UNIVERSITY OF CALIFORNIA, SAN DIEGO

Multiscale modeling of metalloproteins and protonation equilibria in biomolecular systems

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Chemistry

by

Patrick G. Blachly

Committee in charge:

Professor James A. McCammon, Chair
Professor Michael K. Gilson
Professor Katja Lindenberg
Professor Louis Noodleman
Professor Stanley J. Opella
Professor John H. Weare

2014

The dissertation of Patrick G. Blachly is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2014

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ε – Dielectric constant
Δ + PHS – Hyperstable, acid-resistant Staphylococcal nuclease mutant
ABZ – Albendazole
ADF – Amsterdam Density Functional
AF – Antiferromagnetic
aMD – Accelerated molecular dynamics
APS – Adenosine 5'-phosphosulfate
B3LYP – Becke exchange with Lee, Yang and Par correlation (3-parameter) hybrid functional
BPTI – Bovine pancreatic trypsin inhibitor
BS – Broken-symmetry
BS-DFT – Broken-symmetry density functional theory
BZ – Benzimidazole
CB[7] – Cucurbit[7]uril
CBZ – Carbendazim
CHELPG – CHarges from ELectrostatic Potentials using a Grid-based method
COSMO – COnductor-like Screening MOdel
CpHaMD – Constant pH accelerated molecular dynamics
CpHMD – Constant pH molecular dynamics
CS – Conformational selection
DFT – Density functional theory
DMAPP – Dimethylallyl diphosphate
e^- – Electron
EFG – Electric field gradient
ENDOR – Electron nuclear double resonance
EPR – Electron paramagnetic resonance
ESP – Electrostatic potential
FBZ – Fuberidazole
FEP – Free energy perturbation
FFT – Fast Fourier transform
FTMAP – Protein
GB – Generalized Born
H^+ – Proton
HIPIP – High potential iron-sulfur protein
HF – Hartree-Fock
HMBPP – (E)-4-hydroxy-3-methyl-but-2-enyl diphosphate
IF – Induced fit
IPP – Isopentenyl diphosphate
IS – (Mössbauer) Isomer shift
k_B – Boltzmann's constant
KL – Kullback-Leibler
KL_{res} – Kullback-Leibler divergence of a residue

MAE – Mean absolute error
MC – Monte Carlo
MD – Molecular dynamics
MEAD – Macroscopic Electrostatics with Atomic Detail
MEP – Methylerythritol phosphate (pathway also referred to as nonmevalonate pathway)
MK – Merz-Kollman
n – Hill coefficient
NMR – Nuclear magnetic resonance
NPT – Constant pressure and temperature
NVT – Constant volume and temperature
OLYP – Handy and Cohen exchange with Lee, Yang and Par correlation functional
PC – Principal component
PCA – Principal component analysis
PDB – Protein data bank
PDLD – Protein dipoles Langevin dipoles
pK_a – Logarithmic value of the acid dissociation constant
POVME – POcket Volume MEasurer
PPi – Inorganic diphosphate (pyrophosphate)
PV – Pressure multiplied by volume
QM – Quantum mechanics
QM/MM – Quantum mechanics/molecular mechanics
QS – (Mössbauer) Quadrupole splitting
RESP – Restrained electrostatic potential
RMSD – Root-mean-square deviation
RMSE – Root-mean-square error
RMSF – Root-mean-square fluctuation
RT – Ideal gas constant multiplied by temperature
s – Fraction of deprotonated species
SCRF – Self-consistent reaction field
SNase – Staphylococcal nuclease
STAMP – Structural Alignment of Multiple Proteins
STO – Slater-type orbitals
STRIDE – STRuctural IDEntification
TBZ – Thiabendazole
TI – Thermodynamic integration
TZP – Triple-zeta plus polarization
UV – Ultraviolet
vdW – van der Waals
WT – Wild-type
Z_{LR} – Charge of ligand-receptor complex
Z_L – Charge of ligand
ZPE – Zero-point energy

Acknowledgements

My conscious brain tends to work in a linear manner. For this reason, these words of thanks are presented chronologically and in paragraphs that would earn disapproval from a journal editor.

To my parents, I am forever grateful for your constant efforts to inhibit boredom in my formative years. I never felt pushed by you to achieve anything specific in life, and for this reason, I explored various avenues and stumbled onto the track I am currently on. For this I am happy. Many thanks for your constant love and support. To my sister Gail, thanks for keeping me in check. You were a professional in this task before you ever entered the workforce. To Alysia, thanks for always accompanying me through life's detours and tolerating me through them. I admire the creativity and courage you show in your own life pursuits—being with you, I am never starved for inspiration.

To Joe Francisco, many thanks for showing me that it is OK to showcase quirks in the workplace, for subtly steering me away from the economists, and for adding the phrase, "Hang tough!" to my vernacular. I've carried this phrase with me through graduate school and fully intend to bring it to my next place of work.

To Andy, thanks for your counsel and for giving me the opportunity to learn from you. Through your generosity, I have had the opportunity to explore a multitude of projects and have had the leeway to fail at many of them. For this, I am grateful. Further, thanks for doing the little things you do to make lab enjoyable. This includes

(but is obviously not limited to) your generous sharing of the NY Times, so we can read the Opinion section without paying a premium for it.

To Lou, thanks for your patience in teaching me the tricky art that is dealing with iron-sulfur proteins and for being a entertaining source of scientific trivia. To my labmates, past and present, thanks for joining me in this ride. I am grateful for conversations and projects embarked upon with Cesar de Oliveira, Levi Pierce, Sarah Williams, Paul Gasper, Olivia Kim, Joe Kaus and Nuo Wang. To Patti, thanks for ensuring that my metaphorical shoes were tied throughout my time in the McCammon Group. To Robert Konecny and Brian Fox, thanks for ensuring that the computer problems I have had resulted in solutions, not broken computers.

To my committee, thanks for participating in this journey with me and coming together on a Friday in July to bring it to a close. Lastly, many thanks to the ARCS Foundation and NIH for funding parts of my education.

Chapter 2 is a minimally modified reprint of the material as it appears in Patrick G. Blachly, Gregory M. Sandala, D. Ann Giammona, Tiqing Liu, Donald Bashford, J. Andrew McCammon, and Louis Noodleman, "Use of broken-symmetry density functional theory to characterize the IspH oxidized state: Implications for IspH mechanism and inhibition," submitted to the *Journal of Chemical Theory and Computation*, 2014. The dissertation author was the primary investigator and author of this paper.

Chapter 3 is a minimally modified reprint of the material as it appears in Patrick G. Blachly, Cesar A.F. de Oliveira, Sarah L. Williams, and J. Andrew

McCammon, “Utilizing a Dynamical Description of IspH to Aid in the Development of Novel Antimicrobial Drugs,” *PLoS Computational Biology*, 2013. The dissertation author was the primary investigator and author of this paper.

Chapter 4 is a minimally modified reprint of the material as it appears in Sarah L. Williams, Patrick G. Blachly, and J. Andrew McCammon, “Measuring the Successes and Deficiencies of Constant pH Molecular Dynamics: A Blind Prediction Study,” *Proteins: Structure, Function, and Bioinformatics*, 2011. The dissertation author was one of two primary investigators, who each contributed equally to performing research and writing this paper.

Chapter 5 is a minimally modified reprint of the material as it appears in M. Olivia Kim, Patrick G. Blachly, Joseph W. Kaus, and J. Andrew McCammon, “Protocols Utilizing Constant pH Molecular Dynamics to Compute pH-Dependent Binding Free Energies,” submitted to the *Journal of Physical Chemistry B*, 2014. The dissertation author was one of two primary investigators, each of whom contributed equally to performing research and writing this paper.

Vita

2009	B. A. in Chemistry Grinnell College
2011	M.S. in Chemistry University of California, San Diego
2014	Ph.D. in Chemistry University of California, San Diego

Publications

Blachly P.G., Sandala G.M., Giammona D.A., Liu T., Bashford D., McCammon J.A., and Noodleman L., Substrate conformational changes in the 1e- reduced state of the IspH enzyme revealed using broken-symmetry density functional theory computations, (in preparation).

Martin D.P., **Blachly P.G.***, McCammon J.A., and Cohen S.M., Exploring the influence of protein-ligand interactions on coordination by metal-binding pharmacophores in a metalloprotein active site, *J. Med. Chem.,* (in review).
* This author was the primary contributor of the computational aspects in this work.

Blachly P.G., Sandala G.M., Giammona D.A., Liu T., Bashford D., McCammon J.A., and Noodleman L., Use of broken-symmetry density functional theory to characterize the IspH oxidized state: Implications for IspH mechanism and inhibition, *J. Chem. Theory Comput.,* (in review).

Kim M.O.**, **Blachly P.G.****, Kaus J.W., and McCammon J.A., Protocols utilizing constant pH molecular dynamics to compute pH-dependent binding free energies, *J. Phys. Chem. B,* (in review).
** These authors contributed equally to this work.

Martin D.P., **Blachly P.G.***, Marts A.R., Woodruff T.M., de Oliveira C.A.F., McCammon J.A., Tierney D.L., and Cohen S.M., 'Unconventional' coordination chemistry by metal chelating fragments in a metalloprotein active site, *J. Am. Chem. Soc.* 2014, 136(14): 5400-5406.
* This author was the primary contributor of the computational aspects in this work.

Blachly P.G., de Oliveira C.A.F., Williams S.L., and McCammon J.A., Utilizing a dynamical description of IspH to aid in the development of novel antimicrobial drugs, *PLoS Comput. Biol.* 2013, 9(12): e1003395.

Williams S.L.**, **Blachly P.G.****, and McCammon J.A., Measuring the successes and deficiencies of constant pH molecular dynamics: a blind prediction study, *Proteins.* 2011, 79(12): 3381-3388.
** These authors contributed equally to this work.

Liessman M., Hansmann B., **Blachly P.G.***, Francisco J.S., Abel B., Primary steps in the reaction of OH radicals with biomolecules at low temperatures in Laval nozzle expansions: perspectives from experiment and theory, *J. Phys. Chem. A*. 2009, 113(26): 7570-7575.
* This author was the primary contributor of the computational aspects in this work.

Abstract of the Dissertation

Multiscale modeling of metalloproteins and protonation equilibria in biomolecular systems

by

Patrick G. Blachly

Doctor of Philosophy in Chemistry

University of California San Diego, 2014

Professor James A. McCammon, Chair

A combination of classical and quantum mechanical approaches are described in Chapter 1 and utilized in this dissertation to study catalysis and allostery in the metalloprotein IspH, as well as to probe protonation equilibria in a variety of macromolecules. Chapters 2 and 3 are dedicated to characterizing the oxidized [4Fe-4S] IspH protein, which is biologically important as an antimicrobial drug target. In Chapter 2, the protonation states of active site residues in substrate-bound IspH are characterized using broken-symmetry density functional theory to provide a foundation for exploring IspH catalysis. Subsequently, a more coarse-grained

approach is used in Chapter 3 to assess the internal motions of IspH, both with and without its substrate bound, using classical molecular dynamics. Both these studies reveal rational approaches for the design of novel IspH inhibitors. Chapters 4 and 5 deviate from the metalloprotein theme established in Chapters 2 and 3 to consider classical approaches for treating proton binding and unbinding in the context of molecular dynamics simulations. The ability of the constant pH molecular dynamics method to predict protein pK_a values is assessed in Chapter 4 using an experimental test set comprising Staphylococcal nuclease variants. Building on this work, Chapter 5 provides proof of concept for the constant pH molecular dynamics method to obtain pH-dependent binding free energies in conjunction with Wyman's binding polynomial formalism.

Chapter 1:

Simulating metalloproteins and protonation equilibria

Introduction

To achieve the highest possible accuracy, the dynamics of biomolecules would be described by solving the time-dependent Schrödinger equation for a desired timescale. While such an approach may be theoretically sound, it is computationally intractable. Indeed quantum molecular dynamics simulations can only be performed on small systems (< 100 atoms) and for short timescales (ps to ns). Consequently, coarser physics-based methodologies (*i.e.* molecular mechanics) must be employed to simulate larger biomolecular systems at longer timescales.[1,2] Typically, such approaches involve assigning point charges to atoms, which are represented as balls on springs with the potential energy of the system described by a classical mechanics force field. Using the force field, the atomic coordinates of a biomolecular system can be propagated over time according to Newton's equations of motion in a molecular dynamics (MD) simulation.[3,4] Classical MD simulations are able to describe the time evolution of a biomolecular system at relatively long timescales (up to several ms)[5,6] at the cost of employing the simplified force field as the system Hamiltonian. This

technique has found success in solving problems related to protein folding[5,6] and biomolecular association,[7,8] but its coarse-grained description of atoms is not suited to accurately represent electronic properties in the way that quantum mechanical (QM)-based methods can. As a result, charge transfer, polarization, and the bond breaking/forming processes inherent to enzyme catalysis cannot be described using classical methods.[9]

Due to the tradeoff between accuracy and computational feasibility, computational chemists must carefully choose appropriate simulation techniques for addressing a given research question under study. In this dissertation, I probe two separate biochemical problems that require different levels of theoretical rigor in the computations employed. The first study separately applies both classical and QM approaches to understand the internal dynamics and catalytic mechanism of the [4Fe-4S] IspH enzyme. Subsequently, the effects of protonation in the course of protein dynamics and host-guest association are probed within a classical framework. Briefly, in this introduction I outline the impetus behind studying each of these topics and the motivation behind the choice of computational methodology for addressing it.

IspH: A novel antimicrobial drug target

With increasing drug resistance complicating efforts to treat and eradicate illnesses such as malaria and tuberculosis, novel therapies are needed.[10,11] The biosynthetic pathway for isoprenoid precursors has been targeted for the development of antibacterial and antimalarial drugs. Isoprenoids are a class of vital biomolecules that includes sterols, cholorophylls and quinones. All isoprenoids are solely derived

from the 5-carbon precursors isopentenyl diphosphate (IPP, Figure 1-1) and dimethylallyl diphosphate (DMAPP). Whereas humans synthesize IPP and DMAPP via the mevalonate pathway (Figure 1-1, bottom),[12] most pathogenic bacteria and the Plasmodia parasites that cause malaria obtain these isoprenoid precursors from the methylerythritol phosphate (MEP) pathway (Figure 1-1, top).[13-15]

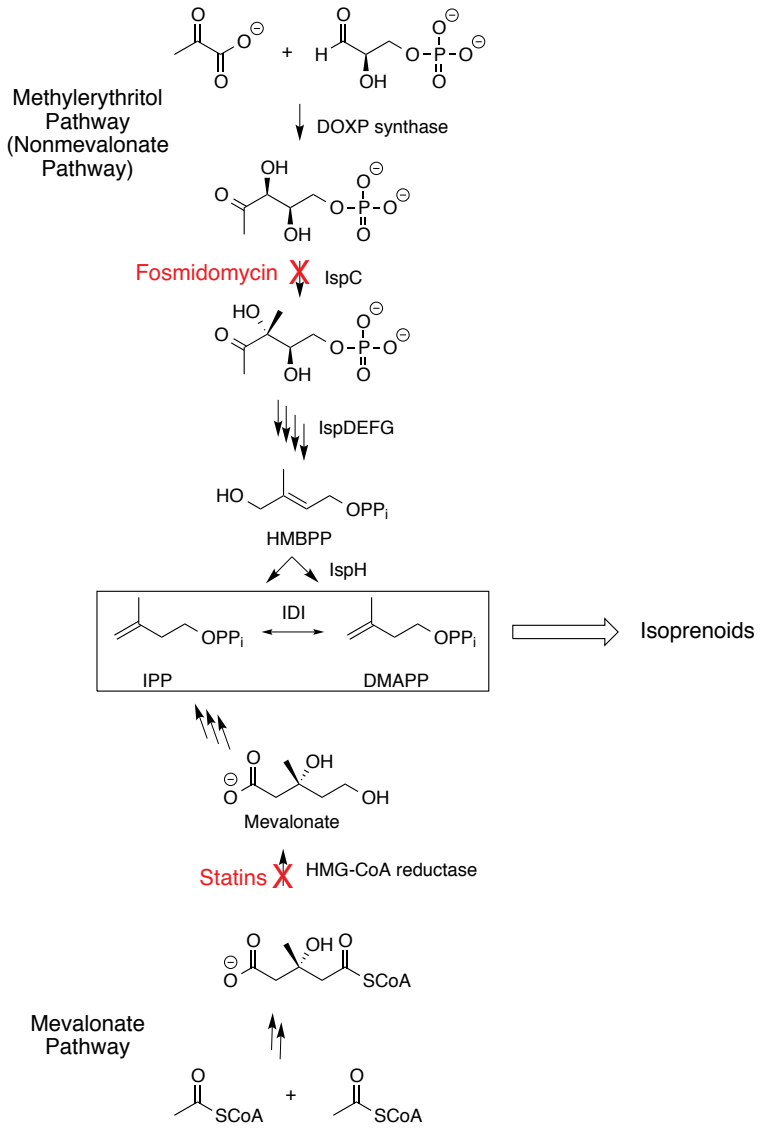

Figure 1-1. The contrasting isoprenoid biosynthetic pathways in most pathogenic bacteria and apicomplexan protozoa (methylerythritol pathway, top) and animals (mevalonate pathway, bottom). Fosmidomycin, which inhibits IspC, is a potent antimalarial drug.

While the mevalonate pathway has been studied for decades, highlighted by the success of statins in combating cholesterol levels in humans (Figure 1-1),[12,16] the MEP pathway was only discovered in the late 1990s.[13-15] Fosmidomycin, which inhibits the third step of the MEP pathway, has recently been approved to treat malaria,[17] demonstrating the utility of the MEP pathway inhibitors for antimicrobial drug development.[18,19] The second and third chapters in this dissertation focus on the [4Fe-4S] IspH protein, which catalyzes the ultimate step of the MEP pathway: a $2e^-/2H^+$ reductive dehydroxylation of HMBPP (Figure 1-1) to afford IPP and DMAPP in a ~5:1 ratio.[20,21]

Modeling the [4Fe-4S] IspH protein

While the IspH protein was discovered a decade ago, its full structural characterization is still incomplete. In fact, consensus on the iron content of the IspH iron-sulfur cluster was only reached upon the solution of a crystal structure of [4Fe-4S] IspH in complex with its substrate HMBPP in 2010.[21] Consequently, efforts to develop effective inhibitors of IspH are still in their infancy. Leading the IspH drug discovery efforts is the Eric Oldfield Group at University of Illinois, Urbana-Champaign. By performing electron paramagnetic resonance (EPR) and electron-nuclear double resonance (ENDOR) spectroscopy on 1-electron ($1e^-$) reduced IspH in complex with HMBPP, Oldfield and co-workers have observed similarities to experiments performed on nitrogenase in complex with acetylene and allyl alcohols.[22] This finding inspired the development of alkyne- and alkene-based inhibitors under the hypothesis that HMBPP may coordinate one of the iron atoms in the [4Fe-4S]

cluster through its olefin (C=C) group.[22,23] Motivated by the Oldfield Group's initial efforts, two questions have arisen that can be addressed using *in silico* methods:

(1) How does IspH catalyze the conversion of HMBPP to IPP and DMAPP? Specifically, how does the protonation of active site residues change over the course of catalysis, during which $2e^-$ and $2H^+$ must be transferred to the substrate? These questions are considered in Chapter 2.

(2) How might additional inhibitors be designed to best interact with the IspH target? This question and the intrinsic motions of the IspH protein are studied in Chapter 3.

Using DFT methods to characterize the IspH oxidized state

Computational modeling of IspH catalysis requires a QM treatment. In addition to the electron additions and molecular rearrangements needed in IspH catalysis, accurate descriptions of iron-sulfur clusters require treatment of electronic effects beyond what classical methodologies are capable of achieving. More specifically, the irons in iron-sulfur clusters are usually high spin, exhibiting large spin-polarization effects. Additionally, there exists significant Fe-S covalency in iron-sulfur clusters, and individual iron sites spin couple through their bridging sulfide (S^{2-}) atoms. The [4Fe-4S] cluster in IspH assumes an antiferromagnetically-coupled state, in which the spins on two iron atoms are aligned parallel to but opposite the spins on the other two irons.[24,25] This spin coupling can occur through different combinations, two of which are illustrated in Figure 1-2. The two systems displayed in Figure 1-2 are

valence isomers, which can have different physicochemical properties and must, therefore, be considered in a computational study.

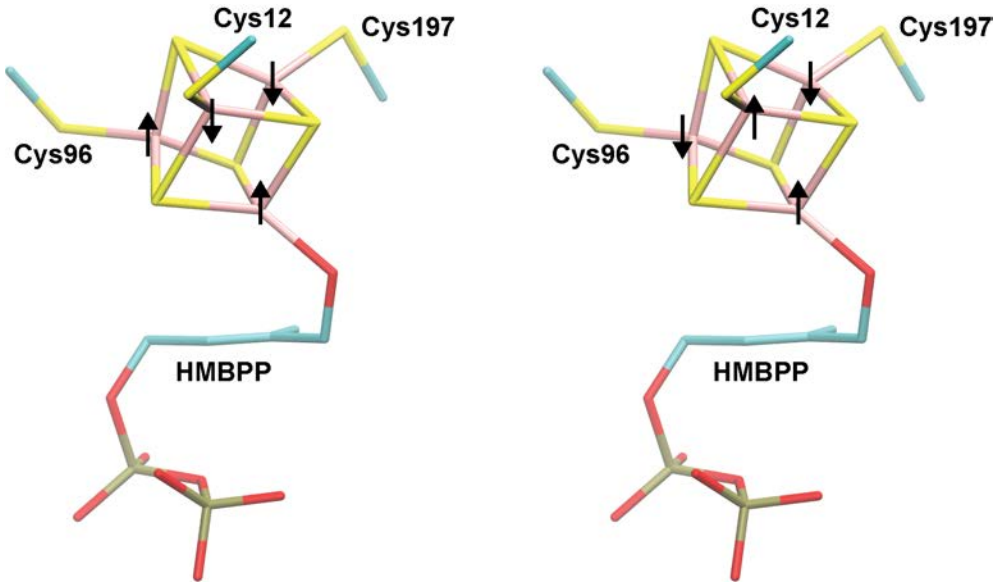

Figure 1-2. Different valence isomers for the IspH [4Fe-4S] cluster. Arrows represent the net spin on the iron atoms (rendered in pink).

Conventional QM methods are unable to recover an antiferromagnetic description of [4Fe-4S] clusters, which motivates the usage of broken symmetry density functional theory (BS-DFT) computations to characterize the [4Fe-4S] cluster in IspH.[26] In BS-DFT computations, a high spin calculation is first performed (with the net spins on all irons aligned). Subsequently, the site spin vectors on two irons are rotated, and geometry optimizations and energy computations are performed to obtain an antiferromagnetic solution.[27]

In practice, BS-DFT computations are performed on a large IspH active site model that is needed due to the highly-charged natures of both the [4Fe-4S] cluster and the substrate HMBPP. To best account for charge transfer effects, third and fourth shell ligands are included in the quantum cluster. Rather than including the rest of the

IspH protein in QM/MM computations, we perform geometry optimizations on the large active site quantum cluster and post-process the optimized geometry by embedding it in an environment containing surrounding protein point charges and regions of different dielectric constants in a Poisson-Boltzmann self-consistent reaction field calculation.[28] This approach effectively accounts for the effects of the surrounding protein electrostatics while removing the ruggedness of the protein energy landscape that can affect QM/MM calculations. Performing these computations on different protonation and oxidation states of the IspH active site, we are able to characterize reaction intermediates and comment on the catalytic mechanism of IspH. A complete characterization of the IspH oxidized state and its implications for reduction are discussed in Chapter 2.

Classical MD simulations of IspH

While the BS-DFT characterization of the IspH active site aids the characterization of the IspH catalytic mechanism, it is incapable of describing the global protein motions that play a role in binding the substrate HMBPP at long timescales. Furthermore, developing competitive inhibitors of metalloproteins is a difficult task, as conventional docking approaches fail to accurately rank interactions between potential inhibitors and the metal site.[29] One opportunity to circumvent this limitation of conventional docking is to find allosteric sites, which may provide opportunities for inhibitors to bind IspH without directly interacting with the metal site. In Chapter 3, we utilize a classical MD approach to characterize the

conformational space available to the IspH protein and find alternative opportunities for the design of novel inhibitors.

Describing protonation phenomena in biomolecular simulations

Typically, the protonation states for a biomolecular system are assigned prior to performing MD simulations. This assignment is made according to the starting coordinates being used for the simulation, which are usually taken from a high-resolution structure provided by X-ray crystallography or NMR. Once assigned, the protonation states are held constant through the duration of the simulation. This approach removes the ability of ionizable sites to change their protonation state as their chemical environments change during the course of the simulation. When the true ensemble of conformations for a given ionizable group contains multiple electrostatic environments, the use of a single, fixed protonation state may preclude accurate sampling of the ensemble.[30] Furthermore, when considering the association of two biomolecules, which inherently results in altered electrostatic environments for the two binding partners, changes in protonation may be necessary to effectively quantitate the binding free energy.

In Chapters 4 and 5, the constant pH molecular dynamics (CpHMD) method is, respectively, assessed and used to compute pH-dependent binding free energies. In CpHMD simulations, protein conformations are sampled according to the semi-grand canonical ensemble, where protons may be added or removed from ionizable residues in the system during Monte Carlo steps that intermittently interrupt the MD simulation (Figure 1-3). More specifically, every 10 fs in the MD simulation, a random residue is

chosen for considering a change in its protonation state, where a change in protonation is represented by changing the partial charges on the ionizable residue. A rapid electrostatic computation is performed to assess the energetic favorability of switching protonation states (partial charges) for the titrating residue at a given level of pH. The Metropolis criterion is applied to either reject or accept the change in protonation, and the MD simulation continues (Figure 1-3). Following this approach, the MD simulation samples both protonation space and conformational space.[31,32] The CpHMD framework is thus beneficial in that it reasonably accounts for proton (un)binding in simulations that can reach long timescales.

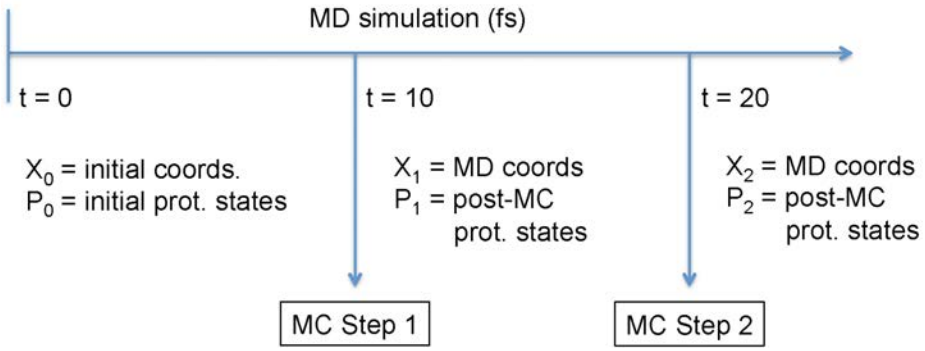

Figure 1-3. Representation of CpHMD simulations. Every 10 fs of MD are interrupted by a Monte Carlo step, in which a random residue is chosen to consider a change in protonation. Change in protonation is assessed using the Metropolis criterion and, subsequently, MD resumes with a potentially new set of protonation states. This sequence of events is repeated for the duration of the MD simulation.

Blind predictions of protein pK_a values

In Chapter 4, the CpHMD method is used to blindly predict the pK_a values of ionizable residues in a series of Staphylococcal nuclease (SNase) proteins. This work was largely inspired by the experimental efforts of the Bertrand Garcia-Moreno Laboratory. Garcia-Moreno and co-workers employed NMR spectroscopy to obtain

pK_a values for all ionizable groups in SNase and, subsequently, generated SNase mutants having a single internal (typically hydrophobic) residues mutated to an ionizable residue.[33-37] The experimental pK_a's of these point mutations, as well as the pK_a values of the wild-type (WT) SNase protein were withheld from computational groups, who were then allowed to compete to assess how well their pK_a prediction techniques perform without the introduction of bias from knowing the experimental results ahead of time.[38] In Chapter 4, we use CpHMD simulations to predict SNase pK_a values and identify key areas for improving the method.

Computing pH-dependent binding free energies

As most drugs are either weak acids or weak bases,[39] the linkage between proton binding and ligand binding must be considered in computer-aided drug discovery.[40] As conventional approaches for computing binding free energies utilize fixed protonation states, the CpHMD method provides a framework for describing an ensemble of protonation states in the binding process. Using CpHMD simulations, pH-dependent binding free energy profiles can be obtained for biomolecular association. This approach is described in Chapter 5 and is applied to various host-guest systems with good success. Moving forward, we believe CpHMD simulations will have great utility in drug design workflows and in considering pH-dependent conformational changes in proteins.

References

(1) Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H. F.; Shaw, D. E. *Annu Rev Biophys* **2012**, *41*, 429.

(2) Steinbrecher, T.; Elstner, M. *Methods Mol Biol* **2013**, *924*, 91.

(3) Hug, S. *Methods Mol Biol* **2013**, *924*, 127.

(4) Monticelli, L.; Tieleman, D. P. *Methods Mol Biol* **2013**, *924*, 197.

(5) Pierce, L. C. T.; Salomon-Ferrer, R.; de Oliveira, C. A. F.; McCammon, J. A.; Walker, R. C. *J Chem Theory Comput* **2012**, *8*, 2997.

(6) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. B.; Wriggers, W. *Science* **2010**, *330*, 341.

(7) Kollman, P. *Chem Rev* **1993**, *93*, 2395.

(8) Shan, Y. B.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. *J Am Chem Soc* **2011**, *133*, 9181.

(9) Warshel, A. *Annu Rev Bioph Biom* **2003**, *32*, 425.

(10) Boucher, H. W.; Talbot, G. H.; Bradley, J. S.; Edwards, J. E.; Gilbert, D.; Rice, L. B.; Scheld, M.; Spellberg, B.; Bartlett, J. *Clin Infect Dis* **2009**, *48*, 1.

(11) Ridley, R. G. *Nature* **2002**, *415*, 686.

(12) Goldstein, J. L.; Brown, M. S. *Nature* **1990**, *343*, 425.

(13) Rohmer, M. *Nat Prod Rep* **1999**, *16*, 565.

(14) Rohmer, M.; Knani, M.; Simonin, P.; Sutter, B.; Sahm, H. *Biochem J* **1993**, *295*, 517.

(15) Sacchettini, J. C.; Poulter, C. D. *Science* **1997**, *277*, 1788.

(16) Buhaescu, I.; Izzedine, H. *Clin Biochem* **2007**, *40*, 575.

(17) Missinou, M. A.; Borrmann, S.; Schindler, A.; Issifou, S.; Adegnika, A. A.; Matsiegui, P. B.; Binder, R.; Lell, B.; Wiesner, J.; Baranek, T.; Jomaa, H.; Kremsner, P. G. *Lancet* **2002**, *360*, 1941.

(18) Jomaa, H.; Wiesner, J.; Sanderbrand, S.; Altincicek, B.; Weidemeyer, C.; Hintz, M.; Turbachova, I.; Eberl, M.; Zeidler, J.; Lichtenthaler, H. K.; Soldati, D.; Beck, E. *Science* **1999**, *285*, 1573.

(19) Oldfield, E. *Accounts Chem Res* **2010**, *43*, 1216.

(20) Grawert, T.; Kaiser, J.; Zepeck, F.; Laupitz, R.; Hecht, S.; Amslinger, S.; Schramek, N.; Schleicher, E.; Weber, S.; Haslbeck, M.; Buchner, J.; Rieder, C.; Arigoni, D.; Bacher, A.; Eisenreich, W.; Rohdich, F. *J Am Chem Soc* **2004**, *126*, 12847.

(21) Grawert, T.; Span, I.; Eisenreich, W.; Rohdich, F.; Eppinger, J.; Bacher, A.; Groll, M. *P Natl Acad Sci USA* **2010**, *107*, 1077.

(22) Wang, W. X.; Wang, K.; Liu, Y. L.; No, J. H.; Li, J. K.; Nilges, M. J.; Oldfield, E. *P Natl Acad Sci USA* **2010**, *107*, 4522.

(23) Wang, K.; Wang, W. X.; No, J. H.; Zhang, Y. H.; Zhang, Y.; Oldfield, E. *J Am Chem Soc* **2010**, *132*, 6719.

(24) Beinert, H.; Holm, R. H.; Munck, E. *Science* **1997**, *277*, 653.

(25) Noodleman, L.; Lovell, T.; Liu, T. Q.; Himo, F.; Torres, R. A. *Curr Opin Chem Biol* **2002**, *6*, 259.

(26) Noodleman, L.; Peng, C. Y.; Case, D. A.; Mouesca, J. M. *Coordin Chem Rev* **1995**, *144*, 199.

(27) Sandala, G. M.; Noodleman, L. *Methods Mol Biol* **2011**, *766*, 293.

(28) Li, J.; Nelson, M. R.; Peng, C. Y.; Bashford, D.; Noodleman, L. *J Phys Chem A* **1998**, *102*, 6311.

(29) Irwin, J. J.; Raushel, F. M.; Shoichet, B. K. *Biochemistry-Us* **2005**, *44*, 12316.

(30) Mongan, J.; Case, D. A. *Curr Opin Struc Biol* **2005**, *15*, 157.

(31) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. *J Chem Phys* **2002**, *117*, 4184.

(32) Mongan, J.; Case, D. A.; McCammon, J. A. *J Comput Chem* **2004**, *25*, 2038.

(33) Castaneda, C. A.; Fitch, C. A.; Majumdar, A.; Khangulov, V.; Schlessman, J. L.; Garcia-Moreno, B. E. *Proteins* **2009**, *77*, 570.

(34) Chimenti, M. S.; Khangulov, V. S.; Robinson, A. C.; Heroux, A.; Majumdar, A.; Schlessman, J. L.; Garcia-Moreno, B. *Structure* **2012**, *20*, 1071.

(35) Isom, D. G.; Cannon, B. R.; Castaneda, C. A.; Robinson, A.; Bertrand, G. M. E. *P Natl Acad Sci USA* **2008**, *105*, 17784.

(36) Isom, D. G.; Castaneda, C. A.; Cannon, B. R.; Garcia-Moreno, B. E. *P Natl Acad Sci USA* **2011**, *108*, 5260.

(37) Isom, D. G.; Castaneda, C. A.; Velu, P. D.; Garcia-Moreno, B. *P Natl Acad Sci USA* **2010**, *107*, 16096.

(38) Nielsen, J. E.; Gunner, M. R.; Garcia-Moreno, E. B. *Proteins* **2011**, *79*, 3249.

(39) Manallack, D. T. *SAR QSAR Environ Res* **2009**, *20*, 611.

(40) Onufriev, A. V.; Alexov, E. *Q Rev Biophys* **2013**, *46*, 181.

Chapter 2:

Use of broken-symmetry density functional theory to characterize the IspH oxidized state: Implications for IspH mechanism and inhibition

Abstract

With current therapies becoming less efficacious due to increased drug resistance, new inhibitors of both bacterial and malarial targets are desperately needed. The recently discovered methylerythritol phosphate (MEP) pathway for isoprenoid synthesis provides novel targets for the development of such drugs. Particular attention has focused on the IspH protein, the final enzyme in the MEP pathway, which uses its [4Fe–4S] cluster to catalyze the formation of the isoprenoid precursors IPP and DMAPP from HMBPP. IspH catalysis is achieved via a $2e^-/2H^+$ reductive dehydroxylation of HMBPP; the mechanism by which catalysis is achieved, however, is highly controversial. The work presented herein provides the first step in assessing different routes to catalysis by using computational methods. By performing broken symmetry density functional theory (BS–DFT) calculations that employ both the conductor–like screening solvation model (DFT/COSMO) and a finite-difference

Poisson–Boltzmann self-consistent reaction field methodology (DFT/SCRF), we evaluate geometries, energies and Mössbauer signatures of the different protonation states that may exist in the oxidized state of the IspH catalytic cycle. From DFT/SCRF computations performed on the oxidized state, we find a state where the substrate, HMBPP, coordinates the apical iron in the [4Fe–4S] cluster as an alcohol group (ROH) to be one of two, iso–energetic, lowest–energy states. In this state, the HMBPP pyrophosphate moiety and an adjacent glutamate residue (E126) are both fully deprotonated, making the active site highly anionic. Our findings that this low–energy state also matches the experimental geometry of the active site and that its computed isomer shifts agree with experiment validate the use of the DFT/SCRF method to assess relative energies along the IspH reaction pathway. Additional studies of IspH catalytic intermediates are currently being pursued.

Introduction

In 2012, the World Health Organization reported 207 million cases of malaria[1] and 8.6 million cases of tuberculosis[2] worldwide—problems that are further complicated by the emergence of drug–resistant pathogens.[3-9] In order to address these global health problems, the development of new drugs with novel modes of action is desperately needed.

Isoprenoid biosynthesis has been determined to be an attractive enzymatic pathway for the development of novel antibacterial and antimalarial drugs.[10] Isoprenoids are a class of essential biomolecules, including sterols, prenyl side chains of chlorophylls, and quinones, which are all derived from the 5-carbon precursors

isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP, Figure 2-1).[11,12] Whereas in archaebacteria, fungi and animals, as well as in the cytosol of plants, IPP and DMAPP are synthesized through a mevalonate-dependent pathway,[12,13] Rohmer and Arigoni independently discovered an alternative, mevalonate–*independent* isoprenoid synthesis pathway that is specific to most pathogenic eubacteria (*e.g. H. pylori* and *M. tuberculosis*) and apicomplexan parasites (*e.g.* the malaria–causing parasite, *P. falciparum*) and also is employed in the plastids of plants.[14-17] Because humans lack this non–mevalonate pathway, which is also referred to as the methylerythritol phosphate (MEP) pathway, the inhibition of enzymes in this pathway provides a desirable route for the development of novel antibacterials, antimalarials and herbicides.[10] To this end, fosmidomycin has recently entered the ranks of antimalarial drugs as an inhibitor of 1–deoxy–D–xylulose 5–phosphate reductoisomerase (DOXP reductase, also known as IspC) and has been shown to be effective in concert with clindamycin for treating multidrug–resistant strains of *P. falciparum*.[18-20]

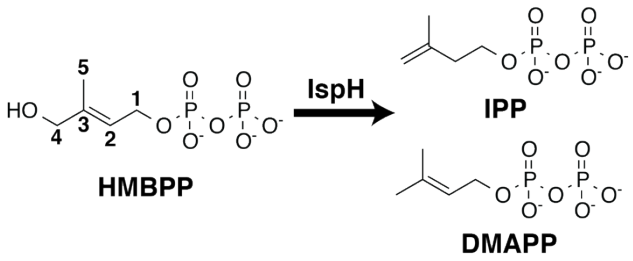

Figure 2-1. The IspH-catalyzed $2e^{-}/2H^{+}$ reductive dehydroxylation of HMBPP to afford DMAPP and IPP.

