Spring 2010

Conformational Change and Topological Stability of Proteins

Jeffrey Andrew Tibbitt
Old Dominion University

Recommended Citation

Tibbitt, Jeffrey A.. "Conformational Change and Topological Stability of Proteins" (2010). Doctor of Philosophy (PhD), Dissertation, Chemistry & Biochemistry, Old Dominion University, DOI: 10.25777/3vc2-0633
https://digitalcommons.odu.edu/chemistry_etds/34

CONFORMATIONAL CHANGE AND TOPOLOGICAL STABILITY OF PROTEINS

by

Jeffrey Andrew Tibbitt
B.S. May 2001, Virginia Polytechnic Institute and State University

A dissertation submitted to the faculty of Old Dominion University
in partial fulfillment of the requirement for the degree of

DOCTOR OF PHILOSOPHY

CHEMISTRY

OLD DOMINION UNIVERSITY
May 2010

Approved by:

Jennifer Poutsma (Co-Director)

Lesley Greene (Co-Director)

Craig Bayse (Member)

Kenneth Brown (Member)

Christopher Osgood (Member)

Patricia Pleban (Member)

ABSTRACT

CONFORMATIONAL CHANGE AND TOPOLOGICAL STABILITY OF PROTEINS

Jeffrey Andrew Tibbitt
Virginia Polytechnic Institute and State University, 2001
Directors: Drs. Jennifer Poutsma and Lesley Greene

The conformation and topology of a protein changes when stabilizing forces are absent, but the mechanisms by which these changes occur remains elusive. This dissertation aims to broaden the understandings. On the conformational level, the M20 loop conformers of E. coli dihydrofolate reductase are interrogated to identify factors responsible for their stability as well as to determine how one conformer might change into another. Molecular dynamics is used to simulate the open, closed and occluded conformers (observed in X-ray crystal structures) under a series of different single ligand conditions. Analysis shows that all open conformers move to a similar new conformation. Free energy methods examine the stability of the new loop conformer relative to the others. External perturbation molecular dynamics simulations and normal mode analysis methods examine possible M20 loop pathways occurring either when one loop conformer is forced to change into another or when a ligand is pulled out of its binding site.

On the topological level, conserved residue-residue interaction networks found among three different protein superfamilies (the all α-helix death domains, the α/β-plaits and the all β-sheet immunoglobulins), each of different secondary structure but sharing the Greek-Key topology, are assessed for any inherent stability they might contain relative to randomly selected interaction networks. This assessment is achieved by simulating one protein from each family at different temperatures, ranging from 300 to 600 K, and observing that adding thermal energy to the system causes the random interaction networks to fall apart more easily than the conserved networks.

When considered together, the conformational and topological projects, although very different from each other, both demonstrate the same idea - that regardless of scale, instability causes change and vice versa. This dissertation is divided into five chapters: Introduction, Theoretical Background, M20 Loop Conformers of Dihydrofolate Reductase, Conserved Contact Networks of Greek-Key Proteins and Summary.

This dissertation is dedicated to my father, Nelson Paul Tibbitt Jr,
because he showed me that education is the road to knowledge.
I love you Dad.

ACKNOWLEDGMENTS

I thank God for being my guide through this adventure and for helping me up when I fell down. My family; Dad, Mom, Sonny, Mark, Christie, Julie and all the rest, has been there the entire way offering unyielding support. Thank-you all so much! My advisor Jennifer Poutsma started me off with a few simple UNIX commands and eventually passed the reigns over. I really appreciate her guidance and trust. She has been the greatest. A large thanks is due to my other advisor, Lesley Greene, for allowing me to join her group to take part in the large multidisciplinary study on Greek-Key proteins. This project was a collaboration, in which I contributed knowledge of and work on MD simulations using CHARMM. She provided the research project aims on the *in silico* studies of network stability, including the supervision and monetary support. She has enthusiasm and support that never fails. Thank you to Joshua Pothen for providing help in this collaboration, including among many other things, advice on how the structural alignments and the generation of networks worked. A great many thanks goes out to all the present and past coworkers of rooms 310 and 406 of the Alfriend chemistry building, Chuanyin, Sonya, Anna, Josh, Randy, Jennifer, Amy, Jessica, Donna, Christine, Joe, Courtney, Cliff, Laura, Deborah, Rajaa and others, for all the help and interesting conversations and arguments. Thank-you to the excellent Chemistry Department staff, including the two departmental secretaries, Lennis and Janice, for helping me out of many different jams, the graduate program secretary, Valerie Decosta, for helping out with much paperwork, and of course Tammy and Alicia, for providing me with much help and directions concerning everything I may and may not do in the lab. The National Institute of Health (NIH) is acknowledged funding for a portion of this research.

