Fall 11-22-2017

Bioinformatic and Biophysical Analyses of Proteins

Jonathan Catazaro
University of Nebraska-Lincoln, jcatazaro@huskers.unl.edu

BIOINFORMATIC AND BIOPHYSICAL ANALYSES OF PROTEINS

by

Jonathan Catazaro

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

November, 2017

BIOINFORMATIC AND BIOPHYSICAL ANALYSES OF PROTEINS

Jonathan Catazaro, Ph.D.

University of Nebraska, 2017

Advisor: Robert Powers

The prevailing dogma in structural genomics is the existence of a strong correlation between protein sequence, structure, and biological function. Proteins with high sequence similarity typically have a similar, if not the same, structure and function. In many cases this logic can fail due to distantly related proteins having very low sequence similarity, a lack of a representative structure, structural novelty, or the absence of a characterized function. Further, the paradigm fails to account for dynamics, which have a significant effect on structural stability and enzymatic efficacy.

Nuclear magnetic resonance (NMR) spectroscopy is uniquely capable of solving the structure, assisting with annotation, and deriving the dynamics of previously unstudied proteins. Historically, NMR has been used to calculate structures and dynamics of small or disordered proteins, which could then be used with computational methods to predict function. Predicted annotations are then confirmed by further experimentation such as ligand screens or titrations. The combination of NMR and bioinformatics, therefore, works synergistically to yield significant results, which has the ability to characterize highly complex proteins and fill gaps in the sequence to structure to function paradigm.

This dissertation begins with work accomplished using the Comparison of Active Site Structures (CPASS) software to show the functional evolution of a class of cofactor dependent enzymes and also expands on the utility of CPASS with the implementation of

a functional clustering of its database. Described next is an emphasis on protein and peptide structure and the relationship between the experimentally derived ensembles and biological function and dynamics. Recent improvements in the calculation of protein fast-timescale dynamics are then introduced before a final concluding chapter.

Acknowledgements

Graduate school has been, without a doubt, one of the most trying times in my life. I have never been one to complain about adversity or hardships because I know I have amazing people in my life to help me when I need it most. As I write this, I find it immensely difficult to truly express how thankful I am to have the support, encouragement, and love of those closest to me. To my family, you have not only enabled me to pursue my passion but to succeed in doing so. Mom and dad, your unwavering support and cultivation of my curiosity has made me the man I am today. Ricky, you inspire me to go out on adventures instead of spending my weekends stuck in front of a computer. To my future wife Alissa, thank you for accompanying me on this endeavor and your unshakable commitment to our happiness.

I have been fortunate to work with many remarkable scientists in my group and in the department, many of whom I consider good friends. Our fruitful and sometimes off-topic discussions made it enjoyable to go to work everyday. To all of the Powers group members both past and present, I am proud to say we share an academic legacy together. Dr. Martha Morton, thank you for the reassurance and advice these last few months as my time at UNL nears its end. I am most especially gracious to have had the opportunity to work for Dr. Robert Powers. You have given me to opportunity to learn, make mistakes, and grow in what now seems like a very short time. Lastly, I would like to acknowledge my supervisory committee, who has guided me with their wisdom, patience and leadership during my graduate education.

Preface

The chapters included within this dissertation have been adapted from articles and communications published in peer-reviewed journals, which follow below:

Chapter 2

- J. Catazaro, A. Caprez, A. Guru, D. Swanson, and R. Powers (2014) "Functional Evolution of PLP-dependent Enzymes based on Active site Structural Similarities", *Proteins,* 82(10):2597-2608 PMC4177364 (cover).

  Reproduced with permission from *Proteins: Structure, Function, and Bioinformatics.*

Chapter 3

- J. Catazaro, A. Caprez, D. Swanson, and R. Powers (2017) "Functional Evolution of Proteins", (*in preparation*).

Chapter 4

- J. Catazaro, M. Bin Samad, N. Rodrigues de Almeida, R. Powers, and M. Conda Sheridan (2017) "Analysis of the Mechanism of Action of Helical Antimicrobial Peptides", (*in preparation*).

Chapter 5

- J. Catazaro, A. J. Lowe, R. L. Cerny, and R. Powers (2017) "The NMR Solution Structure and Function of RPA3313: A Putative Ribosomal Transport Protein from *R. palustris*", *Proteins,* 85(1):93-102 PMC5167650.

  Reproduced with permission from *Proteins: Structure, Function, and Bioinformatics*.

Chapter 6

- J. Catazaro, J. Periago, M. Shortridge, B. Worley, A. Kirchner, R. Powers and M. A. Griep (2017) "Identification of a Ligand Binding Site on the *Staphylococcus aureus* DnaG Primase C-Terminal Domain.", *Biochemistry*, 56(7):932-943.

  Reproduced with permission from *Biochemistry*.

Chapter 7

- N. M. Milkovic, J. Catazaro,, J. Lin, S. Halouska, J. L. Kizziah, S. Basiaga, R. L. Cerny, R. Powers, and M. A. Wilson (2015) "Transient Sampling of Aggregation-prone Conformations Causes Pathogenic Instability of a Parkinsonian Mutant of DJ-1 at Physiological Temperature", *Protein Science*, 24(10):1671-1685 PMC4594666.

  Reproduced with permission from *Protein Science*.

Chapter 8

- J. Catazaro, T. Andrews, N. M. Milkovic, J. Lin, A. J. Lowe, M. A. Wilson, and R. Powers (2017) ^{15}N CEST Data and Traditional Model-Free Analysis Capture Fast Internal Dynamics of DJ-1, *Analytical Biochemistry*, accepted.

Table of Contents

Chapter 1

Introduction

1.1 The Application of Nuclear Magnetic Resonance Spectroscopy and Bioinformatics in Structural Genomics

Upon completion of the human genome project and other genomics initiatives, it became possible to find relationships between organisms using their whole genome sequences.[1, 2] The relationships are leveraged to discover homologous proteins with similar functions. However, an enormous amount of genomic data now exists which contains “gaps” where genes are structurally and functionally uncharacterized. The field of structural genomics has arisen to meet this exceptional challenge. Rather than solving the three-dimensional structure of every protein encoded within a particular genome, the aim is to concentrate on representative domain targets for each globular fold.[2] The known structures can subsequently be used in computational homology modeling to then predict the remainder of the structures.

Thus far, structural genomics has greatly diversified the range of structural information available for many different protein families.[3] High-throughput methods have been used to steadily produce structural representatives of previously uncharacterized genes. Although the new structures have been deposited in the Protein Data Bank (PDB)[4], relatively few papers have been published by structural genomics centers and a distant lack of novel folds have been discovered.[3] In contrast, structural biology groups produce fewer protein structures, but the structures typically contain novel folds and are deposited with a corresponding publication. The authors go further in their analysis and also add detail pertaining to the protein's function and physiological importance. Therefore, the roles of individual structural biology groups and structural genomics centers are synergistic. Structural genomics builds a comprehensive profile of the structural landscape while structural biology characterizes protein function.

Protein structures are traditionally determined by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy.[5] Most of the structures are determined by crystallography, although not all protein targets are amenable to crystallization. Further, crystal structures are static snapshots of the protein in a non-physiological environment, which may contain structural elements that are not present in the native conformation. For example, crystallization has the tendency to produce non-biological symmetry in proteins.[6] NMR complements X-ray crystallography in that it is capable of solving structures in the solution state. The aqueous solution can be additionally adjusted to mimic the physiologic conditions for the protein. Conditions can be matched to the pH, ionic strength, and temperature of the physiological environment of the protein to be studied. Rather than producing a single conformer, NMR structures are presented as an ensemble of low-energy conformers. Therefore, the ensemble represents a set of structures that the protein may transiently sample while in solution. A significant limitation to NMR structure determination is the limitation to proteins less than 25kDa.[7] Higher molecular weight proteins are possible, but require more advanced methods and significantly more time to solve.[7] Nonetheless, it has been estimated that at least 25% of uncharacterized open reading frames (ORFs) are under the molecular weight cutoff for an

NMR solution structure.[8] This ensures that NMR will remain an essential tool in structural genomics for years to come.

The insurmountable amount of genomic data that now exists cannot possibly be completely analyzed using only experimentation. Bioinformatics, the study of complex biological data using computational methods, can be employed to quickly determine function and structure for uncharacterized genes. Many methods have been developed to predict structure and function from amino acid sequences and function from structure.[9-11] These tools have been essential to the field of structural genomics for annotating families of proteins from very little preliminary information. As with most tools and methods, there are limitations to what can be predicted and inferred. There is significant difficulty and error associated with comparative studies involving proteins at or below 30% sequence similarity.[12] For sequence comparison methods, low similarity leads too poor functional characterization and poor evolutionary relationship construction. It becomes necessary to compare proteins at the structural level using their tertiary structure or active sites to infer function and evolution.[13] However, the problem is further compounded in de novo structure prediction where low accuracy models are generated for proteins with less than 30% similarity.[14] This instance highlights the mutually beneficial relationship between bioinformatics and structural biology. A representative protein from an uncharacterized and ancient family can be structurally elucidated, and the resulting structure can be used to model the remainder of the family.

1.2 Overview of Work

This dissertation describes a body of work focused on the application of bioinformatics and structural biology towards the understanding of novel and physiologically relevant proteins. It begins with an original way to phylogenetically model distantly related proteins and describes the advantageousness of using active site structures to cluster proteins. Next are practical applications of NMR toward the structure determination of previously unsolved peptides and proteins. A deeper understanding of protein dynamics is then developed in the subsequent chapters, which includes new methods in their derivation.

Chapter 2 introduces the utility of the Comparison of Protein Active Sites Software (CPASS)[15, 16] in the discovery of evolutionary relationships between proteins.[13] Typically evolutionary relationships between proteins are inferred using sequence comparison methods. These methods, however, fail for families of distantly related proteins due to very low sequence identity. This further hinders functional annotation and characterization of the protein family. Slowly evolving structural moieties of proteins, such as an active site, are therefore valuable for annotating putative and distantly related proteins. CPASS compares the similarity and geometry of active site residues and acts as a sensitive measure of active site evolution. Similar methods compare putative cavities,[17] local structures,[18] active site shapes,[19] and have been extensively reviewed.[20] CPASS differs from these other methods because it uniquely uses ligand-defined active sites. An active site must have a known and bound ligand in its 3D structure file to be considered for CPASS analysis. Thus, this is both an advantage and disadvantage of the software. It compares real ligand binding sites, rather than putative, but is also limited by the necessity to have a bound ligand.

Pyridoxal-5'-phosphate (PLP) dependent enzymes are primordial enzymes that diversified in the last universal ancestor. Molecular evolution efforts focusing on these enzymes have been limited to small subgroups due to the low sequence identity within the family.[21-23] Using CPASS, we show that the active site structures of PLP-dependent enzymes can be used to infer evolutionary relationships based on functional similarity. The enzymes successfully clustered together based on substrate specificity, function, and three-dimensional fold. This chapter reveals the first ever successful clustering of the entire PLP-dependent enzyme family. Furthermore, it demonstrates the value of using active site structures for functional evolutionary analysis and the effectiveness of CPASS. Adam Caprez, Ashu Guru, and David Swanson assisted with the "all versus all" comparison used in this project and were instrumental in the organization of the data.

Chapter 3 further expands upon the expediency of CPASS with the addition of an improved active site comparison search. A standard CPASS search compares the query active site to all members in the database. This search, although optimized, takes several hours due to the enormity of the database (approximately 40,000 active sites), which will continue to increase as the number of deposits in the Protein Data Bank (PDB) grows. While an exhaustive search may produce the top result, the time required may be detrimental when doing multiple queries. Therefore, a search against a representative subset of the CPASS database, approximately 10% in size, would be advantageous and would only become more optimized upon further addition to the PDB. Enzyme classification (EC) numbers are generally used to group proteins and enzymes according to function. The application of these EC numbers to CPASS actives sites creates functional clusters for each ligand, which can then be used for subset generation. Cluster representatives are then found by applying principal component analysis (PCA) to the CPASS distance matrix for each ligand where 95% confidence ellipses are calculated for each EC number. The active site with the closest Euclidean distance to the center of each confidence ellipse is chosen as the cluster representative. This chapter demonstrates the effectiveness of using the CPASS representative subset, the effect it has on computational time, and novelty of functional clustering based on active site structures. For this project, Adam Caprez and David Swanson helped with the organization of the data and with the implementation of the representative tree into the CPASS webserver.

Chapter 4 pivots to more biophysical and structural studies of peptides and the relationship between the structural moieties and antibacterial properties. The rapid emergence of multidrug resistant bacterial strains and the development of permanent biofilms have made the search for new antibacterial agents of paramount importance.[24] Newer approaches have been developed which employ existing antibiotics in combination with other antibiotics, adjuvants, and other previously approved drugs.[24] Nonetheless, the dire need still exists for the discovery of new antibiotics. Cationic antimicrobial peptides (AMPs) are a class of agents that have received a lot of attention lately due to an ability to eliminate resistant pathogens and to inhibit bacterial biofilms.[25] The mechanism of action of AMPs differs from current antibiotics, which makes their potential use appealing. In general, the AMPs act as detergents that are selective for bacterial membranes and this mode of action greatly reduces the ability of infectious bacterial to gain resistance. However, beyond gross cationic, helical and hydrophobic properties of AMPs little is known about the structural requirements for activity. Citropin 1.1 is a helical AMP found on the skin of amphibians that exhibits both antibacterial and anticancer activity.[26] Previous studies suggested that Citropin 1.1 acts through an α-helical agglomerative mechanism known as carpeting.[25, 26] Employing finely tuned analogs of Citropin 1.1, we explore how slight structural variations have a remarkable influence on the AMP secondary structure and activity. Our study identified the AMP regions responsible for essential interactions that also regulate antibacterial activity and go beyond the "charge-only paradigm". Projection of the residue side chains was found to be essential to the activity of the peptides. Surprisingly, shortening the lysine side chains by one carbon led to a significant decrease in antimicrobial activity. Additionally, we found that the hydrogen bond network at the N-terminus of the peptides was crucial in helix stabilization and in the subsequent monomer to aggregate equilibrium. Active Citropin 1.1 analogs displayed a greater tendency to form α-helices and to consequently dimerize when in contact with a negatively charged membrane. In this regard, our results support the carpet-model mechanism of AMP biological activity. AMPs may be used to augment conventional antibacterial therapeutics to overcome drug resistance as part of a multidrug treatment. However, significant challenges must be solved, such as *in vivo* stability and toxicity, before they can be approved for systemic or topical use.[27] This study considerably contributes to the understanding of the biophysical properties of α-helical AMPs and will hopefully assist with future endeavors towards producing a novel antibacterial therapy. Mehdi Bin Samad, Nathalia Rodrigues de Almeida, and Martin Conda-Sheridan synthesized the peptides, performed the MIC assays, and completed the microscopy experiments during this collaborative project.

Sequence analysis was essential to the work done with Citropin 1.1. We compared sequences of known peptides with high antimicrobial activity and were able to make hypotheses about the effects of mutations or modifications. In the end, the sequence analysis was the first, and probably most significant, step towards the synthesis of an enhanced AMP. Protein function elucidation also relies heavily on amino acid sequence analysis and other bioinformatics approaches. The reliance is extended to structure homology modeling for ligand docking and protein-protein interaction mapping. Structural genomics initiatives, such as the Northeast Structure Genomics Consortium (NESG),[28] identify protein families that are sequentially similar but lack functional annotation and a representative structure. The NESG identified RPA3313, a putative protein from *Rhodopseudomonas palustris*, as a potential target. Sequence analysis of RPA3313 exposes a large, unannotated class of hypothetical proteins mostly from the *Rhizobiales* order. In the absence of sequence and structure information, further functional elucidation of this class of proteins has been significantly hindered. In chapter 5 we reveal the NMR structure of RPA3313, which adopts a novel split ββαβ fold with a conserved ligand binding pocket between the first β-strand and the N-terminus of the α-helix.[29] Conserved residue analysis and protein-protein interaction prediction analyses reveal multiple protein binding sites and conserved functional residues. Results of a mass spectrometry proteomic analysis strongly point toward interaction with the ribosome and its subunits. The combined structural and proteomic analyses suggest that RPA3313 by itself or in a larger complex may assist in the transportation of substrates to or from the ribosome for further processing. Importantly, this analysis of RPA3313 can be further extended to the remainder of the previously uncharacterized protein family identified by the NESG and aptly fulfills the core purpose of structural genomics. This work could not have been done without the assistance of Austin Lowe and Ronald Cerny, who helped with protein purification and mass spectrometry analysis, respectively.

Chapter 6 continues with structural biology in the analysis of the C-terminal domain (CTD) of DnaG primase.[30] Bacterial DNA replication initiates through a crucial interaction between two highly conserved proteins, DnaG primase and DnaB helicase. The interface between the DnaG primase C-terminal domain and the N-terminal domain of DnaB helicase is essential for bacterial DNA replication because it allows coordinated priming of DNA synthesis at the replication fork while the DNA is being unwound. Since these two proteins are conserved in all bacteria and distinct from those in eukaryotes, their interface is an attractive antibiotic target. To learn more about this interface, we determined the solution structure and dynamics of the DnaG primase CTD from *Staphylococcus aureus*, a medically important bacterial species. Comparison with the known primase CTD structures shows there are two biologically relevant conformations – an open conformation that likely binds to DnaB helicase and a closed conformation that does not. The *S. aureus* primase CTD is in the closed conformation, but NMR dynamic studies indicates there is considerable movement in the linker between the two subdomains and that N564 is the most dynamic residue within the linker. A high-throughput NMR ligand-affinity screen identified potential binding compounds, among which were acycloguanosine and myricetin. Although the affinity for these compounds and adenosine was in the millimolar range, all three bind to a common pocket that is only present on the closed conformation of CTD. This binding pocket is at the opposite end of helices 6 and 7 from N564, the key hinge residue. The identification of this binding pocket should allow the development of stronger-binding ligands that can prevent formation of the CTD open conformation that binds to DnaB helicase. Preventing the interaction between the primase CTD and helicase would significantly hinder bacterial DNA synthesis and, ultimately, reproduction. Although blocking this interaction would not result in immediate cell death, in combination with another antibiotic the treatment could potentially overcome multidrug resistance. This work was done in collaboration with Jessica Periago and Mark Griep.

Proteins are inherently dynamic and adopt various conformations at physiological conditions. The study of their dynamic behavior and discovering all of their hidden conformations is, consequently, essential to completely understanding their function. This was especially true for the primase CTD from *S. aureus*. From the dynamics study we were able to identify a key hinge residue and binding site that may provide a novel therapeutic target. Knowing firsthand the importance of dynamics relative to protein function we accomplished a similar study on human protein DJ-1, which is presented in chapter 7.

Various missense mutations in the cytoprotective protein DJ-1 (PARK7) cause rare forms of inherited Parkinsonism. One such mutation, M26I, diminishes DJ-1 protein levels in the cell but does not result in large changes in the three-dimensional structure or thermal stability of the protein. Therefore, the molecular basis for M26I DJ-1 instability and dysfunction is unknown. Chapter 7 reveals that by using NMR spectroscopy the picosecond-nanosecond timescale dynamics of wild-type and M26I DJ-1 are similar, but longer timescale (minutes to hours) dynamics are enhanced in M26I DJ-1 near physiological temperature.[31] These slow conformational fluctuations transiently expose part of the hydrophobic core of M26I DJ-1 to solvent and cause the protein to aggregate *in vitro* at 37°C. Aggregation of the protein effectively diminishes its function and removes the protein from the cellular composition. The non-pathogenic M26L mutation, in contrast, does not aggregate at 37°C and is less dynamic than M26I at physiological temperature. Therefore, M26I DJ-1 is an example of a pathogenic change in protein conformational dynamics causing a temperature-dependent loss of stability in the cell. Furthermore, the onset of M26I DJ-1 instability at physiological temperature illustrates the pitfalls of characterizing proteins exclusively at room temperature or below, as key aspects of their behavior may not be apparent. Given the prevalence of structurally conservative disease-causing mutations in proteins, we suggest that disruption of protein dynamics could be a more common pathogenic event than is currently appreciated.

Moving forward with the dynamics of DJ-1, chapter 8 introduces a methodological development that expands on the amount of interpretable data that can be obtained from an NMR experiment. NMR dynamics experiments have traditionally been collected using ^{1}H-^{15}N HSQC R_1, R_2, and heteronuclear NOE experiments with the Carr-Purcell-Meiboom-Gill (CPMG) relaxation dispersion approach.[32] However, these approaches take significant time to complete and often require are large amount of protein divided over multiple samples. In the case of M26I DJ-1, the protein would aggregate at physiological temperature and required a fresh sample for each experiment. Accordingly, there is an immediate need for new methods that reduce experimental time or extract additional information from existing datasets. Previous studies have shown that relaxation parameters and fast protein dynamics can be quickly elucidated from ^{15}N-

CEST experiments. Longitudinal R_1 and transverse R_2 values were reliably derived from fitting of CEST profiles. We show that ^{15}N-CEST experiments and traditional modelfree analysis provide the internal dynamics of three states of human protein DJ-1 at physiological temperature. The chemical exchange profiles show the absence of a minor state conformation and, in conjunction with ^{1}H-^{15}N NOEs, show increased mobility. R_1 and R_2 values remained relatively unchanged at the three naturally occurring oxidation states of DJ-1, but exhibit striking NOE differences. The NOE data was, therefore, essential in determining the internal motions of the DJ-1 proteins. To our knowledge, we present the first study that combines ^{15}N CEST data with traditional model-free analyses in the study of a biological system and affirm that more ‘lean’ model-free approaches should be used cautiously. The DJ-1 work was done in collaboration with Nicole Milkovic, Mark Wilson, and Tessa Andrews.

1.3 References

[1] Venter, J. C., Adams, M. D., Meyers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., and Wortman, J. R. (2001) The Sequence of the Human Genome, *Science 291*, 1304-1351.

[2] Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W., and Swaminathan, S. (1999) Structural Genomics: Beyond the Human Genome Project, *Nature 23*, 151-157.

[3] Chandonia, J.-M., and Brenner, S. E. (2006) The Impact of Structural Genomics: Expectations and Outcomes, *Science 311*, 347-351.

[4] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, *Nucleic Acids Research 28*, 235-242.

[5] Wuthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley & Sons.

[6] Sondergaard, C. R., Garrett, A. E., Carstensen, T., Pollastri, G., and Nielsen, J. E. (2009) Structural artifacts in protein-ligand X-ray structures: implications for the development of docking scoring functions, *J Med Chem 52*, 5673-5684.

[7] Kanelis, V., Forman-Kay, J. D., and Kay, L. E. (2001) Multidimensional Methods for Protein Structure Determination, *Life 52*, 291-302.

[8] Montelione, G. T., Zheng, D., Huang, Y. J., Gunsalus, K. C., and Szyperski, T. (2000) Protein NMR Spectroscopy in Structural Genomics, *Nature Structural Biology Structural Genomics Supplement*, 982.

[9] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016) KEGG as a reference resource for gene and protein annotation, *Nucleic Acids Res 44*, D457-462.

[10] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate,

J., and Bateman, A. (2016) The Pfam Protein Families Database: Towards a more Sustainable Future, *Nucleic Acids Res 44*, D279-285.

[11] Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G., Lehtinen, S., Studer, R. A., Thornton, J., and Orengo, C. A. (2015) CATH: comprehensive structural and functional annotations for genome sequences, *Nucleic Acids Res 43*, D376-381.

[12] Rost, B. (1999) Twilight Zone of Protein Sequence Alignments, *Protein Eng 12*, 85-94.

[13] Catazaro, J., Caprez, A., Guru, A., Swanson, D., and Powers, R. (2014) Functional Evolution of PLP-Dependent Enzymes Based on Active Site Structural Similarities *Proteins: Struct., Funct., Bioinf. 82*, 2597-2608.

[14] Baker, D., and Sali, A. (2001) Protein Structure Prediction and Structural Genomics, *Science 294*.

[15] Powers, R., Copeland, J. C., Germer, K., Mercier, K. A., Ramanathan, V., and Revesz, P. (2006) Comparison of protein active site structures for functional annotation of proteins and drug design, *Proteins 65*, 124-135.

[16] Powers, R., Copeland, J. C., Stark, J. L., Caprez, A., Guru, A., and Swanson, D. (2011) Searching the protein structure database for ligand-binding site similarities using CPASS v.2, *BMC Research Notes 4*, 17.

[17] Schmitt, S., Kuhn, D., and Klebe, G. (2002) A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology, *Journal of Molecular Biology 323*, 387-406.

[18] Gherardini, P. F., Ausiello, G., and Helmer-Citterich, M. (2010) Superpose3D: a local structural comparison program that allows for user-defined structure representations, *PLoS One 5*, e11988.

[19] Batista, J., Hawkins, P. C. D., Tolbert, R., and Geballe, M. T. (2014) SiteHopper - a unique tool for binding site comparison, *Journal of Cheminformatics 6*, P57.

[20] Nisius, B., Sha, F., and Gohlke, H. (2012) Structure-based computational analysis of protein binding sites for function and druggability prediction, *J Biotechnol 159*, 123-134.

[21] Christen, P., and Mehta, P. K. (2001) From Cofactor to Enzymes. The Molecular Evolution of Pyridoxal-5'-Phosphate-Dependent Enyzmes, *The Chemical Record 1*, 436-447.

[22] Mehta, P. K., and Christen, P. (2000) The molecular evolution of pyridoxal-5'-phosphate-dependent enzymes, *Adv. Enzymol. Relat. Areas Mol. Biol. 74*, 129-184.

[23] Alexander, F. W., Sandmeier, E., Mehta, P. K., and Christen, P. (1994) Evolutionary relationships among pyridoxal-5'-phosphate-dependent enzymes. Regio-specific α,β and γ families, *Eur. J. Biochem. 219*, 953-960.

[24] Worthington, R. J., and Melander, C. (2013) Combination approaches to combat multidrug-resistant bacteria, *Trends in biotechnology 31*, 177-184.

[25] Shai, Y., and Oren, Z. (2001) From "Carpet" Mechanism to De-novo Designed Diastereomeric Cell-Selective Antimicrobial Peptides, *Peptides 22*, 1629-1641.

[26] Doyle, J., Brinkworth, C. S., Wegener, K. L., Carver, J. A., Llewellyn, L. E., Olver, I. N., Bowie, J. H., Wabnitz, P. A., and Tyler, M. J. (2003) nNOS inhibition, antimicrobial and anticancer activity of the amphibian skin peptide, citropin 1.1 and synthetic modifications. The solution structure of a modified citropin 1.1, *European Journal of Biochemistry 270*, 1141-1153.

[27] Marr, A. K., Gooderham, W. J., and Hancock, R. E. (2006) Antibacterial peptides for therapeutic use: obstacles and realistic outlook, *Curr Opin Pharmacol 6*, 468-472.

[28] Acton, T. B., Gunsalus, K. C., Xiao, R., Ma, L. C., Aramini, J. M., Baran, M., Chiang, Y., Climent, T., Cooper, B., Denissova, N. G., douglas, S. M., Everett, J. K., Ho, C. K., Macapagal, D., Rajan, P. K., Shastry, R., Shih, L., Swapna, G. V. T., Wilson, M., Wu, M., Gerstein, M., Inouye, M., Hunt, J. F., and Montelione, G. T. (2005) Robotic Cloning and Protein Production Platform of the Northeast Structural Genomics Consortium, *Methods in enzymology 394,* 210-243.

[29] Catazaro, J., Lowe, A. J., Cerny, R. L., and Powers, R. (2017) The NMR solution structure and function of RPA3313: a putative ribosomal transport protein from Rhodopseudomonas palustris, *Proteins 85*, 93-102.

[30] Catazaro, J., Periago, J., Shortridge, M. D., Worley, B., Kirchner, A., Powers, R., and Griep, M. A. (2017) Identification of a Ligand-Binding Site on the Staphylococcus aureus DnaG Primase C-Terminal Domain, *Biochemistry 56*, 932-943.

[31] Milkovic, N. M., Catazaro, J., Lin, J., Halouska, S., Kizziah, J. L., Basiaga, S., Cerny, R. L., Powers, R., and Wilson, M. A. (2015) Transient sampling of aggregation-prone conformations causes pathogenic instability of a parkinsonian mutant of DJ-1 at physiological temperature, *Protein Sci 24*, 1671-1685.

[32] Kay, L. E. (2016) New Views of Functionally Dynamic Proteins by Solution NMR Spectroscopy, *J Mol Biol 428*, 323-331.

Chapter 2

Functional Evolution of PLP-dependent Enzymes Based on Active site Structural Similarities

2.1 Introduction

Sequence comparison methods and structural homology models are often used to imply evolutionary relationships between proteins in the same family or of the same function.[1, 2] These analyses are additionally used to infer the function of uncharacterized proteins.[3, 4] Enhancing our understanding of protein evolution furthers our insights into cellular processes, enables the discovery of drug targets and disease markers, while validating or establishing phylogenies of organisms, among numerous other beneficial impacts.[5] However, evolutionary studies are limited for distantly related proteins in which sequence and structural similarity may not be easily deduced. Sequence comparison methods assume a constant rate of amino acid substitution and are not suitable for all phylogenetic reconstructions of ancestral sequences. Correspondingly, for sequences with minimal identity local conformations can vary greatly and may inhibit structural studies.[6]

It is understood that although the rate of sequence evolution is highly variable, residues essential to function and active site chemistry will remain highly conserved over time.[7] Slight changes in the position and identity of an active site residue can lead to changes in both substrate and reaction specificity.[1, 8] In fact, it has been proposed that catalytic residues maintain their position and identity longer than the configuration of some secondary structures.[9] The importance of these residues has spawned many computational approaches to assist with the prediction of active site residues,[10-12] the functional annotation of proteins based on active sites,[13-15] as well as gene functional annotation.[16, 17] Additionally, the evolution of homologous proteins sharing low sequence identities has been inferred using active site structure geometries.[9, 18, 19] These observations are consistent with the premise that molecular evolution and functional evolution proceed through distinct processes. Molecular evolution is driven by random mutational events where maintaining function is paramount.[1, 2] Conversely, protein functional evolution requires a new gene through duplication, acquisition or creation that enables protein functional drift while avoiding a negative impact on cell fitness.[20-22] Correspondingly, analyzing functional evolution may require a different approach from the global sequence comparisons routinely used for molecular evolution.[1, 2] In this manner, active site structures may be a suitable alternative for inferring an evolutionary relationship based on function.[23]

Some of the most diverse and versatile classes of enzymes are those that utilize pyridoxal-5'-phosphate (PLP) as a cofactor.[24] These enzymes are found in five of the six classes defined by the Enzyme Commission and catalyze a multitude of different reactions.[25] The diversity of the reactions can be attributed to the electron sink properties of the cofactor[26] and the stereochemical restraints in the active site of the enzyme.[27-29] PLP-dependent enzymes acting on amino acid substrates can be separated into four families of paralogous proteins annotated as fold-types I-IV.[26, 30] Interestingly, each of the fold-types are highly promiscuous in terms of function, and multiple examples of convergent evolution have been discovered.[24] These observations have led to the hypothesis that the four lineages of PLP-dependent enzymes developed in the last common universal ancestor.

Extensive work has been done with sequence-based methods to establish a phylogenetic relationship among PLP-dependent enzymes.[24, 31, 32] These studies, however, have been limited to small subgroups due to the low sequence identity within each fold-type and across all PLP-dependent enzymes. Remarkably, the functions of these enzymes are highly dependent on the structure of their active sites. Although these enzymes lend themselves to an evolutionary study based on active site structures, one has not yet been undertaken.

Here, we use a novel method to compare the active site structures of 204 PLP-dependent enzymes from the four different fold-types that are specifically involved in amino acid metabolism. The Comparison of Protein Active Site Structures (CPASS) software and database was developed to aid in the functional annotation of uncharacterized proteins by comparing ligand defined active sites.[15, 33] As a sensitive measure of active site geometry, it was proposed that CPASS could also be used to model families of divergent proteins. In this work we present the functional evolution of PLP-dependent enzymes based on active site structures. To the authors' knowledge, this is the first study of its kind that models the functional evolution of a superfamily of proteins by active site comparisons, and the first to model all PLP-dependent fold-types together on one network. Results were compared to previous studies using both sequence and structure-guided methods. We find that the active site comparisons successfully cluster the enzymes based on both function and three-dimensional fold-type.

2.2 Materials and Methods

2.2.1 Active Site Structure Comparison

Three-dimensional structures with ligand-defined active sites were gathered from the Protein Data Bank on May 9, 2014.[34] An all versus all comparison was carried out between the active site structures using the Comparison of Protein Active Site Structures (CPASS) software. CPASS finds the optimum structural alignment between two active sites by maximizing a scoring function based on structure ($C\alpha$ and $C\beta$ RMSD) and sequence (BLOSUM62) similarity.[35, 36] The scoring function is the summation of an ideal pairing of residues between each of the active sites rather than a summation of all possible pairings. Therefore, it is possible that active site residues remain unpaired during analysis. Since the CPASS algorithm is dependent upon the size of the active site query, an asymmetrical similarity matrix was generated from the all versus all comparison. The lower score of the two pairwise comparisons was extracted and converted to a distance by subtracting from 100% similarity. From the resulting distance matrix, only the PLP bound proteins that acted on amino acid substrates were extracted. Redundancies in the PLP-distance matrix were removed. A total of 204 X-ray crystal structures and corresponding active sites were analyzed in this study.

2.2.2 Protein Annotation

A total of 204 PLP-dependent proteins were functionally annotated with Enzyme Commission (EC) numbers using the UniProt database[37] and in-house software. Enzymes that did not have an EC number in the UniProt database were assigned manually using the PDB. For enzymes with more than one number or classification, only the first was taken. Species, phylogenetic kingdom, and fold-type annotations were done manually using the PDB and UniProt databases as well as literature sources.

2.2.3 Phylogenetic Analysis

A phylogenetic network using all of the PLP-dependent enzymes was generated with Splitstree4[38] using the Neighbor-Net algorithm.[39] The network represents a deterministic model of the distances obtained from the CPASS algorithm and software. Identification of the four fold-type clusters was accomplished by visual inspection of the network. Each of the clusters is continuous and does not contain enzymes that are not of the same fold-type. Phylogenetic trees were created for each of the fold-types using Dendroscope,[40] CPASS distances, and the Neighbor-join algorithm. Neighbor joining is an agglomerative clustering method in which a tree is built from a distance matrix. Briefly, the two closest taxa in the distance matrix are joined to create a new node. Distances from each of the remaining taxa to the new node are recalculated and the matrix is updated. As this process is repeated, the distance matrix shrinks until only a pair of nodes remains.

2.2.4 Sequence Identity

The PDB IDs for the 204 PLP-dependent enzymes in this study were submitted to the UniProt database for primary sequence retrieval. The sequences were subsequently submitted to Clustal Omega[41] for multiple sequence alignment (MSA) using the default parameters. The MSA and percent identity matrix were retrieved upon completion and visualized in the R software environment. A histogram was generated with intervals of 0.2 percent identity and represents the entire, non-condensed percent identity matrix. A heat map of the entire matrix was also generated with a white to red gradient corresponding to low and high percent identity. The ordering of the sequences in the heat map corresponds to the ordering in Table S1.

2.2.5 Structure Alignment

The 204 PLP-dependent enzymes were further subjected to an all versus all structural alignment using TM-align.[42] Results from the pairwise comparisons were organized into a matrix and were visualized in the R software environment. A histogram was generated with TM-align score intervals of 0.001 and represents the entire symmetrical matrix. Correspondingly, a heat map of the matrix was generated with a white to red gradient representing low and high TM-align scores, respectively. Ordering of the heat map is the same as the sequence identity ordering and follows the order in Table S1.

2.2.6 Comparison of Type-I Transaminase Active Sites

From the 204 PLP-dependent enzymes in the CPASS dataset, 4 enzymes were selected to elucidate the active site differences that encode substrate specificity. An acetylornithine (ACO) transaminase (PDB ID: 2ORD), a γ-aminobutyric acid (GABA) transaminase (PDB ID: 1OHW), an aspartate (Asp) transaminase (PDB ID: 1ARG), and a tyrosine (Tyr) transaminase (PDB ID: 3DYD) were selected from functional clusters 1, 1, 4b, and 4a, respectively. Active site residues for each of the enzymes were selected based on their proximity to the PLP ligand (6 Å) as defined by CPASS. Active sites were overlaid in Chimera[43] and visually inspected for the changes resulting in different substrate specificity. Additional transaminases with the same function as the four in question were subjected to individual multiple sequence alignments in Clustal Omega. The additional enzymes were manually selected from the PDB and were not present in our original dataset because a bound ligand was not present in the structure.

2.3 Results

2.3.1 Phylogenetic Network Analysis

Phylogenetic analyses were carried out on 204 PLP-dependent enzymes with bound cofactor (Table S1) from each of the four fold-types previously defined in the literature.[26] It is well known that each of the fold-types originated from separate evolutionary lineages. Using sequence-based phylogenetic methods,[24] each of the lineages has been visualized with a molecular evolutionary tree structure. However, due to their low sequence identity visualizing all of the fold-types together has not yet been accomplished. Using active site similarity comparisons, it is possible to display all of the PLP-dependent enzymes in one phylogenetic network that demonstrates an evolutionary relationship based on function (Figure 1). Type-I enzymes make up the largest portion of the PLP enzymes and encompasses most of the enzymes that act at the Cα position of the substrate. Type-II enzymes catalyze reactions at the Cα and Cβ positions and contain fewer members compared to Type-I. The final two folds, Types III and IV, have the fewest members of PLP-dependent enzymes. The function-based phylogenetic network suggests that each of the fold-types originated from an independent lineage. The main branches of the network all spawn from the center, indicating a universal organism contained ancestral enzymes for each of the fold-types. Also evident are clear separations within each of the groups that form function-specific clusters (numbered 1 through 13).

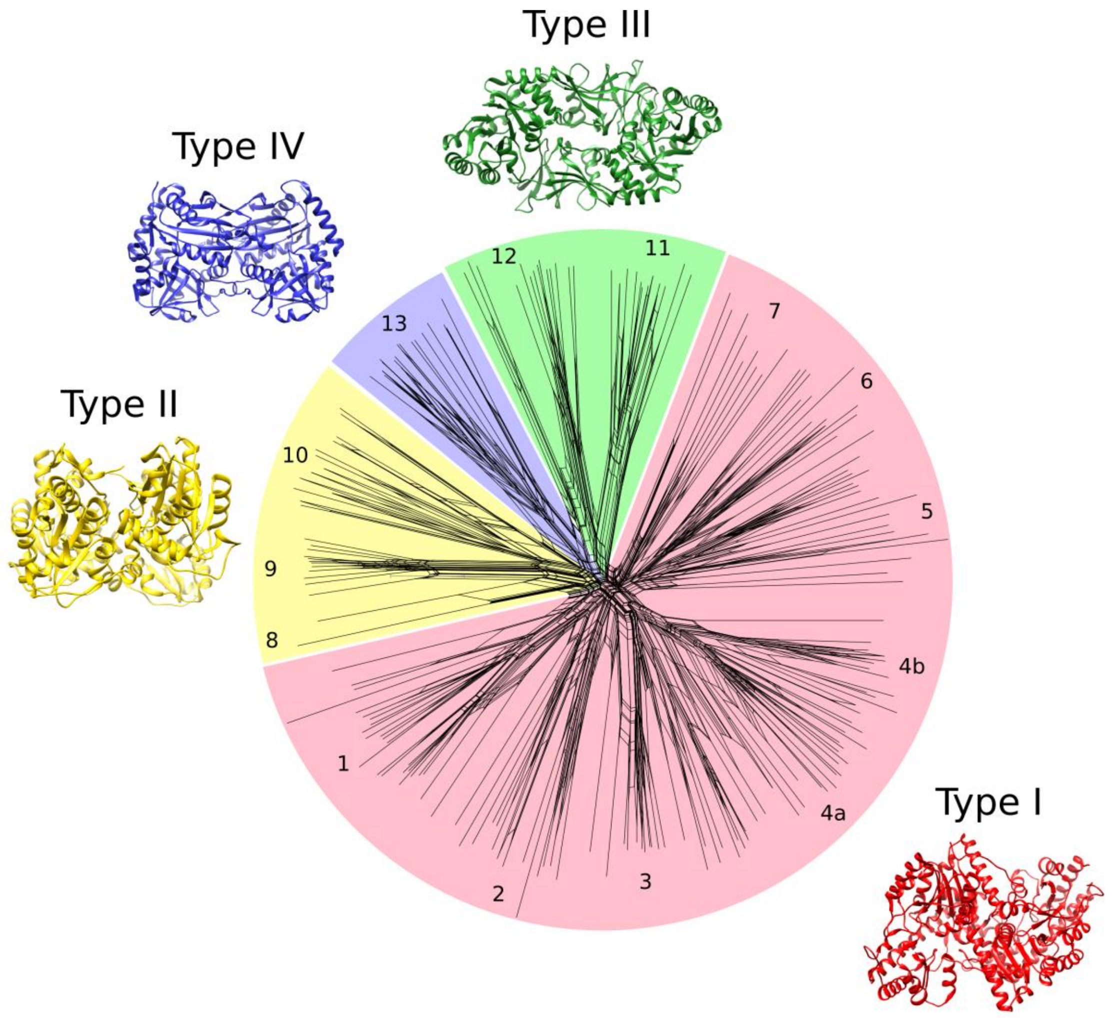


Figure 1: Active site network analysis of PLP-dependent enzymes. A phylogenetic network analysis of 204 PLP-dependent enzymes with ligand defined active sites in the PDB. The red, yellow, green and blue colored regions and protein structures correspond to fold-types I, II, III and IV, respectively. Individual clusters found within each of the

folds represent groups of function specific enzymes and are numbered 1-13. The network was generated using CPASS similarity scores and the Neighbor-net method.

The functional network also highlights the advantages of using an enzyme’s active site to elucidate an evolutionary relationship. Each of the PLP-dependent fold-types has a defined active site geometry, and within each type, each function-specific cluster has a unique variation in the active site geometry. However, the functional network only gives a general overview of the relatedness between each of the functions in the fold-types and between the fold-types. To make valid inferences of functional evolutions, further analysis is still necessary.

2.3.2 Phylogenetic Analysis of Type-I Enzymes

Active site similarity scores from the enzymes in the Type-I fold-type were analyzed in detail to find a phylogenetic relationship among functional clusters 1-7 (Figure 2). The dominating function of fold-type I involves transaminase activity (EC 2.6.1), which is found in clusters 1, 2, 4 and 7. Clusters 3, 5 and 6 are primarily comprised of carbon-sulfur lyases (EC 4.4.1), decarboxylases (EC 4.1.1), and glycine hydroxymethyltransferases (EC 2.1.2.1), respectively. Although clusters 1, 2, 4 and 7 share a broad transaminase activity, these clusters are punctuated in the phylogenetic tree by the remaining functional clusters. This observation suggests multiple Type-I enzymes existed in a common ancestor with specialized function. Conversely, a continuous grouping of all the transaminases would have implied the existence of a single transaminase with broad substrate specificity. Previous studies have recognized similar divisions within the Type-I fold using sequence homology methods and structural alignment tools; and came to the same conclusion.[44-46] The punctuation in the active site similarity tree highlights the sensitivity of CPASS for comparing residues and geometries of active sites; and supports its use as a phylogenetic tool.

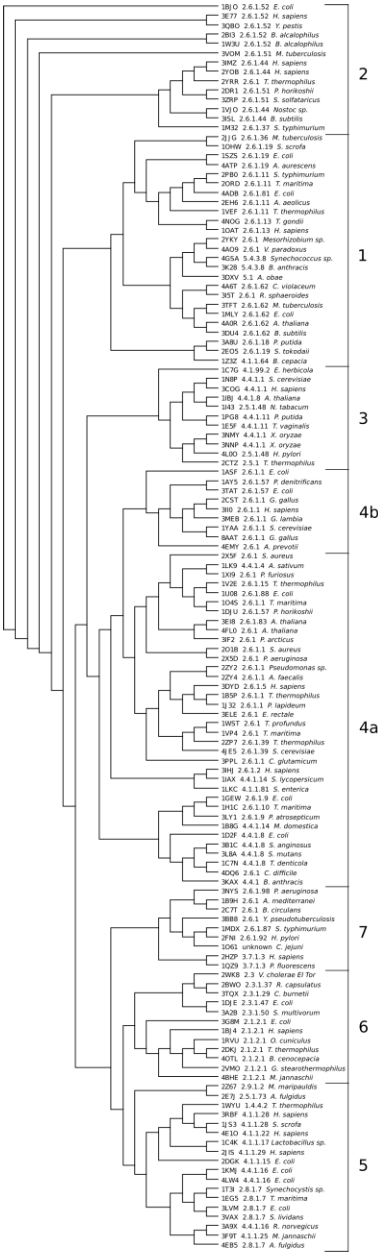

1BJO 2.6.1.52 *E. coli*
3E77 2.6.1.52 *H. sapiens*
3QBO 2.6.1.52 *Y. pestis*
2BI3 2.6.1.52 *B. alcalophilus*
1W3U 2.6.1.52 *B. alcalophilus*
3VOM 2.6.1.51 *M. tuberculosis*
3IMZ 2.6.1.44 *H. sapiens*
2YOB 2.6.1.44 *H. sapiens*
2YRR 2.6.1 *T. thermophilus*
2DR1 2.6.1.51 *P. horikoshii*
3ZRP 2.6.1.51 *S. solfataricus*
1VJO 2.6.1.44 *Nostoc sp.*
3ISL 2.6.1.44 *B. subtilis*
1M32 2.6.1.37 *S. typhimurium*
2JJG 2.6.1.36 *M. tuberculosis*
1OHW 2.6.1.19 *S. scrofa*
1SZS 2.6.1.19 *E. coli*
4ATP 2.6.1.19 *A. aurescens*
2PB0 2.6.1.11 *S. typhimurium*
2ORD 2.6.1.11 *T. maritima*
4ADB 2.6.1.81 *E. coli*
2EH6 2.6.1.11 *A. aeolicus*
1VEF 2.6.1.11 *T. thermophilus*
4NOG 2.6.1.13 *T. gondii*
1OAT 2.6.1.13 *H. sapiens*
2YKY 2.6.1 *Mesorhizobium sp.*
4AO9 2.6.1 *V. paradoxus*
4GSA 5.4.3.8 *Synechococcus sp.*
3K28 5.4.3.8 *B. anthracis*
3DXV 5.1 *A. obae*
4A6T 2.6.1.62 *C. violaceum*
3I5T 2.6.1 *R. sphaeroides*
3TFT 2.6.1.62 *M. tuberculosis*
1MLY 2.6.1.62 *E. coli*
4A0R 2.6.1.62 *A. thaliana*
3DU4 2.6.1.62 *B. subtilis*
3A8U 2.6.1.18 *P. putida*
2EO5 2.6.1.19 *S. tokodaii*
1Z3Z 4.1.1.64 *B. cepacia*
1C7G 4.1.99.2 *E. herbicola*
1N8P 4.4.1.1 *S. cerevisiae*
3COG 4.4.1.1 *H. sapiens*
1IBJ 4.4.1.8 *A. thaliana*
1I43 2.5.1.48 *N. tabacum*
1PG8 4.4.1.11 *P. putida*
1E5F 4.4.1.11 *T. vaginalis*
3NMY 4.4.1.1 *X. oryzae*
3NNP 4.4.1.1 *X. oryzae*
4L0O 2.5.1.48 *H. pylori*
2CTZ 2.5.1 *T. thermophilus*
1ASF 2.6.1.1 *E. coli*
1AY5 2.6.1.57 *P. denitrificans*
3TAT 2.6.1.57 *E. coli*
2CST 2.6.1.1 *G. gallus*
3II0 2.6.1.1 *H. sapiens*
3MEB 2.6.1.1 *G. lambia*
1YAA 2.6.1.1 *S. cerevisiae*
8AAT 2.6.1.1 *G. gallus*
4EMY 2.6.1 *A. prevotii*
2X5F 2.6.1 *S. aureus*
1LK9 4.4.1.4 *A. sativum*
1XI9 2.6.1 *P. furiosus*
1V2E 2.6.1.15 *T. thermophilus*
1U08 2.6.1.88 *E. coli*
1O4S 2.6.1.1 *T. maritima*
1DJU 2.6.1.57 *P. horikoshii*
3EI8 2.6.1.83 *A. thaliana*
4FL0 2.6.1 *A. thaliana*
3IF2 2.6.1 *P. arcticus*
2O1B 2.6.1.1 *S. aureus*
2X5D 2.6.1 *P. aeruginosa*
2ZY2 2.6.1.1 *Pseudomonas sp.*
2ZY4 2.6.1.1 *A. faecalis*
3DYD 2.6.1.5 *H. sapiens*
1B5P 2.6.1.1 *T. thermophilus*
1J32 2.6.1.1 *P. lapideum*
3ELE 2.6.1 *E. rectale*
1WST 2.6.1 *T. profundus*
1VP4 2.6.1 *T. maritima*
2ZP7 2.6.1.39 *T. thermophilus*
4JE5 2.6.1.39 *S. cerevisiae*
3PPL 2.6.1.1 *C. glutamicum*
3IHJ 2.6.1.2 *H. sapiens*
1IAX 4.4.1.14 *S. lycopersicum*
1LKC 4.1.1.81 *S. enterica*
1GEW 2.6.1.9 *E. coli*
1H1C 2.6.1.10 *T. maritima*
3LY1 2.6.1.9 *P. atrosepticum*
1B8G 4.4.1.14 *M. domestica*
1D2F 4.4.1.8 *E. coli*
3B1C 4.4.1.8 *S. anginosus*
3L8A 4.4.1.8 *S. mutans*
1C7N 4.4.1.8 *T. denticola*
4DQ6 2.6.1 *C. difficile*
3KAX 4.4.1 *B. anthracis*
3NYS 2.6.1.98 *P. aeruginosa*
1B9H 2.6.1 *A. mediterranei*
2C7T 2.6.1 *B. circulans*
3BB8 2.6.1 *Y. pseudotuberculosis*
1MDX 2.6.1.87 *S. typhimurium*
2FNI 2.6.1.92 *H. pylori*
1O61 unknown *C. jejuni*
2HZP 3.7.1.3 *H. sapiens*
1QZ9 3.7.1.3 *P. fluorescens*
2WK8 2.3 *V. cholerae El Tor*
2BWO 2.3.1.37 *R. capsulatus*
3TQX 2.3.1.29 *C. burnetii*
1DJE 2.3.1.47 *E. coli*
3A2B 2.3.1.50 *S. multivorum*
3G8M 2.1.2.1 *E. coli*
1BJ4 2.1.2.1 *H. sapiens*
1RVU 2.1.2.1 *O. cuniculus*
2DKJ 2.1.2.1 *T. thermophilus*
4OTL 2.1.2.1 *B. cenocepacia*
2VMO 2.1.2.1 *G. stearothermophilus*
4BHE 2.1.2.1 *M. jannaschii*
2Z67 2.9.1.2 *M. maripauldis*
2E7J 2.5.1.73 *A. fulgidus*
1WYU 1.4.4.2 *T. thermophilus*
3RBF 4.1.1.28 *H. sapiens*
1JS3 4.1.1.28 *S. scrofa*
4E1O 4.1.1.22 *H. sapiens*
1C4K 4.1.1.17 *Lactobacillus sp.*
2JIS 4.1.1.29 *H. sapiens*
2DGK 4.1.1.15 *E. coli*
1KMJ 4.4.1.16 *E. coli*
4LW4 4.4.1.16 *E. coli*
1T3I 2.8.1.7 *Synechocystis sp.*
1EG5 2.8.1.7 *T. maritima*
3LVM 2.8.1.7 *E. coli*
3VAX 2.8.1.7 *S. lividans*
3A9X 4.4.1.16 *R. norvegicus*
3F9T 4.1.1.25 *M. jannaschii*
4EB5 2.8.1.7 *A. fulgidus*
2
1
3
4b
4a
7
6
5

Figure 2: Type-I PLP-dependent enzyme phylogenetic analysis. The phylogenetic tree

of PLP-dependent enzymes belonging to the Type-I fold. Each leaf of the tree is annotated with PDB ID, EC number and species. Branch lengths represent the evolutionary distance between enzymes based on active site structure comparisons. Functional clusters 1-7 correspond to the numbers in the network analysis of all PLP-dependent enzymes (Figure 1). The tree was constructed using CPASS similarity scores and the Neighbor-join algorithm.

There are multiple examples of bacterial and eukaryotic species with homologous enzymes within the Type-I functional clusters (Table S1). A general idea of the extent of enzyme divergence from a universal ancestor can be elucidated based on how these homologous enzymes group within the Type-I functional clusters. In cluster 3, the methionine gamma-lyases (EC 4.4.1.11) belonging to eukaryotic *T. vaginalis* and bacterial *P. putida* were found to be closely related. Directly following in the phylogentic tree are the cystathionine gamma-lyases (EC 4.4.1.1) for both eukaryotes and bacteria. This indicates that a methionine gamma-lyase and a cystathionine gamma-lyase existed in a universal ancestor. Similar examples are also found in clusters 2, 6, and 7 for phosphoserine transaminases (EC 2.6.1.52), kynureninases (EC 3.7.1.3) and glycine hydroxymethyltransferases (EC 2.1.2.1), respectively.

Aspartate transaminases are present in both clusters 4a and 4b, but a clear separation between these two clusters is apparent in the phylogenetic tree shown in Figure 1. This is due to the fact that eukaryotic aspartate transaminases are only found in cluster 4b, while bacterial aspartate transaminases are found in both clusters 4a and 4b. The eukaryotic enzymes in cluster 4a are comprised of carboxylate synthases (EC 4.4.1) from the kingdom *Plantae*, and tyrosine and alanine transaminases (EC 2.6.1.5 and 2.6.1.2) from the kingdom *Animalia*. This finding does not support the idea that all Type-I enzymes became highly specialized in the last universal ancestor. Instead, our results suggest that a promiscuous aspartate transaminase was passed on to emergent species and, through subsequent speciation and gene duplication events, the enzymes in clusters 4a and 4b became more functionally specialized and more substrate specific.

2.3.3 Phylogenetic Analysis of Type-II Enzymes

A detailed phylogenetic analysis was also carried out on the Type-II PLP-dependent enzymes using CPASS similarity scores and the Neighbor-Join algorithm (Figure 3a). Three functional clusters based on active site similarity were identified for this fold-type and are labeled as clusters 8, 9 and 10. Enzymes catalyzing deaminase reactions on 1-aminocyclopropane-1-carboxylate (EC 3.5.99.7) are found exclusively in cluster 8. These enzymes belong to species in archaeal and eukaryal domains indicating that their function was specialized in a common ancestor. The same observation was made for the enzymes in cluster 10. Threonine synthase (EC 4.2.3.1), tryptophan synthase (EC 4.2.1.20), and serine dehydratase/threonine deaminase (EC 4.3.1.17/4.3.1.19) enzymes from archaeal, bacterial, and eukaryal species cluster by function rather than by domain. Cluster 9 contains cysteine synthases (EC 2.5.1.47) from the bacterial, archaeal, and eukaryal domains and cystathionine β-synthases (EC 4.2.1.22) from the eukaryal domain. The clustering of the cysteine synthases by function rather than by domain once again indicates specialization in a common ancestor; however, an insufficient number of cystathionine β-synthases are represented in the dataset to elucidate its functional evolution.

2.3.4 Phylogenetic Analysis of Type-III Enzymes

Phylogenetic analysis was also undertaken to explain the emergence of enzymes belonging to the Type-III fold-type (Figure 3b). Clusters 11 and 12 were found to be discernibly different from one another in the phylogenetic network (Figure 1) and, correspondingly, separate from one another on the phylogenetic tree created with the Neighbor-Join algorithm. Alanine racemases (EC 5.1.1.1) exclusively populate functional cluster 12; whereas, functional cluster 11 is made up of three comparable decarboxylases. Unlike most of the enzymes in fold-types I and II, it does not appear the Type-III enzymes specialized in the last common ancestor.

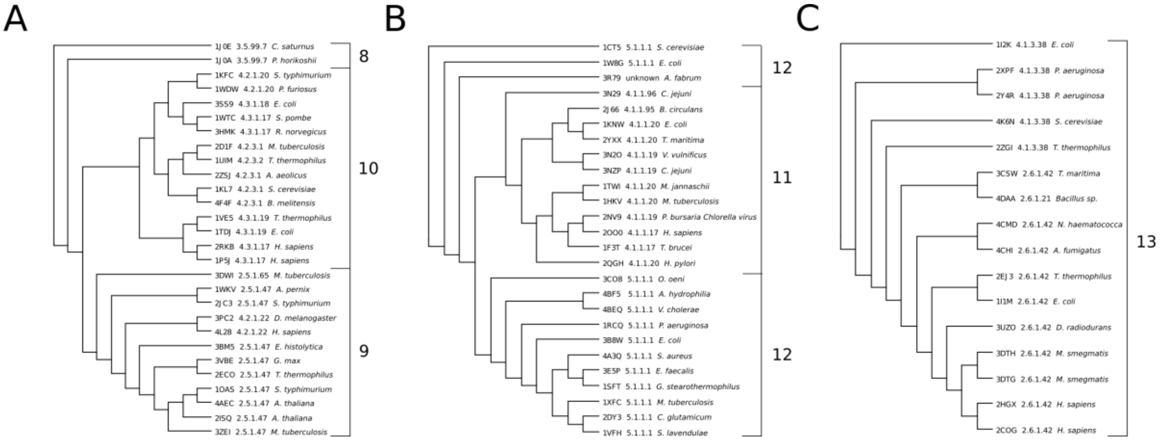

A
1J0E 3.5.99.7 C. saturnus
1J0A 3.5.99.7 P. horikoshii
1KFC 4.2.1.20 S. typhimurium
1WDW 4.2.1.20 P. furiosus
3SS9 4.3.1.18 E. coli
1WTC 4.3.1.17 S. pombe
3HMK 4.3.1.17 R. norvegicus
2D1F 4.2.3.1 M. tuberculosis
1UIM 4.2.3.2 T. thermophilus
2ZSJ 4.2.3.1 A. aeolicus
1KL7 4.2.3.1 S. cerevisiae
4F4F 4.2.3.1 B. melitensis
1VE5 4.3.1.19 T. thermophilus
1TDJ 4.3.1.19 E. coli
2RKB 4.3.1.17 H. sapiens
1P5J 4.3.1.17 H. sapiens
3DWI 2.5.1.65 M. tuberculosis
1WKV 2.5.1.47 A. pernix
2JC3 2.5.1.47 S. typhimurium
3PC2 4.2.1.22 D. melanogaster
4L28 4.2.1.22 H. sapiens
3BM5 2.5.1.47 E. histolytica
3VBE 2.5.1.47 G. max
2ECO 2.5.1.47 T. thermophilus
1OAS 2.5.1.47 S. typhimurium
4AEC 2.5.1.47 A. thaliana
2ISQ 2.5.1.47 A. thaliana
3ZEI 2.5.1.47 M. tuberculosis
8
10
9
B
1CT5 5.1.1.1 S. cerevisiae
1W8G 5.1.1.1 E. coli
3R79 unknown A. fabrum
3N29 4.1.1.96 C. jejuni
2J66 4.1.1.95 B. circulans
1KNW 4.1.1.20 E. coli
2YXX 4.1.1.20 T. maritima
3N2O 4.1.1.19 V. vulnificus
3NZP 4.1.1.19 C. jejuni
1TWI 4.1.1.20 M. jannaschii
1HKV 4.1.1.20 M. tuberculosis
2NV9 4.1.1.19 P. bursaria Chlorella virus
2OO0 4.1.1.17 H. sapiens
1F3T 4.1.1.17 T. brucei
2QGH 4.1.1.20 H. pylori
3CO8 5.1.1.1 O. oeni
4BF5 5.1.1.1 A. hydrophilia
4BEQ 5.1.1.1 V. cholerae
1RCQ 5.1.1.1 P. aeruginosa
3B8W 5.1.1.1 E. coli
4A3Q 5.1.1.1 S. aureus
3E5P 5.1.1.1 E. faecalis
1SFT 5.1.1.1 G. stearothermophilus
1XFC 5.1.1.1 M. tuberculosis
2DY3 5.1.1.1 C. glutamicum
1VFH 5.1.1.1 S. lavendulae
12
11
12
C
1I2K 4.1.3.38 E. coli
2XPF 4.1.3.38 P. aeruginosa
2Y4R 4.1.3.38 P. aeruginosa
4K6N 4.1.3.38 S. cerevisiae
2ZGI 4.1.3.38 T. thermophilus
3CSW 2.6.1.42 T. maritima
4DAA 2.6.1.21 Bacillus sp.
4CMD 2.6.1.42 N. haematococca
4CHI 2.6.1.42 A. fumigatus
2EJ3 2.6.1.42 T. thermophilus
1I1M 2.6.1.42 E. coli
3UZO 2.6.1.42 D. radiodurans
3DTH 2.6.1.42 M. smegmatis
3DTG 2.6.1.42 M. smegmatis
2HGX 2.6.1.42 H. sapiens
2COG 2.6.1.42 H. sapiens
13

Figure 3: Types-II, -III, and -IV PLP-dependent enzyme phylogenetic analysis. Phylogenetic reconstructions of PLP-dependent enzymes from Types-II (A), -III (B), and -IV (C) folds were built using CPASS similarity scores and the Neighbor-join algorithm. Enzymes are labeled with their PDB ID, an EC number and organism. Tree branch

lengths correspond to the evolutionary distance between the active sites of the enzymes. Numbers 8-13 represent the functional clusters found in the network analysis (Figure 1).

In our dataset, diaminopimelate decarboxylases (DAPDC, EC 4.1.1.20) and arginine decarboxylases (ADC, EC 4.1.1.19) from bacterial and archaeal species are present while ornithine decarboxylases (OCD, EC 4.1.1.17) appear only from eukaryal species. However, function specific clustering of the enzymes in cluster 11 is not observed. It is likely that a decarboxylase with a very generic active site existed in a common ancestor due to the kingdom specific characteristics of the current enzymes.[47, 48] Additionally, ODCs for bacterial species can be found in functional cluster 5 of the Type-I enzymes. This example of convergent evolution further supports the idea of a broad-based decarboxylase belonging to the Type-III fold. A viral ADC (PDB ID: 2NV9) is also present within functional cluster 11, but does not cluster with the bacterial ADCs. It was previously hypothesized that this enzyme was acquired from a bacterial ornithine decarboxylase and subsequently mutated toward arginine specificity.[49] CPASS analysis supports this proposal as the enzyme clusters with eukaryal ODCs and a bacterial DAPDC instead of the bacterial ADCs. Specifically, this finding represents an example of convergent evolution within the Type-III fold.

2.3.5 Phylogenetic Analysis of Type-IV Enzymes

Type-IV enzymes occupy the smallest portion of all PLP-dependent enzymes and form only one functional cluster (Figure 3c). Cluster 13 contains aminodeoxychorismate lyases (EC 4.1.3.38), branched chain amino acid transaminases (EC 2.6.1.42) and a D-amino acid transaminase (EC 2.6.1.21). Similar to the other three fold-types, the function specific enzymes in this family group together. However, because there is only one D-amino acid transaminase represented in the dataset, the enzyme is grouped with the branched chain amino acid transaminases. Interestingly, D-amino acid transaminases share similar active site characteristics with the Type-I transaminases.[50] The geometry and residue identity is different enough, however, to distinguish between Types I and IV enzymes without having to look at the three-dimensional structure.

2.3.6 Sequence Identity and Structural Similarity

A heat map and corresponding histogram was generated from a multiple sequence alignment (MSA) of the 204 PLP-dependent enzymes (Figure 4a-b). Values above 20% identity belong to self-comparisons and comparisons of closely related homologous enzymes within the same organism. This can be seen visually as the diagonal line in the corresponding heat map. As expected, the histogram shows that the majority of the pairwise comparisons are well-below 20% sequence identity. This is consistent with prior observations that PLP-dependent enzymes have very low sequence identity.[24, 31, 32] Importantly, establishing a protein function based solely on sequence homology significantly decreases in reliability as sequence identity falls below 50%.[51] Presumably, a similar high level of uncertainty in establishing a functional evolutionary relationship would occur for proteins with sequence identity well-below 20%.

The 3D structures of PLP-dependent enzymes were also analyzed to ascertain the reliability of using structures of distantly related proteins to generate a functional evolutionary model. Figure 5a-b illustrates the resulting histogram and heat map of the pairwise structural alignments of the 204 PLP-dependent enzymes. The histogram shows a bimodal distribution of TM-align scores. TM-align scores greater than 0.5 are consistent with two proteins sharing the same CATH or SCOP fold. Correspondingly, the ~0.6 TM-align scores occur for PLP-dependent enzymes within the same fold-type. Conversely, the lower TM-align scores result from comparisons between fold-types.

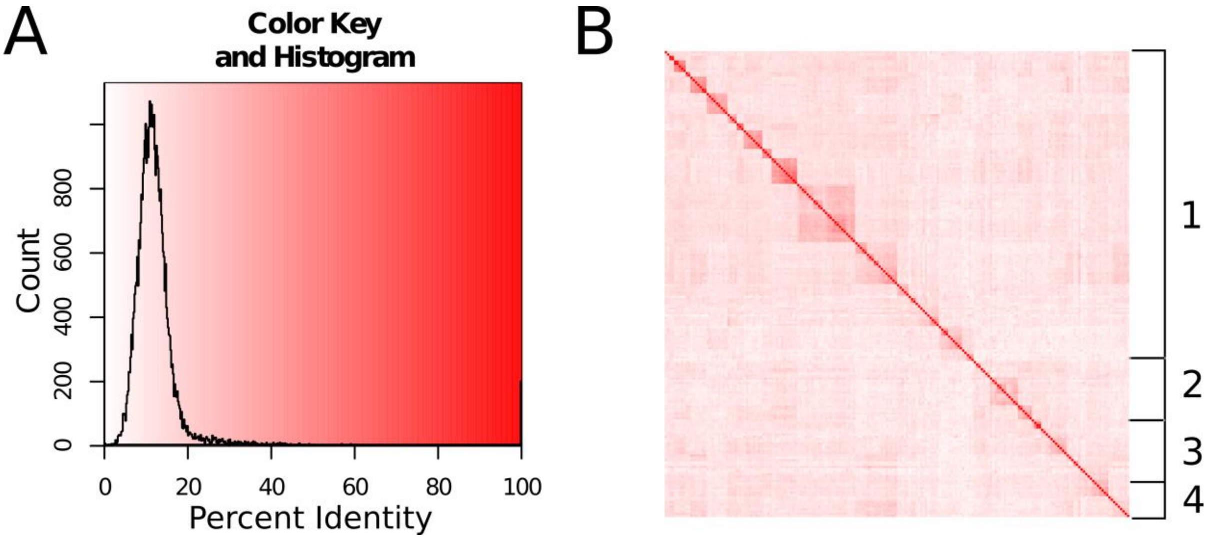


Figure 4: Sequence identity histogram and heat map. (A) A histogram of the percent similarity matrix representing 204 PLP-dependent protein sequences. The entire similarity matrix was analyzed in 0.2% intervals to generate the plot. (B) A heat map of the percent similarity matrix used in (A). The white to red coloring of the heat map corresponds to the percent similarity and color gradient in (A). The order of the

sequences is the same as in Table S1 and the associated network map is illustrated in Figure S1.

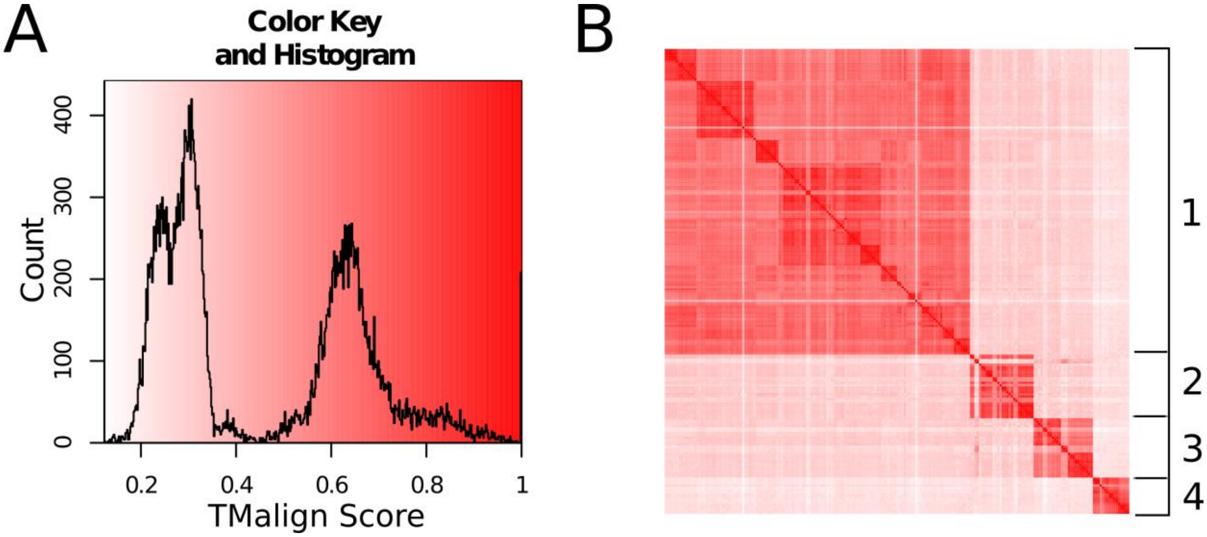


Figure 5: Structural alignment histogram and heat map. (A) A histogram of the pairwise TMalign scores for the 204 PLP-dependent proteins in the dataset. The entire symmetrical score matrix was evaluated in 0.002 intervals to produce the plot. (B) A heat map of the corresponding TMalign score matrix used in (A). The white to red coloring scheme of the heat map corresponds to the score and gradient in the respective histogram. The order of the sequences is the same as in Table S1 and the associated network map is illustrated in Figure S2.

The associated heat map of TM-align scores clarifies this point; high TM-align scores are clustered by fold-type. Although structural alignments are able to separate the fold-types and determine proper PLP-dependent enzyme membership within a fold-type, the global structural information is not sensitive enough to cluster the PLP-dependent enzymes based on function. In essence, the global structural similarity between enzymes within a fold-type over-whelms any subtle differences related to functional divergence. Similarly, the large structural differences between fold-types negate any local similarities that may be present due to a common functional ancestor.

2.3.7 Comparison of Type-I Transaminase Active Sites

The evolutionary network based on active site similarity depicted in Figure 1 identified acetylornithine (ACO) transaminase (EC 2.6.1.11, PDB ID: 2ORD) and a γ-aminobutyric acid (GABA) transaminase (EC 2.6.1.19, PDB ID: 1OHW) as nearest neighbors in the network. Both enzymes correspond to fold-type I and are located in cluster 1, but the enzymes have different EC numbers and substrate specificity. Thus, the active sites were compared in order to reveal structural change(s) that contribute to the different substrate specificities that led to a functional divergence (Figures 6a-b). The goal was to further explore the validity of the phylogenetic network based on CPASS similarity scores to provide insights on the pathway of functional evolution. Further pairwise comparisons were also made with enzymes from functional clusters 4a and 4b in order to show a progressive, step-wise transition of one active site structure to the next. An aspartate (Asp) transaminase belonging to functional cluster 4b (EC 2.6.1.1, PDB ID: 1ARG) was overlaid with the GABA transaminase to show the active site structural change between functional clusters (Figures 6b, d). The final comparison between the Asp transaminase and a tyrosine (Tyr) transaminase (EC 2.6.1.5, PDB ID: 3DYD) was undertaken to show the active site differences between functional clusters 4b and 4a as well as the structural changes necessary to accommodate the larger substrate (Figures 6c-d).

