACILLUS THURINGIENSIS*
491	1BC2	*Firmicutes*	*BACILLUS CEREUS*
500	2PKW	*Proteobacteria*	*SALMONELLA TYPHIMURIUM*
500	2P7I	*Proteobacteria*	*ERWINIA CAROTOVORA SUBSP. ATROSEPTICA SCRI1043*
500	2OYR	*Proteobacteria*	*SHIGELLA FLEXNERI 2A*
500	2IP2	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
500	1PJZ	*Proteobacteria*	*PSEUDOMONAS SYRINGAE PV. PISI*
500	1NKV	*Proteobacteria*	*ESCHERICHIA COLI*
500	1IM8	*Proteobacteria*	*HAEMOPHILUS INFLUENZAE*
500	2P8J	*Firmicutes*	*CLOSTRIDIUM ACETOBUTYLICUM*
500	2GH1	*Firmicutes*	*BACILLUS CEREUS*
500	1XXL	*Firmicutes*	*BACILLUS SUBTILIS*
500	1VL5	*Firmicutes*	*BACILLUS HALODURANS CProteobacteria25*
526	2TRX	*Proteobacteria*	*ESCHERICHIA COLI*

526	2I4A	*Proteobacteria*	*ACETOBACTER ACETI*
526	2O7K	*Firmicutes*	*STAPHYLOCOCCUS AUREUS*
526	2GZY	*Firmicutes*	*BACILLUS SUBTILIS*
526	1NW2	*Firmicutes*	*ALICYCLOBACILLUS ACIDOCALDARIUS*
590	2G84	*Proteobacteria*	*NITROSOMONAS EUROPAEA*
590	2A8N	*Proteobacteria*	*AGROBACTERIUM TUMEFACIENS*
590	1Z3A	*Proteobacteria*	*ESCHERICHIA COLI*
590	2NX8	*Firmicutes*	*STREPTOCOCCUS PYOGENES SEROTYPE M6*
590	1WKQ	*Firmicutes*	*BACILLUS SUBTILIS*
637	2FDR	*Proteobacteria*	*AGROBACTERIUM TUMEFACIENS*
637	1TE2	*Proteobacteria*	*ESCHERICHIA COLI O157:H7*
637	1RQN	*Firmicutes*	*BACILLUS CEREUS*
637	1LVH	*Firmicutes*	*LACTOCOCCUS LACTIS*
664	1VP6	*Proteobacteria*	*RHIZOBIUM LOTI*
664	1U12	*Proteobacteria*	*MESORHIZOBIUM LOTI MAFF303099*
664	1G6N	*Proteobacteria*	*ESCHERICHIA COLI*
664	1FT9	*Proteobacteria*	*RHODOSPIRILLUM RUBRUM*
664	2HKX	*Firmicutes*	*CARBOXYDOTHERMUS HYDROGENOFORMANS*
664	1OMI	*Firmicutes*	*LISTERIA MONOCYTOGENES*
745	2PLN	*Proteobacteria*	*HELICOBACTER PYLORI*
745	1XHF	*Proteobacteria*	*ESCHERICHIA COLI*
745	2A9O	*Firmicutes*	*STREPTOCOCCUS PNEUMONIAE*
745	1MVO	*Firmicutes*	*BACILLUS SUBTILIS*
753	2ISA	*Proteobacteria*	*VIBRIO SALMONICIDA*
753	1QWL	*Proteobacteria*	*HELICOBACTER PYLORI*
753	1M85	*Proteobacteria*	*PROTEUS MIRABILIS*
753	1GGE	*Proteobacteria*	*ESCHERICHIA COLI*
753	2J2M	*Firmicutes*	*EXIGUOBACTERIUM OXIDOTOLERANS*
753	1SI8	*Firmicutes*	*ENTEROCOCCUS FAECALIS*
778	1VFR	*Proteobacteria*	*VIBRIO FISCHERI*
778	2ISJ	*Proteobacteria*	*SINORHIZOBIUM MELILOTI*
778	1KQD	*Proteobacteria*	*ENTEROBACTER CLOACAE*
778	1F5V	*Proteobacteria*	*ESCHERICHIA COLI*
778	2H0U	*Proteobacteria*	*HELICOBACTER PYLORI*
778	2HAY	*Firmicutes*	*STREPTOCOCCUS PYOGENES SEROTYPE M1*
778	2B67	*Firmicutes*	*STREPTOCOCCUS PNEUMONIAE TIGR4*
778	1ZCH	*Firmicutes*	*BACILLUS SUBTILIS*

778	2I7H	*Firmicutes*	*BACILLUS CEREUS*
784	2FKA	*Proteobacteria*	*SALMONELLA ENTERICA SUBSP. ENTERICA SEROVAR TYPHIMURIUM*
784	1P6Q	*Proteobacteria*	*RHIZOBIUM MELILOTI*
784	1M5T	*Proteobacteria*	*CAULOBACTER CRESCENTUS*
784	6CHY	*Proteobacteria*	*ESCHERICHIA COLI*
784	2I6F	*Proteobacteria*	*MYXOCOCCUS XANTHUS*
784	1F51	*Firmicutes*	*BACILLUS SUBTILIS*
784	1QMP	*Firmicutes*	*BACILLUS STEAROTHERMOPHILUS*
796	2JFN	*Proteobacteria*	*ESCHERICHIA COLI*
796	2JFX	*Proteobacteria*	*HELICOBACTER PYLORI*
796	2JFO	*Firmicutes*	*ENTEROCOCCUS FAECALIS*
796	2JFQ	*Firmicutes*	*STAPHYLOCOCCUS AUREUS*
796	2GZM	*Firmicutes*	*BACILLUS ANTHRACIS*
1028	2EWM	*Proteobacteria*	*AZOARCUS*
1028	2DKN	*Proteobacteria*	*PSEUDOMONAS SP.*
1028	2CFC	*Proteobacteria*	*XANTHOBACTER AUTOTROPHICUS*
1028	2B4Q	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
1028	1ZEM	*Proteobacteria*	*GLUCONOBACTER OXYDANS*
1028	1WMB	*Proteobacteria*	*PSEUDOMONAS FRAGI*
1028	1PWX	*Proteobacteria*	*AGROBACTERIUM TUMEFACIENS*
1028	1K2W	*Proteobacteria*	*RHODOBACTER SPHAEROIDES*
1028	1GEG	*Proteobacteria*	*KLEBSIELLA PNEUMONIAE*
1028	1FJH	*Proteobacteria*	*COMAMONAS TESTOSTERONI*
1028	1AHH	*Proteobacteria*	*ESCHERICHIA COLI*
1028	2UVD	*Firmicutes*	*BACILLUS ANTHRACIS*
1028	2HQ1	*Firmicutes*	*CLOSTRIDIUM THERMOCELLUM*
1028	1NXQ	*Firmicutes*	*LACTOBACILLUS BREVIS*
1028	1G6K	*Firmicutes*	*BACILLUS MEGATERIUM*
1151	1JQK	*Proteobacteria*	*RHODOSPIRILLUM RUBRUM*
1151	1E2U	*Proteobacteria*	*DESULFOVIBRIO VULGARIS*
1151	1OA0	*Proteobacteria*	*DESULFOVIBRIO DESULFURICANS*
1151	1SU6	*Firmicutes*	*CARBOXYDOTHERMUS HYDROGENOFORMANS*
1151	1OAO	*Firmicutes*	*MOORELLA THERMOACETICA*
1309	2UXH	*Proteobacteria*	*PSEUDOMONAS PUTIDA*
1309	2HYT	*Proteobacteria*	*ERWINIA CAROTOVORA SUBSP. ATROSEPTICA*