Recognizing the difficulties in combating rapid drug resistance and attempting to build on the success of fosmidomycin, there is considerable interest in discovering

additional inhibitors for the non–mevalonate pathway. This strategy has led to multiple studies aimed at understanding the role of IspH (*lytb*), the final enzyme in the MEP pathway. IspH catalyzes the $2e^{-}/2H^{+}$ reductive dehydroxylation of (E)–4–hydroxy–3–methyl–but–2–enyl diphosphate (HMBPP) to give the isoprenoid precursors IPP and DMAPP in a ~5:1 ratio.[21-26] Only recently has IspH been structurally characterized, with a major hurdle in the process being the elucidation of the character of its central iron–sulfur cluster. Although initial electron paramagnetic resonance (EPR) spectroscopy experiments,[24] as well as two independently solved crystal structures,[27,28] characterized the IspH cluster composition as a [3Fe–4S] center, various spectroscopic studies using different preparation techniques have instead found IspH to be catalytically active with a $[4Fe–4S]^{2+/1+}$ cluster.[24,25,29-31]

Notably, Groll *et al.* have obtained an X–ray structure of the oxidized form of IspH in complex with HMBPP in which the $[4Fe–4S]^{2+}$ cluster is complete.[32] The previously missing Fe atom, designated as Fe1 or the apical Fe, is not ligated by a Cys residue as the other three Fe atoms are, but instead is ligated by the oxygen atom of the HMBPP hydroxyl group (C_4–OH). In this [4Fe–4S] IspH:HMBPP crystal structure,[32] HMBPP sits in a highly polar active site and is coordinated by H41, H74, H124, S225, N227, S269 and a conserved water (W1). HMBPP is also near two residues (E126 and T167) that are proposed to participate in catalysis.[34] Indeed, mutation of the active site E126 leads to undetectable IspH catalytic activity, implying that this residue may be responsible for transferring protons to HMBPP during catalysis.[28,29]

The oxidation state of the [4Fe–4S] cluster and its coordination to HMBPP during IspH catalysis have been investigated using Mössbauer,[30,31,33] electron paramagnetic resonance (EPR)[25,29,34-36] and electron nuclear double resonance (ENDOR) spectroscopies.[29,35-37] These studies, as well as biochemical analyses of IspH catalysis involving analogs of HMBPP[26,38-40] and ^{13}C feeding experiments,[41] have generated complementary data that lend support to two possible catalytic mechanisms for IspH: a Birch reduction mechanism[22,26,31,32,40,42] and an organometallic mechanism where HMBPP forms either a ferraoxetane[36] or a metallacycle intermediate complex involving the Fe1.[10,29,34,35,37,41]

Proponents of the Birch reduction mechanism suggest that upon reduction of the [4Fe–4S] cluster, the reducing electron is transferred to HMBPP concurrent with cleavage of the C_4–OH bond, affording a carbon-centered radical intermediate.[22,26,31,32,38,40,42] Subsequent addition of an electron and proton to this radical intermediate affords IPP or DMAPP. In contrast, Oldfield and co-workers have hypothesized an organometallic mechanism,[29] where, upon reduction of the [4Fe–4S] cluster, HMBPP forms a π–complex between its olefinic carbons (C2 and C3, Figure 2-1) and Fe1 by rotating about the HMBPP C_4–C_3 bond.[29,35,37,41] Following this rotation, two electrons are transferred from the [4Fe–4S] cluster to the HMBPP intermediate concurrent with the breaking of the C_4–OH bond to yield a radical anion with the [4Fe–4S] cluster in a HIPIP–like redox state.[43-45] A subsequent proton transfer affords IPP and DMAPP, while addition of an electron to the [4Fe–4S] cluster regenerates the its oxidized (resting) state.[10,29,34,35,37,41]

Both the Birch reduction and organometallic mechanisms are highly controversial. A more thorough examination of the differences between these possible mechanisms and the data cited to support them has been the subject of several recent review articles.[10,42,46,47]

Here, starting from the Groll *et al.* X–ray crystal structure of the [4Fe–4S] IspH:HMBPP complex, we use density functional theory (DFT)–based computational methods to explore active–site geometries, protonation state energetics and Mössbauer properties. Full DFT treatment is used for the [4Fe–4S] cluster, the substrate, and a number of surrounding, catalytically–important protein sidechains. The remaining environment around the active site is modeled by continuum–based methods, either with the COnductor–like Screening MOdel (COSMO) or with a self–consistent reaction field (SCRF) technique that explicitly includes the protein/solvent environment, including the field produced by the permanent charges and dipoles of the protein. The protonation states considered include all possible states of the HMBPP hydroxyl, which must ultimately be cleaved, and two moieties that are candidates for involvement in acid/base catalysis: the HMBPP pyrophosphate (PP_i) and the sidechain of E126. Because the [4Fe–4S] cluster is in the oxidized (2+) state, this corresponds to the Michaelis complex in either of the proposed catalytic mechanisms. The protonation states of these groups have not been established experimentally but, clearly, they must be part of any detailed description of the mechanism. For each of the protonation states considered, we compare the computed geometry and Mössbauer properties with experimental measurements. Using these comparisons and the relative

computed energetics of the protonation states, we show that the hydroxyl group of HMBPP is protonated (exists as ROH) and that E126 is deprotonated. We also demonstrate that the effects of the permanent charges of the surrounding protein environment, which are included in the SCRF model, as well as the reaction field, are crucial to the correct energetic ordering of protonation states.

Methods

Generation of an IspH active site model

We constructed an active site model for DFT studies (Figure 2-2) using the $[4Fe–4S]^{2+}$ IspH:HMBPP crystal structure solved to 1.7 Å resolution by Grawert *et al.* (PDB ID: 3KE8).[32] Included in this model are the $[4Fe–4S]^{2+}$ cluster, its coordinating thiolates (C13, C96 and C197), the substrate HMBPP and a few key residues that may play a role in catalysis (T167, E126, H41 and H124). Furthermore, due to the large anionic characters of both the [4Fe-4S] cluster and the pyrophosphate (PP_i) tail of HMBPP, we included additional moieties to more realistically interact with these groups. With regard to the PP_i tail, we also include in the model H74, S225, S226, N227 and S269, as these five residues are ideally positioned to donate hydrogen bonds to the PP_i group. Additionally, since hydrogen bonds are known to tune the redox potentials of [4Fe–4S] complexes,[48-51] we include the backbone chains of A199, T200, P97, L98, G14, and V15, as well as the side chain of T200 and a crystallographic water. These groups cumulatively donate the only five hydrogen bonds to the [4Fe–4S] cluster in the [4Fe–4S] IspH:HMBPP crystal structure.[32]

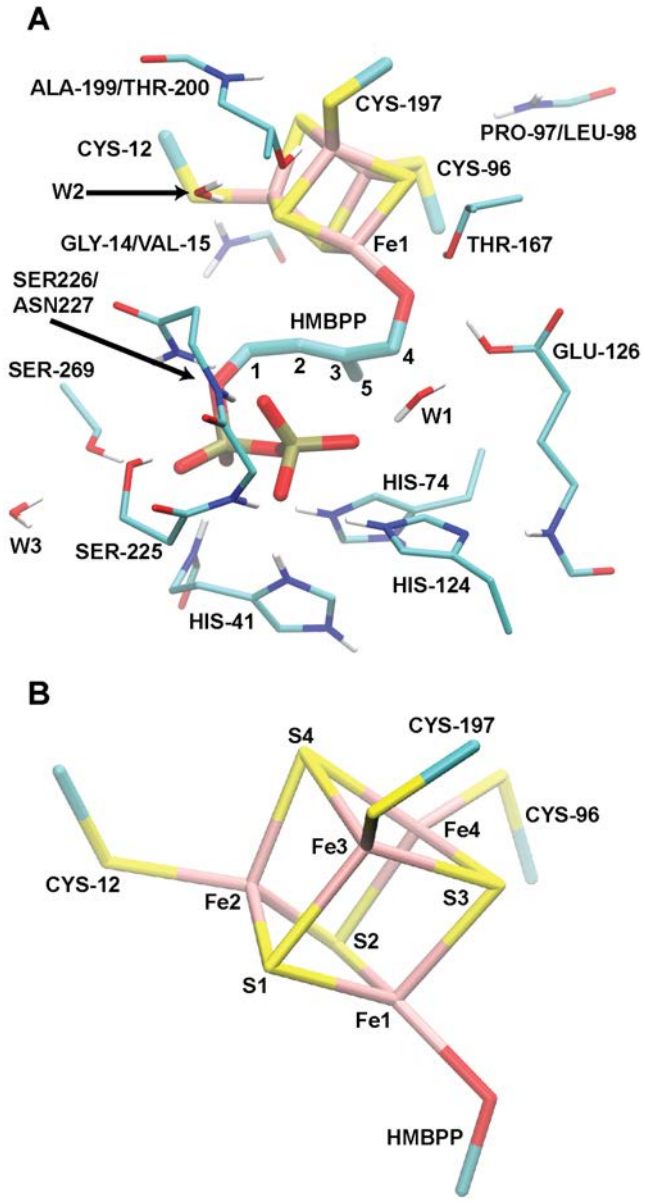

Figure 2-2. (A) Full active site model employed in the DFT/COSMO geometry optimizations and DFT/SCRF single–point energy calculations. (B) Atom labeling used in Table 1 and throughout the text.

In total, the model active site used in these studies contains 203–205 atoms, depending on the protonation state. Adopting such a large active site complex is beneficial because it reduces the dependence of computed geometries and

spectroscopic data on the chosen dielectric environment[52] and allows for a more accurate computation of relative energies in reaction pathways.[53]

Geometry optimizations

Having chosen a model active site from the crystal structure of IspH, the input structure to be used for our BS–DFT calculations is made complete with the addition of hydrogen atoms using Schrödinger's Maestro program.[54] Input structures varying the protonation states of the C_4–OH group of HMBPP, the PP_i moiety of HMBPP and E126 are considered in this study. It is important to note that in order to ensure the active site geometry is reasonable (given the constraints of the surrounding protein), the C_α atoms of all residues, with the exception of the thiolate residues, are constrained to their crystallographic positions.[55] In the case of the thiolates, since they are cut from the protein and capped at the C_β–C_α bond, one hydrogen atom attached to the C_β atom is constrained to lie along the C_β–C_α bond vector.

Because [4Fe–4S] clusters display a high degree of spin polarization and spin–coupling between Fe sites, which individually tend to be high spin, broken-symmetry DFT (BS–DFT) computations are utilized in this study to assess the energetics of different electronic and protonation states of the IspH active site.[56,57] BS–DFT computations are performed, as described previously, to circumvent the inability of conventional DFT methods to obtain an antiferromagnetically (AF)–coupled state. In the case of the IspH [4Fe–4S] cluster, the AF-coupled state has spins on two iron atoms aligned parallel to but opposite the spins on the other two irons.

In practice, BS–DFT solutions are obtained by first computing a ferromagnetically–coupled state, where all Fe atoms are high spin with their spins aligned (in the case of the oxidized [4Fe–4S] cluster considered in this study, S = 18/2). Following this calculation, the spin vector on two of the four Fe atoms is rotated to generate the AF–coupled, "broken–symmetry" (BS) state (S = 0). Geometry optimizations are then started from this BS state.[58]

Given there are four irons that participate in the BS scheme, there exist "4 choose 2" ways to denote Fe–Fe pairs that spin couple. Explicitly, there are six possible states (valence isomers) that can be characterized by the net spin on a specific Fe site: αααββ, αβαβ, αββα, βααβ, βαβα and ββαα. In the case of the αααββ state, Fe_1 and Fe_2 couple and each has an α net spin, while Fe_3 and Fe_4 couple with a β net spin. In the oxidized state investigated here, the two Fe–Fe spin-coupled pairs have identical numbers of unpaired electrons ($S_\alpha = S_\beta$), so only three spin pairs require consideration. Our study includes geometry optimizations of the βααβ, βαβα, αααββ, and αββα states (note that the βααβ state, in principle, is identical to the αββα state). All subsequent results include only the lowest energy valence isomer. A summary of the energies for the different valence isomers considered is given in the Supporting Information (Table 2-S1).

DFT computations are performed using the Amsterdam Density Functional (ADF) 2009 program.[59] All BS–DFT calculations use the OLYP exchange-correlation functional,[60,61] which was chosen due to its ability to generate accurate geometries, to correctly order the energies of different spin states, and to obtain accurate

spectroscopic parameters for various Fe complexes.[62-65] The geometry optimizations employ the Slater-type triple–ζ plus polarization (STO–TZP) basis set on all atoms with the core electrons frozen. The numerical integration accuracy was set to 4.0. Solvent effects in the geometry optimizations are estimated using the COnductor–like Screening MOdel (COSMO)[66,67] with a dielectric constant (ε) of 20. This value for ε is chosen based upon the observation that various geometric and spectroscopic properties appear converged at ε = 20 in studies of active site models for ribonucleotide reductase intermediate X that are similar in size to the IspH active model constructed here.[52] A similar COSMO model (ε = 20) was used in our DFT studies of a different [4Fe–4S] cluster enzyme, adenosine 5'–phosphosulfate (APS) reductase.[55] Following geometry optimization, a single–point energy calculation is performed using the OLYP/STO–TZP level of theory that employs full electron treatment and uses a higher value for the integration accuracy parameter than what is used in the geometry optimizations (i.e., 5.5 versus 4.0). Henceforward, the results from these single–point calculations are referred to as the DFT/COSMO result.

System preparation for DFT/SCRF calculations

Following geometry optimizations conducted in COSMO implicit solvent, all BS states considered in this study are subject to a single–point self–consistent reaction field (SCRF) calculation (henceforth referred to as DFT/SCRF) using the OLYP/STO-TZP basis set with frozen core electrons. In these calculations, the active site region that has been geometry optimized in COSMO (referred to as the DFT/COSMO result) is embedded within the $[4Fe–4S]^{2+}$ IspH:HMBPP crystal structure. All atoms included

in the DFT active site model are deleted from the protein region to avoid overlap between the two regions.[32] Hydrogen atoms are then added to the protein structure in Schrödinger's Maestro program,[54,68,69] employing the PROpKa algorithm[70-73] for assigning protonation states. Following hydrogen addition, all hydrogen bonds are optimized using the IMPACT program.[74] The resulting structure is provided as input to the PDB2PQR server for conversion to a PQR file,[75,76] which is necessary for DFT/SCRF calculations. The PQR file assigns PARSE charges and radii to the protein atoms for use in the continuum electrostatics calculations in the DFT/SCRF scheme.[77]

DFT/SCRF calculations

The first step in DFT/SCRF calculations is a gas-phase single–point energy calculation performed on the BS state. This calculation provides a baseline value from which the effect of stabilizing the active site complex in a protein/solvent environment can be quantified. From the gas-phase computation, point charges are obtained for the active site atoms by fitting to the DFT–derived molecular electrostatic potential (ESP) using the CHELPG algorithm[78] combined with singular value decomposition.[79] In this charge fitting, the total net charge and dipole moment vector are employed as Lagrange constraints, while the link atom (H) charges connecting the active site cluster to the protein are set to zero.

Upon deriving ESP charges for the active site cluster, a three-tiered dielectric environment is defined, whereby the values of 1, 4 and 80 are assigned to the active site quantum cluster, protein and solvent regions, respectively. This environment gives rise to two electrostatic potentials that impinge on the quantum cluster: a reaction

potential arising from the dielectric response of the environment from the cluster charges, and a protein potential due to the permanent partial charges of the protein that are not included in the quantum cluster. These potentials are calculated as numerical solutions to the multi–dielectric Poisson equation using the MEAD (Macroscopic Electrostatics with Atomic Detail) programming suite, as implemented in the interface between ADF2012[80] and MEAD.[81,82] (We note that all calculations were performed using a development version of ADF2011, which yields identical results to SCRF calculations performed with the official release of ADF2012.) These computed protein field and reaction field potentials are subsequently added to the Coulomb interaction operator in the density functional Hamiltonian, and a single–point electronic energy calculation is again performed. This procedure of extracting ESP charges for the active site region in order to solve the Poisson equation for the protein and reaction field potentials is iterated until the sum of the electronic and protein/reaction field energies converges (ΔE between cycles < 0.01 kcal mol^{-1}). A more detailed description of this method can be found elsewhere.[83]

Calculations of relative energetics

Using energies obtained from either the DFT/COSMO or DFT/SCRF methods, the relative free energy of deprotonation at pH = 7, ΔG_{deprot}, can be obtained for a single titratable site through the application of Eq. 1:

$$\Delta G_{deprot} = 1.37\left(pK_a - 7\right) = \left[E(A^-) - E(HA)\right] + \Delta G_{ref}(H^+) + \Delta ZPE + \Delta E_{corr} \quad (1)$$

When computing ΔG_{deprot} with DFT/COSMO, the total energies for the deprotonated and protonated states, $E(A^-)$ and E(HA), respectively, are taken from the BS state energies obtained following the geometry optimization with the DFT/COSMO method. These values follow the usual ADF convention, where $E(A^-)$ and E(HA) are "total" energies with respect to a sum of atomic fragments (spin–restricted atoms).[84] Alternatively, ΔG_{deprot} is computed with the DFT/SCRF method using the total free energies including the protein/solvent environment interaction obtained in the SCRF procedure described previously. We approximate the difference in zero–point energy between protonated and deprotonated states, ΔZPE, as the difference in zero–point energy of the fragment that is titrated (e.g. the carboxylate of E126, the C_4–OH group of HMBPP, or the PP_i tail of HMBPP), as obtained from an OLYP frequency calculation performed on the two protonation states of that fragment. These values are computed to be –8.7, –10.4 and –8.8 kcal mol^{-1} for E126, the C_4–O(H) group of HMBPP and the PP_i moiety on HMBPP, respectively. The free energy of the titrating proton is computed using Eq. 2:

$$\Delta G_{ref}(H^+) = E(H^+) + \Delta G_{sol}(H^+, 1\,atm) - T\Delta S_{gas}(H^+) + \frac{5}{2}RT \quad . \tag{2}$$

While the energy of a proton computed from a gas–phase OLYP calculation with respect to a spin-restricted hydrogen atom is found to be 291.1 kcal mol^{-1}, the value for $E(H^+)$ used in Eq. 2 (292.7 kcal mol^{-1}) is obtained upon the addition of an empirical correction term to the OLYP result. This ensures the computed standard

hydrogen electrode matches experiment exactly.[85,86] For $\Delta G_{sol}(H^+, 1\ atm)$, the solvation free energy of a proton, we use the "best available" value of –264.0 kcal mol^{-1}, based on analysis of cluster-ion solvation data.[87,88] The translational entropy of a proton, $-T\Delta S_{gas}(H^+)$ is taken to be –7.76 kcal mol^{-1}, its value computed theoretically at 298 K and 1 atm.[89] The final term in computing the free energy of the titrating proton, 5/2 RT (1.5 kcal mol^{-1}), arises from the sum of the proton translational energy (3/2 RT) and the work term PV=RT.[87,89] Finally, the term ΔE_{corr} corrects ΔG_{deprot} for a neutral solvent environment (pH = 7), equal to –1.37 x 7 (–9.6 kcal mol^{-1}).

A more general form of Eq. 1 accounts for differences in free energy when any number of protons are titrated:

$$\Delta G_{deprot} = \left[E(A) - E(H_n A) \right] + n\Delta G_{ref}(H^+) + \sum_i (\Delta ZPE)_i + n\Delta E_{corr} \quad (3)$$

where again ΔE_{corr} corrects for a neutral solvent environment (pH = 7), and the number of protons exerts a multiplicative effect on $\Delta G(H^+)$ and ΔE_{corr}. In cases where two states are tautomers (*i.e.* where two states have same number of protons, which themselves reside on different titratable sites), the difference in energy between states is corrected for ΔZPE for the sites that are (de)protonated.

Calculations of ^{57}Fe Mössbauer parameters

Multiple groups have computed ^{57}Fe Mössbauer isomer shifts (ISs) to help identify the oxidation state of the Fe sites considered, drawing on the property that the values of these ISs are linearly proportional to the electron density at the ^{57}Fe nucleus,

ρ(0).[63,65,90-94] In practice, obtaining a value for ρ(0) for a given BS state requires a single–point energy calculation that employs a basis set that includes core electrons and that uses a higher value for the integration accuracy parameter than what is used in the geometry optimizations (i.e., 5.5 versus 4.0).[46] With the hypers2003 program,[95] ρ(0) can then be obtained from the ADF calculation.[59]

Using the linear regression given in Eq. 4, each ^{57}Fe IS can be computed given its ρ(0), using the fitting parameters A, α and C:[65,90]

(4) $$\delta = \alpha\left[\rho(0) - A\right] + C \quad .$$

This linear fit is dependent on the level of theory employed. As all IS computations in this report use the OLYP/STO–TZP level of theory with either the COSMO or SCRF solvation method, we apply appropriately calibrated values of α = –0.323 e^{-1} a_0^3 mm s^{-1}, C = 0.428 mm s^{-1}, and A = 11877 e a_0^{-3}.[65] We note that the DFT/COSMO and DFT/SCRF methods give nearly identical results. The similarity between these two methods is consistent with what has been observed in previous benchmark studies.[65] Consequently, we only present results obtained using the DFT/SCRF method (results from the DFT/COSMO method are given in the Supporting Information, Table 2-S2).

Quadrupole splittings (QSs) arise from the interaction between the ^{57}Fe (S = 3/2) nuclear quadrupole moment and the electric field gradient (EFG) at the nucleus. For this reason, QSs provide useful information on the oxidation state and ligand environment surrounding each ^{57}Fe atom. To obtain values for QSs from our

DFT/COSMO and DFT/SCRF calculations, the EFG must be evaluated at the ^{57}Fe nucleus. Upon reordering of the components of the EFG such that $|V_{zz}| \geq |V_{yy}| \geq |V_{xx}|$, the asymmetry parameter η can be obtained and the QS may be computed (Eq. 5–6) using the constants, e (the electric charge of a positron) and Q (the nuclear excited-state quadrupole moment).

$$\eta = \left| \frac{V_{xx} - V_{yy}}{V_{zz}} \right| \tag{5}$$

$$QS = \frac{1}{2} eQV_{zz} \sqrt{1 + \frac{\eta^2}{3}} \tag{6}$$

Protonation considerations for the active site model

The active site of IspH contains many titratable residues, including E126 (typical pK_a ~ 4.3), H41, H74 and H124 (typical pK_a ~ 6.8).[96,97] Additionally, the C_4–OH group of HMBPP, if complexed strongly to Fe, can be deprotonated (typical pK_a of $[Fe(OH_2)_6]^{3+}$ ~ 2.4),[32] and the PP_i group of HMBPP (typical pK_a values ~ 0.9, 2.0, 6.7, 9.0) can likely assume different protonation states, as well.[32,98] In sum, there are 13 titratable sites, leaving 2^{13} possible protonation states to consider in order to fully characterize the system. To reduce the number of explored states, we make use of some simplifying assumptions, namely that H41 and H74 will likely assume their imidazolium form (doubly protonated) given their close proximity to the highly anionic PP_i tail of HMBPP. Similarly, we choose to fix the protonation state of H124

to be neutral (protonated at N_ε), as its N_ε is in position to donate a single hydrogen bond to the substrate, while receiving a hydrogen bond at its N_δ position from the backbone of E126 in the [4Fe–4S] IspH:HMBPP crystal structure.[32] For E126 we only consider *cis*-protonation at its carboxylate oxygen nearer the HMBPP RO^-/ROH group and Fe1, as this site would allow for participation in an active site hydrogen bond network, as proposed by others (Figure 2-2).[27,29,32] With respect to the substrate, we consider protonation at the C_4–OH alkoxide group, as well as protonation of an oxygen on the terminal phosphate of the PP_i moiety. We consider only single protonation of this group due to the likelihood that the rich hydrogen bond network surrounding the PP_i group will stabilize either its –2 or –3 form. While this still leaves a considerable number of protonation states to consider (2^3), evaluation of the energies of states having both E126 and the HMBPP ROH group deprotonated are found to be energetically unfavorable and, therefore, have been omitted from our discussion.

Naming scheme for protonation states

To facilitate our discussion of the different active site protonation states, we utilize the naming scheme, $RO^XP^YE^Z$, where X, Y and Z are assigned either a minus sign "–" or the letter "H" to signify whether a proton resides on the C_4-OH (RO^-/ROH) group of HMBPP, the PP_i of HMBPP, or E126, respectively (Table 2-1). For instance, in the state $RO^-P^-E^H$, HMBPP has both its terminal C_4–OH group and its PP_i moiety deprotonated (giving the ligand a net charge of –4), while E126 is protonated. Combined with the oxidized $[Fe_4S_4Cys_3]^{1-}$ cluster and two imidazoles H41 and H74 (q = +1, each), the model active site carries a total charge of –3 in this $RO^-P^-E^H$ state

(Table 2-1). We stress that within this naming scheme the "–" is representative of the charges of the RO^{-}/ROH and the E126 groups; however, it does not indicate the net charge of the PP_i group, which maintains a charge of –3 when deprotonated (P^{-}) and –2 when protonated (P^{H}).

Table 2-1. A description of the nomenclature scheme used throughout this report. Unless an H is listed, the moiety described by each column is assumed to be fully deprotonated. The total charge (q) of the active site quantum cluster is given for each state.

Protonation state	q	C_4–O(H)	PP_i	E126
$\mathbf{RO^{-}P^{-}E^{H}}$	–3	–	–	H
$\mathbf{RO^{-}P^{H}E^{H}}$	–2	–	H	H
$\mathbf{RO^{H}P^{-}E^{-}}$	–3	H	–	–
$\mathbf{RO^{H}P^{H}E^{-}}$	–2	H	H	–
$\mathbf{RO^{H}P^{-}E^{H}}$	–2	H	–	H
$\mathbf{RO^{H}P^{H}E^{H}}$	–1	H	H	H

Results

The HMBPP-bound [4Fe-4S] IspH crystal structure

The IspH:HMBPP crystal structure referenced in this study contains a $[4Fe–4S]^{2+}$ cluster with significant asymmetry (Table 2-2, with atom numbering defined in Figure 2-2B).[32] The apical iron, Fe1, has longer Fe–S^{2-} bond lengths (Fe–S^{2-} distances of 2.34, 2.39 and 2.36 Å) than the other Fe atoms. In contrast, a second, ferric–like iron (Fe2 in Figure 2-2B) contains significantly shorter Fe–S^{2-} bonds, with Fe–S^{2-} distances of 2.22, 2.19 and 2.18 Å. The other two irons have intermediate bond lengths (mean Fe-S distance of 2.28 ± 0.04 Å), representative of a delocalized, mixed–valence pair ($Fe^{2.5+}$).[32]

Table 2-2. Tabulation of various [4Fe–4S] cluster distances (in Å) obtained from geometry optimizations of various protonation conformers of an active site model of the oxidized IspH system. For comparison, the corresponding bond lengths from the IspH:HMBPP crystal structure are given.[32] The root–mean–square deviation (RMSD, given in Å) between different protonation conformers computed for oxidized IspH and the IspH:HMPP crystal structure are tabulated.

	Exp[32]	**$RO^{-}P^{-}E^{H}$**	**$RO^{-}P^{H}E^{H}$**	**$RO^{H}P^{-}E^{-}$**	**$RO^{H}P^{H}E^{-}$**	**$RO^{H}P^{-}E^{H}$**	**$RO^{H}P^{H}E^{H}$**
Cluster distances							
Fe1–S1	2.344	2.275	2.281	2.218	2.215	2.200	2.209
Fe1–S2	2.393	2.351	2.351	2.301	2.298	2.282	2.280
Fe1–S3	2.364	2.378	2.370	2.328	2.317	2.319	2.313
Fe2–S1	2.217	2.340	2.347	2.323	2.324	2.328	2.333
Fe2–S2	2.186	2.212	2.213	2.203	2.201	2.206	2.210
Fe2–S4	2.181	2.385	2.380	2.362	2.359	2.357	2.353
Fe3–S1	2.319	2.328	2.330	2.313	2.317	2.305	2.305
Fe3–S3	2.281	2.240	2.239	2.236	2.237	2.240	2.238
Fe3–S4	2.306	2.377	2.372	2.357	2.350	2.361	2.360
Fe4–S2	2.308	2.321	2.322	2.314	2.319	2.326	2.325
Fe4–S3	2.217	2.320	2.323	2.320	2.324	2.322	2.326
Fe4–S4	2.276	2.279	2.280	2.245	2.242	2.241	2.242
$Fe2–S_{C12}$	2.283	2.305	2.303	2.263	2.258	2.256	2.257
$Fe3–S_{C197}$	2.285	2.314	2.306	2.283	2.274	2.275	2.271
$Fe4–S_{C96}$	2.264	2.321	2.314	2.295	2.290	2.295	2.291
Fe1–Fe2	2.792	2.815	2.822	2.692	2.674	2.681	2.692
Fe1–Fe3	2.780	2.757	2.747	2.676	2.656	2.643	2.641
Fe1–Fe4	2.820	2.752	2.728	2.672	2.658	2.645	2.631
Fe2–Fe3	2.717	2.764	2.762	2.752	2.738	2.797	2.797
Fe2–Fe4	2.694	2.767	2.770	2.729	2.723	2.729	2.730
Fe3–Fe4	2.749	2.751	2.755	2.719	2.720	2.722	2.725

The HMBPP molecule coordinates Fe1 through its oxygen atom attached to C4 at a distance of 2.05 Å ($Fe1–O_{C4}$, Table 2-3), which Groll and co-workers propose to be indicative of HMBPP binding as an alkoxide (RO^{-}) rather than as an alcohol (ROH).[32] In addition to the interaction between HMBPP and Fe1 of the [4Fe–4S]

cluster, a significant hydrogen bond network further supports HMBPP within the protein active site. Indeed, several histidines and alcohol side chains forming (charged) hydrogen bonds to the PP_i tail of HMBPP are present. Further, the active site residues T167, E126 and the crystallographic water W1 link the oxygen bonded to C4 in HMBPP to its PP_i tail through a series of hydrogen bonds (Table 2-3, Figure 2-3A). This latter group of moieties has been suggested to participate in a proton relay network for catalysis.[32]

Table 2-3. Key HMBPP bond lengths and distances (in Å) between hydrogen bonding atoms in the active site of oxidized IspH as computed by geometry optimizations of different protonation conformers. O_{C4} refers to the oxygen attached to C4 on HMBPP, O_{T167} refers to the T167 hydroxyl oxygen, O_{E126} refers to the E126 carboxylate oxygen directed inward toward the active site, O_{W1} is a conserved active site water, and O_{PPi} is the oxygen on the PP_i group that interacts with W1. For comparison, the corresponding distances from the IspH:HMBPP crystal structure are given,[32] and the state computed to have the best agreement with each geometric parameter from experiment is set in bold–type font.[32]

	Exp[32]	**$RO^-P^-E^H$**	**$RO^-P^HE^H$**	**$RO^HP^-E^-$**	**$RO^HP^HE^-$**	**$RO^HP^-E^H$**	**$RO^HP^HE^H$**
HMBPP distances							
Fe1–O_{C4}	2.046	1.891	1.899	**2.108**	2.133	2.214	2.254
Fe1–C2	2.913	3.136	**3.107**	3.267	3.266	3.272	3.220
Fe1–C3	3.039	3.497	3.406	3.627	3.551	3.531	**3.399**
Active site H-bonds							
O_{C4}–O_{Thr}	2.702	3.203	3.182	**2.816**	2.914	3.189	3.179
O_{Thr}–O_{Glu}	2.761	2.628	2.639	**2.770**	2.771	2.661	2.667
O_{Glu}–O_{W1}	2.578	3.618	3.745	3.133	**2.907**	3.687	3.711
O_{W1}–O_{PPi}	2.548	**2.723**	2.858	2.746	3.006	2.734	2.836
RMSD	-	0.57	0.59	**0.40**	**0.40**	0.57	0.58

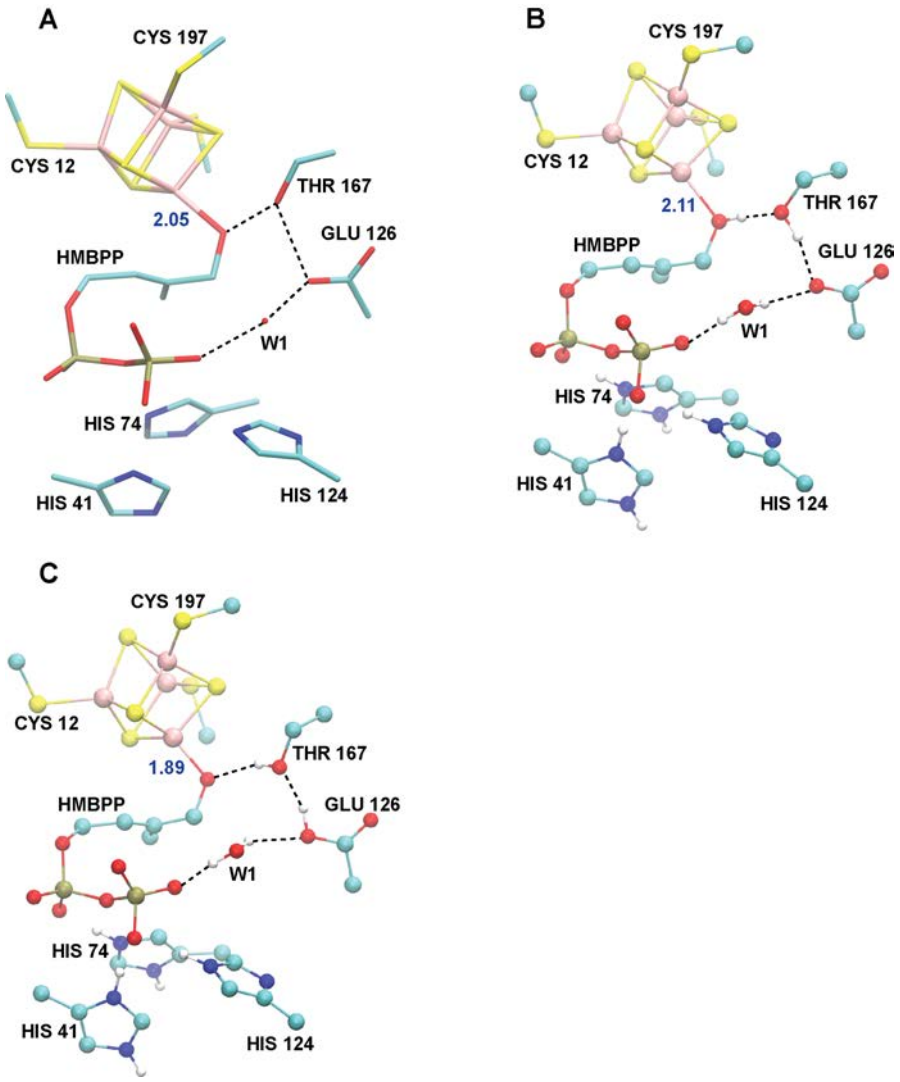

Figure 2-3. (A) Crystal structure active site environment of oxidized IspH (PDB ID: 3KE8).[32] (B) Optimized active site geometry for the $RO^{H}P^{-}E^{-}$ state. (C) Geometry optimized active site geometry of the $RO^{-}P^{-}E^{H}$ state. Active site hydrogen bond networks are indicated with dotted lines.

Computed [4Fe-4S] cluster geometries

Regardless of the protonation state of our model, none of the computed geometries for the IspH active site fully capture the asymmetry observed in the crystal structure (Table 2-2). More specifically, the Fe1–S^{2-} and Fe2–S^{2-} distances in our

computed geometries are not uniformly elongated and shortened, respectively. Instead we observe a four-fold compression of the oxidized [4Fe–4S] cluster, whereby the planes involving the two 2Fe–2S fragments are compressed along one direction characterized by short Fe–S^{2-} bonds. For instance, in the case of the lowest energy structure computed for the $RO^{-}P^{-}E^{H}$ state (Table 2-2, Figure 2-3B), Fe1 has Fe–S^{2-} bond lengths of 2.28, 2.35 and 2.38 Å; Fe2 has Fe–S^{2-} bond lengths of 2.21, 2.34 and 2.39 Å; Fe3 has Fe-S^{2-} bond lengths of 2.24, 2.33 and 2.38 Å; and Fe4 has Fe–S^{2-} bond lengths of 2.28, 2.32 and 2.32 Å (Table 2-2). Thus, for the $RO^{-}P^{-}E^{H}$ state, each iron atom has one short and two long Fe–S^{2-} bonds. This trend is evident for all other protonation states as well (Table 2-2).

Geometries of different HMBPP protonation states

The geometries of the different protonation states considered can first be categorized by whether the HMBPP C_4–OH group exists as an alkoxide (RO^{-}) or alcohol moiety (RO^{H}, in our nomenclature). Two protonation states are computed for the RO^{-} state—specifically, the $RO^{-}P^{-}E^{H}$ and $RO^{-}P^{H}E^{H}$ states, which differ only by a single proton on the PP_i group. Consequently, these two RO^{-} states give total charges of –4 and –3 for HMBPP, respectively. These computed structures display Fe1–O distances (~1.90 Å, Table 2-3) that are appreciably shorter than the Fe1–O distance observed in the IspH:HMBPP crystal structure (2.05 Å).[32] While the geometries of the $RO^{-}P^{-}E^{H}$ and $RO^{-}P^{H}E^{H}$ states are largely similar, displaying a root-mean-square deviation (RMSD) of 0.19 Å, it is clear that protonation of the PP_i group does alter the position of the conserved active site water, W1. In turn, the oxygen-oxygen distances

between E126/W1 and W1/PP_i hydrogen bond partners in the active site differ between the $RO^{-}P^{-}E^{H}$ and $RO^{-}P^{H}E^{H}$ states (Table 2-3). Both states, however, display reasonable similarity to the [4Fe–4S] IspH:HMBPP crystal structure (RMSD ~ 0.6 Å, Table 2-3).[32]

When HMBPP binds Fe1 via an alcohol group (as in the $RO^{H}P^{-}E^{-}$, $RO^{H}P^{-}E^{H}$, $RO^{H}P^{H}E^{-}$, $RO^{H}P^{H}E^{H}$ states), the computed Fe1–O bond lengths (ranging from 2.11 to 2.25 Å) are longer than what is observed in the [4Fe–4S] IspH:HMBPP crystal structure (2.05 Å).[32] These ROH–bound states also possess shorter Fe1–S^{2-} bond lengths compared to those observed in the RO^{-}–bound states; the average Fe1–S^{2-} bond length is 2.27 Å in the ROH states, while it is 2.33 Å in the RO^{-} states (Table 2-2). We note, however, that these differences in cluster geometry are restricted to Fe1 and not the other three Fe atoms. Indeed, the average Fe2–S^{2-}, Fe3–S^{2-} and Fe4–S^{2-} distances in the RO^{-} and ROH states are similar to within 0.01 Å (Table 2-2).

Despite the differences in the bond distances involving Fe1, we observe strong structural similarity between the RO^{-} and ROH states when all other protonation states are maintained, as is the case for $RO^{-}P^{H}E^{-}$/$RO^{H}P^{H}E^{-}$ and $RO^{-}P^{-}E^{H}$/$RO^{H}P^{-}E^{H}$ (Table 2-3). This observation is further supported by measurements of RMSD between the $RO^{-}P^{H}E^{-}$/$RO^{H}P^{H}E^{-}$ and $RO^{-}P^{-}E^{H}$/$RO^{H}P^{-}E^{H}$ pairs, giving values of 0.05 and 0.04 Å, respectively. Similar to the RO^{-} states, (de)protonation of the PP_i group in the HMBPP ROH-bound states perturbs the position of W1 and its corresponding hydrogen bond interactions with PP_i and E126 (Table 2-3).

While the active site geometries following (de)protonation of the RO^{-}/ROH and PP_i groups show only local variation around these titratable groups, significantly larger deviations in active site structure are observed upon the (de)protonation of E126. With respect to the coordination of HMBPP to Fe1, the Fe1-OH distances are significantly shorter in E126 deprotonated states (2.11 and 2.13 Å for the $RO^{H}P^{-}E^{-}$ and $RO^{H}P^{H}E^{-}$ states, respectively) than their protonated counterparts (Fe1-O distances of 2.21 and 2.25 Å for $RO^{H}P^{-}E^{H}$ and $RO^{H}P^{H}E^{H}$, respectively). Aside from the HMBPP ROH group, E126 is the only titratable group whose protonation significantly alters all hydrogen bond partners in the active site (Table 2-3). As can be seen in Table 2-3, the oxygen–oxygen distances for active site hydrogen bond partners are mostly similar between the [4Fe–4S] IspH:HMBPP crystal structure and the ROH/E^{-} states, regardless of PP_i protonation. For instance, the distance between the C_4–OH group of HMBPP and the alcohol side chain of T167 is ~2.8–2.9 Å in the RO^{H}/E^{-} states, ~3.2 Å in all RO^{-} and RO^{H}/E^{H} states, and 2.70 Å in the crystal structure (Table 2-3).[32]

Visually, the closer agreement between the E^{-} states with the crystal structure can be attributed to different orientations of the T167 and E126 sidechains, as well as the position of the active site water, W1 (Figure 2-3B,C). When E126 is deprotonated (Figure 2-3B), the hydrogen bond network originates at the ROH group in HMBPP, which donates a hydrogen bond to T167, which then donates a hydrogen bond to the deprotonated E126. In contrast, the hydrogen bond network in the protonated E126 states (Figure 2-3C) requires that the T167 sidechain accept a hydrogen bond from E126 and donate a hydrogen bond to the HMBPP ROH group. To quantify these

observations, we note that geometries computed with E126 deprotonated better agree with the [4Fe–4S] IspH:HMBPP crystal structure[32] (RMSD = 0.40 Å, Table 2-3) than structures involving protonated forms of E126 (RMSD = 0.57–0.59 Å). From visualizing the superpositions of computed $RO^{-}P^{-}E^{H}$ and $RO^{H}P^{-}E^{-}$ geometries and the IspH:HMBPP crystal structure,[32] it is clear better agreement with experiment is achieved when E126 is deprotonated (Supporting Information, Figure 2-S1).

Relative energetics in the oxidized state

For all states considered in this study, we implement two methods for computing the energetics of the system: (1) the DFT/COSMO method with ε = 20; and (2) the DFT/SCRF method, which allows the active site model described using DFT to be perturbed by the electrostatic effects of protein charges and the dielectric response of the protein and solvent regions (ε = 4 and 80, respectively). Although both methods have shown good agreement with experiment for calculated energies and reduction potentials of model systems,[51,79,99] the DFT/SCRF method has the advantage that it may provide a more realistic description of the protein and solvent environments.[83] While only the energies of the lowest energy valence isomer for each protonation state are presented here, a summary of the relative energies for all valence isomers considered in this study is given in the Supporting Information (Table 2-S1).