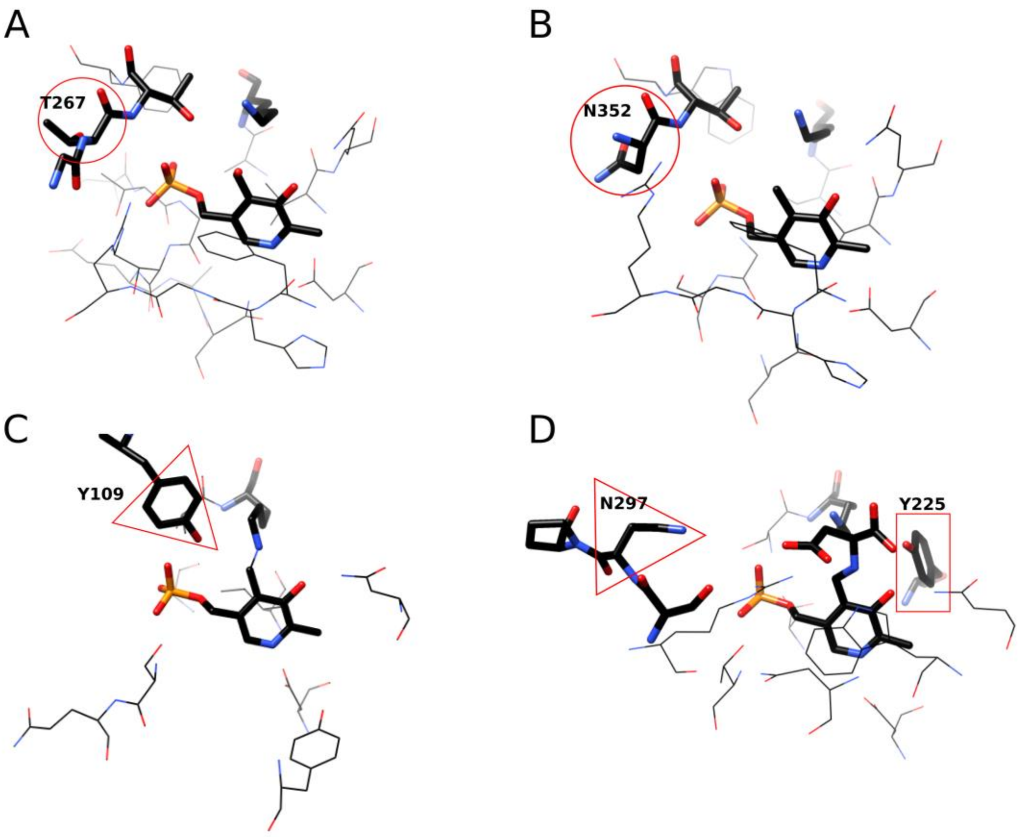


Figure 6: Side by side alignment of ACO, GABA, Asp, and Tyr transaminase active sites. (A) The active site of 2ORD, an ACO transaminase. Circled in red is the threonine residue that stabilizes the substrate (not shown) in the active site. (B) The active site of 1OHW, a GABA transaminase. The red circle indicates the position of the asparagine residue relative to the position of the threonine residue from (A). (C) The active site of 3DYD, a Tyr transaminase. Indicated in the red triangle is the conserved tyrosine residue relative to the position of the asparagine residue in (D). (D) The active site structure of 1ARG, an Asp transaminase. The red box indicates the addition of a tyrosine residue in the active site relative to the GABA transaminase active site in (B).

2.4 Discussion

Sequence and structural homology models are typically used to infer descent from a common ancestor – *molecular evolution*. Many examples of protein evolution have been constructed based solely on sequence information. The advent of computational algorithms based on probability theory has provided a powerful means for inferring sequence homology and evolutionary relationships.[52, 53] However, these methods encounter difficulties when dealing with very divergent sequences. At low sequence identity, the incorporation of gaps for a suitable alignment may produce results that are not genuine.[54] The spatial arrangement of local structures begins to define the function of a protein and represents an important, but distinct, evolutionary signal – *functional evolution*.[23] Several methods have been developed that use structural templates to help guide and ultimately improve the alignment of divergent sequences.[55, 56] Conversely, when both sequence and structure are divergent these methods become less reliable.[51] Critically, while molecular evolution and functional evolution are related, they proceed through distinctly different mechanisms.[1, 2, 20] Correspondingly, different approaches are needed to measure the evolution of homologous proteins across multiple species compared to investigating the evolutionary relationship of various protein functions. Thus, instead of applying global sequence similarities to infer an evolutionary relationship between distantly related proteins, investigating functional evolution requires a targeted approach by analyzing the sequence and structural characteristics of functional epitopes or protein active sites.

The CPASS software and database was initially designed to functionally annotate novel proteins based on ligand defined active site structures and to act as a complement to drug discovery programs.[15, 33] CPASS finds the optimal sequence and structural alignment between two active sites, exclusive of sequence connectivity, and returns a similarity score. It was found that CPASS could effectively group and annotate proteins that bound the same ligand based on specific function.[15] Further analysis of the results led to the hypothesis that CPASS may also infer an evolutionary relationship based on function since active site structure and geometry is primarily functionally dependent. CPASS does, however, have an inherent disadvantage since it requires a 3D structure of the enzyme with a bound ligand.

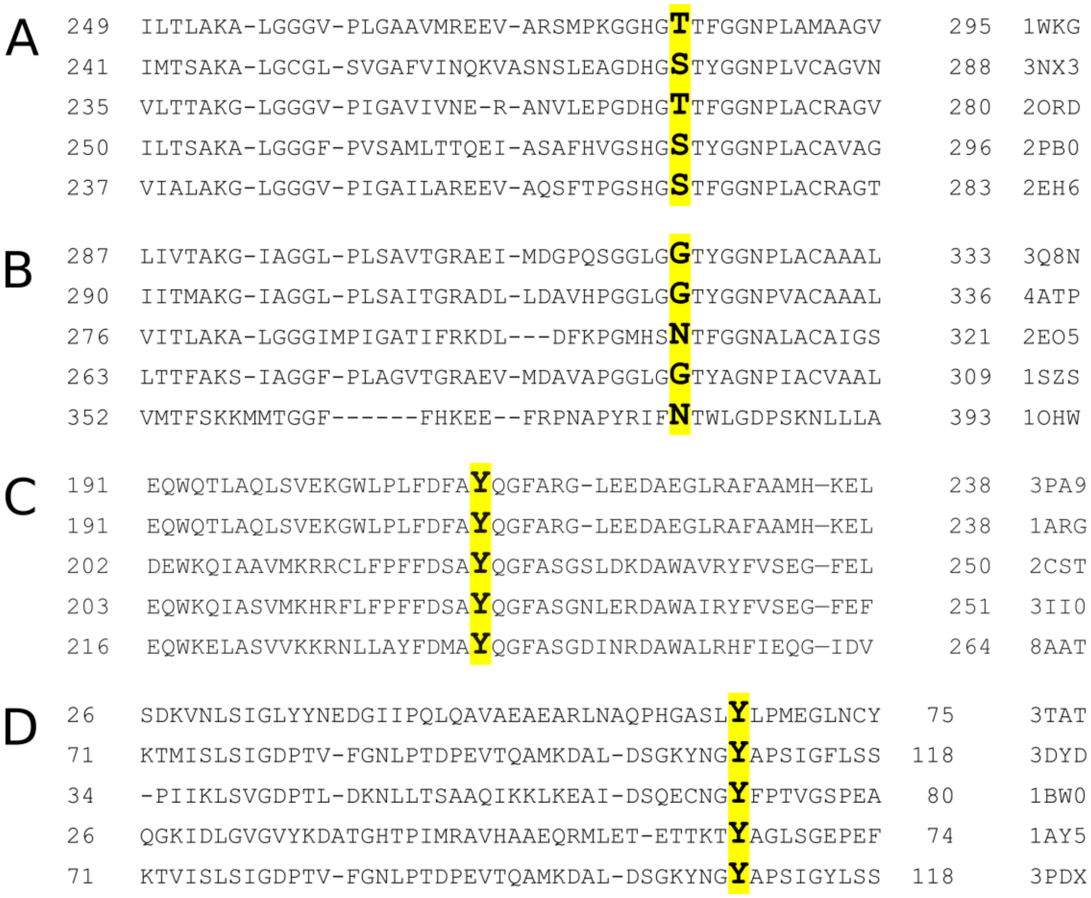
PLP-dependent enzymes encompass a broad-range of catalytic functions and represent a significant percentage of an organism's genome.[57] Correspondingly, PLP-dependent enzymes are a common target for investigating protein evolution.[24, 31, 32] Complicating these analyses is the fact that PLP-dependent enzymes suffer from low sequence identity within and between the four different fold-types.[45, 58] For these very distantly related proteins, any sequence similarity that still remains will only be found in the regions of the proteins that evolve at a significantly slower rate (functionally important) rather than over the entire sequence length. The incorporation of additional sequence information, which is not associated with these functionally conserved regions, is likely to negatively impact a global analysis. Instead, CPASS presents an alternative approach to investigate the functional evolution of PLP-dependent enzymes by restricting the analysis to the conserved, functionally relevant active site sequence and structure. A total of 204 PLP-dependent enzymes had a structure deposited in the RCSB PDB[34] that contained a bound PLP and could be used for a CPASS analysis of functional evolution.

The results from the CPASS analysis of the PLP-dependent enzymes agree with previous findings that the enzymes developed from four different fold-types in the last universal ancestor (Figure 1). CPASS further clustered the enzymes together by specific function within each of the four fold-types. Again, this is consistent with the existing functional annotation for the PLP-dependent enzymes. It is critical to point out that the CPASS analysis did not utilize the fold-type classification or the available functional annotation to generate the phylogenetic network depicted in Figure 1. The network was generated strictly based on the CPASS active site similarity scores. The fold-type classification and functional cluster labels were simply added to the network after the fact.

Correspondingly, these results demonstrate the reliability of using enzyme active sites for comparing distantly related proteins to infer an evolutionary and functional relationship. Conversely, it was not possible to generate a similar function-based network for PLP-dependent enzymes for the four fold types using either global sequence or structure similarity. A network generated from sequence alignment data results in an essentially random, nonsensical map because of the extremely low sequence alignment within and between the fold-types (Figure S1). Similarly, a network generated from structure homology data groups the PLP-dependent enzymes into the four fold-types, which of course is not surprising since the fold types were defined by structure homology (Figure S2). But, any further refinement into functional clusters or any association between the four fold-types is, again, essentially random and irrelevant.

In order to maintain function, the active site of a protein remains highly conserved. Similarly, as the function of an enzyme evolves to accommodate a different substrate or catalyze a different reaction, only subtle changes in the geometry of the active site or moderate sequence substitutions, deletions or insertions are necessary. These active site changes are likely to occur in a slow and step-wise process utilizing the same mechanisms that drive overall molecular evolution. Maintaining an active site change will depend on how the new (if any) enzymatic activity is contributing to cell fitness. In essence, a sequential path from one active site to another is expected where the observed sequence and structural changes directly contribute to the functional differences between the enzymes. This is exactly what was observed in the CPASS derived phylogenetic network. Nearest neighbors identifies subtle differences in active site sequences and structures that occur as the enzymatic activity and substrate specificity diverges. In effect, these changes identify an evolutionary path for the divergence of enzyme function. As an example, ACO transaminase, GABA transaminase, L-tyrosine transaminase, and aspartate transaminase are fold-type I enzymes belonging to clusters 1, 1, 4a, and 4b, respectively, in our CPASS-based network (Figures 1 and 2).

A direct comparison between the ACO and GABA transaminase active site structures identified a single amino acid substitution from a threonine to an asparagine, respectively, at the entrance to the active site (Figures 6a-b). A MSA of several ACO transaminase enzymes further verifies that a serine/threonine is conserved at this position (Figure 7a). The serine/threonine is most likely necessary to stabilize the larger ACO ligand through the formation of additional hydrogen bonds. Previous work has recognized the function of the conserved serine/threonine,[59, 60] but has not identified the residue as being a primary difference between ACO and GABA transaminases. Conversely, a MSA of GABA transaminases confirms the presence of a conserved glycine/asparagine at the entrance to the binding pocket (Figure 7b).



```
A  249  ILTLAKA-LGGGV-PLGAAVMREEV-ARSMPKGGHGTTFGGNPLAMAAGV   295  1WKG
   241  IMTSAKA-LGCGL-SVGAFVINQKVASNSLEAGDHGSTYGGNPLVCAGVN   288  3NX3
   235  VLTTAKG-LGGGV-PIGAVIVNE-R-ANVLEPGDHGTTFGGNPLACRAGV   280  2ORD
   250  ILTSAKA-LGGGF-PVSAMLTTQEI-ASAFHVGSHGSTYGGNPLACAVAG   296  2PB0
   237  VIALAKG-LGGGV-PIGAILAREEV-AQSFTPGSHGSTFGGNPLACRAGT   283  2EH6

B  287  LIVTAKG-IAGGL-PLSAVTGRAEI-MDGPQSGGLGGTYGGNPLACAAAL   333  3Q8N
   290  IITMAKG-IAGGL-PLSAITGRADL-LDAVHPGGLGGTYGGNPVACAAAL   336  4ATP
   276  VITLAKA-LGGGIMPIGATIFRKDL---DFKPGMHSNTFGGNALACAIGS   321  2EO5
   263  LTTFAKS-IAGGF-PLAGVTGRAEV-MDAVAPGGLGGTYAGNPIACVAAL   309  1SZS
   352  VMTFSKKMMTGGF------FHKEE--FRPNAPYRIFNTWLGDPSKNLLLA   393  1OHW

C  191   EQWQTLAQLSVEKGWLPLFDFAYQGFARG-LEEDAEGLRAFAAMH—KEL   238   3PA9
   191   EQWQTLAQLSVEKGWLPLFDFAYQGFARG-LEEDAEGLRAFAAMH—KEL   238   1ARG
   202   DEWKQIAAVMKRRCLFPFFDSAYQGFASGSLDKDAWAVRYFVSEG—FEL   250   2CST
   203   EQWKQIASVMKHRFLFPFFDSAYQGFASGNLERDAWAIRYFVSEG—FEF   251   3II0
   216   EQWKELASVVKKRNLLAYFDMAYQGFASGDINRDAWALRHFIEQG—IDV   264   8AAT

D  26   SDKVNLSIGLYYNEDGIIPQLQAVAEAEARLNAQPHGASLYLPMEGLNCY   75    3TAT
   71   KTMISLSIGDPTV-FGNLPTDPEVTQAMKDAL-DSGKYNGYAPSIGFLSS  118    3DYD
   34   -PIIKLSVGDPTL-DKNLLTSAAQIKKLKEAI-DSQECNGYFPTVGSPEA   80    1BW0
   26   QGKIDLGVGVYKDATGHTPIMRAVHAAEQRMLET-ETTKTYAGLSGEPEF   74    1AY5
   71   KTVISLSIGDPTV-FGNLPTDPEVTQAMKDAL-DSGKYNGYAPSIGYLSS  118    3PDX
```

Figure 7: MSAs of ACO, GABA, Asp, and Tyr transaminases. Segments of individual MSAs from ACO (A), GABA (B), Asp (C), and Tyr (D) transaminases. Each segment highlights a conserved residue corresponding to the active sites in Figure 7. PDB IDs of the selected enzymes are on the far right of the alignment.

A small amino-acid at this location may open-up the binding pocket and/or increase flexibility, which, in turn, may lead to substrate promiscuity.[57] Thus, the glycine/asparagine to serine/threonine substitution may simply reduce substrate promiscuity and increase the selectivity to ACO. Inspection of the differences between the GABA and Asp active sites reveals the addition of a Tyr side chain in the active site of the Asp transaminase (Figures 6b, d). The hydroxyl group on the Tyr side chain is known to stabilize the internal aldimine between the PLP cofactor and the catalytic lysine at physiological pH and to increase the k_{cat} for the reaction in Asp transaminases.[61] Binding pocket hydrophobicity is also increased by the presence of the Tyr side chain, which may explain the moderate affinity of Asp transaminases for aromatic amino acids.[62] Importantly, ACO and GABA transaminases substitute a smaller glutamine residue to hydrogen bond with the cofactor.[59, 63] A MSA of several Asp transaminases from functional clusters 4a and 4b shows that the Tyr residue is highly conserved (Figure 7c). Lastly, the progression from an Asp to Tyr transaminase follows a spatial substitution of Asn to Tyr as opposed to a sequence substitution. The Tyr at the entrance of the binding pocket fulfills the same role as the Asn by forming hydrogen bonds with the PLP phosphate group. However, the effects of the spatial rearrangement are twofold. Mutation of the Asn in Asp transaminases removes a coordinated water molecule from the binding pocket to make room for aromatic substrates,[64] and the presence of the Tyr side chain increases the hydrophobicity of the active site.[62] Therefore, a single spatial substitution can greatly change the affinity for certain substrates in Type-I enzymes. An MSA of five Tyr transaminases confirms the conservation of the Tyr residue in the active site (Figure 7d)

CPASS recognized these subtle, sequential changes that occurred between these four, nearest-neighbor active sites while also acknowledging the overall active site similarity. This enabled the close grouping of the ACO transaminase, GABA transaminase, L-tyrosine decarboxylase, and aspartate aminotransferase within the same functional cluster; and, importantly, the identification of a clear evolutionary path to achieve the observed substrate and functional divergence.

2.5 Conclusion

PLP-dependent enzymes have been a common target of molecular evolution studies, but prior efforts have been limited to small subgroups due to the low sequence identity within each fold-type and across all PLP-dependent enzymes.[24, 31, 32] Instead, we have demonstrated the first phylogenetic network based on functional evolution to model all PLP-dependent fold-types together on a *single* network. We were able to correctly cluster all the PLP enzymes into their previously assigned fold-type simply based on the sequence and structure similarity of the PLP-binding site. Furthermore, the enzymes clustered based on their EC numbers within each fold-type cluster. Our CPASS-derived functional network allowed us to follow the step-wise evolution of substrate specificity and catalytic activity. Effectively, nearest neighbors in the CPASS-derived functional network were related by single-amino-acid changes for highly-conserved active site residues. It was not possible to achieve a similar network using either global sequence or structure similarity. Thus, the study herein successfully determined an evolutionary relationship based on function between many diverse and highly divergent PLP-dependent enzymes using the available active site structures. Also, our results clearly verifies that CPASS is an effective approach for identifying an evolutionary relationship based on function using similarities in active site structure and sequence.

2.6 References

[1] Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001) Evolution of Function in Protein Superfamilies, from a Structural Perspective, *Journal of Molecular Biology 307*, 1113-1143.

[2] Zuckerkandl, E., and Pauling, L. (1965) Evolutionary divergence and convergence in proteins, *Evolving Genes Proteins, Symp. Rutgers, State Univ. 1964*, 97-166.

[3] Morozova, O., and Marra, M. A. (2008) Applications of next-generation sequencing technologies in functional genomics, *Genomics 92*, 255-264.

[4] Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W., and Swaminathan, S. (1999) Structural genomics: beyond the human genome project, *Nat. Genet. 23*, 151-157.

[5] Pál, C., Papp, B., and Lercher, M. J. (2006) An integrated view of protein evolution, *Nature Reviews Genetics 7*, 337-348.

[6] Chothia, C., and Gough, J. (2009) Genomic and structural aspects of protein evolution, *Biochemical Journal 419*, 15.

[7] Martincorena, I., Seshasayee, A. S. N., and Luscombe, N. M. (2012) Evidence of non-random mutation rates suggests an evolutionary risk management strategy, *Nature (London, U. K.) 485*, 95-98.

[8] Eliot, A. C., and Kirsch, J. F. (2004) Pyridoxal Phosphate Enzymes: Mechanistic, Structural, and Evolutionary Considerations, *Annual Review of Biochemistry 73*, 383-415.

[9] Hasson, M. S., Schlichting, I., Moulai, J., Taylor, K., Barrett, W., Kenyon, G. L., Babbitt, P. C., Gerlt, J. A., Petsko, G. A., and Ringe, D. (1998) Evolution of an

enzyme active site: The structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase, *Proceedings of the National Academy of Sciences 95*, 10396-10401.

[10] La, D., Sutch, B., and Livesay, D. R. (2004) Predicting protein functional sites with phylogenetic motifs, *Proteins: Structure, Function, and Bioinformatics 58*, 309-320.

[11] Georgi, B., Schultz, J., and Schliep, A. (2009) Partially-supervised protein subclass discovery with simultaneous annotation of functional residues, *BMC Structural Biology 9*, 68.

[12] Gutteridge, A., Bartlett, G. J., and Thornton, J. M. (2003) Using A Neural Network and Spatial Clustering to Predict the Location of Active Sites in Enzymes, *Journal of Molecular Biology 330*, 719-734.

[13] Russell, R. B., Sasieni, P. D., and Sternberg, M. J. E. (1998) Supersites Withing Superfolds. Binding Site Similarity in the Absence of Homology, *Journal of Molecular Biology 282*, 903-918.

[14] Schmitt, S., Kuhn, D., and Klebe, G. (2002) A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology, *Journal of Molecular Biology 323*, 387-406.

[15] Powers, R., Copeland, J. C., Germer, K., Mercier, K. A., Ramanathan, V., and Revesz, P. (2006) Comparison of protein active site structures for functional annotation of proteins and drug design, *Proteins: Structure, Function, and Bioinformatics 65*, 124-135.

[16] Fetrow, J. S., Godzik, A., and Skolnick, J. (1998) Functional Analysis of the Escherichia coli Genome Using the Sequence-to-Structure-to-Function Paradigm: Identification of Proteins Exhibiting the Glutaredoxin/Thioredoxin Disulfide Oxidoreductase Activity, *Journal of Molecular Biology 282*, 703-711.

[17] Kolker, E., Picone, A. F., Galperin, M. Y., Romine, M. F., Higdon, R., Makarova, K. S., Kolker, N., Anderson, G. A., Qiu, X., Auberry, K. J., Babnigg, G., Beliaev, A. S., Edlefsen, P., Elias, D. A., Gorby, Y. A., Holzman, T., Klappenbach, J. A., Konstantinidis, K. T., Land, M. L., Lipton, M. S., McCue, L. A., Monroe, M., Pasa-Tolic, L., Pinchuk, G., Purvine, S., Serres, M. H., Tsapin, S., Zakrajsek, B. A., Zhu, W., Zhou, J., Larimer, F. W., Lawrence, C. E., Riley, M., Collart, F. R., Yates, J. R., Smith, R. D., Giometti, C. S., Nealson, K. H., Fredrickson, J. K., and Tiedje, J. M. (2005) Global profiling of Shewanella oneidensis MR-1: Expression of hypothetical genes and improved functional annotations, *Proceedings of the National Academy of Sciences 102*, 2099-2104.

[18] Hemrika, W., Renirie, R., Dekker, H. L., Barnett, P., and Wever, R. (1997) From phosphatases to vanadium peroxidases: A similar architecture of the active site, *Proceedings of the National Academy of Sciences 94*, 2145-2149.

[19] Kull, F. J., Vale, R. D., and Fletterick, R. J. (1998) The case for a common ancestor: kinesin and myosin motor proteins and G proteins, *Journal of Muscle Research and Cell Motility 19*, 877-886.

[20] Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes, *Genome Res. 20*, 1313-1326.

[21] Hughes, A. L. (1994) The evolution of functionally novel proteins after gene duplication, *Proc. R. Soc. London, Ser. B 256*, 119-124.

[22] Prince, V. E., and Pickett, F. B. (2002) Splitting pairs: The diverging fates of duplicated genes, *Nat. Rev. Genet. 3*, 827-837.

[23] Jones, S., and Thornton, J. M. (2004) Searching for functional sites in protein structures, *Curr. Opin. Chem. Biol. 8*, 3-7.

[24] Christen, P., and Mehta, P. K. (2001) From Cofactor to Enzymes. The Molecular Evolution of Pyridoxal-5'-Phosphate-Dependent Enyzmes, *The Chemical Record 1*, 436-447.

[25] Bairoch, A. (2000) The ENZYME database in 2000, *Nucleic Acids Research 28*, 304-305.

[26] Jansonius, J. N. (1998) Structure, evolution and action of vitamin B_6-dependent enzymes, *Current Opinion in Structural Biology 8*, 759-769.

[27] Dunathan, H. C. (1966) Conformation and Reaction Specificity in Pyridoxal Phosphate Enzymes, *Proceedings of the National Academy of Sciences 55*, 712-716.

[28] Hayashi, H. (1995) Pyridoxal Enzymes: Mechanistic Diversity and Uniformity, *Journal of Biochemistry 118*, 463-473.

[29] Christen, P., Kasper, P., Gehring, H., and Sterk, M. (1996) Stereochemical constraint in the evolution of pyridoxal-5'-phosphate-dependent enzymes. A hypothesis, *FEBS Letters 389*, 12-14.

[30] Schneider, G., Kack, H., and Lindqvist, Y. (2000) The manifold of vitamin B_6 dependent enzymes, *Structure 8*, 1-6.

[31] Alexander, F. W., Sandmeier, E., Mehta, P. K., and Christen, P. (1994) Evolutionary relationships among pyridoxal-5'-phosphate-dependent enzymes. Regio-specific α,β and γ families, *Eur. J. Biochem. 219*, 953-960.

[32] Mehta, P. K., and Christen, P. (2000) The molecular evolution of pyridoxal-5'-phosphate-dependent enzymes, *Adv. Enzymol. Relat. Areas Mol. Biol. 74*, 129-184.

[33] Powers, R., Copeland, J. C., Stark, J. L., Caprez, A., Guru, A., and Swanson, D. (2011) Searching the protein structure database for ligand-binding site similarities using CPASS v.2, *BMC Research Notes 4*, 17.

[34] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, *Nucleic Acids Research 28*, 235-242.

[35] Henikoff, S., and Henikoff, J. G. (1992) Amino acid substitution matrixes from protein blocks, *Proceedings of the National Academy of Sciences of the United States of America 89*, 10915-10919.

[36] Henikoff, S., and Henikoff, J. G. (1993) Performance evaluation of amino acid substitution matrixes, *Proteins: Structure, Function, and Genetics 17*, 49-61.

[37] Consortium, T. U. (2007) The Universal Protein Resource (UniProt), *Nucleic Acids Research 35*, D193-D197.

[38] Huson, D. H. (2005) Application of Phylogenetic Networks in Evolutionary Studies, *Molecular Biology and Evolution 23*, 254-267.

[39] Bryant, D. (2003) Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks, *Molecular Biology and Evolution 21*, 255-265.

[40] Huson, D. H., and Scornavacca, C. (2012) Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks, *Systematic Biology 61*, 1061-1067.

[41] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, *Molecular Systems Biology 7*.

[42] Zhang, Y. (2005) TM-align: a protein structure alignment algorithm based on the TM-score, *Nucleic Acids Research 33*, 2302-2309.

[43] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera-A Visualization System for Exploratory Research and Analysis, *Journal of Computational Chemistry 25*, 1605-1612.

[44] Jensen, R. A., and Gu, W. (1996) Evolutionary Recruitment of Biochemically Specialized Subdivisions of Family I within the Protein Superfamily of Aminotransferases, *Journal of Bacteriology 178*, 2161-2171.

[45] Grishin, N. V., Phillips, M. A., and Goldsmith, E. J. (1995) Modeling of the spatial structure of eukaryotic ornithine decarboxylases, *Protein Science 4*, 1291-1304.

[46] Rausch, C., Lerchner, A., Schiefner, A., and Skerra, A. (2013) Crystal structure of the ω-aminotransferase from Paracoccus denitrificans and its phylogenetic relationship with other class III aminotransferases that have biotechnological potential, *Proteins: Structure, Function, and Bioinformatics 81*, 774-787.

[47] Liu, L., Iwata, K., Yohda, M., and Miki, K. (2002) Structural insight into gene duplication, gene fusion and domain swapping in the evolution of PLP-independent amino acid racemases, *FEBS Letters 528*, 114-118.

[48] Kidron, H., Repo, S., Johnson, M. S., and Salminen, T. A. (2006) Functional Classification of Amino Acid Decarboxylases from the Alanine Racemase Structural Family by Phylogenetic Studies, *Molecular Biology and Evolution 24*, 79-89.

[49] Shah, R., Coleman, C. S., Mir, K., Baldwin, J., Van Etten, J. L., Grishin, N. V., Pegg, A. E., Stanley, B. A., and Phillips, M. A. (2004) Paramecium bursaria chlorella virus-1 encodes an unusual arginine decarboxylase that is a close homolog of eukaryotic ornithine decarboxylases, *J Biol Chem 279*, 35760-35767.

[50] Sugio, S., Petsko, G. A., Manning, J. M., Soda, K., and Ringe, D. (1995) Crystal Structure of a D-Amino Acid Aminotransferase: How the Protein Controls Stereoselectivity, *Biochemistry 34*, 9661-9669.

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

[52] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research 25*, 3389-3402.

[53] Karplus, K., Barrett, W., and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies, *Bioinformatics 14*, 846-856.

[54] Doolittle, R. F. (1986) *Of URFs and ORFs: A primer on how to analyze derived amino acid sequences.*, University Science Books.

[55] Pei, J., Kim, B. H., and Grishin, N. V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments, *Nucleic Acids Research 36*, 2295-2300.

[56] Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V., and Notredame, C. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee, *Nucleic Acids Research 34*, W604-W608.

[57] Percudani, R., and Peracchi, A. (2003) A genomic overview of pyridoxal-phosphate-dependent enzymes, *EMBO reports 4*, 850-854.

[58] Percudani, R., and Peracchi, A. (2009) The B6 database: a tool for the description and classification of vitamin B6-dependent enzymatic activities and of the corresponding protein families, *BMC Bioinformatics 10*, 273.

[59] Rajaram, V., Ratna Prasuna, P., Savithri, H. S., and Murthy, M. R. N. (2007) Structure of biosynthetic N-acetylornithine aminotransferase from Salmonella typhimurium: Studies on substrate specificity and inhibitor binding, *Proteins: Structure, Function, and Bioinformatics 70*, 429-441.

[60] Kack, H., Sandmark, J., Gibson, K., Schneider, G., and Lindqvist, Y. (1999) Crystal Structure of Diaminopelargonic Acid Synthase: Evolutionary Relationships between Pyridoxal-5'-phosphatedependent Enzymes, *Journal of Molecular Biology 291*, 857-876.

[61] Goldberg, J. M., Swanson, R. V., Goodman, H. S., and Kirsch, J. F. (1991) The Tyrosine-225 to Phenylalanine Mutation of Escherichia coli Aspartate Aminotransferase Results in an Alkaline Transition in the Spectrophotometric and

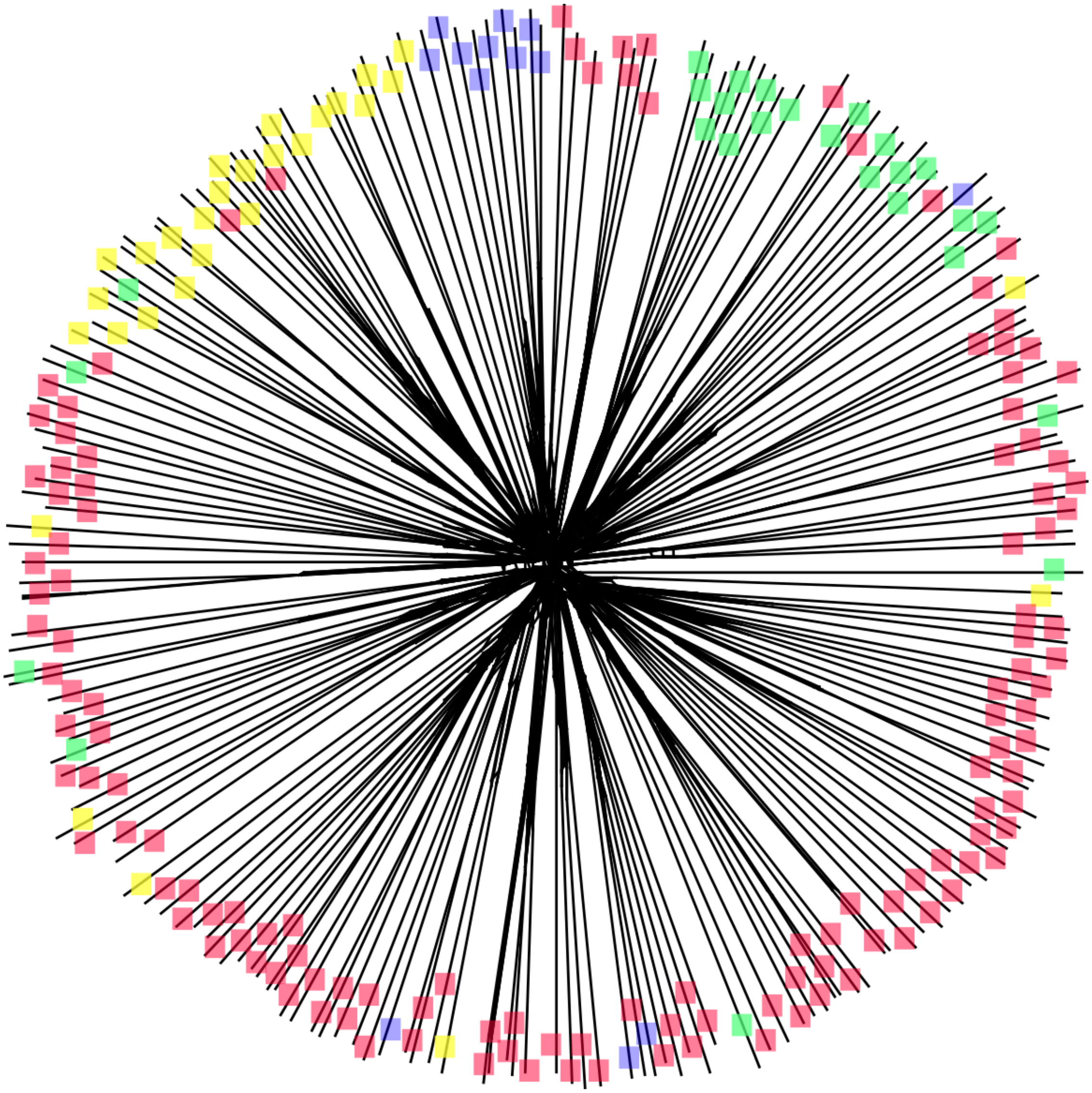
Kinetic pK_a Values and Reduced Values of both k_{cat} and K_m, *Biochemistry 30*, 305-312.

[62] Kamitori, S., Okamoto, A., Hirotsu, K., Higuchi, T., Kuramitsu, S., Kagamiyama, H., Matsuura, Y., and Katsube, Y. (1990) Three-Dimensional Structures of Aspartate Aminotransferase from *Escherichia coli* and Its Mutant Enzyme at 2.5A Resolution, *Journal of Biochemistry 108*, 175-184.

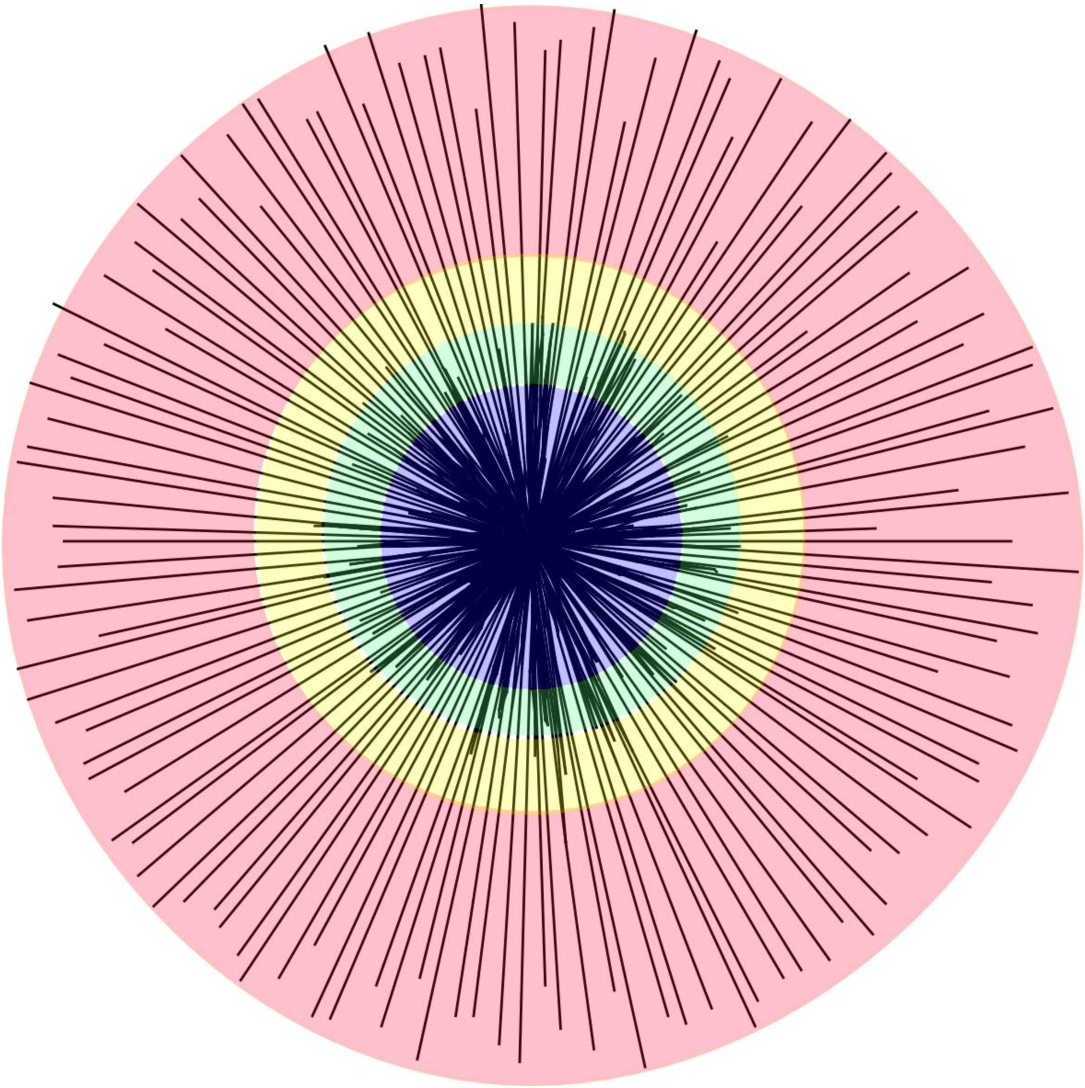
[63] Liu, W., Peterson, P. E., Carter, R. J., Zhou, X., Langston, J. A., Fisher, A. J., and Toney, M. D. (2004) Crystal Structures of Unbound and Aminooxyacetate-Bound *Escherichia coli* y-Aminobutyrate Aminotransferase, *Biochemistry 43*, 10896-10905.

[64] Malashkevich, V. N., Onuffer, J. J., Kirsch, J. F., and Jansonius, J. N. (1995) Alternating argninie-modulated substrate specificity in an engineered tyrosine aminotransferase, *Nature Structural Biology 2*, 548-553.

2.7 Supplemental Figures and Tables



Supplemental Figure 1: Sequence based network analysis of PLP-dependent enzymes. A phylogenetic network of 204 PLP-dependent enzymes with ligand defined active sites in the PDB. The red, yellow, green, and blue markers indicate enzymes belonging to fold-types I, II, III, and IV, respectively. The network was built using a MSA from Clustal Omega and the Neighbor-net algorithm.



Supplemental Figure 2: Structure based network analysis of PLP-dependent enzymes. A phylogenetic network of 139 PLP-dependent enzymes with ligand defined active sites in the PDB. The network was built using pairwise TMalign scores and the Neighbor-net algorithm. Red, yellow, green, and blue regions correspond to fold-types I, II, III, and IV, respectively.

Fold-type	**Cluster**	**PDB**	**EC**	**Organism**	**Domain**
I	1	2JJG	2.6.1.36	*M. tuberculosis*	Bacteria
		1OHW	2.6.1.19	*S. scrofa*	Eukaryota
		1SZS	2.6.1.19	*E. coli*	Bacteria
		4ATP	2.6.1.19	*A. aurescens*	Bacteria
		2PB0	2.6.1.11	*S. typhimurium*	Bacteria
		2ORD	2.6.1.11	*T. maritima*	Bacteria
		4ADB	2.6.1.81	*E. coli*	Bacteria
		2EH6	2.6.1.11	*A. aeolicus*	Bacteria
		1VEF	2.6.1.11	*T. thermophilus*	Bacteria
		4NOG	2.6.1.13	*T. gondii*	Eukaryota
		1OAT	2.6.1.13	*H. sapiens*	Eukaryota
		2YKY	2.6.1	*Mesorhizobium sp.*	Bacteria
		4AO9	2.6.1	*V. paradoxus*	Bacteria
		4GSA	5.4.3.8	*Synechococcus sp.*	Bacteria
		3K28	5.4.3.8	*B. anthracis*	Bacteria
		3DXV	5.1	*A. obae*	Bacteria
		4A6T	2.6.1.62	*C. violaceum*	Bacteria
		3I5T	2.6.1	*R. sphaeroides*	Bacteria
		3TFT	2.6.1.62	*M. tuberculosis*	Bacteria
		1MLY	2.6.1.62	*E. coli*	Bacteria
		4A0R	2.6.1.62	*A. thaliana*	Eukaryota
		3DU4	2.6.1.62	*B. subtilis*	Bacteria
		3A8U	2.6.1.18	*P. putida*	Bacteria
		2EO5	2.6.1.19	*S. tokodaii*	Archaea
		1Z3Z	4.1.1.64	*B. cepacia*	Bacteria
I	2	1BJO	2.6.1.52	*E. coli*	Bacteria
		3E77	2.6.1.52	*H. sapiens*	Eukaryota
		3QBO	2.6.1.52	*Y. pestis*	Bacteria
		2BI3	2.6.1.52	*B. alcalophilus*	Bacteria
		1W3U	2.6.1.52	*B. alcalophilus*	Bacteria
		3VOM	2.6.1.52	*M. tuberculosis*	Bacteria
		3IMZ	2.6.1.44	*H. sapiens*	Eukaryota
		2YOB	2.6.1.44	*H. sapiens*	Eukaryota
		2YRR	2.6.1	*T. thermophilus*	Bacteria
		2DR1	2.6.1.51	*P. horikoshii*	Bacteria
		3ZRP	2.6.1.51	*S. solfataricus*	Archaea
		1VJO	2.6.1.44	*Nostoc sp.*	Bacteria
		3ISL	2.6.1.44	*B. subtilis*	Bacteria
		1M32	2.6.1.37	*S. typhimurium*	Bacteria
I	3	1C7G	4.1.99.2	*E. herbicola*	Bacteria
		1N8P	4.4.1.1	*S. cerevisiae*	Eukaryota
		3COG	4.4.1.1	*H. sapiens*	Eukaryota
		1IBJ	4.4.1.8	*A. thaliana*	Eukaryota

		1I43	2.5.1.48	*N. tabacum*	Eukaryota
		1PG8	4.4.1.11	*P. putida*	Bacteria
		1E5F	4.4.1.11	*T. vaginalis*	Eukaryota
		3NMY	4.4.1.1	*X. oryzae*	Bacteria
		3NNP	4.4.1.1	*X. oryzae*	Bacteria
		4L0O	2.5.1.48	*H. pylori*	Bacteria
		2CTZ	2.5.1	*T. thermophilus*	Bacteria
I	4a	2X5F	2.6.1	*S. aureus*	Bacteria
		1LK9	4.4.1.4	*A. sativum*	Eukaryota
		1XI9	2.6.1	*P. furiosus*	Archaea
		1V2E	2.6.1.15	*T. thermophilus*	Bacteria
		1U08	2.6.1.88	*E. coli*	Bacteria
		1O4S	2.6.1.1	*T. maritima*	Bacteria
		1DJU	2.6.1.57	*P. horikoshii*	Archaea
		3EI8	2.6.1.83	*A. thaliana*	Eukaryota
		4FL0	2.6.1	*A. thaliana*	Eukaryota
		3IF2	2.6.1	*P. arcticus*	Bacteria
		2O1B	2.6.1.1	*S. aureus*	Bacteria
		2X5D	2.6.1	*P. aeruginosa*	Bacteria
		2ZY2	2.6.1.1	*Pseudomonas sp.*	Bacteria
		2ZY4	2.6.1.1	*A. faecalis*	Bacteria
		3DYD	2.6.1.5	*H. sapiens*	Eukaryota
		1B5P	2.6.1.1	*T. thermophilus*	Bacteria
		1J32	2.6.1.1	*P. lapideum*	Bacteria
		3ELE	2.6.1	*E. rectale*	Bacteria
		1WST	2.6.1	*T. profundus*	Archaea
		1VP4	2.6.1	*T. maritima*	Bacteria
		2ZP7	2.6.1.39	*T. thermophilus*	Bacteria
		4JE5	2.6.1.39	*S. cerevisiae*	Eukaryota
		3PPL	2.6.1.1	*C. glutamicum*	Bacteria
		3IHJ	2.6.1.2	*H. sapiens*	Eukaryota
		1IAX	4.4.1.14	*S. lycopersicum*	Eukaryota
		1LKC	4.1.1.81	*S. enterica*	Bacteria
		1GEW	2.6.1.9	*E. coli*	Bacteria
		1H1C	2.6.1.10	*T. maritima*	Bacteria
		3LY1	2.6.1.9	*P. atrosepticum*	Bacteria
		1B8G	4.4.1.14	*M. domestica*	Eukaryota
		1D2F	4.4.1.8	*E. coli*	Bacteria
		3B1C	4.4.1.8	*S. anginosus*	Bacteria
		3L8A	4.4.1.8	*S. mutans*	Bacteria
		1C7N	4.4.1.8	*T. denticola*	Bacteria
		4DQ6	2.6.1	*C. difficile*	Bacteria
		3KAX	4.4.1	*B. anthracis*	Bacteria
I	4b	1ASF	2.6.1.1	*E. coli*	Bacteria

		1AY5	2.6.1.57	*P. denitrificans*	Bacteria
		3TAT	2.6.1.57	*E. coli*	Bacteria
		2CST	2.6.1.1	*G. gallus*	Eukaryota
		3II0	2.6.1.1	*H. sapiens*	Eukaryota
		3MEB	2.6.1.1	*G. lamblia*	Eukaryota
		1YAA	2.6.1.1	*S. cerevisiae*	Eukaryota
		8AAT	2.6.1.1	*G. gallus*	Eukaryota
		4EMY	2.6.1	*A. prevotii*	Bacteria
I	5	2Z67	2.9.1.2	*M. maripauldis*	Archaea
		2E7J	2.5.1.73	*A. fulgidus*	Archaea
		1WYU	1.4.4.2	*T. thermophilus*	Bacteria
		3RBF	4.1.1.28	*H. sapiens*	Eukaryota
		1JS3	4.1.1.28	*S. scrofa*	Eukaryota
		4E1O	4.1.1.22	*H. sapiens*	Eukaryota
		1C4K	4.1.1.17	*Lactobacillus sp.*	Bacteria
		2JIS	4.1.1.29	*H. sapiens*	Eukaryota
		2DGK	4.1.1.15	*E. coli*	Bacteria
		1KMJ	4.4.1.16	*E. coli*	Bacteria
		4LW4	4.4.1.16	*E. coli*	Bacteria
		1T3I	2.8.1.7	*Synechocystis sp.*	Bacteria
		1EG5	2.8.1.7	*T. maritima*	Bacteria
		3LVM	2.8.1.7	*E. coli*	Bacteria
		3VAX	2.8.1.7	*S. lividans*	Bacteria
		3A9X	4.4.1.16	*R. norvegicus*	Eukaryota
		3F9T	4.1.1.25	*M. jannaschii*	Archaea
		4EB5	2.8.1.7	*A. fulgidus*	Archaea
I	6	2WK8	2.3	*V. cholerae El Tor*	Bacteria
		2BWO	2.3.1.37	*R. capsulatus*	Bacteria
		3TQX	2.3.1.29	*C. burnetii*	Bacteria
		1DJE	2.3.1.47	*E. coli*	Bacteria
		3A2B	2.3.1.50	*S. multivorum*	Bacteria
		3G8M	2.1.2.1	*E. coli*	Bacteria
		1BJ4	2.1.2.1	*H. sapiens*	Eukaryota
		1RVU	2.1.2.1	*O. cuniculus*	Eukaryota
		2DKJ	2.1.2.1	*T. thermophilus*	Bacteria
		4OTL	2.1.2.1	*B. cenocepacia*	Bacteria
		2VMO	2.1.2.1	*G. stearothermophilus*	Bacteria
		4BHE	2.1.2.1	*M. jannaschii*	Archaea
I	7	3NYS	2.6.1.98	*P. aeruginosa*	Bacteria
		1B9H	2.6.1	*A. mediterranei*	Bacteria
		2C7T	2.6.1	*B. circulans*	Bacteria
		3BB8	2.6.1	*Y. pseudotuberculosis*	Bacteria

		1MDX	2.6.1.87	*S. typhimurium*	Bacteria
		2FNI	2.6.1.92	*H. pylori*	Bacteria
		1O61	unknown	*C. jejuni*	Bacteria
		2HZP	3.7.1.3	*H. sapiens*	Eukaryota
		1QZ9	3.7.1.3	*P. fluorescens*	Bacteria
II	8	1J0E	3.5.99.7	*C. saturnus*	Eukaryota
		1J0A	3.5.99.7	*P. horikoshii*	Archaea
II	9	3DWI	2.5.1.65	*M. tuberculosis*	Bacteria
		1WKV	2.5.1.47	*A. pernix*	Archaea
		2JC3	2.5.1.47	*S. typhimurium*	Bacteria
		3PC2	4.2.1.22	*D. melanogaster*	Eukaryota
		4L28	4.2.1.22	*H. sapiens*	Eukaryota
		3BM5	2.5.1.47	*E. histolytica*	Eukaryota
		3VBE	2.5.1.47	*G. max*	Eukaryota
		2ECO	2.5.1.47	*T. thermophilus*	Bacteria
		1OAS	2.5.1.47	*S. typhimurium*	Bacteria
		4AEC	2.5.1.47	*A. thaliana*	Eukaryota
		2ISQ	2.5.1.47	*A. thaliana*	Eukaryota
		3ZEI	2.5.1.47	*M. tuberculosis*	Bacteria
II	10	1KFC	4.2.1.20	*S. typhimurium*	Bacteria
		1WDW	4.2.1.20	*P. furiosus*	Archaea
		3SS9	4.3.1.18	*E. coli*	Bacteria
		1WTC	4.3.1.17	*S. pombe*	Eukaryota
		3HMK	4.3.1.17	*R. norvegicus*	Eukaryota
		2D1F	4.2.3.1	*M. tuberculosis*	Bacteria
		1UIM	4.2.3.2	*T. thermophilus*	Bacteria
		2ZSJ	4.2.3.1	*A. aeolicus*	Bacteria
		1KL7	4.2.3.1	*S. cerevisiae*	Eukaryota
		4F4F	4.2.3.1	*B. melitensis*	Bacteria
		1VE5	4.3.1.19	*T. thermophilus*	Bacteria
		1TDJ	4.3.1.19	*E. coli*	Bacteria
		2RKB	4.3.1.17	*H. sapiens*	Eukaryota
		1P5J	4.3.1.17	*H. sapiens*	Eukaryota
III	11	3N29	4.1.1.96	*C. jejuni*	Bacteria
		2J66	4.1.1.95	*B. circulans*	Bacteria
		1KNW	4.1.1.20	*E. coli*	Bacteria
		2YXX	4.1.1.20	*T. maritima*	Bacteria
		3N2O	4.1.1.19	*V. vulnificus*	Bacteria
		3NZP	4.1.1.19	*C. jejuni*	Bacteria
		1TWI	4.1.1.20	*M. jannaschii*	Archaea
		1HKV	4.1.1.20	*M. tuberculosis*	Bacteria
		2NV9	4.1.1.19	*P. bursaria Chlorella*	Virus

				virus	
		2OO0	4.1.1.17	*H. sapiens*	Eukaryota
		1F3T	4.1.1.17	*T. brucei*	Eukaryota
		2QGH	4.1.1.20	*H. pylori*	Bacteria
III	12	1CT5	5.1.1.1	*S. cerevisiae*	Bacteria
		1W8G	5.1.1.1	*E. coli*	Bacteria
		3R79	unknown	*A. fabrum*	Bacteria
		3CO8	5.1.1.1	*O. oeni*	Bacteria
		4BF5	5.1.1.1	*A. hydrophilia*	Bacteria
		4BEQ	5.1.1.1	*V. cholerae*	Bacteria
		1RCQ	5.1.1.1	*P. aeruginosa*	Bacteria
		3B8W	5.1.1.1	*E. coli*	Bacteria
		4A3Q	5.1.1.1	*S. aureus*	Bacteria
		3E5P	5.1.1.1	*E. faecalis*	Bacteria
		1SFT	5.1.1.1	*G. stearothermophilus*	Bacteria
		1XFC	5.1.1.1	*M. tuberculosis*	Bacteria
		2DY3	5.1.1.1	*C. glutamicum*	Bacteria
		1VFH	5.1.1.1	*S. lavendulae*	Bacteria
IV	13	1I2K	4.1.3.38	*E. coli*	Bacteria
		2XPF	4.1.3.38	*P. aeruginosa*	Bacteria
		2Y4R	4.1.3.38	*P. aeruginosa*	Bacteria
		4K6N	4.1.3.38	*S. cerevisiae*	Bacteria
		2ZGI	4.1.3.38	*T. thermophilus*	Bacteria
		3CSW	2.6.1.42	*T. maritima*	Bacteria
		4DAA	2.6.1.21	*Bacillus sp.*	Bacteria
		4CMD	2.6.1.42	*N. haematococca*	Eukaryota
		4CHI	2.6.1.42	*A. fumigatus*	Eukaryota
		2EJ3	2.6.1.42	*T. thermophilus*	Bacteria
		1I1M	2.6.1.42	*E. coli*	Bacteria
		3UZO	2.6.1.42	*D. radiodurans*	Bacteria
		3DTH	2.6.1.42	*M. smegmatis*	Bacteria
		3DTG	2.6.1.42	*M. smegmatis*	Bacteria
		2HGX	2.6.1.42	*H. sapiens*	Eukaryota
		2COG	2.6.1.42	*H. sapiens*	Eukaryota

Supplemental Table 1: PLP-dependent enzymes selected for active site evolution analysis. A table of the 204 PLP-dependent enzymes used in this study annotated by PDB ID. Enzymes are grouped by fold-type and then by functional cluster in the network analysis. The enzyme classification number, organism, and domain are also listed for each enzyme.

Chapter 3

Functional Evolution of Proteins

3.1 Introduction

A functional clustering of ligand-defined active sites in the RCSB Protein Data Bank (RCSB PDB)[1] was undertaken to infer an evolutionary lineage of enzymatic function. Sequence based phylogenetic methods are typically utilized to produce evolutionary tree structures.[2] This is possible because molecular evolution occurs through random mutations at a constant rate.[3] Nevertheless, the reliability of sequence-based evolutionary measurements becomes suspect when protein sequence homology enters the "twilight-zone" and falls below 25% sequence identity.[4] Simply, the accuracy of a phylogenetic tree is directly dependent on the accuracy of the sequence alignment, which becomes undependable at low sequence identity.[5, 6] Therefore, due to the large sequential dissimilarity for the entirety of proteins deposited in the RCSB PDB, sequence-based evolutionary methods are not easily or reliably employed across an all-inclusive set of protein functional classes.

Structure-based alignment is an alternative to sequence based alignments, especially considering the tremendous reduction in structure space relative to sequence space. Recent estimates suggest that only a few-thousand distinct protein folds exists,[7, 8] which is consistent with the 1391 protein topologies currently identified by CATH.[9] Nevertheless, the alignment of protein structures is even more challenging than sequence alignment, and fails for completely dissimilar structures.[8] Like sequence, the arrangement of tertiary structures is extremely evolutionarily labile when considering the entirety of known protein functions. While global protein sequence and structure may drift without detrimental consequences, dramatic changes to an active site or functional epitope of a protein may negatively impact the survivability of an organism. Instead, functional evolution progresses slowly through gene duplication and functional drift to avoid negative influence on cellular fitness.[10, 11] This occurs because even minor changes in the spatial orientation or amino acid composition within an active site may lead to dramatic changes in substrate and reaction specificity. Consequently, protein active sites mutate at a much slower rate relative to other structural elements and remain highly conserved over time.[12] In effect, a similarity in protein active sites may remain even though the overall sequence or structure of a protein has completely diverged. Thus, it may be possible to infer an evolutionary relationship based on similarities in protein active sites in situations when global sequence or structure similarities no longer exist.

Several methods and databases have been previously published describing the clustering of proteins from the RCSB PDB. These include sequence,[13] structure,[14] ligand conformation,[15] atomic properties,[16] and putative cavity[17] based approaches. However, a clustering and subsequent phylogenetic analysis based on ligand-defined active sites has not been done. The Comparison of Protein Active site Structures (CPASS) software and database compares the geometry and amino acid similarity between pairs of experimentally determined ligand-defined active sites. CPASS is distinctly different from protein cavity approaches because it focuses on known binding sites rather than putative pocket detection. Further, substrate conformation is only used in the determination of active site residues and not in the CPASS scoring function. Consequently, the evolutionary analysis of proteins in the RCSB PDB based on active site similarity is a novel approach.

We previously demonstrated the utility of CPASS to decipher the functional evolution of proteins by comparing the active sites of 204 PLP-dependent enzymes.[18] We produced the first-ever phylogenetic tree that contained all four families or fold-types (I to IV) for PLP-dependent enzymes. The resulting phylogenetic tree correctly distinguished between the four individual folds and further sorted the enzymes by substrate specificity and function. Critically, no functional information was utilized to produce the phylogenetic tree of PLP-dependent enzymes, yet the enzymes were clustered perfectly based on EC number (*i.e.*, branches were comprised of enzymes with the same EC number). Furthermore, examining individual branches of the phylogenetic tree illustrates the step-wise evolution of function through a series of single amino-acid substitutions. In effect, nearest neighbors in the CPASS derived phylogenetic tree identified subtle differences in active site sequences and structures that led to changes in enzymatic activity and substrate specificity. Importantly, we were able to produce a phylogenetic tree for the PLP-dependent enzymes despite sequence identity well-below 20% and poor structural alignments between folds (TM-align[19] score of ~ 0.3).

Based on this prior success, we expanded upon the phylogenetic tree of PLP-dependent enzymes by using CPASS to functionally cluster all ligand-containing proteins present in the RCSB PDB. In essence, CPASS was used to produce a phylogenetic tree containing essentially all protein functional classes present in the RCSB PDB. CPASS was used to make a pair-wise comparison between all of the ligand-defined binding sites within the

RCSB PDB to produce an all-versus-all CPASS similarity score matrix. The proteins were then clustered by the identity of the bound ligand. Principal component analysis of the CPASS scores was employed to identify a representative structure for each functional class (*i.e.*, same ligand and EC number) in order to reduce the overall size of the dataset. The representative structure for each functional class was then successfully modeled into a single phylogenetic tree based on the CPASS similarity score matrix. The resulting phylogenetic tree demonstrates the functional evolution across all of the protein functional classes within the RCSB PDB. To further illustrate the effectiveness of our approach, we also highlight two specific regions of the phylogenetic tree that demonstrate the stepwise substrate and enzymatic evolution of fructose-6-phosphate (F6P) bound active sites.

3.2 Methods

3.2.1 Active site Structure Comparison

Protein structures with ligand defined active sites were collected from the Protein Data Bank[20] and subjected to an all versus all comparison using CPASS.[21, 22] CPASS scores were subsequently converted to relative distances by subtracting from 100% CPASS similarity. The distance matrix, due to size and computational constraints, was divided into smaller matrices based on bound ligand. Principal component analysis was applied to the smaller ligand defined matrices using MVAPACK[23] where functional clusters were generated based on Enzyme Commission (EC) number.[24] For each EC number cluster, a

95% confidence ellipse was calculated which was used to find the representative active site with the shortest Euclidean distance to its center. A total of 3996 representative active sites were found and utilized in the remainder of this study.

3.2.2 Phylogenetic Analysis of Representative Active sites

The CPASS distance matrix for the representative active sites was input into FastME for tree generation using the Neighbor-join algorithm.[25] Briefly, the neighbor-join algorithm joins the two closest taxa or nodes in the distance matrix and creates a new node, which has recalculated distances to the remaining taxa and nodes. Multiple iterations of this process build the tree until only a pair of nodes remains. Identification and investigation of the resulting tree structure was accomplished through visual inspection using the Interactive Tree of Life online tool.[26]

3.2.3 Active site Overlays

From the CPASS representative dataset and tree, 8 enzymes were selected for additional investigation. Structural and sequential differences between the active sites of the enzymes (PDB IDs: 1DLJ, 1DLI, 2O2D, 1MV8, 2P3V, 1LBY, 1JP4, 1KA1) were elucidated by visual inspection using Chimera.[27] In each case, residues were considered to be in the active site based on their relative proximity to the bound ligand in their respective crystal structure (6Å). The orientation of the active sites relative to one another was also determined by CPASS, as a standard 3D overlay of the tertiary structures would result in misalignment of the active sites.

3.3 Results

3.3.1 Functional Clustering and Principal Component Analysis of Ligand Defined Active sites

The Comparison of Protein Active site Structures (CPASS) software and database (http://cpass.unl.edu/) was used to compare all protein active sites from the RCSB PDB that contained a bound ligand. CPASS performs a pairwise comparison between two protein active sites, where active site residues were determined based on a defined distance to the bound ligand (6Å). CPASS similarity scores are determined by similarities in both amino acid composition and by the relative amino acid positions between the two compared active site. An "all versus all" distance matrix derived from CPASS similarity scores was initially calculated for all of the ligand defined active sites in the RCSB PDB.

The protein structures were then clustered based on the identity of the bound ligand in order to create function specific protein groupings and to reduce the size of the dataset. A total of 169 protein function groups were created based on a shared identity of the bound ligand. Consequently, a total of 169 principal component analyses (PCA) were then performed using our MVAPACK[23] software for each of these ligand defined protein groups. Group membership within the PCA scores plot was further defined by Enzyme Commission (EC)[24] number and demarcated by a 95% confidence ellipse. A representative example of a PCA scores plot for the collection of fructose-6-phosphate

(F6P) bound active sites is shown in Figure 1. There are 7 different enzymatic functional classes (EC numbers: 2.7.1.105, 2.7.1.11, 3.1.3.11, 3.1.3.25, 3.5.1.25, 5.3.1.8, and 5.3.1.9) and one unannotated group in the PCA scores plot.

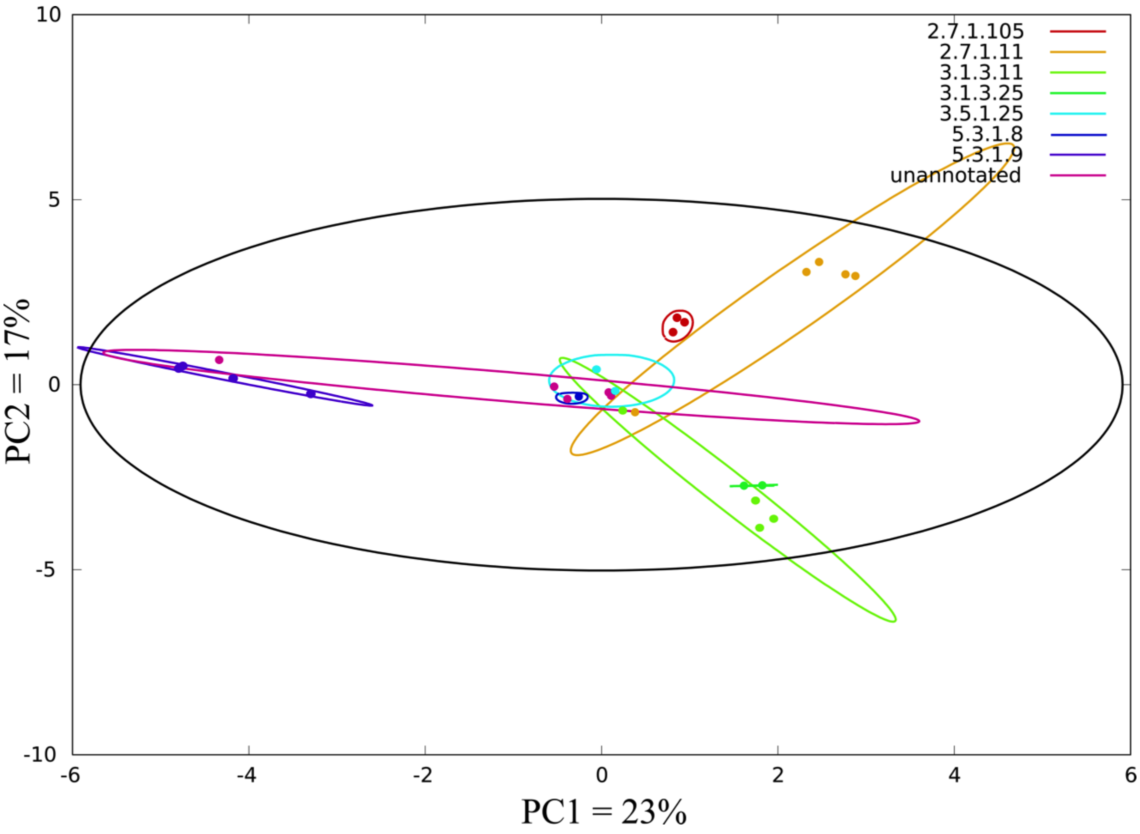


Figure 1. The PCA scores plot of a CPASS distance matrix for fructose-6-phosphate bound proteins. Active sites are clustered by Enzyme Commission number, which refers to a specific function. Ellipses correspond to the 95% confidence intervals for each of the functional clusters (colored) and the dataset (black).

3.3.2 Structural Representatives of Functional Classes

The PCA scores plots were leveraged to find a representative protein structure for each functional class based on EC number the type of bound ligand. For each functional class in the PCA scores plot, the protein active site with the shortest Euclidean distance to the center of the 95% confidence ellipse was chosen as a representative structure. Again, the 95% confidence ellipse defines the membership for a given functional class. Accordingly, the selected protein active site should have a high CPASS similarity score or a small variance relative to the other protein active sites in the functional class. In effect, the selected protein active site is expected to serve as a structural "average" for the functional class. In total, the 169 PCA score plots identified a representative structure for 3996 EC functional classes.

3.3.3 Phylogenetic Analysis

A phylogenetic analysis was conducted using a distance matrix based on CPASS similarity scores for the 3996 protein active sites. The phylogenetic analysis used the neighbor-join algorithm and the resulting phylogenetic tree is shown in Figure 2. The phylogenetic tree is shown in a circular display with leaves colored according to the function defined by the first EC number [oxidoreductases (red), transferases (blue), hydrolases (yellow), lyases (green), isomerases (purple), ligases (orange), not annotated (black)]. Importantly, the functional classification was not used as part of the phylogenetic analysis. Instead, the resulting phylogenetic tree was simply annotated with the known functional classifications.

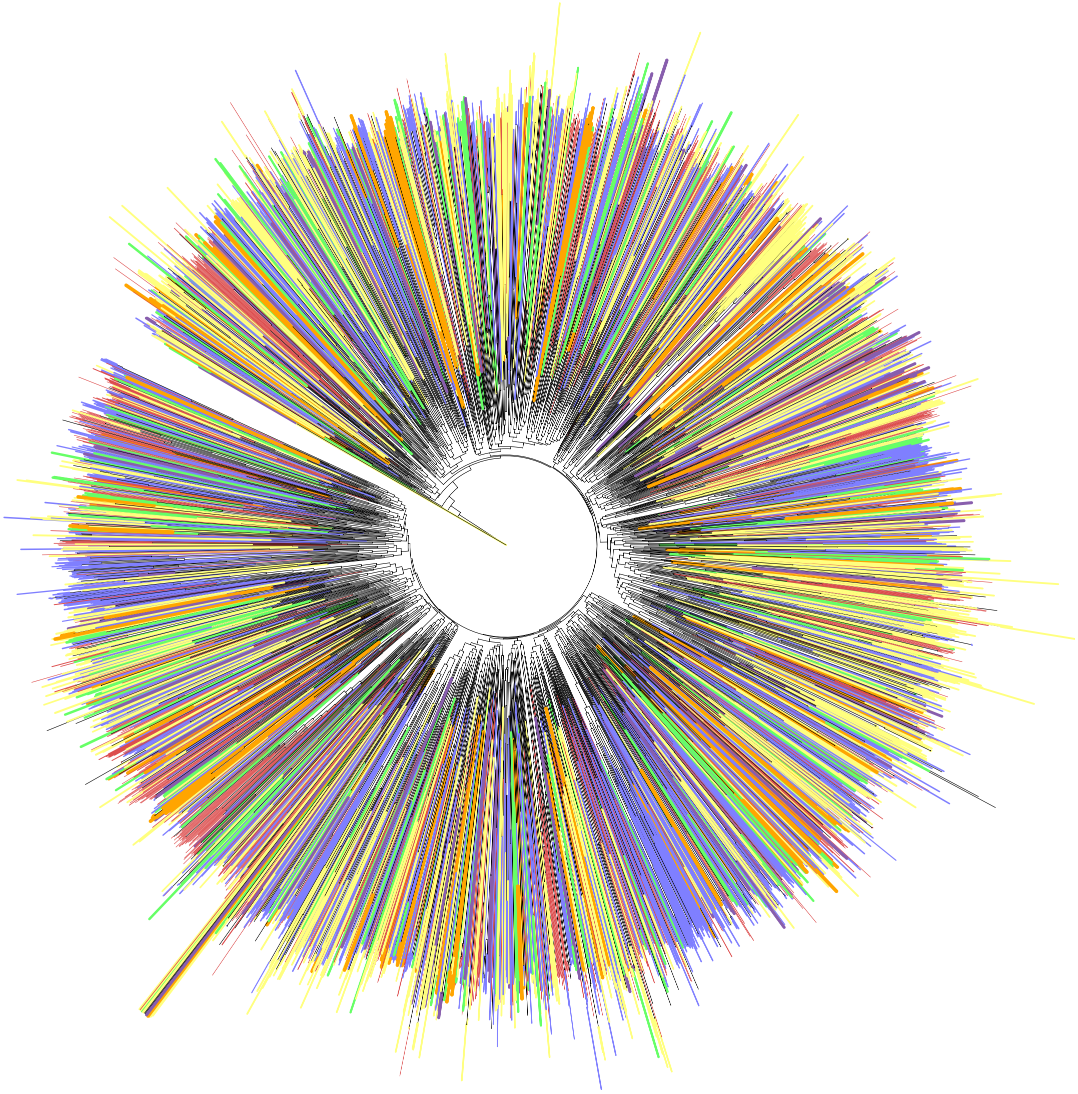


Figure 2. A phylogenetic tree of 3996 representative CPASS active sites from the RCSB PDB is presented. The phylogenetic tree highlights the functional evolutionary relationships between protein active site structures. Leaves are colored according to the first EC number of the annotated active site (1: oxidoreductases, red; 2: transferases, blue; 3: hydrolases, yellow; 4: lyases, green; 5: isomerases, purple; 6: ligases, orange; not annotated: black).

3.4 Discussion

Herein, we report the first functional clustering and evolutionary analysis of the entirety of proteins deposited in the RCSB PDB with a bound ligand. A functional evolution was based on active site similarities determined by our CPASS software and database. Protein active sites were first divided into functional classes based on the type of bound ligand. PCA of the CPASS similarity scores was then used to visualize the relative similarities of the functional class membership. The resulting PCA scores plot was then annotated with EC numbers and the 95% confidence ellipses (Figure 1) were used to define the membership of each functional class within the scores plot.

PCA has been extensively used in chemometrics and various ‘omics’ fields for fingerprint analysis.[28] In this study, PCA was used to reduce the variance within each functional class while also reducing the size of the dataset used for the phylogenetic analysis. The PCA scores plot for the collection of F6P bound active sites (Figure 1) yielded several important observations. First, a number of the 95% confidence ellipses partially overlap in the PCA scores plot. This suggests that there are structural elements that remain consistent within the active site even though the enzymatic functions vary considerably. Second, a complete separation of two functional classes would indicate that the active sites have either diverged significantly over time or have converged to act upon the same substrate. An evolutionary drift is also apparent when considering the shape and positions of the ellipses in the scores plot. The various clusters appear to drift away from the center of the scores plot. Assuming the center of the PCA scores plot is the structural average of all active sites bound to a ligand, the movement of an ellipse or active site toward or away from the center would indicate convergent or divergent evolution, respectively. In effect, the substrate specificity and/or enzymatic activity is diverging as the enzyme moves away from the center of the scores plot or converging as it moves towards its center.

The resulting scores plot also suggests that a combination of CPASS and PCA may be a useful approach for annotating functionally unclassified or hypothetical proteins. For example, the PCA scores plot of the active sites bound to F6P contained a functionally unannotated class of proteins. While the 95% confidence ellipse for this unannotated class was understandably broad (presumably because it contains different functions), individual proteins were found to be near or within other functional classes suggesting possible annotation(s) (Table 1). Furthermore, proteins found within the intersection of two or more ellipses may suggest a bi-functionality or a functionally promiscuous active site.

Table 1.

PDB ID:	Nearest EC Cluster[a]
1UXR	3.5.1.25
2R66	3.5.1.25
3BXH	5.3.1.8
3PR3	5.3.1.9
3PT1	5.3.1.8

[a] Shortest Euclidean distance to the center of the 95% confidence ellipse

In this study, PCA was primarily utilized to identify a representative protein structure for each functional class. Simply, the protein structure closest to the center of each ellipse was identified as the representative active site for the functional class. For example, a total of eight protein structures were identified from the PCA scores plot of the F6P bound active sites shown in Figure 1. This corresponds to one protein representative for each of the seven EC functional classes and one protein for the single unannotated class. In this manner, PCA allowed for a drastic reduction in the size of the dataset to about 10% of its initial size. Consequently, a distance matrix was generated from an all-vs-all comparison of the 3996 representative protein active sites from each functional class. The matrix of CPASS similarity scores were then subjected to the neighbor-join algorithm for a phylogenetic analysis (Figure 2). The resulting phylogenetic tree captures the stepwise functional evolution of essentially all of the protein functional classes present in the RCSB PDB.