1309	2G7S	*Proteobacteria*	*AGROBACTERIUM TUMEFACIENS*
1309	2FBQ	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
1309	1T33	*Proteobacteria*	*SALMONELLA TYPHIMURIUM*
1309	1PB6	*Proteobacteria*	*ESCHERICHIA COLI*
1309	2IU5	*Firmicutes*	*LACTOCOCCUS LACTIS SUBSP. LACTIS IL1403*
1309	2FX0	*Firmicutes*	*BACILLUS CEREUS*
1309	1Z0X	*Firmicutes*	*ENTEROCOCCUS FAECALIS V583*
1309	1VI0	*Firmicutes*	*BACILLUS SUBTILIS*
1396	1Y9Q	*Proteobacteria*	*VIBRIO CHOLERAE*
1396	1Y7Y	*Proteobacteria*	*AEROMONAS HYDROPHILA*
1396	2P5T	*Firmicutes*	*STREPTOCOCCUS PNEUMONIAE*
1396	2B5A	*Firmicutes*	*BACILLUS CALDOLYTICUS*
1396	1B0N	*Firmicutes*	*BACILLUS SUBTILIS*
1404	2B6N	*Proteobacteria*	*SERRATIA SP.*
1404	1V6C	*Proteobacteria*	*PSEUDOALTEROMONAS SP. ASProteobacteria1*
1404	1S2N	*Proteobacteria*	*VIBRIO SP. PA-44*
1404	3TEC	*Firmicutes*	*HIRUDINARIA MANILLENSIS*
1404	2SIC	*Firmicutes*	*BACILLUS AMYLOLIQUEFACIENS*
1404	2IXT	*Firmicutes*	*BACILLUS SPHAERICUS*
1404	1YU6	*Firmicutes*	*MELEAGRIS GALLOPAVO*
1404	1XF1	*Firmicutes*	*STREPTOCOCCUS PYOGENES*
1404	1V5I	*Firmicutes*	*PLEUROTUS OSTREATUS*
1404	1TEC	*Firmicutes*	*THERMOACTINOMYCES VULGARIS*
1404	1SEL	*Firmicutes*	*BACILLUS SUBTILIS*
1404	1SBN	*Firmicutes*	*HIRUDO MEDICINALIS*
1404	1MEE	*Firmicutes*	*BACILLUS PUMILUS*
1404	1IAV	*Firmicutes*	*BACILLUS LENTUS*
1404	1DBI	*Firmicutes*	*BACILLUS SP.*
1404	1BH6	*Firmicutes*	*BACILLUS LICHENIFORMIS*
1733	2F2E	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
1733	1YYV	*Proteobacteria*	*SALMONELLA TYPHIMURIUM*
1733	2HZT	*Firmicutes*	*BACILLUS SUBTILIS*
1733	1Z7U	*Firmicutes*	*ENTEROCOCCUS FAECALIS V583*
1846	2FBH	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
1846	2FA5	*Proteobacteria*	*XANTHOMONAS CAMPESTRIS*
1846	1JGS	*Proteobacteria*	*ESCHERICHIA COLI*
1846	2QWW	*Firmicutes*	*LISTERIA MONOCYTOGENES STR. 4B F2365*
1846	2BV6	*Firmicutes*	*STAPHYLOCOCCUS AUREUS*

1846	1Z91	*Firmicutes*	*BACILLUS SUBTILIS*
1846	1LJ9	*Firmicutes*	*ENTEROCOCCUS FAECALIS*
2159	2HBV	*Proteobacteria*	*PSEUDOMONAS FLUORESCENS*
2159	2DVT	*Proteobacteria*	*RHIZOBIUM SP.*
2159	2QPX	*Firmicutes*	*LACTOBACILLUS CASEI ATCC 334*
2159	2F6K	*Firmicutes*	*LACTOBACILLUS PLANTARUM*
2367	1N4O	*Proteobacteria*	*XANTHOMONAS MALTOPHILIA*
2367	1JTG	*Proteobacteria*	*STREPTOMYCES CLAVULIGERUS*
2367	1JTD	*Proteobacteria*	*STREPTOMYCES EXFOLIATUS*
2367	1HZO	*Proteobacteria*	*PROTEUS VULGARIS*
2367	1HTZ	*Proteobacteria*	*KLEBSIELLA PNEUMONIAE*
2367	1G68	*Proteobacteria*	*PSEUDOMONAS AERUGINOSA*
2367	1FQG	*Proteobacteria*	*ESCHERICHIA COLI*
2367	1DY6	*Proteobacteria*	*SERRATIA MARCESCENS*
2367	1BUE	*Proteobacteria*	*ENTEROBACTER CLOACAE*
2367	1KGG	*Firmicutes*	*STAPHYLOCOCCUS AUREUS*
2367	1I2S	*Firmicutes*	*BACILLUS LICHENIFORMIS*
2730	1TVN	*Proteobacteria*	*PSEUDOALTEROMONAS HALOPLANKTIS*
2730	1EGZ	*Proteobacteria*	*ERWINIA CHRYSANTHEMI*
2730	2JEP	*Firmicutes*	*PAENIBACILLUS PABULI*
2730	1QHZ	*Firmicutes*	*BACILLUS AGARADHAERENS*
2730	1LF1	*Firmicutes*	*BACILLUS SUBTILIS*
2730	1G01	*Firmicutes*	*BACILLUS SP.*

Starburst

COG	PDB	Phylum	Source
3693	2CNC	*Proteobacteria*	*CELLVIBRIO MIXTUS*
3693	1US3	*Proteobacteria*	*CELLVIBRIO JAPONICUS*
3693	1E5N	*Proteobacteria*	*PSEUDOMONAS FLUORESCENS*
3693	2F8Q	*Firmicutes*	*BACILLUS SP. NG-27*
3693	2DEP	*Firmicutes*	*CLOSTRIDIUM STERCORARIUM*
3693	1R85	*Firmicutes*	*BACILLUS STEAROTHERMOPHILUS*
4948	2QDE	*Proteobacteria*	*AZOARCUS SP. EBN1*
4948	2PPG	*Proteobacteria*	*SINORHIZOBIUM MELILOTI*
4948	2PMQ	*Proteobacteria*	*ROSEOVARIUS SP. HTCC2601*
4948	2PGE	*Proteobacteria*	*DESULFOTALEA PSYCHROPHILA LSV54*
4948	2PCE	*Proteobacteria*	*ROSEOVARIUS NUBINHIBENS ISM*
4948	2OZ8	*Proteobacteria*	*MESORHIZOBIUM LOTI*
4948	2OZ3	*Proteobacteria*	*AZOTOBACTER VINELANDII AVOP*

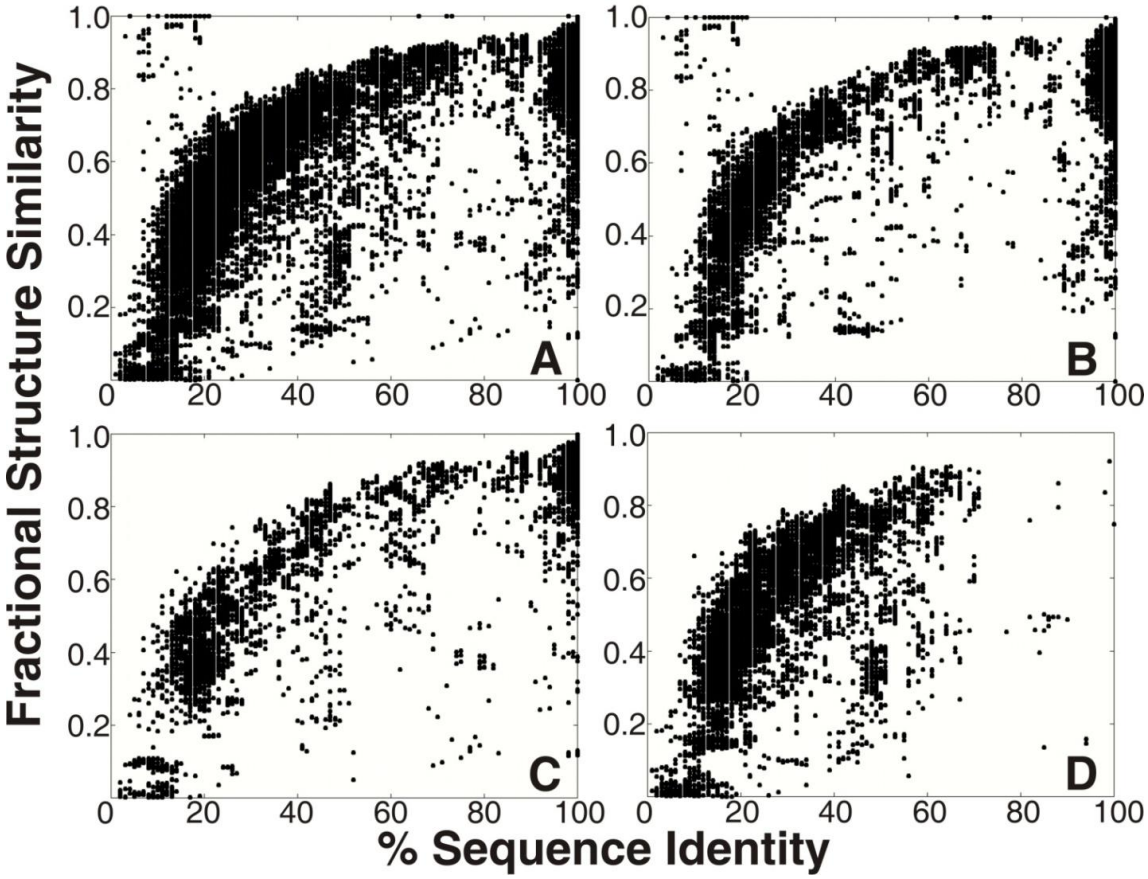
4948	2OX4	*Proteobacteria*	*ZYMOMONAS MOBILIS*
4948	2OO6	*Proteobacteria*	*BURKHOLDERIA XENOVORANS*
4948	2OG9	*Proteobacteria*	*POLAROMONAS SP. JS666*
4948	2NQL	*Proteobacteria*	*AGROBACTERIUM TUMEFACIENS*
4948	2HZG	*Proteobacteria*	*RHODOBACTER SPHAEROIDES*
4948	2GSH	*Proteobacteria*	*SALMONELLA TYPHIMURIUM*
4948	2DW6	*Proteobacteria*	*BRADYRHIZOBIUM JAPONICUM*
4948	1YEY	*Proteobacteria*	*XANTHOMONAS CAMPESTRIS PV. CAMPESTRIS*
4948	1NU5	*Proteobacteria*	*PSEUDOMONAS SP.*
4948	1MUC	*Proteobacteria*	*PSEUDOMONAS PUTIDA*
4948	1EC7	*Proteobacteria*	*ESCHERICHIA COLI*
4948	1CHR	*Proteobacteria*	*RALSTONIA EUTROPHA*
4948	2P88	*Firmicutes*	*BACILLUS CEREUS ATCC 14579*
4948	2OQY	*Firmicutes*	*OCEANOBACILLUS IHEYENSIS*
4948	2OKT	*Firmicutes*	*STAPHYLOCOCCUS AUREUS*
4948	2GGE	*Firmicutes*	*BACILLUS SUBTILIS*
4948	2GDQ	*Firmicutes*	*BACILLUS SUBTILIS SUBSP. SUBTILIS*
4948	1WUF	*Firmicutes*	*LISTERIA INNOCUA CLIP11262*
4948	1WUE	*Firmicutes*	*ENTEROCOCCUS FAECALIS*
4948	1JPM	*Firmicutes*	*BACILLUS SUBTILIS*

