Molecular Modeling: Elucidation of Structure/Function Relationships of Proteins and DNA at the Atomic Resolution

Jory Zmuda Ruscio

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Genetics, Bioinformatics & Computational Biology

David R. Bevan, Co-Chair
Alexey Onufriev, Co-Chair
Asim Esen
T. M. Murali
Adrian Sandu

April 6, 2007
Blacksburg, Virginia

Keywords: Molecular Dynamics, Nucleosome Core Particle, β-Glucosidases, Myoglobin
Copyright 2007, Jory Zmuda Ruscio

Molecular Modeling: Elucidation of Structure/Function Relationships of Proteins and DNA at the Atomic Resolution

Jory Zmuda Ruscio

(ABSTRACT)

While experiments provide valuable information about biological molecules, current technology cannot yet monitor atomic fluctuations at relevant time scales. Theoretical computational simulations are able to model the appropriate interactions at atomic resolution. Computational techniques have become widely used for identifying interactions in biological systems. Such methods have proven quite accurate in their ability to reproduce experimental data and also in screening and predicting pertinent activities. Molecular modeling employs theoretical and computational techniques to elucidate biologically relevant information from macromolecular structures. Three biological systems, the nucleosome core particle, myoglobin and glycosyl hydrolase family 1 β-glucosidases will be examined with molecular modeling methods. Results of our analyses provide information about DNA flexibility and packaging, internal migration of ligands in a small protein, and substrate specificity of an enzyme system.

Attribution

I, Jory Z. Ruscio, performed all of the work reported in this dissertation except for that which is reported below.

Chapter 2
Dr. Alexey Onufriev oversaw the project and contributed to the electrostatic discussion

Chapter 3
Deept Kumar and Maulik Shukla wrote the PathFinder algorithm used in part of the analysis. Drs. Alexey Onufriev and T. M. Murali oversaw the project.

Chapter 4
Drs. David Bevan and Asim Esen oversaw the project. Onufriev.

Acknowledgments

This section was the last portion of this document that I wrote, because I find it the most difficult. I have spent almost five of my 25 years studying and living at Virginia Tech in Blacksburg. Numerous people have touched and affected my life during this time. I can only attempt to acknowledge the influences these individuals have had on me.

First and foremost, I want to acknowledge the members of my committee, David Bevan, Alexey Onufriev, Asim Esen, T. M. Murali and Adrian Sandu. I especially want to thank Dr. Bevan, who has also my adviser, mentor, teacher and traveling companion. He gave me to opportunity to work in his lab, and I appreciate the time and resources he's provided to me. I also feel extremely fortunate to have for a co-adviser Dr. Onufriev. Not only has Dr. Onufriev taught me much about science, he always had the time to discuss other topics related to my development as a scientist. The mentorship that Dr. Bevan and Dr. Onufriev have shown me is truly greater than the sum of the parts. I am also grateful for the the experience I had during my rotation in Dr. Esen's lab: it solidified my belief that I did not want to be a wet lab scientist; but Dr. Esen's passion for science was extremely inspiring. The skill sets I learned from Murali and Dr. Sandu, Java and Perl, were extremely instrumental to my research, and if not for the strong base they taught me, I'm not sure I would have my data analyses finished.

Secondly, I want to thank my friends, with whom I've shared the joys and challenges of graduate school. Many nights of stress relief were spent at Rivermill, Bourdreax's, Bdubs or someones apartment. Without all of you, life in Blacksburg would not have been as enjoyable. Thank, you: Ken Hurley, Graham Jack, Craig Tollin, Matt Lieber, Ina and Andrew O'Carroll, Laura and James Freeman, Corban Rivera, Andrew Fenley, Ramu Anandakrishnan, Curtis Dahn, John Gordon, Mihaela Babiceanu, Diego Cortes, GSA Board Members, poker companions, and Evergrid friends. I also want to thank the VT faculty, administrators and staff who have supported me in my

non-research endeavors: Monika Gibson, Dean Karen DePauw, Donna Sanzebach, Dr. Edward Spencer, Dr. Dorris Zallen and Elaine Quenseberry.

I also want to thank my family, who have always supported me in everything that I've done . Words cannot express the love and gratitude I have for all of you: my parents, who always believed in my abilities; Ryan, a great brother who has persevered with a smile on his face; Kira, with whom I've become much closer through our shared Hokie experience; and Randy, my younger, not little, brother, who has been just down to the road and contributed much humor to many of the fun gatherings.

Finally, and most importantly, I want to thank my husband Joseph Ruscio. I cannot begin to describe how important you are to me. I would not be here now if it were not for you. I cherish the last seven years we have been together and look forward to the rest of our lives together.

In light of the recent tragic event on Monday, April 16, 2007, I also want to thank the Virginia Tech community as a whole. The manner in which the community has come together truly makes me proud to be associated with the Virginia Tech Hokies.

Contents

1 Introduction **1**

2 A Computational Study of Nucleosomal DNA Flexibility **4**

2.1 Abstract 4

2.2 Introduction 5

2.3 Methods 8

2.4 Results and Discussion 13

2.5 Conclusions 28

2.6 Acknowledgments 31

3 Atomic level identification of ligand migration pathways between solvent and heme in myoglobin **32**

3.1 Abstract 32

3.2 Results and Discussion 33

3.3 Supplementary Material 43

4 Computational Investigation of Mechanism and Substrate Specificities of Two Family 1 β-Glucosidases **52**

4.1 Abstract 52

4.2 Introduction 53

4.3 Methods 59
4.4 Results 62
4.5 Discussion 74

Bibliography **81**

List of Figures

1.1 Explicitly solvated system (a) compared with an implicitly solvated system (b) 2

2.1 Plots of (a) RMSD vs. time of every 10 ps and (b) Atomic Positional Fluctuation by residue. 12

2.2 Localization of kinked base-pair steps during the MD simulation of the whole nucleosome core particle 14

2.3 Minor groove width of the bases of the nucleosomal DNA 15

2.4 Axes of curvature 17

2.5 End-to-end distance and energy vs. time. 20

2.6 Conformational energy as a function of the end-to-end distance of the nucleosomal DNA 22

2.7 Standard deviation of the base-pair step parameters roll, twist and slide values 27

2.8 An example of two plausible structural scenarios of cyclization of short DNA fragments. 28

3.1 A schematic representation of the two channels of ligand migration and exit/entry in native myoglobin. 35

3.2 Comparison of ligand trajectories and free volume fluctuation analysis. 36

3.3 Effect of conformation of residue 138 in native myoglobin. 37

3.4 Effects of residue 68 in native myoglobin. 40

3.5 Effect of residue 68 in V68F myoglobin. 41

3.6 Structural origins of ligand migration channels and exit/entry portals in myoglobin. 42

3.7 The starting positions of the 8 ligands used in the "CO entry from solvent" MD simulations. 45

3.8 Ligand entry and exit times by channel and portal. 48

3.9 RMSD of backbone atoms during the 90 ns simulations of carbonmonoxy myoglobin (black) and myoglobin with photodissociated CO (red). 49

3.10 RMSD per residue during two native myoglobin trajectories. 50

3.11 The points in native myoglobin that occur with the highest frequency 51

4.1 First step (glycosylation) of double-replacement retaining mechanism. 55

4.2 Structures of DIMBOA-glucoside and dhurrin 56

4.3 Orientation of the glucosyl moiety of the ligand in the simulations of Glu1_DIMglc and Dhr1_DIMglc 65

4.4 Number of water molecules within 3.5 Å of acid/base catalytic glutamate over the course of the simulations. 68

4.5 Effect of dhurrin on active site of Glu1. 69

4.6 The active sites of Glu1_DIMglc, Glu1_dhr and Y473F_dhr after minimization and after 3.5 ns of MD. 70

4.7 Different conformations of glucose. 71

4.8 Residues included in the three different Active Site Profiles. 73

4.9 Clustering of active site residues that line the aglycone binding site (Profile20) of the β-glucosidases. 75

4.10 Profile20 of the three sequences that did not cluster in the same subfamilies based on the full sequences. 76

List of Tables

2.1 Computation time and processor utilization of the simulations. 11

2.2 Comparison of the nucleosomal DNA bending energies predicted within the classical elastic rod theory with those observed in the simulations reported here. 24

2.3 Destabilization effects of monovalent salt on the "bent" DNA state relative to the "straight" one as a function of solvent salt concentration. 25

3.1 Surface residues that close to the nine ligand exit/entry portals in native myoglobin shown in Figure 1. 34

3.2 Simulations in which CO escapes/enters native myoglobin and the residue number of the CO. 47

4.1 Distances between the catalytic residues in other Family 1 glycosyl hydrolases. 54

4.2 Hydrogen bonding between glucose atoms and protein atoms. 57

4.3 Simulations performed and abbreviations used in the paper. 61

4.4 Initial and average distances between two distances that are affected by substrate specificity. 63

4.5 Persistence of hydrogen bonds between glucose moiety and protein complexes over the 3.5 ns simulations. 66

4.6 The change in the dihedral angle that characterizes the conformation of the glucose and the charges of the C1, O5 and O1 atoms when the glycosidic bond (between C1 and O1) is lengthened. 72

Chapter 1

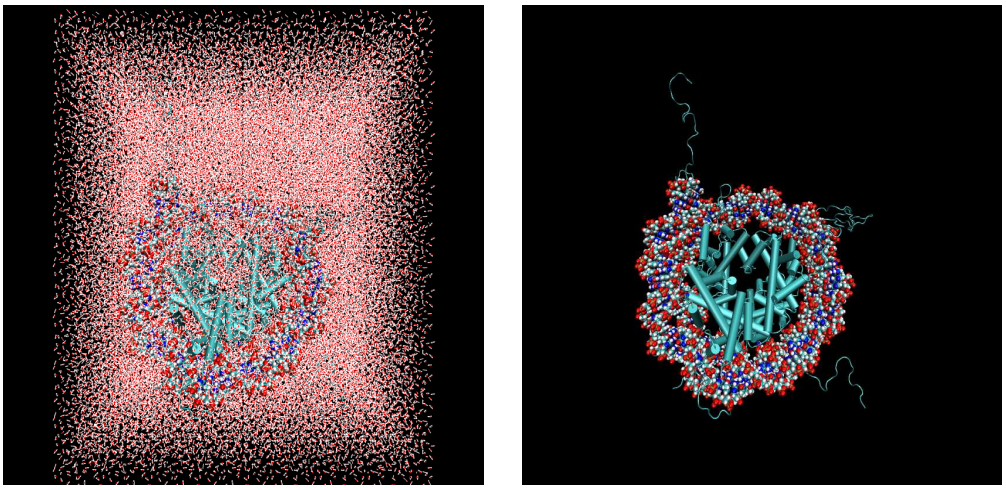
Introduction

The term "molecular modeling" refers to the use of computational and theoretical techniques to model the dynamics of molecules. The use of molecular modeling techniques allows for researchers to interpret or predict interactions at a resolution that cannot be detected with current experimental approaches. The availability of high-performance computing clusters and the increasing accuracy of theoretical methods have allowed molecular modeling and experiment to complement each other in an attempt to answer biological questions (82).

Molecular dynamics (MD) is the chosen technique for the analysis of the biological systems presented in this work. MD simulations use Newtonian physics to determine time-dependent behavior of the atomic system. The AMBER (Assisted Model Building with Energy Refinement) suite of programs is used to simulate and analyze the systems of interest (20). The AMBER MD simulations use the following potential function to calculate the interactions between atoms:

$$E_{\text{total}} = \sum_{\text{bonds}} \frac{K_i}{2}(l_i - l_{i,0})^2 + \sum_{\text{angles}} \frac{K_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{\text{torsions}} \frac{V_n}{2}(1 + cos\,[n\omega - \gamma])$$
$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} (4\epsilon_{ij} \left[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}})$$

Bond stretching, angle bending, bond rotation and non-bonded interactions contribute to the energy of the system. The last term, $\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$, corresponds to the explicit solvent treatment. The simulations are typically conducted under NVT conditions



(a) **(b)**

Figure 1.1: Explicitly solvated system (a) compared with an implicitly solvated system (b)

- constant Number of atoms, constant Volume and constant Temperature. Either explicit solvation (Figure 1.1a) or implicit solvation (Figure 1.1b) is used to model water molecules. Implicit solvent simulations are described in detail in Sec. 2.3.4.

Atomistic molecular simulations have progressed tremendously over the past couple of decades. The first simulation of the dynamics of molecules was reported 50 years ago (3). This hard sphere model, though rather simple, did demonstrate the utility of using theoretical calculations to model the dynamic behavior of a system. Since 1957, the use of molecular dynamics simulations has greatly increased, and the application of this technique has allowed for the understanding, interpretation and prediction of properties and dynamics of many biological systems. The year 1977 saw the first MD simulation of a biological molecule, bovine pancreatic trypsin inhibitor (85). This 500 atom system was simulated for 9.2 ps. Recently in 2006, the complete satellite tobacco mosaic virus, a system of over 1 million atoms, was simulated for over 50 ns (51).

This document presents the results of MD simulations of three different biological systems: DNA; a carrier protein; and enzyme-ligand complexes.

Chapter two presents research on the nucleosome core particle (NCP), which is a

DNA-protein complex, and the associated DNA. The NCP is the primary unit in DNA packaging in the eukaryotic cell. It is composed of ∼147 bp of DNA wrapped around a histone octamer protein core. Static crystal structures of the NCP indicate two different modes of DNA distortion: local kinking between base-pairs and global bending of the 147 bp of the DNA. We attempt to understand the mechanism of these distortions through the use of implicit solvation MD of the NCP and its associated DNA free in solution.

The work in chapter three examines ligand migration pathways in myoglobin. Located in muscle cells, this small globular protein stores oxygen, which binds to an internal heme iron atom. Although myoglobin was one of the first proteins to have its three-dimensional structure solved, uncertainty still exists as to how oxygen migrates between the exterior and the interior of the protein; no pathway is evident in any of the over 250 structures solved to date. We used numerous explicit solvation MD simulations and a novel algorithm to determine what appears to be a fairly complete picture of ligand migration pathways in myoglobin.

Finally, chapter four describes the investigation into the substrate specificity and mechanism of β-glucosidases. β-glucosidases are proteins involved in the hydrolysis of glucosides. Some of these enzymes exhibit a wide range of substrate specificity, while others are very narrow in specificity. We use two experimentally well characterized β-glucosidases, Glu1 and Dhr1, to examine the types of interactions that are important in determining substrate specificity and the mechanism of the proteins. Explicit solvation MD simulations are conducted to identify amino acids responsible for substrate recognition in Glu1 and Dhr1, which may lead to a better understanding of substrate specificity in other β-glucosidases.

Chapter 2

A Computational Study of Nucleosomal DNA Flexibility

Jory Z. Ruscio and Alexey Onufriev

Permission to reprint granted by the Biophysical Journal for this article originally published in the Biopysical Journal
***Biophys. J.* 91:4121-4132. 2006**

2.1 Abstract

Molecular dynamics simulations of the nucleosome core particle and its isolated DNA free in solution are reported. The simulations are based on the implicit solvent methodology and provide insights into the nature of large-scale structural fluctuations and flexibility of the nucleosomal DNA. In addition to the kinked regions previously identified in the X-ray structure of the nucleosome, the simulations support the existence of a biochemically identified distorted region of the DNA. Comparison of computed relative free energies shows that formation of the kinks is associated with little, if any, energy cost relative to a smooth, ideal conformation of the DNA superhelix. Isolated nucleosomal DNA is found to be considerably more flexible than expected for a 147 bp stretch of DNA based on its canonical persistence length of 500 Å. Notably, the significant bending of the DNA observed in our simulations occurs

without breaking of Watson-Crick bonds. The computed relative stability of bent conformations is sensitive to the ionic strength of the solution in the physiological range; the sensitivity suggests possible experiments that might provide further insights into the structural origins of the unusual flexibility of the DNA.

Key words: molecular dynamics, nucleosome, DNA flexibility

2.2 Introduction

Evidence is now overwhelming that not only the sequence, but also the details of DNA packaging inside the cell are an important part of the genetic message. The primary level of DNA compaction in eukareotic organisms *in vivo* is the *nucleosome*. At this level, a stretch of 147 base-pairs of the DNA is tightly wrapped ($\sim$1.65 times) around a set of eight proteins (histones) that carry the charge opposite of that of the DNA. Details of the nucleosome dynamics are vital for understanding key cellular processes such as DNA replication, repair and transcription (55, 70, 71, 139, 140). Cell differentiation is also intimately linked with DNA compaction. Despite its importance, the nucleosome system is far from being fully understood. One of the key unanswered questions is the following: how can the whole nucleosome be highly stable, protective of its genetic material, while at the same time its tightly wrapped DNA be highly accessible, easily revealing its information content?

This dual nature of the DNA packaging in the nucleosome is supported by experimental studies; these reveal the high stability of the whole nucleosome at physiological conditions (122), and, at the same time, suggest that small fragments ($\sim$50 bp) of the DNA helix can transiently "peel off" (4, 97). The latter observation implies a relatively small free energy barrier associated with such partial unraveling and is suggestive of possible mechanisms behind processes such as transcription, in which fragments of the nucleosomal DNA become accessible sequentially (97). Since the atomic resolution structures of the nucleosome have become available, there is likely not much ambiguity left about its static conformation; however, details of nucleosomal dynamics are not as clear, especially at atomic resolution. These are important for developing molecular mechanisms of the key biological processes involving the nucleosome. Theoretical studies using atomistic molecular dynamics simulations may provide much needed insights in this area.

A general question that can be addressed by such simulations is exactly how flexible

is the DNA wrapped around the nucleosome? Severely restricted mobility of the double helix would be suggestive of tight binding, possibly inconsistent with the idea of transient DNA dissociation in processes such as transcription. A related question is what is the origin of structural distortions that arise when the straight DNA double-helix is forced to adopt the conformation found on the nucleosome. Past experimental studies have characterized the persistence length of DNA to be ∼500 Å, or 150 base pairs (33, 112, 133). Based on this finding, one would expect that considerable force would have to be exerted on the DNA for it to adopt the highly bent superhelical conformation of nucleosomal DNA. Not only would this force need to wrap the DNA around the histone core, but, based on experimental data, it would also have to locally distort the DNA itself. In particular, the histone core was shown to have a preference for DNA with an altered helical periodicity (58). Additionally, Hayes et al. showed that DNA sequences with different structural properties in solution all adopt a similar, slightly perturbed conformation in the nucleosome (57). The 1.9 Å resolution nucleosome core particle crystal structure affirms that the structure of the DNA on the histone core deviates significantly from the best fit ideal superhelical DNA (100). Not only does nucleosomal DNA have twice the base-pair-step curvature needed for the superhelical conformation, but it is kinked in several regions. Structurally distorted regions, or "kinks", have also been identified biochemically. Regions ±1.5 helical turns from the dyad have been shown to be the sites of DNA distortion. Some of these kinked regions have been shown to have potential biological significance. For example, the DNA located at ±15 bp from the dyad have increased sensitivity to attack by singlet oxygen (61), which preferentially attacks denatured or wedge-shaped DNA structures. This region of the DNA is also recognized as being distorted by HIV integrase (98) and permanganate (48).

The idea that considerable force is needed to create the large overall bending and local distortions of the nucleosomal DNA appears common sense from the classical picture of the rather inflexible DNA. However, it may not be all that simple, according to recent intriguing experimental findings (28, 29). Namely, on short length scales (∼100 bp) the DNA double helix was not found to behave as a relatively stiff rod, as might have been expected based on its classical (33, 112, 133) persistence length value of ∼150 bp. In contrast, short DNA fragments were found to cyclize spontaneously, with an appreciable probability. Exactly how this unusual flexibility is accommodated structurally is not known (29). Since relatively short, bent DNA fragments participate in many vital biological processes, the issue is important; it has already provoked considerable interest and debate in the community. Various, mutually exclusive (29, 44, 143), explanations for the phenomenon have been proposed. While the debate does not yet appear to have been settled (44), recent experimental

evidence based on techniques such as FRET (113) and AFM(138) – different from those used in the original study(28) – provide more evidence in support of enhanced DNA flexibility on short length scales.

In order to investigate the flexibility of nucleosomal DNA and examine the origins of possible structural distortions contributing to this flexibility, we have performed several all-atom molecular dynamics simulations. These simulations explore, at the full atomic resolution, the dynamics of the entire nucleosome core particle (histone core and the DNA) as well as its isolated DNA free in solution. In some of the simulations, constraints have been used to model the DNA winding around the nucleosome.

The key to our approach is the use of the "implicit solvent" (93, 126) technique in which the effects of aqueous solvation are represented implicitly, via a continuum medium with the properties of bulk water. The electrostatic screening effects of the high dielectric medium as well as those of salt ions enter implicitly, via appropriate terms added to the system's configurational energy. The "hydrophobicity" is also included as a separate contribution. Within the approach, the system's configurational energy in the presence of solvent becomes an analytical function of the coordinates of the macromolecule only (and solvent parameters such as ionic strength), which contributes to the method's computational efficiency. The technique effectively eliminates the need to keep track of the individual water molecules, and focuses the computational power on the macromolecule of interest (*e.g.* the protein and the DNA), often resulting in considerable gains in computational efficiency. Also, compared to the traditional explicit solvent simulations, the implicit solvent methodology effectively eliminates the drag of viscosity, leading to greatly enhanced conformational sampling. The approach is also particularly well suited for estimations of relative free energies of various molecular conformations. Over the past decade the methodology has enjoyed considerable success, especially in the computationally challenging applications such the protein folding problem(72, 114).

The paper has the following structure. First, we orient the reader by briefly describing the simulations we have performed. Then we present the results of the molecular dynamics simulation of the entire nucleosome, highlighting the structural fluctuations and "kinking" of the DNA. Next we describe the simulations and energetic analyses of two superhelical DNA structures and show that formation of the kinked structures on the nucleosomal DNA are unlikely to entail considerable energetic costs, on average. Finally, we discuss the flexibility of isolated nucleosomal DNA observed in our simulations. Analysis of the relative free energies of the conformations with various degrees of bending is presented. Computational protocols and comparative validation of the methodology are presented Section 2.3.

2.3 Methods

2.3.1 Structures

The 1.9 Å crystal structure of the nucleosome core particle (PDB code 1KX5) (37) has been used as the initial structure for molecular dynamics of the *NCP-DNA* and *WholeNCP-DNA* simulations. The same sequence is used to build DNA in standard B-form and also the *Ideal-DNA*. The structures are built with NAB (83). The *Free-DNA* is built using standard B-form parameters of 35.87^o as the twist value and 3.33 Å as the rise value. The parameters of the *WholeNCP-DNA* are first analyzed with X3DNA (81), and the corresponding parameter values of 34.65^o for twist and 3.38 Å for rise are used to build the *Ideal-DNA*.

2.3.2 Molecular Dynamics

All MD trajectories have been obtained with AMBER 8 (19), using ff99 force-field. The SHAKE method is used to restrain hydrogen – heavy atom bond distances. The integration time-step is 2 fs. The average temperature of the system is maintained at $300K$ by weak coupling (via the Berendsen algorithm) to a heat bath with coupling constant of 2 ps.

Simulations in Implicit Solvent

The implicit solvent methodology based on the modified GB model(93) (GB^{OBC}, igb = 5) is used to describe solvation effects in MD simulations of the *WholeNCP-DNA*, *Free-DNA*, *NCP-DNA* and *Ideal-DNA*. The mbondi2 radii are set. The non-polar contribution is computed via $\Delta G_{\text{surf}} = 0.005$[kcal/mol]xA[Å^2], where A is the solvent accessible surface of the molecule estimated by a fast analytical routine within AMBER. A reaction field "cut-off" (`rgbmax = 15`) is employed to speed-up the calculation of effective Born radii(20). No cut-off is used for the long-range interactions. The salt concentration is set to 0.2 M during the *WholeNCP-DNA* and the *Free-DNA* simulations. A salt concentration of 10 M is used for the *NCP-DNA* and *Ideal-DNA* simulations in order to dampen the electrostatic repulsions between the "coils" of the nucleosomal DNA and reduce possible artifacts associated with the use of the restraints. All simulations undergo 100 steps of minimization with a 5 kcal/mol all atom constraint relative to the x-ray positions. Two 10 ps equilibration

steps are performed, the first using a constraint of 1.0 kcal/mol/Å^2, the second using a 0.1 kcal/mol constraint. Following equilibration, the simulations continued without constraints, or with 0.1 or 0.001 kcal/mol/Å^2 constraints, as in the *NCP-DNA* and *Ideal-DNA* simulations. All simulations continue for 1 ns, except for the *Free-DNA* for which a total of 5 ns is performed.

Simulations in Explicit Solvent

Explicit simulations of the *WholeNCP-DNA* and *Free-DNA* have been performed to provide further validation and a reference point for implicit solvent simulations. Detailed comparisons are presented in the Validation subsection below.

The initial minimization steps are as follows. The systems are immersed in a box of TIP3 water and are neutralized with Na^+ ions. The default PME parameters of AMBER 8 are used (20). The water and ions are first equilibrated to 300K for 100 ps while the solute is held frozen. Next the water and ions undergo 300 steps of minimization. Lastly, the solvent and solute are minimized for 300 steps. The system is then gradually heated to 300 K over 40 ps at constant volume and then an additional 40 ps of constant volume MD at 300K is performed. The constant pressure explicit solvent MD simulations with Particle Mesh Ewald (PME) of the *WholeNCP-DNA* and *Free-DNA* are performed. The following conditions are used: SHAKE on the hydrogen atoms, a 2 fs timestep, temperature of 300 K, a 9 Å cutoff applied to the long-range interactions.

2.3.3 Calculation of structural signatures

The minor groove has been measured with the `ptraj` module of the AMBER package, using the same procedure as reported by El Hassan and Calladine (46). The DNA parameter values of roll, twist and slide have been calculated with the package `Curves` (77).

The error margins are computed by dividing the 500 snapshots into 5 equal sized bins in order of the time. The average of each bin is computed, and the error is computed from the standard deviations of the 5 averages.

2.3.4 Calculation of energies and forces in implicit solvent.

All energies are calculated using the MM-GBSA module in AMBER8, which uses the continuum solvent approximation. The same scheme is used to compute configurational energy $E(x_i, y_i, z_i)$ as a function of atomic coordinates (x_i, y_i, z_i) for each step of molecular dynamics. The forces are computed as $F_i = \partial E/\partial x_i$, see *e.g.* Ref. (106). Within the implicit solvent approach, the total energy E of the solvated system is calculated as the sum of gas-phase energy plus the free energy of solvation that includes the electrostatic and non-polar parts. Namely, $E = E_{\rm int} + E_{\rm elec} + E_{\rm vdw} + E_{\rm surf}$. Here, $E_{\rm int}$ represents the additive bond, angular and dihedral degrees of freedom respectively,

$$E_{\rm int} = \sum_{\rm bonds} K_r(r - r_{eq})^2 + \sum_{\rm angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{\rm dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)]$$

with the constants and their specific values defined in the Cornell *et al.* force field (32); a general discussion of this type of decomposition can be found in (106)

The van der Waals interactions between the atoms are described by $E_{\rm vdw}$, and $E_{\rm surf}$ mimics the hydrophobic effect. We use [kcal/mol] $E_{\rm surf} = 0.005\mathbf{A}$, where $\mathbf{A}$ [Å^2] is the calculated solvent accessible area of the molecule. $E_{\rm elec}$ is computed as $E_{\rm elec} = E_{\rm vac} + E_{\rm solv}$, where $E_{\rm vac}$ is the protein's Coulomb energy in vacuum, and $E_{\rm solv}$ is the electrostatic component of the free energy of solvation, computed here within the generalized Born approximation. All components of the total energy are computed for each snapshot using the AMBER force field parameters. Note that the energy computed by the MM-GBSA approach includes the free energy of solvent rearrangement implicitly. For the MM-GBSA we use the same force-field, GB model and non-polar surface parameters as described above in the MD simulation section. The differences are as follows. For the energetic analysis we do not use any reaction field cut-offs (when calculating effective Born radii we set `rgbmax` > system size). Also, we set the salt concentration to 0.2 M in all cases except in the calculation of the salt dependence of the relative stability of the "bent" and "straight" conformations. Naturally, no restraining potentials are included for the "post-MD" analysis of the individual snapshots.

The relative stability is calculated as $\Delta\Delta G(salt) = \Delta G(salt \to \infty) + [\Delta E_{\rm elec}(salt) - \Delta E_{\rm elec}(salt \to \infty)]$. Since we are interested only in how the destabilization effects of salt change as the salt concentration is decreased, we set $\Delta G(salt \to \infty) = 0$. The GB model we use here was demonstrated to describe the electrostatic effects of monovalent salt adequately(117).

2.3.5 Computational resources and times

Table 2.1: Computation time and processor utilization of the simulations.