Using the DFT/COSMO method, the $RO^{-}P^{H}E^{H}$ and $RO^{H}P^{H}E^{H}$ states are found to be iso-energetic and the lowest–energy oxidized states computed at pH = 7 (Table 2-4A, Figure 2-4). The $RO^{H}P^{-}E^{H}$ state is 2.9 kcal mol^{-1} above the energy of the $RO^{-}P^{H}E^{H}$ and $RO^{H}P^{H}E^{H}$ states, while the $RO^{-}P^{-}E^{H}$ and $RO^{H}P^{-}E^{-}$ states are the two

highest–energy states computed with the DFT/COSMO energy model, at energies that are, respectively, 7.9 and 8.4 kcal mol^{-1} above the $RO^{-}P^{H}E^{H}$ and $RO^{H}P^{H}E^{H}$ states (Figure 2-4). The DFT/COSMO method thus preferentially stabilizes states where either two or three of the titratable sites considered carry a proton.

In contrast to the DFT/COSMO results, use of the DFT/SCRF method shows the $RO^{-}P^{-}E^{H}$ and $RO^{H}P^{-}E^{-}$ states to be iso–energetic and lowest–energy (Table 2-4C, Figure 2-4). The $RO^{-}P^{H}E^{H}$ and $RO^{H}P^{-}E^{H}$ states are next lowest in energy at 2.2 and 5.5 kcal mol^{-1} above the $RO^{-}P^{-}E^{H}$ and $RO^{H}P^{-}E^{-}$ states, while the fully protonated $RO^{H}P^{H}E^{H}$ state is the highest energy state (at 13.3 kcal mol^{-1} above the $RO^{-}P^{-}E^{H}$ and $RO^{H}P^{-}E^{-}$ states).

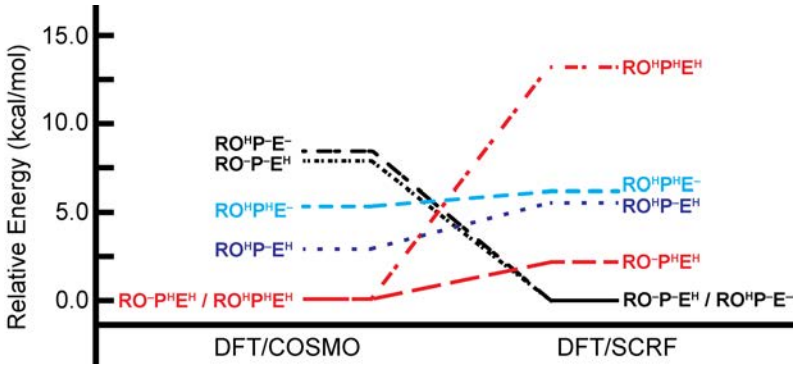

Figure 2-4. Relative energies of the different protonation conformers computed from Eq. 3 using (Left) DFT/COSMO and (Right) DFT/SCRF.

Table 2-4. Relative energies computed at pH = 7 for the different protonation states of the oxidized state of the IspH active site cluster, using Eq. 3 with energies obtained from (A) the DFT/COSMO method; (B) the total free energy stemming from the gas–phase energy plus the reaction field (RF) component of the DFT/SCRF method (*i.e.* the sum of E_0 and E_{RF} from Table 5); (C) the *full* DFT/SCRF method. All energies are given in kcal mol^{-1}. [a–d]Decomposed terms used to compute relative energies with Eq. 3: [a]The total energy of the protonation state considered, as obtained from (A) DFT/COSMO, (B) the RF component of the DFT/SCRF method or (C) the *full* DFT/SCRF method. [b]The relative free energy of the titrating proton(s), taken from Eq. 2. [c]Energies correcting for differences in zero–point energy between protonation states. [d]Correction term applied to obtain relative energies at pH = 7.

A.

	DFT/COSMO				
State (Charge)	$\Delta E_{Tot,COSMO}$ [a]	$n\Delta G_{ref}(H^+)$ [b]	$\Sigma(ZPE)_i$ [c]	$n\Delta E_{Corr,pH=7}$ [d]	$\Delta E_{Tot,COSMO,pH=7}$ [e]
$RO^-P^-E^H$ (–3)	3.9	22.4	-8.8	-9.6	7.9
$RO^-P^HE^H$ (–2)	0.0	0.0	0.0	0.0	0.0
$RO^HP^-E^-$ (–3)	2.6	22.4	-7.1	-9.6	8.4
$RO^HP^HE^-$ (–2)	3.6	0.0	1.7	0.0	5.3
$RO^HP^-E^H$ (–2)	1.3	0.0	1.6	0.0	2.9
$RO^HP^HE^H$ (–1)	2.4	-22.4	10.4	9.6	0.0

B.

	DFT/SCRF Reaction Field				
State (Charge)	$\Delta E_{Tot,RF}$	$n\Delta G_{ref}(H^+)$	$\Sigma(ZPE)_i$	$n\Delta E_{Corr,pH=7}$	$\Delta E_{Tot,RF,pH=7}$
$RO^-P^-E^H$ (–3)	9.5	44.9	-19.2	-19.2	16.0
$RO^-P^HE^H$ (–2)	-1.4	22.4	-10.4	-9.6	1.1
$RO^HP^-E^-$ (–3)	10.3	44.9	-17.5	-19.2	18.5
$RO^HP^HE^-$ (–2)	4.2	22.4	-8.7	-9.6	8.4
$RO^HP^-E^H$ (–2)	2.4	22.4	-8.8	-9.6	6.5
$RO^HP^HE^H$ (–1)	0.0	0.0	0.0	0.0	0.0

C.

	***Full* DFT/SCRF**				
State (Charge)	$\Delta E_{Tot,SCRF}$	$n\Delta G_{ref}(H^+)$	$\Sigma(ZPE)_i$	$n\Delta E_{Corr,pH=7}$	$\Delta E_{Tot,SCRF,pH=7}$
$RO^-P^-E^H$ (–3)	1.7	0.0	-1.7	0.0	0.0
$RO^-P^HE^H$ (–2)	8.0	-22.4	7.1	9.6	2.2
$RO^HP^-E^-$ (–3)	0.0	0.0	0.0	0.0	0.0
$RO^HP^HE^-$ (–2)	10.1	-22.4	8.8	9.6	6.1
$RO^HP^-E^H$ (–2)	9.6	-22.4	8.7	9.6	5.5
$RO^HP^HE^H$ (–1)	21.5	-44.9	17.5	19.2	13.3

Comparison of DFT/COSMO and DFT/SCRF

It is clear the relative energetics computed with DFT/COSMO and DFT/SCRF descriptions of the IspH active site model differ. While the $RO^{-}P^{H}E^{H}$ and $RO^{H}P^{H}E^{H}$ states, with their respective net charges (q) of –2 or –1, are preferred energetically when using the DFT/COSMO method, the DFT/SCRF method favors a highly anionic (q = –3) quantum cluster, where two of the three considered titratable sites are deprotonated ($RO^{-}P^{-}E^{H}$ and $RO^{H}P^{-}E^{-}$ states, Figure 2-4). In the gas phase, the energy of the $RO^{H}P^{H}E^{H}$ state (q = –1) is 152 kcal mol^{-1} lower than that of the more negatively–charged (q = –3) $RO^{H}P^{-}E^{-}$ state (Table 2-5A,B). Solvation with either the COSMO or SCRF methods can accommodate greater charge in the quantum cluster and lessens (or even reverses) the trend in energetics obtained from gas–phase energies alone. From DFT/COSMO and DFT/SCRF computations, it is possible to isolate the energetic contributions arising from embedding the active site in an environment containing regions of different dielectric constants and, in the case of the DFT/SCRF computations, protein point charges. Understanding these energetic contributions provides insight into why certain protonation states (or, alternatively, states with specific net charges in the active site quantum cluster) are stabilized relative to others when using DFT/COSMO or DFT/SCRF methods.

Table 2-5. A summary of the components of the total electrostatic energy computed by the DFT/COSMO and DFT/SCRF[83] methods for different protonation states of the oxidized IspH active site. (A) Energies (in kcal mol^{-1}) are presented as total energies. (B) Relative energies (in kcal mol^{-1}) are given with respect to the energy of the $RO^{H}P^{-}E^{-}$ state. [a]The total gas–phase electronic energy of the active site quantum cluster obtained following geometry optimization with the DFT/COSMO method. [b]The stabilization of the quantum cluster from COSMO solvation with ε = 20 (obtained from Eq. 7). $E_{solv,COSMO}$ contains both quantum cluster polarization, solvent interaction and electronic strain terms, analogous to those in $E_{env,SCRF}$. [c]The total energy computed using the DFT/COSMO method (*i.e.* the sum of E_0 and $E_{Solv,COSMO}$) performed on the COSMO optimized geometry and used to compute relative energies in Table 4A. [d]The energetic cost of polarizing the active site quantum cluster in response to the protein and reaction potentials in the DFT/SCRF scheme. [e]The total protein field energy, including electronic polarization of the quantum cluster, resulting from interactions between active site and protein charges that are screened by the three dielectric media (ε = 1, 4 and 80 for the quantum cluster, protein region and solvent, respectively). [f]The total reaction field energy, including electronic polarization of the quantum cluster, arising from the dielectric response of the protein (ε = 4) and solvent (ε = 80) environments from the cluster charges. [g]The total environmental (env) energy from the DFT/SCRF method (i.e. the sum of E_{Strain}, E_P and E_{RF}). [h]The total free energy associated with the total gas–phase electronic energy of the quantum cluster and the reaction field component of the DFT/SCRF method (*i.e.* the sum of E_0 and E_{RF}), also used to generate the relative energies given in Table 4B. [i]The electrostatic free energy of a given state computed by the *full* DFT/SCRF method (*i.e.* the sum of E_0, E_{Strain}, E_P and E_{RF}) and used to obtain the relative energies in Table 4C.

A.

	DFT/COSMO			DFT/SCRF						
State (Charge)	E_0[a]	$E_{Solv,COSMO}$[b]	$E_{Tot,COSMO}$[c]	E_0	E_{Strain}[d]	E_P[e]	E_{RF}[f]	$E_{env,SCRF}$[g]	$E_{Tot,RF}$[h]	$E_{Tot,SCRF}$[i]
$RO^{-}P^{-}E^{H}$ (–3)	-26445.7	-286.7	-26732.4	-26445.7	53.9	-129.9	-304.3	-380.3	-26750.0	-26826.0
$RO^{-}P^{H}E^{H}$ (–2)	-26541.2	-195.0	-26736.2	-26542.8	52.1	-110.9	-218.1	-276.9	-26760.9	-26819.7
$RO^{H}P^{-}E^{-}$ (–3)	-26440.5	-293.1	-26733.6	-26440.5	54.8	-133.3	-308.7	-387.2	-26749.2	-26827.7
$RO^{H}P^{H}E^{-}$ (–2)	-26528.4	-204.2	-26732.7	-26529.1	52.9	-115.3	-226.2	-288.5	-26755.3	-26817.6
$RO^{H}P^{-}E^{H}$ (–2)	-26537.7	-197.3	-26735.0	-26537.7	53.2	-114.2	-219.4	-280.4	-26757.1	-26818.1
$RO^{H}P^{H}E^{H}$ (–1)	-26592.2	-141.6	-26733.8	-26592.2	49.8	-96.5	-167.3	-214.0	-26759.5	-26806.2

B.

	DFT/COSMO			DFT/SCRF						
State (Charge)	ΔE_0	$\Delta E_{Solv,COSMO}$	$\Delta E_{Tot,COSMO}$	ΔE_0	ΔE_{Strain}	ΔE_P	ΔE_{RF}	$\Delta E_{env,SCRF}$	$\Delta E_{Tot,RF}$	$\Delta E_{Tot,SCRF}$
$RO^{-}P^{-}E^{H}$ (–3)	-5.2	6.4	1.3	-5.2	-0.9	3.4	4.4	6.9	-0.8	1.7
$RO^{-}P^{H}E^{H}$ (–2)	-100.7	98.1	-2.6	-102.3	-2.7	22.4	90.6	110.3	-11.7	8.0
$RO^{H}P^{-}E^{-}$ (–3)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$RO^{H}P^{H}E^{-}$ (–2)	-87.9	88.9	1.0	-88.6	-1.9	18.0	82.5	98.7	-6.1	10.1
$RO^{H}P^{-}E^{H}$ (–2)	-97.2	95.9	-1.3	-97.2	-1.6	19.1	89.3	106.8	-7.9	9.6
$RO^{H}P^{H}E^{H}$ (–1)	-151.7	151.5	-0.2	-151.7	-5.0	36.8	141.4	173.2	-10.3	21.5

The stabilization of the gas–phase active site cluster by COSMO, $E_{Solv,COSMO}$ is computed as

(7) $$E_{Solv,COSMO} = E_{Tot,COSMO} - E_0 \ ,$$

where E_0 and $E_{Tot,COSMO}$ are the total electronic energies of the system in the gas phase and as solvated by COSMO, respectively. The values of $E_{Solv,COSMO}$ range from –142 kcal mol^{-1}, in the case of the $RO^HP^HE^H$ state (q = –1), to –293 kcal mol^{-1}, in the case of the $RO^HP^-E^-$ state (q = –3). Unsurprisingly, the more highly–charged states under study induce a larger reaction field and are, consequently, more stabilized by the COSMO solvation environment (ε = 20). This greater stabilization of states with q = –3 over those with q = –1 is approximately equal in magnitude (151 kcal mol^{-1}) to the greater gas–phase stabilization of states with q = –1 over states with q = –3. Consequently, the values of $E_{Tot,COSMO}$ obtained using the DFT/COSMO method result from the cancellation of additive terms, falling into a range of relative energies that is considerably smaller than the range of the individual energetic contributions.

To better understand the differences between the DFT/COSMO and DFT/SCRF results, we now consider the latter method in a stepwise manner. First, we quantify the effects of only the reaction field energies on the relative energies of the different protonation states (Table 2-4B). Then, we consider the *full* DFT/SCRF energies, which, in addition to the reaction field, also include the effects of the protein field and electronic strain, which is the energetic cost of distorting the quantum cluster

electron density in response to the protein and reaction fields (Table 2-4C). Similar to COSMO, the SCRF reaction field (RF) potential gives greater stabilization to the more highly–charged states. From Table 2-5B, it is evident the RF stabilization (ΔE_{RF}) of the $RO^{H}P^{-}E^{-}$ state (q = –3) is 141 kcal mol^{-1} greater (more negative) than that of the $RO^{H}P^{H}E^{H}$ state (q = –1). Using a Hamiltonian comprising only the gas–phase energies and the RF potential, we obtain the relative energies given in Table 2-4B (last column), which appear similar to those computed with the DFT/COSMO method (Table 2-4A, last column): the $RO^{H}P^{H}E^{H}$ state (q = –1) is lowest in energy, with the $RO^{-}P^{H}E^{H}$ state (q = –2) only 1.1 kcal mol^{-1} above it.

The feature of the full SCRF treatment that re-orders the energy rankings and makes the q = –3 states more favorable (Table 2-4C) is the protein field term, E_P. Without it, the $\Delta E_{Tot,RF}$ values (second column of Table 2-4B) of the two q = –3 states, $RO^{-}P^{-}E^{H}$ and $RO^{H}P^{-}E^{-}$, are 9.5 and 10.3 kcal mol^{-1}, respectively, higher than that of the lowest-energy q = –1 state ($RO^{H}P^{H}E^{H}$). But as Table 2-5 shows, the protein field term (E_p) favors $RO^{-}P^{-}E^{H}$ and $RO^{H}P^{-}E^{-}$ over the $RO^{H}P^{H}E^{H}$ state by 36.8 and 33.4 kcal mol^{-1}, respectively. This stabilization of q = –3 states by the E_P term is more than enough to change the ordering of relative energies (Table 2-4C). The E_{strain} term slightly favors the less charged states, but is not large enough to change the ordering. Further, the corrections for proton free energy, ZPE and pH are significant but do not change the ordering with respect to overall charge (Table 2-4C).

Computed Mössbauer isomer shifts

Mössbauer isomer shifts (ISs) are computed for the various protonation states considered in our oxidized IspH model (Table 2-6). The average IS, δ_{ave}, for all protonation states is computed to be ~0.43 mm s^{-1}. The range of ISs present in the different states, however, differs depending on the protonation of the HMBPP C_4–OH group. For all RO^- states, Fe1 displays an IS that is nearly indistinguishable from those of all other irons in the cluster (0.44, compared to 0.42 or 0.43 mm s^{-1}). In contrast, the apical Fe in ROH–bound states has an IS of 0.53 – 0.54 mm s^{-1}, which is significantly larger than those computed for the other Fe atoms (Table 2-6). In these states, there is also a Fe site that gives an IS that is significantly lower than the mean value (0.38 – 0.39 mm s^{-1}).

Table 2-6. Tabulation of individual and average isomer shifts (ISs) computed using the DFT/SCRF method for different protonation state conformers (given in mm s^{-1}) and compared with experiment.[33] The error with respect to the experimental values is given as mean absolute error (MAE).

	Exp[33]	**$RO^-P^-E^H$**	**$RO^-P^HE^H$**	**$RO^HP^-E^-$**	**$RO^HP^HE^-$**	**$RO^HP^-E^H$**	**$RO^HP^HE^H$**
Fe1	0.64	0.44	0.45	0.54	0.53	0.53	0.54
Fe2	0.38	0.42	0.42	0.39	0.38	0.38	0.38
Fe3	0.42	0.43	0.43	0.42	0.41	0.41	0.40
Fe4	0.42	0.43	0.42	0.40	0.40	0.40	0.39
Ave	0.47	0.43	0.43	0.44	0.43	0.43	0.43
MAE	-	0.06	0.06	0.03	0.04	0.04	0.04

Both low–[30] and high–field[33] Mössbauer spectroscopic parameters are available for the $[4Fe–4S]^{2+}$ (oxidized) IspH:HMBPP complex. Our computed isomer shifts are reasonably consistent with the experiments performed at low–field;[30]

however, the high–field Mössbauer parameters are considered to be of superior quality.[100] The IS values obtained from their study display significant asymmetry, ranging from 0.38 to 0.64 mm s^{-1}, with the IS of 0.64 mm s^{-1} attributed to the ferrous, HMBPP-bound Fe1.[33] When comparing our computed IS values to these experimental results, it is clear the asymmetric ISs computed for ROH–bound states better agree with experiment, having a mean absolute error (MAE) of 0.03–0.04 mm s^{-1} (compared with 0.06 mm s^{-1} for the RO^-–bound states). These MAEs for the ROH–bound states are also within the systematic error associated with computing ISs using the OLYP functional.[65] While the IS corresponding to the ferric site (Fe2 in Figure 2-2) in our ROH calculations matches its experimental counterpart, it is clear that the IS computed for Fe1 (with a ferrous IS of 0.53–0.54 mm s^{-1}) underestimates the experimental value (0.64 mm s^{-1}).[33]

Computed Mössbauer quadrupole splittings

The Mössbauer quadrupole splittings (QSs) corresponding to the ISs displayed in Table 2-6 are given in Table 2-7. We note that the MAEs (relative to experiment) of the QSs computed with the DFT/COSMO and DFT/SCRF methods are almost identical (Table 2-7, Table 2-S3). We observe this despite the latter approach containing a more extensive description of the surrounding electrostatics, which would be expected to affect the QS values through the values of V_{zz} and η (Eqs. 5–6). Here, we only discuss QSs computed using the DFT/SCRF method.

Table 2-7. Tabulation of individual and average quadrupole splittings (QSs) computed using the DFT/SCRF method for different protonation state conformers (given in mm s^{-1}) and compared with experiment.[33] The error with respect to the experimental values is given as mean absolute error (MAE).

	Exp[33]	$RO^{-}P^{-}E^{H}$	$RO^{-}P^{H}E^{H}$	$RO^{H}P^{-}E^{-}$	$RO^{H}P^{H}E^{-}$	$RO^{H}P^{-}E^{H}$	$RO^{H}P^{H}E^{H}$
Fe1	1.22	-0.67	-0.70	0.94	0.99	1.02	1.11
Fe2	0.92	1.01	1.00	1.09	1.06	1.15	1.15
Fe3	1.33	1.11	1.11	1.13	1.15	1.24	1.25
Fe4	1.33	0.83	0.79	0.65	0.62	0.54	0.53
Ave	1.21	0.91	0.91	0.96	0.96	1.00	1.02
MAE	-	0.34	0.34	0.33	0.32	0.33	0.31

The average QS (QS_{ave}) of the RO^{-} states (0.91 mm s^{-1}) is slightly lower in magnitude than that computed for the ROH–bound states (0.96 – 1.02 mm s^{-1}). We note that the values of QS_{ave} obtained for all states considered are lower than experiment (1.21 mm s^{-1}).[33] In the case of the RO^{-} states, the absolute value of the QS for Fe1 is ~ 0.7 mm s^{-1}, which is significantly lower than the experimental value of 1.22 mm s^{-1}.[33] In contrast, there exists good agreement between the QS computed for Fe2 in the RO^{-} states (1.00 mm s^{-1}) and the experimental value (0.92 mm s^{-1}). The DFT/SCRF method, however, fails to capture the elevated QSs found for the delocalized, mixed–valence pair, Fe3/Fe4 (computed to be 0.8 to 1.1 mm s^{-1}; experimental value is 1.33 mm s^{-1}).[33]

All QS values computed for Fe1 in the ROH-bound states agree better with experiment than their RO^{-} counterparts (QS_{comp} = 0.94 – 1.11 mm s^{-1}; QS_{exp} = 1.22 mm s^{-1}), with the ROH/E^{-} states giving the best agreement with experiment (Table 2-7). Similar to the RO^{-} states, the QS values computed for Fe2 (QS = 1.06 – 1.15 mm s^{-1}) match experiment (0.92 mm s^{-1}) reasonably well, but those computed for Fe3/Fe4 are underestimated. In the case of Fe4, in particular, the QS computed for the ROH-

bound states (0.52 – 0.64 mm s^{-1}) are much lower than the experimental value (1.33 mm s^{-1}).[33] In sum, the results from computation of QSs are largely inconclusive, with both RO^- and ROH states (MAEs of 0.34 and 0.31 – 0.33 mm s^{-1}, respectively) giving good agreement with the experimentally determined QSs for Fe1 and Fe2, while underestimating the QSs for Fe3 and Fe4. The MAEs computed for the QSs, however, are only slightly larger than those obtained for synthetic Fe–S complexes of known structure (MAE ~ 0.25 mm s^{-1}).[65]

Discussion

By performing DFT calculations on a large active site model of the oxidized form of the [4Fe–4S] cluster in IspH, we are able to characterize geometries, relative energies and Mössbauer signatures of different protonation states. A thorough understanding of the interplay between [4Fe–4S]–HMBPP interactions and protein electrostatics in this state are of value both to understanding the unique catalytic mechanism of IspH and assessing factors that determine how new competitive inhibitors may best interact with the IspH active site.

The geometries of the [4Fe–4S] clusters computed for the various $RO^XP^YE^Z$ states do not fully match the asymmetry seen in the IspH:HMBPP crystal structure at 1.7 Å resolution,[32] although we do observe a four-fold compression of the cluster that is consistent with what is seen in other [4Fe–4S] protein systems.[101] In our calculations, a short Fe1–O bond length of 1.9 Å differentiates the RO^- states from their ROH counterparts; however, this distance is significantly shorter than what is observed in the IspH:HMBPP crystal structure (2.05 Å).[32] Instead, $RO^HP^YE^-$ states

(regardless of PP_i protonation) give better agreement with the crystal structure, displaying Fe1-OH distances of 2.11 – 2.13 Å. These $RO^{H}P^{Y}E^{-}$ states also better match the hydrogen bonding network implied by the crystal structure and, consequently, display higher structural similarity to the crystal structure, as measured by RMSD (Table 2-3).

An analysis of the relative energies of the different protonation states in our model provides different possibilities as to which states are energetically favorable in the oxidized state, depending on the solvation method employed (Figure 2-4). While the DFT/COSMO method indicates the $RO^{-}P^{H}E^{H}$ and $RO^{H}P^{H}E^{H}$ states (q = –2 and –1, respectively) are lowest in energy, the DFT/SCRF method favors the $RO^{-}P^{-}E^{H}$ and $RO^{H}P^{-}E^{-}$ states (q = –3). Further, the computed Mössbauer isomer shifts show preference for HMBPP to bind Fe1 as an alcohol group rather than an alkoxide. The ROH–bound states all display asymmetric isomer shift values, similar to what is seen experimentally.[30,33] In contrast, all states where an RO^{-} group coordinates Fe1 display symmetric isomer shifts. This finding indicates the dominant oxidized state involves HMBPP with its C_4–OH group protonated (ROH).

The finding that RO^{-}– and ROH–bound states give different Mössbauer signatures follows from the different ligand environments of the apical Fe atom. The "harder" RO^{-} form of HMBPP binds Fe1 more tightly than the ROH form, with concomitant lengthening of the Fe1–S^{2-} bonds. The net effect of these RO^{-} geometries is an isomer shift for Fe1 that appears similar to what is observed in related $[Fe_4S_4(SR)_4]^{2-}$ systems[65,102,103] and is more indicative of a valence delocalized $Fe^{2.5+}$

oxidation state. In contrast, when bound as an alcohol, the HMBPP–Fe1 distance is elongated with shortened Fe1–S^{2-} distances. This results in a net effect of giving Fe1 more ferrous character, as evidenced by its greater isomer shift (Table 2-6).

Overall, the quadrupole splittings (QSs) obtained for active site clusters containing ROH–bound HMBPP are slightly more accurate than those computed for RO^{-}–bound states. In both RO^{-}– and ROH–bound states, the QSs computed for Fe1, Fe2 and Fe3 better match experiment than for those computed for Fe4. It is worth noting that the active site cluster utilized in all calculations is constructed from the [4Fe–4S] cluster on the face containing Fe1 and Fe2, in order to properly stabilize the highly charged PP_i moiety on HMBPP. This approach largely omits the protein environment around Fe3 and Fe4, except the backbones amides of A199, T200, P97, and L98, as well as the side chain of T200, which cumulatively donate three hydrogen bonds to the S^{2-} atoms and thiolates bound to Fe3 and Fe4. Despite including nearby point charges in the DFT/SCRF scheme, it is possible that the electric field gradients at Fe3 and Fe4 suffer from excluding nearby protein residues from the quantum region in our computations.

We note that a previous study has also computed Mössbauer isomer shifts and quadrupole splittings for a [4Fe–4S]:HMBPP complex. Ahrens-Botzong *et al.* computed Mössbauer ISs and QSs for comparison with Mössbauer experiments they performed on complexes of [4Fe-4S] IspH with HMBPP and its amino and thiol analogues.[33] Their work utilized a DFT/MM approach, employing the B3LYP/LANL2DZ level of theory to describe the $[4Fe–4S]^{2+}$ cluster, the SCH_2

groups of its coordinating cysteines and a singly protonated (ROH) HMBPP ligand in the DFT region. Following the DFT/MM minimization of the IspH:HMBPP complex, the authors compute Mössbauer parameters at the B3LYP/CP(PPP)/TZVP level of theory (where the CP[PPP] basis set is used for Fe) using the ORCA program with the closed shell approach. The resulting computed isomer shifts, which assumed ROH coordination to the [4Fe–4S] cluster, are found to be 0.48, 0.56, 0.30 and 0.81 mm s^{-1}, with corresponding QSs of –1.18, 1.30, –2.02 and –0.75 mm s^{-1}.[33]

Our work differs from this previous work in two principal ways: (1) we utilize a large active site model to compute Mössbauer isomer shifts using BS–DFT/COSMO and BS–DFT/SCRF approaches; and (2) we do not make assumptions regarding the protonation state of HMBPP when coordinated to the [4Fe–4S] cluster. In regard to the computational approach employed, it is evident that BS–DFT approaches (both DFT/COSMO and DFT/SCRF) better predict IS (MAE with respect to experiment of 0.03 mm s^{-1}, compared to 0.11 mm s^{-1}) and QS values (MAE with respect to experiment of 0.33 mm s^{-1}, compared to 0.44 mm s^{-1}). Regardless, after our examination of different HMBPP protonation states, we similarly conclude that the ROH-bound geometries of HMBPP give better agreement with experiment, consistent with the assumption and conclusions of Ahrens–Botzong *et al*.[33]

Both DFT/COSMO and DFT/SCRF results show ROH–bound states, whose computed Mossbauer properties are consistent with experiment, to be iso–energetic with the lowest–energy oxidized states. While DFT/COSMO results show preference for protonation of E126 and the HMBPP PPi moiety ($RO^{H}P^{H}E^{H}$, q = –1), the $RO^{H}P^{-}E^{-}$

state is preferred with the DFT/SCRF method. Since computed geometries with a protonated E126 show less similarity to the IspH:HMBPP crystal structure, we propose the preferred oxidized state for IspH to be $RO^{H}P^{-}E^{-}$. The ability of the DFT/SCRF method to select the most relevant state, as verified through comparisons with experimental results, indicates that the inclusion of both the protein and reaction field effects is critical to computing accurate energies along the IspH catalytic pathway.

Of additional interest, we find that when the IspH active site is described with the DFT/SCRF method, the different charge (protonation) states considered cluster together energetically. With q = –3, the $RO^{-}P^{-}E^{H}$ and $RO^{H}P^{-}E^{-}$ states are lowest in energy; however, states with q = –2 are both energetically accessible from these lowest-energy states and also possess an additional proton that can play a role in catalysis. For instance, the $RO^{-}P^{H}E^{H}$ and $RO^{H}P^{-}E^{H}$ states (q = –2) are 2.2 and 5.5 kcal mol^{-1} above the energies of the $RO^{-}P^{-}E^{H}$ and $RO^{H}P^{-}E^{-}$ states (Table 2-4, Figure 2-4C). Since the active site stabilizes q = –3 charge states, protonation to generate these q = –2 states may precede subsequent electron addition to reinforce a q = –3 active site, suggesting a mechanism by which protons and electrons can be shuttled into the active site for catalysis. It is worth noting the stabilization of a highly–anionic active site in DFT/SCRF computations can be attributed to their inclusion of the highly–polar IspH protein environment. Indeed, the active site quantum cluster is within 8 Å of the of seven Arg/Lys (cationic) residues, 10 Asp/Glu (anionic) residues and 95

crystallographic waters, which are all included as point charges in the protein region within the DFT/SCRF scheme.

In this work, we have demonstrated the ability of the DFT/SCRF method to describe protonation state changes in the IspH oxidized state, which suggests that similar efforts can be applied to reasonably characterize the later stages of IspH catalysis, as well. Drawing from the results presented here, we have used the $RO^{H}P^{-}E^{-}$ state computed for oxidized IspH as a starting point to examine the proposed IspH catalytic mechanism. Initial computations performed on the $1e^{-}$ reduced state suggest that reduction is not easily achieved when HMBPP coordinates Fe1 through its C_4–OH group, as computed reduction potentials are ~ 0.3 V more negative than the typical values associated with ferredoxin [4Fe–4S] proteins. Additional work has been performed to assess alternative routes for reduction.

Further evaluation of the energetics along the IspH catalytic pathway will both increase the knowledge of how these [4Fe–4S] clusters facilitate catalysis, as well as reveal differences between the oxidized and reduced forms of IspH that may be exploited in the development of novel antibacterial and antimalarial drugs.

Acknowledgements

P.G.B. acknowledges support from the National Institute of Health (NIH) Molecular Biophysics Training Grant (2T32GM008326-21) and the ARCS Foundation. D.A.G. and D.B. acknowledge support from the NIH (GM086243). J.A.M. acknowledges support from the NIH (GM31749), National Science Foundation (NSF, MCB-1020765), Howard Hughes Medical Institute, National Biomedical

Computation Resource and NSF supercomputer centers. L.N. acknowledges support from the NIH (GM100934). The authors also thank Dr. Wenge Han for many helpful discussions.

Chapter 2 is a minimally modified reprint of the material as it appears in Patrick G. Blachly, Gregory M. Sandala, D. Ann Giammona, Tiqing Liu, Donald Bashford, J. Andrew McCammon, and Louis Noodleman, "Use of broken-symmetry density functional theory to characterize the IspH oxidized state: Implications for IspH mechanism and inhibition," submitted to the *Journal of Chemical Theory and Computation*, 2014. The dissertation author was the primary investigator and author of this paper.

Supporting Information

Table 2-S1. Relative energies (kcal mol^{-1}) of the four valence isomers considered in this study, computed using DFT/COSMO and DFT/SCRF methods. The net spins on the different Fe atoms are given, relative to the Fe atom numbering in Figure 2B (e.g. βααβ indicates that Fe2 and Fe3 are each net spin α, while Fe1 and Fe4 are each net spin β). The energies of the lowest-energy states are set to zero and are indicated with bold–type font. The energy values of the lowest–energy valence isomers correspond to the relative energies given in Table 4A and 4C ($\Delta E_{Tot,COSMO,pH=7}$ and $\Delta E_{Tot,SCRF,pH=7}$).

	Relative energies (kcal mol^{-1})	
	DFT/COSMO	**DFT/SCRF**
$RO^{-}P^{-}E^{H}$		
βαβα	11.8	5.7
βααβ	7.9	**0.0**
αββα	7.9	**0.0**
ααββ	12.6	5.2
$RO^{-}P^{H}E^{H}$		
βαβα	3.8	7.1
βααβ	0.3	2.2
αββα	**0.0**	2.4
ααββ	3.9	8.1
$RO^{H}P^{-}E^{-}$		
βαβα	15.4	7.7
βααβ	8.4	**0.0**
αββα	10.5	0.6
ααββ	13.3	4.6
$RO^{H}P^{H}E^{-}$		
βαβα	10.4	11.7
βααβ	5.3	6.1
αββα	9.5	10.2
ααββ	12.2	13.4
$RO^{H}P^{-}E^{H}$		
βαβα	6.3	9.2
βααβ	2.9	5.5
αββα	2.9	5.5
ααββ	8.0	10.3
$RO^{H}P^{H}E^{H}$		
βαβα	2.1	17.1
βααβ	**0.0**	13.3
αββα	**0.0**	13.3
ααββ	4.9	18.8

Table 2-S2. Tabulation of isomer shifts (given in mm s^{-1}) for different protonation state conformers computed using the DFT/COSMO method and compared with experiment.[33] The error with respect to the experimental values is given as mean absolute error (MAE).

	Exp	$RO^{-}P^{-}E^{H}$	$RO^{-}P^{H}E^{H}$	$RO^{H}P^{-}E^{-}$	$RO^{H}P^{H}E^{-}$	$RO^{H}P^{-}E^{H}$	$RO^{H}P^{H}E^{H}$
Fe1	0.64	0.44	0.44	0.54	0.53	0.53	0.54
Fe2	0.38	0.42	0.42	0.39	0.38	0.38	0.38
Fe3	0.42	0.43	0.42	0.41	0.40	0.40	0.40
Fe4	0.42	0.43	0.42	0.41	0.40	0.40	0.39
AVG	0.47	0.43	0.43	0.44	0.43	0.43	0.43
MAE	-	0.07	0.06	0.03	0.04	0.04	0.04

Table 2-S3. Tabulation of quadrupole splittings (given in mm s^{-1}) for different protonation state conformers computed using the DFT/COSMO method and compared with experiment.[33] The error with respect to the experimental values is given as mean absolute error (MAE).

	Exp	$RO^{-}P^{-}E^{H}$	$RO^{-}P^{H}E^{H}$	$RO^{H}P^{-}G^{-}$	$RO^{H}P^{H}E^{-}$	$RO^{H}P^{-}E^{H}$	$RO^{H}P^{H}E^{H}$
Fe1	1.22	-0.64	-0.66	0.93	0.97	1.04	1.12
Fe2	0.92	1.03	1.03	1.11	1.08	1.18	1.17
Fe3	1.33	1.10	1.10	1.16	1.17	1.26	1.26
Fe4	1.33	0.88	0.83	0.65	0.61	0.53	0.52
AVG	1.21	0.91	0.91	0.96	0.96	1.00	1.02
MAE	-	0.34	0.35	0.33	0.32	0.33	0.31

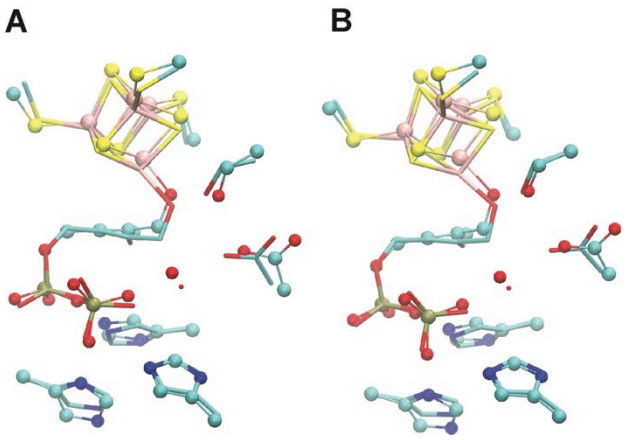

Figure 2-S1. Superposition of the geometry optimized (A) $RO^{-}P^{-}E^{H}$ and (B) $RO^{H}P^{-}E^{-}$ states (each represented by the ball and stick format) onto the [4Fe–4S] IspH:HMBPP crystal structure (licorice).[32]

References

(1) World Health Organization. *World Malaria Report 2013*. WHO Press: Geneva, Switzerland, 2013.

(2) World Health Organization. *World Tuberculosis Report 2013*. WHO Press: Geneva, Switzerland, 2013.

(3) Alonso, P. L.; Tanner, M. *Nat Med* **2013**, *19*, 150.

(4) Boucher, H. W.; Talbot, G. H.; Bradley, J. S.; Edwards, J. E.; Gilbert, D.; Rice, L. B.; Scheld, M.; Spellberg, B.; Bartlett, J. *Clin Infect Dis* **2009**, *48*, 1.

(5) Levy, S. B.; Marshall, B. *Nat Med* **2004**, *10*, S122.

(6) Payne, D. J.; Gwynn, M. N.; Holmes, D. J.; Pompliano, D. L. *Nat Rev Drug Discov* **2007**, *6*, 29.

(7) Ridley, R. G. *Nature* **2002**, *415*, 686.

(8) Wells, T. N. C.; Alonso, P. L.; Gutteridge, W. E. *Nat Rev Drug Discov* **2009**, *8*, 879.

(9) World Health Organization. *The evolving threat of antimicrobial resistance: options for action*. WHO Press: Geneva, Switzerland, 2012.

(10) Oldfield, E. *Accounts Chem Res* **2010**, *43*, 1216.

(11) Christianson, D. W. *Science* **2007**, *316*, 60.

(12) Christianson, D. W. *Curr Opin Chem Biol* **2008**, *12*, 141.

(13) Bloch, K. *Steroids* **1992**, *57*, 378.

(14) Eisenreich, W.; Bacher, A.; Arigoni, D.; Rohdich, F. *Cell Mol Life Sci* **2004**, *61*, 1401.

(15) Kuzuyama, T.; Seto, H. *Nat Prod Rep* **2003**, *20*, 171.

(16) Rohmer, M. *Nat Prod Rep* **1999**, *16*, 565.

(17) Rohmer, M.; Knani, M.; Simonin, P.; Sutter, B.; Sahm, H. *Biochem J* **1993**, *295*, 517.

(18) Jomaa, H.; Wiesner, J.; Sanderbrand, S.; Altincicek, B.; Weidemeyer, C.; Hintz, M.; Turbachova, I.; Eberl, M.; Zeidler, J.; Lichtenthaler, H. K.; Soldati, D.; Beck, E. *Science* **1999**, *285*, 1573.

(19) Missinou, M. A.; Borrmann, S.; Schindler, A.; Issifou, S.; Adegnika, A. A.; Matsiegui, P. B.; Binder, R.; Lell, B.; Wiesner, J.; Baranek, T.; Jomaa, H.; Kremsner, P. G. *Lancet* **2002**, *360*, 1941.

(20) Borrmann, S.; Issifou, S.; Esser, G.; Adegnika, A. A.; Ramharter, M.; Matsiegui, P. B.; Oyakhirome, S.; Mawili-Mboumba, D. P.; Missinou, M. A.; Kun, J. F. J.; Jomaa, H.; Kremsner, P. G. *J Infect Dis* **2004**, *190*, 1534.

(21) Rohdich, F.; Hecht, S.; Gartner, K.; Adam, P.; Krieger, C.; Amslinger, S.; Arigoni, D.; Bacher, A.; Eisenreich, W. *P Natl Acad Sci USA* **2002**, *99*, 1158.

