AN ABSTRACT OF THE DISSERTATION OF

Donald S. Berkholz for the degree of Doctor of Philosophy in Biochemistry and Biophysics presented on August 20, 2009.

Title: Modeling Protein Structure at Atomic Resolution

Abstract approved:

_____

P. Andrew Karplus

This thesis includes three studies involving different aspects of modeling protein structure. The first study illustrates the levels of insight available from atomic-resolution protein structures. The second study derives general trends of protein geometry from atomic-resolution structures and shows their implications for modeling. The third study creates a model of a protein and uses it to derive new biological insights.

In the first study, a series of structures were analyzed from human glutathione reductase, a biomedically relevant enzyme. Newly accessible at atomic resolution is structural evidence showing the catalytic importance of active-site compression, which additionally causes distortions from standard geometry that further enhance catalytic power. Another aspect of geometry visible at atomic resolution is the remarkably ideal positioning of atoms for catalysis. The stereoelectronic control displayed by compression and geometric preorganization provides insight into the origins of catalytic power.

The second study builds upon quantum-mechanics calculations and empirical analyses of protein structure from the 1990s that showed the concept of a single ideal value for backbone geometry was wrong. Here, a nonredundant set of protein structures

at atomic resolution is probed to better define the dependence of backbone geometry upon the conformation of the backbone torsion angles Φ and Ψ. The set was taken from the Protein Geometry Database created here (http://pgd.science.oregonstate.edu/). The trends seen make structural sense and lay the groundwork for a paradigm shift in the concept of ideal geometry. A conformation-dependent library accounting for these trends has the potential to improve modeling accuracy.

In the third study, a model of the tumor-suppressor merlin is created and used to gain new understanding of merlin's function. Merlin is the only known cytoskeletal tumor suppressor, and loss of functional merlin results in neurofibromatosis 2, characterized by nervous-system tumors, cataracts, and skin tumors. Clear errors were evident in available automatically created models, driving the need for a reliable structure. Merlin and its homologs have distinct functions, so the differences between them were probed, suggesting critical functional clusters. A new technique developed here for discovering gains and losses of function should be generally applicable to any two protein subfamilies with distinct functions.

Modeling Protein Structure at Atomic Resolution

by
Donald S. Berkholz

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented August 20, 2009
Commencement June 2010

Doctor of Philosophy dissertation of Donald S. Berkholz presented on August 20, 2009.

APPROVED:

Major Professor, representing Biochemistry and Biophysics

Chair of the Department of Biochemistry and Biophysics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Donald S. Berkholz, Author

ACKNOWLEDGEMENTS

First and foremost, I thank my adviser, Andy Karplus, for turning me into a real scientist. He helped me develop better scientific communication, a stronger work ethic, and the beginnings of the same sense of detail he has. Through his enthusiasm and dedication to what he does every day, he embodied my role model of a great scientist.

I thank all my labmates, particularly these: Rick Faber, who shared his in-depth knowledge of the depths of crystallography; Ganapathy Sarma, who showed me what the end of grad school looked like when I was just beginning; Dale Tronrud, who every day exemplified his requirements for perfection; Andrea Hall, who helped me to become a more organized person through her example; and Scott Hollingsworth and Camden Driggers, who let me inflict my inexperienced mentoring upon them.

I thank my graduate committee: Joe Beckman, Elisar Barbar, Weng-Keen Wong, Kaichang Li, and my adviser, Andy. Over the years, all of them have helped me at some level, and I'm grateful for all of their time and advice.

I thank the funding agencies that made it possible for me me to undertake this work: the National Institutes of Health, the National Science Foundation, and the Department of Defense.

Finally, I thank my family for helping me cope with the demands of grad school.

CONTRIBUTION OF AUTHORS

Savvas Savvides performed the crystallization and data collection for Chapter 2. Rick Faber processed and refined the atomic-resolution structures in Chapter 2, except for the final rounds. Peter Krenesky designed the database backend for Chapter 3 and made its graphical interface follow specifications. Both Peter Krenesky and John Davidson wrote code for Chapter 3. Max Shapovalov used our data to create the kernel-regression graphs in Chapter 4. Roland Dunbrack, Jr., and Max Shapovalov wrote the methods section describing the kernel methods in Chapter 4. Roland Dunbrack, Jr., also contributed minor suggestions to the rest of the chapter. Tony Bretscher contributed suggestions to Chapter 5. Andy Karplus was involved with the design, interpretation, and writing of all chapters.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF TABLES

Modeling Protein Structure at Atomic Resolution

Chapter 1

General Introduction

Introduction

Life depends on proteins whose structures and functions are intimately intertwined. Fine details of protein structure on the scale of 0.1 Å determine how enzymes catalyze chemical reactions, the role of mutations in diseases, and the difference between a potential drug and a failure (e.g., "orbital steering," Mesecar et al., 1997). Structure-based drug design has brought a rational approach to designing drugs to the pharmaceutical industry, which historically discovered drugs using brute force by assaying an enzyme against a library of millions of compounds, then optimizing with essentially random guesses because no knowledge of the drug-bound enzyme complex existed (Congreve et al., 2005).

The triumphs of structure-based drug design began in the 1970s with attempts to create a drug that controlled high blood pressure and heart failure by targeting the renin-angiotensin system—one early target was the angiotensin-converting enzyme (ACE). Although the structure of ACE was unavailable, the structure of the related protein carboxypeptidase A allowed construction of a hypothetical model of the ACE active site (Ondetti et al., 1977). Based on the loss of specificity for hydrophobic amino acids vs carboxypeptidase A, the authors proposed a modification to the active site that removed a hydrophobic binding pocket and replaced it with a chemical group that would interact with the terminal COOH group of a peptide. Designed ACE inhibitors based on this structural model of the active site were orally active (unlike the original inhibitor) and highly effective in animal models, indicating the viability of using structure to guide drug design (Ondetti et al., 1977).

Another significant example of rational drug design was the design of drugs to treat AIDS by targeting HIV protease. Retroviral proteases are key proteins in viral propagation because they cleave polyprotein precursors to form mature, functional viral proteins. When the structure of HIV protease was solved, it revealed a dimer with the active site shared between two monomers (Wlodawer et al., 1989). Soon afterward, a

structure was solved with a bound substrate-based inhibitor (Miller et al., 1989). Using the symmetric structure of the active site, numerous drugs were designed that were partial mimetics of the peptides that HIV protease normally cleaves, with the normally proteolyzed bond modified to be unbreakable by the protease. By 1996, three of them had been approved by the FDA: saquinavir, ritonavir, and indinavir—only 8 years after the identification of HIV protease as a target, compared to 10-15 years for a typical drug. The detailed knowledge of the active site provided by crystal structures enabled this acceleration. Many other examples of the importance of detailed structural knowledge to drug design exist, such as neuroaminidine in influenza, COX-2 inhibitors in inflammation, and the tyrosine kinase BCR-Abl in cancers (these and others are reviewed by Congreve et al., 2005), but they will not be discussed here.

Effective biomedical research thus requires that we gain a detailed picture of protein structures. But without experimentally solving the structure of every protein in existence at atomic resolution, how can we reach an accuracy where we can draw conclusions about these fine details? Improvement in modeling protein structure, both *de novo* and via homology, is necessary to realize this vision of a detailed structure for every protein. In recent years, significant advances have been made in improving modeling accuracy, with the best predictive approaches using knowledge-based minimization functions grounded in empirical data rather than *ab initio* functions based purely on theory (Rohl et al., 2004). For a knowledge-based approach to succeed, it requires accurate knowledge based on high-resolution, experimentally determined structures. The existence of atomic-resolution structures to provide an accurate starting point for homology modeling is a prerequisite to obtaining atomic-resolution models. Therefore, a clear need exists for high accuracy in both experimentally determined and predictively modeled protein structures. The NIH Protein Structure Initiative aims to obtain an accurate structure for every protein, either experimentally determined or modeled, underscoring the importance of improving the accuracy of both methods.

The accuracy of predictive modeling is limited, however, even in the best-performing programs to an RMSD of 0.5 Å–1 Å from “truth,” as shown by atomic-

resolution crystallography (Bradley et al., 2005). This discrepancy between model and truth is on the same scale as the error caused by assuming rigid, ideal geometry (Holmes and Tsai, 2004), suggesting this may be a major limiting factor of modeling accuracy. The rigid-geometry approximation uses fixed bond lengths and angles for the backbone covalent geometry. That leaves torsion angles, which describe the peptide conformation, as the only variables needed to define the structure. The influence of the rigid-geometry paradigm also appears in experimentally derived crystal structures at medium and low resolutions, where the ability to obtain a structure closer to truth would make possible new insights into function.

Using the best-determined crystallographic protein structures available in 1996, Karplus (1996) showed that systematic variations in covalent geometry do occur as a function of the conformation of the backbone torsion angles. That study was performed using structures at resolutions as low as 1.75 Å, which allowed accurate determination of torsion angles but was limited in terms of the accuracy of the definition of bond angles and bond lengths. Now, a vast number of structures are known at resolutions of 1 Å or better, allowing much more accurate determination of these systematic variations. Over the past 15 years, the number of atomic-resolution structures has exploded (Schmidt and Lamzin, 2002), with an ever-accelerating rate of new structures being solved (Figure 1.1). The increase has been steady, going from negligible numbers in 1994 to today's counts of ~100 structures/year determined at 1.2 Å resolution or better, with ~40 of those at 1.0 Å resolution or better. This explosion happened because of dual innovations in crystallography leading to the use of both synchrotron radiation and cryotemperatures. Both of these innovations have benefits and liabilities that must be understood to know which conclusions can be drawn from atomic-resolution structures.

Synchrotrons produce high-intensity X-rays that have many benefits to crystallography (Moffat and Ren, 1997). The benefit most relevant to this work is that synchrotrons allow for atomic-resolution structures to be solved, which happens for two reasons. First, synchrotron radiation is 10^3–10^9 more intense than laboratory-based X-ray sources, which allows weaker diffraction to be measured. One liability of the brilliant X-

rays available at synchrotrons is that they can cause significant radiation damage to the protein crystals, resulting in poor diffraction and specific synchrotron-induced structural artifacts (Ravelli and Garman, 2006; Garman and Owen, 2006). Second, synchrotron radiation is available at lower wavelengths which allows more convenient data collection at higher resolution simply due to geometric considerations. Typical laboratory-based X-ray home sources for protein crystallography use copper rotating anodes, which produce intense characteristic radiation at 1.54 Å. According to Bragg's law ($2d \sin \theta = \lambda$), for λ = 1.54 Å, one would need to collect diffraction data out to a scattering angle of nearly 100˚ for 1 Å resolution information, yet for λ = 1.0 Å radiation, 1 Å resolution data can be collected at a much more accessible scattering angle of $2\theta \approx 60°$. As synchrotrons can produce radiation at tunable wavelengths, down to those far below 1.0 Å, this allows for collection to resolutions as high as any protein crystal has diffracted—the protein known at the highest resolution, crambin, was analyzed at 0.54 Å resolution (Jelsch et al., 2000).

Cryocrystallography involves freezing protein crystals at liquid-nitrogen temperatures (~75K) to stabilize them so they are both more resistant to radiation damage and diffract to higher resolution (Dauter et al., 1995). At low temperature, thermal vibrations decrease, resulting in a more ordered structure. This is reflected by a drop in the crystallographic B-factors, which represent the spread in atomic positions caused by thermal motion and disorder ($B = 8\pi^2 \bar{u}^2$, where $\bar{u}$ is the mean atomic displacement, assumed to be isotropic). The two benefits of cryocrystallography are synergistic, as the higher-resolution data require increased time for data collection that would result in extensive radiation damage, but the cryotemperatures dramatically slow the rate of this damage and allow for collection of atomic-resolution data (Garman and Owen, 2006). One liability of cryocrystallography is that structures solved at cryotemperatures can have systematic deviations from the more physiologically relevant room-temperature structures. These vary from small but real differences in the core (~0.2 Å) to possibly more extensive changes on the surface (Dunlop et al., 2005). Dunlop et al. observed almost 5-fold more water molecules in the second solvation shell around the protein and 2-fold more in the first solvation shell at cryotemperatures than at room temperature. The

low temperature often causes multiple conformations of a given residue to become distinct, making modeling much more complex because two or more partially occupied positions may exist for the same atoms. The population distribution of conformations, described by the Boltzmann distribution, depends on the free-energy difference and the temperature. As the temperature increases, the distribution flattens out and widens, increasing the structural disorder at the same free-energy difference. These differences could impact biological interpretations of structures, so care should be taken when drawing conclusions based on cryostructures, particularly when multiple conformations or water molecules are involved.

The scope of this work encompasses accurate, atomic-level modeling of protein structure. Toward that end, this dissertation begins with a focused study of a single protein at atomic resolution. From there, the study expands to explore broader structural trends of backbone geometry in atomic-resolution proteins and their implications for high-accuracy modeling. Finally, modern modeling techniques are applied to create a homology model, which is used to illustrate the types of information available from these models.

An enzyme at atomic resolution

At atomic resolution (typically defined as 1.2 Å resolution or better, with half of the data in the highest resolution shells having intensities of at least 2σ; Dauter et al., 1995), many new possibilities become available. It requires different approaches, such as anisotropic refinement (Longhi et al., 1998). Most atomic-resolution structures are solved at cryotemperatures, which better defines the atomic positions via lowering the B-factors. First, the atomic displacements of atoms can be seen to be distinctly nonspherical, which violates an assumption of a single B-factor reflecting equal displacement in all directions. Consequently, a generalization of B-factors to atomic displacement parameters (ADPs) is used, which allows for the disorder to be nonspherical, or anisotropic, with three axes of

displacement. The importance of modeling the anisotropic behavior of atoms is shown by a typical drop in the crystallographic R-factor and R_{free} by 5%. Considered across multiple atoms, such anisotropy reveals experimental information about the directionality of group motion through crystallography, a feature that is commonly thought to be limited to NMR.

Second, atomic-resolution structures provide an extremely low coordinate uncertainty, on the order of ~0.02 Å, which lowers errors in derived parameters such as bond lengths and angles (Dauter et al., 1997). Consequently, some structural features are visible at this resolution that were not apparent or could only be guessed about at lower resolutions. For example, the difference in bond lengths in the asparagine side chain's carboxamide group can be visualized to directly determine the correct flip of the side chain; this also applies to glutamine and histidine. At lower resolution, the orientation of these side chains are often inferred from the predicted hydrogen-bonding network. Additionally, a similar difference in bond lengths can discriminate between protonated and unprotonated side chains involved in acid-base catalysis, potentially providing direct evidence of their protonation states. Third, at the highest resolutions, the hydrogens themselves are visible in the electron-density maps (Figure 1.2).

Chapter 2 is a case study of atomic-resolution crystallography. Using the flavoenzyme human glutathione reductase, a series of unliganded and substrate-bound structures providing snapshots of the catalytic cycle were analyzed at atomic resolution using cryostructures (Berkholz et al., 2008). The crystals diffracted to ~1 Å resolution at a synchrotron, allowing extension of previous work at ~2 Å to reveal further details about structure and catalysis. These structures illustrate many of the features common to atomic-resolution crystallographic analyses—multiple conformations existed for ~20% of the side chains, radiation damage from the bright synchrotron light source reduced a critical part of the active site, and anisotropic B-factors helped the refinement significantly and provided insight into motions relevant to catalysis.

Additionally, numerous active-site distortions in geometry were visible, and their

presence implied potentially important contributions to catalysis. One of these distortions was an active-site loop that was seen at lower resolution to have nonplanar peptide bonds systematically deviating in the same direction; an in-house Protein Geometry Database was used to assess the rarity of this motif among all protein structures. The atomic-resolution structures presented in this chapter highlight the value added to our biological knowledge by extending the resolution of known structures from medium (~2 Å) to atomic (~1 Å).

A database for mining geometric features of protein structure and a conformation-dependent library

Structures at all but the highest resolutions lack sufficient information to solve the structure based on the experimental data alone, so external knowledge must also be used. A major source of that knowledge is a geometry library that provides ideal values for many parameters including bond angles and bond lengths. The weighting given to this geometry library is varied depending on the amount of experimental data available. Because of the large amount of data at atomic resolution, the weight of the geometry library is very low. This allows the experimental data to make distortions from standard geometry apparent, which can then be modeled. These distortions are not visible at lower resolutions because there are not enough data to prove a difference from the standard value in the geometry library.

The low coordinate uncertainty and low weighting of geometry restraints at atomic resolution revealed something unexpected and frustrating. At improving resolutions, the standard geometry libraries used to restrain the bond angles and lengths matched increasingly poorly with the experimentally determined protein structures (EU 3-D Validation Network, 1998). In the thought that the ideal values must not be quite right, efforts were undertaken to find more correct single values (Jaskolski et al., 2007a). This resulted in a debate within the structural biology community. The initial paper used

the 10 highest-resolution structures to propose slight modifications to ideal values and weights, showed that medium-resolution structures were restrained too tightly to the geometry library, and suggested that the RMSD of the bond lengths from ideal values should be ~0.015–0.020 Å for structures with R-factors of 15%–20% (Jaskolski et al., 2007a). One response covered a number of additional factors, including the difficulty of using existing refinement methods on ultrahigh-resolution structures and the point that ideal values are context-dependent (Stec, 2007). A second response to the original paper refutes the suggested RMSDs by finding the optimal value of the log-likelihood function while varying the weight of the geometry library, then calculating the corresponding RMSD at that weight, which turned out to be much looser than the original recommendation and more in line with current restraints (Tickle, 2007). The original authors replied by suggesting Tickle had used elaborate numerology inconsistent with reality (Jaskolski et al., 2007b). We contributed to this debate by suggesting that the problem could be solved by noting that there is not a single perfect ideal value but instead a range of values appropriate for different structural contexts, defined by a residue's backbone conformation (Karplus et al., 2008).

In chapter 3, we have described in more detail the Protein Geometry Database we have developed to correlate conformation and covalent geometry, providing some specifics of its design, implementation, and use. The driving force behind the creation of this database was the lack of any straightforward way to ask questions about the backbone geometry of known protein structures and its relationship to backbone conformation. This database is useful for exploring the broader existence of particular motifs seen in a structure of interest, as was done in chapter 2. Additionally, the database is a useful tool for providing the data needed to deduce general features and trends in protein structure such as those presented in chapter 4.

In chapter 4, we use the Protein Geometry Database to probe a nonredundant set of atomic-resolution protein structures to empirically define the dependence of backbone bond angles and lengths upon the conformation of a residue (i.e., the backbone torsion angles). The trends seen in the conformation dependence of backbone geometry make

structural sense and should lead to a paradigm shift in how ideal geometry is thought about and applied. In this light, we describe and make available a conformation-dependent library and show how its use can improve the accuracy of crystallographic refinement and homology modeling.

A homology model for a tumor-suppressor protein

Homology modeling combines the amino-acid sequence of a target protein with the structure of a related (homologous) protein, which is used as a template to create a modeled structure of the target. Modern modeling techniques commonly allow models to reach within 1–2 Å of the experimentally determined structure; accuracy is limited primarily by how closely related the template and target structures are and how well the two sequences can be aligned, so the target sequence can be threaded onto the template structure (Zhang, 2009). At levels of sequence identity above 35–40%, the RMSD between a consensus-based automatically created model and the experimental structure rarely exceeds 2 Å (Zhang, 2009). Below 35% sequence identity, no correlation exists between sequence identity and automated model quality; the majority of RMSDs are below 5 Å, but it is not unusual for them to go as high as 15 Å (Zhang, 2009). Interestingly, one of the limiting problems is the inability to discover the best template structure using only the target sequence; a postdictive approach based on the experimentally determined target structure tended to find a better template structure than was found with the sequence alone (Zhang, 2009).

Chapter 5 describes a case study of homology modeling that leverages the same modern tools and background knowledge I used to test the implications of the work in chapter 4. In this chapter, I create a model for a tumor-suppressing protein called merlin, validate it, compare it with existing automated models, and use it to make new proposals about the determinants of merlin's function. This chapter additionally brings my crystallographic experiences from chapter 2 to bear upon a modeling problem by

providing stringent validation techniques rarely used by pure modelers. To propose functional differences between merlin and the other members of its family, which all function differently from merlin, existing and novel comparison methods are used to examine changes in the electrostatic potential as well as losses, gains, and changes of function between the two protein subfamilies.

Conclusion

Finally, in chapter 6, I present general conclusions of this work, its impact upon the field, and an outlook of future work to expand upon the insights gained by the research in this dissertation.

References

Berkholz, D.S., Faber, H.R., Savvides, S.N., and Karplus, P.A. (2008). Catalytic cycle of human glutathione reductase near 1 A resolution. J. Mol. Biol. *382*, 371-384.

Bradley, P., Misura, K.M.S., and Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. Science *309*, 1868-1871.

Congreve, M., Murray, C.W., and Blundell, T.L. (2005). Structural biology and drug discovery. Drug Discov. Today *10*, 895-907.

Dauter, Z., Lamzin, V.S., and Wilson, K.S. (1995). Proteins at atomic resolution. Curr. Opin. Struct. Biol. *5*, 784-790.

Dunlop, K.V., Irvin, R.T., and Hazes, B. (2005). Pros and cons of cryocrystallography: should we also collect a room-temperature data set? Acta Crystallogr. D Biol. Crystallogr. *61*, 80-87.

Garman, E.F., and Owen, R.L. (2006). Cryocooling and radiation damage in macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. *62*, 32-47.

Holmes, J.B., and Tsai, J. (2004). Some fundamental aspects of building protein structures from fragment libraries. Protein Sci. *13*, 1636–1650.

Jaskolski, M., Gilski, M., Dauter, Z., and Wlodawer, A. (2007a). Numerology versus reality: a voice in a recent dispute. Acta Crystallogr. D Biol. Crystallogr. *63*, 1282-1283.

Jaskolski, M., Gilski, M., Dauter, Z., and Wlodawer, A. (2007b). Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? Acta Crystallogr. D Biol. Crystallogr. *63*, 611-620.

Jelsch, C., Teeter, M.M., Lamzin, V., Pichon-Pesme, V., Blessing, R.H., and Lecomte, C. (2000). Accurate protein crystallography at ultra-high resolution: valence electron distribution in crambin. Proc. Natl. Acad. Sci. USA *28*, 3171-3176.

Karplus, P.A. (1996). Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci. *5*, 1406-1420.

Karplus, P.A., Shapovalov, M.V., Dunbrack, R.L., and Berkholz, D.S. (2008). A forward-looking suggestion for resolving the stereochemical restraints debate: ideal geometry functions. Acta Crystallogr. D Biol. Crystallogr. *64*, 335-336.

Mesecar, A.D., Stoddard, B.L., and Koshland, D.E. (1997). Orbital steering in the catalytic power of enzymes: small structural changes with large catalytic consequences.

Science *277*, 202-206.

Miller, M., Schneider, J., Sathyanarayana, B.K., Toth, M.V., Marshall, G.R., Clawson, L., Selk, L., Kent, S.B., and Wlodawer, A. (1989). Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution. Science *246*, 1149-1152.

Moffat, K., and Ren, Z. (1997). Synchrotron radiation applications to macromolecular crystallography. Curr. Opin. Struct. Biol. *7*, 689-696.

Ondetti, M.A., Rubin, B., and Cushman, D.W. (1977). Design of specific inhibitors of angiotensin-converting enzyme: new class of orally active antihypertensive agents. Science *196*, 441-444.

Ravelli, R.B.G., and Garman, E.F. (2006). Radiation damage in macromolecular cryocrystallography. Curr. Opin. Struct. Biol. *16*, 624-629.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., and Baker, D. (2004). Protein structure prediction using Rosetta. Methods Enzymol. *383*, 66-93.

Schmidt, A., and Lamzin, V.S. (2002). Veni, vidi, vici - atomic resolution unravelling the mysteries of protein function. Curr. Opin. Struct. Biol. *12*, 698-703.

Stec, B. (2007). Comment on Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? by Jaskolski, Gilski, Dauter & Wlodawer (2007). Acta Crystallogr. D Biol. Crystallogr. *63*, 1113-1114.

Tickle, I.J. (2007). Experimental determination of optimal root-mean-square deviations of macromolecular bond lengths and angles from their restrained ideal values. Acta Crystallogr. D Biol. Crystallogr. *63*, 1274-1281; author reply 1282-1283.

Wilson, K.S., Dauter, Z., Lamsin, V.S., Walsh, M., Wodak, S., Richelle, J., Pontius, J., Vaguine, A., Sander, R.W.W., and Hooft, V.G. (1998). Who checks the checkers? Four validation tools applied to eight atomic resolution structures. J. Mol. Biol. *276*, 417-36.

Wlodawer, A., Miller, M., Jaskólski, M., Sathyanarayana, B.K., Baldwin, E., Weber, I.T., Selk, L.M., Clawson, L., et al. (1989). Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. Science *245*, 616-621.

Zhang, Y. (2009). Protein structure prediction: when is it useful? Curr. Opin. Struct. Biol. *19*, 145-155.

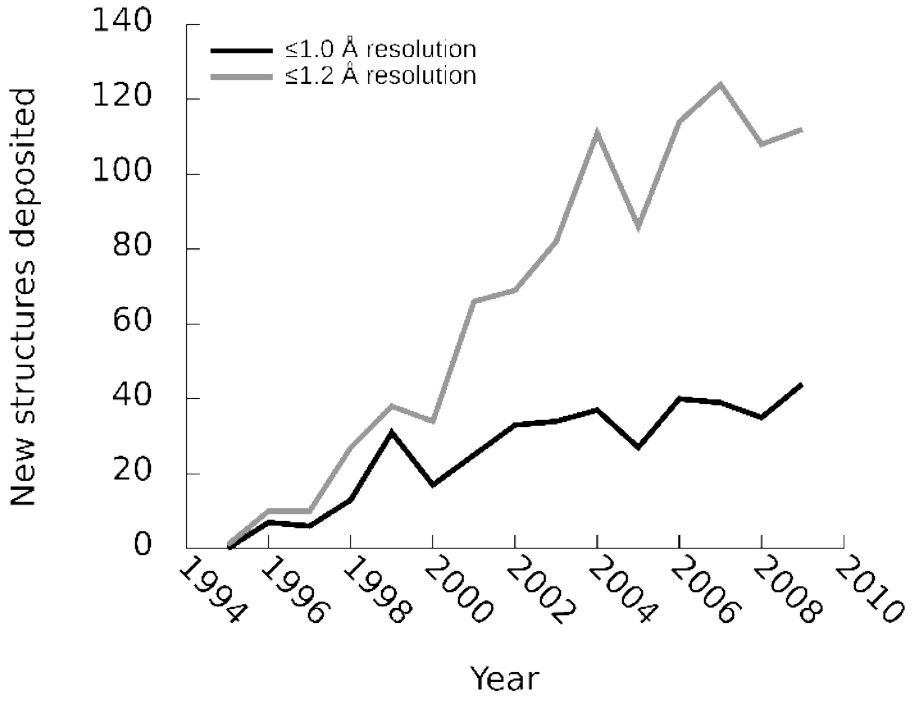


Figure 1.1. Growth in the number of atomic-resolution structures deposited in the Protein Data Bank. The lines indicated structures publicly released during each 12-month period, with the 1995 point indicating all structures from 1995 and earlier. The 2009 depositions are extrapolated by doubling the released PDBs between 1 January and 31 June. Lines are as indicated in the key.

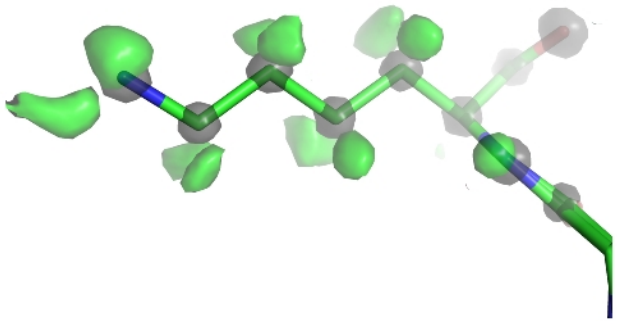


Figure 1.2. Atomic-resolution electron density can make hydrogen atoms visible. Electron density is shown from Lys203 in the 0.73 Å resolution structure of the PDZ domain of syntenin, which has the best R-factor and R_{free} of any known protein (7.5% and 8.7%, respectively). Gray semitransparent surfaces are 2Fo-Fc density at $5.0\rho_{rms}$, which indicates the positions of all modeled atoms. Green semitransparent surfaces are Fo-Fc density at 2.5σ, which indicates the positions of hydrogens.

Chapter 2

Catalytic cycle of human glutathione reductase near 1 Å resolution

Donald S. Berkholz, H. Richard Faber, Savvas N. Savvides, and P. Andrew Karplus

Published in the *Journal of Molecular Biology*, **382**(2):371 (2008)

Summary

Efficient enzyme catalysis depends on exquisite details of structure beyond those resolvable in typical medium- and high-resolution crystallographic analyses. Here we report synchrotron-based cryocrystallographic studies of natural substrate complexes of the flavoenzyme human glutathione reductase (GR) at nominal resolutions between 1.1 and 0.95 Å that reveal new aspects of its mechanism. Compression in the active site causes overlapping van der Waals radii and distortion in the nicotinamide ring of the NADPH substrate, which enhances catalysis via stereoelectronic effects. The bound NADPH and redox-active disulfide are positioned optimally on opposite sides of the flavin for a 1,2-addition across a flavin double bond. The new structures extend earlier observations to reveal that the redox-active disulfide loop in GR is an extreme case of sequential peptide bonds systematically deviating from planarity, a net deviation of 53º across 5 residues. But this apparent strain is not a factor in catalysis as it is present in both oxidized and reduced structures. Intriguingly, the flavin bond lengths in oxidized GR are intermediate between those expected for oxidized and reduced flavin, but we present evidence that this may not be due to the protein environment but instead to partial synchrotron reduction of the flavin by the synchrotron beam. Finally, of more general relevance, we present evidence that the structures of synchrotron-reduced disulfide bonds cannot generally be used as reliable models for naturally reduced disulfide bonds.

Introduction

The short lifetimes of reaction intermediates makes difficult any detailed structural study of steps involved in enzyme catalysis. The use of transition-state analogs and other inhibitors has allowed for much useful insight into enzyme mechanism, but because the inhibitors are not true substrates, the results may not always apply in detail. Only in a few cases has it been possible to examine authentic reaction intermediates in sufficient detail

to visualize fine details of catalysis at atomic resolution, commonly defined as 1.2 Å resolution or better (Dauter et al., 1997). These cases include horseradish peroxidase (Berglund et al., 2002), which used time-resolved crystallography, and D-2-deoxyribose-5-phosphate aldolase (Heine et al., 2001).

Understanding flavoenzymes—enzymes using FAD or FMN in catalysis—is particularly complex because of flavin's involvement in a wide diversity of chemical reactions and its many possible redox and protonation states, each of which has unique properties and can be stabilized or destabilized by the protein environment (De Colibus and Mattevi, 2006). Proteins apparently modulate flavin reactivity via a variety of mechanisms, including bending the flavin away from planarity and varying the degree of stabilization of negative charges at the flavin N1/O2α locus and at other loci (Miura, 2001; Lennon et al., 1999; Fraaije and Mattevi, 2000; Massey, 1995; Massey, 2000; Fox and Karplus, 1999). Nevertheless, despite many investigations of flavoenzymes in their native state and after reconstitution with modified flavins (De Colibus and Mattevi, 2006; Miura, 2001; Lennon et al., 1999; Fraaije and Mattevi, 2000; Massey, 1995; Massey, 2000; Fox and Karplus, 1999; Ghisla and Massey, 1989; Karplus, 1999), much remains poorly understood about how proteins modulate flavin reactivity.

The flavoenzyme glutathione reductase (GR) is a dimeric disulfide oxidoreductase that converts oxidized glutathione (GSSG) to two molecules of reduced glutathione (GSH) using an NADPH cofactor and an FAD prosthetic group. Glutathione plays a critical role in maintaining the cell's reducing environment and battling oxidative stress. Human erythrocyte GR is a homodimer of 52 kD monomers, each with three domains: an NADPH-binding domain, an FAD-binding domain, and a dimerization domain (Karplus and Schulz, 1987). The NADPH- and FAD-binding domains meet at the active site, in which both monomers participate.

Much of our understanding of how GR works comes from combining steady-state and presteady-state analyses of catalysis by GR and related disulfide reductases (Huber and Brandt, 1980; Thorpe and Williams, 1976; Krauth-Siegel et al., 1998; Argyrou et al.,

2002; Bohme et al., 2000; Vanoni et al., 1990; Rietveld et al., 1994) with a series of 2 Å resolution structures of GR in various redox states and bound to combinations of natural substrates: GR_{Native}, $GR_{GSSG/NADP}$, GR_{GSH} and GR_{NADPH} (Karplus and Schulz, 1989). Figure 2.1 illustrates the consensus mechanism derived from these studies. First, NADPH binds and transiently reduces the flavin. Tyr197, which swings out of the way so NADPH can bind, is proposed to act as a spring in forcing the nicotinamide into the flavin. The reduced flavin then reduces the Cys58-Cys63 disulfide bond by forming a short-lived covalent intermediate with Cys63, followed by formation of a stable charge-transfer complex between the flavin and the Cys63 thiolate. Calculations have shown that the pairwise overlap of molecular orbitals in the X-ray structure is optimal for hydride transfer between the nicotinamide and the flavin and also for covalent catalysis of electron transfer between the flavin and Cys63 (Sustmann et al., 1989). After formation of the charge-transfer complex, $NADP^+$ dissociates and is replaced by another NADPH. This constitutes the reductive half-reaction leading to the enzyme form known as EH_2. The oxidative half-reaction begins with the binding of GSSG. Cys58 in GR, which is activated similarly to serine or cysteine proteases by the His467′-Glu472′ pair (primes denote residues coming from the second subunit of the dimer), attacks Cys_I of GSSG to form a mixed disulfide between GS_I and Cys58. After the freed GSH_{II} leaves, Cys58 and Cys63 re-form a disulfide releasing the second molecule of GSH product.

Despite the extensive enzymatic and structural studies of GR catalysis, uncertainties remain that more detailed pictures of catalysis might resolve. For example, in the native structure determined at 1.54 Å resolution (Karplus and Schulz, 1987), marginally reliable deviations in peptide planarity of the active-site disulfide loop were suggested to indicate strain that would favor disulfide reduction. Similarly, very small deviations in covalent flavin geometry were suggested as possible evidence as to how the protein modulates flavin reactivity. But both of these observations were near the limits of coordinate accuracy and need confirmation. Also, the more accurate determination of the geometry of the nicotinamide-flavin approach provided by atomic-resolution analysis will help us to better understand the hydride-transfer step. Finally, atomic-resolution

analysis could yield additional insight by providing a direct visualization of the protonation states of the active site Cys and His residues at various stages of catalysis.

The advent of synchrotron sources and cryocrystallography has induced an explosion in atomic-resolution structures: over 220 structures with >45 residues determined at resolutions of ≤1 Å now exist in the Protein Data Bank (Berman et al., 2000) as compared to only five in 1996. Of these proteins, only two are flavoenzymes: cholesterol oxidase (Lario et al., 2003) and pentaerythritol tetranitrate reductase (Khan et al., 2004). Here, we use the same GR crystal form previously studied (Karplus and Schulz, 1989) to extend the structural analysis of the GR catalytic cycle to near 1 Å resolution.

Results and Discussion

Structure determination

The four complexes of GR_{Native}, $GR_{GSSG/NADP}$, GR_{GSH}, and GR_{NADPH} previously studied at 2 Å resolution at room temperature were structurally analyzed at cryotemperatures and refined with SHELXL to R-factors near 12% and R_{free} near 15% at nominal resolutions of 0.95, 1.1, 1.0 and 1.0 Å, respectively (Table 2.2). For each model, riding hydrogen atoms were included and individual anisotropic B-factors were refined. The inclusion of each of these led to drops in R_{free} of >2%, indicating that their inclusions were justified (see Methods). In addition, multiple conformations were modeled for about 20% of the side chains in all of the atomic-resolution structures and for a few stretches of the backbone (see below). For the highest-resolution, GR_{Native} structure, leaving out the hydrogens resulted in >2σ difference peaks for only about 30% of the peptide backbone NH atoms, so we conclude that this structure does not have sufficient information content to provide reliable evidence for the presence or absence of specific hydrogen atoms. Based on Cruickshank's DPI error estimator (Cruickshank, 2001), the coordinate error for atoms

with average B factors is ~0.02 Å. Consistent with this level of coordinate error and the nominal resolution of the analysis, atoms are clearly discernible as discrete peaks in the final electron density map (Figure 2.2).

In addition, a 1.8 Å, room-temperature GR_{NADPH} structure based on data collected using a laboratory X-ray source was solved and refined to R and R_{free} near 14% and 19%, respectively (Table 2.2). This structure provides a direct image of NADPH binding at room temperature that replaces the previous best-resolved model of NADPH binding, which was derived by combining information from structures with a bound NADH or a bound $NADP^+$ (Karplus and Schulz, 1987).

Because insights into catalysis are often based on small differences between the structures, we note that our refinement strategy involved fully refining the highest-resolution GR_{Native} structure first and then using that as a starting model for generating the three other structures. This means that differences between each individual model and GR_{Native} will tend to be underestimated, enhancing the confidence that can be placed in any significant structural differences observed.

The ultrahigh-resolution structures gave us no new insights into the binding of GSSG or GSH or the transfer of electrons from the redox-active disulfide to GSSG. In particular, we were unable to determine the protonation state of His467′ important for the oxidative half-reaction. Thus, we do not use space here to describe those aspects of the structures. Instead, the focus in this presentation is on novel structural results and catalytic insights related to the reductive half-reaction involving NADPH, FAD and the redox-active disulfide. In the following sections, we will describe the structural results, followed by insights relevant to catalysis.

Temperature-dependent changes in structure

Each of these new structures has been previously determined at lower resolution at room

temperature, and comparisons were carried out to assess any consistent changes that appeared to be due to a change in temperature. As expected, the temperature factors in the cryo-structures were consistently lower, typically about 75% as large. In only a few segments does the cryo-structure have significantly higher temperature factors, and these are all surface loops involved in crystal contacts that shift somewhat and apparently become less ordered during the unit-cell changes that occur upon freezing. Interestingly, the most systematic exception to the general drop in B-factors is the set of residues that are the most ordered in the room temperature structures. For these residues with B~8 $Å^2$, the B-factors stay largely the same, implying that these B-factors may be an indicator of lattice disorder in the crystal rather than intrinsic thermal motion of the atoms themselves.

Also as expected, the cryo-structures had many more ordered water molecules, with the GR_{Native} structure going from 523 modeled water sites at room temperature to 832 at low temperature, including many partially occupied water sites involved in definable alternate hydrogen-bonding networks. With regard to alternate conformations of protein atoms, the 1.54 Å resolution room temperature structure of GR_{Native}, showed evidence for alternate conformations of 12 side chains (see Table 8 of Karplus & Schulz 1987). In the 0.95 Å resolution cryo-stucture of the same crystal form, the discrete disorder of 8 of these residues is confirmed, but for 4 residues—Thr119, Ser190, Ser231 and Lys420—it is not. Discrete disorder is also modeled for an additional 69 residues. As these residues mostly had relatively high B-factors in the room-temperature structure, the observation of discrete disorder could be simply a resolution effect rather than a temperature effect. The one exception is the redox-active disulfide, which in the cryo-structure is modeled in both an open (reduced) conformation and a closed (oxidized) conformation. This is a result of radiation-induced opening of the disulfide rather than a temperature effect and is discussed further below. As no examples were found of atoms that were well-defined in both structures but with distinctly different conformations, we conclude that there are no notable conformational changes due to freezing itself.