Protein active sites are paired together in the tree according to enzymatic function, which was also seen in our previous study of PLP-dependent enzymes.[18] The structure and amino-acid composition of a protein active site is typically highly conserved in order to maintain function and retain cellular fitness. Thus, functional evolution of a protein progresses slowly and likely follows a step-wise process of single-amino substitutions that also involves gene duplication. The process proceeds until a new function or substrate specificity is achieved. Importantly, this step-wise evolution of function is clearly evident in our phylogenetic tree of protein active sites. Nearest-neighbors, even those from different organisms, have very subtle differences in active site structures and/or sequence. Simply, as an active site progresses towards the next node, a change in substrate specificity or enzymatic activity may result from a few amino-acid substitutions and/or minor conformational change (Figures 4, 5). Interestingly, while nearest neighbors share similar function, an overall view of the functional distribution throughout the entire phylogenetic tree is much more complex and diverse. This is apparent from the relatively random distribution of color throughout the phylogenetic tree, where leaves are colored according to the first EC number for each representative protein. The phylogenetic tree is not uniformly divided or colored into six contiguous functional classes. Instead, there are many small pairings and subgroupings of similar functional classes that are evenly distributed throughout the tree. This mixing of function is likely due to multiple ancestral active site scaffolds that have evolutionarily diverged and then expanded their biological roles. The organization of the phylogenetic tree is also consistent with convergent evolution where distant active site architectures have slowly mutated toward the same enzymatic function. In essence, the dramatic dispersion of color throughout the phylogenetic tree is further evidence of the multitude of divergent and convergent events that have occurred in the evolution of protein functions.

Two representative regions of the phylogenetic tree have been highlighted to further illustrate the effectiveness of an evolutionary clustering of function based on CPASS similarity scores (Figure 3). It is important to note that two branches highlighted in Figure 3 come from distinct regions of the phylogenetic tree. Nevertheless, both branches contain a protein active site bound to fructose-6-phosphate (F6P), where two proteins (1LBY[29] 3.1.3.25, 3M5P 5.3.1.9) were representative structures identified from the PCA scores plot displayed in Figure 1. Using these two proteins as arbitrary evolutionary starting points, a step-wise evolution of substrate specificity and enzymatic activity is easily observed. The active site of 1LBY has inositol-phosphate phosphatase activity and was found to be most similar to 2P3V,[30] which shares the same function as 1LBY (Figure 3A). An overlay of the 1LBY and 2P3V CPASS determined active sites reveals an almost identical match in terms of both amino acid identity and geometry (Figure 4A). This is to be expected as nearest neighbors have the closest distance (highest CPASS similarity). Furthermore, the primary difference between the two active sites is the identity of the bound ligand. 2P3V is bound to S,R meso-tartaric acid instead of F6P, which was simply a result of the crystallization conditions. This outcome also demonstrates an important feature, the robustness of CPASS to identify highly similar active sites independent of the identity of the bound ligand.

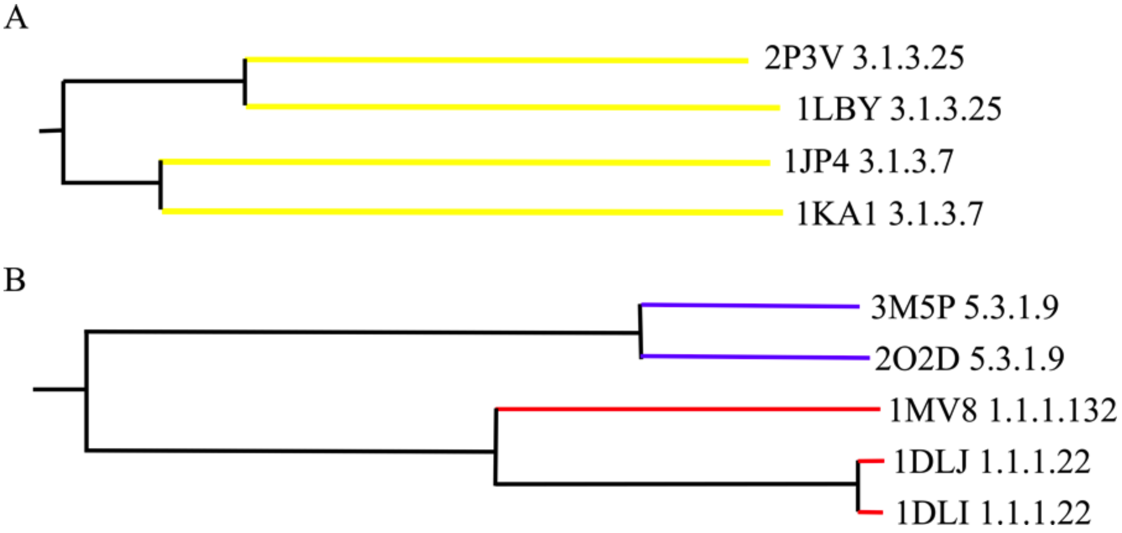


Figure 3. Two regions of the phylogenetic tree from Figure 2 were selected for further detailed analysis. (**A**) The protein active sites in this branch of the phylogenetic tree illustrate protein functional evolution that results in changes in substrate specificity. (**B**)

The protein active sites in this branch of the phylogenetic tree illustrate protein functional evolution that results in changes in both enzymatic activity and substrate specificity. Proteins are listed by their PDB IDs and EC functions (3.1.3.25: inositol-phosphate phosphatase; 3.1.3.7: 3'(2'), 3’ phosphoadenosine-5’-phosphate phosphatase; 5.3.1.9: glucose-6-phosphate isomerase; 1.1.1.132: GDP-mannose 6-dehydrogenase; 1.1.1.22: UDP-glucose 6-dehydrogenase).

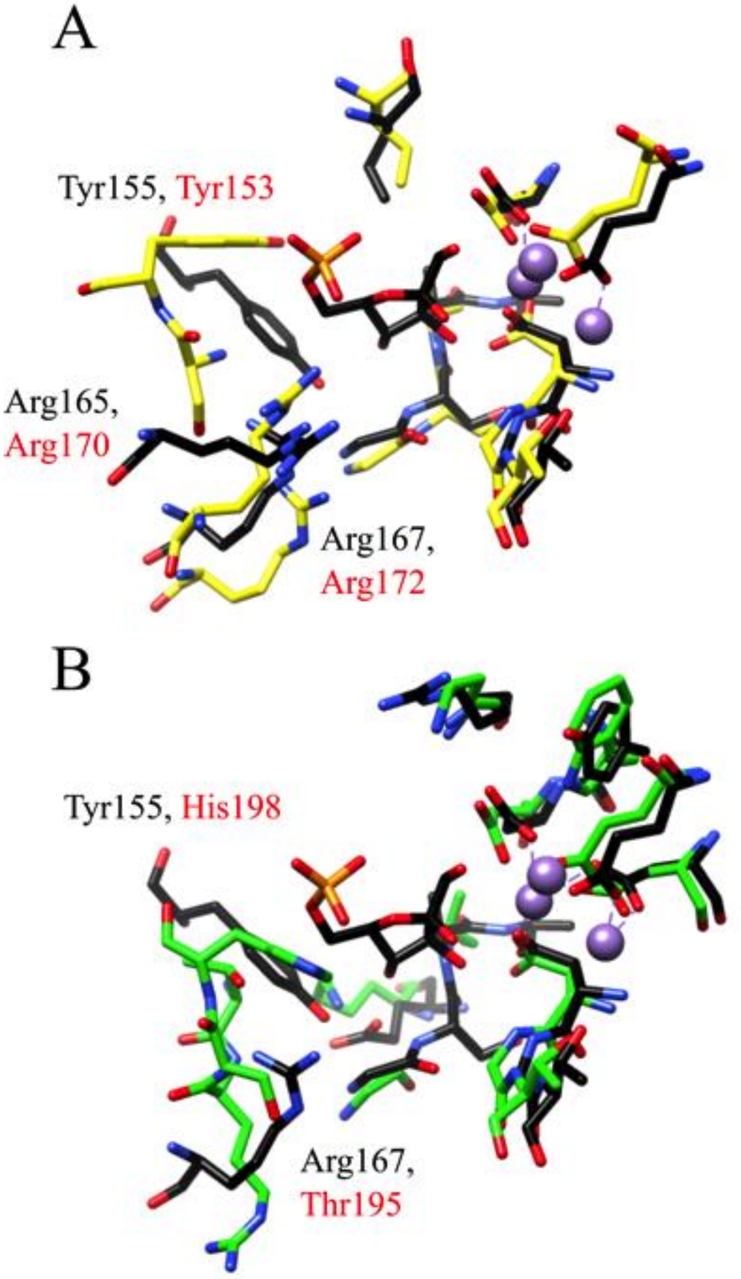


Figure 4. Structural overlays of the active sites for (**A**) 1LBY (black) and 2P3V (yellow), and (**B**) 1LBY (black) and 1JP4 (green). Residues are labeled by type and sequence position with those from 1LBY in black and those from 2P3V and 1JP4 in red. Overlays are oriented relative to the bound F6P in 1LBY with the coordinated magnesium ions displayed in purple. Residues were chosen for the comparative analysis if they were

within 6Å of the bound ligand and were used in the CPASS similarity scoring.

The next nearest node to 1LBY in Figure 3A includes two protein active sites (1JP4[31] and 1KA1[32]) with a similar function (identical for the first three EC numbers), but that act on different substrates. The CPASS determined active sites for 1JP4 and 1KA1 are quite similar (*not shown*), where the primary difference is the identity of the bound ligand (adenosine monophosphate vs. adenosine-3'-5'-diphosphate). In effect, these two nearest-neighbor nodes (Figure 3A) contain a pair of proteins with similar functional classification (3.1.3.25 or 3.1.3.7), but with different bound ligands in the experimental structures deposited in the RCSB PDB. A comparison of the 1LBY active site, an inositol-phosphate phosphatase, with 1JP4, a 3' phosphoadenosine-5'-phosphate phosphatase, reveals minor structural and amino acid differences between the two active sites (Figure 4B). Since the active sites have a similar function but different substrate specificity, the observed changes in amino acid composition and active site geometry are most likely related to substrate binding.

The four proteins (1LBY, 2P3V, 1JP4, 1KA1) comprising this node are magnesium dependent phosphatases, which have an evolutionarily conserved active site and coordinate 2 to 3 metal ions.[32] The metal ions specifically enable the catalytic dephosphorylation of bound substrates and are essential to enzyme activity. Interestingly, the metal coordination sites are strictly conserved even though the metal ions do not participate in substrate binding. Critical to our study, the sequence identity for members of the Mg^{2+}-dependent superfamily is below 25%,[31] which makes sequence-based evolutionary analysis extremely challenging and further highlights the benefits of our CPASS approach. In fact, the CPASS analysis further confirms the high conservation of the metal coordination site. This is apparent in the structural overlays in Figure 4. The active site residues identified by CPASS around the coordination sites deviate very little in position while sequence identity is absolutely maintained. Conversely, the residues opposite the metal coordination sites, which do change between Figures 4A to 4B, are involved in substrate recognition.

Since the active sites for 1LBY and 2P3V have the same function and act upon the same substrate, the tyrosine (1LBY:Tyr155, 2P3V:Tyr153) and arginine (1LBY:Arg165, Arg167, 2P3V: Arg170, Arg172) residues on the distal side of the active site relative to the metal ions are conserved. These residues assist in the coordination of the substrate sugar and phosphate moieties, respectively. For 1JP4, the substrate is changed to 3’-phosphoadenosine 5’-phosphate (PAP), which induces spatial changes and amino-acid substitutions in the active site (Figure 4B). Specifically a tyrosine is replaced by a histidine (1JP4: His198) and an arginine is replaced by a threonine (1JP4: Thr195). These amino-acid substitutions form a new hydrogen bond network around the PAP 5’-phosphate moiety.[31] A reorientation of the side chain of the remaining arginine creates space to accommodate the increase in size of the PAP ligand. Interestingly, the PAP phosphatase maintains some of its inositol 1-phosphate/fructose-1,6-bisphosphate phosphatase activity.[31] Considering how close the active sites are to one another in the phylogenetic tree and the similarity of the active site structures, the residual enzymatic activity is understandable.

A similar comparative analysis of protein evolution is illustrated by examining another branch from the phylogenetic tree (Figure 3B). Unlike the first illustrated example that lead to an evolution of substrate specificity (Figure 3A), this branch leads to the proteins adopting new functions in addition to changes to substrate specificities. The evolutionary focus of this branch is the active site of 3M5P, which is a glucose-6-phosphate (G6P) isomerase and was identified as a representative structure from the PCA scores plot in Figure 1.

The active site of 3M5P was found to be most similar to 2O2D,[33] which has the same function as 3M5P (EC 5.3.1.9, G6P isomerases). An overlay of the two CPASS active site structures in Figure 5A indicates near identity in regards to both amino-acid composition and structure. Again, the only difference in these two active site structures is the nature of the bound ligand. 3M5P is bound to F6P; whereas, 2O2D is bound to citrate. This difference is likely just a result of differences in the crystallization buffers. Critical residues in the active sites of 3M5P and 2O2D that are directly responsible for enzymatic activity are a lysine (3M5P: Lys505, 2O2D: Lys571), a glutamate (3M5P: Glu346, 2O2D: Glu411), and an arginine (3M5P: Arg261, 2O2D: Arg326). These residues are positioned directly around the substrate sugar moiety,[33] facilitate proton transfer (Lysine, Glutamate), and stabilize the intermediate structure.

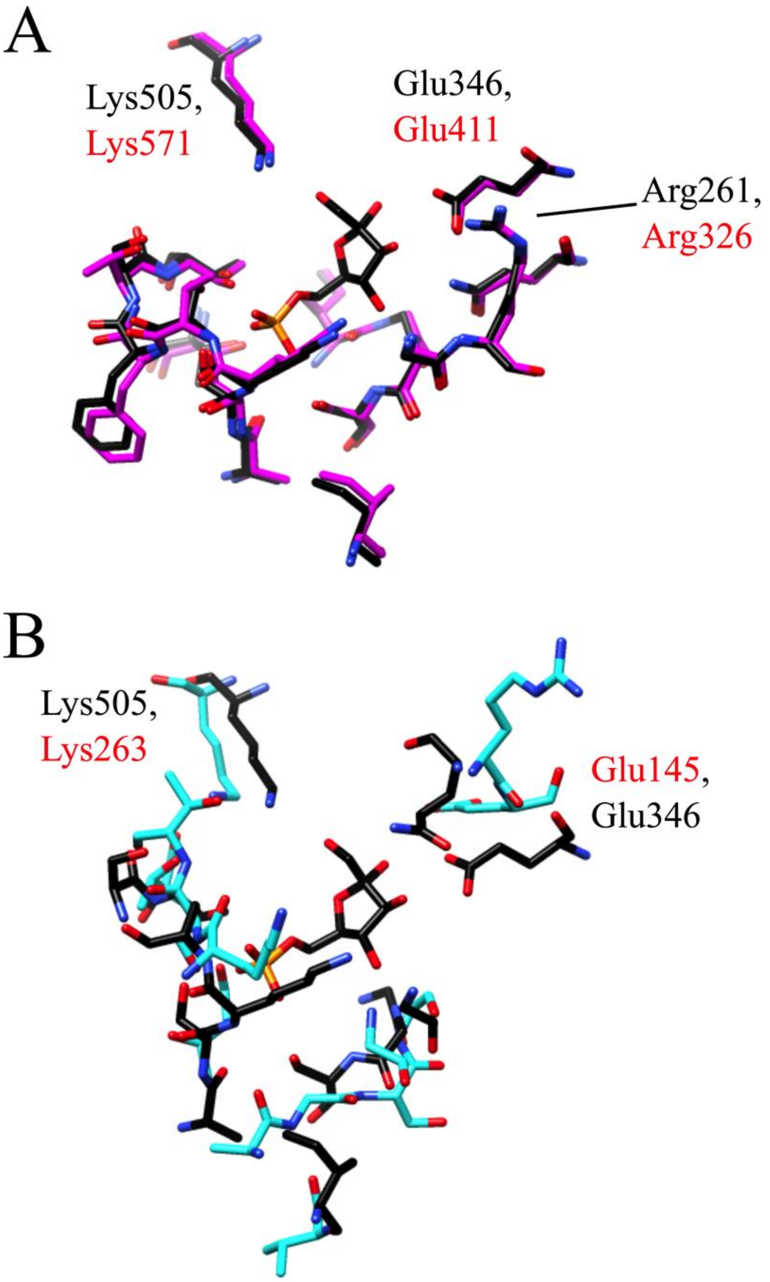


Figure 5. Structural overlays of the active sites for (**A**) 3M5P (black) and 2O2D (pink), and (**B**) 3M5P (black) and 1DLI (cyan). Residues are labeled by type and sequence position with those from 3M5P in black and those from 2O2D and 1DLI in red. Overlays are oriented relative to the bound F6P in 3M5P. Residues were chosen for the comparative analysis if they were within 6Å of the bound ligand and were used in the

CPASS similarity scoring.

The next nearest node to 3M5P includes three protein active sites (1MV8, 1DLJ, 1DLI)[34, 35] with a similar function (Identical for the first three EC numbers), but act on different substrates. The CPASS determined active sites for 1DLJ and 1DLI are essentially identical (*not shown*), where once again the primary difference is the identity of the bound ligand [nicotinamide-adenine-dinucleotide (NAD) vs. 1,4-dihydronicotinamide adenine dinucleotide (NADH)]. However, the active sites are no longer conserved when comparing 3M5P, a G6P isomerase, to 1DLI (or 1DLJ), a UDP-glucose 6-dehydrogenase (U6DH) (Figure 5B). The lysine and glutamate residues, which are important to enzymatic activity remain in 3M5P, but now occupy different positions within the active site. Of particularly note, the importance of these residues to the enzymatic activity of 1DLI has been diminished. Now, the residues only assist in hydrogen bonding to the substrate rather than serving a more integral role in the enzymatic activity of the protein.[35] Moreover, the critical arginine in the G6P isomerase active site is not in the CPASS determined active site for U6DH. Simply, there is no longer a reaction intermediate in U6DH that requires stabilization by an arginine. As a result, the arginine is not conserved in U6DH and the space it occupied has been better utilized. In essence, the overlay of active sites in Figure 5 demonstrates the stepwise evolution from a glucose-6-phosphate isomerase to a UDP-glucose 6-dehydrogenase A similar result is obtained when 3M5P is compared to 1MV8, in which a G6P isomerase is converted into a GDP-mannose dehydrogenase instead of UDP-glucose 6-dehydrogenase (*not shown*).

Comparing these active site structures provides a clear understanding of the slow, step-wise evolution of protein function that is essential to the survivability and adaptability of a cell or organism.

3.5 References

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

[2] Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001) Evolution of Function in Protein Superfamilies, from a Structural Perspective, *Journal of Molecular Biology 307*, 1113-1143.

[3] Fay, J. C., and Wu, C.-i. (2003) Sequence divergence, functional constraint, and selection in protein evolution, *Annu. Rev. Genomics Hum. Genet. 4*, 213-235.

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

[5] Cantarel, B. L., Morrison, H. G., and Pearson, W. (2006) Exploring the relationship between sequence similarity and accurate phylogenetic trees, *Mol. Biol. Evol. 23*, 2090-2100.

[6] Liu, K., Linder, C. R., and Warnow, T. (2010) Multiple sequence alignment: a major challenge to large-scale phylogenetics, *PLoS Curr 2*, RRN1198.

[7] Schaeffer, R. D., and Daggett, V. (2011) Protein folds and protein folding, *Protein Eng Des Sel 24*, 11-19.

[8] Kolodny, R., Pereyaslavets, L., Samson, A. O., and Levitt, M. (2013) On the universe of protein folds, *Annu. Rev. Biophys. 42*, 559-582.

[9] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997) CATH - a hierarchic classification of protein domain structures, *Structure (London) 5*, 1093-1108.

[10] Hughes, A. L. (1994) The evolution of functionally novel proteins after gene duplication, *Proc. R. Soc. London, Ser. B 256*, 119-124.

[11] Prince, V. E., and Pickett, F. B. (2002) Splitting pairs: The diverging fates of duplicated genes, *Nat. Rev. Genet. 3*, 827-837.

[12] Martincorena, I., Seshasayee, A. S. N., and Luscombe, N. M. (2012) Evidence of non-random mutation rates suggests an evolutionary risk management strategy, *Nature (London, U. K.) 485*, 95-98.

[13] Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics 22*, 1658-1659.

[14] Holm, L., Kaariainen, S., Rosenstrom, P., and Schenkel, A. (2008) Searching protein structure databases with DaliLite v.3, *Bioinformatics 24*, 2780-2781.

[15] Shin, J. M., and Cho, D. H. (2005) PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures, *Nucleic Acids Res 33*, D238-241.

[16] Totrov, M. (2011) Ligand binding site superposition and comparison base on Atomic Property Fields: identification of distant homologues, convergent evolution, and PDB-wide clustering of binding sites, *BMC Bioinformatics 12*.

[17] Leinweber, M., Fober, T., Strickert, M., Baumgartner, L., Klebe, G., Freisleben, B., and Hullermeier, E. (2016) CavSimBase: A Database for Large Scale Comparison of Protein Binding Sites, *IEEE Transactions on Knowledge and Data Engineering 28*, 1423-1434.

[18] Catazaro, J., Caprez, A., Guru, A., Swanson, D., and Powers, R. (2014) Functional Evolution of PLP-Dependent Enzymes Based on Active Site Structural Similarities *Proteins: Struct., Funct., Bioinf. 82*, 2597-2608.

[19] Zhang, Y., and Skolnick, J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score, *Nucleic Acids Res. 33*, 2302-2309.

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

[21] Powers, R., Copeland, J. C., Germer, K., Mercier, K. A., Ramanathan, V., and Revesz, P. (2006) Comparison of protein active site structures for functional annotation of proteins and drug design, *Proteins 65*, 124-135.

[22] Powers, R., Copeland, J. C., Stark, J. L., Caprez, A., Guru, A., and Swanson, D. (2011) Searching the protein structure database for ligand-binding site similarities using CPASS v.2, *BMC Research Notes 4*, 17.

[23] Worley, B., and Powers, R. (2014) MVAPACK: a complete data handling package for NMR metabolomics, *ACS Chem Biol 9*, 1138-1144.

[24] Cornish-Bowden, A. (2014) Current IUBMB recommendations on enzyme nomenclature and kinetics, *Perspectives in Science 1*, 74-87.

[25] Lefort, V., Desper, R., and Gascuel, O. (2015) FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program, *Mol Biol Evol 32*, 2798-2800.

[26] Letunic, I., and Bork, P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees, *Nucleic Acids Res 44*, W242-245.

[27] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera-A Visualization System for Exploratory Research and Analysis, *Journal of Computational Chemistry 25*, 1605-1612.

[28] Worley, B., and Powers, R. (2013) Multivariate Analysis in Metabolomics, *Current Metabolomics 1*, 92-107.

[29] Stieglitz, K. A., Johnson, K. A., Yang, H., Roberts, M. F., Seaton, B. A., Head, J. F., and Stec, B. (2002) Crystal structure of a dual activity IMPase/FBPase (AF2372) from Archaeoglobus fulgidus. The story of a mobile loop, *J Biol Chem 277*, 22863-22874.

[30] Stieglitz, K. A., Roberts, M. F., Li, W., and Stec, B. (2007) Crystal Structure of the Tetrameric Inositol 1-phosphate phosphatase (TM1415) from the Hyperthermophile, Thermotoga maritima, *The FEBS journal 274*, 2461-2469.

[31] Patel, S., Yenush, L., Rodriguez, P. L., Serrano, R., and Blundell, T. L. (2002) Crystal structure of an enzyme displaying both inositol-polyphosphate-1-phosphatase and 3'-phosphoadenosine-5'-phosphate phosphatase activities: a novel target of lithium therapy, *J Mol Biol 315*, 677-685.

[32] Patel, S., Martínez-Ripoll, M., Blundell, T. L., and Albert, A. (2002) Structural Enzymology of Li+-sensitive/Mg2+-dependent Phosphatases, *Journal of Molecular Biology 320*, 1087-1094.

[33] Arsenieva, D., Appavu, B. L., Mazock, G. H., and Jeffery, C. J. (2009) Crystal structure of phosphoglucose isomerase from Trypanosoma brucei complexed with glucose-6-phosphate at 1.6 A resolution, *Proteins: Struct., Funct., Bioinf. 74*, 72-80.

[34] Snook, C. F., Tipton, P. A., and Beamer, L. J. (2003) Crystal Structure of GDP-Mannose Dehydrogenase: A Key Enzyme of Alginate Biosynthesis in P. aeruginosa, *Biochemistry 42*, 4658-4668.

[35] Campbell, R. E., Mosimann, S. C., van de Rijn, I., Tanner, M. E., and Strynadka, N. C. J. (2000) The First Structure of UDP-Glucose Dehydrogenase Reveals the Catalytic Residues Necessary for the Two-fold Oxidation, *Biochemistry 39*, 7012-7023.

Chapter 4

Analysis of the Mechanism of Action of Helical Antimicrobial Peptides

4.1 Introduction

The emergence of drug resistant bacteria has made the development of novel antibacterial agents a pressing issue.[1] In fact, most experts consider bacterial infections as one of the greatest threats to human health, and may result in upwards of 10 million deaths worldwide by 2050.[2] Bacterial resistance to common antibiotics occurs through a variety of mechanisms.[3, 4] For example, the gram-positive pathogen methicillin resistant *Staphylococcus aureus* (MRSA) copes with antibiotics by increasing membrane thickness or altering penicillin binding proteins (PBP).[5] Alternatively, gram-negative bacteria such as E*nterobacter* spp., *K. pneumoniae*, and *P. aeruginosa* acquire resistance by developing efflux pumps, producing degrading enzymes, or by altering drug binding targets.[6-8]

An interesting alternative to combat resistant bacteria are antimicrobial peptides (AMPs).[9] AMPs are a class of molecules capable of killing a wide range of microorganisms by membrane pore formation, membrane disruption, or through a "multi-hit" mechanism. AMPs can act on the cell membrane or on cytoplasmic targets such as the inhibition of protein synthesis, enzymatic activity, or cell wall synthesis.[10-12] Perhaps the most attractive feature of AMPs as a source for new antimicrobials is the observation that bacteria are less likely to develop resistance to AMPs as compared to non-peptide antibiotics.[13] This occurs because the fundamental structure of the membrane is not easily altered by mutations. Thus, determining the precise mechanisms of action and the structural features that give rise to AMP antimicrobial activities is key to improving and evolving their efficacy.

Citropin 1.1 (Cit 1.1) is an AMP derived from the skin of an Australian tree frog of the *Litora* genus. Cit 1.1 has shown promising activity against a large number of gram-positive and gram-negative species that includes: *S. aureus*, *S. epidermis*, *S. uberis*, *E. fecalis*, *E. coli*, and *K. pneumonia*.[14] Truncation studies have shown that deletion of Gly1, Leu2, or Phe3 completely ablated Cit 1.1 activity. Conversely, deleting Gly14, Gly15, or Leu6 had no impact on Cit 1.1 activity.[15] Tyler and co-workers suggested that the α-helical conformation of Cit 1.1 was critical to antimicrobial activity.[16] In fact, a similar biological activity was achieved when the secondary structure was maintained, but the L-amino acids were replaced with D-amino acids. However, reversing the peptide sequence did yield a less potent antibacterial compound.[17] In a follow up study, the same group found that if an amide bond is formed between the amine of Leu5 and the carboxylic acid side chain of Asp4 (an isoAsp bond, which alters the linear conformation of Cit 1.1) the biological activity is effectively abolished while the metabolic stability of the peptide is greatly enhanced. This conformational change does not affect the total charge of the peptide, and only marginally changes the isoelectric point and hydrophobicity, but the isoAsp bond significantly alters its secondary structure.[18] In addition, it has recently been proposed that two α-helices are formed upon contact of Cit 1.1 with bacterial membranes.[15]

Taken together, the α-helical structure of Cit 1.1 appears to play an important role in the peptide's antimicrobial activity. Of particular note, these prior studies relied on the introduction of relatively large structural changes to investigate a relationship between amino acid composition and peptide conformation on biological activity. In this report, we expand upon these initial investigations by further characterizing the structure and biological activity the Cit 1.1 peptide in a membrane mimetic environment. The synthesis and antimicrobial evaluation of new Cit 1.1 AMP analogs are also reported. The membrane-bound solution structures of the Cit 1.1 peptides were determined using circular dichroism (CD) and nuclear magnetic resonance (NMR) spectroscopy. The mode of membrane disruption was visualized using microscopy techniques and validated by NMR. The toxicity of the most potent AMP was studied *in vitro*, and the *in vivo* efficacy was studied using a well-establish wax-moth larva assay. The resulting structural and biological data has enabled us to define a mechanism of action for Cit 1.1, to describe the structural arrangement of the peptide when in contact with membranes, and to understand the impact of subtle conformational changes on peptide activity.

4.2 Materials and Methods

4.2.1 Peptide synthesis and purification

AMPs were synthesized using solid phase peptide synthesis (SPPS) on a AAPPTec Focus XC automated peptide synthesizer (Louisville, KY) on a 0.25 mmol scale. All peptides were synthesized using Fmoc-Rink Amide AM resin support (AAPPTec, Louisville, KY). The resin was wetted in dichloromechane (DCM) for 30 min to maximize swelling. The Fmoc- group was then deprotected twice using 20% 4-methyl piperidine (10 mL) each time. The resin was thoroughly washed twice for 10 mins each time using a 1:1 mixture of DCM and dimethylformamide (DMF). The desired amino acids were then added to obtain a predetermined sequence in the peptide synthesizer. Molecular weight of the AMPs was confirmed by MALDI using a SCIEX 4800 MALDI TOF/TOF analyzer (Ontario, Canada). The AMPs were purified using a preparative reverse-phase high performance liquid chromatography (HPLC) using a water (0.1% Trifluoroacetic acid (TFA)) and acetonitrile (0.1% TFA) gradient on an Agilent HPLC system (Santa Clara, CA) using a reverse phase column (Kinetex 5u XB-C18 100 A, 150 × 30.0 mm column). Purity for each AMP was confirmed to be >95% with an Agilent analytical HPLC (Santa Clara, CA). To determine the relative hydrophobicity, all AMPs were run through analytical HPLC using the same column, on the same day, using the same solvent reservoir.

4.2.2 Minimum inhibitory concentration assay (MIC)

Bacteria strains used in this experiment were *Staphylococcus aureus* JE2, *Escherichia coli* K12, and *Klebsiella pneumoniae* ATCC 13883. The MICs of the amphiphilic peptides was determined using a broth microdilution method as described in the third edition of the ASM Clinical Microbiology Procedures Handbook.[19] A stock solution of each peptide was prepared in deionized water and then serial 2-fold dilutions were made in Muller Hinton broth (Difco^{TM}, BD Diagnostics, Becton Drive, NJ), containing 5% DMSO in Cellstar 96-well microtiter plates (Greiner Bio-One, Austria). Bacterial cultures were prepared using the direct colony suspension method to 1.5 X 10^8 colony forming unit (CFU)/ml (0.5 McFarland) and diluted for 2 mL into 40 mL of MHB. Each well was inoculated with 10 μl of bacterial cultures. Plates were statically incubated at 37°C for 16-24 hr. MIC values were taken at the lowest concentration at which no growth was observed with the unaided eye or the microplate reader. The optical density (O.D.) value at 600 nm was recorded using a Thermo Fisher Scientific AccuSkan, MultiSkan FC (Waltham, MA). Vancomycin (Sigma, St. Louis, MO) and Gentamicin (Alfa Aesar, Ward Hill, MA) was used as positive control and blank media was used as negative control. All assays were performed in triplicates.

4.2.3 Circular dichroism (CD) and secondary structure analyses

The AMPs were fully dissolved in Milli-Q water + 20% trifluoroethanol (TFE) to a final concentration of 2 mM. CD spectroscopy was performed at 25 °C using a J-815 Jasco Circular Dichroism Spectrometer (Easton, MD) in a 0.1 mm quartz cuvette. The solvent system (H_2O + TFE) was kept as the reference, which was then subtracted from the AMP spectra. The wavelengths for CD analysis were ranged from 190 - 300 nm with a scan speed of 100 nm/min. The response time was set at 2 s and the bandwidth was set to 1 nm. The spectra were analyzed for their relative proportions of secondary structure using the Dichroweb online server for protein secondary structure analysis employing the K2D analytical program.[20] CD experiments in presence of *E. Coli* were done following a previously published method.[21] Briefly, *E. Coli* K12 was cultured in Muller Hinton Broth and incubated at 37°C. The resultant mid-log phase cultures were diluted to a final concentration of $1.5\text{x}10^8$ CFU/ml using phosphate buffer saline (0.5 McFarland). The spectrum of the bacterial suspension was subtracted from the spectrum collected after addition of the AMP (2 mM) to yield the final spectrum. Helical wheel projections were calculated using the Heliquest online interface.[22]

4.2.4 NMR Structure Determination

Unlabeled peptide samples were prepared similarly to previous studies.[23] Briefly, 2 mM of peptide was dissolved in 90% H_2O and 10% D_2O with 120 mM of SDS-d_{25} (Cambridge Isotope), 50 mM trimethylsilylpropanoic acid (TMSP), and phosphate buffered saline (PBS) buffer at pH 7 (uncorrected). NMR experiments were collected at 25^{o}C on a 700 MHz Bruker Avance III spectrometer equipped with a 5 mm QCI-P probe with cryogenically cooled carbon and proton channels. Proton assignments and NOEs were accomplished with TOCSY, DQF-COSY, and NOESY experiments with shifts referenced to TMSP. NMR spectra were collected with 32 transients and 512 points in the indirect dimension. NOESY spectra were collected with mixing times of 100 ms and 200 ms. Interstrand NOEs were identified using a 3D X-filtered experiment with a mixing time of 120 ms and a 50/50 mixture of unlabeled peptide and a peptide uniformly ^{13}C, ^{15}N labeled at Leu2.[24] Initial models of the peptides were generated from Hα chemical shifts using the CS-Rosetta webserver at the BMRB.[25] Structural refinement was carried out with XPLOR-NIH with dimer symmetry enforced.[26] 400 refined structures were generated for each peptide and the lowest 20 energy structures were subsequently subjected to water refinement in XPLOR-NIH. Validation of the peptide ensembles was done using the Protein Structure Validation Software (PSVS) suite from the Northeast Structural Genomics Consortium.[27]

4.2.5 One-Dimensional NMR Temperature Titration

The unlabeled NMR peptide samples used for structure determination were also subjected to a temperature titration to monitor structure stability. 1D ^{1}H NMR spectra were recorded at temperatures of 25, 30, 40, 50, 60, and 70^{o}C with 64 transients, 128K data points, and a spectral width of 12.87ppm. Between each experiment, the samples were allowed to rest at the target temperature for 5 min before data collection. Following the titration, the 25^{o}C spectrum was recollected to ensure that the sample did not undergo irreversible degradation upon heating.

4.2.6 Paramagnetic Relaxation Enhancement Titration

The unlabeled NMR peptide samples used for structure determination were also used to observe the effects of a diethylenetriaminepentaacetic acid gadolinium (III) dihydrogen salt hydrate (Gd-DTPA) titration. Gd-DTPA was titrated into the samples to concentrations of 0, 2, 4, 8, 16, and 24 mM. 1D ^{1}H NMR spectra were recorded with 128 transients, 64K data points, and a spectral width of 12.87 ppm. Effects resulting from sample dilution were accounted for by normalizing peak intensities to TMSP.

4.2.7 Transmission electron microscopy

E. coli K12 and *S. aureus* JE2 were cultured in Muller Hinton Broth and incubated at 37 °C. The resultant mid-log phase cultures were diluted to a final concentration of $1.5x10^8$ CFU/ml (0.5 McFarland). The bacteria cells were treated with the AMP at 2x the MIC value and incubated for 1 h at 37°C. A control was prepared by adding the media only. After AMP treatment, the cells were immediately washed thrice with HyClone™ Dulbecco's Phosphate Buffered Saline solution (GE Healthcare Life Science, Marlborough, MA) and fixed with 2.0% (v/v) glutaraldehyde[28] Imaging was done on a FEI Tecnai G2 Spirit transmission electron microscope (Japan). 10 µL of the AMP-treated or control bacterial suspension was placed on a copper grid and was allowed to dry for 5 min. The excess suspension was blotted off of the grid and it was dried for 5 mins to remove the residual moisture.

4.2.8 Scanning electron microscopy

Bacterial suspensions for SEM were prepared identically as they were for TEM. The suspension was then placed on 0.1% Poly-L Lysine coated glass slide and allowed to adhere for 30 minutes. The slides were then washed thrice with PBS to remove excess fixative. Samples were post-fixed in a 1% solution of OsO_4 for 30 mins to facilitate conductivity. They were then dehydrated in a graded ethanol series (50, 70, 90, 95, and 100%). Ethanol was removed by washing the slides thrice with hexamethyldisilazane (HMDS). The HMDS was allowed to evaporate overnight to dry the samples. The glass slides were attached to aluminum SEM stubs with double-sided carbon tape. Silver paste was applied to enhance conductivity, which was allowed to dry overnight. Samples then were coated with ≈ 50nm gold-palladium alloy in a Hummer VI Sputter Coater (Anatech, Battle Creek, MI) and imaged at 30kV in a FEI Quanta 200 SEM (Japan) operating in high vacuum mode.

4.2.9 Atomic force microscopy and nano-indentation

Escherichia coli K12 and *Staphylococcus aureus* JE2 were cultured in Muller Hinton Broth and incubated at 37°C. The resultant mid-log phase cultures were adjusted to a final concentration of $1.6\text{x}10^9$ CFU/ml (5.0 McFarland). Cells were treated with AMP at 10 times the MIC (30µM) and incubated for 30 minutes at 37°C. An untreated control was prepared in the same way by adding blank media. Cells were then washed thrice in HyClone™ Dulbecco's Phosphate Buffered Saline solution, and centrifuged at 8000 rpm for 10 mins. AFM imaging of the cells and nano-indentation measurements were performed at ambient conditions with relative humidity of 75% using the MFP 3D system (Oxford Instruments Asylum Research, Santa Barbara, CA). A silicon nitride cantilever MSNL-F (MSNL, Bruker) with nominal value of spring constant in the range of 0.6 N/m was used for both imaging and indentation. The spring constant of the cantilever was obtained using a thermal method prior to each experiment. Indentation was performed in the “Go There” mode after the image of the cells was obtained on various specified points on the cell surface. Collected force-distance curves were analyzed with the Igor Pro 6.23 software package (Wavemetrics, Portland, OR). Mechanical properties were obtained by fitting of the force curves to a cone geometry model of the tip. Fitting was done with calibrated parameters of the tip geometry, which were obtained before the experiment by indentation of the probe against the PDMS Gel reference sample (PDMS-SOFT-1-12M, Bruker) with a known Young's modulus of 2.5 MPa.

4.2.10 FITC uptake kinetics

E. coli K12 and *S. aureus* JE2 were cultured in Muller Hinton Broth and incubated at 37 °C. The resultant mid-log phase cultures were diluted to a final concentration of 1.5x10^{8} CFU/ml (0.5 McFarland). The bacteria cells were treated with the AMP at 2x the MIC value and incubated for 1 h at 37°C. A control was prepared with solvent used to dissolve the AMP. 100 µL of the bacterial suspension was loaded onto a 96-well plate and fluorescein isothiocyanate (FITC) (6 µg/ml) was added and thoroughly mixed with the bacterial cells. The initial fluorescence immediately after FITC addition was recorded and subsequent fluorescence intensities were recorded at 30, 60, 120 and 240 min intervals. The decrease in fluorescence intensity were expressed as % of initial fluorescence and recorded as the means ± SD.

4.2.11 Membrane Internalization Kinetics

E. coli K12 and *S. aureus* JE2 were cultured in Muller Hinton Broth and incubated at 37 °C. The resultant mid-log phase cultures were diluted to a final concentration of 1.5x10^{8} CFU/ml (0.5 McFarland). The bacteria cells were treated with the AMP at 2x the MIC value and the fluorescence value was immediately determined. A control was prepared by adding only the solvent. The initial fluorescence immediately after AMP addition was recorded and subsequent fluorescence intensities were recorded at 30, 60, 120 and 240 min intervals. The decrease in fluorescence intensity were expressed as % of initial fluorescence and recorded as the means ± SD.

4.2.12 *In vivo* assay on *G. mellonella* model of Staphylococcal infection

500 µL of the bacterial suspension was prepared and adjusted to 2.5×10^{8} CFU/10 µL in PBS buffer. 10µL of the bacterial suspension was injected to the last left pro-leg of the *G. mellonella* larva. Following injection, the infected larva was incubated in top-side perforated petri dishes for 2 hours. All larvae were found to be alive 2 hours post infection (n=15). Following this the AMP-treated group and Vancomycin-treated groups were treated with AMP-016 (15mg/kg) and Vancomycin (15mg/kg) respectively. The larvae were observed every 24 hr for 5 days. A survival graph was plotted representing % survival over time. The survival data were plotted using the Kaplan Meyer method and comparisons were made between groups using the log rank test.

4.2.13 *In vitro* toxicity studies

AMPs were dissolved in Dulbecco's Modified Eagle Medium (DMEM) containing 10% FBS and 1% penicillin streptomycin to a final concentration of 4 mM. An aliquot of this stock solution was taken and serial dilutions were performed using DMEM buffer to obtain the desired concentration. Cytotoxicity was performed using a celltiter blue assay and following the manufacturer’s protocols. Cell death was determined by observing the change of absorbance of resazurin using a spectrophotometer. The results were plotted as Absorbance vs AMP concentration and the IC50 (concentration that kills 50% of the cell population) was calculated.

4.3 Results

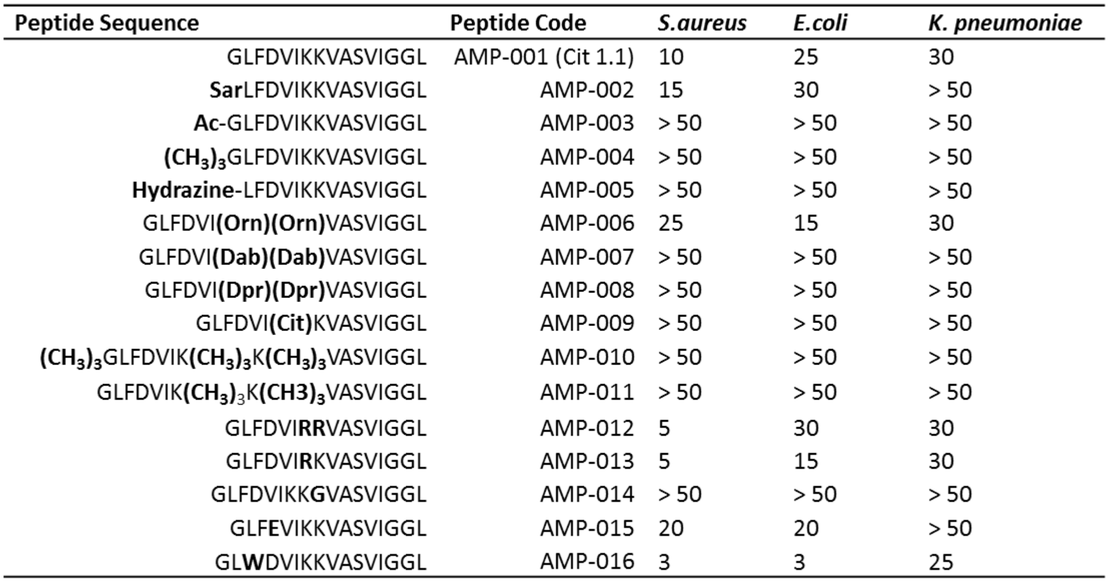
4.3.1 Peptide Synthesis and Characterization

The crude AMPs were identified by Matrix Assisted Laser Desorption Ionization-Time of Flight (MALDI-TOF) spectrometry (Supporting information). The purities of the AMPs were determined to be >95% and the relative hydrophobicities of the AMPs were determined using analytical High-Performance Liquid Chromatography (HPLC) (Table 1).

	HaCat (uM)	***Retention Time* (mins)**	***MW***
AMP-001	50	17.48	1615
AMP-002	120	17.78	1629.03
AMP-003	>310	20.09	1657.04
AMP-004	>310	17.35	1658.09
AMP-005	>310	19.52	1615.99

AMP-006	45	16.86	1586.94
AMP-007	140	16.63	1558.89
AMP-008	>310	16.44	1530.84
AMP-009	>310	18.89	1644
AMP-010	>310	21.76	1724.26
AMP-011	>310	17.88	1700.17
AMP-012	>40	17.83	1671.03
AMP-013	60	17.91	1643.01
AMP-014	>310	16.05	1673
AMP-015	>310	18.18	1629.03
AMP-016	45	18.22	1654.04

Table 1. Cytotoxicity against HaCat cell lines, relative retention times and calculated molecular weights of the antimicrobial peptides.



Peptide Sequence	Peptide Code	*S.aureus*	*E.coli*	*K. pneumoniae*
GLFDVIKKVASVIGGL	AMP-001 (Cit 1.1)	10	25	30
SarLFDVIKKVASVIGGL	AMP-002	15	30	> 50
Ac-GLFDVIKKVASVIGGL	AMP-003	> 50	> 50	> 50
(CH$_3$)$_3$GLFDVIKKVASVIGGL	AMP-004	> 50	> 50	> 50
Hydrazine-LFDVIKKVASVIGGL	AMP-005	> 50	> 50	> 50
GLFDVI**(Orn)(Orn)**VASVIGGL	AMP-006	25	15	30
GLFDVI**(Dab)(Dab)**VASVIGGL	AMP-007	> 50	> 50	> 50
GLFDVI**(Dpr)(Dpr)**VASVIGGL	AMP-008	> 50	> 50	> 50
GLFDVI**(Cit)**KVASVIGGL	AMP-009	> 50	> 50	> 50
(CH$_3$)$_3$GLFDVIK**(CH$_3$)$_3$**K**(CH$_3$)$_3$**VASVIGGL	AMP-010	> 50	> 50	> 50
GLFDVIK**(CH$_3$)**$_3$K**(CH3)$_3$**VASVIGGL	AMP-011	> 50	> 50	> 50
GLFDVI**RR**VASVIGGL	AMP-012	5	30	30
GLFDVI**R**KVASVIGGL	AMP-013	5	15	30
GLFDVIKK**G**VASVIGGL	AMP-014	> 50	> 50	> 50
GLF**E**VIKKVASVIGGL	AMP-015	20	20	> 50
GL**W**DVIKKVASVIGGL	AMP-016	3	3	25

Table 2. Antimicrobial activities of Citropin and its analogues against *S. aureus*, *E. coli*, *P. aeruginosa* and *K. pneumoniae*. Activities are listed by concentration (μM).

4.3.2 Minimum Inhibitory Concentration (MIC Assay)

Cit 1.1 showed high activity against *S. aureus* (10 μM), *E. coli* (25 μM), and *K. pneumoniae* (30 μM) corroborating findings from previous studies (Table 2).[16, 29-31] The AMP-002 analog was synthesized by substituting a methyl glycine (Sar1) for Gly1. AMP-002 retained activity against *S. aureus* and *E. coli*, but exhibited diminished activity against *K. pneumoniae* (>50 μM). The N-terminus acetylation of Cit 1.1 yielded the AMP-003 analog, which resulted in completely diminished activity against *S. aureus*, *E. coli* and *K. pneumoniae*. Methylation of the N-terminus amine of Cit 1.1 yielded the AMP-004 analog, which also resulted in full ablation of activity across all of the tested bacterial strains. The AMP-005 analog was synthesized by substituting Gly1 with a hydrazine group, which resulted in complete loss of antibacterial activity.

Replacing two Lys residues (Lys7, Lys8) with ornithine (Orn, AMP-006) resulted in a 2.5-fold decrease in the MIC value against *S. aureus* (25 µM) and a 1.7-fold decrease against *E. coli* (15 µM). The activity of AMP-006 against *K. pneumoniae* remained unchanged (30 µM) relative to Cit 1.1. Replacement of the lysine residues with 2,4-diaminobutyric acid (Dab) or 2,3-diaminopropionic acid (Dpr) yielded analogs AMP-007 and AMP-008, respectively. Both Dab and Dpr substitutions resulted in a complete loss of antibacterial activity (> 50 µM). An AMP-009 analog was constructed in which Lys7 was replaced with a citrulline, which is a structural mimic of arginine and contains a urea functional group instead of the charged guanidium group. The citrulline substitution resulted in a complete lack of antibacterial activity for AMP-009. AMP-010 was synthesized by methylating all of the available Cit 1.1 amine groups. AMP-011 was generated by methylating only the Lys amine moieties to produce a peptide analog that still contains a positive charge. Nevertheless, both AMP-010 and AMP-011 exhibited a complete lack of antibacterial activity. AMP-012 was constructed by replacing Lys7 and Lys8 with an arginine. Similarly, an AMP-013 analog was generated by substituting Lys7 for an arginine. The MICs for both AMP-012 and AMP-013 decreased 2-fold against *S. aureus* (5 µM). However, the potency of AMP-012 and AMP-013 against gram-negative organisms was mixed. AMP-013 exhibited a 1.7-fold higher activity against *E. coli* (15 µM) compared to both Cit 1.1 and AMP-012. Conversely, all three peptides exhibited the same activity against *K. pneumoniae* (30 µM). A glycine residue was inserted into the Cit 1.1 sequence following Lys 8 to create the AMP-014 analog. The AMP-014 analog was observed to be inactive against all of the bacterial strains tested. The remaining Cit 1.1 analog (AMP-015) was constructed with a modest substitution of Asp4 for a Glu. As expected, this substitution retained activity comparable to Cit 1.1 against *S. aureus* and *E. coli*, but the activity against *K. pneumoniae* was completely lost. An AMP-016 analog was constructed where Phe3 was replaced with a Trp. The AMP-016 analog exhibited the best antibacterial activity against both *S. aureus* (3 µM) and *E. coli* (3 µM) relative to Cit 1.1 and the other analogs. This corresponded to 3-fold and 8-fold increase in activity, respectively. The activity against *K. pneumoniae* was the same as Cit 1.1 (30 µM).

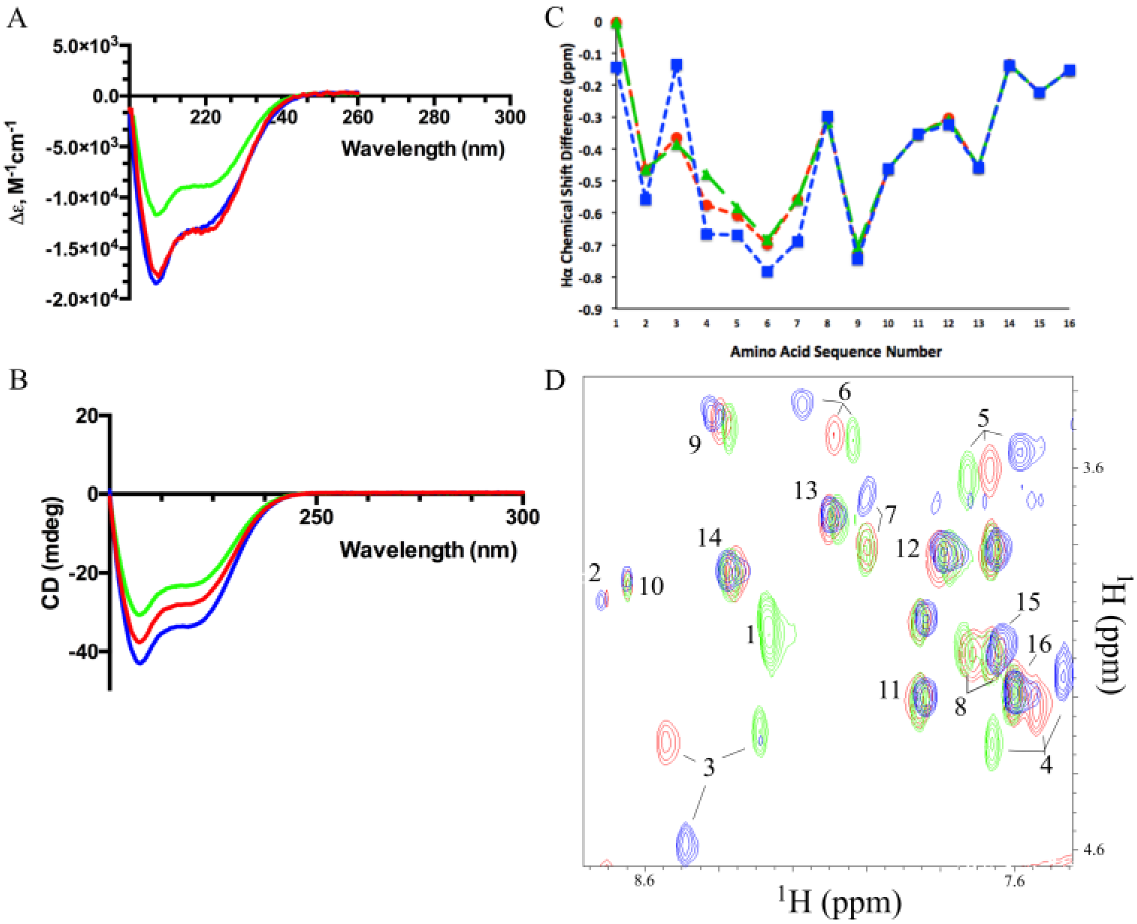
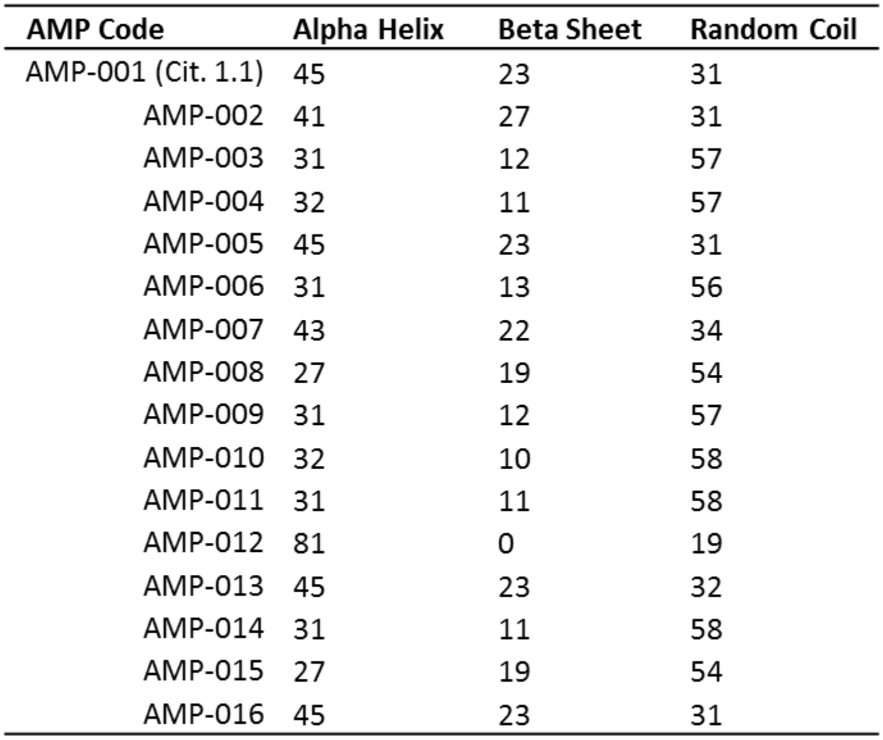


Figure 1. A) Representative CD spectra of Cit 1.1 (red), AMP-003 (green) and AMP-016 (blue) in a solution of 20% TFE+H_2O at a concentration of 2 mM. B) Subtracted CD

Spectrum of Cit 1.1 (red), AMP-003 (green) and AMP-016 (blue), incubated with live *E.coli* cells. Cells (0.1 OD) were suspended in 10 mM phosphate buffer (pH 7) and AMP-016 (25µM) was added to the cuvette, and the spectrum was immediately collected. The background spectrum of the *E. coli* cells was subtracted from the *E. coli* AMP spectrum to derive the final spectrum. (C) Plot showing the Hα deviations (ppm) from statistical-coil chemical shifts. Amino acids are numbered according to their sequence in Cit 1.1 (red), AMP-003 (green), AMP-016 (blue). (D) TOCSY Hα region of Cit 1.1 (red), AMP-003 (green), and AMP-016 (blue). Peaks are labeled according to their respective amino acid position.

4.3.3 Circular Dichroism (CD) Spectroscopy

CD Spectroscopy is routinely employed to determine the secondary structure of proteins and peptides.[32] Accordingly, the antimicrobial activity of the Cit 1.1 analogs was compared to α-helical content using CD spectroscopy (Figure 1A, B). The most active peptides Cit 1.1, AMP-002, -012, -013, and -016 had the highest α-helical content corresponding to ≥ 45% (Table 3). Conversely, the moderately active (AMP-006 and -015) or inactive peptides (AMP-003, -004, -008, -009, -010, -011, and -014) contained a lower α-helical content of ≤ 31%. The inactive AMP-005 and -007 analogs were exceptions and exhibited an α-helical content of approximately 40%. In general, a high α-helical content appears to be a prerequisite for Cit 1.1 activity. The secondary structures for the Cit 1.1 analogs were also analyzed in the presence of bacteria to mimic a native environment (Figure 1B). The Cit 1.1, AMP-003 and -016 were all observed to adopt α-helical structures upon incubation with *E. coli* despite differences in antimicrobial activity. It has been previously demonstrated that the minima of a CD spectrum is an indicator of the proportion of the peptides that have folded into an ordered secondary structure.[21] Accordingly, an observed difference in the minimum depth of CD spectra may identify relative differences in the proportion of Cit 1.1 analogs that adopt α-helical conformations. Comparing the CD spectra of Cit 1.1, AMP-003 and -016 in the presence of *E. coli* cells indicates that AMP-016 has the lowest minima and, consequently, the highest proportion of α-helical peptides. Cit 1.1 was the next lowest and AMP-003 exhibited the least α-helical content. Importantly, this trend correlates with the relative activity of the Cit 1.1 analogs.



AMP Code	Alpha Helix	Beta Sheet	Random Coil
AMP-001 (Cit. 1.1)	45	23	31
AMP-002	41	27	31
AMP-003	31	12	57
AMP-004	32	11	57
AMP-005	45	23	31
AMP-006	31	13	56
AMP-007	43	22	34
AMP-008	27	19	54
AMP-009	31	12	57
AMP-010	32	10	58
AMP-011	31	11	58
AMP-012	81	0	19
AMP-013	45	23	32
AMP-014	31	11	58
AMP-015	27	19	54
AMP-016	45	23	31

Table 3. Relative content of α-helix, β-sheet and random coil in the structure of Cit 1.1 and its analogs. The relative proportion of secondary structures was calculated using the

Dichroweb online interface using K2D analytical algorithm.[20]

4.3.4 NMR Spectroscopy

Negatively charged sodium dodecyl sulfate (SDS) micelles closely mimic a bacterial membrane and were therefore chosen as a model for the structural elucidation of Cit 1.1, AMP-003, and AMP-016 by NMR.[33] Standard two-dimensional (2D) TOCSY and NOESY datasets were collected for backbone and side-chain resonance assignments.[34] NMR assignments for Cit 1.1, AMP-003, and AMP-016 were nearly complete with 94% HN, 100% Hα, 100% Hβ, 88% Hγ, 90-100% Hδ, 60-100% Hε, 0-33% Hζ, and 100% Hη of the assignments made. Figure 1C shows the Hα chemical shift deviations from statistical-coil values for the three peptides bound to SDS micelles.[35, 36] The three peptides show negative chemical shift differences across the entirety of the sequence, which is strongly suggestive of an α-helical conformation along the SDS micelle. Chemical shift differences between the Cit 1.1, AMP-003, and AMP-016 peptides are more apparent when the HN chemical shifts are included. This is illustrated by an overlay of the Hα region of the 2D TOCSY spectra for the three peptides as shown in Figure 1D. Major chemical shifts changes are observed for residues 3 to 7, which are consistent with the Hα chemical shift changes plotted in Figure 1C.

The NMR structures for the Cit 1.1, AMP-003, and AMP-0016 peptides in SDS micelles were determined to further elucidate the role that structure plays in antimicrobial activity. An initial dimer model was generated for each peptide using CS-ROSETTA.[37] The three peptide dimer structures were then refined with XPLOR-NIH[38] based on 536 to 622 intramolecular NOE restraints, and 16 intermolecular (peptide-peptide) NOE restraints. A total of 400 models were calculated per peptide of which the 20 lowest energy structures were further water refined. The structural statistics for the three peptides (Table 4) indicates that the experimental data agrees well with the calculated structures. Backbone RMSDs for each of the peptides was 0.3Å, 0.7Å, and 0.4Å for Cit 1.1, AMP-003, and AMP-016, respectively. The structures did not contain any NOE violations >0.5Å or dihedral angle violations >5°. The high quality of the NMR structures is also evident by the PROCHECK z-scores of -0.3, -0.95, and -0.47 for Cit 1.1, AMP-003, and AMP-016, respectively.[39] Further, 92-94% of all residues are located in the most favored region of the Ramachandran plot. An overlay of the water refined 20 lowest energy NMR structures for the Cit 1.1, AMP-003, and AMP-016 peptides are shown in Figure 2A-C.

Structure statistics for:	Cit 1.1	AMP-003	AMP-016
NOE Restraints	622	552	536
Short Range (|i-j|<=1)	414	390	362
Medium Range (1<|i-j|<5)	208	162	174
Long Range (peptide-peptide)	16	16	16
ϕ/ψ in most favored region	91.70%	94.20%	92.9
ϕ/ψ in additionally allowed region	8.30%	5.80%	7.1
ϕ/ψ in generously allowed	0%	0%	0
ϕ/ψ in disallowed region	0%	0%	0

Backbone atoms rmsd (Å)	0.3	0.7	0.4
Heavy atoms rmsd (Å)	0.5	0.7	0.6
NOE violations (>0.5Å)	0	0	0
PROCHECK (all) Z-score	-0.3	-0.95	-0.47

Table 4. Structure statistics for Cit 1.1, AMP-003, and AMP-016.

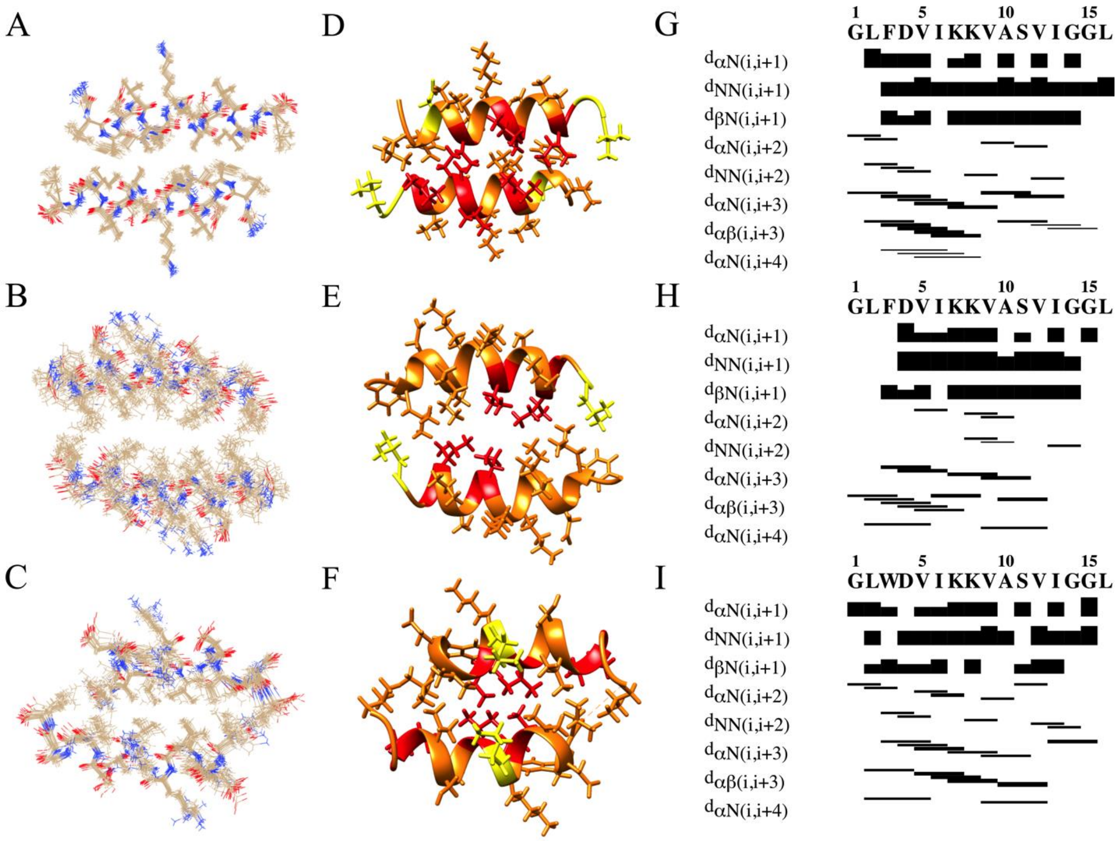


Figure 2. The ensemble of the lowest energy, water refined structures of (A) Cit 1.1, (B) AMP-003, and (C) AMP-016. Peptide ribbon structures for (D) Cit 1.1, (E) AMP-003, and (F) AMP-016 with binned amide proton temperature titration data according to

the ΔH ranges 0 to -0.05 (minor chemical shift, yellow), -0.05 to -0.25 (intermediate chemical shift, orange), and -0.25 to -0.45 (major chemical shift, red). NOE connectivity plots for (G) Cit 1.1, (H) AMP-003, and (I) AMP-016. Weights of the connectivities are normalized to the NOE intensities.

The resulting NMR structures indicate that the Cit 1.1, AMP-003, and AMP-016 peptides adopt a head to tail helical dimer (Figure 2D-F). The head-to-tail orientation of the peptides was confirmed by the presence of NOEs from Leu2 of strand A to Ile13/Leu16 of strand B. These long-range structural restraints are only possible for a head to tail dimer. Of particular note, a significant number of NOEs exist at the N-terminus (Figure 2G, I), which is consistent with the presence of both an α- and 3_{10}-helix for the biologically active peptides (Cit 1.1 and AMP-016). It is possible that the peptides adopt both conformations (or are in equilibrium) when bound to a SDS micelle. In this case, the NMR structure would simply represent an overall average of the conformational exchange. Importantly, the inactive AMP-003 lacks a similar NOE pattern (and number of NOEs) at the N-terminus and, thus, it exclusively adopts an α-helical structure (Figure 2H). Since the only difference between Cit 1.1 and AMP-003 is the N-terminal acetylation, this clear difference in helix composition between the peptides is due to the acetylation.

Amide proton temperature titrations up to 70^{o}C were performed on the three peptides to characterize helical stability (Figures S1-3). The titration data was binned according to the relative magnitude of the HN chemical shift changes ($\Delta^{H}N$): minor (0 to -0.05, yellow), intermediate (-0.05 to -0.25, orange), and major (-0.25 to -0.45, red). The peptide ribbon diagrams in Figure 2D-F are colored based on these HN chemical shift changes in order to qualitatively assess the thermal stability of the helices. Amino acid residues exhibiting major chemical shift changes upon heating are transitioning from a stable and structurally-ordered state to a disordered and unstructured state. Conversely, residues with relatively unchanged chemical shifts are likely already in a transiently disordered state at room temperature.

The peptides were titrated with gadolinium-diethylenetriamine pentaacetic acid (Gd-DTPA) and monitored by NMR to identify the solvent exposed residues while the peptide was bound to SDS micelles (Figure S4-6). The addition of a paramagnetic species will reduce the NMR signal intensity for residues in contact with the solvent proportionally to the amount of the added paramagnetic species. Upon the addition of Gd-DTPA, the majority of the amino acids exhibited an approximate linear decrease in peak intensity regardless of location in the dimer structure. This is most likely due to multiple binding sites for Gd-DTPA on each peptide, an aggregation of the peptides along the surfaces of the micelles, and a dynamic association/dissociation of the dimer.[40] Leu2 in Cit 1.1 and AMP-016 was an exception to this general observation. The NMR peak intensity for Leu2 increased, at first, with the addition of Gd-DTPA. Of note, the peak intensity for Leu2 was relatively weak compared to the other residues (Figure 1D), which is suggestive of a high exchange rate and/or a transiently disordered structure.

4.3.5 Transmission Electron Microscopy (TEM), Scanning Electron Microscopy (SEM) and Atomic Force Microscope (AFM) imaging

Treatment of *S. aureus* and *E. coli* with AMP-016 lead to membrane disruption accompanied by the formation of distinct vesicle-like structures budding as seen by SEM, TEM, and AFM (Figures 3, 4A-D).

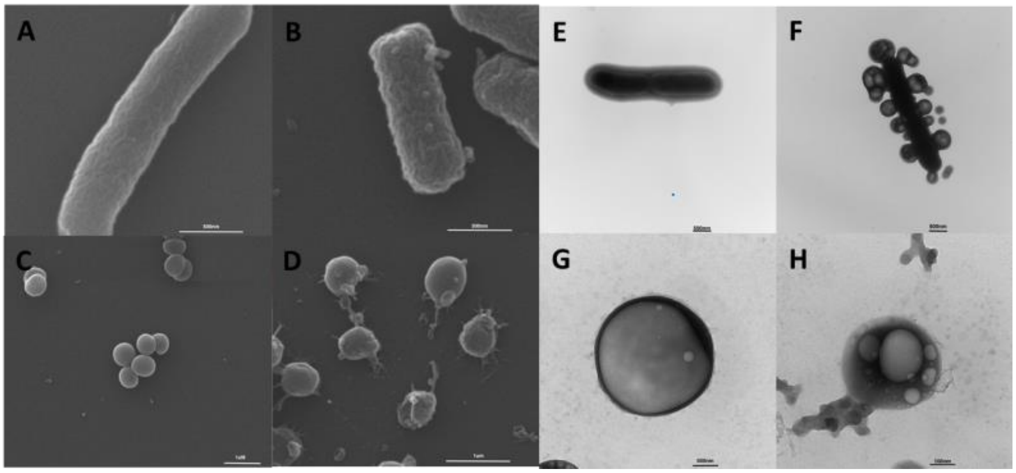


Figure 3. SEM Micrograph of (A) *E. coli* (control) (B) *E. coli* (treated) (C) *S. aureus* (control) (D) *S. aureus* (treated). TEM Micrograph of (E) *E. coli* (control) (F) *E. coli* (treated) (G) *S. aureus* (control) (H) *S. aureus* (treated). For both TEM and SEM, bacteria (10^6 CFU) were treated with AMP-016 (6uM) for 30 mins.

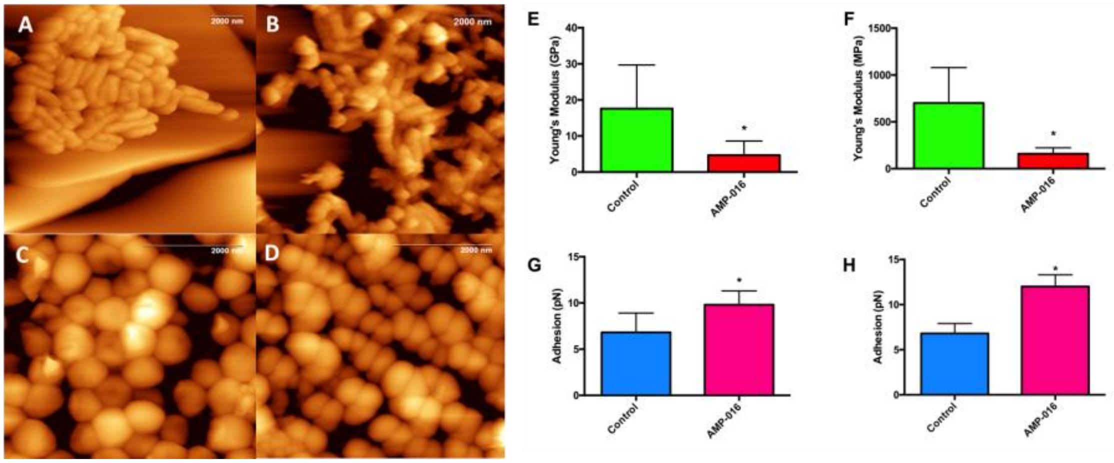


Figure 4. AFM Micrograph of (A) *E. coli* (control) (B) *E. coli* (treated) (C) *S. aureus* (control) (D) *S. aureus* (treated). Young's Modulus of (E) *E. coli* (F) *S. aureus*. Adhesion forces of (G) *E. coli* (H) *S. aureus.* Young's modulus and adhesion forces are presented as mean±SD and were collected over 50 randomly selected points. Statistical significance was determined by Student's t-test. $P < 0.05$ was considered to be statistically significant. The bacterial cells were first imaged and the mechanical properties were subsequently determined from the same samples. Bacteria at 10^6 CFU were treated with AMP-016 (6uM) for 30 mins.