(22) Rohdich, F.; Zepeck, F.; Adam, P.; Hecht, S.; Kaiser, J.; Laupitz, R.; Grawert, T.; Amslinger, S.; Eisenreich, W.; Bacher, A.; Arigoni, D. *P Natl Acad Sci USA* **2003**, *100*, 1586.

(23) Altincicek, B.; Duin, E. C.; Reichenberg, A.; Hedderich, R.; Kollas, A. K.; Hintz, M.; Wagner, S.; Wiesner, J.; Beck, E.; Jomaa, H. *Febs Lett* **2002**, *532*, 437.

(24) Grawert, T.; Kaiser, J.; Zepeck, F.; Laupitz, R.; Hecht, S.; Amslinger, S.; Schramek, N.; Schleicher, E.; Weber, S.; Haslbeck, M.; Buchner, J.; Rieder, C.; Arigoni, D.; Bacher, A.; Eisenreich, W.; Rohdich, F. *J Am Chem Soc* **2004**, *126*, 12847.

(25) Wolff, M.; Seemann, M.; Bui, B. T. S.; Frapart, Y.; Tritsch, D.; Estrabot, A. G.; Rodriguez-Concepcion, M.; Boronat, A.; Marquet, A.; Rohmer, M. *Febs Lett* **2003**, *541*, 115.

(26) Xiao, Y. L.; Zhao, Z. K.; Liu, P. H. *J Am Chem Soc* **2008**, *130*, 2164.

(27) Rekittke, I.; Wiesner, J.; Rohrich, R.; Demmer, U.; Warkentin, E.; Xu, W. Y.; Troschke, K.; Hintz, M.; No, J. H.; Duin, E. C.; Oldfield, E.; Jomaa, H.; Ermler, U. *J Am Chem Soc* **2008**, *130*, 17206.

(28) Grawert, T.; Rohdich, F.; Span, I.; Bacher, A.; Eisenreich, W.; Eppinger, J.; Groll, M. *Angew Chem Int Edit* **2009**, *48*, 5756.

(29) Wang, W. X.; Wang, K.; Liu, Y. L.; No, J. H.; Li, J. K.; Nilges, M. J.; Oldfield, E. *P Natl Acad Sci USA* **2010**, *107*, 4522.

(30) Seemann, M.; Janthawornpong, K.; Schweizer, J.; Bottger, L. H.; Janoschka, A.; Ahrens-Botzong, A.; Tambou, M. N.; Rotthaus, O.; Trautwein, A. X.; Rohmer, M.; Schunemann, V. *J Am Chem Soc* **2009**, *131*, 13184.

(31) Xiao, Y. L.; Chu, L.; Sanakis, Y.; Liu, P. H. *J Am Chem Soc* **2009**, *131*, 9931.

(32) Grawert, T.; Span, I.; Eisenreich, W.; Rohdich, F.; Eppinger, J.; Bacher, A.; Groll, M. *P Natl Acad Sci USA* **2010**, *107*, 1077.

(33) Ahrens-Botzong, A.; Janthawornpong, K.; Wolny, J. A.; Tambou, E. N.; Rohmer, M.; Krasutsky, S.; Poulter, C. D.; Schunemann, V.; Seemann, M. *Angew Chem Int Edit* **2011**, *50*, 11976.

(34) Wang, K.; Wang, W. X.; No, J. H.; Zhang, Y. H.; Zhang, Y.; Oldfield, E. *J Am Chem Soc* **2010**, *132*, 6719.

(35) Wang, W. X.; Wang, K.; Span, I.; Jauch, J.; Bacher, A.; Groll, M.; Oldfield, E. *J Am Chem Soc* **2012**, *134*, 11225.

(36) Xu, W. Y.; Lees, N. S.; Hall, D.; Welideniya, D.; Hoffman, B. M.; Duin, E. C. *Biochemistry-US* **2012**, *51*, 4835.

(37) Li, J. K.; Wang, K.; Smirnova, T. I.; Khade, R. L.; Zhang, Y.; Oldfield, E. *Angew Chem Int Edit* **2013**, *52*, 6522.

(38) Chang, W. C.; Xiao, Y. L.; Liu, H. W.; Liu, P. H. *Angew Chem Int Edit* **2011**, *50*, 12304.

(39) Xiao, Y. L.; Chang, W. C.; Liu, H. W.; Liu, P. H. *Org Lett* **2011**, *13*, 5912.

(40) Xiao, Y. L.; Liu, P. H. *Angew Chem Int Edit* **2008**, *47*, 9722.

(41) Citron, C. A.; Brock, N. L.; Rabe, P.; Dickschat, J. S. *Angew Chem Int Edit* **2012**, *51*, 4053.

(42) Grawert, T.; Span, I.; Bacher, A.; Groll, M. *Angew Chem Int Edit* **2010**, *49*, 8802.

(43) Gloux, J.; Gloux, P.; Lamotte, B.; Mouesca, J. M.; Rius, G. *J Am Chem Soc* **1994**, *116*, 1953.

(44) Middleton, P.; Dickson, D. P. E.; Johnson, C. E.; Rush, J. D. *Eur J Biochem* **1980**, *104*, 289.

(45) Noodleman, L. *Inorg Chem* **1988**, *27*, 3677.

(46) Chang, W. C.; Song, H.; Liu, H. W.; Lu, P. H. *Curr Opin Chem Biol* **2013**, *17*, 571.

(47) Zhao, L. S.; Chang, W. C.; Xiao, Y. L.; Liu, H. W.; Liu, P. H. *Annu Rev Biochem* **2013**, *82*, 497.

(48) Yang, X.; Niu, S. Q.; Ichiye, T.; Wang, L. S. *J Am Chem Soc* **2004**, *126*, 15790.

(49) Backes, G.; Mino, Y.; Loehr, T. M.; Meyer, T. E.; Cusanovich, M. A.; Sweeney, W. V.; Adman, E. T.; Sandersloehr, J. *J Am Chem Soc* **1991**, *113*, 2055.

(50) Glaser, T.; Hedman, B.; Hodgson, K. O.; Solomon, E. I. *Accounts Chem Res* **2000**, *33*, 859.

(51) Torres, R. A.; Lovell, T.; Noodleman, L.; Case, D. A. *J Am Chem Soc* **2003**, *125*, 1923.

(52) Han, W.-G.; Noodleman, L. *Theor Chem Acc* **2009**, *125*, 305.

(53) Sumner, S.; Söderhjelm, P.; Ryde, U. *J Chem Theory Comput* **2013**, *9*, 4205.

(54) Maestro, In *Schrodinger Suite 2012*; version 9.3; Schrödinger, LLC: New York, NY, 2012.

(55) Bhave, D. P.; Han, W. G.; Pazicni, S.; Penner-Hahn, J. E.; Carroll, K. S.; Noodleman, L. *Inorg Chem* **2011**, *50*, 6610.

(56) Noodleman, L. *J Chem Phys* **1981**, *74*, 5737.

(57) Noodleman, L.; Peng, C. Y.; Case, D. A.; Mouesca, J. M. *Coordin Chem Rev* **1995**, *144*, 199.

(58) Sandala, G. M.; Noodleman, L. *Methods Mol Biol* **2011**, *766*, 293.

(59) ADF2009; SCM, Theoretical Chemistry, Vrije Universiteit: Amsterdam, The Netherlands, 2009.

(60) Handy, N. C.; Cohen, A. J. *Mol Phys* **2001**, *99*, 403.

(61) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys Rev B* **1988**, *37*, 785.

(62) Conradie, J.; Ghosh, A. *J Chem Theory Comput* **2007**, *3*, 689.

(63) Han, W. G.; Noodleman, L. *Inorg Chim Acta* **2008**, *361*, 973.

(64) Swart, M. *J Chem Theory Comput* **2008**, *4*, 2057.

(65) Sandala, G. M.; Hopmann, K. H.; Ghosh, A.; Noodleman, L. *J Chem Theory Comput* **2011**, *7*, 3232.

(66) Klamt, A. *J Phys Chem-US* **1995**, *99*, 2224.

(67) Klamt, A.; Schuurmann, G. *J Chem Soc Perk T 2* **1993**, 799.

(68) Protein Preparation Wizard, In *Schrodinger Suite 2012*; Schrödinger, LLC: New York, NY, 2012.

(69) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. *J Comput Aided Mol Des* **2013**, *27*, 221.

(70) Bas, D. C.; Rogers, D. M.; Jensen, J. H. *Proteins: Structure, Function, and Bioinformatics* **2008**, *73*, 765.

(71) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704.

(72) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. *J Chem Theory Comput* **2011**, *7*, 525.

(73) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. *J Chem Theory Comput* **2011**, *7*, 2284.

(74) Impact, In *Schrodinger Suite 2012*; version 5.8; Schrödinger, LLC: New York, NY, 2012.

(75) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. *Nucleic Acids Res* **2007**, *35*, W522.

(76) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. *Nucleic Acids Res* **2004**, *32*, W665.

(77) Sitkoff, D.; Sharp, K. A.; Honig, B. *J Phys Chem-US* **1994**, *98*, 1978.

(78) Breneman, C. M.; Wiberg, K. B. *J Comput Chem* **1990**, *11*, 361.

(79) Mouesca, J. M.; Chen, J. L.; Noodleman, L.; Bashford, D.; Case, D. A. *J Am Chem Soc* **1994**, *116*, 11898.

(80) ADF2012; SCM, Theoretical Chemistry, Vrije Universiteit: Amsterdam, The Netherlands, 2012.

(81) Bashford, D. In *Scientific Computing in Object-Oriented Parallel Environments*; Ishikawa, Y., Oldehoeft, R. R., Reynders, J. V. W., Tholburn, M., Eds.; Springer: Berlin, Germany, 1997; Vol. 1343, p 233.

(82) Bashford, D.; Gerwert, K. *J Mol Biol* **1992**, *224*, 473.

(83) Li, J.; Nelson, M. R.; Peng, C. Y.; Bashford, D.; Noodleman, L. *J Phys Chem A* **1998**, *102*, 6311.

(84) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Guerra, C. F.; Van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J Comput Chem* **2001**, *22*, 931.

(85) Fee, J. A.; Case, D. A.; Noodleman, L. *J Am Chem Soc* **2008**, *130*, 15002.

(86) Noodleman, L.; Han Du, W.-G.; Fee, J. A.; Götz, A. W.; Walker, R. C. *Inorg Chem* **2014**, *In press.*

(87) Lewis, A.; Bumpus, J. A.; Truhlar, D. G.; Cramer, C. J. *J Chem Educ* **2004**, *81*, 596.

(88) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R. *J Phys Chem A* **1998**, *102*, 7787.

(89) Tawa, G. J.; Topol, I. A.; Burt, S. K.; Caldwell, R. A.; Rashin, A. A. *J Chem Phys* **1998**, *109*, 4852.

(90) Hopmann, K. H.; Ghosh, A.; Noodleman, L. *Inorg Chem* **2009**, *48*, 9155.

(91) Lovell, T.; Han, W. G.; Liu, T. Q.; Noodleman, L. *J Am Chem Soc* **2002**, *124*, 5890.

(92) Lovell, T.; Li, J.; Liu, T. Q.; Case, D. A.; Noodleman, L. *J Am Chem Soc* **2001**, *123*, 12392.

(93) Neese, F. *Inorg Chim Acta* **2002**, *337*, 181.

(94) Zhang, Y.; Mao, J. H.; Oldfield, E. *J Am Chem Soc* **2002**, *124*, 7829.

(95) Liu, T.; Noodleman, L.; Case, D. A.; The Scripps Research Institute: La Jolla, CA.

(96) Bashford, D.; Case, D. A.; Dalvit, C.; Tennant, L.; Wright, P. E. *Biochemistry-US* **1993**, *32*, 8045.

(97) Kyte, J. *Structure in protein chemistry*; Garland Pub.: New York, 1995.

(98) McElroy, W. D.; Glass, B. *Phosphorus Metabolism* Baltimore, MD, 1951; Vol. 1.

(99) Li, J.; Fisher, C. L.; Chen, J. L.; Bashford, D.; Noodleman, L. *Inorg Chem* **1996**, *35*, 4694.

(100) Seemann, M. Université de Strasbourg/CNRS, Strasbourg, France, Personal Communication, 2013.

(101) Fukuyama, K.; Okada, T.; Kakuta, Y.; Takahashi, Y. *J Mol Biol* **2002**, *315*, 1155.

(102) Kanatzidis, M. G.; Baenziger, N. C.; Coucouvanis, D.; Simopoulos, A.; Kostikas, A. *J Am Chem Soc* **1984**, *106*, 4500.

(103) Silver, J.; Fern, G. R.; Miller, J. R.; McCammon, C. A.; Evans, D. J.; Leigh, G. J. *Inorg Chem* **1999**, *38*, 4256.

Chapter 3:

Utilizing a dynamical description of IspH to aid in the development of novel antimicrobial drugs

Abstract

The nonmevalonate pathway is responsible for isoprenoid production in microbes, including *H. pylori*, *M. tuberculosis* and *P. falciparum*, but is nonexistent in humans, thus providing a desirable route for antibacterial and antimalarial drug discovery. We coordinate a structural study of IspH, a [4Fe-4S] protein responsible for converting HMBPP to IPP and DMAPP in the ultimate step in the nonmevalonate pathway. By performing accelerated molecular dynamics simulations on both substrate-free and HMBPP-bound $[Fe_4S_4]^{2+}$ IspH, we elucidate how substrate binding alters the dynamics of the protein. Using principal component analysis, we note that while substrate-free IspH samples various open and closed conformations, the closed conformation observed experimentally for HMBPP-bound IspH is inaccessible in the absence of HMBPP. In contrast, simulations with HMBPP bound are restricted from accessing the open states sampled by the substrate-free simulations. Further investigation of the substrate-free simulations reveals large fluctuations in the HMBPP binding pocket, as well as allosteric pocket openings – both of which are achieved

through the hinge motions of the individual domains in IspH. Coupling these findings with solvent mapping and various structural analyses reveals alternative druggable sites that may be exploited in future drug design efforts.

Introduction

In the past couple decades, antimicrobial drug resistance has risen dramatically and greatly hampered the efficacy of currently available therapies for bacterial and malarial infections.[1-9] Whereas (multiple-)drug-resistant bacterial infections are a ubiquitous problem, affecting both the Western world and developing nations, the burdens of malaria fall disproportionately on the poorest regions of the world, with over 219 millions cases and 666,000 deaths reported in 2010.[3] Beyond the common problems associated with decreased lifetimes for drug efficacy due to rapid development of resistance,[1,2,5,6,9] advances in the fight against bacterial and malarial infections have also been plagued by diminished attention from major pharmaceutical companies toward the development of new therapies and drugs.[4,5,10] Consequently, there is urgent need for the development of new drugs with novel modes of action, for administration either independently or in combination with established regimen, both to combat bacterial and malarial infections, as well as to address the propensity of each for rapidly developing drug resistance.[1,4,6,7,9]

The nonmevalonate (methylerythritol) pathway for isoprenoid biosynthesis has recently been revealed as a novel target for both antibacterial and antimalarial drugs. Isoprenoids comprise essential metabolites derived from the 5-carbon biomolecules, isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP, Figure 3-1),

examples of which include sterols that provide structural support to membranes, chlorophylls used in photosynthesis, and quinones that participate in electron transport chains.[11-14] In contrast, animals acquire IPP and DMAPP in a distinctive manner via a mevalonate-dependent pathway. Given this metabolic difference, the proteins involved in the nonmevalonate pathway provide novel targets for the development of antibacterial and antimalarial drugs that are both broadly specific to pathogenic species such as *H. pylori*, *M. tuberculosis* and *P. falciparum* and without known human analogs.[15-18]

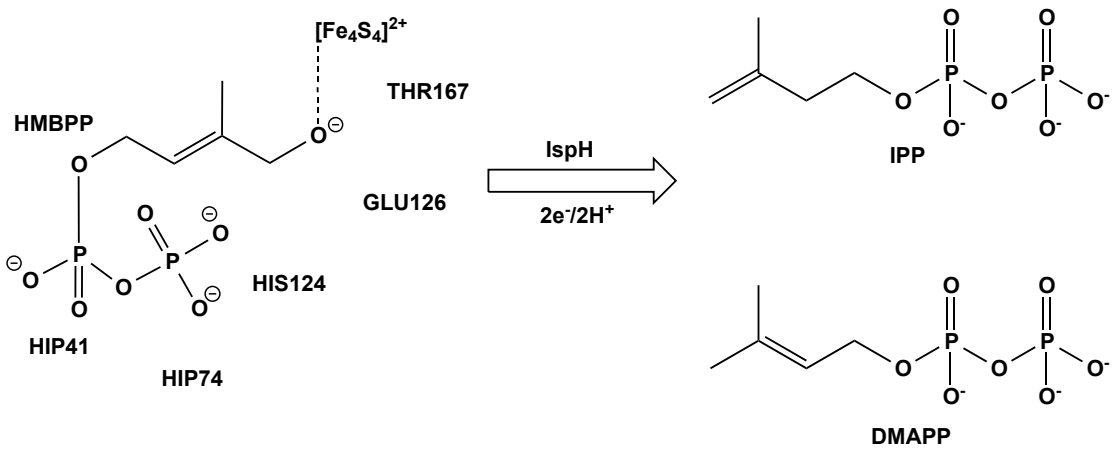

Figure 3-1. Reductive dehydroxylation of HMBPP affords the isoprenoid precursors, IPP and DMAPP. The relative positions of active site residues suggested to play a role in either substrate binding or catalysis are labeled.

The ultimate step in the nonmevalonate pathway is the generation of IPP and DMAPP through a 2-electron reductive dehydroxylation of (E)-1-hydroxy-2-methyl-but-2-enyl pyrophosphate (HMBPP) by IspH, a [4Fe-4S] protein (Figure 3-1).[11,19-21] The catalytic mechanism of IspH has been a topic of great debate, largely due to uncertainties introduced by the iron-sulfur cluster.[18,22] Initial structures of IspH from

Aquifex aeolicus[23] and *Escherichia coli*[24] solved by X-ray crystallography resemble cloverleaves and comprise three sequentially different domains with pseudo-C3 symmetry, each tethered to a $[Fe_3S_4]^+$ cluster via a conserved cysteine residue. The *A. aeolicus* $[Fe_3S_4]^+$ IspH structure (PDB ID: 3DNF; henceforth referred to as $[Fe_3S_4]^+_{(open,\ substrate\text{-}free)}$ IspH) assumes an open conformation, with a 10 x 20 Å cavity where the HMBPP molecule is expected to bind at the cluster.[23] In contrast to the *A. aeolicus* crystal structure, the $[Fe_3S_4]^+$ *E. coli* counterpart is closed around an inorganic diphosphate molecule (PP_i) that sits in the vicinity of the centrally located $[Fe_3S_4]^+$ cluster. Various conserved polar and charged residues, including Glu-126, Thr-167, Asn-227, His-41, His-74, His-124, Ser-225, Ser-226 and Ser-269 (*E. coli* numbering scheme), coordinate the PP_i molecule, likely via hydrogen bonding or salt bridge interactions.[24] The orientations of these conserved residues in the *E. coli* structure are distinct from their *A. aeolicus* counterparts due to a tilt of a single domain that enables co-localization of charged and polar residues around the PP_i in the case of the former.

While results from electron paramagnetic resonance (EPR) spectroscopy have shown $[Fe_3S_4]^+$ IspH to be catalytically active,[25] reconstituted IspH displays EPR and Mossbauer signatures of a $[Fe_4S_4]^{2+}$ cluster.[26,27] Groll and co-workers provide further support for the catalytically relevant form of IspH containing a $[Fe_4S_4]^{2+}$ cluster with their work in crystallizing IspH in the presence of its substrate, HMBPP. This HMBPP-bound crystal structure (PDB ID: 3KE8, henceforth referred to as $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ IspH) assumes a closed conformation having a domain tilt

similar to that of the $[Fe_3S_4]^+$ *E. coli* structure, with HMBPP bound via its terminal hydroxyl moiety to an unliganded iron of a $[Fe_4S_4]^{2+}$ cluster (Figure 3-2).[28] The coordination sphere of the HMBPP ligand is virtually identical to the inorganic diphosphate molecule, while its terminal hydroxyl moiety interacts with Glu-126, Thr-167 (*E. coli* numbering) and an ordered water molecule to make a hydrogen bond network that is proposed to facilitate proton transfer during catalysis.[28] While these structural data provide a good picture of the $[Fe_4S_4]^{2+}$ IspH structure with HMBPP bound, the structure of the 4Fe-form in the absence of substrate, as well as a detailed understanding of how IspH changes conformation upon ligand binding, are not fully understood.

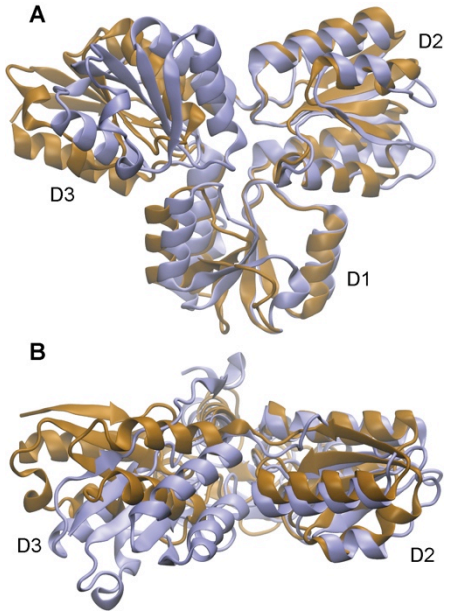

Figure 3-2. Superposition of $[Fe_3S_4]^+_{(open,\ substrate\text{-}free)}$ (bronze)[23] and $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ (purple)[28] IspH crystal structures, viewed (A) head-on toward the binding site and (B) from a top-view highlighting the domain tilt of D3.

Drawing from insight gained from the aforementioned structural work, as well as various spectroscopic and mutational studies, multiple groups have contributed to

drug discovery efforts on the IspH target.[29-34] To the best of our knowledge, IspH inhibitor development has fallen under two classes: (1) HMBPP analogues[29-31] and (2) pyridine or alkenyl/alkynyl diphosphates and bisphosphonates.[32-34] In the case of HMBPP analogues, inhibitor binding emulates the natural substrate, while leveraging improved interactions with the Fe-site (*e.g.* binding of a thiol instead of an alcohol).[30,31] Alternatively, Oldfield and co-workers have created novel inhibitors of IspH by utilizing olefinic and pyridine groups to form π/σ "metallacycle" complexes and η^1-complexes, respectively, coupling these metal binding groups to phosphate skeletons that preserve the hydrogen bond and salt bridge interactions present in IspH-HMBPP complexation.[32-34] These initial drug discovery efforts may be enhanced, both in terms of finding new lead compounds and developing already discovered leads, by obtaining a better description of the IspH binding pocket and possible allosteric sites that may be targeted.

Given that there exists no high-resolution structural data for substrate-free, $[Fe_4S_4]^{2+}$ IspH, this work employs accelerated molecular dynamics (aMD) simulations to describe the dominant conformations available to IspH having a fourth iron atom in the absence of HMBPP. Characterization of these dominant conformations reveals an expanded binding pocket and allosteric sites that may be targeted with future rational drug design efforts. Additional attention is directed toward understanding how IspH dynamics are altered upon ligand binding, allowing us to propose a mechanism for how IspH-HMBPP complexation is achieved.

Results

aMD simulations of open, substrate-free IspH

Consistent with the nomenclature used by Gräwert, *et al*.,[28] descriptions of IspH from this point forward will use the nomenclature D1, D2 and D3 to describe the domains containing residues 14-96; 97-193; and 194-281, 1-13, respectively (*A. aeolicus* numbering, Figure 3-2). We perform 3 x 100 ns aMD simulations of $[Fe_4S_4]^{2+}_{(open,\ substrate\text{-}free)}$ IspH, starting from the *A. aeolicus* crystal structure with a fourth iron modeled into the cluster, as described in the Methods. All trajectories are aligned to the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ IspH crystal structure by the backbone atoms of all D1 residues, since these residues display significantly lower fluctuation throughout the simulation than those in D2 and D3.[23] The root-mean-square deviation (RMSD) for the backbone atoms of all residues after alignment is given in Figure 3-3A. From this RMSD analysis, it is apparent that each independent trajectory samples conformational space differently. The large changes in RMSD correspond to opening and closing motions of the D2 and D3 domains, providing a more dynamic description of the $[Fe_4S_4]^{2+}_{(open,\ substrate\text{-}free)}$ state than is acquired from a static X-ray structure. While all three simulations extensively sample conformational space near the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ IspH crystal structure for the first ~20 ns of the simulation, one simulation diverges from this experimental reference, implying that other distinctive, low-energy conformational states exist for substrate-free IspH.

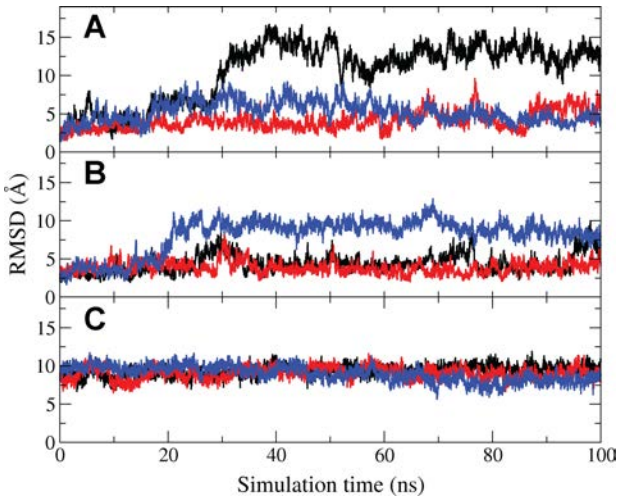

Figure 3-3. Plots of RMSD relative to the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ crystal structure over the course of 3 x 100 ns aMD simulations of (A) $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$, (B) $[Fe_4S_4]^{2+}$/$HMBPP_{(open,docked)}$, and (C) $[Fe_4S_4]^{2+}$/$HMBPP_{(closed)}$ IspH.

Docking of HMBPP to open IspH

Using Schrodinger's Glide program,[35-37] we dock HMBPP to the unique iron site in IspH. Docked poses are filtered applying knowledge from experiment that the terminal alkoxide/alcohol group of HMBPP directly chelates the apical Fe site.[26,28,38,39] The docked pose used in our MD studies is found by constraining the position of the terminal alkoxide moiety to within a 2.5 Å radius of the apical iron. While the orientation of the PP_i moiety in our docked pose differs from the $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ IspH crystal structure (3KE8),[28] it is worth mentioning that the cyclic structure of HMBPP observed in the crystal structure likely results from "induced fit" effects, with polar and charged groups closing around the PP_i moiety. Given these effects are absent from our docking procedure, we use the Glide geometry as a starting point for elucidating how open, substrate-free IspH responds to the formation of an encounter complex with HMBPP bound to its unliganded Fe.

aMD simulations of apo-IspH with docked HMBPP

Similar to the $[Fe_4S_4]^{2+}_{(open,\ substrate\text{-}free)}$ simulations, three independent, 100 ns aMD simulations of HMBPP docked into the open, $[Fe_4S_4]^{2+}$-IspH structure (henceforth referred to as $[Fe_4S_4]^{2+}/HMBPP_{(open,\ docked)}$) are aligned to the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ IspH crystal structure, with the RMSD of all backbone atoms to the crystal structure given in Figure 3-3B. Both seeds one and three (Figure 3-3B, black and blue, respectively) approach an RMSD of ~ 8-10 Å, with respect to the crystal structure. This jump occurs rapidly for seed three (in the first 20 ns of simulation), while seed one only appears to approach this level in the last 10 ns of simulation. This shift from the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ IspH crystal structure results from the closing of D2 and D3 around the docked HMBPP, matching the conformation assumed by the $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ IspH crystal structure (Figure 3-S1).

To gain insight into the dominant conformations sampled by these $[Fe_4S_4]^{2+}/HMBPP_{(open,\ docked)}$ simulations, we cluster the frames of each trajectory according to pairwise RMSD comparing C_α atoms, as described in the Methods. The dominant cluster (58%) corresponds to an open conformation, similar to the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ IspH crystal structure.[28] The second most populated cluster (18%) contains closed structures resembling the $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ IspH crystal structure (Figure 3-4). When considering the structures in this closed cluster, it is notable that the ligand does not form a ring structure consistent with its pose in the crystal structure. Nevertheless, the closing of the D2 and D3 domains around the substrate is consistent with the $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ experimental reference.[28] A

more detailed inspection of the HMBPP environment in a representative structure from this closed cluster reveals the three key active site histidines, as well as the conserved Thr-165, Thr-166, Glu-126, Ser-221, Asn-223 and Ser-265 forming contacts with HMBPP that appear identical to those seen in the $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ IspH crystal structure (Figure 3-4B,C). While Glu-126 and Thr-167 are co-localized with the iron-sulfur cluster in the active site of $[Fe_4S_4]^{2+}_{(open,\ substrate\text{-}free)}$ IspH (Figure 3-4D), the other contacts mentioned are unique to substrate-bound IspH, as seen in the $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ crystal structure (Figure 3-4B). These findings demonstrate that aMD simulations have effectively captured the closing of loops from D2 and D3 around HMBPP—confirming earlier hypotheses for how conformational change occurs upon substrate binding.[28]

Inconsistent with the RMSD results for seed three, both seeds one and two (black and red, Figure 3-3B) are trapped in a basin near the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ IspH crystal structure for a majority of their respective simulations. These differing trajectories arise, in part, because the residues in D3 that coordinate the PP_i of HMBPP in seeds one and two do not coordinate the bound HMBPP. For instance, the side chain of Ser-265 in seed two does not extend inward toward the bound substrate, instead interacting with loop residues at the interfaces of D3 with D1 (Phe-12, Asn-43 and Thr-266). The local conformations of these residues are more consistent with those observed in $[Fe_4S_4]^{2+}_{(open,\ substrate\text{-}free)}$ simulations of IspH. Coupled with the observed closing event in seed three, these findings demonstrate the presence of a barrier between the open and closed states, requiring the intramolecular interactions

present in the substrate-free state to break in order to form interactions with bound HMBPP.

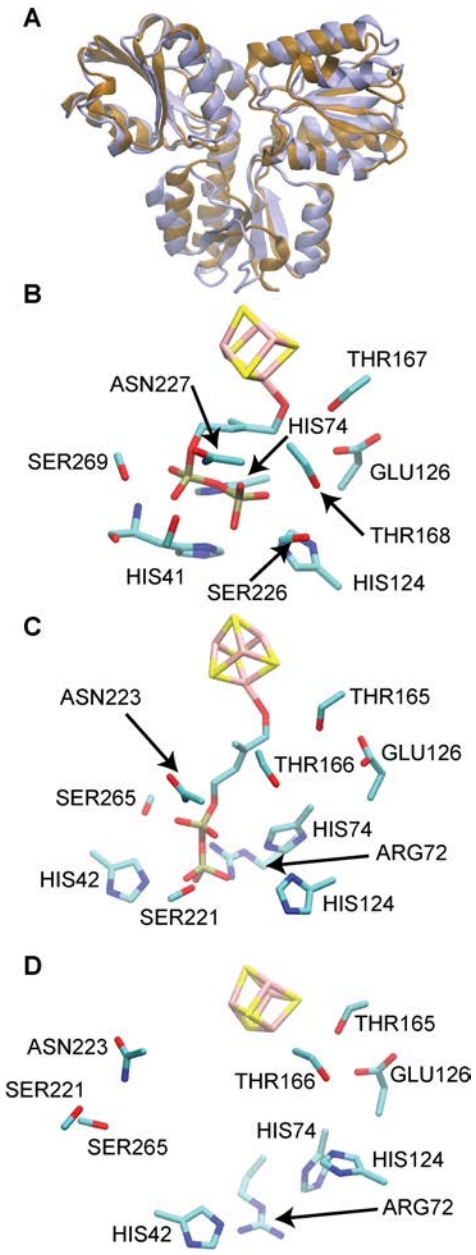

Figure 3-4. Comparison of HMBPP-bound IspH structures from experiment and simulation. (A) shows the superposition of the $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ IspH crystal structure (bronze, [28]) and a representative structure from the dominant closed cluster from $[Fe_4S_4]^{2+}/HMBPP_{(open,docked)}$ aMD simulations (purple). (B) and (C) correspond to the active site microenvironments of the crystal structure and the representative closed structure from aMD, respectively, while (D) illustrates the positions of the residues present in (B) and (C) in the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ crystal structure [23].

Simulations of the closed IspH:HMBPP complex

The second most populated cluster from $[Fe_4S_4]^{2+}$/$HMBPP_{(open,\ docked)}$ IspH simulations, which corresponds to the most populated closed conformation, is used as a starting point for three additional 100 ns aMD simulations (henceforth referred to as $[Fe_4S_4]^{2+}$/$HMBPP_{(closed)}$ simulations). Plots of the computed RMSD with respect to the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ IspH crystal structure for these simulations are marked by their lack of change, not deviating more than ~3 Å from the closed conformations sampled in $[Fe_4S_4]^{2+}$/$HMBPP_{(open,\ docked)}$ simulations (Figure 3-3C, Figure 3-S1). Similar to what is seen in the $[Fe_4S_4]^{2+}$/$HMBPP_{(open,\ docked)}$ aMD simulations, we note that HMBPP never fully reaches its ring conformation seen crystallographically.[28] From these simulations, it is evident that substrate-bound IspH, once folded around HMBPP, has less conformational space accessible to it and does not access open states.

Assessing sampling using principal component analysis

In constructing principal component (PC) space from all $[Fe_4S_4]^{2+}_{(open,\ substrate\text{-}free)}$ and $[Fe_4S_4]^{2+}$/$HMBPP_{(open,\ docked)}$ simulations, as described in the Methods, we observe that the first two principal components account for 83% of the variance. Using Bio3D,[40] the motions that correspond to movement along PC1 and PC2 are visualized (Figure 3-S2) and are shown to correspond to opening and closing motions achieved through the hinge-like properties of the loops that connect D3 to D1 and D2 and D2 to D1 and D3, as suggested by Groll and co-workers.[28]

All simulations ($[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$, $[Fe_4S_4]^{2+}/HMBPP_{(open,docked)}$, and $[Fe_4S_4]^{2+}/HMBPP_{(closed)}$), as well as the coordinates from the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ and the $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ IspH crystal structures, are projected onto the PC space to assess how the simulations sample configuration space with respect to the crystal structures within this coordinate system (Figure 3-5). Viewing these projections, it is clear that the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ simulations (Figure 3-5A) sample significantly greater conformational space than the $[Fe_4S_4]^{2+}/HMBPP_{(open,docked)}$ and $[Fe_4S_4]^{2+}/HMBPP_{(closed)}$ simulations (Figure 3-5B,C). While other local minima are present, the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ simulations sample energy wells near both the open (PDB ID: 3DNF) and closed (PDB ID: 3KE8) crystal structures along PC1 but do not overlap with the latter, HMBPP-bound crystal structure. This finding suggests that the precise closing motions that accompany ligand binding are absent without HMBPP bound to IspH, despite the intrinsic ability of $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ IspH to sample alternative closed states.

Volume analysis of the states sampled in the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ IspH simulations demonstrates the extent to which various open and closed states are sampled within this PC framework. Using the Pocket Volume MEasurer (POVME) program,[41] the volumes of representative structures from the clusters generated from $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ aMD trajectories are obtained and given in Figure 3-5A. Using this algorithm, it is notable that the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ IspH crystal structure[23] has a binding pocket volume of 451 $Å^3$, whereas the $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ IspH crystal structure[28] has a volume of 6 $Å^3$ (71 $Å^3$, in the absence of HMBPP). Movement

along PC1 generally accompanies a decrease in binding pocket size in the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ aMD simulations (from 612 $Å^3$ at the most negative values of PC1 to 319 $Å^3$ at the most positive values, Figure 3-5A). The characteristics of these different pockets are probed later in this report.

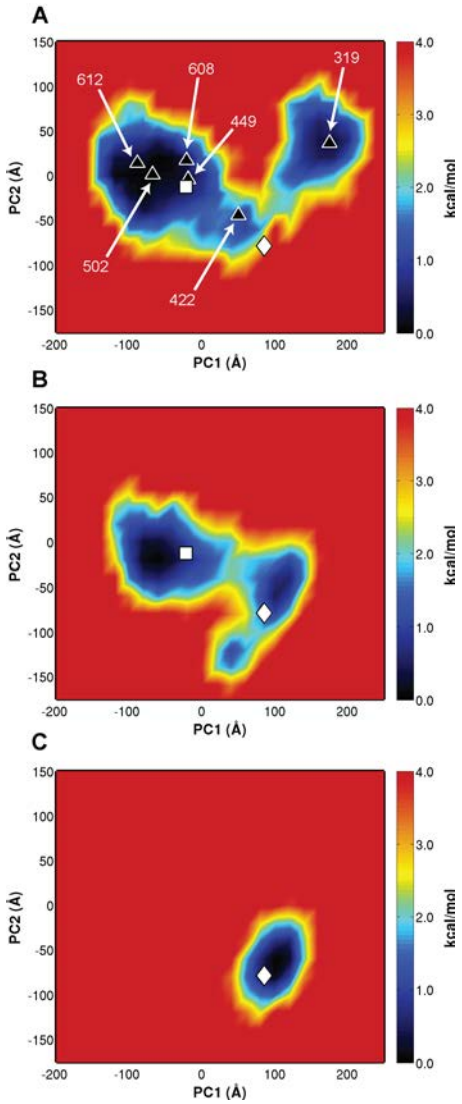

Figure 3-5. Projections of (A) $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$, (B) $[Fe_4S_4]^{2+}$/$HMBPP_{(open,docked)}$, and (C) $[Fe_4S_4]^{2+}$/$HMBPP_{(closed)}$ trajectories onto principal component (PC) space constructed from $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ and $[Fe_4S_4]^{2+}$/$HMBPP_{(open,docked)}$ aMD simulations. Crystal structures corresponding to $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ (white square, PDB ID: 3DNF) and $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ IspH (white diamond, PDB ID: 3KE8) are also projected onto PC space. Numbers in (A) correspond to POVME volumes ($Å^3$), as described in the text.

Projection of the $[Fe_4S_4]^{2+}$/$HMBPP_{(open,docked)}$ simulations onto PC space reveals a single, clear pathway for the transition between open and closed states (Figure 3-5B). Three minima are apparent in the projections, one centered near the substrate-free crystal structure and two near the HMBPP-bound crystal structures that differ slightly in the specific contacts made between the protein and ligand. The extent to which bound-HMBPP restricts IspH dynamics is highlighted from the projection of the $[Fe_4S_4]^{2+}$/$HMBPP_{(closed)}$ simulations onto PC space. When simulated from a closed conformation, it is clear that bound-HMBPP effectively locks the protein in a closed conformation, unable to access open states—evident by a well present only around the closed, HMBPP-bound IspH crystal structure (Figure 3-5C).

RMSF analyses of different HMBPP-bound states

Combining the three trajectories for each individual system simulated, we performed a root-mean-square fluctuation (RMSF) analysis to quantify the extent to which each residue fluctuates in the different systems (Figure 3-6). In the case of the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ simulations (Figure 3-6, black curve), the fluctuations in D2 and D3 are slightly greater than what is seen in the $[Fe_4S_4]^{2+}$/$HMBPP_{(open,docked)}$ simulations. These fluctuations are abolished when the simulations are started from a closed conformation with HMBPP bound ($[Fe_4S_4]^{2+}$/$HMBPP_{(closed)}$ simulations).

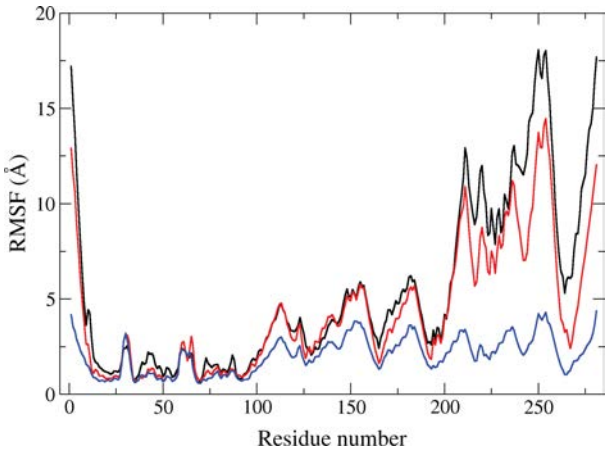

Figure 3-6. RMSF analysis of $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ (black), $[Fe_4S_4]^{2+}/HMBPP_{(open,docked)}$ (red), and $[Fe_4S_4]^{2+}/HMBPP_{(closed)}$ (blue) aMD simulations.