TABLE OF CONTENTS

Page

Chapter

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Proteins constantly change shape. From the smallest bond stretches, to the larger conformational shifting of loops and domains, to the complete unfolding, these changes are all controlled by the energy available. The malleable polymeric structure of proteins allows for almost countless shapes; hence, they possess very complex energy landscapes, and the binding of ligands magnifies the complexity. The paths proteins take as they transform, either by themselves or upon binding to other molecules, are of immense interest to the scientific community and in many cases are still not well understood.

As an example, the quaternary structure of hemoglobin undergoes a conformational shift as oxygen binds to one of its four identical heme groups (Figure 1) (1-2). Prior to binding, the heme is dome-shaped, with the iron sitting slightly out of

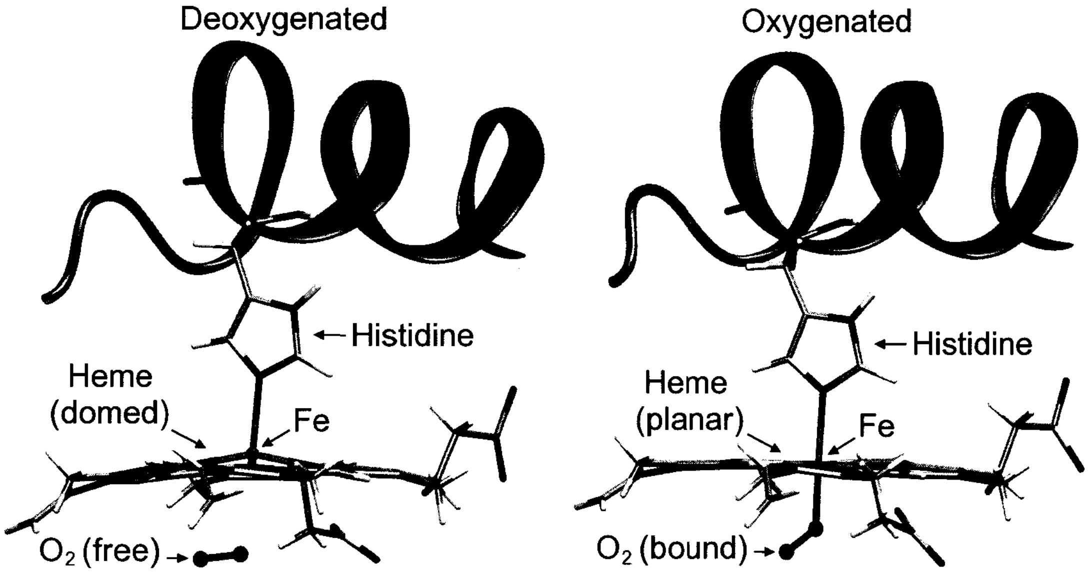


Figure 1. Conformational change of hemoglobin as oxygen binds. The molecules here (PDB codes: left=2DN2, right=2DN1) and in all other images were drawn in VMD (3).

The journal model followed here is the *Proceedings of the National Academy of Sciences*.

plane towards the histidine residue; upon binding, the metal atom attains an octahedral geometry and the heme becomes planar. The change in the local geometry surrounding the iron causes a shift in the entire subunit. The subunit shift triggers similar shifts in the other three subunits that allow them to bind oxygen more easily. This process involves changes in bond lengths, bond angles, dihedral angles and so on, but the exact order of events (i.e. the pathway of conformational change) remains elusive (4-5).

Today, two big problems in protein science are mapping the pathway of conformational change between two different structures and solving the three-dimensional structure of a protein given only its amino-acid residue sequence (a.k.a. the protein folding problem). A major obstacle is the experimental difficulty involved with observing proteins in real time; this is due to their sub-microscopic size and picosecond-level motions. Ideally, the positions of all atoms in the protein would need to be known at each point in time. Nuclear magnetic resonance (NMR) and X-ray crystallography are the two most popular experimental methods that can map atomic positions (6-8). The timescale of motions within proteins ranges from picosecond atomic fluctuations up to millisecond and second foldings (9). This is too fast for either conventional X-ray crystallography or NMR to handle, since they both capture time-averaged data.