Overall anisotropic motions

The large data-to-parameter ratio at atomic resolution allows for consideration of nonspherical anisotropic displacement parameters (ADPs) rather than a single isotropic B factor. Anisotropy can be quantified by a single number that is the ratio of the smallest to the largest elements of the 3 x 3 ADP matrix (Trueblood et al., 1996). This measure ranges from 1 (perfectly isotropic) decreasing toward zero with increasing anisotropy. Atomic anisotropies in GR_{Native} were roughly normally distributed with a mean of 0.32 and σ=0.1, and the three other GR structures showed higher means of ~0.43 and σ=0.1. In all distributions, very few atoms had anisotropies >0.8. This indicates that although high-resolution data are needed to reveal anisotropy, its presence is the rule and isotropically vibrating atoms are the rare exceptions. Separate distributions for just protein atoms, just heteroatoms or just solvent atoms are similar (means within 0.03 of the mean for all atoms). Except for GR_{Native}, these results are consistent with Merritt's analysis of all structures known at 1.4 Å or better (mean 0.45 and σ=0.15; Merritt, 1999). The degree of anisotropy in GR_{Native} is slightly higher than any protein in Merritt's study; the closest was lysozyme (PDB ID 1lks) with a mean of 0.35. Described later are details of anisotropic motions in the active site and related to catalysis.

Synchrotron reduction fails to model natural reduction

Synchrotron radiation has been seen not only to cause a generic gradual decay in the diffraction strength of protein crystals but also to cause specific structural changes such as the cleavage of disulfide bridges (Burmeister, 2000; Weik and Sussman, 2000; Ravelli and Garman, 2006; Ravelli and McSweeney, 2000) and reduction of active site metallocenters (Carugo and Carugo, 2005). These changes are thought to be caused by X-ray generated solvated electrons (Carugo and Carugo, 2005). Many groups, including ours, have taken advantage of this "radiation-induced reduction," assuming it provided a view of catalytically relevant reduced enzyme forms that allowed insights into enzyme

mechanism (Berglund et al., 2002; Carugo and Carugo, 2005; Alphey et al., 2003; Roberts et al., 2005; Kort et al., 2004).

In both oxidized GR crystal forms analyzed here, GR_{Native} and $GR_{GSSG/NADP}$, synchrotron radiation partially cleaved the active-site (Cys58-Cys63) disulfide. Refinement gave occupancies of 0.53/0.47 (GR_{Native}) and 0.38/0.62 ($GR_{GSSG/NADP}$) for disulfide/open forms of the two structures. Interestingly, when these synchrotron-reduced structures are compared with the structures of GR_{GSH} and GR_{NADPH} that have been chemically reduced at room temperature, the positions of Cys58 differ significantly (Figure 2.3). The synchrotron-based reduction merely involved a χ_1 sidechain rotation such that the Cys58 sulfur moved 1.0-1.3 Å away from Cys63; in contrast, the chemically reduced structures show an accompanying shift of up to 0.75 Å by the backbone atoms of Cys58 and nearby residues. Our explanation is that at the cryotemperatures of data collection, any larger-scale motions involving the protein backbone are blocked. We note that a recent published structure of the thioredoxin-like protein cDsbD (Stirnimann et al., 2006) showed a similar discrepancy between synchrotron reduction and chemical reduction, although the authors did not point out this difference. Also supporting the hypothesis that motions are limited at the cryotemperatures of data collection is that at temperatures near 200 K, the motions required for enzyme catalysis are hindered (Rasmussen et al., 1992). In any case, independent of the explanation, the discrepancy between the synchrotron-cleaved and chemically reduced conformations seen for GR and cDsbD proves that one cannot assume a synchrotron-generated reduced form of an enzyme accurately reflects active-site changes that occur during normal catalysis. Consequently, any insights into reaction mechanisms based on radiation-reduced structures at cryotemperatures must be re-examined.

The redox-active disulfide loop has large ω-angle deviations

One of the goals of these atomic resolution refinements was to better assess the

preliminary observation of conformational strain in the redox-active disulfide loop seen in native GR at 1.54 Å resolution, with the six peptide bonds from residue 58 to 64 all having negative ω-values with an average value of –175˚ (Karplus and Schulz, 1987) even though the maximal deviations from planarity were only about 10˚. It is now well-documented that ω-deviations are strongly underestimated in lower-resolution structures but can be determined with about a 3˚ accuracy in atomic-resolution structures (Sevcik et al., 1996). Consistent with this, all four of the atomic-resolution GR structures show ω-values that deviate from planarity by up to ~20˚, and the rms deviations in the four structures are all less than 3˚. Considering the redox-active disulfide loop (Figure 2.4a), the six peptides from Cys58 to Cys63 all deviate from planarity in the same direction with an average nonplanarity of ~10˚ (Figure 2.4b). The systematic non-planarity of near 50˚ for a 5-residue segment (Figure 2.4c) is nearly twice as large as seen for any other region in GR.

To explore how unusual this was among all proteins, we surveyed atomic resolution structures in the Protein Data Bank (see Methods; Berman et al., 2000). 516 five-residue segments in 142 proteins were selected by this search. Among 48,428 residues in 249 proteins, the ~45–55° net deviation from planarity in GR is an extreme case matched by only two other structures (Figure 2.4d). One of those two is the flavoenzyme cholesterol oxidase (PDB ID 1n4w), where the region of deviation occurs in a poorly conserved loop distant from the active site. In the other structure, deoxyribose-phosphate aldolase (PDB ID 1p1x), the deviation occurs in a well-conserved loop (residues 169–174) implicated in binding the phosphate group of the substrate (Heine et al., 2004).

Karplus & Schulz (1987, 1989) hypothesized that the systematic deviation in peptide planarity in GR would stabilize the reduced form of the enzyme if upon reduction, the loop opening allowed a more relaxed conformation to be adopted. Also, the presence of conformational strain could enhance the kinetics of reduction because the reduction rate of disulfide bonds is exponentially dependent upon the force applied to

those bonds (Wite et al., 2006). However, these hypotheses are not supported because the GR structures with an open disulfide loop (GR_{GSH} and GR_{NADPH}) harbor the same level of peptide nonplanarity as oxidized GR (Figure 2.4c). Since the nonplanarity remains in the open loop, we hypothesize it is related to the local conformation, as this has been shown to have a systematic influence on peptide planarity (Karplus et al., 2008; Karplus, 1996). Another unfavorable aspect of the disulfide loop conformation is that the hydrogen-bonding potential of some of its backbone atoms is rather poorly satisfied (Figure 2.4a). In two cases, for a peptide NH group, the closest potential hydrogen-bonding partners are quite distant, 4.5–5.0 Å away (red dotted lines in Figure 2.4a). As was seen for the peptide nonplanarity, the quality of hydrogen bonding does not improve upon disulfide reduction.

One additional observation is that the Cys58-Cys63 disulfide bond is unusually long in both atomic-resolution oxidized structures: 2.22 Å in GR_{Native} and 2.32 Å in $GR_{GSSG/NADP}$, compared with the standard value of 2.04 Å. We suspect this is an artifact of the radiation-induced partial disulfide reduction. In GR_{GSH}, Cys63-SG moves 0.15 Å compared with its position in GR_{Native}, suggesting that the single position modeled for Cys63-SG in GR_{Native} ($GR_{GSSG/NADP}$) is actually an average of two conformations ~0.30 Å apart. In this case, the true oxidized position need not have an unusual bond length.

NADPH binding and catalysis of hydride transfer

Karplus & Schulz (1989) derived the NADPH binding mode at ca. 2 Å resolution by making a composite of the NADH and $NADP^+$ bound forms. Here, the direct analysis of the NADPH complexes at room temperature and at cryotemperatures not only confirms the general features of the composite model, but the ultrahigh-resolution cryo-structure also gives novel information about the detailed interactions at the catalytic center that provide insight into the roles played in hydride-transfer catalysis by compression and stereoelectronic effects. It is well-known that steric compression can in principle

contribute to enzyme catalysis (Rajagopalan and Benkovic, 2002; Bruice and Lightstone, 1999; Almarsson and Bruice, 1993; Bruice and Pandit, 1960), but its involvement in particular enzymes is difficult to document without ultrahigh-resolution structural analyses of catalytically relevant complexes. At 1.8 Å resolution, the X-ray structure of the NADPH complex of GR shows the C4 atom to be only 3.3 Å from the flavin N5 atom, closer than normal van der Waals interactions predict. Now at atomic resolution, we confirm this close approach and see additional evidence of compression that would facilitate hydride transfer.

The first evidence of compression is a strong decrease in the level of motion of the active-site atoms in the NADPH complex. In the empty oxidized active site, the anisotropic thermal ellipsoids show that the flavin has significant freedom to shift perpendicular to the plane of the flavin, and Tyr197, the residue filling the nicotinamide pocket, also has freedom to anisotropically wag around its average position (Figure 2.5a). In contrast, when the nicotinamide displaces Tyr197 upon NADPH binding, both the absolute mobility (defined by the B-factor) and the anisotropy diminish markedly, indicating that the four groups—Tyr197, nicotinamide, flavin and the Cys63 thiolate—are tightly juxtaposed against one another (Figure 2.5b). The existence of compression is further supported by overlapping van der Waals radii in the active site (Figures 2.6 & 2.8). With a flavin N5 to nicotinamide C4 distance of 3.29 Å and a flavin C4a to Cys63-SG distance of 3.29 Å, the flavin is tightly fixed through compression from both sides. A third line of evidence for compression in the NADPH-bound complex is the shifting of the flavin ring system in the two reduced structures (Figure 2.6). In the GSH-reduced structure, which has the Cys-63 thiolate-flavin charge-transfer interaction but no bound nicotinamide, the flavin N5 is pushed toward the nicotinamide pocket by about 0.3 Å even though the Cys63 thiolate only moves half that distance. This specifies the relaxed distance of approach for the thiolate-flavin charge-transfer interaction. In contrast, in the NADPH-reduced structure, the flavin N5 is actually pushed by the presence of the nicotinamide 0.3 Å the other direction (toward the still-unmoved thiolate). These changes result in a 0.15 Å compression in the flavin-thiolate interaction (flavin C4a and Cys63S).

This provides a structural explanation for how NADPH binding intensifies the thiolate-flavin charge-transfer intensity (Bohme et al., 2000). Assuming that the level of compression is evenly spread between the players, this implies that the nicotinamide/flavin interaction is similarly compressed. Finally, the fourth line of evidence for compression is a clearly visible distortion in the planarity of the nicotinamide group at atom N1 (Figures 2.2a & 2.7). Based on inspection of the structure, it appears that the pyramidalization of N1 occurs because if the nicotinamide ring were not puckered, it would extend in the direction of the ribose-N1 bond and the nicotinamide C4 atom would collide with the flavin N5 atom (Figures 2.6–2.8). This creates a loaded-spring effect, with the nicotinamide not only tightly packed but presumably strongly forced toward the flavin, aided by the backstop of Tyr197 (Figure 2.5b).

Interestingly, according to stereoelectronic theory, N1 pyramidalization can serve to optimize the nicotinamide for hydride transfer (Nambiar et al., 1983), so that the observed distortion is likely to be a mechanism to chemically enhance the rate of hydride transfer. Normally, in the planar nicotinamide, the N1 lone-pair electrons form a conjugated system with the C=C bonds of the rings stabilizing NADPH (Young and Post, 1996; Wu and Houk, 1991; Wu and Houk, 1993; Benner, 1982). In GR, however, the out-of-plane N1-C1´ bond (Figure 2.7a) pulls the N1 lone-pair electrons out of the resonance and moves them into a pseudoaxial orbital on the hindered flavin side. Importantly, this favors hydride transfer when the C4 atom fluctuates out of the plane, creating a boat-like conformation with the hydride to be transferred pseudoaxial and on the same side of the nicotinamide ring as the N1 lone pair (Almarsson and Bruice, 1993; Young and Post, 1996; Benner, 1982). This type of reaction, a 1,4-*syn* elimination, has been well-studied in the chemistry literature (Yates et al., 1975). Furthermore, the N1 pyramidalization with the lone pair facing the flavin entropically restricts by half the conformations available to nicotinamide (N1 can only pucker in one direction). This makes the required boat-like conformation more likely and thus enhances the propensity for hydride transfer.

A second factor improving catalysis is the conformation of the ribose moiety relative to the nicotinamide, which has been shown by Wu and Houk (1991) to influence

the NADPH redox potential. In GR, the ribose C-O bond is parallel to the nicotinamide ring, which stabilizes $NADP^+$ by allowing hyperconjugative donation by the ribose σ_{C-H} and σ_{C-C} orbitals into the electron-deficient ring of $NADP^+$ (Figure 2.7b; Wu and Houk, 1991). Additionally, this parallel conformation minimizes NADPH stabilization via the anomeric effect—the ribose C-O is a better acceptor than C-H, so if the C-O were anti to the N1 lone pair, it would stabilize it with the σ^*_{C-O} orbital (Wu and Houk, 1991). These two factors together serve to increase the NADPH reduction potential, favoring hydride transfer.

A third factor favoring catalysis proposed in the lower (1.54 Å) resolution analysis was that deviations in key flavin bond lengths indicated the protein environment predisposed the flavin toward reduction. The data in Table 2.1 indicate that changes in the redox states of small-molecule flavins are associated with changes in four bond lengths by >0.05 Å (Karplus, 1999). None of the four structures of the enzyme are expected to have formally reduced flavins. However, at atomic resolution, the two EH_2 structures have all bond lengths close to those of reduced flavin, and the two oxidized state structures have some bonds that are reduced-like and others that are intermediate between oxidized and reduced states (Table 2.1). The strong tendency toward reduced-like bond lengths in all structures led us to ask whether they were influenced by synchrotron radiation. Synchrotron-induced reduction of FAD has been seen in DNA photolyase (Kort et al., 2004), but in that case the flavin is naturally light-sensitive. An assessment of the other two flavoenzymes with structures known at ≤1 Å resolution, cholesterol oxidase and PETN reductase, showed that both appear reduced based on their bond lengths (Table 2.1). The PETN reductase structure was of a reduced form of the enzyme, but cholesterol oxidase was not, so the reduced bond lengths seen are unexpected (Lario et al., 2003). Given these results, we are suspicious that X-ray reduction of these flavins has occurred to at least some extent and hesitate to draw conclusions about how the protein environment in GR influences flavin electronic structure. We further conclude that although ultrahigh-resolution structure determination has the potential to yield accurate bond-length information that can give insight into detailed influences of the protein

environment on the flavin, because synchrotron radiation itself can influence the flavin structure, conclusions must be tempered unless the changes due to the protein environment and those due to perturbation by the experiment can be dissected.

Finally, a fourth factor favoring catalysis is that the nicotinamide C4 hydride and Cys63-SG are on opposite sides of the flavin plane and roughly in line with the N5-C4a bond (Figure 2.8), in position for good HOMO-LUMO overlap (Sussman et al., 1989; Cavelier and Amzel, 2001; Rivas et al., 2004) and also consistent with optimal geometry for 1,2-addition of the hydride and the sulfur across a double bond (Wu and Houk, 1991; Liotta et al., 1984; Miessler and Tarr, 2004; Matthews et al., 1979; Miller et al., 1990). This optimal geometry for a 1,2-addition raises the interesting possibility that the reduction of GR is not a distinct two-step process with hydride transfer to the flavin occurring first, followed by disulfide reduction, but instead is concerted, with hydride transfer linked to and enhanced by disulfide reduction. If this is the case, stopped-flow studies would not show presence of a reduced flavin. Indeed, Huber and Brandt (1980) wrote of yeast GR: “It is possible that electron transfer is relatively concerted so that discrete intermediates are not detected, as suggested by Matthews et al. (1979) for lipoamide dehydrogenase.” Furthermore, in wild-type *E. coli* GR, reoxidation of FAD obscured direct observation of FAD reduction (Rietveld et al., 1994), with both steps at rates greater than 100 s^{-1}. However, mutants can disrupt this tight linkage. Reduced flavin intermediates were observed in active-site mutants of *E. coli* GR by Rietveld et al. (1994) and mercuric reductase by Miller et al. (1990). In another *E. coli* GR mutant of the Cys63 equivalent to alanine to block the reaction at flavin reduction, only 10% of the flavin was reduced, indicating a possible influence of the disulfide acceptor on the thermodynamics of hydride transfer (Rietveld et al., 1994). These points suggest a reductive half-reaction in disulfide reductases that, if not concerted, is at least tightly linked to hydride transfer both kinetically and thermodynamically.

Materials and Methods

Protein expression, purification, and crystallization

Recombinant human GR was provided by R.H. Schirmer, prepared in *Escherichia coli* as previously described (Krauth-Siegel et al., 1998). Crystals were grown as previously described (Savvides and Karplus, 1996). Briefly, they were grown reproducibly at room temperature using 10 μL hanging drops containing 20 mg/mL GR, 3% ammonium sulfate, 0.1 M potassium phosphate and 0.1% β-octyl glucoside at pH 7.0. The reservoir was 700 μL containing 21-23% ammonium sulfate and 0.1 M potassium phosphate at pH 8. Crystals used at cryotemperatures were frozen by soaking them for 1 minute in each of 5%, 10%, 15%, 25% glycerol solutions of the mother liquor before cryocooling in liquid nitrogen.

Crystals of GR grew to 1.0 x 0.7 x 0.4 mm^3, but the larger crystals tended to lose diffraction quality under cryogenic conditions. The best data were measured from crystals measuring 0.5 x 0.3 x 0.2 mm^3. To prepare complexes of GR with natural substrates, crystals of oxidized GR were soaked in mother liquor supplemented with 10 mM GSSG / 5 mM $NADP^+$ (24 hours for $GR_{GSSG/NADP}$), 1-5 mM GSH (24 hours for GR_{GSH}), or 1-5 mM NADPH (1-2 hours for GR_{NADPH}). Reduction of crystalline GR was monitored by the ensuing color change from golden yellow to orange-red. Reduced crystals of GR were immediately cryocooled in liquid nitrogen and stored for data collection. Cryocooling was performed according to the method developed for native crystals and with the cryosolutions supplemented with the appropriate substrates to maximize ligand occupancies.

X-ray data collection and refinement for atomic-resolution structures

Crystals of GR belong to the space group C2. Diffraction data were collected at the

Advanced Photon Source on BioCARS beamline 14-BM-C using an ADSC Quantum 4 detector at 100 K with λ=1.0 Å radiation. For each structure, data sets were collected from a single crystal. Using a detector offset in 2θ, data from several oscillation runs were collected to cover reciprocal space at high resolution, then data from a low-resolution pass was also collected with no detector offset and with a shortened exposure time to minimize the number of overloaded reflections. Data were processed and scaled using Denzo and Scalepack (Otwinowski and Minor, 1997). R_{meas} and R_{merge} were calculated from unmerged data and R_{meas} >50% and I/σ <2.0 were used in defining resolution limit cutoffs (Figure 2.9). In some cases the highest resolution bin completeness is low (near 50%), but in all cases the completeness climbs to be above 75% within 0.1 Å of the reported resolution (Table 2.2). Finally, 5% of the native data were selected to set aside as a test set for R_{free} calculations. R_{free} test sets for all of the complexes were based on the native test set.

Refinement of GR_{Native} began with PDB model 3grs transformed into the C2 unit cell and all water molecules removed. This 1.54 Å room temperature model was used a starting point (R=0.347, R_{free}=0.352) for 40 cycles of conjugant gradient refinement at 1.5 Å resolution with the TNT package (Tronrud, 1997). A quick visual inspection/correction of obvious changes and addition of 192 waters followed by an additional 40 cycles of refinement yielded a model with excellent geometry and R=0.253 R_{free}=0.305 . All subsequent refinement was performed with SHELX (Sheldrick and Schneider, 1997).

SHELX refinement for GR_{Native} followed the scheme described in the SHELX manual for high-resolution structures (Sheldrick and Schneider, 1997). The resolution limit was extended to 1.0 Å in SHELX with a diffuse solvent correction (SWAT). Following this, individual atoms were refined anisotropically using the suggested SIMU/DELU restraints. The default value for DELU standard deviation was used, and the standard deviation of SIMU was moved up to 0.1 from 0.02. The ISOR restraint was only applied to solvent atoms and was increased to a standard deviation of 0.1 from 0.02. As refinement progressed, alternate conformations of sidechains and loops were modeled as indicated by the electron density maps. Water molecules were added as manually

identified throughout refinement; in later rounds, this was supplemented with automated methods of SHELX. These suggested water molecules were manually checked for reasonable geometry before they were added to the model. In the final steps of refinement, some water molecules were reduced to half occupancy, riding hydrogens were added, and the resolution was extended to 0.95 Å. All of these steps consisted of multiple cycles of refinement and manual model building. $2F_o$-F_c and F_o-F_c maps were monitored along with the peak list from SHELX, and geometry outliers were identified by ProCheck (Laskowski et al., 1993). Iterative rounds of model building/refinement continued until convergence.

This produced the final GR_{Native} model (R=0.122, R_{free}=0.151) at 0.95 Å. This model was used as a starting point for refinement of the GR_{NADPH}, GR_{GSH} and $GR_{GSSG/NADP}$ complex structures. Further refinement and manual model building were performed as needed for each of these structures independently from this common starting point. Refinement results are shown in Table 2.2.

X-ray data collection and refinement for the room-temperature 1.8 Å $GR_{NADPH,1.8}$ structure

This data set used another NADPH-soaked crystal, prepared identically to the above-described crystals but not frozen. The crystal was mounted in a glass capillary filled with the soak solution and G25 Sephadex to immobilize the crystal (as in Karplus & Schulz 1989). Data for the 1.8 Å, room-temperature GR_{NADPH} was collected at our in-house X-ray source with Cu-Kα radiation (Rigaku RU-H3R rotating anode operating at 50 kV and 100 mA and an R-Axis IV image plate detector; λ=1.54 Å, Δφ=0.3º, 400 10-min images) (Table 2.2). The structure was solved using molecular replacement of the 1.54 Å GR_{Native} structure and refined with first CNS (Brunger et al., 1998), then Refmac (Winn et al., 2003). Between refinement cycles, F_o-F_c and $2F_o$-F_c electron-density maps were used for manual rebuilding with O (Jones, 1978), and later with Coot (Emsley and Cowtan, 2004).

To account for large-scale anisotropy shared among residue groups, TLS (translation, libration and screw) domains were added in Refmac based on a TLS refinement of malarial GR (Sarma et al., 2003) and checked using TLSMD (Painter and Merritt, 2006). The addition of TLS domains to the refinement resulted in decreases in R and R_{free} of 0.7% and 1.3%, respectively.

Structural comparisons and analyses

Noncovalent interactions were analyzed using Coot's (Emsley and Cowtan, 2004) interface to Reduce 3.0 and Probe 2.11 (Word et al., 1999), which displays favorable and unfavorable van der Waals interactions as well as hydrogen bonds visually within Coot.

Global anisotropy was analyzed using PARVATI (Merritt, 1999), which produced a distribution of atomic displacement parameters (ADPs) across each structure and highlighted the most anisotropic residues. Local ADPs were analyzed using Coot or PyMol to display anisotropy with ellipsoids at 50% or 67% probability. Figure 2.5 shows anisotropy at 67% probability. Structural figures were also generated with PyMol (DeLano, 2002).

Structural overlays were created using the McLachlan algorithm (McLachlan, 1982) as implemented in the program ProFit by iteratively overlaying structures using a subset of Cα atoms with a maximum per-atom RMSD of 0.1 Å until convergence was reached. When overlaying more than 2 structures, ProFit's multiple overlay function was used, and structures were overlaid in order of decreasing resolution to minimize bias to the average structure. For overlays containing multiple conformations and substrates, the original ProFit-overlaid structure (lacking multiple conformations) had a complete structure (including multiple conformations) overlaid upon it in Coot using least-squares fit, resulting in an identically overlaid set of structures.

Protein Geometry Database

The Protein Geometry Database (PGD; Berkholz et al., unpublished), an in-house database derived from the PDBselect list of protein structures with nonredundant sequences (Hobohm and Sander, 1994), was used to compare the GR structure with the rest of the Protein Data Bank. The PGD contains primarily covalent backbone geometry and is flexibly searchable, with results available via data dumps or a graphing module. Searches were performed on structures determined at 1.2 Å resolution or better, using a subset of the PGD containing no proteins with >90% sequence identity. Figures of the results were generated using gnuplot.

Protein Data Bank entry codes

Models were deposited into the RCSB Protein Data Bank with accession codes 3dk9, 3dk4, 3dk8, 3djj and 3djg for GR_{Native}, $GR_{GSSG/NADP}$, GR_{GSH}, GR_{NADPH}, and $GR_{NADPH,1.8}$, respectively.

Acknowledgement

This work was sponsored by National Science Foundation grant MCB-9982727 to P.A.K.

References

Almarsson, O., and Bruice, T.C. (1993). Evaluation of the factors influencing reactivity and stereospecificity in NAD(P)H dependent dehydrogenase enzymes. Journal of the American Chemical Society *115*, 2125-2138.

Alphey, M.S., Attrill, H., Crocker, P.R., and van Aalten, D.M.F. (2003). High resolution crystal structures of Siglec-7: Insights Into ligand specificity in the Siglec family. Journal of Biological Chemistry *278*, 3372-3377.

Argyrou, A., Blanchard, J.S., and Palfey, B.A. (2002). The lipoamide dehydrogenase from Mycobacterium tuberculosis permits the direct observation of flavin intermediates in catalysis. Biochemistry *41*, 14580-14590.

Benner, S.A. (1982). The stereoselectivity of alcohol dehydrogenases: A stereochemical imperative? Cellular and Molecular Life Sciences (CMLS) *38*, 633-637.

Berglund, G.I., Carlsson, G.H., Smith, A.T., Szöke, H., Henriksen, A., and Hajdu, J. (2002). The catalytic pathway of horseradish peroxidase at high resolution. Nature *417*, 463-8.

Böhme, C.C., Arscott, L.D., Becker, K., Schirmer, R.H., and Williams, C.H. (2000). Kinetic characterization of glutathione reductase from the malarial parasite Plasmodium falciparum. Comparison with the human enzyme. Journal of Biological Chemistry *275*, 37317-23.

Bruice, T., and Lightstone, F. (1999). Ground state and transition state contributions to the rates of intramolecular and enzymatic reactions. Accounts of Chemical Research *32*, 127-136.

Bruice, T.C., and Pandit, U.K. (1960). Intramolecular models depicting the kinetic importance of "fit" in enzymatic catalysis. Proceedings of the National Academy of Sciences *46*, 402-404.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallographica Section D Biological Crystallography *54*, 905-21.

Burmeister, W.P. (2000). Structural changes in a cryo-cooled protein crystal owing to radiation damage. Acta Crystallographica Section D Biological Crystallography *56*, 328-341.

Carugo, O., and Carugo, K.D. (2005). When X-rays modify the protein structure: radiation damage at work. Trends in Biochemical Sciences *30*, 213-219.

Cavelier, G., and Amzel, L.M. (2001). Mechanism of NAD (P) H: Quinone reductase: Ab

initio studies of reduced flavin. Proteins Structure Function and Genetics *43*, 420-432.

Cruickshank, D.W.J. (2001). International Tables for Crystallography. In International Tables for Crystallography, M G Rossmann, and E Arnold, eds. (: Dordrecht: Kluwer Academic Publishers), pp. 403-418.

Dauter, Z., Lamzin, V.S., and Wilson, K.S. (1997). The benefits of atomic resolution. Current Opinion in Structural Biology *7*, 681-688.

De Colibus, L., and Mattevi, A. (2006). New frontiers in structural flavoenzymology. Current Opinion in Structural Biology *16*, 722-728.

DeLano, W.L. (2002). The PyMOL Molecular Graphics System (: DeLano Scientific, San Carlos, CA).

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallographica. Section D, Biological Crystallography *60*, 2126-2132.

Fox, K.M., and Karplus, P.A. (1999). The flavin environment in old yellow enzyme: An evaluation of insights from spectroscopic and artificial flavin studies. Journal of Biological Chemistry *274*, 9357-9362.

Fraaije, M.W., and Mattevi, A. (2000). Flavoenzymes: diverse catalysts with recurrent features. Trends in Biochemical Sciences *25*, 126-132.

Ghisla, S., and Massey, V. (1989). Mechanisms of flavoprotein-catalyzed reactions. FEBS Journal *181*, 1-17.

G. L. Miessler, and D. A. Tarr (2004). Inorganic Chemistry. In Inorganic Chemistry (: Pearson Prentice Hall), pp. 116-164.

Heine, A., DeSantis, G., Luz, J.G., Mitchell, M., Wong, C., and Wilson, I.A. (2001). Observation of covalent intermediates in an enzyme mechanism at atomic resolution. Science *294*, 369-374.

Heine, A., Luz, J.G., Wong, C., and Wilson, I.A. (2004). Analysis of the class I aldolase binding site architecture based on the crystal structure of 2-deoxyribose-5-phosphate aldolase at 0.99 Å resolution. Journal of Molecular Biology *343*, 1019-1034.

Hobohm, U., and Sander, C. (1994). Enlarged representative set of protein structures. Protein Sci. *3*, 522-524.

Huber, P.W., and Brandt, K.G. (1980). Kinetic studies of the mechanism of pyridine nucleotide dependent reduction of yeast glutathione reductase. Biochemistry *19*, 4568-4575.

Jones, T.A. (1978). A graphics model building and refinement system for macromolecules. Journal of Applied Crystallography *11*, 268-272.

Karplus, P.A. (1999). Flavins and flavoproteins 1999. In Flavins and flavoproteins 1999, S. Ghisla, P. Kroneck, P. Macheroux, and H. Sund, eds. (: Agency for Scientific Publ.), pp. 233-238.

Karplus, P.A. (1996). Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci. *5*, 1406-1420.

Karplus, P.A., Pai, E.F., and Schulz, G.E. (1989). A crystallographic study of the glutathione binding site of glutathione reductase at 0.3-nm resolution. Eur J Biochem *178*, 693-703.

Karplus, P.A., and Schulz, G.E. (1987). Refined structure of glutathione reductase at 1.54 Å resolution. Journal of Molecular Biology *195*, 701-29.

Karplus, P.A., and Schulz, G.E. (1989). Substrate binding and catalysis by glutathione reductase as derived from refined enzyme: substrate crystal structures at 2 Å resolution. Journal of Molecular Biology *210*, 163-180.

Khan, H., Barna, T., Harris, R.J., Bruce, N.C., Barsukov, I., Munro, A.W., Moody, P.C.E., and Scrutton, N.S. (2004). Atomic resolution structures and solution behavior of enzyme-substrate complexes of Enterobacter cloacae PB2 pentaerythritol tetranitrate reductase: Multiple conformational states and implications for the mechanism of nitroaromatic explosive degradation. Journal of Biological Chemistry *279*, 30563-30572.

Kort, R., Hellingwerf, K.J., and Ravelli, R.B.G. (2004a). Initial events in the photocycle of photoactive yellow protein. Journal of Biological Chemistry *279*, 26417-26424.

Kort, R., Komori, H., Adachi, S., Miki, K., and Eker, A. (2004b). DNA apophotolyase from Anacystis nidulans: 1.8 Å structure, 8-HDF reconstitution and X-ray-induced FAD reduction. Acta Crystallogr D Biol Crystallogr *60*, 1205-1213.

Krauth-Siegel, R.L., Arscott, L.D., Schoenleben-Janas, A., Schirmer, R.H., and Williams, C.H. (1998). Role of active site tyrosine residues in catalysis by human glutathione reductase. Biochemistry *37*, 13968-13977.

Lario, P.I., Sampson, N., and Vrielink, A. (2003). Sub-atomic resolution crystal structure of cholesterol oxidase: What atomic resolution crystallography reveals about enzyme mechanism and the role of the FAD cofactor in redox activity. Journal of Molecular Biology *326*, 1635-1650.

Lennon, B.W., Williams, C.H., and Ludwig, M.L. (1999). Crystal structure of reduced thioredoxin reductase from Escherichia coli: structural flexibility in the isoalloxazine ring of the flavin adenine dinucleotide cofactor. Protein science : a publication of the Protein Society *8*, 2366-2379.

Liotta, C.L., Burgess, E.M., and Eberhardt, W.H. (1984). Trajectory analysis. 1. Theoretical model for nucleophilic attack at pi-systems. The stabilizing and destabilizing

orbital terms. Journal of the American Chemical Society *106*, 4849-4852.

Massey, V. (1995). Introduction: flavoprotein structure and mechanism. FASEB Journal *9*, 473-5.

Massey, V. (2000). The chemical and biological versatility of riboflavin. Biochemical Society Transactions *28*, 283-296.

Matthews, R.G., Ballou, D.P., and Williams, C.H. (1979). Reactions of pig heart lipoamide dehydrogenase with pyridine nucleotides. Evidence for an effector role for bound oxidized pyridine nucleotide. The Journal of biological chemistry *254*, 4974-81.

Merritt, E.A. (1999). Expanding the model: anisotropic displacement parameters in protein structure refinement. Acta Crystallographica Section D Biological Crystallography *55*, 1109-1117.

Miller, S.M., Massey, V., Ballou, D., Williams, C.H., Distefano, M.D., Moore, M.J., and Walsh, C.T. (1990). Use of a site-directed triple mutant to trap intermediates: demonstration that the flavin C(4a)-thiol adduct and reduced flavin are kinetically competent intermediates in mercuric ion reductase. Biochemistry *29*, 2831-2841.

Miura, R. (2001). Versatility and specificity in flavoenzymes: Control mechanisms of flavin reactivity. The Chemical Record *1*, 183-194.

Nambiar, K.P., Stauffer, D.M., Kolodziej, P.A., and Benner, S.A. (1983). A mechanistic basis for the stereoselectivity of enzymic transfer of hydrogen from nicotinamide cofactors. Journal of the American Chemical Society *105*, 5886-5890.

Otwinowski, Z., and Minor, W. (1997). Processing of X-Ray Diffraction Data Collected in Oscillation Mode. Academic Press, New York, 307–326.

Painter, J., and Merritt, E.A. (2006). TLSMD web server for the generation of multi-group TLS models. Journal of Applied Crystallography *39*, 109-111.

Rajagopalan, P.T.R., and Benkovic, S.J. (2002). Preorganization and protein dynamics in enzyme catalysis. The Chemical Record *2*, 24-36.

Rasmussen, B.F., Stock, A.M., Ringe, D., and Petsko, G.A. (1992). Crystalline ribonuclease A loses function below the dynamical transition at 220 K. Nature *357*, 423-424.

Ravelli, R.B.G., and Garman, E.F. (2006). Radiation damage in macromolecular cryocrystallography. Current Opinion in Structural Biology *16*, 624-629.

Ravelli, R.B.G., and McSweeney, S.M. (2000). The ‘fingerprint’ that X-rays can leave on structures. Structure *8*, 315-328.

Rietveld, P., Arscott, L.D., Berry, A., Scrutton, N.S., Deonarain, M.P., Perham, R.N., and

Williams Jr, C.H. (1994). Reductive and oxidative half-reactions of glutathione reductase from Escherichia coli. Biochemistry *33*, 13888-13895.

Rivas, P., Zapata-Torres, G., Melin, J., and Contreras, R. (2004). Probing the hydride transfer process in the lumiflavine–1-methylnicotinamide model system using group softness. Tetrahedron *60*, 4189-4196.

Roberts, B.R., Wood, Z.A., Jonsson, T.J., Poole, L.B., and Karplus, P.A. (2005). Oxidized and synchrotron cleaved structures of the disulfide redox center in the N-terminal domain of Salmonella typhimurium AhpF. Protein Science *14*, 2414.

Sarma, G.N., Savvides, S.N., Becker, K., Schirmer, M., Schirmer, R.H., and Karplus, P.A. (2003). Glutathione reductase of the malarial parasite Plasmodium falciparum: crystal structure and inhibitor development. Journal of Molecular Biology *328*, 893-907.

Savvides, S.N., and Karplus, P.A. (1996). Kinetics and crystallographic analysis of human glutathione reductase in complex with a xanthene inhibitor. J Biol Chem *271*, 8101-8107.

Sem, D.S., and Kasper, C.B. (1992). Geometric relationship between the nicotinamide and isoalloxazine rings in NADPH-cytochrome P-450 oxidoreductase: implications for the classification of evolutionarily and functionally related flavoproteins. Biochemistry *31*, 3391-3398.

Sevcik, J., Dauter, Z., Lamzin, V.S., and Wilson, K.S. (1996). Ribonuclease from Streptomyces aureofaciens at Atomic Resolution. Acta Crystallogr. D. Biol. Crystallogr. *52*, 327-344.

Sheldrick, G., and Schneider, T. (1997). SHELXL: High-resolution refinement. Methods in Enzymology *277*, 319-343.

Stehle, T., Claiborne, A., and Schulz, G.E. (1993). NADH binding site and catalysis of NADH peroxidase. FEBS Journal *211*, 221-226.

Stirnimann, C.U., Rozhkova, A., Grauschopf, U., Böckmann, R.A., Glockshuber, R., Capitani, G., and Grütter, M.G. (2006). High-resolution structures of Escherichia coli cDsbD in different redox states: A combined crystallographic, biochemical and computational study. Journal of Molecular Biology *358*, 829-845.

Sustmann, R., Sicking, W., and Schulz, G.E. (1989). The Active Site of Glutathione Reductase: An Example of Near Transition-State Structures. Angewandte Chemie International Edition in English *28*, 1023-1025.

Thorpe, C., and Williams, C.H. (1976). Spectral evidence for a flavin adduct in a monoalkylated derivative of pig heart lipoamide dehydrogenase. J. Biol. Chem. *251*, 7726-7728.

Tronrud, D.E. (1997). TNT refinement package. Methods Enzymol *277*, 306-319.

Trueblood, K.N., Burgi, H.B., Burzlaff, H., Dunitz, J.D., Gramaccioli, C.M., Schulz, H.H., Shmueli, U., and Abrahams, S.C. (1996). Atomic dispacement parameter nomenclature. Report of a subcommittee on atomic displacement parameter nomenclature. Acta Crystallographica Section A Foundations of Crystallography *52*, 770-781.

Vanoni, M.A., Wong, K.K., Ballou, D.P., and Blanchard, J.S. (1990). Glutathione reductase: comparison of steady-state and rapid reaction primary kinetic isotope effects exhibited by the yeast, spinach, and Escherichia coli enzymes. Biochemistry *29*, 5790-5796.

Weik, M., and Sussman, J. (2000). Synchotron X-ray radiation produces specific chemical and structural damage to protein structures. Proceedings of the National Academy of Sciences *97*, 623–628.

Wiita, A.P., Ainavarapu, S.R.K., Huang, H.H., and Fernandez, J.M. (2006). Force-dependent chemical kinetics of disulfide bond reduction observed with single-molecule techniques. Proceedings of the National Academy of Sciences *103*, 7222-7227.

Winn, M.D., Murshudov, G.N., and Papiz, M.Z. (2003). Macromolecular TLS refinement in REFMAC at moderate resolutions. Methods in Enzymology *374*, 300-21.

Word, J.M., Lovell, S.C., LaBean, T.H., Taylor, H.C., Zalis, M.E., Presley, B.K., Richardson, J.S., and Richardson, D.C. (1999). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. Journal of Molecular Biology *285*, 1711-1733.

Wu, Y.D., and Houk, K.N. (1991). Theoretical evaluation of conformational preferences of NAD+ and NADH: an approach to understanding the stereospecificity of NAD+/NADH-dependent dehydrogenases. Journal of the American Chemical Society *113*, 2353-2358.

Wu, Y.D., and Houk, K.N. (1993). Theoretical study of conformational features of NAD+ and NADH analogs: protonated nicotinamide and 1, 4-dihydronicotinamide. The Journal of Organic Chemistry *58*, 2043-2045.

Yates, R.L., Epiotis, N.D., and Bernardi, F. (1975). Importance of nonbonded attraction in the stereochemistry of the SN2' reaction. Journal of the American Chemical Society *97*, 6615-6621.

Young, L., and Post, C.B. (1996). Catalysis by entropic guidance from enzymes. Biochemistry *35*, 15129-15133.