Appendix 6B: The median E-value for each COG. A median of 0 represents all proteins within the COG gave a perfect match to the PDB.

COG	Median E-value
28	1E-175
39	1E-121
110	1E-74
171	2.5E-147
242	7E-90
251	4.9E-66
346	1E-56
366	1E-138
394	1.51E-80
446	5E-96
454	4E-72
491	4.5E-26
500	6E-136
526	7E-55
590	8E-78
604	3.5E-113
605	5E-91
637	1E-120
664	4E-16
742	2.5E-89
745	1E-61
753	0
778	1E-117
784	2E-58
796	7E-128
813	4.5E-97
1012	0
1028	3E-66
1052	1E-56
1057	3.5E-109
1075	1E-96
1151	1.5E-147
1309	5E-96
1396	4E-58
1404	1.35E-87
1607	3.5E-72
1733	3E-52
1846	1E-69
1940	4E-155

2124	2E-48
2141	1.5E-58
2159	1.5E-21
2188	5.02E-84
2367	2E-58
2730	2.1E-59
3693	2E-48
3832	3.75E-50
4948	1E-140

Appednix 6C: The complete Fractional Structure Similarity (FSS) compared to sequence identity prior to manual filtering. As in figure 6.3, (A) is all vs. all comparisons, (B) is the comparisons of *Proteobacteria* structure against *Proteobacteria* structure, (C) is the *Firmicutes* against *Firmicutes* and (D) is the *Proteobacteria* against the *Firmicutes*. As stated in the text above the comparisons between *Proteobacteria* and *Firmicutes* show an abrupt cutoff at about 65% sequence identity and 0.85 Fraction Structure Similarity. Outliers were shown to be comparisons of the same protein from the same organism solved under non-uniform conditions. The large density of structures a 100% sequence identity illustrates the propensity of solving structures redundantly from the same organism and the large spread of data shows the need for manual curation of the dataset



CHAPTER 7:

A SEQUENCE AND STRUCTURE INDEPENDENT METHOD TO PREDICT PROTEIN FUNCTION

7.1 INTRODUCTION

The recent explosion in sequenced genomes has revealed a vast number of proteins that lack a functional annotation.[1] Many of these unannotated proteins may play an important role in human disease and correspondingly, are critical for developing new therapeutics. Protein sequence and structure similarity methods are currently the most robust and widely-used tools to annotate a protein of unknown function.[2] Nevertheless, these methods are limited in scope, prone to errors, and based on a small set of experimentally characterized proteins.[3] Only 40 to 60% of sequences suggest a potential functional assignment. Moreover, error rates of < 30% occur even with conservative sequence identities of > 60%. The accuracy of functional annotations decreases substantially in the twilight zone of 20-35% sequence identity.

Recent attempts to extend functional prediction beyond global sequence and structure similarity have led to the development of active-site similarity search methods.[4-6] These methods try to identify protein surface structures that interact with biologically important compounds or other proteins. Protein active-sites that share similar sequence, structure and bind similar ligands are predicted to be functionally related. While promising, current active-site similarity techniques still rely on high-resolution protein structures to identify and measure functional similarity.[7, 8] The availability of structures for the entire proteome remains a significant bottleneck for high-throughput functional annotation of hypothetical proteins.

In the previous chapters, functional annotation of proteins was discussed in the presence of sequence, structure and active site information. As shown in chapters 1,3 and 4 these methods are powerful, but have limitations that prevent complete annotation of a specific genome in a high-throughput manner. Specifically, sequence similarity methods often fail below 30% sequence identity[9, 10] and structure similarity or active site similarity methods require a high-resolution protein structure. Additionally, functional similarity is not necessarily dependent on homology. This can lead to similar sequences having different functions or significantly different sequences with similar functions.[9, 11] The issues raised above suggest a new approach to function annotation that is independent of sequence or structure is needed.

Proteins interact with biological compounds to perform specific yet versatile functions. Identifying and comparing which compounds bind a target protein provides an alternative method to predict function. In this chapter I discuss the development of a quantifiable and rapidly adaptable model for protein functional analysis using experimentally derived ligand binding profiles (LBP). This new approach is independent of sequence, structural or evolutionary information; therefore, extending the current analysis of novel genes and predicting ligand binding. A ligand binding profile is defined as a set of ligands that bind a protein from a high-throughput ligand affinity screen. The hypothesis is that proteins with similar function will bind a similar set of compounds from the same high-throughput screening library. A general functional similarity is identified by clustering proteins similar binding profiles.

In this chapter, I discuss the theory behind the ligand binding profile method and report screening and similarity results from 19 proteins with a range of functions defined

by Gene Ontology (GO) terms.[12] With the availability of GO terms, many studies relate functional similarity to protein-protein interactions,[13] network prediction,[14] prediction of cellular localization,[15] pathway modeling,[16] and improving the quality of microarray data.[17] This chapter is the first attempt to relate ligand binding similarity to functional similarity.

7.2 THEORY

7.2.1 Development of a ligand binding profile scoring function. Measuring a significant similarity between two ligand binding profiles requires the development or adaptation of a robust scoring function. Current similarity scoring methods used for sequence analysis, such as the E-value developed by Karlin and Altschul,[18] are also well-suited for measuring a similarity between ligand binding profiles.

$$E = Kmne^{-\lambda S} \qquad [7.1]$$

Here, the E-value is only dependent on the total number of compounds that bind each protein (m and n) and the total number of compounds that bind both proteins (S). Additionally, the probability of finding a significant similarity is proportional to the probability search space (K) and scoring function (λ).

$$K = \frac{(q-p)^2}{q} \text{ and } \lambda = \ln \frac{q}{p} \qquad [7.2]$$

Unlike sequence similarity, a similarity between ligand binding can be thought of as a binary system (binding vs. non-binding) therefore the probabilities p and q simply become the probability of finding a hit within a library:

$$p = \frac{1}{\text{library size}} \qquad [7.3]$$

and the probability of finding a ligand that binds both proteins:

$$q = \frac{S}{mn} \qquad [7.4]$$

The standard E-value also provides a robust measure of the probability. This shows a significant ligand binding similarity is not due to chance using the standard P-value.

$$P = 1 - e^{-E} \qquad [7.5]$$

As expected, the ligand binding profile E-value rapidly becomes non-significant (P > 0.0001) as the probability of finding a ligand that binds both proteins (q) decreases (figure 7.1). Binding profiles that have a P < 0.0001 are significant at the 99.99% confidence interval (~E=10^{-5}).

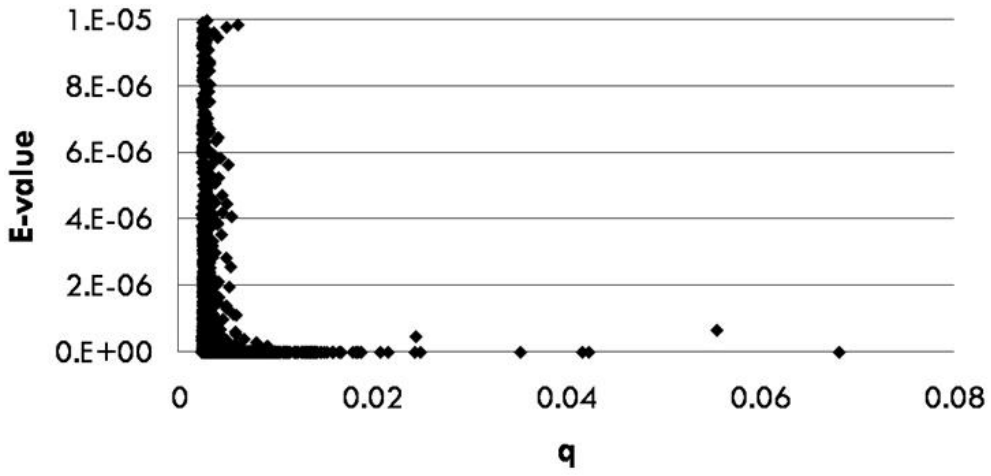


Figure 7.1. E-value response to the probability of finding overlapping ligands between two proteins. A set of 33,207 randomly generated hypothetical binding profiles was generated to observe the response of the E-value similarity with probability of overlapping ligands (q=S/mn) as the probability of finding an overlapping ligand decreases the E-value rapidly becomes non-significant (E>1x10^{-5}).