Simulation	# Atoms	Simulation Time	Computation Time (hours	# CPUs
Implicit - *WholeNCP-DNA*	25,086	1 ns	720	16
Explicit - *WholeNCP-DNA*	223,132	500 ps	70	16
Implicit - *NCP-DNA*	9,346	1 ns	122	32
Implicit - *Free-DNA*	9,346	5 ns	115	128
Explicit - *Free-DNA*	111,629	500 ps	35	16
Explicit - DNA decamer	10,277	10 ns	56	16
Implicit - DNA decamer	632	10 ns	49	8

All of the MD simulations have been performed on Virginia Tech's SYSTEM X, the 2200 CPU supercomputer (`http://www.tcf.vt.edu/systemX.html`). The computational times and CPUs used for each simulation are noted in table 2.1. Clearly, for the large systems such as the entire nucleosome, the implicit simulations reported here appear to be more computationally expensive than the explicit simulations. This is due to the fact that we have not used any approximations to speed up the calculation of the charge-charge interactions to avoid introducing any potential artifacts into the implicit solvent (GB) approach. As a result, the time complexity of the method is $O(N^2)$, where N is the number of atoms. On the other hand, the traditional explicit solvent simulations typically employ the Particle-Mesh Ewald (PME) approximation, which reduces the time complexity of computing the electrostatic interactions to $O(Nlog(N))$ by introducing an artificial periodicity into the system. Still, the seemingly unfavorable computational expense of the implicit simulations are completely offset by drastically enhanced conformational sampling: for the DNA system, the enhancement is believed to be a factor of 20 or even 100, see Refs. (125, 141). Thus, a 5 ns implicit solvent simulation of a DNA fragment effectively corresponds to 0.1 - 0.5 μs in explicit solvent.

2.3.6 Additional Validation

The structures from the implicit solvent MD simulations described above have been compared to the ones obtained by the traditional, explicit solvent approach. The latter, in conjunction with modern force-fields, is known to provide a fairly good

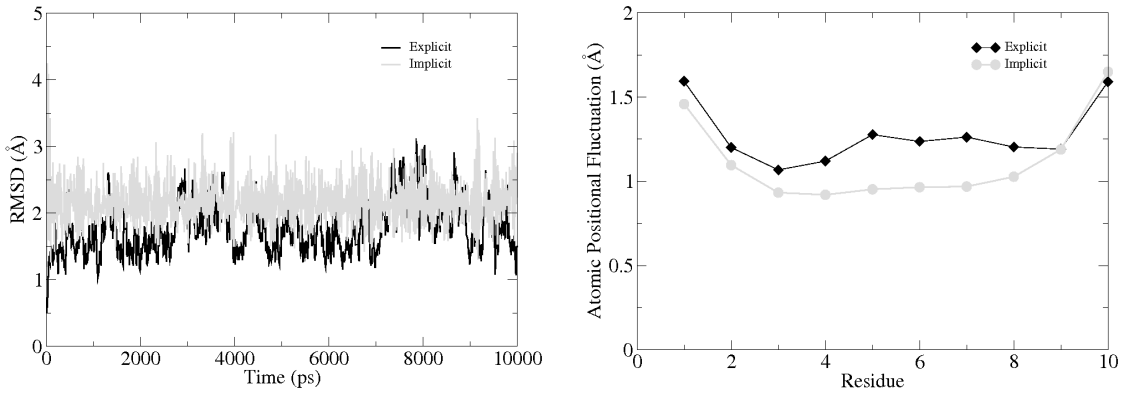


Figure 2.1: Plots of (a) RMSD vs. time of every 10 ps and (b) Atomic Positional Fluctuation by residue. Only the last 5 ns of the explicit (black) and implicit (gray) decamer simulations are shown.

representation of nucleic acids as compared to experiment(23, 24). In our implicit solvent simulation, the all-atom rmsd to the X-ray structure of the *WholeNCP-DNA* system stabilizes soon after equilibration to around 4.0 Å and exhibits only minor fluctuations around this value for the rest of the simulation. The all-atom rmsd of the explicit solvent trajectory is 3.6 Å at the end of the 500 ps simulation. Flexible histone tails are excluded from the rmsd calculations in all cases.

To provide an additional support for the use of the implicit solvent approach in the context of the nucleosomal DNA, the number of Watson-Crick base pairs have been calculated for each snapshot of the implicit solvent based *WholeNCP-DNA* and *Free-DNA* simulations. (The `3DNA` package(81) is used). The average number of pairs for the *WholeNCP-DNA* is 143. The average number of pairs for the *Free-DNA* simulation is 146. When the 10 base-pairs from both ends are excluded from the calculation, on average, all included base pairs are persistent throughout both the *WholeNCP-DNA* and *Free-DNA* simulations.

Relatively short DNA fragments present an opportunity to directly compare the explicit and the implicit solvent MD on multi-nanosecond time scale. To this end we have used a 10 bp fragment (decamer GCGCGCGCGC created with the NAB program using standard B-form DNA parameters). Both the implicit and explicit water simulations are run for 10 ns using the protocols described above. All Watson-

Crick base-pairs have persisted throughout both the implicit and explicit solvent trajectories. In the explicit solvent simulation the backbone rmsd to the $t = 0$ conformation is 1.8 ± 0.4, averaged over the last 5 ns; for the same time frame, the rmsd of the implicit simulation is only slightly larger, 2.2 ± 0.4. As is evident form Fig.2.1, the decamer in the implicit simulation is as stable as in the in the explicit solvent simulation. We attribute the fluctuations to the shortness of the fragment, and note that similar results and rmsd values were reported earlier by Tsui and Case for a DNA fragment of similar size (125). The agreement between the explicit and the implicit solvent results provides an additional support for the use of GB-based implicit solvent model for the DNA simulations, at least on the time scale of 10 ns.

2.4 Results and Discussion

The molecular dynamics (MD) simulations we have performed can be divided into three groups: 1) MD of the whole nucleosome core particle; 2) MD of just the nucleosomal DNA constrained to the wound-up conformation found on the nucleosome; and 3) MD of the "free" nucleosomal DNA in solution, initially set in a straight, classical B-form. All structures used have the same sequence as the DNA in the 1.9 Å NCP crystal structure of Richmond and Davey (100). In this section we describe results and analysis of the simulations in each group identified above. In all simulations, an implicit solvent model based on the generalized Born approximation has been used to represent solvent effects, including the screening effects of salt.

2.4.1 The Whole Nucleosome

The 1 ns long simulation of the whole nucleosome core particle (*WholeNCP-DNA*: histones + the DNA) is based on the X-ray structure of Richmond and Davey as the starting conformation. The main motivation is to estimate the size of the DNA fluctuations and explore the nature of the "kinked" regions originally identified in the X-ray structure of the NCP (100). In particular, Richmond and Davey identified six kinked base-pair steps (16/17, 26/27, 38/39, 109/110, 121/122, and 131/132), three on either side of the dyad. Minor groove bending causes kinking of the DNA in the 1.9 Å crystal structure of the NCP and is defined by, relative to the classical B-form, increased slide values, increased twist values, and decreased roll values. The results of our simulation are summarized in Fig. 2.2. Our definition of a kink is slightly less stringent than Richmond's (roll $< -11.8^o$; twist $> 40^o$; and slide > 1.25Å) in order

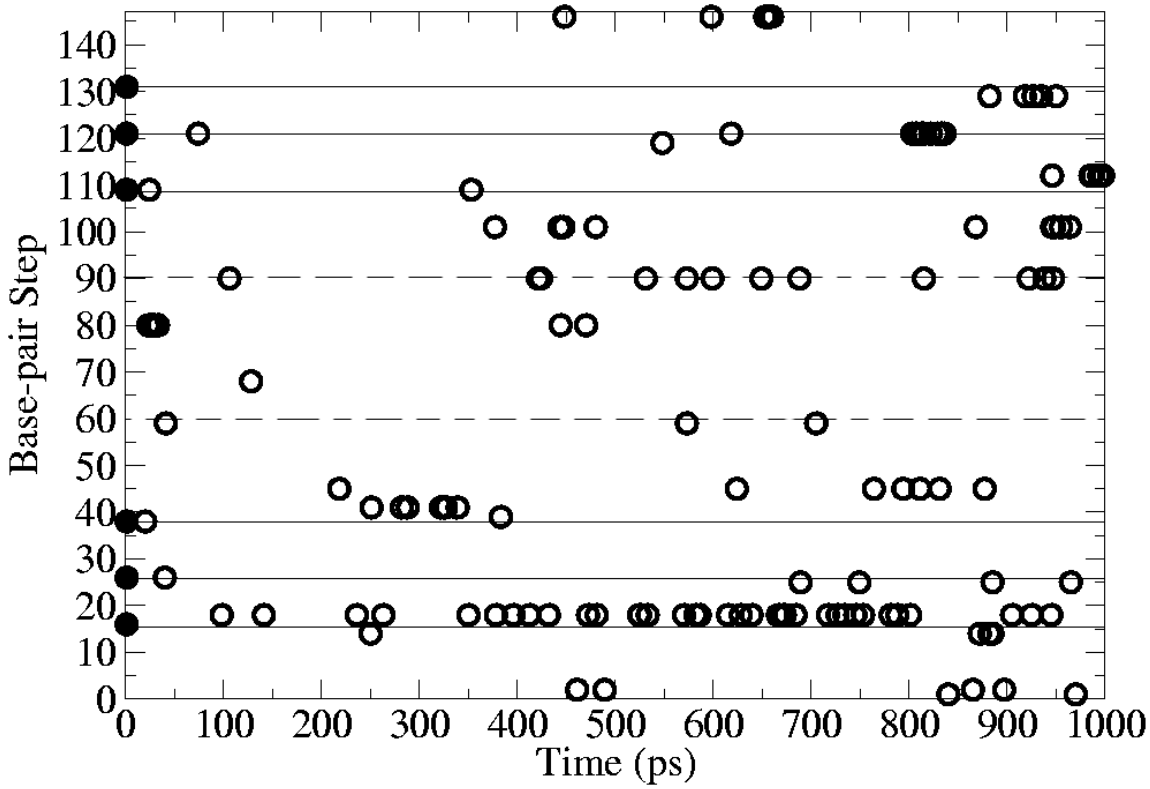


Figure 2.2: Localization of kinked base-pair steps during the MD simulation of the whole nucleosome core particle (*WholeNCP-DNA*), represented here by 1000 equidistant snapshots. The solid lines (shown here to guide the eye), correspond to the kinks in the original structure as identified by Richmond and Davey (100). We adopt a slightly more relaxed definition of a kink (open circles): Roll $<$ -11.8^o, Twist $>$ 40^o, and Slide $>$ 1.25Å. Solid circles represent kinks at t=0. The dashed line represents a region that has been identified as a spot of DNA distortion ("biochemical kink").

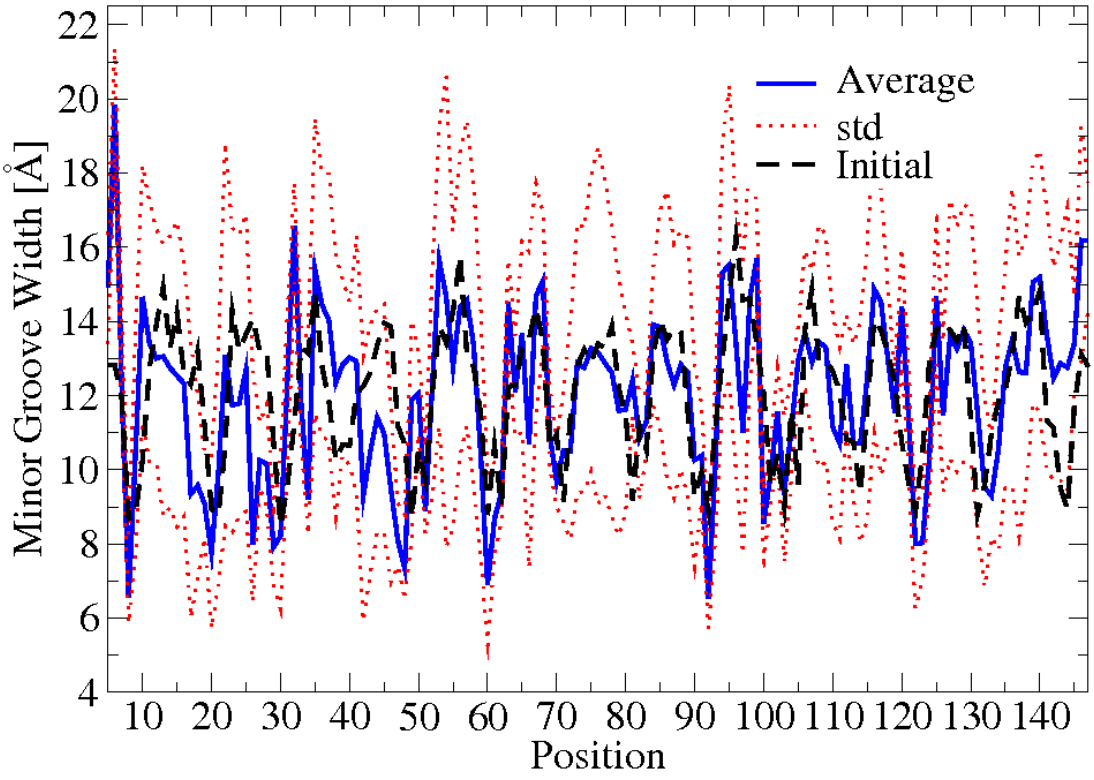


Figure 2.3: Minor groove width of the bases of the nucleosomal DNA (*WholeNCP-DNA*) during the 1 ns MD simulation. Thick blue solid line: values averaged over last 500 ns of the MD, the corresponding standard deviations are shown as red dotted lines. For reference, the initial values of minor grove width are also shown (black dashed line). An extreme narrowing of the minor groove occurs near a region located at ± 15 base pairs from the dyad.

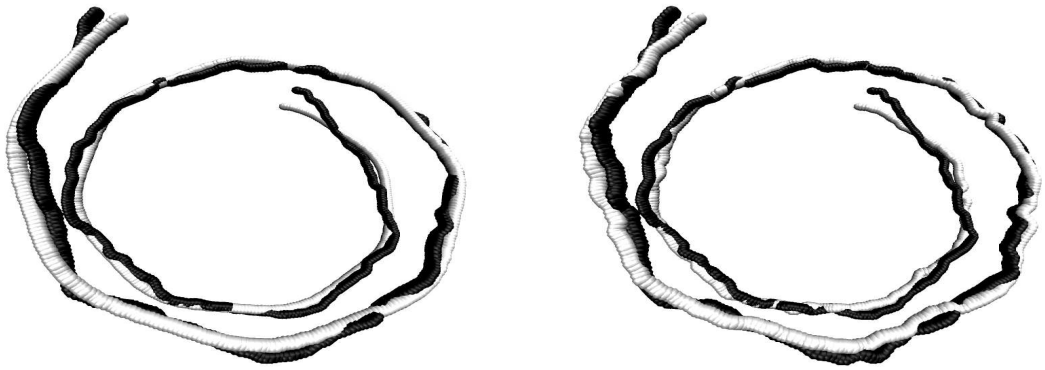
to ensure the inclusion of all of the six kinked base-pairs from the X-ray structure. Overall, during the course of the 1 ns long MD simulation, the kinks do appear to persist predominantly at the locations found in the X-ray structure, providing support for the methodology we use. The simulation also reveals the presence of a kinked region (dashed line at base-pair step 90 in Fig. 2.2) previously not identified in the X-ray structure. This region corresponds with the biochemically identified distorted structure (48, 61, 98). Analysis of the minor groove width during the MD simultion supports the existence of a distorted structure located at ±15 bp from the dyad (Fig 2.3). Near residues 60 and 90, the minor groove width decreases from ∼9 Å the initial value in the crystal structure, to an average of ∼6.5 Å over the last 500 ps of the simulation. During DNA bending, the minor groove will become compressed when it faces the histone octamer. Sharp bending is evident ±15 bp from the dyad, as the width of the minor groove at this location is much less than the other minor grooves that also face the histone octamer.

2.4.2 The Superhelical DNA: energetics of the kink formation

Experimental determination of the amount of free energy stored in the kinks of the nucleosomal DNA is difficult, if not impossible; certainly, this information is not directly available from the X-ray structure of the nucleosome. At the same time, MD simulations based on implicit solvent offer an opportunity to estimate this additional conformational energy, relative to a hypothetical "smooth" superhelical DNA, in a fairly straightforward manner (125, 126). To this end, we have performed MD simulations of two 147 bp fragments of DNA starting from the structures described below and illustrated in Fig. 2.4 (a):

1. *NCP-DNA* - nucleosome core particle DNA in the same conformation as in the core particle, but stripped of histones.
2. *Ideal-DNA* - an "ideal" ("smooth", no kinks) superhelical DNA with structural parameters corresponding to the average values of *NCP-DNA* (see "Methods" for details). No histones.

The *Ideal-DNA* and *NCP-DNA* have each undergone 1 ns of MD in implicit solvent. Cartesian harmonic restraints (0.1 $kcal/mol/\mathring{A}^2$) have been used to hold all the atoms close to their original positions. The results are illustrated in Fig. 2.4 (b),



(a) **(b)**

Figure 2.4: Axes of curvature of the initial **(a)**, and final **(b)** snapshots of the *NCP-DNA* (dark) and the *Ideal-DNA* (light) simulations. Both structures represent the winding of the DNA in the nucleosome, but the "kinks" are initially present only in the *NCP-DNA* ; the *Ideal-DNA* is in an ideal, smooth superhelical conformation at the start of the simulation. No histone core is present in either case, but all-atom harmonic restraints (0.1 $kcal/mol/\mathring{A}^2$) are used to keep the structures close to their respective initial conformations.

which shows the qualitative difference between the two structures at the beginning and at the end of the simulation. Clearly, the overall superhelical shape of the DNA found on the nucleosome is preserved by the constraints throughout the simulation. A detailed analysis reveals the persistence of most of the kinks in the *NCP-DNA* and almost a complete absence of those in the *Ideal-DNA* throughout the simulation. Namely, an average of 379 ± 27 kinked base-pair steps occur in a 100 ps interval in the *NCP-DNA* while only an average of 2.6 ± 0.4 kinks occur in the *Ideal-DNA* during the same time interval. To put these numbers in perspective: if all six kinks identified by Richmond and Davey persisted in every snapshot throughout the entire simulation, the number of kinked base-pair steps would be 600 per a 100 ps interval.

To determine whether an ideal, smooth superhelical conformation of the nucleosomal DNA is energetically preferred over the conformation (with kinks) found in the actual nucleosome, the difference in (free) energy of the two systems has been estimated on the ensemble of snapshots from the last 500 ps of the MD trajectories described above. The result is that the energy of the *NCP-DNA* is 172.42 ± 7.09 kcal/mol *less* than that of the *Ideal-DNA*. Therefore, it appears that, if anything, kinks do not cost additional energy relative to an ideal, smooth DNA constrained to the overall conformation found in the nucleosome. One has to be careful though as to the exact meaning of this finding: it does not guarantee that no extra energy (conformational strain) is stored locally in the kinked regions, but if the strain does build up, it is compensated for by conformational adjustments elsewhere on the structure so that the overall energy of the 147 bp fragment is still lowered. Of course, force-field artifacts are always a possibility, but we are reasonably confident that this is not the case: in an earlier work by Tsui and Case (125), a similar approach based on AMBER force-field was reported to be adequate to correctly describe the subtle free energy differences between A and B forms of the DNA.

To assess the effects of the strength of theharmonic constraints, we have performed another set of simulations, this time with the restraining force holding the DNA reduced by two orders of magnitude, from 0.1 to 0.001 $kcal/mol/\mathring{A}^2$. The same structures have again undergone 1 ns of MD with all other conditions being the same as described above. With these much softer restraints the overall global shape of the DNA still remains superhelical and similar to what is shown in Fig. 2.4, however the kinks practically disappear in the NCP-DNA during the course of the simulations. The structures become similar to each other; namely, an average of 2.4 ±1.2 kinked base-pair steps occur over a 100 ps interval of the *NCP-DNA* while an average of 4.8 ±4.6 kinked base-pair steps occur in a 100 ps interval of the *Ideal-DNA*. The structure that has evolved from the *Ideal-DNA* appears to be slightly

more favorable energetically, by 21.96 ± 9.16 kcal/mol, than the one evolved from the *NCP-DNA*. Again, formation of the small local distortions (kinks) appear to be energetically favorable for the entire structure.

2.4.3 The Free B-form nucleosomal DNA

Accepted theories of DNA flexibility (111, 145) backed by classic experiments (112, 133) predict the persistence length of DNA to be about 150 bp (∼500 Å) at physiological conditions. Therefore, DNA fragments shorter than this perceived length, were, until very recently, believed to be generally inflexible. Nucleosome core particle DNA contains approximately 147 bp of DNA that is wound ∼1.65 times around the histone core, and therefore one might expect that considerable force would be needed to bend the DNA to assume its orientation on the nucleosome. Recent experiments, however, suggested that DNA as short as 94 bp can spontaneously cyclize (28), indicating that at the length scales on an order of or smaller than 500 Å, the DNA may be much more flexible than previously thought.

To investigate the flexibility of nucleosomal DNA free in solution, we have performed a 5 ns MD simulation of the DNA segment 147 bp long (*Free-DNA*) with the same sequence as the DNA in the 1.9 Å NCP structure of Richmond and Davey The main results are summarized in Fig. 2.5. A natural question arises: is it realistic to observe the large fluctuations seen in Fig. 2.5 on the seemingly short time scale of the MD reported here, only 5 ns? Of course, in an experiment, the relevant time scale for the structural fluctuations of this size would be much larger, but note that in our simulations the drag of viscosity is eliminated through the use of the implicit solvent methodology, and so the *Free-DNA* can sample conformations on a much faster time scale than it would in an explicit solvent simulation or experiment. Therefore, while the simulations reported here are only several nanosecond long, the conformational sampling for the DNA systems is expected to be enhanced 20-100 fold (125, 141), which would make these simulations effectively correspond to hundreds of nanoseconds or more of "real" time. The implicit solvent approach can be considered as a sampling enhancement technique, which preserves the thermodynamic properties of the system but changes its dynamic time-scales. The presence of quasi-periodic motions, Fig. 2.5, on time-scales atypical for traditional simulations in viscous solvents is therefore not surprising. In what follows, we will not discuss any dynamical properties of the system.

Several qualitative conclusions can be made by inspecting Fig. 2.5. First, the nucleo-

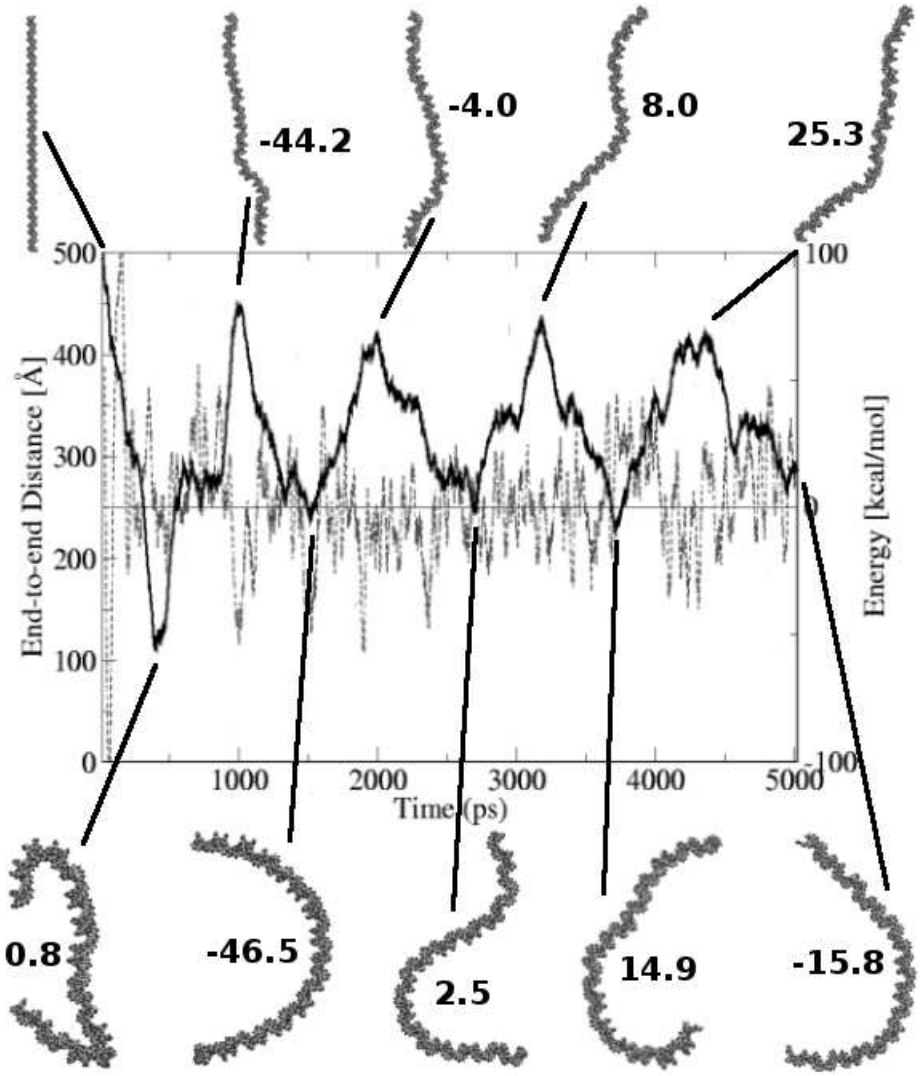


Figure 2.5: End-to-end distance and energy vs. time. End-to-end distance (thick solid line) of the the free B-DNA during the course of the 5 ns MD simulation at room temperature. Horizontal line represents the average potential energy of the last 4 ns. Dashed line is a running average of the potential energy over a window of 25 data points; this energy includes the molecular mechanical energy and the free energy of solvent rearrangement (see "Methods"). Some of the structures and their corresponding energies relative to the 4 ns average are also shown. The DNA helix is shown as VDW representation. The complete MD trajectory is available as a movie at http://people.cs.vt.edu/%7Eonufriev/RESEARCH/dna.html.

somal DNA appears quite flexible on the length scale of 150 bp. Its shape undergoes substantial fluctuations, indicating that DNA that forms a nucleosome does not behave like a fairly rigid rod. Assuming contour length of the nucleosomal DNA to be roughly equal to its canonical persistence length of L_p ∼500 Å (∼150 bp), the standard polymer theory (43, 49) predicts the average end-to-end distance to be $500 \times \sqrt{2/e} \sim 425$Å. However, according to Fig. 2.5, the end-to-end distance barely reaches this value (apart from the starting point). These results are difficult to reconcile with the classic models of the DNA flexibility; they are in better qualitative agreement with experiments of Cloutier and Widom, who estimated the persistence length for a nucleosome positioning sequence to be ∼350Å (105 bp) (29). We avoid making quantitative comparisons here because, on one hand, the measured persistence length appears to be sequence-dependent (28, 29), but also because its reliable estimation through computed end-to-end distance would require longer MD simulations than reported here. Still, we are confident that our qualitative conclusion – that the persistence length of the nucleosomal DNA is appreciably shorter than the classical value of 500 Å is valid. Given the careful equilibration protocol, a relatively tight binding to the thermostat (see "Methods") and the analysis of the energy fluctuations presented below, we are reasonably sure that during the simulated dynamics the system has not traveled far from the state of thermal equilibrium at 300K – the simulation temperature.

A more detailed analysis of the relative free energies of the DNA snapshots will now be presented to confirm the qualitative observations made above. Again, the time scales are irrelevant here: the only requirement is that enough of the conformational space has been sampled to make a statistically meaningful analysis. The implicit solvent methodology (MM-GBSA) (69, 116) allows straightforward calculation of the energy of each snapshot as a sum of the full molecular mechanics energy of the macromolecule plus the free energy of solvent rearrangement (see "Methods").

As with the size of the conformational fluctuations, one may wonder if the considerable energy fluctuations seen in Fig. 2.5 are reasonable to expect? In fact, they are quite reasonable for the system of this size. That is if the DNA fragment used here is approximated by a classical solid with 3N degrees of freedom, general thermodynamics argument predict, in the Dulong-Petit limit, the size of the energy fluctuations $\sqrt{\langle (E - < E >)^2 \rangle} = kT\sqrt{3N} \sim 100$ kcal/mol at 300 K and for the number of atoms in the nucleosomal DNA $N \sim 9000$. This is in qualitative agreement with the value of 52 kcal/mol calculated directly from the MD trajectory.