Advances in time-resolved crystallography now allow X-ray snapshots to be taken on the picosecond level (10-13), and there have been many reported 'molecular movies' of protein conformational changes (14-19) and more recently of an actual protein folding (19). One study examined the nature of ligand binding to myoglobin (a protein that contains the same type of oxygen-binding heme group as hemoglobin does) (15). Mutational studies on this protein showed that Leu29 replaced with phenylalanine resulted in the elevation of O_2-binding affinity by an order of magnitude (20). The time-resolved structures of the ligand-binding process of both the wild-type and L29F proteins revealed that the size of the added phenylalanine residue was the mechanistic origin of this functional difference. Previous studies on wild-type myoglobin had indicated that entering/exiting ligands were trapped in a 'holding cell' in close proximity to the heme group before being released (21-22). The time-resolved structures of the wild-type binding confirm this. In the mutant binding, they reveal that the large benzene ring in the side chain of Phe29 pushed aside the nearby His64, which in the wild-type enzyme would normally reside next to the binding site. Thus in the mutant enzyme, ligand rebinding is easier due to less steric hindrance from the His64 and the destruction of the holding cell. The time-resolved snapshots in this study provided several missing mechanistic details.

However, several problems with this experimental technique still remain. Even at 100-ps, the time-resolution was not quick enough to show the ligand moving from the binding site over to the holding cell. Other problems are that the resolution remains low when compared to conventional crystallography techniques (19), many proteins cannot be easily studied by X-ray crystallography (15) and the technique is very time-consuming and expensive (16). Regardless of these issues, experimental tools such as time-resolved X-ray crystallography are invaluable for observing phenomena.

The two problems in protein science mentioned above will be understood once models correctly predict their solutions without need for prior experimental observation. In other words, the process of building models to correctly predict protein behavior is a process of understanding. Computer simulations are currently used to 'observe' protein conformational changes and foldings, however their accuracy needs to be confirmed by comparison with experimental data. Probably the most powerful and most popular simulation method is molecular dynamics (23-25). This technique models molecular motion by placing the atoms inside of a programmable force field, and allowing them to move through Newtonian mechanics. Molecular dynamics force fields are always being improved by tuning them to agree with newer, more detailed experimental data. Experimental methods (e.g. time-resolved X-ray crystallography) provide extremely detailed data, which apart from being useful in itself, helps to test the accuracy of theoretical methods. So, more accurate experimental methods allow scientists to improve their models, which in turn, allows them to better predict the conformational changes and folding of any unknown protein.

This dissertation uses molecular dynamics in two separate studies, the first involving protein conformational changes and the second involving protein (un)foldings. In the first study (Chapter III), several different conformations of Escherichia coli dihydrofolate reductase are modeled under equilibrium conditions. Results indicate the existence of a new conformation. Simulated conformational changes show that this new conformer plays a role when one conformer transitions to another. In the second study

(Chapter IV), proteins sharing a Greek-Key topology are simulated at several different high temperatures, with the highest ones causing the proteins to unfold. The effect on long-range conserved contacts is closely monitored. Results support a novel hypothesis, which could eventually lead to the ability to predict protein structure from an amino-acid sequence. The next chapter discusses the fundamentals of the main theoretical methods used in these two studies.

CHAPTER II

THEORETICAL BACKGROUND

The research methods used in this dissertation are theoretical in nature. The four major tools used are molecular dynamics, targeted molecular dynamics, free energy analysis and normal mode analysis. Here, the framework of each is reviewed. These theories are mathematically complex and using them for the purposes of the work in this dissertation requires both the speed of computers, as well as numerical approximation methods for implementing them. Except for the integration scheme involved in the theory of molecular dynamics, the rest of the approximation methods lie outside the scope of this review. More information can be found in the literature (26-28). Theories are concisely developed from first principles, with lighter treatment offered where appropriate. Mathematical derivations assume familiarity with calculus and linear algebra. Vectors and matrices are represented as bold letters.