Local phenomena driving IspH conformational changes

Changes in various peptide dihedral angles (phi, psi and chi) typically accompany global changes in protein conformation.[42,43] In other words, certain dihedral angles may select for specific conformations in proteins.[42,43] Recently, McClendon *et al.* contributed a method that quantifies differences in probability distributions of protein dihedral angles between a reference and altered state of a protein by using an expansion of the Kullback-Leibler (KL) Divergence.[42] This application assigns a value for the KL divergence of each residue (KL_{res})—a measure of the extent to which the distributions of dihedral angles differ between the two states. Using the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ simulation as a reference, we compute KL_{res} upon substrate binding to IspH using the MutInf suite of programs[42,44] in an attempt to isolate local changes in protein structure that give rise to globally different conformational ensembles between the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ and $[Fe_4S_4]^{2+}/HMBPP_{(closed)}$ aMD simulations.

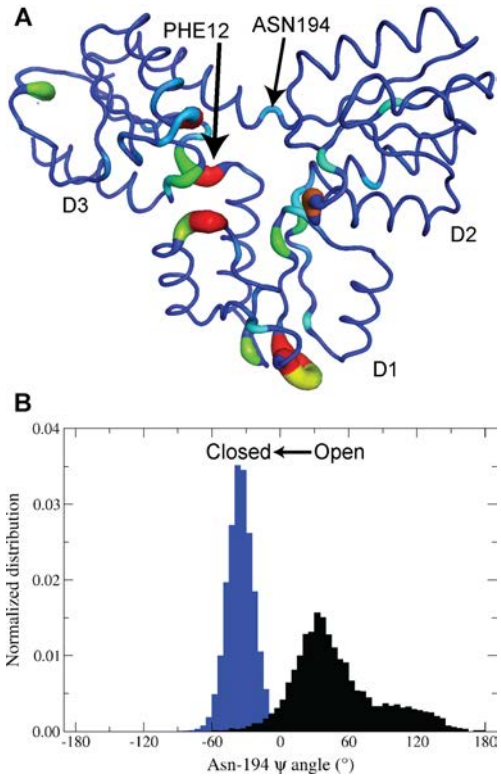

Figure 3-7. (A) Use of the Kullback-Leibler (KL) divergence to highlight residues with distinct dihedral distributions between $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ and $[Fe_4S_4]^{2+}$/HMBPP$_{(closed)}$ simulations of IspH. Visualization of residues with high values of KL_{res} in the IspH structure, ranging from blue (low) to red (high). (B) Illustration of the different dihedral distributions of the Asn-194 ψ-angles sampled in open (black) and closed (blue) conformations of IspH.

A visual representation of KL_{res} values is provided in Figure 3-7A, with the highest scoring residues shown in Table 3-1. We present these data together with measures of sequence conservation, computed as Shannon entropy,[40,45,46] as both these metrics are suggested to highlight residues of functional importance.[42,45] The link between sequence conservation and functionality is obvious—residues that are highly conserved are usually conserved for some purpose, *e.g.* to bestow certain structural features to a protein or to participate in catalysis. Similarly, residues whose conformations change dramatically upon some natural perturbation to the system, ligand binding in our case, are likely responsible for the functionality of that protein.

Consequently, we propose that residues that both are highly conserved and display high KL_{res} upon ligand binding are critical to the structure and function of IspH.

Table 3-1. Residues with KL_{res} greater than 1.0 and their respective sequence conservation scores, as computed by Shannon entropy.[40,45,46]

Residue	KL_{res}	Sequence conservation score
GLY11	**1.01**	**0.99**
PHE12	**3.06**	**0.83**
LYS33	1.17	0.31
HIS42	**2.44**	**0.99**
ASN43	**1.26**	**0.95**
PHE63	2.25	0.37
LYS64	1.46	0.00
GLU65	1.35	0.17
GLY66	2.42	0.24
ASP67	5.18	0.18
ARG72	**1.03**	**0.77**
HIS124	**1.79**	**1.00**
ASN223	**2.31**	**1.00**
TRP251	1.16	0.38
SER265	**1.03**	**1.00**

Interestingly, five residues displaying higher KL_{res} (Phe-63, Lys-64, Glu-65, Gly-66 and Asp-67) are located in a loop region in D1 and are neither conserved nor directly interacting with bound-HMBPP (Figure 3-7, Table 3-1). Arg-72 and His-74 are positioned at the opposite end of this loop region and form hydrogen bonds with the PP_i tail of HMBPP in the $[Fe_4S_4]^{2+}/HMBPP_{(closed)}$ simulations. From these observations, it can be reasoned that the conformations of Arg-72 and His-74, altered upon HMBPP binding, in turn disrupt the conformations of the residues at the end of the loop.

Most other residues with high KL_{res} can be characterized by one of two distinct environments in the protein: either (a) coordinating HMBPP when it is bound (*e.g.* His-42, His-124, Asn-223 and Ser-265); or (b) structurally flanking the thiolates that anchor the $[Fe_4S_4]^{2+}$ cluster to the protein (as is the case for Phe-12, which is adjacent to Cys-13).

High KL_{res} is seen for residues that occupy the first coordination shell of HMBPP when it is bound. These residues assume different conformations based upon whether they are coordinating the substrate. For instance, both His-42 and Arg-72 from D1, His-124 from D2 and both Asn-223 and Ser-265 from D3 all assume different main and side chain dihedral angle distributions in the $[Fe_4S_4]^{2+}$/HMBPP$_{(closed)}$ state compared to the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ state. These differences derive from the reorientation of these residues about the PP_i of HMBPP in order to participate in hydrogen bonds or salt bridges.

The other class of residues with significant KL_{res} reside adjacent to the thiolates tethered to the $[Fe_4S_4]^{2+}$ cluster. Phe-12 exemplifies this finding, in that it maintains altered ϕ/ψ angle distributions, contingent on whether HMBPP is bound (Figure 3-S3). The case of Phe-12 suggests similar behavior may exist in other thiolate-adjacent residues. In inspecting the dihedral angle distributions of Thr-95 and Asn-194 in the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ and $[Fe_4S_4]^{2+}$/HMBPP$_{(closed)}$ simulations (both having more modest KL_{res} of 0.77 and 0.32, respectively; Figure 3-S3), it is evident that while the ϕ/ψ angle distributions are virtually identical for Thr-95, Asn-194 samples different distributions in HMBPP-free and bound states, much like Phe-12.

Unlike Phe-12, however, the ϕ/ψ angles of Asn-194 are unimodal in the closed simulations, indicating that closed conformations require that Asn-194 maintain certain backbone dihedral angles. Indeed, when the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ and $[Fe_4S_4]^{2+}$/$HMBPP_{(closed)}$ simulations are combined and clustered together into open and closed conformations, it is clear that Asn-194 samples entirely different psi angles, contingent on whether D2 and D3 are open or closed (Figure 3-7B).

Whereas the psi angle for Asn-194 in open states contributes to the residue's disordered secondary structure, as computed by STRIDE calculations,[47] Asn-194 in all closed states is strictly α-helical with a mean psi angle of -34°. It is clear from these distributions that dihedral angles near -34° select for the closed conformations of IspH and contribute to the helicity of Asn-194, unseen in the open conformations that only exist in the ensemble of states sampled in $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ simulations.

Moving from the dihedral angle to the global structure of D3, it is evident that the helicity of Asn-194 is achieved via cranking motions that pull the helix, comprised of residues 195 to 207 and anchored by Asn-194, behind the [4Fe-4S] cluster in all closed states. This "crank" motion effectively compresses the D3 domain and also draws inward the loops needed to corral HMBPP into a closed active site. In contrast, Asn-194 samples states with no ordered secondary structure in $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ IspH simulations, while extended in the open conformation.

Implications of an expanded binding pocket

Clustering of the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ IspH aMD simulations reveals dominant structures with substrate pockets of differing volumes and chemical

environments. Using representative structures from each of the clusters, we investigate the druggability of these different pockets by performing solvent mapping with the FTMAP program.[48]

Taking the fragment positions as they are docked by FTMAP into each representative structure from the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ simulations, we synthesize information regarding where the docked fragments congregate by generating a probe occupancy map for IspH. Probe occupancy is highest at the pocket corresponding to the substrate-binding site (Figure 3-8A, Figure 3-S4). In the more voluminous clusters, as well as the most dominant cluster, probes expand beyond the HMBPP-binding site at the iron, into a crevice between D1 and D3 (Figure 3-8A, Figure 3-S4). This finding suggests that inhibitors capable of occupying this expanded pocket while locking the protein in a state that is more open with respect to the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ IspH crystal structure may provide a feasible route toward novel inhibitor design.

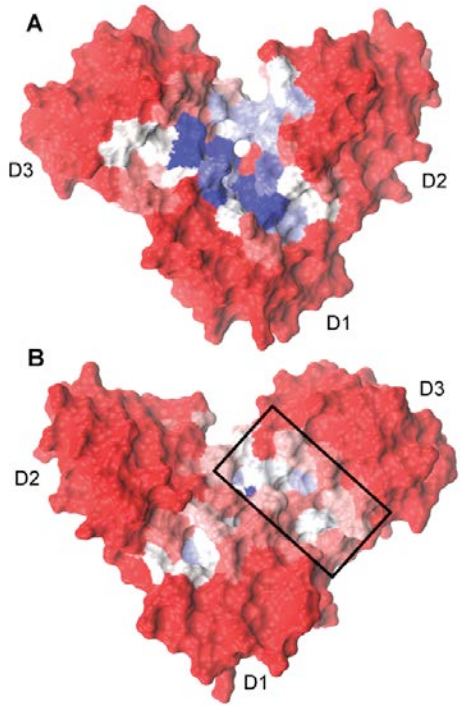

Figure 3-8. IspH structures as seen from the front (A) and the back (B) with respect to the binding pocket location, colored by normalized FTMAP probe occupancy (red to white to blue follows occupancies of 0.0 to 0.5 to 1.0). (A) illustrates the high propensities of FTMAP fragments to bind to the active site, as well as to the interface between D1 and D3. (B) reveals an allosteric pocket between D1 and D3 (highlighted by the black box).

An unanticipated finding from solvent mapping concerns the side of IspH opposite the substrate-binding pocket. When the protein opens fully, as seen in the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ simulations, the hinge-like quality of the interface between D3 and D1/D2 hyperextends, creating a druggable pocket found opposite the side of the HMBPP-binding site (Figure 3-8B, black rectangle; Figure 3-S4). When the hinge is opened, this pocket occupies a POVME-measured volume of 330-500 $Å^3$ and accommodates a variety of polar and nonpolar probes. This result, stemming from the opening motions intrinsic to substrate-free, $[Fe_4S_4]^{2+}$ IspH, may provide an allosteric target for inhibitor design.

Discussion

Application of the aMD method to sample conformational space in both $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ and $[Fe_4S_4]^{2+}/HMBPP_{(closed)}$ states of IspH increases our understanding of how HMBPP binding affects IspH structure and dynamics, as well as highlights alternative routes for the design of novel IspH inhibitors.

In regard to IspH dynamics, our $[Fe_4S_4]^{2+}/HMBPP_{(open,docked)}$ aMD simulations are able to capture the closing event that accompanies ligand binding in two out of three simulations. In these simulations, residues in D1 that are needed to coordinate HMBPP (His-42 and His-74) are already properly positioned to interact with the pyrophosphate tail of HMBPP, whereas residues in D2 and D3 that coordinate HMBPP require domain motions to bring them in proximity of the substrate. Once D2 and D3 close around HMBPP, it is apparent from our PCA that IspH is unable to reopen, with the fluctuations of residues from D2 and D3 largely suppressed as these domains engage in multiple electrostatic and hydrogen bond interactions with HMBPP (*e.g.* His-124, and Ser-226). These observations underscore the suggestions by others that both electron addition to the substrate and changes in active site and substrate titration states are necessary, not only for catalysis, but also to alter the electrostatics in the active site to enable IspH opening and release of the catalytic product, IPP or DMAPP.[24]

In contrast with the $[Fe_4S_4]^{2+}/HMBPP_{(open,docked)}$ and $[Fe_4S_4]^{2+}/HMBPP_{(closed)}$ systems, $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ IspH is much more flexible and thus able to access both closed states and conformations that open beyond what is seen in the $[Fe_3S_4]^{+}_{(open,}$

substrate-free) IspH crystal structure. When closed in our simulations, projections of $[Fe_4S_4]^{2+}_{(open,substrate-free)}$ IspH onto PC space show no overlap with the $[Fe_4S_4]^{2+}_{(closed,}$ HMBPP-bound) IspH crystal structure, indicating that substrate binding allows IspH to sample a closed state that is inaccessible in the absence of HMBPP.

In our simulations of the substrate-free state, IspH accesses both open and closed conformations. Since closed states preexist in the substrate-free ensemble, it is tempting to suggest that conformational selection (CS)[49] is responsible for ligand recognition in IspH. Following the logic of Sullivan and Holyoak,[50] however, induced fit (IF) likely better describes the conformational changes occurring upon ligand binding since HMBPP cannot actually bind to the closed state that preexists in substrate-free IspH (due to occlusion of the active site by D2 and D3).[51] We propose that ligand binding may still be described as a combination of CS and IF, where the ligand initially selects open conformations for formation of an encounter complex. Once initially bound, HMBPP induces closure of D2 and D3 via motions that are also intrinsic to IspH in the absence of ligand. This ligand recognition mechanism, drawing from both CS and IF, is not unique to IspH, but rather gives further support to the suggestions of others that ligand binding can contain elements of both CS and IF.[52-54]

Returning to the structures observed in the $[Fe_4S_4]^{2+}_{(open,substrate-free)}$ IspH aMD simulations, it is interesting that the active site volume is subject to significant fluctuations—largely due to the flexibility of the loop regions connecting D3 to D1 and D2 and, to a lesser extent, D2 to D1 and D3. These fluctuations are expected, as HMBPP likely binds initially to an enlarged binding pocket that may accommodate

the expansive hydration shell expected for pyrophosphate-containing molecules[55] that HMBPP carries from solution into an encounter complex with IspH. The larger pocket stemming from the super-open state seen in the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ simulations would allow for this initial complex to form. Given the presence of these larger pockets in our simulations and this mechanistic rationale, it is reasonable to hypothesize that a variety of differently sized ligands may also be accommodated in the binding pocket.

Combining these volume data for the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ state with the results from our KL divergence analysis and FTMAP solvent mapping of IspH, we can build on the work of others[29-34] in suggesting a novel framework for future IspH inhibitor design. HMBPP binding to IspH can be regarded the first step in the catalytic process vital to most microbes for production of IPP and DMAPP. Preventing this binding event is thus the goal of competitive inhibitor development.

From our KL divergence analysis, we find that in addition to conserved residues that coordinate HMBPP upon its binding, residues that are adjacent to thiolate residues achieve high KL_{res} due to their distinct dihedral distributions when IspH is open and closed. Given its position adjacent to the fully conserved Cys-193, Asn-194 likely coordinates the hinge motions of D3 that give way to the necessary closing events that accompany HMBPP binding. Preventing the closing of the D3 hinge and, consequently, locking the Asn-194 backbone dihedrals in their disordered, open conformations may provide a novel mode of inhibiting IspH.

From our aMD simulations, two differing mechanisms for disrupting the hinge motions of D3 are apparent. The first targets the outward motion of D3 from the HMBPP binding site that creates an enlarged cavity that extends from the active site to the interface between D3 and D1 (Figure 3-8A). Either design of larger competitive inhibitors that interact with the apical iron and the D3/D1 interface or design of ligands that interact allosterically with the D3/D1 interface could successfully exploit the enlarged pocket on the active site side of IspH. Alternatively, the presence of an allosteric pocket opposite the side of the HMBPP binding site may be targeted for inhibitor design (Figure 3-8B, Figure 3-S4). Both these proposed sites for inhibitor design are "hot spots" found by solvent probes with FTMAP. Noting that probe occupancy correlates well with sequence conservation as measured by Shannon entropy ($r = 0.49$) provides further support for these suggested modes of inhibition.

Given the documented difficulties of rational drug design for metalloproteins, notably from a computational perspective,[56] allosteric sites that do not require a detailed description of metal binding (*e.g.* orbital interactions, polarization and charge transfer) are highly desirable if existent. Furthermore, it has been shown that perturbations to allosteric networks in redox-active metalloproteins may affect the redox potential of these proteins and, consequently, alter their activities.[57] These factors motivate us to include the different pockets revealed by aMD simulations, particularly those that may provide routes to noncompetitive inhibition, in future computer aided drug design workflows.

Conclusion

Using aMD simulations, we are able to capture the closing event that accompanies the binding of HMBPP to IspH when starting from the substrate-free crystal structure. Drawing from PCA and visual analyses of the different trajectories considered, we propose that ligand binding occurs via a combination of induced fit and conformational selection. We note that a single dihedral angle, the ψ angle in Asn-194, selects for either open or closed conformations of IspH, the latter being achieved via a crank motion that draws D3 inward to corral the active site. Furthermore, our aMD simulations reveal both an expanded active site pocket encompassing a crevice between D1 and D3, as well as an allosteric pocket between D1 and D3 on the side opposite the substrate binding pocket that may be utilized for the design of novel IspH inhibitors.

Methods

Ligand parameterization

Since the questions under consideration in this study begin with open, substrate-free IspH protein, we use the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ IspH crystal structure from Rekittke, *et al* (PDB ID: 3DNF) as a starting point.[23] Applying insight from the $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ IspH crystal structure from Gräwert, *et al.* (PDB ID: 3KE8),[28] we model the apical iron into the cluster by superposition.

Using the Amsterdam Density Functional program,[58] a model $[Fe_4S_4]^{2+}$ cluster is geometry optimized using broken symmetry density functional theory (BS-DFT)[59,60] at the OLYP/TZP level of theory.[61,62] With the Gaussian 09 suite of programs,[63] we

optimize the geometry of HMBPP and compute the electrostatic potentials of both geometry optimized HMBPP and the model $[Fe_4S_4]^{2+}$ cluster using MK radii[64] at the HF/6-31G(d) level of theory. The antechamber program[65] in the AmberTools 13 suite of programs[66] is then used to apply the restrained electrostatic potential (RESP) procedure to derive point charges for use in MD simulations. In the case of the $[Fe_4S_4]^{2+}$ cluster, parameters for nonbonded terms are taken from the AMBER GAFF force field,[67] and bonds and angles between atoms are implicitly accounted for by harmonic restraints applied to these terms, using parameters from the $[Fe_4S_4]^{2+}_{(closed, HMBPP\text{-}bound)}$ IspH crystal structure.[28] For HMBPP, all force field parameters are taken from the AMBER GAFF force field.[67] All charge and nonbonded parameters, as well as, a more detailed discussion of the ligand parameterization process, are included in the Supporting Information.

System preparation for molecular dynamics simulations

Hydrogens are added using PDB2PQR,[68,69] with protonation states assigned using the PROpKa program.[70] In our setup, His-42 and His-124 are set to their imidazolium states, and Glu-126 is protonated. Following hydrogen addition, the protein systems are minimized for 2000 steps in the gas phase using the sander module in AMBER12,[66] to remove problematic steric clashes. The systems are solvated in a box of TIP3P waters[71] that extends 12 Å beyond the closest solute atom, with counterions added to enforce electroneutrality. Non-water bonds to hydrogen atoms are constrained using the SHAKE algorithm,[72] while the O-H bonds in water are constrained using the SETTLE algorithm.[73] All protein force field parameters are

taken from the AMBER ff99SB force field,[74] while the ligand parameters referred to above are taken from the AMBER GAFF force field.[67] Subsequent 2000 step minimizations are performed (a) to relax the water with protein fixed by positional constraints, (b) to relax the protein with all waters constrained, and (c) relax the whole system. Following this minimization protocol, all systems are equilibrated at constant pressure and temperature (NPT) conditions for 1 ns, with the protein fixed by positional constraints. The pressure is regulated using the Berendsen barostat[75] with isotropic position scaling (ntp=1) and a pressure relaxation time of 2.0 ps, while a Langevin thermostat[76] with collision frequency of 2.0 ps^{-1} is used to increase the temperature of the system from 0 to 300K. The protein constraints are then lifted and a subsequent 2 ns NPT equilibration is performed at 300K to verify the density of the system is reasonable and stable. The last equilibration step is performed at constant volume and temperature (NVT) for 5 ns at 300K to prepare the system for production MD simulations. All dynamics are conducted using the pmemd.cuda engine,[66,77] with Particle Mesh Ewald summations used for computing long-range electrostatic interactions and short-range nonbonded interactions truncated beyond a cutoff of 10 Å.[78,79]

Accelerated molecular dynamics (aMD) simulations

Given current computational power, most MD simulations are limited to sampling timescales on the order of 10-1000 ns. Since many biomolecular processes, including, for example, protein folding, ligand binding, and cis/trans isomerization events, may occur on the order of milliseconds to days, enhanced sampling techniques

that facilitate traversing of configuration space efficiently are often implemented to provide information about the relevant conformations of biomolecules.[80,81] Accelerated molecular dynamics (aMD) simulations promote enhanced sampling of systems without the need for defining a reaction coordinate. In aMD simulations, when the potential energy of the system, V(r), is below a threshold energy level, E, a boost energy, ΔV(r), is applied to encourage exploration of other areas of phase space (Eq. 1). The parameter α modulates the aggressiveness of this boost by altering the depth of the wells in the modified potential.

$$(1) \qquad \Delta V(r) = \begin{cases} 0, & V(r) \geq E \\ \dfrac{\left(E - V(r)\right)^2}{\alpha + \left(E - V(r)\right)}, & V(r) < E \end{cases}$$

We employ the dual-boost implementation of aMD to boost both dihedral and total potential energy force field terms to promote side chain dihedral angle rotations and diffusive transitions, respectively.[82,83] We set the parameters E and α for our systems by defining these variables for the dihedral and total potential energy components with respect to the number of residues in the system, N_{res}, and the number of atoms in the system, N_{atoms}, respectively (Eq. 2-5):

$$(2) \qquad E_{DIHED} = \left\langle V_{DIHED}(r) \right\rangle + 2.5 \times N_{res}$$

$$(3) \qquad \alpha_{DIHED} = 1.25 \times N_{res}$$

(4) $$E_{TOT} = \langle V_{TOT}(r) \rangle + 0.17 \times N_{atoms}$$

(5) $$\alpha_{TOT} = 0.17 \times N_{atoms} \ .$$

Subsequent reweighting of the trajectory frames from the aMD simulations using a tenth-order Maclaurin series expansion allows us to extract canonical ensemble averages of the system (further details included in Text S2). Recently, both these methodologies for obtaining aMD parameters and reweighting aMD results were successfully applied to bovine pancreatic trypsin inhibitor (BPTI) to properly obtain the relative populations of relevant, low-lying energetic states.[84] For some semblance of statistics, 3 x 100 ns aMD simulations are performed on all systems explored in this study.

Molecular dynamics analysis

RMSD, RMSF, clustering, and dihedral angle analyses are all performed using the AmberTools 12 suite of programs.[66] Alignment procedures implemented in the RMSD and RMSF calculations are performed with respect to the $[Fe_3S_4]^+_{(\text{open, substrate-free})}$ IspH crystal structure (PDB ID: 3DNF), aligning to the backbone atoms of D1, as this domain is the most rigid in all simulations. Clustering analyses for each of the simulations use pairwise RMSD computed for C_α atoms between frames to divide the cumulative trajectories for each system simulated into eight clusters using the average-linkage algorithm.[85]

Principal component analysis

Principal component analysis (PCA) reduces atomic fluctuations in the various trajectories into vectors that represent the dominant correlated motions present in the simulations.[86,87] Since we want our PCA to assess how well the different simulations sample conformational space with respect to the $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ and $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ IspH crystal structures (PDB ID: 3DNF and 3KE8, respectively), we first align the two crystal structures using the STructural Alignment of Multiple Proteins (STAMP) procedure,[88] as implemented in the VMD MultiSeq plugin.[89,90] The indices of aligned residues are then used in subsequent PCA.

Principal component (PC) space is constructed from the three $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ simulations and three $[Fe_4S_4]^{2+}/HMBPP_{(open,docked)}$ simulations. The trajectories for each set of simulations ($[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$, $[Fe_4S_4]^{2+}/HMBPP_{(open,docked)}$, $[Fe_4S_4]^{2+}/HMBPP_{(closed)}$) are then projected onto the first and second principal components. Additionally, $[Fe_3S_4]^{+}_{(open,\ substrate\text{-}free)}$ and $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ IspH crystal structures are projected onto PC space to assess overlap between the simulations and these structures along the PC1 and PC2 coordinates. The modes that correspond to PC1 and PC2 are visualized using the Bio3D suite of programs.[40]

Application of the Kullback-Leibler Divergence

We quantify differences in IspH structure upon ligand binding by applying the Kullback-Leibler (KL) Divergence expansion, also referred to as relative entropy, to assess differences in the distributions of φ, ψ, and χ dihedral angles in

$[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ and $[Fe_4S_4]^{2+}/HMBPP_{(closed)}$ ensembles generated by aMD simulations. To obtain the KL divergence for each residue (KL_{res}), we first split the 3 x 100 ns sets of simulations for the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ and $[Fe_4S_4]^{2+}/HMBPP_{(closed)}$ systems into 6 sets of 50 ns to provide statistical robustness to the calculations. The MutInf program[42,44] processes the dihedral angle distributions for each of these 50 ns blocks as computed by the g_torsion program from the GROMACS suite of programs,[91] computing the KL Divergence for a specific dihedral angle using Eq. 6:

$$(6) \qquad KL = \sum_{i}^{n\,bins} p_i \ln \frac{p_i}{p_i^*} \ .$$

In this equation, p_i refers to the probability that a particular dihedral angle from the $[Fe_4S_4]^{2+}/HMBPP_{(closed)}$ simulations falls into a specific range of torsional space, which has been divided into 12° bins. The term p_i^* is the corresponding probability that the same dihedral angle from the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ simulation falls into the same bin. Combining the KL terms for each of the dihedral angles (φ, ψ, and χ's) of a given residue provides a value for the KL divergence of a specific residue (KL_{res}):

$$(7) \qquad KL_{res} = \sum_{\varphi,\psi,\,and\,\chi's} \sum_{i}^{n\,bins} p_i \ln \frac{p_i}{p_i^*} \ .$$

This value for the KL divergence of a given residue provides a measure of the difference between the dihedral angle probability distribution functions of the

$[Fe_4S_4]^{2+}$/HMBPP$_{(closed)}$ simulations with respect to the $[Fe_4S_4]^{2+}_{(open,\ substrate\text{-}free)}$ reference simulations.

Sequence conservation analysis

Using the Bio3D suite of programs, we compute the Shannon entropy[45,46] according to Eq. 8 for all residues in the *A. aeolicus* IspH structure with a 22-letter alphabet, where the 20 amino acids are included, as well as a gap character '-' and a mask character 'X'.[40] After normalizing the Shannon entropy score, residues that are fully conserved assume the value 1, while residues with no conservation have a Shannon entropy of 0.

$$(8) \qquad s = -\sum_{1}^{22} p_i \ln p_i$$

FTMAP

We employ FTMAP[48] to allow many drug-like, organic fragments to bind to representative structures from the dominant clusters from $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ aMD simulations. FTMAP utilizes a fast Fourier transform (FFT) algorithm to allow the organic probes to sample many positions along the protein surface. Prevalence of fragment hits along the protein surface signifies "hot spots" that correspond to potentially druggable pockets.[48]

We measure the ability of residues in substrate-free IspH to bind FTMAP probes by first defining binding of the residue by the probe as existent if the distance between their respective heavy atoms is less than 5 Å. We then combine these binding

results across all dominant clusters from the $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ aMD simulations and count the number of probes that bind each residue. Normalization of these data indicates the relative propensity of each residue to bind drug-like molecules.[92]

Acknowledgements

We would like to thank Prof. Eric Oldfield (University of Illinois, Urbana-Champaign) for his direction and active work in the area of IspH. We are grateful for helpful conversations regarding iron-sulfur proteins and IspH with Prof. Louis Noodleman (The Scripps Research Institute). Additionally, we appreciate computational help from Dr. Levi Pierce.

Chapter 3 is a minimally modified reprint of the material as it appears in Patrick G. Blachly, Cesar A.F. de Oliveira, Sarah L. Williams, and J. Andrew McCammon, "Utilizing a Dynamical Description of IspH to Aid in the Development of Novel Antimicrobial Drugs," *PLoS Computational Biology*, 2013. The dissertation author was the primary investigator and author of this paper.

Supporting Information

Obtaining force field parameters

To obtain charges for the [4Fe-4S] cluster, as well as the three thiolate residues that coordinate it, we construct a model cluster of the form $[Fe_4S_4(SCH_3)_3OH_2]^{1-}$ (Figure 3-S5). The formal charges of this model cluster are as follows: a spin-coupled ferric/ferrous iron pair (1 x Fe^{3+}; 1 x Fe^{2+}), a spin-coupled, delocalized mixed-valence

iron pair (2 x $Fe^{2.5+}$), four inorganic sulfides (4 x S^{2-}) and three methyl thiolates (3 x $^{-}SCH_3$). A water is placed in the model to coordinate the unliganded iron.

The [4Fe-4S] model cluster is geometry optimized using the Amsterdam Density Functional program[58,93,94] with broken symmetry density functional theory (BS-DFT).[59,95] In practice, this is achieved by following the procedure implemented by Noodleman and co-workers in studying Fe-S clusters, which involves three separate calculations: (1) a high spin (ferromagnetic) single point calculation is performed, with all iron net spins aligned parallel; (2) the spins on one Fe-Fe pair are flipped, such that there are two sets of Fe-Fe pairs containing spins that are oppositely aligned, allowing a broken symmetry single point calculation to be performed; (3) utilizing the broken symmetry wavefunction, the model cluster is geometry optimized.[60] Geometry optimization is achieved using the OLYP functional[61,62] with the TZP basis set. This method has previously been shown to give good agreement with experimental geometries and spectroscopic parameters for Fe-S systems.[96]

Following geometry optimization, the electrostatic potential (ESP) is computed at the HF/6-31G(d) level of theory using the Gaussian 09 suite of programs.[63] From the ESP, charges are obtained using the standard RESP procedure within the AMBER suite of programs.[66] The values of these charges, which correspond to the atom labels shown in Figure 3-S5, are given in Table 3-S1. For the $Fe^{2+/3+}$ and S^{2-} nonbonded parameters, we employ the values obtained and utilized by others for heme groups[97] and thiolates,[98] respectively (Table 3-S2).

To parameterize HMBPP, the molecule is first geometry optimized at the B3LYP/6-31G(d) level of theory[62,99-101] using the Gaussian 09 suite of programs.[63] The electrostatic potential is obtained at the HF/6-31G(d) level of theory, from which point charges are extracted with RESP. The charges for all HMBPP atoms are included along with their respective AMBER GAFF force field atom types (Table 3-S3), the latter determining the HMBPP nonbonded parameters used in this work.[102] For reference, atom names given in Table 3-S3 are included in a visual representation of HMBPP in Figure 3-S6.

Reweighting aMD trajectories for PCA

The principal component analysis (PCA) plots presented in this work follow the procedure outlined by Pierce *et al.* and employ a 2D histogram across the PC1 and PC2 coordinates.[103] PC space is divided into bins of dimension 1Å x 1Å, and points sampled during the aMD simulations are placed into their respective bins. To facilitate this procedure computationally, we utilize an indicator function, $\delta_{k,ij}$, that accounts for whether each point in PC space, (PC1, PC2), falls into the bin ($PC1_i$, $PC2_j$) for a given trajectory frame, k (Eq. S1):

$$\delta_{k,ij} = \begin{Bmatrix} 1, & (PC1, PC2)_k \in (PC1_i, PC2_j) \\ 0, & otherwise \end{Bmatrix} . \quad \text{(S1)}$$

Using $\delta_{k,ij}$, the histogram at bin ($PC1_i$, $PC2_j$) can be reweighted using Eq. S2, where K is the total number of trajectory frames and ΔV_k is the total boost potential applied at frame k:

$$H_{ij} = \sum_{k=1}^{K} \delta_{k,ij} * \exp\left(\beta \Delta V_k\right) . \quad \text{(S2)}$$

In our analyses, we perform this reweighting using a tenth-order Maclaurin series expansion of the exponential (Eq. S3):

$$\exp\left(\beta \Delta V_k\right) \approx 1 + \frac{\beta \Delta V_k}{1!} + \frac{\left(\beta \Delta V_k\right)^2}{2!} + \ldots + \frac{\left(\beta \Delta V_k\right)^{10}}{10!} . \quad \text{(S3)}$$

After reweighting all the histograms in PC space, we obtain the PC1-PC2 free energy plot shown in Figure 5 using Eq. S4,

$$W_{ij} = -k_B T \ln H_{ij} + W_0 , \quad \text{(S4)}$$

where k_B is the Boltzmann constant, T is temperature (taken to be 300K) and W_0 is a constant chosen to set the free energy minimum to zero.

Table 3-S1. Charge parameters for $[4Fe4S]^{2+}$ cluster and its liganded cysteines.

[4Fe-4S]	
Fe1	0.703
Fe2	0.772
Fe3	0.981
Fe4	0.652
S1	-0.447
S2	-0.738
S3	-0.736
S4	-0.788

	CYS13	**CYS96**	**CYS193**
N	-0.463	-0.463	-0.463
HN	0.252	0.252	0.252
CA	0.136	0.145	0.143
HA	0.048	0.048	0.048
CB	-0.177	-0.177	-0.177
HB3	0.085	0.094	0.092
HB2	0.085	0.094	0.092
SG	-0.677	-0.529	-0.477
C	0.616	0.616	0.616
O	-0.504	-0.504	-0.504

Table 3-S2. Nonbonded parameters used for the $[4Fe-4S]^{2+}$ cluster in simulations of IspH.

	r	ε
Fe	1.20	0.05
S	2.00	0.25

Table 3-S3. Force field parameters used for HMBPP. The atom types listed are assigned their respective nonbonded parameters in the AMBER GAFF force field.

Atom Name	Charge	AMBER GAFF Atom type
O1	-0.905	o
O2	-0.905	o
O3	-0.905	o
P1	1.206	p5
O4	-0.568	os
O5	-0.893	o
O6	-0.893	o
P2	1.345	p5
O7	-0.651	os
C1	0.493	c3
H1	-0.056	h1
H2	-0.056	h1
C2	-0.408	c2
H3	0.128	ha
C3	0.009	c2
C4	-0.101	c3
H4	-0.004	hc
H5	-0.004	hc
H6	-0.004	hc
C5	0.563	c3
H7	-0.214	h1
H8	-0.214	h1
O8	-0.963	o

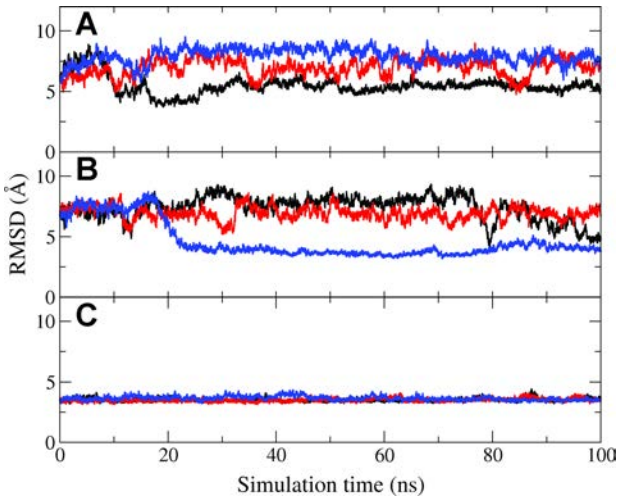

Figure 3-S1. Plots of RMSD relative to the $[Fe_4S_4]^{2+}_{(closed,\ HMBPP\text{-}bound)}$ IspH crystal structure (PDB ID: 3KE8) over the course of 3 x 100 ns aMD simulations of (A) $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$, (B) $[Fe_4S_4]^{2+}$/$HMBPP_{(open,docked)}$, and (C) $[Fe_4S_4]^{2+}$/$HMBPP_{(closed)}$ IspH.

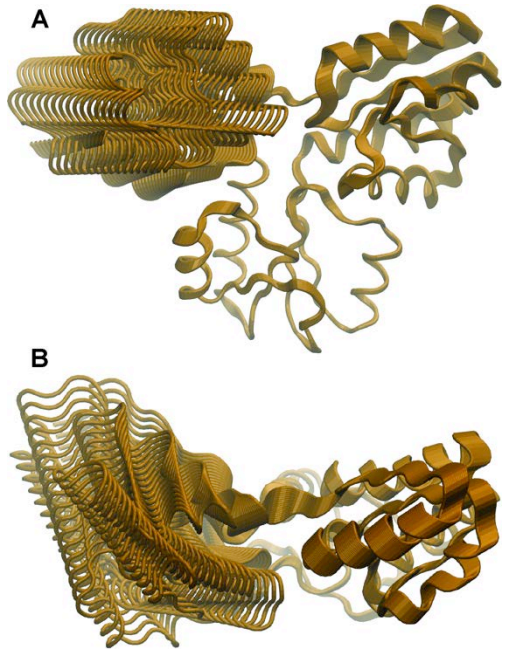

Figure 3-S2. Visualization of IspH motions along PC1 from (A) head-on toward the binding site and (B) from a top-view. As the principal components are constructed from both $[Fe_4S_4]^{2+}_{(open,substrate\text{-}free)}$ and $[Fe_4S_4]^{2+}$/$HMBPP_{(open,docked)}$ simulations, the dominant motion is the super-opening to closing of D3, with smaller closing motions of D2.

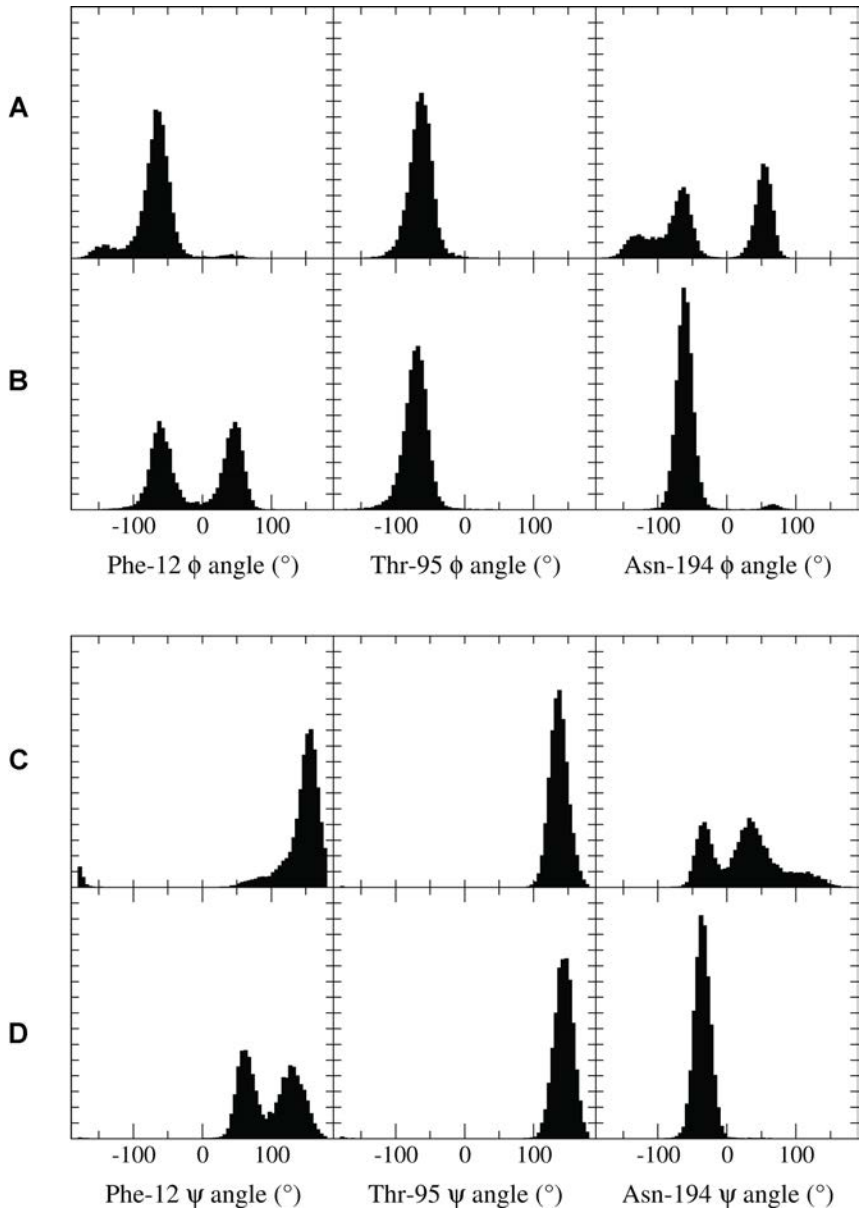

Figure 3-S3. Distributions of ϕ and ψ angles for Phe-12, Thr-95 and Asn-194 (the three having large KL_{res} values of 3.06, 0.77 and 0.32, respectively). (A) ϕ angle distributions in $[Fe_4S_4]^{2+}/HMBPP_{(open,docked)}$ simulations; (B) ϕ angle distributions in $[Fe_4S_4]^{2+}/HMBPP_{(closed)}$ simulations; (C) ψ angle distributions in $[Fe_4S_4]^{2+}/HMBPP_{(open,docked)}$ simulations; (D) ψ angle distributions in $[Fe_4S_4]^{2+}/HMBPP_{(closed)}$ simulations.