Looking at the conformational snapshots presented in Fig. 2.5 and their associated (free) energies, there appears to be no large penalty for significant (semi- circle or

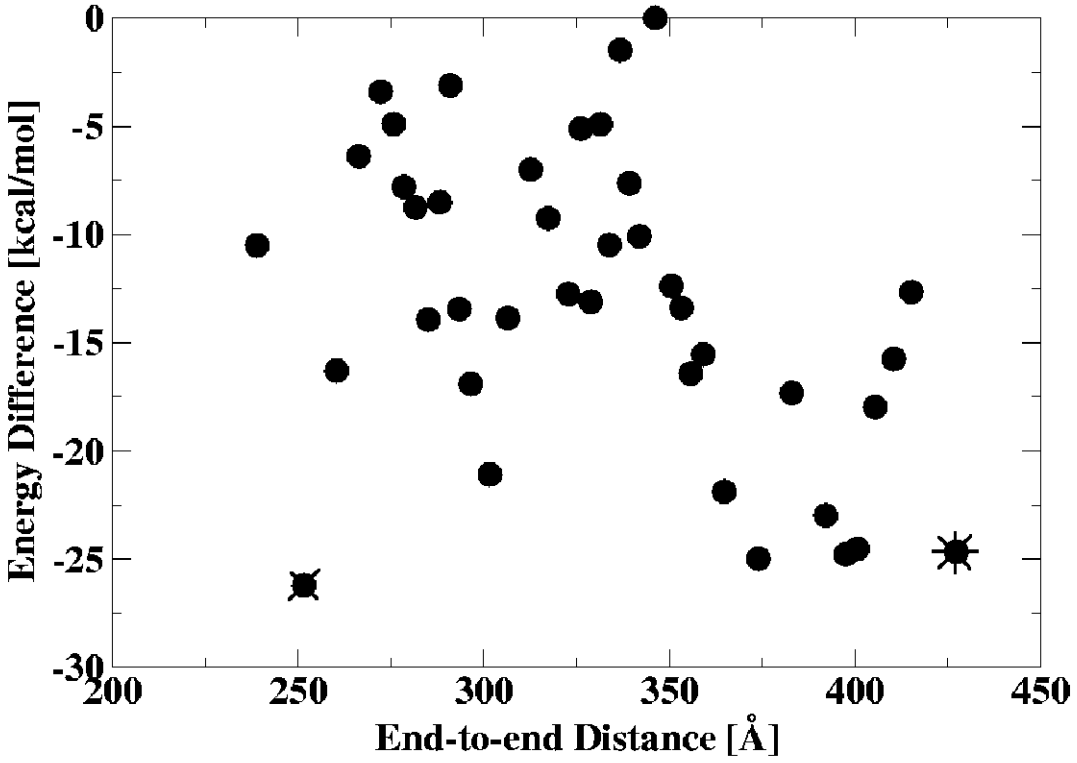


Figure 2.6: Conformational energy as a function of the end-to-end distance of the nucleosomal DNA free in solution (*Free-DNA*). Each point represents an average over a bin of 100 structures: the average energy and distance for each bin is shown. The lowest energy "bent" and "straight" conformations are denoted by an "x" and "*", respectively. All the energies are offset by a constant (the highest energy).

even more) bending of nucleosomal DNA at room temperature. This conclusion is consistent with a recent theoretical model proposed by Wiggins *et al.* (137) and with experimental observations on the contours of DNA adsorbed to a surface (138). The point is investigated further in Fig. 2.6. Clearly, the expected trend is present: conformations with smaller end-to-end distances are generally less energetically favorable. The scatter is also not surprising, as more than one spacial conformation corresponds to a given end-to-end distance, with the exception of the completely straight polymer. At the same time, some obviously bent conformations (end-to-end distance as low as 250 Å) have virtually the same energy as relatively straight ones (those with end-to-end distances around 400 Å). In fact, among the 5 lowest energy conformations (all within the error margin from each other) in Fig. 2.6, one corresponds to snapshots with an end-to-end distance of ∼250 Å.

It is worth mentioning that there appears to be no obvious sampling bias in our simulation due to possible dynamic effects. As evident from Fig. 2.6, the number of bent conformations (end-to-end distance less than 318 Å) is roughly equal to the number of stretched ones (end-to-end distance >318 Å). It does not appear that highly bent conformations of the nucleosomal DNA are very rare, as one might expect from the classical picture of the "stiff" DNA on the length scale of 150 bp. A more accurate quantitative analysis of the relative populations would require considerably longer simulations to generate an adequate ensemble of states; and the numbers may still be somewhat dependent on the molecular mechanical force-fields used.

Still, it is instructive to compare the relative energies of the snapshots in Fig. 2.6 with the predictions of the "classical" picture of DNA flexibility based on the semi-flexible polymer model(43). The latter assumes that the DNA can be represented as an elastic rod with quadratic dependence of the elastic energy upon the bending degree of freedom. Within Landau's model of thin elastic rod, the bending energy is(75):

$$E_{bend} \sim kTL_p \int_0^{L_p} ds \left(\frac{\partial \vec{t}}{\partial s}\right)^2 \geq kT(\phi)^2 \tag{2.1}$$

where the rod's axis is given by parametric curve $\vec{t}(s)$ ($\vec{t}$ is the tangent unit vector at position s). The inequality reflects the fact that the quadratic function strongly penalizes sharp bends, and so the most energetically favorable conformation for a given end-to-end distance is that of a perfect circular arc of arclength ϕ radians. The inequality becomes an equality only for an unbent rod. In the case of the nucleosomal DNA, *contour length* $\sim L_p \sim$ 500 Å which gives about $\sim kT$ of elastic energy for the classical DNA bent into a circular arc of exactly 1 radian.

Table 2.2: Comparison of the nucleosomal DNA bending energies predicted within the classical elastic rod theory with those observed in the simulations reported here. The classical values are computed assuming bending through perfectly circular arcs of specified end-to-end distance; Eq. 2.1 is used with the persistence length of $L_p = 500$ Å. Conformational energies are computed as averages over the following representative points in Fig. 2.6, as one moves along the horizontal axis: five rightmost points, five in the middle (interval [300, 325] Å), and 5 leftmost points. Their average energies $\langle E \rangle$ and end-to-end distances are reported. Note that each interval contains 500 of the original molecular dynamics conformations, 100 per each point. Since only the snapshots from the last 4 ns of the MD are included into the energy analysis, the energy of the initially fully stretched DNA conformation (end-to-end distance ~ 500 Å, $\phi = 0$) is not computed, and can be considered as an unknown constant offset.

arc length, ϕ, rad	2.0	π	3.7
end-to-end distance, Å	411	318	257
"classical" E_{bend}, kcal/mol	2.3	5.8	8.0
$\langle E \rangle$ of snapshots in Fig. 2.6, kcal/mol	-19.1	-12.8	-12.5

In Table 2.2 we compare predictions of the classical theory, Eq 2.1, with the averages of points in Fig. 2.6 taken over three intervals: the two extremes of the observed range of end-to-end distances and the middle of the range. Due to a constant (unknown) offset of the energy of the snapshots in Fig. 2.6 relative to that of the completely straight conformation of the nucleosomal DNA, it is only meaningful to compare the differences in energies as the end-to-end distance changes. We start from the relatively lightly bent conformations. The extra cost in the DNA bending energy in going from the end-to-end distance of 411 Å to 318 Å, assuming bending through perfect circular arcs, is $5.8 - 2.3 = 3.5$ kcal/mol according to the classical theory, Table 2.2. This is to be compared with $-12.8-(-19.1) = 6.3$ kcal/mol corresponding to the same change in end-to-end distance of the conformations observed in our simulation. In view of the inequality in Eq.2.1, the latter value is reasonable, and does not contradict the classical picture. However, the classical picture predicts an almost equal increase in bending energy of $8.0 - 5.8 = 2.2$ kcal/mol in going from the half-circle (318 Å or $\phi = \pi$) to the most bent conformations in Fig. 2.6, with the average end-to-end distance of 257 Å. This predicted lower bound of the energy increment is more difficult to reconcile with the much smaller value observed in our simulations $-12.5 - (-12.8) = 0.3$ kcal/mol, especially since the strength of the

inequality in Eq.2.1 is only expected to increase as the contour shape of the polymer deviates more and more from the straight line. We would like to stress, however, that generalizing this observation made for the specific nucleosomal sequence used in this work to arbitrary DNA may not be appropriate, as the the flexibility of the DNA is known to be sequence dependent.

Quantitative predictions can be made for the ionic strength dependence of the relative free energies of conformers (Table 2.3). Success of the the atomic-detail continuum electrostatics approach in describing the salt dependent conformational changes was recently reported on various molecular systems such as proteins (94), the nucleosome(8), and chromatin fibers (120). The analysis can potentially provide clues into the nature of the enhanced DNA flexibility and how it is consistent with the strong repulsion between parts of this highly charged polymer.

Table 2.3: Destabilization effects of monovalent salt on the "bent" DNA state relative to the "straight" one as a function of solvent salt concentration. The bent state corresponds to an average of the 100 conformations represented by (x) and the straight state corresponds to the average of the 100 conformations represented by (*) in Fig.2.6. Progressively higher values of $\Delta\Delta G$ indicate decreasing relative stability of the bent conformations as ionic strength of the solution decreases. Computational details are in the "Methods" Section.

Salt concentration, M	10.0	1.0	0.4	0.2	0.1	0.05	0.01
Debye screening length, Å	1.0	3.1	5.0	7.0	10.0	14.1	31.6
$\Delta\Delta G$, kcal/mol	0.0	0.0	1.85	1.48	3.22	5.19	19.17

Note that the presence of salt introduces an important length scale into the system – the Debye screening length, $D[\mathring{A}\ ] \approx 3.161/(\sqrt{[\text{salt}][mol/L]})$; electrostatic interactions over distance L scale as $\sim exp(-L/D)$. From Table 2.3 there is no apparent salt dependence above 1M salt concentration, or $D < 3$ Å, which makes sense given that the distance between consecutive negative charges along the backbone is about 6.6 Å. In the physiological range of salt concentrations, there is a small, but noticeable difference in relative free energies of straight and bent conformations. For example, changing the salt concentration from 0.2 to 0.1 M will result in destabilization of the "bent" state by 1.74 kcal/mol, corresponding to an ∼18 fold decrease in the relative population. The observation suggests that particular care of ionic strength conditions must be taken when comparing quantitative results from experiments that measure end-to-end distance of DNA fragments comparable in length to the nucleo-

somal DNA. When the salt concentration is reduced to 0.01 M, the low end-to-end distance conformations become virtually non-existing, separated by at least a 10 to 15 kcal/mol barrier from the lowest energy "straight" conformations. Interestingly, the 0.01 M salt corresponds to Debye length of only $\sim 30\mathring{A} \ll$ lowest end-to-end distance observed in our simulations (250 Å). This suggests that electrostatic repulsion on length scales of $\sim$ 30 Å become important. The observation also suggests that salt dependence of end-to-end distance fluctuations, or related experimentally accessible quantities such as salt-dependence of cyclization probability, may be used to gain insights into the mechanism of the unusual DNA flexibility discussed above.

The idea is that these salt-dependenct quantities should be strongly affected by the details of the molecular mechanism behind the observed phenomenon of unusually high bending probability of short DNA fragments. In particular, if formation of sharp "kinks" (conformation (1) in Fig. 2.8 on left) is key, one should expect a noticeably stronger dependence on the screening effects of salt compared to the case of a more uniform bending, as in conformation (2) in the same figure. This is because, for a polyion, the electrostatic repulsion disfavors sharply bent conformations much more strongly compared to the uniformly bent.

The analysis presented above and Fig. 2.6 also suggest that the elastic energy component associated with the curvature of the nucleosomal DNA may be small. This, by itself, does not necessarily mean that the over-all energy cost of creating the DNA in the conformation found in the nucleosome is small: the cost of bringing the two strands right "on top" of each other to create the 1.65 turns found in the nucleosome is unlikely to be small, due to strong electrostatic repulsion between the two strands at physiological conditions. This cost is expected to be compensated by the attractive interactions with the oppositely charged histone core.

Atomic-level MD simulations provide an opportunity to complement experiment in narrowing down the list of competing theories that claim to explain the unusual persistence length of the DNA. For example, in Ref. (29) several different theoretical models were found consistent with the available experimental data on DNA cyclization, in particular the "melted-bubble" model of Yan and Marko (143). Note that the melted-bubble mechanism, as proposed, postulates the breaking of at least 2 consecutive WC bonds to create a sharp bend in the DNA. Analysis of our *Free-DNA* MD trajectory shows that two adjacent base-pair unpairings never occurs (and even single breaks occur rarely, and only at the ends of the fragment) in the simulation that nevertheless exhibits sharp bends. Therefore, the melted-bubble may not be an appropriate model to describe the unusually high flexibility of the DNA, at least on length scales of $\sim$150 bp. This conclusion is likely to be model independent: the

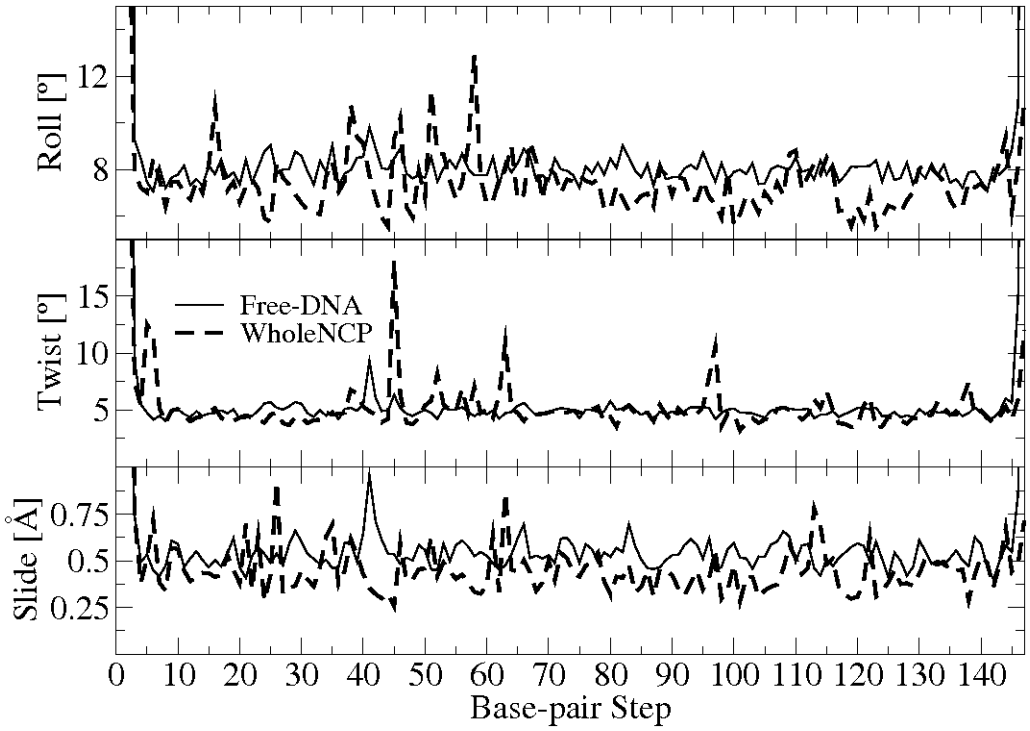


Figure 2.7: Standard deviation of the base-pair step parameters roll, twist and slide values for the *WholeNCP-DNA* (thick dashed line) and the *Free-DNA*(thin solid line) simulations. Values used to compute the standard deviation are from the last 500 ps of the *WholeNCP-DNA* simulation and that last 4.5 ns of the *Free-DNA* simulations.

drastic conformational changes such as formation of "bubbles" of broken WC bonds, should not be sensitive to details of modern force-fields used here. Similarly, while the fine-grain conformational details may be force-field dependent, the salt dependence computed within the implicit solvent approach appears to be insensitive to them (94). This is because non-electrostatic components cancel out, and what remains are "$\Delta\Delta$" estimates, which are typically more accurate than just "Δ" involved in computing of the electrostatic part of solvation energy. Our previous experience with a similar size system—unfolding apomyoglobin—shows that the computed salt dependence of free energy can be within kT of the experimental numbers (94).

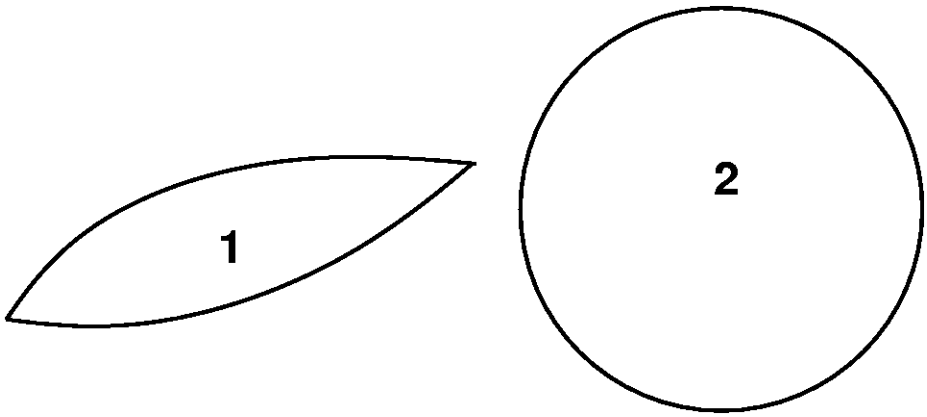


Figure 2.8: An example of two plausible structural scenarios of cyclization of short DNA fragments. Scenario (1) implies stronger electrostatic repulsion between the strands and is therefore expected to be more sensitive to the screening effects of salt.

2.4.4 Fluctuations of the DNA on the nucleosome

The many interactions between the histone core and the nucleosomal DNA are thought to severely constrain the motion of the DNA. Our MD simulations provide evidence that these interactions may not be so restrictive after all. To support our point, we compare structural fluctuations between the DNA on the nucleosome and fluctuations of the same fragment free in solution. Comparison of the standard deviation of the base-pair step parameters roll, twist and slide of the *Free-DNA* with those of the *WholeNCP-DNA* trajectory shows that these are similar, (Fig. 2.7). That is, the degrees of freedom of the nucleosomal DNA determined by these parameters do not appear to be noticeably restricted by interactions with the core histones. The similarity of the structural fluctuations between the DNA bound in the nucleosome and the corresponding free DNA was also reported in a recent molecular dynamics study that used explicit solvent (121).

2.5 Conclusions

This work is motivated by the biological importance of structural fluctuations in the nucleosome; these are related to the inherent flexibility of its DNA. Several multi-

nanosecond molecular dynamics simulations of the nucleosome core particle and its corresponding DNA in isolated form in solution have been performed. Implicit solvent methodology has been used to represent effects of aqueous solvation at physiological conditions, including the electrostatic screening effects of monovalent salt. Despite the relatively large molecular size of the system, the use of the continuum solvent methodology has allowed us to observe considerable structural fluctuations, such as significant bending of the 147 bp stretch of isolated nucleosomal DNA. The implicit solvent approach (MM-GBSA) has also been used to estimate relative free energies of conformations observed in the simulation. The analysis of the computed structural signatures and energies has led to a number conclusions.

A biochemically identified region of DNA distortion localized at ± 15 base-pairs away from the dyad becomes apparent during the simulation of the entire nucleosome core particle. This region was not identified as distorted in the 1.9 Å crystal structure of the nucleosome, but, during the simulation the distortion of the DNA in this region becomes clear. This region may have importance in the binding of the DNA to the histones; short DNA sequences that can be easily distorted may indicate where DNA binds to the histones. At the same time, the six kinked base-pair steps identified in the 1.9 Å crystal structure by Richmond and Davey, three on either side of the dyad, are not always present throughout the entire simulation. Rather, the six regions are the predominant regions of kink persistence, dynamically adapting to the fluctuations of the superhelical DNA. The result is consistent with our finding that the free energy cost of forming the kinks cannot be large: in fact, the formation of kinks is likely to be energetically favorable over-all. This result follows from comparing free energy of kinked superhelical nucleosomal DNA conformation to the corresponding "smooth" conformation making the same number of turns around the histone core. Our finding does not exclude the possibility that additional conformational strain builds up locally in the kinked regions, but if it the case, this extra energy appears to be compensated by structural adjustments elsewhere in the nucleosomal DNA that lower the overall energy of the fragment, relative to the corresponding "smooth" conformation.

A related observation is that the nucleosomal DNA bound to the histones experiences substantial structural fluctuations. In particular, the fluctuations of the key structural parameters such as roll, twist and tilt are essentially the same as those observed in the simulations of the isolated nucleosomal DNA. This observation supports the idea that the histone core does not impose strong structural constraints on the nucleosomal DNA. The histones most likely stabilize the superhelical structure of the DNA through electrostatic interactions, which are long-range and therefore

more conducive of substantial structural rearrangements in the core particle compared to strong short-range van der Waals forces. Thus, the structure of the core particle appears to be more "fluid" than one might expect from its "static" X-ray representation. From a biological standpoint, an ability of the nucleosomal DNA to easily change conformation locally while preserving the overall global structure of the nucleosome may be important. In particular, it could hint at how the entire nucleosome can be highly stable, protective of its genetic material, while at the same time its tightly wrapped DNA be highly accessible, easily revealing its information content.

The simulations of the isolated nucleosomal DNA free in solution have provided interesting insights into the nature of its flexibility. First, we have found the nucleosomal DNA to be considerably more bendable than might be expected from the the classical picture of a DNA polymer with a persistence length of about 150 bp. The conclusion is further supported by the analysis of relative free energies of the conformers seen along the simulated trajectory: the energetic cost associated with substantial degrees of bending is smaller than expected from the classical picture. Our results are in qualitative agreement with the recent cyclization experiments of Cloutier and Widom, which suggest that small DNA fragments are capable of sharp bending unaccounted for by the classical models. Notably, no breaking of Watson-Crick bonds is observed in our simulations, suggesting that the unusual flexibility of nucleosomal DNA does not imply formation of local "bubbles" of single stranded DNA. This conclusion relies on the observed absence of drastic conformational changes during the simulation, and is robust to subtle details of the methodology and the force-fields used. Computed dependence of relative free energies of the simulated conformations upon ionic strength of the solution suggests that the relative population of the highly bent and relatively straight conformations is sensitive to the salt conditions in the physiological range. This relative population is also predicted to be sensitive to the structural details responsible for the enhanced DNA flexibility. This observation suggests that by measuring this dependence experimentally one might gain further insights into the molecular mechanism underlying this newly discovered phenomenon. For example, a scenario that involved formation of sharp bends would produce a much stronger salt dependence of some experimentally accessible quantities (*e.g.* cyclization probability) than would a mechanism based on smooth, uniform bending. The methodology that we have used here to compute the salt effects was shown before to give quantitatively reliable predictions for molecular systems of similar size. Insight into the molecular mechanism underlying DNA flexibility can better our understanding of processes such as gene regulation and nucleosome positioning. Flexible DNA can form loops with much lower energetic consequence, thus providing some explanation

for the many looped regulatory complexes, and regions of DNA that are of higher flexibility may be indicators for nucleosome positioning.

Perhaps one of the main methodological conclusion of this work is that the use of implicit solvent methodology allows one to observe and analyze conformational changes on a scale that would have been difficult or impossible with the traditional explicit solvent approach. This is mainly due to the fact that the implicit solvent approach eliminates the viscous forces, and so a considerably larger part of the available conformational space is explored. Also, since solvent degrees of freedom are taken into account implicitly, estimating free energies of solvated structures is much more straightforward than with explicit water models: for the systems as large as the nucleosome, an approach based on implicit solvation maybe the only practical way to estimate relative stabilities of various conformations. Potential artifacts of periodic boundary conditions (often used to speed up explicit solvent simulations) are also avoided. This may be important for "free moving", highly charged systems such as the DNA in our simulation. While obtaining atomic-resolution pictures of conformational fluctuations in such large systems experimentally is difficult, they come within reach through the methodology described here.

2.6 Acknowledgments

The authors thank Jonathan Widom and Philip Nelson for many helpful suggestions. The authors also thank Alexander Birger and *Intelligent Software Solutions, Inc.* for the help with the animation and artistic rendering issues. This work was supported by NIH grant GM076121 and ASPIRES seed grant from Virginia Tech.

Chapter 3

Atomic level identification of ligand migration pathways between solvent and heme in myoglobin

Jory Z. Ruscio, Deept Kumar, Maulik Shukla, T. M. Murali and Alexey Onufriev

To be submitted to *Science*

3.1 Abstract

Myoglobin, the "hydrogen atom" of structural biology, is a small globular protein involved in oxygen storage in muscle cells. The oxygen binds to an internal heme iron. For several decades there has been significant interest in identifying the pathway of ligand (e.g. O_2 or CO) diffusion in myoglobin. Does the ligand travel through myoglobin along a single dominant pathway, or is it able to reach the heme iron by many different routes? Our study provides strong support for the existence of only two distinct ligand migration channels going through the protein bulk with a small number of distinct exit/entry points (portals). We identify one dominant, primary channel and one secondary channel. This conclusion is based upon the analysis of numerous room temperature explicit solvent molecular dynamics simulations, totaling over four microseconds, of two native forms of myoglobin and its V68F mutant.

We identify the two pathways by studying a) the diffusion of ligand (CO, in our MD simulations) both in and out of the protein; and b) the connectivity pattern of computed transient cavities that connect the iron atom to the surface of the protein.

3.2 Results and Discussion

The route through which small molecular ligands move to reach the heme iron in myoglobin, a process vital for its function in oxygen transport, has remained unresolved for many years, despite extensive study (50). None of the over 250 static crystallographic structures in the Protein Data Bank show an obvious path between the external solvent to the heme iron. Precisely identifying how ligands exit/enter the protein is important, as this small protein is integral for oxygen transport. Is there a single dominant pathway or do multiple pathways exist? Atomistic simulations may have the potential to fill the gap of ligand migration, but have so far failed to directly simulate ligand escape, or entry, at experimentally relevant conditions.

Forty years ago, Perutz and Matthews proposed that the distal histidine may rotate in order for the heme iron to become accessible in hemoglobin (96). The histidine acts like a gate to the sole entrance, controlling the accessibility of the heme. In Case and Karplus' simulations, their high-energy modeled ligands escaped primarily between the distal histidine and residues 67 and 68 (18). Crystal structures of myoglobin with large bound ligands have shown displacement of the distal histidine (68, 101). The size of the distal histidine has been shown to affect ligand entry and exit (9, 92, 109, 115).

Many X-ray crystallography studies and computational studies provide alternatives to the histidine-gate hypothesis for ligand entry/escape. Tilton et al. identified four xenon binding cavities, Xe1, Xe2, Xe3 and Xe4, in the interior of myoglobin (123). These four cavities are the basis of many studies contending that they are the intermediates in one of several exit pathways for a ligand. Elber and Karplus (47) observed through LES MD simulations at least five different pathways through which the high temperature ligand escaped from the protein. Experimental results have also correlated with MD results: random mutagenesis identified clusters of residues, not located near the distal histidine, that affect ligand binding kinetics (63). Based on observed CO escape dependence on Xe pressure, Nishihara *et al.* proposed a branched escape model, with 3 intermediate states (91).

The existence of internal cavities in myoglobin and hemoglobin have been analyzed

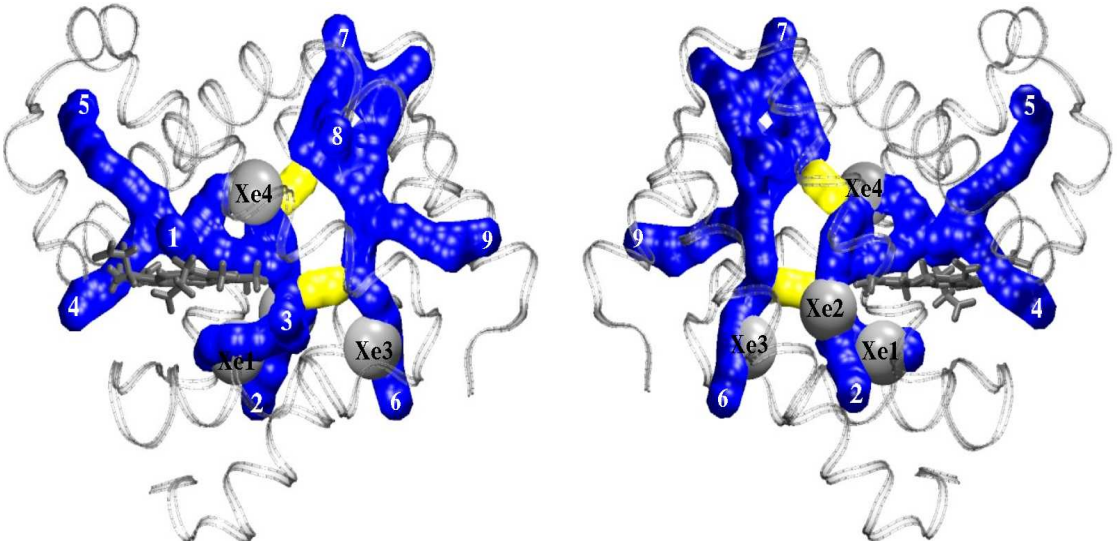
in snapshots from MD simulations to investigate ligand migration (12, 13, 64, 87). While the correlation of these cavities identified in single frames of MD simulations to experimentally identified cavities is beneficial for understanding internal ligand migration, the question of how the ligands enter/escape from myoglobin is still not completely understood. Very recently, Cohen *et al.* used three-dimensional maps of potential of mean forces of ligands over MD simulations to identify regions of favorable ligand placement, which supported multiple escape points (30).