MOLECULAR DYNAMICS

The most popular method for simulating the motion of large molecules is molecular dynamics (29-33). The algorithm involves solving the classical equations of motion. The solution generates the trajectory of the molecule as it moves through the 3N-dimensional configuration space, R^{3N}, where N is the number of atoms in the molecule and each point in the space represents an entire atomic configuration. During this motion, the atomic velocities change in reaction to forces acting on them, which in turn alters the forces. Assuming the forces in the system depend only on the current configuration of the molecule (i.e. they are path independent, or conservative), then they are derivable from a potential energy function as its negative gradient. A potential energy surface of the molecule can then be defined. The valleys and peaks of the surface represent relatively stable and unstable configurations, respectively. The interplay between velocities and forces can be treated as an exchange between the kinetic and potential energies possessed by the molecule. The passage of time funnels the molecule along an energetically feasible path on this surface. The mechanical concepts described above form the theoretical basis of molecular dynamics. The following derives the dynamic equations of motion used for simulating molecules.

Given (at time t) the atomic positions, **x**, and their respective velocities, **v**, the positions at a slightly later time (t + dt) are approximated by the second degree Taylor expansion

$$\mathbf{x}(t+dt)=\mathbf{x}(t)+\frac{\partial\mathbf{x}(t)}{\partial t}dt+\tfrac{1}{2}\frac{\partial^2\mathbf{x}(t)}{\partial t^2}dt^2=\mathbf{x}(t)+\mathbf{v}(t)dt+\tfrac{1}{2}\mathbf{a}(t)dt^2, \qquad [1]$$

where **a** is the atomic acceleration vector and t is the time. The acceleration vector is known if the forces on each atom (i.e. the force vector) are known. Using Newton's law relating force to acceleration, **F=ma**, and extending the infinitesimal time (dt) to a finite time-step (Δt), Equation 1 becomes

$$\mathbf{x}(t+\Delta t)=\mathbf{x}(t)+\mathbf{v}(t)\Delta t+\frac{\mathbf{F}(t)}{2\mathbf{m}}\Delta t^2, \qquad [2]$$

where **F** and **m** are the atomic force and mass vectors, respectively. Similarly, the positions at a previous time-step are

$$\mathbf{x}(t-\Delta t)=\mathbf{x}(t)-\mathbf{v}(t)\Delta t+\frac{\mathbf{F}(t)}{2\mathbf{m}}\Delta t^2. \qquad [3]$$

Adding Equations 2 and 3 produces the Verlet integration algorithm,

$$\mathbf{x}(t+\Delta t)=2\mathbf{x}(t)-\mathbf{x}(t-\Delta t)+\frac{\mathbf{F}(t)}{\mathbf{m}}\Delta t^2. \qquad [4]$$

The velocities can be estimated by

$$\mathbf{v}(t) = \frac{\mathbf{x}(t+\Delta t) - \mathbf{x}(t-\Delta t)}{2\Delta t} . \qquad [5]$$

Verlet integration can become numerically unstable. This is because a small number ($\sim\Delta t^2$) is added to the difference of two large numbers (Equation 4). The Leapfrog algorithm, a more numerically well-behaved variation of the Verlet scheme (and directly derivable from it), is applied in this dissertation and uses the following relationships:

$$\mathbf{x}(t+\Delta t) = \mathbf{x}(t) - \Delta t \mathbf{v}(t+\tfrac{1}{2}\Delta t) \qquad [6]$$

$$\mathbf{v}(t+\tfrac{1}{2}\Delta t) = \mathbf{v}(t-\tfrac{1}{2}\Delta t) + \frac{\Delta t \mathbf{F}(t)}{\mathbf{m}} , \qquad [7]$$

where the evaluation of the positions and velocities are off by a half time-step.

Starting a molecular dynamics simulation requires three items: initial positions, initial velocities and the forces acting on the atoms. Coordinates for proteins are obtained from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB). Initial velocities are randomly assigned to all atoms according to a Boltzmann distribution at the desired simulation temperature. The forces are obtained from a potential energy function, V, (also known as a force field) by

$$\mathbf{F}(t) = -\nabla V . \qquad [8]$$

Common force fields for proteins are CHARMM22 (34-35), AMBER-99 (36), GROMOS96 (37) and OPLS-AA (38-40). All are similar, as exemplified by the CHARMM22 potential,

$$V = \sum_{\text{bonds}} k_b(b-b_0)^2 + \sum_{\text{angles}} k_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} k_\phi\left[1+\cos(n\phi-\delta)\right] \qquad [9]$$
$$+ \sum_{\text{impropers}} k_\omega(\omega-\omega_0)^2 + \sum_{\text{Urey-Bradley}} k_\mu(\mu-\mu_0)^2$$
$$+ \sum_{\text{nonbonded}} \varepsilon\left[\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{12} - \left(\frac{R_{\min,ij}}{r_{ij}}\right)^{6}\right] + \sum_{\text{nonbonded}} \frac{q_i q_j}{\varepsilon r_{ij}}.$$