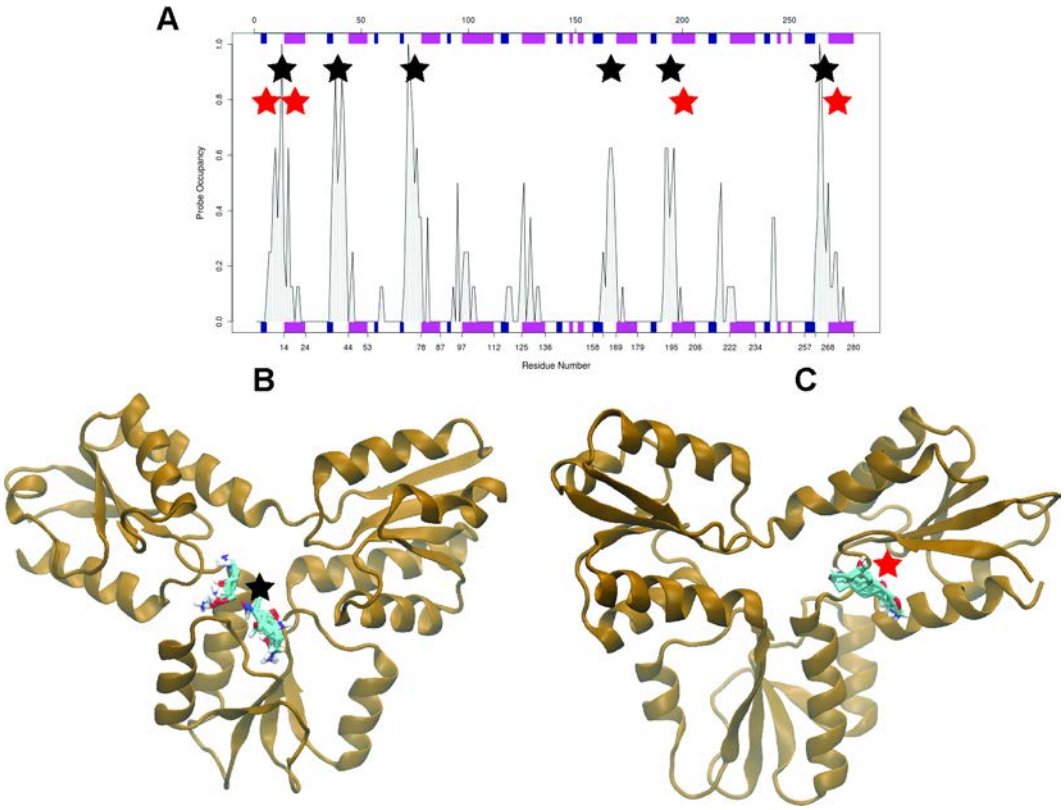

Figure 3-S4. (A) Plot of normalized FTMAP probe occupancy with respect to individual residues of IspH. Probes binding to the expanded substrate binding pocket (B) are marked by black stars, whereas probes that stick to the allosteric site (C), opposite the side of the substrate binding site, are marked by red stars.

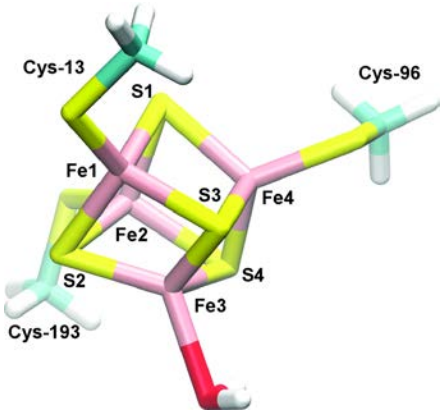

Figure 3-S5. Visual representation of the $[Fe_4S_4(SCH_3)_3OH_2]^{1-}$ model cluster utilized to obtain charges for the $[4Fe\text{-}4S]^{2+}$ cluster and its coordinating thiolate residues. Atom labels correspond to those accompanying charges in Table S1.

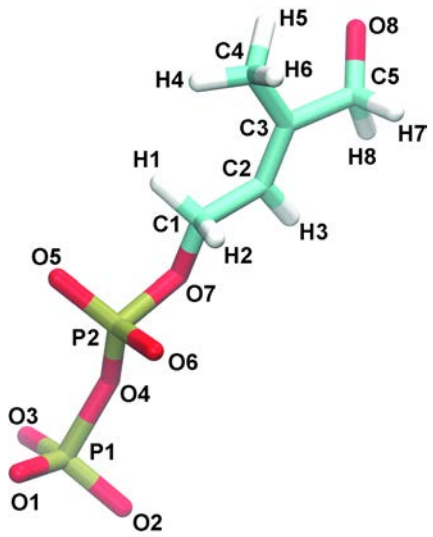

Figure 3-S6. Atom labels that correspond to the charges and atom types for the HMBPP molecule given in Table S3.

References

(1) Boucher, H. W.; Talbot, G. H.; Bradley, J. S.; Edwards, J. E.; Gilbert, D.; Rice, L. B.; Scheld, M.; Spellberg, B.; Bartlett, J. *Clin Infect Dis* **2009**, *48*, 1.

(2) Levy, S. B.; Marshall, B. *Nat Med* **2004**, *10*, S122.

(3) *World Malaria Report: 2012*, World Health Organization, 2012.

(4) Overbye, K. M.; Barrett, J. F. *Drug Discov Today* **2005**, *10*, 45.

(5) Payne, D. J.; Gwynn, M. N.; Holmes, D. J.; Pompliano, D. L. *Nat Rev Drug Discov* **2007**, *6*, 29.

(6) Renslo, A. R.; McKerrow, J. H. *Nat Chem Biol* **2006**, *2*, 701.

(7) Ridley, R. G. *Nature* **2002**, *415*, 686.

(8) Walsh, C. *Nat Rev Microbiol* **2003**, *1*, 65.

(9) Wells, T. N. C.; Alonso, P. L.; Gutteridge, W. E. *Nat Rev Drug Discov* **2009**, *8*, 879.

(10) Projan, S. J. *Curr Opin Microbiol* **2003**, *6*, 427.

(11) Eisenreich, W.; Bacher, A.; Arigoni, D.; Rohdich, F. *Cellular and Molecular Life Sciences* **2004**, *61*, 1401.

(12) Kuzuyama, T.; Seto, H. *Nat Prod Rep* **2003**, *20*, 171.

(13) Rohmer, M. *Nat Prod Rep* **1999**, *16*, 565.

(14) Rohmer, M. *Pure Appl Chem* **2007**, *79*, 739.

(15) Hale, I.; O'Neill, P. M.; Berry, N. G.; Odom, A.; Sharma, R. *Medchemcomm* **2012**, *3*, 418.

(16) Jomaa, H.; Wiesner, J.; Sanderbrand, S.; Altincicek, B.; Weidemeyer, C.; Hintz, M.; Turbachova, I.; Eberl, M.; Zeidler, J.; Lichtenthaler, H. K.; Soldati, D.; Beck, E. *Science* **1999**, *285*, 1573.

(17) Obiol-Pardo, C.; Rubio-Martinez, J.; Imperial, S. *Curr Med Chem* **2011**, *18*, 1325.

(18) Oldfield, E. *Accounts Chem Res* **2010**, *43*, 1216.

(19) Rohdich, F.; Bacher, A.; Eisenreich, W. *Bioorg Chem* **2004**, *32*, 292.

(20) Altincicek, B.; Duin, E. C.; Reichenberg, A.; Hedderich, R.; Kollas, A. K.; Hintz, M.; Wagner, S.; Wiesner, J.; Beck, E.; Jomaa, H. *Febs Lett* **2002**, *532*, 437.

(21) Rohdich, F.; Zepeck, F.; Adam, P.; Hecht, S.; Kaiser, J.; Laupitz, R.; Grawert, T.; Amslinger, S.; Eisenreich, W.; Bacher, A.; Arigoni, D. *P Natl Acad Sci USA* **2003**, *100*, 1586.

(22) Grawert, T.; Span, I.; Bacher, A.; Groll, M. *Angew Chem Int Edit* **2010**, *49*, 8802.

(23) Rekittke, I.; Wiesner, J.; Rohrich, R.; Demmer, U.; Warkentin, E.; Xu, W. Y.; Troschke, K.; Hintz, M.; No, J. H.; Duin, E. C.; Oldfield, E.; Jomaa, H.; Ermler, U. *J Am Chem Soc* **2008**, *130*, 17206.

(24) Grawert, T.; Rohdich, F.; Span, I.; Bacher, A.; Eisenreich, W.; Eppinger, J.; Groll, M. *Angew Chem Int Edit* **2009**, *48*, 5756.

(25) Grawert, T.; Kaiser, J.; Zepeck, F.; Laupitz, R.; Hecht, S.; Amslinger, S.; Schramek, N.; Schleicher, E.; Weber, S.; Haslbeck, M.; Buchner, J.; Rieder, C.; Arigoni, D.; Bacher, A.; Eisenreich, W.; Rohdich, F. *J Am Chem Soc* **2004**, *126*, 12847.

(26) Seemann, M.; Janthawornpong, K.; Schweizer, J.; Bottger, L. H.; Janoschka, A.; Ahrens-Botzong, A.; Tambou, M. N.; Rotthaus, O.; Trautwein, A. X.; Rohmer, M.; Schunemann, V. *J Am Chem Soc* **2009**, *131*, 13184.

(27) Wolff, M.; Seemann, M.; Bui, B. T. S.; Frapart, Y.; Tritsch, D.; Estrabot, A. G.; Rodriguez-Concepcion, M.; Boronat, A.; Marquet, A.; Rohmer, M. *Febs Lett* **2003**, *541*, 115.

(28) Grawert, T.; Span, I.; Eisenreich, W.; Rohdich, F.; Eppinger, J.; Bacher, A.; Groll, M. *P Natl Acad Sci USA* **2010**, *107*, 1077.

(29) Ahrens-Botzong, A.; Janthawornpong, K.; Wolny, J. A.; Tambou, E. N.; Rohmer, M.; Krasutsky, S.; Poulter, C. D.; Schunemann, V.; Seemann, M. *Angew Chem Int Edit* **2011**, *50*, 11976.

(30) Janthawornpong, K.; Krasutsky, S.; Chaignon, P.; Rohmer, M.; Poulter, C. D.; Seemann, M. *J Am Chem Soc* **2013**, *135*, 1816.

(31) Span, I.; Wang, K.; Wang, W. X.; Jauch, J.; Eisenreich, W.; Bacher, A.; Oldfield, E.; Groll, M. *Angew Chem Int Edit* **2013**, *52*, 2118.

(32) Wang, K.; Wang, W. X.; No, J. H.; Zhang, Y. H.; Zhang, Y.; Oldfield, E. *J Am Chem Soc* **2010**, *132*, 6719.

(33) Wang, W. X.; Li, J. K.; Wang, K.; Smirnova, T. I.; Oldfield, E. *J Am Chem Soc* **2011**, *133*, 6525.

(34) Wang, W. X.; Wang, K.; Liu, Y. L.; No, J. H.; Li, J. K.; Nilges, M. J.; Oldfield, E. *P Natl Acad Sci USA* **2010**, *107*, 4522.

(35) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J Med Chem* **2004**, *47*, 1739.

(36) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. *J Med Chem* **2004**, *47*, 1750.

(37) Maestro, In *Schrodinger Suite 2012*; version 9.3; Schrödinger, LLC: New York, NY, 2012.

(38) Xiao, Y. L.; Liu, P. H. *Angew Chem Int Edit* **2008**, *47*, 9722.

(39) Xiao, Y. L.; Zhao, Z. K.; Liu, P. H. *J Am Chem Soc* **2008**, *130*, 2164.

(40) Grant, B. J.; Rodrigues, A. P. C.; ElSawy, K. M.; McCammon, J. A.; Caves, L. S. D. *Bioinformatics* **2006**, *22*, 2695.

(41) Durrant, J. D.; de Oliveira, C. A. F.; McCammon, J. A. *J Mol Graph Model* **2011**, *29*, 773.

(42) McClendon, C. L.; Hua, L.; Barreiro, G.; Jacobson, M. P. *J Chem Theory Comput* **2012**, *8*, 2115.

(43) Fenley, A. T.; Muddana, H. S.; Gilson, M. K. *P Natl Acad Sci USA* **2012**, *109*, 20006.

(44) McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P. *J Chem Theory Comput* **2009**, *5*, 2486.

(45) Mirny, L. A.; Shakhnovich, E. I. *J Mol Biol* **1999**, *291*, 177.

(46) Shannon, C. E. *At&T Tech J* **1948**, *27*, 379.

(47) Frishman, D.; Argos, P. *Proteins-Structure Function and Genetics* **1995**, *23*, 566.

(48) Brenke, R.; Kozakov, D.; Chuang, G. Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. *Bioinformatics* **2009**, *25*, 621.

(49) Kumar, S.; Ma, B. Y.; Tsai, C. J.; Sinha, N.; Nussinov, R. *Protein Sci* **2000**, *9*, 10.

(50) Sullivan, S. M.; Holyoak, T. *P Natl Acad Sci USA* **2008**, *105*, 13829.

(51) Koshland, D. E. *P Natl Acad Sci USA* **1958**, *44*, 98.

(52) Boehr, D. D.; Nussinov, R.; Wright, P. E. *Nat Chem Biol* **2009**, *5*, 789.

(53) Bucher, D.; Grant, B. J.; McCammon, J. A. *Biochemistry-Us* **2011**, *50*, 10530.

(54) Silva, D. A.; Bowman, G. R.; Sosa-Peinado, A.; Huang, X. H. *Plos Comput Biol* **2011**, *7*, e1002054.

(55) Saint Martin, H.; Ortegablake, I.; Les, A.; Adamowicz, L. *Bba-Protein Struct M* **1994**, *1207*, 12.

(56) Irwin, J. J.; Raushel, F. M.; Shoichet, B. K. *Biochemistry-Us* **2005**, *44*, 12316.

(57) Baxter, E. L.; Zuris, J. A.; Wang, C.; Vo, P. L. T.; Axelrod, H. L.; Cohen, A. E.; Paddock, M. L.; Nechushtai, R.; Onuchic, J. N.; Jennings, P. A. *P Natl Acad Sci USA* **2013**, *110*, 948.

(58) ADF2009; SCM, Theoretical Chemistry, Vrije Universiteit: Amsterdam, The Netherlands, 2009.

(59) Noodleman, L.; Case, D. A. *Adv Inorg Chem* **1992**, *38*, 423.

(60) Sandala, G. M.; Noodleman, L. *Methods Mol Biol* **2011**, *766*, 293.

(61) Handy, N. C.; Cohen, A. J. *Mol Phys* **2001**, *99*, 403.

(62) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys Rev B* **1988**, *37*, 785.

(63) Gaussian 09, Revision A.1., Gaussian, Inc.: Wallingford, CT, 2009.

(64) Besler, B. H.; Merz, K. M.; Kollman, P. A. *J Comput Chem* **1990**, *11*, 431.

(65) Wang, J. M.; Wang, W.; Kollman, P. A.; Case, D. A. *J Mol Graph Model* **2006**, *25*, 247.

(66) AMBER 12, University of California, San Francisco: San Francisco, CA, 2012.

(67) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J Comput Chem* **2004**, *25*, 1157.

(68) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. *Nucleic Acids Res* **2007**, *35*, W522.

(69) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. *Nucleic Acids Res* **2004**, *32*, W665.

(70) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704.

(71) Price, D. J.; Brooks, C. L. *J Chem Phys* **2004**, *121*, 10096.

(72) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J Comput Phys* **1977**, *23*, 327.

(73) Miyamoto, S.; Kollman, P. A. *J Comput Chem* **1992**, *13*, 952.

(74) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712.

(75) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J Chem Phys* **1984**, *81*, 3684.

(76) Uberuaga, B. P.; Anghel, M.; Voter, A. F. *J Chem Phys* **2004**, *120*, 6363.

(77) Gotz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. *J Chem Theory Comput* **2012**, *8*, 1542.

(78) Darden, T.; York, D.; Pedersen, L. *J Chem Phys* **1993**, *98*, 10089.

(79) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J Chem Phys* **1995**, *103*, 8577.

(80) Christen, M.; Van Gunsteren, W. F. *J Comput Chem* **2008**, *29*, 157.

(81) Hamelberg, D.; Shen, T.; McCammon, J. A. *J Am Chem Soc* **2005**, *127*, 1969.

(82) Hamelberg, D.; de Oliveira, C. A. F.; McCammon, J. A. *J Chem Phys* **2007**, *127*, 155102.

(83) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J Chem Phys* **2004**, *120*, 11919.

(84) Pierce, L. C. T.; Salomon-Ferrer, R.; de Oliveira, C. A. F.; McCammon, J. A.; Walker, R. C. *J Chem Theory Comput* **2012**, *8*, 2997.

(85) Shao, J. Y.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. *J Chem Theory Comput* **2007**, *3*, 2312.

(86) Grossfield, A.; Zuckerman, D. M. *Annu Rep Comput Chem* **2009**, *5*, 23.

(87) Romo, T. D.; Clarage, J. B.; Sorensen, D. C.; Phillips, G. N., Jr. *Proteins* **1995**, *22*, 311.

(88) Russell, R. B.; Barton, G. J. *Proteins-Structure Function and Genetics* **1992**, *14*, 309.

(89) Humphrey, W.; Dalke, A.; Schulten, K. *J Mol Graph Model* **1996**, *14*, 33.

(90) Roberts, E.; Eargle, J.; Wright, D.; Luthey-Schulten, Z. *Bmc Bioinformatics* **2006**, *7*.

(91) Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013**, *29*, 845.

(92) de Oliveira, C. A. F.; Grant, B. J.; Zhou, M.; McCammon, J. A. *Plos Comput Biol* **2011**, *7*, e1002178.

(93) Guerra, C. F.; Snijders, J. G.; te Velde, G.; Baerends, E. J. *Theor Chem Acc* **1998**, *99*, 391.

(94) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Guerra, C. F.; Van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J Comput Chem* **2001**, *22*, 931.

(95) Torres, R. A.; Lovell, T.; Noodleman, L.; Case, D. A. *J Am Chem Soc* **2003**, *125*, 1923.

(96) Noodleman, L.; Han, W. G. *J Biol Inorg Chem* **2006**, *11*, 674.

(97) Giammona, D., University of California, Davis, 1984.

(98) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712.

(99) Becke, A. D. *J Chem Phys* **1993**, *98*, 5648.

(100) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J Phys Chem-Us* **1994**, *98*, 11623.

(101) Vosko, S. H.; Wilk, L.; Nusair, M. *Can J Phys* **1980**, *58*, 1200.

(102) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J Comput Chem* **2004**, *25*, 1157.

(103) Pierce, L.C.T.; Salomon-Ferrer, R.; de Oliveira, C.A.F.; McCammon, J.A.; Walker, R.C. *J. Chem. Theory Comput* **2012**, *8*, 2997-3002.

Chapter 4:

Measuring the Successes and Deficiencies of Constant pH Molecular Dynamics: A Blind Prediction Study

Abstract

A constant pH molecular dynamics method has been used in the blind prediction of pKa values of titratable residues in wild type and mutated structures of the Staphylococcal nuclease (SNase) protein. The predicted values have been subsequently compared to experimental values provided by the laboratory of García-Moreno. CpHMD performs well in predicting the pKa of solvent-exposed residues but The CpHMD method encounters difficulties in reaching convergence and predicting the pKa values for residues that are either located in the protein interior or are involved in strong interactions with neighboring residues. These results show the need to accurately and sufficiently sample conformational space in order to obtain pKa values consistent with experimental results.

Introduction

It is well established that the structure and function of a protein are highly dependent on the pH of its surrounding environment. The pKa of a titratable residue, which is heavily influenced by interactions with neighboring residues within a protein, governs the protonation state of that residue for a given solution pH. Changes in protonation state within a protein manifest as alterations to the charge distribution of the titratable residue, influencing the electrostatics of the protein environment. Protonation equilibria are thus closely linked with protein conformation, evidence of which is the sensitivity of proteins to denaturation at extreme pH.

The interplay between protonation state and protein conformation is not accounted for in conventional molecular dynamics (MD) simulations. Currently, these simulations employ fixed, predetermined protonation states for titratable residues, which are generally chosen according to the pKa value of the respective residue when isolated in solution. This method of protonation state assignment can be a severe approximation, as the pKa values of titratable residues are frequently shifted from that of the isolated residue in solution. Furthermore, protonation states are not constant, but rather exist in equilibria, subject to the changing electrostatic environment surrounding the titratable group. Therefore, incorporating pH as an input variable in MD simulations is highly desirable, as it would allow a more accurate study of pH-coupled protein dynamics, such as ligand binding and protein folding.

Over the past few decades, a number of theoretical methods have been developed to try to accurately determine the protonation states of titratable residues in

proteins. One class of methods utilizes static protein structures and employs a Poisson-Boltzmann approach for the calculation of electrostatics.[1-3] However, the use of static structures is thought to be a major contributor to discrepancies observed in the calculation of pKa shifts, as the conformational changes in the protein induced by change in residue protonation state are not taken into account. More recently, these methods have been improved by including descriptions of conformational variability, with adaptations to account for dielectric heterogeneity[4,5] and inclusion of conformational flexibility.[6-9] Notably, Warshel and co-workers were the first to employ MD methods to improve calculation of pKa values in proteins, with their electrostatic protein dipoles Langevin dipoles (PDLD) model.[10] Other groups have incorporated MD and QM/MM methods coupled with free energy perturbation techniques for pKa calculations.[11-13] A drawback of these techniques is their high reliance on the resolution of the input structure, which renders these methods incapable of calculating pKa shifts where protonation is accompanied by large conformational change.

Another class of methods incorporates the important coupling of conformation and protonation state through the use of computational simulations that employ pH as an external thermodynamic parameter.[14-26] These methods are often described as either continuous or discrete constant pH methods, contingent on how titratable protons are considered within the simulation. The former treats protonation state as a continuous titration parameter that advances simultaneously with the atomic coordinates of the system.[19-21] Originally, the implementation of this method used a mean-field

approximation, and protonation sites could exist as fractionally occupied. More recently, Lee *et al.* have developed methods to overcome issues with fractional protonation states, using λ-dynamics with an artificial titration barrier to discourage fractional protonation.[22] Extensions to the work of Lee *et al.* have incorporated proton tautomerism[23] and enhanced sampling methods to improve convergence.[24]

Discrete constant pH methods avoid non-physical intermediate charge states. These methods use MD simulations for conformational sampling, while sampling different discrete protonation states with periodic Monte Carlo (MC) steps interspersed throughout the MD trajectory.[14-18] The methods employed in this paper utilize the constant pH MD (CpHMD) method, originally developed by Mongan *et al*,[14] which uses generalized Born (GB) implicit solvent. Differences among these methods arise from choice of solvation model and protocols for updating protonation states within the simulation. While these methods have achieved good results for small protein systems, they can be computationally expensive, and long convergence times have been reported for systems with multiple titration sites. In an attempt to overcome these issues of convergence, use of enhanced sampling methods coupled with constant pH MD, such as constant pH accelerated MD (CpHaMD)[25] and constant pH replica-exchange MD (REX-CPHMD)[26] have been investigated. Results from simulations employing these methods indicate the increased sampling provides improvement over the conventional method.

The previous paragraphs provide only a brief summary of the computational methods available for pKa prediction. Further details of these and other methods can

be found in the literature, and several reviews have been published.[27-29] In this study, the successes and deficiencies of the CpHMD method have been investigated in the blind prediction of pKa values of titratable residues of the WT and mutant forms of the Staphylococcal nuclease (SNase) enzyme based upon comparison to experimental results released after submission to the pKa cooperative[30] by García-Moreno and co-workers.[31-36] Particular attention is paid to the differences in electrostatics and, consequently, acid/base properties of exterior and interior residues.

Theory

Constant pH molecular dynamics

CpHMD employs MD with GB implicit solvent.[14] Within the simulation, the MD simulation is periodically halted, and a MC step is taken, randomly considering a titratable residue for change in protonation. The transition energy corresponding to this MC step is evaluated according to Eq. 1:

$$\Delta G = k_B T\left(pH - pK_{a,ref}\right)\ln 10 + \Delta G_{elec} - \Delta G_{elec,ref} \quad (1)$$

which calculates pKa with respect to a reference compound for the residue of interest. Reference compounds are the isolated titratable residues solvated in water (reference pKa values are 3.8 for ASP, 4.3 for GLU, 6.8 for HIS, 9.6 for TYR, and 10.5 for LYS).[14,37,38] In equation 1, k_B is the Boltzmann constant, T is the temperature, pH is the specified solvent pH, $pK_{a,ref}$ is the pKa of the reference compound, ΔG_{elec} is the

electrostatic energy change for protonation state change of the titratable residue, and $\Delta G_{elec,ref}$ is the corresponding electrostatic transition energy for the reference compound. The same GB electrostatics employed in the MD is used for calculating this transition energy, with acceptance of the change in protonation determined by the Metropolis criterion. If the MC move is accepted, the protonation state of the residue will change to the new state, and MD is continued. If not, the simulation will continue with the residue remaining in the unchanged protonation state. CpHMD has been successfully applied in the pKa prediction of titratable residues in the Hen Egg White Lysozyme (HEWL) enzyme.[14]

Titration curve construction and pKa calculation

The predicted pKa values are calculated from performing CpHMD simulations over a range of solution pH values. Assuming the system is ergodic, we further assume fractional protonation is given by the amount of time a particular titratable residue spends in its protonated state.[15] Thus the fraction of deprotonated species, *s*, for a residue at a specific pH value can be used to predict the pKa from a Hill plot:[16,20,22,39]

$$(2) \qquad s = \frac{1}{1+10^{n(pK_a - pH)}} \ ,$$

Fits to this curve allow for estimation of both the pKa value as a midpoint of titration, as well as the Hill coefficient, *n*, which describes the cooperativity of various sites

with respect to titration.[40] Illustrated in ref. 39, for example, the usefulness of the Hill equation resides in its ability to provide a good prediction of the midpoint pKa value, even when the fit is inaccurate at the tails of the titration curve.[39]

Methods

Test system: Staphylococcal nuclease

Staphylococcal nuclease (SNase) is a highly charged protein, which has generated difficulty in obtaining accurate structure-based pKa predictions.[41] The structures of the wild-type and mutant proteins of the *SNase* system are provided for this study by the lab of Garcia-Moreno *et al.,* who measured the pKa shifts of the titratable residues using NMR spectroscopy.[31-36] Along with other computational groups, we have computed blind pKa predictions for residues of wild-type SNase (PDB ID: 1STN or 1SNC), the SNase mutant Δ + PHS (PDB ID: 3BDC), and various mutants from the Δ + PHS parent protein (referred to as calculated results in this study). Δ + PHS is unique in that it is a hyperstable, acid-resistant SNase mutant with five substitutions (G50F, V51N, P117G, H124L and S128A) and a deletion of residues 44 - 49.[31,32] Garcia-Moreno directed this effort, holding experimental pKa determinations from those making predictions and picking residues of interest for pKa prediction (hereafter referred to as experimental results in this study). In total, approximately 93 structures of the wild-type (WT) and mutant SNase have been provided.[30] Owing to time constraints, however, our CpHMD pKa predictions have not been carried out on the entire set of provided structures; the subset of which were studied and submitted as blind predictions are shown in Table 4-1.

CpHMD simulations

The standard CpHMD method has been implemented in AMBER10 molecular dynamics program. All simulations are conducted with the AMBER99SB force field[42] and the GB solvent model igb=2,[43-45] using a 30 Å cutoff value for nonbonded interactions and computation of effective Born radii calculations. Similar to experimental conditions, salt concentrations are set to either 0.1 or 1.0 M. The SHAKE algorithm constrains all bonds involving hydrogen with a time step of 2 fs,[46] and temperature is maintained at 300 K using the Berendsen temperature coupling method with a time constant of 2 ps.[47] A period of 10 fs of MD separates the MC trials. With these parameters, a 10 ns CpHMD simulation takes approximately 72 hours using 16 Xeon X5650 2.67GHz processors.

All simulations begin from the crystal structure coordinates of the WT and Δ+PHS SNase systems provided by Garcia-Moreno *et al*, from which specific titratable residues were chosen for the blind prediction study (Table 4-1). For the blind predictions performed on SNase systems where a single ASP or GLU residue has been highlighted as the residue of interest, CpHMD simulations of 10 ns in length have been performed in the solution pH range 2.0-7.0 at 0.5 pH unit intervals, titrating only acidic residues. For these simulations, HIS residues are allowed to titrate from pH 4.5 to pH 7.0. In systems where a LYS or TYR is highlighted as the residue of interest, simulations have been carried out in the pH range 7 to 10.5, where HIS, LYS and TYR residues are set to titrate. The exclusion of HIS residues from the most acidic simulations is justified, as the pKa of the HIS reference is around 6 - 7.[38] In most

cases, it is safe to assume all ASP and GLU residues are deprotonated above pH 7 and all LYS and TYR are protonated below pH 7, allowing exclusion of these residues from titration in these respective pH regions. Models for the terminal residues have not yet been developed for this system, so these residues are set to their most likely protonation states at neutral pH, with the N-terminus protonated and the C-terminus deprotonated. All non-titrating residues are set to their expected protonation states.

Simulations conducted after publishing of experimental results

To understand why some predictions fail to reproduce the experimental results, further CpHMD simulations have been conducted, as indicated in the proceeding sections. In particular, the pH range at which simulations were originally performed is extended to account for residues that deviate the most from their reference value. In cases where convergence has been determined to be problematic, extended simulations (>10 ns) do not appear to improve predictions (results not shown).

Results

Titration curves

Titration curves are obtained from CpHMD simulations for 32 titratable residues of the WT, Δ+PHS, and mutant Δ+PHS SNase systems. The experimental pKa values for the WT protein were published prior to the predictions, but those for the mutant proteins were withheld until blind predictions were made (Table 4-1). From Eq. 2, pKa values are calculated along with the standard errors of regression for curve fits to the Hill equation in Eq. 2 (Table 4-1).[48] In the instances of ASP21 and L37D,

pKa values cannot be computed due to the lack of transitions between protonated and deprotonated forms.

A representative plot of calculated pKa over time is given in Figure 4-1 for both a surface residue for which the CpHMD predicts pKa accurately (Δ+PHS, GLU52) and the Δ+PHS mutant L36D, for which the pKa prediction of the interior residue ASP36 deviates by more than three pKa units from the experimental result. In the former case, the pKa converges rapidly, whereas ASP36 in Δ+PHS L36D is indicated to not achieve convergence over the duration of the simulations.

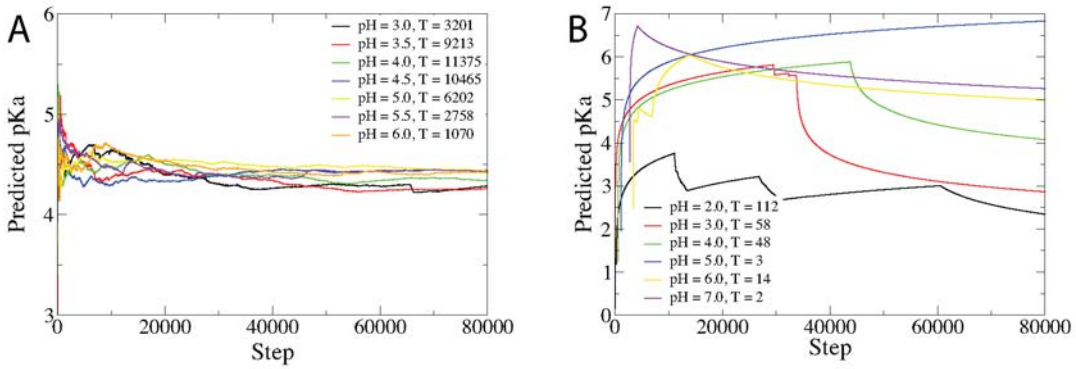

Figure 4-1. Plots of predicted pKa over the duration of CpHMD simulations for (a) Δ+PHS GLU52 (experimental pKa = 3.93)[31] and (b) Δ+PHS L36D (experimental pKa = 7.90).[31] The number of protonation state transitions (T) are given in the figure legend for each system.

In assessing the convergence of a system, it is interesting to observe the trend in the number of transitions between protonated and deprotonated states as a function of pH. In systems that are well converged (e.g. Δ+PHS, GLU52), the greatest number of transitions between deprotonated and protonated states within the CpHMD scheme

are found for the simulation conducted at a pH nearest to the calculated pKa value. Simulations conducted at pH values far from the predicted pKa encounter fewer transitions between protonation states, as is to be expected from the acceptance criteria defined in Eq. 1. This is not the case for certain systems (e.g. Δ+PHS L36D), where convergence is a problem. Thus the presence of a clear distribution of transitions across the different pH values simulated, peaked at the pH nearest the predicted pKa, may be an indicator of how well converged the system is.

Experimental validation

Following the blind predictions, García-Moreno and coworkers have released experimental results for comparison to predicted pKa values (Table 4-1).[30] In summary, CpHMD simulations calculate the pKa values of 17 residues to within 1 pKa unit of the experimental value, 9 residues within 2 units, 1 residue within 3 units, and 2 residues within 4 units. Residues in the Δ+PHS protein chosen for analysis are surface residues. All of these residues remain solvent-exposed throughout the CpHMD simulation and are generally well predicted with respect to experiment (Table 4-1). From Figure 4-2, it is clear that the predictions with the largest deviation from the experimental values are for Δ+PHS variants that have residues located within the hydrophobic interior of the protein (e.g. L37D, L36D). These residues have also been found experimentally to have the largest shifts in pKa from their reference values (Table 4-1). Errors with respect to the experimental pKa are shown in Table 4-1, with predictions within ranges of experimental pKa considered to have zero error.

Table 4-1. Predicted and experimental values for various residues from the WT SNase, Δ+PHS and Δ+PHS mutant proteins.[30,31] The difference between experimental and model compound pKa values for *Δ+PHS* mutants. Model compound pKa values: 3.8 (ASP), 4.3 (GLU), 6.8 (HIS) and 10.4 (LYS).[14,37,38]

Protein	Residue	Experimental pKa	Predicted pKa	(Pred. – Exp) pKa offset	(Pred. – Model) pKa offset
WT	**HIS8**	6.52	5.67 ± 0.04	-0.85	-1.1
	HIS46	5.86	6.8 ± 0.3	0.7	0.0
	HIS121	5.30	7.0 ± 0.1	1.7	0.2
	HIS124	5.73	6.0 ± 0.1	0.3	-0.8
Δ+PHS	**ASP19**	2.21	4.1 ± 1.1	0.9	0.3
	ASP21	6.54	-	-	-
	ASP40	3.87	3.1 ± 0.1	-0.8	0.7
	ASP77	<2.2	3.6 ± 0.2	>1.1	-0.2
	ASP83	<2.2	2 ± 8	-	-2
	ASP95	2.16	3.6 ± 0.1	1.4	-0.2
	GLU10	2.82	4.4 ± 0.2	1.6	0.1
	GLU43	4.32	1 ± 5	-0.9	-3
	GLU52	3.93	4.3 ± 0.2	0.4	0.0
	GLU57	3.49	4.3 ± 0.1	0.8	0.0
	GLU67	3.76	4.39 ± 0.03	0.6	0.09
	GLU73	3.31	4.2 ± 0.1	0.9	-0.1
	GLU75	3.26	4.0 ± 0.1	0.7	-0.3
	GLU101	3.81	3.5 ± 0.2	-0.3	-0.8
	GLU122	3.89	3.8 ± 0.1	-0.1	-0.5
	GLU129	3.75	4.28 ± 0.04	0.6	-0.02
	GLU135	3.76	4.2 ± 0.1	0.4	-0.1
F34E	**GLU34**	7.30	5.9 ± 0.1	-1.4	1.6
F34K	**LYS34**	7.10	2 ± 5	-5	-3
G20D	**ASP20**	<4.0	2 ± 2	-2	-2
G20E	**GLU20**	<4.5	4.1 ± 0.3	-	-0.2
G20K	**LYS20**	>10.4	8.6 ± 0.2	< -1.8	-1.8
L25D	**ASP25**	6.80	4.8 ± 0.3	-2.0	1.0
L36D	**ASP36**	7.90	5 ± 3	-3	1
L37D	**ASP37**	<4.0	-	-	-
V23D	**ASP23**	6.8	3 ± 2	-4	-1
V23E	**GLU23**	7.1	6.4 ± 0.1	-0.7	2.1
V23K	**LYS23**	7.40	7.3 ± 0.6	-0.1	-3.1

As described in Table 4-1, CpHMD simulations have correctly predicted the experimental trends for the majority of these buried residues, three within 1 pKa unit of the experimental result (G20E, V23E and V23K). The simulations predict the pKa values of F34E/K, L25D, L36D, and V23E/K to be shifted from their model values in

the direction of favoring the neutral residue at physiological pH; although, for some of these residues, the shift in the predicted pKa is not as large as that found experimentally (Table 4-1).

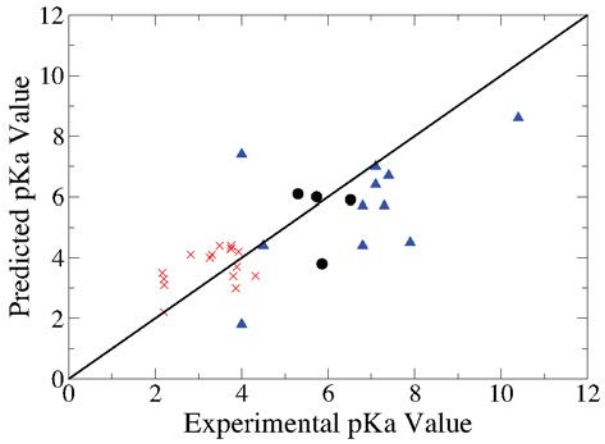

Figure 4-2. Plot of predicted versus experimental pKa values for WT SNase (•), Δ+PHS (x, exterior residues), and Δ+PHS mutants (Δ, internal residues).[31-36] The line y = x represents accurate prediction of the experimental pKa.

Discussion

Since the release of experimental results, further simulations have been carried out to investigate why our methods predict pKa values that deviate more than 1 pKa unit from experimental results. For this paper, we have chosen a selection of residues that illustrate problems with the application of the CpHMD method to these specific systems.

Δ+PHS: ASP21

In the Δ+PHS mutant, the residue ASP21 is a notable exception to the good performance of CpHMD in predicting pKa values of surface residues. This problem

arises from a lack of transitions between protonated and deprotonated states. In this case, longer simulations fail to alleviate the problem, likely due to the existence of a strong, charged hydrogen bond interaction between ASP19 and ASP21 preventing changes in protonation state from occurring. Consistent with our results, García-Moreno and co-workers have needed to apply two-site binding isotherms to properly describe the experimental titration of these interacting residues and have also noted the difficulty in predicting the pKa for ASP21 computationally.[31] Similar problems arise in the simulation of L37D, indicating that sampling of protonation states is critical to the performance of the CpHMD method. Use of enhanced sampling techniques to allow the system to sample other protonation states may be necessary for accurate pKa predictions in conventional simulations where strong interactions persist.

Δ+PHS: G20K

For the mutant Δ+PHS G20K, the CpHMD method predicts a pKa of 8.6, nearly two pKa units lower than the experimental value (>10.4).[33] This lysine residue sufficiently sampled protonation space, encountering more than 600 transitions over the duration of each simulation. The trajectories of these simulations incur large motions indicative of protein instability. The root mean square distances (RMSD) with respect to the starting structure for these simulations do not converge at any solution pH, largely influenced by the winding and helical motion of the last 20 residues of C-terminus (Figure 4-3). To further probe this conformational change, we performed conventional MD simulations with set protonation states computed by the program PROPKA[49] for pH 7-10 at intervals of 0.5. The conventional MD simulations

similarly suffer from protein instability near neutral pH (pH 7 and pH 8), although take longer to encounter it than CpHMD simulations. At higher pH, the terminal helix in the G20K protein does not incur the same motion observed at neutral pH and in CpHMD simulations. These simulations show the sensitivity of the G20K protein toward change in protonation state, and in order to achieve results closer to experiment with the CpHMD method, it may be necessary to spatially constrain the termini. These findings indicate that although increased sampling is desirable and may be achieved in certain systems, it is important that the correct conformational space is sampled in order to attain an accurate prediction of pKa.