This work tries reconciles the two seemingly mutually exclusive views by examining ligand migration into myoglobin at the atomic resolution. Our molecular dynamics simulations attempts to provide what other techniques, experimental or theoretical, have yet to resolve: the path of ligand migration between external solvent and the internal heme iron (Figure 1). Our conclusion is that ligands migrate along two major channels, from which they can escape or into which they can enter the molecule through several portals located near the surface (Figure 1). This conclusion is based on: (a) direct observation of ligand (CO) migration in and out in MD simulations of native myoglobin on relevant time scales; (b) direct observation of ligand migration and protein dynamics of mutant V68F myoglobin; and (c) a computational analysis of free volume fluctuations in native myoglobin using our PathFinder algorithm.

Table 3.1: Surface residues that close to the nine ligand exit/entry portals in native myoglobin shown in Figure 1. Highlighted residues are identified by mutagenesis (63) as greatly affecting ligand binding kinetics.

Channel	Portal	Residues near portal	Escapes	Entries	Total
1	1	**64 67 154**	7	5	31
	2	101 104 139 143 **146**	4	1	
	3	**71** 74 75 82 85 89	3	5	
	4	**43 45**	1	3	
	5	**34** 51 55	2	0	
2	6	7 134 **137**	2	2	7
	7	**16** 20 **24** 118 **119**	1	0	
	8	18 21 70	1	0	
	9	8 11 79	1	0	

We analyzed a total of 68 CO trajectories in 38 explicit solvent all-atom MD simulations of ligand entry and escape from native myoglobin. Each simulation is 90 ns



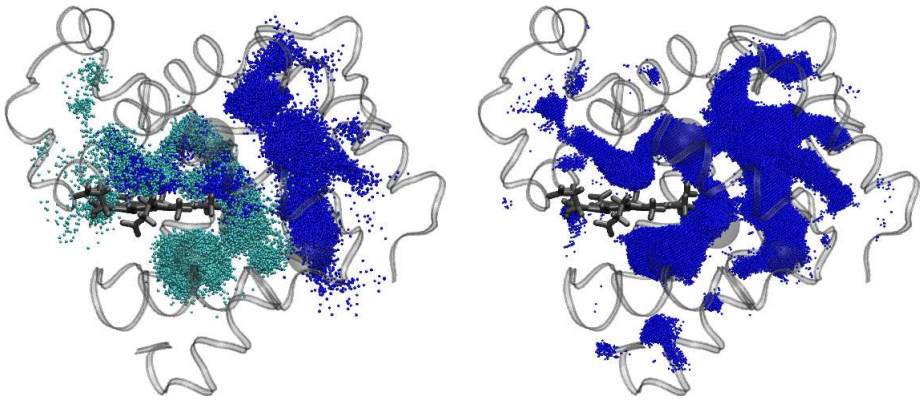
View 1 **View 2**

Figure 3.1: A schematic representation of the two channels of ligand migration and exit/entry in native myoglobin. The figures show two views (rotated 180^o) of CH1 (View 1 -left side), CH2 (View 1 - right side) and the connection between the two channels (yellow). Xenon cavities are labeled and shown as grey spheres. The heme is shown for orientation. The numbers in white correspond to the exit/entry portals described in Table 1.

long. In 20 simulations, we traced ligand migration after CO photodissociation. In each of the remaining eight simulations, we traced the diffusion of eight free ligands from the solvent into the protein matrix. In total, our MD simulations identified nine unique entrance/exit points, which we call portals. The subdivision of the pathways/portals into two channels - one major and one minor - comes naturally if one considers individual ligand trajectories (Figure 3.2(a)). Migration through CH1 occurs more often than through CH2: 31 of the 38 exits/entries occur through CH1 (Table 1). The few observed exits/entries via CH2 were also much slower, on average, by a factor of 2.4 relative to CH1. These trajectories agree with experimental data on ligand migration (108). Although only twelve of the 33 portal residues are identified in the Huang and Boxer study (63) as affecting ligand binding kinetics, the portal residues are located on the surface while many of the residues studied by mutagenesis are internal residues.

In 55% of our native myoglobin simulations, the CO escaped. In 33% of the native simulations in which CO was originally in the solvent, CO successfully migrated to distal heme pocket; 69% of these COs later escaped. The migration pathways traced by the COs that never escaped do in fact coincide with the ones from trajectories in which CO escaped/reached the heme.

Even though we have traced the trajectories of 38 COs that exited/entered myoglobin, MD simulation may not completely explore the universe available for ligand migration. To address this possibility, we developed the PathFinder algorithm, which systematically analyzes the dynamic free volume fluctuations of the protein. This analysis, PathFinder, computes cavities present in each static frame and integrates cavities across multiple frames to reveal connected migratory channels. Portions of the channel transiently occur in internal cavities, influenced by protein motion. By carefully connecting these cavities across time, PathFinder reveals the ligand migration pathways.



(a) (b)

Figure 3.2: Comparison of ligand trajectories and free volume fluctuation analysis. (a) The positions occupied by CO during the trajectories in which CO escapes/enters myoglobin. Light blue points are positions of CO that exited/entered via portals in CH1. Darker blue points are positions of CO that exited/entered via portals in CH2. (b) The results of PathFinder analysis in the native protein. Each point represents a cavity large enough to hold a ligand molecule.

The results of PathFinder analysis in native myoglobin shown in Figure 3.2(b) are in qualitative agreement with the migration points of the MD simulations (Figure 3.2(a)). Most free volume fluctuations occur in regions previously identified as

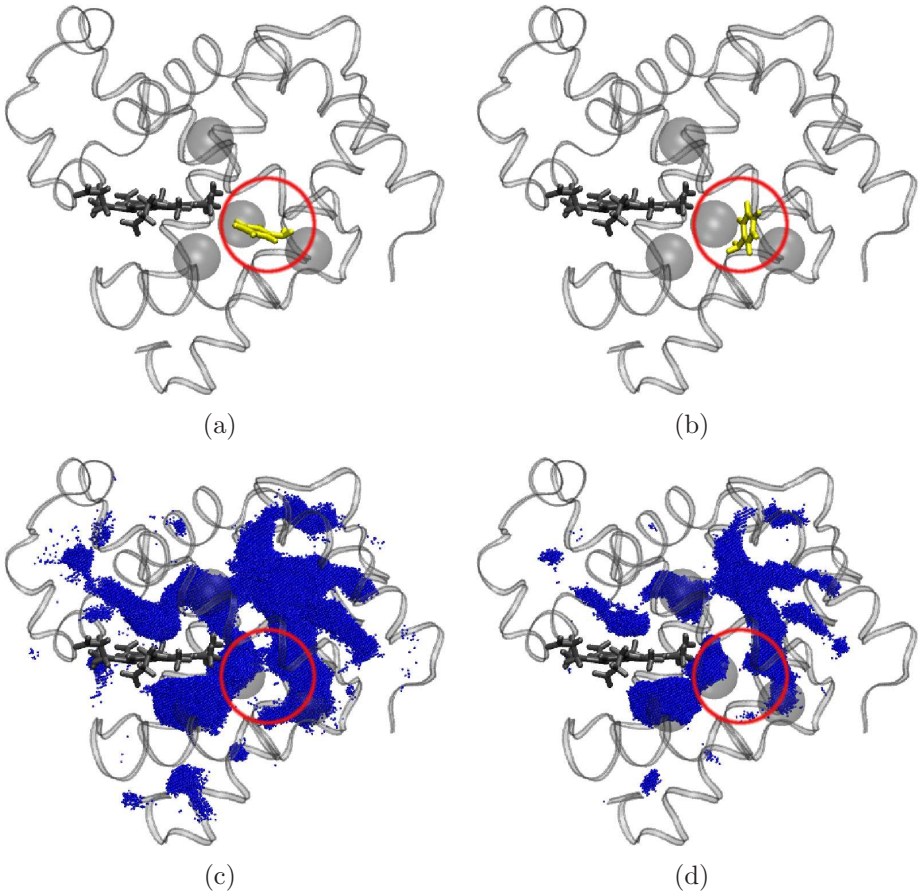


Figure 3.3: Effect of conformation of residue 138 in native myoglobin. (a) More frequent, and crystal structure, orientation of F138, which is parallel to the heme. (b) Less frequent orientation of F138, which is perpendicular to the heme. (c) The results of PathFinder analysis of all snapshots of the more frequent orientation of F138, which allows access between CH1 and CH2. (d) The results of PathFinder analysis of all snapshots of the less frequent orientation of F138, which blocks access between CH1 and CH2.

Xe1, Xe3, Xe4 and a fourth region that corresponds to a xenon cavity observed under alkaline conditions (107) (Supplementary Material). It is interesting to note that the protein fluctuations allow numerous points to accommodate a ligand molecule during the course of the 15 ns simulation.

In our MD simulations, the pattern of paths along which CO entered/exited through

CH1 is different from the pattern of paths along which CO entered/exited through CH2 (Figure 3.2(a)). The COs that exit/enter in CH1 traverse Xe1, Xe3 and Xe4 whereas the COs that exit/enter in CH2 traverse Xe2, Xe3 and Xe4. Since COs that exit/enter via CH1 portals do not enter Xe2 and COs that enter/exit via CH2 portals do not enter Xe1, we reasoned that a gating mechanism may control the access between CH1 and CH2. Visualization of the MD trajectories implicates residue F138 as the gate between CH1 and CH2. We divided the snapshots from the trajectories into two groups based on the orientation of F138, and then used PathFinder to identify the effect of the position of F138 on the connectivity between CH1 and CH2. When the phenyl ring of F138 is parallel to the heme (Figure 3.3(a)), CO is able to travel between CH1 and CH2 (Figure 3.3(c)). In about 5% of the simulation, though, the phenyl ring is oriented perpendicular to the heme (Figure 3.3(b)), effectively blocking access between CH1 and CH2 (Figure 3.3(d)). The result is consistent with the observation by Huang and Boxer that F138 affects ligand binding kinetics (63).

In native myoglobin, the analysis usingPathFinder identified numerous cavities that connect the distal pocket to the rest of the protein. The visualization of the MD simulations indicated that the orientation of V68 may influence the connection between the distal pocket and the rest of the protein. V68 adopts two different configurations: "open" in which the side chain methyl groups are oriented parallel to the heme (Figure 3.4(a)); and "closed" in which the side chain methyl groups are not parallel to the heme (Figure 3.4(b)). The PathFinder analysis on the snapshots of V68 in the "open" position show that migration between the distal pocket and the rest of the protein funnels through Xe4 but ligands are able to migrate from the distal pocket to the external solvent near the distal histidine (Figure 3.4(c)). The PathFinder analysis on the snapshots of V68 in the "closed" position show a more direct route becomes accessible, but the access to the external solvent near the distal histidine decreases (Figure 3.4(d)).

Additional support for our picture of ligand migration in the native myoglobin comes from the analysis of the V68F mutant. While the kinetics of tens of myoglobin mutants have been studied, only a select few can provide a clear-cut test case to study ligand migration. We chose to simulate the V68F mutant because V68 residue is highly conserved in myoglobin and the V68F mutant only affects the ligand pathways, not binding affinities (45, 99, 135). In the analysis of free volume fluctuations in the V68F mutant, the bulky phenylalanine side chain (Figure 3.5(a)) blocks the route between the distal pocket and the rest of the interior of the protein and also the access between the distal pocket and the external solvent near the distal histidine is decreased (Figure 3.5(b)). The orientation of the CB, CB1 and CG2 atoms

of F68 in the V68F mutant are similar to the orientation of these atoms in the "closed" conformation of the V68 of the native structure. The V68F MD simulations confirm this result: all of the ligands that do not escape within 90 ns remain in the distal pocket; of the ones that do escape the distal pocket, the F68 must rotate approximately 90^o from the position in the V68F crystal structure.

One possible explanation for the ligand migration channels is to look at the highly alpha-helical structure of the myoglobin protein. These helices are relatively rigid and therefore the volume fluctuations must occur in-between the helices (Figure 3.6). The access to the interior may only be accessible near the helix joints.

This study integrates (i) an analysis of free volume fluctuations in native myoglobin using the PathFinder algorithm, (ii) extensive MD simulations of both photodissociated and free CO in native myoglobin, and (iii) MD simulations of the V68F mutant myoglobin to provide a new characterization of migration pathways by which ligands reach the internal heme iron of myoglobin. This analysis reveals that ligand migration is characterized by two distinct channels with multiple portals. Migration via CH1 dominates migration via CH2. In CH1, exit/entry via portals near the distal histidine and near the Xe1 cavity dominate other portals. This atomic level characterization of ligand exit, entry and migration leads to the conclusion that there exists one major channel, with two dominant portals, and one minor channel for ligand migration in myoglobin.

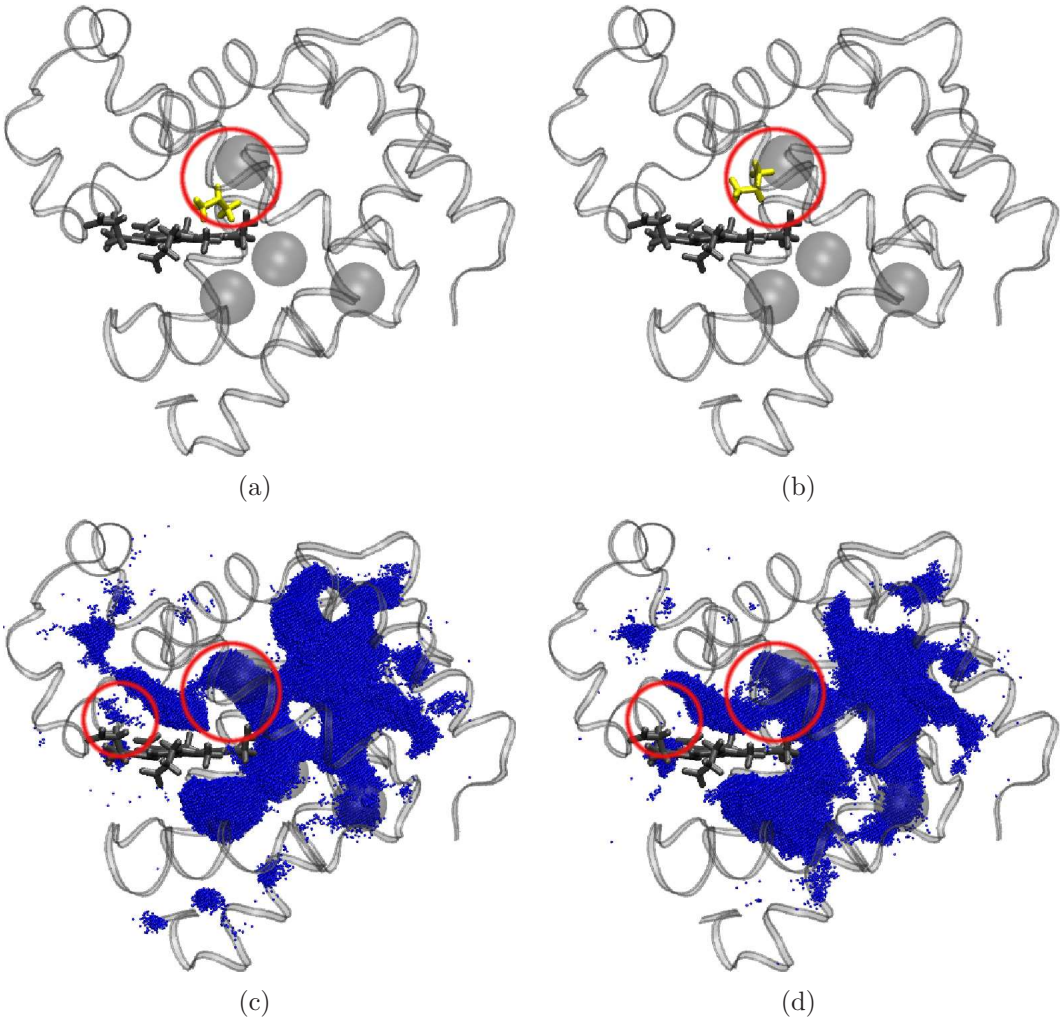
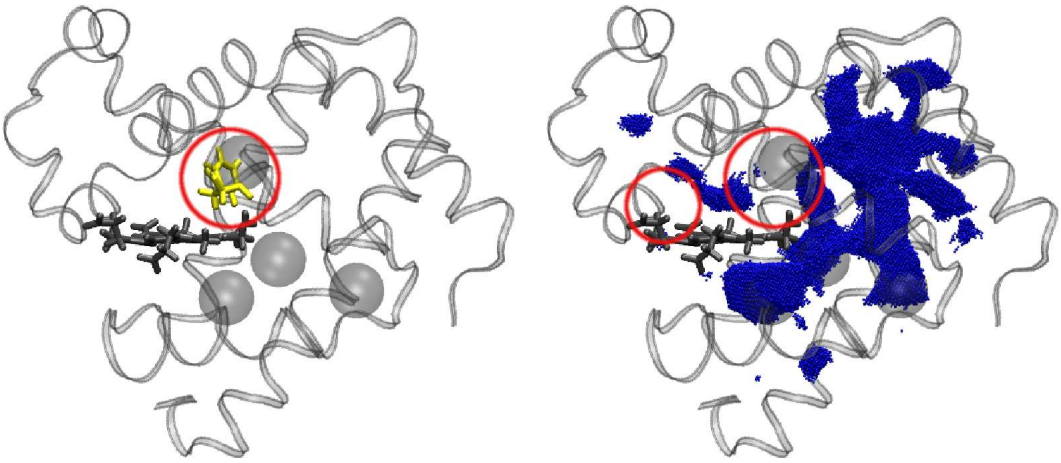


Figure 3.4: Effects of residue 68 in native myoglobin. (a) Sidechain conformation of V68 (yellow) in native myoglobin in crystal structure and in 18% of the MD simulation. (b) Sidechain conformation of V68 (yellow) in native myoglobin with a rotation of C-CA-CB-CG2 compared with crystal structure conformation (82% of the MD simulation). (c) PathFinder identifies a narrow connection, which goes through Xe4, between the distal pocket and the rest of the protein (right circle), and an opening exists between the distal pocket and the solvent (left circle). (d) PathFinder identifies more free volume between the distal pocket and the rest of the protein interior (right circle), but the opening between the distal pocket and the solvent is less open (left circle).



(a) (b)

Figure 3.5: Effect of residue 68 in V68F myoglobin. (a) Sidechain conformation of F68 (yellow) in V68F myoglobin. (b) PathFinder confirms that bulky phenyl side chain of F68 blocks access between the distal pocket and the rest of CH1 and CH2 (right circle) and cuts off all access between the distal pocket and the solvent (left circle).

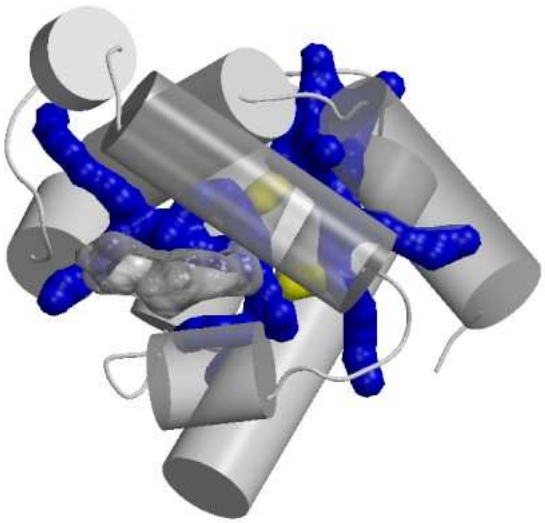


Figure 3.6: Structural origins of ligand migration channels and exit/entry portals in myoglobin. Migration occurs between the helices and the portals occur near the joints of the helices. Alpha-helices A, E and F and the heme molecular surface are transparent. Alpha-helices are rendered with a radius of 4.0 Å.

3.3 Supplementary Material

Methods

Structures

Two different structures were used for the native myoglobin: PDB ID 1A6G - carbonmonoxy myoglobin solved at 1.15Å resolution by X-ray diffraction (132); and PDB ID 2MB5 - carbonmonoxy myoglobin solved by neutron diffraction (25). The V68F mutant was represented by PDB ID 1MLJ - carbonmonoxy myoglobin solved at 2.0Å resolution by X-ray diffraction.

All water molecules found in the PDB 2MB5 structure were retained. In the other two structures the original water molecules found in the PDB files were stripped off. The initial placement of protons for 1A6G and 1MLJ X-ray structures was performed by MOLPROBITY (41) server; no dihedral flips were suggested at the validation step. As the 2MB5 structure already had most of the protons, these were retained. In agreement with earlier computational prediction and experiment(7), the HIS residues were set in the following states of protonation: HIS 12, 36, 81, 116 in the doubly protonated (charged) form, HIS 64 and 93 as singly protonated in the δ position, and HIS 24, 48, 82, 97, 113, 119 as singly protonated in the ϵ position. All other titratable groups were set in their standard protonation states, in agreement with the predictions based on the continuum solvent methodology(7). Missing heavy atoms and protons were added, and pKs of titratable groups computed with the help of `H++`(54) – web-based tool for structure completion and pK prediction.

3.3.1 Molecular Dynamics Simulations

AMBER (version 9) suite of programs(19) was used to further prepare the structures and perform all of the MD simulations described here. Each structure was solvated in a box of pre-equilibrated TIP3P water, with a 7 Å solvent buffer. A bond between the "C" atom of the CO ligand and the "Fe" atom of the heme was set during the preparation stage (*tleap* module of AMBER). All of the MD simulations were performed using *pmemd* module. Periodic boundary conditions were used, with the default parameters for the PME. The ff99 force-field was employed. (Parameters of the heme group and CO ligand are not part of the standard ff99 and are discussed below.) The SHAKE method was used to restrain hydrogen – heavy atom bond distances. The systems first underwent 100 steps of minimization, during which

all non-solvent atoms were restrained to their original (PDB) positions with a 5.0 kcal/mol/Å^2 harmonic potential. This was followed by two steps of equilibration, 200 ps each, under NPT conditions. All non-solvent atoms were constrained to their original positions with 1.0 kcal/mol/Å^2 and 0.1 kcal/mol/Å^2 harmonic potentials respectively. The temperature was gradually raised to 300 K during the first 200 ps step of the equilibration, and then kept constant during the second step. The integration time-step was 2 fs. The average temperature of the system was maintained by weak coupling (via the Berendsen algorithm) to a heat bath with coupling constant of 2 ps. The pressure coupling constant was 2 ps. All of the production MD runs were performed at 300 K and constant pressure. All of them share the common preparation, minimization and equilibration protocol. The production MD simulations can be separated into three groups. Features of the MD protocols common to all of the simulations are given above, and the differences are presented below.

Generation of snapshots for the PathFinder analysis

The ensemble of structures used in the PathFinder analysis was generated by 3 MD simulations, each corresponding to one of the structures described above. Each simulation is 15 ns in length, with the protein and solvent coordinates saved every picosecond. The initial structures were prepared as described above. To model the bound ligand we used the parameters corresponding to the 6-coordinated Fe, see below. The first 200 ps of the production run for each simulation was not included in the PathFinder analysis.

Simulation of Photodissociated CO

We use the "instantaneous" model for photodissociation(47, 86), in which the Fe-C bond is cut and the heme force field is switched instantly from that of the 6-coordinated Fe to that of the 5-coordinated Fe, see below. The LJ interactions with the ligand also change instantly at $t = 0$, corresponding to the absence of the bond between the "C" and "Fe" atoms. Each complete trajectory is 90 ns. All of the trajectories were different, which was achieved by using randomized initial velocities (drawn from a thermal distribution at $300K$) at $t = 0$. Solvent and solute coordinates were saved every 10 ps.

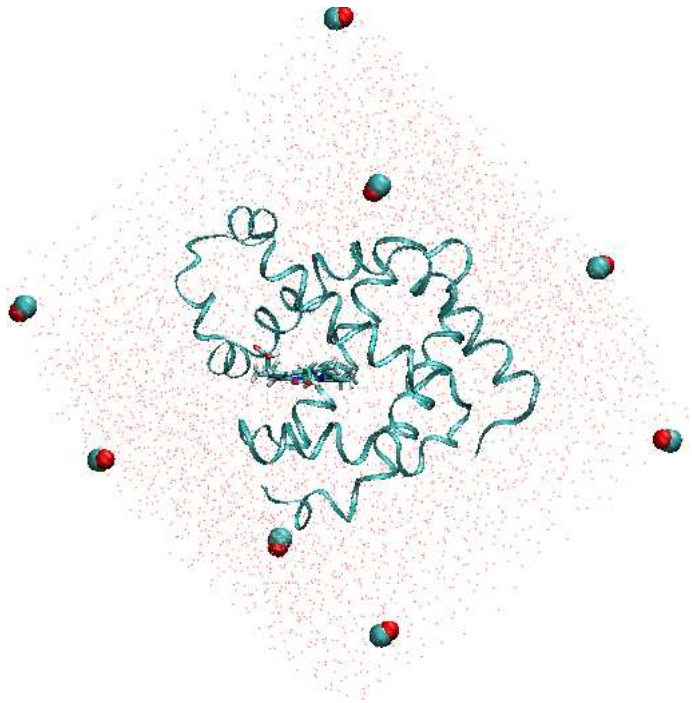


Figure 3.7: The starting positions of the 8 ligands used in the "CO entry from solvent" MD simulations. A planar view. Each ligand is initially placed at least 15 Å from the protein surface.

Simulation of ligand entry from solvent

The starting structures for these simulations were snapshots from the photodissociated CO trajectories of each of the two native myoglobin structures described above. The snapshots in which the ligand had escaped and moved at least 15 Å away from the protein were chosen. Seven additional CO molecules were added to each solute box by replacing the water molecules at the edges of the box, see Fig. 3.7. Including the originally escaped CO, the eight COs are initially located outside of the protein, at least 15Å from the protein surface. All water molecules were then deleted and the protein along with the 8 added ligands were re-solvated in a box of pre-equilibrated TIP3P water as described above. The minimization, equilibration, and production MD protocols are described above, and correspond to the 5-coordinated heme iron since the heme is now in "deoxy" state. Each MD simulation is 90 ns long, with the solvent and solute coordinates saved every 10 ps. All of the trajectories were

different, which was achieved by using randomized initial velocities (drawn from a thermal distribution at $300K$) at $t = 0$. A total of 6 MD simulations of this type were performed, 3 per each native structure. This gives a total of (3+3)×8 = 48 individual ligand diffusion trajectories.

heme and ligand force field parameters

Force field parameters for the ligated (6-coordinated Fe) are available with the AMBER distribution, these include the LJ parameters of the CO ligand. To model the unligated (5-coordinated Fe) heme, we modified the above parameters as follows: Fe-NB and FE-NO bond lengths were increased by 0.1A, and Fe-Np bond lengths were increased by 0.1431A. These changes correspond to the difference between Fe-N bonds for 5- and 6- coordinated Fe published in Table 1 of Ref. (79). We employ the two-site carbon monoxide model that was used earlier in a similar context(86). It consists of two atoms bearing small charges (0.021 au) of opposite sign, located at the ligand atomic positions. This model approximately reproduces the experimental dipole moment of carbon monoxide as well as several other experimental trends. In a systematic comparative analysis(86), Meller *at al.* did not find a clear evidence that a more sophisticated "three site" model should be preferred to the two-site one used here, at least as long as one is not interested in short time scale dynamics. The general conclusion of insensitivity of the ligand migration pathways to this level of detail of the ligand charge model is consistent with recent recent work of of Cohen *et al.*(30).

3.3.2 CO escapes/entries

The simulations in which CO escaped/entered and reached the iron atom are noted in Table 3.2. The amount of time in which CO traveled between the surface and the iron is noted for each entry/escape. These times are plotted in Figure 3.8. The simulations are divided into groups by the portal in which they enter/escape.

Table 3.2: Simulations in which CO escapes/enters native myoglobin and the residue number of the CO. The times (in ns) at which the CO reached the surface and at which the CO reached the iron atom, along with the total time of travel between the two, and the portal through which it escapes/enters are shown.