Table 2.1. Flavin covalent bond geometry[a]

	N5-C4a	C4a-C4	C4a-C10a	C10a-N1
Small-molecule flavins[b]				
Oxidized	1.30	1.49	1.47	1.32
Oxidized-H+	1.30	1.49	1.42	1.36
Oxidized(-)	1.30	1.49	1.45	1.31
Reduced	1.42	1.40	1.37	1.38
GR 2.0 Å structures[c]				
GR_{Native}[d]	1.33	1.44	1.45	1.35
$GR_{GSSG/NADP}$	1.30	1.49	1.44	1.32
GR_{GSH}	1.32	1.48	1.47	1.33
GR_{NADPH}	1.30	1.49	1.47	1.33
GR 1.0 Å structures[e]				
GR_{Native}	1.36	1.47	1.38	1.33
$GR_{GSSG/NADP}$	1.39	1.47	1.36	1.38
GR_{GSH}	1.38	1.39	1.40	1.36
GR_{NADPH}	1.37	1.43	1.40	1.37
Atomic-resolution flavoenzymes				
Cholesterol oxidase[f]	1.40	1.41	1.41	1.36
PETN reductase[g]	1.39	1.40	1.40	1.37

[a]All numbers in Å
[b]From Karplus, 1999
[c]From Karplus, 1989
[d]GR_{Native} is 1.54 Å
[e]From this paper
[f]PDB ID 1n4w
[g]PDB ID 2abb

Table 2.2. Data collection and refinement statistics[a]

	GR_{Native}	$GR_{GSSG/NADP}$	GR_{GSH}	GR_{NADPH}
Cell dimensions (Å)				
a	120.4	119.4	120.0	119.7
b	62.4	62.2	62.3	62.6
c	84.0	83.9	84.0	84.2
β (°)	122.0	121.9	121.9	122.3
Resolution limit (Å)	0.95	1.1	1.0	1.0
Unique observations	290866	192429	232314	249970
Multiplicity	4.1	3.5	2.8	3.2
Average I/σ	11.3 (2.1)	7.6 (1.9)	7.1 (3.0)	7.2 (2.0)
R_{meas} (%)	7.6 (45.2)	8.1 (42.4)	9.4 (29.4)	9.8 (30.7)
Completeness (%)	88 (45.2)[b]	91 (74.8)	84.1 (55.7)[b]	90.9 (64.5)[b]
Refinement				
Reflections with F > 0 σ	276635	147475	184585	194474
Protein atoms	3835	3849	3842	3842
Heteroatoms	78	208	178	131
Solvent atoms	825	916	906	858
Hydrogen atoms	3317.8	3359	3351.5	3359
rmsd bonds (Å)	0.018	0.015	0.016	0.016
rmsd angles (Å)	0.038	0.034	0.033	0.033
$<B_{protein}>$ (Å^2)	14.2	20.4	17.4	17.8
$<B_{ligand}>$ (Å^2)	9.0	20.1	13.8	12.1
R_{cryst} (R_{free}) (%)	12.2 (15.1)	11.4 (16.4)	11.3 (14.7)	12.33 (15.7)
Coordinate error (Å)[c]	0.018	0.028	0.022	0.022

[a] Numbers in parentheses refer to the highest-resolution shell.
[b] The resolution cutoffs for which local completeness exceeds 75% are 0.97 Å, 1.05 Å and 1.1 Å for GRNative, GR_{GSH} and GRNADPH, respectively.
[c] Coordinate estimated standard uncertainty calculated using the R_{free} variant of Cruickshank's DPI (2001), with the equation $\sigma(x, B_{avg}) = 1.0(N_i/n_{obs})^{1/2}C^{-1/3}R_{free}d_{min}$, where N_i is the number of non-hydrogen atoms, n_{obs} is the number of unique observations, and C is the completeness.

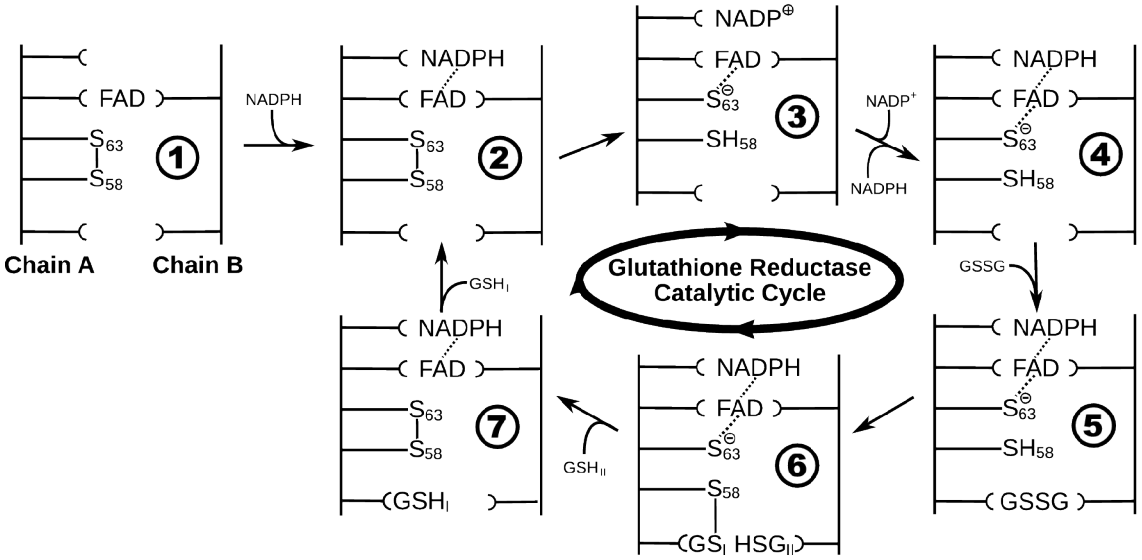


Figure 2.1. Catalytic cycle of glutathione reductase. The native state with no substrate bound is not part of the cycle but merely forms an entrance point. Dotted lines indicate charge-transfer complexes between NADPH, FAD, and the sulfur of Cys63. The substrate and product binding and dissociation may occur with different timing than that shown. The four crystal structures reported here provide information about the catalytic intermediates as follows: **1** is GR_{Native} ; **2** are derived from GR_{Native} and GR_{NADPH} ; **3** is derived from GR_{NADPH} and $GR_{GSSG/NADP}$; **4** is GR_{NADPH} ; **5** is derived from $GR_{GSSG/NADP}$ and GR_{NADPH} ; **6** is derived from GR_{GSH} and GR_{NADPH} ; **7** is derived from GR_{Native} and GR_{NADPH} , with $GR_{GSSG/NADP}$ and GR_{GSH} providing an idea for the GSH_I binding site. Created in Inkscape.

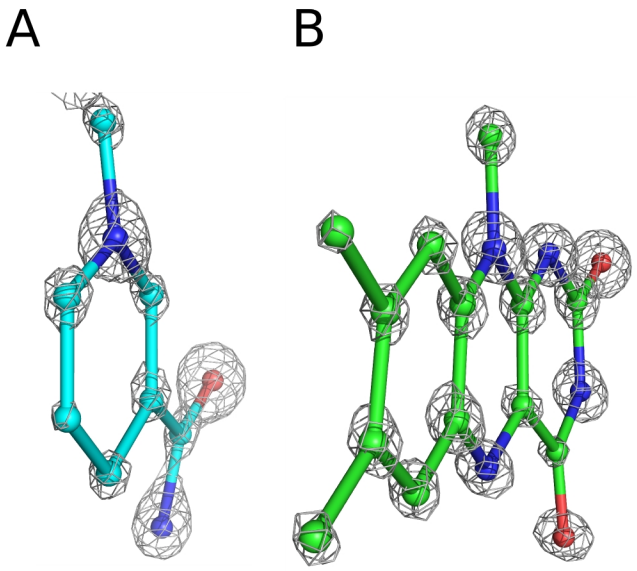


Figure 2.2. Atomic-resolution electron density for the active-site cofactors. a) The nicotinamide ring of NADPH at 1.0 Å resolution (contour level 3.2*ρ_{rms}) from GR_{NADPH}, with carbons (cyan), nitrogens (blue) and oxygens (red) having distinct electron density levels. The C4 atom, which transfers a hydride to FAD, is at the bottom. Pyramidalization of atom N1, at the top of the ring, is visible. A slight twist in the carboxamide relative to the ring is also visible. b) The flavin ring system of FAD at 0.95 Å resolution (contour level 3.8*ρ_{rms}) from GR_{Native}, with coloring as in (a) except carbons are green. The N5 atom, where FAD receives electrons from NADPH, is at the bottom center. A small twist in the flavin is evident.

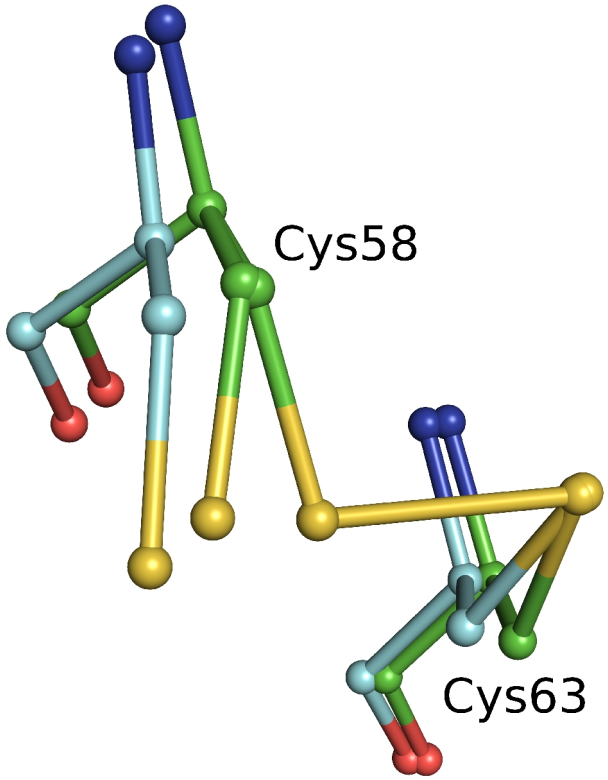


Figure 2.3. Disulfide bonds reduced by radiation at cryotemperatures are different from those reduced chemically. The GR_{Native} (green carbons) structure shows both the native disulfide and an alternate conformation for Cys58 due to radiation damage at cryotemperatures. GR_{NADPH} (cyan carbons) shows the structure resulting from chemical reduction at room temperature. The overlay shows a clear difference in the backbone relaxation of Cys58 that depends upon the mode of reduction.

Figure 2.4. Peptide non-planarity in the active-site disulfide loop. a) Stereoview of the disulfide loop with standard hydrogen bonds (green dotted lines) and unusually long "hydrogen bonds" (red dotted lines) shown. b) Views down each peptide bond in the loop in GR_{Native} visually reveal the magnitude of omega deviations from planarity, which are 4º, 13º, 7º, 10º, 5º, and 11º for residues 58-63. c) A plot of smoothed (N=5) omega deviations from planarity shows the disulfide loop (residues 59-63 in particular) as the most consistently non-planar region in GR. Omega deviations in this loop are similar in all four structures. d) Histogram of pentapeptide stretches in atomic-resolution structures with deviations from planarity (see Methods). The level of nonplanarity of this GR loop (ranging from 46˚ to 53˚ in the four GR structures) is unusual, with only two other examples of similarly deviating loops (see Results & Discussion).

Figure 2.4 (continued)

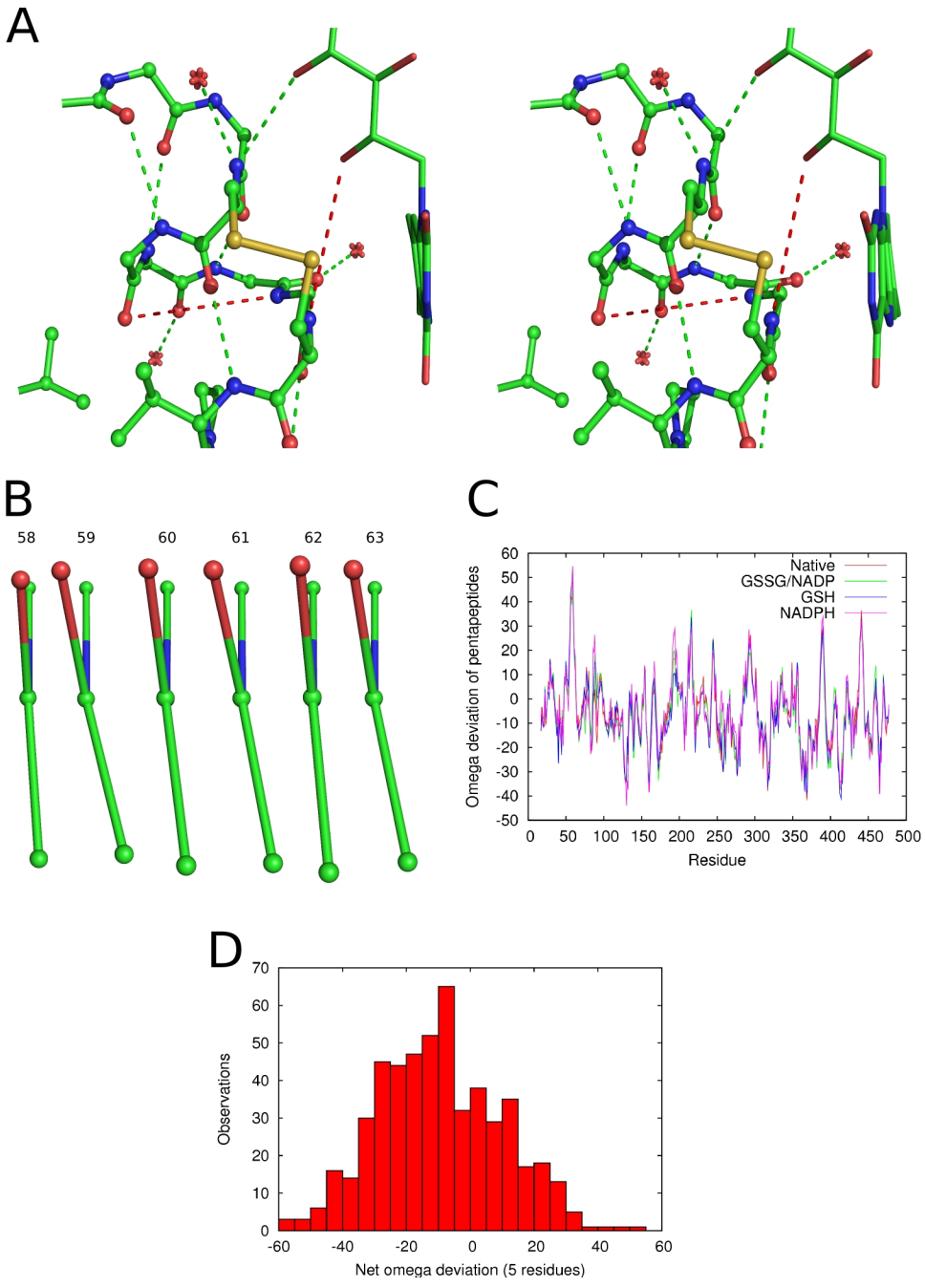
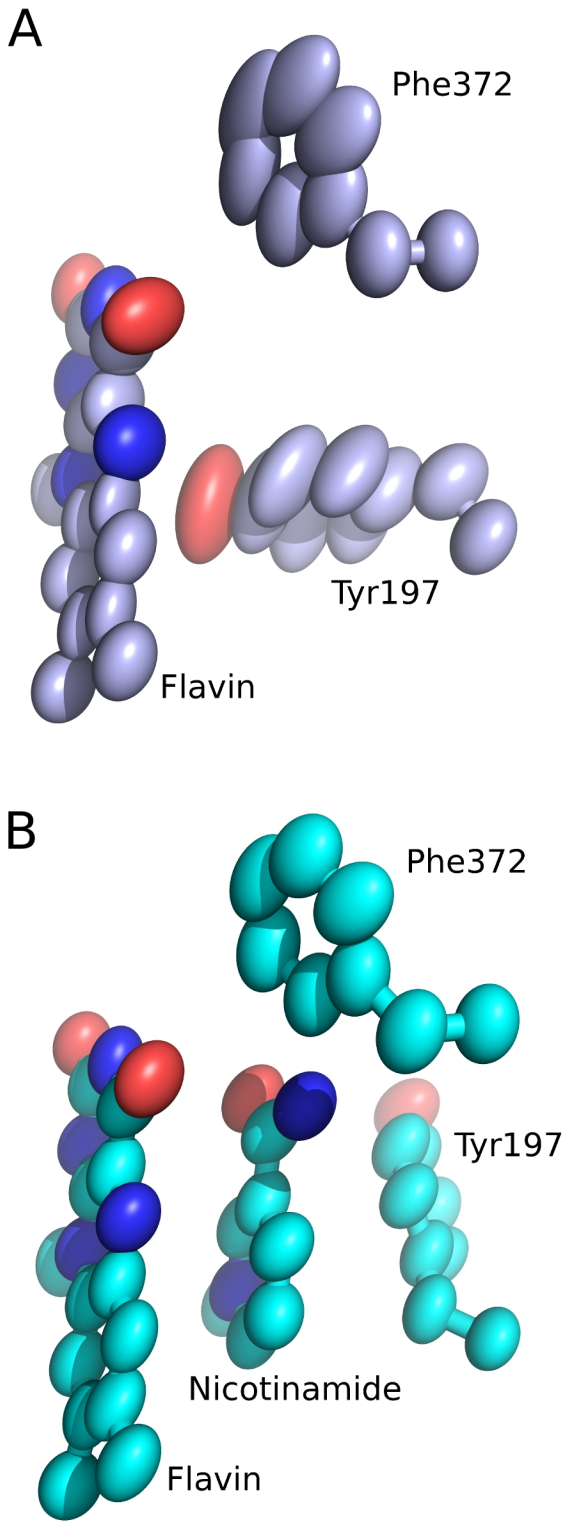


Figure 2.5. Nicotinamide binding tightens the active site. a) Anisotropic mobility is shown as thermal ellipsoids for residues as seen in the active site of the $GR_{GSSG/NADP}$ complex (carbons violet). b) The same view for the GR_{NADPH} complex (carbons cyan). In the NADPH complex, the motion is much lower and more isotropic.

Figure 2.5 (continued)



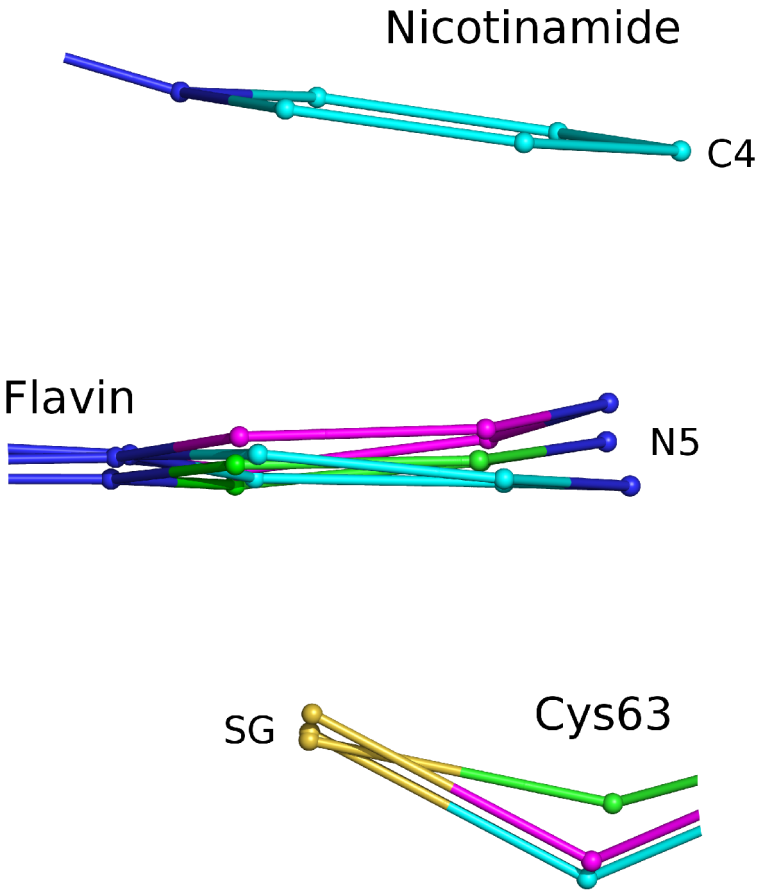


Figure 2.6. Steric compression in nicotinamide-flavin interaction. Side view of overlaid active site centered on flavin, showing GR_{Native} (green), GR_{GSH} (magenta) and GR_{NADPH} (cyan). In GR_{NADPH}, NADPH binding above the flavin pushes it down into Cys63, and in GR_{GSH}, GSH and the Cys63 thiolate below the flavin push it up. $GR_{GSSG/NADP}$ is not shown because its atoms are in the same positions as in GR_{Native}. To conserve space the flavin-nicotinamide and the flavin-Cys63 separations are not to scale. Figure 8 shows these distances to scale.

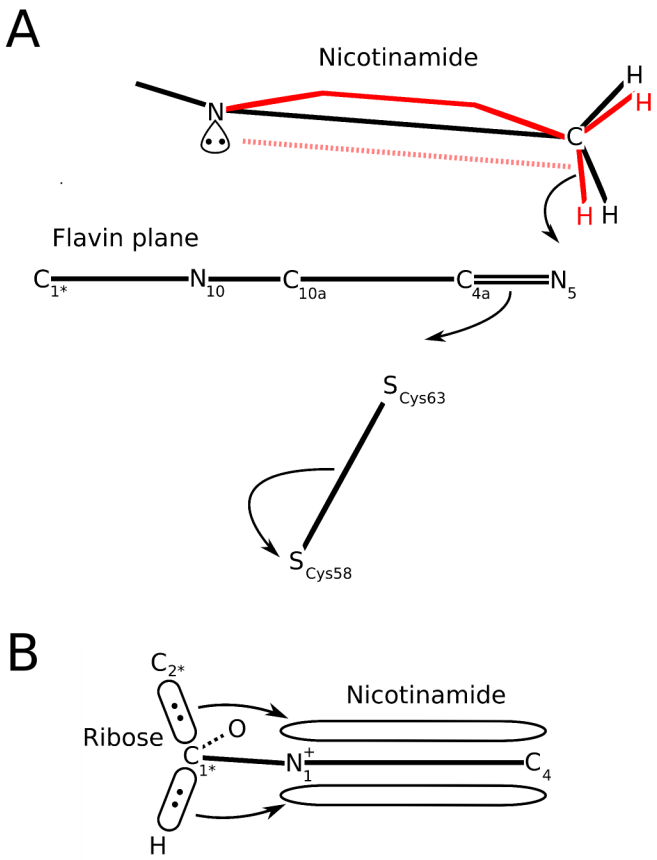


Figure 2.7. The nicotinamide distortion and ribose conformation favor catalysis. a) The schematic shows the planes of the nicotinamide and flavin (solid black lines). The hypothesized partial boat is shown as a solid red line. Pyramidalization at the nicotinamide N1 places the lone pair on the flavin side where it (i) entropically favors the productive boat conformation to form and (ii) repels the hydride to be transferred (dashed red line). b) The ribose conformation relative to the nicotinamide stabilizes the electron-deficient $NADP^+$ ring orbitals via hyperconjugative electron donation from the ribose. The glycosidic C-O bond position parallel to the nicotinamide ring also favors $NADP^+$ over NADPH (see Results & Discussion).

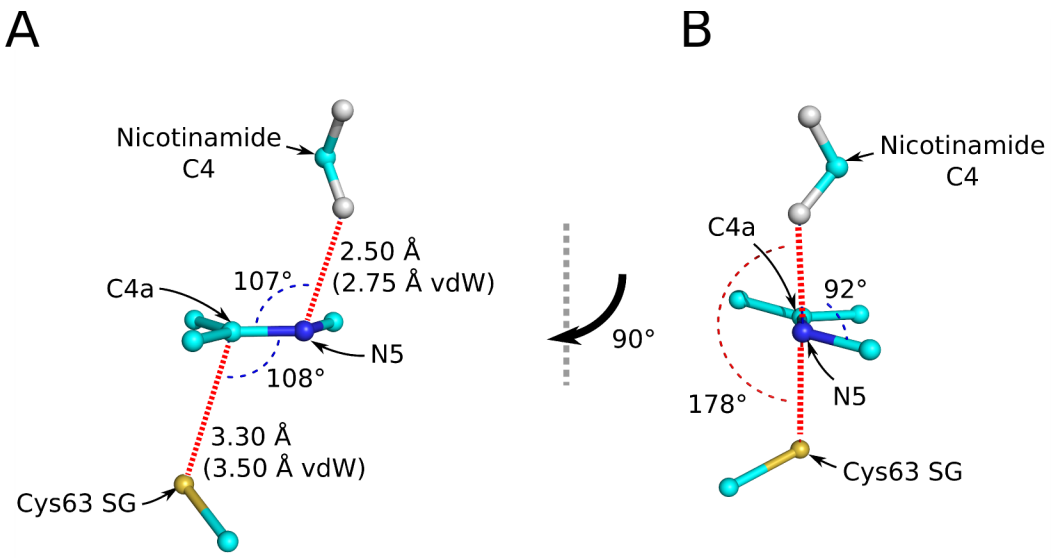


Figure 2.8. Stereoelectronic control in nicotinamide-flavin interaction. (a) A side view with the flavin N5-C4a bond in the plane of the paper and (b) a view down the flavin N5-C4a bond together show the optimal geometry for concerted 1-2 addition across the double bond. Compression in the form of shorter than van der Waals interactions is also shown in (a).

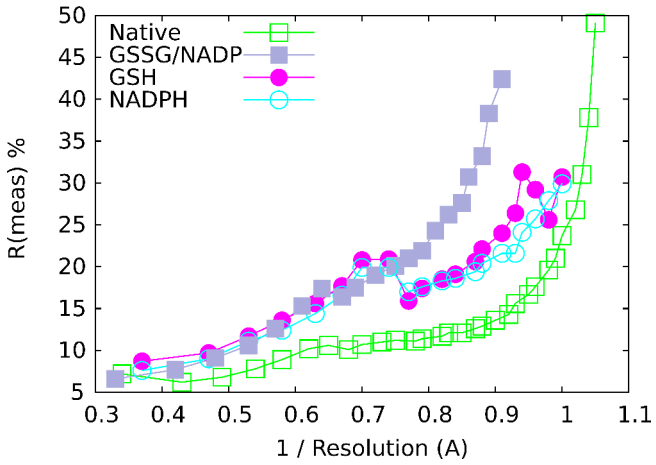


Figure 2.9. Data quality as a function of resolution. Observed data quality is indicated by Rmeas values and plotted as a function of resolution-1. The four structures are colored as indicated in the key. These data were used to determine where to set the high-resolution cutoff (see Methods).

Chapter 3

Protein Geometry Database: A flexible engine to explore backbone conformations and their relationships to covalent geometry

Donald S. Berkholz, Peter B. Krenesky, John R. Davidson, and P. Andrew Karplus

Submitted to *Nucleic Acids Research*, Database Issue on August 14, 2009

Abstract

Peptide geometry—the protein's bond lengths and bond angles—is tightly intermingled with the backbone conformation (Φ,Ψ), but no tool exists to explore these relationships, leaving this area as a reservoir of untapped information about protein structure and function. The Protein Geometry Database (PGD) enables biologists to easily and flexibly query information about the conformation alone, the backbone geometry alone, and the relationships between them. The capabilities the PGD provides are invaluable in assessing the uniqueness of observed conformational or geometric features in protein structure as well as discovering novel features and principles of protein structure. The PGD server is available at http://pgd.science.oregonstate.edu/ and the data and code underlying it are freely available to use and extend.

Introduction

With the explosion of atomic-resolution protein structures in the past decade, the possibility to accurately determine the details of protein geometry from proteins themselves rather than from small-molecule peptides has finally become a reality. The importance of bond angles and bond lengths in validating structures as well as discovering real and functionally important deviations from standard geometry has become increasingly apparent (Lawson, 1996); (Dobson et al., 2008); (Merritt et al., 1998); (Davis et al., 2007); (Esposito et al., 2000); (Laidig and Cameron, 1993); (Karplus, 1996). To our knowledge, no database has existed until now to search peptide geometry either on a large scale to discover trends or on an individual basis to explore unusual features. Highly unusual and significant features often pass unrecognized even by the structural biologists who solved the structure (Figure 1).

The protein backbone is defined primarily by three dihedral angles: Φ, Ψ, and ω.

The first two are highly rotatable and define the residue's conformation. The third, ω, defines the planarity of the peptide bond and generally is near 180° (for trans peptides) or 0° (for cis peptides). In addition, bond angles and lengths determine the details of peptide geometry, and their averages vary in a conformation-dependent manner, reflecting an intimate relationship between conformation and geometry (Karplus, 1996). The prevalent misconception that bond angles and lengths are static has been caused in part by the lack of any straightforward way to examine their dependence on local conformation. The Protein Geometry Database is a unique resource that now makes it possible for biologists to explore peptide geometry, conformation, and the ties between them. Other databases exist to allow searching conformation alone [e.g., SPASM server (Kleywegt, 1999), Fragment Finder (Ananthalakshmi et al., 2005), Protein Segment Finder (Samson and Levitt, 2009), Conformational Angles Database (Sheik et al., 2003), PDBeMotif (Golovin and Henrick, 2008)], but even in this arena, the Protein Geometry Database offers a uniquely enabling level of convenience and flexibility compared to these other databases.

Implementation

The intent of the PGD is to contain derived data for a complete, representative data set of protein structures that are relevant to discovering reliable instances of conformations and peptide geometries. This allows users to set thresholds specific to their queries at search time without being unduly limited by the cutoffs chosen during database creation. To ensure that the PGD data are representative of conformational and geometric space rather than being biased by multiple highly similar structures, the PGD contains data derived from a nonredundant set of proteins. As is common, the nonredundancy is defined by the maximum allowed sequence identity between any pair of proteins in the data set. Two thresholds of 25% and 90% are available in the PGD. The nonredundant set is taken from PISCES (Wang and Dunbrack, 2003). Because different resolution ranges are suitable for different queries, the PGD maximizes structural data by using a generous low-resolution cutoff of 3.0 Å resolution or better and no cutoff for the crystallographic R-factor. The 3

Å resolution cutoff exists because structures determined at resolutions poorer than this have increased torsion-angle uncertainty that makes them less useful.

The PGD contains data on per-chain and per-residue levels. For each chain, stored parameters include the PDB code, the chain ID, the sequence-identity threshold, the resolution, and the crystallographic R-factor. The sequence-identity threshold, resolution, and R-factor are all useful in parameters to define the independence and quality of the data searched. For each residue, stored parameters include a mapping back to the chain and protein, the residue number, the torsion angles Φ, Ψ, ω, and χ_1, the improper dihedral ζ [describing the chirality of the C_α (Laskowski et al., 1993)], all seven backbone bond angles, all five backbone bond lengths, the DSSP-defined (Kabsch and Sander, 1983) secondary-structure type, and three B-factors: the mainchain average, the sidechain average, and the C_γ atom. The B-factors are intended for use as thresholds to ensure that atomic positions and thus conformation and geometry are well-defined [e.g., (Karplus, 1996), (Lovell et al., 2003)].

The PGD uses the Python-based Django framework for both populating and searching a MySQL database. Using Django allows us to follow the DRY principle ("Don't Repeat Yourself") by only having one description of the database format. This reduces the difficulty of changes, increases the clarity of code, and avoids potential conflicts between multiple descriptions. A single change can transform the database schema for all applications that use it. The database is populated by interfacing a tool written with BioPython (Cock et al., 2009) to calculate PDB-derived information. The tool, Splicer, splices derived data from the PDB files together into all possible consecutive segments from 1-10 residues long. This approach speeds searching because segments do not need to be constructed during every search.

A single run of Splicer to populate the PGD can take ~16 hours on a current single-processor compute node, so we constructed a new Python framework for distributed, parallel data processing called Pydra <http://pydra-project.osuosl.org/> that provides significant enhancements over all other Python-based alternatives. Using Pydra,

a parallel Splicer run across 20 CPUs on 4 nodes takes ~1 hour, providing nearly linear speedup over the single-processor method.

The current version of the PGD contains ~3.8 million residues from nearly 16,000 protein chains, with all amino acids and secondary-structure types being well-represented (Figure 2). New PISCES lists containing the nonredundant set of chains are generated weekly, and we intend to continue updating the PGD at that frequency.

Searching and Analyzing Results

The search page

The PGD has a professionally designed, user-friendly, highly flexible graphical interface for mining protein conformational and geometric space. Upon proceeding beyond the introductory entry page, users encounter the search page (Figure 3). On this page, users define all parts of their queries. Help is available for each section by clicking the adjacent question mark. Each of the criteria can be defined positively (e.g., Gly and Pro) or negatively (e.g., all but Gly, Pro).

At the top of the page is the length of the motif to be searched (from 1-10 residues), followed by protein-chain properties and residue properties. The protein-chain properties are the length of the motif, the resolution range for selecting crystal structures, the sequence-identity threshold, and specific PDB codes to search (defaults to the full PGD). Changing the motif length will cause the corresponding set of residue properties to appear.

The residue properties are grouped into five sections: composition, conformation, mobility, angles, and lengths. The composition section allows users to indicate any grouping of specific amino-acid types to search (i.e. with no limitation to predefined categories such as hydrophobic or acidic). The conformation section allows users to

restrict searches to specific classes of DSSP-defined secondary structure, defined as follows: 'H' — α-helix; 'G' — 3_{10} helix; 'E' — β-strand; 'T' — hydrogen-bonded turn; 'S' — non-hydrogen-bonded turn; 'I' — π-helix; and 'B' — β-bridge. The long names are used on the search page, and the short names are used when space is at a minimum (e.g., on the statistics page). The conformation section additionally offers options for more fine-grained conformational searches using ranges of Φ, Ψ, and the peptide planarity, ω. The mobility, angles, and lengths sections are collapsed by default to simplify the search page for first-time users and for those who only want to perform conformational searches; clicking the titles will expand them (other sections can be expanded or hidden in the same manner). The mobility section allows searching ranges of the three B-factors of the mainchain, sidechain, and C_γ-atom (B^m, B^{sc}, and B^γ, respectively). The angles are defined by three atoms and proceed in order from N- to C-terminus of the residue, with '-1' indicating an atom from the previous residue and '+1' indicating an atom from the next residue. The lengths are defined by two atoms and are otherwise named and searched identically to angles.

To allow for additional flexibility and convenience in searches, we made two significant enhancements beyond what is typically allowed in similar databases. First, we created a special query syntax for ranges that allows multiple ranges to be specified (using commas), which enables searches wrapping around circular angles in either direction (search ranges must always be specified as negative to positive from left to right). This is quite useful for searches of conformations (the β region extends beyond Ψ = +180°/-180°) or peptide planarity (which peaks at +180°/-180°). To make it difficult for users to create an invalid search, we also provide on-the-fly validation that highlights valid syntax in green and invalid syntax in red. Second, we created a special exclusion feature for selections (a green plus sign indicates when selections are included, and clicking it reverses the search to exclusion and displays a red minus sign) that allows users to easily exclude a small number of selections instead of tediously selecting almost all of them. This is useful for common cases like excluding Gly or Pro from a search.

Once a search is fully defined, clicking the "Submit" button passes the query to

the PGD, which immediately indicates that a search is in progress and displays results for simple searches on the initial output page in a matter of seconds.

The initial output page

Immediately following a search, the total number of results is reported in the upper left-hand corner and the numbers of results are displayed as a function of Φ and Ψ on an interactive Ramachandran plot (Figure 4). The plot is colored by observation density within 10° × 10° bins. To maximize the visual contrast, coloring uses a logarithmic scale derived from the plotted values. Moving the mouse cursor over any bin produces a JavaScript popup indicating the Φ and Ψ ranges and the observation count. The Ramachandran plot is not limited to displaying the number of observations but instead can show any of the PGD residue attributes, using colors (like a contour plot) or even on the X and Y axes (replacing Φ/Ψ). Attributes from any position of the search (i-4 to i+5) can be plotted by changing the "residue" parameter. Additionally, plots can be zoomed by changing the minima and maxima, and bin sizes can be modified. Further flexibility is available in the colors/dimensions section, hidden by default. To re-plot after changing any parameter, click the "Re-Plot" button. Plots or the summary data used to create them (bin definitions, observation counts, averages and standard deviations for each attribute) are downloadable using the "Save Plot Image" or "Save Plot Data" butttons, respectively.

Additional tools and analysis

A key feature of the PGD is that it enables extensive, unlimited analysis beyond its built-in capabilities by allowing users to download a complete set of search results. Clicking "Data Dump" will prompt download of a plain-text dump of the raw results for each matching motif in tab-separated value format, ideal for importing into other applications.

In addition to the summary data provided by the Ramachandran plot, the individual motifs found by the search are viewable online by clicking "Browse Results" at the top of the page. Highlighting of each column and row under the mouse cursor eases comparison within a residue or attribute. To reduce load time and maximize responsiveness, pagination splits up the potentially large result sets.

The "Statistics" link at the top of the page leads to a page of summary statistics about residue *i*, including a breakdown of observations by amino-acid type, secondary-structure types, and the average backbone covalent geometry. Scrolling a mouse cursor over the covalent-geometry values produces a pop-up window that displays the standard deviations and ranges. Automatic highlighting of the column and row under the mouse cursor eases comparisons within residue types and attributes.

Examples

The PGD enables searches of conformational and geometric space in a powerful, flexible manner that allows for a wide variety of uses, from understanding of large-scale patterns in protein structure to analyzing the significance and/or rarity of a feature in an individual structure. Here, we describe two examples from papers published by our group using the PGD that illustrate its two primary aspects of conformational and geometric searching.

Conformational searching

The first example is a large-scale analysis of protein conformation that asked a simple question: What linear groups with repeating Φ,Ψ pairs exist in proteins (Hollingsworth et al., 2009)? To answer this question, we used the PGD to search for well-defined ($B^m < 25$ $Å^2$) three-residue segments from structures solved at 1.2 Å resolution or better. At this resolution, the atomic positions and thus the torsion angles have high accuracy so if linear

groups are truly tightly grouped, they should be observed as such. To ensure a maximally representative result, we chose the 25% sequence-identity threshold and included all amino-acid types but only *trans* peptides (all three ω values limited to '-180--90,90-180'). To identify linear groups in specific regions, we required all three residues to be in the same 20° × 20° box and systematically searched all such boxes using a 10° sliding window. We found that only three true clusters of linear groups exist in proteins: the right-handed α-/3_{10}-helix, the β-strand (with no substantive difference between parallel and antiparallel), and the P_{II} helix (occupied by many nonproline residues, despite the misconception that only polyproline populates it). The 2.2_7 ribbon, π-helix, and left-handed α-/3_{10}-helical conformations only occur for isolated residues and rare short segments.

Geometric searching

The second example is a small-scale analysis investigating the commonality of a specific five-residue geometric motif (Berkholz et al., 2008). In glutathione reductase, a key active-site loop bridges two cysteines forming a redox-active disulfide bond. This loop has five consecutive residues with nonplanar peptide bonds, and intriguingly, they are all bent in the same direction with a summed deviation of 55° across the pentapeptide. We suspected this highly strained loop was involved in the enzyme's function. To find out how common such an ω-deviation was in proteins, we searched the PGD for all *trans* pentapeptides for which each residue was at least 5° away from planarity (ω delimiter: '-175--90,90-175'), using cutoffs of 1.2 Å resolution and 90% sequence identity. By downloading a data dump and creating a histogram of the net deviations, we discovered that the strained active-site loop of glutathione reductase was not just unique within its own structure but also nearly unique among all proteins, with only two other examples in the PGD.

Conclusions

As the examples illustrate, the ability to search peptide geometry and conformation can provide important insights into protein structure and function. The PGD is the only database to connect peptide geometry and conformation, and provides a highly flexible yet intuitive search interface. The PGD provides a user-friendly tool that allows biologists to explore important details of protein structure that are often missed or ignored.

Funding

This work was supported by the National Institutes of Health [R01-GM083136 to P.A.K.] and the National Science Foundation [MCB-9982727 to P.A.K.].