4.3.6 Stiffness of Bacterial Membrane by AFM Nano-Indentation

Untreated *E. coli* membrane was found to have a Young's modulus of about 17.6 GPa, which reduced to only 4.7 GPa upon AMP-016 treatment, a 3.7-fold reduction in stiffness (Figure 4E, G). Young's modulus of *S. aureus* reduced from 700 MPa to 160 MPa, a 4.4-fold reduction in membrane stiffness (Figure 4F, H). Previous AFM nano-indentation studies on *E. coli* have yielded stiffness values ranging between 200 MPa and 3.700

GPa.[41, 42] Although our reported values are higher, these variations might result from differences in bacterial strains, sample processing, or the state of the cells at the time of the study (vegetative vs spore).[43, 44]

4.3.7 Fluorescein Isothiocyanate (FITC) Uptake Kinetics

Membrane disruption causes cellular internalization of FITC and an associated decrease of fluorescence intensity outside the bacterial cells.[45] Thus, a correlation between an increase in antimicrobial activity and the rate and extent of FITC uptake is expected for active Cit 1.1 analogs. AMP-012, -013, and -016 peptides exhibited a greater rate and extent of FITC uptake when compared to Cit 1.1 (Figure 5A).

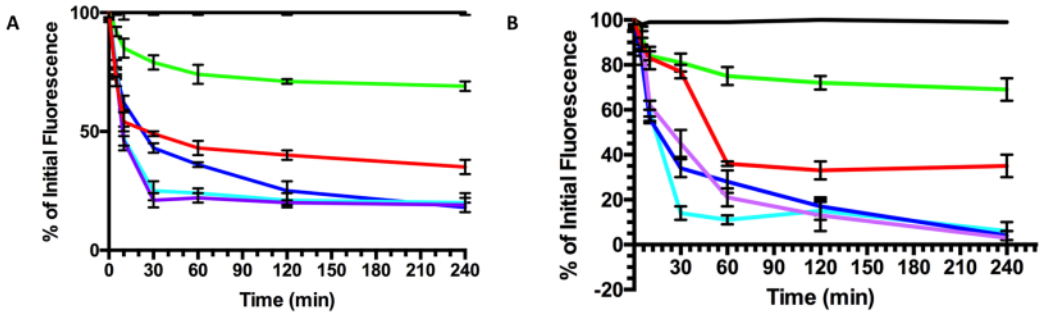


Figure 5. A) FITC uptake kinetics of AMPs. *E. coli* (10^6 CFU) was treated with AMPs (2X MIC) for 240 mins and the percent decrease in fluorescence was determined. The data are presented as mean percent of the initial fluorescence. B) Membrane internalization kinetics of AMPs. *E. coli* (10^6 CFU) was treated with the AMPs and incubated for 240 mins. Data are presented as mean percent of initial fluorescence.

Untreated control (black), Cit 1.1 (red), AMP-003 (green) AMP-012 (turquoise), AMP-013 (purple), AMP-016 (blue).

4.3.8 AMP Internalization Kinetics

Treatment of *E. coli* with AMP-012, -013, or -016 resulted in a significant decrease in fluorescence when compared to Cit 1.1 (Figure 5B). This is consistent with the observation that AMP-012, -013, and -016 displayed a similar or higher antimicrobial activity relative to Cit 1.1 (Table 2). Minimal internalization was observed with AMP-003, which corresponds with its lower antibacterial activity. Introduction of tryptophan or arginine residues into Cit 1.1 appears to increase the rate of membrane binding or the affinity of the peptide dimer to the membrane.

4.3.9 *In-vivo* Efficacy Study

In vivo efficacy of the most active compound, AMP-016, was determined in a wax moth model for staphylococcal infection due to its maximum activity. The AMP-016 treated group displayed a comparable survivability to the vancomycin treated group at a dose of 15 mg/kg (Figure 6).

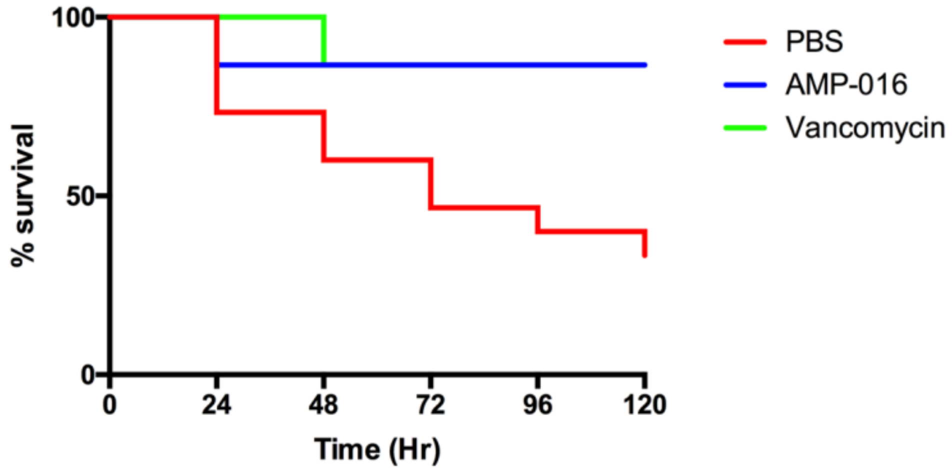


Figure 6: *In-vivo* survival analyses of *G. mellonella* larva following *S. aureus* infection and drug treatment. Larvae were infected with a *S. aureus* suspension (1.5×10^7 CFU). The larvae were then treated with AMP-016 (15 mg/kg) and vancomycin (15 mg/kg), 2 hr post-infection. A log rank test for trend was performed to determine the statistical difference of the curve (p = 0.017).

4.4 Discussion

4.4.1 Antimicrobial Activity

Helical AMPs have been extensively studied as potential antimicrobial candidates for topical and systemic uses; and utilized as the antimicrobial component of anti-biofouling surfaces. However, little attention has been given towards the structural determinants of AMP activity beyond properties such as charge, helicity and hydrophobicity. To this end, we have rationally designed analogs of the Cit 1.1 AMP to look beyond conventional factors thought to be major determinants of antibacterial activity. The N-terminus of

AMPs tend to be polar or ironic, which is likely to play a key role in membrane interactions.[13] For example, a number of helical AMPs contain either a glycine, a positively charged residue (Arg, Lys) or a polar residue (Ser, Asp, Gln) at the N-terminus (*e.g.,* magainin-2, pexiganan, thanatin, esculetin). In effect, all of these helical peptides have an exposed N-terminal amine which are capable of forming hydrogen bonds and are known to participate in electrostatic interactions with negatively charged bacterial membranes. However, the actual role of the N-terminus in AMP activity or the nature of its potential interaction with the membrane remains unclear. To understand the potential importance of electrostatic interactions with the N-terminus of AMPs, several Cit 1.1 analogs were synthesized with various modifications at the N-terminus. The AMP-002 analog, in which Gly1 was replaced with a methyl glycine, exhibited comparable activity to Cit 1.1 against *S. aureus* and *E. coli*. However, the activity of AMP-002 against *K. pneumoniae* was completely abolished. The methylation of a primary amine changes the lipophilicity at the N-terminus. Thus, the reduction of activity due to the Sar substitution also suggests multiple H-bond donors may be needed at the N-terminus of AMPs for full activity. This is also consistent with the observation that the AMP-003 analog, which is acetylated at the N-terminus, was completely inactive. Acetylation of amines alters the pKa and the isoelectric point of a peptide while also reducing the net charge of the peptide.[46] In effect, the acetylation eliminated the ability of the N-terminus of the AMP-003 to hydrogen bond with the membrane.

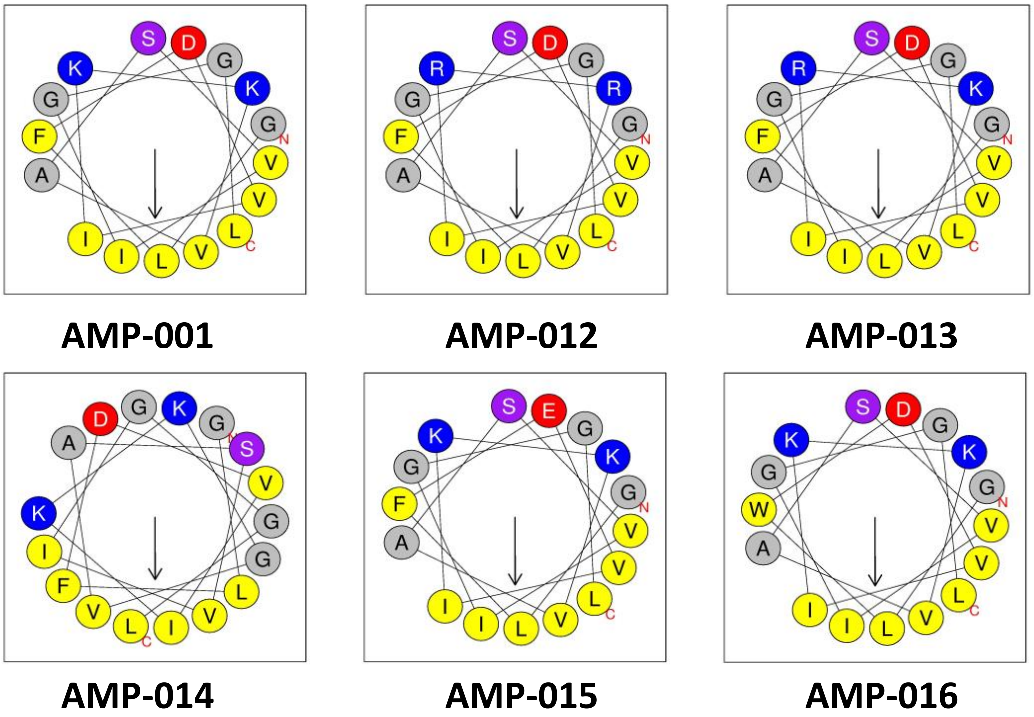
To further investigate the importance of the primary amine on Cit 1.1 activity, the N-terminus was converted into a quaternary amine (AMP-004). This modification was expected to have two distinct and positive impacts on Cit 1.1 activity. First, a permanent positive charge was incorporated into the Cit 1.1 structure that cannot form a hydrogen bond. The role of a resident positive charge in the activity of many antibacterials is well established.[47-49] Second, the quaternary amine makes AMP more resistant to proteolysis.[50] Improved antimicrobial activity would be expected if the peptide is stabilized and has a longer bioavailability. Thus, based on previous observations, AMP-004 was expected to exhibit a better activity than Cit 1.1. Unexpectedly, AMP-004 was observed to be inactive against all tested bacterial strains**.** This indicates that a positive charge on the N-terminal amine is not sufficient to interact with the negatively charged bacterial membrane. Similarly, replacing Gly1 with a hydrazine group resulted in the complete loss of antibacterial activity for the AMP-005 analog. Hydrazines have a lower pKa (~8.3) than amines (~9.6), which is further lowered when near an electron-withdrawing group.[51, 52] As a result, the alkylated hydrazine nitrogen is likely not protonated (note that the alkylated nitrogen is the most likely site of protonation[53]) under the conditions of the MIC assay. Furthermore, an internal hydrogen bond may form between the hydrogen on the hydrazine and the amide carbonyl. In addition to preventing a hydrogen bond with the bacterial membrane, this internal hydrogen bond may also affect the overall peptide structure and alter its biological activity. The inactive AMP-003, AMP-004 and AMP-005 analogs clearly emphasize the importance of the N-terminus in AMP activity and the potential critical role that hydrogen bonds play in the interaction with the bacterial membrane.

In addition to the N-terminal amine, most AMPs contain two or more ionizable amino acids such as Lys and Arg. It is widely recognized that Lys and Arg side-chains are positively charged at biological pH (6.9 to 7.4), which allows for a potential electrostatic interact with the bacterial membrane. This important structural feature has also led to the alternative common name for AMPs of cationic antibacterial peptides, which emphasizes the peptides charge over other properties essential to activity. Accordingly, several modifications to Cit 1.1 were made to evaluate the role of these charged residues to antibacterial activity. Lys7 and Lys8 were replaced with progressively shorter amine bearing sidechains: ornithine (Orn), 2,4-diaminobutyric acid (Dab), and 2,3-diaminopropionic acid (Dpr). Orn, Dab, and Dpr are successively one methylene shorter. Replacing the two Lys with Orn (AMP-006**)** resulted in a peptide with a 2.5-fold decrease in the MIC value against *S. aureus* (25 µM), and a 1.7-fold decrease against *E. coli* (15 µM). The activity against *K. pneumoniae* remained unchanged (30 µM) relative to Cit 1.1. The Dab (AMP-007) and Dpr (AMP-008) substitutions resulted in complete loss of antibacterial activity (>50 µM). These findings suggest the $–NH_3^+$ moiety or the longer side-chain of Lys is necessary for membrane interaction and the antibacterial activity of Cit 1.1. Another possibility is that the decrease in $–CH_2$ units results in a lower hydrophobicity. This is supported by the shorter HPLC retention times for Orn, Dap and Dpr (16.44 to 16.86 mins.) compared to Cit 1.1 (17.48 min). This finding suggests that the incorporation of lysine analogs with longer side-chains may enhance the antibacterial activity of AMPs.

To further explore the relative contribution of the electrostatic interaction and hydrogen-bond formation by the Lys side chain to Cit 1.1 activity, the Lys residues were methylated (AMP-010 and AMP-011) to produce analogs that still contain a positive charge. Unexpectedly, and despite having a permanent +3 charge, both AMP-010 and AMP-011 were found to be inactive. Increasing the methyl substitution of the lysine side chain amine has been found to produce a less favorable cation-π stacking interactions in terms of energy and geometry.[54] This observation raises the possibility that a crucial cation-π stacking interaction exists between a lysine side chain and the phenylalanine at position 3. Additionally, it is possible that the inability of the lysine -NH_2 to form a hydrogen bond with the bacterial membrane might be the cause of the peptide's inactivity.

Arginine-rich peptides have been previously shown to exhibit potent antibacterial activity.[55-57] In fact, a number of studies have indicated that an increase in arginine versus lysine content has resulted in enhanced antimicrobial activity.[58-64] Nevertheless, others have observed a deviation from this arginine-to-lysine enrichment rule.[65-68] It appears that a "sweet spot" exists in the choice of arginine or lysine based on hydrophobicity, hydrophobic ratio, secondary structure, and isoelectric point. To investigate this issue, a Cit 1.1 analog (AMP-012) was constructed where Lys7 and Lys8 were replaced with an arginine. An AMP-013 analog was also constructed that only replaced Lys7 with an arginine. The MICs for both AMP-012 and AMP-013 decreased 2-fold against *S. aureus* (5 µM). The increase in antimicrobial activity may be explained by an enhanced translocation of peptides within the microbial cells.[58, 60, 69-71] However, the potency of AMP-012 and AMP-013 against gram-negative organisms was mixed. AMP-013 exhibited a 1.7-fold higher activity against *E. coli* (15 µM) compared to both Cit 1.1 and AMP-012. Conversely, all three peptides exhibited the same activity against *K. pneumoniae* (30 µM). To further explore the importance of Lys or Arg to Cit 1.1 activity, an AMP-009 analog was constructed in which Lys7 was replaced with a citrulline, which is a structural mimic of arginine, but contains a urea functional group instead of the charged guanidium group. The citrulline substitution resulted in a complete lack of antibacterial activity for AMP-009, further establishing the importance of a Lys at position 7 in the activity of Cit 1.1. Although the citrulline urea group is capable of hydrogen bond formation, its side chain is uncharged. Accordingly, the citrulline substitution reduces the net charge of AMP-009 to only +1. This suggests that the overall net charge of the peptide may also contribute to the antimicrobial activity of AMPs.

Previous studies have shown that Cit 1.1 adopts an α-helical conformation, but there has also been some debate about the exact details of the Cit 1.1 structure. Instead of a single α-helical chain, two helices separated by a β-turn has been suggested for the Cit 1.1 structure.[15] To address this possibility, a glycine residue was inserted into the Cit 1.1 sequence following Lys8 to create the AMP-014 analog. The insertion of Gly9 was expected to provide added flexibility to the peptide[72] and potentially enhance its activity if the two helix model was correct. AMP-014 was observed to be inactive against all of the bacterial strains tested and suggests the two helix model for Cit 1.1 may be incorrect. Nevertheless, it is important to note that the Gly insertion does lengthen the peptide and shifts the sequence beyond Lys8. The sequence shift does affect the amino acids found on each Cit 1.1 surface (Figure 7)**.** Thus, the loss in activity may also be explained by an interruption in the hydrophilic α-helix surface, which leads to a disruption in membrane binding. Of course, if the Cit 1.1 structure is comprised of two α-helices separated by a flexible link, then the added flexibility of the inserted Gly would allow enough mobility for the correct surface of the α-helices to interact with the membrane.


AMP-001

AMP-012

AMP-013

AMP-014

AMP-015

AMP-016

S D G K G V V L V L I I A F G K
N
C
S D G R G V V L V L I I A F G R
S D G K G V V L V L I I A F G R
D G K G S V G G L V I L V F I K A
S E G K G V V L V L I I A F G K
S D G K G V V L V L I I A W G K

Figure 7. Alpha helical properties and position of the amino acid residues were predicted using the Heliquest online analysis tool.[22]

Tryptophan-rich peptides have also been shown to enhance the activity of AMPs.[55, 56, 73] Thus, the potential impact of incorporating a tryptophan into the Cit 1.1 peptide was also investigated. The AMP-016 analog was constructed where Phe3 was replaced with a Trp. The AMP-016 analog exhibited the best antibacterial activity against both *S. aureus* (3 µM) and *E. coli* (3 µM) relative to Cit 1.1 and the other analogs. This corresponded to 3-fold and 8-fold increase in activity, respectively. The activity against *K. pneumoniae* was the same as Cit 1.1 (30 µM). The indole ring in tryptophan is likely acting as a better anchor to the bacterial membranes than the smaller phenyl ring.[73] Biologically active tryptophan-rich peptides typically have multiple tryptophan residues. Therefore, the dramatic increase in antimicrobial activity following such a modest change was remarkable and unexpected.

The remaining Cit 1.1 analog AMP-015 was constructed with a modest substitution of a Glu for Asp4. As expected, this substitution retained an activity comparable to Cit 1.1 against *S. aureus* and *E. coli*, but the activity against *K. pneumoniae* was completely lost. Given the similarities between the peptides, an explanation for this observation is not apparent at this time.

4.4.2 NMR Spectroscopy

Cit 1.1, AMP-003, and -016 were analyzed by NMR to elucidate the relationship between structure and antimicrobial activity. Interestingly, Figure 1C indicates that the largest Hα chemical shift difference between the three peptides occurs at the N-terminus between residues Asp3 and Lys7. Although the dissimilarity between AMP-016 and Cit 1.1/AMP-003 can be attributed to the tryptophan substitution at position 3, the minor deviation of Asp4 is likely a result of the acetylation of the N-terminus. In the case of AMP-016, the steric and electronic effects from the indole side chain are likely to increase the local flexibility at position 3 and potentially lead to a significant decrease in helicity. N-terminal acetylation does not induce any significant effect on the overall helicity of AMP-003 except for a slight decrease at Asp4. The removal of the charged N-terminus may slightly change the polarity or the hydrogen-bond network around Asp4 that may contribute to a decrease in helicity. Figure 1D shows the overlay of the Hα region of the

2D TOCSY for the three peptides. The significant chemical shift differences of the N-terminal residues highlight a changing local environment for residues 3 to 7 that is likely a result of minor alterations in the structure of the three peptides. Accordingly, these subtle changes in the peptide's structure may be directly related to differences in antimicrobial activity.

Our NMR derived structures for Cit 1.1, AMP-003, and -016 differ significantly from the previous results of Sikorska *et al.* that describe the presence of two helices separated by a β-turn.[23] Experimentally determined NOEs were found along the full length of each of the peptides and (Figure 2G-I) and between the helices in a dimer orientation. Although the Cit 1.1, AMP-003, and AMP-016 dimers are not completely superimposable, there are no major structural differences between the three peptide structures that appear to explain the difference in activity. In all cases, the dimers pack together along a hydrophobic face, which then exposes charged residues to the solvent. Consequently, the peptides' NMR structures are consistent with the prior observation that the number and distribution of hydrophobic residues is crucial for antimicrobial activity.[74] Increasing peptide hydrophobicity increases antimicrobial activity, but only if the overall hydrophobicity remains below an optimal upper threshold. If too many hydrophobic residues are inserted into the peptide sequence then unintended peptide aggregation occurs, which leads to a loss in activity. Therefore, the Cit 1.1 peptide must maintain a hydrophobic surface to allow for a stable dimer in order to carpet a bacterial cell membrane, but to also avoid the formation of inactive aggregates. This, however, does not explain the differences in activity between the active (Cit 1.1/AMP-016) and inactive (AMP-003) peptides since all the peptides form a similar head-to-tail dimer. Instead, the primary difference between the active and inactive peptide dimers is the N-terminal acetylation.

Acetylation of the N-terminus in Cit 1.1 most likely changes the polarity and hydrogen bond network near the modification site, which appears to have a profound effect on antimicrobial activity. It has been previously observed that N-terminal acetylation increases the propensity of α-helix formation in peptides regardless of the amino acid sequence.[75] N-terminal acetylation also increases the overall stability of an α-helix. In the case of Cit 1.1, the apparent increase in helical stability due to acetylation resulted in a loss of activity. Consequently, antimicrobial activity may be correlated with a certain amount of structural disorder or mobility within the peptide. This is consistent with a mechanism previously observed for α-synuclein. N-terminal acetylation increased the helical propensity of α-synuclein and substantially decreased protein aggregation.[76, 77] Thus, the antimicrobial activity of Cit 1.1 may likely depend on an equilibrium between a disordered monomer in solution and helical dimers coating the membrane surface. In the presence of bacteria, the equilibrium shifts from disordered monomer to the helical dimer. The dimer is then able to associate with other dimers on the membrane surface and promote antimicrobial activity through a proposed carpet mechanism. However, if the peptide is stabilized in solution than the equilibrium is shifted away from interacting with the bacterial membrane and the antimicrobial activity is abolished. We hypothesize that Cit 1.1 exists as a disordered monomer in solution and dimerizes and aggregates along the anionic surface of the bacterial membrane. In the inactive form, the N-terminal acetylation stabilizes the transiently formed monomer α-helix in solution and the aggregation on the membrane surface, which promotes antimicrobial activity, is greatly diminished. Again, the observation that the inactive AMP-003 dimer is exclusively α-helical and the Cit 1.1/AMP-016 dimers are a mixture of α- and 3_{10}-helix is supportive of the proposal that helix stability affects antimicrobial activity. To further investigate the role of α-helical stability in regards to antimicrobial activity, NMR was used to measure the thermal stability and solvent accessibility of the Cit 1.1, AMP-003, and AMP-016 peptides.

The temperature-dependent chemical shift changes clearly define the dimer interface as the residues with the largest ΔHN values (Figures 2D-F, S1-3). The residues in the dimer interface correspond to: Ile6, Val9, Ala10, Ile13, and Gly14. Importantly, the inactive AMP-003 peptide exhibited the smallest HN chemical shift changes for these residues. This suggests the inactive AMP-003 is relatively less stable as a dimer in the SDS micelle compared to the active peptides. Interestingly, there are additional AMP-003 residues outside of the dimer interface that also exhibit a temperature-dependent chemical shift change. This was not observed for the Cit 1.1 or AMP-016 peptides. One possible explanation for this difference is that the AMP-003 peptide still maintains an α-helical structure as a monomer, which undergoes further temperature-dependent denaturation. This does not appear to occur for Cit 1.1 or AMP-016. In fact, more residues in the Cit 1.1 peptide compared to AMP-003 are disordered based on the lack of a temperature-dependent chemical shift change. This further suggests that a certain amount of structural disorder or dynamics is necessary for antimicrobial activity.

The presence of a helical dimer is also suggested by the Gd-DTPA titration experiments. At low concentrations of Gd-DTPA, the amide peak intensity for Leu2 was found to increase, which indicates a reduction in the chemical exchange with the solvent. It is plausible that the coordination of Gd-DTPA by the active peptides stabilizes the dimer structure and reduces the amide chemical exchange of Leu2 causing an initial sharpening of the peak. Only at higher Gd-DTPA concentrations does the Leu2 NMR resonance broaden like the other amino acid residues. This effect is not seen in the inactive peptide (AMP-003), suggestive of a more stable α-helix along the entirety of the peptide monomer.

4.4.3 Transmission Electron Microscopy (TEM), Scanning Electron Microscopy (SEM) and Atomic Force Microscope (AFM)

Bacterial membrane disruption can be described either by the worm-hole, the barrel-stave, or the carpet model.[78, 79] Previously reported data on the Cit 1.1 interaction with negatively charged lipid vesicles hinted towards involvement of the carpet mechanism.[14] In the carpet-model of membrane disruption, the AMPs bind parallel to the membrane and cover the entire surface. This causes cell membrane permeation and eventually release of small vesicles from the membrane.[78, 79] A separate study involving giant unilamellar vesicles (GUVs) corroborated the carpet mechanism shown in the previous studies.[72] However, these prior studies relied on negatively charged phospholipid model membranes, which are structurally simple compared to bacterial membranes. These model membranes do not accurately represent the complexity of single or double-membraned gram-negative organisms. Furthermore, these studies relied on indirect fluorometric observations instead of direct microscopic imaging.[14, 78] Treatment of *S.*

aureus and *E. coli* with AMP-016 lead to membrane disruption accompanied by the formation of protruding blebs as visualized by SEM, TEM, and AFM (Figures 3, 4A-D). Previous electron micrographs of *S. aureus* showed discontinuous membrane surfaces that did not conform to any of the prevailing modes of membrane disruption,[31, 78] The absence of blebs in the prior studies is most likely due to damaged membrane architecture caused during cell-sectioning.

The carpet mechanism of membrane disruption is commonly associated with detergent-like antibacterials displaying indiscriminant cell lytic properties. It has been also suggested that peptides acting by this mechanism do not require a specific peptide structure.[11] Contrary to these beliefs, we have shown that even those AMPs that disrupt membranes in a detergent like fashion are highly sensitive to subtle structural modifications. A human host defense peptide, LL-37, has also been shown to disrupt membranes following the "carpet model".[80] However, their innocuous presence within human bodies illustrates a high selectivity between prokaryotic and eukaryotic cells.

4.4.4 Stiffness of Bacterial Membrane by AFM Nano-Indentation

Although AMPs are widely thought to disrupt bacterial membranes, few studies exist that have investigated their influence on bacterial membrane stiffness. Consequently, little is known about the mechanical properties of the disrupted membrane. AMP-006 treatment lowered the membrane stiffness by four-fold in both gram-positive and gram-negative bacteria. Based on the electron micrographs from this study and those from others, membrane disruption appears to be a strictly localized phenomenon occurring at specific points of the membrane. However, the highly narrow standard deviation of the Young's modulus indicates the uniformity with which membranes were disrupted over the entire cell. Thus, we conclude cell membrane softening occurred globally in the bacterial membrane. As previously stated, Cit 1.1 appears to disrupt membranes by the carpet model, requiring peptide molecules to bind parallel to the bacterial membrane and then inserting a hydrophobic face into the membrane to cause a disruption.[31] This would require the AMPs to cover the entire bacterial surface, which most likely explains the global reduction in membrane stiffness on AMP treatment even though we observe only localized features of disruption in the form of blebs and leakage.

AMP mediated membrane softening however, has greater implications in augmenting bacterial chemotherapy and overcoming drug resistance. It has been recently shown that disruption of the outer membrane of *E. coli* by pentamidine facilitates overcoming drug resistance in gram-negative bacteria.[81] Pentamidine, however, requires a much higher concentration (25 μM) of drug to disrupt the outer membrane. AMP-016 on the other hand, not only disrupts both the outer and inner membrane at a concentration 5 times lower than pentamidine, it also lowers the membrane stiffness about four-fold on average in both gram-positive and gram-negative bacteria (Figure 4E-H). Depletion of factors which are responsible for maintaining membrane fluidity have been shown to reduce membrane stiffness.[82] Others have shown that altered membrane stiffness has a profound influence on cell channel functions.[83] Reduced membrane fluidity has been a prime mechanism for acquiring drug resistance.[84, 85] Therefore, AMP mediated lowering of membrane rigidity may be an attractive means of facilitating drug penetration and overcoming resistance.

4.4.5 AMP Internalization Kinetics

The rate of membrane internalization was used to measure the relative interaction of Cit 1.1 analogs with bacterial membranes by monitoring changes in peptide fluorescence. Aromatic groups like phenylalanine and tryptophan may anchor a peptide into the bacterial membrane,[73] where the insertion of the side chain is expected to lead to fluorescence quenching. Treatment of *E. coli* with AMP-012, -013, or -016 resulted in a significant decrease in fluorescence when compared to Cit 1.1 (Figure 5B). This is consistent with the observation that AMP-012, -013, and -016 displayed a similar or higher antimicrobial activity relative to Cit 1.1 (Table 2). Introduction of tryptophan or arginine residues into Cit 1.1 appears to increase the rate of membrane binding or the affinity of the peptide dimer to the membrane. These findings may be explained by the ability of arginine and tryptophan to allow the Cit 1.1 analogs to penetrate deeper into the bacterial membrane.[86]

4.4.6 FITC Uptake kinetics

Membrane disruption caused cellular internalization of FITC and an associated decrease of fluorescence intensity outside the bacterial cells.[45] Thus, a correlation between an increase in antimicrobial activity and the rate and extent of FITC uptake is expected for active Cit 1.1 analogs. AMP-012, -013, and -016 peptides exhibited a greater rate and extent of FITC uptake when compared to Cit 1.1 (Figure 5A). Again, this is consistent with the increased rate of membrane internalization (Figure 5B) and the relative increase in antimicrobial activity (Table 2) of these peptides**.** Thus, the arginine and tryptophan modifications to Cit 1.1 appear to induce greater membrane damage compared to the wild-type peptide. These results are consistent with prior observations that arginine and tryptophan-rich peptides enhance antibacterial activity by imparting greater amount of damage to the cell membrane.[55-57]

4.4.7 *In-vivo* efficacy study

In vivo efficacy of the most active compound, AMP-016, was determined in a wax moth model for staphylococcal infection. Cit 1.1 has been shown to have a protective effect against both gram-negative and gram-positive bacteria in *in-vivo* models of sepsis.[87, 88] The AMP-016 treated group displayed comparable survivability to the vancomycin treated group at a dose of 15 mg/kg. AMP activity is often diminished in the presence of salts and serum, which are major obstacles to successful translation to a clinical trial.[89, 90] Our results (Figure 6) indicate that AMP-016 activity is not sensitive to physiological salt or serum proteins and holds significant potential as a new antimicrobial treatment.

4.5 Conclusion

Using finely tuned Cit 1.1 analogs, we have probed the various interactions essential for membrane disruption and antibacterial activity of helical AMPs. Cationic N-terminal residues form crucial hydrogen bonds that were found to be indispensable for antibacterial action. Ablation of either the hydrogen bond donors or the positive charge resulted in inactivity, which suggests hydrogen bonding and electrostatic interactions synergistically cooperate in membrane disruption. We further discovered that the projection of lysine side chains were essential to antibacterial activity. A shortening of the side chain and corresponding reduction of the amine projection led to decreased activity against all tested bacterial strains. Increased bulk of the hydrophobic anchor was found to have the most pronounced influence on the AMPs. Mutation of Phe3 to a Trp produced a peptide with significantly increase antimicrobial activity. NMR studies revealed the propensity of AMPs to form head to tail dimers on the surface of anionic membranes and bacterial surfaces. Both NMR and CD spectroscopy show a positive correlation between a-helicity of the AMPs and their antibacterial activity. Further, equilibrium between a disordered monomer and the helical dimer structures was proposed that supports the carpet model of membrane disruption. This was subsequently confirmed by electron microscopy in both gram-positive and gram-negative bacteria. Fluorometric studies revealed increased internalization of the AMPs and FITC, which further supports membrane disruption. AFM nano-indentation showed a four-fold global reduction membrane rigidity upon AMP treatment. Reduction in membrane rigidity has far-reaching implications for augmenting antibacterial therapeutics and in overcoming drug resistance by using AMPs as part of multidrug therapy.

4.6 References

[1] Livermore, D. M. (2003) Bacterial resistance: origins, epidemiology, and impact, *Clin. Infect. Dis. 36*, S11-S23.

[2] O’Neil, J. (2014) Review on Antimicrobial Resistance. Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations 2014.

[3] Walsh, C. (2000) Molecular mechanisms that confer antibacterial drug resistance, *Nature 406*, 775-781.

[4] Stewart, P. S., and Costerton, J. W. (2001) Antibiotic resistance of bacteria in biofilms, *The lancet 358*, 135-138.

[5] Tenover, F. C. (2006) Mechanisms of antimicrobial resistance in bacteria, *The American journal of medicine 119*, S3-S10.

[6] Paterson, D. L. (2006) Resistance in gram-negative bacteria: Enterobacteriaceae, *The American journal of medicine 119*, S20-S28.

[7] Lin, J., Nishino, K., Roberts, M. C., Tolmasky, M., Aminov, R. I., and Zhang, L. (2015) Mechanisms of antibiotic resistance, *Front. Microbiol. 6*.

[8] Silva, J. (1996) Mechanisms of antibiotic resistance, *Current therapeutic research 57*, 30-35.

[9] Yeaman, M. R., and Yount, N. Y. (2003) Mechanisms of antimicrobial peptide action and resistance, *Pharmacological reviews 55*, 27-55.

[10] Brogden, N. K., and Brogden, K. A. (2011) Will new generations of modified antimicrobial peptides improve their potential as pharmaceuticals?, *Int. J. Antimicrob. Agents 38*, 217-225.

[11] Shai, Y. (2002) Mode of action of membrane active antimicrobial peptides, *Pept. Sci. 66*, 236-248.

[12] Brogden, K. A. (2005) Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria?, *Nature Reviews Microbiology 3*, 238-250.

[13] Zasloff, M. (2002) Antimicrobial peptides of multicellular organisms, *nature 415*, 389-395.

[14] Boland, M. P., and Separovic, F. (2006) Membrane interactions of antimicrobial peptides from Australian tree frogs, *Biochimica Et Biophysica Acta (BBA)-Biomembranes 1758*, 1178-1183.

[15] Sikorska, E., Greber, K., Rodziewicz-Motowidło, S., Szultka, Ł., Łukasiak, J., and Kamysz, W. (2009) Synthesis and antimicrobial activity of truncated fragments and analogs of citropin 1.1: The solution structure of the SDS micelle-bound citropin-like peptides, *Journal of structural biology 168*, 250-258.

[16] Doyle, J., Brinkworth, C. S., Wegener, K. L., Carver, J. A., Llewellyn, L. E., Olver, I. N., Bowie, J. H., Wabnitz, P. A., and Tyler, M. J. (2003) nNOS inhibition, antimicrobial and anticancer activity of the amphibian skin peptide, citropin 1.1 and synthetic modifications, *Eur. J. Biochem. 270*, 1141-1153.

[17] Doyle, J., Brinkworth, C. S., Wegener, K. L., Carver, J. A., Llewellyn, L. E., Olver, I. N., Bowie, J. H., Wabnitz, P. A., and Tyler, M. J. (2003) nNOS inhibition, antimicrobial and anticancer activity of the amphibian skin peptide, citropin 1.1

and synthetic modifications. The solution structure of a modified citropin 1.1, *European journal of biochemistry / FEBS 270*, 1141-1153.

[18] Calabrese, A. N., Markulic, K., Musgrave, I. F., Guo, H., Zhang, L., and Bowie, J. H. (2012) Structural and activity changes in three bioactive anuran peptides when Asp is replaced by isoAsp, *Peptides 38*, 427-436.

[19] Leber, A. L. (2004) Clinical microbiology procedures handbook.

[20] Whitmore, L., and Wallace, B. (2004) DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data, *Nucleic Acids Res. 32*, W668-W673.

[21] Avitabile, C., D'andrea, L. D., and Romanelli, A. (2014) Circular dichroism studies on the interactions of antimicrobial peptides with bacterial cells, *Sci. Rep. 4*.

[22] Gautier, R., Douguet, D., Antonny, B., and Drin, G. (2008) HELIQUEST: a web server to screen sequences with specific α-helical properties, *Bioinformatics 24*, 2101-2102.

[23] Sikorska, E., Greber, K., Rodziewicz-Motowidlo, S., Szultka, L., Lukasiak, J., and Kamysz, W. (2009) Synthesis and antimicrobial activity of truncated fragments and analogs of citropin 1.1: The solution structure of the SDS micelle-bound citropin-like peptides, *J Struct Biol 168*, 250-258.

[24] Breeze, A. L. (1999) Isotope-filtered NMR Methods for the Study of Biomolecular Structure and Interactions, *Prog Nucl Magn Reson Spectrosc 36*, 323-372.

[25] Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J. M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K. K., Lemak, A., Ignatchenko, A., Arrowsmith, C. H., Szyperski, T., Montelione, G. T., Baker, D., and Bax, A. (2008) Consistent Blind Protein Structure Generation from NMR Chemical Shift Data, *PNAS 105*, 4685-4690.

[26] Schwieters, C. D., Kuszewski, J. J., and Clore, G. M. (2006) Using Xplor-NIH for NMR Molecular Structure Determination, *Prog Nucl Magn Reson Spectrosc 48*, 47-62.

[27] Bhattacharya, A., Tejero, R., and Montelione, G. T. (2007) Evaluating Protein Structures Determined by Structural Genomics Consortia Tools for Structure Quality Evaluation, *Proteins: Structure, Function and Bioinformatics 66*, 778-795.

[28] Hartmann, M., Berditsch, M., Hawecker, J., Ardakani, M. F., Gerthsen, D., and Ulrich, A. S. (2010) Damage of the bacterial cell envelope by antimicrobial peptides gramicidin S and PGLa as revealed by transmission and scanning electron microscopy, *Antimicrob. Agents Chemother. 54*, 3132-3142.

[29] Giacometti, A., Cirioni, O., Kamysz, W., Silvestri, C., Del Prete, M. S., Licci, A., D'amato, G., Łukasiak, J., and Scalise, G. (2005) In vitro activity of citropin 1.1 alone and in combination with clinically used antimicrobial agents against Rhodococcus equi, *J. Antimicrob. Chemother. 56*, 410-412.

[30] Giacometti, A., Cirioni, O., Kamysz, W., Silvestri, C., Del Prete, M. S., Licci, A., D'Amato, G., Łukasiak, J., and Scalise, G. (2005) In vitro activity and killing effect of citropin 1.1 against Gram-positive pathogens causing skin and soft tissue infections, *Antimicrob. Agents Chemother. 49*, 2507-2509.

[31] Chia, B., Gong, Y., Bowie, J. H., Zuegg, J., and Cooper, M. A. (2011) Membrane binding and perturbation studies of the antimicrobial peptides caerin, citropin, and maculatin, *Pept. Sci. 96*, 147-157.

[32] Sreerama, N., and Woody, R. W. (2000) Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set, *Anal. Biochem. 287*, 252-260.

[33] Strandberg, E., and Ulrich, A. S. (2004) NMR Methods for Studying Membrane-Active Antimicrobial Peptides, *Concepts in Magnetic Resonance Part A 23A*, 89-120.

[34] Bax, A. (1989) Two-dimensional NMR and protein structure, *Annual review of biochemistry 58*, 223-256.

[35] Wishart, D. S., Bigam, C. G., Holm, A., Hodges, R. S., and Sykes, B. D. (1995) 1H, 13C and 15N Random Coil NMR Chemical Shifts of the Commom Amino Acids. I. Investigations of Nearest-Neighbor Effects, *J Biomol NMR 5*, 67-81.

[36] Schwarzinger, S., Kroon, G. J. A., Foss, T. R., Chung, J., Wright, P. E., and Dyson, H. J. (2001) Sequence-Dependent Correction of Random Coil NMR Chemical Shifts, *JACS 123*, 2970-2978.

[37] Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J. M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K. K., and Lemak, A. (2008) Consistent blind protein structure generation from NMR chemical shift data, *Proc. Natl. Acad. Sci. 105*, 4685-4690.

[38] Schweizer, F. (2009) Cationic amphiphilic peptides with cancer-selective toxicity, *Eur. J. Pharmacol. 625*, 190-194.

[39] Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures, *J. Appl. Crystallogr. 26*, 283-291.

[40] Clore, G. M., and Iwahara, J. (2009) Theory, Practice, and Applications of Paramagnetic Relaxation Enhancement for the Characterization of Transient Low-Population States of Biological Macromolecules and Their Complexes, *Chem. Rev. 109*, 4108-4139.

[41] Eaton, P., Fernandes, J. C., Pereira, E., Pintado, M. E., and Malcata, F. X. (2008) Atomic force microscopy study of the antibacterial effects of chitosans on Escherichia coli and Staphylococcus aureus, *Ultramicroscopy 108*, 1128-1134.

[42] Yao, X., Jericho, M., Pink, D., and Beveridge, T. (1999) Thickness and elasticity of gram-negative murein sacculi measured by atomic force microscopy, *J. Bacteriol. 181*, 6865-6875.

[43] Fernandes, J. C., Eaton, P., Gomes, A. M., Pintado, M. E., and Malcata, F. X. (2009) Study of the antibacterial effects of chitosans on Bacillus cereus (and its spores)

by atomic force microscopy imaging and nanoindentation, *Ultramicroscopy 109*, 854-860.

[44] Touhami, A., Nysten, B., and Dufrêne, Y. F. (2003) Nanoscale mapping of the elasticity of microbial cells by atomic force microscopy, *Langmuir 19,* 4539-4543.

[45] Makovitzki, A., Avrahami, D., and Shai, Y. (2006) Ultrashort antibacterial and antifungal lipopeptides, *Proc. Natl. Acad. Sci. 103*, 15997-16002.

[46] Spencer, E. M. (1971) Isoelectric heterogeneity of bovine plasma albumin, *J. Biol. Chem. 246*, 201-208.

[47] Chen, C. Z., Beck-Tan, N. C., Dhurjati, P., van Dyk, T. K., LaRossa, R. A., and Cooper, S. L. (2000) Quaternary ammonium functionalized poly (propylene imine) dendrimers as effective antimicrobials: Structure– activity studies, *Biomacromolecules 1*, 473-480.

[48] Thorsteinsson, T., Másson, M., Kristinsson, K. G., Hjálmarsdóttir, M. A., Hilmarsson, H., and Loftsson, T. (2003) Soft antimicrobial agents: synthesis and activity of labile environmentally friendly long chain quaternary ammonium compounds, *J. Med. Chem 46*, 4173-4181.

[49] Timofeeva, L., and Kleshcheva, N. (2011) Antimicrobial polymers: mechanism of action, factors of activity, and applications, *Appl. Microbiol. Biotechnol. 89*, 475-492.

[50] Gentilucci, L., De Marco, R., and Cerisoli, L. (2010) Chemical modifications designed to improve peptide stability: incorporation of non-natural amino acids, pseudo-peptide bonds, and cyclization, *Curr. Pharm. Des. 16*, 3185-3203.

[51] McCleskey, E., and Almers, W. (1985) The Ca channel in skeletal muscle is a large pore, *Proc. Natl. Acad. Sci. 82*, 7149-7153.

[52] Nelson, D. L., Lehninger, A. L., and Cox, M. M. (2008) *Lehninger principles of biochemistry*, Macmillan.

[53] Bagno, A., Menna, E., Mezzina, E., Scorrano, G., and Spinelli, D. (1998) Site of Protonation of Alkyl- and Arylhydrazines Probed by 14N, 15N, and 13C NMR Relaxation and Quantum Chemical Calculations, *The Journal of Physical Chemistry A 102*, 2888-2892.

[54] Rapp, C., Goldberger, E., Tishbi, N., and Kirshenbaum, R. (2014) Cation-pi interactions of methylated ammonium ions: a quantum mechanical study, *Proteins 82*, 1494-1502.

[55] Chan, D. I., Prenner, E. J., and Vogel, H. J. (2006) Tryptophan-and arginine-rich antimicrobial peptides: structures and mechanisms of action, *Biochimica et Biophysica Acta (BBA)-Biomembranes 1758*, 1184-1202.

[56] Jin, L., Bai, X., Luan, N., Yao, H., Zhang, Z., Liu, W., Chen, Y., Yan, X., Rong, M., and Lai, R. (2016) A Designed Tryptophan-and Lysine/Arginine-Rich Antimicrobial Peptide with Therapeutic Potential for Clinical Antibiotic-Resistant Candida albicans Vaginitis, *J. Med. Chem 59*, 1791-1799.

[57] Strøm, M. B., Rekdal, Ø., and Svendsen, J. S. (2002) Antimicrobial activity of short arginine- and tryptophan- rich peptides, *J. Pept. Sci. 8,* 431-437.

[58] Gabriel, G. J., Madkour, A. E., Dabkowski, J. M., Nelson, C. F., Nüsslein, K., and Tew, G. N. (2008) Synthetic mimic of antimicrobial peptide with nonmembrane-disrupting antibacterial properties, *Biomacromolecules 9,* 2980-2983.

[59] Svenson, J., Karstad, R., Flaten, G. E., Brandsdal, B.-O., Brandl, M., and Svendsen, J. S. (2009) Altered activity and physicochemical properties of short cationic antimicrobial peptides by incorporation of arginine analogues, *Molecular pharmaceutics 6,* 996-1005.

[60] Bahnsen, J. S., Franzyk, H., Sandberg-Schaal, A., and Nielsen, H. M. (2013) Antimicrobial and cell-penetrating properties of penetratin analogs: effect of sequence and secondary structure, *Biochimica et Biophysica Acta (BBA)-Biomembranes 1828*, 223-232.

[61] Schmidt, N. W., Tai, K. P., Kamdar, K., Mishra, A., Lai, G. H., Zhao, K., Ouellette, A. J., and Wong, G. C. L. (2012) Arginine in α-defensins differential effects on bactericidal activity correspond to geometry of membrane curvature generation and peptide-lipid phase behavior, *J. Biol. Chem. 287*, 21866-21872.

[62] Locock, K. E. S., Michl, T. D., Valentin, J. D. P., Vasilev, K., Hayball, J. D., Qu, Y., Traven, A., Griesser, H. J., Meagher, L., and Haeussler, M. (2013) Guanylated polymethacrylates: a class of potent antimicrobial polymers with low hemolytic activity, *Biomacromolecules 14,* 4021-4031.

[63] Vedel, L., Bonke, G., Foged, C., Ziegler, H., Franzyk, H., Jaroszewski, J. W., and Olsen, C. A. (2007) Antiplasmodial and Prehemolytic Activities of α-Peptide–β- Peptoid Chimeras, *ChemBioChem 8*, 1781-1784.

[64] Andreev, K., Bianchi, C., Laursen, J. S., Citterio, L., Hein-Kristensen, L., Gram, L., Kuzmenko, I., Olsen, C. A., and Gidalevitz, D. (2014) Guanidino groups greatly enhance the action of antimicrobial peptidomimetics against bacterial cytoplasmic membranes, *Biochimica et Biophysica Acta (BBA)-Biomembranes 1838*, 2492-2502.

[65] Strøm, M. B., Haug, B. E., Skar, M. L., Stensen, W., Stiberg, T., and Svendsen, J. S. (2003) The pharmacophore of short cationic antibacterial peptides, *J. Med. Chem 46*, 1567-1570.

[66] Schibli, D. J., Nguyen, L. T., Kernaghan, S. D., Rekdal, Ø., and Vogel, H. J. (2006) Structure-function analysis of tritrpticin analogs: potential relationships between antimicrobial activities, model membrane interactions, and their micelle-bound NMR structures, *Biophysical journal 91*, 4413-4426.

[67] Nguyen, L. T., de Boer, L., Zaat, S. A. J., and Vogel, H. J. (2011) Investigating the cationic side chains of the antimicrobial peptide tritrpticin: hydrogen bonding properties govern its membrane-disruptive activities, *Biochimica et Biophysica Acta (BBA)-Biomembranes 1808*, 2297-2303.

[68] Chen, P.-W., Shyu, C.-L., and Mao, F. C. (2003) Antibacterial activity of short hydrophobic and basic-rich peptides, *American journal of veterinary research 64*, 1088-1092.

[69] Mitchell, D. J., Steinman, L., Kim, D. T., Fathman, C. G., and Rothbard, J. B. (2000) Polyarginine enters cells more efficiently than other polycationic homopolymers, *The Journal of Peptide Research 56*, 318-325.

[70] Stanzl, E. G., Trantow, B. M., Vargas, J. R., and Wender, P. A. (2013) Fifteen years of cell-penetrating, guanidinium-rich molecular transporters: basic science, research tools, and clinical applications, *Accounts of chemical research 46*, 2944-2954.

[71] Nakase, I., Okumura, S., Katayama, S., Hirose, H., Pujals, S., Yamaguchi, H., Arakawa, S., Shimizu, S., and Futaki, S. (2012) Transformation of an antimicrobial peptide into a plasma membrane-permeable, mitochondria-targeted peptide via the substitution of lysine with arginine, *Chem. Commun. 48*, 11097-11099.

[72] Ambroggio, E. E., Separovic, F., Bowie, J. H., Fidelio, G. D., and Bagatolli, L. A. (2005) Direct visualization of membrane leakage induced by the antibiotic peptides: maculatin, citropin, and aurein, *Biophysical journal 89*, 1874-1881.

[73] Haug, B. E., and Svendsen, J. S. (2001) The role of tryptophan in the antibacterial activity of a 15- residue bovine lactoferricin peptide, *J. Pept. Sci. 7*, 190-196.

[74] Chen, Y., Guarnieri, M. T., Vasil, A. I., Vasil, M. L., Mant, C. T., and Hodges, R. S. (2007) Role of peptide hydrophobicity in the mechanism of action of alpha-helical antimicrobial peptides, *Antimicrob Agents Chemother 51*, 1398-1406.

[75] Chakrabartty, A., Doig, A. J., and Baldwin, R. L. (1993) Helix Capping Propensities in Peptides Parallel Those in Proteins, *PNAS 90*, 11332-11336.

[76] Kang, L., Moriarty, G. M., Woods, L. A., Ashcroft, A. E., Radford, S. E., and Baum, J. (2012) N-terminal acetylation of alpha-synuclein induces increased transient helical propensity and decreased aggregation rates in the intrinsically disordered monomer, *Protein Sci 21*, 911-917.

[77] Trexler, A. J., and Rhoades, E. (2012) N-Terminal acetylation is critical for forming alpha-helical oligomer of alpha-synuclein, *Protein Sci 21*, 601-605.

[78] Salditt, T., Li, C., and Spaar, A. (2006) Structure of antimicrobial peptides and lipid membranes probed by interface-sensitive X-ray scattering, *Biochimica et Biophysica Acta (BBA)-Biomembranes 1758*, 1483-1498.

[79] Giuliani, A., Pirri, G., Bozzi, A., Di Giulio, A., Aschi, M., and Rinaldi, A. (2008) Antimicrobial peptides: natural templates for synthetic membrane-active compounds, *Cell. Mol. Life Sci. 65*, 2450-2460.

[80] Wang, G., Hanke, M. L., Mishra, B., Lushnikova, T., Heim, C. E., Chittezham Thomas, V., Bayles, K. W., and Kielian, T. (2014) Transformation of human cathelicidin LL-37 into selective, stable, and potent antimicrobial compounds, *ACS Chem. Biol. 9*, 1997-2002.

[81] Stokes, J. M., MacNair, C. R., Ilyas, B., French, S., Côté, J.-P., Bouwman, C., Farha, M. A., Sieron, A. O., Whitfield, C., and Coombes, B. K. (2017) Pentamidine sensitizes Gram-negative pathogens to antibiotics and overcomes acquired colistin resistance, *Nature microbiology 2*, 17028.

[82] Byfield, F. J., Aranda-Espinoza, H., Romanenko, V. G., Rothblat, G. H., and Levitan, I. (2004) Cholesterol depletion increases membrane stiffness of aortic endothelial cells, *Biophysical journal 87*, 3336-3343.

[83] Lundbaek, J. A., Birn, P., Girshman, J., Hansen, A. J., and Andersen, O. S. (1996) Membrane Stiffness and Channel Function, *Biochemistry 35*, 3825-3830.

[84] Bayer, A. S., Prasad, R., Chandra, J., Koul, A., Smriti, M., Varma, A., Skurray, R. A., Firth, N., Brown, M. H., Koo, S., and Yeaman, M. R. (2000) In Vitro Resistance of Staphylococcus aureus to Thrombin-Induced Platelet Microbicidal Protein Is Associated with Alterations in Cytoplasmic Membrane Fluidity, *Infection and Immunity 68*, 3548-3553.

[85] Rauch, C., Blanchard, A., Wood, E., Dillon, E., Wahl, M. L., and Harguindey, S. (2009) Cell Membranes, Cytosolic pH and Drug Transport in Cancer and MDR: Physics, Biochemistry and Molecular Biology, In *Multiple Drug Resistance* (Meszaros, A., and Balogh, G., Eds.), Nova Science Publishers.

[86] Jing, W., Demcoe, A. R., and Vogel, H. J. (2003) Conformation of a bactericidal domain of puroindoline a: structure and mechanism of action of a 13-residue antimicrobial peptide, *J. Bacteriol. 185*, 4938-4947.

[87] Cirioni, O., Giacometti, A., Ghiselli, R., Kamysz, W., Orlando, F., Mocchegiani, F., Silvestri, C., Licci, A., Chiodi, L., and Łukasiak, J. (2006) Citropin 1.1-treated central venous catheters improve the efficacy of hydrophobic antibiotics in the treatment of experimental staphylococcal catheter-related infection, *Peptides 27*, 1210-1216.

[88] Ghiselli, R., Silvestri, C., Cirioni, O., Kamysz, W., Orlando, F., Calcinari, A., Kamysz, E., Casteletti, S., Rimini, M., and Tocchini, M. (2011) Protective effect of citropin 1.1 and tazobactam-piperacillin against oxidative damage and lethality in mice models of Gram-negative sepsis, *J. Surg. Res. 171*, 726-733.

[89] Friedrich, C., Scott, M. G., Karunaratne, N., Yan, H., and Hancock, R. E. (1999) Salt-resistant alpha-helical cationic antimicrobial peptides, *Antimicrob. Agents Chemother. 43*, 1542-1548.

[90] Marr, A. K., Gooderham, W. J., and Hancock, R. E. (2006) Antibacterial peptides for therapeutic use: obstacles and realistic outlook, *Curr. Opin. Pharmacol. 6*, 468-472.

4.7 Supplemental Figures

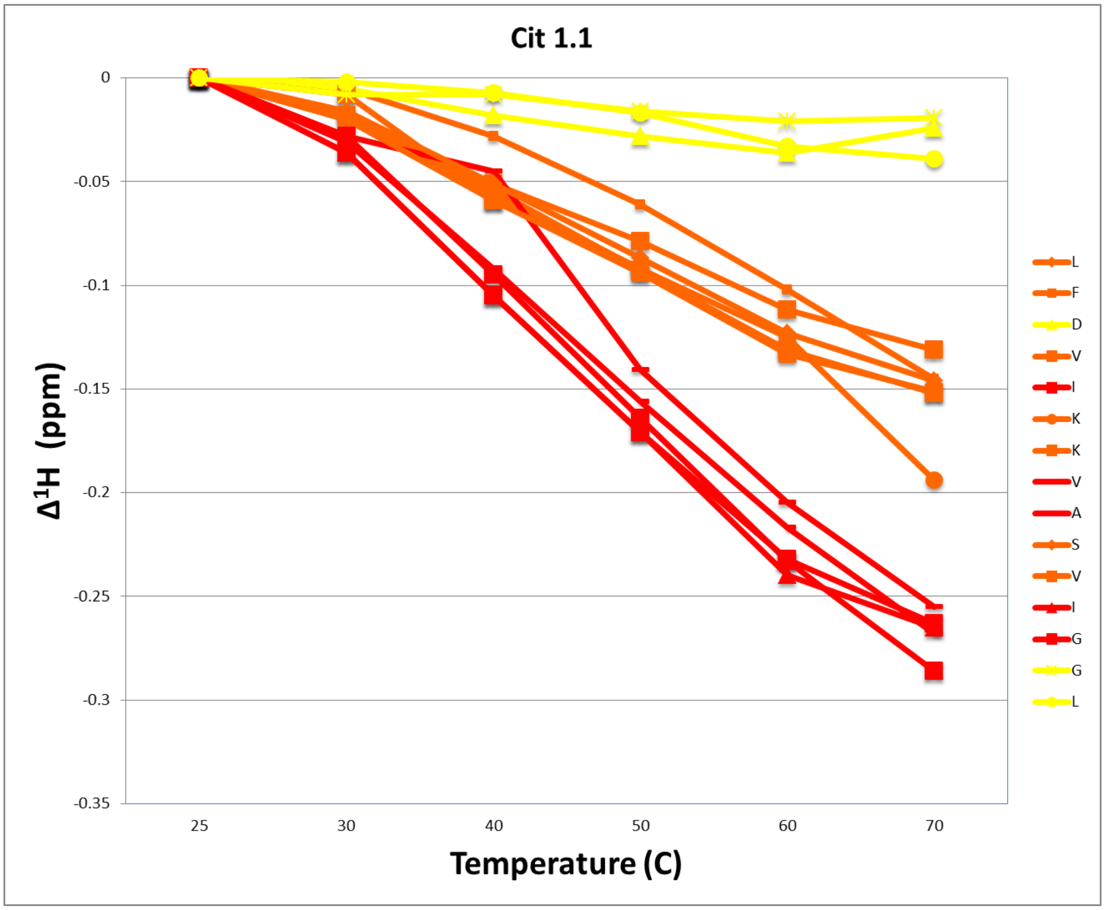


Figure S1. Temperature titration data for Cit 1.1. The change in the amide proton chemical shift (ΔH) was monitored as a function of temperature. Labels are found right of the graph and follow the peptide sequence in order. The largest changes in chemical shifts correspond to stable amino acids becoming disordered at elevated temperature. Coloring corresponds to binned amide proton temperature titration data according to the ΔH ranges 0 to -0.05 (minor chemical shift, yellow), -0.05 to -0.25 (intermediate chemical shift, orange), and -0.25 to -0.45 (major chemical shift, red).

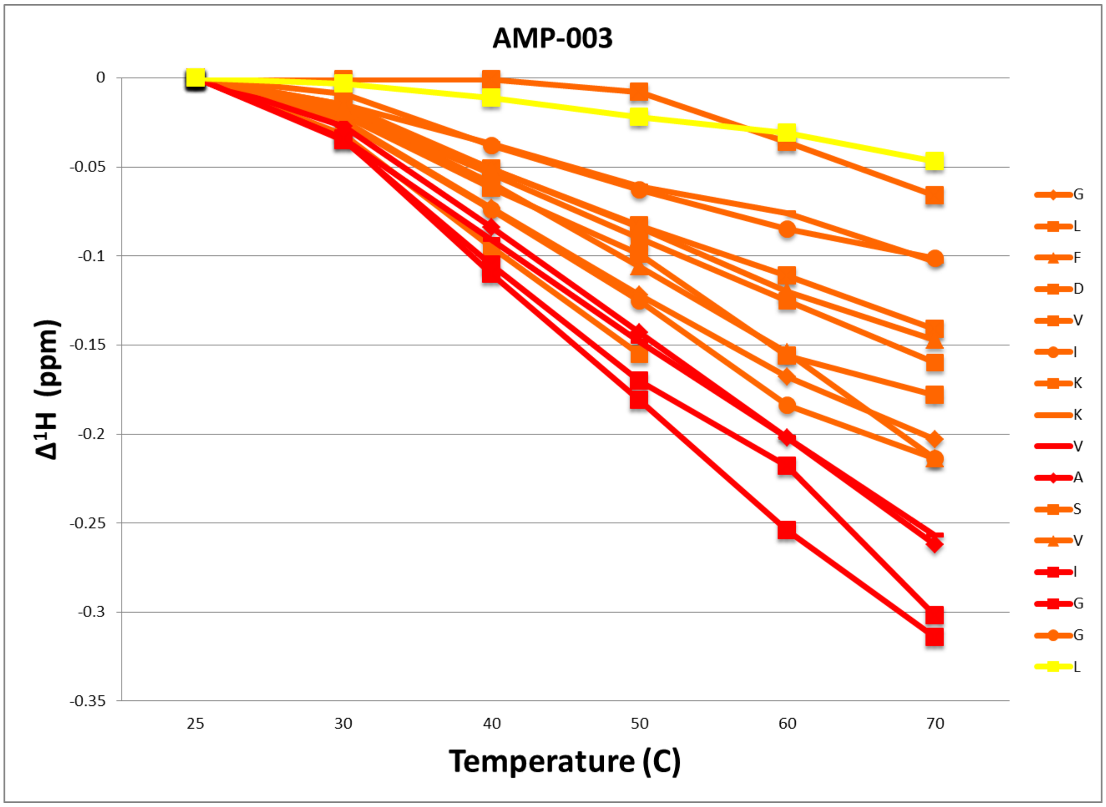


Figure S2. Temperature titration data for AMP-003. The change in the amide proton chemical shift (ΔH) was monitored as a function of temperature. Labels are found right of the graph and follow the peptide sequence in order. The largest changes in chemical shifts correspond to stable amino acids becoming disordered at elevated temperature. Coloring corresponds to binned amide proton temperature titration data according to the ΔH ranges 0 to -0.05 (minor chemical shift, yellow), -0.05 to -0.25 (intermediate chemical shift, orange), and -0.25 to -0.45 (major chemical shift, red).

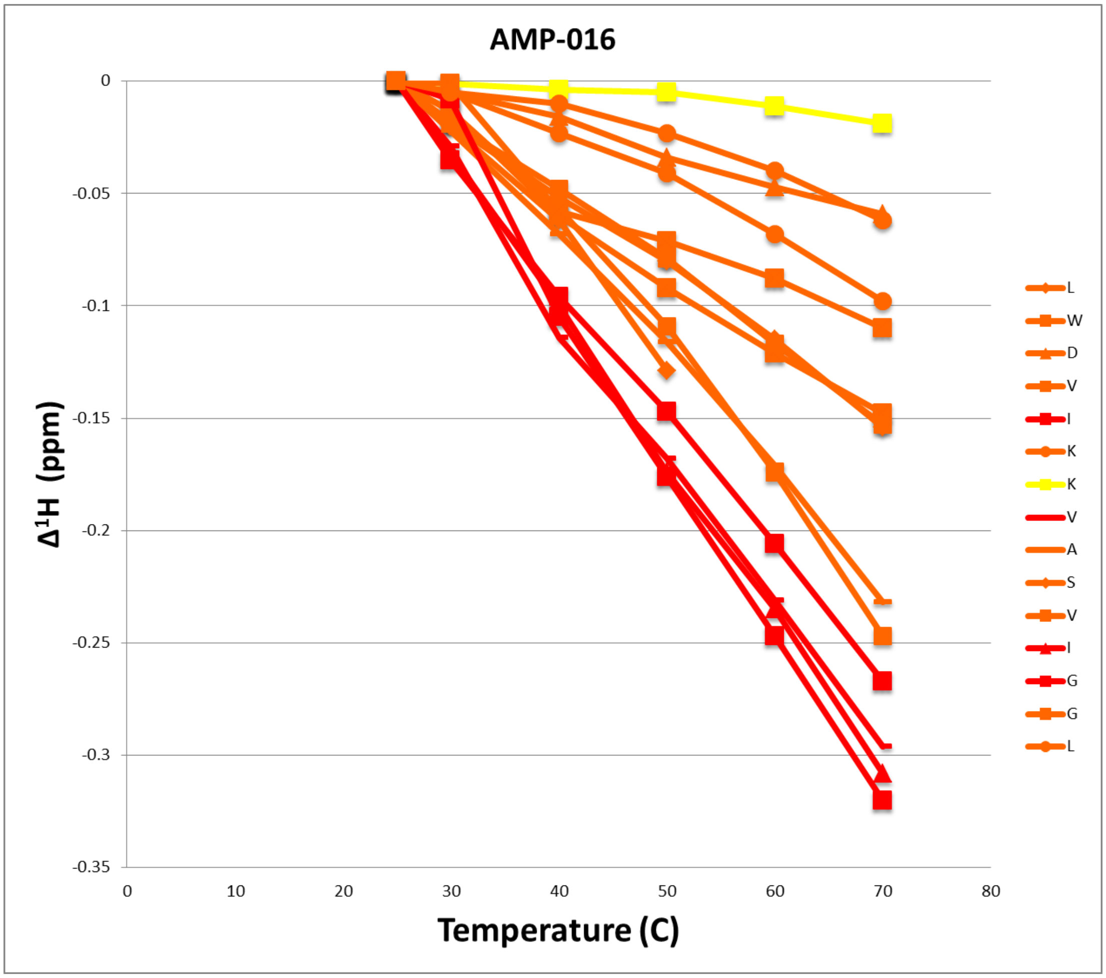


Figure S3: Temperature titration data for AMP-016. The change in the amide proton chemical shift (ΔH) was monitored as a function of temperature. Labels are found right of the graph and follow the peptide sequence in order. The largest changes in chemical shifts correspond to stable amino acids becoming disordered at elevated temperature. Coloring corresponds to binned amide proton temperature titration data according to the ΔH ranges 0 to -0.05 (minor chemical shift, yellow), -0.05 to -0.25 (intermediate chemical shift, orange), and -0.25 to -0.45 (major chemical shift, red).

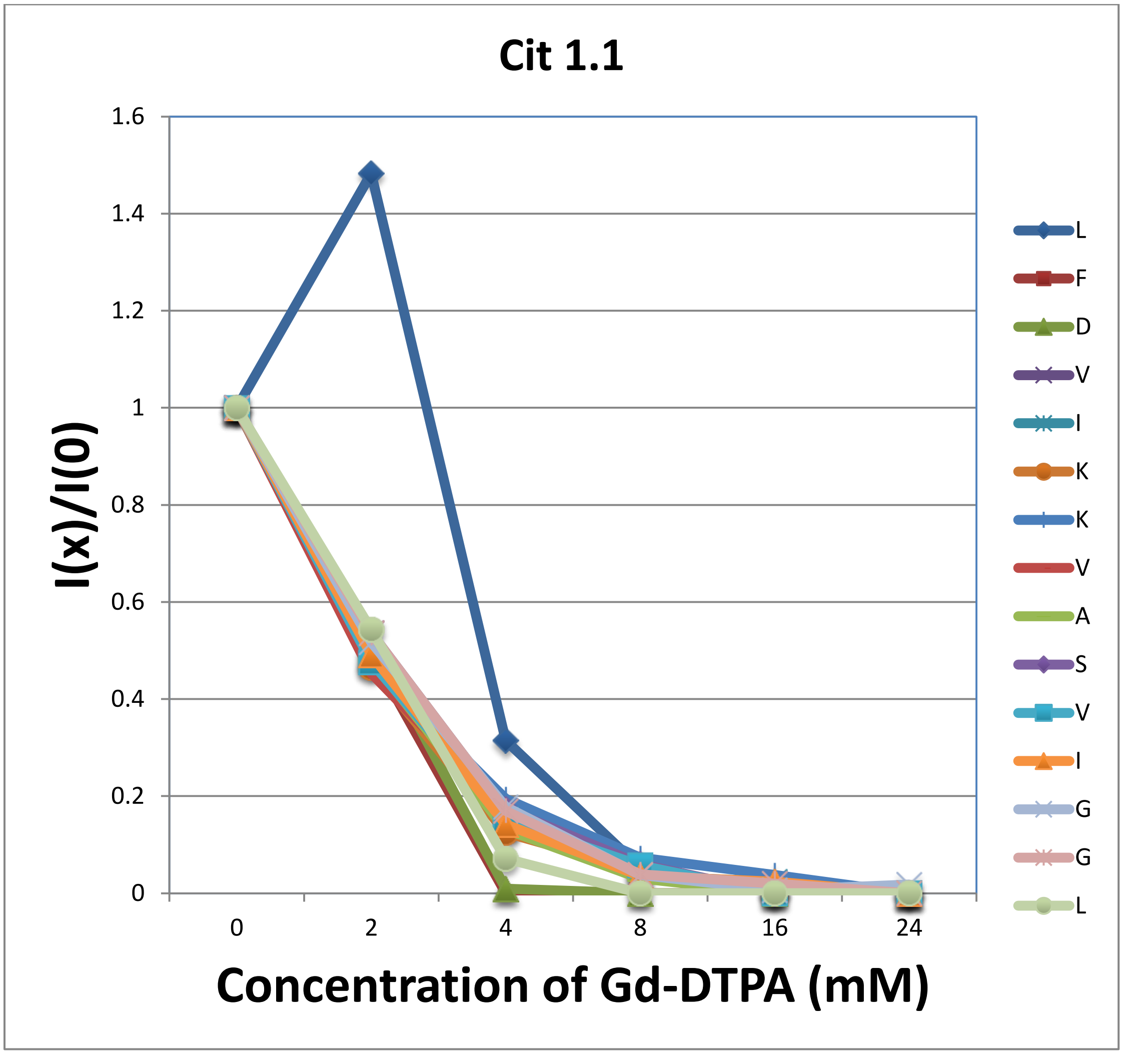


Figure S4: Gd-DTPA titration of Cit 1.1. The intensities of the amide proton peaks for each amino acid were monitored upon titration with Gd-DTPA at the listed concentrations. Intensities for each peak ($I_{(x)}$) are normalized to the corresponding peak in the initial spectrum ($I_{(0)}$).

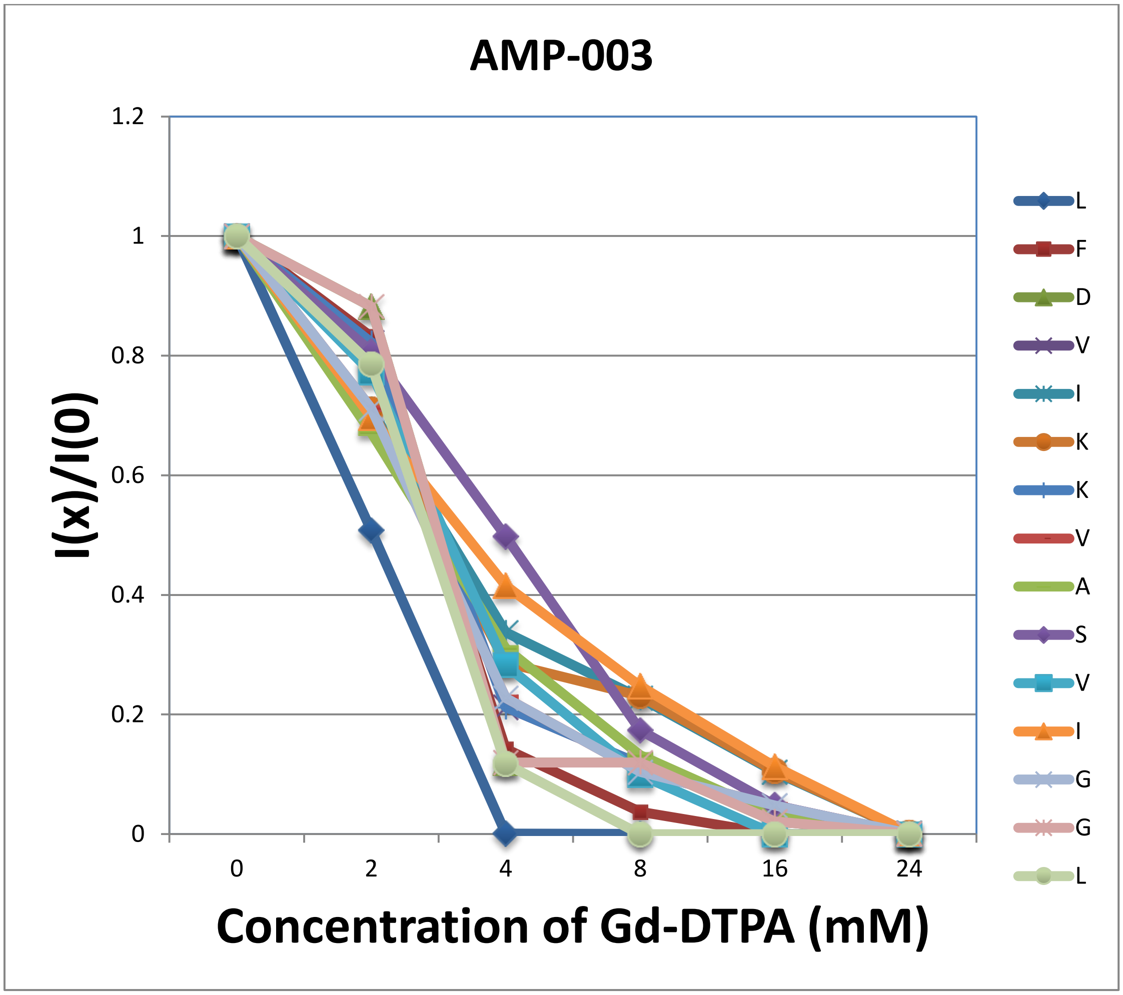


Figure S5: Gd-DTPA titration of AMP-003. The intensities of the amide proton peaks for each amino acid were monitored upon titration with Gd-DTPA at the listed concentrations. Intensities for each peak ($I_{(x)}$) are normalized to the corresponding peak in the initial spectrum ($I_{(0)}$).