Simulation	Res.	Entry				Exit			
		Surf	Fe	Time	Portal	Fe	Surf	Time	Portal
1A6G.IG10	153	-	-	-	-	0	36.71	36.71	1
1A6G.IG20	153	-	-	-	-	0	0.20	0.20	5
1A6G.IG30	153	-	-	-	-	0	36.70	36.70	2
1A6G.IG40	153	-	-	-	-	0	12.44	12.44	2
1A6G.IG60	153	-	-	-	-	0	1.4	1.4	5
1A6G.IG0.eq2	153	-	-	-	-	0	42.7	42.7	3
1A6G.IG50.eq2	153	-	-	-	-	0	38.58	38.58	6
1A6G.8CO	158	11.05	11.26	0.21	4	11.26	29.33	18.07	4
1A6G.8CO.prop	153	14.39	15.21	0.82	1	15.21	33.51	18.31	1
1A6G.8CO.prop	160	87.78	89.99	2.22	1	-	-	-	-
1A6G.8CO.prop.IG10	153	15.72	16.57	0.86	1	16.57	22.33	5.77	1
1A6G.8CO.prop.IG10	158	3.86	4.10	0.26	1	4.10	23.82	19.73	1
1A6G.8CO.prop.IG10	160	1.31	1.93	0.03	1	1.92	4.59	2.67	1
2MB5.IG0	155	-	-	-	-	0	76.15	76.15	7
2MB5.IG20	155	-	-	-	-	0	17.10	17.10	1
2MB5.IG30	155	-	-	-	-	0	11.62	11.62	1
2MB5.IG50.eq2	155	-	-	-	-	0	45.91	45.91	3
2MB5.8CO	156	12.74	37.35	24.62	3	37.35	53.58	16.24	9
2MB5.8CO	157	24.30	48.24	23.99	6	-	-	-	-
2MB5.8CO.prop	156	37.70	38.40	0.70	6	38.40	48.00	9.55	2
2MB5.8CO.prop	159	69.60	72.70	3.31	3	-	-	-	-
2MB5.8CO.prop	160	45.80	53.50	7.90	3	53.50	65.40	11.87	3
2MB5.8CO.prop	161	33.74	33.88	0.14	2	33.88	50.84	16.96	6
2MB5.8CO.prop	161	64.10	66.60	2.62	3	66.60	78.05	11.34	2
2MB5.8CO.prop	162	33.80	34.10	0.50	4	34.10	63.20	28.81	8
2MB5.8CO.prop.IG10	158	8.61	12.00	3.40	3	-	-	-	-
2MB5.8CO.prop.IG10	162	44.80	52.38	7.59	4	-	-	-	-

3.3.3 Validation

RMSD

The RMSD of backbone atoms during a 90 ns simulation in which CO was bound to the iron atom and the RMSD of backbone atoms during a 90 ns simulation in which 8 COs were initially outside the protein are shown in Figure 3.9. The overall RMSDs shown the the simulation is relatively stable and does not deviate much from the

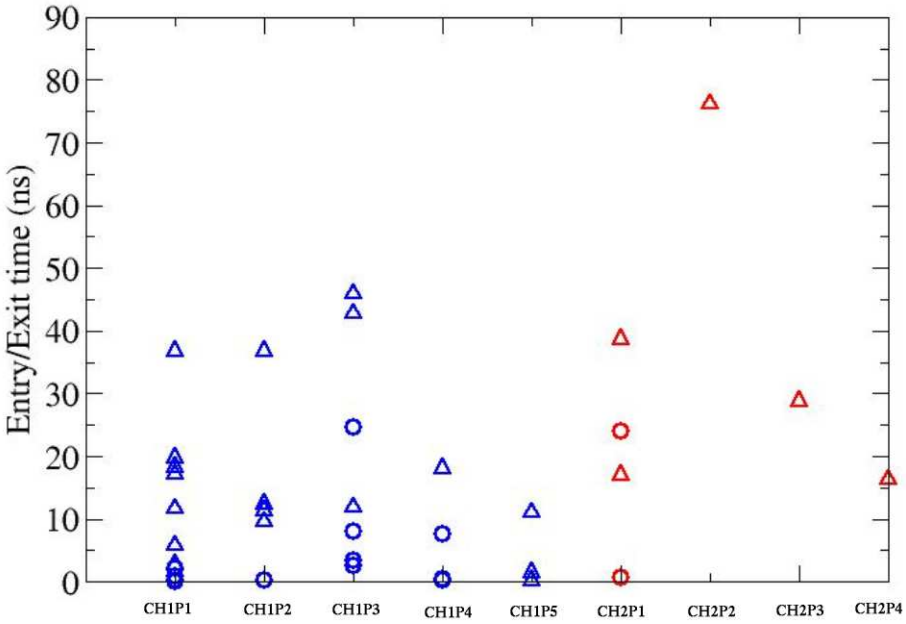


Figure 3.8: Ligand entry and exit times by channel and portal. Circles - time it took for the ligand to enter the protein and reach the distal pocket. Triangles - time it took for the ligand to escape from the distal pocket to the protein surface. The string 'CHiPj' denotes channel i and portal j.

initial crystal structures.

The RMSD per residue during the two native myoglobin trajectories were measured (Figure 3.10). The residues from both trajectories that exhibited the largest fluctuations (RMSD *gt* 1) were: 1, 2, 4, 6, 8, 11, 12, 21, 24, 31, 40, 41, 47, 49, 50, 59, 63, 64, 68, 69, 75, 76, 83, 85, 89, 91, 97, 98, 101, 102, 104, 109, 118, 119, 122, 126, 136, 138, 140, 141, 146, 151, and 152. The residues located in the interior of the protein are: 24, 69, 76, 89, 104, 138. Residues 76 and 138 are located near the connection between CH1 and CH2. The conformation of residue 138 is shown to affect the accessibility between CH1 and CH2 at the lower connection point; the conformation of residue 76 does slightly affect the accessibility between CH1 and CH2 at the upper connection, but not as clearly as residue 138 affects the lower connection point.

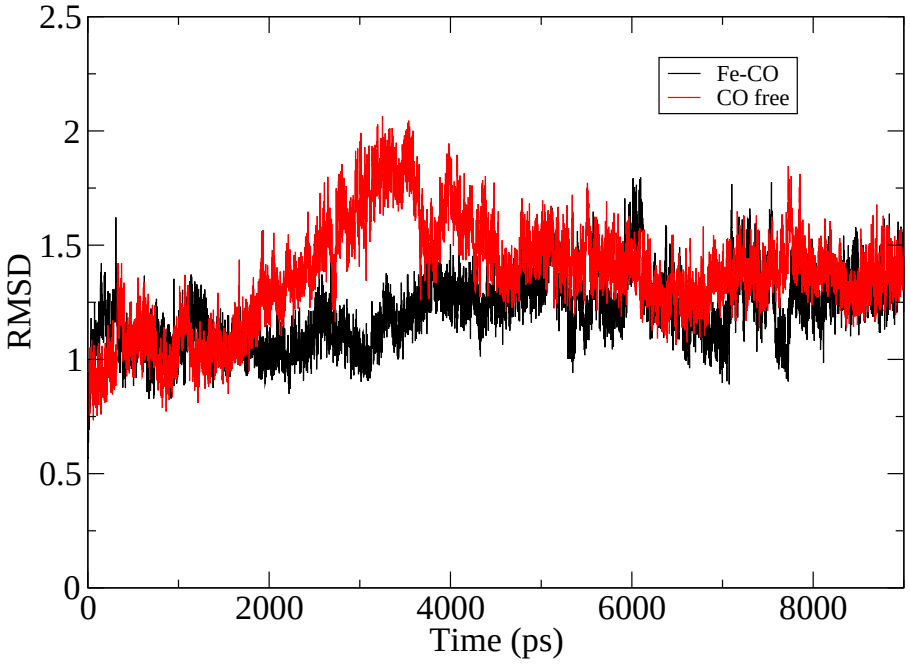


Figure 3.9: RMSD of backbone atoms during the 90 ns simulations of carbonmonoxy myoglobin (black) and myoglobin with photodissociated CO (red).

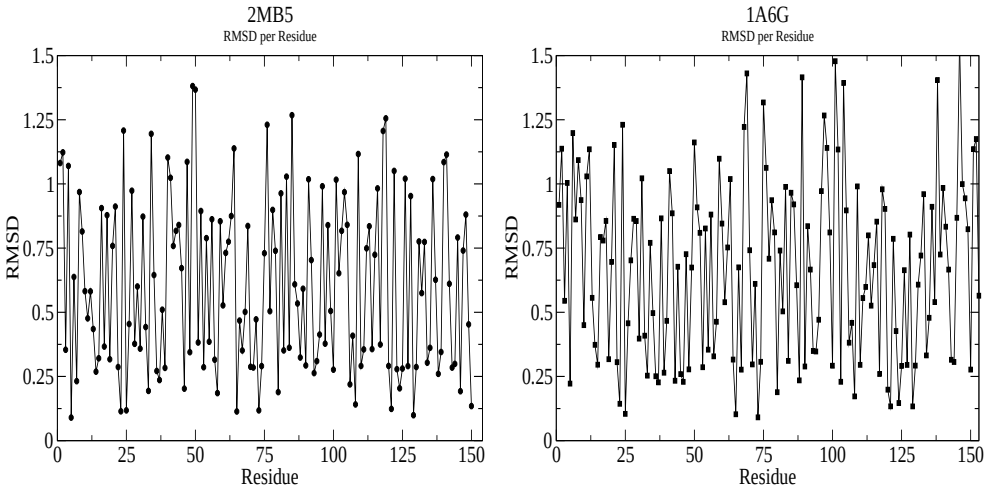


Figure 3.10: RMSD per residue during two native myoglobin trajectories.

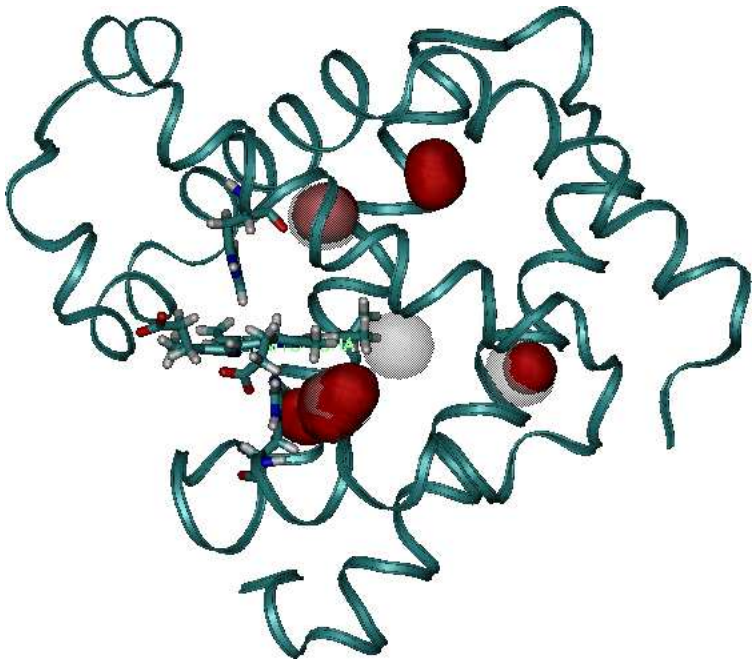


Figure 3.11: The points in native myoglobin that occur with the highest frequency (over 2500 times in 11,200 snapshots) are shown in red. Three of the four regions correspond to Xe1, Xe3 and Xe4, while the fourth region is near a xenon binding site observed by Schoenborn (107).

Chapter 4

Computational Investigation of Mechanism and Substrate Specificities of Two Family 1 β-Glucosidases

Jory Z. Ruscio, Asim Esen and David R. Bevan

4.1 Abstract

Computational methods have been used to characterize substrate specificity and investigate the catalytic mechanism in two family 1 β-glucosidases, Glu1 and Dhr1. Glu1, from *Zea mays*, hydrolyzes DIMBOA-glc, its natural substrate, and is inhibited by dhurrin. Dhr1, from Sorghum, hydrolyzes dhurrin, its natural substrate, and is inhibited by DIMBOA-glc. Experimental data do not provide a complete understanding as to how these two proteins interact with their respective substrate and inhibitor. Molecular dynamics (MD) simulations have been performed on Glu1 with DIMBOA-glc, Glu1 with dhurrin, Dhr1 with DIMBOA-glc, Dhr1 with dhurrin, and a Glu1-Y473F mutant with dhurrin. Our MD simulations also have implications on the current view of the mechanism of retaining glucosidases. A distance of $\sim$5.5Å between the catalytic glutamates has generally been accepted. The results

of the MD simulations indicate that the distance between the catalytic glutamates is closer to 3.0Å which supports an in-plane protonation hypothesis. Additionally, the observed interactions between ligands and proteins provide information on the substrate specificities of each protein.

4.2 Introduction

β-glucosidases (β-D-glucoside glucohydrolase, EC 3.2.1.21) occur in all organisms and play essential roles in numerous biological processes. Their function in plants is of basic importance, as they are involved in functions such as defense, lignification (42), growth, signaling and cell wall metabolism. β-glucosidases are members of families 1 and 3 of glycosyl hydrolases (60) and hydrolyze substrates of type G-O/S-X, where G is a glycosyl residue and X is either another glycosyl residue or a non-glycosyl aglycone moiety. G and X are connected in the β-configuration by either an oxygen(O) or sulfur(S) atom. Two glutamate residues, which are highly conserved in the FNEP and ITENG motifs, act as a general acid/base and a nucleophile, respectively. The substrate is cleaved by a two-step retaining mechanism. The first step involves nucleophilic attack at the anomeric carbon (C1) of the glycosyl moiety; the concerted protonation of the glycosidic oxygen by the general acid yields a covalent glycosyl-enzyme intermediate and release of the leaving group (131). The deprotonated glutamate then acts as a general base, extracting a hydrogen from a water molecule. The activated water molecule attacks at the anomeric carbon, causing release of the glycosyl moiety from the enzyme, and the catalytic glutamate residues return to their original protonation states. The anomeric configuration of the glycosyl residue is retained from substrate to product (144).

In order for enzymatic catalysis to occur, a substrate must be oriented such that the anomeric carbon of the glycosyl moiety and the catalytic glutamates are positioned appropriately. These distances have been approximated based on crystal structures of retaining enzymes. The distance between the catalytic glutamates was determined to be $\sim$5.5Å (40). To date, eleven Family 1 glycosyl hydrolase unligated native structures have been solved (Table 4.1). While the distance between the δ carbons of the glutamate residues is about 5Å, the ϵ oxygens are much closer (around 3.5Å).

Crystal structures of glucosidases complexed with transition state inhibitors depict the protonation occurring above the plane of the glucose moiety, rather than above (127, 130) (Figure 4.1). For in-plane protonation, the catalytic glutamates would need to be positioned much closer than the conventionally accepted 5.5Å (59).

Table 4.1: Distances between the catalytic residues in other Family 1 glycosyl hydrolases.

Protein	Organism	PDB ID	Catalytic Residues	Cδ-Cδ	Oϵ-Oϵ	Year
cyanogenic β-glucosidase	*Trifolium repens*	1CBG	183/397	5.21	3.75	1995 (6)
6-phospho-β-galactosidase	*Lactococcus lactis*	1PBG	160/375	4.82	3.35	1995 (136)
β-glucosidase	*Sulfolobus solfataricus*	1GOW	206/387	4.93	4.06	1997 (1)
β-glucosidase A	*Bacillus Polymyxa*	1BGA	166/352	5.1	3.22	1998 (104)
myrosinase	*Sinapis alba*	1MYR	187/409	5.02	3.61	1997 (16)
β-glucosidase	*Zea mays*	1E1E	191/406	4.97	3.49	2000 (35)
β-glucosidase	*Bacillus circulans*	1QOX	166/355	4.24	3.23	2000 (56)
β-glycosidase	*Thermus nonproteolyticus*	1NP2	164/338	5.2	3.65	2003 (134)
β-glucosidase	*Sorghum bicolor*	1V02	189/404	4.92	3.47	2004 (129)
β-glucosidase	*Pyrococcus horikoshii*	1VFF	155/324	5.23	4.37	2004 (2)
myrosinase	*Brevicoryne brassicae*	1WCG	167/374	5.18	3.89	2005 (66)

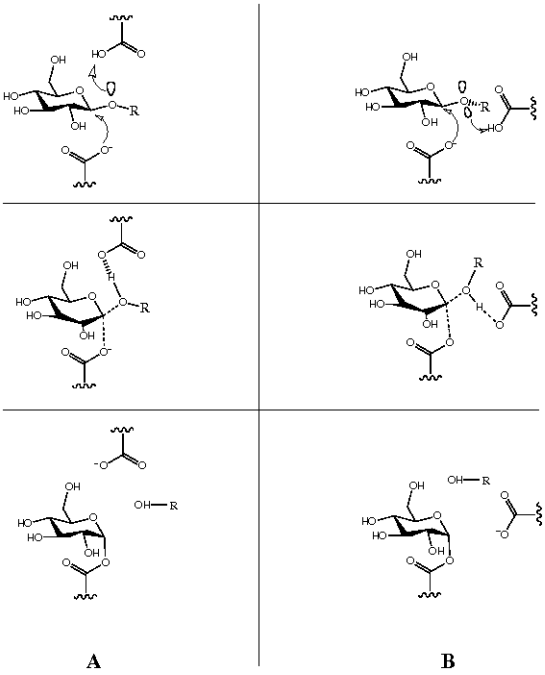


Figure 4.1: First step (glycosylation) of double-replacement retaining mechanism. The catalytic residues may be oriented for (A) above plane protonation or (B) in plane protonation (39, 59).

A shorter distance indicates that the catalytic glutamates would be on the same side of the glycosidic bond, and not on opposite sides, as sometimes depicted. The distance between the ϵ oxygens of the catalytic glutamic acids in the proteins with solved structures provides support for in-plane protonation. These distances are similar to those described by Heightman et al (59). Additionally, the optimal distance between the anomeric carbon and nucleophile for nucleophilic attack should be $<$ 3.25Å (14, 80, 105, 124).

The discrepancies of the catalytic glutamate distance between each other in models with inhibitor and those with substrate indicate that the mechanism of glycosyl hydrolases, and more specifically, family 1 β-glucosidases, needs to be examined in more depth. If the interaction distance of the catalytic glutamates can be defined when substrate is present, it will be much easier to define whether or not a small molecule is a potential substrate.

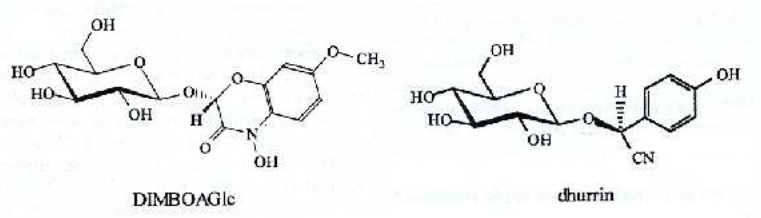


Figure 4.2: Structures of DIMBOA-glucoside and dhurrin

Some β-glucosidases exhibit a wide range of substrate specificities (5, 6, 10, 134) while others have a quite narrow substrate specificity (27, 62). Kinetic studies with native and chimeric enzymes provide much useful information about the substrate specificities of Glu1 and Dhr1. Glu1 from *Zea mays* and Dhr1 from *Sorghum bicolor* have $\sim$70% sequence identity (27), yet their substrate specificities are extremely different. In addition to its natural substrate, 2-O-β-D-glucopyranosyl-4-hydroxy-7-methoxy-1,4-benzoxanzin-3-one (DIMBOA-glc), Glu1 can hydrolyze a wide variety of artificial substrates (5, 34), but is inhibited by dhurrin (K_i *of 0.076mM* (26))(Figure 4.2) . Dhr1 is highly specific for its natural substrate dhurrin (β-D-glucopyranosyl-oxy-(S)-p-hydroxymandelonitrile) and is inhibited by DIMBOA-glc (K_i *of 0.009mM* (26)).

The crystal structures of Glu1 and Dhr1 provide information that can help identify residues important in determining substrate specificity. Czjzek et al. (35) showed that the aglycone binding pocket of Glu1 is defined by Phe198, Phe205, Phe466 and W378, which form a hydrophobic slot. The aglycone moiety is believed to be

Table 4.2: Hydrogen bonding between glucose atoms and protein atoms in the glycosyl-enzyme myrosinase structure (16), glucotetrazole in myrosinase (15), dhurrin in Dhr1-E189D (129), and DIMBOA-glc in Glu1-E191D (35) (hydrogen bonds in glucotetrazole-myrosinase complex, dhurrin-Dhr1-E189D complex and DIMBOA-glc-Glu1-E191D complex from Verdoucq *et. al* (129)).

Glucose atom	Glycosylated MYR	Glucotetrazole in MYR	Dhurrin in Dhr1-E189D	DIMBOA-Glc in Glu1-E191D
O2	Asn186-$N_\delta 2$	Gln187-$O_\epsilon 1$ Asn186-$N_\delta 2$	Asp189-$O_\delta 2$ Asn188-$N_\delta 2$ His143-$N_\epsilon 2$	Asp191-$O_\delta 2$
O3	His151-$N_\epsilon 2$ Gln39-$O_\epsilon 1$	His141-$N_\epsilon 2$ Gln39-$O_\epsilon 1$ Gln39-$N_\epsilon 2$	His143-$N_\epsilon 2$ Gln39-$O_\epsilon 1$ Gln39-$N_\epsilon 2$ Trp461-$N_\epsilon 2$	His152-$N_\epsilon 2$ Glu406-$O_\epsilon 1$ Glu406-$O_\epsilon 2$
O4	Gln39-$N_\epsilon 2$ Glu464-$O_\epsilon 1$	Gln39-$N_\epsilon 2$ Glu464-$O_\epsilon 1$ Trp457-$N_\epsilon 1$	Gln39-$N_\epsilon 2$ Glu464-$O_\epsilon 1$ Trp453-$N_\epsilon 1$	Gln38-$O_\epsilon 1$ Trp465-$N_\epsilon 1$ His142-$N_\epsilon 2$
O6	Glu464-$O_\epsilon 2$	Glu464-$O_\epsilon 2$	Glu460-$O_\epsilon 2$	Glu464-$O_\epsilon 2$ Glu464-$O_\epsilon 1$ Gln38-$N_\epsilon 2$ Trp457-$N_\epsilon 1$

positioned by Trp378 through aromatic stacking interactions (35, 36, 146). This positioning is thought to determine the orientation of the glucosyl moiety in relation to the catalytic glutamates. Incorrect positioning of the glucosyl moiety in relation to the catalytic glutamates will inhibit catalysis, as is evident with Glu1 inhibitors dhurrin and *p*-nitrophenyl β-D-thioglucoside (35, 36).

The residues involved in aglycone recognition are different in Dhr1 than in Glu1. The three phenylalanines that compose one side of the active site opening are replaced with valine, leucine and serine. The opening of the active site is small and circular in Dhr1, compared to the slot-like opening in Glu1. Although native Glu1 cannot hydrolyze dhurrin, some chimeric and altered Glu1 proteins have gained the ability to hydrolyze dhurrin (26). Analysis of chimeric protein with the fewest residues different between Glu1 and Dhr1 eventually helped identify residue Tyr473 (Glu1) and Phe469 (Dhr1) as important for conferring dhurrin hydrolysis ability to Glu1. The opposite substitution in Dhr1 does not confer DIMBOA-glc hydrolysis ability, however. Glu1 appears uses aromatic stacking to interact with the aglycone while Dhr1 utilizes polar interactions (129), and therefore the difference may be due to the fact the two proteins use different modes of aglycone recognition.

With the availability of protein-ligand complexes, the conformation of the ligand in the active site can now be examined. To some extent, the interactions between the protein and ligand are similar to the interactions in the glycosylated protein (Table 4.2). But, for instance, the O6 atom in DIMBOA-glc in Glu1-E191D makes several interactions with the protein that are not seen in glycosylated MYR, glucotetrazole in MYR or dhurrin in Dhr1-E189D. Do atoms of DIMBOA-glc in Glu1 really interact with several protein atoms, or do the definition of these interactions need to be further studied, taking into account the dynamics of the protein-ligand complex?

Additionally, the conformation of the glucose moiety appears to have implications on the mechanism and the protein residues with which it interacts. Several of the crystal structures of protein-ligand complexes show distortion of the glycosyl moiety of the ligand. A proposed pathway, based on these distortions, describes a $^{1}S_{3} \rightarrow {}^{4}H_{3} \rightarrow {}^{4}C_{1}$ glucose conformation (Fig 4.1) in which the $^{4}H_{3}$ is a transition state structure (38) . The electron density of the glycone binding pocket of Glu1 with DIMBOA-glc was not clear enough to accurately define the glucose conformation (35). The conformation could be modeled as either the $^{4}C_{1}$ chair, which is the ground state conformation, or the $^{1}S_{3}$ skew-boat, which is a proposed Michaelis conformation (40). The crystal structure of Glu1 with glucotetrazole, which is a transition state-mimicking inhibitor, depicts the glucose in the $^{4}H_{3}$ conformation. Additionally, the crystal structure of Dhr1 with dhurrin depicts the glucose in the $^{1}S_{3}$ skew-boat conformation (129). It

has been proposed that Glu1 can bind substrate with the glucose as $^4C_1$ or as the skew-boat, and the glycone binding pocket can stabilize either one, while Dhr1 can only bind substrate with the glucose in the $^1S_3$ conformation. It has been proposed that for retaining glycosyl hydrolases that bind with a distorted $^1S_3$ glycosyl moiety conformation, the glycosyl moiety in the transition state would adopt a $^4H_3$ conformation (38). A computational study involving a retaining family 11 xylanase showed that the enzyme binds the substrate more tightly when the sugar is the in skewed conformation than in the $^4C_1$ chain conformation (74).

While the kinetic data show whether or not an enzyme can hydrolyze a ligand, it does not provide any information as to how or why one ligand is a substrate and one is an inhibitor. Likewise, while the crystal structures have increased our understanding of the enzyme-ligand complexes, these are still only static snapshots of dynamic systems. Additionally, all of the enzyme-ligand complexes were crystallized with an altered form of the enzyme (the catalytic acid/base glutamate changed to an aspartate), rendering the enzyme inactive, and thus not a mechanistically relevant representation.

In order to gain a better understanding of what is occurring at the atomic level, we have applied computational techniques to the Glu1 and Dhr1 systems. The use of molecular dynamics simulations has enabled us to model the protein-ligand interactions and help answer the question of how ligands affect protein dynamics and in turn, determine catalytic activity. We first quantify the importance of the distance between the catalytic glutamate residues, as well as the distance between the nucleophilic glutamate and the anomeric carbon. Then, we determine how interactions between the ligand and the protein and the ligand and the solvent may affect binding. The atomic fluctuations of the active site are then examined so that we may have a better understanding of the changes that are caused by the inhibition of dhurrin in Glu1. We then investigate the affects of catalysis on the conformation of the glycosyl moiety. Finally, we extend some of the substrate specificity based information gained from Glu1 and Dhr1 to 41 putative β-glucosidases in *Arabidopsis thaliana*.

4.3 Methods

In this investigation, molecular dynamics was the primary tool used to characterize substrate specificity in Glu1 and Dhr1. The crystal structures of Glu1 and Dhr1 with and without ligands are available. All MD setup and MD simulations were performed using the AMBER 7 suite of programs (21). To fully examine the differ-

ences in the active site architecture, many perturbations of the two proteins were used for simulations. Catalytically inactive forms, such as when an aspartic acid is substituted for one of the glutamic acids and when the acid/base catalytic glutamic acid is unprotonated, as well as protein-inhibitor complexes were simulated in order to compare interactions with catalytically active forms. The simulations performed as well as the abbreviations used in the rest of the paper are noted in Table 4.3.

Preparation of Structures for MD: The structures of Glu1 with ligand (1E56) and Dhr1 with ligand (1V03) obtained from the Protein Data Bank were of altered forms (acid/base glutamate residue changed to an aspartate residue) or of the proteins with no ligand (1E1E and 1V02). Y473F was created by deleting the hydroxyl group of tyrosine 473 and renaming the residue PHE. The DIMBOA-glc and dhurrin structures were geometry optimized and partial charges calculated using Hartree-Fock 3-21G in Gaussian98 (52). To prepare for MD, the ligand was docked into the native enzyme by superimposing the structures of an altered enzyme (e.g., 1E56) with ligand and a native enzyme (e.g., 1E1E) and then removing the altered enzyme. To alter the protonation state of the acid/base catalyst, in order to accurately describe the catalytic state, residue 191 was changed from GLU to GLH, to adhere to AMBER notation. The cysteine residues involved in a disulfide bond (103) were changed from CYS to CYX, and connected using the bond command in Leap. All the systems were processed with Leap to add hydrogens, add counterions and solvate in a box of TIP3P water.