Acknowledgements

We would like to thank our beta testers, who made suggestions for improving the PGD. We would also like to thank all members of the development team at the Open Source Lab at Oregon State University who assisted with the PGD's coding.

References

Ananthalakshmi, P., Kumar, C.K., Jeyasimhan, M., Sumathi, K., and Sekar, K. (2005). Fragment Finder: a web-based software to identify similar three-dimensional structural motif. Nucleic Acids Res. *33*, W85-88.

Antonyuk, S.V., Strange, R.W., Sawers, G., Eady, R.R., and Hasnain, S.S. (2005). Atomic resolution structures of resting-state, substrate- and product-complexed Cu-nitrite reductase provide insight into catalytic mechanism. Proc. Natl. Acad. Sci. USA *102*, 12041-12046.

Berkholz, D.S., Faber, H.R., Savvides, S.N., and Karplus, P.A. (2008). Catalytic cycle of human glutathione reductase near 1 A resolution. J. Mol. Biol. *382*, 371-384.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J.L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422-1423.

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., 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.

Dobson, R.C.J., Griffin, M.D.W., Devenish, S.R.A., Pearce, F.G., Hutton, C.A., Gerrard, J.A., Jameson, G.B., and Perugini, M.A. (2008). Conserved main-chain peptide distortions: a proposed role for Ile203 in catalysis by dihydrodipicolinate synthase. Protein Sci. *17*, 2080-2090.

Esposito, L., Vitagliano, L., Zagari, A., and Mazzarella, L. (2000). Experimental evidence for the correlation of bond distances in peptide groups detected in ultrahigh-resolution protein structures. Protein Eng. *13*, 825-828.

Golovin, A., and Henrick, K. (2008). MSDmotif: exploring protein sites and motifs. BMC Bioinformatics *9*, 312.

Hollingsworth, S.A., Berkholz, D.S., and Karplus, P.A. (2009). On the occurrence of linear groups in proteins. Protein Sci. *18*, 1321-1325.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577-637.

Karplus, P.A. (1996). Experimentally observed conformation-dependent geometry and

hidden strain in proteins. Protein Sci. *5*, 1406-1420.

Kleywegt, G.J. (1999). Recognition of spatial motifs in protein structures. J. Mol. Biol. *285*, 1887-1897.

Laidig, K.E., and Cameron, L.M. (1993). What happens to formamide during C—N bond rotation? Atomic and molecular energetics and molecular reactivity as a function of internal rotation. Can. J. Chem. *71*, 872-879.

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. Journal of Applied Crystallography *26*, 283-291.

Lawson, C.L. (1996). An atomic view of the L-tryptophan binding site of trp repressor. Nat. Struct. Biol. *3*, 986-987.

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: ϕ, ψ and Cβ deviation. Proteins: Struct. Func. Genet. *50*, 437-450.

Merritt, E.A., Kuhn, P., Sarfaty, S., Erbe, J.L., Holmes, R.K., and Hol, W.G. (1998). The 1.25 A resolution refinement of the cholera toxin B-pentamer: evidence of peptide backbone strain at the receptor-binding site. J. Mol. Biol. *282*, 1043-1059.

Samson, A.O., and Levitt, M. (2009). Protein segment finder: an online search engine for segment motifs in the PDB. Nucleic Acids Res. *37*, D224-228.

Sheik, S.S., Ananthalakshmi, P., Bhargavi, G.R., and Sekar, K. (2003). CADB: Conformation Angles DataBase of proteins. Nucleic Acids Res. *31*, 448-451.

Wang, G., and Dunbrack, R.L. (2003). PISCES: a protein sequence culling server. Bioinformatics *19*, 1589-1591.

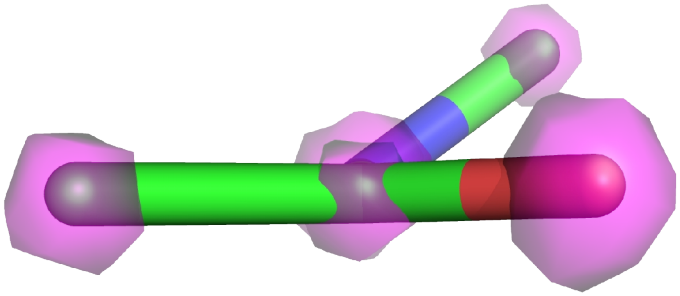


Figure 3.1. One example of a highly unusual, active-site peptide geometry feature discovered by use of the PGD. Shown is a view down the peptide bond between residues His306 and Asn307 in the 0.90 Å resolution structure of Cu-nitrite reductase [PDB code 2bw4 (Antonyuk et al., 2005)]. This peptide bond is 37° from planar ($\omega = 143°$). $2F_o$-F_c electron density shown at 6.0 ρ_{rms} (magenta transparent surface) indicates that the atoms are reliably positioned; the estimated coordinate uncertainty is 0.018 Å. His306 ligates the Cu^{2+}, so this is an important structure-function feature that was overlooked. This example is not unique; residues with highly unusual and unrecognized yet real and important deviations in peptide geometry are relatively common.

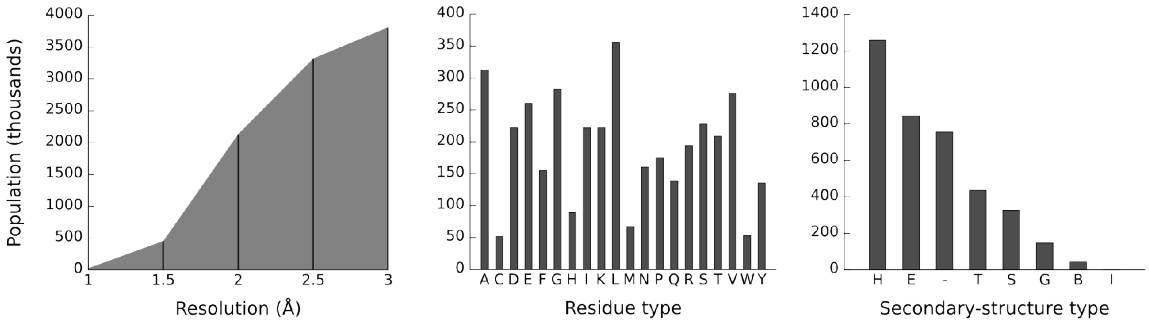
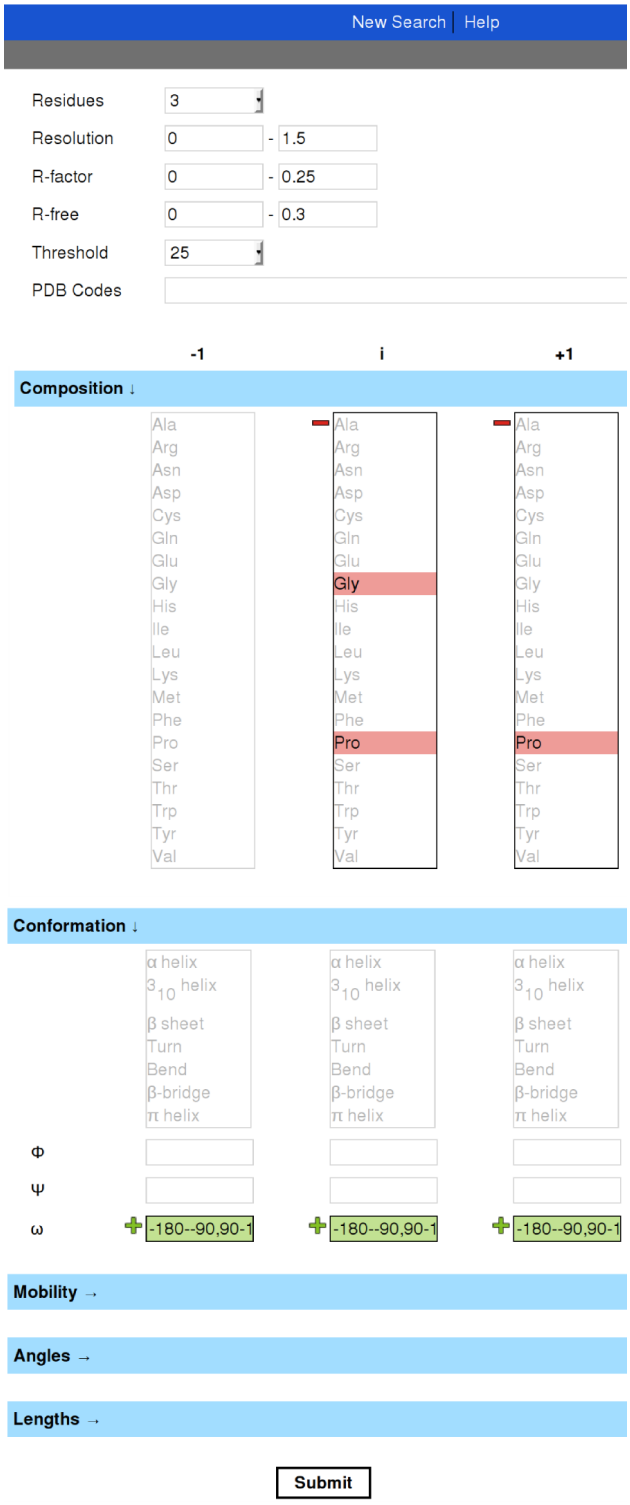


Figure 3.2. Extent and diversity of the database. The residue population of the PGD is shown as a function of resolution (left panel), amino-acid composition (middle panel), and secondary-structure type (right panel). The population as a function of resolution is cumulative. At 1.0 Å resolution or better, the PGD contains 30,256 residues. Secondary-structure types are named as described in the text. The 'I' type (π-helix) bar is too small to be visible, with only 687 observations.

Figure 3.3. Excerpt from a representative query. The query form defines a search for three-residue motifs that do not include Gly, Pro, or prePro residues at 1.5 Å resolution or better. For residue composition, red highlights indicate excluded residues. For flexible-syntax boxes, green highlights indicate valid input, and examples of the flexible query syntax are visible: '<25' to restrict B-factors to be less than 25 $Å^2$, and '-180--90,90-180' for ω, describing a search for *trans* peptides.

Figure 3.3 (continued)



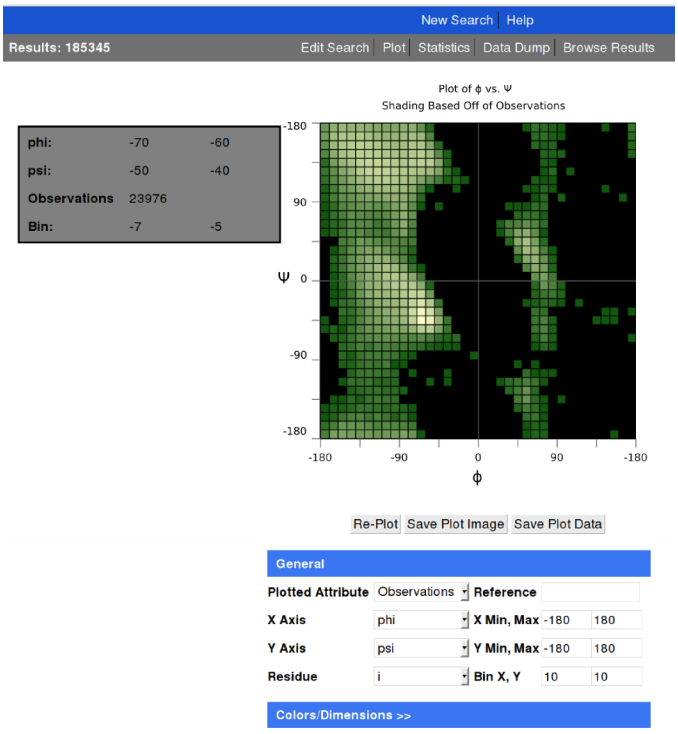


Figure 3.4. Excerpt from a representative output. The Ramachandran plot shows results of a search for three-residue motifs that do not include Gly, Pro, or prePro residues at 1.5 Å resolution or better with other settings left at their defaults. Coloration of the plot in green indicates the observation density in each bin, from low (dark) to high (light). The gray popup box on the left gives information for the pixel over which the cursor is placed. It is one of the most highly populated bins in the α region. The total result count is visible at the left edge of the top navigation bar.

Chapter 4

Conformation dependence of backbone geometry in proteins

Donald S. Berkholz, Maxim V. Shapovalov, Roland L. Dunbrack, Jr.,
and P. Andrew Karplus

Accepted to *Structure*

Summary

Protein structure determination and predictive modeling have long been guided by the paradigm that the peptide backbone has a single, context-independent ideal geometry. Both quantum-mechanics calculations and empirical analyses have shown this is an incorrect simplification in that backbone covalent geometry actually varies systematically as a function of the Φ and Ψ backbone dihedral angles. Here, we use a nonredundant set of ultrahigh-resolution protein structures to define these conformation-dependent variations. The trends have a rational, structural basis that can be explained by avoidance of atomic clashes or optimization of favorable electrostatic interactions. To facilitate adoption of this new paradigm, we have created a conformation-dependent library of covalent bond lengths and bond angles and shown that it has improved accuracy over existing methods without any additional variables to optimize. Protein structures derived both from crystallographic refinement and predictive modeling both stand to benefit from incorporation of the new paradigm.

Introduction

Structural details at the 0.1 Å scale guide our understanding of enzyme catalysis, how mutations cause disease, and what makes a good inhibitor and potential drug. Since the work of Pauling and Corey (1951), protein model building at all levels has been guided by the assumption that the peptide backbone has a certain ideal geometry independent of context (Figure 4.1). This paradigm underlies the restraints used to guide protein structure refinement (e.g., Evans, 2007) and is also the basis of the rigid-geometry approximation used to simplify model building in the most successful structure-prediction packages such as Rosetta and I-TASSER (Rohl et al., 2004; Zhang, 2009). The rigid-geometry approximation uses fixed bond lengths and angles, leaving torsion angles as the only variables needed to define the structure. Ideal target values for the peptide backbone have varied little over the years, and a set of values most recently updated in 1999 (EH; Engh

and Huber, 1991; Engh and Huber, 2001) is commonly used (Figure 4.1).

Experimentally derived crystal structures at all but the highest resolutions reflect the influence of the single-value ideal-geometry paradigm that is applied in the form of geometric restraints. However, strong evidence exists that this paradigm is flawed. Quantum-mechanics calculations and empirical analyses of high-resolution protein structures from over a decade ago suggested that the concept of a single, context-independent ideal value for backbone bond angles and lengths was wrong (Schäfer et al., 1995; Karplus, 1996). Instead, both approaches showed that backbone covalent geometry varies systematically as a function of the conformation of the backbone torsion angles. The systematic conformation dependence of ideal geometry was most notable for the N-C_α-C bond angle ($\angle NC_\alpha C$) that varied by 8.8°, from 105.7° to 114.5° (Karplus, 1996). Similarly, systematic distortions of geometry are known to occur for classically disallowed but experimentally observed conformations (e.g., Gunasekaran 1996, Ramakrishnan 2007). And finally, particularly intriguing has been the observation that at increasingly higher resolution, protein structures are in progressively worse agreement with the supposedly "ideal" values (e.g., Longhi et al., 1998). This observation resulted in a recent literature debate about how to adjust the target values used for geometric restraints and how heavily to weight them (Jaskolski et al., 2007a; Tickle, 2007; Jaskolski et al., 2007b; Stec, 2007). We contributed to this debate with the suggestion that the root of the problem is not simply a matter of incorrect ideal target values or weights but instead is a matter of an incorrect paradigm in that ideal geometry should be a function, not a single value (Karplus et al., 2008).

With the explosion of protein structures solved at 1.0 Å resolution or better, the time is ripe to extend the earlier analysis (Karplus, 1996) and more accurately determine the nature and extent of the systematic variations of peptide geometry with conformation. To accomplish this, we created a nonredundant database of atomic-resolution structures that has nearly 20,000 residues. Here, we use this database to analyze conformation-dependent trends in backbone geometry in all bond angles and lengths. We also show that

accounting for these trends has the potential to improve both crystallographic refinement and homology modeling.

Results and Discussion

Data Source and Analysis Strategy

To accurately characterize the nature and extent of conformation-dependent variations in geometry, we used a data set of 16,682 well-defined three-residue segments from 108 diverse protein chains determined at 1.0 Å resolution or better (see Experimental Procedures). A three-residue segment includes all of the atoms in two complete peptide units, and the data set included the bond lengths and bond angles for the peptide units uniquely identified by whether they mostly involve atoms from residue -1, 0, or +1 in the three-residue segment (Figure 4.1). Based on previous work (Karplus, 1996) indicating distinct geometric behavior of Gly, Pro, the β-branched residues Ile and Val (Thr behaves more like a general residue because of stabilizing sidechain-backbone hydrogen bonds) and residues preceding proline (prePro), we carried out separate statistical analyses for those five groups. The data set used here included 1,379 Gly, 639 Pro, 511 general prePro (644 before exclusion of Gly/Pro/Ile/Val), 1,822 Ile/Val, and 10,921 general residues (the 16 other residue types taken together). All prePro residues are excluded from the other classes. As seen in Figure 4.2, these residues were distributed in Φ,Ψ as has been seen for many well-filtered data sets (Karplus, 1996; Kleywegt and Jones, 1996, Lovell et al., 2003). Figure 4.2 also provides the shorthand nomenclature we will use for certain regions of the Ramachandran plot.

We analyzed these data to visualize and to document the Φ,Ψ-dependent variations in bond lengths and angles. Our approach was to use kernel-regression methods to smooth the data and to produce continuously variable functions for each parameter (see Experimental Procedures). The figures and tables in this paper are based

on the kernel-regression analysis and only include regions of the Ramachandran plot having an observation density of at least 0.03 residues/degree^2 (i.e., 3 residues in a 10° × 10° area) and a finite standard error of the mean.

Ubiquitous, Systematic, Φ,Ψ-Dependent Variations Exist in Peptide Geometry

Bond angles. The data reveal that for general residues, all 15 bond angles in the two peptides adjacent to the central residue vary systematically with Φ and Ψ (Figure 4.5 and Table 4.1). The most prominent observation is that the variations do not occur only in rare outlier conformations, but they occur throughout even the most populated areas of the plot for all residue types (Figure 4.5-4.9). Consistent with the lower-resolution analysis (Karplus, 1996), ∠NC$_\alpha$C varies the most (6.5°), but four other angles also vary by ≥5°. An important difference from the results of the earlier study is that the conformation-dependent standard deviations of the bond angles are about half what was seen previously (Karplus, 1996), ranging from 1.3°-1.8° (Table 1). These are also substantially smaller than the standard deviations of ~2.5° used for the single ideal values defined by Engh and Huber (1991) based on small-molecule structures. It is notable that ultrahigh-resolution crystal structures are generally refined using geometric restraints that do not match the local averages, so the narrow (small σ) distributions cannot be an artifact of the restraints used. Interestingly, the variations in the averages are 2-4 times the standard deviations (Table 1), implying that current modeling restraints would work to wrongly pull angles away from their actual optimal values in many regions. Dramatically, the distributions at the extremes can even be completely non-overlapping because of the small standard deviations (Figure 4.10). The standard errors of the Φ,Ψ-dependent means (i.e., σ/√N) for bond angles are less than 0.5° in nearly all regions and typically less than 0.2° in the highly populated regions (Figures 4.11-4.15)—however, errors should be considered when examining averages for the lowest-populated edges and other regions, such as the prePro region for general residues. In comparison, the 2°-7° ranges seen for the expected

values are 10-50 times greater than their uncertainties. This shows that the variations are well-determined and backbone geometry in no way obeys the single ideal value paradigm.

Bond lengths. In the 1996 study, the resolution of the data did not allow reliable visualization of bond-length variations. Here at atomic resolution, systematic Φ,Ψ-dependent trends are now visible in bond lengths (Figure 4.16) but the variation ranges (0.01 Å-0.02 Å) are only on par with the standard deviations (0.012 Å-0.016 Å), meaning the distributions are highly overlapping. The standard errors of the mean are smaller (~0.002 Å), so the variations in the means seen are nevertheless significant (~10-fold larger). Given that the standard deviations are on par with the expected coordinate accuracy, we hypothesize that the true underlying bond lengths are distributed more narrowly and thus will require still higher resolution analyses to determine accurately. Because of this limitation and the expectation that, because of the very small distances involved, the bond-length variations will have little impact on modeling accuracy, we will not further describe the bond-length trends here. Nevertheless, we suspect the variations involved will be chemically informative (e.g., Esposito et al., 2000; Figure 4.16).

Variations are Correlated with Local Interactions

Comparison of conformation-dependent trends across the two sequential peptide units reveals that the trends are largely locally influenced. For each of the seven angles associated with the central residue, the range is larger than the range for the same angle associated with the previous or subsequent residue (Table 1). For instance, $\angle N_{-1}C_{\alpha-1}C_{-1}$ and $\angle N_{+1}C_{\alpha+1}C_{+1}$ have ranges of 5.5° and 3.0°, whereas $\angle NC_{\alpha}C$ has a range of 6.5°. This implies that the angles in Table 1 associated with residues -1 and +1 show highly local effects, being more influenced by the Φ,Ψ values of their respective residues than the Φ,Ψ values of residue 0 (the central residue). For modeling purposes, it makes sense to assign the "ideal" target values for all angles based only on Φ,Ψ of the central residue.

Furthermore, among these seven angles, additional evidence of the dominance of local effects is seen as each angle is influenced mostly by the single closest torsion angle, whether it is Φ or Ψ. Starting at the N-terminal end, $\angle C_{-1}NC_{\alpha}$ is heavily Φ-dependent as is seen in the vertical pattern of variation, then the C_{α}-centered angles are a mixture, displaying diagonal patterning, and the angles at the C-terminal end, such as $\angle C_{\alpha}CN_{+1}$, have Ψ-dependent horizontal patterning. Even among the C_{α}-centered angles, $\angle NC_{\alpha}C_{\beta}$ shows enhanced dependence on Φ and $\angle C_{\beta}C_{\alpha}C$ shows enhanced dependence on Ψ. This extreme locality agrees with much prior work noting that local steric interactions are critical factors in determining observed conformational and secondary-structure preferences (e.g., Dunbrack and Karplus, 1994; Baldwin and Rose, 1999).

Comparison of Trends with Quantum Mechanics

As noted in the introduction, quantum-mechanical (QM) calculations of isolated alanine peptides (Jiang et al., 1997; Yu et al., 2001) also produce conformation-dependent trends in bond angles and bond lengths. The QM calculations are computationally intensive and they have only been carried out at 30° resolution in Φ,Ψ (Jiang et al., 1997; Yu et al., 2001), making detailed features of the trends unavailable. Beyond a certain level, the method and basis set used in QM calculations is unimportant to this analysis because they produce trends on the same scale with a nearly constant offset (Yu et al., 2001). As was reported by Karplus (1996), the QM results have similar trends, but now it is apparent that QM results show larger deviations, ranging farther both positively and negatively than experimental protein structures. For example, the empirical deviations from the central value for $\angle NC_{\alpha}C$ are roughly 70% of the calculated deviations. Additionally, QM calculations show serious discrepancies in some less populated regions, such as a false global maximum for $\angle O_{-1}C_{-1}N$ in Lδ (Figure 4.2). The mis-scaling seen in QM-calculated angles has been suggested by others to be caused by a lack of long-distance structural effects (Jiang et al., 1997; Yu et al., 2001; Feig, 2008). However, if that were the case, comparison of residues in secondary structure versus those in loops should show this same difference, but Karplus (1996) did not see a difference, and here we confirm that

observation (Figures 4.17-4.18). One potential underlying cause is the difference between a protein environment and vacuum rather than a long-distance effect caused by repeating secondary structure, but the reason that calculations in small peptides fail to predict the correct details of conformation-dependent geometry for proteins is uncertain.

Local Variations Make Structural Sense

The bond-angle trends for five classes of residues for all Φ,Ψ possibilities comprise a massive amount of information that cannot be exhaustively described in the context of this article. A survey of the results, however, reveals a general principle that the observed trends in geometry make structural sense in terms of accommodating local steric and electrostatic interactions. In Karplus (1996), the behavior of $\angle NC_\alpha C$ in the well-populated α, β, and δ regions (Figure 4.2) was rationalized in these terms, including the proposal of a π-peptide interaction in the δ region optimized by the opening of $\angle NC_\alpha C$ (see Figure 8 of Karplus, 1996). Instead of rehashing those observations, here we present four illustrative examples of Φ,Ψ regions with significant distortions. The conformations are shown in Figure 4.2, the relevant bond-angle values can be seen in Figure 4.5, and the specific collisions being ameliorated are illustrated in Figure 4.19.

In the Lα/Lδ region (Figure 4.2), non-Gly residues are disfavored because when using single ideal values for bond angles and lengths, there is a close-contact collision between O_{-1} and $C_\beta H$. As Φ increases, this collision becomes worse. The conformation-dependent trends show that these conformations become accessible by a systematic increase in $\angle O_{-1}C_{-1}N$, $\angle C_{-1}NC_\alpha$, and $\angle NC_\alpha C_\beta$ that opens the ring between O_{-1} and C_β. At the extreme tip of the region near (+90°, 0°), these angles open compared to the EH values (Figure 4.1) by 0.4°, 4.3°, and 2.8°, respectively, to increase the $O_{-1}...C_\beta$ distance from 2.59 Å to 2.85 Å. Although this difference and the others described in this section are small in distance, they can make large energetic differences by transforming an unfavorable atomic overlap to a close contact.

The II′ region is adopted by the i+1 residue of type II′ turns, a tight turn with a

hydrogen bond between O_{-1} and $N_{+2}H$. In this conformation, C_β is quite close to both O_{-1} and N_{+1}, which results in this region being unfavorable for nonglycine residues. Under the rigid-geometry approximation, the entire region should be disallowed because of this clash, but distortions in covalent geometry allow it to be accessible. The variations seen in Figure 4.5 show that the distortions relative to EH values (Table 4.1) include a large opening in $\angle C_\beta C_\alpha C$ (5.9°) as well as opening of $\angle C_\alpha C N_{+1}$ (3.3°) to reduce the $C_\beta...N_{+1}$ clash. This also reduces the $O_{-1}...C_\beta$ clash, where the $\angle C_\beta C_\alpha C$ distortion acts like opening jaws to move C_β away from O_{-1}. The extreme bond openings are enabled by a closing of $\angle N C_\alpha C$ (2.5°), $\angle C_\alpha C O$ (1.8°), and $\angle O C N_{+1}$ (2.0°). The $C_\beta...N_{+1}$ distance increases from 2.65 Å to 2.71 Å, and the $O_{-1}...C_\beta$ distance increases from 3.06 Å to 3.09 Å.

Left of the δ region is a Ramachandran-allowed but sparsely populated region. The primary clash is between HN and HN_{+1}. This clash is prevented through a combination of distortions relative to EH values: the dominant increases are in $\angle N C_\alpha C$ (3.5°) and $\angle C_\alpha C N_{+1}$ (2.8°) that both exhibit their extreme values (Figure 4.5), coupled with a decrease in $\angle C_\alpha C O$ (2.0°). The combined effect is to open and twist a nearly planar ring between NH and $N_{+1}H$ to prevent a van der Waals overlap by increasing the $HN...HN_{+1}$ distance from 1.78 Å to 1.92 Å and the $N...N_{+1}$ distance from 2.66 Å to 2.76 Å.

As a final example, we illustrate the importance of treating prePro as a special residue type. Preproline residues are classically disallowed in the α region, yet they are experimentally observed with low populations (Hurley et al., 1992). The primary clash occurs between N and $C_{\delta+1}$ with a secondary clash between $C_\beta H$ and $C_{\delta+1}H$ (Figure 4.19). To alleviate this clash, the Pro ring bends away from the prePro residue through increases in $\angle N C_\alpha C$ (2.0°), $\angle C_\beta C_\alpha C$ (2.4°), and $\angle C_\alpha C N_{+1}$ (3.3°), enabled by decreases in $\angle C_\alpha C O$ (2.3°), $\angle O C N_{+1}$ (2.6°), and $\angle C N_{+1} C_{\alpha+1}$ (3.8°). In comparison to calculations by Hurley et al. (1992) that suggested a single, very large deviation of 8.5° in $\angle C_\beta C_\alpha C$, here we observe that the distortions have diffused across all of the angles between the sterically hindered atoms. These distortions increase the $N...C_{\delta+1}$ distance from 2.65 Å to 2.85 Å

and the $C_{\beta}H...C_{\delta+1}H$ distance from 1.86 Å to 1.90 Å to reduce the van der Waals overlap. $\angle CN_{+1}C_{\delta+1}$ was not included in the database, but we expect it also opens to further alleviate the collision.

A 10°-Resolution Conformation-Dependent Library

With the knowledge of these systematic trends comes the possibility of leveraging them to improve the accuracy of crystallographic refinement and homology modeling. To provide a convenient form in which the documented systematic variations can be used in modeling applications, we created a binned conformation-dependent library (CDL) for distribution. Similar to the approach taken by Karplus (1996), we divided Φ,Ψ space into 1296 10° × 10° bins and calculated the averages and standard deviations for each bin for each of the five residue-type categories (Gly, Pro, prePro, Ile/Val, General). This first-generation CDL (v1.0), available from the authors or at http://proteingeometry.sourceforge.net/, uses a simple precalculated lookup table that accepts conformations and returns the appropriate target value for each bond angle and length. When considering crystallographic refinement and homology modeling, it is important to note that using more accurate CDL values in place of EH values does not increase the number of variable parameters used in the modeling.

Conformation-Dependent Angles are More Accurate

A variety of simple control calculations can be carried out to show that the CDL is an improvement over the single-value paradigm (EH values) and even context-dependent values derived from molecular mechanics (MM) force fields. Because an MM force field allows bond angles and lengths to vary with conformation, it could in theory provide better conformation-dependent values than the empirical approach.

As one simple assessment, we compared how well the $\angle NC_{\alpha}C$ values in a 1.15 Å

ribonuclease structure (PDB code 1rge; Sevcik et al., 1996) matched with EH values, the CDL, and bond-angle values from the structure after minimization using a molecular mechanics force field (see Experimental Procedures). As seen in Figure 4.20, the conformation-dependent library outperforms both the single ideal value and molecular mechanics. Importantly, the CDL produces more angles with very close (<1°) agreement with the reference crystal structure as well as fewer extremely large deviations. In terms of modeling accuracy, there appears to be no downside to using the CDL.

For a more thorough comparison of the CDL with EH values, we compared how well each matched the $\angle NC_\alpha C$ values for the set of protein structures used to generate the CDL, with each protein jackknifed out during its comparison. Averaged over the whole data set, the median deviation from the native bond angles for the EH single-value paradigm was 1.5°/residue and the median deviation for the CDL dropped to 1.1°/residue. This amounts to an improvement of ~25% in $\angle NC_\alpha C$ accuracy relative to the old paradigm.

To understand the impact this difference could have upon protein modeling, coordinates for each jackknifed structure were rebuilt from torsion and bond angles using EH or CDL values. Holmes and Tsai (2004) have shown that the replacement of experimental bond angles with ideal ones while holding Φ and Ψ fixed shifts coordinates by an average of 6 Å (unnormalized by protein length), and this limits model-building accuracy. Here, applying the same approach, we find that the median C_α $RMSD_{100}$ (normalized to the length of a 100-residue protein) from the native structure for EH values was 3.23 Å, and for CDL values it was 2.85 Å. The CDL produced a significant improvement in the C_α $RMSD_{100}$ of ~0.4 Å over the old single-value paradigm.

Potential Applications: Crystallographic Refinement and Homology Modeling

To assess the potential impact of accounting for Φ,Ψ-dependent variations upon X-ray

crystal structures at various resolutions, we evaluated how much the experimental $\angle NC_\alpha C$ values deviated from those in the CDL as a function of resolution (Figure 4.21). To avoid bias, none of the structures used in the survey were used in the generation of the CDL. Analysis of the data shows that for structures solved at near 1 Å resolution, the RMSD of $\angle NC_\alpha C$ from the CDL is ~1.6°. This matches well with the standard deviation seen in the CDL for this angle and serves as an effective validation of the CDL. Additionally, the small standard deviation of the RMSDs at this resolution shows that each of the individual high-resolution structures is well-described by the CDL. Already at a resolution of 1.5 Å, normally considered very high resolution, the match of $\angle NC_\alpha C$ values to the CDL is nearly twice as poor as for the 1.0 Å resolution structures. This loss of accuracy became steadily more pronounced in lower-resolution structures, rising to nearly 4° at 3.0 Å resolution. We conclude that by using the CDL, high-, medium-, and low-resolution structures could all be improved. We suspect that at resolutions worse than 3 Å, the CDL would have less impact because dihedral angles would be less reliable.

To understand the potential benefit of accounting for Φ,Ψ-dependent geometry variations in predictive modeling of protein structure, we carried out a test using the Rosetta modeling program (Rohl et al., 2004). A standard control calculation for homology modeling is to ask how far a crystal structure moves from the experimental structure when minimized by the force field. This provides a lower limit on how accurately a structure can be predicted (e.g., Bradley et al., 2005). For our test, we performed a series of 100 Monte Carlo energy minimizations starting with different random seeds using both native and "ideal" bond geometries for two ultrahigh-resolution protein structures: ribonuclease chain A at 1.15 Å resolution (PDB code 1rge; Sevcik et al., 1996; Figure 4.22) and the PDZ domain of syntenin at 0.73 Å (PDB code 1r6j; Kang et al., 2004; data not shown). "Native" geometry refers to the bond lengths and angles as seen in the crystal structure. As seen in Figure 4.22A, minimizations using the "native" bond geometry instead of the idealized geometry resulted in better convergence (tighter grouping) and allowed the minimized structure to be about 30% closer to the true structure (~0.6 Å vs ~0.9 Å). One notable feature is that the improved behavior occurs

despite the force field's optimization based on the traditional "ideal" geometry values. We conclude from this that the use of the rigid-geometry approximation with standard single ideal values limits modeling accuracy substantially. Thus, it is worthwhile to adapt modeling programs to account for the new conformation-dependent geometry paradigm.

To pinpoint exactly where in the structure the improvements occurred, we calculated the deviations between the crystal structure and the energy-minimized structures using native versus ideal geometry (Figure 4.22B). As an indication of the variation that can occur for this protein in two environments, the deviations with chain B from the same structure is also shown. The largest differences between "ideal" and "native" geometry occur in loops rather than regular secondary structure (Figure 4.22B). This meets the expectation that the largest systematic deviations from single ideal values should occur in parts of the protein with less observed, more diverse Φ,Ψ values. This result was expected because the most highly populated regions dominate the global averages, resulting in the illusion of single ideal values assumed in EH, whereas more conformationally diverse, less populated regions contribute less to the global average. Importantly, the two loops that were highly improved by using experimental geometry are at the active site of the protein, so the accuracy with which they are modeled would significantly degrade the ability of this mock homology model to provide insight.

Outlook

The studies here show that the dependence of backbone geometry on conformation is unmistakably real, significant, systematic, and has a rational structural basis. These systematic distortions in covalent geometry additionally explain how some conformations are accessible to amino-acid residues despite being theoretically disallowed by modeling based on single ideal values for backbone geometry. The conformation-dependent library we derived from the database represents the first step toward implementing the new paradigm of “ideal-geometry functions.” With their much-improved agreement to ultrahigh-resolution crystal structures, the ideal-geometry functions provide an

intellectually satisfying resolution to the debate among crystallographers as to what ideal values should be used during refinement. Also, because the ideal-geometry functions captured in the CDL are simply a highly enlarged set of immutable ideal values, their use in the place of single ideal values represents no increase in algorithmic complexity. Use of the CDL thus offers the potential for improved modeling accuracy in a wide variety of experimentally based and predictive modeling applications without increasing the risk of overfitting.

Experimental Procedures

Data Set Construction

A Protein Geometry Database being developed in our laboratory (DSB, PAK, unpublished) was used to generate our data set of atomic-resolution geometry information. To optimally analyze Φ,Ψ-dependent geometry trends, the data set must be large but also have independent and accurate information about geometry. The plethora of new atomic-resolution protein structures allowed us to use stringent criteria for independence and accuracy, yet still have sufficient observations for reasonable statistics. To ensure independence, we used the PDBSelect (Hobohm and Sander, 1994) list from March 2006 to choose protein chains with less than 90% sequence identity to any other chain in the data set. To ensure high accuracy, we only used structures determined at 1.0 Å or better. At this resolution, we estimate Φ and Ψ dihedral angle accuracy to be better than 3° (see next paragraph). Also, as in Karplus (1996), to ensure that individual residues used were well-resolved, we required that all residues in a five-residue segment were all well-ordered, having B-factors <25 $Å^2$ for the mainchain average, the sidechain average, and C_γ, and alternative conformations were discarded.

To estimate the experimental uncertainty in Φ and Ψ for 1 Å resolution structures, we chose to use a straightforward, empirical approach—randomize and re-refine a test

structure multiple times and then examine the spread of the dihedral angles among the structures. Specifically, we applied 10 coordinate randomizations with a mean shift of 0.2 Å using phenix.pdbtools (Adams et al., 2002) to the coordinates of glutathione reductase at 0.95 Å resolution (PDB ID: 3dk9; Berkholz et al., 2008) and re-refined each in SHELX (Sheldrick, 2007). Dihedral RMSDs for the vast majority of residues were between 1°-2°. The 90th percentile of the per-residue RMSDs in both Φ and Ψ was 2.2°, and the RMSD values of the per-residue RMSDs for Φ and Ψ were 1.7° and 2.4°, respectively.

Kernel Regression for the Bond Lengths and Bond Angles

The data value of any structural parameter *a* of residue *i* (or of the left or right neighbor of residue *i*) may be expressed:

$$a_i = m(\phi_i, \psi_i) + v^{\frac{1}{2}}(\phi_i, \psi_i)\varepsilon_i$$

where *m* is a regression function, and ε are random Gaussian-distributed errors with mean 0 and σ=1:

$$m(x,y) = E(a|\phi = x, \psi = y)$$
$$v(x,y) = \mathrm{Var}(a|\phi = x, \psi = y)$$

In these expressions, *E* is the expectation value of *a* and Var is the variance of *a*.

To obtain an estimate of *m* and ν, we use a zeroth-order or Nadaraya-Watson kernel regression (Nadaraya, 1964) by summing over *N* data points:

$$\hat{m}(\phi,\psi) = \frac{\sum_{i=1,N} K(\phi_i - \phi, \psi_i - \psi) a_i}{\sum_{i=1,N} K(\phi_i - \phi, \psi_i - \psi)}$$

$$\hat{v}(\phi,\psi) = \frac{\sum_{i=1,N} K(\phi_i - \phi, \psi_i - \psi)\left(a_i - \hat{m}(\phi_i,\psi_i)\right)^2}{\sum_{i=1,N} K(\phi_i - \phi, \psi_i - \psi)}$$

The latter is Var(a|Φ,Ψ), an estimate of the heteroscedastic data variance as a function of Φ and Ψ.

The functions K are kernels that weight the data points based on how far away they are from the query Φ,Ψ value. Since Φ and Ψ are angles, we use the product of two von Mises kernel functions (Mardia and Zamroch, 1975)

$$K(\phi - \phi_i, \psi - \psi_i) = \frac{1}{4\pi^2}\frac{1}{\left(I_0(\kappa)\right)^2}\exp\left(\kappa\left(\cos(\phi_i - \phi) + \cos(\psi_i - \psi)\right)\right)$$

At large values of κ, these functions behave very similarly to Gaussian distributions, except that they are periodic. We investigated several values of κ and plotted the resulting regressions as a function of Φ and Ψ. We empirically chose a value of κ=50 to produce distributions that varied smoothly with Φ and Ψ in a reasonable way.

The Φ,Ψ map is not uniformly populated by data points, each of them representing a single residue backbone conformation. Therefore, for the unpopulated regions of the map, the kernel regression analysis generates non-local estimates of m and ν. A query point (Φ,Ψ) in which we estimate expectation and variance values of a, can be surrounded by an effective radius r, equal to half of a bandwidth, b of the kernel function, K. We can count the effective number of data points, N_{eff} within the radius, r around any query point. These points will have an impact on the estimated local values of m and ν .