Figure 2 describes the terms in Equation 9. The first term is harmonic, and accounts for the bond stretches, where k_b is the bond force constant and b-b_0 is the distance from equilibrium for a given bond (Figure 2A). The second term, also harmonic, accounts for the bond angles, where k_θ is the angle force constant and $\theta-\theta_0$ is the degrees from equilibrium for a given angle (Figure 2B). The third term is periodic, and accounts for the dihedral (or torsion) angles, where k_ϕ is the dihedral force constant, n is the multiplicity, ϕ is the dihedral angle and δ is the phase shift (Figure 2C). The fourth term is harmonic, and accounts for the improper dihedral (i.e. the out of plane bending), where k_ω is the force constant and ω-ω_0 is the out of plane angle (Figure 2D). The fifth term, also harmonic, accounts for the Urey-Bradley potentials (or 1,3 nonbonded angle bending interactions), where k_μ is the force constant and μ-μ_0 is the distance from equilibrium of the 1,3-nonbonded (i.e. atoms bonded to a common atom) interactions (Figure 2E). Term six represents the classic far-attractive close-repulsive Lennard-Jones potential, where ε is the electric permittivity constant, r_{ij} is the distance between the two nonbonded atoms (i.e. atoms not bonded to each other or to another common atom) in the configuration and $R_{\min,ij}$ is the constant distance at which the potential is zero (Figure 2F). The seventh term is the electrostatic energy between the two atoms. This interaction arises from the partial charges, q_i and q_j, residing on the two atoms. The last two terms take up the vast bulk of simulation time. To minimize computing time, the nonbonded interactions (van der Waal and electrostatic) are ignored outside of a certain cutoff distance. Switching and shifting functions are used to smoothly truncate these interactions at that cutoff.

In order to run a molecular dynamics simulation, the force field must be given every parameter required by the seven terms of the energy function. Developing