Running MD: The following conditions were used: default particle mesh Ewald (PME) parameters; SHAKE on the hydrogen atoms; a 2fs timestep; and a 9Å cutoff applied to long-range interactions. Before running MD, the systems were equilibrated and minimized. The first step involved 100 ps equilibration of all the water molecules and ions at 300K while holding the protein and ligand atoms frozen. Next the water molecules and ions were subjected to 300 steps of steepest descent minimization while again holding the protein and ligand atoms frozen. Finally, the entire system underwent 300 steps of steepest descent minimization. To begin MD, the entire system was slowly heated from 0 to 300K in 30ps at constant volume. Constant volume MD was continued for an additional 50ps, after which the system was run under constant pressure conditions. Twelve simulations were carried out to 3.5 ns. Eight simulations involving catalytically inactive states were only of 1.5 ns. These simulations were not carried out to 3.5 ns because the results at 1.5 ns were already sufficient to show why they are not catalytically active and had provided enough data to compare with catalytically active complexes.

Analysis of MD: All MD analysis was performed using the ptraj module of the

Table 4.3: Simulations performed and abbreviations used in the paper. †Catalytically active

	Simulation	Protein Description	Ligand
3.5 ns	Glu1	Glu1	N/A
	†Glu1_DIMglc	Glu1	DIMBOA-glc
	Glu1_dhr	Glu1	dhurrin
	Dhr1	Dhr1	N/A
	†Dhr1_dhr	Dhr1	dhurrin
	Dhr1_dhr_tors	Dhr1 w/torsion angle restraint on C2 C1 O5 C5 of glucosyl moiety	dhurrin
	Dhr1_DIMglc	Dhr1	DIMBOA-glc
	Dhr1_DIMglc_tors	Dhr1 w/torsion angle restraint on C2 C1 O5 C5 of glucosyl moiety	DIMBOA-glc
	Y473F	Glu1 w/Tyr473 changed to Phe	N/A
	†Y473F_DIMglc	Glu1 w/Tyr473 changed to Phe	DIMBOA-glc
	†Y473F_dhr	Glu1 w/Tyr473 changed to Phe	dhurrin
	†Chim21_dhr	Glu1 w/^{466}FAGFTERY^{473} w/Dhr1 ^{462}SSGYTERF^{469}	dhurrin
	F198V	Glu1 w/residue Phe198 changed to Val	N/A
	F198V_DIMglc	Glu1 w/residue Phe198 changed to Val	DIMBOA-glc
	F198V_dhr	Glu1 w/residue Phe198 changed to Val	dhurrin
1.5 ns	Glu1_a/b_un	Glu1 w/acid/base catalyst unprotonated	N/A
	Glu1_a/b_un_DIMglc	Glu1 w/acid/base catalyst unprotonated	DIMBOA-glc
	Glu1_nuc_prot_DIMglc	Glu1 w/nucleophile protonated (acid/base unprotonated)	DIMBOA-glc
	Glu1_E/D_DIMglc	Glu1 w/Glu191 changed to Asp	DIMBOA-glc
	Glu1_E/D_un	Glu1 w/Glu191 changed to Asp (unprotonated)	N/A
	Glu1_E/D_un_DIMglc	Glu1 w/Glu191 changed to Asp (unprotonated)	DIMBOA-glc
	Glu1_E/D_un_dhr	Glu1 w/Glu191 changed to Asp (unprotonated)	dhurrin
	Dhr1_E/D_dhr	Dhr1 w/Glu189 changed to Asp	dhurrin

AMBER package. Calculation of hydrogen bonds used the default values of a distance cutoff of 3.0Å and angle cutoff of 120°.

Glucosyl moiety conformation analysis: The charges and conformations of the glucosyl moiety were calculated using Jaguar (67). The DIMBOA-glucoside after sander minimization of the Glu1_DIMglc complex was used as the initial structure. A geometry optimization was performed while the glycosidic bond (between the anomeric carbon and the glycosidic oxygen) was held constant. After each optimization, the glycosidic bond length was increased by 0.2Å, and the structure was subjected to another geometry optimization. This cycle was continued until the energy stabilized (final bond length = 3.5Å). The geometry optimizations were done with B3LYP and basis set 6-31G**.

Active Site Profiling: Three Active Site Profiles of the Glu1 and Dhr1 active sites were created using the two catalytic glutamate residues in Glu1 and Dhr1 (191/406 and 189/404, respectively) as the reference residues. The Glu1 and Dhr1 sequences were aligned with 39 putative β-glucosidases from *Arabidopsis thaliana*. The residues of the *A. thaliana* sequences that aligned with the residues defining the Glu1/Dhr1 Active Site Profiles were included in the Active Site Profile for β-glucosidases. Clustering of these alignments was done using the ClustalW server (http://www.ebi.ac.uk/clustalw/).

The first, and largest, Profile59, based on 59 amino acids from Glu1 and Dhr1, included the glycone and aglycone binding pockets. The second, Profile39, included the glycone binding pocket (red residues in sequences and Figure 4.8). The third Active Site Profile, Profile20, contained only the aglycone binding pocket (yellow residues in sequences and Figure 4.8). The clustering of the partial sequences is compared with the clustering of the full sequences (142). The seven putative pseudogenes (At5g16580,At1g60260, At1g60270, At2g44480, Ag1g51490, Ag5g48375 and At3g18070) were not included in the clustering.

```
Glu1AATQRFHWFNEPQTFTFLAFDVNFGLNYYNWIITENGIGDVDRYFAWSLDNEATERYGI
Dhr1AATQRFHWFNEPETFCVLALNVCFGINYYNWIITENGMGDIDRYFAWSLDNESTERFGI
```

4.4 Results

Orientation of the catalytic glutamates: Several structural features within the active site of the glucosidases examined in these simulations were analyzed. In each of

	Simulation	191O_ϵ2-406O_ϵ2		C1-406O_ϵ2	
		Initial	Final	Initial	Final
3.5 ns	Glu1	4.94	3.32 ±0.54	n/a	n/a
	†Glu1_DIMglc	**4.94**	**2.60±0.13**	**5.92**	**3.74±0.25**
	Glu1_dhr	**4.94**	**2.76±0.32**	**3.95**	**3.60±0.25**
	Dhr1	3.24	5.31 ±0.51	n/a	n/a
	†Dhr1_dhr	**3.24**	**3.80±0.34**	**3.56**	**3.38±0.20**
	Dhr1_DIMglc	3.24	4.34±0.40	4.19	4.61±0.14
	Y473F	4.94	3.65±0.61	n/a	n/a
	†Y473F_DIMglc	3.49	3.51±0.29	3.53	5.00±0.39
	†Y473F_dhr	**4.64**	**3.13±0.29**	**3.95**	**3.25±0.14**
	†Chim21_dhr	4.08	3.62±0.43	5.03	4.84±0.21
	F198V	3.64	3.41±0.60	n/a	n/a
	F198V_DIMglc	3.64	3.73 ±0.55	5.21	6.57 ±1.05
	F198V_dhr	3.64	4.17 ±0.21	4.65	5.35 ±0.27
1.5 ns	Glu1_a/b_unprot	4.94	3.51±0.25	n/a	n/a
	Glu1_a/b_unprot_DIMglc	4.94	4.81±0.42	5.92	4.66±0.15
	Glu1_nuc_prot_DIMglc	4.94	3.12±0.45	5.92	4.77±0.36
	Glu1_E/D_DIMglc	6.24	4.40±0.47	4.46	5.21±0.25
	Glu1_E/D_unprot	6.24	4.12±0.47	n/a	n/a
	Glu1_E/D_unprot_DIMglc	6.24	7.92±0.94	4.46	5.82±0.81
	Glu1_E/D_unprot_dhr	3.46	7.58±1.29	4.59	6.52±1.25
	Dhr1_E/D_dhr	4.16	4.95±0.61	3.63	5.19±0.39

Table 4.4: Initial and average distances between two distances that are affected by substrate specificity: the O_ϵ2 oxygen atoms of catalytic glutamate residues; and the nucleophilic O_ϵ2 and C1 of the glucosyl moiety over the course of the simulations. Bold rows indicate simulations in which the distance between the catalytic glutamates and the distance between the nucleophile and C1 are close to the expected values for catalysis to occur. †Indicates catalytically active complex.

the 20 different simulations, two key distances were measured: the distance between the catalytic glutamates (O_ϵ2 to O_ϵ2) and, if relevant, the distance between the nucleophilic glutamate and C1 of the glucosyl moiety (O_ϵ2 to C1) (Table 4.4). These distances provide information as to how the residues are oriented in catalytically active complexes versus catalytically inactive complexes and may distinguish between the above-plane and in-plane protonation hypotheses. Also, since the nucleophile

must be close enough to attack at the anomeric carbon, the C1-O2 distance may distinguish between catalytically active and catalytically inactive complexes. For a complex to be considered catalytically active, the distance between the catalytic glutamates must be less than 5Å and the distance between the nucleophilic glutamate and the anomeric carbon should be around 3.25Å.

The distances measured in the systems that only underwent MD for 1.5 ns were sufficient to determine that these complexes did not represent active enzymes. In Glu1_a/b_unprot_DIMglc, Glu1_E/D_DIMglc, Glu1_E/D_unprot, Glu1_E/D_unprot_DIMglc, Glu1_E/D_unprot_dhr and Dhr1_E/D_dhr, the distances between the catalytic glutamates are too great for catalysis to occur. In Glu1_nuc_prot_DIMglc, although the distance between the catalytic glutamates may be appropriate, the distance between the nucleophilic glutamate and the anomeric carbon is too great.

Of the 12 simulations (eight of which contain a ligand) carried out to 3.5 ns, only four of the protein-ligand complexes show the distance between the catalytic glutamates and the distance between the nucleophile and C1 of the ligand to be close enough for catalysis to occur: Glu1_DIMglc, Glu1_dhr, Dhr1_dhr and Y473F_dhr. Three of these four complexes are known to be catalytically active. Glu1_dhr is not catalytically active, but the effective binding is consistent with the ability of dhurrin to inhibit Glu1. Although the initial distances between the anomeric carbon and the nucleophilic glutamate in these four complexes prior to MD vary between ∼3.5Å to 5.9Å, only the Dhr1_DIMglc distance is significantly different than the others after MD (4.61Å vs. ∼3.4Å). After MD, the glucosyl moieties in Glu1_DIMglc, Glu1_dhr, Dhr1_dhr and Y473F_dhr are positioned at a similar distance from the nucleophilic glutamate, most likely the position optimal for catalysis; conversely, the glucosyl moiety in Dhr1_DIMglc is rotated, such that the anomeric carbon aligns with the ring oxygen of the other simulations (Figure 4.3).

One thing to note is that the distances between the catalytic glutamates in Glu1_DIMglc and Glu1_dhr are slightly closer than the distances in Dhr1_DIMglc and Dhr1_dhr, regardless of ligand. The distances in Chim21_dhr, Y473F_DIMglc and Y473F_dhr are intermediate. This suggests that enzyme specific interactions, rather than protein-ligand interactions, have greater influences on the slight variations in this distance.

In the five catalytically active complexes (Glu1_DIMglc, Dhr1_dhr, Y473_DIMglc, Y473F_dhr and Chim21_dhr), the distances between the catalytic glutamates support in-plane protonation of the glycosidic oxygen, as this distance is much closer than

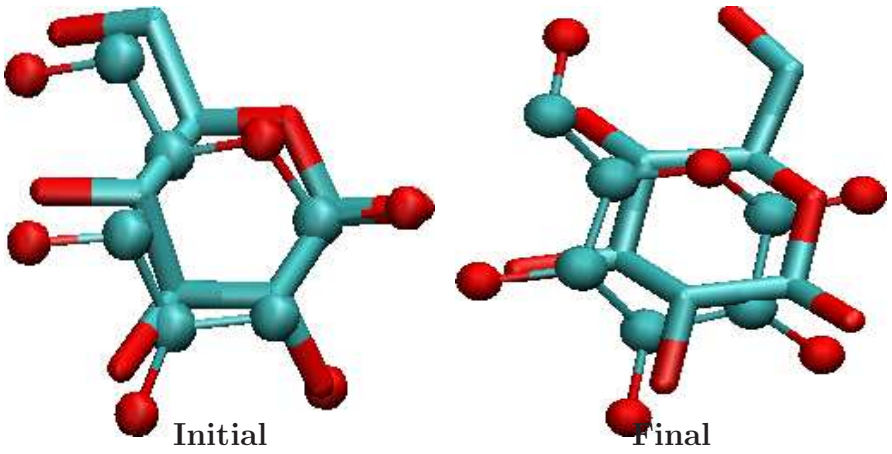


Figure 4.3: Orientation of the glucosyl moiety of the ligand of Glu1_DIMglc (licorice) and Dhr1_DIMglc (cpk) after minimization (left) and after 3.5 ns of MD (right). The anomeric carbon in Glu1_DIMglc is better oriented for catalysis to occur, while the anomeric carbon in Dhr1_DIMglc is too far from the nucleophilic glutamic acid. While not shown, the anomeric carbons in the Glu1_dhr, Dhr1_dhr and Y473F_dhr simulations align well with Glu1_DIMglc.

the conventionally accepted distance of ∼5Å for retaining glycosyl hydrolases. The distance between the catalytic glutamates in these five simulations was significantly less than 5Å (2.60, 3.80, 3.51, 3.13, 3.62).

In the simulations of the protein in absence of ligand (Glu1, Dhr1, Y473F, F198V, Glu1_a/b_unprot and Glu1_E/D_unprot), the distances between the catalytic glutamates is slightly larger than when a ligand is present (e.g. Glu1 - 3.32 vs Glu1_DIMglc - 2.63 and Glu1_dhr - 2.76), but still much less than 5Å. Additionally, the standard deviation of the distance between the catalytic glutamates is slightly higher (Table 4.4) than when is ligand is present, indicating that they are able to fluctuate more freely.

The F198V complex is known to have drastically decreased catalytic ability (128). In the F198V simulations (with and without ligand), the distances between the catalytic glutamates are over 1Å greater than of the native enzyme in complex with a ligand. Additionally, the nucleophilic glutamate is located too far away from the anomeric carbon for catalysis. Taken together, these distances would indicate the the F198V complexes are catalytically inactive, correlating with the experimental data.

Table 4.5: Persistence of hydrogen bonds between glucose moiety and protein complexes over the 3.5 ns simulations (percentage). Only interactions that occur in at least 20% of the simulation are reported.

Glucose atom	Protein atom	Glu1		Dhr1		Y473F		F198V	
		DIMglc	dhr	DIMglc	dhr	DIMglc	dhr	DIMglc	dhr
O-2	Glu191-$H_\epsilon 2$	25.5							
	Glu406-$O_\epsilon 2$	97.1	95.4	99.1	97.9	99.9	96.5		
O-3	Gln38-$O_\epsilon 1$				50.8		20.6	35.9	81.5
	Glu406-$O_\epsilon 1$			55.8					
	Trp465-$N_\epsilon 1$	45.8	59.1				59.9	22.2	78.2
O-4	Gln38-$O_\epsilon 1$			47.8					
	Gln38-$N_\epsilon 2$	69.7	48.3				22.5	29.1	68.4
	Trp457-$N_\epsilon 1$	58.2				24.4	65.5	33.1	
	Glu464-$O_\epsilon 1$		48.3		81.1		90.1	72.1	70.7
	Glu464-$O_\epsilon 2$		25.8					21.4	33.0
O-6	Asn462-$N_\delta 2$	28.7				21.0			
	Glu464-$O_\epsilon 1$	97.6		63.4				35.8	77.2
	Glu464-$O_\epsilon 2$		91.0		85.7	90.7	97.7		

Hydrogen bonding networks between ligand and protein: The measurement of the persistence of hydrogen bonds reveals information about how the protein binds the ligand. Evident by Table 4.5, several interactions between the ligand and protein are present in catalytically active complexes. The nucleophilic $O_\epsilon 2$ remains hydrogen bonded to the O-2 of the glycosyl moiety; this interaction is maintained throughout all of the simulations except F198V complexes, regardless of ligand. The absence of this hydrogen bond in the F198V simulations provides further confirmation of the inactivity of these complexes.

Y473F_dhr, which is able to hydrolyze dhurrin, maintains a hydrogen bonding pattern similar to both DIMBOA-glc in Glu1 and dhurrin in Dhr1. The Y473F_dhr interactions similar to Glu1_DIMglc are: O-3 with Trp465-$N_\epsilon 1$ and O-4 with Gln38-$N_\epsilon 2$ and Trp-457-$N_\epsilon 1$. The Y473F_dhr interactions similar to Dhr1_dhr are: O-3 with Gln38-$O_\epsilon 1$, O-4 with Glu464-$O_\epsilon 1$ and O-6 with Glu464-$O_\epsilon 2$.

Interactions with solvent: Several studies have shown that hydrogen bonds with the solvent may provide additional information about the mechanism of binding (31, 73, 110). In the enzyme-substrate complexes, hydrogen bonding to solvent molecules occurs less frequently than in the enzyme-inhibitor complexes. DIMBOA-glucoside can form two hydrogen bonds, one through the 4-hydroxy group and one through the 3-one group. Glu1_DIMglc maintains one hydrogen bond between the 4-hydroxy group and Trp378 and one hydrogen bond between the 3-one group and water. The aglycone in Dhr1_DIMglc maintains two hydrogen bonds with water, with the 3-one group and the 4-hydroxy group. When DIMBOA-glucoside is an inhibitor of Dhr1, it hydrogen bonds with two solvent molecules, rather than when it is a substrate and only hydrogen bonds with one solvent molecule. Similarly, dhurrin can form two hydrogen bonds, one through the nitrile and one through the phenyl hydroxyl group. In Dhr1_dhr and Y473F_dhr, only the phenyl hydroxyl group hydrogen bonds with water; in Glu1_dhr, both the phenyl hydroxyl group and the nitrile hydrogen bond with water. When dhurrin is an inhibitor, it hydrogen bonds with two solvent molecules, rather than when it is a substrate and only hydrogen bonds with one solvent molecule. More hydrogen bonding with the solvent can contributed favorably to the enthalpy, but the unfavorable entropy contribution may overcome the enthalpy contributation, and therefore increased solvent ordering may contribute to unfavorable binding, contributing to inhibition.

Additionally, the number of solvent molecules that are able to move close to the acid/base catalytic glutamate in Dhr1_DIMglc (Figure 4.4) is much higher than in any of the other simulations. An average of three water molecules are able to occupy space within 3.5 Å of the acid/base catalytic glutamate in Dhr1_DIMglc, whereas the

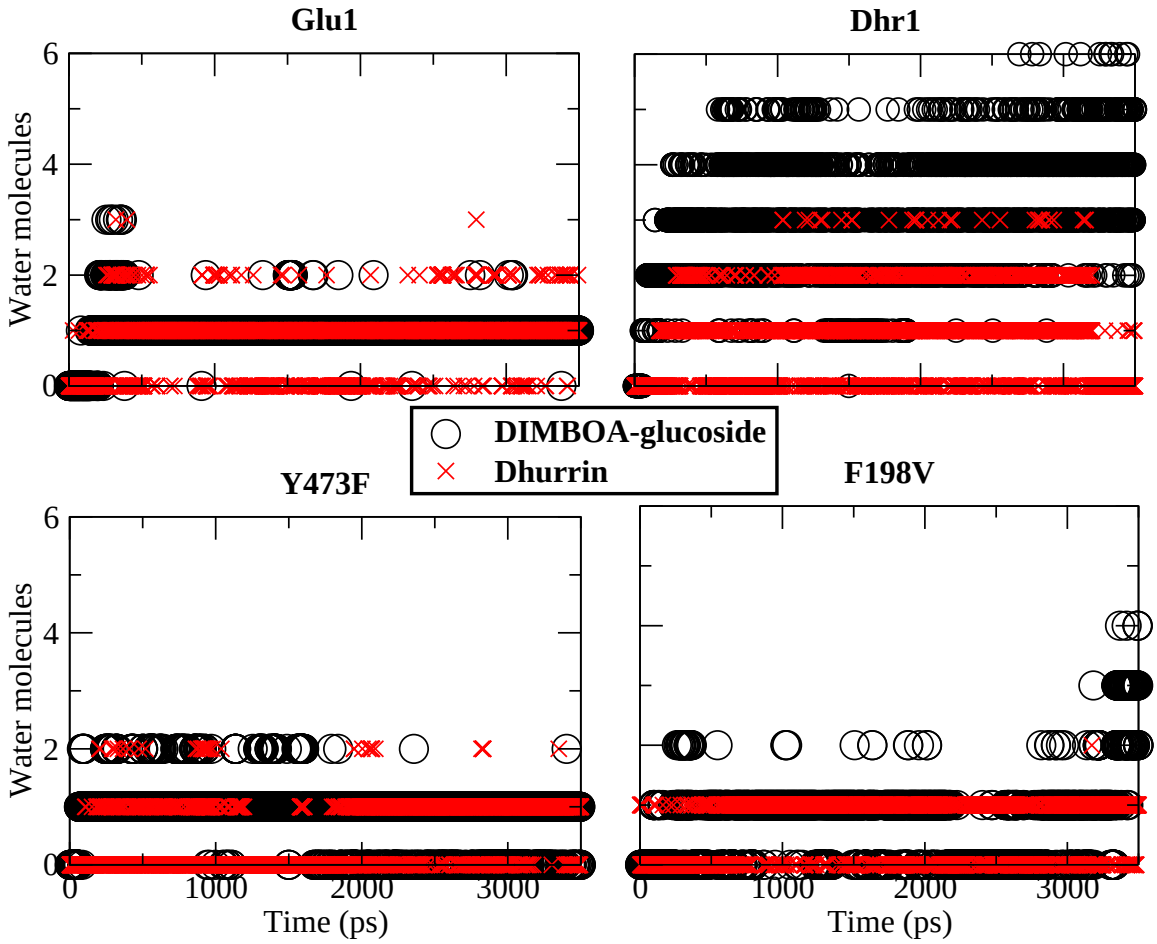


Figure 4.4: Number of water molecules within 3.5 Å of acid/base catalytic glutamate over the course of the simulations.

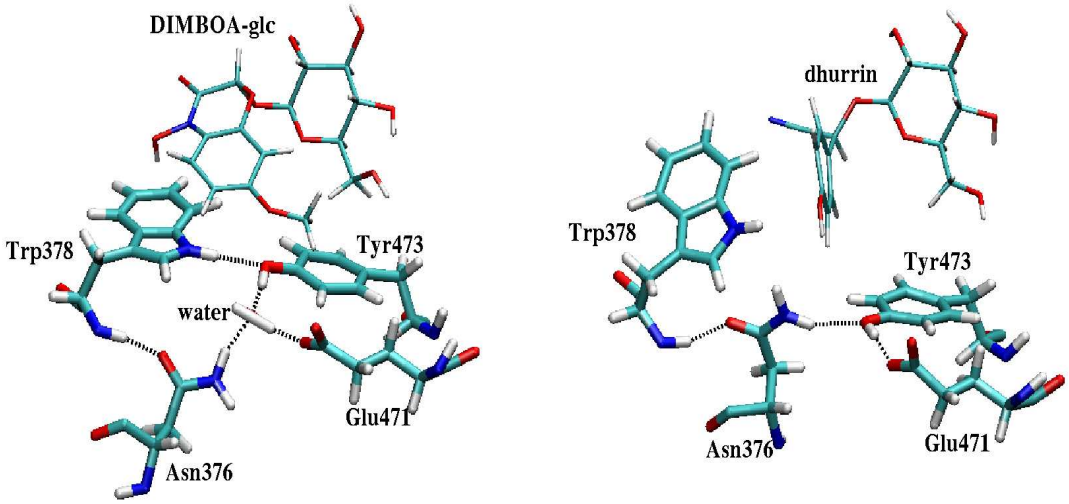
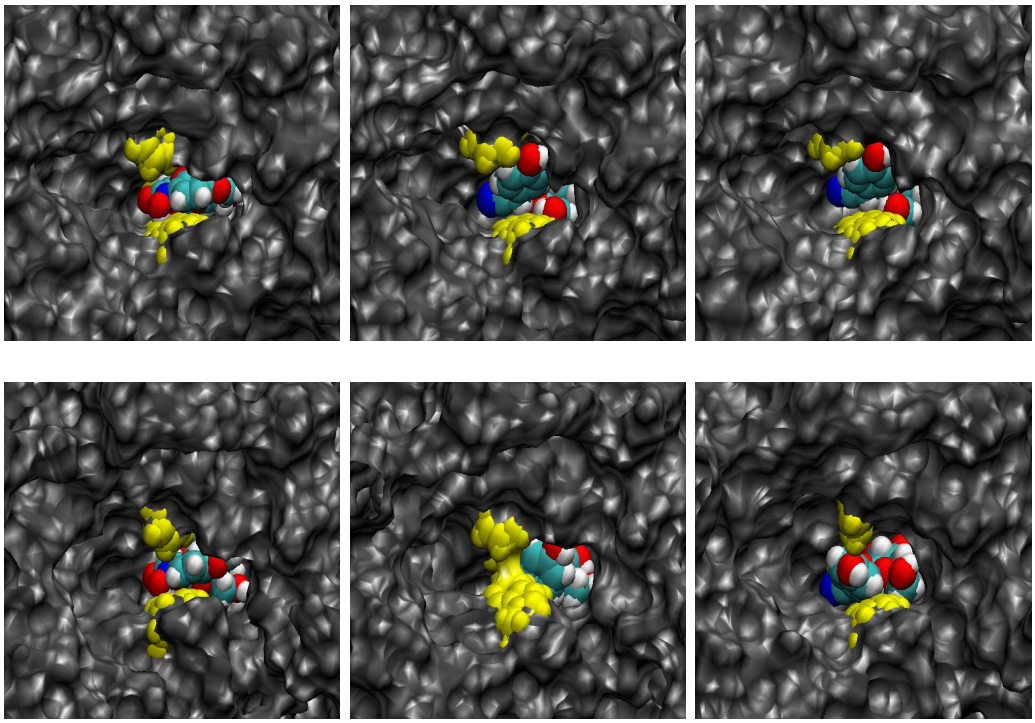


Figure 4.5: Effect of dhurrin on active site of Glu1. **A** - In Glu1_DIMglc simulation, Tyr473 directly hydrogen bonds with Trp378 and can either form direct or indirect hydrogen bonds with Asn376 and Glu471. **B** - In Glu1_dhr simulation, due to lack of interaction between aglycone and Trp378, Tyr473 cannot directly hydrogen bond with Trp378. Tyr473 can then shift slightly and form direct hydrogen bonds with Asn376 and Glu471.

average number of water molecules within that distance in Glu1_DIMglc, Glu1_dhr, Dhr1_dhr, Y473F_DIMglc, Y473F_dhr, F198V_DIMglc and F198V_dhr is about one, which is consistent with the need for a water molecule to be near the acid/base glutamic acid during the second step of the mechanism. .

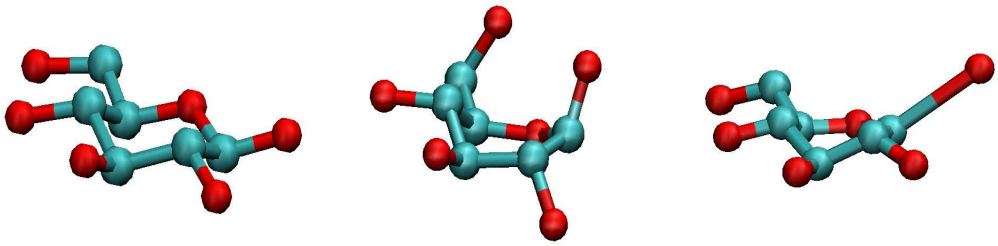
Insight into dhurrin inhibition in Glu1: The ligand-protein interactions greatly affect catalysis ability. Glu1 residue Y473 has been identified in several studies as contributing to the conformational flexibility of the active site (119, 128, 146). In Glu1, this tyrosine residue's hydroxyl group has the ability to form a hydrogen bond with other close residues or water; in Dhr1, this residue is a phenylalanine and does not have the ability to form a hydrogen bond due to the lack of the hydroxyl group.

During the Glu1_DIMglc simulation, Trp378 is stabilized through its interaction with the 4-hydroxy group of the DIMBOA moiety, and a hydrogen bond between Trp378 and Tyr473 stabilizes the active site opening. Tyr473, situated near the C-6 of



Glu1_DIMglc **Glu1_dhr** **Y473F_dhr**

Figure 4.6: The active sites of Glu1_DIMglc, Glu1_dhr and Y473F_dhr after minimization (top) and after 3.5 ns of MD (bottom). The molecular surface of the proteins is shown in grey, residues 198 and 378 are colored yellow and the substrates are shown as van Der Waals in atom coloring In Glu1_dhr, the active site appears to become closed by the interaction of Trp378 and Phe198. Pictures were generated in VMD (65).