7.3 EXPERIMENTAL

7.3.1 Hypothetical binding profiles. A set of hypothetical binding profiles was generated to test the E-value scoring method for the ligand binding profiles. To generate the hypothetical binding profiles, an Excel program was written to generate random values for m, n and S for 100,000 hypothetical binding profiles. The hypothetical library size was 437 compounds; random numbers were generated between 0 and 437. The data set was filtered such that $S \leq m$ and $S \leq n$ giving 33,207 comparisons. The data set was used to compare the E-value response to probablilty of finding an overlapping ligand (figure 7.1).

7.3.2 Materials. The human serum albumin (HSA) (essentially fatty acid free, ≥ 96 % pure), bovine serum albumin (BSA) (minimum 98% agarose gel electrophoresis, lyophilized), α-amylase from *Bacillus lincheniformis* (Bli) (500-1,500 units/mg protein, 93-100% (SDS page)), α-amylase from *Aspergillus oryzae* (Aor) (powder, ~30 units/mg), α-amylase from *Bacillus amyloliquefaciens* (Bam) (liquid, ≥250 units/g protein), β-amylase from barley (Hvu) (type II-B 20-80 units/mg protein), and β-amylase from sweet potato (Iba) (Type I-B, ammonium sulfate suspension, ≥750 units/mg protein) protein samples were all purchased from Sigma (St. Louis, MO). The *S. typhimurium* PrgI protein samples and assigned 2D ^{1}H-^{15}N HSQC spectrum were generously provided by Dr. Roberto DeGuzman (University of Kansas). *Staphylococcus aureus* primase C-Terminal domain (CTD) protein sample was purchased from Nature Technologies Corporation (Lincoln, NE). *H. sapiens* diacylglycerol kinase alpha (DGKA), *P. aeruginosa* unannotated protein PA1324, *S. aureus* unannotated protein SAV1430, *S. typhimurium* unannotated protein STM1790, *H. sapiens* ubiquitin-fold modifier-

conjugating enzyme 1 (UFC1), *E. coli* unannotated protein YjbR, *E. coli* unannotated protein YkfF, *B. subtilis* unannotated protein YkvR and *E. coli* unannotated protein YtfP protein samples were provided by Dr. Gaetano Montelione, Director of the Northeast Structural Genomics Consortium (NESG, www.nesg.org). The *S. aureus* nuclease was over-expressed in house from a cell stock of *E. coli* Bl21 DE3 codon+ (Stratagene) containing the pET28(a)+plasmid with the *dnuc* gene provided by Dr. Greg Somerville (University of Nebraska-Lincoln) grown in LB broth and purified using a Talon cobalt affinity resin (Clontech). The deuterium oxide (99.9% D) and the dimethyl sulfoxide-d6 (99.9% D) were purchased from Aldrich (Milwaukee, WI) The 3-(trimethylsilyl)propionic acid-2,2,3,3-d4 (TMS) was purchased from Cambridge Isotope (Andover, MA). The bis-Tris-d19 (98% D) was purchased from Isotec (Milwaukee, WI). The compound library was previously complied as described elsewhere [19] .

7.3.2 Apparatus. All NMR data was collected on a Bruker 500 MHz Avance spectrometer (Billercia, MA) equipped with a triple resonance, Z-axis gradient cryoprobe and using a Bruker BACS-120 sample changer and IconNMR software for automated data collection. The screening data for this study was compiled over a 5 year time span in which two different 1D ^{1}H solvent suppression pulse sequences were used for the measurement of ligand 1D ^{1}H NMR line broadening. High-throughput NMR screening spectra for the HSA, BSA, *S. aureus* primase CTD, PrgI, PA1324, and SAV1430 were collected at 298 K using 64 transients with a sweep-width of 6009 Hz with 8 K data points and a 2.0 s relaxation delay using the using a presaturation solvent suppression pulse sequence (chapter 3 & 5).[4, 20-22] High-throughput NMR screening spectra for DGKA, STM1790, UFC1, YjbR, YkfF, YkvR and YtfP, the 5 amylases and *S. aureus*

nuclease proteins were collected at 298 K using 64 transients with a sweep-width of 6009 Hz with 8 K data points and a 1.0 sec relaxation delay using the excitation sculpting[23] method for solvent suppression of the residual H_2O resonance signal (chapter 4).

7.3.3 Sample preparation. All NMR ligand affinity assays were completed by screening each protein individually with a library of biologically active compounds. The compound library is composed of 113 mixtures with 3-4 ligands per mixture and is described in detail elsewhere.[19] The screens of HSA, BSA, *S. aureus* primase CTD, PrgI, PA1324, and SAV1430 were prepared as previously described.[4, 20-22] *S. aureus* nuclease, DGKA, STM1790, UFC1, YjbR, YkfF, YkvR, YtfP, and the 5 amylases were screened at 5 μM protein concentration and 100 μM ligand concentration in a screening buffer of 2% DMSO-d6, 20 mM Bis-Tris pH 7.0 (uncorrected), 11.1 μM TMSP in "!00%"D_2O.

7.3.4 Binding assay. Ligand binding was identified by a decrease in free ligand signal upon the addition of protein. The methods for data processing and identifying binding ligands have been previously discussed in detail in the previous chapters 2, 3, and 4 and references.[4, 21, 24] Briefly, data was Fourier transformed, auto-phase and baseline corrected. Each 1D ^1H NMR spectrum were compared to the corresponding free ligand mixture reference spectrum and visually analyzed to identify binding ligands. A binding event was identified by the decrease in ligand intensity of the nuclease-mixture relative to the free ligand mixture.

7.3.5 Ligand binding profiles. A ligand binding profile score was measured for each protein comparison using equation 7.1. Overlapping binding ligands (S) for every protein were identified in a pairwise manner for a total of 171 comparisons. The probability of finding overlapping ligands between two proteins was calculated using eq

7.5 Each pairwise E-value was calculated using a library size of 437 compounds (with p = 1/437 = 0.00229).

7.3.6 Functional similarity measurement. The Uniprot accession number was obtained for each protein in the study (http://www.uniprot.org/). The list of Uniprot accession numbers was uploaded to the semantic similarity tool FunSimMat.[25] All reported functional similarities are expressed as the *funsim* score measured as described.[25] Briefly, the funsim score is measure of relative functional similarity between GO terms at the *biological process* and *molecular function* levels of the gene ontology. It ranges from 0 for no functional similarity to 1 for maximal functional similarity

$$funSim = \frac{1}{2}\left[\left(\frac{BPscore}{\max\,(BPscore)}\right)^2 + \left(\frac{MFscore}{\max\,(MFscore)}\right)^2\right]$$

Where, max(*BPscore*) and max(*MFscore*) denote the maximal similarity scores for biological process and molecular function, respectively. The max(*BPscore*) and max(*MFscore*) scores for the *funSim* score is computed using *simRel*. *simRel* is a combination of Resnik's and Lin's measure of semantic similarity[25-27]

$$simRel = \max\,(c1, c2)\left[\frac{2(logp(c)}{logp(c1) + logp(c2)} * (1 - p(c))\right]$$

Where, c1 and c2 are the semantic similarity terms of a protein, maxS(c1,c2) is the set of common terms, $p(c)$ is the relative frequency of occurrence of a term.