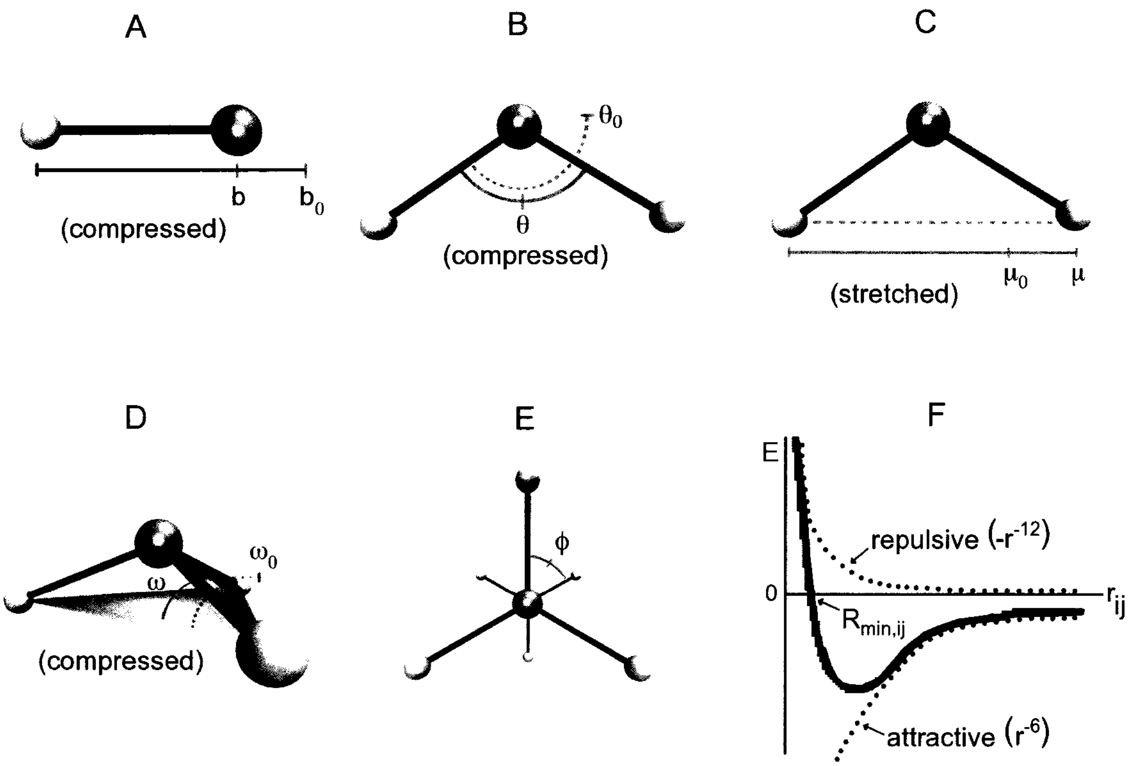


Figure 2. CHARMM22 potential energy function terms. Bond lengths (A), bond angles (B), Urey-Bradley 1,3 distances (C) and improper dihedrals (D) are all harmonic. Dihedral angle potentials (E) are cosine cyclic (with period 2π). van der Waals potentials (F) are treated as Lennard-Jones.

parameters for molecules is usually accomplished via intensive quantum mechanical (ab initio or density functional theory) calculations, or through the use of experimental data. Parameters are developed that satisfactorily reproduce experimental values or the quantum mechanical data. Good parameter sets exist for all protein amino acids and countless other types of molecules.

As proteins are in solution *in vivo*, so they must be *in silico*. Thus the protein is placed inside of a box filled with water molecules, then the overlapping water molecules are removed. Apart from questioning whether the chosen water model accurately portrays the effects of real water, there remain two problems. First, it increases the number of atoms, which in turn, lengthens the simulation time. Therefore, the protein is solvated in just enough water to cover the nonbonded cutoff space surrounding the protein atoms. Secondly, the use of a water box creates boundary conditions. Since the protein atoms smoothly truncate their nonbonded forces inside the box, none of these atoms directly 'feel' the boundary via the force field. But this is not the case for the water molecules, especially those close to the boundary. So, the protein indirectly 'feels' the boundary by directly interacting with affected waters.

The second problem is overcome by using periodic boundary conditions. The simulation box is infinitely replicated in all directions. Each box is equivalent to its neighbor, with their respective atoms occupying the exact same relative positions. Figure 3 shows the two-dimensional version of such a periodic system. If a water molecule leaves on the right side of the center cell, its image enters on the left side. It is really just the same molecule being wrapped around. In other words, the atoms feel no edges