A **B** **C**

Figure 4.7: Different conformations of glucose. Initial conformation of glucose moiety of **A)** DIMBOA-glucoside in chair $^4C_1$, **B)** dhurrin in skew boat $^1S_3$ **C)** and geometry optimized DIMBOA-glucoside with the C1-O1 bond length extended to 2.7Å in half-chair $^4H_3$. The aglycone moieties are omitted for clarity.

glucose, also forms direct or indirect hydrogen bonds, through water molecules, with Asn376 and Glu471 (Figure 4.5A). These four interactions stabilize the active site opening (Figure 4.6), so the entrance to the active site remains accessible.

In the Glu1_dhr simulation, the aglycone does not interact with Trp378, and therefore Trp378 is not positioned to hydrogen bond with Tyr473. Tyr473 then has enough freedom to shift slightly from beneath the C-6 of glucose and form a direct hydrogen bond with Asn376 and Glu471 (Figure 4.5B). Without the Trp378-Tyr473 hydrogen bond, Trp378 is free to fluctuate, and Asn376 shifts slightly upwards. These motions cause the active site opening to become distorted, as Trp378 and Phe198 become closer and block the entrance to the active site (Figure 4.6).

In the Y473F simulation, the lack of a side chain hydroxyl group on residue Phe473 has an impact on the dynamics of the protein with affect the ability of the protein to hydrolyze dhurrin. Phe473 is unable to form hydrogen bonds and therefore Asn378, Trp378 and Glu471 are not constrained by Phe473. The entrance to the active site remains open, and the fluctuations of Trp378 do not distort the ligand or accessibility of the active site(Figure 4.6).

Glucosyl moiety conformations: The conformation of the glucosyl moiety in the initial DIMBOA-glucoside structure from the crystal structure was $^4C_1$ and the conformation of the glucosyl moiety in the initial dhurrin structure was $^1S_3$ (Figure 4.7). It was proposed the Glu1 can bind the ligand in the $^4C_1$ or a skewed conformation but Dhr1 may only bind the ligand in the skewed conformation, as it is closer to the

Table 4.6: The change in the dihedral angle that characterizes the conformation of the glucose and the charges of the C1, O5 and O1 atoms when the glycosidic bond (between C1 and O1) is lengthened.

dist	C2-C1-O5-C5	C1	O5	O1
1.42	-53.9	0.50468	-0.44677	-0.36869
1.70	-55.9	0.43459	-0.41143	-0.38214
1.90	-53.8	0.38239	-0.37148	-0.39585
2.10	-46.2	0.39260	-0.35481	-0.43091
2.30	-23.7	0.37428	-0.33299	-0.47995
2.50	-11.2	0.31414	-0.30074	-0.51648
2.70	-8.6	0.31962	-0.27899	-0.52675
2.90	-6.3	0.30315	-0.24979	-0.54209
3.10	-4.0	0.31222	-0.25770	-0.55773
3.30	-4.0	0.28763	-0.25580	-0.56028
3.50	-4.2	0.27480	-0.25430	-0.57216

transition state conformation (35, 129).

To investigate the conformation of the glucose in the transition state, quantum mechanical calculations were performed on DIMBOA-glucoside, with the glucosyl moiety in the $^4C_1$ conformation. The glycosidic bond was lengthened to mimic the reaction, and the dihedral (C2-C1-O5-C5) that characterizes the conformation of the glucose was measured, as well as the change in charges on the C1, O5 and O1 atoms (Table 4.6).

As the glycosidic bond lengthens, the glucose conformation changes from $^4C_1$ to $^4H_3$. In the $^4H_3$ conformation, the C2, C1, O5, C5 atoms lie in a plane, which must occur at the transition state. The $^4H_3$ glucose conformation and the lower negative charge on the O_5 is consistent with the oxocarbenium ion transition state (38, 78, 118).

Active Site Profiling: Conventionally, entire sequences of proteins are used for clustering the proteins to determine functionality. Several groups have recognized, though, that clustering using the entire sequences, where identities are often as low as 20%, may not be the most accurate (17, 22, 102). Clustering of active site residues may be more effective in providing informative clustering of sequences. This clustering analysis requires many sequences, so we have used the putative β-glucosidases

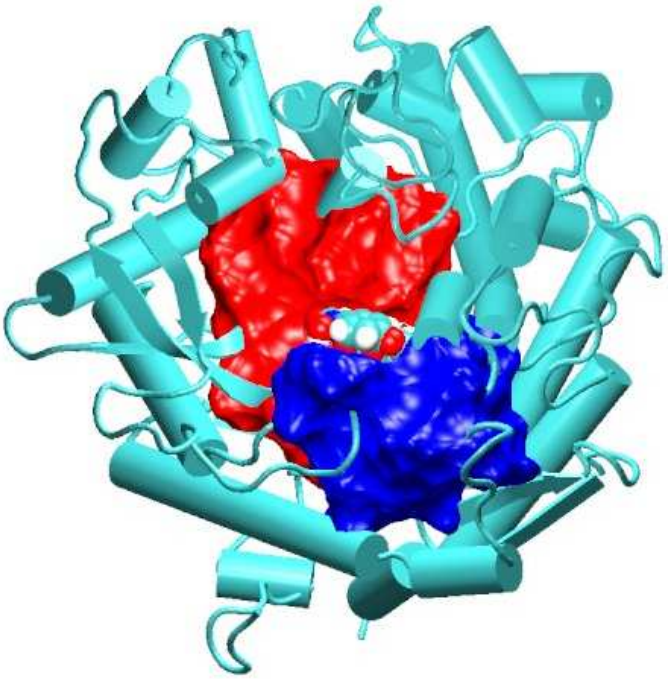


Figure 4.8: Residues included in the three different Active Site Profiles. Profile59 - Red & Blue, Profile39 - Red, and Profile20 - Blue. DIMBOA-glucoside shown as van der Waals in atom coloring.

from *A. thaliana* and the Glu1, Dhr1 and Myr proteins to possibly gain insight into the substrate specificities of the *A. thaliana* proteins.

The clustering of the alignments based on the full sequences (142), residues in the glycone binding site (Profile39), aglycone binding site (Profile20) and both the glycone and aglycone binding site (Profile59) were compared to determine if residues in the active site will yield information about substrate specificities. Surprisingly, only including active site residues did not change the overall clustering of the sequences when compared with the full sequences. Possibly this is because 40 of the 59 residues are in the glycone binding pocket, which one would expect to be fairly conserved in β-glucosidases.

Profile20, Profile39 and Profile59 identified four putative β-glucosidases that cluster differently than when the full sequences are used. At2g44480, in the full sequence subclassification 2, clusters near sequences in either full sequence subclassification 1

(Profile20) or 7 (Profile39 and Profile59) . While the sequences in subclassification 2 and At2g44480 have three insertions compared to the sequences in subclassification 1, the residues in At2g44480 aglycone binding pocket appear more similar to the aglycone binding pockets in subclassification 1 (Figure 4.10). The second deviation from the full sequence clustering occurs with subclassifications 5 and 7. The putative myrosinases compose subclassification 7, so it is interesting that At5g25980 and At5g26000 cluster separately from At1g47600 and At1g51470. Thirdly, At5g36890 was not clustered with the other two proteins from subclassification 8. Instead, At3g18080 (original subclassification 9) clustered with the two subclassification 8 proteins. At5g36890 has several charged residues (E,D,Q and K) aligning with uncharged residues. Finally, Profile39, which contained the residues in the glycone binding site, showed that At3g03640 clustered differently from the full sequence subclassification. At3g03640 is extremely different from all of the other sequences, as instead of a glutamic acid in the nucleophilic position, it contains a glycine. Examining the effect of this alternation on the functionality, if any, of this protein would be interesting.

4.5 Discussion

This study confirms the ability of MD simulations to model the dynamics of protein-ligand interactions. The crystal structures of Glu1 and Dhr1 with and without ligand have provided much information about some of these interactions. A major caveat about the experimentally determined crystal structures is that the structures of the complexes were solved with the acid/base catalytic glutamic acid substituted with an aspartic acid. This exchange eliminates the catalytic functioning of these enzymes, and therefore the interactions in these crystal structure complexes may be slightly different from interactions that occur with the native protein. Additionally, the experimental structures do not provide information about the dynamic behavior of these systems, only a static picture. We are now able to better understand the dynamics and interactions of the native structures because of the application of extensive MD of several different complexes of the Glu1 and Dhr1 systems. The MD simulations of Glu1_DIMglc, Glu1_dhr, Dhr1_dhr, Dhr1_DIMglc and Y473F_dhr provide valuable information about the protein-ligand interactions that affect substrate specificity in these closely related proteins.

The protonation states of the catalytic glutamic acids are integral to the mechanism of these proteins. The MD simulation results confirm that only when the

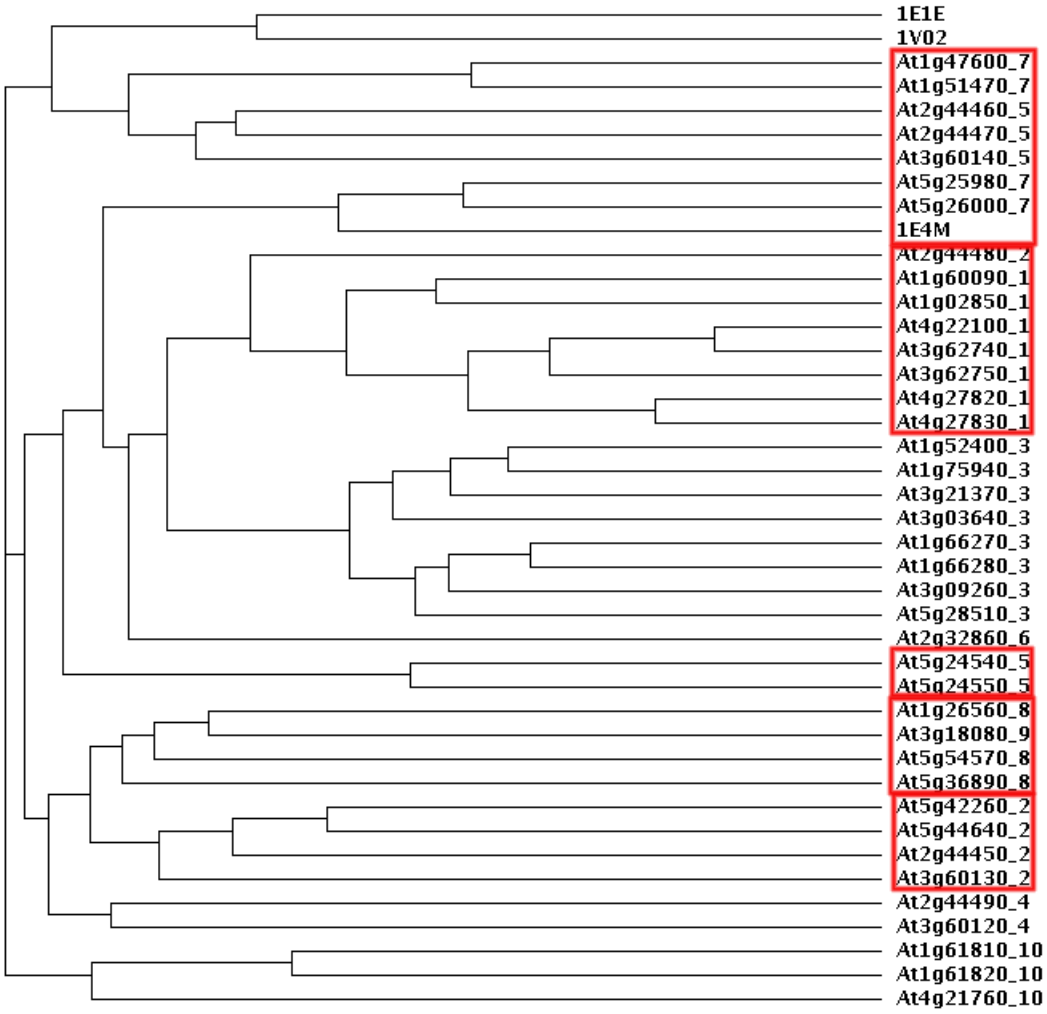


Figure 4.9: Clustering of active site residues that line the aglycone binding site (Profile20) of the β-glucosidases. The putative subfamilies based on the entire sequences in Xu *et. al* (142) are appended to each *A. thaliana* gene ID. Clusters that are different from the full sequence subfamilies are boxed in red. The alignments of the residues of Profile20 are shown in Figure 4.10.

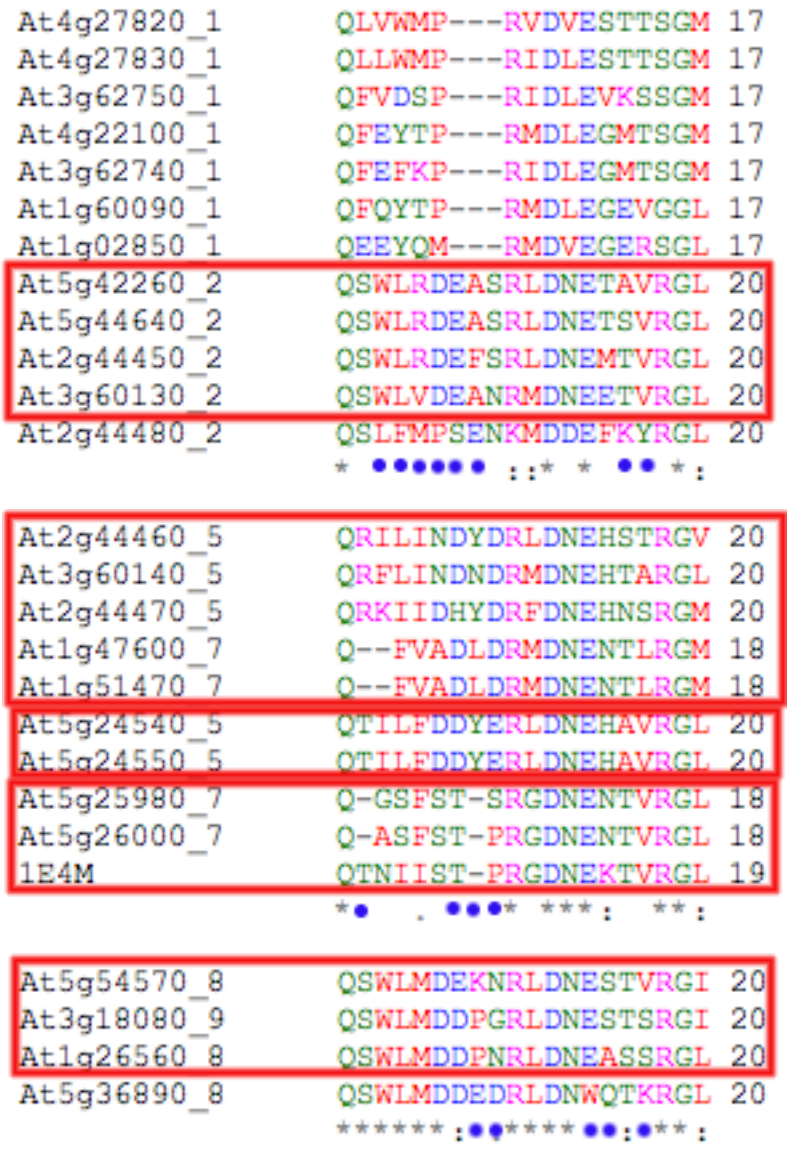


Figure 4.10: Profile20 of the three sequences that did not cluster in the same subfamilies based on the full sequences. The gene identifiers are appended with the full sequences subfamily number At2g44480 clustered with Profile20 subcluster 1, not with subfamily 1. Full sequences subfamilies 5 and 7 were intermixed in the Profile20 clustering. At3g18080, from full sequences subfamily 9, clustered with Profile20 cluster 8. The ClustalW physio-chemical coloring (red - small + hydrophobic (AVFPMILW); blue - acidic (DE); magenta - basic (RHK); and green - hydroxyl + amine + basic (STYHCNGQ)) and conservation symbols are shown in each clustering ('*' - identical; ':' - conserved substitutions; '.' - semi-conserved substitutions)

acid/base catalyst is protonated and the nucleophile is unprotonated are the catalytic residues oriented properly. When the acid/base catalyst is unprotonated, the catalytic residues repel each other, probably due to unfavorable electrostatic interactions, and are oriented too far away from each other for catalysis to occur. Several other computational studies have demonstrated the importance of accounting for the proper protonation states when observing dynamics (76, 95).

The distances between the catalytic residues in this study support the in-plane protonation mechanism for glycosyl hydrolases. Our results show that the carboxyl oxygens are ∼3 Å apart in the protein-ligand complexes and not the current proposed distance of 5.5 Å. This closer distance would require a different orientation between the catalytic glutamic acids and the substrate, which leads to the in-plane protonation hypothesis. Under the in-plane protonation hypothesis, members of family 1 glycosyl hydrolases have been shown to be *anti* protonators, which is consistent with the orientation of the acid/base residue with respect to the glycosyl moiety plane (59). In Glu1, Dhr1 and in all of the putative *A. thaliana* β-glucosidases, a tyrosine near the ring oxygen of the glucose moiety is highly conserved and believed to be involved in the *anti* in-plane protonation mechanism (90). This tyrosine is also thought to be a part of a hydrophobic platform, with a tryptophan and a phenylalanine, that stabilizes the transition state (89). The *A. thaliana* putative β-glucosidases all show conservation of these residues.

Only one of the interactions between the glucosyl moiety and the protein is observed in all of the simulations of catalytically active complexes. The O-2 oxygen of glucose and the nucleophilic glutamic acid form a hydrogen bond persistent throughout the simulations, and this hydrogen bond is observed in the crystal structure complexes of Glu1-191D with DIMBOA-glucoside, Dhr1-189D with dhurrin and MYR with glucotetrazole (129). This interaction is known to stabilize the transition state (53, 88).

The interaction between DIMBOA-glucoside and Glu464 in Glu1 and Glu460 in Dhr1 is slightly different than the interactions between dhurrin and Glu464 in Glu1 and Glu460 in Dhr1. This glutamic acid has been shown to have a role in discriminating between glucosides and galactosides (84). The O-6 of the glucosyl moiety of DIMBOA-glucoside interacts with Glu464-O_ϵ1 in Glu1 and Dhr1, but it is the O-4 of the glucosyl moiety of dhurrin that interacts with Glu464-O_ϵ1 atom of Glu1 and Dhr1. The O-6 of the glucosyl moiety of dhurrin interacts with Glu464-O_ϵ2 atom of Glu1 and Dhr1. The Glu1-E191D with DIMBOA-glucoside crystal structure shows the O-6 of the glucosyl moiety of the DIMBOA-glucoside forming a hydrogen bond with both Glu464-O_ϵ1 and Glu464-O_ϵ2 atoms. The Dhr1-E189D with dhurrin crys-

tal structure shows the same hydrogen bonding pattern of Glu464 carboxyl oxygen atoms as the MD simulation of Dhr1_dhr.

Also, the interactions between Glu1 and DIMBOA-glucoside substrate are different from the interactions between Dhr1 and dhurrin. Besides the O-2 atom of the glucosyl moiety interacting with nucleophilic glutamic acid, none of the Glu1 with DIMBOA-glucoside interactions are observed in the Dhr1 with dhurrin interactions. These differences in the glycone binding suggest possibly slightly different modes of binding between these two ligands. Additional support for different binding modes between Glu1 and Dhr1 comes from the differences in the distance between the catalytic glutamic acids, which, in the catalytically active complexes, is much closer in Glu1 than in Dhr1. Also, while substituting a Dhr1 residue for a Glu1 residue (e.g. Y473/F471) allows Glu1 to hydrolyze dhurrin, the opposite substitution in Dhr1 does not confer DIMBOA-glucoside hydrolysis ability to Dhr1 (26, 128).

The conformation of the glucosyl moiety may also contribute to the binding of the substrate. Based on the lack of electron density of several atoms of the glucosyl moiety, it was proposed that Glu1 can bind the glucose moiety in both the chair and the skew boat conformation (35); electron density of Dhr1 shows binding of the glucosyl moiety in the skew boat conformation (129). While dhurrin is initially in the skew boat position, the computationally derived parameters used during the MD simulations are for glucose in the chair position, as that was the conformation of the glucosyl moiety in the dhurrin structure available at the time of the study. During the MD simulations, the glucosyl moiety of dhurrin switched from the skew boat to the chair conformation; the distances between the anomeric carbon and the nucleophilic glutamate may change if the glucosyl moiety was in the skew boat conformation. A study of the affect of the glucose moiety was done by rerunning the MD simulations of Dhr1_DIMglc and Dhr1_dhr with the dihedral restraint on the C2-C2-O5-C5 atoms to keep the sugar in the skew boat conformation. While the distance between the anomeric carbon and the nucleophilic glutamic acid were closer in the the Dhr1_DIMglc simulation, this distance is slightly larger in the Dhr1_dhr simulation. More relevant results may be obtained by rerunning the simulations with the ligand reparameterized for the skew boat conformation.

A recent Car-Parrinello MD (CPMD) and QM/MM CPMD study of $^4C_1$ versus the $^1S_3$ glucose conformation presented an interesting proposal about the effect of atomic charge of the anomeric carbon on the glucose conformation (11). They noted that the glucose remained in the $^1S_3$ only if the charge on the anomeric carbon was increased by 0.6 electrons, though no experimental or theoretical reason was provided for choosing this number. Additionally, they saw the charge on the anomeric carbon

increase and the charge on the ring oxygen decrease from the $^4C_1$ to $^1S_3$ conformations. This change in the charges of the anomeric carbon and ring oxygen is difficult to reconcile with our charge calculations during the $^4C_1$ to $^4H_3$ conformational change. We see the charge on the anomeric carbon decrease and the charge on the ring oxygen increase. An increase on the charge of the ring oxygen, though, is consistent with the oxocarbenium ion intermediate of the proposed mechanism.

The interaction with solvent may play an important role in differentiating substrates from inhibitors. When DIMBOA-glucoside was in Dhr1 and dhurrin was in Glu1, the aglycone portion of each ligand hydrogen bonded with two water molecules. But when DIMBOA-glucoside was in Glu1 and dhurrin was in Dhr1, the aglycone portion of each ligand hydrogen bonded with only one water molecule. The number of water molecules within 3.5 Å of the acid/base catalytic glutamate in Dhr1_DIMglc is three times greater than the number of water molecules within 3.5 Å of the acid/base catalytic in all of the other simulations. The increased solvent ordering caused by the inhibitors may contribute to unfavorable binding.

Moreover, water appears to play an important role in stabilizing the active site architecture (73). A water molecule bridges interactions between Asn376, Trp378, Glu471 and Tyr473 in Glu1_DIMglc. The absence of this water molecule in Glu1_dhr causes shifts in the above mentioned residues, and distortion of the active site. The ability of the Tyr473 phenolic hydroxyl to form hydrogen bonds with water also appears to affect the efficiency of Glu1. Phe473 cannot form a hydrogen bond with water, as is seen in the Glu1_DIMglc simulations. Several studies have shown that, while the loss of a hydrogen bond contributes to unfavorable binding enthalpy, the increased entropy may not compensate for the unfavorable binding enthalpy associated with the loss of that hydrogen bond (31, 110). A Glu1/Dhr1 chimera in which the Tyr473 is changed to a Phe473 shows only 65% of the relative efficiency of the native Glu1 (26).

One major finding of this work is that, even assuming no prior knowledge about the substrate specificities of Glu1 and Dhr1, by evaluating all of the interactions between the substrate and active site residues, we could computationally differentiate between catalytically active and inactive complexes. Fifteen MD simulations of protein-ligand complexes were conducted. Based on the distance between the catalytic glutamates and the distance between the nucleophilic glutamic acid and the anomeric carbon, we would propose that eight of the fifteen simulations did not model a catalytically relevant system, as these distances were too large. The skewed position of the glucosyl moiety of DIMglc and the high number of water molecules that were able to be within 3.5 Å of the acid/base catalytic glutamic acid would indicate that the Dhr1_DIMglc may not be catalytically active. The distortion of the opening to the active site in the

Glu1_dhr simulations would also indicate that this complex may not be catalytically active. In the simulations of Glu1_DIMglc, Dhr1_dhr, Y473F_DIMglc, Y473F_dhr and Chim21_dhr, we do not observe any unusual interactions or structural distortions. The five proteins in these simulations have demonstrated activity towards the ligand with which they are complexed.

While it was known that dhurrin inhibits Glu1 and DIMBOA-glucoside inhibits Dhr1, how this inhibition occurred was not known. DIMBOA-glucoside is a much more potent inhibitor of Dhr1 than dhurrin is of Glu1 (K_i 9μM vs 76μM). While the dhurrin inhibition of Glu1 is associated with conformational change of the active site, no obvious protein conformational changes occur during the act of DIMBOA-glucoside inhibition of Dhr1. The lack of major conformational changes in Dhr1 with DIMBOA-glucoside may explain why DIMBOA-glucoside is more potent then dhurrin in their respective proteins, as the residues in the active site of Dhr1 are not greatly perturbed.

The MD simulations of Glu1 and Dhr1 allow for atomic level insight into the protein fluctuations of protein-ligand complexes. These simulations help to bridge our understanding as to what interactions occur in these different complexes. Hopefully the application of this atomic level information of known systems can be applied to additional proteins, especially the putative β-glucosidases in *A. thaliana*.

Bibliography

[1] C F Aguilar, I Sanderson, M Moracci, M Ciaramella, R Nucci, M Rossi, and L H Pearl. Crystal structure of the beta-glycosidase from the hyperthermophilic archeon sulfolobus solfataricus: resilience as a key factor in thermostability. *J Mol Biol*, 271(5):789–802, Sep 1997.

[2] T Akiba, M Nishio, I Matsui, and K Harata. X-ray structure of a membrane-bound beta-glycosidase from the hyperthermophilic archaeon pyrococcus horikoshii. *Proteins*, 57(2):422–431, Nov 2004.

[3] B.J. Alder and T.E. Wainwright. Phase transition for a hard sphere system. *J. Chem. Phys.*, 27(3):1208–1209, 1957.

[4] J.D. Anderson, A. Thåström, and J. Widom. Spontaneous access of proteins to buried nucleosomal dna target sites occurs via a mechanism that is distinct from nucleosome translocation. *Mol. Cell. Biol.*, 22(20):7417–7157, 2002.

[5] G.D. Babcock and A. Esen. Substrate specificity of maize β-glucosidase. *Plant Science*, 101(1):31–39, 1994.

[6] T. Barrett, C.G. Suresh, S.P. Tolley, E.J. Dodson, and M.A. Hughes. The crystal structure of a cyanogenic beta-glucosidase from white clover, a family 1 glycosyl hydrolase. *Structure*, 3:951–60, Sep 1995.

[7] D Bashford, D A Case, C Dalvit, L Tennant, and P E Wright. Electrostatic calculations of side-chain pk(a) values in myoglobin and comparison with nmr data for histidines. *Biochemistry*, 32(31):8045–8056, Aug 1993.

[8] D. Beard and T. Schlick. Modeling salt-mediated electrostatics of macromolecules: The discrete surface charge optimization algorithm and its application to the nucleosome. *Biopolymers*, 58:106–115, 2001.

[9] A. Bellelli, R.S. Blackmore, and Q.H. Gibson. Ligand binding to a hemoprotein lacking the distal histidine. *J. Biol. Chem.*, 265:13595–13600, 1990.

[10] J G Berrin, M Czjzek, P A Kroon, W R McLauchlan, A Puigserver, G Williamson, and N Juge. Substrate (aglycone) specificity of human cytosolic beta-glucosidase. *Biochem J*, 373(Pt 1):41–48, Jul 2003.

[11] X Biarnés, J Nieto, A Planas, and C Rovira. Substrate distortion in the michaelis complex of bacillus 1,3-1,4-beta-glucanase. insight from first principles molecular dynamics simulations. *J Biol Chem*, 281(3):1432–1441, Jan 2006.

[12] C. Bossa, A. Amadie, I. Daidone, M. Anselmi, B. Vallone, M. Brunori, and A. Di Nola. Molecular dynamics simulation of sperm whale myoglobin: effects of mutations and trapped co on the structure and dynamics of cavities. *Biophys. J.*, 89:465–474, 2005.

[13] C. Bossa, M. Anselmi, D. Roccatano, A. Amadei, B. Vallone, M. Brunori, and A. Di Nola. Extended molecular dynamics simulation of the carbon monoxide migration in sperm whale myoglobin. *Biophys. J.*, 86:3855–3862, 2004.