We define the bandwidth, $b(\kappa)$ as a diameter of the circle centered on the query point (Φ_0,Ψ_0) within which the von Mises kernel function integrates to 68.2% (the value of integral of the normal distribution PDF within one standard deviation from its center):

$$\int_{\sqrt{\varphi^2+\psi^2}<b(\kappa)} K(\varphi-\varphi_0,\psi-\psi_0)\,d\varphi\,d\psi = 0.682$$

The bandwidth of the von Mises kernel at κ=50 is approximately 16°.

In order to depict the trends of $\hat{m}(\phi,\psi)$ and $\hat{v}(\phi,\psi)$, we only plot their estimates at Φ,Ψ grid points where $N_{eff}(\Phi_0,\Psi_0) \geq 3$ within a circle with a diameter equal to the bandwidth $b(\kappa=50) = 16°$.

In the sparsely populated areas of the Φ,Ψ map the threshold of at least 3 data points within the effective bandwidth may lead to estimates with high standard errors of mean (SEM). We calculated an estimate of SEM, as

$$SEM(a|\phi,\psi) = \sqrt{\frac{v(\phi,\psi)}{N_{eff}(\phi,\psi)}}$$

It is very important to analyze the trends of *m* and ν as a function of Φ,Ψ together with SEM(*a*|Φ,Ψ). The values of SEM will indicate the significance of the trend in the more sparsely populated areas.

Creation of the Binned Conformation-Dependent Library

To create a binned conformation-dependent library (CDL) for each residue class, averages and standard deviations were calculated in 10° × 10° bins in Φ,Ψ. The results were stored in a set of files, one per residue class. Python scripts provide an interface to the CDL, allowing easy retrieval of the conformation-dependent values when given a residue name and conformation. Additional tools building upon this simple interface are also part of the distributed code, including a tool that will compare the bond angles and lengths in any PDB coordinate file with CDL values, EH values, or another PDB coordinate file. The CDL and accessory tools are available under an open-source license

from http://proteingeometry.sourceforge.net/.

Molecular Mechanics Calculations

Molecular mechanics-derived context-dependent values for bond angles for two test cases (PDB codes 1rge (Sevcik et al., 1996) and 1r6j (Kang et al., 2004)) were generated using the following protocol: the structures were minimized in CHARMM (Brooks et al., 1983) using the parm_all22_prot force field with the CMAP correction (MacKerell, 2004) using the GBMV implicit solvent model (Lee et al., 2003). The protocol used cycles of 100 steps of steepest-descent minimization with heavy-atom restraints of 5, 3, 1 and 0 * atomic mass kcal/mol/Å2. Following the last cycle (which had no restraints), 1000 steps of adopted basis Newton-Raphson minimization were performed, and the typical gradient RMS was about 0.05 kcal/mol/Å.

CDL Assessments

Building Ideal Models and Analysis of Nonbonded Interactions

Ideal peptides with EH or CDL backbone geometry were built using PyRosetta (http://graylab.jhu.edu/~sid/pyrosetta/), Python bindings to Rosetta (Rohl et al., 2004). To account for the length dependence of RMSD calculations (e.g., Holmes and Tsai, 2004), we linearly rescaled all RMSDs to those of 100-residue proteins using the EH RMSDs and the assumption that RMSDs intersect the origin. Based on the linear fit of EH RMSDs versus length produced, we calculated a scaling factor of (0.0332519/100) / (0.0332519/length). To understand the structural basis of variations between these theoretical peptides, van der Waals clashes were visually analyzed using the Coot (Emsley and Cowtan, 2004) interface to MolProbity (Davis et al., 2007).

Crystal Structure ∠$NC_{\alpha}C$ Angles

Nonredundant structures with a 25% sequence-identity threshold were taken from PDBSelect (Hobohm and Sander, 1994). Among these, 50 structures were selected from each of five resolution ranges: 1.0-1.1 Å, 1.5-1.6 Å, 2.0-2.1 Å, 2.5-2.6 Å, 3.0-3.1 Å. For each residue in these structures, we then calculated the difference in the observed ∠$NC_{\alpha}C$ and the CDL value. These were used to calculate the per-structure RMSDs, which were then used to calculate averages, standard deviations, and standard errors of the mean for each of the five resolution shells.

Acknowledgements

We thank Charles L. Brooks III (University of Michigan) for performing the molecular-mechanics minimizations used in this study. We additionally thank the David Baker lab (University of Washington at Seattle), in particular Srivatsan Raman, James Thompson, and Elizabeth Kellogg, for their help with Rosetta. We thank Jeffrey Gray (Johns Hopkins University) for providing PyRosetta, the Python bindings to Rosetta. We thank Lothar Schäfer (University of Arkansas) for providing a database of QM-calculated dipeptides and an extrapolation program to obtain values for conformation-dependent bond angles and lengths. This work was supported in part by NIH grant R01-GM083136 (to PAK), NSF grant MCB-9982727 (to PAK), and NIH grant P20-GM76222 (to RLD).

References

Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Read, R.J., Sacchettini, J.C., Sauter, N.K., and Terwilliger, T.C. (2002). PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr. D. Biol. Crystallogr. *58*, 1948-1954.

Baldwin, R.L., and Rose, G.D. (1999). Is protein folding hierarchic? I. Local structure and peptide folding. Trends Biochem. Sci. *24*, 26-33.

Berkholz, D.S., Faber, H.R., Savvides, S.N., and Karplus, P.A. (2008). Catalytic cycle of human glutathione reductase near 1 A resolution. J. Mol. Biol. *382*, 371-384.

Bradley, P., Misura, K.M.S., and Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. Science *309*, 1868-1871.

Brooks, B.R., Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. *4*, 187-217.

Corey, R.B., and Donohue, J. (1950). Interatomic distances and bond angles in the polypeptide chain of proteins. J. Am. Chem. Soc. *72*, 2899-2900.

Dunbrack, R.L., and Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat. Struct. Biol. *1*, 334-340.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D. Biol. Crystallogr. *60*, 2126-2132.

Engh, R.A., and Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallogr. A Found. Crystallogr. *47*, 392-400.

Engh, R.A., and Huber, R. (2001). International Tables for Crystallography.
In International Tables for Crystallography, M.G. Rossmann, and E.
Arnold, eds. (Dordrecht, The Netherlands: Kluwer Academic Publishers), pp. 382-392.

Esposito, L., Vitagliano, L., Zagari, A., and Mazzarella, L. (2000). Experimental evidence for the correlation of bond distances in peptide groups detected in ultrahigh-resolution protein structures. Protein Eng. *13*, 825-828.

Evans, P.R. (2007). An introduction to stereochemical restraints. Acta Crystallogr. D. Biol. Crystallogr. *63*, 58-61.

Feig, M. (2008). Is alanine dipeptide a good model for representing the torsional preferences of protein backbones? J. Chem. Theory Comput. *4*, 1555-1564.

Gunasekaran, K., Ramakrishnan, C., and Balaram, P. (1996). Disallowed Ramachandran conformations of amino acid residues in protein structures. J. Mol. Biol. *264*, 191-198.

Hobohm, U., and Sander, C. (1994). Enlarged representative set of protein structures. Protein Sci. *3*, 522-524.

Hollingsworth, S.A., Berkholz, D.S., and Karplus, P.A. (2009). On the occurrence of linear groups in proteins. Protein Sci. *18*, 1321-1325.

Holmes, J.B., and Tsai, J. (2004). Some fundamental aspects of building protein structures from fragment libraries. Protein Sci. *13*, 1636–1650.

Hurley, J.H., Mason, D.A., and Matthews, B.W. (1992). Flexible-geometry conformational energy maps for the amino acid residue preceding a proline. Biopolymers *32*, 1443-1446.

Jaskolski, M., Gilski, M., Dauter, Z., and Wlodawer, A. (2007a). Numerology versus reality: a voice in a recent dispute. Acta Crystallogr. D. Biol. Crystallogr. *63*, 1282-1283.

Jaskolski, M., Gilski, M., Dauter, Z., and Wlodawer, A. (2007b). Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? Acta Crystallogr. D. Biol. Crystallogr. *63*, 611-620.

Jiang, X., Yu, C., Cao, M., Newton, S.Q., Paulus, E.F., and Schäfer, L. (1997). ø/ψ-Torsional dependence of peptide backbone bond-lengths and bond-angles: comparison of crystallographic and calculated parameters. J. Mol. Struct. *403*, 83-93.

Kang, B.S., Devedjiev, Y., Derewenda, U., and Derewenda, Z.S. (2004). The PDZ2 domain of syntenin at ultra-high resolution: bridging the gap between macromolecular and small molecule crystallography. J. Mol. Biol. *338*, 483-493.

Karplus, P.A. (1996). Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci. *5*, 1406-1420.

Karplus, P., Shapovalov, M., Dunbrack Jr, R., and Berkholz, D.S. (2008). A forward-looking suggestion for resolving the stereochemical restraints debate: ideal geometry functions. Acta Crystallogr. D. Biol. Crystallogr. *64*, 335-336.

Kleywegt, G.J., and Jones, T.A. (1996). Phi/psi-chology: Ramachandran revisited. Structure *4*, 1395-1400.

Laskowski, R.A., Chistyakov, V.V., and Thornton, J.M. (2005). PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. Nuc. Acids Res. *33*, D266-D268.

Lee, M.S., Feig, M., Salsbury, F.R., and Brooks, C.L. (2003). New analytic

approximation to the standard molecular volume definition and its application to generalized Born calculations. J. Comput. Chem. *24*, 1348-1356.

Longhi, S., Czjzek, M., and Cambillau, C. (1998). Messages from ultrahigh resolution crystal structures. Curr. Opin. Struct. Biol. *8*, 730-737.

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: ϕ, ψ and Cβ deviation. Proteins: Struct. Func. Genet. *50*, 437-450.

Mackerell, A.D. (2004). Empirical force fields for biological macromolecules: overview and issues. J. Comput. Chem. *25*, 1584-1604.

Mardia, K.V., and Zemroch, P.J. (1975). Algorithm AS 86: The Von Mises distribution function. Applied Statistics *24*, 268-272.

Naradaya, E. (1964). On estimating regression. Theory of Probability and its Applications *9*, 141–142.

Pauling, L., Corey, R.B., and Branson, H.R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. USA *37*, 205-211.

Ramakrishnan, C., Lakshmi, B., Kurien, A., Devipriya, D., and Srinivasan, N. (2007). Structural compromise of disallowed conformations in peptide and protein structures. Protein Pept. Lett. *14*, 672-682.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., and Baker, D. (2004). Protein structure prediction using Rosetta. Methods Enzymol. *383*, 66-93.

Schäfer, L., and Cao, M. (1995). Predictions of protein backbone bond distances and angles from first principles. J. Mol. Struct. *333*, 201-208.

Sevcik, J., Dauter, Z., Lamzin, V.S., and Wilson, K.S. (1996). Ribonuclease from Streptomyces aureofaciens at Atomic Resolution. Acta Crystallogr. D. Biol. Crystallogr. *52*, 327-344.

Sheldrick, G.M. (2007). A short history of SHELX. Acta Crystallogr. A. Found. Crystallogr. *64*, 112-122.

Stec, B. (2007). Comment on Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? by Jaskolski, Gilski, Dauter and Wlodawer (2007). Acta Crystallogr. D. Biol. Crystallogr. *63*, 1113-1114.

Tickle, I.J. (2007). Experimental determination of optimal root-mean-square deviations of macromolecular bond lengths and angles from their restrained ideal values. Acta

Crystallogr. D. Biol. Crystallogr. *63*, 1274-1281.

Yu, C.H., Norman, M.A., Schäfer, L., Ramek, M., Peeters, A., and van Alsenoy, C. (2001). Ab initio conformational analysis of N-formyl L-alanine amide including electron correlation. J. Mol. Struct. *567*, 361-374.

Zhang, Y. (2009). Protein structure prediction: when is it useful? Curr. Opin. Struct. Biol. *19*, 145-155.

Table 4.1. Expected and observed ranges for peptide geometries[a]

Residue	Angle	EH[b]	Min(CDL)	Max(CDL)	**Range**	σ(EH)	σ(CDL)[c]
-1	∠$NC_\alpha C$	111.0	107.0	112.5	**5.5**		
	∠$C_\beta C_\alpha C$	110.6	108.5	111.5	**3.0**		
	∠$C_\alpha CO$	120.1	119.3	121.2	**1.9**		
	∠$C_\alpha CN_{+1}$	117.2	115.3	117.6	**2.3**		
	∠OCN_{+1}	122.7	121.8	123.5	**1.7**		
0	∠$C_{-1}NC_\alpha$	121.7	120.0	126.0	**6.0**	1.8	1.7
	∠$NC_\alpha C_\beta$	110.6	109.0	114.0	**5.0**	1.7	1.6
	∠$NC_\alpha C$	111.0	107.5	114.0	**6.5**	2.8	1.5
	∠$C_\beta C_\alpha C$	110.6	109.5	116.0	**6.5**	1.9	1.8
	∠$C_\alpha CO$	120.1	118.5	122.0	**3.5**	1.7	1.3
	∠$C_\alpha CN_{+1}$	117.2	114.5	119.5	**5.0**	2.0	1.3
	∠OCN_{+1}	122.7	121.0	123.5	**2.5**	1.6	1.3
+1	∠$C_{-1}NC_\alpha$	121.7	120.5	122.7	**2.2**		
	∠$NC_\alpha C_\beta$	110.6	109.8	111.2	**1.4**		
	∠$NC_\alpha C$	111.0	109.5	112.5	**3.0**		

[a]All values are in degrees. CDL indicates the conformation-dependent kernel regressions from this work.
[b]Values are from Engh and Huber (2001)
[c]Values are typical for the majority of the plot, although they are greater in the least populated regions. See Figure S5 for details.

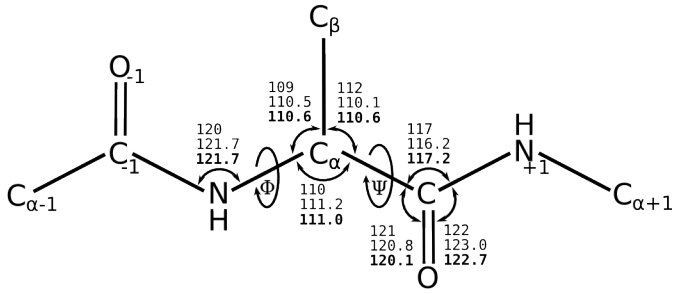


Figure 4.1. Evolution of the ideal values for backbone geometry used in the single-value paradigm. A central residue (residue 0) is shown with atoms from residues -1 and +1 that contribute to its two adjacent peptide units. For each of the seven bond angles associated with residue 0, three ideal values from earlier work are shown from oldest (top) to most recent (bottom). They are from Corey and Donohue (1950), Engh and Huber (1991), and Engh and Huber (2001). Most refinement and modeling programs use one of the Engh and Huber sets or a slight variation on them. Rotatable bonds defining the backbone torsion angles Φ and Ψ are indicated. Figure created with Inkscape.

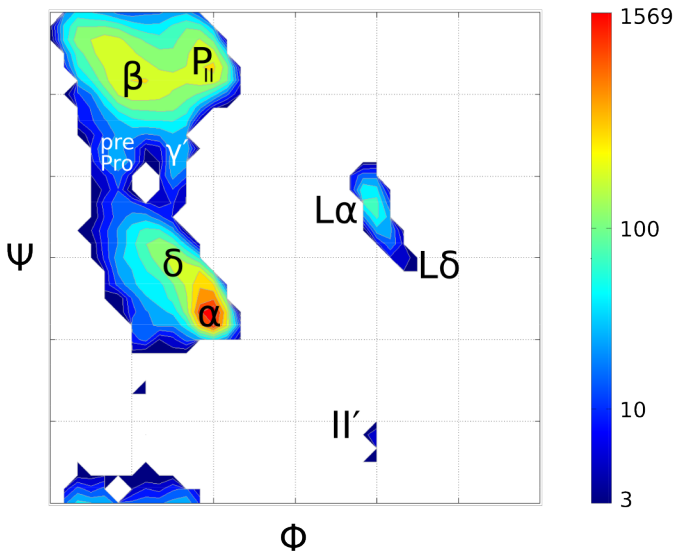


Figure 4.2. Protein backbone conformations of non-Gly residues. This Ramachandran plot is colored by empirical observation density in atomic-resolution proteins. Labels indicate regions of particular interest (Karplus, 1996; Lovell et al., 2003; Hollingsworth et al., 2009). Coloring uses a logarithmic function to allow lower- and higher-population regions to be seen simultaneously. Observation density was calculated using kernel regressions (see Experimental Procedures). Unlabeled versions of this plot and another for only Gly residues are available as Figures 4.3 and 4.4. Figure created with Inkscape.

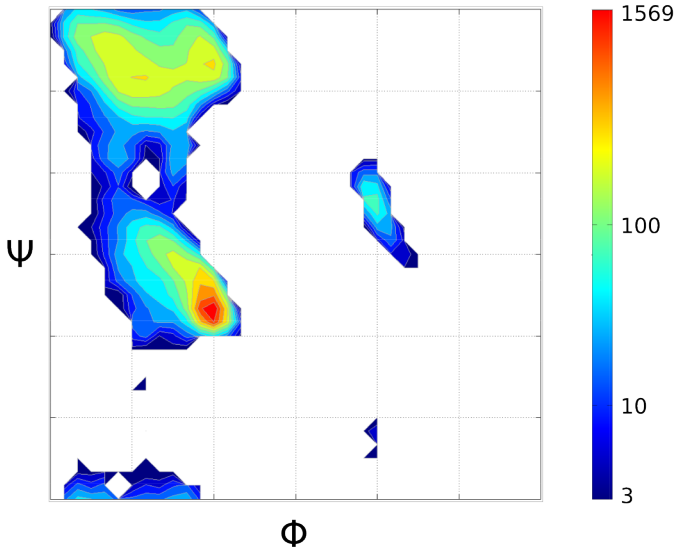


Figure 4.3. Protein backbone conformations of non-Gly residues, unlabeled. This is identical to Figure 4.2 but unlabeled.

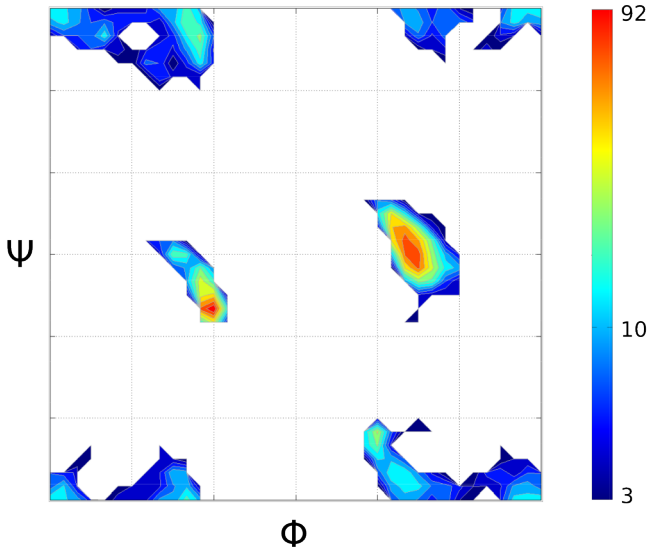
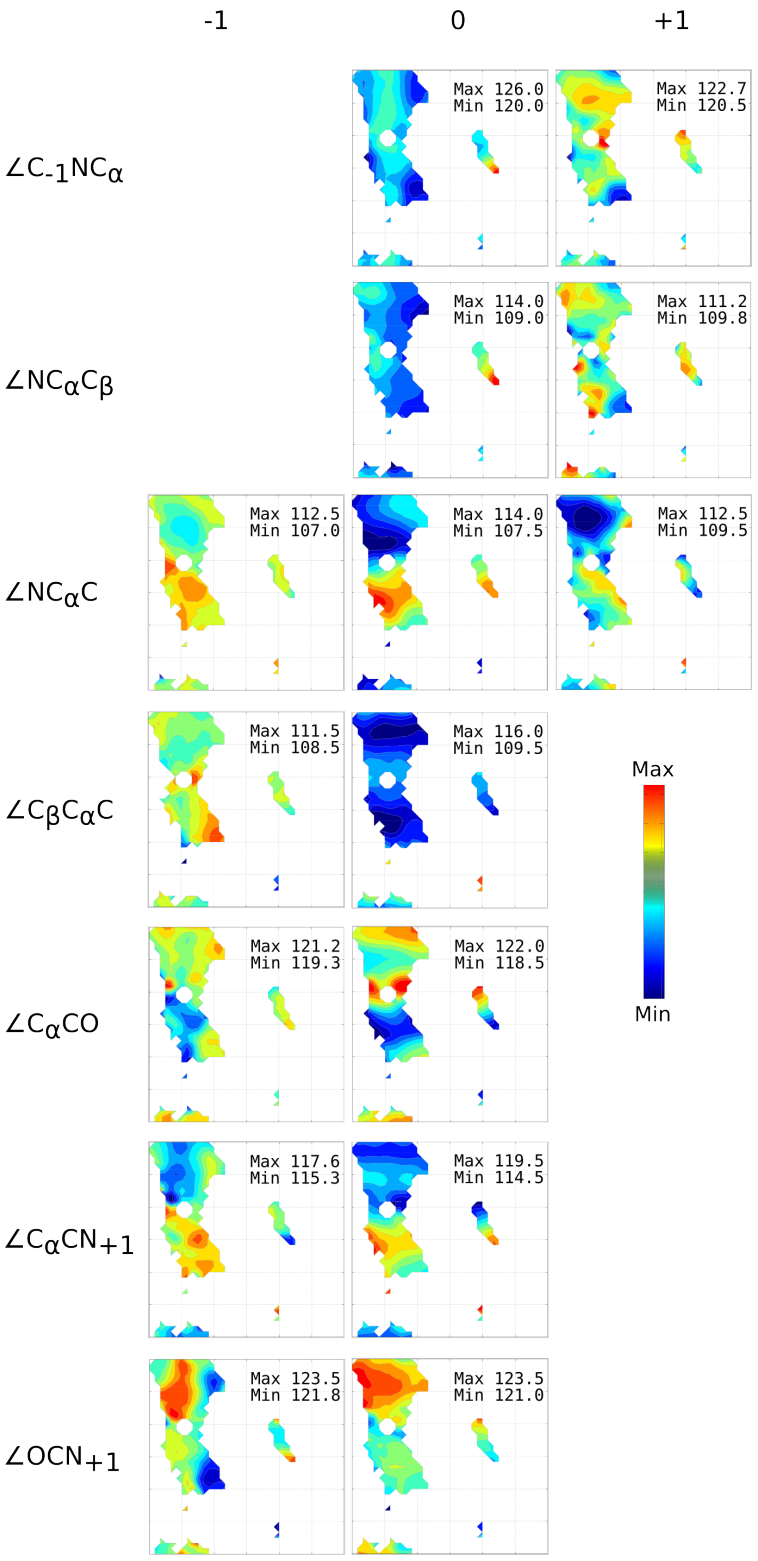


Figure 4.4. Protein backbone conformations of Gly residues. This is otherwise identical to Figure 4.2.

Figure 4.5. Conformation-dependent variation in bond angles of general residues as a function of the Φ,Ψ of the central residue. A Ramachandran plot is shown for each backbone bond angle in the two peptide units surrounding the central residue of the tripeptide. The seven unique peptide bond angles are associated with either residue -1, 0, or +1 based on which residue contributes at least two atoms to the angle. Φ and Ψ in each plot, however, refer to the central residue, 0. Within each plot, colors indicate averages ranging from the global minimum (blue) to the global maximum (red) as calculated using kernel regressions (see Experimental Procedures). The global minima and maxima are provided in each plot. Figure created with Matlab.

Figure 4.5 (continued)



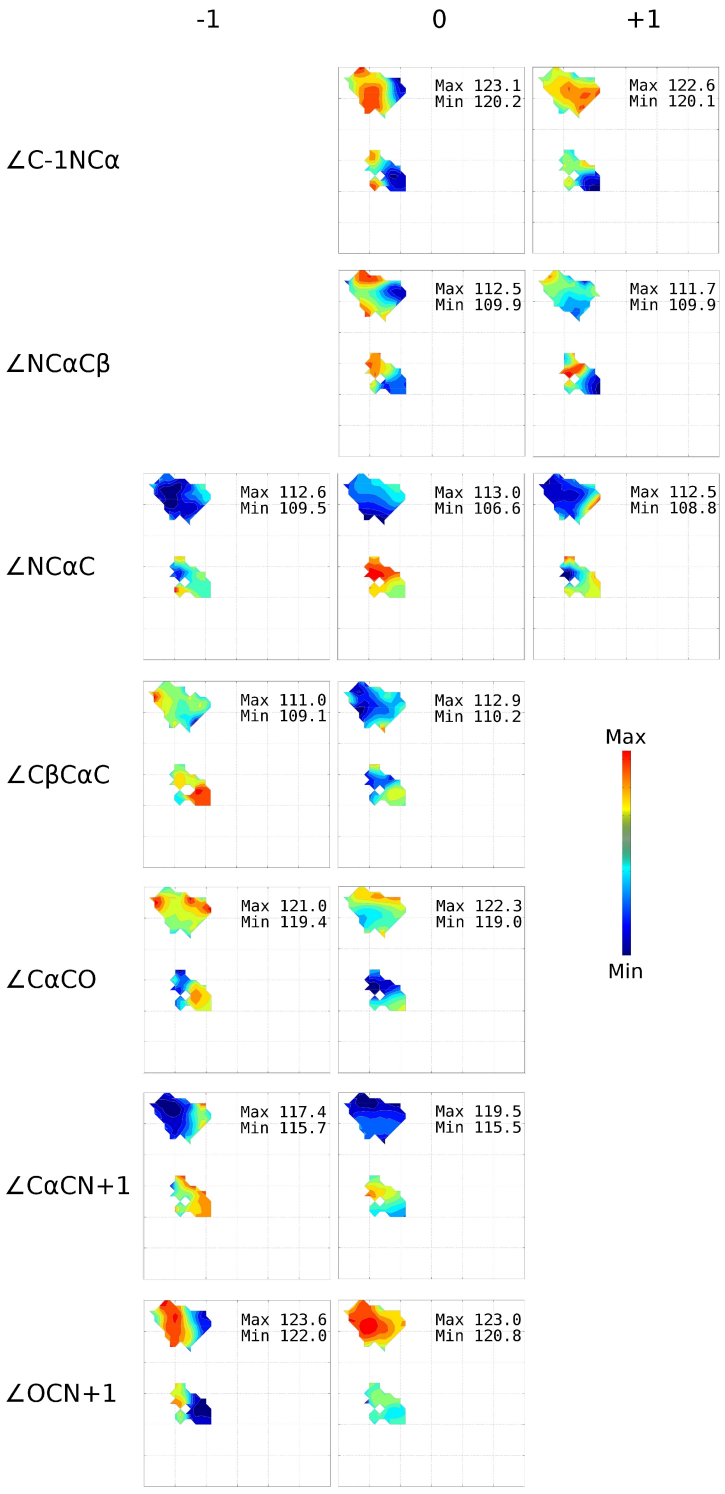


Figure 4.6. Conformation-dependent variation in bond angles of Ile/Val residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 4.5.

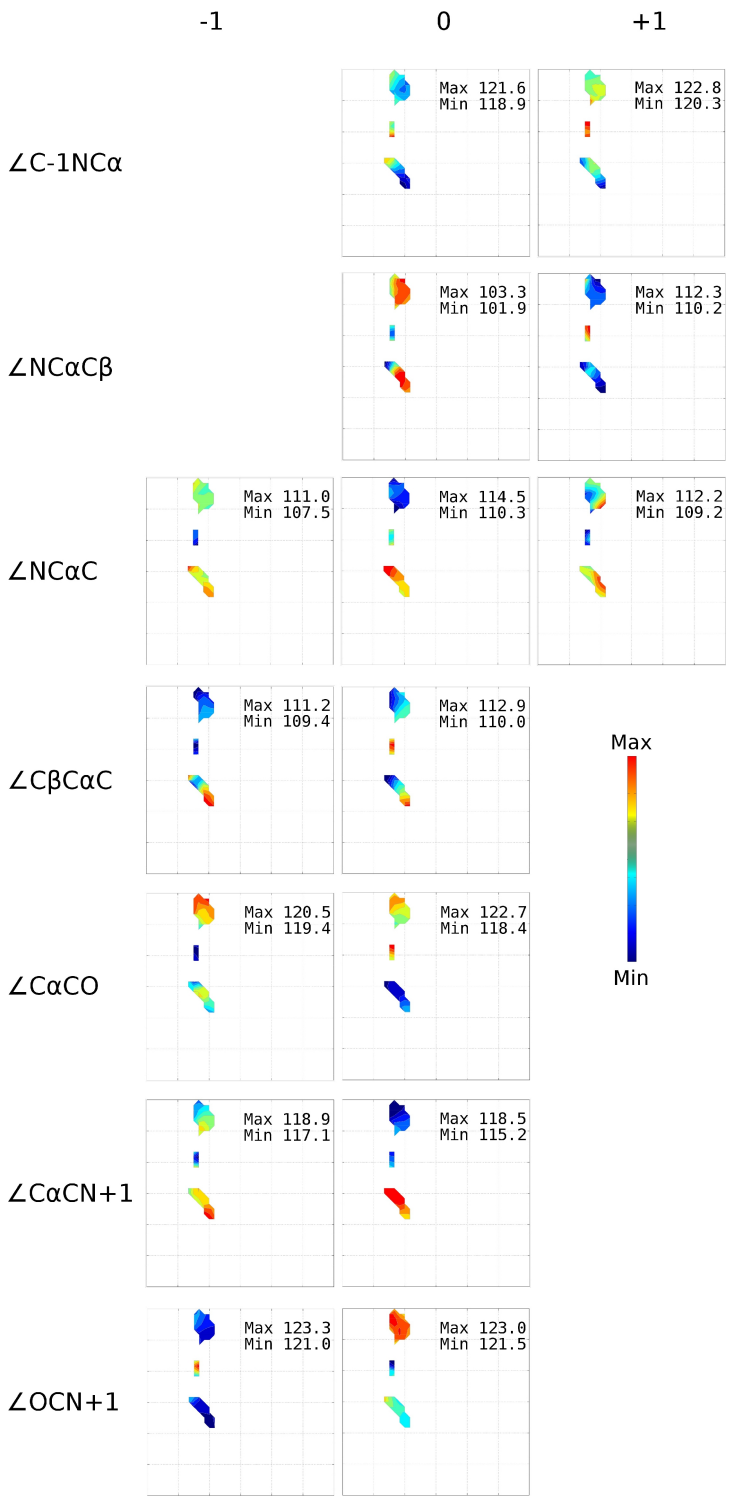


Figure 4.7. Conformation-dependent variation in bond angles of Pro residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 4.5.

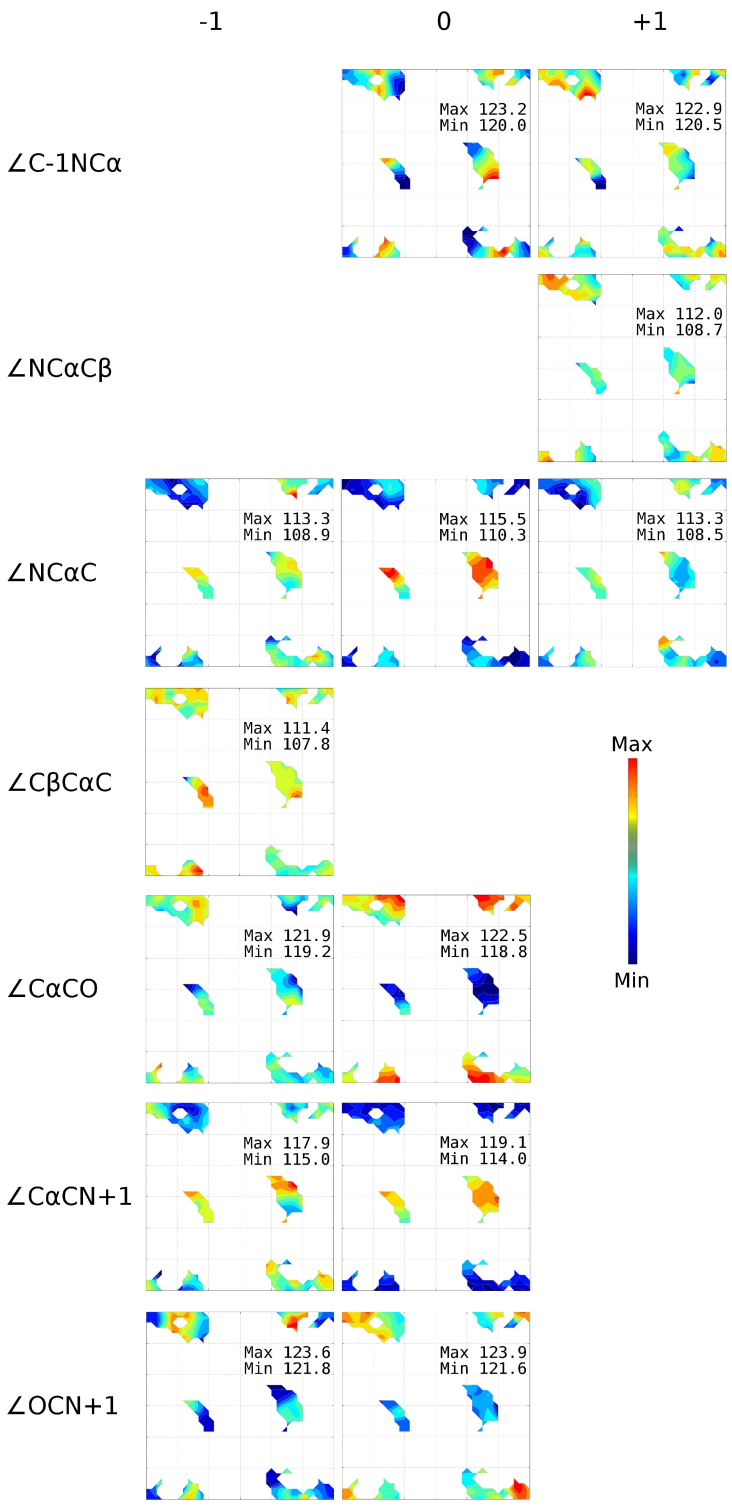


Figure 4.8. Conformation-dependent variation in bond angles of Gly residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 4.5.

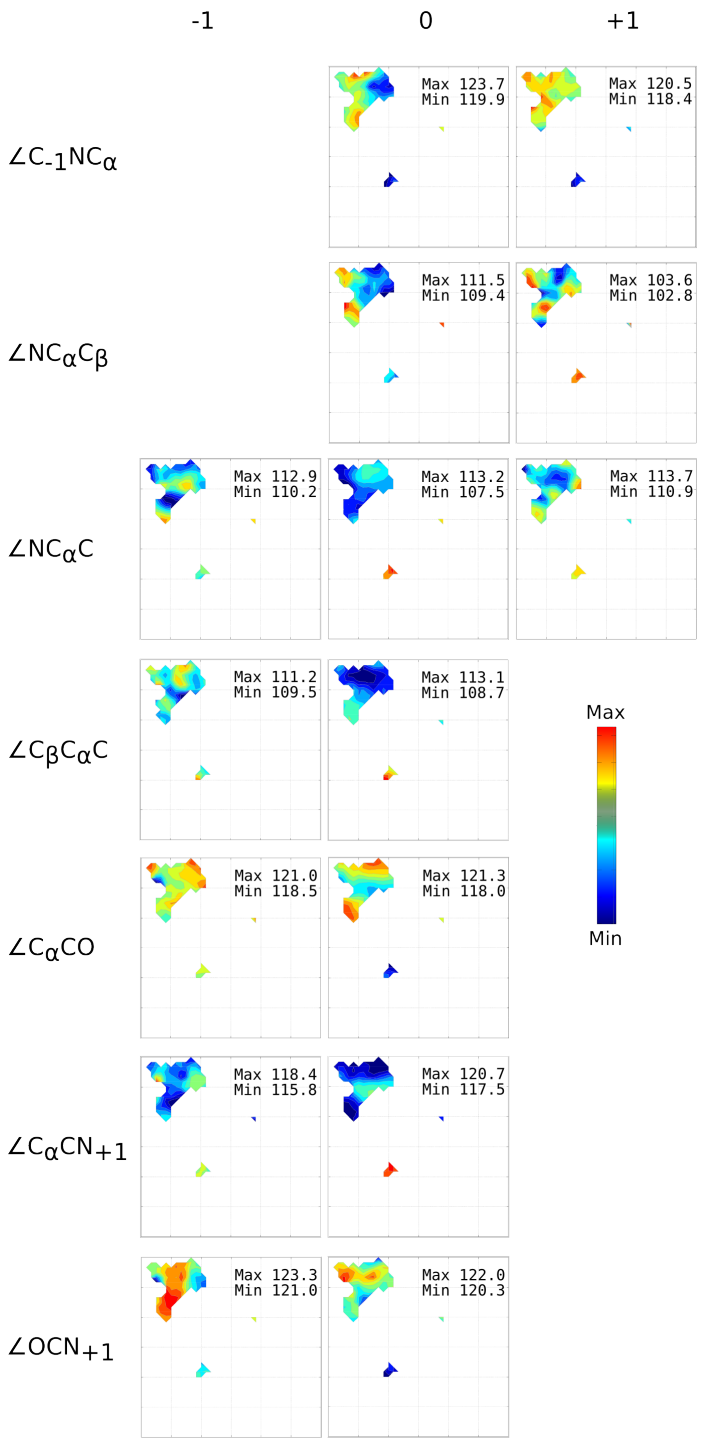


Figure 4.9. Conformation-dependent variation in bond angles of general prePro (i.e., nonGly, nonPro, nonIle/Val residues preceding proline) residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 4.5.

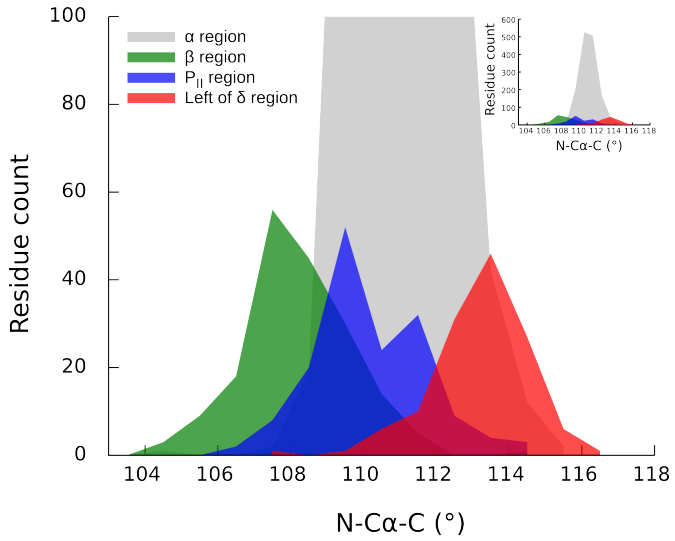


Figure 4.10. ∠NC$_\alpha$C distributions are well-defined and distinct. Shown are observations from selected 10° × 10° bins in each of four conformations: α (gray), β (green), P_{II} (blue), and a region left of δ at (-125°, -5°) (red). The Y-axis range is set to visualize the distributions in non-α bins. Histograms were calculated using 1° bins and plotted as frequency polygons. Distributions are visibly separate and thus conformation-dependent. Inset: The same plot, with the Y-axis range set to display the full height of the α distribution. If not broken out by conformation, the non-α bins would be indistinguishable from tails of the α distribution. Figure created with gnuplot and Inkscape.