7.4 RESULTS AND DISCUSSION

7.4.1 Establishing a set of functionally diverse proteins. Chapter 1 discussed the difficulties with non-uniform methods for functional annotation. The Gene Ontology Annotation project[28, 29] is becoming the standard representation for functional annotating of individual proteins. The success of the GO method lies in the hierarchical approach to protein annotation. Each protein or gene product is annotated with three levels of functional similarity, *biological process*, *molecular function* and *cellular component*. This approach annotates a specific GO number for each level of function which allows for development of computational functional similarity scoring methods. A number of methods have been developed to measure functional similarity with the majority of the methods based on semantic similarity of GO terms.[25, 26, 30-32] In this study the functional similarity score from FunSimMat[25, 32] was used to measure functional similarly between 19 proteins with a range of functional similarity (appendix 7A). FunSimMat is a composite average method for semantic similarity. The composite methods are generally more biologically accurate[33]

For the 19 proteins screened in the NMR ligand affinity assay, 13 proteins have a previously annotated function based on GO terms and 6 proteins have an unknown function. As a positive control, two sets of functionally related proteins (2 serum albumins and 5 amylases) were evaluated. A functional similarity score between each pair of proteins was measured by the semantic similarity tool of GO annotations FunSimMat (table 7.1).[25] The FunSimMat similarity for HSA and BSA was 0.98 and an average FunSimMat similarity score of 0.69 ± 0.01 was calculated for the amylases. The remaining 12 proteins exhibited no functional relationship to any other protein in the

screening set, yielding an average FunSimMat similarity score of 0.1 ± 0.1. A weak functional similarity was observed between the two albumins and the human protein ubiquitin-fold modifier-conjugating enzyme 1 (UFC1, Uniprot: Q9Y3C8). However, this similarity is limited to one overlapping and generic "protein binding" GO number (GO:0005515).

7.4.2 High-throughput ligand screening of a set of functionally diverse proteins. To experimentally support the ligand binding profile hypothesis, 19 proteins were screened by NMR using a chemical library of biologically active compounds.[19] Binding events were identified as previously described by measuring a decrease in ligand ^{1}H NMR peak intensities in the presence of a protein (figure 7.4).[4, 21] As an example, figure 7.4 shows the relative responses in binding for HSA and BSA to the non-steroidal anti-inflammatory drug naproxen. Naproxen was identified from a screen of the entire ligand library as a binder for both proteins. The relative change in linewidth for naproxen binding HSA was comparable to naproxen binding BSA. The ligand binding profile method only uses the identification of binding ligands (hit vs. no hit) to compare functional similarities. The binary mode of measuring ligand binding similarities makes the ligand binding profile a high-throughput method for functional annotation.

Table 7.1 A diverse set of proteins have been screened by 1D ^{1}H NMR line broadening experiments (see methods 7.3.3). The set of 19 proteins is comprised of 2 sets of positive controls (set1=albumins, set2=amylases). Functional similarity between each protein was measured by the semantic similarity tool FunSimMat.[25] The 6 unannotated proteins in the data were removed from the table for clarity; there was no measured functional similarity due to the lack of Gene Ontology[12] annotations for the proteins. The nuclease protein was also removed for clarity because there was no functional similarity to any protein in the dataset.

	HSA	**BSA**	**Primase**	**PrgI**	**Aor-A**	**Hvu-B**	**Bam-A**	**Bli-A**	**Iba-B**	**DGKA**	**UFC1**
HSA		0.98	0.07	0.28	0.05	0.02	0.04	0.04	0.03	0.23	0.49
BSA			0.07	0.28	0.05	0.03	0.04	0.04	0.04	0.25	0.49
Primase				-	0.2	0.15	0.19	0.19	0.17	0.24	0.15
PrgI					0	0	0	0	0	0	0
Aor-A						0.64	0.68	0.68	0.67	-	-
Hvu-B							0.63	0.63	0.71	-	-
Bam-A								0.68	0.63	-	-
Bli-A									0.63	-	-
Iba-B										0.07	0.22
STM1790										-	-
DGKA											0.03
YjbR											-
UFC1											

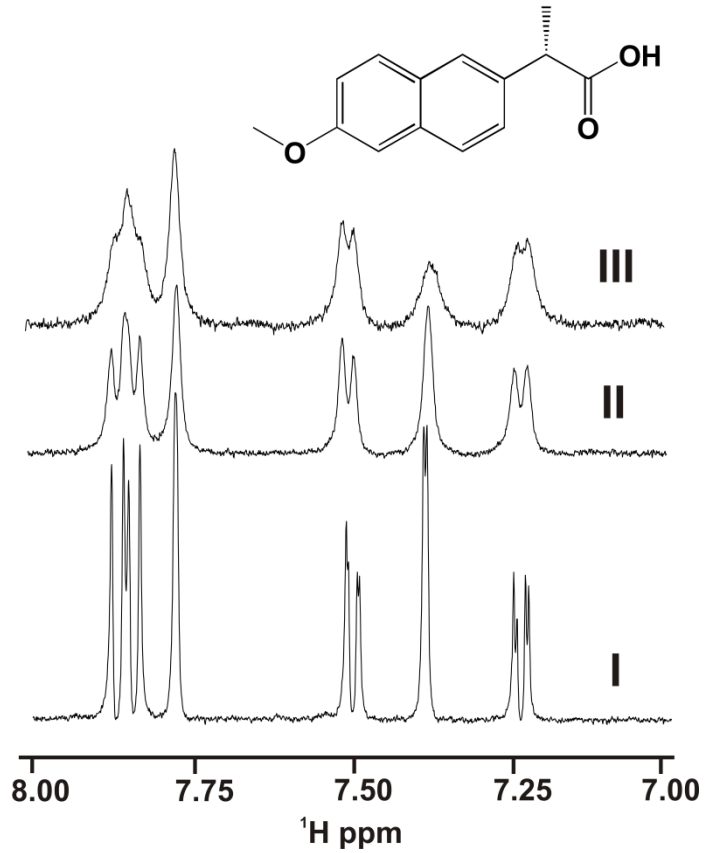


Figure 7.4. Proteins with similar function bind similar ligands. Ligand binding is identified by a decrease in ligand peak intensity upon addition of a target protein. The 1D ^{1}H NMR spectrum of the non-steroidal anti-inflammatory drug naproxen (I) is shown to broaden in the presence of *H. sapiens* serum albumin (HSA) (II) and *B. taurus* serum albumin (BSA) (III) indicating a positive binding event. The NMR line broadening experiments used 100 μM ligand and 5 μM protein as described in the methods section.

7.4.3 Ligand binding profiles for a set of functionally diverse proteins. An all-vs-all pairwise comparison of the 19 proteins gave a total of 171 ligand binding profile comparisons with only 11 comparisons giving a significant similarity score ($P < 0.0001$). The comparisons with the highest similarity scores corresponded to the set of albumins (E-value $1x10^{-58}$) and the set of amylases (average E-value ~$1x10^{-11}$). Table 7.2 lists all protein pairs with a significant ligand binding similarity score along with the corresponding FunSimMat functional similarity score. The complete list of ligand binding similarity scores (appendix 7A) shows an abrupt decrease in significance for the remaining proteins. This correlates with the remaining proteins having no functional similarity to one another.

As shown in table 7.2, human serum albumin (HSA) and bovine serum albumin (BSA) had a large number of binding ligands (178 and 171, respectively) compared to the overall size of the library. The relative hit rate for these two proteins was 40.7% and 39.1%, respectively. With a large hit rate, false similarities may arise if a second protein serendipitously bound to a small subset of compounds that were shown to bind HSA or BSA. However, the ligand binding similarity score (eq 7.2) effectively eliminates this concern by scaling the score based on both the total number of compounds found to bind each protein and by the number of overlapping binding ligands. As an example, the *S. typhimurium* type III secretion system protein PrgI bound to a total of five compounds, where each compound was also shown to bind HSA and BSA. The corresponding E-values for the ligand binding profile comparisons between PrgI and HSA (6.9×10^{-2}) and BSA (6.4×10^{-2}) were not significant.

Table 7.2 The number of hits per protein (m and n), overlapping ligands (S), E-values and functional similarity scores (FunSim) are reported for the significant ligand binding profiles at 99.99% confidence interval from a comparison of 19 proteins. The set of serum albumins from *H. sapiens* (HSA) and *B. taurus* (BSA) and amylases (Aor, Bam, Bli, Hvu, and Iba) gave significant similarity. The set of amylases was composed of 3 α-amylases from *A. oryzae* (Aor), *B. amyloliquefaciens* (Bam), *and B. licheniformis* (Bli) and 2 β-amylases *H. vulgare* (Hvu) and *I. batatas* (Iba). A complete list of binding profiles is reported in the appendix 7A.

Comparison	m/n	S	E-value	Funsim Score
HSA-BSA	178/171	162	$2.16x10^{-58}$	0.98
Bam-Aor	35/36	22	$6.38x10^{-19}$	0.68
Bam -Hvu	35/29	14	$1.17x10^{-10}$	0.63
Bli- Aor	28/36	18	$1.19x10^{-15}$	0.68
Bli - Bam	28/35	16	$1.42x10^{-14}$	0.68
Bli - Hvu	28/29	9	$3.86x10^{-06}$	0.63
Hvu - Aor	29/36	13	$2.98x10^{-08}$	0.64
Iba- Aor	29/36	12	$2.98x10^{-08}$	0.67
Iba - Bam	29/35	15	$7.56x10^{-12}$	0.63
Iba - Bli	29/28	11	$2.43x10^{-08}$	0.63
Iba - Hvu	29/29	12	$2.45x10^{-09}$	0.71

There was an observed similarity in ligand binding profiles between *S. aureus* nuclease and the α-amylases from *A. oryzae* and *B. amyloliquefaciens*. However, the similarity in the ligand binding profiles was limited to the nucleosides in the library. Additionally, the remaining 3 amylases did not bind these ligands or exhibit a significant ligand binding similarity to nuclease. The observed ligand binding similarity between the nuclease and two of the α-amylases is potentially due to trace amounts of a nuclease that may be present in the *A. oryzae* and *B. amyloliquefaciens* α-amylases samples. This is a likely occurrence since the samples were purchased as crude mixtures, where size-exclusion chromatography only yielded a modest improvement in purity.