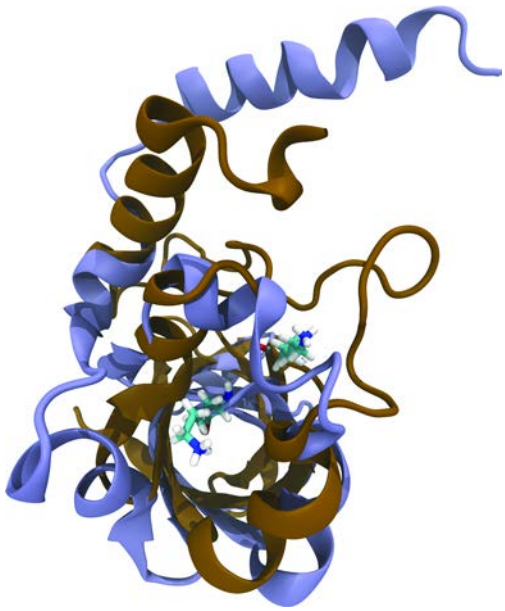

Figure 4-3. Conformational change encountered by the Δ+PHS G20K protein at the start (copper) and end (purple) of CpHMD simulation performed at pH = 8.5.

In further probing the problems involving predicting the pKa for G20K, it is noteworthy that other mutations at site 20 generate similar instabilities (e.g. G20D and G20E). Having an acidic residue at site 20, however, does not affect the pKa prediction to the same extent. Visual analysis of trajectories for the G20D protein reveals that hydrogen bonds from Thr-29 persist throughout the simulations, likely lowering its pKa (Table 4-1). Similarly, G20E forms transient hydrogen bonds with Thr-29. All mutated residues at site 20 sample conformational space that is solvent-exposed, in addition to time spent buried in the protein interior. While this explains the propensities for G20D and G20E to exist in their charged states, it fails to explain the shift in pKa for G20K.

Δ+PHS: F34E

CpHMD simulations performed on the Δ+PHS F34E mutant consistently obtain a predicted pKa (5.7) lower than experiment (7.30).[32] The stability of this particular mutant shows great sensitivity to the pH of the simulation, with large conformational changes occurring at acidic pH (Figure 4-4). Nevertheless, the pKa values calculated at neutral pH—closer to the pKa of the residue—still underestimate the experimental pKa despite undergoing a large number of transitions between protonated and deprotonated states. Upon visualization of this structure, it is notable that the carboxylate of GLU34 forms salt bridges with an adjacent arginine residue (ARG81), which causes this residue to favor its deprotonated state. The simulations may not sample enough conformational space owing to the persistence of this salt bridge, therefore leading to a predicted pKa value lower than the experimental result.

Enhanced sampling techniques may provide the means to allow the system to escape this GLU34-ARG81 salt bridge and give a more representative prediction of pKa.

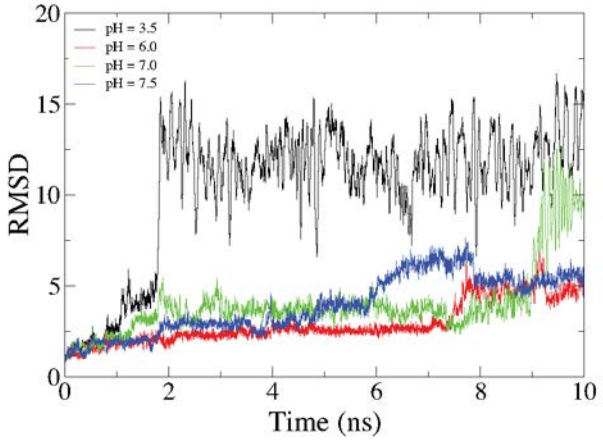

Figure 4-4. RMSD (Å) as a function of MD time step for the Δ+PHS mutant F34E protein at varying pH values.

Δ+PHS: L36D

The mutant L36D suffers from sampling problems, both of conformational and protonation space (Figure 4-1). While at certain pH values CpHMD simulations correctly predict the pKa, which experimentally is found to be 7.90,[31] there is no clear trend in pKa prediction for simulations conducted at different levels of pH (Figure 4-1). From visualization of the various trajectories, it is suggested that ASP36 may form a strong hydrogen bond with ASP21 in the MD simulations, stabilizing the deprotonated form. This scenario is seen at pH 4.5, where the simulation more accurately predicts a pKa of 7.4. In other cases, ASP36 becomes buried in the hydrophobic interior of the protein, again leading to insufficient sampling of different

protonation states. It is therefore likely that L36D needs to better sample conformational space in order to more effectively predict the pKa of ASP36.

Analysis of CpHMD performance

It is clear the CpHMD method performs better at predicting the pKa values of solvent-exposed residues, which possess pKa values closer to their reference compounds (Table 4-1). This is evident from the calculation of the root mean square error (RMSE) of predicted pKa values, measured against the experimental work of García-Moreno to quantify this result, showing that residues on the surface of Δ+PHS deviate from experiment with an RMSE of 1.23, whereas the RMSE for interior residues of the various Δ+PHS mutants is 2.42 (Table 4-2).[31] Most residues found at the surface of the protein encounter an increased number of transitions between protonated and deprotonated forms, and tend to converge relatively quickly (~6-8 ns). The counterexample to this trend is ASP21, which likely fails to transition due to sampling problems derived from the persistence of its hydrogen bond with ASP19.

Table 4-2. RMS errors of predicted pKa values against experimental values for residues located in different regions of the Δ+PHS protein (exterior residues) and Δ+PHS mutants (interior residues).[31] Errors are computed with zero error if the predicted pKa falls within the bounds of experimental pKa with limiting values. Residues that do not incur transitions (ASP21 and L37D) are omitted from this calculation.

	Surface	**Interior**
All Residues	1.23	2.42
Aspartates	1.22	2.59
Glutamates	1.23	0.90
Lysines	-	3.12

The importance of selecting a suitable pH range for titration and difficulties in achieving proper sampling of conformational space are illustrated in some pKa predictions of interior residues for the various Δ+PHS mutants. Given that many internal residues are found experimentally to have pKa values shifted considerably from their reference pKa, it is thus important to set up simulations over a wide pH range to conduct the titration. For example, in the case of L36D, simulations were performed at acidic pH under the assumption that the pKa of the aspartic acid would exist closer to its reference value of 3.8. In fact, experimental results show this residue to titrate at a pKa of 7.80. The selection of the pH range is also important for the stability of the system when performing CpHMD simulations, as illustrated by Δ+PHS F34E.

Analyses of computed pKa over time show internal residues to be far less converged compared to surface residues, making the prediction of accurate pKa values more challenging. While the CpHMD method applied in this study usually predicts the direction of the pKa shift from the reference compounds correctly, there is still room for improvement in accurately predicting pKa for internal residues.

Residues buried within the protein environment experience dielectric environments quite different from those at the surface of the protein, with their pKa properties very susceptible to the nature of the residues in their vicinity and thus more difficult to treat computationally.[31,41] This difficulty is illustrated by the Δ+PHS L36D and F34E proteins, where strong hydrogen bonds or salt bridges involving these titratable residues affect their protonation equilibria. Simulations of L36D do not

contain any transitions between deprotonated and protonated forms owing to the persistence of an interaction between the ASP36 and ASP21 residues. For residues such as this, the use of an enhanced sampling method, such as accelerated MD, may assist in the sampling of relevant conformations and thus protonation states. The requirement for increased sampling is also highlighted in instances where salt bridges persist throughout the simulation, as in the case of ARG81-GLU34 in the F34E mutant protein. The CpHMD method severely under-predicts the pKa of this glutamate, suggesting it spends more time in its deprotonated form than experiment predicts.[31] It is possible, although not proven in our studies, that the strength of salt bridges sampled in our CpHMD method is overestimated under the GB implicit solvation, thus leading to error in predicting protonation state.[50]

Although not specifically quantified in this study, errors likely exist in CpHMD simulations due to the use of implicit solvation and conventional (non-polarizable) force fields. With regards to implicit solvation, issues regarding global protein movements would likely be dampened by the presence of explicit solvent molecules. Despite this generally accepted point, we believe CpHMD simulations employing implicit solvation still merit further study due to the simplicity of protonation changes and transition energy calculations. With regards to force fields, polarizable force fields would likely better capture the sensitivity of neighboring groups to changes in the protonation state. The topic of force field effects on constant pH MD simulations is further investigated by others in this issue.[51]

While there exist problems with both implicit solvation and conventional force fields, the CpHMD method has been successful in predicting the pKa of a significant number of residues from the test set from García-Moreno. This study has highlighted areas that may add significant improvement in the pKa prediction capability of the method, such as enhanced conformational sampling and implementation of an improved solvation model. Future work will focus on testing other solvation models and the implementation of different accelerated molecular dynamics techniques, with the goal of achieving better sampling of physically meaningful conformations and protonation states.

Acknowledgements

This work was supported in part by the NSF, NIH, HHMI, CTBP, NBCR and the NSF supercomputer centers.

Chapter 4 is a minimally modified reprint of the material as it appears in Sarah L. Williams, Patrick G. Blachly, and J. Andrew McCammon, "Measuring the Successes and Deficiencies of Constant pH Molecular Dynamics: A Blind Prediction Study," *Proteins: Structure, Function, and Bioinformatics*, 2011. The dissertation author was one of two primary investigators, who each contributed equally to performing research and writing this paper.

References

(1) Antosiewicz, J.; Mccammon, J. A.; Gilson, M. K. *J Mol Biol* **1994**, *238*, 415.

(2) Bashford, D.; Karplus, M. *Biochemistry-US* **1990**, *29*, 10219.

(3) Yang, A. S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. *Proteins* **1993**, *15*, 252.

(4) Mehler, E. L.; Guarnieri, F. *Biophys J* **1999**, *77*, 3.

(5) Wisz, M. S.; Hellinga, H. W. *Proteins* **2003**, *51*, 360.

(6) Archontis, G.; Simonson, T. *Biophys J* **2005**, *88*, 3888.

(7) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. *Biophys J* **2002**, *83*, 1731.

(8) Pokala, N.; Handel, T. M. *Protein Sci* **2004**, *13*, 925.

(9) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J Am Chem Soc* **1990**, *112*, 6127.

(10) Sham, Y. Y.; Chu, Z. T.; Warshel, A. *J Phys Chem B* **1997**, *101*, 4458.

(11) Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. *J Phys Chem A* **2005**, *109*, 6634.

(12) Merz, K. M. *J Am Chem Soc* **1991**, *113*, 3572.

(13) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H. B.; Ghosh, N.; Prat-Resina, X.; Konig, P.; Li, G. H.; Xu, D. G.; Guo, H.; Elstner, M.; Cui, Q. *J Phys Chem B* **2006**, *110*, 6458.

(14) Mongan, J.; Case, D. A.; McCammon, J. A. *J Comput Chem* **2004**, *25*, 2038.

(15) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. *J Chem Phys* **2002**, *117*, 4184.

(16) Machuqueiro, M.; Baptista, A. M. *Proteins-Structure Function and Bioinformatics* **2008**, *72*, 289.

(17) Dlugosz, M.; Antosiewicz, J. M. *Chem Phys* **2004**, *302*, 161.

(18) Dlugosz, M.; Antosiewicz, J. M.; Robertson, A. D. *Phys Rev E* **2004**, *69*.

(19) Baptista, A. M.; Martel, P. J.; Petersen, S. B. *Proteins* **1997**, *27*, 523.

(20) Borjesson, U.; Hunenberger, P. H. *J Chem Phys* **2001**, *114*, 9706.

(21) Borjesson, U.; Hunenberger, P. H. *J Phys Chem B* **2004**, *108*, 13551.

(22) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. *Proteins-Structure Function and Bioinformatics* **2004**, *56*, 738.

(23) Khandogin, J.; Brooks, C. L. *Biophys J* **2005**, *89*, 141.

(24) Khandogin, J.; Chen, J. H.; Brooks, C. L. *Proc Natl Acad Sci USA* **2006**, *103*, 18546.

(25) Williams, S. L.; de Oliveira, C. A. F.; McCammon, J. A. *J Chem Theory Comput* **2010**, *6*, 560.

(26) Meng, Y. L.; Roitberg, A. E. *J Chem Theory Comput* **2010**, *6*, 1401.

(27) Baker, N. A.; Bashford, D.; Case, D. A. In *New Algorithms for Macromolecular Simulation*; Leimkuhler, B., Chipot, C., Elber, R., Laaksonen, A., Mark, A., Schlick, T., Schutte, C., R., S., Eds.; Springer: New York, 2006; Vol. 49, p 263.

(28) Chen, J. H.; Brooks, C. L.; Khandogin, J. *Curr Opin Struc Biol* **2008**, *18*, 140.

(29) Mongan, J.; Case, D. A. *Curr Opin Struc Biol* **2005**, *15*, 157.

(30) http://amylase.ucd.ie/pKacoop/

(31) Castaneda, C. A.; Fitch, C. A.; Majumdar, A.; Khangulov, V.; Schlessman, J. L.; Garcia-Moreno, B. E. *Proteins-Structure Function and Bioinformatics* **2009**, *77*, 570.

(32) Isom, D. G.; Cannon, B. R.; Castaneda, C. A.; Robinson, A.; Bertrand, G. M. E. *Proc Natl Acad Sci USA* **2008**, *105*, 17784.

(33) Isom, D. G.; Castaneda, C. A.; Velu, P. D.; Garcia-Moreno, B. *Proc Natl Acad Sci USA* **2010**, *107*, 16096.

(34) Isom, D. G.; Castaneda, C. A.; Cannon, B. R.; Garcia-Moreno, B. E. *Proc Natl Acad Sci USA* **2011**, *108*, 5260.

(35) Harms, M. J.; Schlessman, J. L.; Sue, G. R.; Garcia-Moreno, B. *Proc Natl Acad Sci USA* **2011**, *108*, 18954.

(36) Chimenti, M. S.; Khangulov, V. S.; Robinson, A. C.; Heroux, A.; Majumdar, A.; Schlessman, J. L.; Garcia-Moreno, B. *Structure* **2012**, *20*, 1071.

(37) Bashford, D.; Case, D. A.; Dalvit, C.; Tennant, L.; Wright, P. E. *Biochemistry-US* **1993**, *32*, 8045.

(38) Kyte, J. *Structure in Protein Chemistry*; Garland Publishing Inc.: New York, 1995.

(39) Onufriev, A.; Case, D. A.; Ullmann, G. M. *Biochemistry-US* **2001**, *40*, 3413.

(40) Cantor, C. R.; Schimmel, P. R. *Biophysical Chemistry: Part III The Behavior of Biological Macromolecules*; W.H. Freeman and Co.: San Francisco, 1980.

(41) Fitch, C. A.; Whitten, S. T.; Hilser, V. J.; Garcia-Moreno, B. *Proteins-Structure Function and Bioinformatics* **2006**, *63*, 113.

(42) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J Comput Chem* **2000**, *21*, 1049.

(43) Onufriev, A.; Case, D. A.; Bashford, D. *J Comput Chem* **2002**, *23*, 1297.

(44) Onufriev, A.; Bashford, D.; Case, D. A. *J Phys Chem B* **2000**, *104*, 3712.

(45) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins-Structure Function and Bioinformatics* **2004**, *55*, 383.

(46) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *Journal of Computational Physics* **1977**, *23*, 327.

(47) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J Chem Phys* **1984**, *81*, 3684.

(48) MATLAB, Version 7.11.0; The MathWorks Inc.: Natick, Massachusetts, 2010.

(49) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins-Structure Function and Bioinformatics* **2005**, *61*, 704.

(50) Chen, J. H.; Im, W. P.; Brooks, C. L. *J Am Chem Soc* **2006**, *128*, 3728.

(51) Machuqueiro, M.; Baptista, A. M. *Proteins-Structure Function and Bioinformatics* **2011**, *79*, 3437.

Chapter 5:

Protocols Utilizing Constant pH Molecular Dynamics to Compute pH-Dependent Binding Free Energies

Abstract

In protein-ligand binding, the electrostatic environments of the two binding partners may vary significantly in bound and unbound states, which may lead to protonation changes upon binding. In cases where ligand binding results in a net uptake or release of protons, the free energy of binding is pH-dependent. Nevertheless, conventional free energy calculations and molecular docking protocols typically do not rigorously account for changes in protonation that may occur upon ligand binding. To address these shortcomings, we present a simple methodology based on Wyman's binding polynomial formalism to account for the pH dependence of binding free energies and demonstrate its use on cucurbit[7]uril (CB[7]) host-guest systems. Using constant pH molecular dynamics and a reference binding free energy that is taken either from experiment or from thermodynamic integration computations, the pH-dependent binding free energy is determined. This computational protocol accurately

captures the large pK_a shifts observed experimentally upon CB[7]:guest association and reproduces experimental binding free energies at different levels of pH. We show that incorrect assignment of fixed protonation states in free energy computations can give errors of > 2 kcal/mol in these host-guest systems. Use of the methods presented here avoids such errors, thus suggesting their utility in computing binding free energies for protein-ligand complexes.

Introduction

The changes in the electrostatic environment that accompany binding of small molecules, nucleic acids, or other proteins may thus induce changes in the protonation states of titratable groups in the protein.[1-8] Recently, Aguilar *et al*. conducted a computational survey of various protein-protein, protein-small molecule, and protein-nucleic acid complexes to ascertain the prevalence of protonation change in the protein receptor upon biomolecular association. Notably, in 60 percent of the protein-small molecule complexes considered, at least one titratable residue in the protein was found to assume different protonation states in its free and bound states.[9] Furthermore, protonation changes that accompany small molecule binding to proteins are not limited to the protein partner: an estimated 60-80 percent of orally-administered drugs are weak acids or bases, whose protonation states can also be tuned by the cellular pH and electrostatic environment of their protein binding partners.[10-13] In cases where protein-ligand binding accompanies a net transfer of protons to either binding partner, the binding process is pH-dependent, *i.e.* the observed binding free energy is a function of pH.

Conventionally, both computational docking and more rigorous free energy computations, such as the thermodynamic integration (TI) and free energy perturbation (FEP) methods, employ fixed protonation states that are identical for free and bound states in the computation of binding affinities. Clearly, in cases where ligand binding is linked to the (un)binding of protons, such approximations will lead to error. Improper assignment of protonation states in binding free energy computations may result in significant errors, making correct assignment of pK_a and protonation state essential to obtaining accurate free energies.

Simulations of protein-ligand systems are typically preceded by the assignment of fixed protonation states to titratable groups on the two binding partners, often using programs such as H++[14-16] and PROPKA[17-20] to do so. Further, docking studies often employ empirical prediction algorithms, which often use Hammett and Taft relations to assign fixed protonation states to the free ligands being docked.[21,22] These approaches, however, fail to account for changes in protonation that may follow from the altered electrostatic environment surrounding the two binding partners upon complex formation. Several computational methods, however, have been developed that permit the protonation of titratable residues to respond to changes in the electrostatic environment.[2,23-28] For instance, various flavors of constant pH molecular dynamics (CpHMD) methodologies have emerged to incorporate pH as an added external thermodynamic parameter to conventional molecular dynamics (MD) simulations, allowing fluctuations in the protonation of titratable residues to accompany conformational sampling.[29-34] To date, CpHMD simulations have been

used to successfully predict pK_a values of titratable groups in proteins[29-37] and nucleic acids,[38-40] as well as to explain the mechanism behind the pH-dependent conformational changes critical to the function of proteins such as nitrophorin[41] and rhodopsin.[42]

The CpHMD method provides a framework through which the pH dependence of binding processes can be examined. To the best of our knowledge, there is currently no standard protocol available to rigorously account for proton-linked ligand binding. Multiple experimental and computational groups, however, have utilized the binding polynomial formalism devised by Wyman[43] to calculate the changes in binding free energy that accompany binding-induced protonation changes for both protein-protein[3,4,44] and protein-nucleic acid binding.[39,45,46] Motivated by Mason and Jensen's usage of this binding polynomial formalism to estimate the free energies of binding for protein-protein complexes using the PROPKA web server,[4] we adopt a similar approach in conjunction with the CpHMD method by Mongan *et al*.[33] to obtain pH-dependent free energy profiles *in silico* for the binding of small molecules to the cucurbit[7]uril (CB[7]) host.

CB[7] is a synthetic molecule with seven repeating glycoluril units bridged by methylene groups (Figure 5-1A).[47,48] This 7-fold symmetric host has gained much attention due to its ability to encapsulate drug-like small molecules with high affinity as a stable host-guest complex.[49-55] Benzimidazole (BZ) and a series of its derivatives (Figure 5-1B) comprise a class of widely used fungicides and anthelmintic drugs[56-58] that have been shown to bind to the CB[7] host and undergo the pK_a shifts as large as

4 pK units upon complex formation (Table 5-1).[59] At neutral pH, these weakly acidic guests are predominantly deprotonated when free in solution, but each binds a single proton upon encapsulation by CB[7]. Both the acid/base behaviors of BZ-derived guests and the small size and relative rigidity of CB[7] compared to a typical biomolecule make the CB[7]:BZ complexes ideal model systems to test theoretical methods for computing pH-dependent binding free energies.

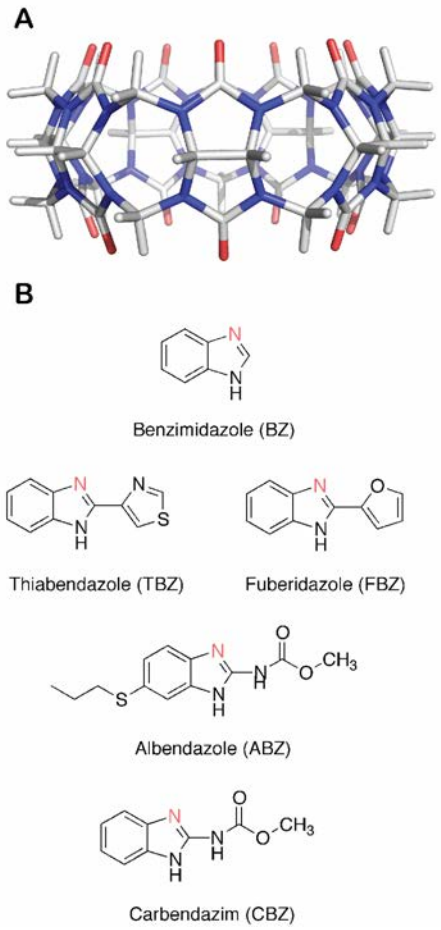

Figure 5-1. (A) Structure of the cucurbit[7]uril (CB[7]) host. (B) Structures of benzimidazole (BZ) and its derivatives.

Table 5-1. Experimental pK_a shifts of benzimidazole guests upon binding to CB[7].[59] pK_a^F denotes the pK_a of the free guest and pK_a^C represents the pK_a of the guest in complex with CB[7].

Guest	pK_a^F	$pK_a^{C,exp}$	ΔpK
BZ	5.5	9.0	3.5
TBZ	4.6	8.6	4.0
FBZ	4.8	8.6	3.8
ABZ	3.5	6.1	2.6
CBZ	4.5	7.0	2.5

In this work, we accurately reproduce the pK_a shifts of the various BZ derivatives upon binding to CB[7], using CpHMD simulations. Coupling these pK_a data with reference binding free energies taken either from experiment or from thermodynamic integration (TI) computations allows us to obtain a full description of CB[7]:guest binding free energies as functions of pH. Additionally, we show that improper assignment of guest protonation states in binding free energy computations can produce errors in excess of 2 kcal/mol at neutral pH, highlighting the importance of accurately accounting for the pH effects in free energy calculations or docking.

Theory

Binding Polynomial Formalism

Mason and Jensen recently examined the pH dependence of protein-protein binding[4] through an application of the binding polynomial formalism developed by Wyman[43] and used by Tanford to describe protein folding/unfolding.[60] Following the theoretical foundations of these groups, the binding of a titratable ligand (L) to a general macromolecular receptor (R) can be considered through a general equation for ligand association governed by the apparent equilibrium constant, K_{app}:

(1) $$R + \{L\} \overset{K_{app}}{\rightleftharpoons} \{RL\},$$

where the brackets indicate that the ligand (L) and complex (RL) ensembles may contain different protonated forms of the titratable ligand species. In the case of a ligand with a single titratable site binding to CB[7], which itself does not titrate in the biological range of pH levels, K_{app} can be written as

(2) $$K_{app} = \frac{[LR] + [HLR^+]}{[R]([L] + [HL^+])},$$

where the concentrations, rather than activities, of the given species are reported assuming ideal dilute solutions. Building from the thermodynamic cycle used to describe the proton-linked ligand binding to CB[7] (Figure 5-2), K_{app} can be rewritten according to Eq. 3, in which the concentrations of all species are presented in binding polynomials with respect to the concentrations of the deprotonated complex and ligand species:

(3) $$K_{app} = \frac{[LR]\left(1 + \frac{[HLR^+]}{[LR]}\right)}{[R][L]\left(1 + \frac{[HL^+]}{[L]}\right)} = K_b^\circ \frac{\left(1 + \frac{[HLR^+]}{[LR]}\right)}{\left(1 + \frac{[HL^+]}{[L]}\right)}.$$

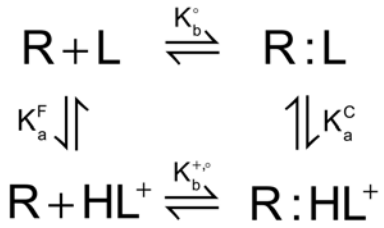

Figure 5-2. Thermodynamic cycle for complex formation between a receptor (R) and a titratable ligand (L).

Using the acid dissociation constants for the free ligand (K_a^F) and ligand-receptor complex (K_a^C), as illustrated by the vertical reactions in Figure 5-2 (Eqs. 4 and 5):

$$(4) \qquad K_a^F = \frac{[HL^+]}{[L]a_{H^+}}$$

$$(5) \qquad K_a^C = \frac{[HLR^+]}{[LR]a_{H^+}},$$

where the proton activity is denoted by a_{H+}, Eq. 3 can be rewritten in terms of the overall free energy of binding for the ligand L to the receptor R (ΔG_{bind}):

$$(6) \qquad \Delta G^\circ(pH) = -RT\ln K_b^\circ - k_B T\ln\left(\frac{1+10^{pK_a^C-pH}}{1+10^{pK_a^F-pH}}\right) = \Delta G_{ref}^\circ - k_B T\ln\left(\frac{1+10^{pK_a^C-pH}}{1+10^{pK_a^F-pH}}\right),$$

where the proton activity and acid dissociation constants have been converted to their respective logarithmic constants, pH and pK_a. The pH dependence of the binding free

energy can thus be obtained having only the pK_a values of the ligand molecule free in solution (pK_a^F) and in complex with the receptor (pK_a^C), as well as the free energy of binding for a reference reaction shown in Eq. 6 (the top reaction in Figure 5-2), $\Delta G°_{ref}$, in which there is no net uptake or release of protons. This formalism for obtaining $\Delta G°$ as a function of pH can further be applied to cases where multiple ligand and receptor groups titrate in the pH range considered, assuming that proton binding occurs independently. In other words, Eq. 6 can only be applied when all titratable groups are uncoupled from each other.

As protein active sites often contain multiple titratable groups whose protonation states are coupled to perform a given function, it will sometimes be wrong to assume that all titratable groups remain uncoupled upon ligand binding. In such cases, Wyman[43,61] derived a relation between K_{app} and pH such that

$$\frac{\partial \ln K_{app}}{\partial \ln[H^+]} = \Delta\nu_{H^+} = Z_{LR} - (Z_L + Z_R), \quad (7)$$

where, using the notation used by Tanford[60], $\Delta\nu_{H+}$ is the change in the number of bound protons in the receptor-ligand complex, relative to the number of protons bound to the ligand and receptor individually. Utilizing the unit charge of a proton, this relation is equivalent to the difference in total charge, Z, between reactants and product in Eq. 1. With $\Delta Z = Z_{LR} - (Z_L + Z_R)$, integration of Eq. 7 provides a thermodynamic relation that holds for proton-linked ligand binding in cases where titratable sites may interact (Eq. 8):

$$(8) \qquad \Delta G^{\circ}(pH) = \Delta G^{\circ}_{ref,pH} - k_B T \ln(10) \int_{pH_{ref}}^{pH} \left\{ Z_{LR}(pH) - Z_L(pH) \right\} dpH ,$$

where Z_R is omitted since the CB[7] receptor under consideration does not titrate in the pH range considered in this study. Since the integration is performed with respect to pH in the second term in Eq. 8, the reference binding free energy corresponds to the binding free energy at a specific pH.

Both Eqs. 6 and 8 thus provide frameworks for computing the pH-dependent binding free energy by adding a correction term to the reference free energy of binding. In the case of Eq. 6, the reference free energy, $\Delta G°_{ref}$, is obtained for receptor-ligand binding with protonation states fixed, such that no net change of protonation occurs. Analogously, the reference free energy in Eq. 8 is required to be the free energy of binding at a given value of pH. These two reference free energies are not necessarily equivalent; however, the reference reaction can be chosen such that they have the same value.

Constant pH Molecular Dynamics

Baptista and co-workers developed constant pH molecular dynamics (CpHMD) with stochastic titration to enable concurrent sampling of both conformational and protonation spaces according to the semi-grand canonical ensemble.[32] Here, we use the simplified CpHMD formulation implemented in the standard release of AMBER 12[62] that is similar to Baptista's formulation except that the simulation is performed in implicit solvent with generalized Born electrostatics.[33] In this method, an MD simulation is propagated from initial sets of coordinates and

protonation states. After a chosen number of MD steps, the simulation is halted, at which point a Monte Carlo (MC) step evaluates whether a random titratable residue in the system should change protonation states. The acceptance of this new protonation state is contingent on the application of the Metropolis criterion to the computed transition free energy, ΔG_{trans}, obtained using Eq. 9:

$$\Delta G_{trans} = k_B T(pH - pK_{a,ref})\ln 10 + \Delta G_{elec} - \Delta G_{elec,ref} \quad (9)$$

where pH enters as an external thermodynamic parameter and k_BT is the Boltzmann constant multiplied by the temperature of the system. For the value of pH at which the simulation is conducted, the difference in electrostatic free energy that accompanies the change in protonation being considered, ΔG_{elec}, is computed with respect to the difference in electrostatic free energy that accompanies the analogous change in protonation for a model compound, $\Delta G_{elec,ref}$, which has a known pK_a value ($pK_{a,ref}$). In this manner, any non-classical contributions to the transition free energy cancel. For a given CB[7]:guest system, the model compound that enters Eq. 9 is the guest molecule free in solution, its $pK_{a,ref}$ is the experimentally obtained pK_a value of the free guest (pK_a^F, Table 5-1), and $\Delta G_{elec,ref}$ is defined to be the electrostatic free energy that equally populates the protonated and deprotonated forms of the free guest when the solution pH is equal to the experimental pK_a of the free guest. If the transition is accepted, then MD is continued with the new protonation state for the titratable residue; otherwise, MD continues without change in the protonation state. Repeated

application of these steps builds an ensemble of protonation states along the MD trajectory.

Combining CpHMD with the Binding Polynomial Scheme

The CpHMD method is applied to obtain values for pK_a^C in Eq. 6 and ΔZ in Eq. 8 to provide pH-dependent correction terms to the reference binding free energies. In the case of Eq. 6, values of pK_a^C are obtained from simulating the CB[7]:guest system at a range of pH values. Each CpHMD simulation obtains a fractional protonation for the titratable guest being considered. By tabulating the fraction of deprotonated guest species (s) computed at each value of pH, application of the Hill equation can be used to predict pK_a^C as the midpoint of the titration (Eq. 10):

$$s = \frac{1}{1+10^{n(pK_a-pH)}} \quad (10)$$

in which n is the Hill coefficient. This method can reliably extract the pKa when the titratable residue exhibits typical titration behavior.[63]

In the case of Eq. 8, the partial charges for the guest free in solution, Z_L, and the partial charges for the guest in complex with CB[7], Z_{LR}, can similarly be obtained from CpHMD simulations. BZ and its derivatives have charges of +1 when protonated and 0 when deprotonated. Consequently, Z_{LR} and Z_L are equivalent to the fraction of protonated species (1 - s) obtained from CpHMD simulations performed on the CB[7]:guest complex and the free guest, respectively.

Methods

Parameterization of CB[7] and Benzimidazole Ligands

Partial charges for the CB[7] host have previously been derived[64] using the restrained electrostatic potential (RESP) procedure,[65-67] conventionally used to parameterize nonstandard residues for molecular simulations performed with AMBER force fields. Analogously, the geometries of benzimidazole (BZ), albendazole (ABZ), carbendazim (CBZ), fuberidizole (FBZ) and thiabendazole (TBZ, Figure 5-1), are optimized at the B3LYP/6-31G(d) level of theory[68-71] using the Gaussian 09 suite of programs.[72] Subsequently, the electrostatic potentials (ESP) associated with the optimized geometries of these guests are computed using MK radii[73] at the HF/6-31G(d) level of theory. The ESPs of the different guest molecules are submitted to the antechamber module[67] in the AmberTools 12 suite of programs,[62] which applies the RESP procedure to extract atomic point charges for use in molecular dynamics (MD) simulations. All other CB[7] and guest ligand force field terms, including Lennard-Jones parameters, are taken from the general AMBER force field (GAFF).[74]

Docking of Guest Molecules to CB[7]

To generate starting coordinates for MD simulations of different CB[7]:guest complexes, the various BZ-derivatives are docked rigidly into the CB[7] cavity using the extra precision mode (XP) in Schrodinger's Glide program.[75-77] Each CB[7]:guest docking experiment yields a single pose for the CB[7]:guest complex, and all guest molecules bind CB[7] similarly. For illustrative purposes, the resulting CB[7]:FBZ complex obtained from Glide is shown in Figure 5-3. The BZ core of each guests is

encapsulated in the CB[7] cavity, while additional furanyl, thiazole, amido, or thioether R-groups seen in the BZ derivatives protrude outside of the entrance to CB[7]. All poses give good agreement with experiment.[59]

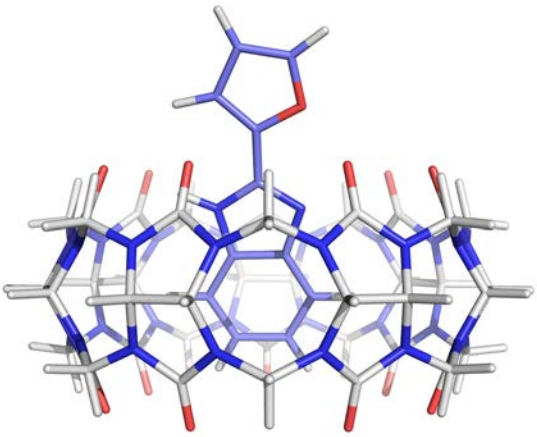

Figure 5-3. Structure of CB[7]:FBZ complex generated by docking.

Constant pH Molecular Dynamics Simulation Details

CpHMD simulations are performed using the AMBER 12 suite of programs for the range of pH values between 2 and 12 at increments of 0.5.[33,62] All simulations employ the OBC generalized Born (GB) implicit solvent model (igb=5)[78] with a salt concentration of 0.1 M. Starting from the docked CB[7]:guest structures, all systems are minimized for 5,000 steps while applying positional constraints to all heavy atoms with a force constant of 20 kcal/mol Å^2. Following minimization, the system is heated to 300 K over the course of 500 ps using a Langevin thermostat[79] while maintaining the positional constraints applied to all heavy atoms with a force constant of 5 kcal/mol Å^2. After heating, a 1 ns equilibration simulation is performed at 300K.

Production simulations are then performed for 5 ns, with MC steps taken every 10 fs. In all equilibration and production steps, the bonds involving hydrogen are constrained using the SHAKE algorithm[80], and a cutoff of 30 Å for the computation of nonbonded interactions is enforced.

Thermodynamic Integration computations

The calculation of the pH-dependent binding free energy requires a reference binding energy obtained either in the absence of protonation change (Eq. 6) or at a specified pH value (Eq. 8). TI computations are thus performed to obtain the absolute binding free energy between CB[7] and guest molecules that are deprotonated both free in solution and in complex. In TI, the free energy change is evaluated as

$$\Delta G_{0\to 1} = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle d\lambda \,, \tag{11}$$

where U is the total potential energy of the system coupled to λ, which varies smoothly between the initial state of λ=0 and the final state of λ=1.[81] The reference binding free energy is obtained from the thermodynamic cycle shown in Figure 5-4 and is calculated using Eq. 12:

$$\Delta G^{\circ}_{\mathrm{ref,TI}} = -\Delta G_1 - \Delta G_2 - \Delta G^{\circ}_3 + \Delta G_4 \,, \tag{12}$$

where ΔG_1 is the free energy for gradually turning on restraints (see below), ΔG_2 is the free energy for decoupling the guest while bound to the host in the presence of the restraints, $\Delta G_3°$ is the free energy for turning off the restraint and correcting for the standard state, and ΔG_4 is the solvation free energy for the decoupled guest (Figure 5-4).

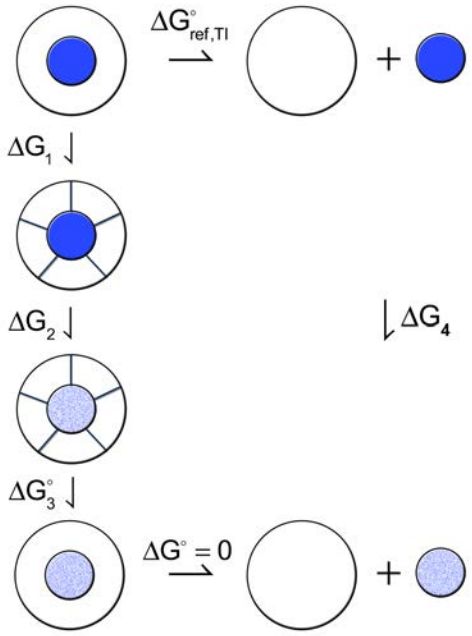

Figure 5-4. Thermodynamic cycle for an absolute binding free energy calculation. The outer circle represents a CB[7] host and the inner blue circle shows a guest molecule in the reference deprotonated state. ΔG_1 is the free energy for gradually turning on the restraints; ΔG_2 is for decoupling the guest from the host in the presence of the restraints; $\Delta G_3°$ is the analytical correction for removing the restraints; and ΔG_4 is the solvation free energy for the guest.

The electrostatic and van der Waals (vdW) contributions to ΔG_2 and ΔG_4 are computed separately, the latter using the softcore potential algorithm.[82-84] To improve the convergence for these computations, the virtual bond algorithm developed by Karplus and co-workers is applied, where a set of restraints are used to fix the position

and orientation of the guest relative to CB[7]. The free energy for turning on the restraint, ΔG_1, is computed using TI. The free energy for turning off the restraint, $\Delta G_3°$, is calculated using an analytical expression, which corrects for the presence of restraints and also accounts for the standard state:[85]

$$(13) \qquad \Delta G_3^\circ = -k_B T \ln\left[\frac{8\pi^2 V^\circ}{r_{aA,^\circ}^2 \sin\theta_{A,^\circ} \sin\theta_{B,^\circ}} \frac{(K_r K_{\theta_A} K_{\theta_B} K_{\phi_A} K_{\phi_B} K_{\phi_C})^{1/2}}{(2\pi k_B T)^3}\right] - k_B T \ln\frac{\sigma_{R\cdots L}}{\sigma_R \sigma_L}.$$

Here V° is the standard state volume of 1661 Å^3 for ideal gas, r_{aA}, $\sin\theta_A$, $\sin\theta_B$ are the distance and angle values used for each restraint, having corresponding harmonic force constants (K's in Eq. 13), which are 5 kcal/mol Å^2 for the distance restraint and 20 kcal/mol rad^2 for the angle and dihedral restraints. The second term in Eq. 13 accounts for the symmetry in the system, where $\sigma_{R\cdots L}$, σ_R and σ_L are the symmetry numbers for the host-guest complex, CB[7] and the guest molecule, respectively. For our system, $\sigma_{R\cdots L}$ and σ_L are 1, and σ_R is 14.