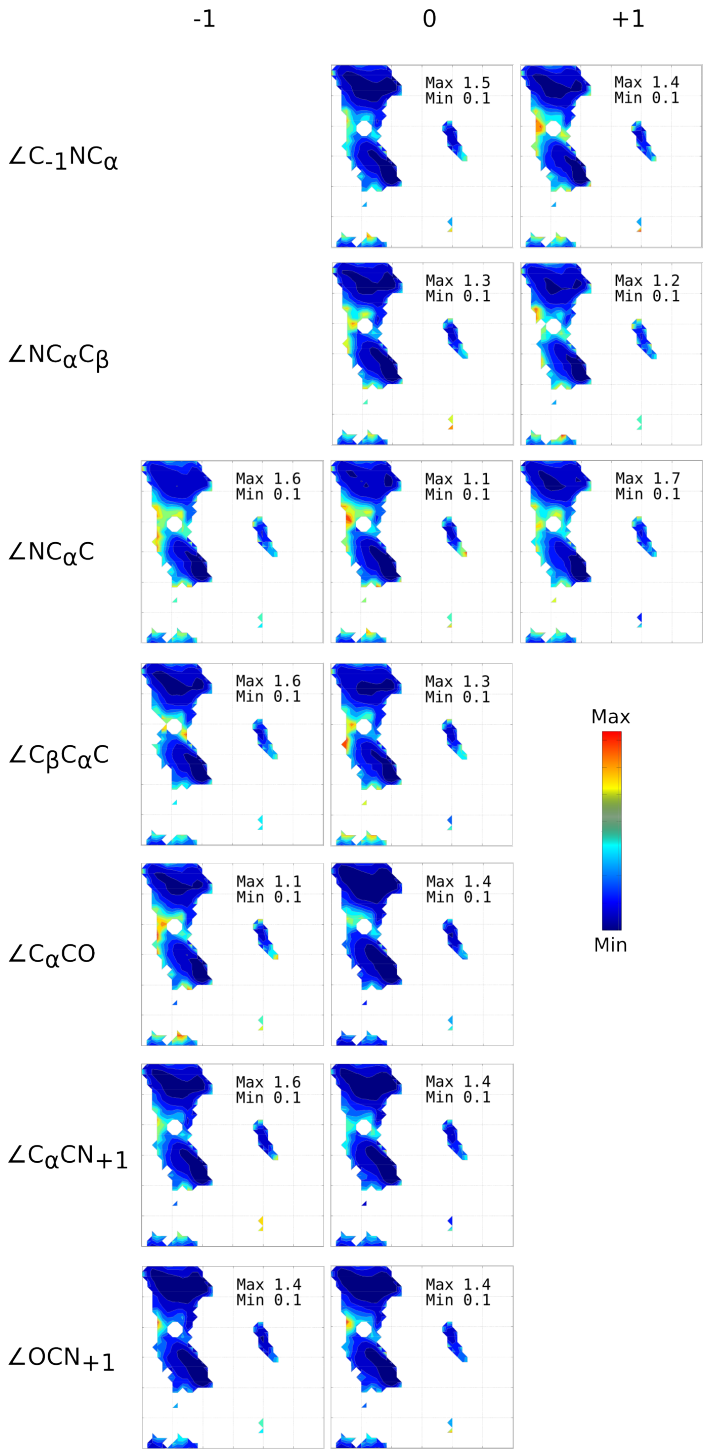


Figure 4.11. Conformation-dependent variation in the standard errors of the means of bond angles of general residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 4.5.

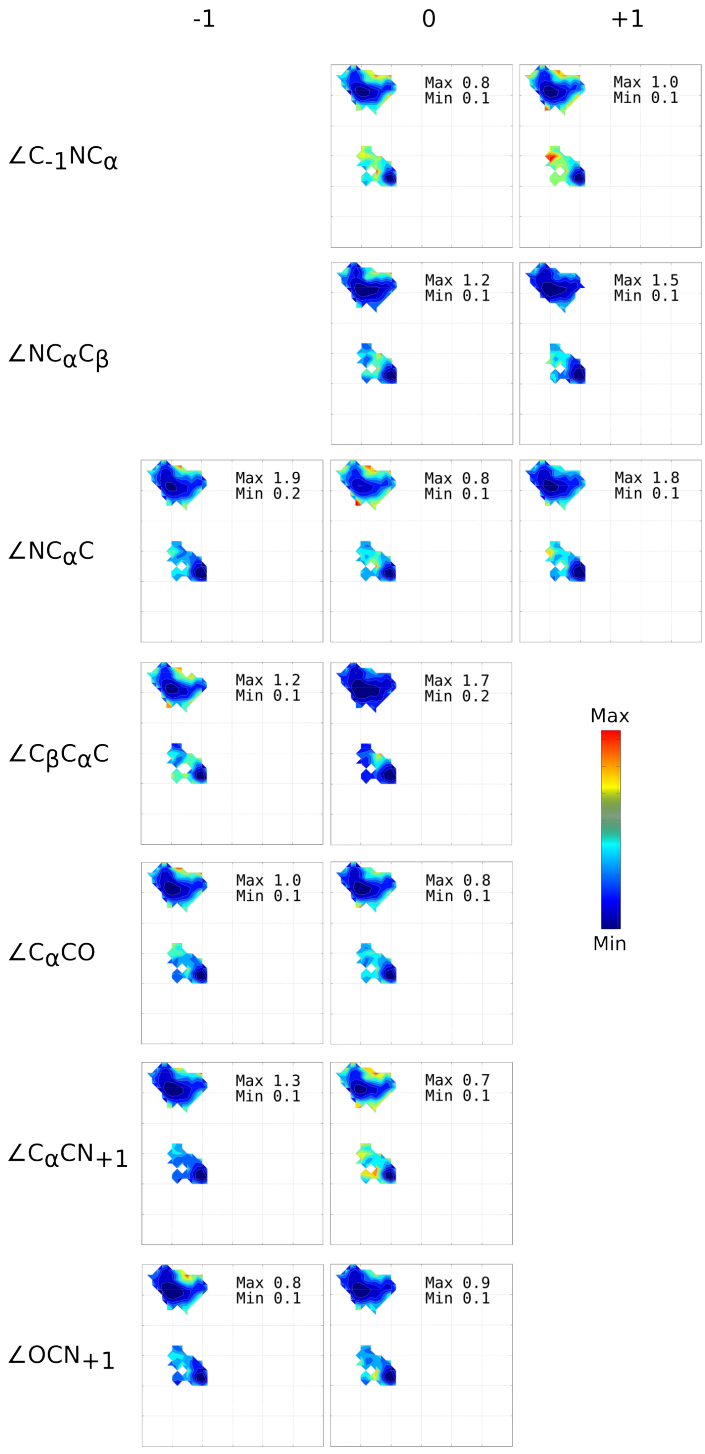


Figure 4.12. Conformation-dependent variation in the standard errors of the means of bond angles of Ile/Val residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 4.5.

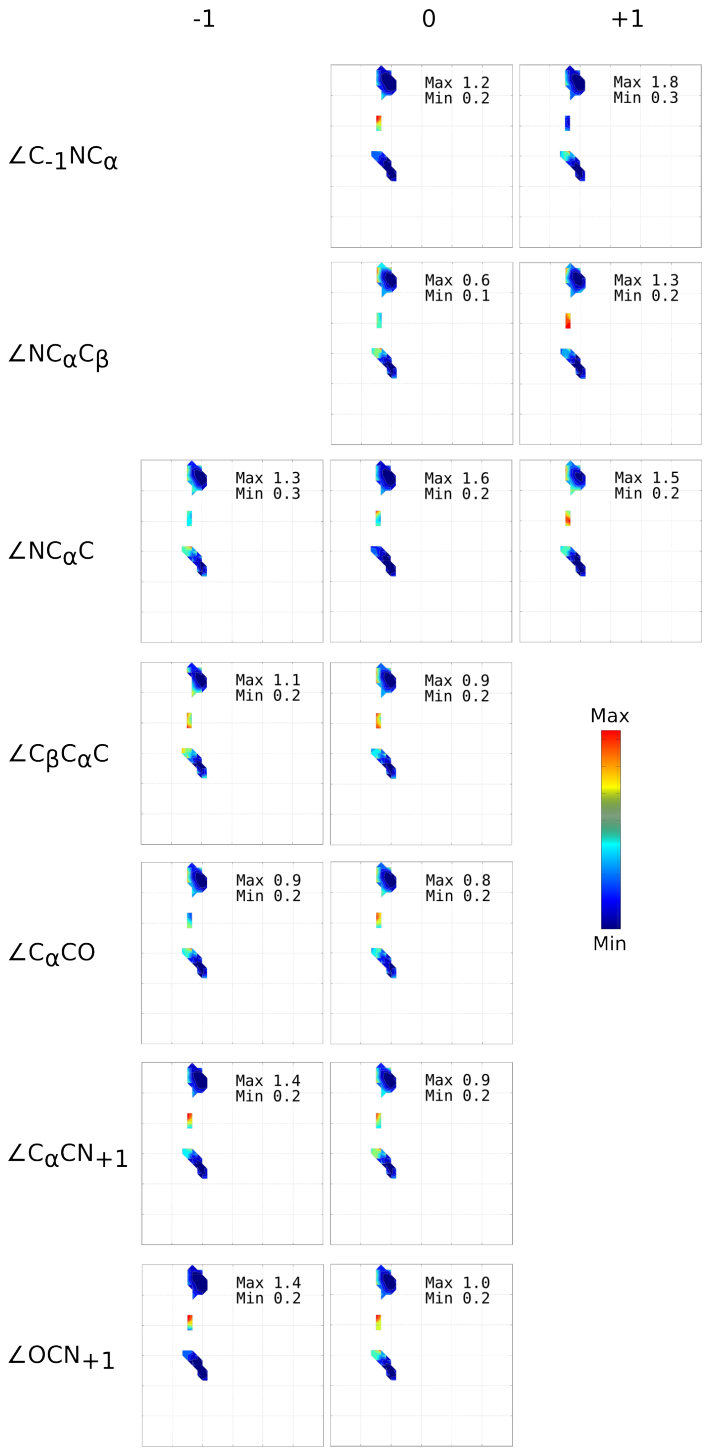


Figure 4.13. Conformation-dependent variation in the standard errors of the means of bond angles of Pro residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 4.5.

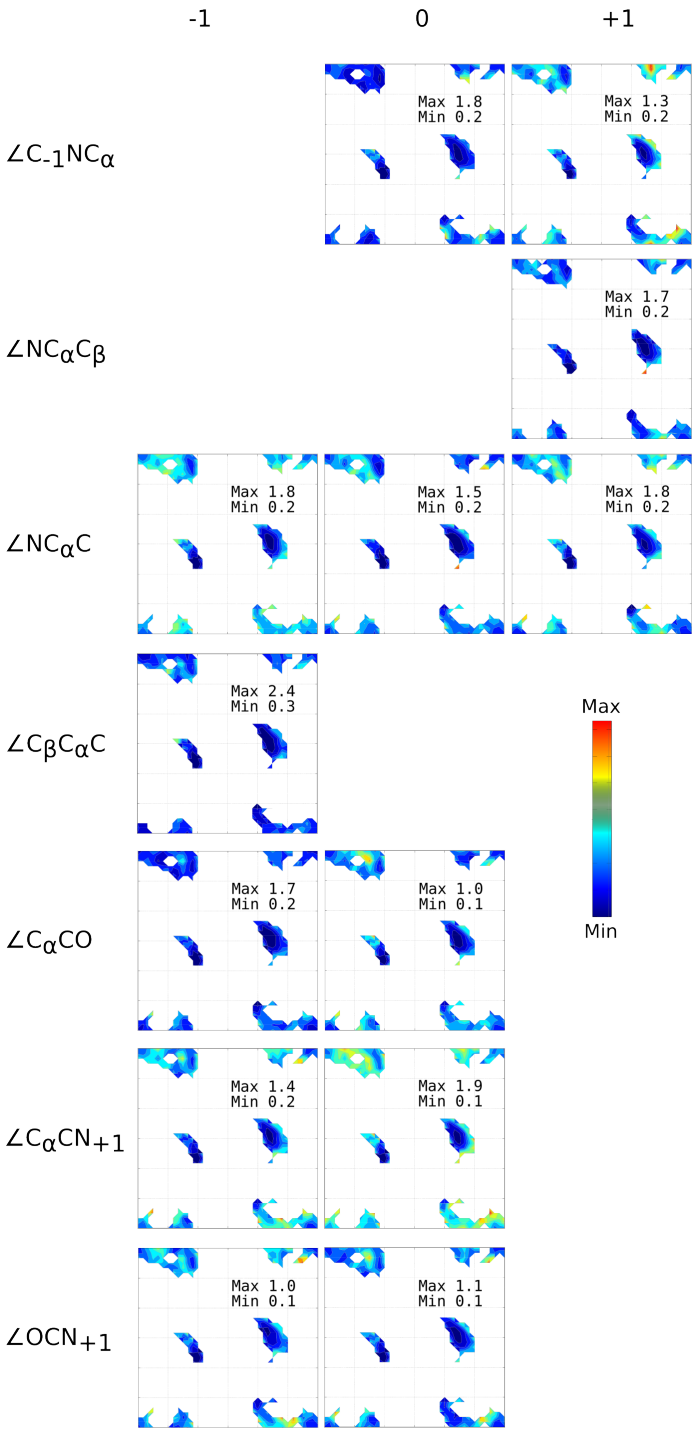


Figure 4.14. Conformation-dependent variation in the standard errors of the means of bond angles of Gly residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 4.5.

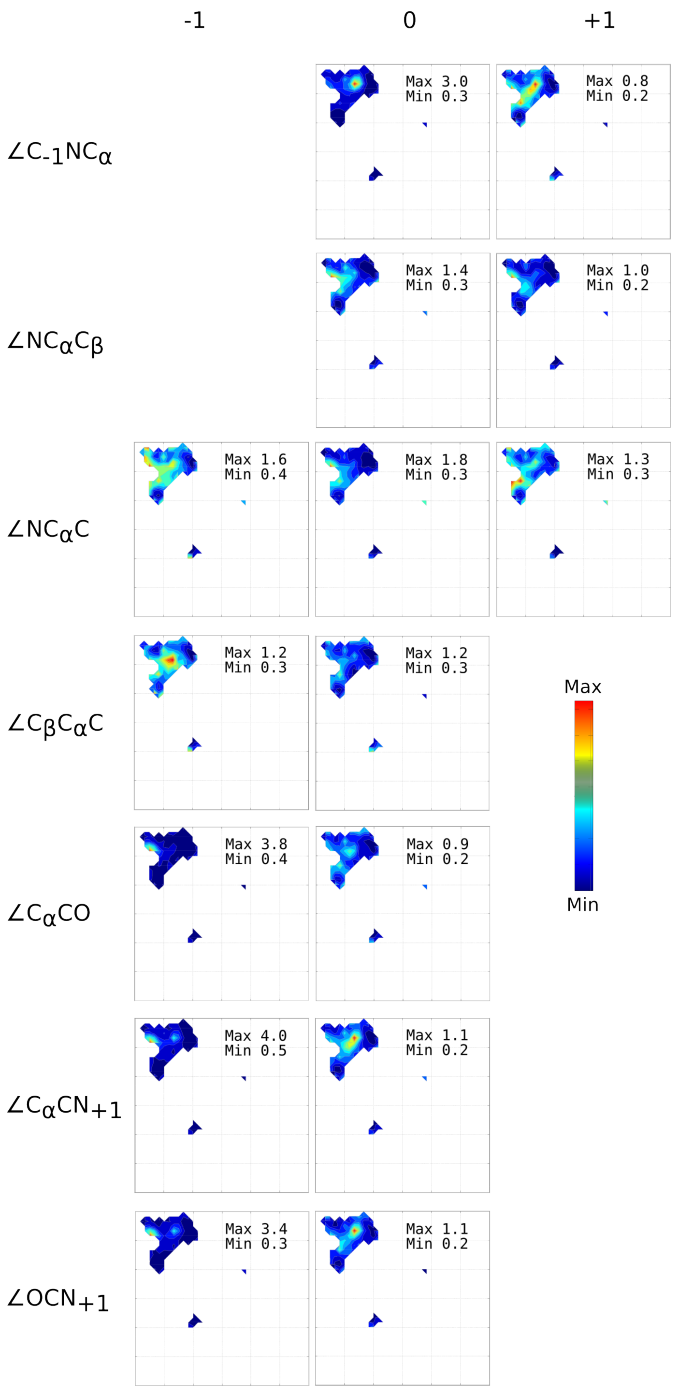


Figure 4.15. Conformation-dependent variation in the standard errors of the means of bond angles of general prePro (i.e., nonGly, nonPro, nonIle/Val residues preceding proline) residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 4.5.

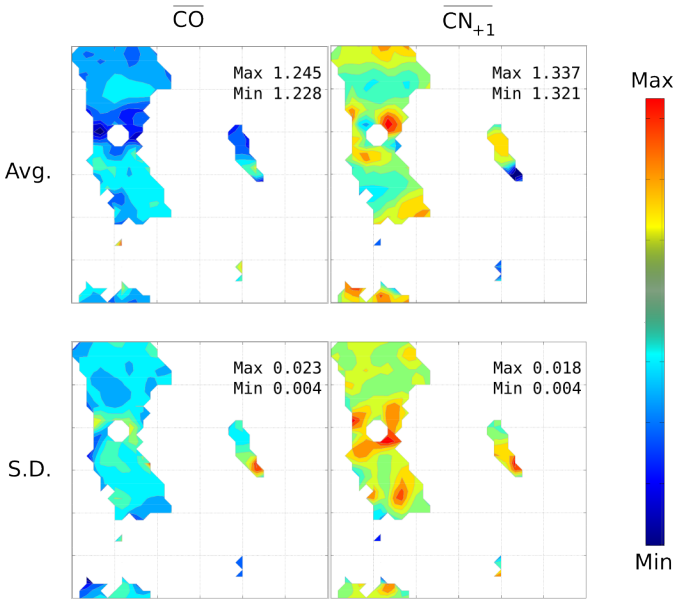


Figure 4.16. Conformation-dependent variation in bond lengths is partially masked by experimental uncertainty. Ramachandran plots are shown for average lengths and standard deviations of the C=O bond (left panels) and the C-N bond (right panels) using colors as in Figure 3. These bonds are involved in the partial double-bond character of the peptide bond, so the expectation is for them to be anticorrelated as electron density shifts between them. Some such anticorrelation is visible as a Ψ-dependent effect in averages (shown in the top panels) but it is not as clear as trends seen in bond angles, possibly because the standard deviations (shown in the bottom panels) are near the level of experimental uncertainty.

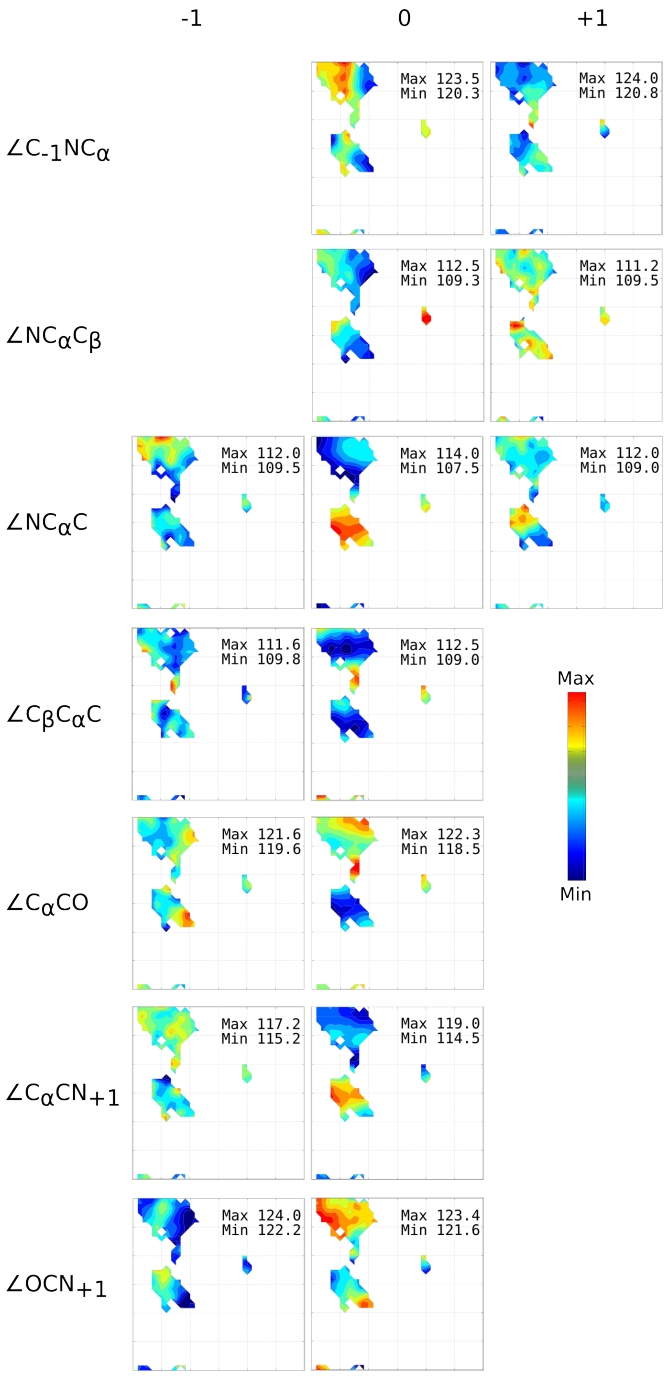


Figure 4.17. Conformation-dependent variation in bond angles of general residues without defined secondary structure as a function of the Φ,Ψ of the central residue. This can be compared with residues including those with defined secondary structure (Figure 4.5). The lack of secondary structure is defined by DSSP codes 'S' or '-'. The calculations used 1,342 residues. Otherwise, all is as in Figure 4.5.

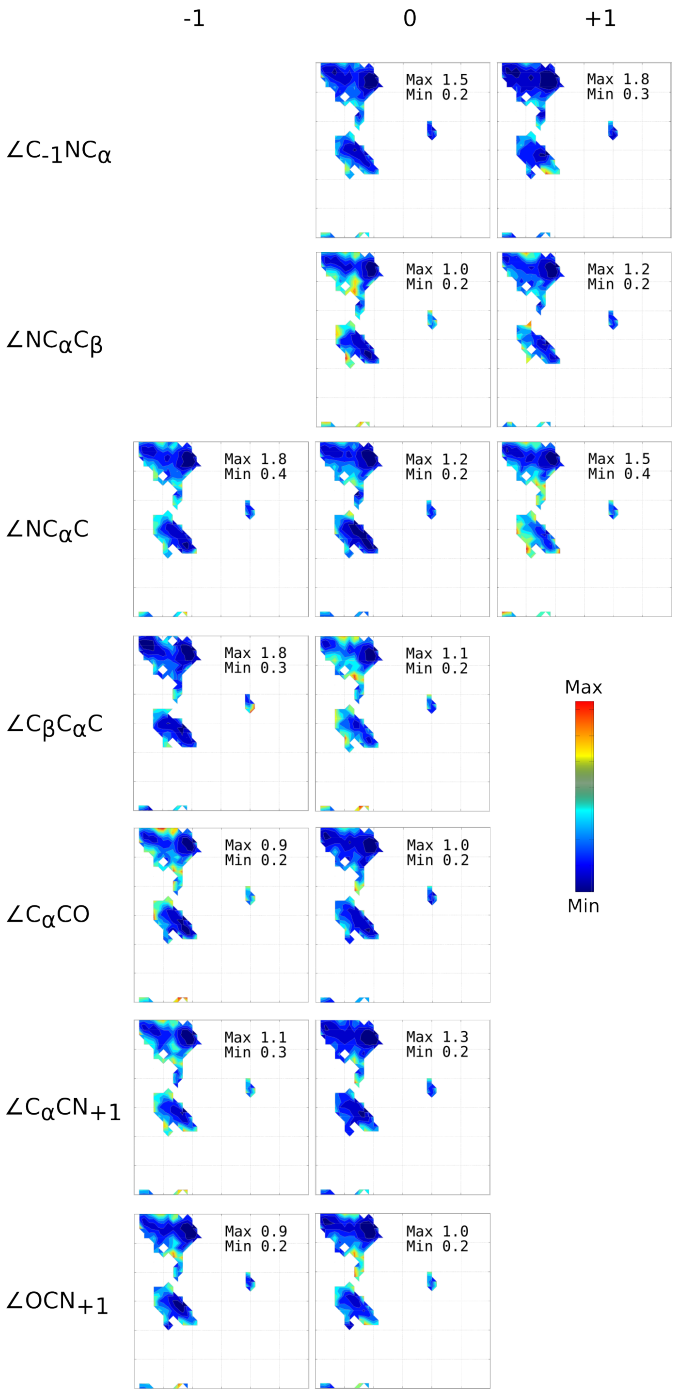


Figure 4.18. Conformation-dependent variation in the standard errors of the means of bond angles for general residues without defined secondary structure as a function of the Φ,Ψ of the central residue. This can be compared with residues including those with defined secondary structure (Figure 4.5). The lack of secondary structure is defined by DSSP codes 'S' or '-'. The calculations used 1,342 residues. Otherwise, all is as in Figure 4.5.

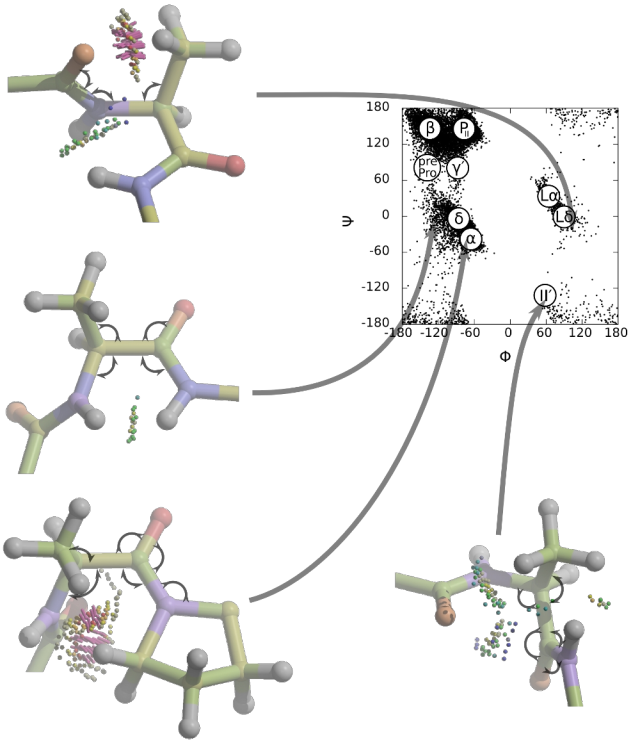


Figure 4.19. Structural basis for geometry variations of selected conformations. Four Ala residues with adjacent peptides are shown, built using EH values and with dots showing van der Waals overlap between atoms: blue (wide contact), green (close contact), yellow (small overlap), and red (bad overlap). Clockwise from top left: tip of the Lα/Lδ region; left of the δ region; a prePro-Pro dipeptide in the α region; and the II′ region. Arrows indicate angles that open or close substantially relative to EH values. Note that all of these distortions serve to alleviate atomic clashes. The overlaps were calculated by MolProbity (Davis et al., 2007) and are shown in Coot (Emsley and Cowtan, 2004).

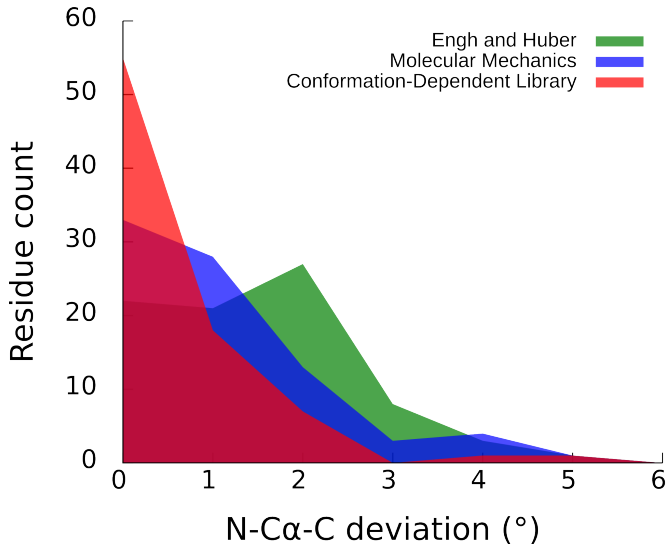


Figure 4.20. CDL ∠NC$_\alpha$C values match ultrahigh-resolution structures best. Deviations of predicted angles from the experimental ones for atomic-resolution ribonuclease (PDB code 1rge; Sevcik et al., 1996) with various methods are shown: EH single ideal values (blue), molecular mechanics (green), and the CDL (red). Results are shown in a histogram-like manner using 1° bins and frequency polygons. Of these three methods, the CDL matches best, followed by molecular mechanics, then single ideal values. Figure created with gnuplot.

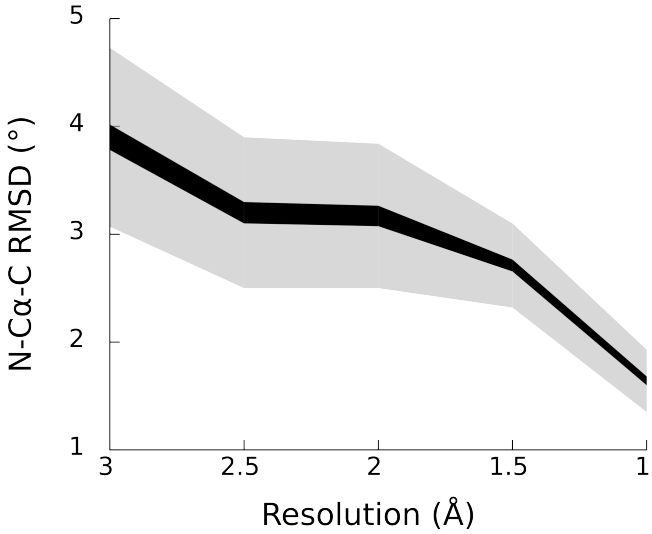


Figure 4.21. ∠$NC_{\alpha}C$ deviation of the CDL values from crystal structures as a function of resolution of the analysis. At each of five resolutions ranging from 1.0-3.0 Å, the ∠$NC_{\alpha}C$ RMSDs from the CDL were calculated for 50 nonredundant structures. The distributions of RMSDs at each resolution are shown. The thickness of the black line indicates the standard error of the mean, and the thickness of the gray line indicates the standard deviation. Figure created with gnuplot.

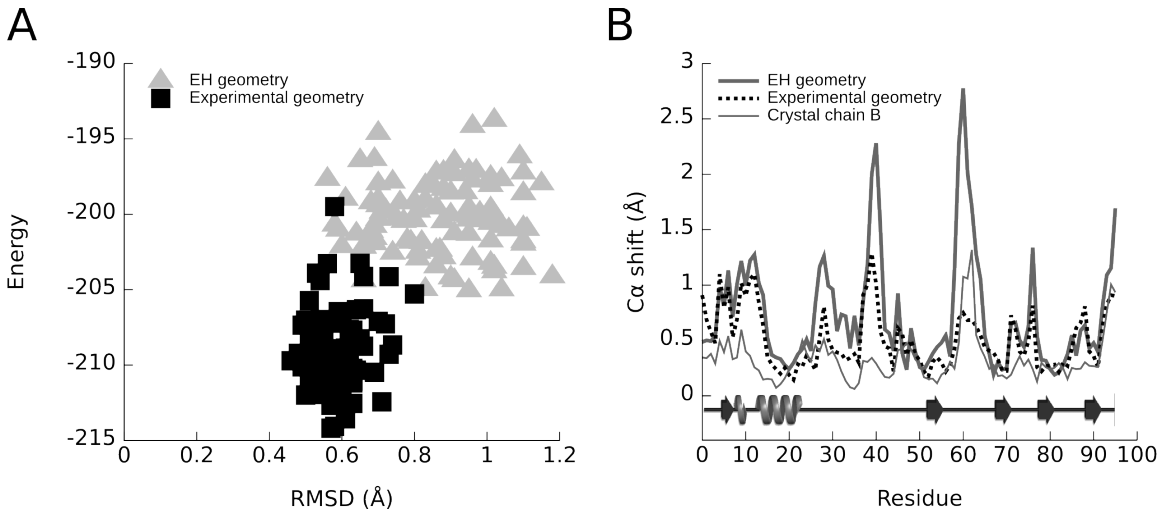


Figure 4.22. Energy minimization behaves better using experimental geometry as opposed to the rigid-geometry approximation. (A) Shown are 100 trials minimized with experimental (squares) and with EH (triangles) geometries. They are plotted as Rosetta energy versus the C_α RMSD from the crystal structure (as calculated by Rosetta). Figure created with gnuplot. (B) Shown are C_α shifts between the crystal structure chain A and a structure selected from the cluster center using experimental (dashed line) or EH (solid thick line) geometry, C_α shifts for chain B vs chain A from the same crystal (solid thin line), and a schematic of the secondary structure (using spirals for helices and arrows for β-strands). The crystal structure chain B reflects the differences in the same protein in two environments. Overlays were created using the McLachlan algorithm (McLachlan, 1982) as implemented in ProFit by iteratively overlaying structures using a subset of C_α atoms with a maximum per-atom RMSD of 0.1 Å until convergence was reached. The secondary structure is taken from PDBsum (Laskowski et al., 2005). Figure created with gnuplot and Inkscape.

Chapter 5

An expert-created model of the tumor-suppressor merlin suggests critical functional regions

Donald S. Berkholz, Anthony Bretscher, and P. Andrew Karplus

Formatted for submission

Abstract

Neurofibromatosis 2 (NF2) is a autosomal-dominant syndrome resulting in tumors, which occur most often as schwannomas in the bilateral vestibular nerve. NF2 results from a mutation in merlin (moesin-ezrin-radixin-like protein), unique among tumor suppressors as a cytoskeletal protein. Merlin's distinct functions relative to moesin, ezrin, and radixin must arise from distinct structural/dynamical characteristics, but understanding these has been hindered because no reliable structure exists for full-length merlin. To meet this need, here we present a carefully constructed homology model of merlin and point out shortcomings of available automatically constructed models. In combination with sequence comparisons, we use this model to propose critical regions contributing to merlin's distinct functions. These proposals will guide new experiments, and the model itself can guide understanding how merlin interacts with other proteins. A novel approach of using sequence comparisons to discover clustered gains and losses of function should be broadly useful for locating functionally critical regions of any two protein subfamilies having distinct functions.

Introduction

Neurofibromatosis 2 (NF2) is an autosomal, dominantly inherited genetic disease that affects about 250,000 people worldwide (Evans et al., 2005). Tumors characterize the disease, most commonly bilateral vestibular schwannomas, followed by schwannomas of other nerves, meningiomas and ependymomas (Ahronowitz et al., 2007). People affected by NF2 inherit heterozygous mutations at the *NF2* locus with homozygous loss of the wild-type function in tumor cells. Thus, the *NF2* locus is a tumor-suppressing gene. The protein expressed by this gene is merlin, also called schwannomin (Sanson et al., 1993; MacCollin et al., 1993).

Merlin is a novel class of tumor suppressor because it lacks regulatory DNA-

binding domains but instead is homologous to the ERM (ezrin-radixin-moesin) family of membrane cytoskeletal proteins. The structures of merlin and other ERM proteins are expected to be similar based on sequence similarity, with an ~300-residue N-terminal FERM domain, an ~200 residue central coiled-coil domain, and an ~80-residue C-terminal tail domain that can bind to the FERM domain to form a closed conformation (Pearson et al., 2000; Kang et al., 2002; Li et al., 2007). For the ERM proteins, the closed form is dormant, but evidence suggests that for merlin it is the closed form that is needed for tumor-suppressor activity (LaJeunesse et al., 1998; Nguyen et al., 2001). Also, the loss of merlin function versus ERM function have distinct phenotypes (LaJeunesse et al., 1998; McClatchey et al., 1997; Speck et al., 2003), implying that they must have distinct, important functions (McClatchey and Feton, 2009).

These distinct functions must derive from distinct structural/dynamical differences, but no structure exists for full-length merlin. The only known structures of merlin contain just the FERM domain (Kang et al., 2002; Shimizu et al., 2002), which is the domain most similar to the other ERM proteins. This lack of structure blocks structure-guided interpretation of the large number of biochemical experiments that have been performed on merlin and its various interacting partners, which are of unknown importance (Bretscher et al., 2002; Okada et al., 2007; Scoles et al., 2008). A structure of merlin is required for rational, structure-informed design of new experiments (Schwede et al., 2009), yet structural studies of full-length merlin have been hampered by difficulties with its expression and purification. Because a complete structure of merlin has not yet been solved experimentally, to fill this gap we have produced a homology model of merlin. This model is patently better than existing, automatically created merlin models, and its combination with sequence comparisons allows us to propose potential critical regions contributing to the differential behavior of merlin relative to the other ERM proteins.

The largest value added from homology modeling comes from properties depending not on highly accurate geometric characteristics but instead on the sequence changes, such as electrostatic potential and presence of polar areas (Chakravarty and

Sanchez, 2004). Although sequence alignments alone can give some clue about the position of specific residues, it is difficult to discover the existence or understand the meaning of broader trends across multiple residue substitutions without a structural model. Even single-residue changes require significant effort to map to structural knowledge without this merlin model. Here, we describe insights generated from this model that would not be easily apparent without it. The resolution power of these insights is far greater than the resolution with which binding partners are known from other biochemical studies (often as large as 100-200 residues), so this model provides a guide for future mutational studies to hone in on merlin-partner interactions.

Results and Discussion

Creation and validation of the homology model

To create the merlin model, we used a template based on moesin, which is the most similar known structure to merlin, about 45% identical in sequence and the only ERM protein with a structure for the full-length protein. The template structure was a composite of two moesin crystal structures from the insect *Spodoptera frugiperda* reported by Li et al. (2007). The 2.1 Å resolution structure was used (PDB code 2i1j), with the addition of parts of the structure only visible at 3.0 Å resolution (PDB code 2i1k). We refer to this composite template as *Sf*moesin. Following a conservative strategy, the merlin model was built using Rosetta (allowing the side chains to repack but, with one exception, restraining the backbone to the template positions (see Methods). Also, we chose not to model residues in merlin lacking an equivalent residue in the template. Only three unmodeled regions are longer than three residues (1-20, 415-426, and 486-505), and the ones with *Sf*moesin equivalents correspond to disordered residues, making it reasonable to predict that they will also be natively unstructured in merlin. The final model has reasonable geometry quality.

Existing automated models have serious errors

Two popular databases of automatically created models, ModBase (Pieper et al., 2004) and SwissModel (Kopp and Schwede, 2006), each contain a model of merlin. For both of the models, however, problems with alignment accuracy exist that seriously limit their utility. For the ModBase model, misalignment causes the coiled-coil to turn at Leu383, which leaves half the coiled-coil unmodeled and creates a 36-residue frameshift that resolves at residue 508 through the looping out of a coil (Fig. 5.1B). The SwissModel errors are more distributed, being due to four shorter misalignments (Fig. 5.1C). The first occurs because of failure to recognize a 10-residue gap in the sequence alignment from residues 400-409, which resolves at residue 431 through the looping out of a coil. The second occurs at residues 494-504, which form a loop misaligned to a beta strand in *Sf*moesin. In actuality, residues 487-504 are an insertion in merlin relative to *Sf*moesin. The third is near the launching pad of the coiled-coil, where residues 341-349 form a protrusion of random loop wrapping around the coiled-coil. The fourth is near the landing pad of the coiled-coil, where residues 476-482 form a random loop.

The incorrect features of both of these models underscore that obtaining an accurate sequence alignment is crucial to success, yet it remains a limiting problem (Zhang, 2009).

Experimental results support our model

The most direct way to validate our model is by comparison with the crystal structure of the merlin FERM domain (PDB entry 1h4r; Kang et al., 2002). This structure was intentionally left out of our modeling so it could serve as an independent source for validation.