The ligand binding profiles for all 19 proteins is represented as a heat map in figure 7.5. Each binding ligand was colored red while each non-binding ligand was colored white. The heat map shows the overall clustering patterns for each binding profile. The heat map correlates well with table 7.2 showing that functionally similar proteins bind a consensus set of ligands from a standardized library of compounds. Ligands that are not included in the consensus set could either be due to non-specific binding, differences between sample preparation, or potentially unique and specific binders.

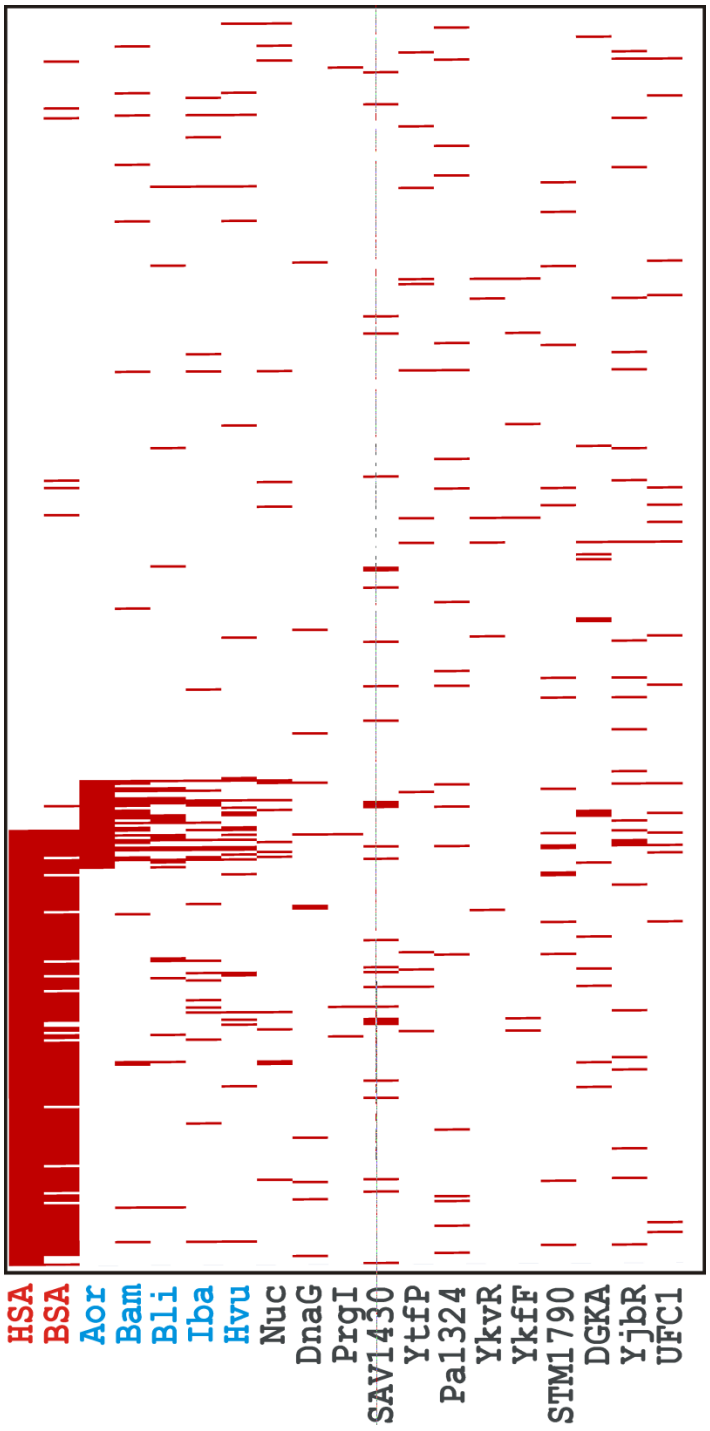

HSA
BSA
Aor
Bam
Bli
Iba
Hvu
Nuc
DnaG
PrgI
SAV1430
YtfP
Pa1324
YkvR
YkfF
STM1790
DGKA
YjbR
UFC1

Figure 7.5 Heat map summarizing the NMR ligand affinity screens. For 19 proteins: *H. sapiens* serum albumin (HSA), *B. taurus* serum albumin (BSA), *A. oryzae* α-amylase (Aor), *B. amyloliquefaciens* α-amylase (Bam), *B. licheniformis amyloliquefaciens* α-amylase (Bli), *I. batatas* β-amylase (Iba), *H. vulgare* β-amylase (Hvu), *S. aureus* nuclease, *S. aureus* primase C-terminal domain, *S. typhimurium* type III secretion system protein (PrgI), *S. aureus* unannotated protein SAV1430, *E. coli* unannotated protein YtfP, *P. aeruginosa* unannotated protein PA1324, *B. subtilis* unannotated protein YkvR, *E. coli* unannotated protein YkfF, *S. typhimurium* unannotated protein STM1790, *H. sapiens* diacylglycerol kinase alpha (DGKA), *E. coli* unannotated protein YjbR, *H. sapiens* ubiquitin-fold modifier-conjugating enzyme 1 (UFC1), where the albumins are colored red, the amylases cyan and the remainder of the proteins grey. A binding ligand is indicated by a red line. The 437 ligands were sorted to maximize the clustering of binding ligands for the albumins and amylases.

7.4.4 Future developments to the ligand binding profile method. Ligand binding profiles are independent of sequence and structural information and thus provide an experimental-based approach to predict protein function in a relatively robust and high-throughput fashion. The results reported herein demonstrate a clear correlation between ligand binding similarity scores and FunSimMat functional similarity scores. Specifically, only the set of albumins and amylases gave significant ligand binding similarity scores. Unfortunately, the ligand binding profile method was unable to differentiate between the α and β amylase families. A further refinement of the functional annotation would require a second screening step using a focused library to differentiate these functional classes. In the case of the amylases, this would involve screening the proteins with a carbohydrate library, where a subset of the compounds would selectively bind to the α- or β-amylase proteins.

The success of the ligand binding profile approach to annotate a protein depends on a functionally diverse and modestly sized chemical library that differentiates between various functional classes. Importantly, the methodology used to identify binding ligands must efficiently eliminate non-specific or irrelevant interactions. This is not the case with traditional high-throughput screening (HTS) methods that encounter significant false-positive and false-negative rates. Applying the ligand binding profile technique to HTS data sets from the High Throughput Screening Laboratory at the University of Kansas were unsuccessful. Alternatively, NMR ligand-affinity screens provide a direct observation of a specific interaction between the ligand and protein. As demonstrated, the preponderance of binding ligands identified from the 19 NMR ligand affinity screens was

uniquely associated with each functional class and were shown to correlate with the protein's GO terms.

7.5 REFERENCES

1. Janitz, M., Assigning functions to genes-the main challenge of the post-genomics era. *Rev. Physiol., Biochem. Pharmacol.* **2007,** 159, 115-129.
2. Rentzsch, R.; Orengo, C. A., Protein function prediction--the power of multiplicity. *Trends Biotechnol* **2009,** 27, (4), 210-9.
3. Rost, B.; Liu, J.; Nair, R.; Wrzeszczynski, K. O.; Ofran, Y., Automatic prediction of protein function. *Cell. Mol. Life Sci.* **2003,** 60, (12), 2637-2650.
4. Mercier, K. A.; Baran, M.; Ramanathan, V.; Revesz, P.; Xiao, R.; Montelione, G. T.; Powers, R., FAST-NMR: functional annotation screening technology using NMR spectroscopy. *J Am Chem Soc* **2006,** 128, (47), 15292-9.
5. Najmanovich, R. J.; Allali-Hassani, A.; Morris, R. J.; Dombrovsky, L.; Pan, P. W.; Vedadi, M.; Plotnikov, A. N.; Edwards, A.; Arrowsmith, C.; Thornton, J. M., Analysis of binding site similarity, small-molecule similarity and experimental binding profiles in the human cytosolic sulfotransferase family. *Bioinformatics* **2007,** 23, (2), e104-9.
6. Kinnings, S. L.; Jackson, R. M., Binding site similarity analysis for the functional classification of the protein kinase family. *J Chem Inf Model* **2009,** 49, (2), 318-29.

7. Mercier, K. A.; Baran, M.; Ramanathan, V.; Revesz, P.; Xiao, R.; Montelione, G. T.; Powers, R., FAST-NMR: Functional Annotation Screening Technology Using NMR Spectroscopy. *J Am Chem Soc* **2006,** 128, (47), 15292-15299.