[14] H. B. Búrgi, J. D. Dunitz, and E. Shefter. Chemical reaction paths. IV. Aspects of $O\cdots C=O$ interactions in crystals. *Acta Crystallographica Section B*, 30(6):1517–1527, Jun 1974.

[15] W P Burmeister, S Cottaz, P Rollin, A Vasella, and B Henrissat. High resolution x-ray crystallography shows that ascorbate is a cofactor for myrosinase and substitutes for the function of the catalytic base. *J Biol Chem*, 275(50):39385–39393, Dec 2000.

[16] W.P. Burmeister, S. Cottaz, H. Druquez, R. Iori, Palmieri S., and B. Henrissat. The crystal structures of sinapis alba myrosinase and a covalnd glycosyl-enzyme intermediate provide insights into th substrate recognition and active-site machinery of an s-glycosidase. *Structure*, 5:663–75, 1997.

[17] S.A. Cammer, B.T. Hoffman, J.A. Speir, M.A. Canady, M.R. Nelson, S. Knutson, M. Gallian, S.M. Baxter, and J.S. Fetrow. Structure-based active site profiles for genome analysis and functional family subclassification. *J.Mol.Biol.*, 334:387–401, 2003.

[18] D A Case and M Karplus. Dynamics of ligand binding to heme proteins. *J Mol Biol*, 132(3):343–368, Aug 1979.

[19] D.A. Case, T. Darden, T. Cheatham, H. Gohlke, K. Merz, A. Onufriev, R. Luo, and R. Woods. The amber biomolecular simulation programs. *J. Comp. Chem.*, 26:1668–1688, 2005.

[20] D.A Case, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, B. Wang, D.A. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J.W. Caldwell, W.S. Ross, and P.A. Kollman. Amber 8. *University of California, San Francisco*, 2004.

[21] D.A. Case, D.A. Pearlman, J.W. Caldwell, T.E. III Cheatham, J. Wang, W.S. Ross, C.L. Simmerling, T.A. Darden, K.M. Merz, R.V. Stanton, A.L. Cheng, J.J. Vincent, M. Crowley, V. Tsui, H. Gohlke, R.J. Radmer, Y. Duan, J. Pitera, I. Massova, G.L. Seible, U.C. Singh, P.K. Weiner, and P.A. Kollman. *AMBER7*, 2002.

[22] R Chakrabarti, A M Klibanov, and R A Friesner. Computational prediction of native protein ligand-binding and enzyme active site sequences. *Proc Natl Acad Sci U S A*, 102(29):10153–10158, Jul 2005.

[23] TE Cheatham. Simulation and modeling of nucleic acid structure, dynamics and interactions. *Curr. Opin. Struct. Biol.*, 14(3):360–367, Jun 2004.

[24] TE Cheatham and MA Young. Molecular dynamics simulation of nucleic acids: successes, limitations, and promise. *Biopolymers*, 56(4):232–256, Dec 2001.

[25] X D Cheng and B P Schoenborn. Neutron diffraction study of carbonmonoxymyoglobin. *J Mol Biol*, 220(2):381–399, Jul 1991.

[26] M. Cicek, D. Blanchard, D. R. Bevan, and A. Esen. The aglycone specificity-determining sites are different in 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA)-glucosidase (Maize beta-Glucosidase) and Dhurrinase (Sorghum beta-Glucosidase). *J. Biol. Chem.*, 275(26):20002–20011, 2000.

[27] M. Cicek and A. Esen. Structure and expression of a dhurrinase (β-glucosidase) from sorghum. *Plant Physiol.*, 116:1469–1478, 1998.

[28] T.E. Cloutier and J. Widom. Spontaneous sharp beinding of double-stranded dna. *Molecular Cell*, 14:355–362, 2004.

[29] TE. Cloutier and J. Widom. DNA twisting flexibility and the formation of sharply looped protein-DNA complexes. *PNAS*, 102(10):3645–3650, Mar 2005.

[30] Jordi Cohen, Anton Arkhipov, Rosemary Braun, and Klaus Schulten. Imaging the migration pathways for o2, co, no, and xe inside myoglobin. *Biophys. J.*, 91(5):1844–1857, 2006.

[31] PR Connelly, RA Aldape, FJ Bruzzese, SP Chambers, MJ Fitzgibbon, MA Fleming, S Itoh, DJ Livingston, mA Navia, JA Thomson, and KP Wilson. Enthalpy of Hydrogen Bond Formation in a Protein-Ligand Binding Reaction. *PNAS*, 91(5):1964–1968, 1994.

[32] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[33] DM Crothers, J Drak, JD Kahn, and SD Levene. DNA bending, flexibility, and helical repeat by cyclization kinetics. *Methods Enzymol.*, 212:3–29, 1992.

[34] L. Cuevas, H.M. Niemeyer, and L.M.V. Jonsson. Partial purification and characterization of a hydroxamic acid glucoside β-glucosidase from maize. *Phytochem.*, 31:2609–2612, 1992.

[35] M. Czjzek, M. Cicek, V. Zamboni, D. R. Bevan, B. Henrissat, and A. Esen. The mechanism of substrate (aglycone) specificity in beta -glucosidases is revealed by crystal structures of mutant maize beta-glucosidase-DIMBOA, -DIMBOAGlc, and -dhurrin complexes. *PNAS*, 97(25):13555–13560, 2000.

[36] M. Czjzek, M. Cicek, V. Zamboni, W.P. Burmeister, D.R. Bevan, B. Henrissat, and A. Esen. Crystal structure of a monocotyledon (maize zmglu1) beta-glucosidase and a model of its complex with p-nitrophenyl beta-d-thioglucoside. *Biochem. J.*, 354:37–46, 2001.

[37] C.A. Davey, D.F. Sargent, K. Luger, A.W. Maeder, and T.J. Richmond. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9å resolution. *J.Mol.Biol.*, 319:1097–1113, 2002.

[38] G J Davies, V M Ducros, A Varrot, and D L Zechel. Mapping the conformational itinerary of beta-glycosidases by x-ray crystallography. *Biochem Soc Trans*, 31(Pt 3):523–527, Jun 2003.

[39] G.J. Davies and B. Henrissat. Structures and mechanisms of glycosyl hydrolases. *Structure*, 3:853–859, September 1995.

[40] G.J. Davies, L. Mackenzie, A. Varrot, M. Dauter, A. M. Brzozowski, M. Schülein, and S.G. Withers. Snapshots along an enzymatic reaction coordinate: analysis of a retaining β-glycoside hydrolase. *Biochemistry*, 37(34):11707–11713, 1998.

[41] IW. Davis, LW. Murray, JS. Richardson, and DC. Richardson. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res*, 32(Web Server issue):615–619, Jul 2004.

[42] D P Dharmawardhana, B E Ellis, and J E Carlson. A beta-glucosidase from lodgepole pine xylem specific for the lignin precursor coniferin. *Plant Physiol*, 107(2):331–339, Feb 1995.

[43] M. Doi and S.F. Edwards. *Theory of Polymer Dynamics*. Oxford Press, New York, 1985.

[44] Q. Du, C. Smith, N. Shiffeldrim, M. Vologodskaia, and A. Vologodskii. Cyclization of short DNA fragments and bending fluctuations of the double helix. *Proc Natl Acad Sci U S A*, 102(15):5397–5402, Apr 2005.

[45] K.D. Egeberg, B.A. Springer, S.G. Sligar, T.E. Carver, R.J. Rohlfs, and J.S. Olson. The role of val^{68}(e11) in ligand binding to sperm whale myoglobin. *J. Biol. Chem.*, 265:11788–11795, 1990.

[46] M.A. El Hassan and C.R. Calladine. Two distinct modes of protein-induced bending in dna. *J. Mol. Biol.*, 282:331–343, 1998.

[47] R. Elber and M. Karplus. Enhanced sampling in molecular dynamics: use of the time-dependent hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *J. Am. Chem. Soc.*, 112:9161–9175, 1990.

[48] D.J. Fitzgerald and J.N. Anderson. Dna distortion as a factor in nucleosome positioning. *J. Mol. Biol.*, 293:477–491, 1999.

[49] P.J. Flory. *Statistical Mechanics of Chain Molecules*. Interscience Publishers, New York, 1969.

[50] H. Frauenfelder, B. H. McMahon, and P. W. Fenimore. Myoglobin: The hydrogen atom of biology and a paradigm of complexity. *PNAS*, 100(15):8615–8617, 2003.

[51] Peter L. Freddolino, Anton S. Arkhipov, Steven B. Larson, Alexander McPherson, and Klaus Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14(3):437–449, 2006.

[52] M.J. Frisch. et. al, Gaussian98, 1998.

[53] T M Gloster, S Roberts, V M Ducros, G Perugino, M Rossi, R Hoos, M Moracci, A Vasella, and G J Davies. Structural studies of the beta-glycosidase from sulfolobus solfataricus in complex with covalently and non-covalently bound inhibitors. *Biochemistry*, 43(20):6101–6109, May 2004.

[54] J.C. Gordon, J.B. Myers, T. Folta, V. Shoja, L.S. Heath, and A. Onufriev. H++: a server for estimating pk_as and adding missing hydrogens to macromolecules. *Nucleic Acids Research*, 33:368–371, 2005.

[55] J.M. Gottesfeld, J.M. Belitsky, C. Melander, P.B. Dervan, and K. Luger. Blocking transcription through a nucleosome with synthetic dna ligands. *J. Mol. Biol.*, 321:263, 2002.

[56] N Hakulinen, S. Paavilainen, T. Korpela, and J. Rouvinen. The crystal structure of β-glucosidase from *Bacillus circulans* sp. *alkalophilus*: ability to form long polymeric assemblies. *J Struct Biol*, 129:69–79, 2000.

[57] JJ. Hayes, J. Bashkin, TD. Tullius, and AP. Wolffe. The histone core exerts a dominant constraint on the structure of DNA in a nucleosome. *Biochemistry*, 30(34):8434–8440, Aug 1991.

[58] JJ. Hayes, TD. Tullius, and AP. Wolffe. The structure of DNA in a nucleosome. *Proc Natl Acad Sci U S A*, 87(19):7405–7409, Oct 1990.

[59] T.D. Heightman and A.T. Vasella. Recent insights into inhibition, structure, and mechanism of configuration retaining glycosidases. *Angew. Chem. Int. Ed.*, 38:750–770, 1999.

[60] B. Hernrissat. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.*, 280:309–316, 1991.

[61] M.E. Hogan, T.F. Rooney, and R.H. Austin. Evidence for kinks in dna folding in the nucleosome. *Nature*, 328:554–557, 1987.

[62] W Hösel, I Tober, S H Eklund, and E E Conn. Characterization of beta-glucosidases with high specificity for the cyanogenic glucoside dhurrin in sorghum bicolor (l.) moench seedlings. *Arch Biochem Biophys*, 252(1):152–162, Jan 1987.

[63] X. Huang and S.G. Boxer. Discover of new ligand binding pathways in myoglobin by random mutagenesis. *Nat. Struct. Biol.*, 1(4):226–229, 1994.

[64] G. Hummer, F. Schotte, and P.A. Anfinrud. Unveiling functional protein motions with picosecond x-ray crystallography and molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*, 101:15330–15334, 2004.

[65] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[66] H. Husebye, S. Arzt, W.P. Burmeister, F.V. Härtel, A. Brandt, J.T. Rossiter, and A.M. Bones. Crystal structure at 1.1 åresolution of an insect myrosinase from *Brevicoryne brassicae* shows its cloe relationship to β-glucosidases. *Insect Biochem Mol Biol*, 35:1311–1320, 2005.

[67] Jaguar. version 6.0 Schrodinger, LLC, New York, NY, 2005.

[68] K A Johnson, J S Olson, and G N Phillips. Structure of myoglobin-ethyl isocyanide. histidine as a swinging door for ligand entry. *J Mol Biol*, 207(2):459–463, May 1989.

[69] P.A. Kollman, I.R. Massova, C.A. Reyes, B. Kuhn, S.A. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, and T.E. III. Cheatham. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.*, 33:889–897, 2000.

[70] R.D. Kornberg and Y. Lorch. Interplay between chromatin structure and transcription. *Curr. Opin. Cell Biol.*, 7:371–375, 1995.

[71] R.D. Kornberg and Y. Lorch. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, 98:285–294, 1999.

[72] B Kuhlman, G Dantas, G C Ireton, G Varani, B L Stoddard, and D Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368, Nov 2003.

[73] J E Ladbury. Just add water! the effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem Biol*, 3(12):973–980, Dec 1996.

[74] T Laitinen, J Rouvinen, and M Peräkylä. Mm-pbsa free energy analysis of endo-1,4-xylanase ii (xynii)-substrate complexes: binding of the reactive sugar in a skew boat and chair conformation. *Org Biomol Chem*, 1(20):3535–3540, Oct 2003.

[75] L. D. Landau and E. M. Lifshitz. *Theory of Elasticity*. Pergamon, New York, 1986.

[76] E Langella, R Improta, O Crescenzi, and V Barone. Assessing the acid-base and conformational properties of histidine residues in human prion protein (125-228) by means of pk(a) calculations and molecular dynamics simulations. *Proteins*, 64(1):167–177, Jul 2006.

[77] R. Lavery and H. Sklenar. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.*, 6(1):63–91, 1988.

[78] J K Lee, A D Bain, and P J Berti. Probing the transition states of four glucoside hydrolyses with 13c kinetic isotope effects measured at natural abundance by nmr spectroscopy. *J Am Chem Soc*, 126(12):3769–3776, Mar 2004.

[79] H Li, R Elber, and JE Straub. Molecular dynamics simulation of NO recombination to myoglobin mutants. *J. Biol. Chem.*, 268(24):17908–17916, 1993.

[80] F.C. Lightstone and T.C. Bruice. Ground state conformations and entropic and enthalpic factors in the efficiency of intramolecular and enzymatic reactions. 1. cyclic anhydride formation by substituted glutarates, succinate, and 3,6-endoxo-4-tetrahydrophthalate monophenyl esters. *Journal of the American Chemical Society*, 118(11):2595–2605, 1996.

[81] Xiang-Jun Lu and Wilma K. Olson. 3dna: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nuc. Acids Res.*, 31:5108–5121, 2003.

[82] Buyong Ma and Ruth Nussinov. From computational quantum chemistry to computational biology: experiments and computations are (full) partners. *Physical Biology*, 1(4):P23–P26, 2004.

[83] T. Macke and D.A Case. Modeling unusual nucleic acid structures. Molecular Modeling of Nucleic Acids, pages 379–393. American Chemical Society, 1998.

[84] S.R. Marana, W.R. Terra, and C. Ferreira. The role of amino-acid residues Q39 and E451 in the determination of substrate specificity of the *Spodoptera frugiperda* β-glycosidase. *Eur. J. Biochem.*, 269:3705–3714, 2002.

[85] J. Andrew McCammon, Bruce R. Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977.

[86] J Meller and R Elber. Computer simulations of carbon monoxide photodissociation in myoglobin: structural interpretation of the b states. *Biophys J*, 74(2 Pt 1):789–802, Feb 1998.

[87] L. Mouawad, J-D. Marechal, and D. Perahia. Internal cavities and ligand passageways in human homoglobin characterized by molecular dynamics simulations. *Biochim. Biophys. Acta*, 1724:385–393, 2005.

[88] M N Namchuk and S G Withers. Mechanism of agrobacterium beta-glucosidase: kinetic analysis of the role of noncovalent enzyme/substrate interactions. *Biochemistry*, 34(49):16194–16202, Dec 1995.

[89] W Nerinckx, T Desmet, and M Claeyssens. A hydrophobic platform as a mechanistically relevant transition state stabilising factor appears to be present in the active centre of all glycoside hydrolases. *FEBS Letters*, 538(1):1–7, 2003.

[90] W Nerinckx, T Desmet, K Piens, and M Claeyssens. An elaboration on the syn–anti proton donor concept of glycoside hydrolases: electrostatic stabilisation of the transition state as a general strategy. *FEBS Letters*, 579(2):302–312, 2005.

[91] Y. Nishihara, M. Sakakura, Y. Kimura, and M. Terazima. The escape process of carbon monoxide to solution at physiological temperature. *J. Am. Chem. Soc.*, 126:11877–11888, 2004.

[92] J.S. Olson, A.J. Mathews, R.J. Rohlfs, B.A. Springer, K.D. Egeberg, S.G. Sligar, J. Tame, J.P. Renaud, and K. Nagai. The role of the distal histidine in myoglobin and haemoglobin. *Nature*, 336:265–266, 1988.

[93] A. Onufriev, D. Bashford, and D.A Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*, 55:383–394, 2004.

[94] A Onufriev, DA Case, and D Bashford. Structural details, pathways, and energetics of unfolding apomyoglobin. *J. Mol. Biol.*, 325:555–67, 2003.

[95] H Park and S Lee. Determination of the active site protonation state of beta-secretase from molecular dynamics simulation and docking experiment: implications for structure-based inhibitor design. *J Am Chem Soc*, 125(52):16416–16422, Dec 2003.

[96] M.F. Perutz and F.S. Mathews. An x-ray study of azide methaemoblogin. *J. Mol. Biol.*, 21:199–202, 1966.

[97] K.J. Polach and J. Widom. Mechanism of protein access to specific dna sequences in chromati: a dynamic equilibrium model for gene regulation. *J. Mol. Biol.*, pages 130–149, 1995.

[98] D. Pruss, FD. Bushman, and AP. Wolffe. Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc Natl Acad Sci U S A*, 91(13):5913–5917, Jun 1994.

[99] M.L. Quillin, T. Li, J.S. Olson, G.N. Phillips Jr, Yi Dou, M. Ikeda-Saito, R. Regan, M. Carlson, Q.H. Gibson, H. Li, and R. Elber. Structural and functional effecs of apolar mutations of the distalvaline in myoglobin. *J. Mol. Biol.*, 245:416–436, 1995.

[100] T.J. Richmond and C.A. Davey. The structure of dna in the nucleosome core. *Nature*, 423:145–150, 2003.

[101] Dagmar Ringe, Gregory A. Petsko, David E. Kerr, and Paul R. Ortiz de Montellano. Reaction of myoglobin with phenylhydrazine: a molecular doorstop. *Biochemistry*, 23(1):2–4, 1984.

[102] B Rost. Enzyme function less conserved than anticipated. *J Mol Biol*, 318(2):595–608, Apr 2002.

[103] Vladimir Rotrekl, Eliska Nejedla, Igor Kucera, Fuad Abdallah, Klaus Palme, and Bretislav Brzobohaty. The role of cysteine residues in structure and enzyme activity of a maize beta-glucosidase. *Eur J Biochem*, 266(3):1056–1065, 1999.

[104] J. Sanz-Aparicio, J.A. Hermoso, J.L. Martínez-Ripoll, Lequerica, and J. Polaina. Crystal structure of β-glucosidase a from *Bacillus polymyxa*: insights into the catalytic activity in family 1 glucosyl hydrolases. *J Mol Biol*, 275:491–502, 1998.

[105] Steve Scheiner, William N. Lipscomb, and Daniel A. Kleier. Molecular orbital studies of enzyme activity. 2. nucleophilic attack on carbonyl systems with comments on orbital steering. *Journal of the American Chemical Society*, 98(16):4770–4777, 1976.

[106] T. Schlick. *Molecular Modeling and Simulation*. Springer, 2002.

[107] B P Schoenborn. Structure of alkaline metmyoglobin-xenon complex. *J Mol Biol*, 45(2):297–303, Oct 1969.

[108] F Schotte, M Lim, T A Jackson, A V Smirnov, J Soman, J S Olson, G N Phillips, M Wulff, and P A Anfinrud. Watching a protein as it functions with 150-ps time-resolved x-ray crystallography. *Science*, 300(5627):1944–1947, Jun 2003.

[109] E.E. Scott, Q.H. Gibson, and J.S. Olson. Mapping the pathways for o_2 entry into and exit for myoglobin. *J. Biol. Chem.*, 276(7):5177–5188, 2001.

[110] Scott D. Sharrow, Katherine A. Edmonds, Michael A. Goodman, Milos V. Novotny, and Martin J Stone. Thermodynamic consequences of disrupting a water-mediated hydrogen bond network in a protein:pheromone complex. *Protein Sci*, 14(1):249–256, 2005.

[111] J. Shimado and H. Yamakawa. Ring-closure probabilities for twisted wormlike chains. application to dna. *Macromolecules*, 17:689–698, 1984.

[112] D. Shore, J. Langowski, and RL. Baldwin. DNA flexibility studied by covalent closure of short fragments into circles. *Proc Natl Acad Sci U S A*, 78(8):4833–4837, Aug 1981.

[113] H Shroff, B M Reinhard, M Siu, H Agarwal, A Spakowitz, and J Liphardt. Biocompatible force sensor with optical readout and dimensions of 6 nm3. *Nano Lett*, 5(7):1509–1514, Jul 2005.

[114] C Simmerling, B Strockbine, and A E Roitberg. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.*, 124(38):11258–11259, Sep 2002.

[115] B.A. Springer, K.D. Egeberg, S.G. Sligar, R.J. Rohlfs, A.J. Mathews, and J.S. Olson. Discrimination between oxygen and carbon monoxide and inhibition of autooxidation by myoglobin. *J. Biol. Chem.*, 264:3057–3060, 1989.

[116] J Srinivasan, T.E Cheatham, P Cieplak, PA Kollman, and DA Case. Continuum solvent studies of the stabiity of dna, rna and phosphoramidate-dna helices. *J. Am. Chem. Soc.*, 120(37):9401–9409, 1998.

[117] J Srinivasan, MW Trevathan, P Beroza, and DA Case. Application of a pairwise generalized Born model to proteins and nucleic acids: Inclusion of salt effects. *Theor. Chem. Acc.*, 101:426–434, 1999.

[118] J M Stubbs and D Marx. Aspects of glycosidic bond formation in aqueous solution: chemical bonding and the role of water. *Chemistry*, 11(9):2651–2659, Apr 2005.

[119] Masayuki Sue, Kana Yamazaki, Shunsuke Yajima, Taiji Nomura, Tetsuya Matsukawa, Hajime Iwamura, and Toru Miyamoto. Molecular and Structural Characterization of Hexameric beta-D-Glucosidases in Wheat and Rye. *Plant Physiol.*, 141(4):1237–1247, 2006.

[120] J. Sun, Q. Zhang, and T. Schlick. Electrostatic mechanism of nucleosomal array folding revealed by computer simulation. *PNAS*, 102:8180–8185, 2005.

[121] Bishop T. Molecular dynamics simulations of a nucleosome and free dna. *J Biomol Struct Dyn*, 22(6):673–685, 2005.

[122] A. Thåström, J.M. Gottesfeld, K. Luger, and J. Widom. Histone-dna binding free energy cannot be measured in dilution-driven dissociation experiments. *Biochemistry*, 43:736–741, 2004.

[123] R.F. Tilton, I.D. Kuntz, and G.A. Petsko. Cavities in proteins: Structure of a metmyoglobin-xenon complex solved to 1.9å. *Biochemistry*, 23:2849–2857, 1984.

[124] Rhonda A. Torres and Thomas C. Bruice. Molecular dynamics study displays near in-line attack conformations in the hammerhead ribozyme self-cleavage reaction. *PNAS*, 95(19):11077–11082, 1998.

[125] V. Tsui and D.A. Case. Molecular dynamics simulations of nucleic acids with a generalized born solvation model. *J. Am. Chem. Soc.*, 122(11):2489–2498, 2000.

[126] V. Tsui and D.A. Case. Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers*, 56:275–291, 2001.

[127] A. Varrot, M. Schulein, M. Pipelier, A. Vasella, and G.J. Davies. Lateral protonation of a glycosidase inhibitor. structure of the bacillus agaradhaerens cel5a in complex with a cellobiose-derived imidazole at 0.97 å resolution. *Journal of the American Chemical Society*, 121(11):2621–2622, 1999.

[128] L. Verdoucq, M. Czjzek, J. Moriniere, D. R. Bevan, and A. Esen. Mutational and structural analysis of aglycone specificity in maize and sorghum beta-glucosidases. *J. Biol. Chem.*, 278(27):25055–25062, 2003.

[129] L. Verdoucq, J. Moriniere, D.R. Bevan, A. Esen, A. Vasella, B. Henrissat, and M. Czjek. Structural determinants of substrate specificity in family 1 beta-glucosidases: novel insights from the crystal structure of sorghum dhurrinase-1, a plant beta-glucosidase with strict specificity, in complex with its natural substrate. *J. Biol. Chem*, 279(30):31796–31803, 2004.

[130] F Vincent, T M Gloster, J Macdonald, C Morland, R V Stick, F M Dias, J A Prates, C M Fontes, H J Gilbert, and G J Davies. Common inhibition of both beta-glucosidases and beta-mannosidases by isofagomine lactam reflects different conformational itineraries for pyranoside hydrolysis. *Chembiochem*, 5(11):1596–1599, Nov 2004.

[131] D.J. Vocadlo, G.J. Davies, R. Laine, and S.G. Withers. Catalysis by hen egg-white lysozyme proceeds via a covalent intermediate. *Nature*, 412:835–838, 2001.

[132] Jaroslav Vojtchovský, Kelvin Chu, Joel Berendzen, Robert M. Sweet, and Ilme Schlichting. Crystal Structures of Myoglobin-Ligand Complexes at Near-Atomic Resolution. *Biophys. J.*, 77(4):2153–2174, 1999.

[133] M. Vologodskaia and A. Vologodskii. Contribution of the intrinsic curvature to measured DNA persistence length. *J Mol Biol*, 317(2):205–213, Mar 2002.

[134] Xinquan Wang, Xiangyuan He, Shoujun Yang, Xiaomin An, Wenrui Chang, and Dongcai Liang. Structural Basis for Thermostability of beta-Glycosidase from the Thermophilic Eubacterium Thermus nonproteolyticus HG102. *J. Bacteriol.*, 185(14):4248–4255, 2003.

[135] Yuhong Wang, J. Spencer Baskin, Tianbing Xia, and Ahmed H. Zewail. Human myoglobin recognition of oxygen: Dynamics of the energy landscape. *PNAS*, 101(52):18000–18005, 2004.

[136] C. Wiesmann, G. Beste, W. Hengestenberg, and G. E. Schulz. The three-dimensional structure of 6-phospho-β-galactosidase from *Lactococcus lactis*. *Structure*, 3:961–968, Sep 1995.

[137] P. A. Wiggins and P. C.Nelson. Generalized theory of semiflexible polymers. *Phys. Rev. E*, 73:031906, 2006.

[138] Paul A. Wiggins, Thijn van der Heijden, Fernando Moreno-Herrero, Andrew Spakowitz, Rob Phillips, Jonathan Widom, Cees Dekker, and Philip C. Nelson. High flexibility of dna on short length scales probed by atomic force microscopy. *Nat Nano*, 1(2):137–141, 2006.

[139] A.P. Wolffe and D. Guschin. Review: chromatin structural features and targets that regulate transcription. *J. Struct. Biol.*, 129:102–122, 2000.

[140] J.L. Workman and R.E. Kingston. Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu. Rev. Biochem.*, 67:545–79, 1998.

[141] B. Xia, V. Tsui, D.A. Case, H.J. Dyson, and P.E. Wright. Comparison of protein solution structures refined by molecular dynamics simultion in vacuum, with a generalized born model, and with explicit water. *J. Biomol. NMR*, pages 317–331, 2002.

[142] Zhiwi Xu, Luis L. EscamillaTrevino, Lihui Zeng, Mallikarjun Lalgonar, D. R. Bevan, Brenda S. J. Winkel, Ali Mohamed, Chi-Lien Cheng, Ming-Che Shih, Jonathan E. Poulton, and Asim Esen. Functional genomic analysis of *Arabidopsis thaliana* glycoside hydrolase family 1. *Plant Mol. Biol.*, 55:343–367, 2004.

[143] J. Yan and JF. Marko. Localized single-stranded bubble mechanism for cyclization of short double helix DNA. *Phys. Rev. Lett.*, 93(10):108108–108108, Sep 2004.

[144] D.L. Zechel and S. Withers. Glycosidase mechanisms: anatomy of a finely tuned catalyst. *Acc. Chem. Res.*, 33(1):11–18, 2000.

[145] Y Zhang and DM Crothers. Statistical mechanics of sequence-dependent circular DNA and its application for DNA cyclization. *Biophys J*, 84(1):136–153, Jan 2003.

[146] J. Zouhar, J. Vévodová, J. Marek, J. Damborský, X.D. Su, and B. Brzobohatý. Insights into the functional architecture of the catalytic center of a maize β-glucosidase zm-p60.1. *Plant Physiology*, 127:973–985, 2001.