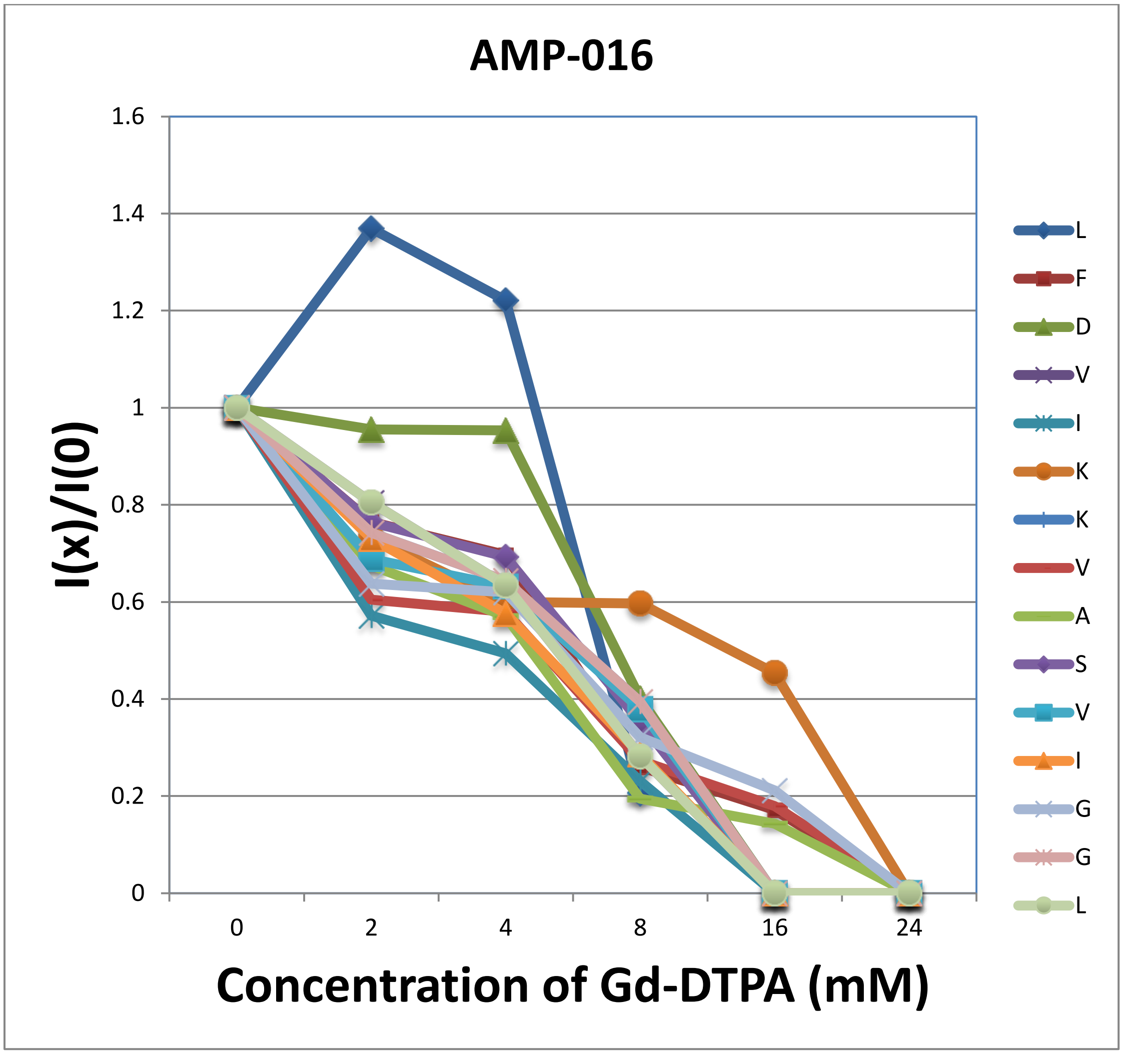


Figure S6: Gd-DTPA titration of AMP-016. The intensities of the amide proton peaks for each amino acid were monitored upon titration with Gd-DTPA at the listed concentrations. Intensities for each peak ($I_{(x)}$) are normalized to the corresponding peak in the initial spectrum ($I_{(0)}$).

Chapter 5

The NMR Solution Structure and Function of RPA3313: A Putative Ribosomal Transport Protein from *Rhodopseudomonas palustris*

5.1 Introduction

Rhodopseudomonas palustris is a unique organism known for its metabolic diversity and extensive distribution throughout the environment.[1] It has the ability to grow under four distinct modes of metabolism (photoautotrophic, photoheterotrophic, chemoautotrophic, and chemoheterotrophic) on a wide assortment of carbon sources. *R. palustris* is typically found in soil and freshwater sources, but has also been discovered in swine waste and coastal sediments.[2] As a purple non-sulfur photosynthetic bacterium, it belongs to the *alphaproteobacterium* order.[3] Within this order exists many species which are similarly metabolically versatile, yet there are clear phylogenetic differences. In fact, based on 16S rRNA sequencing the inherent divergence in the order is not based on phototrophic ability but rather demonstrates a mixing of phototrophs and non-phototrophs.[3] *R. palustris* is of particular biotechnological interest because it utilizes aromatic hydrocarbons as a carbon source under both aerobic and anaerobic growth conditions. Also, *R. palustris* fixes more nitrogen when grown on aromatic hydrocarbons relative to aliphatic substrates.[4] Furthermore, growth of the organism can occur on aromatic substrates containing a range of functional moieties. The combination of all of these factors makes *R. palustris* a model organism for bioremediation, energy production, and other biotechnological applications.[5-9]

In 2004, the genome of *R. palustris* was sequenced and published along with a prediction of general gene functional classes.[2] Approximately 15% of the genome is believed to be devoted solely to transport, which is surprising since prokaryotes usually commit only a third of this amount (5-10%) to transport.[10] Nevertheless, this greater commitment of *R. palustris* to transport is consistent with its observed metabolic diversity. A larger assortment of transport proteins would be necessary for *R. palustris* to readily adapt to various carbon and energy sources; or to proliferate under changing respiration or environmental states. Conversely, 29% of the genome has been tentatively labeled as hypothetical or of unknown function. An additional 8% of the genome is only annotated with a general function. A follow-up LC-ES-MS/MS proteomics analysis of *R. palustris* included a more detailed functional annotation based on protein sequence analysis.[11] But, the percentage of functionally uncharacterized or partially annotated proteins remained unchanged. Of particular note, the proteome of *R. palustris* was analyzed for each metabolic mode of growth. Thus, the relative expression rates of *R. palustris* proteins under each metabolic mode of growth are known.[11] The large fraction of unannotated or partially annotated *R. palustris* genes presents a significant obstacle for the further development of biotechnological applications and hinders additional biochemical studies.

The *R. palustris* protein RPA3313 (7.45 kDa, 70 amino acids) is a hypothetical protein targeted for structural elucidation by the Structural Genomics Consortium at the University of Toronto (http://www.thesgc.org/). RPA3313 is currently classified by UniProtKB[12] (Q6N4M4) as an uncharacterized protein. A BLAST search of the RPA3313 sequence reveals a group of 93 hypothetical yet conserved proteins (> 32%

identity) from *only* the *alphaproteobacterium* order (Figure 1). Unfortunately, minimal structural or functional information was obtained from the sequence analysis since no structures of homologous proteins are present in the Protein Data Bank (http://www.rcsb.org/).[14]

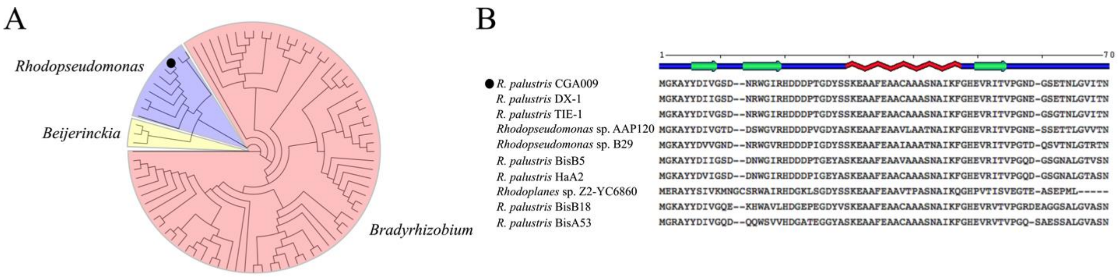

A
Rhodopseudomonas
Beijerinckia
Bradyrhizobium
B
1
70
● R. palustris CGA009 MGKAYYDIVGSD--NRWGIRHDDDPTGDYSSKEAAFEAACAAASNAIKFGHEVRITVPGND-GSETNLGVITN
R. palustris DX-1 MGKAYYDIIGSD--NRWGIRHDDDPTGDYSSKEAAFEAACAAASNAIKFGHEVRITVPGNE-GSETNLGVITN
R. palustris TIE-1 MGKAYYDIVGSD--NRWGIRHDDDPTGDYSSKEAAFEAACAAASNAIKFGHEVRITVPGND-GSGTNLGVITN
Rhodopseudomonas sp. AAP120 MGKAYYDIVGTD--DSWGVRHDDDPVGDYASKEAAFEAAVLAATNAIKFGHEVRITVPGNE-SSETTLGVVTN
Rhodopseudomonas sp. B29 MGKAYYDVVGND--NRWGVRHDDDPIGDYSSKEAAFEAAIAAATNAIKFGHEVRITVPGTD-QSVTNLGTRTN
R. palustris BisB5 MGKAYYDIIGSD--DNWGIRHDDDPTGEYASKEAAFEAAVAAASNAIKFGHEVRITVPGQD-GSGNALGTVSN
R. palustris HaA2 MGKAYYDVIGSD--DNWGIRHDDDPIGEYASKEAAFEAACAAASNAIKFGHEVRITVPGQD-GSGNALGTASN
Rhodoplanes sp. Z2-YC6860 MERAYYSIVKMNGCSRWAIRHDGKLSGDYSSKEAAFEAAVTPASNAIKQGHPVTISVEGTE-ASEPML-----
R. palustris BisB18 MGKAYYDIVGQE--KHWAVLHDGEPEGDYVSKEAAFEAACAAASNAIKFGHEVRVTVPGRDEAGGSALGVASN
R. palustris BisA53 MGRAYYDIVGQD--QQWSVVHDGATEGGYASKEAAFEAACAAASNAIKFGHEVRVTVPGQ-SAESSALGVASN

Figure 1. (A) A neighbor-join tree of the protein BLAST results of RPA3313 against non-redundant protein sequences. All of the sequence hits belong the *alphaproteobacteria* order and have identities >32%. The three colored groups are dominated by species belonging to the genera listed next to them. The tree highlights the fact that this protein is a member of an unannotated and a structurally uncharacterized class of proteins. (B) The BLAST sequence alignment of RPA3313 with the nine other *Rhodopseudomonas* proteins. The secondary structure is indicated above the sequence alignment and was generated with Polyview-2D.[13] RPA3313 is indicated in the phylogenetic tree and sequence alignment with a dot.

RPA3313 was identified in the previously reported *R. palustris* proteomics study[11] and was only observed to be expressed during photoautotrophic growth. Photoautotrophic organisms, such as *R. palustris*, sequester atmospheric CO_2 and convert it to energy rich carbon sources. Ribulose 1,5-bisphosphate carboxylase/oxygenase (RubisCO) are found in photoautotrophic organisms and are responsible for most of the organic carbon in the environment. As the most abundant protein in nature, RubisCO is found in plants, bacteria, and archaea in at least four molecular forms.[15] The genome of *R. palustris* contains multiple forms of RubisCO, which further contributes to its adaptability to diverse environmental conditions.[16] The combined adaptability and RubisCO activity of *R. palustris* may be beneficial to biotechnological applications involving bulk removal of CO_2 from the atmosphere. However, the photoautotrophic mode of metabolism in *R. palustris* and other *alphaproteobacteria* remains relatively unknown.[17]

Although the hypothetical protein RPA3313 has been experimentally verified as an expressed protein during photoautotrophic growth; the function and structure of RPA3313 still remains elusive. Structural approaches are a valuable alternative to obtaining a functional annotation when sequence similarity techniques fail and leaves a large class of functionally uncharacterized proteins.[18,19] Thus, obtaining an NMR solution structure for *R. palustris* protein RPA3313 is expected to provide a better understanding of its general biological role and also provide a putative structure and function for the 93 homologous proteins (Figure 1). RPA3313 forms a novel split ββαβ fold with a conserved ligand binding pocket. The NMR structure combined with a bioinformatics analysis and mass spectrometry proteomics suggest RPA3313 may assist in the transportation of substrates to or from the ribosome for further processing.

5.2 Methods

5.2.1 Protein expression and purification

Uniformly ^{15}N and ^{13}C labeled RPA3313 samples were prepared for NMR structural studies as follows. The target sequence for RPA3313 (70 amino acids with a 21 amino acid histidine tag for purification, MGSSHHHHHHSSGRENLYFQG) was expressed from a pRI952 with *glyT* construct transformed into BL21(DE3) cells.[20] Cells were grown in Luria-Bertani (LB) media at 37°C until an approximate optical density (OD_{600}) of 0.6 and then spun down and transferred to M9 minimal media at 37°C containing 4% U-^{13}C glucose and 1% U-^{15}N NH_4Cl. Expression of RPA3313 was induced after one hour of equilibration in the M9 media with isopropyl β-D-1-thiogalactopyranoside (IPTG). Cell lysates were collected 4 hours after induction with IPTG and purified with a Co^{2+} affinity column (HisPur Cobalt Resin, Thermo Scientific). Sample homogeneity was assessed by SDS-PAGE. ESI-MS and size exclusion chromatography was used to confirm the monomeric solution state and exact mass of RPA3313 (Supplemental Figure S1). The protein sample was stored in an NMR sample tube in 18 mM 2-(4-morpholino)ethanesulfonic acid (MES) buffer with 0.01% sodium azide, 80 mM sodium chloride and 10% D_2O at a pH of 5.6 (uncorrected).

5.2.2 NMR structure determination

All NMR experiments were collected with non-uniform sampling at 20% sparsity using a Poisson-gap schedule[21] at 298K on a 700 MHz Bruker Avance III spectrometer equipped with a 5 mm QCI-P probe with cryogenically cooled carbon and proton channels. Backbone and side-chain assignments were completed using the standard triple resonance approach consisting of the following experiments: ^{1}H-^{15}N HSQC, ^{1}H-^{13}C HSQC, HNCO, HN(CA)CO, HNCA, HN(CO)CA, CCANH, CBCA(CO)NH, HNHA, HBHA(CO)NH, CC(CO)NH, HCC(CO)NH, HCCH-COSY, and HCCH-TOCSY.[22,23] Identification of nuclear Overhauser effects (NOEs) was accomplished with ^{15}N-edited NOESY-HSQC and ^{13}C-edited NOESY-HSQC experiments using a mixing time of 150 ms. The resulting data was reconstructed using multidimensional decomposition (MDD) and processed in TopSpin 3.2 followed by evaluation in CCPNMR Analysis.[24] Initial model generation according to backbone chemical shifts was undertaken using CS-ROSETTA[25-27] on the open webserver at the BMRB (https://csrosetta.bmrb.wisc.edu/). The CS-ROSETTA software was only used for the creation of an initial model for RPA3313. CS-ROSETTA was not used to further refine the RPA3313 ensemble.

XPLOR-NIH version 2.37 was used to refine the initial model of target RPA3313.[28,29] Briefly, the refinement involved 912 manually assigned NOE distance restraints, 66 hydrogen bond distance restraints, 30 $^{3}J_{\mathrm{NH}\alpha}$ coupling constants, 128 ^{13}Cα/^{13}Cβ chemical shifts, and 102 predicted dihedral angle restraints from TALOS+.[30] 1000 total structures were generated during the XPLOR-NIH structure refinement and the 20 lowest energy structures were subsequently subjected to water refinement according to the RECOORD conventions. The coordinate average structure for the water-refined models was further subjected to the same explicit water refinement method for energy minimization. The water-refined ensemble and average structure for target RPA3313 was analyzed with the PSVS software suite, which is comprised of commonly used structural validation packages.[31-35] UCSF Chimera was used for the structural visualization and surface representation of RPA3313.[36]

5.2.3 Chemical crosslinking and in-gel digestion

Approximately 3 hours following the induction of RPA3313 in Lysogeny broth (LB) media, the *E. coli* culture was pelleted and resuspended in crosslinking buffer (1% paraformaldehyde, 1X PBS, pH 8, 37°C). Crosslinking was allowed to proceed for 15 minutes before quenching with 1.25 M glycine. The *E. coli* cells were lysed by sonication and RPA3313 with crosslinked binding partners was purified using the RPA3313 histidine tag and a Co^{2+} affinity column as described above in the protein purification section. The sample preparation procedure also efficiently removes the formaldehyde crosslinking. The purified proteins were then visualized by SDS-PAGE. Protein bands were excised before submission to the Nebraska Center for Mass Spectrometry for MS/MS analysis.

Rhodopseudomonas palustris (ATCC) was propagated in a filled flask of 500 mL of 112 medium at 30°C for several days until reaching stationary growth. Bacterial growth was red in color indicating that photoautotrophic had occurred. The culture was pelleted by centrifugation and resuspended in water prior to lysis by sonication. Extracted proteins were frozen and lyophilized overnight. Approximately 2 mg of pure RPA3313 was added to the *R. palustris* protein extract before the addition of crosslinking buffer. The crosslinking was performed similar to the *E. coli* crosslinking above. RPA3313 with crosslinked binding partners was purified using the RPA3313 histidine tag and a Co^{2+} affinity column as described above in the protein purification section.

5.2.4 MS proteomics

Protein bands separated with SDS–PAGE were digested *in situ* using a slightly modified version of a published method.[37] Briefly, the samples were washed with 100 mM ammonium bicarbonate, reduced with 10 mM DTT, alkylated with 55 mM iodoacetamide, washed twice with 100 mM ammonium bicarbonate, and digested *in situ* with 10 ng/µL trypsin (Promega, Madison, WI, USA). Peptides were extracted with two 60 µL aliquots of 1:1 acetonitrile:water containing 1% formic acid. The extracts were dried down using a SpeedVac and then reconstituted into 15 uL of water + 0.1% formic acid. Four microliters of the extract solution was injected onto a trapping column (300 µm × 1 mm) in line with a 75 µm × 15 cm C18 reversed phase LC column (Waters, Milford, MA, USA). Peptides were eluted from the column using a water + 0.1% formic acid (A)/ acetonitrile + 0.1% formic acid (B) gradient with a flow rate of 500 nL/min. The gradient was developed with the following time profile: 0 min, 5% B; 5 min, 5% B; 35 min, 35% B; 40 min, 45% B; 42 min, 60% B; 45 min, 90% B; 48 min, 90% B; and 50 min, 5% B.

The eluting peptides were analyzed using a Synapt G2S Q-TOF tandem mass spectrometer (Waters, Milford, MA, USA) with electrospray ionization. Analyses were performed using data-dependent acquisition (DDA) with the following parameters: 0.7 sec. survey scan (380–200 Da) followed by up to four MS/MS acquisitions (50-2000 Da).

The instrument was operated at a mass resolution of 18000. The instrument was calibrated using a solution of NaI in 1:1 water:acetonitrile. The MS/MS data were processed using Masslynx software (Micromass, Milford, MA, USA) to produce peak lists for database searching. Mascot (Matrix Science Ltd, London, UK) was used as the search engine. Data were searched against the National Centre for Biotechnology Information (NCBI) non-redundant database. The following search parameters were used: mass accuracy 20 ppm, enzyme specificity trypsin, fixed modification carboxyamidomethylcysteine (CAM), variable modification oxidized methionine. Protein identifications were based on random probability scores with a minimum value of 25.

5.2.5 Bioinformatics Analyses

The RPA3313 sequence (excluding the 21 residue histidine tag) and the NMR structure were submitted to the ConSurf webserver to identify evolutionary conserved residues.[38-41] Structural comparison was done with PDBeFold[42] and protein-protein interaction residues were predicted with cons-PPISP.[43,44] Results from the ConSurf and cons-PPISP analyses were mapped onto the surface of the protein with UCSF Chimera.[45] Surface hydrophobicity was also calculated within Chimera. The BLAST hits were visualized with the neighbor-join algorithm using Dendroscope.[46]

5.3 Results and Discussion

5.3.1 Solution Structure of *R. palustris* protein RPA3313

Backbone and side-chain resonance assignments for RPA3313 were made for the 68 assignable residues excluding the N-terminus histidine tag (Figure 2). The NMR assignments are nearly complete with 68 of 68 N, 68 of 68 HN, 68 of 68 Cα, 76 of 76 Hα, 60 of 60 Cβ, 93 of 97 Hβ, 33 of 55 Cγ, 57 of 60 Hγ, 13 of 28 Cδ, 32 of 39 Hδ, 3 of 12 Cε, 16 of 21 Hε, 0 of 9 Cζ, and 4 of 7 Hζ. The monomeric solution structure of RPA3313 was calculated using 912 distance restraints, 102 angle restraints, 30 $^{3}J_{NH\alpha}$ coupling constants, 128 $^{13}C\alpha/^{13}C\beta$ chemical shifts, and an initial model generated using CS-ROSETTA.[25-27] During structure generation, 1000 structures were initially created and the 20 lowest energy models were selected for further water refinement. A coordinate average of the 20 water-refined structures was subjected to water refinement for additional minimization. The water refined ensemble structures did not contain any distance violations >0.5 Å or dihedral angle violations >5°. Also, the NMR data agrees well with the calculated structures since the RMSD of the backbone secondary structure residues is 0.70±0.07 Å and the RMSD for heavy atoms is 1.2±0.12 Å. Complete structural statistics for the RPA3313 NMR structures are listed in Table 1. Chemical shift assignments have been submitted to the BMRB as entry 30070 and coordinate files have been uploaded to the PDB as entry 5JN6.

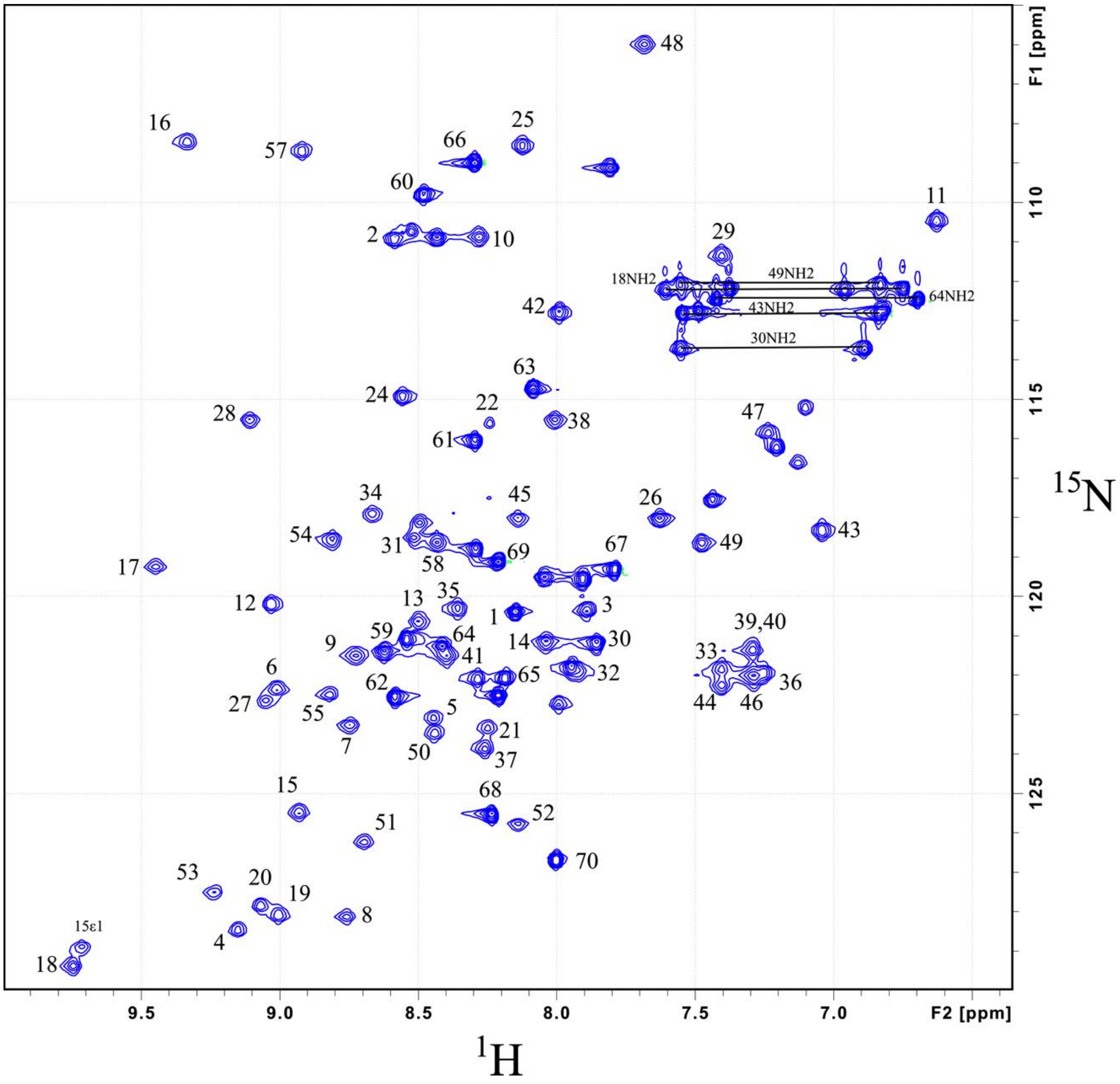


Figure 2. A 2D ^{1}H-^{15}N HSQC NMR spectrum of RPA3313 with peaks labeled according to their respective residue numbers or side-chain identification. Unlabeled peaks correspond to the 6x His tag, which was not included in the structure generation or analysis.

The overall quality of the RPA3313 NMR structure was assessed with the PSVS software suite (Table 2). All but one residue was located in the most favored region (98.3%) of the

Ramachandran plot with the remaining residue in the allowed region (1.7%). PROCHECK further supported the dihedral angle quality of the RPA3313 NMR structure with Z-scores of 0.12 and -0.71 for ϕ, ψ angles and all angles, respectively. Overall model quality was further assessed with ProsaII that produced a good Z-score of -0.58. An excellent quality score of -0.45 was also obtained from a MolProbity analysis, which evaluates atom clashes in the 3D structure. The ProsaII, PROCHECK and MolProbity scores are consistent with other high-quality NMR structures deposited in the PDB. Conversely, the Verify3D structure assessment yielded only a modest score of -1.93, but the analysis is still within an acceptable range compared to other NMR structures. Verify3D measures agreement between the 3D structure and the primary sequence. The novel fold for the RPA3313 structure may be a factor in the relatively low Verify3D score.

Table 1. Structure Calculation Statistics[α]

rmsd for distance restraints (experimental) (Å)	<SA>	$(\overline{SA})r$
all (912)	0.048±0.004	0.039
inter-residue sequential (|i-j| = 1) (269)	0.019±0.009	0.002
inter-residue short-range (1 < |i-j| < 5) (238)	0.075±0.006	0.067
inter-residue long-range (|i-j| ≥ 5) (83)	0.070±0.019	0.054
intraresidue (256)	0.007±0.003	0.005
H-bonds (66)	0.022±0.007	0.024
rmsd for dihedral angle restraints (deg) (102)	0.654±0.032	0.626

rmsd for $^{3}J_{HN\alpha}$ restraints (Hz) (30)	0.515±0.050	0.569
rmsd (covalent geometry)		
bonds (Å)	0.007±0.000	0.008
angles (deg)	0.716±0.027	0.461
impropers (deg)	1.075±0.090	0.977
energy (kcal/mol)		
total	-1900.73±89.02	-2161.44
bond	29.05±3.24	30.99
angle	87.92±8.88	96.49
dihedral	3.58±1.76	2.43
impropers	49.58±8.67	39.10
van der Waals	-184.53±10.91	-191.49
NOE	63.30±10.65	41.50
$^{3}J_{HN\alpha}$	8.03±1.51	9.73
Cα and Cβ shifts	77.89±10.00	65.75
RMSD from mean (residues 2-21, 27-55) (Å)		
Backbone	0.70±0.07	
Heavy Atoms	1.20±0.12	

[α] <SA> represents the ensemble of the 20 water-refined simulated annealing structures. ($\overline{\text{SA}}$)r represents the water refined average of the ensemble.

Table 2. Structure Evaluation

PSVS Z-score (residues 6-54)	
Verify3D	-1.93

ProsaII (-ve)	-0.58
Procheck (ϕ and ψ)	0.12
Procheck (all)	-0.71
MolProbity	-0.45
Ramachandran Space (all residues)	
most favored regions	98.30%
allowed regions	1.70%
disallowed regions	0.00%

The structure of RPA3313 adopts a split ββαβ motif formed by 3 β-strands (β1-3) packed against an α-helix (Figure 3). There are no known structures of homologs to RPA3313 and a search against the PDB using PDBeFold did not yield any significant results. Although the ββαβ motif is ubiquitous, when the RPA3313 structure is compared to proteins with similar motifs there is either a different handedness, or the orientation of the β–sheet along the α-helix is askew. This is not uncommon for this type of fold, as the β–sheet typically curls or flexes to cover the hydrophobic core of the protein.[47]

Starting at the N-terminus, the first 2 β-strands are formed antiparallel to one another and are connected by a β-hairpin turn. The initial residues at this terminus do not contribute to β1 and are disordered. At the beginning of β2, Trp15 creates significant bulk in the core of the protein near the β-hairpin. The indole side chain reaches from β2 toward the surface of the protein, which forces Gly10 to accommodate this structural perturbation. Both β1 and β2 have branched side chains forming the center of the protein. Connecting β2 to the α-helix is an extended loop region comprised mostly of negatively charged, polar residues. This loop outlines the top of a cavity formed with β1 and the N-terminus of the α-helix. An additional Tyr residue in the loop has its side chain in close proximity to Lys30, which marks the beginning of the α-helix. The length of the α-helix is approximately 17 residues and is terminated by Gly48. Alanine residues line the inside of the helix and polar residues, including one cysteine, create the solvent exposed surface. A γ-turn links the α-helix to β3, which runs parallel to β2. The side chain of Arg52 on β3 is angled toward the center of the β-sheet and creates a stacking interaction with Arg14 on β2 (Figure 4). This interaction is stabilized by Glu50, which may explain why the α-helix to β3 turn contains only 2 residues. The bottom of β3 is hydrophobic and consists of branched chain amino acids.

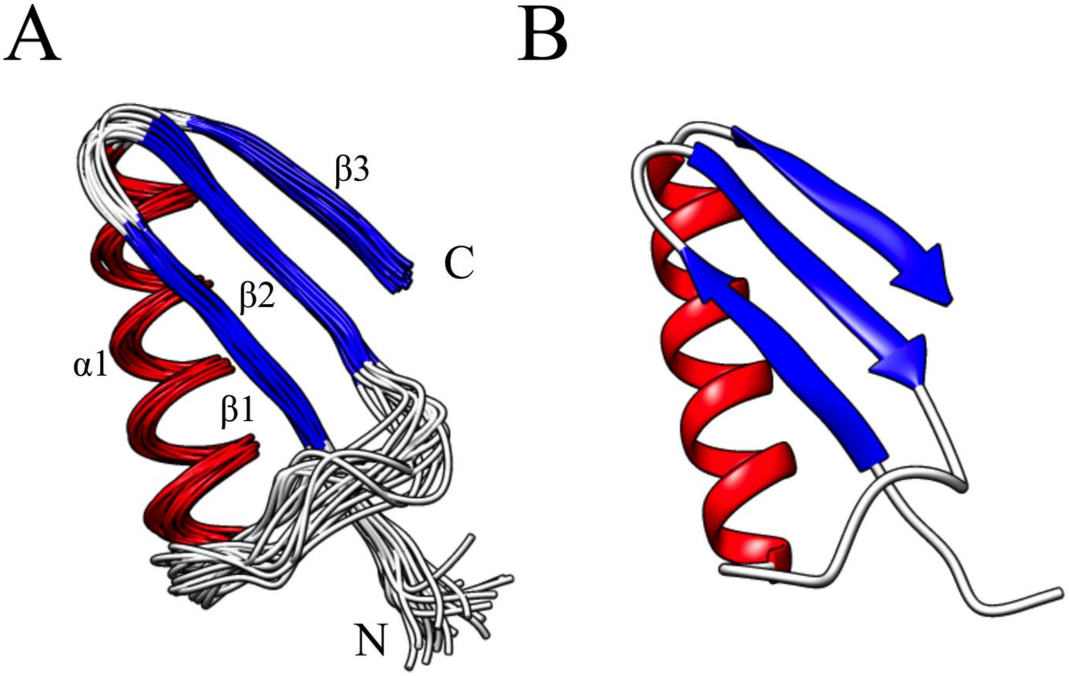


Figure 3. (A) An ensemble of the 20 best water-refined structures of RPA3313 for

residues 6-54 and (B) the coordinate average structure of the ensemble after water refinement. The structures are colored according to secondary structure: red for α-helix, blue for β-strand, and white for loops and disordered regions. The structural elements and the N- and C-terminus are labeled. The disordered C-terminus has been excluded for clarity.

Following the last β-strand is the disordered C-terminus. At approximately 15 residues in length, this tail is mostly unstructured except for a small α-helical propensity centered on Val67. Seemingly uninteresting at first, the disordered C-terminus probably has a significant physiological function. Disordered termini are known to serve in a broad range of roles such as protein-protein interaction sites, chaperones, and signal processing.[48] The proximity of the terminus to the large cavity on the surface of RPA3313 also suggests that it may potentially serve a role in activity (Figure 3b). Single-stranded DNA (ssDNA) binding proteins in prokaryotes maintain evolutionary conserved disordered C-termini that compete with the DNA binding site in order to exclude unwanted binders.[49] Although it is not known if RPA3313 binds ssDNA, the mechanism of the competition between the disordered tail and ligand remains a possibility. Also, many photosynthetic organisms possess globular proteins that have extended termini and are involved in a wide array of functions.[50] These extensions are highly variable and show little conservation between homologous species. However, they are necessary for host protein regulation and function. Since RPA3313 is expressed during photoautotrophic growth, it is possible that the disordered tail is involved in a light dependent mechanism.

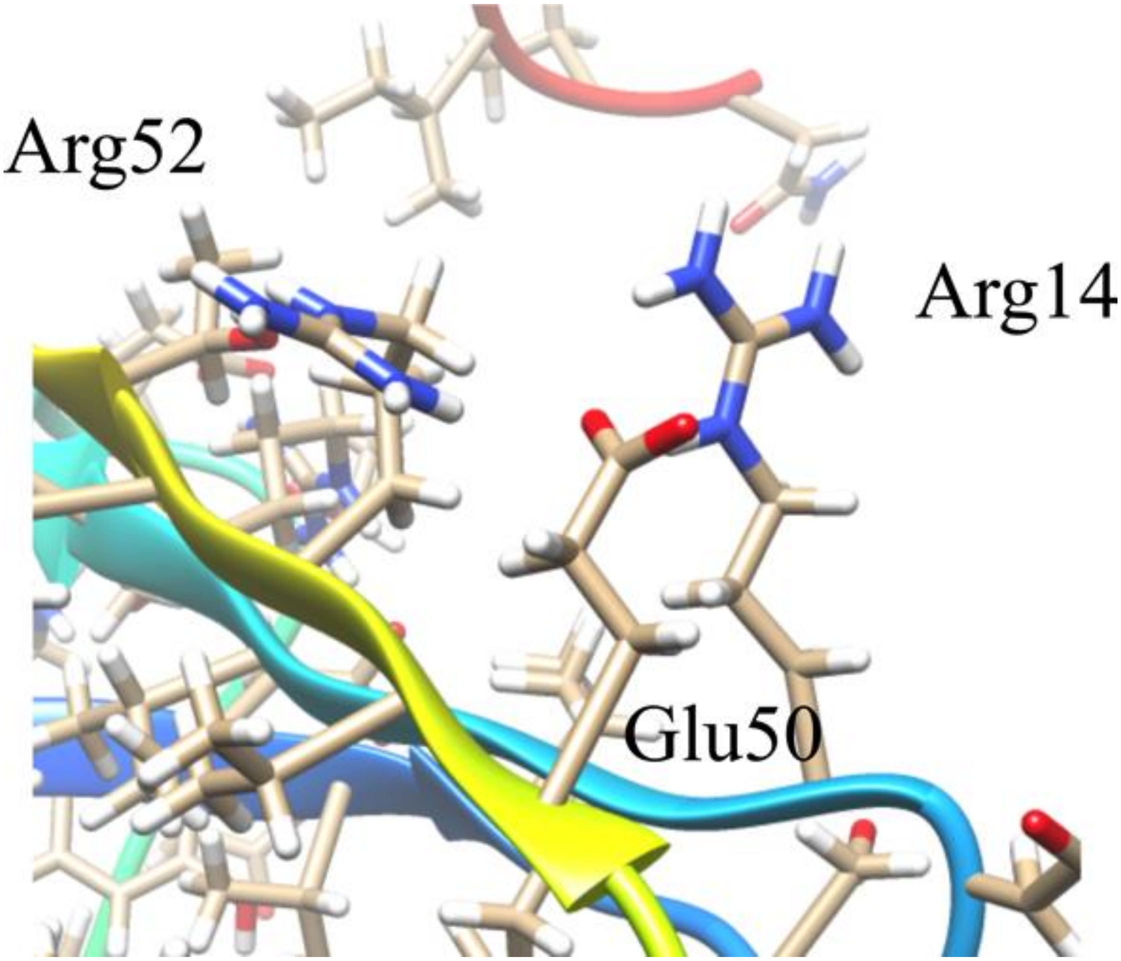


Figure 4. Arginine stacking interaction on the surface of RPA3313. Arginines 14 and 52 interact across the top of the β-sheet and are stabilized by glutamate 50. Each of the residues is evolutionarily stable indicating the occurrence of a conserved interaction.

5.3.2 Conserved Residue Analysis

The structure of RPA3313 was submitted to the ConSurf server for conserved residue analysis. ConSurf identifies and scores residue conservation based on a BLAST search and a subsequent multiple sequence alignment. Plotted on a surface representation of the RPA3313 NMR structure are the ConSurf scores, which range from 0 (blue) to 1 (red) with 1 signifying high conservation (Figure 5). Clearly visible is a conserved pocket between the extended loop and the top of β1. The deepest region of the cavity is defined by the peptide backbone of the α-helix and Tyr6. Conserved residues with side chains pointing into the pocket are Asp20, Tyr27, Lys30, and Phe34 (Figure 6). The Asp, Tyr, and Lys residues have the ability to form hydrogen bonds with a ligand. Additionally, Lys and Asp are possible metal coordinators and Tyr and Phe may be involved in π-π interactions with a ligand. Also conserved are small flexible residues Gly2, Ala4, and Gly25. Each residue is either at the top or the bottom of the pocket and likely contributes to important structure flexibility. These small residues would enable the protein to bend in order to accommodate a larger ligand, or to change the size of the entrance to the pocket based on other structural perturbations or modifications. Gly2 is doubly important as it follows the N-terminal start methionine. Small flexible residues trailing methionine enable truncation by an aminopeptidase and it is anticipated that the physiological form of RPA3313 lacks this initial methionine residue.[51] Distal to the pocket, the γ-turn between α-helix and β3 is also highly conserved. It is possible that this turn also acts like a hinge between the β-sheet and α-helix to allow the protein to adjust to a possible change in the hydrophobic core resulting from binding a ligand or a protein-protein interaction. The aforementioned stacked arginine residues (14, 52) are also moderately conserved.

The ConSurf score of approximately 0.5 indicate that this could be an evolutionary newer interaction or function. The stabilizing Glu50 residue has a higher conservation score (~0.7) than the arginines, but lower than the other residues in the γ-turn. This suggests that Glu50 may have a dual role in providing flexibility at the hinge in addition to stabilizing the arginine stacking interaction. Lastly, a highly conserved proline residue (56) exists at the end of β3. Proline is known to disrupt secondary structure formation and is most often found in disordered regions or turns. In this case, proline is acting as a terminator of a β-strand, which may assist in keeping the C-terminus residues in a disordered state. Furthermore, proline residues are associated with protein-protein interactions involving disordered protein regions.[52] In these situations the disordered tail or region adopts an induced fit upon interaction or binding.

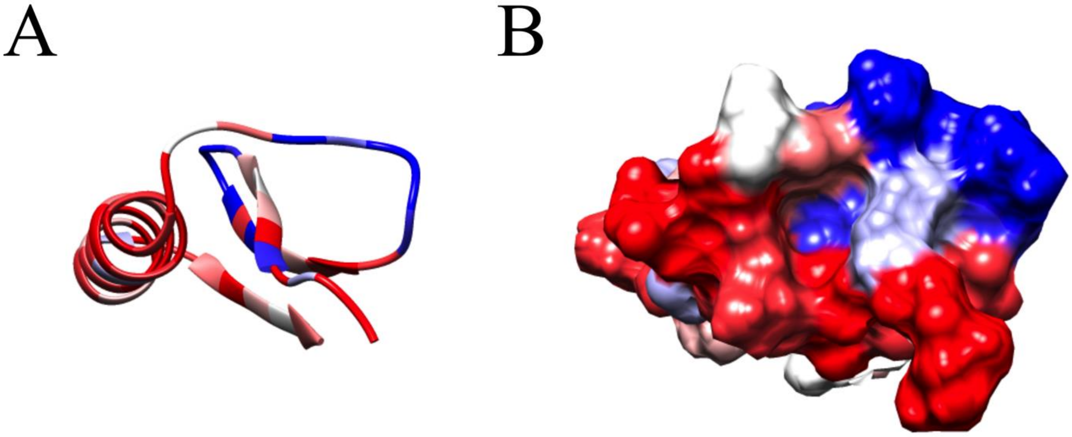


Figure 5. ConSurf Analysis. (A) ConSurf per residue scores for the structure RPA3313. Conserved residues are red and non-conserved residues are blue. (B) ConSurf scores mapped onto the molecular surface representation of RPA3313. A conserved pocket is

formed between the N-terminus of the α-helix and the first β-strand. Residues found to be less conserved are mostly found in the loop regions and the disordered tail (C-terminus, not shown).

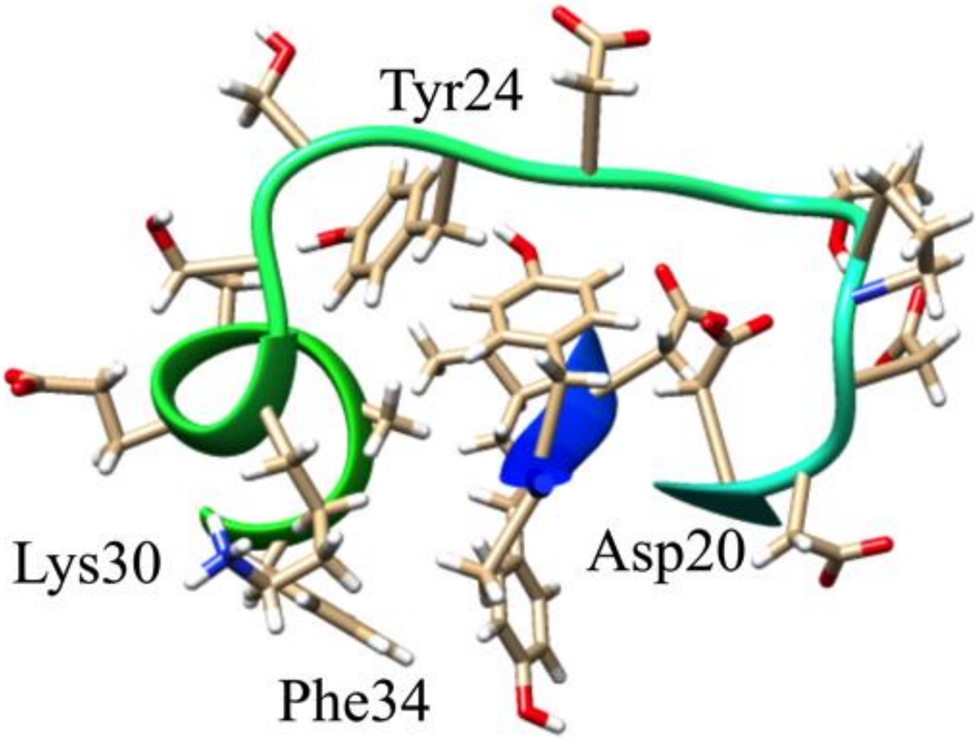


Figure 6. Surface pocket and possible ligand binding site on RPA3313. Shown are the residues with side chains that point into the pocket of the protein. The evolutionarily conserved residues are labeled. The remaining residues may also participate in the function of the pocket even though they are not highly conserved.

5.3.3 Protein-Protein Interaction Site Prediction

A further bioinformatics analysis of the structure of RPA3313 was carried out with cons-PPISP. The cons-PPISP server utilizes a neural network to predict position specific interaction sites on protein surfaces. Based on the output of cons-PPISP, it is possible to reliably identify clusters of residues that suggest a potential protein binding site. Two large sites were successfully identified that, when visualized on the surface of the RPA3313 NMR structure, lie opposite of one another (Figure 7). One potential protein binding site is between the β-sheet and α-helix on the bottom of RPA3313, while the other crosses the width of the β-sheet on the top of the protein. The bottom protein biding site consists of side chains from residues Tyr6, Trp15, Phe34, Cys38, Ser42, Ile45, Lys46, Glu50, Val51, Arg52, Ile53, and Thr54 (Figure 7a). Although mostly hydrophobic in composition (Figure 7c), these residues form a likely interaction hotspot due to their high abundance in other known protein-protein interactions.[53] Furthermore, the surfaces of β-sheets are known to commonly participate in protein binding. A protein binding event at this bottom location on the RPA3313 surface could induce a significant reshuffling of the hydrophobic core as discussed earlier. The second predicted protein binding site runs perpendicular to the β-sheet and lies directly opposite the first predicted binding site (Figure 7b). Solvent exposed side chains from residues Val9, Tyr27, Ala32, Ala36, Ala39, and Asn43 populate the surface of the top protein binding site. This putative protein interaction site has both hydrophobic and hydrophilic regions (Figure 7d) indicating that multiple binding partners are possible.

.

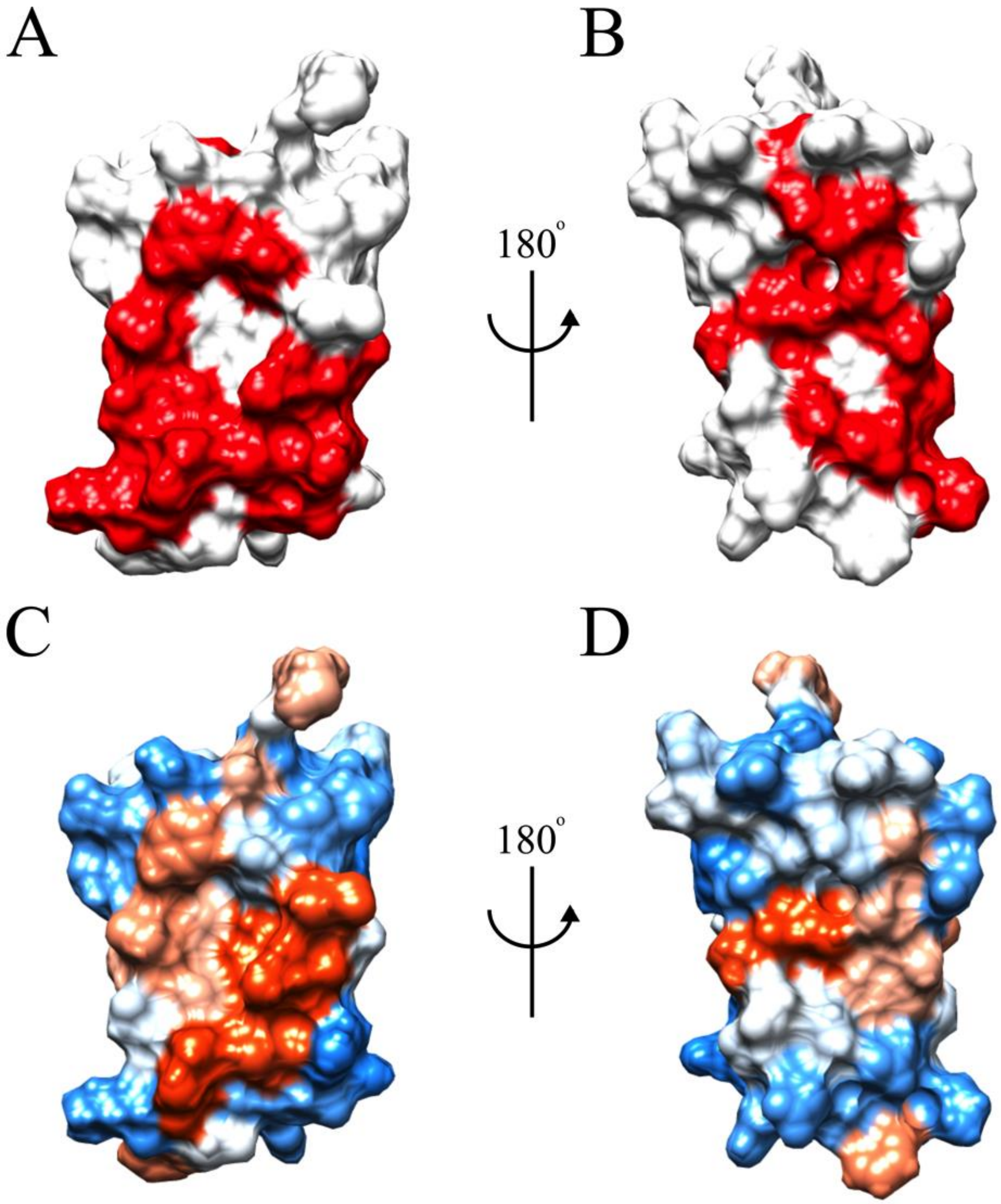


Figure 7. cons-PPISP predictions and surface hydrophobicity. (A) Front and (B) back orientations of RPA3313 with predicted protein interaction sites colored red. (C) Front and (D) back surface hydrophobicity colored from blue (hydrophilic) to orange (hydrophobic).

5.3.4 Protein-Protein Crosslinking

A simple crosslinking experiment was carried out in order to determine possible protein binders to RPA3313. The crosslinking experiment was performed both *in vivo* in *E. coli* and *in vitro* with a proteome extract from *R. palustris*. The replicate crosslinking experiments were performed to reliably identify physiologically-relevant interaction partners to RPA3313. RPA3313 was overexpressed in *E. coli* and prior to the two distinct crosslinking experiments the cell culture was split into two separate samples. A small aliquot of the total cell culture was removed for the *in vivo* crosslinking experiment, and the remaining cell culture was then used to extract and purify the overexpressed RPA3313 protein.

Purified RPA3313 was spiked into a total protein extract from an *R. palustris* cell culture and the formaldehyde crosslinking was then performed *in vitro.* In contrast, the *E. coli* cell culture overexpressing RPA3313 was simply treated with formaldehyde for an *in vivo* crosslinking experiment. Formaldehyde was used to covalently link lysine side chains through amide bond formation and subsequently removed by heating the sample after purification. Following purification, the crosslinked proteins were resolved by SDS-PAGE and identified by MS/MS analysis. Proteins found to be crosslinked to RPA3313 belonged to ribosomal subunits in both *E. coli* and *R. palustris* (Table 3). Moreover, most of the protein component of the ribosome from both organisms was identified to bind RPA3313. Thus, RPA3313 appears likely to bind to the ribosome at one or multiple points. Since it is also possible that the ribosome remained intact during the crosslinking and purification process, the number of binding sites for RPA3313 on the ribosome remains undetermined. It is important to note that RPA3313 was only overexpressed in *E.*

coli and not in *R. palustris*. This insures that the results are physiologically relevant and not a simple artifact of an overexpressed protein being crosslinked to equally abundant ribosomal proteins. In fact, an additional *in vivo* crosslinking experiment with a second overexpressed protein (human DJ-1) served as a negative control and verified that the RPA3313 results were not an artifact of a protein overexpression system. Despite identical experimental conditions and unlike RPA3313, human protein DJ-1 did not crosslink with any (as expected) *E. coli* proteins *in vivo*.

A previous study successfully sequenced and identified the ribosomal subunits from *R. palustris*.[54] During the study, other uncharacterized proteins were purified with the ribosome, but none of them were identified as RPA3313. Like many ribosomal subunits from other organisms, some of the subunits from *R. palustris* contained disordered C-termini. The disordered C-terminus acts as an anchor and buries itself into the RNA core and also promotes proper assembly of the ribosome.[55,56] Globular portions of the proteins are then exposed to the solvent to interact with other proteins. While RPA3313 is not part of the ribosome, its tertiary structure mimics a ribosomal subunit with the multiple protein-protein binding sites and a disordered C-terminus. RPA3313 may instead act as a chaperone or transporter for substrates traveling to or from the ribosome.

Table 3. *In vivo* and *in vitro* Crosslinking Results

Organism	Identified Ribosomal Proteins
E. coli	S2-11, S13, S15, S16, S18, S19, S21, L2-6, L9-11, L13-22, L24, L27, L28, L32
R. palustris	S6-13, S15-17, S19-21, L1-3, L6, L7/12, L9, L10, L15, L17-19, L21, L23, L24, L30, L32, L33
E. coli *(DJ-1 control)*	NA

5.4 Conclusion

RPA3313 is a conserved protein from *R. palustris* and a member of functionally unannotated class of proteins in *alphaproteobacteria.* The purpose of this study was to structurally characterize this class of proteins and provide an initial functional characterization. An NMR solution structure reveals that RPA3313 adopts a novel globular split ββαβ motif followed by a disordered C-terminus tail. PSVS evaluation of the ensemble of the 20 lowest energy structures of RPA3313 produced generally good quality scores consistent with other high-quality NMR structures deposited in the PDB. Bioinformatics analyses led to the identification of several possible protein-protein interaction sites on the surface of RPA3313 and a large conserved pocket sandwiched between the β-sheet and α-helix. Crosslinking analysis revealed that RPA3313 interacts with the ribosome both *in vivo* and *in vitro*. Multiple ribosomal subunits were pulled down with RPA3313 in *E. coli* and in *R. palustris*, so the exact nature of the interaction between the two is unknown. *In silico* dockings, ^{15}N NMR titrations, and ligand screenings were done in an attempt to determine the physiological role of RPA3313 (data not shown). However, a binder with a sub-millimolar binding constant was not found. It is possible that the tertiary structure of RPA3313 changes the shape of its binding pocket when in contact with another protein or in a much larger complex. Also, the C-terminus tail could be blocking or competing for the binding site and the lack of an N-terminus methionine truncation could also impede the binding site. The expression conditions of RPA3313 also remain unknown. It is possible that the protein is only expressed during certain metabolic modes of growth, as this protein is not found in evolutionary distant bacterial species with more limited metabolism. The combined structural and proteomic analyses in this study strongly suggests that RPA3313 by itself or in a larger complex may serve as a ribosomal transport protein.

5.5 References

[1] Evans, K., Georgiou, T., Hillon, T., Fordham-Skelton, A., and Papiz, M. (2009) Bacteriophytochromes control photosynthesis in Rhodopseudomonas palustris, *Adv. Photosynth. Respir. 28*, 799-809.

[2] Larimer, F. W., Chain, P., Hauser, L., Lamerdin, J., Malfatti, S., Do, L., Land, M. L., Pelletier, D. A., Beatty, J. T., Lang, A. S., Tabita, F. R., Gibson, J. L., Hanson, T. E., Bobst, C., Torres, J. L. T. y., Peres, C., Harrison, F. H., Gibson, J., and Harwood, C. S. (2004) Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris, *Nature Biotechnology 22*, 55-61.

[3] Kawasaki, H., Hoshino, Y., and Yamasato, K. (1993) Phylogenetic diversity of phototrophic purple non-sulfur bacteria in the Proteobacteria α group, *FEMS Microbiology Letters 112*, 61-66.

[4] Sasikala, C., Ramana, C. V., and Rao, P. R. (1994) Nitrogen fixation by Rhodopseudomonas palustris OU 11 with aromatic compounds as carbon source/electron donors, *FEMS Microbiology Letters 122*, 75-78.

[5] Akkerman, I., Janssen, M., Rocha, J., and Wijffels, R. H. (2002) Photobiological hydrogen production: photochemical efficiency and bioreactor design, *Int. J. Hydrogen Energy 27*, 1195-1208.

[6] Dagley, S. (1971) Catabolism of aromatic compounds by microorganisms, *Advan. Microbial Physiol. 6*, 1-46.

[7] Romagnoli, S., and Tabita, F. R. (2009) Carbon dioxide metabolism and its regulation in nonsulfur purple photosynthetic bacteria, *Adv. Photosynth. Respir. 28*, 563-576.

[8] McKinlay, J. B. (2014) Systems Biology of Photobiological Hydrogen Production by Purple Non-sulfur Bacteria, *Adv. Photosynth. Respir. 38*, 155-176.

[9] Tanawade, S. S., Bapat, B. A., and Naikwade, N. S. (2011) Biofuels: use of Biotechnology to meet energy challenges, *Int. J. Biomed. Res. 2*, 25-31.

[10] Paulsen, I. T., Sliwinski, M. K., and Saier, M. H. (1998) Microbial Genome Analyses : Global Comparisons of Transport Capabilities Based on Phylogenies, Bioenergetics and Substrate Specificities, *Journal of Molecular Biology 277*, 573-592.

[11] VerBerkmoes, N. C., Shah, M. B., Lankford, P. K., Pelletier, D. A., Strader, M. B., Tabb, D. L., McDonald, W. H., Barton, J. W., Hurst, G. B., Hauser, L., Davison, B. H., Beatty, J. T., Harwood, C. S., Tabita, F. R., Hettich, R. L., and Larimer, F. W. (2006) Determination and comparison of the baseline proteomes of the versatile microbe Rhodopseudomonas palustris under its major metabolic states, *Journal of Proteome Research 5*, 287-298.

[12] Apweiler, R., Bairoch, A., and Wu, C. H. (2004) Protein sequence databases, *Curr. Opin. Chem. Biol. 8*, 76-80.

[13] Porollo, A. A., Adamczak, R., and Meller, J. (2004) POLYVIEW: a flexible visualization tool for structural and functional annotations of proteins, *Bioinformatics 20*, 2460-2462.

[14] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, *Nucleic Acids Research 28*, 235-242.

[15] Tabita, F. R., Hanson, T. E., Satagopan, S., Witte, B. H., and Kreel, N. E. (2008) Phylogenetic and evolutionary relationships of RubisCO and the RubisCO-like proteins and the functional lessons provided by diverse molecular forms, *Philos. Trans. R. Soc., B 363*, 2629-2640.

[16] Badger, M. R., and Bek, E. J. (2008) Multiple Rubisco forms in proteobacteria: their functional significance in relation to CO2 acquisition by the CBB cycle, *J. Exp. Bot. 59*, 1525-1541.

[17] Bryant, D. A., and Frigaard, N.-U. (2006) Prokaryotic photosynthesis and phototrophy illuminated, *Trends Microbiol. 14*, 488-496.

[18] Skolnick, J., Fetrow, J. S., and Kolinski, A. (2000) Structural genomics and its importance for gene function analysis, *Nature Biotechnology 18*, 283-287.

[19] Baker, D., and Sali, A. (2001) Protein Structure Prediction and Structural Genomics, *Science 294*.

[20] Baca, A. M., and Hol, W. G. J. (2000) Overcoming codon bias: A method for high-level overexpression of Plasmodium and other AT-rich parasite genes in Escherichia coli, *Int. J. Parasitol. 30*, 113-118.

[21] Hyberts, S. G., Takeuchi, K., and Wagner, G. (2010) Poisson-Gap Sampling and FM Reconstruction for Enhancing Resolution and Sensitivity of Protein NMR Data, *Journal of the American Chemical Society 132*, 2145-2147.

[22] Ikura, M., Kay, L. E., and Bax, A. (1990) A Novel Approach for Sequential Assignment of 1H, 13C, and 15N Spectra of Larger Proteins: Heteronuclear Triple-Resonance Three-Dimensional NMR Spectroscopy. Application to Calmodulint, *Biochemistry 29*, 4659-4667.

[23] Kay, L. E., Ikura, M., Tschudin, R., and Bax, A. (1990) Three-Dimensional Triple-Resonance NMR Spectroscopy of Isotopically Enriched Proteins, *Journal of Magnetic Resonance 89*, 496-514.

[24] Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, M., Ulrich, E. L., Markley, J. L., Ionides, J., and Laue, E. D. (2005) The CCPN Data Model for NMR Spectroscopy : Development of a Software Pipeline, *Proteins: Structure, Function and Bioinformatics 59*, 687-696.

[25] Shen, Y., Bryan, P. N., He, Y., Orban, J., Baker, D., and Bax, A. (2010) De novo structure generation using chemical shifts for proteins with high-sequence identity but different folds, *Protein Science 19*, 349-356.

[26] Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J. M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K. K., Lemak, A., Ignatchenko, A., Arrowsmith, C. H., Szyperski, T., Montelione, G. T., Baker, D., and Bax, A. (2008) Consistent blind protein structure generation from NMR chemical shift data, *Proceedings of the National Academy of Sciences of the United States of America 105*, 4685-4690.

[27] Shen, Y., Vernon, R., Baker, D., and Bax, A. (2009) De novo protein structure generation from incomplete chemical shift assignments, *Journal of Biomolecular NMR 43*, 63-78.

[28] Schwieters, C. D., Kuszewski, J. J., Tjandra, N., and Clore, G. M. (2003) The Xplor-NIH NMR molecular structure determination package, *Journal of Magnetic Resonance 160*, 65-73.

[29] Schwieters, C. D., Kuszewski, J. J., and Clore, G. M. (2006) Using Xplor–NIH for NMR molecular structure determination, *Progress in Nuclear Magnetic Resonance Spectroscopy 48*, 47-62.

[30] Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009) TALOS + : a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts, *Journal of Biomolecular NMR 44*, 213-223.

[31] Bhattacharya, A., Tejero, R., and Montelione, G. T. (2007) Evaluating Protein Structures Determined by Structural Genomics Consortia Tools for Structure Quality Evaluation, *Proteins: Structure, Function and Bioinformatics 66*, 778-795.

[32] Sippl, M. J. (1993) Recognition of Errors in Three-Dimensional Structures of Proteins, *Proteins: Structure, Function, and Genetics 17*, 355-362.

[33] Luthy, R., Bowie, J. U., and Eisenberg, D. (1992) Assessment of protein models with three-dimensional profiles, *Nature 356*, 83-85.

[34] Lovell, S. C., Davis, I. W., Arendall III, W. B., de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C. (2003) Structure Validation by Ca Geometry: phi, psi and CB, Deviation, *Proteins: Structure, Function, and Genetics 50*, 437-450.

[35] Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) PROCHECK: a program to check stereochemical quality of protein structures, *Journal of Applied Crystallography 26*, 283-291.

[36] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera — A Visualization System for Exploratory Research and Analysis, *Journal of Computational Chemistry 25*, 1605-1612.

[37] Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996) Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels, *Anal Chem 68*, 850-858.

[38] Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., and Ben-tal, N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures, *Nucleic Acids Research 33*, 299-302.

[39] Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T., and Ben-Tal, N. (2013) ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function, *Israel Journal of Chemistry 53*, 199-206.

[40] Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003) ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information, *Bioinformatics 19*, 163-164.

[41] Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-tal, N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids, *Nucleic Acids Research 38,* 529-533.

[42] Krissinel, E., and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, *Acta Crystallographica Section D 60,* 2256-2268.

[43] Chen, H., and Zhou, H. X. (2005) Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data, *Proteins: Structure, Function and Bioinformatics 61*, 21-35.

[44] Zhou, H.-x., and Shan, Y. (2001) Prediction of Protein Interaction Sites From Sequence Profile and Residue Neighbor List, *Proteins: Structure, Function, and Genetics 44*, 336-343.

[45] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera-A visualization system for exploratory research and analysis, *J. Comput. Chem. 25*, 1605-1612.

[46] Huson, D. H., and Scornavacca, C. (2012) Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks, *Systematic Biology 0*, 1-7.

[47] Orengo, C. A., and Thornton, J. M. (1993) Alpha plus beta folds revisited: some favoured motifs, *Current Biology 1*, 105-120.

[48] Uversky, V. N. (2013) The most important thing is the tail: Multitudinous functionalities of intrinsically disordered protein termini, *FEBS Letters 587*, 1891-1901.

[49] Marintcheva, B., Marintchev, A., Wagner, G., and Richardson, C. C. (2008) Acidic C-terminal tail of the ssDNA-binding protein of bacteriophage T7 and ssDNA compete for the same binding surface, *Proceedings of the National Academy of Sciences of the United States of America 105*, 1855-1860.

[50] Thieulin-pardo, G., Avilan, L., Kojadinovic, M., and Gontero, B. (2015) Fairy “tails”: flexibility and function of intrinsically disordered extensions in the photosynthetic world, *Frontiers in Molecular Biosciences 2*, 1-18.

[51] Ben-bassat, A., Bauer, K., Chang, S.-y., Myambo, K. E. N., and Boosman, A. (1987) Processing of the Initiation Methionine from Proteins: Properties of the Escherichia coli Methionine Aminopeptidase and Its Gene Structure, *Journal of Bacteriology 169*, 751-757.

[52] Theillet, F.-x., Kalmar, L., Tompa, P., Han, K.-h., Dunker, A. K., Daughdrill, G. W., Uversky, V. N., Theillet, F.-x., Kalmar, L., Tompa, P., Han, K.-h., Selenko, P., and Dunker, A. K. (2013) The alphabet of intrinsic disorder I . Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins, *Intrinsically Disordered Proteins 1*.

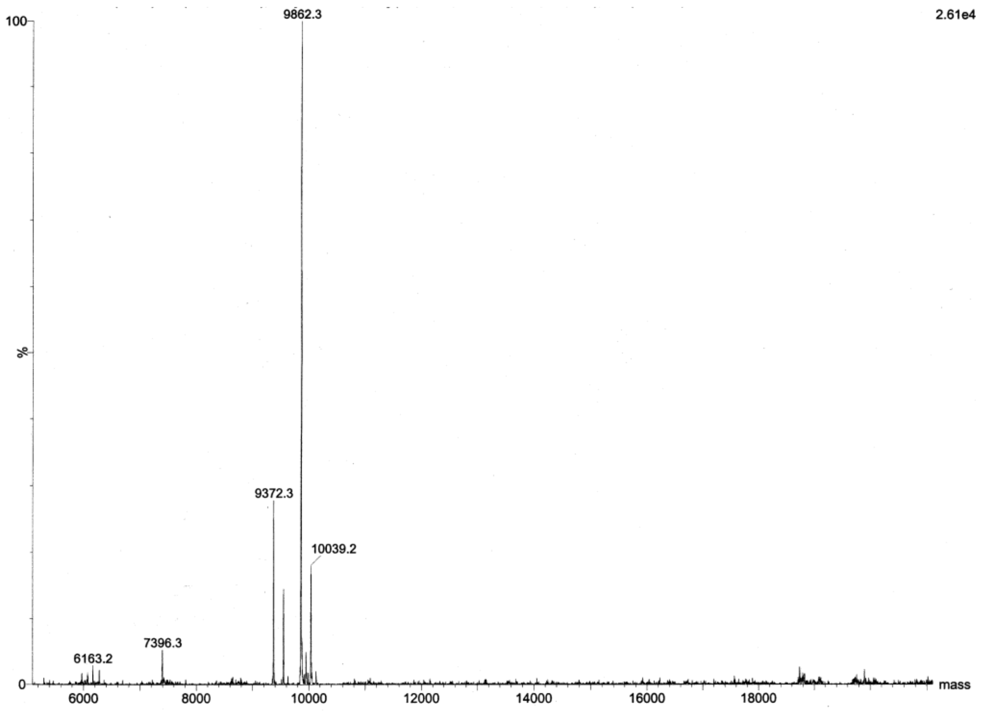
[53] Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003) Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces., *Proceedings of the National Academy of Sciences of the United States of America 100*, 5772-5777.

[54] Strader, M. B., VerBerkmoes, N. C., Tabb, D. L., Connelly, H. M., Barton, J. W., Bruce, B. D., Pelletier, D. A., Davison, B. H., Hettich, R. L., Larimer, F. W., and Hurst, G. B. (2008) Characterization of the 70S Ribosome from Rhodopseudomonas palustris Using and Integrated “Top-Down” and “Bottom-Up” Mass Spectrometric Approach, *Journal of Proteome Research 3,* 965-978.

[55] Peng, Z., Oldfield, C. J., Xue, B., Mizianty, M. J., Dunker, A. K., Kurgan, L., and Uversky, V. N. (2014) A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome, *Cellular and Molecular Life Sciences 71,* 1477-1504.

[56] Brodersen, D. E., Jr, W. M. C., Carter, A. P., Wimberly, B. T., and Ramakrishnan, V. (2002) Crystal Structure of the 30 S Ribosomal Subunit from Thermus thermophilus: Structure of the Proteins and their Interactions with 16 S RNA, *Journal of Molecular Biology 316*, 725-768.

5.6 Supplemental Figures



Supplemental Figure S1. Intact mass analysis using ESI-MS of RPA3313 from Figure 1. ESI-MS confirms the major state of RPA3313 is a monomer.

Chapter 6

Identification of a Ligand-Binding site on the *Staphylococcus aureus* DnaG Primase C-Terminal Domain

6.1 Introduction

The initiation of DNA synthesis in living organisms and many large viruses results from the coordinated interaction of two enzymes - primase and helicase.[1, 2] In bacteria, DnaG primase is the specialized DNA-dependent RNA polymerase that synthesizes short oligoribonucleotide polymers called primers. DNA polymerase, an enzyme that lacks the ability to initiate chain synthesis, elongates the primers. DnaG primase is stimulated by the homohexameric DnaB helicase. (Note: The name DnaB helicase is used here for the protein that is actually named DnaC helicase in *Staphylococcus aureus* and *Bacillus subtilis* because, in all other bacteria, DnaB is the name of the replicative helicase and DnaC is the name of the helicase-loading enzyme). DnaB helicase dissociates the two strands of duplex DNA during DNA replication while hydrolyzing ATP, and travels processively in the 5’-3’ direction along the single-stranded lagging template toward the replication fork, communicating allosterically with the multi-subunit replicative DNA polymerase. This action keeps it in proximity of the replication fork and ensures the primers are synthesized on the exposed single-stranded DNA nearest to the replication fork. Since primase activity is weak, the stimulation by DnaB causes primase to synthesize primers only at the replication fork when and where they are needed.[3, 4] Finally, the DnaG-DnaB interaction is an attractive antibiotic target because it is conserved in bacteria, essential for DNA replication, and is distinctly different from that of viruses, archaea, and eukaryotes.[5-7]

DnaG primase is composed of three functional domains: the N-terminal zinc-binding domain (ZBD, Pfam: PF08275) responsible for DNA binding specificity; the RNA polymerase domain (RPD, Pfam: PF01751) responsible for enzymatic activity; and the C-terminal domain (CTD, Pfam: PF10410) responsible for the interaction with the DnaB helicase.[8-10] The 110-residue ZBD contains the highest percentage of conserved residues and structural conservation, whereas the 140-residue CTD has the least sequence conservation and the most structural diversity (Table S1). Of these domains, the ZBD is unique to bacterial primases, the magnesium-binding residues of the RPD shares some similarity to the magnesium-binding residues of topoisomerases and some DNA-hydrolyzing enzymes, while the CTD is similar only to the N-terminal domain of DnaB helicases, the domain to which it binds.[11-14] Currently, no high-resolution structure has been solved for a full, intact bacterial primase, but structures are available for all three individual domains. The primase CTD is the domain with the most experimental structures.[6, 15-18] There are X-ray crystal structures and NMR solution structures (Table 1) from *S. aureus* (2LZN, current paper), *E. coli* (2HAJ, 1T3W), *G. stearothermophilus* (1Z8S, 2R6A), and *H. pylori* (4EHS). In addition, the ZBD-RPD domain pair from *Aquifex aeolicus* has been experimentally determined.[11] It has been proposed that the conformations of the two *E. coli* structures in 1T3W may not be biologically relevant due to crystallization conditions and packing effects.[17]

To develop the DnaG-DnaB interface as an antibiotic target, we determined the solution structure of the DnaG CTD from *S. aureus*. This organism was chosen because there is an urgent need for new antibiotic targets due to the rapid rise in antibiotic resistance.[19] A comparison to known structures from other bacteria indicated that the *S. aureus* CTD was composed of two subdomains. The CTD from *Firmicutes* and *Proteobacteria* may adopt similar conformations when bound to helicase, but may adopt different free conformations. A study of the structural dynamics of the free *S. aureus* CTD confirmed that the junction between the two subdomains is the focal point for the greatest movement. Also, one of the subdomains is significantly more ordered than the other. The structure was used to identify three small compounds that bound to the close conformation, which is likely to reduce the primase-helicase interaction.

Table 1: Primase CTD structures used for analysis

Species	PDB ID – chain	Reference
S. aureus	2LZN – 14	Current paper
G. stearothermophilus	1Z8S – 5	[6]
E. coli	2HAJ – 19	[17]
G. stearothermophilus	2R6A – C	[16]
H. pylori	4EHS – A and 4EHS – B	[20]
E.coli	1T3W – A and 1T3W – B	[18]

6.2 Experimental Procedures

6.2.1 Chemicals

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 Isotopes (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 earlier.[21] Briefly, the compound library is composed of 437 known biologically active compounds distributed across 113 mixtures with 3-4 compounds in each mixture.