The *pmemd* implementation of TI in AMBER 14 is used to calculate the reference binding free energy $\Delta G°_{ref,\ TI}$ for each guest to CB[7].[86,87] The reference ionization state is chosen to be deprotonated because all experimental values of $\Delta G°_{ref}$ were measured with the guests deprotonated.[59] For the calculation of the electrostatic contribution to ΔG_2 and ΔG_4, 11 equally spaced λ values are used (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0). For the calculation of the vdW contribution to ΔG_2 and ΔG_4, 21 λ values are used (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.725, 0.75, 0.775, 0.8,

0.825, 0.85, 0.875, 0.9, 0.925, 0.95, 0.975, 1.0). For the computation of the free energy for turning on the restraints, ΔG_1, 16 λ values are used (0.0, 0.01, 0.02, 0.03, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0). The unequal spacing of λ windows is needed to capture a smoother transition of $\partial U(\lambda)/\partial\lambda$ along the λ parameter and reduce errors in integration. Integration is performed numerically using the trapezoidal rule, and uncertainties in the free energies are propagated as standard deviations.

Each CB[7]:guest complex was solvated with TIP3P water[88] with a region of 12 Å in any direction using tleap program.[89] The system was minimized for 5,000 steps and heated to 300 K over 500 ps in the NVT ensemble using a Langevin thermostat[79], followed by an equilibration for 500 ps in the NPT ensemble using a Berendsen barostat[90] with isotropic position scaling to bring the system to a stable density. All production simulations are performed in the NVT ensemble and are extended until the cumulative free energy computed for each individual transformation converges (changes in $\Delta G < 0.01$ kcal/mol).

Results

Review of Experimental Results

Previously, Koner *et al*. observed enhancements in stabilities and solubilities of benzimidazole (BZ) derivatives upon encapsulation by the cucurbit[7]uril (CB[7]) host.[59] The authors obtained values of pK_a^F and pK_a^C by fitting the data from UV titrations and ^{1}H NMR spectroscopy.[91] Henceforth $pK_a^{F,exp}$ and $pK_a^{C,exp}$ will differentiate experimental pK_a values from their respective computed values, $pK_a^{F,calc}$ and $pK_a^{C,calc}$. The experimental data showed large shifts in pK_a ranging between 2.5 to

4 pK units upon complex formation with CB[7] (Table 5-1). Additionally, association constants of the complexes were obtained at basic pH where guests were presumably deprotonated in both bound and unbound states; association constants were also obtained for the binding of protonated guests through application of the thermodynamic cycle (see Scheme 2 in ref. 59). In all cases, measurements of the binding free energies for different CB[7]:guest complexes indicated that the protonated guests are favored in the CB[7] cavity.

pK_a Shifts Upon CB[7]:Guest Complex Formation

To compute the pK_a values of various BZ derivatives in complex with CB[7], we perform CpHMD simulations on five CB[7]:guest complexes. In Figure 5-5, representative titration curves are shown for benzimidazole (BZ) and albendazole (ABZ), both in complex with CB[7] and free in solution.

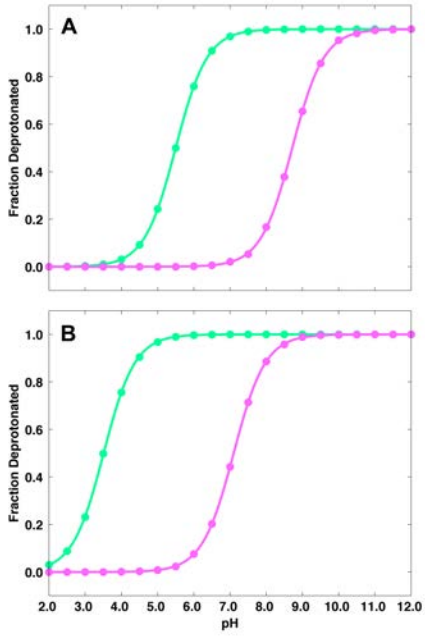

Figure 5-5. Titration curves from constant pH MD simulations of the guests free in solution (green) and in complex with CB[7] (purple). (A) Benzimidazole (BZ). (B) Albendazole (ABZ).

In the case of BZ, the $pK_a^{F,calc}$ value matches $pK_a^{F,exp}$ value of 5.5, indicating proper calibration of the CpHMD method. From the titration curve of BZ free in solution (Figure 5-5A, green curve), it is apparent that free BZ is protonated at values of pH less than 4.5 and deprotonated at pH levels above 6.5. Between these pH levels, an ensemble of protonated and deprotonated states exists. Relative to the titration curve for free BZ, the titration curve for the CB[7]:BZ complex is shifted toward more basic values of pH (Figure 5-5A, purple curve). Indeed, the value of $pK_a^{C,calc}$ for BZ is found to be 8.7 – a shift of more than 4 pK units above its pK_a^F (Table 5-2); consequently, complexed BZ is protonated at pH below 7.5, indicating the preferred protonation state of BZ at neutral (typical physiological) pH differs depending on its bound state. The observed preference for the protonated guest in the cavity of CB[7] is due to the additional hydrogen bond between the titratable proton on BZ and one of the carbonyl oxygens at the entrance to the CB[7] cavity (Figure 5-5). It is worth noting that the Hill equation provides a reasonable estimate of the pK_a values for BZ both free and in complex with CB[7], with fitting errors of ~0.01 pK units. Furthermore, the $pK_a^{C,calc}$ value for BZ underestimates its $pK_a^{C,exp}$ by only 0.3 pK units (Table 5-2).

Table 5-2. Comparison of pK_a^C values obtained from CpHMD simulations ($pK_a^{C,calc}$) with experimental data ($pK_a^{C,exp}$).[59] Fitting errors in obtaining $pK_a^{C,calc}$ from application of the Hill equation are shown.

Guest	$pK_a^{C,exp}$	$pK_a^{C,calc}$
BZ	9.0	8.71 ± 0.01
TBZ	8.6	8.19 ± 0.01
FBZ	8.6	8.61 ± 0.01
ABZ	6.1	7.10 ± 0.01
CBZ	7.0	7.40 ± 0.01

The chemical structure of ABZ differs from that of BZ by the presence amido and thioether R-groups attached to the BZ core. Additionally, the experimentally determined pK_a for the CB[7]:ABZ complex remains acidic ($pK_a^{C,exp}$ = 6.1). The titration curves obtained from CpHMD simulations of free and complexed ABZ are shown in Figure 5-5B. Qualitatively, the titration behavior of ABZ appears similar to that of BZ, as its $pK_a^{C,calc}$ of 7.1 is shifted toward a more basic value from its $pK_a^{F,exp}$ of 3.5. At neutral pH, these data suggest that ABZ is fully deprotonated when free in solution, whereas both its protonated and deprotonated forms are significantly populated when in complex with CB[7]. While the errors obtained for fitting the Hill equation to the titration data are minimal with errors observed for BZ and all other guests of less than 0.01 pK units, the value of $pK_a^{C,calc}$ for the CB[7]:ABZ complex has the greatest deviation from the experiment (ΔpK = 1.0, Table 5-2).

The titration curves for the other guests follow a similar trend, where formation of the CB[7]:guest complex increases the pK_a of the guest (data not shown). These shifts are in line with the experimentally determined pK_a values (Table 5-2), with the largest deviation seen for ABZ as stated above. Overall, the CpHMD method provides accurate predictions of $pK_a^{C,calc}$ values, with a mean average error (MAE) of 0.42 pK units with respect to experiment ($pK_a^{C,exp}$, Table 5-2).

pH Dependence of the Binding Free Energy

As discussed above, the pK_a values of the BZ-derived guests differ when bound to CB[7] and when free in solution (Figure 5-5, Table 5-2). Since there are no other titratable groups in the CB[7]:guest complexes in the pH ranges studied here, the

binding of the guests to CB[7] can have a net uptake of protons, which makes their binding free energies depend on the solution pH. In this section, we compute binding free energies as functions of pH using Eqs. 6 and 8. Both of these equations can be used to obtain the pH-dependent binding free energy by adding a pH-dependent correction term to a reference binding free energy. In Eq. 6, this reference free energy corresponds to the free energy of binding in the absence of proton binding. In contrast, Eq. 8 requires that the reference free energy be obtained at a specific pH. The reference binding free energies in these two equations can be identical if obtained at a specific value of pH where the protonation states do not change. Since experimental association constants were obtained at pH levels where both the free and bound guests are deprotonated,[59] we use reference binding free energies for the association of deprotonated guests with CB[7] in this work.

As a simple illustration of how pH-dependent binding free energies may be obtained, we first use the binding free energy measured experimentally for each of the different CB[7]:guest systems ($\Delta G°_{exp,ref}$) as the reference free energy term in Eqs. 6 and 8. We refer to this as a "hybrid" approach, as it obtains a pH-dependent binding free energy ($\Delta G°_{hybrid}$) from the experimental reference binding free energy ($\Delta G°_{exp,ref}$) and CpHMD-derived terms (either $pK_a^{C,calc}$ or ΔZ when using Eqs. 6 and 8, respectively). While all results described here have been obtained using Eq. 6 with pH-dependent corrections requiring values of $pK_a^{C,calc}$, identical results have also been obtained using Eq. 8 (data not shown).

Plots of binding free energies as functions of pH for CB[7] complexes with BZ, FBZ and ABZ are shown Figure 5-6A-C. While these binding free energies are referenced to $\Delta G°_{exp,ref}$ (Figure 5-6, red line), the use of $pK_a^{C,calc}$ in Eq. 6 and ΔZ in Eq. 8 to generate the full curve as a function of pH can be assessed by how well the computed binding free energy at acidic pH ($\Delta G°^{+}_{hybrid}$ in Table 5-3) matches the analogous value derived from experiment ($\Delta G°^{+}_{exp}$, blue line in Figure 5-6). For all CB[7]:guest complexes, the values of $\Delta G°^{+}_{hybrid}$ deviate less than 1.35 kcal/mol from the respective experimental values (Table 5-3), with the greatest error observed for ABZ. These errors are entirely due to the errors in computing values of $pK_a^{C,calc}$, as the value for $\Delta G°^{+}_{exp}$ was derived using experimentally obtained values of pK_a^{C}, pK_a^{F} and $\Delta G°_{ref}$.[59]

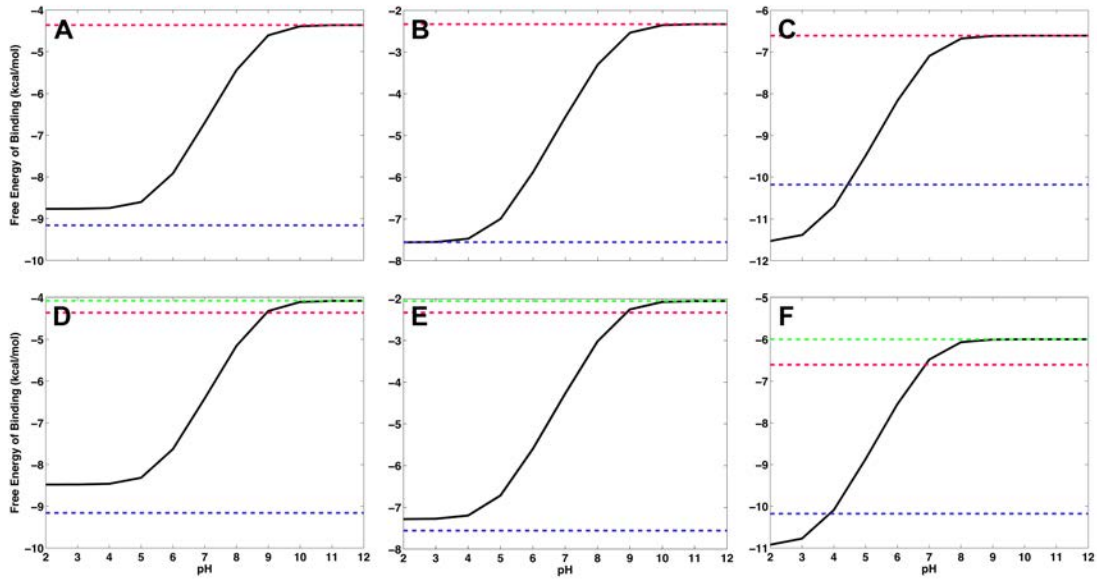

Figure 5-6. Binding free energies as functions of pH (black line). The top row (A-C) is computed by the hybrid approach using the experimental reference binding energies ($\Delta G°_{ref,exp}$, red line) and the bottom row (D-F) uses the full computational approach with the reference binding energies computed by thermodynamic integration ($\Delta G°_{ref,TI}$, green line). Experimentally derived binding free energies for the protonated guests are shown in blue. (A, D) CB[7]:BZ. (B, E) CB[7]:FBZ. (C, F) CB[7]:ABZ.

Table 5-3. Binding free energies of the guests upon complex formation with CB[7], computed using the hybrid approach with Eq. 6. All energies are reported in kcal/mol. $\Delta G°_{ref,exp}$ is the experimental[59] binding free energy for the reference deprotonated guest; $\Delta G°^{+}_{exp}$ is the binding free energy for the protonated guest derived from the $\Delta G°_{ref,exp}$; and $\Delta G°^{+}_{hybrid}$ is the free energy obtained by using $pK_a^{C,calc}$ with $\Delta G°_{ref,exp}$ in Eq. 6.

Guest	$\Delta G°_{ref,exp}$	$\Delta G°^{+}_{exp}$	$\Delta G°^{+}_{hybrid}$	$\Delta G°_{ref,exp,pH=7}$	$\Delta G°_{hybrid,pH=7}$
BZ	-4.4	-9.2	-8.8	-7.1	-6.7
TBZ	-3.0	-8.6	-7.9	-5.2	-4.7
FBZ	-2.3	-7.6	-7.6	-4.5	-4.6
ABZ	-6.6	-10.2	-11.5	-6.7	-7.1
CBZ	-6.0	-9.5	-10.0	-6.4	-6.8

From Figure 5-6A-C, it is evident that all guests bind more favorably when protonated. Indeed, the binding free energies observed for deprotonated guests (at extremely basic pH) are 3.4 to 5.6 kcal/mol more positive (less favorable) than those obtained at acidic pH when the guests are protonated. This tendency is most pronounced in CB[7]:FBZ complex (Figure 5-6B), for which the binding free energy obtained when FBZ is predominantly protonated (-7.56 kcal/mol) is over 5 kcal/mol more favorable than its respective value when FBZ is deprotonated (-2.33 kcal/mol). This observation is consistent with experiment[59] and stems from the additional hydrogen bond formed between the protonated guest and CB[7] (Figure 5-7).

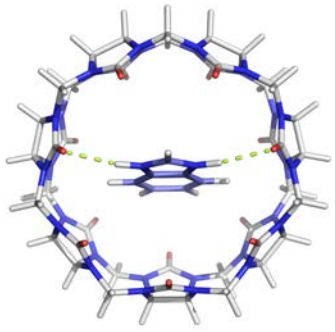

Figure 5-7. Hydrogen bonds formed between the protonated benzimidazole with the carbonyl oxygens of CB[7].

Taking a closer look at the pH-dependent binding free energies of the CB[7]:BZ complex, it is apparent that the binding free energy spans 4.8 kcal/mol between pH levels 4.5 to 10, a range that essentially encompasses the pH levels of most biological reactions (Figure 5-6A).[92] At physiological pH (~ 7), free BZ is predominantly deprotonated, whereas BZ in complex with CB[7] is protonated (Figure 5-6A). In conventional free energy computations, ligand protonation states are typically assigned as the preferred protonation state for the free ligand. Consistent with this convention, BZ would be considered deprotonated in free energy computations performed at pH 7. In making this assumption, the binding free energy deviates from the pH-dependent binding free energy obtained here by ~2.3 kcal/mol for the CB[7]:BZ complex. Similar deviations are noted for the binding of other guests as well, with the magnitudes ranging between 0.4 to 2.3 kcal/mol (Table 5-3).

Full Prediction of the pH-Dependent Free Energy Profile

To demonstrate the utility of our method when experimental binding free energies are unavailable, we perform TI computations based on the thermodynamic cycle shown in Figure 5-4 to obtain $\Delta G°_{ref,TI}$, the reference binding free energies of the CB[7]:guest complexes with the guests deprotonated. The pH-dependent correction terms obtained either with Eq. 6 or 8 are then referenced to $\Delta G°_{ref,TI}$ to obtain a full computational prediction (CpHMD/TI) of the pH-dependent free energy profiles.

Table 5-4. Binding free energies of the guests, computed using full computational approach (CpHMD/TI) and compared to experiment.[59] All energies are reported in kcal/mol. $\Delta G°_{ref,exp}$ is the experimental[59] binding free energy for the reference deprotonated guest; $\Delta G°_{ref,TI}$ is the absolute binding free energy obtained from TI computations for the reference state; $\Delta G°^{+}_{exp}$ is the binding free energy for the protonated guest derived from the $\Delta G°_{ref,exp}$; and $\Delta G°^{+}_{TI}$ is the binding free energy obtained by using $pK_a^{C,calc}$ with $\Delta G°_{ref,TI}$ in Eq. 6.

Guest	$\Delta G°_{ref,exp}$	$\Delta G°_{ref,TI}$	$\Delta G°^{+}_{exp}$	$\Delta G°^{+}_{TI}$	$\Delta G°_{ref,exp,pH=7}$	$\Delta G°_{ref,TI,pH=7}$
BZ	-4.4	-4.1 ± 2.0	-9.2	-8.5	-7.1	-6.4
TBZ	-3.0	-2.3 ± 2.6	-8.6	-7.3	-5.2	-4.0
FBZ	-2.3	-2.1 ± 2.6	-7.6	-7.3	-4.5	-4.3
ABZ	-6.6	-6.0 ± 3.0	-10.2	-11.0	-6.7	-6.5
CBZ	-6.0	-4.8 ± 2.7	-9.5	-8.7	-6.4	-5.5

The free energy profiles of CB[7]:guest complexes using $\Delta G°_{ref,TI}$ are shown in Figure 5-6 (D-F), while the computed values of $\Delta G°_{ref,TI}$ are reported in Table 5-4 for comparison with experiment ($\Delta G°_{ref,exp}$, Table 5-4). All values of $\Delta G°_{ref,TI}$ agree well with experiment, showing absolute errors that are less than 1.3 kcal/mol. Further, the error with respect to experiment ($\Delta G°^{+}_{exp}$) in the predicted values for the binding free energy of protonated guests ($\Delta G°^{+}_{TI}$) are similarly low (< 1.4 kcal/mol). Errors in $\Delta G°^{+}_{TI}$ arise from both the computation of $pK_a^{C,calc}$ (or ΔZ, when using Eq. 8) using the CpHMD method and from the binding free energy computation with TI. These errors are not always additive for the CB[7]:BZ guest systems; for example, the deviation from $\Delta G°^{+}_{exp}$ obtained for ABZ decreased from 1.4 kcal/mol when using $\Delta G°_{ref,exp}$ in the hybrid approach to 0.7 kcal/mol when using the $\Delta G°_{ref,TI}$ reference. In contrast, the deviation for TBZ increased from 0.7 kcal/mol when using the experimental $\Delta G°_{ref,exp}$ reference to 1.4 kcal/mol when using $\Delta G°_{ref,TI}$. Regardless, results obtained using the full computational approach with $\Delta G°_{ref,TI}$ show errors that

are similar in range to those observed using $\Delta G°_{ref,exp}$ in the hybrid experimental/computational approach. Furthermore, the errors associated with the full computational protocol can be lower than the errors that arise from performing binding free energy computations with fixed protonation states assigned to the unbound CB[7] and guest molecules (Table 5-4).

Discussion

Changes in the pK_a values and, consequently, the protonation states of ionizable species participating in biomolecular association processes are well documented. To address this phenomenon, we present a simple methodology for obtaining the pH dependence of binding free energies for a series of cucurbit[7]uril (CB[7]):guest complexes. Based on Wyman's binding polynomial formalism[43], binding free energies are computed as pH-dependent corrections to a reference binding free energy. Combining this formalism with constant pH molecular dynamics (CpHMD) simulations and free energy computations yields a reasonable protocol for evaluating the pH-dependent binding free energies of biomolecular systems.

Focusing on the application of CpHMD to Eq. 6 in order to obtain pH-dependent relative binding free energies, we assess how well CpHMD simulations can capture the pK_a of BZ guests in complex with CB[7] (Table 5-2). With the exception of albendazole (ABZ), the values of $pK_a^{C,calc}$ obtained for the different CB[7]:guest complexes deviate from experiment by less than 0.41 pK units. The $pK_a^{C,calc}$ value obtained for the CB[7]:ABZ complex, however, exhibits an error of 1.0 pK unit. Since the CpHMD simulations conducted in this study are only 5 ns long, we extended the

simulation of CB[7]:ABZ to 25 ns to ascertain whether the value of $pK_a^{C,calc}$ had converged; however, the resulting $pK_a^{C,calc}$ remains unchanged. Further, the process of fitting titration data obtained from CpHMD simulations of CB[7]:guest complexes to the Hill equation is achieved with very little statistical error (< 0.01 pK units, Table 5-2). Both of these findings indicate that the error in the $pK_a^{C,calc}$ values is not due to convergence problems in the CpHMD simulations. Instead, it is possible, though not explicitly demonstrated in this work, that inaccuracies in the computed pK_a^{C} values stem from problems with the force field due to the accuracies of similar magnitude to those seen in the previous CpHMD runs.[93,94]

Since CpHMD simulations can reliably compute the values of $pK_a^{C,calc}$ (and, similarly, ΔZ in the case of Eq. 8) for the CB[7]:guest systems, we proceed to incorporate these $pK_a^{C,calc}$ values in Eq. 6 along with a reference experimental binding free energy ($\Delta G°_{ref}$) to obtain binding free energies as functions of pH. This hybrid experimental/computational approach is followed and shown for CB[7]:guest systems (Figure 5-5A-C). To evaluate the accuracy of this approach, we compare the computed value of $\Delta G°^{+}_{hybrid}$ to experiment ($\Delta G°^{+}_{exp}$) and observe good agreement, with errors of < 1.4 kcal/mol arising from the computation of $pK_a^{C,calc}$.

Having established that Eq. 6 can successfully recapitulate pH-dependent binding free energies with an experimentally determined reference binding free energy, we consider the use of TI computations to remove this dependence on experiment. TI computations effectively reproduce the reference binding free energies observed from experiment ($\Delta G°_{ref,exp}$) with absolute errors less than 1.3 kcal/mol

($\Delta G°_{ref,TI}$, Table 5-4). The resulting pH-dependent free energy profiles using $\Delta G°_{ref,TI}$ are similar to those computed with $\Delta G°_{ref,exp}$, as shown in Figure 5-6. Furthermore, the absolute errors in predicting the free energies of the protonated guests, $\Delta G°^{+}_{TI}$, using the $\Delta G°_{ref,TI}$ reference are less than 1.3 kcal/mol. These errors arise from both $\Delta G°_{ref,TI}$ and the use of CpHMD simulations to obtain $pK_a^{C,calc}$ values. In regard to the computation of $\Delta G°_{ref,TI}$, we do observe large statistical uncertainties for all CB[7]:guest complexes considered, which stem largely from the van der Waals decoupling simulations (see Supporting Information); however, the free energies computed for every transformation in the thermodynamic cycle shown in Figure 5-6 have all converged, with the cumulative computed $\Delta G < 0.01$ kcal/mol.

The use of our CpHMD/TI approach to provide a full computational prediction of pH-dependent binding free energies is particularly advantageous when experimental association constants are not available, as most experimental measurements face the limitations at extreme pH levels due to the highly possible destabilization or denaturation of the proteins under such severe conditions. Therefore, when combined with computational free energy calculations, our method is free from such concerns, eliminating the reliance on the availability of experimental data. While the CpHMD/TI computation of pH-dependent binding free energies is prone to greater error than the hybrid experimental/computational approach described previously, we find the absolute errors in the CpHMD-derived pK_a^C values and reference binding free energies obtained from TI computations are not necessarily additive for the CB[7]:guest systems considered. Further, the observed errors with respect to the

$\Delta G^{\circ +}_{exp}$ are relatively low (< 1.4 kcal/mol). In contrast, the error in assigning incorrect protonation states in free energy computations without correcting for the pH dependence of the binding free energy can give errors in excess of 2 kcal/mol (Table 5-3). This observation underscores the importance of accounting for the linkage of proton binding or release to ligand binding in free energy computations and demonstrates the high utility of the CpHMD/TI approach.

Our results highlight the significant changes in pK_a and free energy of binding upon complex formation that accompanies a net proton uptake. Noting that the guests used here have a single titratable site, corresponding changes in free energy may sometimes be larger in protein-ligand binding, where multiple titratable groups exist. Therefore, we believe that our method will have great utility in computer-aided drug discovery, where early stages of the structure-based drug design often focus on finding a high-affinity binder to a target protein. Extensions of our methodology to such more complex protein systems may require improvements to the computational protocols employed. The simple framework developed here allows for trivial incorporation of CpHMD methods that incorporate explicit solvent models[35] and/or enhanced sampling techniques, such as accelerated molecular dynamics[37] or replica exchange,[36] to improve pK_a computations in systems where convergence is difficult.[94] Similarly, our protocol accommodates the use of alternative methods for obtaining the reference binding free energy required by Eqs. 6 and 8. Thus the computational methodology for performing CpHMD/TI computations can be chosen to best address the system under consideration.

While we have focused on the results obtained using Eq. 6, which assumes all titratable groups are decoupled, the CpHMD/TI method is also compatible with the expression for obtaining the pH-dependent binding free energy given in Eq. 8, and these two expressions yield identical results in the case of the CB[7]:guest systems considered here. We intend to build on the computational protocol developed here, with a natural extension being the application of the CpHMD/TI method to obtain pH-dependent binding free energies of protein-ligand complexes. As protein-ligand systems are more complicated than the CB[7]:guest systems considered in this work, we believe the use of Eq. 8 will have high utility to address potential interactions between titratable groups.

Given the magnitude of errors in computed binding free energies obtained with fixed protonation states in the CB[7]:guest systems, our computational protocol represents a promising approach to remove these errors, thus implicating its utility in drug discovery workflows.[95] Though not specifically addressed in this work, similar philosophies may also be applicable to the scoring functions in docking protocols.

Conclusion

In this work, we determined the pH-dependent changes in binding free energies for complex formation between cucurbit[7]uril (CB[7]) and a series of benzimidazole guests. Using constant pH molecular dynamics simulations combined with experimental data, we developed a hybrid protocol that could capture the significant changes in the CB[7]:guest binding free energies with high accuracy. Subsequently, we combined our method with thermodynamic integration (TI) to

enable a full computational prediction of the pH-dependent free energy profiles. This protocol successfully accounted for the pH-dependent changes in the binding free energies during complex formation. Future work will include examination of pH-dependent binding free energies for protein-ligand complexes.

Acknowledgements

This work was supported by the NSF (MCB-1020765), NIH (NIH GM31749), Howard Hughes Medical Institute, Center for Theoretical Biological Physics (CTBP), National Biomedical Computation Resource (NBCR) and NSF supercomputer centers. PGB and JWK acknowledge support from the NIH Molecular Biophysical Training Grant (2T32GM008326-21).

Chapter 5 is a minimally modified reprint of the material as it appears in M. Olivia Kim, Patrick G. Blachly, Joseph W. Kaus, and J. Andrew McCammon, "Protocols Utilizing Constant pH Molecular Dynamics to Compute pH-Dependent Binding Free Energies," submitted to the *Journal of Physical Chemistry B*, 2014. The dissertation author was one of two primary investigators, each of whom contributed equally to performing research and writing this paper.

Supporting Information

Table 5-S1. Free energy for each segment in the thermodynamic cycle (Scheme 2) for absolute binding free energy computations for the guests. All energies are reported in kcal/mol. ΔG_1 is the free energy for gradually turning on restraints on; ΔG_2 is the free energy for decoupling the guest while bound to the host in the presence of the restraints; $\Delta G_3°$ is the free energy for turning off the restraint and correcting for the standard state; and ΔG_4 is the solvation free energy for the decoupled guest. Statistical errors are reported as standard deviations. Following Scheme 2, $\Delta G°_{ref,TI}$ is computed as - ΔG_1 - ΔG_2 - $\Delta G_3°$ + ΔG_4.

System	Segment	ΔG_{elec}	ΔG_{vdw}	ΔG_1	$\Delta G_3°$	$\Delta G°_{ref,TI}$	$\Delta G°_{ref,exp}$
CB[7]:BZN	ΔG_2	6.1 ± 0.6	16.2 ± 1.1	6.3 ± 0.7	-14.28	-4.0 ± 2.0	-4.36
	ΔG_4	10.1 ± 0.9	0.2 ± 1.2				
CB[7]:TBN	ΔG_2	-36.5 ± 0.8	16.1 ± 1.5	4.5 ± 0.6	-14.36	-2.3 ± 2.6	-2.99
	ΔG_4	-33.7 ± 1.0	1.1 ± 1.6				
CB[7]:FBN	ΔG_2	-10.6 ± 0.8	15.4 ± 1.5	4.4 ± 0.6	-14.36	-2.1 ± 2.5	-2.33
	ΔG_4	-7.9 ± 0.9	0.6 ± 1.5				
CB[7]:ABN	ΔG_2	249.1 ± 0.9	17.7 ± 1.8	3.6 ± 0.6	-14.85	-6.0 ± 3.0	-6.61
	ΔG_4	248.8 ± 1.1	0.8 ± 1.8				
CB[7]:CBN	ΔG_2	257.8 ± 0.9	16.6 ± 1.5	4.4 ± 0.6	-14.69	-4.8 ± 2.7	-6.01
	ΔG_4	258.2 ± 1.1	1.2 ± 1.6				

References

(1) Onufriev, A. V.; Alexov, E. *Q Rev Biophys* **2013**, *46*, 181.

(2) Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. *J Phys Chem A* **2005**, *109*, 6634.

(3) Jensen, J. H. *Curr Pharm Biotechnol* **2008**, *9*, 96.

(4) Mason, A. C.; Jensen, J. H. *Proteins* **2008**, *71*, 81.

(5) Cassidy, C. S.; Lin, J.; Frey, P. A. *Biochemistry* **1997**, *36*, 4576.

(6) Stranzl, G. R.; Gruber, K.; Steinkellner, G.; Zangger, K.; Schwab, H.; Kratky, C. *J Biol Chem* **2004**, *279*, 2699.

(7) Brandsdal, B. O.; Smalås, A. O.; Åqvist, J. *Proteins* **2006**, *64*, 740.

(8) Trylska, J.; Antosiewicz, J.; Geller, M.; Hodge, C. N.; Klabe, R. M.; Head, M. S.; Gilson, M. K. *Protein Sci* **1999**, *8*, 180.

(9) Aguilar, B.; Anandakrishnan, R.; Ruscio, J. Z.; Onufriev, A. V. *Biophys J* **2010**, *98*, 872.

(10) Manallack, D. T. *SAR QSAR Environ Res* **2009**, *20*, 611.

(11) Leeson, P. D.; St-Gallay, S. A.; Wenlock, M. C. *Med Chem Commun* **2011**, *2*, 91.

(12) Meanwell, N. A. *Chem Res Toxicol* **2011**, *24*, 1420.

(13) Manallack, D. T.; Prankerd, R. J.; Yuriev, E.; Oprea, T. I.; Chalmers, D. K. *Chem Soc Rev* **2013**, *42*, 485.

(14) Gordon, J. C.; Myers, J. B.; Folta, T.; Soja, V.; Heath, L. S.; Onufriev, A. V. *Nucleic Acids Res* **2005**, *33*.

(15) Myers, J. B.; Grothaus, G.; Narayana, S.; Onufriev, A. V. *Proteins* **2006**, *63*, 928.

(16) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. *Nucleic Acids Res* **2012**, *40 (W1)*, W537.

(17) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704.

(18) Bas, D. C.; Rogers, D. M.; Jensen, J. H. *Proteins* **2008**, *73*, 765.

(19) Olsson, M. H. M.; Søndergard, C. R.; Rostkowski, M.; Jensen, J. H. *J Chem Theory Comput* **2011**, *7*, 525.

(20) Søndergard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. *J Chem Theory Comput* **2011**, *7*, 2284.

(21) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. *pKa prediction for organic acids and bases*; Chapman and Hall: London, 1981.

(22) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. *J Comput Aided Mol Des* **2007**, *21*, 681.

(23) Alexov, E.; Gunner, M. R. *Biophys J* **1997**, *74*, 2075.

(24) Georgescu, R. E.; Alexov, E.; Gunner, M. R. *Biophys J* **2002**, *83*, 1731.

(25) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J Am Chem Soc* **1990**, *112*, 6127.

(26) Warshel, A.; Russel, S. *Q Rev Biophys* **1984**, *17*, 283.

(27) Sham, Y. Y.; Chu, Z. T.; Warshel, A. *J Phys Chem B* **1997**, *101*, 4458.

(28) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2004**, *55*, 689.

(29) Baptista, A. M.; Martel, P. J.; Petersen, S. B. *Proteins* **1997**, *27*, 523.

(30) Borjesson, U.; Hunenberger, P. H. *J Chem Phys* **2001**, *114*, 9706.

(31) Lee, M. S.; Salsbury Jr., F. R.; Brooks III, C. L. *Proteins* **2004**, *56*, 738.

(32) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. *J Chem Phys* **2002**, *117*, 4184.

(33) Mongan, J.; Case, D. A.; McCammon, J. A. *J Comput Chem* **2004**, *25*, 2038.

(34) Bürgi, R.; Kollman, P. A.; van Gunsteren, W. F. *Proteins* **2002**, *47*, 469.

(35) Swails, J.; York, D. M.; Roitberg, A. *J Chem Theory Comput* **2014**, *10*, 1341.

(36) Meng, Y. L.; Roitberg, A. *J Chem Theory Comput* **2010**, *6*, 1401.

(37) Williams, S. L.; de Oliveira, C. A. F.; McCammon, J. A. *J Chem Theory Comput* **2010**, *6*, 560.

(38) Goh, G. B.; Knight, J. L.; Brooks III, C. L. *J Chem Theory Comput* **2012**, *8*, 36.

(39) Goh, G. B.; Knight, J. L.; Brooks, I., C. L. *J Chem Theory Comput* **2013**, *9*, 935.

(40) Goh, G. B.; Knight, J. L.; Brooks III, C. L. *J Phys Chem Lett* **2013**, *4*, 760.

(41) Di Russo, N. V.; Estrin, D. A.; Marti, M. A.; Roitberg, A. *PLoS Comput Biol* **2012**, *8*, e1002761.

(42) Laricheva, E. N.; Arora, K.; Knight, J. L.; Brooks, I., C. L. *J Am Chem Soc* **2013**, *135*, 10906.

(43) Wyman, J. *Adv Protein Chem* **1948**, *4*, 407.

(44) Mitra, R. C.; Zhang, Z.; Alexov, E. *Proteins* **2011**, *79*, 925.

(45) Misra, V. K.; Hecht, J. L.; Yang, A. S.; Honig, B. *Biophys J* **1998**, *75*, 2262.

(46) Moody, E. M.; Lecompte, J. T. J.; Belivacqua, P. C. *RNA* **2005**, *11*, 157.

(47) Freeman, W. A.; Mock, W. L.; Shih, N. Y. *J Am Chem Soc* **1981**, *103*, 7367.

(48) Kim, J.; Jung, I. S.; Kim, S. Y.; Lee, E.; Kang, J. K.; Sakamoto, S.; Yamaguchi, K.; Kim, K. *J Am Chem Soc* **2000**, *122*, 540.

(49) Liu, S.; Ruspic, C.; Mukhopadhyay, P.; Chakrabarti, S.; Zavalij, P. Y.; Isaacs, L. *J Am Chem Soc* **2005**, *127*, 15959.

(50) Rekharsky, M. V.; Mori, T.; Yang, C.; Ko, Y. H.; Selvapalam, N.; Kim, H.; Sobransingh, D.; Kaifer, A. E.; Liu, S.; Isaacs, L.; Chen, W.; Moghaddam, S.; Gilson, M. K.; Kim, K.; Inoue, Y. *Proc Natl Acad Sci USA* **2007**, *109*, 20006.

(51) Macartney, D. H. *Isr J Chem* **2011**, *51*, 600.

(52) Ghosh, I.; Nau, W. M. *Adv Drug Deliv Rev* **2012**, *64*, 764.

(53) Ma, D.; Hettiarachchi, G.; Nguyen, D.; Zhang, B.; Wittenberg, J. B.; Zavalij, P. Y.; Briken, V.; Isaacs, L. *Nat Chem* **2012**, *4*, 503.

(54) Hettiarachchi, G.; Nguyen, D.; Wu, J.; Lucas, D.; Ma, D.; Isaacs, L.; Briken, V. *PLoS One* **2010**, *5*, e10514.

(55) Saleh, N.; Meetani, M.; Al-Kaabi, L.; Ghosh, I.; Nau, W. M. *Supramol Chem* **2011**, *23*, 654.

(56) Horton, J. *Parasitology* **2000**, *121*, S113.

(57) Gould, M.; Nelson, L.; Waterere, D.; Hynes, R. *Biocontrol Sci Technol* **2008**, *18*, 1005.

(58) Gallidis, E.; Papadopoulos, E.; Ptochos, S.; Arsenos, G. *Vet Parasitol* **2009**, *164*, 53.

(59) Koner, A. L.; Ghosh, I.; Saleh, N.; Nau, W. M. *Can J Chem* **2011**, *89*, 139.

(60) Tanford, C. *Adv Protein Chem* **1970**, *24*, 1.

(61) Wyman, J. *J Mol Biol* **1965**, *11*, 631.

(62) AMBER 12, University of California, San Francisco: San Francisco, CA, 2012.

(63) Onufriev, A. V.; Case, D. A.; Ullmann, G. M. *Biochemistry* **2001**, *40*, 3413.

(64) Rogers, K. E.; Ortiz-Sánchez, J. M.; Baron, R.; Fajer, M.; de Oliveira, C. A. F.; McCammon, J. A. *J Chem Theory Comput* **2013**, *9*, 46.

(65) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J Phys Chem* **1993**, *97*, 10269.

(66) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J Am Chem Soc* **1993**, *115*, 9620.

(67) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J Mol Graphics Modell* **2006**, *25*, 247.

(68) Vosko, S. H.; Nusair, L. W. M. *Can J Phys* **1980**, *58*, 1200.

(69) Lee, C.; Yang, W.; Parr, R. G. *Phys Rev B* **1988**, *37*, 785.

(70) Becke, A. D. *J Chem Phys* **1993**, *98*, 5648.

(71) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J Phys Chem* **1994**, *98*, 11623.

(72) Gaussian 09, Revision A.1., Gaussian, Inc.: Wallingford, CT, 2009.

(73) Besler, B. H.; Merz, K. M.; Kollman, P. A. *J Comput Chem* **1990**, *11*, 431.

(74) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J Comput Chem* **2004**, *25*, 1157.

(75) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J Med Chem* **2004**, *47*, 1739.

(76) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. *J Med Chem* **2006**, *49*, 6177.

(77) Glide, In *Schrodinger Suite 2012*; version 5.8; Schrödinger, LLC: New York, NY, 2012.

(78) Onufriev, A. V.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383.

(79) Uberuaga, B. P.; Anghel, M.; Voter, A. F. *J Chem Phys* **2004**, *120*, 6363.

(80) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J Comput Phys* **1977**, *23*, 327.

(81) Kirkwood, J. G. *J Chem Phys* **1935**, *3*, 300.

(82) Shirts, M. R.; Pande, V. S. *J Chem Phys* **2005**, *122*, 134508.

(83) Steinbrecher, T.; Mobley, D. L.; Case, D. A. *J Chem Phys* **2007**, *127*, 214108.

(84) Steinbrecher, T.; Joung, I.; Case, D. A. *J Comput Chem* **2011**, *32*, 3253.

(85) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *J Phys Chem B* **2003**, *107*, 9535.

(86) Kaus, J. W.; Pierce, L. T.; Walker, R. C.; McCammon, J. A. *J Chem Theory Comput* **2013**, *9*, 4131.

(87) AMBER 14, University of California, San Francisco: San Francisco, CA, 2014.

(88) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J Chem Phys* **1983**, *79*, 926.

(89) Schafmeister, C. E. A. F.; Ross, W. S.; Romanovski, V. University of California, San Francisco, 1995.

(90) Zhang, Y.; Feller, S. E.; Brooks, B. R.; Pastor, R. W. *J Chem Phys* **1995**, *103*, 10252.

(91) Zhang, X.; Gramlich, G.; Wang, X.; Nau, W. M. *J Am Chem Soc* **2002**, *124*, 254.

(92) Casey, J. R.; Grinstein, S.; Orlowski, J. *Nat Rev Mol Cell Bio* **2010**, *11*, 50.

(93) Machuqueiro, M.; Baptista, A. M. *Proteins-Structure Function and Bioinformatics* **2011**, *79*, 3437.

(94) Williams, S. L.; Blachly, P. G.; McCammon, J. A. *Proteins* **2011**, *79*, 3381.

(95) Jorgensen, W. L. *Cell* **2013**, *155*, 1199.