8. Dessailly, B. H.; Lensink, M. F.; Orengo, C. A.; Wodak, S. J., LigASite-a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res.* **2008,** 36, (Database Iss), D667-D673.

9. Gerlt, J. A.; Babbitt, P. C., Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis. *Curr Opin Chem Biol* **1998,** 2, (5), 607-12.

10. Brown, S. D.; Gerlt, J. A.; Seffernick, J. L.; Babbitt, P. C., A gold standard set of mechanistically diverse enzyme superfamilies. *Genome Biol* **2006,** 7, (1), R8.

11. Goonesekere, N. C.; Shipely, K.; O'Connor, K., The challenge of annotating protein sequences: The tale of eight domains of unknown function in Pfam. *Comput Biol Chem* **2010,** 34, (3), 210-214.

12. Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **2000,** 25, (1), 25-9.

13. Wu, X.; Zhu, L.; Guo, J.; Zhang, D. Y.; Lin, K., Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res* **2006,** 34, (7), 2137-50.

14. Lee, P. H.; Lee, D., Modularized learning of genetic interaction networks from biological annotations and mRNA expression data. *Bioinformatics* **2005,** 21, (11), 2739-47.

15. Lei, Z.; Dai, Y., Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinformatics* **2006,** 7, 491.

16. Guo, X.; Liu, R.; Shriver, C. D.; Hu, H.; Liebman, M. N., Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics* **2006,** 22, (8), 967-73.

17. Tuikkala, J.; Elo, L.; Nevalainen, O. S.; Aittokallio, T., Improving missing value estimation in microarray data with gene ontology. *Bioinformatics* **2006,** 22, (5), 566-72.

18. Karlin, S.; Altschul, S. F., Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A* **1990,** 87, (6), 2264-8.

19. Mercier, K. A.; Germer, K.; Powers, R., Design and Characterization of a Functional Library for NMR Screening against Novel Protein Targets. *Combinatorial Chemistry and High Throughput Screening* **2006,** 9, (7), 515-534.

20. Mercier, K. A.; Cort, J. R.; Kennedy, M. A.; Lockert, E. E.; Ni, S.; Shortridge, M. D.; Powers, R., Structure and function of Pseudomonas aeruginosa protein PA1324 (21-170). *Protein Sci* **2009,** 18, (3), 606-18.

21. 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* **2009,** 12, (3), 285-95.

22. 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* **2009,** 4, (10), e7442.

23. Hwang, T.-L.; Shaka, A., Water Suppression That Works. Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients. **1995,** 112, (2), 275-279.

24. 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* **2008,** 10, (6), 948-58.

25. Schlicker, A.; Albrecht, M., FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Res* **2008,** 36, (Database issue), D434-9.

26. Schlicker, A.; Domingues, F. S.; Rahnenfuhrer, J.; Lengauer, T., A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* **2006,** 7, 302.

27. Resnik, P., Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application *J. Artif. Intell. Res.* **1999,** 11, 95-130.

28. Blake, J. A.; Harris, M. A., The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis. *Curr Protoc Bioinformatics* **2008,** Chapter 7, Unit 7 2.

29. Camon, E.; Barrell, D.; Brooksbank, C.; Magrane, M.; Apweiler, R., The Gene Ontology Annotation (GOA) Project--Application of GO in SWISS-PROT, TrEMBL and InterPro. *Comp Funct Genomics* **2003,** 4, (1), 71-4.

30. Lord, P. W.; Stevens, R. D.; Brass, A.; Goble, C. A., Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **2003,** 19, (10), 1275-83.

31. Francisco, M. C.; M; rio, J. S.; Pedro, M. C., Measuring semantic similarity between Gene Ontology terms. *Data Knowl. Eng.* **2007,** 61, (1), 137-152.

32. Schlicker, A.; Albrecht, M., FunSimMat update: new features for exploring functional similarity. *Nucleic Acids Res* **2010,** 38, (Database issue), D244-8.

33. Pesquita, C.; Faria, D.; Bastos, H.; Ferreira, A. E.; Falcao, A. O.; Couto, F. M., Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* **2008,** 9 Suppl 5, S4.

Appendix 7A Complete list of ligand binding profile scores, - marks indicate no overlapping binding ligands

	HSA	BSA	Primase_SA	PrgI	SAV1430	YtfP	PA1324	YkvR	YkfF	Nuclease	Aor	Hvu	Bam	Bli	Iba	STM1790	DGKA	YjbR	UFC1
HSA	2.32E-68	2.16E-58	5.16E-02	6.48E-02	2.38E-02	1.29E+00	1.19E+00	-	-	2.18E-02	6.02E-02	2.43E-03	4.03E-01	2.73E-02	1.88E-02	4.89E-02	4.23E+01	2.22E-02	1.58E-01
BSA	2.16E-58	1.32E-68	1.35E-02	5.98E-02	3.25E-02	7.08E+00	3.17E-03	-	-	2.82E-04	5.89E-04	6.43E-01	7.33E-02	3.48E-02	3.59E-02	2.82E-04	3.22E+01	3.05E-02	2.22E-02
Primase_SA	5.16E-02	1.35E-02	2.09E-18	8.18E-03	-	-	-	-	-	1.26E-01	1.25E-01	3.30E-02	1.25E-01	4.11E-02	-	6.92E-02	-	-	1.08E-02
PrgI	6.48E-02	5.98E-02	8.18E-03	9.58E-10	1.48E-01	-	-	-	-	-	1.42E-01	1.48E-01	1.44E-01	1.48E-01	1.48E-01	1.30E-01	-	-	1.88E-02
SAV1430	2.38E-02	3.25E-02	-	1.48E-01	3.27E-34	1.03E-02	8.46E-02	-	-	6.86E-02	3.77E-02	6.19E-02	8.30E-02	1.02E-01	1.81E-02	6.86E-02	-	2.80E-03	1.03E-01
YtfP	1.29E+00	7.08E+00	-	-	1.03E-02	1.75E-19	-	1.32E-02	-	-	5.39E-03	-	-	-	-	-	1.31E-01	2.32E-02	-
PA1324	5.43E-01	3.17E-03	-	-	8.46E-02	-	2.96E-26	-	-	1.17E-01	4.21E-02	9.97E-02	5.08E-02	1.02E-01	9.97E-02	3.27E-02	-	7.67E-02	5.47E-03
YkvR	5.10E+00	-	-	-	-	5.59E-04	-	3.91E-11	-	-	-	-	-	-	-	-	-	1.46E-01	1.43E-01
YkfF	1.47E-01	4.27E+00	-	-	5.35E-02	5.59E-04	-	3.12E-03	3.91E-11	-	-	5.08E-02	-	-	-	-	-	-	-
Nuclease	2.18E-02	2.82E-04	1.26E-01	-	6.86E-02	-	1.17E-01	-	-	1.93E-24	8.32E-05	1.56E-02	2.84E-08	6.46E-02	2.25E-03	4.96E-02	7.66E-02	1.97E-04	1.13E-01
Aor	6.02E-02	5.89E-04	1.25E-01	1.42E-01	3.77E-02	5.39E-03	4.21E-02	-	-	8.32E-05	2.83E-38	2.98E-08	6.38E-19	1.19E-15	2.98E-08	2.99E-02	2.12E-02	7.34E-03	7.09E-03
Hvu	2.43E-03	6.43E-01	3.30E-02	1.48E-01	6.19E-02	-	9.97E-02	-	-	1.56E-02	2.36E-09	6.69E-32	1.17E-10	3.86E-06	2.45E-09	1.56E-02	1.24E-01	8.69E-03	2.85E-03
Bam	4.03E-01	7.33E-02	1.25E-01	1.44E-01	8.30E-02	-	5.08E-02	-	-	2.84E-08	3.18E-20	1.66E-09	1.25E-37	1.42E-14	7.56E-12	2.76E-02	1.95E-02	7.63E-02	9.51E-02
Bli	2.73E-02	3.48E-02	4.11E-02	1.48E-01	1.02E-01	-	1.02E-01	-	-	6.46E-02	1.19E-15	3.86E-06	2.62E-13	9.48E-33	2.43E-08	1.40E-02	5.19E-02	1.13E+00	7.08E-02
Iba	1.88E-02	3.59E-02	-	1.48E-01	1.81E-02	-	9.97E-02	-	-	2.25E-03	2.98E-08	2.45E-09	7.56E-12	2.43E-08	1.73E-33	6.86E-02	-	1.40E-02	1.09E-01
STM1790	4.89E-02	2.82E-04	6.92E-02	1.30E-01	6.86E-02	-	3.27E-02	-	-	4.96E-02	2.99E-02	1.56E-02	2.76E-02	1.40E-02	6.86E-02	1.93E-24	-	1.40E-02	3.34E-04
DGKA	4.23E+01	3.22E+01	-	-	-	1.31E-01	-	-	-	7.66E-02	2.12E-02	1.24E-01	1.95E-02	5.19E-02	-	-	1.55E-22	6.50E-04	1.03E-01
YjbR	2.22E-02	3.05E-02	-	-	2.80E-03	2.32E-02	7.67E-02	-	-	1.97E-04	7.34E-03	8.69E-03	7.63E-02	1.71E-02	1.40E-02	1.40E-02	6.50E-04	3.79E-31	1.66E-02
UFC1	1.58E-01	2.22E-02	1.08E-02	1.88E-02	1.03E-01	-	5.47E-03	-	-	1.13E-01	7.09E-03	2.85E-03	9.51E-02	7.08E-02	1.09E-01	3.34E-04	1.03E-01	1.66E-02	2.33E-25

CHAPTER 8:

CONCLUSIONS

8.1 SUMMARY OF WORK

Protein science has always had a long history intertwined with the advancements in chemistry, biology and physics. Today, with nearly 1350 complete genome sequences available, our understanding of biology at the molecular level has never been more complete. While our understanding of biology continues to grow exponentially, we are still at the beginning of having a truly systematic understanding of Mother Nature's most fundamental secrets. This is most evident by the large functionally unannotated segments of each organism's genome.