6.2.2 Proteins

Uniformly ^{13}C,^{15}N-labeled *S. aureus* DnaG primase CTD and uniformly ^{15}N-labeled *S. aureus* DnaG primase CTD were designed and purified by Nature Technologies (Lincoln, NE). The details of protein expression and purification were previously described.[22] The vector added 20 N-terminal residues: an HN metal affinity tag (MGHNHNHNHNHNHNGG) followed by a protease-sensitive DDDD sequence.[22]

6.2.3 NMR data analysis, structure calculations and refinement

The NMR spectra for obtaining the protein backbone assignments were collected at 298 K on a five-channel 600-MHz Bruker Avance spectrometer equipped with a 5-mm TXI probe. The NMR spectra for the protein side chain assignments 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. The nearly complete *S. aureus* DnaG primase CTD NMR resonance assignments have been previously reported.[22] Distance constraints were obtained from 3D ^{15}N-edited NOESY and 3D ^{13}C-edited NOESY that were collected at 900 MHz.[23] Hydrogen bond constraints were determined using the (CLEANEX-PM)-FHSQC experiment.[24] All torsion angle constraints were obtained by chemical shift analysis using the TALOS[25] software program, and measured coupling constants from an HNHA experiment.[26]

For the backbone NMR assignment experiments, ^{13}C,^{15}N-labeled protein was concentrated to 1.2 mM in a 95% H_2O/5% D_2O buffered solution of 100 mM NaCl, 25 mM potassium phosphate at pH 6.64 (uncorrected) using an Amicon ultra centricon (MW cutoff 10,000 Da). Long-term protein stability was enhanced by adding 50 mM arginine and 50 mM glutamine. For the side chain experiments, ^{13}C,^{15}N-labeled protein was concentrated to 1.4 mM in the same buffer. All multidimensional experiments were processed using NMRPipe[27] and analyzed using PIPP[28] and CCPNMR.[29]

Nuclear Overhauser Effect (NOE) assignments were obtained by using 3D ^{15}N-edited NOESY and 3D ^{13}C-edited NOESY. NOE intensities were sorted visually into four classes: strong (1.8–2.5 Å), medium (1.8–3.0 Å), weak (1.8–4.0 Å), and very weak (3.0–5.0 Å). Upper limits for distances involving methyl protons and non-stereospecifically assigned methylene protons were corrected appropriately for center averaging. Initial NOE assignments were completed by the program Autostructure.[30] Despite the high magnetic field of 900 MHz, the extent of cross-peak overlap was great enough to warrant manual refinement of the NOE assignments.

Hydrogen bond constraints were determined using the (CLEANEX-PM)-FHSQC experiment, which identifies amide residues with fast water exchange rates.[24] Hydrogen bond constraints were assigned to amides in secondary structure regions for any 2D ^{1}H-^{15}N HSQC peaks lacking a corresponding peak in the (CLEANEX-PM)-FHSQC spectrum. The carboxyl oxygen to amide nitrogen hydrogen bond distances were set at 2.8 Å while the carbonyl oxygen to amide proton distances were set at 1.8 Å. All carboxyl groups within 2.5 Å of slowly exchanging amide groups were constrained to be involved in a hydrogen bond.

The structures were refined using the hybrid distance geometry dynamic-simulated annealing method[31] using the XPLOR-NIH program[32, 33] adapted to incorporate pseudopotentials for ^{3}J(HN-Hα) coupling constants, secondary ^{13}Cα/^{13}Cβ chemical shift constraints, and a conformational database potential.[34-38] 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.[39] All peptide bonds were constrained to be planar and trans. There were no hydrogen-bonding or electrostatic empirical potential energy terms in the target function. A total of 1000 structures were calculated and the 20 lowest energy structures were selected to become the ensemble of *S. aureus* DnaG primase CTD structures. Each of these twenty structures were subjected to further energy minimization using explicit water with Crystallography and NMR system (CNS) version 1.2.[40] Explicit water solvation was accounted for with Lennard-Jones and electrostatic potentials using a modification of the procedure and force field of Nilges.[41, 42]

6.2.4 Protein backbone dynamics

The NMR experiments for protein dynamics analysis were collected on a Bruker 500 MHz Avance spectrometer (Billerica, MA) equipped with a triple resonance, Z-axis gradient cryoprobe. The sample was uniformly ^{15}N-labeled *S. aureus* DnaG primase CTD concentrated to 1.2 mM in a 95% H_2O/5% D_2O buffered solution of 100 mM NaCl, 50 mM arginine, 50 mM glutamine, 25 mM potassium phosphate, pH 6.64 (uncorrected) using an Amicon ultra centricon (MW cutoff 10 000 Da).

The previously described[43-45] experiments included a 2D ^{15}N-^{1}H 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 ^{15}N-^{1}H 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 ^{15}N-^{1}H HSQC experiment (*hsqcnoef3gpsi*) designed to measure NOE changes.

The relaxation rates (T_1, T_2) for each amino acid within the structure were calculated by fitting the intensity of each peak to the intensity decay curve (eq 1) where I_t is the intensity of each peak at the delay time *t*, I_0 is the initial steady state intensity.

$$I_t = I_0 \exp(-\frac{t}{T_{1,2}}) \qquad \text{eq 1}$$

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

$$NOE = I_{sat}/I_{unsat} \qquad \text{eq 2}$$

The T_1 rates, T_2 rates, and NOE ratios were determined from their fits to their respective equations using Kaleidagraph (Synergy Software, Reading, PA) and exported to the FAST-MODEL FREE program[46] to calculate an overall correlation time (τ_m); and per residue order parameters (S^2), internal motion (τ_e), and chemical exchange (R_{ex}) using the Lipari-Szabo model-free method.[47] Some of the NOE ratios were above the theoretical maximum of 1.0 and were set to 1.0 to run the FAST-MODEL FREE program.

6.2.5 Primase CTD and C1 subdomain similarity phylograms

An all-versus-all root-mean-square deviation (RMSD) matrix table (Table S2) was generated using the MatchMaker tool in UCSF Chimera.[48] All chains from X-ray structures were included in the structural comparison. Only the best representative structure from the NMR ensemble, identified by XPLOR as having the lowest RMSD relative to the mean, was included in the comparison (See Table S3).

The computed RMSD matrices were then inserted into the Splitstree4[49] program to generate the unrooted phylogenic trees depicted in Figure 3. The trees were generated using the Neighbor-join algorithm. The pairwise RMSDs were calculated between the complete C-terminal domains (Figure 3A) as well as between the aligned C1 subdomains (Figure 3C).

6.2.6 Primase CTD sequence similarity

The sequences for the *E. coli*, *G. stearothermophilus*, *H. pylori*, and *S. aureus* primase C-terminal domains were acquired from the Protein Data Bank.[50] Three different sequence alignment programs, Muscle, T-Coffee, and Clustal Omega,[51-53] were used to determine a sequence similarity between the four primase CTD proteins. The results obtained from the three sequence alignment programs were identical and were converted into the Phylip format for analysis using Splitstree4.[49] The unrooted phylogenic tree (Figure 3B) was created using the Neighbor-join algorithm and equal angle method.

6.2.7 NMR ligand affinity assays

Sample preparation and experimental parameters for the NMR ligand affinity screen were performed as described previously.[54] Briefly, each ligand mixture (113 total) was screened using a 1D ^{1}H NMR spectrum with excitation sculpting.[55] Each NMR sample contained 100 □M ligand and 25 □M 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. All 1D ^{1}H NMR spectra were processed with the ACD/1D NMR manager v. 12.0 (Advanced Chemistry Development, Inc., Toronto, Ontario). Each 1D ^{1}H NMR spectrum was compared to the corresponding free ligand mixture reference spectrum and visually analyzed to identify binding ligands based on a decrease in the intensity of ligand NMR resonances. A 2D ^{1}H-^{15}N HSQC spectrum was collected for each ligand with a positive response from the 1D ^{1}H NMR line-broadening screen using 500 □M of ligand and 100 □M of protein under the same buffer conditions as the 1D ^{1}H NMR screen except for 95% H_2O/5% D_2O. A single ligand-free 2D ^{1}H-^{15}N HSQC spectrum was collected as a reference, where a binding event was confirmed based on a clustering of surface residues that incurred a chemical shift change one standard deviation above the average.

6.2.8 Protein-ligand titration and docking

DnaG primase CTD HSQC was collected at a protein concentration of 1 mM and was titrated with adenosine (0.9, 1.8, 2.6, 3.3, and 4.0 mM), acycloguanosine (0.7, 1.9, 3.5, and 5.2 mM), and myricetin (0.9, 1.8, 2.6, 3.3, and 4.0 mM). All NMR data was collected at 25 °C on a 700 MHz Bruker Avance III spectrometer equipped with a 5 mm QCI-P probe with cryogenically cooled carbon and proton channels. The 2D ^{1}H-^{15}N HSQC protein titration data was processed and analyzed in NMRPipe.[27] Briefly, the spectral titration data for each ligand was automatically peak picked and adjusted manually within NMRDraw. The ^{1}H-^{15}N and ^{15}N chemical shifts from the selected peaks were then fit to a Ka (1/Kd) according to a previously published method.[56] Subsequently, the peaks that showed significant binding were manually assigned to residues. AutoDock was then used to individually dock the three ligands into a primase CTD binding site defined by the observed chemical shift perturbations. Chimera was used to visualize the lowest energy docked complexes.[48]

6.3 Results

To fight antibiotic-resistant infections, there is a need to identify new antibiotic targets. The bacterial primase C-terminal domain is an especially promising target because eukaryotic primases do not have a cognate sequence or structure and because its critical function is to interact with DnaB helicase to limit DNA replication to the replication fork. Among a number of factors, evaluating the “drugability” of a protein target requires a fundamental understanding of its structure-function relationship.[57] As a first step towards this goal, we determined the three-dimensional structure of the primase CTD from the pathogen *S. aureus*.

6.3.1 *S. aureus* primase CTD structure determination

Backbone and side chain resonance assignments were completed using standard triple resonance NMR experiments.[22] In summary, the backbone resonance assignment was 93% complete with 133 amino acids out of the 143 primase residues assigned unambiguously in the 2D ^{1}H-^{15}N HSQC spectrum. Nearly all the peaks were uniformly-shaped and separated from the others, which indicates that each residue is in a unique environment and that *S. aureus* primase CTD exists in a single conformation. Unassigned residues include M1-H13, D19, E470, H479, L480, M481, T500, R536, E537, E543, P551, and Y552. Primase CTD residues are numbered relative to the complete DnaG primase sequence, while the N-terminal purification tag is simply numbered from residue 1 to 20. Most of the unassigned residues were located in the purification tag or within highly solvent exposed regions. The structure consists of eight α-helices. Residues H479, L480, and M481 were in a turn region between helix 1 and 2. Residue T500 was in an unstructured loop region between helix 2 and helix 3, whereas residues P551 and Y552 were in an unstructured loop region between helix 5 and helix 6. The only unassigned residues that appear to be localized within a secondary structure were E470, R536, E537, and E543. E470 is the second residue of helix 1, while R536 and E537 are in the middle of helix 5, and E543 is near the end of helix 5. An exhaustive analysis of the NMR data set was unable to yield assignments for these residues, suggesting the end of helices 1 and 5 may undergo partial unfolding leading to chemical shift exchange broadening for these residues.

The solution structure of the *S. aureus* primase CTD was calculated using 1823 distance constraints, 280 dihedral constraints, 256 $^{13}C\alpha$ and $^{13}C\beta$ carbon chemical shift constraints, and 82 $^{3}J_{NH\alpha}$ coupling constant constraints, among others (Table 2). In the initial phase, 1000 structures were calculated from 10 individual sets of 100 structures each using XPLOR-NIH[32, 33] as described previously.[58] Each set of structures was started from a randomly generated seed. The lowest energy structures were consolidated to generate a set of 20 low energy structures that were further refined within a virtual water environment using the RECOORD recalculated coordinates[42] as implemented in CNS.[40] The resulting *S. aureus* primase CTD structures (Figure 1) form a self-consistent set as determined by a range of statistical analyses (Table 2). The root mean square deviation (RMSD) for all 1823 experimental distance constraints is 0.040 (± 0.023) Å, which implies a good agreement between the structural ensemble and the constraints. None of the distance constraints have a violation that exceeds 0.247 Å. There are also low deviations from the constraints for the $^{13}C\alpha$ chemical shifts, $^{13}C\beta$ chemical shifts, and the

^{3}J(HN-Hα) coupling constants, indicating consistent conformations for both the backbone and side chains. Similarly, the deviations from idealized covalent geometries are consistent with good quality structures.

Table 2: Structure Statistics[a]

Root mean square parameter	<SA>	σ	$(SA)_r$
Distance restraints (Å)			
All (1823)	0.040	0.023	0.022
Inter-residue sequential (|i-j| = 1) (460)	0.056	0.021	0.030
Inter-residue short range (1 < |i-j| ≤ 5) (444)	0.062	0.018	0.041
Inter-residue long-range (|i-j| > 5) (140)	0.247	0.093	0.374
Intra-residue (663)	0.001	0.001	0.003
Backbone hydrogen bonds (116)	0.014	0.005	0.008
Dihedral restraints (deg) (280)	1.21	0.23	1.12
C□ chemical shift restraints (ppm) (130)	1.05	0.05	1.08
C□□ chemical shift restraints (ppm) (126)	1.06	0.04	1.09
$^{3}J_{NH□}$ coupling restraints (Hz) (82)	0.011	0.0005	0.011
F_{NOE} (kcal mol^{-1})	87	36	46
$F_{torsion}$ (kcal mol^{-1})	28	20	21
F_{VDW} (kcal mol^{-1})	-509	31	-510
Idealized covalent geometry			
Bond length (Å) (2400)	0.0114	0.0005	0.0108
All angles (deg) (4302)	1.44	0.06	1.43
Improper torsion angles (deg) (1287)	1.95	0.23	1.84

[a]<SA> is the average value from an all-versus-all comparison of the set of 20 annealed structures; σ is the standard deviation for the all-versus-all comparison; and $(SA)_r$ is the value for the restrained minimized mean structure. The number of restraints for each parameter is given in parentheses. For backbone NH-CO hydrogen bonds, the two restraints are r(NH-O) = 1.5 - 2.3 Å and r(N-O) = 2.5 - 3.3 Å. The values of the square-well NOE (F_{NOE}) and torsion angle ($F_{torsional}$) potentials[59] are calculated with force constants of 50 kcal $mol^{-1}Å^{-2}$ and 200 kcal $mol^{-1}rad^{-2}$, respectively. The value of the Lennard-Jones van der Waals term (F_{VDW})[60] is calculated with a force constant of 4 kcal $mol^{-1}Å^{-4}$ with the CHARMM[61] empirical energy function. The improper torsion angle restraints serve to maintain planarity and chirality.

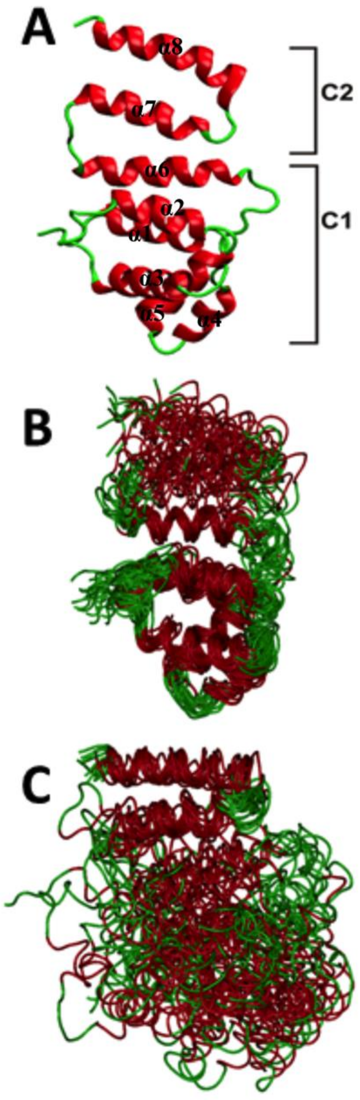


Figure 1: Primase CTD ensemble overlay. (A) A ribbon diagram of the average water refined structure. The C1 subdomain is composed of helices 1-6 and the C2 subdomain is

composed of helices 7-8. All structures were generated with Visual Molecular Dynamics (VMD-XPLOR) and are colored according to the secondary structure: red, α-helix; green, loop. (B) An overlay of the backbone trace of the 20 lowest-energy, water-refined structures aligned with residues 468-566 from the C1 subdomain. (C) An overlay of the backbone trace of the 20 lowest-energy structures aligned with residues 570-603 from the C-terminal C2 subdomain.

The quality of the *S. aureus* primase CTD NMR structure was assessed using two suites of programs (Table 3). PROCHECK[62] indicates the average minimized structure has an overall G-Factor of -0.12 ± 0.04 with no bad contacts, consistent with a good quality structure. The Protein Structure Validation Software suite of programs (PSVS)[63] gave a Verify3D score of -4.58 ± 0.71, which is within the typical range. The Molprobity module indicates that the *S. aureus* primase CTD structure has a very good Z-score (-2.50) compared to the average Z-score (-10.74) for all NMR structures in the Protein Data Bank. Finally, a Ramachandran plot of backbone dihedral angles for all non-glycine residues indicates that 83.4 ± 3.2% lie within the “most favored” region and only 0.7 ± 0.6% lie within “disallowed” regions. That is, one residue in 14 of the 20 structures is found in the disallowed region but never in a consistent location within the structure (Table S4). These residues tend to have a lower number of constraints, which results in a low penalty for unusual conformations during energy minimization.

The final step in the analysis was to compare the 20 energy-minimized structures with the mean structure (Table 4). Within the C1 subdomain comprising the first two-thirds of the residues, the average root-mean square deviation (RMSD) of the 20 lowest energy structures about the mean coordinate position is 1.54 ± 0.19 Å for backbone atoms and 2.37 ± 0.17 Å for all heavy atoms. When only secondary structure elements are used for alignment, the mean coordinate position deviation is 1.18 ± 0.14 Å for backbone atoms and 1.93 ± 0.16 Å for all heavy atoms. These deviations indicate the backbone structure is well defined. The side chain residues have fewer constraints and correspondingly higher structural disorder.

Table 3: Ensemble self-consistency[a]

Parameter	<SA>	σ	$(SA)_r$
PROCHECK[b]			
Overall G-Factor	-0.12	0.04	-0.09
H-bond energy	0.82	0.06	0.8
Number bad contacts/100 residues	0.0		0.0
PSVS Z-scores[c]			
Verify3D	-4.58	0.71	-4.82
ProsaII (-ve)	-1.14	0.34	-1.20
Procheck (phi-psi)	-0.56	0.42	-0.28
Procheck (all)	-1.92	0.32	-2.07
MolProbity clash score	-2.50	0.97	-2.46
Ramachandran space[d]			
Most favored regions	83.4%	3.2	85.5%
Additional allowed regions	14.5%	3.0	13.0%
Generously allowed regions	1.43%	1.2	0.7%

Disallowed regions	0.7%	0.56	0.7%

[a] <SA> is the average value from an all-versus-all comparison of the set of 20 annealed structures; □ is the standard deviation for the all-versus-all comparison; and $(SA)_r$ is the value for the restrained minimized mean structure.

[b] PROCHECK values are in comparison to the likelihood of finding each residue within the derived structure as compared to a database of residues found in similar environments within structures with resolution below 1.8 Å.

[c] PSVS (Protein Structure Validation Suite) provides an analysis of five separate structural analysis programs, each of which measures a different parameter. The PSVS reports Z-scores for each analysis, where a Z-score is the number of standard deviations from the mean. The authors of PSVS suggest that Z-scores lower than -5 should be reevaluated.

[d] Ramachandran space analysis from the PDB sum website uses a database of structures with resolution below 1.8 Å to decide whether a region is considered to be "most favored", etc.

Table 4: Ensemble self-consistency of the C1 subdomain[a]

Comparison	Backbone atoms	All heavy atoms
All residues		
<SA> vs. SA_{mean}	1.54 ± 0.19 Å	2.37 ± 0.17 Å
<SA> vs. $(SA)_r$	1.98 ± 0.28	3.05 ± 0.32
$(SA)_r$ vs. SA_{mean}	1.26	1.94
Secondary structures		
<SA> vs. SA_{mean}	1.18 ± 0.14	1.93 ± 0.16
<SA> vs. $(SA)_r$	1.56 ± 0.22	2.55 ± 0.26
$(SA)_r$ vs. SA_{mean}	1.03	1.66

[a] <SA> is the average value from an all-versus-all comparison of the set of 20 annealed structures; $(SA)_r$ is the value for the restrained minimized mean structure; and SA_{mean} is the Cartesian mean structure created from the geometric mean for each atom of all 20 structures. The SA_{mean} structure does not adhere to any of the restraints for bond lengths, angles, etc. The secondary structure residues are: 469-478□□□□□, 484-493 (□□□, 502-515 (□□□, 524-528 (□□), 533-544 (□□□, 552-564 (□□□.The secondary structure rmsd values were measured by aligning all secondary structure elements and calculating an average and standard deviation.

6.3.2 *S. aureus* primase CTD structure

The 20 energy-minimized *S. aureus* primase CTD structures (Figure 1) are deposited in the Protein Data Bank (PDB) with the identification code 2LZN. The structure is composed of 8 helices arranged into two subdomains as is the primase CTD NMR structure from *G. stearothermophilus.*[6] In our *S. aureus* primase CTD structure, the first six helices create the C1 subdomain and encompass residues 467 to 565 (Figure 1A). The last two helices, 7 and 8, create the C2 subdomain encompassing residues 572 to 603. When the 20 structures are overlaid using the alignment of the C1 subdomain backbone residues as a guide (Figure 1B), the C2 subdomain structures do not superimpose well. The converse overlay based on the alignment of the C2 subdomain shows the same effect (Figure 1C). The poor superimposition arises because of the very low number of structural restraints between the subdomains. However, the linker between the subdomains is constrained by 11 NOEs from the C1 subdomain, 77 NOEs within the linker, and 17 NOEs from the C2 subdomain. On the other hand, the loop region between the subdomains lacks sequential NH-NH NOEs in the ^{1}H-^{15}N HSQC-edited NOESY and the loop residues G567, Q568 and E569 exhibit exchange peaks in the CLEANEX experiment.[24] These residues are likely undergoing rapid exchange with the solvent, a feature indicative of exposed residues that lack protection from hydrogen bonds found in secondary structures. As we shall see below, the dynamics analysis indicates that the subdomains behave as two independent domains.

The surface charge distribution shows the exposed surface is uniformly but weakly electrically positive (Figure 2). The buried interface between the subdomains is electrically negative, as reflected in the moderately large Verify 3D Z-score. This repulsion may destabilize the interaction between the C1 and C2 subdomains, causing them to prefer a more extended conformation. In addition, there are some partially buried electrically negative residues in the C1 subdomain. The overall charge distribution is unlike the uniformly negative CTD from *G. stearothermophilus*,[6] suggesting that it is not a universal feature of primase CTDs.

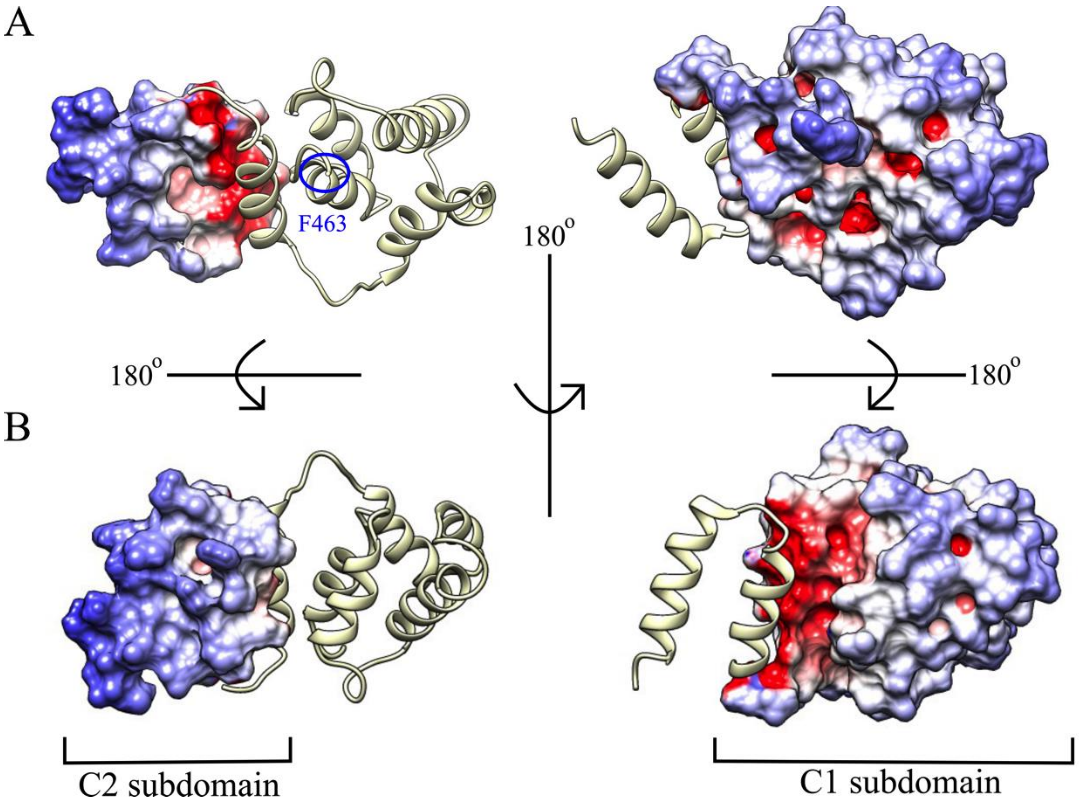


Figure 2: Electrostatic surface potential. Surface charge distribution of the C-terminal domain showing (A) the face toward the RNA polymerase domain (F463 is the N-

terminal residue of this domain) with space-filled C2 subdomain on the left and the ribbon diagrammed C1 subdomain on the right. (B) The solvent-exposed face (180° rotation around the x-axis relative to the structures in panel A) with space-filled C1 subdomain and ribbon diagrammed C2 subdomain.

6.3.3 Primase CTD conformations

Every enzyme family has a distinct sequence/structure relationship linked to the evolutionary pressures on the sequence and structure within each organism.[64] To determine how the *S. aureus* primase CTD structure relates to other known primase CTD structures (Table 1), the MatchMaker tool in UCSF Chimera[48] was used to calculate pairwise RMSDs for the eight known primase CTD structures (Table S2). The resulting unrooted phylogram (Figure 3A) shows three main branches that correlate with distinct conformations (Figure 4). The first branch (conformation A) includes our NMR structure from *S. aureus* (2LZN-14) and the *G. stearothermophilus* NMR structure (1Z8S-5). The second branch (conformation B) includes the *E. coli* NMR structure (2HAJ-19), the *G. stearothermophilus* crystal structure (2R6A-C), and both *H. pylori* crystal structures (4EHS-A and B). The third branch (conformation C) includes both *E. coli* crystal structures (1T3W-A and B). Conformations A and B are more similar to each other than they are to conformation C. Furthermore, it has been proposed[17] that conformation C may not be biologically relevant because the crystals were formed at low pH (4.6), which resulted in a packing effect of swapped dimer conformation.

Since the unrooted phylogram for the full length structure did not split according to sequence phylogeny, we created unrooted phylograms based on CTD sequence similarity (Figure 3B) and the primase CTD C1 subdomain structure (Figure 3C). From the sequence similarity phylogram, we observed two distinct branches corresponding to the expected phylogenetic split between the *Firmicutes* (*G. stearothermophilus*, *S. aureus*) and the *Proteobacteria* (*E. coli*, *H. pylori*). The same phylogenetic split was observed for the C1 subdomain structures, showing that there is a structure-sequence phylogenetic split for this protein.

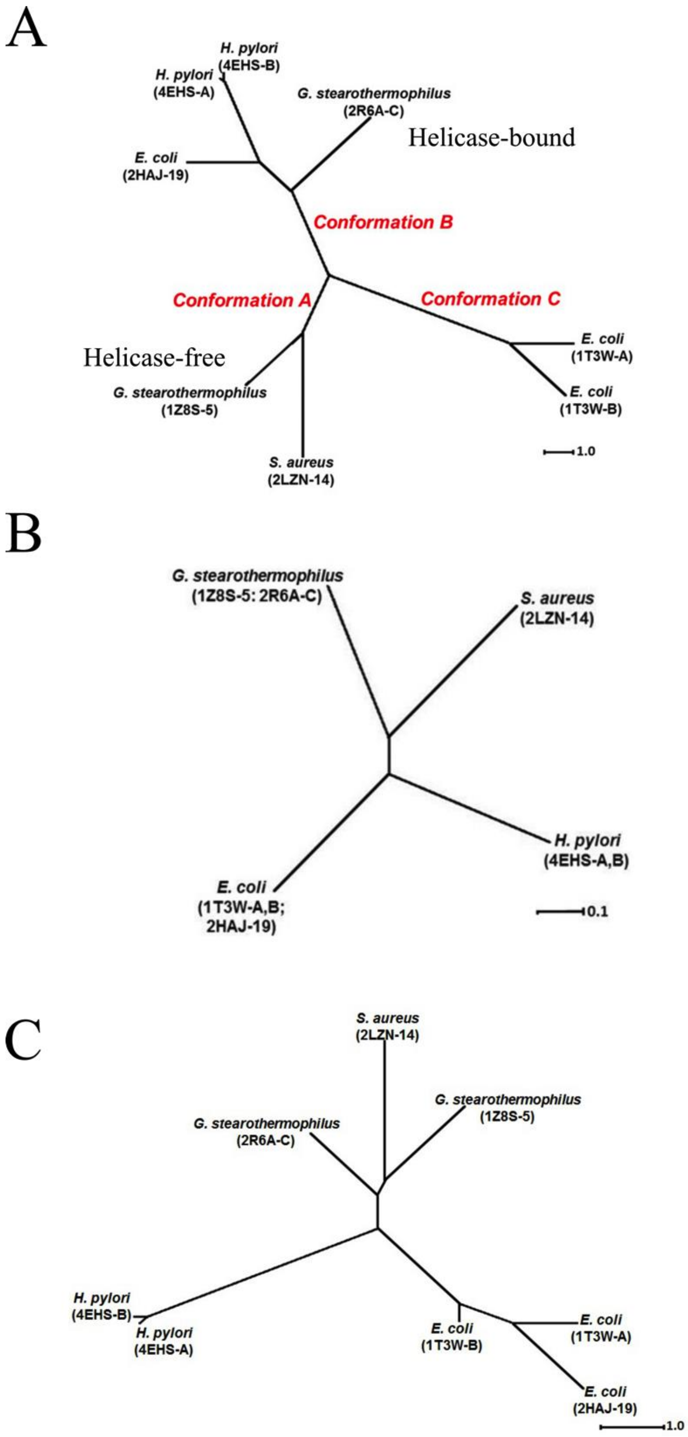


Figure 3: Unrooted phylograms. Representation of the primase CTD structure (A), sequence similarity (B) and primase CTD C1 subdomain structure (C). All trees were generated with SplitsTree4 Each tree contains the following species: *S. aureus* primase

CTD (this study)**,** *G. stearothermophilus* primase CTD**,** *H. pylori* primase CTD, and *E. coli* primase CTD.

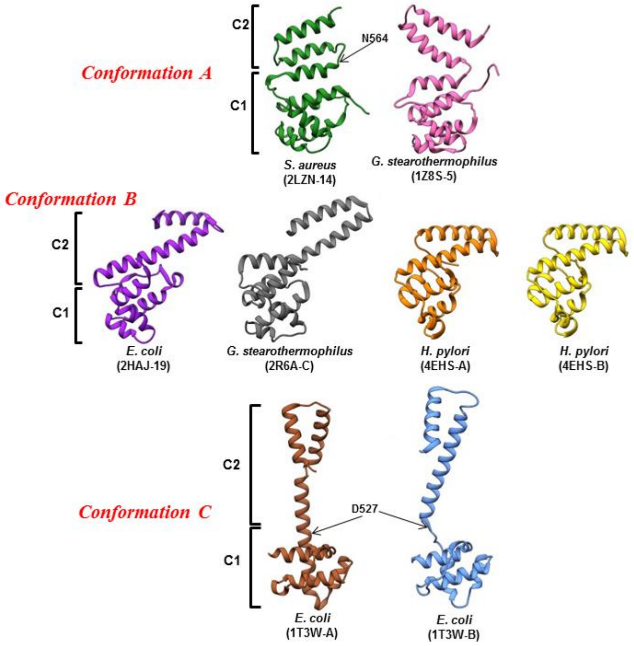


Figure 4: Primase CTD conformation classes. The species structures were categorized in three conformations from Figure 3A based on compactness: A compact, B intermediate, C extended. Conformation C may be biologically irrelevant.[17]

6.3.4 Primase CTD flexibility

An NMR dynamics analysis of the *S. aureus* primase CTD structure was conducted to complement the structural analysis. The NMR relaxation parameters T_1 and T_2, and the relative ratio of NOE enhancement were measured on a per residue basis and analyzed using the Fast-Model free program[46] to calculate order parameters (S^2) and chemical exchange (R_{ex}) values (Figure 5). In general, the T_1 values for the C1 subdomain residues were higher than the T_1 values for the C2 subdomain residues, indicating that the two subdomains exhibit different dynamics. A similar observation was made for the T_2 values, except that the T_2 values for the C1 subdomain were lower than the T_2 values for C2.

It is important to note that Fast-Model free did not converge on a result when using the entire primase CTD structure. This is consistent with the observation that the C1 and C2 subdomains behave as two separate domains and are dynamically independent. A convergent result was only obtained when the NMR dynamics data of the C1 and C2 subdomains were modeled separately. The C1 subdomain included residues 463-572 and the C2 subdomain included residues 565-605. The linker region, corresponding to residues 565-572, was included in the dynamics analysis for both subdomains.

The order parameters (S^2) for the C1 subdomain residues covered a large range (0.240-1.000) whereas the C2 subdomain residues were close to 1 (Figure 5D). The average S^2 values were 0.79 ± 0.02 and 0.97 ± 0.03 for the C1 and C2 domains, respectively. In general, an S^2 value approaching 1.0 indicates a highly ordered residue with limited flexibility, while a lower S^2 value indicates greater flexibility. This implies that the C1 subdomain has a greater overall flexibility relative to C2.

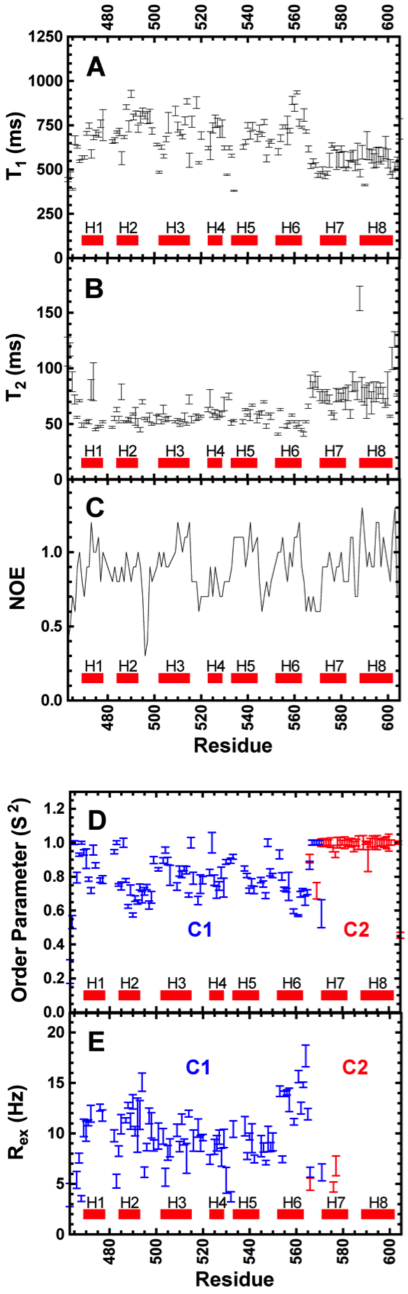


Figure 5: Dynamics of *S. aureus* primase CTD. The NMR relaxation parameters T_1 (A) and T_2 (B), NOE enhancements (C), S^2 order parameter (D), and R_{ex} chemical exchange rate are plotted per residue. The C1 subdomain included residues 463-572 and the C2

subdomain included residues 565-605. The linker region corresponding to residues 565-572 was included in the dynamics analysis for both subdomains.

Within the primase CTD structure there are two stretches of three or more residues with S^2 values less than 0.75: residues 488 to 498; residues 555 to 557 and residues 561 to 563. Residues 488 to 498 lie in helix 2 and the linker between helix 2 and 3. Residues 555 to 557 and 561 to 563 lie in helix 6 (Table S5). The flexibility of the linker between helix 2 and 3 is consistent with the variable lengths for helix 3 observed in the structures from other organisms. It is possible that the reduction in the flexibility of helix 3 is part of an allosteric system that provides an entropic energy change when helix 6 and 8 bind to the helicase NTD.[65] The flexibility of the second half of helix 6 shows how the unstructured linker between helix 6 and 7 may act as a nucleus for unraveling helix 6. Surprisingly, the linker region corresponding to residues 565-572 was observed to be dynamic, indicating that the point of flexibility between the two domains is located at the end of helix 6. This finding further suggests that the dynamic, second half of helix 6 (residues 561-564) compensates by either unraveling or bending for the increased motion between the two subdomains at the linker region.

To identify the hinge between the two subdomains, we searched for a residue with a low order parameter and fast chemical exchange. N564 had one of the lowest order parameters (0.658) and the fastest chemical exchange (17.7 Hz), indicating a high degree of motion on both the ps-ns and ms timescales. It is the final residue of helix 6 and is located between a very rigid C2 subdomain and a flexible C1 domain, potentially implicating N564 as a key residue for a conformational change.

In addition to per-residue order parameters and chemical exchange rates, an overall correlation time (τ_m) was calculated for each subdomain. The correlation times for the C1 and C2 subdomains were 7.2 ns and 6.4 ns, respectively. The predicted value for a hydrated protein with a molecular weight of 17.2 kDa, corresponding to the intact primase CTD structure, is 7.2 ns, where $\square_m \approx MW/2400$.[66] The predicted tumbling times were 5.45 ns and 2.1 ns for C1 (13.1 kDa) and C2 (5.0 kDa). The observed τ_m values for both subdomains are substantially higher than predicted and share a greater degree of similarity than the predicted values. This indicates the C1 and C2 subdomains are both tumbling at approximately the same rate as the intact protein. The model that emerges is one in which the two subdomains can adopt a wide range of relative orientations, with N564 as a pivot point, but share an overall tumbling rate.

6.3.5 Identification of ligands that bind to *S. aureus* primase CTD

A high-throughput NMR ligand affinity screen of the *S. aureus* primase CTD was undertaken to identify potential binding compounds. Out of the 423 compounds tested, a total of 12 compounds were shown to bind *S. aureus* primase CTD using a 1D ^{1}H NMR line-broadening screen: acycloguanosine, 3-aminopropionitrile fumarate, chelerythrine chloride, didecyldimethylammonium bromide, 5,5-diphenylhydantoin, L-histidine, (±)-α-lipoamide, 1-methylimidazole, mitoxantrone dihydrochloride, myricetin, (±)-propranolol hydrochloride, sodium creatine phosphate, and sodium DL-lactate. Of these compounds, acycloguanosine and myricetin were further analyzed for binding with primase CTD from *S. aureus*. Furthermore, adenosine was selected for binding studies due to its structural similarity to acycloguanosine. Quercetin, luteolin and kaempferol were also selected for their structural similarity to myricetin, but their low solubility prevented further study.

A 2D ^{15}N-^{1}H HSQC spectrum was collected for free primase and for the primase-ligand complex using acycloguanosine, adenosine, or myricetin. All three compounds were shown to bind primase CTD based on protein chemical shift perturbations (CSPs) of Y552, N564, T570, L574, E580, I584, G585, L589, and Q590 (Figure 6A). Although initially unassigned, Y552 was discovered during the titration experiment and assigned based on spatial proximity to other perturbed residues. None of the three compounds reached CSP saturation at the highest concentration tested ($\geq$ 4.0 mM), indicating that their K_D's were greater than 4 mM. Acycloguanosine, adenosine and myricetin exhibited CSPs above one standard deviation from the average of all non-binding site residues. Specifically, acycloguanosine, adenosine and myricetin caused CSPs of 0.35 ± 0.09 ppm, 0.16 ± 0.29 ppm and 0.07 ± 0.02 ppm, respectively, for the binding site residues.

When the nine residues incurring CSPs were mapped onto the structure of primase CTD, eight of the residues were located adjacent to one another on helix 6 or 7 (Figure 6B). The residue furthest from the other binding site residues was N564, which was previously identified as a key residue for a C1-C2 conformation change (Figure 5). When adenosine was docked into the groove between helices 6 and 7 (Figure 6B), the adenine moiety filled a hydrophobic pocket while its ribose was exposed to solvent and made specific interactions with three residues: Y552, I584, and G585 (Figure 6C). All three compounds docked into the same primase CTD pocket and adopted a uniform binding conformation.

The hydrophobic fused rings were buried deep into the pocket and the hydrophilic moieties interacted with the same three residues at the opening of the pocket (Figure 6D). Since these compounds bind to the closed conformation of primase CTD, they are likely to act as an inhibitor and reduce the helicase-primase interaction.

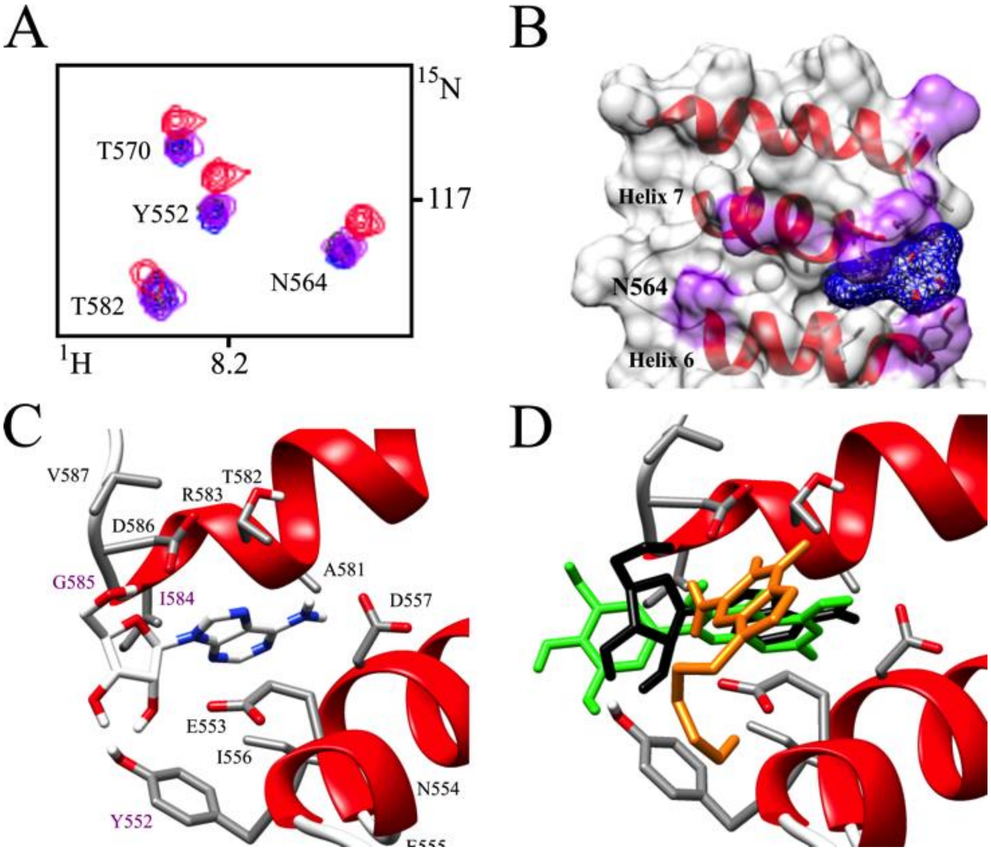


Figure 6: Adenosine binding site identification. When *S. aureus* primase CTD (1 mM) was titrated with up to 4.0 mM adenosine, nine residues exhibited chemical shift perturbations in the 2D ^{1}H-^{13}C HSQC spectrum of primase CTD. A detail from the overlay of the 2D ^{1}H-^{13}C HSQC spectra is colored according to the addition of 0 mM (blue), 2.6 mM (purple) and 4.0 mM (red) of adenosine (A). Expanded view of a surface rendition of the primase CTD NMR structure is superimposed onto a ribbon structure to

highlight secondary structure elements. Adenosine (blue tessellated structure) was docked into the primase CTD region defined by the nine residues with significant chemical shift perturbations (colored purple), which are located between helices 6 and 7. The side chains for all the residues that interact with adenosine are drawn with licorice bonds and are the same side chains displayed in panel C. Residue N564 and □-helices 6 and 7 are labeled (B). An expanded view of the ligand-binding pocket (flipped 180 degree from panel B) represented as a ribbon structure with side chains interacting with adenosine shown as licorice bonds. The residues interacting with adenosine are labeled. The labels for residues with significant CSPs are colored purple (C). Superposition of the docked conformations of myricetin (green), adenosine (black), and acycloguanosine (orange) into the same expanded view of the primase CTD binding site from panel C (D).

6.4 Discussion

We report the first primase CTD structure from a medically relevant, mesophilic *Firmicutes* bacterium. Its structure shares many features with the known thermophilic *Firmicutes* and mesophilic proteobacterial primase CTD. It has an N-terminal 6-helix bundle called the C1 subdomain connected by linker residues to a C-terminal helix-turn-helix called the C2 subdomain. The C1 subdomains from different species share the same overall fold, although there are differences in the lengths of several helices such that *S. aureus* and *G. stearothermophilus* are more similar to each other than to *E. coli* or *H. pylori*. The overall resolution of the *S. aureus* structure was low due to significant peak overlap in the NMR spectra, even at a field of 900 MHz. The high leucine and □-helical content of the CTD, coupled with a high degree of dynamic motion, led to peak broadening such that only 56% of side chain Hγ and 76% Hδ could be unambiguously assigned. The number of long range (>5 Å) NOEs was also lower than anticipated (140) for a protein of 19.6 kDa. Nevertheless, the overall rmsd of the 20 lowest energy structures was reasonable at 1.54 (± 0.19) Å and the ensemble of structures lies in acceptable ranges of common structure validation programs (all Z-scores above -5.0).

When a representative structure for *S. aureus* primase CTD was compared to the x-ray and NMR CTD structures from other organisms, three conformational classes were observed (Figure 3A, 4). Two of the conformations are likely related to the function of primase. The conformation of the *E. coli* 1T3W structure is probably not biologically relevant given the low pH of its crystallization and some strong interface packing effects.[17] Of all the structures, the *G. stearothermophilus* crystal structure 2R6A is the only one bound to helicase.[16] This suggests that the structures of *H. pylori* (4EHS) and *E. coli* (2HAJ) are also helicase-bound forms. Since conformation B contains members from both *Firmicutes* and *Proteobacteria*, it indicates that the helicase-bound form of primase CTD is not phylum-dependent. Since the solution structures of *G. stearothermophilus* and *S. aureus* primase CTD are in conformation A, it appears to be the conformation adopted when primase dissociates from helicase. In the case of *S. aureus*, it has also been possible to use NMR dynamics to identify N564 as the key hinge residue that allows helix 6 to bend into helices 6 and 7.

Three small molecules, acycloguanosine, adenosine, and myricetin, were discovered to bind a common site on the closed conformation of primase CTD. The ligand-binding pocket is only created when helices 6 and 7 are adjacent to one another. Among the nine residues with CSPs that define this ligand-binding site, N564 exhibited the largest CSP and was also identified as a key residue in a conformational change between the open and closed form of primase CTD. As a result, this ligand-binding site is a potential target for the further development of small molecule inhibitors that may “lock” primase CTD in its closed form and prevent its interaction with the NTD of the helicase. In fact, two of the compounds shown to bind primase CTD have known inhibitory effects on DNA replication proteins, although their known mode of actions are different from the one identified here. Acycloguanosine is a known inhibitor of the primase-helicase interaction in herpes simplex virus and acts as a chain terminator.[67, 68] Additionally, myricetin inhibits bacterial helicases with an IC_{50} of 10 □M[69] by competing with the ATPase active sites.

Since the *S. aureus* CTD structure shares many features with the *G. stearothermophilus* solution structure, we used the *G. stearothermophilus* DnaB/primase CTD co-crystal structure[16] to identify potential interface residues. The co-crystal structure of the *G. stearothermophilus* primase CTD C2 subdomain (Figure 7A,C) contains five nonpolar residues (F577, L578, A581, A584, and I588) in the final helix that could make contact with DnaB NTD1 (L67, A72, A75, L80 and V86) (Table S6). Following this idea, we created a homology model of the *S. aureus* primase CTD and DnaB NTD (Figure 7B, D) and found there were conserved but not identical residues in the equivalent positions. The *S. aureus* CTD C2 subdomain residues were V598, L594, K591, E588 and V587, and the *S. aureus* DnaB NTD1 residues were V67, D72, S75, L80, and P86. It has been previously shown that a C-terminal fragment containing the last three helices of the *G. stearothermophilus* C2 subdomain[6] exhibited sub-nanomolar binding affinity to DnaB, but was not capable of stimulating the ATPase activity of DnaB. It could be that DnaG and DnaB from these two organisms interact using spatially similar, hydrophobic, but differently sized residues. The DnaG and DnaB structures may have coevolved to preserve the primase CTD interaction in a manner that led to species-specificity.

The observation that the primase CTD-helicase structural interface is species-specific is consistent with previous studies showing that *S. aureus* helicase will only stimulate primer synthesis when incubated with the cognate primase.[6] Stimulation of DnaB ATPase activity by primase binding has been shown to involve the C1 subdomain Y548 in *G. stearothermophilus*[70] and F534 in the *H. pylori*[20] C1 subdomain. Mutation of these residues to an alanine reduced binding affinity and ATPase stimulation. An examination of the *G. stearothermophilus* primase CTD structure shows that Y548 and the adjacent residue D547 in helix 6 of the C1 subdomain make contact with P97, T98, and N101 in the DnaB NTD2 (Figure 7C, Table S5). The structural equivalent residues in *S. aureus* C1 subdomain are D558 and Y559 and within the *S. aureus* DnaB NTD2 are also P97, T98, and N101 (Figure 7D). Since the residues are identical between *G. stearothermophilus* and *S. aureus*, it indicates the mechanism of ATPase stimulation is highly conserved.

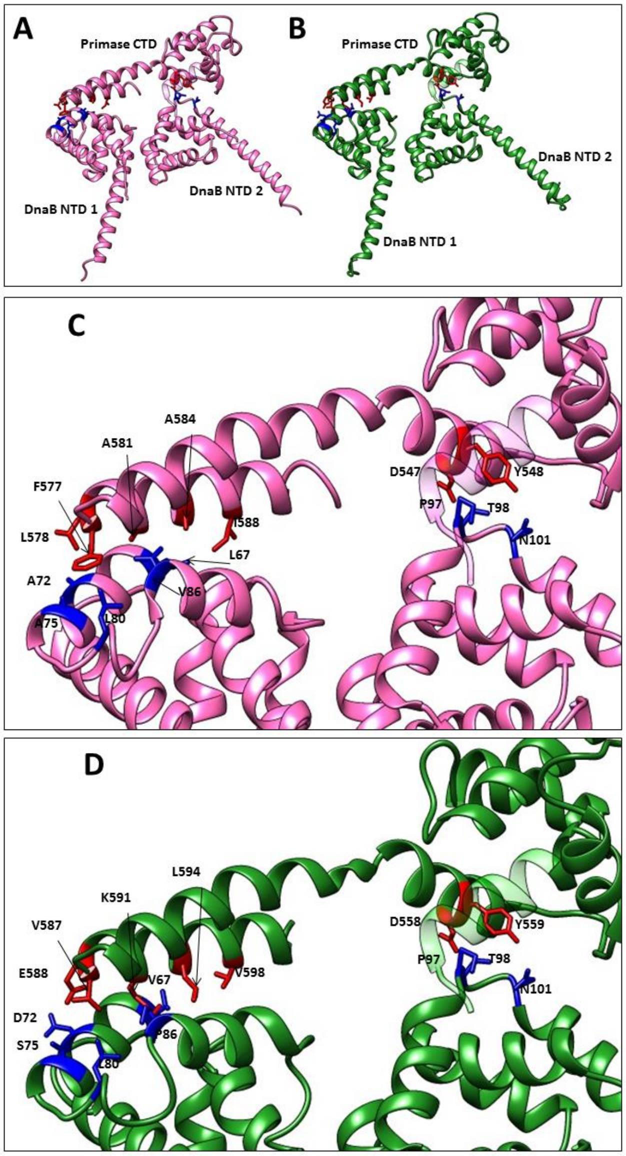


Figure 7: Critical interface residues between primase and helicase. The co-crystal structure of the complex from *G. stearothermophilus* involves two DnaB NTDs, one of which interacts with the C1 subdomain of the primase and the other with the C2 subdomain (A). The specific interface residues are shown in an expanded view (C). This was the template for the homology model of *S. aureus* primase CTD and DnaB NTDs (B), which is also shown in an expanded view (D).

6.5 References

[1] Griep, M. A., and Periago, J. (2017) Primase, In *Reference Module in Life Sciences*, pp 1-6, Elsevier, New York.

[2] Frick, D. N., and Richardson, C. C. (2001) DNA Primases, *Annual review of biochemistry 70*, 39-80.

[3] Johnson, S. K., Bhattacharyya, S., and Griep, M. A. (2000) DnaB Helicase Stimulates Primer Synthesis Activity on Short Oligonucleotide Templates, *Biochemistry 39*, 736-744.

[4] Chintakayala, K., Larson, M. A., Grainger, W. H., Scott, D. J., Griep, M. A., Hinrichs, S. H., and Soultanas, P. (2007) Domain Swapping Reveals that the C- and N-Terminal Domains of DnaG and DnaB, Respectively, are Functional Homologues, *Molecular microbiology 63*, 1629-1639.

[5] Keck, J. L., and Berger, J. M. (2001) Primus Inter Pares (First Among Equals), *Nature structural biology 8*, 2-4.

[6] Syson, K., Thirlway, J., Hounslow, A. M., Soultanas, P., and Waltho, J. P. (2005) Solution Structure of the Helicase-Interaction Domain of the Primase DnaG: A Model for Helicase Activation, *Structure 13*, 609-616.

[7] Koepsell, S. A., Larson, M. A., Griep, M. A., and Hinrichs, S. H. (2006) *Staphylococcus aureus* Helicase but not *Escherichia coli* Helicase Stimulates *S.*

aureus Primase Activity and Maintains Initiation Specificity, *Journal of bacteriology 188*, 4673-4680.

[8] Urlacher, T. M., and Griep, M. A. (1995) Magnesium Acetate Induces a Conformational Change in Escherichia coli Primase, *Biochemistry 34*, 16708-16714.

[9] Tougu, K., and Marians, K. J. (1996) The Extreme C Terminus of Primase is Required for Interaction with DnaB at the Replication Fork, *The Journal of biological chemistry 271*, 21391-21397.

[10] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016) The Pfam Protein Families Database: Towards a More Sustainable Future, *Nucleic Acids Res 44*, D279-285.

[11] Corn, J. E., Pease, P. J., Hura, G. L., and Berger, J. M. (2005) Crosstalk Between Primase Subunits Can Act to Regulate Primer Synthesis in Trans, *Molecular cell 20*, 391-401.

[12] Podobnik, M., McInerney, P., O'Donnell, M., and Kuriyan, J. (2000) A TOPRIM Domain in the Crystal Structure of the Catalytic Core of *Escherichia coli* Primase Confirms a Structural Link to DNA Topoisomerases, *Journal of molecular biology 300*, 353-362.

[13] Aravind, L., Leipe, D. D., and Koonin, E. V. (1998) Toprim--A Conserved Catalytic Domain in Type IA and II Topoisomerases, DnaG-type Primases, OLD Family Nucleases and RecR Proteins, *Nucleic acids research 26*, 4205-4213.

[14] Soultanas, P. (2005) The Bacterial Helicase-Primase Interaction: A Common Structural/Functional Module, *Structure 13*, 839-844.

[15] Pan, H., and Wigley, D. B. (2000) Structure of the Zinc-Binding Domain of *Bacillus stearothermophilus* DNA Primase, *Structure 8*, 231-239.

[16] Bailey, S., Eliason, W. K., and Steitz, T. A. (2007) Structure of Hexameric DnaB Helicase and its Complex with a Domain of DnaG Primase, *Science 318*, 459-463.

[17] Su, X. C., Schaeffer, P. M., Loscha, K. V., Gan, P. H., Dixon, N. E., and Otting, G. (2006) Monomeric Solution Structure of the Helicase-Binding Domain of *Escherichia coli* DnaG Primase, *The FEBS journal 273*, 4997-5009.

[18] Oakley, A. J., Loscha, K. V., Schaeffer, P. M., Liepinsh, E., Pintacuda, G., Wilce, M. C., Otting, G., and Dixon, N. E. (2005) Crystal and Solution Structures of the Helicase-Binding Domain of *Escherichia coli* Primase, *The Journal of biological chemistry 280*, 11495-11504.

[19] Larson, E. (2007) Community Factors in the Development of Antibiotic Resistance, *Annual review of public health 28*, 435-447.

[20] Abdul Rehman, S. A., Verma, V., Mazumder, M., Dhar, S. K., and Gourinath, S. (2013) Crystal Structure and Mode of Helicase Binding of the C-Terminal Domain of Primase from *Helicobacter pylori*, *Journal of bacteriology 195*, 2826-2838.

[21] Mercier, K. A., Germer, K., and Powers, R. (2006) Design and Characterization of a Functional Library for NMR Screening Against Novel Protein Targets, *Combinatorial chemistry & high throughput screening 9*, 515-534.

[22] Shortridge, M. D., Griep, M. A., and Powers, R. (2012) (1)H, (1)(3)C, and (1)(5)N NMR Assignments for the Helicase Interaction Domain of *Staphylococcus aureus* DnaG Primase, *Biomolecular NMR assignments 6*, 35-38.

[23] Sattler, M., Schleucher, J., and Griesinger, C. (1999) Heteronuclear Multidimensional NMR Experiments for the Structure Determination of Proteins in Solution Employing Pulsed Field Gradients, *Prog Nucl Mag Res Sp 34*, 93-158.

[24] Hwang, T. L., Mori, S., Shaka, A. J., and vanZijl, P. C. M. (1997) Application of Phase-Modulated CLEAN Chemical EXchange Spectroscopy (CLEANEX-PM) to Detect Water-Protein Proton Exchange and Intermolecular NOEs, *J Am Chem Soc 119*, 6203-6204.

[25] Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009) TALOS+: A Hybrid Method for Predicting Protein Backbone Torsion Angles from NMR Chemical Shifts, *Journal of biomolecular NMR 44*, 213-223.

[26] Vuister, G. W., Wang, A. C., and Bax, A. (1993) Measurement of 3-Bond Nitrogen Carbon-J Couplings in Proteins Uniformly Enriched in N-15 and C-13, *J Am Chem Soc 115*, 5334-5335.

[27] Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: A Multidimensional Spectral Processing System Based on UNIX Pipes, *Journal of biomolecular NMR 6*, 277-293.

[28] Garrett, D. S., Powers, R., Gronenborn, A. M., and Clore, G. M. (1991) A Common-Sense Approach to Peak Picking in 2-Dimensional, 3-Dimensional, and 4-Dimensional Spectra Using Automatic Computer-Analysis of Contour Diagrams, *J Magn Reson 95*, 214-220.

[29] 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., and Laue, E. (2002) The CCPN Project: An Interim Report on a Data Model for the NMR Community, *Nature structural biology 9*, 416-418.

[30] Huang, Y. J., Tejero, R., Powers, R., and Montelione, G. T. (2006) A Topology-Constrained Distance Network Algorithm for Protein Structure Determination from NOESY Data, *Proteins 62*, 587-603.

[31] Clore, G. M., Appella, E., Yamada, M., Matsushima, K., and Gronenborn, A. M. (1990) Three-Dimensional Structure of Interleukin 8 in Solution, *Biochemistry 29*, 1689-1696.

[32] Schwieters, C. D., Kuszewski, J. J., Tjandra, N., and Clore, G. M. (2003) The Xplor-NIH NMR Molecular Structure Determination Package, *J. Magn. Reson. 160*, 65-73.

[33] Schwieters, C. D., Kuszewski, J. J., and Clore, G. M. (2006) Using Xplor-NIH for NMR Molecular Structure Determination, *Prog Nucl Mag Res Sp 48*, 47-62.

[34] Garrett, D. S., Kuszewski, J., Hancock, T. J., Lodi, P. J., Vuister, G. W., Gronenborn, A. M., and Clore, G. M. (1994) The Impact of Direct Refinement Against Three-Bond HN-CalphaH Coupling Constants on Protein Structure Determination by NMR, *Journal of magnetic resonance. Series B 104*, 99-103.

[35] Kuszewski, J., Qin, J., Gronenborn, A. M., and Clore, G. M. (1995) The Impact of Direct Refinement Against 13Calpha and 13Cbeta Chemical Shifts on Protein Structure Determination by NMR, *Journal of magnetic resonance. Series B 106*, 92-96.

[36] Kuszewski, J., Gronenborn, A. M., and Clore, G. M. (1996) Improving the Quality of NMR and Crystallographic Protein Structures by Means of a Conformational Database Potential Derived from Structure Databases, *Protein science : a publication of the Protein Society 5*, 1067-1080.

[37] Kuszewski, J., Gronenborn, A. M., and Clore, G. M. (1997) 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. 125*, 171-177.

[38] Kuszewski, J., and Clore, G. M. (2000) Sources of and Solutions to Problems in the Refinement of Protein NMR Structures Against Torsion Angle Potentials of Mean Force, *J. Magn. Reson. 146*, 249-254.

[39] 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., and Montelione, G. T. (2005) 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 science : a publication of the Protein Society 14*, 2849-2861.

[40] Brunger, A. T. (2007) Version 1.2 of the Crystallography and NMR System, *Nature protocols 2*, 2728-2733.

[41] Linge, J. P., and Nilges, M. (1999) Influence of Non-Bonded Parameters on the Quality of NMR Structures: A New Force Field for NMR Structure Calculation, *Journal of biomolecular NMR 13*, 51-59.

[42] 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., and Bonvin, A. M. (2005) RECOORD: A Recalculated Coordinate Database of 500+ Proteins from the PDB Using Restraints from the BioMagResBank, *Proteins 59*, 662-672.

[43] Kay, L. E., Torchia, D. A., and Bax, A. (1989) Backbone Dynamics of Proteins as Studied by 15N Inverse Detected Heteronuclear NMR Spectroscopy: Application to Staphylococcal Nuclease, *Biochemistry 28*, 8972-8979.

[44] 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., and Kay, L. E. (1994) Backbone Dynamics of a Free and Phosphopeptide-Complexed Src Homology 2 Domain Studied by 15N NMR Relaxation, *Biochemistry 33*, 5984-6003.

[45] Mandel, A. M., Akke, M., and Palmer, A. G., 3rd. (1995) Backbone Dynamics of *Escherichia coli* Ribonuclease HI: Correlations with Structure and Function in an Active Enzyme, *Journal of molecular biology 246*, 144-163.

[46] Cole, R., and Loria, J. P. (2003) FAST-Modelfree: A Program for Rapid Automated Analysis of Solution NMR Spin-Relaxation Data, *Journal of biomolecular NMR 26*, 203-213.

[47] Lipari, G., and Szabo, A. (1982) Model-Free Approach to the Interpretation of Nuclear Magnetic-Resonance Relaxation in Macromolecules .1. Theory and Range of Validity, *J Am Chem Soc 104*, 4546-4559.

[48] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera--A Visualization System for Exploratory Research and Analysis, *Journal of computational chemistry 25*, 1605-1612.

[49] Huson, D. H., and Bryant, D. (2006) Application of Phylogenetic Networks in Evolutionary Studies, *Molecular biology and evolution 23*, 254-267.

[50] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, *Nucleic Acids Research 28*, 235-242.

[51] Edgar, R. C. (2004) MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput, *Nucleic acids research 32*, 1792-1797.

[52] Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment, *Journal of molecular biology 302*, 205-217.

[53] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D., and Higgins, D. G. (2011) Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega, *Molecular systems biology 7*, 539.

[54] Mercier, K. A., Baran, M., Ramanathan, V., Revesz, P., Xiao, R., Montelione, G. T., and Powers, R. (2006) FAST-NMR: Functional Annotation Screening Technology Using NMR Spectroscopy, *J Am Chem Soc 128*, 15292-15299.

[55] Hwang, T. L., and Shaka, A. J. (1995) Water Suppression That Works - Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients, *J Magn Reson Ser A 112*, 275-279.

[56] Johnson, P. E., Tomme, P., Joshi, M. D., and McIntosh, L. P. (1996) Interaction of Soluble Cellooligosaccharides with the N-Terminal Cellulose-Binding Domain of Cellulomonas fimi CenC. 2. NMR and Ultraviolet Absorption Spectroscopy, *Biochemistry 35*, 13895-13906.

[57] Barril, X. (2013) Druggability Predictions: Methods, Limitations, and Applications, *WIREs Comput. Mol. Sci. 3*, 327-338.

[58] Halouska, S., Zhou, Y., Becker, D. F., and Powers, R. (2009) Solution Structure of the *Pseudomonas putida* Protein PpPutA45 and its DNA Complex, *Proteins 75*, 12-27.

[59] Clore, G. M., Nilges, M., Sukumaran, D. K., Brunger, A. T., Karplus, M., and Gronenborn, A. M. (1986) The Three-Dimensional Structure of Alpha1-purothionin in Solution: Combined Use of Nuclear Magnetic Resonance, Distance Geometry and Restrained Molecular Dynamics, *The EMBO journal 5*, 2729-2735.

[60] Nilges, M., Gronenborn, A. M., Brunger, A. T., and Clore, G. M. (1988) 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 2*, 27-38.

[61] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983) Charmm - A Program for Macromolecular Energy,

Minimization, and Dynamics Calculations, *Journal of computational chemistry 4*, 187-217.

[62] Laskowski, R. A., Macarthur, M. W., Moss, D. S., and Thornton, J. M. (1993) Procheck - A Program to Check the Stereochemical Quality of Protein Structures, *J Appl Crystallogr 26*, 283-291.

[63] Bhattacharya, A., Tejero, R., and Montelione, G. T. (2007) Evaluating Protein Structures Determined by Structural Genomics Consortia, *Proteins 66*, 778-795.

[64] Shortridge, M. D., Triplet, T., Revesz, P., Griep, M. A., and Powers, R. (2011) Bacterial Protein Structures Reveal Phylum Dependent Divergence, *Computational biology and chemistry 35*, 24-33.

[65] Motlagh, H. N., Wrabl, J. O., Li, J., and Hilser, V. J. (2014) The Ensemble Nature of Allostery, *Nature 508*, 331-339.

[66] Cantor, C. R., and Schimmel, P. R. (1980) Biophysical Chemistry Part II: Techniques for the Study of Biological Strcuture and Function, p 461, W. H. Freeman and Co., San Francisco.

[67] Crute, J. J., Grygon, C. A., Hargrave, K. D., Simoneau, B., Faucher, A. M., Bolger, G., Kibler, P., Liuzzi, M., and Cordingley, M. G. (2002) Herpes Simplex Virus Helicase-Primase Inhibitors are Active in Animal Models of Human Disease, *Nature medicine 8*, 386-391.

[68] 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., and Rubsamen-Waigmann, H. (2002) New Helicase-Primase Inhibitors as Drug Candidates for the Treatment of Herpes Simplex Disease, *Nature medicine 8*, 392-398.

[69] Griep, M. A., Blood, S., Larson, M. A., Koepsell, S. A., and Hinrichs, S. H. (2007) Myricetin Inhibits *Escherichia coli* DnaB Helicase but not Primase, *Bioorgan Med Chem 15*, 7203-7208.

[70] Chintakayala, K., Larson, M. A., Griep, M. A., Hinrichs, S. H., and Soultanas, P. (2008) Conserved Residues of the C-terminal p16 Domain of Primase are Involved in Modulating the Activity of the Bacterial Primosome, *Molecular microbiology 68*, 360-371.

6.6 Supplemental Figures and Tables

Table S1**:** Bacterial primase sequence alignment

According to the Clustal Omega FAQ page (http://www.ebi.ac.uk/Tools/msa/clustalo/help/faq.html#23)

* indicates positions which have a single, fully conserved residues.

: indicates conservation between groups of strongly similar properties – scoring > 0.5 in the Gonnet PAM 250 matrix.