An overlay of the merlin FERM domain (yellow) and the complete merlin model (green) shows reasonable agreement between the model and the crystal structure (Fig. 5.2A), with an overall C_α RMSD for the FERM domain of 1.2 Å. Most residues have C_α shifts 0.1 Å–0.5 Å, and only three loops have C_α shifts above 1 Å: residues 70-71, 160-161, and 278-279 at 1.5 Å, 1.2 Å, and 1.2 Å, respectively (Fig. 5.2B). Many differences are small subdomain-level shifts rather than local structural errors. Overlays of the three individual subdomains of the FERM domain tend to show much better agreement than overlays of the entire domain (Kang et al., 2002). The extensive similarity in backbone conformation between moesin (the template for the merlin model) and merlin supports the validity of this model for making biologically relevant insights.

New insights from the complete model of merlin

Electrostatic analyses

Electrostatic potential surfaces were examined for moesin alone (Li et al., 2007) and for the merlin FERM domain alone (Kang et al., 2002), but this report marks the first time is it possible to extend electrostatic analysis of merlin to the full protein, adding the critical α-helical and tail domains, both of which mask the FERM domain in merlin's closed state. Since the closed state of merlin is the active, tumor-suppressing state, the electrostatic potential analyzed here is more biologically relevant than that of the FERM domain alone.

The tail-FERM masking interaction is weaker in merlin than in moesin, resulting in increased levels of the open form of merlin relative to moesin (LaJeunesse et al., 1998; Nguyen et al., 2001). However, the reason for this difference is unknown. In moesin, a strong complementary charge interaction between the FERM and tail domains existed (see Fig. 4 of Pearson et al., 2000). In merlin, the electrostatic potential is much weaker in both the tail and the FERM face (Fig. 5.3), which will significantly weaken the

masking interaction.

Residue-conservation patterns

Additionally, we compared residue-conservation patterns of merlin with those of other ERM proteins to identify three potential classes of residue that help define the distinct functions of merlin versus ERM proteins: change of function are those that are different but are conserved in each subfamily, gain of function are those conserved only in merlin, and loss of function are those conserved only in other ERM proteins.

A mapping of change-of-function residues and NF2-associated missense mutations onto the merlin model highlighted five clusters (Fig. 5.5). The first is along one coil of the helical domain at residues ~432-465 (Fig. 5.5B). The second is the landing pad of the helical domain including residues 59, 62, 64, 77, and 481 (Fig. 5.5D), also observed as part of a much larger conserved region by Kang (2002), who suggested that because E38 and W41 missense NF2 mutants are in there, the effects are manifested by impairing the ability of merlin to bind effectors/activators. Additionally, residues 50-70 are implicated in binding paxillin, involved in focal adhesions (Scoles, 2008). The third is the α-helix of the tail domain (residues ~526-549; Fig. 5.5E), in which the three disease-related mutations were briefly mentioned by Pearson (2000) as involved in association with the FERM domain. The fourth is a large patch next to the final α-helix of the tail domain (residues 272 and ~281-286; Fig. 5.5C), noted by Kang (2002) to have reversed charge for two residues (E270 and K284) but without any proposal for functional relevance. Residues 280-323 are involved in binding SCHIP-1 (Scoles, 2008), which could explain the conservation of this cluster. The fifth and final is another patch on the FERM domain between the beginning of the helical domain and the tail domain (residues 197, 198, 202; Fig. 5.5A).

The implicated gain-of-function and loss-of-function between merlin and other ERM proteins can be shown simultaneously by mapping the difference in conservation

between merlin sequences and other ERM sequences onto the merlin model. Three gain-of-function and two loss-of-function clusters are readily apparent (Fig. 5.4). Three changes serve as positive controls controls that the method works as expected: a loss-of-function change clustered at the tail actin-binding site (Fig. 5.4A), along with a gain-of-function at Ser518 (which is phosphorylated to regulate the tail's masking; Fig. 5.4A) and a loss-of-function at the nonequivalent residue serving the same function in moesin (Fig. 5.4A). The novel methodology we used allows for straightforward visualization of gains and losses of function by mapping them as easily interpretable colors on a single structure, negating the common need for error-prone comparisons using multiple different structures and sequence groups.

The gain-of-function mutations are: (1) the alternate actin-binding site near the end of the helical domain at residues ~466-474 (also implicated in binding RIβ and HRS (Scoles, 2008) and the adjacent helix's residues ~358-373 (Fig. 5.4A); (2) in and shortly after the tail domain's initial β-strand that binds in the same place as ICAM-2 (Fig. 5.4B); and (3) another tail-domain motif, the α-helix at residues ~536-544 (Fig. 5.4A) also seen to have a gain of function relative to other ERM proteins; implicated in binding HRS and CRM1/exportin; Scoles, 2008), along with a FERM domain α-helix adjacent to it, at residues ~182-193 and two very large gains at residues 157 and 159. The major loss-of-function clusters are: (1) the actin-binding site on the tail domain, which has a much stronger signal than anything else (Fig. 5.4A); and (2) near the helical domain's landing pad (Fig. 5.4B). Simultaneously mapping the missense mutations onto the structure provided additional support for the importance of the tail-domain α-helix.

Outlook

The lack of sufficient specificity for most experimentally determined interactions between merlin and its binding partners makes it impossible to propose a one-to-one mapping of all of these critical clusters to specific functions. This was only possible in a

few cases, described above, that have reasonably localized interaction sites. Many merlin-interacting proteins bind to it in locations that are known with an error of 100-200 residues and that overlap with those of other binding partners. The level of accuracy of critical regions discovered in this comparison exceeds the available experimental data, so these clusters can guide follow-up mutational studies to hone in on merlin-partner interactions. We expect the method developed here to discover clustered gains and losses of function to be broadly useful for discovering critical structural regions between any two subfamilies with differential function.

Using an expert-created model of merlin, we have shown a structural basis for the increased propensity of merlin for the open form relative to moesin and suggested putative critical regions of merlin using both known and novel comparison methods. We expect that the existence of this new, expert-created model for merlin's complete closed form will open new avenues of structurally informed experimental design and interpretation.

Materials and Methods

Homology-modeling protocols

The merlin query sequence (GenBank accession number NP_000259.1) was aligned with potential template sequences using an HHPred (Söding et al., 2005) query against the Protein Data Bank (Berman et al., 2000). The HHPred results contained 100 hits above 90% probability of homology, with the top two hits being the moesin structures used as templates.

The composite template was created by making a structure-based superposition, using Theseus (Theobold and Wuttke, 2006), of the moesin structures for each segment taken from the 3.0 Å structure (PDB code 2i1k) using the residues near the edge of each

segment. Following superposition in Theseus, residues were copied from that overlap region into the 2.1 Å coordinate set (PDB code 2i1j).

The Rosetta modeling suite (Rohl et al., 2004), version 2.3.0, was used to create the homology model. Scripts were also used from the BioTools toolbox, available with Rosetta. To create a map of residues in moesin to residues in merlin, the PDB file needed to have all residues in the moesin sequence, including those not seen in the crystal structure. These residues were added with XYZ coordinates of 0.000 and occupancy of -1 using completePdbCoords.pl. Next, a FASTA alignment of moesin and merlin was converted to a Rosetta zones file that aligned residues in four segments (1-49:1-49, 52-476:53-477, 482-547:480-545, and 551-577:548-574), and createTemplate.pl generated a template structure. Parameters included the query FASTA file, '-takeoffpad F', and '-sidechains T', which preserves C_β geometry on identical residues. Of note is that the FASTA file represents residues present in the crystal structure rather than the complete sequence of the protein. This was handled by changing unstructured loop residue IDs to '-' as well as deleting them from the PDB file. At this point, a complete model of the backbone existed.

Once the backbone model was created, sidechains were added using RosettaDesign's option to only pack sidechains and do nothing else (-design -onlypack), and expanding the rotamer library by two standard deviations from the mean (-ex1 -ex2). After that, all-atom sidechains were energy-minimized using Rosetta's relax module (-relax -farlx -minimize -sc_only). This resulted in an atomically complete, minimized structure containing all the residues we intended to model. These residues were not modeled: 1-20 (1-18 are nonexistent in moesin), 68-69, 130-132, 415-426, 486-505, 566-568.

Based on the residue identity, we expect the main-chain RMSD between the moesin template and the modeled portions of merlin to be ~1 Å in the FERM domain, which is 65% identical, and ~1.6 Å in the α-helical and tail domains, which are 25% identical (Chothia and Lesk, 1986). Because this is near the accuracy limit of current

state-of-the-art modeling programs such as Rosetta in best-case scenarios, and because modeling programs have extreme difficulty with proteins the size of merlin at 500+ residues, we chose to be conservative with our use of energy minization. Thus, we only remodeled the side-chains instead of also minimizing the backbone.

We validated the model and compared results with the two template structures and the merlin FERM domain structure at 1.8 Å (PDB code 1h4r; Kang et al., 2002). Validation with MolProbity (Davis et al., 2007) showed that the model has good geometry, with 0.38% outliers on the Ramachandran plot, 2.74% sidechain rotamer outliers, 0.19% bond-length outliers, 0.19% bond-angle outliers. The overall MolProbity score places this model at the 69th percentile of experimentally solved protein structures. The largest discrepancy occurs at residues 85-88, where there is a small deviation of a loop between the two structures.

The MolProbity results revealed one potentially serious problem at residues 508-509 in the initial C-terminal β strand, which change from Gly-Gly in *Sf*moesin to Phe-Asp in merlin but have Φ,Ψ angles only allowed for glycine. Fortuitously, a structure of the human moesin FERM domain (PDB code 1ef1 at 1.9 Å resolution; Pearson et al., 2000) has Lys-Asp instead of Gly-Gly, and these residues occupy the nearby β region of the Ramachandran plot. Residues 489-493 in human moesin were superimposed onto 506-510 in the model, then the backbone geometry was transferred to the merlin model from Ser506C to Lys510C_α, inclusive, because these atoms overlapped within <0.2 Å.

Analyses of electrostatic potential and residue conservation

Electrostatic potential surfaces were created using the APBS (Baker et al., 2001) plugin for PyMol and displayed using a range of +6/-6 kT/e.

For the change-of-function analysis, groups of merlin sequences were obtained by querying the PipeAlign server (Plewniak et al., 2003) with the human merlin sequence,

which then returned a set of pre-grouped sequences. Groups were manually edited to remove sequences lacking large proportions of the sequence (e.g., missing most of the FERM domain). The change-of-function analysis was performed by uploading the PipeAlign groups to the AMAS server (Livingstone and Barton, 1993).

The loss-of-function and gain-of-function analyses were performed using ConSurf (Landau et al., 2005) analyses of the group of 21 merlin sequences from PipeAlign and a group of 20 nonmerlin ERM sequences obtained from a ConSurf query of PDB code 2i1k. PDB files with ConSurf conservation indices in the B-factor field were loaded into PyMol, then we performed a structural alignment and saved the alignment object. With that alignment object, we used the PyMol scripting interface to calculate the difference between the conservation indices and store it in the B-factor field.

Acknowledgements

This work was supported by NIH grant R01-GM083136 (to PAK) and DOD/CDMRP grant NF073094 (to AB and PAK). We would like to thank members of the David Baker lab at the University of Washington at Seattle for their assistance with Rosetta. In particular, we would like to thank Srivatsan Raman, James Thompson and Elizabeth Kellogg for their help in understanding Rosetta's homology-modeling protocols.

References

Baker, N.A., Sept, D., Joseph, S., Holst, M.J., and McCammon, J.A. (2001). Electrostatics of nanosystems: Application to microtubules and the ribosome. Proceedings of the National Academy of Sciences of the United States of America *98*, 10037-10041.

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.

Bretscher, A., Edwards, K., and Fehon, R.G. (2002). ERM proteins and merlin: integrators at the cell cortex. Nature Reviews. Molecular Cell Biology *3*, 586-599.

Chakravarty, S., and Sanchez, R. (2004). Systematic analysis of added-value in simple comparative models of protein structure. Structure (London, England: 1993) *12*, 1461-1470.

Chothia, C., and Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. The EMBO Journal *5*, 823-826.

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., Snoeyink, J., Richardson, J.S., and Richardson, D.C. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Research *35*, W375-383.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallographica. Section D, Biological Crystallography *60*, 2126-2132.

Evans, D.G.R., Moran, A., King, A., Saeed, S., Gurusinghe, N., and Ramsden, R. (2005). Incidence of vestibular schwannoma and neurofibromatosis 2 in the North West of England over a 10-year period: higher incidence than previously thought. Otology & Neurotology: Official Publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology *26*, 93-97.

Iris Ahronowitz, Winnie Xin, Rosemary Kiely, Katherine Sims, Mia MacCollin, and Fabio P. Nunes (2007). Mutational spectrum of the NF2 gene: a meta-analysis of 12 years of research and diagnostic laboratory findings. Human Mutation *28*, 1-12.

Kang, B.S., Cooper, D.R., Devedjiev, Y., Derewenda, U., and Derewenda, Z.S. (2002). The structure of the FERM domain of merlin, the neurofibromatosis type 2 gene product. Acta Crystallographica. Section D, Biological Crystallography *58*, 381-391.

Kopp, J., and Schwede, T. (2006). The SWISS-MODEL Repository: new features and

functionalities. Nucleic Acids Research *34*, D315-318.

LaJeunesse, D.R., McCartney, B.M., and Fehon, R.G. (1998). Structural analysis of Drosophila merlin reveals functional domains important for growth control and subcellular localization. The Journal of Cell Biology *141*, 1589-1599.

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. Nuc. Acids Res. *33*, W299-302.

Li, Q., Nance, M.R., Kulikauskas, R., Nyberg, K., Fehon, R., Karplus, P.A., Bretscher, A., and Tesmer, J.J.G. (2007). Self-masking in an intact ERM-merlin protein: an active role for the central alpha-helical domain. Journal of Molecular Biology *365*, 1446-1459.

Livingstone, C.D., and Barton, G.J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput. Appl. Biosci. *9*, 745-756.

MacCollin, M., Mohney, T., Trofatter, J., Wertelecki, W., Ramesh, V., and Gusella, J. (1993). DNA diagnosis of neurofibromatosis 2. Altered coding sequence of the merlin tumor suppressor in an extended pedigree. JAMA: The Journal of the American Medical Association *270*, 2316-2320.

McClatchey, A.I., Saotome, I., Ramesh, V., Gusella, J.F., and Jacks, T. (1997). The Nf2 tumor suppressor gene product is essential for extraembryonic development immediately prior to gastrulation. Genes & Development *11*, 1253-1265.

McClatchey, A.I., and Fehon, R.G. (2009). Merlin and the ERM proteins--regulators of receptor distribution and signaling at the cell cortex. Trends Cell Biol. *19*, 198-206.

Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallographica. Section D, Biological Crystallography *53*, 240-255.

Nguyen, R., Reczek, D., and Bretscher, A. (2001). Hierarchy of merlin and ezrin N- and C-terminal domain interactions in homo- and heterotypic associations and their relationship to binding of scaffolding proteins EBP50 and E3KARP. J. Biol. Chem. *276*, 7621-7629.

Okada, T., You, L., and Giancotti, F.G. (2007). Shedding light on Merlin's wizardry. Trends in Cell Biology *17*, 222-229.

Pearson, M.A., Reczek, D., Bretscher, A., and Karplus, P.A. (2000). Structure of the ERM protein moesin reveals the FERM domain fold masked by an extended actin binding tail domain. Cell *101*, 259-270.

Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M.S., Davis, F.P., Stuart, A.C.,

Mirkovic, N., Rossi, A., Marti-Renom, M.A., Fiser, A., et al. (2004). MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Research *32*, D217-222.

Plewniak, F., Bianchetti, L., Brelivet, Y., Carles, A., Chalmel, F., Lecompte, O., Mochel, T., Moulinier, L., Muller, A., Muller, J., et al. (2003). PipeAlign: A new toolkit for protein family analysis. Nuc. Acids Res. *31*, 3829-3832.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., and Baker, D. (2004). Protein structure prediction using Rosetta. Methods Enzymol. *383*, 66-93.

Sanson, M., Marineau, C., Desmaze, C., Lutchman, M., Ruttledge, M., Baron, C., Narod, S., Delattre, O., Lenoir, G., and Thomas, G. (1993). Germline deletion in a neurofibromatosis type 2 kindred inactivates the NF2 gene and a candidate meningioma locus. Human Molecular Genetics *2*, 1215-1220.

Schwede, T., Sali, A., Honig, B., Levitt, M., Berman, H.M., Jones, D., Brenner, S.E., Burley, S.K., Das, R., Dokholyan, N.V., et al. (2009). Outcome of a workshop on applications of protein models in biomedical research. Structure (London, England: 1993) *17*, 151-159.

Scoles, D.R. (2008). The merlin interacting proteins reveal multiple targets for NF2 therapy. Biochimica Et Biophysica Acta *1785*, 32-54.

Shimizu, T., Seto, A., Maita, N., Hamada, K., Tsukita, S., Tsukita, S., and Hakoshima, T. (2002). Structural basis for neurofibromatosis type 2. Crystal structure of the merlin FERM domain. The Journal of Biological Chemistry *277*, 10332-10336.

Söding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Research *33*, W244-248.

Speck, O., Hughes, S.C., Noren, N.K., Kulikauskas, R.M., and Fehon, R.G. (2003). Moesin functions antagonistically to the Rho pathway to maintain epithelial integrity. Nature *421*, 83-87.

Theobald, D.L., and Wuttke, D.S. (2006). THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics (Oxford, England) *22*, 2171-2172.

Zhang, Y. (2009). Protein structure prediction: when is it useful? Curr. Opin. Struct. Biol. *19*, 145-155.

Abbreviations list

ERM: Ezrin, radixin, moesin

FERM: 4.1-band protein, ezrin, radixin, moesin.

NF2: Neurofibromatosis 2 (an autosomal, dominantly inherited genetic disease)

PDB: Protein Data Bank

RMSD: Root-mean-square deviation (a way to measure distance between equivalent atoms across an entire protein)

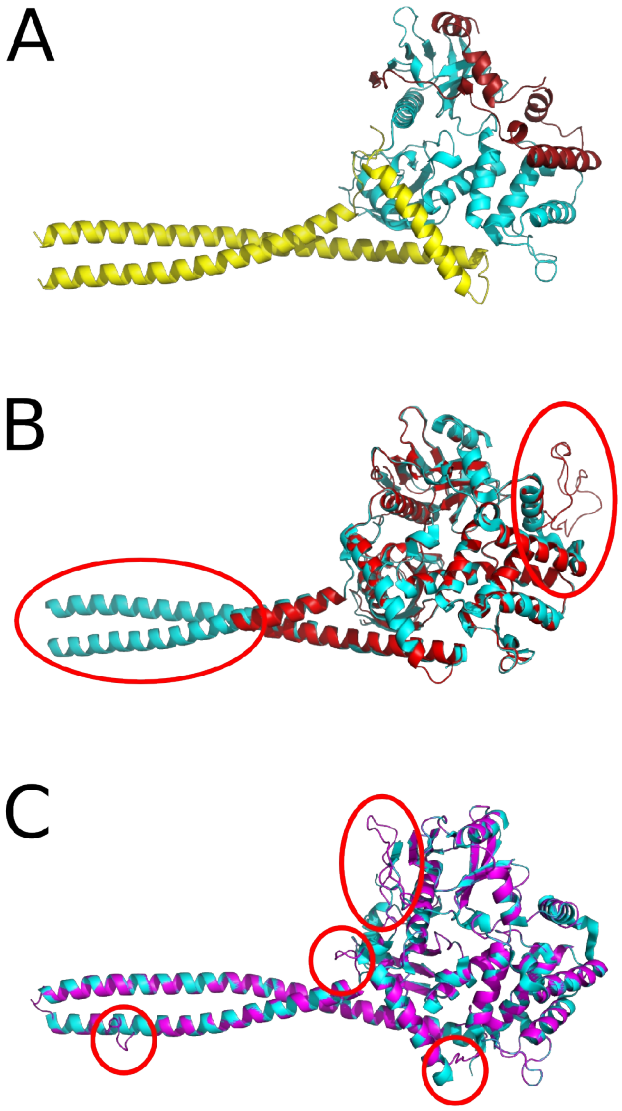


Figure 5.1. Comparison of models and the template used to construct them. (A) Our Rosetta model colored by domain: FERM (cyan), coiled-coil (yellow), and tail (red). (B) An overlay of the ModBase automatically created model (red) and the moesin template (cyan) shows large discrepancies (red circles). (C) An overlay of the SwissModel automatically created model (magenta) and the moesin template structure (cyan) also shows significant discrepancies (red circles).

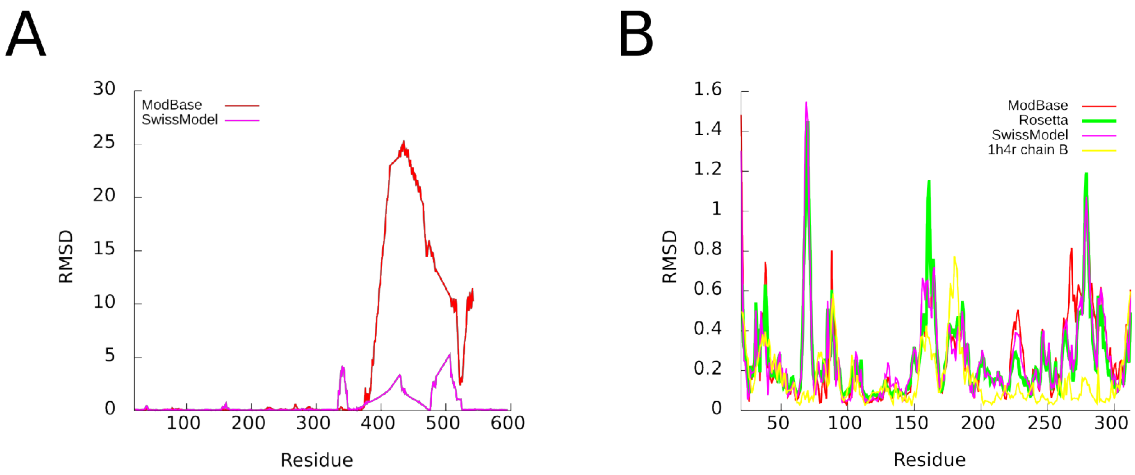


Figure 5.2. Quantitative comparison and validation of our homology model. (A) C_α differences along the chain between the two automated models and our model. Large differences reflect different sequence alignments and are nearly all outside of the highly similar FERM domain, which is easier to align accurately. Colors are indicated in the key. (B) C_α differences along the chains between the crystallographic merlin FERM domain and all models (our Rosetta model and the SwissModel and ModBase automated models) quantitates the errors compared to the experimentally determined structure. Chain B of the crystal structure is shown as a control of the differences that occur for the same protein in different environments. Colors are indicated in the key.

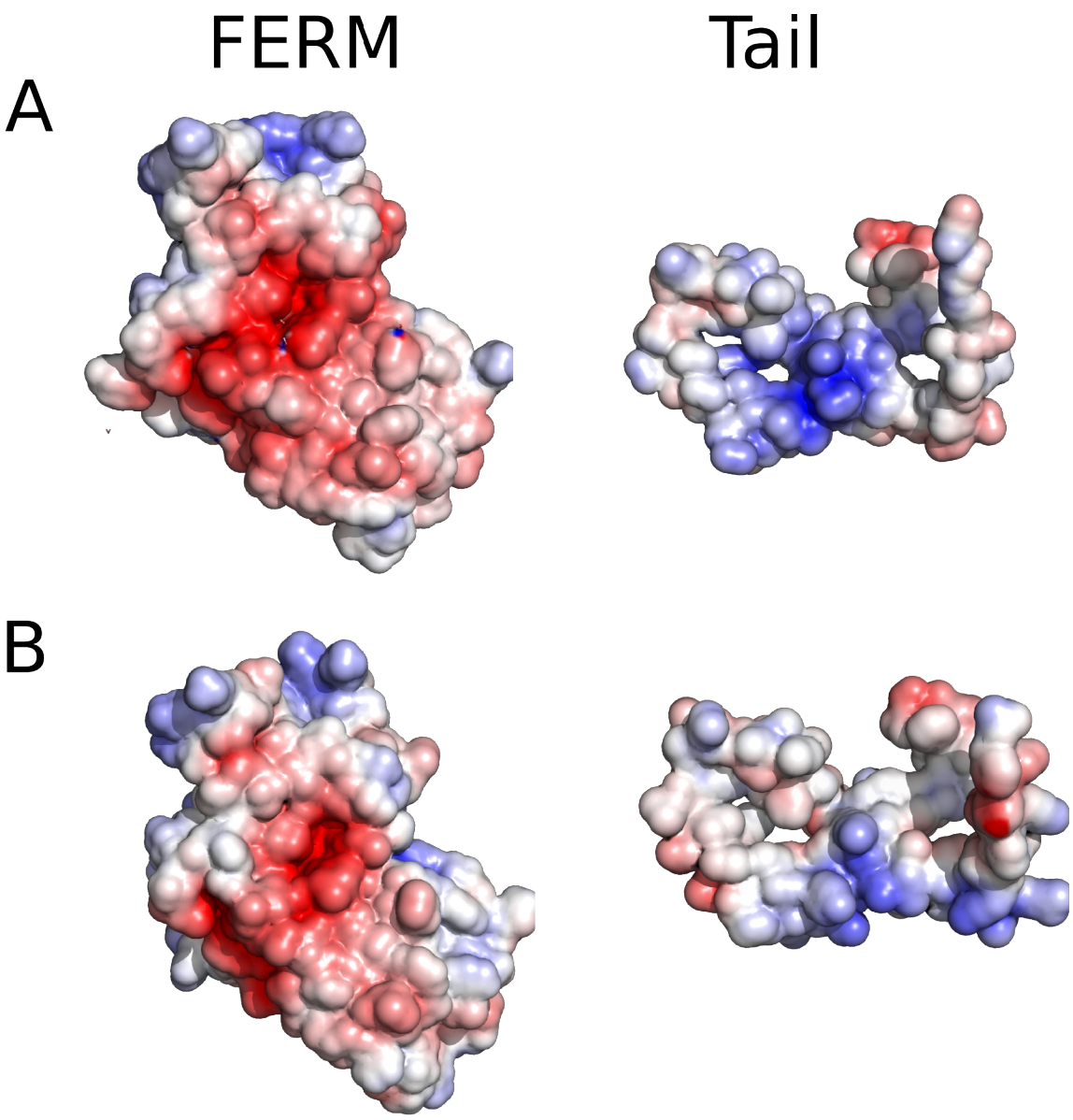


Figure 5.3. Structural basis for weaker closed form of merlin than moesin. The electrostatic potential surface of the inside of an opened FERM-tail interface is shown with the FERM domain on the left and the tail domain on the right. (A) Moesin shows a highly negative FERM face that interacts strongly with the highly negative mask of the tail domain. (B) Merlin shows lower charge potentials in both the FERM face and tail mask, resulting in a more weakly closed form.

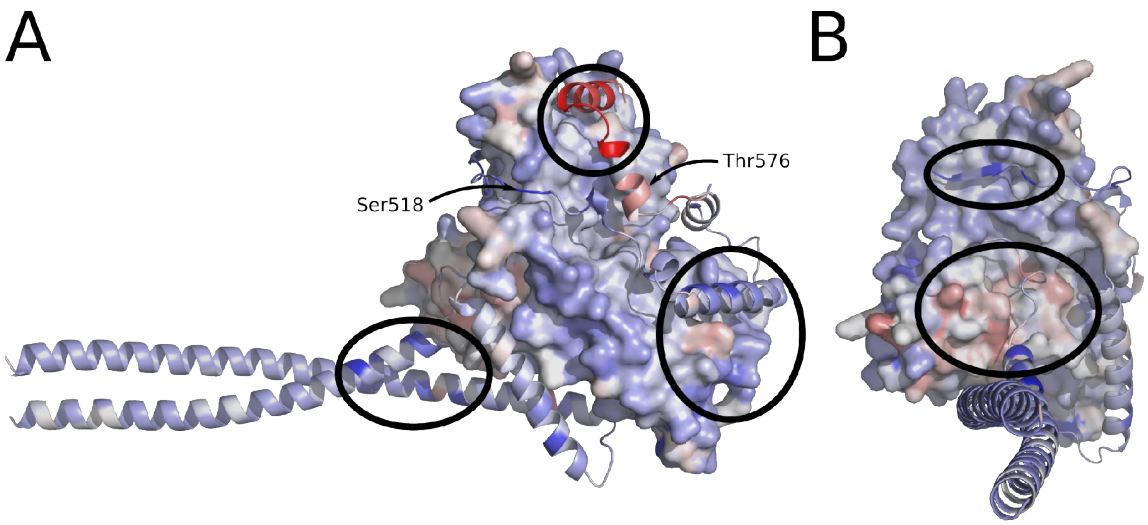


Figure 5.4. Loss-of-function and gain-of-function between merlin and other ERM proteins are revealed by differential conservation. Colors indicate the change in conservation between merlin and other ERM proteins, with red indicating loss of function, blue indicating gain of function, and whiter colors indicating little to no change. (A) A side view of merlin highlights three regions showing differential conservation. Ser518 is phosphorylated in merlin and shows a large gain of function relative to other ERM family members. Thr576 is the equivalent of Thr558 in human moesin, which is phosphorylated there, and it shows a loss of function in merlin. (B). A view head-on into the helical domain highlights two more regions showing differential conservation.

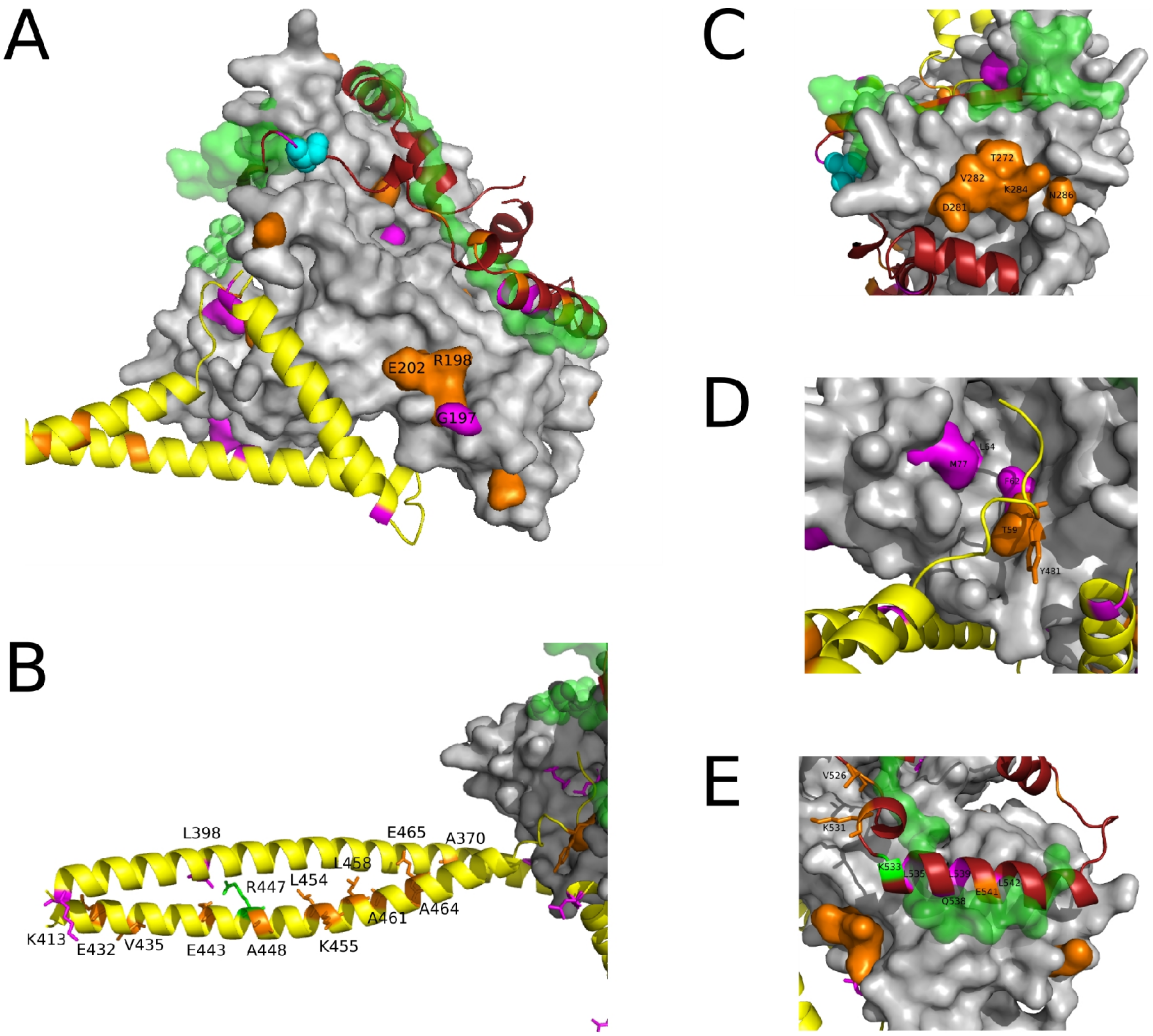


Figure 5.5. Residues implicated in conserved change-of-function between merlin and other ERM proteins and those involved in NF2-related missense mutations produce putative critical clusters, shown in A-E. Function-changing residues that are conserved in both merlin and other ERM proteins but with a different residue are shown in orange. Residues involved in NF2 missense mutations are shown in pink. ERM-interacting small molecules and peptides superimposed onto our merlin model are shown in semitransparent green (IP3, ICAM-2, EBP50). The helical domain is shown in yellow, the tail domain is shown in red, and the FERM domain is shown in gray.

Chapter 6

General Conclusions

Introduction

In this thesis, I have described a series of advances in accurate modeling and understanding of protein structure. Beginning with a focused study of a single protein at atomic resolution, types of information available at this level were illustrated through discovery of catalytically relevant structural insights. From there, the scope expanded to explain broader structural trends of atomic-resolution protein structure and show how they will improve high-accuracy modeling. Finally, accurate modeling techniques were applied to create a homology model that was used to formulate a number of biologically relevant insights, illustrating the types of information available from modeled protein structures.

In this chapter, I will consider how each component of this work contributed to broaden scientific knowledge. For each study in this thesis, I will present general conclusions, its impact upon the field, and an outlook of future studies to expand upon the insights gained here. I will then conclude with a brief discussion of my philosophy of science.

The catalytic importance of compression is supported by its visualization in atomic-resolution structures

This dissertation opened by claiming that details of protein structure on the scale of 0.1 Å regulate our understanding of catalysis, how mutations cause disease, and what makes a good inhibitor and potential drug. Here, I have proven that this is the case for an enzyme at atomic resolution, which revealed important new insights into structure and function that were invisible even at 2 Å resolution. I concluded in chapter 2 that extending known protein structures to atomic resolution for the flavoenzyme glutathione reductase (GR), a protein critical to the viability of malarial parasites, allows visualization of exquisite details that enable insights into the sources of catalytic power. First, compression in the

active site causes overlapping van der Waals radii and distortion in the substrate, which enhance catalysis via stereoelectronic effects. Second, the atoms involved in electron transfer between the substrate and enzyme are optimally positioned for the ensuing redox chemistry. Third, a redox-active disulfide loop is an extreme case of sequential peptide bonds systematically deviating from planarity, as revealed by the Protein Geometry Database. Another feature available at even higher resolution is reliable visualization of hydrogen atoms (e.g., Figure 1.2), but that was not possible in this case. One caution for ultrahigh-resolution structures is that the effects of synchrotron radiation can partially reduce a number of specific motifs, creating artifacts in the disulfide loop and potentially also in the flavin of the FAD cofactor, and these artifacts are not reliable models for naturally reduced forms. This impacts crystallographers attempting to model biologically relevant disulfide reductions using a synchrotron-reduced structure, because they will now be aware of the hazards involved and can take precautions, such as monitoring radiation damage over time.

These conclusions are highly relevant to anyone analyzing an atomic-resolution protein structure because they provide new avenues for investigation. Most other papers describing atomic-resolution structures devote little effort to investigating geometric distortions, which provide evidence of compression in catalysis. To illustrate the lack of appreciation for the importance of compression to enzymatic catalysis, a search on Google Scholar for the terms “catalysis” and “enzyme” produced more than 500,000 results, but adding the term “compression” reduced the results to less than 10,000 (a 98% decrease). However, the advent of quantum-mechanical simulations of proteins allows for an approach to compression and catalytic power from the theoretical side as well, and I expect this bidirectional attack upon the problem to be informative as well as convincing. The experimental data from this chapter show that distortions are biologically relevant because they are direct evidence of compression, and they require attention for a complete description of catalysis.

Future studies to continue this work by applying its general conclusions include enzymology to explore the role of compression biochemically as well as analyses of other

atomic-resolution structures that include a focus on compression and distorted geometry in substrates, cofactors, active-site residues, and other important regions. One example of this is underway in the Karplus lab, where we have solved atomic-resolution structures of ferredoxin-$NADP^+$ reductase (FNR). Like GR, FNR has a bound flavin cofactor and uses NADPH for reducing equivalents, so many of the same distortions that produced significant catalytic insights in GR could have equivalents in FNR. The concept of compression as a source of catalytic power, as evidenced by tight packing and geometric distortions, can be applied more generally in any protein structure where distortions are visible.

Structural trends at atomic resolution and a database for mining protein geometric features of protein structure

In chapter 3, I described the creation and use of the Protein Geometry Database, which is the first and only tool to enable easy and flexible exploration of the relationship between peptide geometry and conformation. The database provides capabilities invaluable in gaining a better understanding of protein structure as well as examining the importance of various distortions or conformations. A web server is online at http://pgd.science.oregonstate.edu/ and the underlying data and code are freely available to use and extend. The availability of the server allows any scientist to ask the same kinds of questions that have contributed to some recent papers for the Karplus lab, and the availability of the underlying code means that anyone can add new features to the database. Because we expect to release the code under an open-source license called the AGPL, anyone making an instance of the server publicly available anywhere must share their modifications so that the entire community benefits from an improved Protein Geometry Database.

In chapter 4, I refuted the paradigm that the peptide backbone has a single ideal geometry independent of context and showed that it instead varies systematically as a

function of the Φ and Ψ backbone dihedral angles, for which the groundwork was set by Lothar Schäfer and Andy Karplus. These trends have a rational, structural basis that is explicable by optimization of atomic overlaps and favorable electrostatic interactions. To ease adoption of this new paradigm, I created a conformation-dependent library of backbone geometry, which improves accuracy over existing methods with negligible cost. This library represents the first step toward the new paradigm of "ideal-geometry functions." With much-improved agreement to ultrahigh-resolution crystal structures, ideal-geometry functions provide an intellectually satisfying resolution to the debate among crystallographers regarding the correct ideal values to use during refinement. Protein structures derived from both crystallographic refinement and predictive modeling are expected to benefit from conversion to the new paradigm. I and others from the Karplus lab are working to bring this new paradigm into the most popular modeling programs with the expectation that developers of other programs will also implement it to stay competitive.

Future studies to further explore these trends in covalent geometry are ongoing in four directions: incorporation of the conformation-dependent library in refinement programs, to test its real impact; examination of nonplanarity of the peptide bond; continuation of the analysis of backbone-geometry trends to further separate residues based on their behavior and extend the database and analysis to additional side-chain information; and a second-generation version of the conformation-dependent library that uses the kernel-regression data instead of simple binned averages, and that may be split into additional residue classes based on further analysis.

Incorporation of the conformation-dependent library into crystallographic refinement programs

Dale Tronrud in the Karplus lab has incorporated the library into the crystallographic refinement program TNT and tested its impact. The results show that our library produces

an improvement on the same scale as that produced by the migration to the last-generation library from its predecessor. To briefly sum up, the geometry library is much closer to the experimental structure than the prior library was, and there is also a small improvement in the crystallographic R-factors. A manuscript describing this work is in preparation for submission to Acta Cryst D, and I will be second author. Additionally, I have a preliminary version of the integration completed in another refinement program, Phenix, which I expect to continue working on after leaving OSU. Phenix has a number of advantages that enable additional types of analysis, such as large-scale testing of the library's impact when re-refining every known protein structure. Because our analysis using TNT relied on picking a small subset of structures, this large-scale experiment could show quite different results, and I expect it will be far more convincing because of the sheer numbers of structures involved.