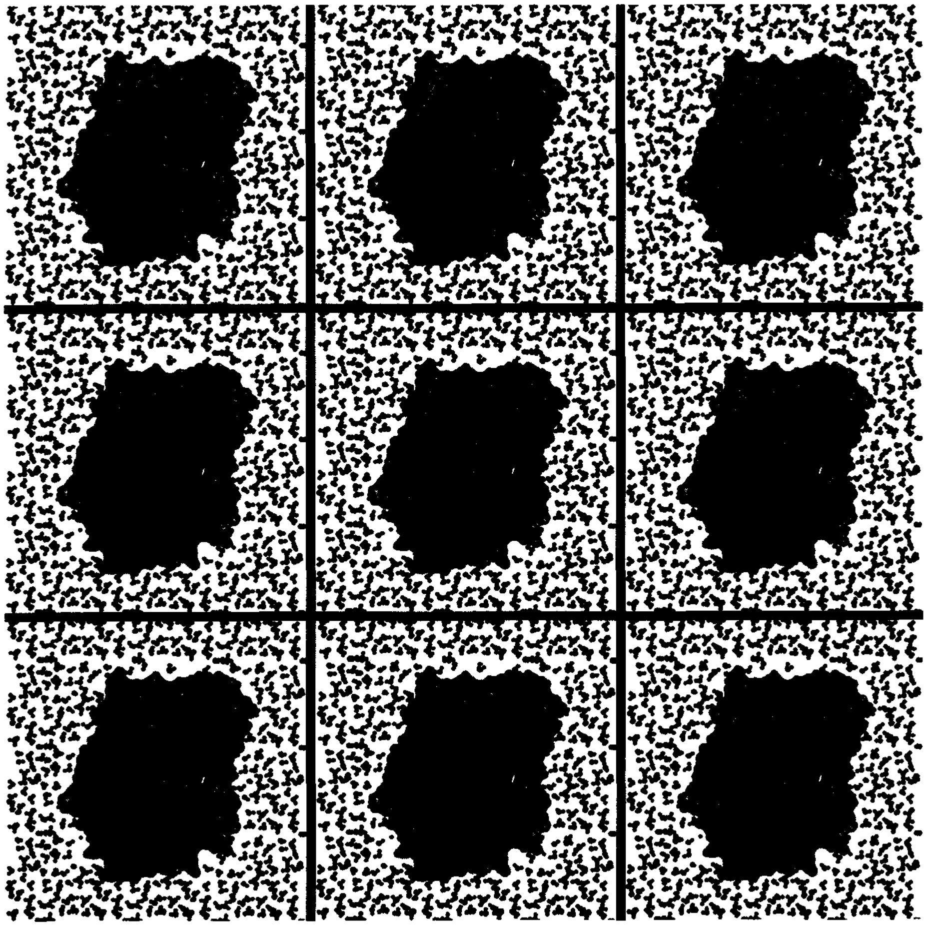


Figure 3. Two-dimensional periodic boundary conditions. Human death domain (1E3Y), solvated in a cubic box of water is drawn as a flattened 2-D image.

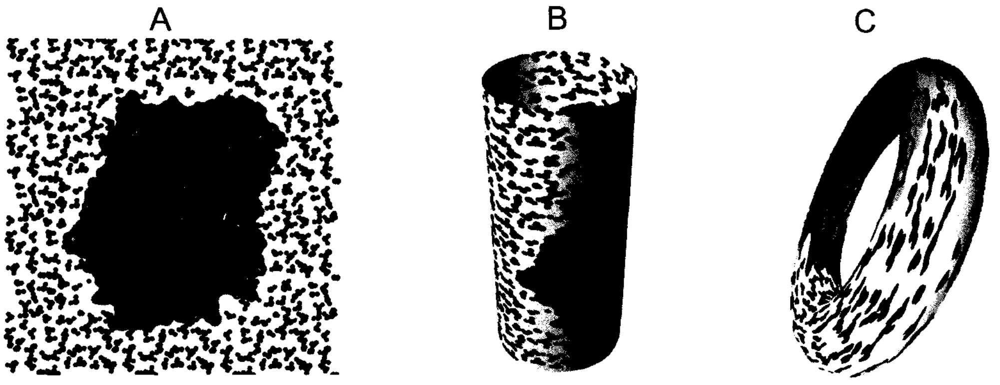


Figure 4. Torus representation of periodic boundary conditions. dihydrofolate reductase (1RD7), solvated in a cubic box of water is drawn as a flattened 2-D rectangle (A). The sides are joined by rolling the sheet into a vertical cylinder (B). The top and bottom ends of the cylinder are joined to form a torus (C).

because there are none. So, in the two-dimensional case, the simulation box (Figure 4A) is first rolled up into a cylinder (Figure 4B), the ends of which are then connected to form a torus (Figure 4C). Thus, the topology of the two-dimensional periodic boundaries is equivalent to that of a three-dimensional torus. The two-dimensional atoms are embedded on the surface of the three-dimensional torus. Similarly, the opposite faces of a three-dimensional box are connected to form a four-dimensional torus. In this case, the three-dimensional atoms lie on the surface of the four-dimensional torus, with no boundaries or image atoms. Therefore when applying periodic boundary conditions, only the atoms in the center cell need to be tracked.

EXTERNAL PERTURBATION MOLECULAR DYNAMICS

Proteins exhibit motions on a wide variety of timescales. These include bond vibrations on the femtosecond timescale, surface sidechain rotations on the picosecond timescale, hinge bending between two domains on the picosecond to nanosecond timescale, conformational transitions on the nanosecond to microsecond timescale and folding on the millisecond to second timescale (41). Unfortunately it takes much longer to simulate these motions using normal molecular dynamics. For example, running an MD simulation of a single protein conformational change can easily require months of computer time, on even the worlds fastest supercomputers. Several methods exist, which speed up this process. This dissertation makes use of methods that facilitate both ligand unbinding and conformational change for the enzyme dihydrofolate reductase.