. indicates conservation between groups of weakly similar properties – scoring =< 0.5 in the Gonnet PAM 250 matrix

Blue – indicates *S. aureus* zinc binding domain

Green – indicates *S. aureus* RNA polymerase domain

Red – indicates *S. aureus* C-terminal domain

```
 CLUSTAL multiple sequence alignment by MUSCLE (3.8)



Staphylococcus      --
MRIDQSIINEIKDKTDILDLVSEYVKLEKRGRNYIGLCPFHDEKTPSFTVSEDKQICH
Bacillus
MGNRIPEEVVEQIRTSSDIVEVIGEYVQLRKQGRNYFGLCPFHGENSPSFSVSSDKQIFH
Geobacillus
MGHRIPEETIEAIRRGVDIVDVIGEYVQLKRQGRNYFGLCPFHGEKTPSFSVSPEKQIFH
Francisella
MAKKVSNSFIKELVATADIVDVVSRYVNLKKTGKNYKGCCPFHNEKTPSFFVNPEKNFYH
Aquifex             -----
MSSDIDELRREIDIVDVISEYLNLEKVGSNYRTNCPFHPDDTPSFYVSPSKQIFK
Pseudomonas
MAGLIPQSFIDDLLNRTDIVEVVSSRIQLKKTGKNYSACCPFHKEKTPSFTVSPDKQFYY
E.coli
MAGRIPRVFINDLLARTDIVDLIDARVKLKKQGKNFHACCPFHNEKTPSFTVNGEKQFYH
Yersinia
MAGRIPRVFINDLLARTDIIDLIDARVKLKKQGKNYHACCPFHHEKTPSFTVNGEKQFYH
                             :. :    **::::.  ::* . * *:   **** :.:*** *. .*::

Staphylococcus      CFGCKKGGNVFQFTQEIKDISFVEAVKELGDRVNVAVDIEATQSNSNVQIASDDLQ-
MIE
Bacillus            CFGCGEGGNVFSFLMKMEGLAFTEAVQKLGERNGIAVA-
EYTSGQGQQEDISDDTVIMQQ
Geobacillus
CFGCGAGGNAFTFLMDIEGIPFVEAAKRLAAKAGVDLSVYELDVRGRDDGQTDEAKAMTE
Francisella         CFGCQASGDALTFVKNINKLEFIDAVKNLAEIVGKPVEYENYS-QEDIQKEQLYNK-
CIS
Aquifex             CFGCGVGGDAIKFVSLYEDISYFEAALELAKRYGKKL---------DLEKISKDEK-
VYV
Pseudomonas         CFGCGAGGNALGFVMDHDQLEFPQAVEELAKRAGMDVPREERGGRGHTPRQPTDSP-
LYP
E.coli              CFGCGAHGNAIDFLMNYDKLEFVETVEELAAMHNLEVPFEAGSGPSQIERHQRQT--
LYQ
Yersinia            CFGCGAHGNAVDFLMNYDRLEFVESIEELATMHGLEVPYEAGSGTTQIERHQRQS--
LYQ
                    ****   *:.. *    . : : ::   *.   .  :

Staphylococcus
MHELIQEFYYYALTKTVEGEQALTYLQERGFTDALIKERGIGFAPDSSHFCHDFLQKKGY
Bacillus
AHELLKKYYHHLLVNTEEGNEALSYLLKRGITKEMIEKFEIGYASPAWDAATKILQKRGL
Geobacillus
AHALLKRFYHHLLVHTKEGQAALDYLQARGWTKETIDRFEIGYAPDAPDAAAKLLESHSF
Francisella
FLAAAQKYYRWNLGNSVTKDKAINYLKKRGIDSSLAKFFGIGYSSEGWNNITELAKSINI
Aquifex             ALDRVCDFYRESLL---KNREASEYVKSRGIDPKVARKFDLGYAPSS-
EALVKVLKENDL
Pseudomonas
LLSAAAEFYKQALKSHPARKAAVNYLKGRGLTGEIARDFGLGFAPPGWDNLLKHLGGDNL
```

```
E.coli              LMDGLNTFYQQSLQ-
QPVATSARQYLEKRGLSHEVIARFAIGFAPPGWDNVLKRFGGNPE
Yersinia            LMESLSAFYQQSLK-
GQNAKQAREYLKHRGLSEEIIQHFAIGFAPPGWDNALKRFGRDGE
                           :*   *        *  *:  **          :*::. .     .

Staphylococcus      DIELAYEAGLLSRNEENFSYYDRFRNRIMFPLKNAQGRIVGYSG-RTYTG-
QEPKYLNSP
Bacillus            SLSSMEQAGLLIRSEKDGSHYDRFRGRVMFPIYTLQGKVIAFSG-RALGD-
DTPKYLNSP
Geobacillus         SLPVMEKAGLLTKK-EDGRYVGRFRNRIMFPIHDHRGETVGFSG-RLLGE-
GHPKYVNSP
Francisella         PEEILVDTGLAIKN-DKGNLYDRFRGRVMFPIRNIQGNVIAYGG-
RVTEDSDGVKYINSP
Aquifex             LEAYLETKNLLSPT--KGVYRDLFLRRVVIPIKDPRGRVIGFGGRRIVED-
KSPKYINSP
Pseudomonas         QLKAMLDAGLLVENSDTGKRYDRFRDRVMFPIRDSRGRIIAFGG-RVLGD-
DKPKYLNSP
E.coli              NRQSLIDAGMLVTN-DQGRSYDRFRERVMFPIRDKRGRVIGFGG-RVLGN-
DTPKYLNSP
Yersinia            SRTALNDAGMLVTN-DTGRTYDRFRERVMFPIRDKRGRVIAFGG-RILGD-
GVPKYLNSP
                            .:   .       . *  *:::*:   .*  :.:.* *
**:***

Staphylococcus      ETPIFQKRKLLYNLDKARKSIRK-----
LDEIVLLEGFMDVIKSDTAGLKNVVATMGTQL
Bacillus            ETPIFHKSKLLYNFHQARPFIRK-----
RGQVVLFEGYADVLAAVKSGVEEAVATMGTAL
Geobacillus         ETPVFRKGAILYHFHAARVPIRK-----
RQEALLVEGFADVISAAQAGIDYAIATMGTSL
Francisella
ETLVFQKNNILYGLYEYRERKKQYPDLINQSLVVVEGYMDVVGLAQHGFYAAVATLGTAF
Aquifex             DSRVFKKGENLFGLYEAKEYIKE-----
EGFAILVEGYFDLLRLFSEGIRNVVAPLGTAL
Pseudomonas         ETPVFHKGQELYGLYEARQKNRD-----
LDEIMVVEGYMDVIALAQQGIRNAVATLGTAT
E.coli              ETDIFHKGRQLYGLYEAQQDNAE-----
PNRLLVVEGYMDVVALAQYGINYAVASLGTST
Yersinia            ETEIFHKGRQLYGLYEAQVNHPN-----
PTRLLVVEGYMDVVALAQFGIDYAVASLGTAT
                    :: :*.*   *: :   .    .         ::.**: *::     *.  .:*.:**

Staphylococcus      SDEHITFIRKLTSNITLMFDGDFAGSEATLKTGQH----
LLQQGLNVFVIQLPSGMDPDE
Bacillus            TEEQAKLLRRNVETVVLCYDGDKAGREATMKAGQL----
LLQVGCQVKVTSLPDKLDPDE
Geobacillus         TEEQARIL-RPCDTITICYDGDRAGIEAAWAAAEQ----
LSALGCRVKVASLPNGLDPDE
Francisella         SPNHAKILFRETSSVILCFDGDEAGQKAALRTIKILLP-
MLDGNKKLKILTLPDKDDPDD
Aquifex             TQNQANLLSKFTKKVYILYDGDDAGR----
KAMKSAIPLLLSAGVEVYPVYLPEGYDPDE
Pseudomonas         SEEHIKRLFRLVPSILFCFDGDQAGRKAAWRALESVLP-
NLQDGKRVRFLFLPEGEDPDS
E.coli              TADHIQLLFRATNNVICCYDGDRAGRDAAWRALETALP-
YMTDGRQLRFMFLPDGEDPDT
Yersinia            TAEHIQLLFRATDNVICCYDGDRAGRDAAWRALETALP-
YLNDGRQLRFMFLPDGEDPDT
                    : ::   : .   .:   :*** **      : :         .  :    **.  ***
```

```
Staphylococcus      YIGKYGNDAFTTFVKNDKKSFAHYKVSILKDEIAHNDLS-
YERYLKELSHDISLMKSSIL
Bacillus            YVQQYGTTAFENLVKS-
SISFVGFKINYLRLGKNLQDESGKEEYVKSVLKELSLLQDAMQ
Geobacillus         YIRVYGGERFAGEAGC-
RRPLVAFKMAYLRRGKNLQHEGERLRYIDEALREIGKLSSPVE
Francisella         YIKKYGLERFLTALDN-
SLAVADFVIDNLIQGKDLRKAEAKAEVLENLKNFLADVEDNIY
Aquifex             FIKEFGKEELRRLINS-----SGELFETLIK----
TARENLEEKTREFRYYLGFISDGVR
Pseudomonas
LVRAEGEDAFRARITQQAQPLAEYFFQQLMLEADPATLEGKAHLATLAAPLLEKIPGNNL
E.coli              LVRKEGKEAFEARMEQ-
AMPLSAFLFNSLMPQVDLSTPDGRARLSTLALPLISQVPGETL
Yersinia            LVRKEGKDAFEQRMEA-
AQPLSTFLFETLMPQVDLSSPDGRAKLSTLALPLISQVPGEAL
                     :   *   :               .  *                      :  :  .

Staphylococcus      QQKAINDVAPFFNVSPEQLAN--------------------------------------
-
Bacillus            AESYLKSLSQEFSYSMETLLN--------------------------------------
-
Geobacillus         QDYYLRQLAEEFSLSLSALHEQLSRSQRE------------------------------
-
Francisella         SESITATIADKIGIKVEQFKN--------------------------------------
-
Aquifex             RFALASEFHTKYKVPMEILLM--------------------------------------
-
Pseudomonas
RLLMRQRLSEITGLSGENIGQLAHHSPPPSSMDHGASGVLDGDDYFAASAYYENEPSHAP
E.coli              RIYLRQELGNKLGILDDSQLE--------------------------------------
-
Yersinia            RLYLRQQLGNKLGLLDDSQLD--------------------------------------
-
                           .        .

Staphylococcus      --------EIQFNQAPANYYPEDEYGGYDEYGGY-
IEPEPIGMAQFDNLSRREKAERAFL
Bacillus            --------QLHQYRKEQKVQQKQVK---------
QVSKPSQIVQTKPKLTGFERAEREII
Geobacillus         --------RTKPREAPDGETARPMLA-------------------
KKLLPAFQNAERLLL
Francisella         --------LLKIRKQTTLNVNR-----------------------
QQNIQQKKLAKNLLL
Aquifex             --------KIEKNSQEKEIK-------------------------------
LSFKEKIFL
Pseudomonas
FDAAPGYVEAQPRKSWNKDKKPWDGKKWDGKKKWDKGGRGDFKAPQRTPVSVESTTLNAL
E.coli              --------RLMPKAAESGVS-------------------------
RPVPQLKRTTMRILI
Yersinia            --------KLMPKQIDNANT-------------------------
YQPPQLKRTTMRILI

:

Staphylococcus
KHLMRDKDTFLNYYESVDKDNFTNQHFKYVFEVLHDFYAENDQYNISDAVQYVNSNELRE
Bacillus            YHMLQSPEVAVRMESHI--
EDFHTEEHKGILYELYAYYEKGNEPSVGTFLSWLSDEKLKN
Geobacillus         AHMMRSRDVALVVQERIG-
GRFNIEEHRALAAYIYAFYEEGHEADPGALISRIPGELQPL
```

```
Francisella         EEFVLA-ELFVNIADFRILQHTND---------
FEIFATSKNLDILAKSLKILKEDSSNQ
Aquifex             KGLI---ELKPKID-------------------LEVLNLSPELKELA------------
-
Pseudomonas
RTLLHHPQLALKVDDAGTLAREQDTYAQLLVSLLEALQKNPRQSSMQLIARWHGTPQGRL
E.coli
GLLVQNPELATLVPPLENLDENKLPGLGLFRELVNTCLSQPGLTTGQLLEHYRGTNNAAT
Yersinia
GLLVQNPQLATLIPSLQGLEQAKLAGLPLFIELVETCLAQPGLTTGQLLELYRDNKFSQQ
                      ::   :                         .     .

Staphylococcus      -TLISLEQYNLNG---------EPYENEIDDYVNVINEK-
GQETIESLNHKLREATRIGD
Bacillus            IITDISTDEFINP---------
EYTEEVLQSHLETLRRHQEKLEKMEIIFKIKQMEKTDP
Geobacillus         ASDVSL--LLIAD---------DVSEQELEDYIRHVLNR--------
PKWLMLKVKEQEK
Francisella         IEAVILIQLLAEDYPDYREYFFELLSYGIRNTQKKYAED----
KYQEQMFVMLKRVENSS
Aquifex             VNALNGEEHLLPK---------EVLEYQV-DNLEKLFNN------------
ILRDLQKSG
Pseudomonas         LQALGEKEWLIVQ---------ENLEKQFFDTITKLSES--------
QRFGEREERLRSV
E.coli              LEKLSMWDDIADK---------NIAEQTFTDSLNHMFDS------------
LLELRQEEL
Yersinia            LETLATWNHMIVE---------DMVEPTFVDTLASLYDS------------
ILEQRQETL
                                          :  .  . .

Staphylococcus      VELQKYYLQQIVAKNKERM----------------
Bacillus            VEAAKYY-----VAYLQNQKARK------------
Geobacillus         TEAERRK----DFLTAAR-IAKEMIEMKKMLSSS-
Francisella         VKKRLKYLGSLPFRSDVQEMERKYLVAKLGNSNII
Aquifex             KKRKKRG-----LKNVNT-----------------
Pseudomonas         MQKSYSE-----LTDEEKALLREHYSVAASSPSQS
E.coli              IARERTH----GLSNEER---LELWTLNQELAKK-
Yersinia            IARDRTH----GLNAEER---KELWSLNLALARKK
```

Table S2: Pairwise structure comparison matrices for all eight primase C-terminal domain structures: (A) root mean square deviation (RMSD) and (B) number of aligned residues

A. RMSD

	E.coli 1T3W_A	E.coli 1T3W_B	E.coli 2HAJ-19	S.aureus 2LZN-14	G.stearo 1Z8S-5	G.stearo 2R6A_C	H.pylori 4EHS_A	H.pylori 4EHS_B
E.coli 1T3W_A	0	4.369	14.444	22.447	12.324	13.578	18.049	17.751
E.coli 1T3W_B	4.369	0	15.897	14.734	12.675	14.268	19.743	19.735
E.coli 2HAJ-19	14.444	15.897	0	12.162	11.226	5.870	13.047	12.891

S.aureus 2LZN-14	22.447	14.734	12.162	0	6.511	13.097	12.254	15.522
G.stearo 1Z8S-5	12.324	12.675	11.226	6.511	0	10.937	11.578	12.535
G.stearo 2R6A_C	13.578	14.268	5.870	13.097	10.937	0	8.780	9.014
H.pylori 4EHS_A	18.049	19.743	13.047	12.254	11.578	8.780	0	0.444
H.pylori 4EHS_B	17.751	19.735	12.891	15.522	12.535	9.014	0.444	0

B. Number of aligned residues

	E.coli 1T3W_A	E.coli 1T3W_B	E.coli 2HAJ-19	S.aureus 2LZN-14	G.stearo 1Z8S-5	G.stearo 2R6A_C	H.pylori 4EHS_A	H.pylori 4EHS_B
E.coli 1T3W_A	**134**	134	134	98	125	126	110	109
E.coli 1T3W_B	134	**135**	135	126	134	126	95	94
E.coli 2HAJ-19	134	135	**135**	125	132	126	95	94
S.aureus 2LZN-14	98	126	125	**143**	139	140	114	85
G.stearo 1Z8S-5	125	134	132	139	**146**	141	108	110
G.stearo 2R6A_C	126	126	126	140	141	**141**	111	110
H.pylori 4EHS_A	110	95	95	114	108	111	**124**	122
H.pylori 4EHS_B	109	94	94	85	110	110	122	**122**

Table S3: Residues of the twenty 2LZN ensemble structures that are in the Ramachandran plot disallowed regions

Model 1: Asn 91

Model 2: none

Model 3: His 37, Glu 127

Model 4: Glu 127, Thr 128

Model 5: His 37

Model 6: none

Model 7: Asp 53, Asn 91

Model 8: Thr 128

Model 9: none

Model 10: His 37

Model 11: Asn 91

Model 12: none

Model 13: none

Model 14: Thr 128

Model 15: Gln 126, Glu 127

Model 16: Thr 128

Model 17: His 37, Asn 91

Model 18: Gln 126

Model 19: Asn 91, Pro 109

Model 20: none

Table S4: Equivalent interface residues in the DnaC helicase N-terminal domain of *G. stearothermophilus* and *S. aureus*, where yellow highlights show interaction with primase C1 subdomain, and green highlight show interaction with primase C2 subdomain

```
Clustal run

  G.stear       MSELFSERIPPQSIEAEQAVLGAVFLDPAALVPASEILIPEDFYRAAHQKIFHAMLRVAD-
  60

  S.aureus      MDRMYEQNQMPHNNEAEQSVLGSIIIDPELINTTQEVLLPESFYRGAHQHIFRAMMHLNE

                *..::.:.  *:. ****:***::::**  :  :.*:*:**.***.***:**:**::: :
```

```
G.stear         RGEPVDLVTVTAELAASEQLEEIGGVSYLSELADAVPTAANVEYYARIVEEKSVLRRLIR-
120

S.aureus        DNKEIDVVTLMDQLSTEGTLNEAGGPQYLAELSTNVPTTRNVQYYTDIVSKHALKRRLIQ

                  : :*:**:  :*::.  *:* ** .**:**:  ***: **:**: **.:::: ****:


G.stear         TATSIAQDGYTREDEIDVLL

S.aureus        TADSIANDGYNDELELDAIL

                ** ***:***. * *:*.:*
```

Chapter 7

Enhanced Conformational Dynamics in a Parkinsonian Mutant of DJ-1 Causes Pathogenic Instability at Physiological Temperature

7.1 Introduction

Parkinson's disease results from the progressive death of dopaminergic neurons in the midbrain. While the cause of most cases of Parkinson's disease is unknown, the study of heritable forms of parkinsonism has helped elucidate the biochemical basis of dopaminergic neurodegeneration. Recessively inherited mutations in three proteins can cause parkinsonism: parkin (PARK2), PINK1 (PARK6) and DJ-1 (PARK7).[1] These three proteins are natively neuroprotective, particularly against mitochondrial dysfunction. Human DJ-1 is a conserved, ubiquitously expressed homodimeric protein that participates in multiple pathways[2-8] to protect cells against oxidative stress[9-12] and to maintain proper mitochondrial function.[13] In addition to parkinsonism, loss of DJ-1 mediated cytoprotection is also implicated in ischemia-reperfusion injury, as occurs in stroke and myocardial infarction.[14, 15] Despite abundant evidence that DJ-1 participates in multiple cytoprotective pathways, the details of its molecular activity remain incompletely understood.

Several pathogenic missense mutations in DJ-1 decrease protein stability, thereby causing disease.[3, 16] For example, the L166P, L10P, and P158Δ parkinsonian mutations disrupt the DJ-1 dimer, causing poor folding and proteolytic degradation of the protein.[1, 17]

However, several other disease-associated mutants in DJ-1 are more structurally benign[16, 18, 19] and thus the molecular basis of their pathogenicity is not understood.

M26I DJ-1 is an example of a mutant whose underlying molecular defect is obscure. Steady state levels of M26I DJ-1 are lower than wild-type protein in cell culture[3, 20] and *in vivo*,[21] indicating reduced protein stability in the cellular environment.[3, 20, 21] However, X-ray crystallography, NMR, and biophysical studies indicate that M26I DJ-1 is highly structurally similar to the wild-type protein and both proteins have comparable stability to thermal and chemical denaturation.[16, 17] Therefore, it is not clear from prior work why M26I DJ-1 is less stable in cells than the wild-type protein.

The absence of a structural explanation for the established pathogenicity of M26I DJ-1 suggests that the conformational dynamics of the protein might be changed by the mutation, potentially decreasing DJ-1 stability. The dynamical hypothesis for M26I DJ-1 pathogenesis is untested, as no comprehensive study of the dynamics of DJ-1 or any of its pathogenic mutants has been reported to date. However, recombinant M26I DJ-1 is prone to spontaneous aggregation *in vitro* that is not observed for the wild-type protein.[19, 22] This enhanced *in vitro* aggregation indicates that M26I DJ-1 has intrinsically reduced stability that would likely result in greater turnover and reduced steady-state levels of the protein *in vivo*. Such an effect in M26I DJ-1 would provide a rare example of a direct connection between aberrant protein dynamics and disease.

Here we show that the M26I mutation does not alter the picosecond-nanosecond timescale dynamics of DJ-1 but does increase protein flexibility on longer timescales. The enhanced conformational dynamics of M26I DJ-1 causes transient exposure of the hydrophobic core to solvent and results in pronounced aggregation near physiological temperatures (35-37°C) but not below. The dynamically driven instability of M26I DJ-1 at physiological temperature provides an explanation for why the M26I mutation destabilizes the protein in cells and causes loss of DJ-1 protection. These results also illustrate a common and underappreciated limitation of *in vitro* experiments performed at room temperature, where the functional consequences of protein dynamics may not be evident.

7.2 Materials and Methods

7.2.1 Protein Expression and Purification

Human wild-type and mutant (M26I, M26V, M26L, L101W, Y141W, M26I Y141W) DJ-1 proteins were expressed from pET15b (Novagen) constructs[23] and purified using Ni^{2+} affinity chromatography as previously described[23] except that all buffers contained freshly added 2.5 mM dithiothreitol (DTT). Uniformly ^{15}N-labeled proteins (wild-type, M26I, M26L DJ-1) were expressed as previously described.[24]

Purified proteins were dialyzed against storage buffer (25 mM HEPES pH 7.5, 100 mM KCl, 2.5mM DTT) overnight at 4°C and concentrated to 1 mM using an ultrafiltration concentrator with a 10 kDa MWCO regenerated cellulose membrane (Millipore), snap-frozen in small aliquots on liquid nitrogen, and stored at -80°C until needed. All proteins ran as a single band in overloaded Coomassie Blue stained SDS-PAGE gel and had correct intact masses determined using electrospray mass spectroscopy (Redox Biology

Center, UNL). Only DJ-1 that was >95% reduced at Cys106 was used for the experiments in this study.

7.2.2 NMR Sample Preparation

For NMR experiments, purified ^{15}N-labeled DJ-1 samples were dialyzed against 25 mM MES pH 6.5, 25 mM NaCl, 2.5 mM DTT and concentrated to 0.9 mM. The ^{15}N labeling efficiency as well as intact protein mass was determined using reverse phase liquid chromatography-mass spectrometry (LC-MS) as detailed previously.[16, 24] Proteins were diluted in water and analyzed using an Agilent 1200 LC system (Agilent Technologies) and a Q-Trap-4000 MS (AB Sciex) at the Metabolomics and Proteomics Core Facility in the Redox Biology Center (RBC) and the University of Nebraska-Lincoln (UNL). Samples for T_1, T_2, and NOE measurements were prepared by adding D_2O to a final concentration of 10% by addition of 99.9% D_2O (Sigma-Aldrich) and the sample was transferred to a 5 mm, high-throughput 7" standard series NMR tube (Norell).

Samples for NMR-detected H/D Exchange (HDX) experiments were dialyzed against 10 mM ammonium acetate pH 6.7 and 5 mM DTT at ~21°C. Samples were flash frozen with liquid nitrogen, lyophilized, and stored at -80°C until needed. To ensure that lyophilization did not cause irreversible changes in DJ-1, we compared the 2D ^{1}H-^{15}N HSQC spectra obtained from protein before and after lyophilization. No differences were observed. HDX was performed by resuspending the lyophilized protein in 25 mM MES pH 6.1, 25 mM NaCl, 2.5 mM DTT in 100% D_2O. The solution pH was adjusted to 6.1

with a 40% solution of NaOD (Sigma-Aldrich) and the pD calculated according to: $pD_{corrected}$ = pH + 0.4 (therefore the final pD of this buffer was 6.5).[25]

7.2.3 T_1, T_2 Relaxation and NOESY Experiments

All 2D ^1H-^{15}N heteronuclear single quantum coherence (HSQC) spectra were collected using a Bruker AVANCE DRX 500 MHz spectrometer equipped with a 5 mm triple resonance, Z-axis gradient cryoprobe with a cold proton channel at the University of Nebraska-Lincoln's Research Instrumentation Facility. 2D ^1H-^{15}N HSQC spectra acquired for T_1 (Bruker pulse sequence *hsqct1etf3gpsi*), T_2 (*hsqct2etf3gpsi*), and NOE (*hsqcnoeetf3gpsi*) measurements were collected at 35°C. The delays for T_1 (0.2696, 0.05392, 0.5392, 0.005392, 0.4044, 1.0784, 0.1348, 0.6740, 0.2696 sec) and T_2 (0.0176, 0.0352, 0.0528, 0.0704, 0.0880, 0.1056, 0.1231, 0.1408, 0.1584, 0.1760, 0.1936, 0.2112 sec) relaxation experiments were randomized to minimize systematic error. All spectra were collected using a total of 2048 and 256 data points and spectral sweep widths of 7309.942 and 2027.164 Hz along the proton and nitrogen dimensions, respectively. Free induction decays (FIDs) were Fourier transformed, zero-filled, phased, and baseline corrected using NMRPipe[26] and analyzed using NMRViewJ.[27] Chemical shift assignments for wild-type DJ-1 were retrieved from the Biological Magnetic Resonance Data Bank with accession number BMRB 17507.[17] Our 35°C spectra are shifted ~0.16 ppm in the hydrogen dimension compared to the previously collected 27°C spectra,[17] as expected with an increase in temperature.

The relaxation times (T_1, T_2) for each amide resonance were calculated by fitting a decay curve (see Eq. 1) to the intensity of each peak plotted against the associated delay (in msec), where I_t is the intensity of each peak at the specified delay time t, and I_0 is the initial steady state intensity. The inverse of the relaxation times produced relaxation rates (R_1, R_2) (converted to $\sec^{-1}$).

$$I_t = I_0 exp(-t/T_{1,2}) \qquad \text{[EQ. 1]}$$

$$R_{1,2} = 1/T_{1,2}$$

Interleaved 2D ^{1}H-^{15}N NOESY spectra were collected using 2048 and 128 data points along the direct and indirect axis, respectively, with 30 msec of proton saturation applied (or not) during a 5 sec relaxation interval. The NOE values were calculated by dividing the peak intensity for the saturated (I_{sat}) spectrum by the unsaturated (I_{unsat}) spectrum (see Eq. 2)

$$NOE = I_{sat} / I_{unsat} \qquad \text{[EQ. 2]}$$

The R_1, R_2 relaxation rates and NOE data were used to calculate the generalized order parameter (S^2), internal motion (τ_e), chemical exchange (R_{ex}), and the overall correlation time (τ_c) using Lipari-Szabo model free parameters in the FAST-Modelfree program

(Facile Analysis and Statistical Testing for Modelfree), interfaced with ModelFree 4.20.[28-30] The PDB coordinate files for the WT and mutant proteins were first input into PDBInertia to translate their centers of mass to the origin. For each R_2/R_1 list and respective PDB file output from PDBInertia, the R_2R_1-Diffusion program was used to predict the axially symmetric diffusion tensor based on the approach of Tjandra, *et al*.[31] The NMR spin-relaxation data (R_1, R_2, NOE) and the PDB obtained from R_2R_1-Diffusion was then input into the FAST-ModelFree software package.[30] To ensure that the dynamics analysis was not converging on a local minimum, the random seed value was changed for each run and the same seed value was never used more than once. Further, each FAST-ModelFree run was done in triplicate to verify that the output values were reproducible.

Temperatures for all NMR experiments were calibrated using 100% ethylene glycol as a temperature standard. Briefly, the separation between the OH resonances and the CH_2 resonances in ethylene glycol were measured (Δ, ppm) and the sample temperature calculated, according to the following equation: $T = (4.637 - \Delta) / 0.009967$, where T is the calculated temperature (Kelvin), and Δ is the shift difference (in ppm) between OH and CH_2 peaks (Bruker Instruments, Inc. VT-calibration manual).

7.2.4 H/D Exchange (HDX) NMR Spectroscopy

For HDX measurements, protonated ^{15}N-labelled proteins were lyophilized and resuspended in 500µL of 25 mM MES pH 6.5, 25 mM NaCl, 2.5 mM DTT in 100% D_2O. 2D ^{1}H-^{15}N HSQC spectra (*sfhmqcf3gpph*) were collected every 10 min for 60 hrs at 35˚C.

Spectra were collected using a total of 2048 and 256 data points with spectral sweep widths of 7309.942 and 3041.362 Hz along the proton and nitrogen dimensions, respectively. The pulse program was edited to use a reburp shape pulse[32] and to include a delay equal to 300 sec. This delay occurs before data collection to allow the cryoprobe coil temperature to equilibrate after each acquisition. The 300 sec delay was only applied for subsequent acquisitions in the first 60 hrs of data acquired. Each spectrum had a reburp pulse duration of 1 msec, receiver gain set to 3k, SP23 set to 24.05 dB, SP24 set to 13.11 dB, PC9 set to 3 msec, thereby making each SoFAST-HSQC's total acquisition time 5 min 8 sec.

HDX rates (k_{ex}) were calculated similarly to T_1 and T_2 relaxation rates (see above). A protection factor (PF) was calculated by dividing an estimated intrinsic rate (k_{int}) for each residue by the measured rate (k_{ex}) for that residue,[33-36] i.e. PF = k_{int}/k_{ex}. The PFs presented within the manuscript are the $\log_{10}(k_{int}/k_{ex})$ referred to as $\log_{10}$PF.[37, 38] Due to rapid exchange, we had to estimate k_{ex} values for several residues for wild-type DJ-1 (47 for the monomer) and M26I DJ-1 (85 for the monomer) as described below. For residues that could be observed in a T_1 HSQC but were not present in the first HDX HSQC, the rate constant for exchange was estimated to be ten times greater than the collection time for the HDX HSQC (e.g. k_{ex}=1/1.4 min^{-1} for an HSQC spectrum obtained after 14 minutes of H/D exchange); this affected 85 residues for M26I and 44 residues for wild-type DJ-1. For all residues present in the first HDX spectrum, but then exchanged before a second spectrum could be collected, the k_{ex} was estimated to be ten times faster than the second HDX HSQC collection time, in min^{-1}; this affected three residues for wild-type DJ-1 only. There were also a total of six residues for M26I and 12 residues for wild-type DJ-1 that did not exchange with D_2O over the 60 hr experiment.

7.2.5 DJ-1 aggregation as a function of temperature

Because Ni^{2+}-NTA purified DJ-1 (referred to as 'As Purified' in Fig. 5a) has been reported to be more susceptible to aggregation than a sample that has not been exposed to transition metals,[39] DJ-1 used for the aggregation experiments was supplemented with 10 mM EDTA, dialyzed overnight against storage buffer containing 1 mM EDTA, then dialysed again against storage buffer alone (referred to as 'EDTA treated and Removed' in Fig. 5b.) Aggregation of 100 μM DJ-1 was monitored by measuring the optical density of the sample at 400 nm every 30 min over 60 continuous hours with a BioTek Synergy 2 multi-mode microplate reader (BioTek Instruments, Inc.). Sample temperature was controlled with the installed Gen5 software and data were collected at 30, 35, and 37°C. WT, M26I, M26L, and M26V DJ-1 were diluted to a volume of 180 μL with degassed storage buffer and placed in a 96-well plate (Costar 3595). 50 µL of light mineral oil (Fisher Scientific) was layered over the sample to prevent evaporation, the 96-well plate was sealed with optically clear polystyrene seals (VWR International), and the samples were allowed to equilibrate for 10 min prior to the start of the assay.

7.2.6 Fluorometry of DJ-1 Tryptophan Mutants

Human DJ-1 lacks a Trp, making it necessary to introduce a Trp by site-directed mutagenesis. To identify minimally disruptive locations for introduction of a Trp residue into DJ-1, sequence alignments of various proteins in the DJ-1/PfpI superfamily that contain Trp residues were used. Proteins in this alignment were 40-46% identical to human DJ-1, maximizing the chances for identifying well-tolerated locations for Trp substitution. Leu101 and Tyr141 were identified as candidate sites for mutation that also corresponded to areas of DJ-1 with markedly different responses to the M26I mutation in HDX. All mutations were verified by DNA sequencing. Hexahistidine-tagged recombinant proteins were expressed using pET15b in BL21(DE3) *E. coli* and purified as previously described.[16, 23] Due to solubility problems, M26I/L101W DJ-1 cells were induced with 0.2 mM IPTG followed by incubation at 20°C overnight with shaking. Additionally, chloramphenicol (Fisher Scientific) was added to a final concentration of 100 μg/mL two hrs before cells were harvested to enhance the recovery of soluble protein.[40, 41]

Intrinsic tryptophan fluorescence emission spectra were collected at 37°C using a Cary Eclipse fluorescence spectrophotometer (Varian) from 300-400 nm with excitation at 280 nm, and excitation and emission slit widths set to 5 nm. A recirculating thermostated water bath (Varian) was used to control temperature. DJ-1 samples were diluted to 500 μM in storage buffer, the cuvette was capped, and the solution was allowed to equilibrate for 5 min. After correction for background scattering using a buffer control, the fluorescence spectra of the samples were measured. All fluorescence intensities were normalized to the concentration of their respective protein.

7.2.7 Chemical Cross-linking and Proteolysis of DJ-1

Disuccinimidyl suberate (DSS) (Thermo-Scientific) was aliquoted into 2 mg portions in a Coy anaerobic chamber (Coy Lab Products) with plastic spatulas. All plastic material was equilibrated in this environment for at least two weeks prior to use in order to minimize metal or oxygen contamination prior to the experiment, which can reduce DSS cross-linking efficiency. DSS aliquots were then stored in desiccant at 4°C until ready to use.

Wild-type and M26I DJ-1 were diluted to 100 μM and subsequently dialyzed against storage buffer lacking DTT. Each protein was divided into two 200 μL aliquots where one aliquot was incubated at room temperature (~22°C) and the other at 37°C. The samples were incubated at their respective temperatures for ~ 2 hrs prior to the initiation of cross-linking by addition of 100 mM DSS dissolved in DMSO to final concentrations of 0, 2, 3, 4, or 5 mM. DMSO alone was added as a negative control. After 30 min, the reaction was quenched with the addition of Tris-HCl, pH 7.5 to a final concentration of 38 mM and incubated for an additional 15 min at their respective temperatures. All samples were then mixed with SDS-loading dye, heated to 95°C for 5 min and analyzed using SDS-PAGE with a 12% gel stained with colloidal Coomassie blue dye (Life Technologies).

Wild-type, M26I and M26L DJ-1 were diluted to 100 μM and subsequently dialyzed against PBS pH 7.4 (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, 5 mM EDTA). Each protein was incubated either at room temperature (~22°C) or at 37°C. The samples incubated at their respective temperatures for 0-3 hr, and at selected time points, a small aliquot was removed and either incubated with Bismaleimidoethane (BMOE) (Thermo-Scientific) (final concentrations: 0, 0.1, or 0.5 mM) or DMSO (negative control) for 45 min. After 45 min, the reaction was quenched with the addition of DTT to a final concentration of 20 mM and incubated for an additional 15 min. Samples were mixed with SDS-loading dye, heated to 100°C for 5 min and analyzed using SDS-PAGE with a 12% gel stained with colloidal Coomassie blue dye (Life Technologies).

7.2.8 Limited protease digestion

Wild-type, M26I and M26L DJ-1 were diluted to 400 μM in storage buffer. Trypsin, dissolved in storage buffer, was added to the protein sample at a final concentration of 4.3 μM. The samples were incubated in a water bath at 37°C for 22.5 hrs. Aliquots were removed at various time points, mixed with SDS-loading dye, and heated to 100°C for 5 min to stop proteolysis. Digested samples were analyzed using SDS-PAGE with a 15% gel and stained with colloidal Coomassie blue (Life Technologies).

7.2.9 Thermal stability determination and far-UV circular dichroism (CD) spectroscopy

The melting temperatures of DJ-1 proteins were determined using differential scanning fluorimetry as previously described.[24, 42] The secondary structural content of the proteins was determined using far-UV CD spectroscopy following a previously reported protocol[24] with minor modifications. CD spectra were measured for 13 μM proteins in 10 mM potassium phosphate pH 7.2, 2.5 mM DTT using a Jasco J-815 CD Spectrometer (Jasco, Inc.). Sample temperature was controlled with a Julabo AWC100 (Julabo USA, Inc.) recirculating water cooler and far-UV CD spectra were collected at 25 and 37°C. Protein concentration in the cuvette was determined using Scopes' method[43] and used to calculate the mean residue molar ellipticity according to the formula below:[24]

$$[\Theta(\lambda)] = \frac{\theta_{obs}(\lambda)}{10ncl} \qquad [\mathrm{EQ.}\,3]$$

where $\Theta(\lambda)$ is the mean residue molar ellipticity as a function of wavelength (deg cm^2 $dmol^{-1}$ $residue^{-1}$), $\theta_{obs}(\lambda)$ is the measured ellipticity as a function of wavelength (nm), *n* is the number of residues in the protein, *c* is the concentration of the protein (M), and *l* is the path length of the cuvette (cm).

7.2.10 Crystal growth, data collection and processing, and structure determination for M26V DJ-1

M26V DJ-1 was crystallized using the sitting drop vapor diffusion method by mixing 2.4 µL of protein at 1.3 mM and 1.8 µL of reservoir solution (100 mM Tris-HCl pH 9, 200 mM sodium acetate trihydrate, 25 % PEG 4000 and 3 mM DTT), followed by equilibration against 1 mL of the reservoir solution. Bipyramidal crystals of M26V DJ-1 in space group $P3_121$ appeared in 2-4 days at room temperature. Crystals were cryoprotected by serial transfer through the reservoir solution supplemented with gradually increasing amounts of ethylene glycol until a final concentration of 25% (v/v) was reached. The cryoprotected crystals were removed from the cryoprotectant in nylon loops and mounted in a nitrogen cryostream for data collection.

X-ray diffraction data were collected to 1.45 Å resolution from a single crystal at 110 K at the University of Nebraska-Lincoln Macromolecular Structural Core Facility using a MicroMax-007 copper rotating anode source (Rigaku) operating at 40 kV and 20 mA with Osmic Blue confocal optics, and a Raxis IV^{++} detector (Rigaku). *In situ* annealing[44] of all crystals was done by blocking the cold nitrogen stream for ~3 sec, which reduces mosaicity and improves data scaling statistics. All data were indexed, scaled, and merged using HKL2000[45] with final data statistics in Table 1.

Phases for M26V DJ-1 were determined by molecular replacement (MR) using the structure of human DJ-1 (PDB: 1P5F, 99% sequence identity[46]) as a search model in Phaser[47] in the CCP4 suite.[48] Manual adjustments to the initial top MR solution were made by inspection of $2mF_O$-DF_C and mF_O-DF_C electron density maps in COOT.[49] The resulting model was refined in Refmac5[50] against a maximum likelihood amplitude-based target function with geometric and anisotropic atomic displacement parameter (ADP) restraints and including riding hydrogens. A test set of 5% randomly chosen reflections were used for the calculation of the R_{free}[51] value. The stereochemical, side chain rotameric, and packing quality of the final model was validated using COOT[49] and the MolProbity server[52]. All residues except Cys106 were in favored regions of the Ramachandran plot, which is commonly observed in DJ-1 crystal structures. Final data and model statistics are provided in Table 1.

Table 1: Data Collection and Refinement Statistics	
Data Collection	
Sample	M26V DJ-1
X-ray source	Rotating Cu Anode
X-ray wavelength (Å)	1.54
Space group	$P3_121$
Cell dimensions *a, b, c* (Å)	75.04, 75.04, 75.27
α, β, γ (degrees)	90, 90, 120
Resolution (Å)^{a}	64.99-1.45 (1.49-1.45)
R_{merge}^{b}	0.062 (0.523)
<I>/<σ(*I*)>	56.6 (5.0)
Completeness (%)	99.7 (97.0)
Redundancy	20.4 (18.2)
Refinement	
Program	Refmac5
Resolution (Å)	64.99-1.45
No. of reflections	41569
R_{work} (%)^{c}	0.110 (0.344)
R_{free} (%)^{d}	0.135 (0.375)
R_{all} (%)^{e}	0.111
No. of protein residues	189
No. of water atoms	309
No. of heteroatoms (ligand EDO)	4
B_{eq} factors (Å^2)	
Protein	18.07
Water	38.07

Heteroatoms	30.50
r.m.s. deviations	
Bond lengths (Å)	0.013
Bond angles (degrees)	1.63

[a] Values in parentheses are for highest resolution shell

[b] R_{merge} is calculated according to Equation 4,

$$\mathrm{R_{merge}} = \sum_{hkl}\sum_{i}\left|I_{hkl}^{i} - \langle I_{hkl}\rangle\right| / \sum_{hkl}\sum_{i} I_{hkl}^{i} \quad [\mathrm{EQ.}\,4]$$

where *i* is the *i*th observation of a reflection with index *h, k, l*, and angle brackets indicate the average over all *i* observations.

[c] R_{work} is calculated according to Equation 5,

$$\mathrm{R_{work}} = \sum_{hkl}\left|F_{hkl}^{O} - F_{hkl}^{C}\right| / \sum_{hkl} F_{hkl}^{O} \quad [\mathrm{EQ.}\,5]$$

where F_{hkl}^{C} is the calculated structure factor amplitude with index *h, k, l*, and F_{hkl}^{O} is the observed structure factor amplitude with index *h, k, l*.

[d] R_{free} is calculated as R_{work}, where the F_{hkl}^{O} values are taken from a test set comprising 5% of the data that were excluded from the refinement.

[e] R_{all} is calculates as R_{work}, where the F_{hkl}^{O} include all measured data (including the R_{free} test set).

7.3 Results

7.3.1 M26I and wild-type DJ-1 have similar picosecond-nanosecond dynamics at 35˚C

Consistent with previous NMR and X-ray crystallographic results, the solution structures of wild-type and M26I DJ-1 are comparable based on their highly similar two dimensional (2D) ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) NMR spectra collected at 35°C (Figure 1).[16, 17] This temperature was chosen because we sought to characterize DJ-1 near physiological temperature, however experiments at 37°C resulted in aggregation of M26I DJ-1 (see below). The similar structures but divergent cellular stabilities of wild-type and M26I DJ-1 suggest the hypothesis that the M26I mutation may alter the conformational dynamics of DJ-1.

We characterized the picosecond-nanosecond (ps-ns) time-scale dynamics of DJ-1 using NMR by measuring the R_1 and R_2 relaxation rates as well as Nuclear Overhauser Effect (NOE) ratios at 500 MHz and 35°C (Figure 2). The R_2/R_1, and NOE ratios were similar for wild-type and M26I DJ-1 at 35°C (Table 2, Figure 3). NOE ratios are bound by 0 and 1, with lower values indicating areas of increased motion in the backbone of a protein.[17] The averaged NOE ratios at 35°C for both wild-type (0.78 ± 0.10) and M26I DJ-1 (0.77 ± 0.087) are lower than previously reported values at 27°C (0.9 ± 0.09),[17] consistent with greater protein motion at higher temperature.

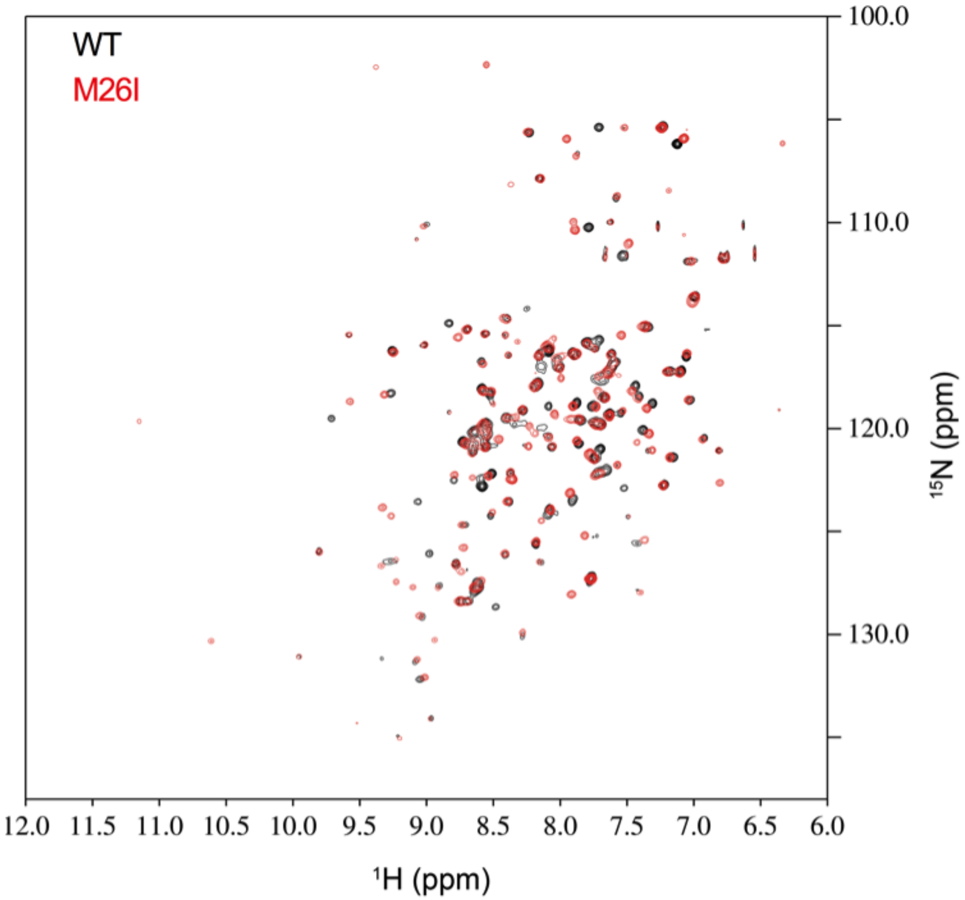


Figure 1: Wild-type and M26I DJ-1 have similar 2D ^{1}H-^{15}N HSQCs at 35°C. The 2D ^{1}H-^{15}N HSQC spectrum for wild-type (WT, black) is overlaid with the spectrum for M26I DJ-1 (red). M26I DJ-1 resonances mostly overlap with those observed for WT DJ-1, as was previously shown by Malgieri and Eliezer.[17]

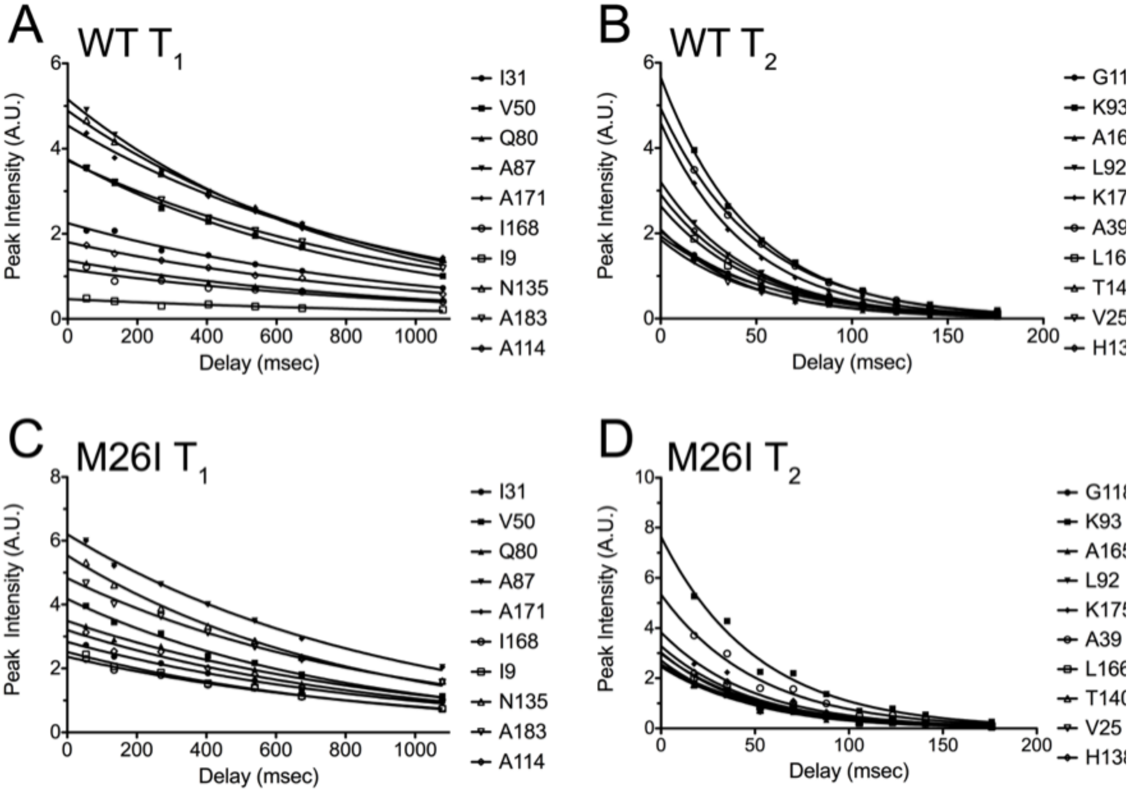


Figure 2: Example fitting for T_1, T_2 relaxation curves of select residues in wild-type and M26I DJ-1 at 35°C. (*A*, *C*) Wild-type (WT) and M26I T_1 relaxation fits for ten randomly chosen amino acids. (*B*, *D*) WT and M26I T_2 relaxation fits for the same amino acids as in (A). (*A*-*D*)

Table 2.

	WT DJ-1	**M26I DJ-1**
Model Free Input		
NOE	0.779 ± 0.0997	0.770 ± 0.0871
R_1	1.158 ± 0.1398	1.219 ± 0.288
R_2	19.303 ± 2.309	18.821 ± 1.727

R_2/R_1	16.877 ± 2.643	15.911 ± 2.271
Model Free Output		
S^2	0.912 ± 0.0788	0.934 ± 0.0544
S^2_f	0.876 ± 0.0492	---
S^2_s	0.916 ± 0.0724	0.934 ± 0.0544
t_e	1313.9 ± 714.923	622.75 ± 656.975
R_{ex}	3.188 ± 1.688	1.356 ± 1.917
SSE	1.977 ± 2.742	1.545 ± 2.238

Values in the table are averages ± SD

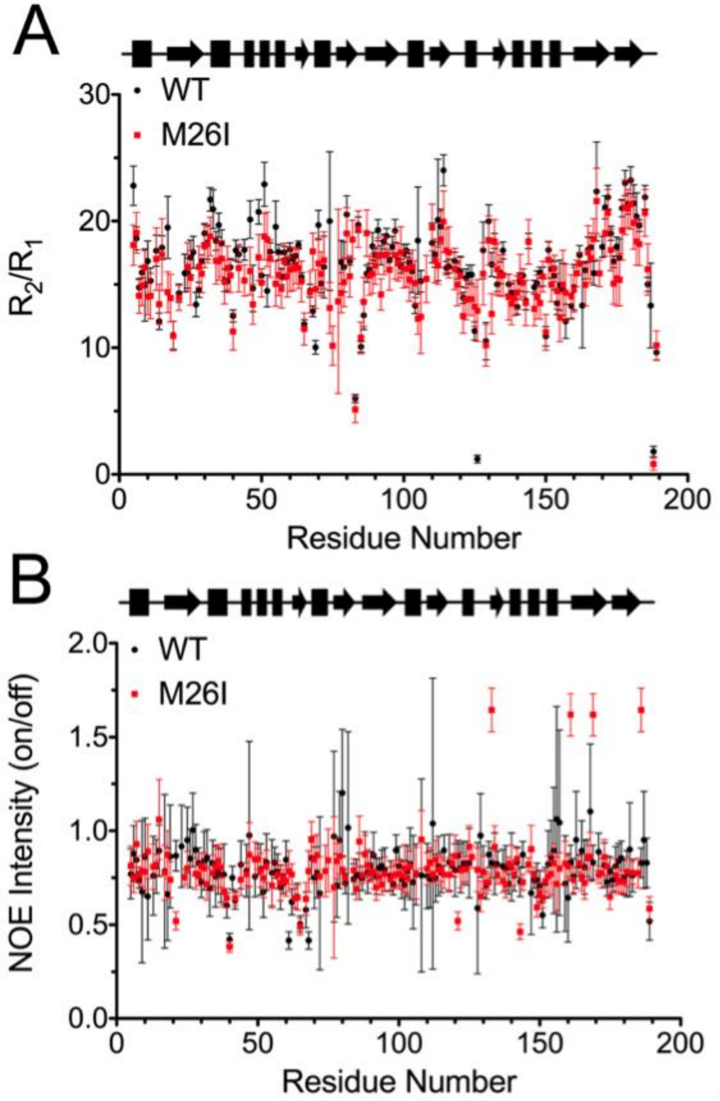


Figure 3: The R_2/R_1 and NOE ratios are similar for WT and M26I DJ-1 at 35°C. (*A*) The ratio of relaxation rates (R_2/R_1) is plotted for every observed backbone amide resonance in both wild-type (WT, black) and M26I (red) DJ-1. The R_2/R_1 ratio is similar for both proteins at 35°C. (*B*) WT (black) and M26I (red) NOE ratios. Four NOE ratio values for M26I DJ-1 are not bound by 0 and 1, and they are Met133, Ser161, Val169, and Val186. This is likely due to inadequate equilibration of NOE saturation for these few residues.

Lipari-Szabo model free analysis was performed using the measured R_1, R_2, and NOE ratios, generating generalized order parameters (S^2) that describe the extent of motion of each assigned amide N-H pair using a "vector in a cone" description. S^2 is bound by 0 and 1, where lower values indicate areas of increased mobility of the backbone peptide groups.[30, 53] The S^2 values for wild-type and M26I DJ-1 at 35°C are similar (Figure 4), with average values of 0.91 ± 0.079 and 0.93 ± 0.054, respectively (Table 2). We note that these values are high, indicating that DJ-1 is fairly rigid on the ps-ns timescale. These calculated S^2 values are robust to changes in the selection of residues used for initial diffusion tensor calculation and different random number seeds, indicating good convergence in the calculations and low sensitivity to initial parameters.

7.3.2 The M26I mutation increases longer time-scale dynamics at 35°C

Slower timescale (minutes to hours) dynamics of DJ-1 were investigated using NMR-detected hydrogen/deuterium exchange (HDX). Because enhanced protein mobility will increase the solvent accessibility of backbone amide groups, NMR-detected HDX is a sensitive and site-specific probe of dynamics that occur on longer timescales than those characterized using NMR relaxation-based methods.

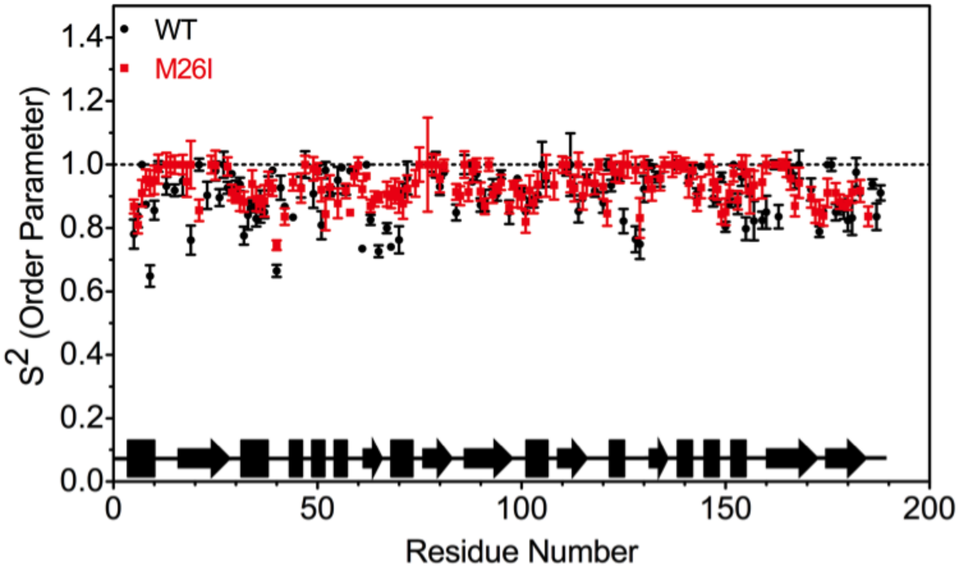


Figure 4: Wild-type and M26I DJ-1 behave similarly on a fast time-scale at 35°C. The general order parameter (S^2) for wild-type (WT, black) and M26I DJ-1 (red) at 35°C obtained from FAST ModelFree are plotted per residue. Picosecond-nanosecond motions for both proteins are nearly identical at 35°C. The regions without a reported value indicate amino acids that could not be analyzed in FAST ModelFree, primarily due to unobserved or overlapping resonances in the NMR spectra. A schematic of the secondary structure of DJ-1 is shown below for reference (rectangles=β-sheet, arrows=α-helix.)

HDX reveals that the hydrophobic core of M26I DJ-1 is substantially less protected (and hence more dynamic) than that of wild-type DJ-1, while both proteins have comparable HDX rates for solvent exposed residues. HDX rates measured at 35˚C were converted to protection factors ($\log_{10}$PF)[33, 34] (Figure 5a, Figure 6) and were then mapped onto the crystal structure of dimeric DJ-1 (Figure 5b). Despite being structurally similar to the wild-type protein, M26I DJ-1 experiences greater conformational fluctuations than wild-type DJ-1, particular for residues 15-35, 45-55, 110-120, 130-135, 140-190. The elevated exchange in the core of M26I DJ-1 is particularly noteworthy, as it indicates that the mutant protein transiently samples conformations that deviate markedly from the crystal structure and exposes the hydrophobic core to solvent. This is highlighted by inspection of difference $\log_{10}$PF factors, which clearly show that the hydrophobic core of DJ-1 is less protected (and hence more solvent exposed) in M26I DJ-1 (Fig. 5c). Because the hydrophobic core of DJ-1 spans both monomers in the dimer, the elevated exchange in the core of M26I DJ-1 could involve transient dimer opening, although M26I is highly dimeric in other studies,[16, 17, 19, 20] suggesting that dissociation into monomers is less likely.

Tryptophan (Trp) fluorescence spectroscopy was used as an additional probe of the influence of the M26I mutation on the flexibility of M26I DJ-1. The quantum yield of Trp is sensitive to its local environment, and thus Trp can serve as a reporter of site-specific environmental changes resulting from mutation of residue 26. Human DJ-1 contains no Trp residues, therefore we introduced Trp at two locations of the protein (residues Tyr141 and Leu101) that exhibited strongly differing HDX behavior in response to the M26I mutation (Fig. 5c). Both the Y141W and L101W mutations have small effects on the thermal stability and secondary structure of DJ-1, indicating that these mutations do not significantly perturb the DJ-1 structure (Figure 7). Tyr141 is located in a region of DJ-1 with low $\log_{10}$PF and whose HDX exchange rate is insensitive to the M26I mutation ($\log_{10}\text{PF}_{\text{WT}}$=2.09 vs. $\log_{10}\text{PF}_{\text{M26I}}$= 2.08). Consistent with the HDX results, the fluorescence quantum yield of Y141W is similar in wild-type and M26I DJ-1

(Figure 8). In contrast, the L101W substitution is located at a position that becomes mobile as a result of the M26I mutation, exhibiting a HDX $\log_{10}$PF decrease from 4.35 in wild-type to 1.75 in M26I DJ-1. Consistent with the HDX results, the fluorescence emission of M26I/L101W DJ-1 is markedly lower than that of L101W alone, indicating that Trp101 fluorescence in the M26I mutant is partially quenched due to a dynamically driven increase in solvent exposure (Figure 8).

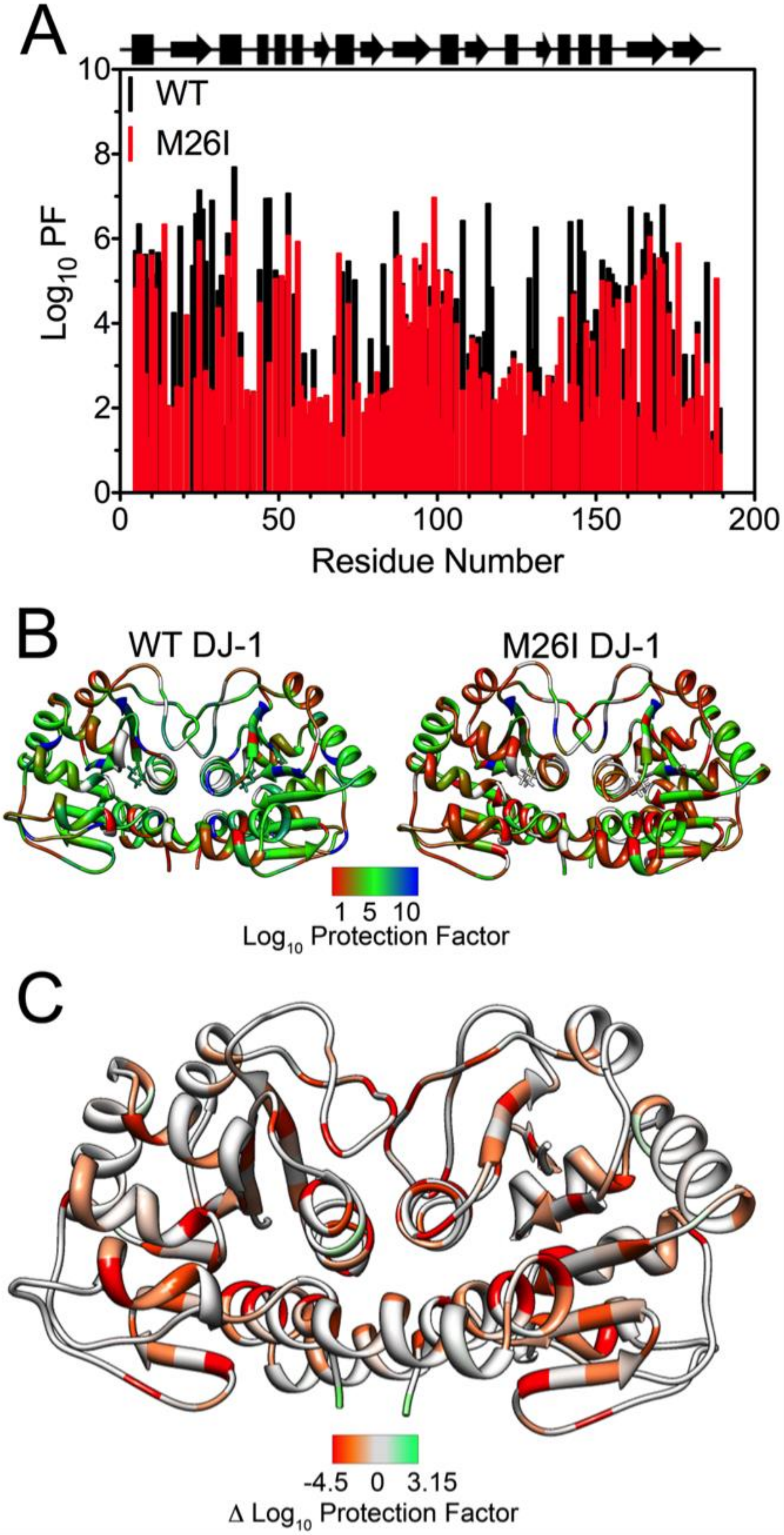

A
10
8
6
4
2
0
Log10 PF
WT
M26I
0
50
100
150
200
Residue Number
B
WT DJ-1
M26I DJ-1
1 5 10
Log10 Protection Factor
C
-4.5 0 3.15
Δ Log10 Protection Factor

Figure 5: HDX-detected dynamics are dramatically changed for M26I DJ-1 at 35°C.

(*A*) Calculated $\log_{10}$ HDX protection factors ($\mathrm{Log}_{10}\mathrm{PF}$) for wild-type (WT, black) and M26I (red) DJ-1. Lower $\mathrm{Log}_{10}\mathrm{PF}$ values correspond to backbone amide hydrogen atoms that exchange more readily with solvent deuterons. (*B*) $\mathrm{Log}_{10}\mathrm{PF}$ values from (*A*) have been mapped onto the structure of wild-type (WT, left) and M26I (right) DJ-1 dimer. All residues for which a $\mathrm{Log}_{10}\mathrm{PF}$ could not be assigned are shown in grey. M26I DJ-1 is more dynamic than WT DJ-1, particularly in the hydrophobic core of the protein. (*C*) Difference $\mathrm{Log}_{10}\mathrm{PF}$ (M26I $\mathrm{Log}_{10}\mathrm{PF}$ – WT $\mathrm{Log}_{10}\mathrm{PF}$) mapped on to the dimer of M26I DJ-1. Red indicates areas that are more dynamic for M26I DJ-1, while green indicates areas that are more dynamic for WT DJ-1. Areas that are similar between the two proteins are grey. The enhanced exposure of the hydrophobic core of M26I DJ-1 to solvent is evident.

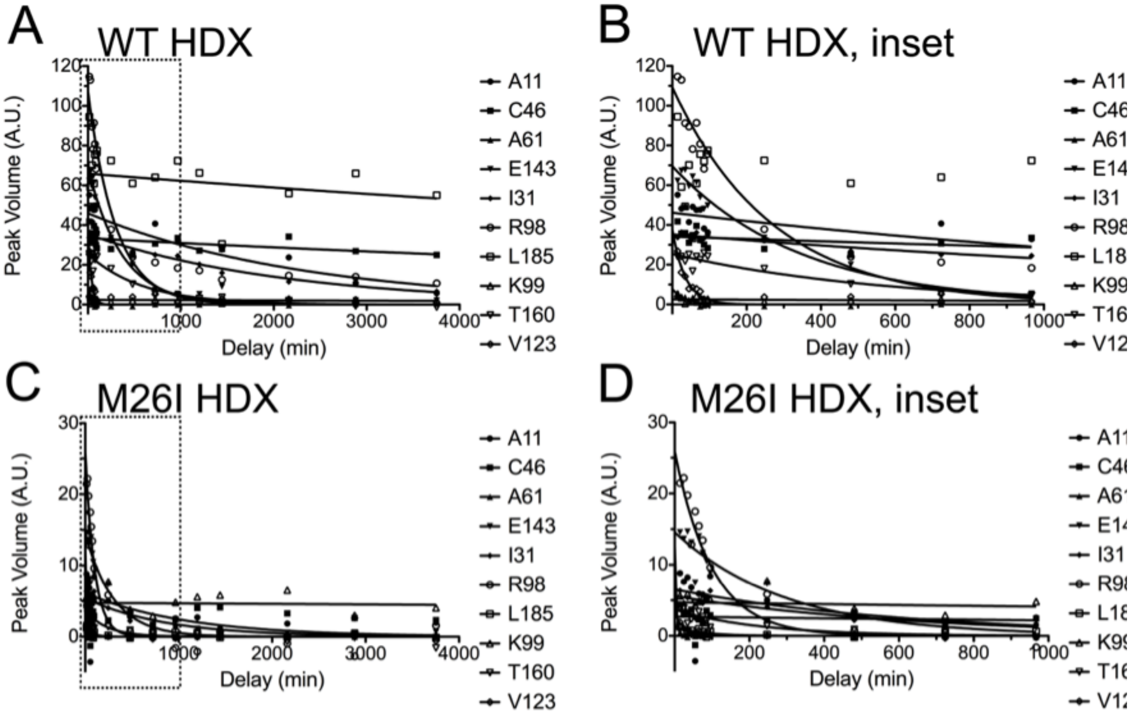


Figure 6: Example fitting for HDX curves of select residues in WT and M26I DJ-1. (*A*, *C*) Fitted decay curves for ten amino acids in wild-type (WT) and M26I DJ-1 HDX, spanning a range of different $\log_{10}$PF values. (*B, D*) Expanded view of the boxed region in (*A*) and (*C*), respectively. For WT HDX, Leu185 (open squares) the data could not be adequately fitted due to chemical exchange (see L185 in (*A*)).

7.3.3 Thiol cross-linking and proteolysis indicate that M26I DJ-1 is more flexible than wild-type DJ-1 at 37°C

We have recently described the use of thiol cross-linking to probe DJ-1 flexibility.[24] DJ-1 contains three free thiols per monomer, only one of which (Cys53) is highly solvent accessible (Figure 9a). Cys53 is ~3.5 Å from its symmetry mate in the DJ-1 dimer. Consequently the Cys53 pair is readily cross-linked, while the other cysteines (Cys106 and Cys46) are not. However, conformational fluctuations in DJ-1 can transiently expose the other cysteines and make them available for reaction, increasing the number of cross-linked DJ-1 species that can be observed in SDS-PAGE.[24] Therefore, thiol cross-linking is a sensitive probe of large-scale conformational plasticity in DJ-1.

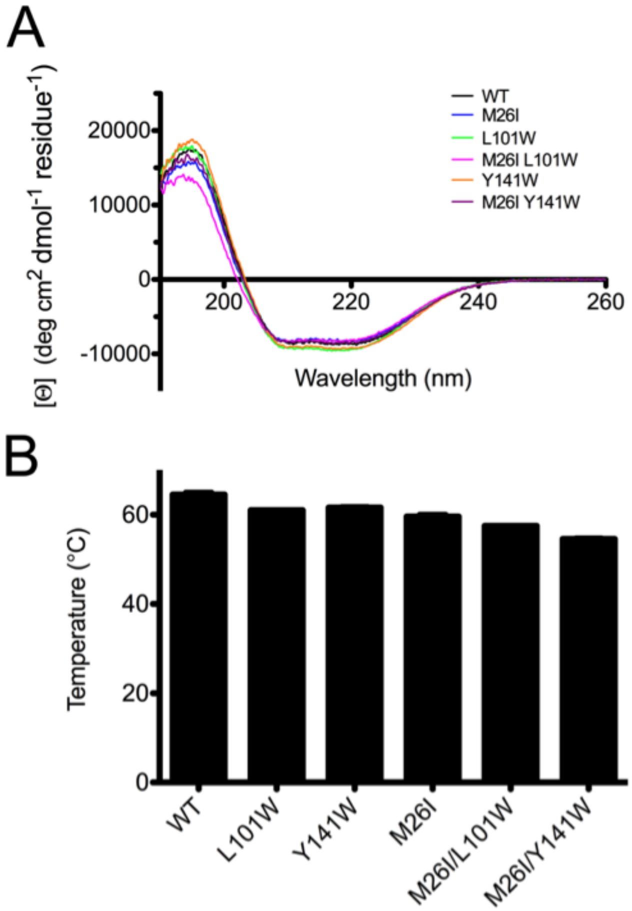


Figure 7: Tryptophan mutations have modest effects on secondary structure and thermal stability of DJ-1 at 37°C. (*A*) Circular dichroism (CD) spectra were collected at 37°C for wild-type (WT) (black), M26I (blue), L101W (green), M26I/L101W (magenta), Y141W (orange), and M26I/Y141W DJ-1 (purple). These proteins show no significant differences in the mean residue molar ellipicity ([Θ]) as a function of wavelength, indicating that the Trp mutations are not highly disruptive to DJ-1 structure. (*B*) Thermal

stability for all proteins in (*A*) was determined by the thermofluor assay. Consistent with the CD spectroscopy data (A), only minor changes are observed. The melting temperature (°C) is plotted for all proteins (n=5, plotted as avg ± SD.)

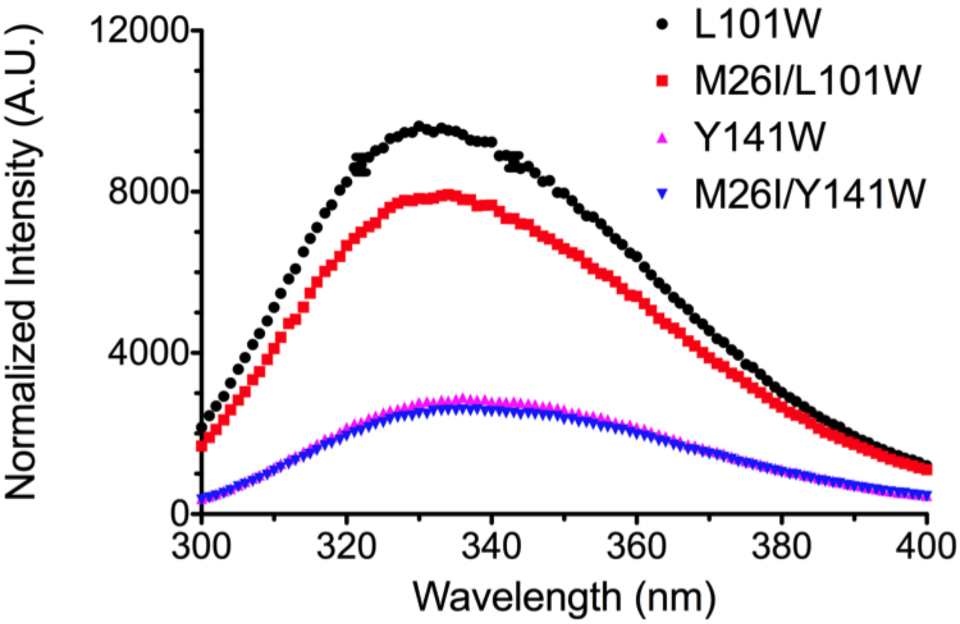


Figure 8: Tryptophan fluorescence emission spectra reveal site-specific differences in M26I DJ-1 solvent exposure. Tryptophan emission spectra for L101W (●), M26I/L101W (■), Y141W (▲), and M26I/Y141W (▼) obtained at 37°C. The M26I mutation results in quenching of L101W DJ-1 fluorescence, indicating Trp101 is more solvent exposed in M26I DJ-1. Leu101 is in a dynamic area of DJ-1 that is sensitive to the M26I substitution in HDX (Log$_{10}$PF WT$_{L101}$=4.35, Log$_{10}$PF M26I$_{L101}$=1.75.) In contrast, no differences are observed in the spectra for Y141W and M26I/Y141W DJ-1. Tyr141 is in a dynamic region of DJ-1 that is not sensitive to the M26I mutation (Log$_{10}$PF WT$_{Y141}$=2.09, Log$_{10}$PF M26I$_{Y141}$=2.08.)

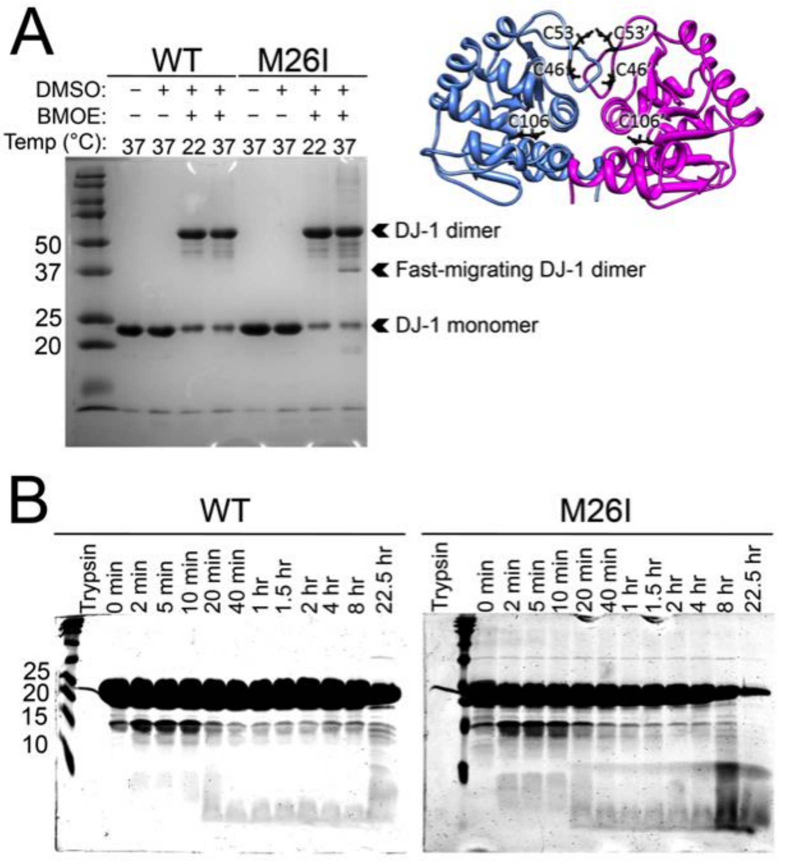


Figure 9: Thiol cross-linking and limited proteolysis show that M26I DJ-1 is more dynamic than wild-type DJ-1 at physiological temperature. (*A*) *Left:* Cysteine cross-linking with BMOE for wild-type (WT), and M26I DJ-1 at 22 and 37°C resolved by SDS-PAGE. M26I DJ-1 forms a faster migrating, more extensively cross-linked dimeric species at 37°C that is absent at 22 °C and for WT DJ-1. Monomer and dimer DJ-1 bands are indicated. *Right:* Location of all six cysteine residues on dimeric DJ-1. (*B*) Limited proteolysis time course with trypsin for wild-type (WT) and M26I DJ-1 at 37°C. M26I DJ-1 is more vulnerable to proteolysis, which is apparent by comparing the 8 and 22.5 hr lanes.