Examination of nonplanarity of the peptide bond

The work in chapter 4 purposely left out analysis of peptide planarity, defined by the ω torsion angle. Because of a number of complicating factors and because its results are a distraction from the remainder of the backbone-geometry story, ω was left for an independent analysis. Complicating factors include the dipeptide dependence of ω because the peptide bond is shared equally by two residues and a much larger collection of previous literature on this specific parameter, which made the story less clear. Additionally, its conformation-dependent variations are on the scale of 15°–20°, dwarfing the variations in bond angles and lengths; this would cause readers to skim over the larger part of the research in chapter 4 because of its smaller scale.

As it is a separate story about a better-known parameter, we are pursuing a more in-depth analysis of the ω angle that uses the most extreme deviations from planarity as a tool to understand the rest, and I will describe our preliminary results. First, we have shown that the true nonplanarity is masked until resolutions as high as 1.0 Å and even

beyond, so analyses on the lower-resolution structures available until recently would not reflect reality. Second, we have been examining its dipeptide dependence; interestingly, it seems that the conformations of two peptides does little beyond a single-peptide conformation to cause ω to deviate (accounting for ~6°–8°), so other factors such as hydrogen bonding or atomic clashes must be involved to create a 25°-degree deviation. Third, we have examined the connection between a highly nonplanar ω and functional significance of the surrounding residues. Intriguingly, I have found that despite popular opinion, there is no enrichment of highly nonplanar peptide bonds in residues with functional roles. Comparison of 116 residues that have 20°-degree ω deviation with more than 500 randomized controls revealed a miniscule difference between the chance that a nonplanar vs control residue would be functionally important. Since it's not functional, and there is a clear energetic cost to bending the peptide bond this far from planarity, there must be some sort of structural role. Fourth, we are examining the conservation of peptide nonplanarity across homologous proteins, although it is too early to know the results. Based on all of these preliminary results, I am writing a short, first-author manuscript, and we hope to finish the research for this work and get the manuscript near submission by the time I leave OSU.

Continued in-depth analysis of protein geometry trends and conversion of the library to finer-grained classes

Another direction in which this project will continue is further analysis of the trends in backbone geometry, based on splitting the residue types into further classes. The current group of residue classes is fairly minimal, and the general class contains 16 residues that could probably be further separated into subsets based on their covalent geometry. One approach for this would be to perform a separate analysis for each of the 21 residue types (the 20 amino acids and preproline), then cluster the ones that behave the same. This could be limited by the lack of sufficient observations, particularly for rarer amino acids such as cysteine and methionine. Since these trends are only available at sufficiently high

resolution, this would require that we simply wait until more structures become available.

This clustering can become very complex; it could be that for each type of backbone angle or length, a different set of residue classes would exist—for example, a β-branched residue may cause certain angles to distort but others could behave just like general residues. A step beyond that, different classes could exist for different conformations, because two residues could behave the same in one region but not another because different atoms clash in the two regions. For the actual comparison of behavior to determine whether two residues were sufficiently different, I would calculate the difference between their averages for a given conformation, then ask whether that difference was larger than the sum of the standard errors of the means. If it was, that conformation of that parameter behaves differently for the two residues, and based on the type of classes used, they should be somehow separated.

A related direction is initial analysis of trends in side-chain geometry. The Protein Geometry Database only contains a single side-chain torsion angle at present (χ_1), so it requires extensions before additional analysis can be performed. These extensions are already planned, and they will take place in three stages. The first, which is nearly working, is the addition of the other χ torsion angles. The second will be the side-chain covalent geometry containing angles and lengths of the atoms that define the χ angles. The third will be the addition of any other angles and lengths that we decide we need. And naturally, the library will be extended to contain the new database parameters. This work will happen as part of the Karplus lab's new collaboration with Roland Dunbrack's lab, initiated by the work in chapter 4. The Dunbrack lab has a major focus on side-chain geometry, and they will take the lead on transferring the same techniques used in chapter 4 to this work.

Insights into function of a tumor suppressor from a model and a novel method to compare residue conservation

In chapter 5, I created a homology model of the tumor-suppressor merlin using the background knowledge gained from earlier research. I gained a number of insights into the protein's function using approaches based on structural clusters when various properties such as conservation and charge were mapped onto the structure. Because of the lack of specificity for most experimentally determined merlin interactions, it was often impossible to propose a one-to-one mapping of the clusters to binding partners. The higher level of accuracy in this study can guide follow-up mutational studies.

Additionally, I developed a method to discover clustered gains and losses of function given two protein families. I described the method in detail in chapter 5, but here is a summary: (1) generate a sequence alignment for each subfamily, with no overlapping sequences between the two alignments, and with at least one sequence in each alignment having a known structure; (2) using the alignments, calculate a conservation score for each residue of each known structure; (3) using a structural alignment of the two structures, calculate the differences in conservation score of equivalent residues—this indicates gains and losses of conservation and thus likely function; and (4) map those differences onto the structure and look for clusters. The final step could be replaced with a more easily automated but less sensitive method that just looks for sequence-based proximity (e.g., by calculating the conservation difference smoothed across five residues). I also developed but was unable to perfect a slightly different application of a similar method to compare electrostatic potentials between merlin and its homolog moesin, with the same theme of directly visualizing the differences that are truly desired instead of inferring them from looking at two structures simultaneously. Like the conservation method, this one should also be extremely useful but needs additional work to perfect.

To build upon the results of this work, efforts to crystallographically determine the structure of merlin are underway in the Karplus lab. Another part of the same project, in

collaboration with Tony Bretscher's lab at Cornell, involves mutations to strengthen the interface between the FERM and tail domains, and the structural basis of the interface strength discovered here as well as key conserved clusters on the tail will guide these mutations. The mutants may even allow for easier crystallization by immobilizing the tail domain through a strengthened interface. Additionally, the same project aims to discover specific partners interacting with merlin and disrupt those interactions with mutations; the work here provides unparalleled precision and accuracy of which residues we propose contribute to merlin's distinct function from other ERM proteins. Finally, the novel methods developed here should prove generally useful for any two protein subfamilies with distinct functions.

Final statements

My scientific approach to this work has some unifying general philosophies. First, this work has a focus on unifying theory and experiment. This happens in chapters 2 and 4 through comparison of our experimental data with quantum mechanical calculations, and it happens in chapter 5 through comparison of the homology model (theoretical) with existing experimental data and proposal of how theory could inform further experiments.

Second, this work has a focus on using new or underappreciated methods of analysis, particularly those that are systematic and can be automated. This allows me to answer key questions that would be difficult or impossible to answer any other way and to perform large numbers of analyses very easily. In chapter 2, the core of the research is a detailed examination of geometric distortions; this approach is very uncommon even for atomic-resolution structures. In chapter 3, I described a new database designed specifically to allow asking questions about backbone geometry that were essentially unanswerable before. Chapter 4 used a combination of new methods (kernel regressions) and automated methods (the jack-knifing tests of the library and the test of crystallographic resolution vs RMSD of bond angles from the library). In chapter 5, I

developed the residue-conservation method that proved invaluable to proposing key regions of merlin.

Third, this work focuses on directly examining what I wanted to compare, instead of performing an indirect comparison across multiple graphs or structures simultaneously. This often manifests as a direct visualization or calculation of a difference instead of showing two plots and forcing the reader to imagine the difference between them. In chapter 4, this is visible in the many RMSD tests vs the library or vs an atomic-resolution crystal structure. One of the best examples of this principle is in chapter 5, when I directly visualize losses and gains of function on a single structure instead of trying to imagine what they would look like by examining two structures at once. However, this technique must be tempered by considering that sometimes the original data are more convincing than just the differences between them, although looking at the differences does aid understanding.

I expect to apply these philosophies throughout the remainder of my career.

Bibliography

Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Read, R.J., Sacchettini, J.C., Sauter, N.K., and Terwilliger, T.C. (2002). PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr. D. Biol. Crystallogr. *58*, 1948-1954.

Almarsson, O., and Bruice, T.C. (1993). Evaluation of the factors influencing reactivity and stereospecificity in NAD(P)H dependent dehydrogenase enzymes. Journal of the American Chemical Society *115*, 2125-2138.

Alphey, M.S., Attrill, H., Crocker, P.R., and van Aalten, D.M.F. (2003). High resolution crystal structures of Siglec-7: Insights Into ligand specificity in the Siglec family. Journal of Biological Chemistry *278*, 3372-3377.

Argyrou, A., Blanchard, J.S., and Palfey, B.A. (2002). The lipoamide dehydrogenase from Mycobacterium tuberculosis permits the direct observation of flavin intermediates in catalysis. Biochemistry *41*, 14580-14590.

Baker, N.A., Sept, D., Joseph, S., Holst, M.J., and McCammon, J.A. (2001). Electrostatics of nanosystems: Application to microtubules and the ribosome. Proceedings of the National Academy of Sciences of the United States of America *98*, 10037-10041.

Baldwin, R.L., and Rose, G.D. (1999). Is protein folding hierarchic? I. Local structure and peptide folding. Trends Biochem. Sci. *24*, 26-33.

Benner, S.A. (1982). The stereoselectivity of alcohol dehydrogenases: A stereochemical imperative? Cellular and Molecular Life Sciences (CMLS) *38*, 633-637.

Berglund, G.I., Carlsson, G.H., Smith, A.T., Szöke, H., Henriksen, A., and Hajdu, J. (2002). The catalytic pathway of horseradish peroxidase at high resolution. Nature *417*, 463-8.

Berkholz, D.S., Faber, H.R., Savvides, S.N., and Karplus, P.A. (2008). Catalytic cycle of human glutathione reductase near 1 A resolution. J. Mol. Biol. *382*, 371-384.

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.

Böhme, C.C., Arscott, L.D., Becker, K., Schirmer, R.H., and Williams, C.H. (2000). Kinetic characterization of glutathione reductase from the malarial parasite Plasmodium falciparum. Comparison with the human enzyme. Journal of Biological Chemistry *275*, 37317-23.

Bradley, P., Misura, K.M.S., and Baker, D. (2005a). Toward high-resolution de novo structure prediction for small proteins. Science *309*, 1868-1871.

Bradley, P., Misura, K.M.S., and Baker, D. (2005b). Toward high-resolution de novo structure prediction for small proteins. Science *309*, 1868-1871.

Bretscher, A., Edwards, K., and Fehon, R.G. (2002). ERM proteins and merlin: integrators at the cell cortex. Nature Reviews. Molecular Cell Biology *3*, 586-599.

Brooks, B.R., Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. *4*, 187-217.

Bruice, T., and Lightstone, F. (1999). Ground state and transition state contributions to the rates of intramolecular and enzymatic reactions. Accounts of Chemical Research *32*, 127-136.

Bruice, T.C., and Pandit, U.K. (1960). Intramolecular models depicting the kinetic importance of "fit" in enzymatic catalysis. Proceedings of the National Academy of Sciences *46*, 402-404.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallographica Section D Biological Crystallography *54*, 905-21.

Burmeister, W.P. (2000). Structural changes in a cryo-cooled protein crystal owing to radiation damage. Acta Crystallographica Section D Biological Crystallography *56*, 328-341.

Carugo, O., and Carugo, K.D. (2005). When X-rays modify the protein structure: radiation damage at work. Trends in Biochemical Sciences *30*, 213-219.

Cavelier, G., and Amzel, L.M. (2001). Mechanism of NAD (P) H: Quinone reductase: Ab initio studies of reduced flavin. Proteins Structure Function and Genetics *43*, 420-432.

Chakravarty, S., and Sanchez, R. (2004). Systematic analysis of added-value in simple comparative models of protein structure. Structure (London, England: 1993) *12*, 1461-1470.

Chothia, C., and Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. The EMBO Journal *5*, 823-826.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J.L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics.

Bioinformatics *25*, 1422-1423.

Congreve, M., Murray, C.W., and Blundell, T.L. (2005). Structural biology and drug discovery. Drug Discov. Today *10*, 895-907.

Corey, R.B., and Donohue, J. (1950). Interatomic distances and bond angles in the polypeptide chain of proteins. J. Am. Chem. Soc. *72*, 2899-2900.

Cruickshank, D.W.J. (2001). International Tables for Crystallography. In International Tables for Crystallography, M G Rossmann, and E Arnold, eds. (: Dordrecht: Kluwer Academic Publishers), pp. 403-418.

Dauter, Z., Lamzin, V.S., and Wilson, K.S. (1995). Proteins at atomic resolution. Curr. Opin. Struct. Biol. *5*, 784-790.

Dauter, Z., Lamzin, V.S., and Wilson, K.S. (1997). The benefits of atomic resolution. Current Opinion in Structural Biology *7*, 681-688.

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., Snoeyink, J., Richardson, J.S., and Richardson, D.C. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Research *35*, W375-383.

De Colibus, L., and Mattevi, A. (2006). New frontiers in structural flavoenzymology. Current Opinion in Structural Biology *16*, 722-728.

DeLano, W.L. (2002). The PyMOL Molecular Graphics System (: DeLano Scientific, San Carlos, CA).

Dunbrack, R.L., and Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat. Struct. Biol. *1*, 334-340.

Dunlop, K.V., Irvin, R.T., and Hazes, B. (2005). Pros and cons of cryocrystallography: should we also collect a room-temperature data set? Acta Crystallogr. D Biol. Crystallogr. *61*, 80-87.

Emsley, P., and Cowtan, K. (2004a). Coot: model-building tools for molecular graphics. Acta Crystallogr. D. Biol. Crystallogr. *60*, 2126-2132.

Emsley, P., and Cowtan, K. (2004b). Coot: model-building tools for molecular graphics. Acta Crystallographica. Section D, Biological Crystallography *60*, 2126-2132.

Engh, R.A., and Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallogr. A Found. Crystallogr. *47*, 392-400.

Engh, R.A., and Huber, R. (2001). International Tables for Crystallography.
In International Tables for Crystallography, M.G. Rossmann, and E.

Arnold, eds. (Dordrecht, The Netherlands: Kluwer Academic Publishers), pp. 382-392.

Esposito, L., Vitagliano, L., Zagari, A., and Mazzarella, L. (2000). Experimental evidence for the correlation of bond distances in peptide groups detected in ultrahigh-resolution protein structures. Protein Eng. *13*, 825-828.

Evans, D.G.R., Moran, A., King, A., Saeed, S., Gurusinghe, N., and Ramsden, R. (2005). Incidence of vestibular schwannoma and neurofibromatosis 2 in the North West of England over a 10-year period: higher incidence than previously thought. Otology & Neurotology: Official Publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology *26*, 93-97.

Evans, P.R. (2007). An introduction to stereochemical restraints. Acta Crystallogr. D. Biol. Crystallogr. *63*, 58-61.

Feig, M. (2008). Is alanine dipeptide a good model for representing the torsional preferences of protein backbones? J. Chem. Theory Comput. *4*, 1555-1564.

Fox, K.M., and Karplus, P.A. (1999). The flavin environment in old yellow enzyme: An evaluation of insights from spectroscopic and artificial flavin studies. Journal of Biological Chemistry *274*, 9357-9362.

Fraaije, M.W., and Mattevi, A. (2000). Flavoenzymes: diverse catalysts with recurrent features. Trends in Biochemical Sciences *25*, 126-132.

Garman, E.F., and Owen, R.L. (2006). Cryocooling and radiation damage in macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. *62*, 32-47.

Ghisla, S., and Massey, V. (1989). Mechanisms of flavoprotein-catalyzed reactions. FEBS Journal *181*, 1-17.

G. L. Miessler, and D. A. Tarr (2004). Inorganic Chemistry. In Inorganic Chemistry (: Pearson Prentice Hall), pp. 116-164.

Gunasekaran, K., Ramakrishnan, C., and Balaram, P. (1996). Disallowed Ramachandran conformations of amino acid residues in protein structures. J. Mol. Biol. *264*, 191-198.

Heine, A., DeSantis, G., Luz, J.G., Mitchell, M., Wong, C., and Wilson, I.A. (2001). Observation of covalent intermediates in an enzyme mechanism at atomic resolution. Science *294*, 369-374.

Heine, A., Luz, J.G., Wong, C., and Wilson, I.A. (2004). Analysis of the class I aldolase binding site architecture based on the crystal structure of 2-deoxyribose-5-phosphate aldolase at 0.99 Å resolution. Journal of Molecular Biology *343*, 1019-1034.

Hobohm, U., and Sander, C. (1994). Enlarged representative set of protein structures. Protein Sci. *3*, 522-524.

Hollingsworth, S.A., Berkholz, D.S., and Karplus, P.A. (2009). On the occurrence of linear groups in proteins. Protein Sci. *18*, 1321-1325.

Holmes, J.B., and Tsai, J. (2004). Some fundamental aspects of building protein structures from fragment libraries. Protein Sci. *13*, 1636–1650.

Huber, P.W., and Brandt, K.G. (1980). Kinetic studies of the mechanism of pyridine nucleotide dependent reduction of yeast glutathione reductase. Biochemistry *19*, 4568-4575.

Hurley, J.H., Mason, D.A., and Matthews, B.W. (1992). Flexible-geometry conformational energy maps for the amino acid residue preceding a proline. Biopolymers *32*, 1443-1446.

Iris Ahronowitz, Winnie Xin, Rosemary Kiely, Katherine Sims, Mia MacCollin, and Fabio P. Nunes (2007). Mutational spectrum of the NF2 gene: a meta-analysis of 12 years of research and diagnostic laboratory findings. Human Mutation *28*, 1-12.

Jaskolski, M., Gilski, M., Dauter, Z., and Wlodawer, A. (2007a). Numerology versus reality: a voice in a recent dispute. Acta Crystallogr. D. Biol. Crystallogr. *63*, 1282-1283.

Jaskolski, M., Gilski, M., Dauter, Z., and Wlodawer, A. (2007b). Numerology versus reality: a voice in a recent dispute. Acta Crystallogr. D Biol. Crystallogr. *63*, 1282-1283.

Jaskolski, M., Gilski, M., Dauter, Z., and Wlodawer, A. (2007c). Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? Acta Crystallogr. D. Biol. Crystallogr. *63*, 611-620.

Jaskolski, M., Gilski, M., Dauter, Z., and Wlodawer, A. (2007d). Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? Acta Crystallogr. D Biol. Crystallogr. *63*, 611-620.

Jelsch, C., Teeter, M.M., Lamzin, V., Pichon-Pesme, V., Blessing, R.H., and Lecomte, C. (2000). Accurate protein crystallography at ultra-high resolution: valence electron distribution in crambin. Proc. Natl. Acad. Sci. USA *28*, 3171-3176.

Jiang, X., Yu, C., Cao, M., Newton, S.Q., Paulus, E.F., and Schäfer, L. (1997). ø/ψ-Torsional dependence of peptide backbone bond-lengths and bond-angles: comparison of crystallographic and calculated parameters. J. Mol. Struct. *403*, 83-93.

Jones, T.A. (1978). A graphics model building and refinement system for macromolecules. Journal of Applied Crystallography *11*, 268-272.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern

recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577-637.

Kang, B.S., Cooper, D.R., Devedjiev, Y., Derewenda, U., and Derewenda, Z.S. (2002). The structure of the FERM domain of merlin, the neurofibromatosis type 2 gene product. Acta Crystallographica. Section D, Biological Crystallography *58*, 381-391.

Kang, B.S., Devedjiev, Y., Derewenda, U., and Derewenda, Z.S. (2004). The PDZ2 domain of syntenin at ultra-high resolution: bridging the gap between macromolecular and small molecule crystallography. J. Mol. Biol. *338*, 483-493.

Karplus, P.A. (1999). Flavins and flavoproteins 1999. In Flavins and flavoproteins 1999, S. Ghisla, P. Kroneck, P. Macheroux, and H. Sund, eds. (: Agency for Scientific Publ.), pp. 233-238.

Karplus, P.A. (1996). Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci. *5*, 1406-1420.

Karplus, P.A., Pai, E.F., and Schulz, G.E. (1989). A crystallographic study of the glutathione binding site of glutathione reductase at 0.3-nm resolution. Eur J Biochem *178*, 693-703.

Karplus, P.A., and Schulz, G.E. (1987). Refined structure of glutathione reductase at 1.54 Å resolution. Journal of Molecular Biology *195*, 701-29.

Karplus, P.A., and Schulz, G.E. (1989). Substrate binding and catalysis by glutathione reductase as derived from refined enzyme: substrate crystal structures at 2 Å resolution. Journal of Molecular Biology *210*, 163-180.

Karplus, P., Shapovalov, M., Dunbrack Jr, R., and Berkholz, D.S. (2008a). A forward-looking suggestion for resolving the stereochemical restraints debate: ideal geometry functions. Acta Crystallogr. D. Biol. Crystallogr. *64*, 335-336.

Karplus, P.A., Shapovalov, M.V., Dunbrack, R.L., and Berkholz, D.S. (2008b). A forward-looking suggestion for resolving the stereochemical restraints debate: ideal geometry functions. Acta Crystallogr. D Biol. Crystallogr. *64*, 335-336.

Khan, H., Barna, T., Harris, R.J., Bruce, N.C., Barsukov, I., Munro, A.W., Moody, P.C.E., and Scrutton, N.S. (2004). Atomic resolution structures and solution behavior of enzyme-substrate complexes of Enterobacter cloacae PB2 pentaerythritol tetranitrate reductase: Multiple conformational states and implications for the mechanism of nitroaromatic explosive degradation. Journal of Biological Chemistry *279*, 30563-30572.

Kleywegt, G.J., and Jones, T.A. (1996). Phi/psi-chology: Ramachandran revisited. Structure *4*, 1395-1400.

Kopp, J., and Schwede, T. (2006). The SWISS-MODEL Repository: new features and

functionalities. Nucleic Acids Research *34*, D315-318.

Kort, R., Hellingwerf, K.J., and Ravelli, R.B.G. (2004a). Initial events in the photocycle of photoactive yellow protein. Journal of Biological Chemistry *279*, 26417-26424.

Kort, R., Komori, H., Adachi, S., Miki, K., and Eker, A. (2004b). DNA apophotolyase from Anacystis nidulans: 1.8 Å structure, 8-HDF reconstitution and X-ray-induced FAD reduction. Acta Crystallogr D Biol Crystallogr *60*, 1205-1213.

Krauth-Siegel, R.L., Arscott, L.D., Schoenleben-Janas, A., Schirmer, R.H., and Williams, C.H. (1998). Role of active site tyrosine residues in catalysis by human glutathione reductase. Biochemistry *37*, 13968-13977.

Laidig, K.E., and Cameron, L.M. (1993). What happens to formamide during C—N bond rotation? Atomic and molecular energetics and molecular reactivity as a function of internal rotation. Can. J. Chem. *71*, 872-879.

LaJeunesse, D.R., McCartney, B.M., and Fehon, R.G. (1998). Structural analysis of Drosophila merlin reveals functional domains important for growth control and subcellular localization. The Journal of Cell Biology *141*, 1589-1599.

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. Nuc. Acids Res. *33*, W299-302.

Lario, P.I., Sampson, N., and Vrielink, A. (2003). Sub-atomic resolution crystal structure of cholesterol oxidase: What atomic resolution crystallography reveals about enzyme mechanism and the role of the FAD cofactor in redox activity. Journal of Molecular Biology *326*, 1635-1650.

Laskowski, R.A., Chistyakov, V.V., and Thornton, J.M. (2005). PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. Nuc. Acids Res. *33*, D266-D268.

Lee, M.S., Feig, M., Salsbury, F.R., and Brooks, C.L. (2003). New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. J. Comput. Chem. *24*, 1348-1356.

Lennon, B.W., Williams, C.H., and Ludwig, M.L. (1999). Crystal structure of reduced thioredoxin reductase from Escherichia coli: structural flexibility in the isoalloxazine ring of the flavin adenine dinucleotide cofactor. Protein science : a publication of the Protein Society *8*, 2366-2379.

Liotta, C.L., Burgess, E.M., and Eberhardt, W.H. (1984). Trajectory analysis. 1. Theoretical model for nucleophilic attack at pi-systems. The stabilizing and destabilizing orbital terms. Journal of the American Chemical Society *106*, 4849-4852.

Li, Q., Nance, M.R., Kulikauskas, R., Nyberg, K., Fehon, R., Karplus, P.A., Bretscher, A., and Tesmer, J.J.G. (2007). Self-masking in an intact ERM-merlin protein: an active role for the central alpha-helical domain. Journal of Molecular Biology *365*, 1446-1459.

Livingstone, C.D., and Barton, G.J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput. Appl. Biosci. *9*, 745-756.

Longhi, S., Czjzek, M., and Cambillau, C. (1998). Messages from ultrahigh resolution crystal structures. Curr. Opin. Struct. Biol. *8*, 730-737.

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: ϕ, ψ and Cβ deviation. Proteins: Struct. Func. Genet. *50*, 437-450.

MacCollin, M., Mohney, T., Trofatter, J., Wertelecki, W., Ramesh, V., and Gusella, J. (1993). DNA diagnosis of neurofibromatosis 2. Altered coding sequence of the merlin tumor suppressor in an extended pedigree. JAMA: The Journal of the American Medical Association *270*, 2316-2320.

Mackerell, A.D. (2004). Empirical force fields for biological macromolecules: overview and issues. J. Comput. Chem. *25*, 1584-1604.

Mardia, K.V., and Zemroch, P.J. (1975). Algorithm AS 86: The Von Mises distribution function. Applied Statistics *24*, 268-272.

Massey, V. (1995). Introduction: flavoprotein structure and mechanism. FASEB Journal *9*, 473-5.

Massey, V. (2000). The chemical and biological versatility of riboflavin. Biochemical Society Transactions *28*, 283-296.

Matthews, R.G., Ballou, D.P., and Williams, C.H. (1979). Reactions of pig heart lipoamide dehydrogenase with pyridine nucleotides. Evidence for an effector role for bound oxidized pyridine nucleotide. The Journal of biological chemistry *254*, 4974-81.

McClatchey, A.I., Saotome, I., Ramesh, V., Gusella, J.F., and Jacks, T. (1997). The Nf2 tumor suppressor gene product is essential for extraembryonic development immediately prior to gastrulation. Genes & Development *11*, 1253-1265.

McClatchey, A.I., and Fehon, R.G. (2009). Merlin and the ERM proteins--regulators of receptor distribution and signaling at the cell cortex. Trends Cell Biol. *19*, 198-206.

Merritt, E.A. (1999). Expanding the model: anisotropic displacement parameters in protein structure refinement. Acta Crystallographica Section D Biological Crystallography *55*, 1109-1117.

Mesecar, A.D., Stoddard, B.L., and Koshland, D.E. (1997). Orbital steering in the

catalytic power of enzymes: small structural changes with large catalytic consequences. Science (New York, N.Y.) *277*, 202-206.

Miller, M., Schneider, J., Sathyanarayana, B.K., Toth, M.V., Marshall, G.R., Clawson, L., Selk, L., Kent, S.B., and Wlodawer, A. (1989). Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution. Science *246*, 1149-1152.

Miller, S.M., Massey, V., Ballou, D., Williams, C.H., Distefano, M.D., Moore, M.J., and Walsh, C.T. (1990). Use of a site-directed triple mutant to trap intermediates: demonstration that the flavin C(4a)-thiol adduct and reduced flavin are kinetically competent intermediates in mercuric ion reductase. Biochemistry *29*, 2831-2841.

Miura, R. (2001). Versatility and specificity in flavoenzymes: Control mechanisms of flavin reactivity. The Chemical Record *1*, 183-194.

Moffat, K., and Ren, Z. (1997). Synchrotron radiation applications to macromolecular crystallography. Curr. Opin. Struct. Biol. *7*, 689-696.

Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallographica. Section D, Biological Crystallography *53*, 240-255.

Nambiar, K.P., Stauffer, D.M., Kolodziej, P.A., and Benner, S.A. (1983). A mechanistic basis for the stereoselectivity of enzymic transfer of hydrogen from nicotinamide cofactors. Journal of the American Chemical Society *105*, 5886-5890.

Naradaya, E. (1964). On estimating regression. Theory of Probability and its Applications *9*, 141–142.

Nguyen, R., Reczek, D., and Bretscher, A. (2001). Hierarchy of merlin and ezrin N- and C-terminal domain interactions in homo- and heterotypic associations and their relationship to binding of scaffolding proteins EBP50 and E3KARP. J. Biol. Chem. *276*, 7621-7629.

Okada, T., You, L., and Giancotti, F.G. (2007). Shedding light on Merlin's wizardry. Trends in Cell Biology *17*, 222-229.

Ondetti, M.A., Rubin, B., and Cushman, D.W. (1977). Design of specific inhibitors of angiotensin-converting enzyme: new class of orally active antihypertensive agents. Science *196*, 441-444.

Otwinowski, Z., and Minor, W. (1997). Processing of X-Ray Diffraction Data Collected in Oscillation Mode. Academic Press, New York, 307–326.

Painter, J., and Merritt, E.A. (2006). TLSMD web server for the generation of multi-group TLS models. Journal of Applied Crystallography *39*, 109-111.

Pauling, L., Corey, R.B., and Branson, H.R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. USA *37*, 205-211.

Pearson, M.A., Reczek, D., Bretscher, A., and Karplus, P.A. (2000). Structure of the ERM protein moesin reveals the FERM domain fold masked by an extended actin binding tail domain. Cell *101*, 259-270.

Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M.S., Davis, F.P., Stuart, A.C., Mirkovic, N., Rossi, A., Marti-Renom, M.A., Fiser, A., et al. (2004). MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Research *32*, D217-222.

Plewniak, F., Bianchetti, L., Brelivet, Y., Carles, A., Chalmel, F., Lecompte, O., Mochel, T., Moulinier, L., Muller, A., Muller, J., et al. (2003). PipeAlign: A new toolkit for protein family analysis. Nuc. Acids Res. *31*, 3829-3832.

Rajagopalan, P.T.R., and Benkovic, S.J. (2002). Preorganization and protein dynamics in enzyme catalysis. The Chemical Record *2*, 24-36.

Ramakrishnan, C., Lakshmi, B., Kurien, A., Devipriya, D., and Srinivasan, N. (2007). Structural compromise of disallowed conformations in peptide and protein structures. Protein Pept. Lett. *14*, 672-682.

Rasmussen, B.F., Stock, A.M., Ringe, D., and Petsko, G.A. (1992). Crystalline ribonuclease A loses function below the dynamical transition at 220 K. Nature *357*, 423-424.

Ravelli, R.B.G., and Garman, E.F. (2006). Radiation damage in macromolecular cryocrystallography. Curr. Opin. Struct. Biol. *16*, 624-629.

Ravelli, R.B.G., and McSweeney, S.M. (2000). The ‘fingerprint’ that X-rays can leave on structures. Structure *8*, 315-328.

Rietveld, P., Arscott, L.D., Berry, A., Scrutton, N.S., Deonarain, M.P., Perham, R.N., and Williams Jr, C.H. (1994). Reductive and oxidative half-reactions of glutathione reductase from Escherichia coli. Biochemistry *33*, 13888-13895.

Rivas, P., Zapata-Torres, G., Melin, J., and Contreras, R. (2004). Probing the hydride transfer process in the lumiflavine–1-methylnicotinamide model system using group softness. Tetrahedron *60*, 4189-4196.

Roberts, B.R., Wood, Z.A., Jonsson, T.J., Poole, L.B., and Karplus, P.A. (2005). Oxidized and synchrotron cleaved structures of the disulfide redox center in the N-terminal domain of Salmonella typhimurium AhpF. Protein Science *14*, 2414.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., and Baker, D. (2004a). Protein structure prediction using Rosetta. Methods Enzymol. *383*, 66-93.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., and Baker, D. (2004b). Protein structure prediction using Rosetta. Methods Enzymol. *383*, 66-93.

Sanson, M., Marineau, C., Desmaze, C., Lutchman, M., Ruttledge, M., Baron, C., Narod, S., Delattre, O., Lenoir, G., and Thomas, G. (1993). Germline deletion in a neurofibromatosis type 2 kindred inactivates the NF2 gene and a candidate meningioma locus. Human Molecular Genetics *2*, 1215-1220.

Sarma, G.N., Savvides, S.N., Becker, K., Schirmer, M., Schirmer, R.H., and Karplus, P.A. (2003). Glutathione reductase of the malarial parasite Plasmodium falciparum: crystal structure and inhibitor development. Journal of Molecular Biology *328*, 893-907.

Savvides, S.N., and Karplus, P.A. (1996). Kinetics and crystallographic analysis of human glutathione reductase in complex with a xanthene inhibitor. J Biol Chem *271*, 8101-8107.

Schäfer, L., and Cao, M. (1995). Predictions of protein backbone bond distances and angles from first principles. J. Mol. Struct. *333*, 201-208.

Schmidt, A., and Lamzin, V.S. (2002). Veni, vidi, vici - atomic resolution unravelling the mysteries of protein function. Curr. Opin. Struct. Biol. *12*, 698-703.

Schwede, T., Sali, A., Honig, B., Levitt, M., Berman, H.M., Jones, D., Brenner, S.E., Burley, S.K., Das, R., Dokholyan, N.V., et al. (2009). Outcome of a workshop on applications of protein models in biomedical research. Structure (London, England: 1993) *17*, 151-159.

Scoles, D.R. (2008). The merlin interacting proteins reveal multiple targets for NF2 therapy. Biochimica Et Biophysica Acta *1785*, 32-54.

Sem, D.S., and Kasper, C.B. (1992). Geometric relationship between the nicotinamide and isoalloxazine rings in NADPH-cytochrome P-450 oxidoreductase: implications for the classification of evolutionarily and functionally related flavoproteins. Biochemistry *31*, 3391-3398.

Sevcik, J., Dauter, Z., Lamzin, V.S., and Wilson, K.S. (1996). Ribonuclease from Streptomyces aureofaciens at Atomic Resolution. Acta Crystallogr. D. Biol. Crystallogr. *52*, 327-344.

Sheldrick, G., and Schneider, T. (1997). SHELXL: High-resolution refinement. Methods in Enzymology *277*, 319-343.

Sheldrick, G.M. (2007). A short history of SHELX. Acta Crystallogr. A. Found.

Crystallogr. *64*, 112-122.

Shimizu, T., Seto, A., Maita, N., Hamada, K., Tsukita, S., Tsukita, S., and Hakoshima, T. (2002). Structural basis for neurofibromatosis type 2. Crystal structure of the merlin FERM domain. The Journal of Biological Chemistry *277*, 10332-10336.

Söding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Research *33*, W244-248.

Speck, O., Hughes, S.C., Noren, N.K., Kulikauskas, R.M., and Fehon, R.G. (2003). Moesin functions antagonistically to the Rho pathway to maintain epithelial integrity. Nature *421*, 83-87.

Stec, B. (2007a). Comment on Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? by Jaskolski, Gilski, Dauter and Wlodawer (2007). Acta Crystallogr. D. Biol. Crystallogr. *63*, 1113-1114.

Stec, B. (2007b). Comment on Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? by Jaskolski, Gilski, Dauter & Wlodawer (2007). Acta Crystallogr. D Biol. Crystallogr. *63*, 1113-1114.

Stehle, T., Claiborne, A., and Schulz, G.E. (1993). NADH binding site and catalysis of NADH peroxidase. FEBS Journal *211*, 221-226.

Stirnimann, C.U., Rozhkova, A., Grauschopf, U., Böckmann, R.A., Glockshuber, R., Capitani, G., and Grütter, M.G. (2006). High-resolution structures of Escherichia coli cDsbD in different redox states: A combined crystallographic, biochemical and computational study. Journal of Molecular Biology *358*, 829-845.

Sustmann, R., Sicking, W., and Schulz, G.E. (1989). The Active Site of Glutathione Reductase: An Example of Near Transition-State Structures. Angewandte Chemie International Edition in English *28*, 1023-1025.

Theobald, D.L., and Wuttke, D.S. (2006). THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics (Oxford, England) *22*, 2171-2172.

Thorpe, C., and Williams, C.H. (1976). Spectral evidence for a flavin adduct in a monoalkylated derivative of pig heart lipoamide dehydrogenase. J. Biol. Chem. *251*, 7726-7728.

Tickle, I.J. (2007a). Experimental determination of optimal root-mean-square deviations of macromolecular bond lengths and angles from their restrained ideal values. Acta Crystallogr. D Biol. Crystallogr. *63*, 1274-1281; author reply 1282-1283.

Tickle, I.J. (2007b). Experimental determination of optimal root-mean-square deviations of macromolecular bond lengths and angles from their restrained ideal values. Acta Crystallogr. D. Biol. Crystallogr. *63*, 1274-1281.

Tronrud, D.E. (1997). TNT refinement package. Methods Enzymol *277*, 306-319.

Trueblood, K.N., Burgi, H.B., Burzlaff, H., Dunitz, J.D., Gramaccioli, C.M., Schulz, H.H., Shmueli, U., and Abrahams, S.C. (1996). Atomic dispacement parameter nomenclature. Report of a subcommittee on atomic displacement parameter nomenclature. Acta Crystallographica Section A Foundations of Crystallography *52*, 770-781.

Vanoni, M.A., Wong, K.K., Ballou, D.P., and Blanchard, J.S. (1990). Glutathione reductase: comparison of steady-state and rapid reaction primary kinetic isotope effects exhibited by the yeast, spinach, and Escherichia coli enzymes. Biochemistry *29*, 5790-5796.

Weik, M., and Sussman, J. (2000). Synchotron X-ray radiation produces specific chemical and structural damage to protein structures. Proceedings of the National Academy of Sciences *97*, 623–628.

Wiita, A.P., Ainavarapu, S.R.K., Huang, H.H., and Fernandez, J.M. (2006). Force-dependent chemical kinetics of disulfide bond reduction observed with single-molecule techniques. Proceedings of the National Academy of Sciences *103*, 7222-7227.

Wilson, K.S., Dauter, Z., Lamsin, V.S., Walsh, M., Wodak, S., Richelle, J., Pontius, J., Vaguine, A., Sander, R.W.W., and Hooft, V.G. (1998). Who checks the checkers? Four validation tools applied to eight atomic resolution structures. J. Mol. Biol. *276*, 417-36.

Winn, M.D., Murshudov, G.N., and Papiz, M.Z. (2003). Macromolecular TLS refinement in REFMAC at moderate resolutions. Methods in Enzymology *374*, 300-21.

Wlodawer, A., Miller, M., Jaskólski, M., Sathyanarayana, B.K., Baldwin, E., Weber, I.T., Selk, L.M., Clawson, L., Schneider, J., and Kent, S.B. (1989). Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. Science *245*, 616-621.

Word, J.M., Lovell, S.C., LaBean, T.H., Taylor, H.C., Zalis, M.E., Presley, B.K., Richardson, J.S., and Richardson, D.C. (1999). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. Journal of Molecular Biology *285*, 1711-1733.

Wu, Y.D., and Houk, K.N. (1991). Theoretical evaluation of conformational preferences of NAD+ and NADH: an approach to understanding the stereospecificity of NAD+/NADH-dependent dehydrogenases. Journal of the American Chemical Society *113*, 2353-2358.

Wu, Y.D., and Houk, K.N. (1993). Theoretical study of conformational features of NAD+ and NADH analogs: protonated nicotinamide and 1, 4-dihydronicotinamide. The Journal of Organic Chemistry *58*, 2043-2045.

Yates, R.L., Epiotis, N.D., and Bernardi, F. (1975). Importance of nonbonded attraction in the stereochemistry of the SN2' reaction. Journal of the American Chemical Society *97*, 6615-6621.

Young, L., and Post, C.B. (1996). Catalysis by entropic guidance from enzymes. Biochemistry *35*, 15129-15133.

Yu, C.H., Norman, M.A., Schäfer, L., Ramek, M., Peeters, A., and van Alsenoy, C. (2001). Ab initio conformational analysis of N-formyl L-alanine amide including electron correlation. J. Mol. Struct. *567*, 361-374.

Zhang, Y. (2009). Protein structure prediction: when is it useful? Curr. Opin. Struct. Biol. *19*, 145-155.