The genes (and proteins they encode) found in these functionally unannotated regions are considered "hypothetical proteins". Current estimates suggest between 12%-50% of the known gene sequences belong to unannotated proteins.[1-3] This is true even for the most highly studied model organisms *Escherichia coli*. An estimated 50% of the genes found in the *E. coli* genome have no experimental annotation.[4, 5] Considering the large degree of biodiversity, it was initially suggested that hypothetical proteins were adaptations to specific environmental niches and therefore species specific.[6, 7] However, many hypothetical proteins are not species-specific and homologous sequences are found in a range of phylogenetic distributions. These evolutionary "conserved hypothetical proteins" significantly limits our understanding of biology.[8]

From a pragmatic viewpoint, identifying the functions of these proteins could lead to new therapeutics; making functional annotation of paramount importance. Considering the large number of unannotated proteins (~1.5 million), the most popular tools for functional annotation rely on homology transfer of sequences and structures to

automatically predict protein function. However, sequence and structure homology does not always imply functional conservation and these automatic methods often lead to spurious annotations. Estimates in the error rates suggested nearly 30% of all automatic functional annotations of enzymes are incorrect.[9] Differences in protein active site structures leading to different ligand specificities and enzyme efficiencies are suspected to be a major source of errors in automatic functional annotations.[9-11]

The large error rate of automatic functional annotation methods strongly supports the need for developing new methods that are independent of homology transfer. In this dissertation I thoroughly tested the hypothesis of using ligand-defined active sites for functional annotation. In chapter 2, I discussed the theory and experimental validation of a method to measure single point binding dissociation constants (K_D) from 1D ^{1}H NMR. The primary goal of the project was to develop a method that would be robust for a broad functional library of compounds to a variety of biological target molecules. The method was intended as a qualitative screening tool to provide accurate ranking of target molecules for both drug discovery and functional annotation using the Functional Annotation Screening Technology by NMR (FAST-NMR) method.

In chapter 3 I used this single point K_D method in concert with the FAST-NMR method to select the best binding ligand to the type three-secretion system protein PrgI (didecyldimethylammonium bromide, DDAB). Didecyldimethylammonium bromide was identified from a compound library using 1D ^{1}H NMR screening techniques and used to identify the active site of PrgI. Finding the active site of PrgI facilitated the identification of a functional similarity between PrgI and Bcl-xL using the Comparison of Protein Active Site Similarities (CPASS) database.[12]

The results from the FAST-NMR screen of *S. aureus* nuclease in chapter 4 confirmed the use of NMR screening to identify a protein active site and the use of active site similarities to identify protein functional similarities. Additionally, the successful identification of a ligand bound *S. aureus* nuclease structure having the best active site similarity validated CPASS and using active site similarity as a functional annotation tool. Finally, the optimization of the initial version of the NMR screening techniques utilized by FAST-NMR significantly improved the efficiency of the high-throughput screen.

The rapid rise in community acquired antibiotic resistance, particularly to *S. aureus,* requires the rapid identification of new antibiotic targets and potential drugs.[13] The interaction between bacterial primase C-terminal domain and replicative helicase N-terminal domain is an attractive antibiotic target because it is functionally conserved in bacteria, essential for DNA replication and distinctly different from eukaryotes.[14, 15] Additionally, the high degree of sequence variability and differences in structure suggest a possible means to tailor antibiotic development to a specific organism. In chapter 5, I reported the NMR solution structure of *S. aureus* primase CTD. I use the structure to show a strong phylum dependency for primase CTD structure similarity and reported a potential drug lead for further antibiotic development.

In chapter 6, I expanded upon the work of phylum dependent structure similarity by thoroughly analyzing functionally conserved, orthologous structures.. I quantify a maximum structure/sequence similarity between the two bacterial phyla, *Proteobacteria* and *Firmicutes*, and discussed the viability of phylogeny as a suitable constraint for

selecting a homology model. This was supported by showing protein folds are not uniformly sensitive changes in sequence.

The problems with automatic functional annotation were thoroughly discussed in this dissertation. The development of the FAST-NMR method is a significant advancement towards high-throughput functional annotation but is limited by the availability of a high-resolution protein structure. In chapter 7, I discussed the development the ligand binding profile (LBP) method for functional annotation. A ligand binding profile is defined as a set of ligands that bind a protein from a high-throughput ligand affinity screen. The hypothesis was proteins with similar function will bind a similar set of compounds from the same high-throughput screening library. I tested the method on a set of 19 proteins with a range of functions and reported only proteins with high degree of functional similarity gave significant LBP scores. The ligand binding profile method is independent of sequence, structure or evolutionary information and therefore not limited by the issues of automatic functional annotation discussed in this dissertation.

As a final thought, the ligand binding profile is not dependent on screening method or chemical library (provided binding profiles are generated from the same chemical library). This opens the door for virtual screening methods to identify binding ligands and compare ligand binding profiles. Virtual screens significantly reduce the time scale of ligand screening relative to experimental based approaches. The continual advancements in virtual screening coupled with the ligand binding profile will help make high-throughput functional annotation a reality.

8.2 REFERENCES

1. Muller, J.; Szklarczyk, D.; Julien, P.; Letunic, I.; Roth, A.; Kuhn, M.; Powell, S.; von Mering, C.; Doerks, T.; Jensen, L. J.; Bork, P., eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res* **2010,** 38, (Database issue), D190-5.
2. Jensen, L. J.; Julien, P.; Kuhn, M.; von Mering, C.; Muller, J.; Doerks, T.; Bork, P., eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* **2008,** 36, (Database issue), D250-4.
3. Sivashankari, S.; Shanmughavel, P., Functional annotation of hypothetical proteins - A review. *Bioinformation* **2006,** 1, (8), 335-8.
4. Galperin, M. Y.; Koonin, E. V., From complete genome sequence to 'complete' understanding? *Trends Biotechnol* **2010,** 28, (8), 398-406.
5. Kolker, E.; Makarova, K. S.; Shabalina, S.; Picone, A. F.; Purvine, S.; Holzman, T.; Cherny, T.; Armbruster, D.; Munson, R. S., Jr.; Kolesov, G.; Frishman, D.; Galperin, M. Y., Identification and functional analysis of 'hypothetical' genes expressed in Haemophilus influenzae. *Nucleic Acids Res* **2004,** 32, (8), 2353-61.
6. Siew, N.; Fischer, D., Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins* **2003,** 53, (2), 241-51.
7. Daubin, V.; Ochman, H., Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. *Genome Res* **2004,** 14, (6), 1036-42.
8. Galperin, M. Y., Conserved 'hypothetical' proteins: new hints and new puzzles. *Comp Funct Genomics* **2001,** 2, (1), 14-8.

9. Devos, D.; Valencia, A., Intrinsic errors in genome annotation. *Trends Genet* **2001,** 17, (8), 429-31.

10. Tian, W.; Skolnick, J., How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* **2003,** 333, (4), 863-82.

11. Rost, B., Enzyme function less conserved than anticipated. *J Mol Biol* **2002,** 318, (2), 595-608.

12. Powers, R.; Copeland, J. C.; Germer, K.; Mercier, K. A.; Ramanathan, V.; Revesz, P., Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **2006,** 65, (1), 124-35.

13. Larson, E., Community factors in the development of antibiotic resistance. *Annu Rev Public Health* **2007,** 28, 435-47.

14. Frick, D. N.; Richardson, C. C., DNA primases. *Annu Rev Biochem* **2001,** 70, 39-80.

15. Koepsell, S. A.; Larson, M. A.; Griep, M. A.; Hinrichs, S. H., Staphylococcus aureus helicase but not Escherichia coli helicase stimulates S. aureus primase activity and maintains initiation specificity. *J Bacteriol* **2006,** 188, (13), 4673-80.