DJ-1 was cross-linked using the homobifunctional thiol cross-linker BMOE both at 21°C and at 37°C. In SDS-PAGE, a dominant dimer band at ~50 kDa is observed for both wild-type and M26I DJ-1 at both temperatures (Figure 9a). However, M26I DJ-1 also exists as a faster-migrating dimeric cross-linked species (~37 kDa) at 37°C that is not observed at 21°C and is never observed in wild-type DJ-1 (Figure 9a). This faster migrating species is likely a multiply cross-linked dimer. Therefore, increased flexibility of M26I DJ-1 allows it to transiently sample conformations at 37°C that expose the more buried Cys46 and Cys106 residues, making them available for additional cross-linking. In contrast, cross-linking with the amine-reactive DSS (Figure 10) shows little difference between wild-type and M26I DJ-1 at room temperature and at 37°C. This is consistent with the larger number of DSS-cross-linked species that can be formed between multiple solvent-exposed locations.

Limited proteolysis with trypsin was used to further characterize DJ-1 flexibility on longer timescales. An increase in DJ-1 flexibility would be expected to result in more extensive digestion by the protease. M26I DJ-1 has an accelerated rate of cleavage by trypsin at 37°C, with a digestion pattern at 8 hours (Figure 9b) that is comparable to that of wild-type protein at 22.5 hrs. In contrast, no loss of either protein was observed at 22°C over the same time-course. This proteolysis result is consistent with the HDX, Trp fluorescence, and cross-linking experiments, all of which demonstrate that longer timescale dynamics of M26I are enhanced at 37°C compared to wild-type DJ-1.

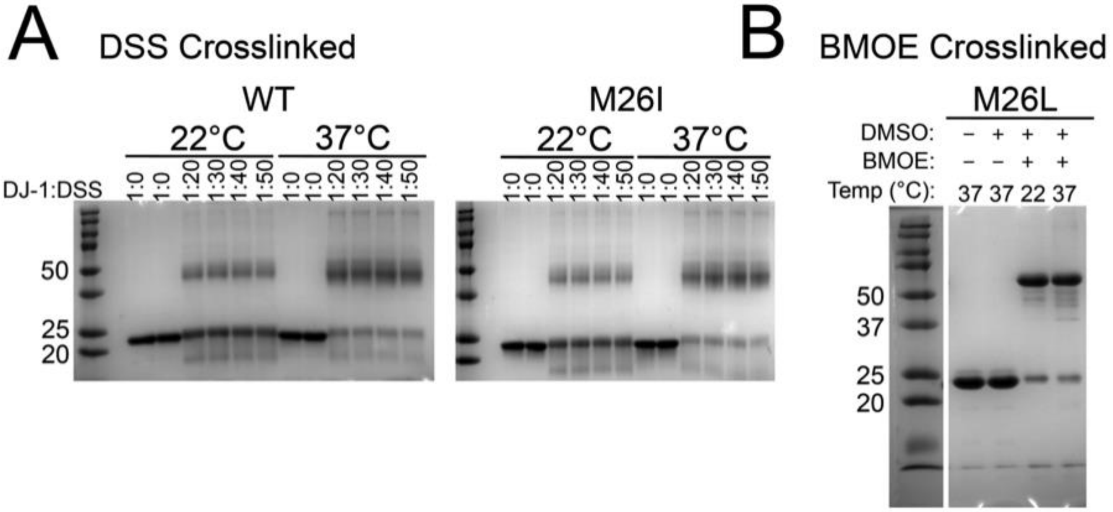


Figure 10: Cross-linking behavior of DJ-1 proteins. (*A*) Primary amine cross-linking with DSS for wild-type (WT, left) and M26I (right) DJ-1 resolved by SDS-PAGE. Cross-linking was performed at 1:20, 1:30, 1:40, and 1:50 (DJ-1:DSS) at both 22 and 37°C. Samples in lanes 1 and 2 for both gels are the input and protein + DMSO controls. There is no difference in DSS cross-linking between WT and M26I DJ-1 at either temperature. (*B*) BMOE cross-linking for M26L DJ-1 is similar to that of wild-type DJ-1 (Figure 9a.), consistent with this mutation being non-disruptive.

7.3.4 M26I DJ-1 is unstable at physiological temperature

Prior studies reported that M26I DJ-1 is more prone to aggregation *in vitro*, although a structural explanation for this aberrant behavior has been elusive. We monitored the aggregation of recombinant M26I *in vitro* by measuring its scattering of 400 nm light as a function of temperature (Figure 11a). Both wild-type and M26I DJ-1 remained soluble from 15 to 30°C. Aggregation of M26I DJ-1 was apparent at 35°C and exacerbated at

37°C, with a lag phase of approximately 20 hrs (Figure 11a). In contrast, wild-type DJ-1 was not observed to aggregate at these temperatures (Figure 11a). The onset of aggregation for M26I at physiological temperature provides an appealing explanation for the reduced stability of the protein in cells (grown at 37°C)[3, 22] and the apparent stability of the recombinant protein, which is typically handled at lower temperatures.[16, 17]

The exposure of DJ-1 to transition metals such as Ni^{2+} and Fe^{3+} has been reported to enhance its instability.[39] As all proteins used in this study were purified by Ni^{2+}-NTA chromatography, we tested the influence of trace metal contamination by treating the samples with a large excess of EDTA followed by dialysis. Consistent with the prior report, the chelation and removal of trace metal diminishes M26I DJ-1 aggregation, although M26I is still unstable and aggregates at 35-37°C (Figure 11b).

The crystal structure of M26I DJ-1 shows that the mutation creates both a small cavity within the hydrophobic core and a steric conflict with a neighboring Ile31.[16] The steric clash with Ile31 is the most obvious structural change resulting from the M26I mutation, but it is minor (0.7 Å displacement). To determine whether the clash at Ile31 in M26I DJ-1 is responsible for the increase in flexibility that drives aggregation of the protein, we created both the M26L and M26V mutations (Figure 12, Table 1). The M26L substitution eliminates the β-branched sidechain at position 26 and alleviates the clash with Ile31.[16] Many close homologues of DJ-1 natively have a Leu at position 26, indicating that this substitution is well-tolerated and preserves DJ-1 function. M26L DJ-1 does not aggregate at 37°C (Figure 11) and is less dynamic as determined by thiol crosslinking (Figure 10). This demonstrates the importance of the steric conflict between I26 and Ile31 that leads to M26I DJ-1 instability. The destabilizing effect of β-branched sidechains at residue 26 was confirmed by the extensive aggregation observed for the M26V mutant at 35-37°C (Figure 11), which, like M26I, places a hydrophobic β-branched amino acid at this position. The crystal structure of M26V DJ-1 shows that the mutation is highly conservative, with a Cα-RMSD of 0.14 Å with M26I DJ-1 (Figure 12). The most significant structural change is the displacement of Ile31 resulting from a clash with the Cγ1 atom of Val26, similar to M26I DJ-1 (Figure 12).

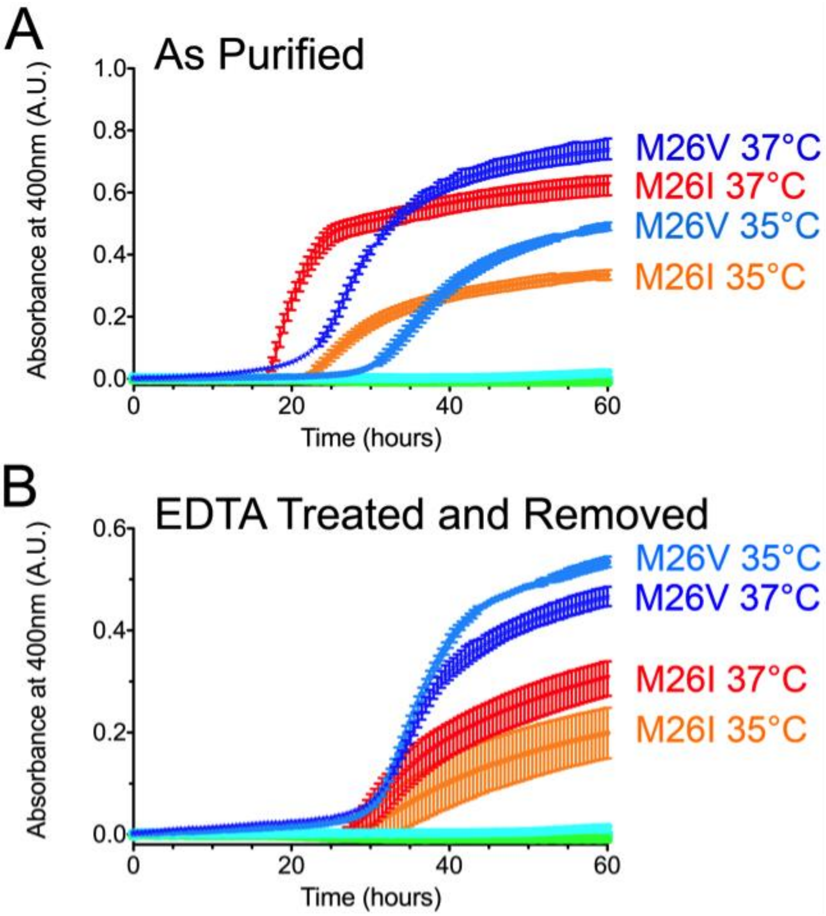


Figure 11: M26I DJ-1 is aggregation-prone at physiological temperature *in vitro*. (*A*) Aggregation of wild-type (WT), M26I, M26L, and M26V at 30, 35, and 37°C. There is no observable aggregation for wild-type and M26L DJ-1 at any of these temperatures (baseline). In contrast, M26I and M26V DJ-1 aggregate extensively at 35 and 37°C (labeled). (*B*) Aggregation was performed as in (*A*), but the samples were treated with 10 mM EDTA and dialyzed to remove any trace metal contamination from the proteins before the start of the experiment. M26I and M26V DJ-1 still aggregate at 35 and 37°C (labeled), although to a lesser extent than in (*A*.) (*A*-*B*) All samples are in triplicate (plotted as avg ± SD.)

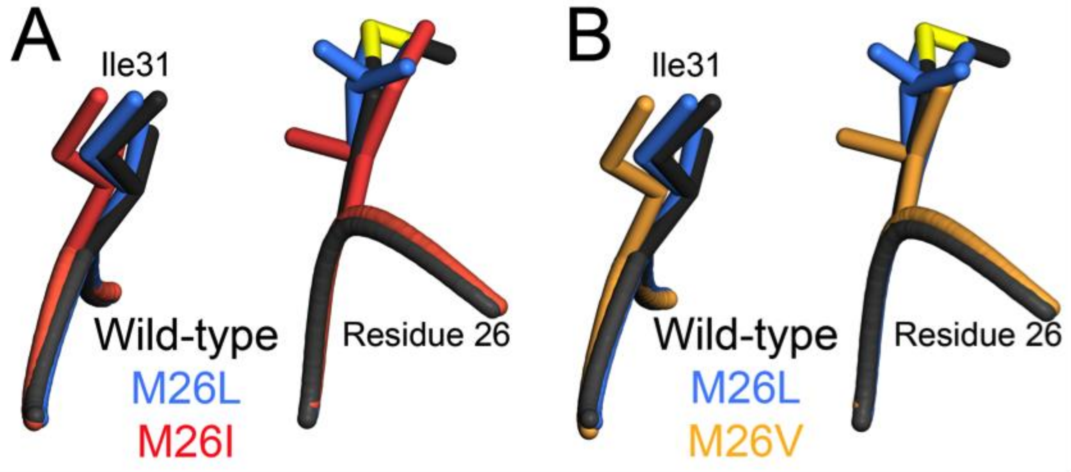


Figure 12: The M26I and M26V mutations cause similar steric conflicts with Ile31. (*A*, *B*) The local environment surrounding residue 26 are shown for wild-type (dark grey), M26L (blue), and either M26I (red) (*A*) or M26V (gold) (*B*). Ile31 and residue 26 are labeled.

7.4 Discussion

This study demonstrates that the structurally conservative parkinsonian M26I mutation in DJ-1 causes an increase in slow protein conformational dynamics near physiological temperature that destabilizes the protein. M26I DJ-1 provides an attractive system in which to explore the connection between protein dynamics and disease, as X-ray crystallography and NMR spectroscopy both indicate a high structural similarity between M26I DJ-1 and the wild-type protein despite being a disease-causing mutation.[16, 17] Therefore, the observed dynamical changes in M26I DJ-1 are not confounded by large structural changes in the protein. Consistent with the structurally conservative character of the mutation, M26I does not alter fast timescale (ps-ns) backbone dynamics. However, this parkinsonian mutation does enhance slower motions detected by HDX as well as by tryptophan fluorescence, chemical cross-linking, and limited proteolysis. These slower motions destabilize M26I DJ-1 and, ultimately, cause parkinsonism.

Considered together, our data suggest a model involving a slow transient opening of the M26I DJ-1 dimer that increases solvent exposure of its hydrophobic core, especially near physiological temperature. Difference $\log_{10}$PF values for HDX show a more exchange-active core in M26I DJ-1, and greater H/D exchange at the edges of the central β-sheet (Figure 5c) are consistent with “fraying” of this central sheet at its edges. Some prior studies have reported modestly reduced secondary structural content for M26I DJ-1,[17, 19] which would be consistent with a transient loss of interstrand hydrogen bonds in these regions. Supporting this interpretation, the Trp fluorescence and thiol cross-linking data both suggest that largely buried regions of the structure become more solvent-exposed in

M26I at 37°C. In particular, the appearance of a distinct BMOE-cross-linked dimer species for M26I at 37°C that is not observed at lower temperature or in the wild-type protein indicates that increased solvent exposure of buried portions of M26I DJ-1 is a temperature-dependent phenomenon.

The exposure of hydrophobic areas of DJ-1 near physiological temperature is the likely cause of the enhanced aggregation that we and others have observed in M26I DJ-1. We find that mutant DJ-1 aggregation is only apparent (over the 60 hour timescale) at 35-37°C, indicating that M26I acts as a temperature-sensitive loss of stability mutant. In addition, M26I DJ-1 aggregation is distinct from thermal denaturation, as the melting temperature for M26I DJ-1 has been measured to be 52-63°C using various techniques.[16, 17, 19, 54] It is important to emphasize that DJ-1 aggregation *in vitro* would likely manifest as protein degradation *in vivo*, and thus would account for the reduced level of M26I DJ-1 observed in cells grown at 37°C. The strong coupling between elevated temperature and M26I DJ-1 instability suggests that the greater flexibility of M26I mutant "softens" DJ-1 and leads to loss of stability at physiological temperature where the wild-type protein is still stable and functional. Attractively, the onset of dynamically-driven instability in M26I DJ-1 at physiological temperature resolved conflicting reports of this mutant's stability in cells (at 37°C) and *in vitro* (at room temperature).

The dynamical changes in DJ-1 resulting from the M26I mutation are initiated by the steric conflict between Ile26 and Ile31[16] that causes a ~0.7 Å displacement of Ile31 in M26I DJ-1. This clash is relieved by the engineered M26L mutant, which also abrogates protein aggregation, indicating that the Ile26-Ile31 clash is an important contributor to

DJ-1 destabilization.[16, 55] The damaging effects of β-branched amino acids at residue 26 in DJ-1 is confirmed by the M26V mutation, which restores this clash and consequently causes aggressive aggregation of the protein. Our data indicates that the minor, local structural change caused by the introduction of a β-branched amino acid at residue 26 causes large-scale, global changes in DJ-1 slow conformational dynamics that result in protein instability. We also note that trace transition metals in the DJ-1 purified in this work seem to enhance aggregation, consistent with a prior finding.[39] This is intriguing, as iron (another transition metal) levels are elevated in parkinsonian nigral tissue [56] and may influence DJ-1 stability. The basis of this effect is a direction for future work.

M26I DJ-1 is a rare case where principally dynamical, rather than structural, changes in a protein can be directly connected with instability and disease. However, other examples where protein dynamics have been connected to pathological dysfunction include the deletion of a single amino acid (F508) in the Cystic fibrosis transmembrane conductance regulator (CFTR) protein, which causes areas of inherent disorder and flexibility resulting in disease.[57, 58] An additional example is copper-zinc superoxide dismutase (Cu-Zn SOD), where mutations associated with heritable forms of amylotropic lateral sclerosis (ALS) cause changes in protein dynamics.[59] DJ-1, which is an easily handled protein amenable to multiple biophysical approaches, makes an attractive model system in which to further explore the connection between disturbed protein dynamics and disease. We propose the aberrant protein dynamics may be a more common contributor to disease than currently appreciated.

More generally, our results indicate that routine *in vitro* characterization of proteins lower than physiological temperature provides an inaccurate view of their behavior. This has been well established in the extreme case of X-ray crystallography of cryo-cooled samples, where proteins are characterized at ~-170°C. In these cases, significant alterations to the degree and correlation of sidechain disorder that effect functionally relevant dynamics has been described.[60] The present study indicates that even a modest ~15°C difference between room temperature and physiological temperature (for humans) is enough to conceal important aspects of protein conformational behavior. As the characterization of protein dynamics becomes more widespread and routine, these results indicate that samples should be studied at their physiologically relevant temperatures wherever possible.

7.5 References

[1] Bonifati, V., Rizzu, P., van Baren, M. J., Schaap, O., Breedveld, G. J., Krieger, E., Dekker, M. C., Squitieri, F., Ibanez, P., Joosse, M., van Dongen, J. W., Vanacore, N., van Swieten, J. C., Brice, A., Meco, G., van Duijn, C. M., Oostra, B. A., and Heutink, P. (2003) Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism, *Science 299*, 256-259.

[2] Greenamyre, J. T., and Hastings, T. G. (2004) Biomedicine. Parkinson's--divergent causes, convergent mechanisms, *Science 304*, 1120-1122.

[3] Waak, J., Weber, S. S., Gorner, K., Schall, C., Ichijo, H., Stehle, T., and Kahle, P. J. (2009) Oxidizable residues mediating protein stability and cytoprotective interaction of DJ-1 with apoptosis signal-regulating kinase 1, *J Biol Chem 284*, 14245-14257.

[4] Jin, J., Li, G. J., Davis, J., Zhu, D., Wang, Y., Pan, C., and Zhang, J. (2007) Identification of novel proteins associated with both alpha-synuclein and DJ-1, *Mol Cell Proteomics 6*, 845-859.

[5] Clements, C. M., McNally, R. S., Conti, B. J., Mak, T. W., and Ting, J. P. (2006) DJ-1, a cancer- and Parkinson's disease-associated protein, stabilizes the antioxidant transcriptional master regulator Nrf2, *Proc Natl Acad Sci U S A 103*, 15091-15096.

[6] Zhou, W., and Freed, C. R. (2005) DJ-1 up-regulates glutathione synthesis during oxidative stress and inhibits A53T alpha-synuclein toxicity, *J Biol Chem 280*, 43150-43158.

[7] Kim, R. H., Peters, M., Jang, Y., Shi, W., Pintilie, M., Fletcher, G. C., DeLuca, C., Liepa, J., Zhou, L., Snow, B., Binari, R. C., Manoukian, A. S., Bray, M. R., Liu, F. F., Tsao, M. S., and Mak, T. W. (2005) DJ-1, a novel regulator of the tumor suppressor PTEN, *Cancer Cell 7*, 263-273.

[8] Tang, B., Xiong, H., Sun, P., Zhang, Y., Wang, D., Hu, Z., Zhu, Z., Ma, H., Pan, Q., Xia, J. H., Xia, K., and Zhang, Z. (2006) Association of PINK1 and DJ-1 confers

digenic inheritance of early-onset Parkinson's disease, *Hum Mol Genet 15*, 1816-1825.

[9] Canet-Aviles, R. M., Wilson, M. A., Miller, D. W., Ahmad, R., McLendon, C., Bandyopadhyay, S., Baptista, M. J., Ringe, D., Petsko, G. A., and Cookson, M. R. (2004) The Parkinson's disease protein DJ-1 is neuroprotective due to cysteine-sulfinic acid-driven mitochondrial localization, *Proc Natl Acad Sci U S A 101*, 9103-9108.

[10] Choi, J., Sullards, M. C., Olzmann, J. A., Rees, H. D., Weintraub, S. T., Bostwick, D. E., Gearing, M., Levey, A. I., Chin, L. S., and Li, L. (2006) Oxidative damage of DJ-1 is linked to sporadic Parkinson and Alzheimer diseases, *J Biol Chem 281*, 10816-10824.

[11] Meulener, M. C., Xu, K., Thomson, L., Ischiropoulos, H., and Bonini, N. M. (2006) Mutational analysis of DJ-1 in Drosophila implicates functional inactivation by oxidative damage and aging, *Proc Natl Acad Sci U S A 103*, 12517-12522.

[12] Mitsumoto, A., and Nakagawa, Y. (2001) DJ-1 is an indicator for endogenous reactive oxygen species elicited by endotoxin, *Free Radic Res 35*, 885-893.

[13] Blackinton, J., Lakshminarasimhan, M., Thomas, K. J., Ahmad, R., Greggio, E., Raza, A. S., Cookson, M. R., and Wilson, M. A. (2009) Formation of a stabilized cysteine sulfinic acid is critical for the mitochondrial function of the parkinsonism protein DJ-1, *J Biol Chem 284*, 6476-6485.

[14] Aleyasin, H., Rousseaux, M. W., Phillips, M., Kim, R. H., Bland, R. J., Callaghan, S., Slack, R. S., During, M. J., Mak, T. W., and Park, D. S. (2007) The Parkinson's disease gene DJ-1 is also a key regulator of stroke-induced damage, *Proc Natl Acad Sci U S A 104*, 18748-18753.

[15] Xu, J., Zhong, N., Wang, H., Elias, J. E., Kim, C. Y., Woldman, I., Pifl, C., Gygi, S. P., Geula, C., and Yankner, B. A. (2005) The Parkinson's disease-associated DJ-1 protein is a transcriptional co-activator that protects against neuronal apoptosis, *Hum Mol Genet 14*, 1231-1241.

[16] Lakshminarasimhan, M., Maldonado, M. T., Zhou, W., Fink, A. L., and Wilson, M. A. (2008) Structural impact of three Parkinsonism-associated missense mutations on human DJ-1, *Biochemistry 47*, 1381-1392.

[17] Malgieri, G., and Eliezer, D. (2008) Structural effects of Parkinson's disease linked DJ-1 mutations, *Protein Sci 17*, 855-868.

[18] Gorner, K., Holtorf, E., Odoy, S., Nuscher, B., Yamamoto, A., Regula, J. T., Beyer, K., Haass, C., and Kahle, P. J. (2004) Differential effects of Parkinson's disease-associated mutations on stability and folding of DJ-1, *J Biol Chem 279*, 6943-6951.

[19] Hulleman, J. D., Mirzaei, H., Guigard, E., Taylor, K. L., Ray, S. S., Kay, C. M., Regnier, F. E., and Rochet, J. C. (2007) Destabilization of DJ-1 by familial substitution and oxidative modifications: implications for Parkinson's disease, *Biochemistry 46*, 5776-5789.

[20] Blackinton, J., Ahmad, R., Miller, D. W., van der Brug, M. P., Canet-Aviles, R. M., Hague, S. M., Kaleem, M., and Cookson, M. R. (2005) Effects of DJ-1 mutations and polymorphisms on protein stability and subcellular localization, *Brain Res Mol Brain Res 134*, 76-83.

[21] Abou-Sleiman, P. M., Healy, D. G., Quinn, N., Lees, A. J., and Wood, N. W. (2003) The role of pathogenic DJ-1 mutations in Parkinson's disease, *Ann Neurol 54*, 283-286.

[22] Repici, M., Straatman, K. R., Balduccio, N., Enguita, F. J., Outeiro, T. F., and Giorgini, F. (2013) Parkinson's disease-associated mutations in DJ-1 modulate its dimerization in living cells, *J Mol Med (Berl) 91*, 599-611.

[23] Lakshminarasimhan, M., Madzelan, P., Nan, R., Milkovic, N. M., and Wilson, M. A. (2010) Evolution of new enzymatic function by structural modulation of cysteine reactivity in Pseudomonas fluorescens isocyanide hydratase, *J Biol Chem 285*, 29651-29661.

[24] Prahlad, J., Hauser, D. N., Milkovic, N. M., Cookson, M. R., and Wilson, M. A. (2014) Use of cysteine-reactive cross-linkers to probe conformational flexibility of human DJ-1 demonstrates that Glu18 mutations are dimers, *J Neurochem*.

[25] Glasoe, P. K., and Long, F. A. (1960) Use of glass electrodes to measure acidities in deuterium oxide, *J Phys Chem 64*, 188-190.

[26] Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes, *J Biomol NMR 6*, 277-293.

[27] Johnson, B., and Blevins, R. (1994) NMRView: A computer program for the visualization and analysis of NMR data, *J Biomol NMR 4*, 603-614.

[28] Mandel, A. M., Akke, M., and Palmer, A. G., 3rd. (1995) Backbone dynamics of Escherichia coli ribonuclease HI: correlations with structure and function in an active enzyme, *J Mol Biol 246*, 144-163.

[29] Palmer, A. r., Rance, M., and PE, W. (1991) Intramolecular motions of a zinc finger DNA-binding domain from Xfin characterized by proton-detected natural abundance carbon-13 heteronuclear NMR spectroscopy, *J. Am. Chem. Soc. 113*, 4371-4380.

[30] Cole, R., and Loria, J. P. (2003) FAST-Modelfree: a program for rapid automated analysis of solution NMR spin-relaxation data, *J Biomol NMR 26*, 203-213.

[31] Tjandra, N., Feller, S. E., Pastor, R. W., and Bax, A. (1995) Rotational diffusion anisotropy of human ubiquitin from 15N NMR relaxation, *J. Am. Chem. Soc. 117*, 12562-12566.

[32] Schanda, P., Kupce, E., and Brutscher, B. (2005) SOFAST-HMQC experiments for recording two-dimensional heteronuclear correlation spectra of proteins within a few seconds, *J Biomol NMR 33*, 199-211.

[33] Bai, Y., Milne, J. S., Mayne, L., and Englander, S. W. (1993) Primary structure effects on peptide group hydrogen exchange, *Proteins 17*, 75-86.

[34] Connelly, G. P., Bai, Y., Jeng, M. F., and Englander, S. W. (1993) Isotope effects in peptide group hydrogen exchange, *Proteins 17*, 87-92.

[35] Andersson, F. I., Werrell, E. F., McMorran, L., Crone, W. J., Das, C., Hsu, S. T., and Jackson, S. E. (2011) The effect of Parkinson's-disease-associated mutations on the deubiquitinating enzyme UCH-L1, *J Mol Biol 407*, 261-272.

[36] Mori, S., van Zijl, P. C., and Shortle, D. (1997) Measurement of water-amide proton exchange rates in the denatured state of staphylococcal nuclease by a magnetization transfer technique, *Proteins 28*, 325-332.

[37] Skinner, J. J., Lim, W. K., Bedard, S., Black, B. E., and Englander, S. W. (2012) Protein hydrogen exchange: testing current models, *Protein Sci 21*, 987-995.

[38] Skinner, J. J., Lim, W. K., Bedard, S., Black, B. E., and Englander, S. W. (2012) Protein dynamics viewed by hydrogen exchange, *Protein Sci 21*, 996-1005.

[39] Hulleman, J. D. (2007) Regulation of DJ-1 structure and function: Implications for Parkinson's Disease, In *Medicinal Chemistry and Molecular Pharmacology*, Purdue University, West Lafayette.

[40] Carrio, M. M., and Villaverde, A. (2001) Protein aggregation as bacterial inclusion bodies is reversible, *FEBS Lett 489*, 29-33.

[41] Carrio, M. M., and Villaverde, A. (2002) Construction and deconstruction of bacterial inclusion bodies, *J Biotechnol 96*, 3-12.

[42] Pantoliano, M. W., Petrella, E. C., Kwasnoski, J. D., Lobanov, V. S., Myslik, J., Graf, E., Carver, T., Asel, E., Springer, B. A., Lane, P., and Salemme, F. R. (2001) High-density miniaturized thermal shift assays as a general strategy for drug discovery, *J Biomol Screen 6*, 429-440.

[43] Scopes, R. K. (1974) Measurement of protein by spectrophotometry at 205 nm, *Anal Biochem 59*, 277-282.

[44] Yeh, J. I., and Hol, W. G. (1998) A flash-annealing technique to improve diffraction limits and lower mosaicity in crystals of glycerol kinase, *Acta Crystallogr D Biol Crystallogr 54*, 479-480.

[45] Otwinowski, Z., and Minor, W. (1997) *Processing of X-ray diffraction data collected in oscillation mode*, Vol. 276, Academic Press, New York.

[46] Wilson, M. A., Collins, J. L., Hod, Y., Ringe, D., and Petsko, G. A. (2003) The 1.1-A resolution crystal structure of DJ-1, the protein mutated in autosomal recessive early onset Parkinson's disease, *Proc Natl Acad Sci U S A 100*, 9256-9261.

[47] McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) Phaser crystallographic software, *J Appl Crystallogr 40*, 658-674.

[48] Collaborative Computational Project, N. (1994) The CCP4 suite: programs for protein crystallography, *Acta Crystallogr D Biol Crystallogr 50*, 760-763.

[49] Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics, *Acta Crystallogr D Biol Crystallogr 60*, 2126-2132.

[50] Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) Refinement of macromolecular structures by the maximum-likelihood method, *Acta Crystallogr D Biol Crystallogr 53*, 240-255.

[51] Brunger, A. T. (1992) Free R value: a novel statistical quantity for assessing the accuracy of crystal structures, *Nature 355*, 472-475.

[52] Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., 3rd, Snoeyink, J., Richardson, J. S., and Richardson, D. C. (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids, *Nucleic Acids Res 35*, W375-383.

[53] Ishima, R., and Torchia, D. A. (2000) Protein dynamics from NMR, *Nat Struct Biol 7*, 740-743.

[54] Logan, T., Clark, L., and Ray, S. S. (2010) Engineered disulfide bonds restore chaperone-like function of DJ-1 mutants linked to familial Parkinson's disease, *Biochemistry 49*, 5624-5633.

[55] Lin, J., Prahlad, J., and Wilson, M. A. (2012) Conservation of oxidative protein stabilization in an insect homologue of parkinsonism-associated protein DJ-1, *Biochemistry 51*, 3799-3807.

[56] Dexter, D. T., Carayon, A., Javoy-Agid, F., Agid, Y., Wells, F. R., Daniel, S. E., Lees, A. J., Jenner, P., and Marsden, C. D. (1991) Alterations in the levels of iron, ferritin and other trace metals in Parkinson's disease and other neurodegenerative diseases affecting the basal ganglia, *Brain 114 (Pt 4)*, 1953-1975.

[57] Kanelis, V., Hudson, R. P., Thibodeau, P. H., Thomas, P. J., and Forman-Kay, J. D. (2010) NMR evidence for differential phosphorylation-dependent interactions in WT and DeltaF508 CFTR, *EMBO J 29*, 263-277.

[58] Lewis, H. A., Wang, C., Zhao, X., Hamuro, Y., Conners, K., Kearins, M. C., Lu, F., Sauder, J. M., Molnar, K. S., Coales, S. J., Maloney, P. C., Guggino, W. B., Wetmore, D. R., Weber, P. C., and Hunt, J. F. (2010) Structure and dynamics of NBD1 from CFTR characterized using crystallography and hydrogen/deuterium exchange mass spectrometry, *J Mol Biol 396*, 406-430.

[59] Museth, A. K., Brorsson, A. C., Lundqvist, M., Tibell, L. A., and Jonsson, B. H. (2009) The ALS-associated mutation G93A in human copper-zinc superoxide dismutase selectively destabilizes the remote metal binding region, *Biochemistry 48*, 8817-8829.

[60] Fraser, J. S., Clarkson, M. W., Degnan, S. C., Erion, R., Kern, D., and Alber, T. (2009) Hidden alternative structures of proline isomerase essential for catalysis, *Nature 462*, 669-673.

Chapter 8

^{15}N CEST Data and Traditional Model-Free Analysis Capture Fast Internal Dynamics of DJ-1

8.1 Introduction

NMR spectroscopy is a powerful tool for the study of protein structures and dynamics in the solution state. Over the years, many NMR methods have been developed to observe protein dynamics for a range of timescales.[1] In which, fast timescale dynamics have been traditionally studied using ^{1}H-^{15}N HSQC R_1, R_2, and heteronuclear NOE experiments with the Carr-Purcell-Meiboom-Gill (CPMG) relaxation dispersion approach.[2] The T_1, T_2 and NOE data obtained from these experiments are routinely used to characterize sub-nano to millisecond protein dynamics with modelfree formalism.[3, 4] The CPMG approach has also been extended to the study of conformational exchange due to its sensitivity to chemical shift differences between ground and excited states.[5] However, CPMG relaxation dispersion fails for proteins undergoing slow conformational exchange or for lowly populated excited states.[6]

Recent advances employing saturation transfer, such as chemical exchange saturation transfer (CEST) and dark-state exchange saturation transfer (DEST), have enabled the detection of these previously invisible protein states.[6, 7] Several studies have already reported the use of CEST to study the invisible conformers of slowly exchanging proteins on the millisecond to second timescale.[8-12] Additionally, the fitting of CEST profiles have been shown to reliably extract R_1 and R_2 parameters that can be used for modelfree analysis of fast timescale dynamics (ps to ns). Thus, the simultaneous measurement of both fast and slow timescale dynamics is possible with the CEST experiment. The extraction of the R_1 and R_2 parameters is particularly advantageous due to the fact that CEST and CPMG experiments can be acquired in a similar amount of experimental time.[12] To date, however, no study has combined CEST-derived R_1 and R_2 parameters with ^{1}H-^{15}N NOE data to establish the picosecond to nanosecond dynamics of a protein. Instead, leaner versions of modelfree have been applied without the NOE data.[9]

The NOE is a sensitive measure of the high frequency motions as it reports directly on the structure of the protein and is strongly associated with its correlation time (□$_c$).[13] Therefore, the heteronuclear NOE experiment has been essential to traditional dynamics analyses in conjunction with R_1 and R_2 values. The importance of the NOE is strengthened by the fact that, at the expense of precise R_1 measurements, only precise NOE and R_2 values are necessary to calculate a reliable S^2.[14] Additionally, the NOE is more sensitive than the R_1 parameter for capturing internal dynamics.[10] The significance of the NOE to the understanding of fast protein dynamics is considerable and we present further evidence to substantiate the use of NOE data for modelfree analysis using ^{15}N-CEST experiments.

8.2 Material and Methods

NMR was used to investigate the dynamics of three DJ-1 oxidation states of Cys106 corresponding to a reduced form (DJ-1 Cys106-H), a sulfinic acid form (DJ-1 Cys106-SO_2) and a sulfonic acid form (DJ-1 Cys106-SO_3). The reduced DJ-1 Cys106 was oxidized to a cysteine sulfinic acid by adding hydrogen peroxide at a molar ratio of 1:7 and incubating on ice for 45 minutes. Similarly, the Cys106 sulfonic acid was prepared by adding hydrogen peroxide at a molar ratio of 1:100 and incubation at room temperature for 2.5 hours. Following incubation, the protein was subjected to a buffer exchange to remove excess hydrogen peroxide. All NMR experiments were collected at 37 or 35^{o}C on a 700 MHz Bruker Avance III spectrometer equipped with a 5 mm QCI-P probe with cryogenically cooled carbon and proton channels. ^{15}N CEST experiments were performed as previously described[6] with B_1 field strengths of 12.5 and 25 Hz. The external field was scanned across the range of 100 to 140 ppm at a step size of 50 Hz (35^{o}C) or 25 Hz (37^{o}C). An interscan delay of 1.5s was employed and ^{15}N saturation was applied for 0.5s. For each experiment a total of 4 transients were collected per 2D plane resulting in a total experiment time of approximately 7 hours. Processing was accomplished with NMRPipe[15] and CEST profiles were fit with ChemEx (Figure 1).[6]

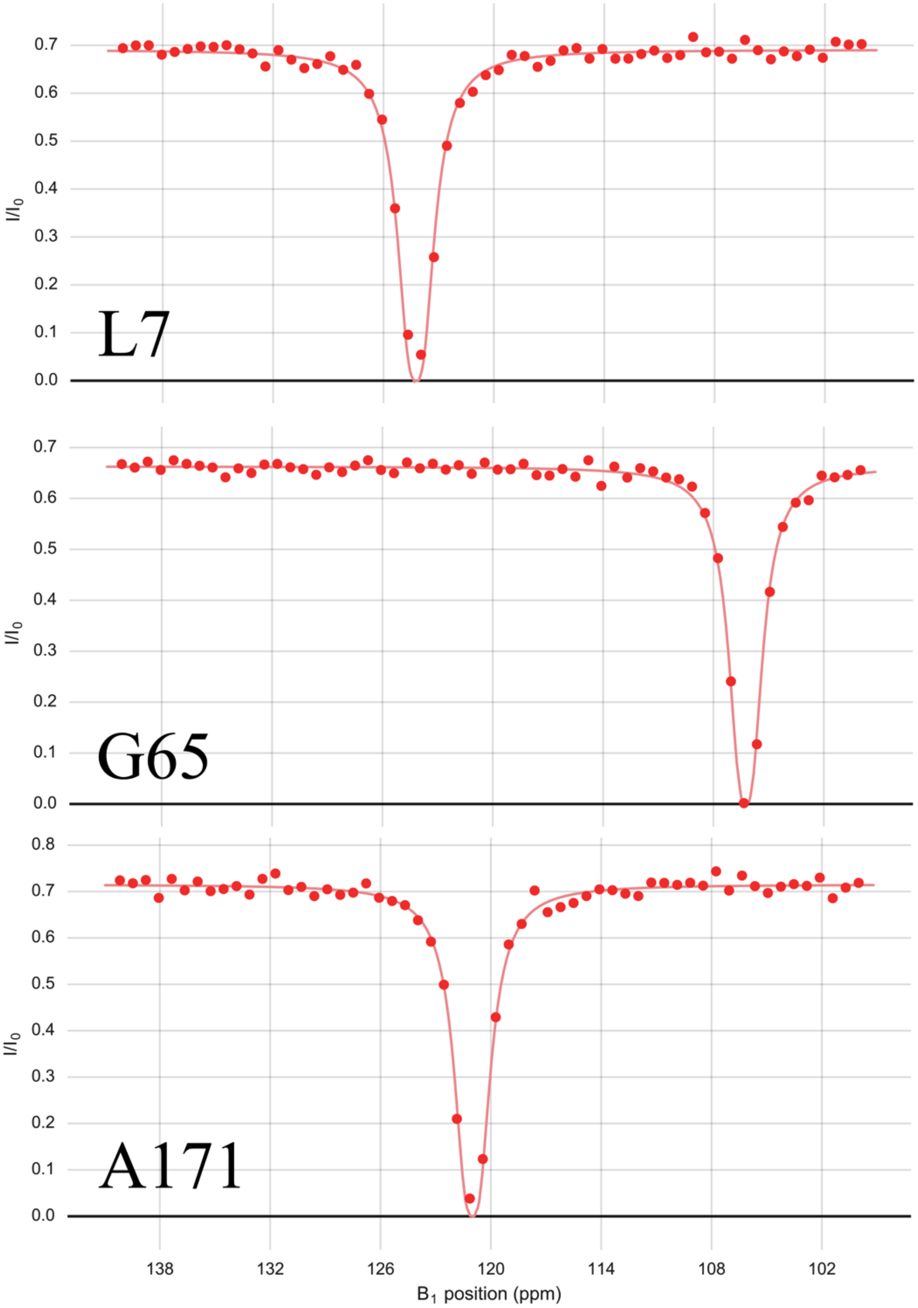


Figure 1. Representative CEST profiles for 3 residues in DJ-1 Cys16-H at 35^{o}C. The profiles show the proper fitting of the dip in intensity and the lack of a noticeable minor state conformation. Residues were chosen based on position in the primary sequence to highlight the consistence of the fitting of the profiles.

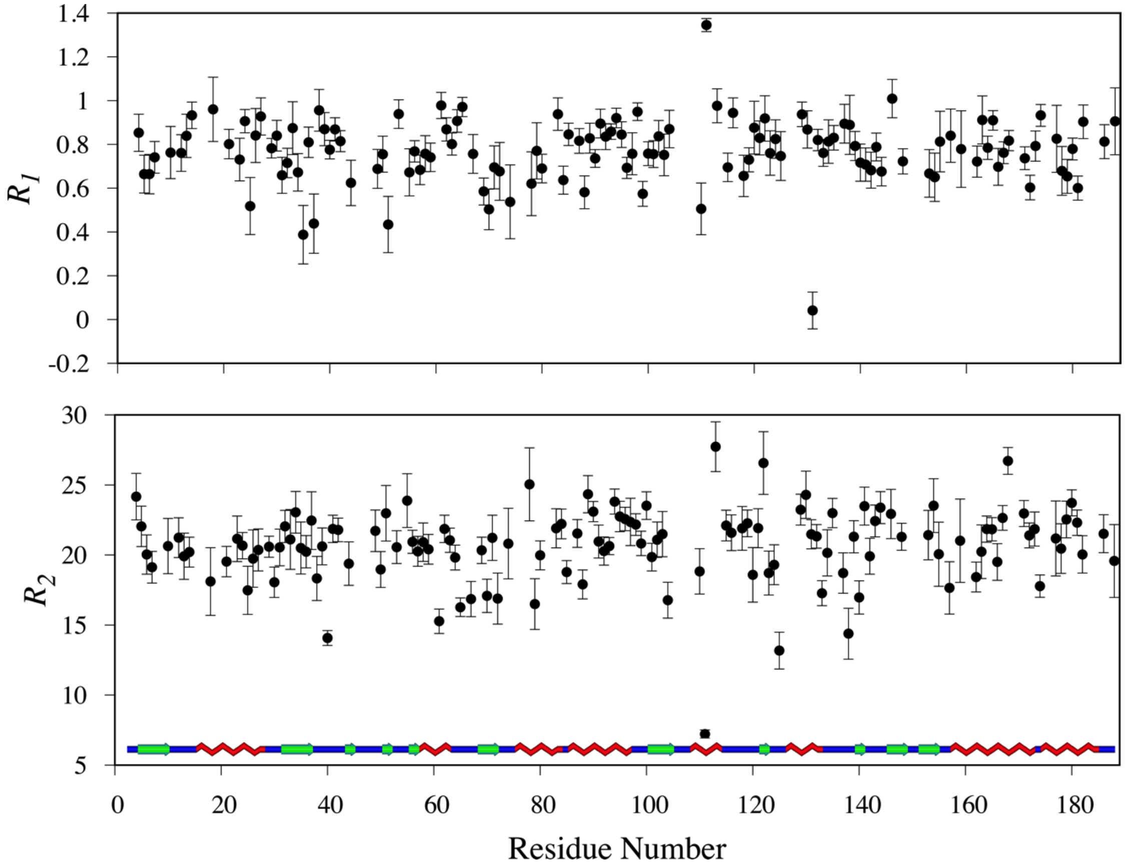


Figure 2. A plot of the CEST derived R_1 and R_2 parameters for DJ-1 Cys106-H at 35°C. The cartoon[5] represents the secondary structure of DJ-1 with helices in red, strands in green, and loops and disordered regions in blue.

The fitting of the CEST profiles by ChemEx produces R_1 and R_2 relaxation parameters as standard output and is explained in detail in Vallurupalli et al., *JACS* (2012).[6] The R_1, R_2, NOE, and S^2 errors reported in Tables 1 and 2 were calculated as the standard error of the mean. Heteronuclear NOE experiments were performed as previously described[16] and processed in NMRPipe. FAST-ModelFree analysis was done identical to our previous study to ensure a proper comparison of the methods.[16]

8.3 Results and Discussion

The fast protein dynamics of reduced DJ-1 have been previously determined using traditional spin relaxation derived R_1 and R_2 values and ^{1}H-^{15}N NOE data.[16] DJ-1 was found to be fairly rigid in solution at 35˚C, with most order parameters (S^2) approaching the maximum theoretical limit ($S^2 = 1$). Importantly, the previous dynamics finding from these traditional relaxation experiments was reproduced with the R_1 and R_2 values obtained from fitting of ^{15}N CEST profiles and from ^{1}H-^{15}N NOE data (Table 1, Figure 2). Per residue CEST profiles were reliably fit simultaneously for both external fields, although only one field was necessary (Figure 1). Unfortunately, the fitting of the CEST profiles did not reveal the presence of a minor state or chemical exchange, but this finding did not negatively affect the quality of the derived relaxation parameters. Our results further establish the robustness of the ^{15}N CEST experiment and justify its adoption as an alternative to traditional relaxation experiments. The observed consistency between two distinct sets of R_1, R_2 and S^2 values substantiates previous claims that CEST data is equivalent to traditional spin relaxation experiments.[9] The calculated RMSDs values (Table 3) comparing the CEST and traditional R_1 and R_2 values are within the observed standard deviations for the data sets (Tables 1 and 2). An NOE RMSD of 0.13 was obtained when we compared our original ^{1}H-^{15}N NOE data[16] with an NOE data set recollected with the CEST experiments. Again, the NOE RMSD is within the observed experimental error and standard deviations (Tables 1 and 2).

Table 1. Average R_1, R_2, NOE, and S^2 values from DJ-1 Cys106-H at 35°C.

Exp. Type	Protein	Observed R_1 [a]	R_1 Error[b]	Observed R_2 [a]	R_2 Error[b]	Observed NOE[a]	NOE Error[b]	Calculated S^2[a]	S^2 Error[b]
Traditional (35°C)	DJ-1, Cys106-H	0.78 (0.10)	0.06	19.30 (2.30)	0.7	0.79 (0.11)	0.15	0.92 (0.10)	0.02
CEST (35°C)	DJ-1, Cys106-H	0.72 (0.21)	0.06	19.33 (3.50)	2.1	0.79 (0.15)	0.11	0.88 (0.12)	0.07

[a]Standard deviations are in parenthesis. [b]The reported errors are the standard error of the mean.

Table 2. Average R_1, R_2, NOE, and S^2 values from different physiological states of DJ-1.

Exp. Type	Protein	Observed R_1 [a]	R_1 Error[b]	Observed R_2 [a]	R_2 Error[b]	Observed NOE [a]	NOE Error	Calculated S^2[a]	S^2 Error[b]
CEST (37°C)	DJ-1, Cys106-H	0.72 (0.22)	0.06	19.50 (3.46)	2.1	0.79 (0.16)	0.11	0.86 (0.15)	0.07
	DJ-1, Cys106-SO_2	0.78 (0.13)	0.08	20.65 (2.66)	1.3	0.80 (0.16)	0.14	0.92 (0.13)	0.04
	DJ-1, Cys106-SO_3	0.71 (0.21)	0.06	18.47 (4.96)	1.0	0.64 (0.42)	0.10	0.76 (0.22)	0.04

[a]Standard deviations are in parenthesis. [b]The reported errors are the standard error of the mean.

Table 3. RMSD values comparing traditionally collected and CEST datasets for DJ-1 Cys106-H at 35°C.

	R_1	R_2	S^2
Traditional vs CEST	0.18	1.95	0.02

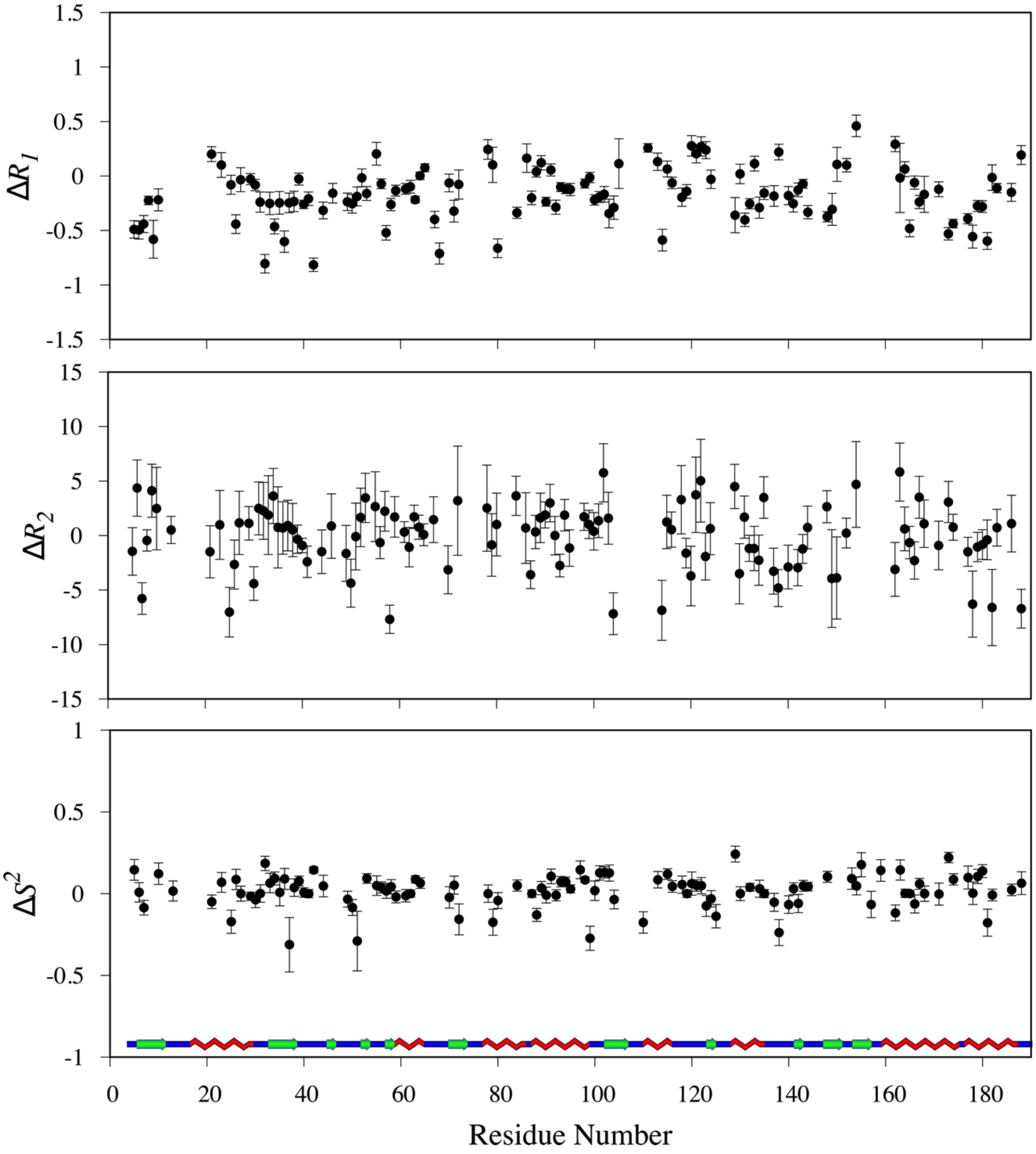


Figure 3. A delta plot of the R_1, R_2, and generalized order parameters comparing traditionally collected data versus CEST derived data for DJ-1 Cys106-H at 35°C. The cartoon[5] represents the secondary structure of DJ-1 with helices in red, strands in green, and loops and disordered regions in blue.

The per residue delta plots of the R_1, R_2 and S^2 values shown in Figure 3 further illustrates the consistency between the two methods. The low variability in the R_2 delta plot indicates that the CEST profile linewidths were robust in the reporting of the true R_2 values and that there was an absence of chemical exchange. Only a slight overall difference of the methods was observed based on an S^2 RMSD of 0.02. This is despite the fact that the prior dynamics measurements were carried out on different NMR spectrometers at different field strengths (*i.e.*, 500 MHz vs. 700 MHz). The largest variations in the observed ΔS^2 values were primarily located in loop regions, which are expected to be more sensitive to subtle temperature and solvent differences. A few residues with large ΔS^2 were removed from the comparison since the ΔS^2 error was likely a result of peak overlap and not representative of the overall quality of the CEST data. It is again important to note that these datasets were collected several years apart on different protein samples, on different spectrometers, and by different spectroscopists. Therefore, although the reproducibility between the two datasets is not on par with prior results,[9] our comparison does represent a more realistic scenario and a more accurate representation of the true uncertainty. Importantly, the final results of the modelfree analyses from both the CEST and CPMG datasets led to the same conclusions regarding the overall dynamic properties of DJ-1.

Leveraging the ^{15}N-CEST experiment, several naturally occurring oxidation states of DJ-1 were studied (Cys106-H, -SO_2, -SO_3) to identify if there were any differences in their dynamic properties. At physiological temperature (37^{o}C), none of the three forms of DJ-1 exhibited a minor state conformer based on analysis of the CEST profiles. The CEST-derived R_1 and R_2 parameters for each of the DJ-1 states were further analyzed to identify any ps-ns timescale dynamics that may differ between the three states. Interestingly, there were only minor differences in the R_1 and R_2 relaxation parameters between the three states in addition to being comparable to our previous results (Tables 1 and 2). ^1H-^{15}N NOE experiments were subsequently performed to further explore the dynamic properties of the three DJ-1 states. A drastic change was readily apparent in DJ-1 Cys106-SO_3, where the average NOE is lower, and the NOE standard deviation is much larger compared to the two other DJ-1 states (Table 2). The CEST derived R_1 and R_2 parameters and the heteronuclear NOE values were then used with FAST-ModelFree[17-19] to determine the per residue order parameters. The dynamic properties for the first two DJ-1 states are quite similar to our previously published results for the reduced form of DJ-1, where the protein exhibits minimal motion on the ps-ns time scale. Conversely, DJ-1 Cys106-SO_3 exhibits a significant increase in dynamics based on an overall lower S^2 value of 0.76 ± 0.22. Importantly, it appears that the main contributor to the increase in dynamics for DJ-1 Cys106-SO_3 is a reduction in the average NOE from approximately 0.80 to 0.64 ± 0.42. Further, it is well known that NOE data is essential for the accurate analysis of protein dynamics since the NOE is very sensitive to fast internal dynamics (e.g., $\square_e$), as previously defined in Kay *et al*. (1989).[10] In addition, the large standard deviation in the NOEs for DJ-1 Cys106-SO_3 is also indicative of a substantial decrease in structural stability. This is consistent with the observation that DJ-1 Cys106-SO_3 was less stable than the other two forms of the protein. DJ-1 Cys106-SO_3 would aggregate and precipitate out of solution after approximately two days at 37°C. Conversely, the reduced and sulfinic forms of DJ-1 were indefinitely stable.

To gain insight on the impact of the NOEs on the per residue order parameters, the NOE data from DJ-1 Cys106-SO_3 was combined with the R_1 and R_2 data from DJ-1 Cys106-SO_2. DJ-1 Cys106-SO_2 is arguably the most rigid form of DJ-1 since it has the highest average order parameter of 0.92 ± 0.13. Moreover, DJ-1 Cys106-SO_2 has the lowest R_1 and R_2 standard deviations of 0.13 and 2.66, respectively. Any changes in the order parameters calculated by combining the DJ-1 Cys106-SO_2 R_1 and R_2 data with the DJ-1 Cys106-SO_3 NOE data will highlight the impact of the more dynamic NOEs on the calculated order parameters. Figure 4 shows the ΔS^2 obtained from comparing the DJ-1 Cys106-H and Cys106-SO_3 NOE data. It is readily apparent that the more dynamic NOEs contribute significantly to the per residue order parameters as approximately 10 residues order parameters decreased by -0.25 or more. These differences are noteworthy because they may change the classification of a residue from "ordered" to "disordered". Interestingly, a majority of the ΔS^2 values are near zero, which indicates essentially identical dynamics between the two states. Not surprisingly, only the residues with a significant difference in NOE values were substantially affected. It is likely that these DJ-1 residues in the Cys106-SO_3 form have sub-ns internal motions, which make them sensitive to the ^1H-^{15}N NOE experiment. These findings are significant because exclusion of any NOE data would have hidden the true dynamics of these particular DJ-1 residues.

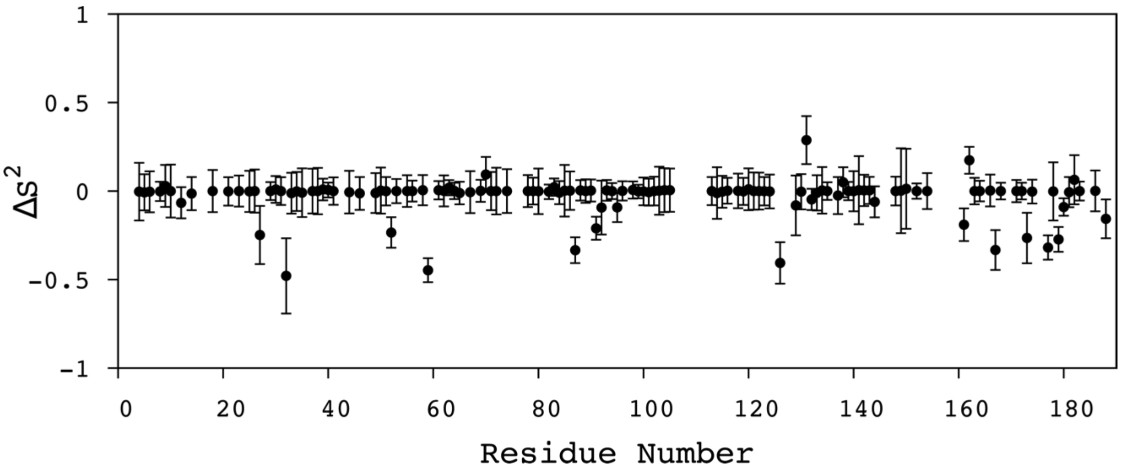


Figure 4. A delta plot of the generalized order parameters comparing DJ-1 Cys106-H and DJ-1 Cys106-SO_3 NOE data. Data points below zero indicate an increase in internal motion due to the DJ-1 Cys106-SO_3 NOE value.

8.4 Conclusion

The purpose of this study was to further evaluate the robustness of the ^{15}N CEST experiments as an alternative method to determine the fast protein dynamics of three physiological states of DJ-1 while combining CEST derived R_1 and R_2 parameters with traditional NOE values. Initial results confirmed that this new method was comparable to traditional dynamics analyses as our results correspond well with previous findings. We have also reaffirmed the importance of the heteronuclear NOE to a dynamics analysis. Extremely low NOE values are strong indicators of protein disorder or unfolding and, in this study, were the only experimental evidence for this phenomenon in DJ-1 since the R_1 and R_2 parameters were similar for the three DJ-1 states. To our knowledge, this is the first study that combines ^{15}N CEST data with a traditional model-free analysis of a biological system and also confirms the necessity of NOE data for exploring fast protein dynamics.

8.5 References

[1] Palmer, A. G., Kroenke, C. D., and Patrick Loria, J. (2001) Nuclear Magnetic Resonance Methods for Quantifying Microsecond-to-Millisecond Motions in Biological Macromolecules, *Proteins 339*, 204-238.

[2] Kay, L. E. (2016) New Views of Functionally Dynamic Proteins by Solution NMR Spectroscopy, *J Mol Biol 428*, 323-331.

[3] Lipari, G., and Szabo, A. (1982) Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 1. Theory and Range of Validity, *J Am Chem Soc 104*, 4546-4559.

[4] Clore, G. M., Szabo, A., Bax, A., Kay, L. E., Driscoll, P. C., and Gronenborn, A. M. (1990) Deviations from the Simple Two-Parameter Model-Free Approach to the Interpretation of Nitrogen-15 Nuclear Magnetic Relaxation of Proteins, *J Am Chem Soc 112*, 4989-4991.

[5] Tollinger, M., Skrynnikov, N. R., Mulder, F. A. A., Forman-Kay, J. D., and Kay, L. E. (2001) Slow dynamics in folded and unfolded states of an SH3 domain, *J. Am. Chem. Soc. 123*, 11341-11352.

[6] Vallurupalli, P., Bouvignies, G., and Kay, L. E. (2012) Studying “Invisible” Excited Protein States in Slow Exchange with a Major State Conformation, *J Am Chem Soc 134*, 8148-8161.

[7] Fawzi, N. L., Ying, J., Ghirlando, R., Torchia, D. A., and Clore, G. M. (2011) Atomic-resolution dynamics on the surface of amyloid-beta protofibrils probed by solution NMR, *Nature 480*, 268-272.

[8] Anthis, N. J., and Clore, G. M. (2015) Visualizing transient dark states by NMR spectroscopy, *Q Rev Biophys 48*, 35-116.

[9] Gu, Y., Hansen, A. L., Peng, Y., and Bruschweiler, R. (2016) Rapid Determination of Fast Protein Dynamics from NMR Chemical Exchange Saturation Transfer Data, *Angew Chem Int Ed Engl 55*, 3117-3119.

[10] Kay, L. E., Torchia, D. A., and Bax, A. (1989) Backbone Dynamics of Proteins As Studied by 15N Inverse Detected Heteronuclear NMR Spectroscopy: Application to Staphylococcal Nuclease, *Biochemistry 28*, 8972-8979.

[11] Abragam, A. (1961) *The Principles of Nuclear Magnetism,*, Clarendon Press, Oxford.

[12] Valluruapalli, P., Sekhar, A., Yuwen, T., and Kay, L. E. (2017) Probing conformational dynamics in biomolecules via chemical exchange saturation transfer: a primer, *J. Biomol. NMR*, Ahead of Print.

[13] Gust, D., Moon, R. B., and Roberts, J. D. (1975) Applications of natural-abundance nitrogen-15 nuclear magnetic resonance to biochemically important molecules, *Proc Natl Acad Soc USA 72*, 4696-4700.

[14] Jin, D., Andrec, M., Montelione, G. T., and Levy, R. M. (1998) Propagation of experimental uncertainties using the Lipari-Szabo model-free analysis of protein dynamics, *J Biomol NMR 12*.

[15] Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: A multidimensional spectral processing system based on UNIX pipes, *Journal of Biomolecular NMR 6*, 277-293.

[16] Milkovic, N. M., Catazaro, J., Lin, J., Halouska, S., Kizziah, J. L., Basiaga, S., Cerny, R. L., Powers, R., and Wilson, M. A. (2015) Transient sampling of aggregation-prone conformations causes pathogenic instability of a parkinsonian mutant of DJ-1 at physiological temperature, *Protein Sci 24*, 1671-1685.

[17] Palmer, A. G., Rance, M., and Wright, P. E. (1991) Intramolecular Motions of a Zinc Finger DNA-Binding Domain from Xfin Characterized by Proton-Detected Natural Abundance 13C Heteronuclear NMR Spectroscopy, *J Am Chem Soc 113*, 4372-4380.

[18] Mandel, A. M., Akke, M., and Palmer, A. G. (1995) Backbone Dynamics of Escherichia coli Ribonuclease HI: Correlations with Structure and Function in an Active Enzyme, *J Mol Biol 246*, 144-163.

[19] Cole, R., and Loria, J. P. (2003) FAST-ModelFree: A program for rapid automated analysis of solution NMR spin-relaxation data, *Journal of Biomolecular NMR 26*, 203-213.

Chapter 9

Summary and Future Directions

9.1 Summary

This dissertation has covered a wide variety of research projects revolving around the sequence, structure, function, and dynamics of proteins. Using both conventional and novel biophysical and bioinformatics approaches, we have made important contributions in the expansive field of structural biology. Structures for therapeutic and putative protein targets have been elucidated and advancements in evolutionary analysis have been discovered. Furthermore, we have reaffirmed the crucial role of NMR in the study of proteins. NMR has and will continue to be a significant instrument for probing proteins in environments closest to their most natural and physiological state.

Beginning with Chapter 2, a novel approach to the evolutionary analysis of distant proteins was introduced. By simply using comparisons of active site structures we quickly and efficiently mapped the functional evolution of proteins regardless of their sequence or structure similarity. This method was first proven on the pyridoxal 5'-phosphate (PLP) enzyme family, which began their divergence in the last common ancestor. Altogether, the PLP family has approximately 10% sequence identity within four structurally independent fold-types. Thus, conventional sequence based studies for the evolutionary analyses of this family have been narrowly focused on small subsets of functionally similar PLP enzymes. The application of CPASS, however, enabled the entirety of the family to be studied simultaneously and presented on a single phylogenetic network. Within the generated network was a perfect clustering of the enzymes by fold-type, enzymatic function, and organism classification. Our bioinformatics investigation of the functional evolution of PLP-enzymes represents the first time ever that a single phylogentic tree was constructed that contained the four PLP fold-types. Consequently, our phylogentic tree captures the step-wise evolution of function for the entire PLP enzyme family.

The success of the PLP active site evolutionary study led to a similar but greatly expanded project aimed at all of the ligand-bound proteins in the PDB. Approximately 40,000 unique active sites exist in the PDB, which is significantly larger than the PLP set used previously (~200). To reduce the overall size of the data and improve visualization of the end results, the unique active sites were first grouped according to the identity of the bound ligand. A statistical tool known as Principal Component Analysis (PCA) was applied to each ligand defined group to identify a representative structure. This proved to be an effective approach to reducing the number of active sites to 10% of the total (~4000) while maintaining essential information since a representative structure was easily identified from the PCA scores plot for each ligand defined group. The resulting phylogenetic tree based on a CPASS similarity scoring matrix for the 4000 representative ligand binding sites resulted in a near-perfect functional clustering of the active sites based on EC function. Importantly, the phylogenetic tree enables the functional evolution at the residue level to be visualized for the entirety of protein functions present in the RCSB PDB. This is a major achievement and advancement in our understanding of the evolution of protein function since it represents the first time the entire RCSB PDB has been functionally mapped according to active site structure similarity.

Obtaining a protein structure is fundamental to our understanding of the biological function of a protein; and its possible role in human disease; and as a either a therapeutic target or a new therapy. Accordingly NMR was used to investigate the structure of the Citropin 1.1 peptide and its analogs, which are potential novel antimicrobial therapies; the structure and dynamics of the *S. aureus* protein primase CTD, which is a potential novel therapeutic target for the development of next generation antibiotics; the dynamics of the human protein DJ-1 and its mutants, which is an important therapeutic target for Parkinson's disease; and the structure and potential function for the functionally uncharacterized *R. palustris* protein RPA3313. The NMR structures of the Citropin 1.1 peptide and its analogs were presented in Chapter 4. We found that the cationic residues at the N-terminus of the peptide formed essential hydrogen bonds and electrostatic interactions that synergistically worked to promote antimicrobial activity. Importantly, our structural analysis of the Citropin 1.1 peptides clearly demonstrated that the peptides adopt a single helical structure and forms a head to tail dimer in the presence of membrane-mimic. Thus, our results correct a mistake propagated in the literature that Citropin 1.1 peptides are a monomer composed of two separate helical regions. Moreover, we discovered an equilibrium between disordered monomer and helical dimer structures that was directly correlated with antimicrobial activity. Fluorescence and microscopy experiments indicated that the peptides mechanism of action involved disrupting the bacterial membrane through a "carpet model". Our results identified sequence and structural features of the Citropin 1.1 peptide that are critical for antimicrobial activity. Furthermore, our results provide support for the peptides' mechanism of antimicrobial activity. This information is essential to evolve the Citropin 1.1 peptide into a drug-like candidate and to develop a new antibiotic.

Chapters 5 and 6 presented the NMR structure determination of the *R. palustris* RPA3313 protein and the *S. aureus* primase C-terminal domain (SaP-CTD), respectively. In both instances, the NMR solution structures provided critical evidence to possible protein function, protein-protein interaction sites, and putative ligand binding sites. RPA3313 was observed to adopt a novel fold and was found to interact with the ribosome in *R. palustris.* Thus, the RPA3313 NMR structure extended our knowledge of protein conformational space through the identification of a new protein fold. Significantly, the RPA3313 structure can be extended to homologs of RPA3313, since the protein family was structurally uncharacterized prior to our study. While the structure of RPA3313 was used to infer a potential function, the NMR solution structure of SaP-CTD was used to identify potential ligands and ligand binding sites. This information is essentially to assessing the potential therapeutic utility of SaP-CTD as a drug discovery target. The NMR structure for SaP-CTD indicates that the protein is composed of two subdomains. Importantly, the relative orientations of the two subdomains (open or closed conformation) are correlated with helicase binding. SaP-CTD adopts the closed conformation, which does not bind helicase. NMR dynamics experiments indicated that the two subdomains of SaP-CTD moved independently of one another with a binding site located between the two subdomains. An *in silico* virtual screen and subsequent NMR titration experiment found significant binders which could stabilize the two subdomains in the closed conformation and prevent proper protein function. In essence, our structural and dynamics analysis of SaP-CTD identified a potential novel ligand binding site, a set of potential ligands and, importantly, a potential novel mechanism of inhibiting the primase and helicase interaction by capitalizing on the dynamic properties of primase. The relative flexibility between the two SaP-CTD subdomains allows the protein to readily adopt the open conformation and bind primase. Conversely, a ligand that binds between the two subdomains and locks the protein in the closed conformation may be a valuable antibiotic. In both projects, the NMR determined structures proved to be essential to the understanding of the proteins' biological functions and therapeutic utility.

The importance of NMR to the study protein dynamics was further demonstrated in our analysis of DJ-1 presented in Chapters 7 and 8. Human protein DJ-1 is a ubiquitously expressed homodimer, however, its molecular activity is largely unknown. M26I DJ-1 is an inheritable mutant that has been implicated in Parkinson's disease; and adopts a fold structurally identical to the wild-type protein. The "benign" mutation does not alter the tertiary structure or disrupt dimer formation, yet it still leads to protein instability and disease. NMR relaxation studies revealed that the wild-type and M26I DJ-1 proteins are both rigid in solution. A hydrogen-deuterium exchange experiment discovered that the hydrophobic core of M26I DJ-1 was transiently exposed to solvent, which enabled it to subsequently aggregate, which, in turn, leads to a loss of function, and, presumably, is a cause of Parkinson's disease. Thus, the NMR dynamics has provided an important insight into a potential molecular mechanism associated with Parkinson's disease. This important finding further illustrates that changes in dynamics, rather than structure, may also be connect to protein instability and the diseased state of the protein.

DJ-1 dynamic properties were also investigated as a result of changes in the protein's oxidative state. It was found that over-oxidized DJ-1 was extremely unstable in solution and began to unfold and aggregate at physiological temperature. This raised the possibility that DJ-1 may be regulated by its oxidation state and may serve as a cellular sensor for oxidative stress. Concurrently, improvements in the NMR experimental methods for probing protein dynamics were accomplished while investigating the dynamics of the oxidative states of DJ-1. Chemical exchange saturation transfer (CEST) experiments can be used to visualize the transient, invisible states of proteins in solution. Although no invisible state of DJ-1 was found, the fitting of the CEST profiles produced relaxation parameters of the fast timescale dynamics that were consistent with our previous findings. This observation demonstrated that the dynamics at multiple timescales can be studied simultaneously and expand the capabilities of NMR spectroscopy. This was particularly important in the case of DJ-1 because the CEST experiments were collected significantly faster than the traditional methods that required individual set of experiments to measure R_1, R_2 and NOEs. Rapid data collection was critical because different forms of DJ-1 had limited long-term stability.

9.2 Future Directions

Improvements to CPASS in the future should be aimed at increasing the utility of the software toward the functional annotation of proteins as well as increasing its usefulness as an evolutionary tool. Each day more active sites are deposited in the PDB, and it will become necessary to re-evaluate the active site representative tree. Although it is unlikely that new representative active sites will need to be chosen, it is very likely that new functions are added or that improved annotation tools are developed. I firmly believe in the tenet, “If its not broken, don’t fix it”. CPASS is proven in its ability to annotate and evolutionarily relate active sites. Future efforts should be aimed at increasing the portability of the software and expanding its capabilities rather than modifications to the scoring function. Merging CPASS with an existing, popular software tool will undoubtedly increase its overall use. In parallel, the development of a process or pipeline to dock ligands *in silico* to both known and putative binding sites prior to CPASS analysis is necessary. This improvement would not only increase the use of CPASS, but could also significantly decrease experimental time in the determination of protein function.

NMR will always maintain a significant role in the field of structural biology. For decades NMR has been used to solve the structures of proteins in the solution state. Improvements in the hardware and software of spectrometers; as well as in the preparation of the samples has significantly expanded the range of proteins that can be studied. However, the future of the field involves studying membrane proteins and their structures in membrane mimics. Significant development is still required to study membrane proteins by NMR. New isotopic labeling schemes, pulse programs, and data collection methods will all be needed to study membrane proteins. Although the task seems insurmountable, the ability to elucidate the structure of a protein in the solution state is valuable, as it closely resembles the physiological state of the protein.

NMR is one of the few techniques that are capable of studying the dynamics of proteins. Already discussed in this dissertation is the possibility that dynamics may play a more significant role in disease than previously considered. That being said, future developments are of the utmost importance. The extraction of even more information from the data rich NMR spectra will provide key clues into the dynamics of invisible or transient conformers. Additionally, it should be possible to collect distance restraints on the invisible states by using a paramagnetic ion or spin label in combination with a saturation transfer experiment. The distance restraints could then be used to refine the invisible state structure, which would be a significant accomplishment. Lastly, these structures would provide researchers with a superior target for drug design and development.