Ligand unbinding is accomplished simply by applying an external force on one of the ligand atoms to push it out of the binding pocket. The force is included with the rest of the forces on that atom. The force magnitude is linearly varied from zero to a predefined value over the course of the simulation. This allows the ligand time to first break the weak interactions holding it in place and then the stronger interactions. The direction is chosen such as to offer a path of least resistance between its current position and the protein exterior.

The targeted molecular dynamics (TMD) method forces structural changes to occur faster than they normally would during a simulation (42). An extra force is added in at each timestep pushing the system towards a target structure. The force arises from a time-dependent holonomic constraint based on the mass-weighted root-mean-square deviation (RMSD) between the current and target structures. The general form of the constraint is

$$\Phi[\mathbf{X}(t)] = \frac{\sum_{i=1}^{N} m_i \left|\mathbf{X}_i(t) - \mathbf{X}_{i,F}\right|^2}{\sum_{i=1}^{N} m_i} - \rho^2(t) = 0 \qquad [10]$$

,

where t is the current time, N is the number of atoms in the protein, $\mathbf{X}(t)$ is the 3N-dimensional time-dependent atomic coordinate vector, m_i is the mass of atom i, $\mathbf{X}_i(t)$ is the position of atom i at time t, $\mathbf{X}_{i,F}$ is the target position of atom i and $\rho(t)$ is the desired mass-weighted RMSD between the structure and target structure. Thus, the constraint is set equal to the difference between the current mass-weighted RMSD and the desired mass-weighted RMSD. Treating this difference as 0 means the constraint is holonomic and that system dynamics are restrained to occur on the confines of the ρ-hypersphere. In other words, adding in the constraint at one time-step forces the system to attain the desired RMSD (ρ) at the next time-step. So decreasing ρ at each successive time-step restrains the system trajectory along a path of decreasing RMSD with respect to the target structure. Since the constraint is holonomic it is incorporated into the original potential energy function using a Lagrange multiplier:

$$V' = V + \lambda\Phi, \qquad [11]$$

where V is the original potential (given by Equation 9) and λ is the Lagrange multiplier to be determined (see below). Hence, the net force on each atom is

$$\mathbf{F}_i' = \mathbf{F}_i + \mathbf{F}_i^C = -\nabla_i V - \lambda\nabla_i\Phi, \qquad [12]$$

where the first and second components on the right are the original and perturbing forces, respectively. Differentiating the second term on the right yields the perturbing force

$$\mathbf{F}_i^C(t) = \frac{2\lambda m_i}{\langle m \rangle}\left[\mathbf{X}_i(t) - \mathbf{X}_{i,F}\right], \qquad [13]$$

where ⟨m⟩ is the average atomic mass. The resulting integrated equations of motion under the Leapfrog algorithm (Equations 6 and 7) are

$$\begin{aligned}\mathbf{X}_i(t) &= \left[\mathbf{X}_i(t-\Delta t) + \mathbf{v}_i(t - \tfrac{3}{2}\Delta t)\Delta t + \frac{\mathbf{F}_i(t-\Delta t)\Delta t^2}{m_i}\right] + \left[\frac{\mathbf{F}_i^C(t-\Delta t)\Delta t^2}{m_i}\right] \\ &= \mathbf{x}_i(t) + \mathbf{p}_i(t),\end{aligned} \qquad [14]$$

where $\mathbf{x}_i(t)$ is the position in the absence of the holonomic constraint and $\mathbf{p}_i(t)$ is the perturbation (change in position) due to the constraint. A simplified expression for $\mathbf{p}_i(t)$ is obtained by incorporating Equation 13 into the $\mathbf{p}_i$ component of Equation 14:

$$\mathbf{p}_i(t) = \frac{2\lambda\left[\mathbf{X}_i(t-\Delta t) - \mathbf{X}_{i,F}\right]}{\langle m \rangle}\Delta t^2 = \gamma\left[\mathbf{X}_i(t-\Delta t) - \mathbf{X}_{i,F}\right], \qquad [15]$$

with γ (implicitly defining the Lagrange multiplier, λ) as

$$\gamma = \frac{2\lambda}{\langle m \rangle}\Delta t^2 \quad . \qquad [16]$$

This shows the perturbation is a simple scaling factor (γ) times the difference between the previous position and the target. To obtain γ, Equation 14 (lower form) is substituted into Equation 10 giving

$$\frac{\sum_{i=1}^{N} m_i\left\{\mathbf{p}_i^2(t) + 2\mathbf{p}_i(t)\left[\mathbf{x}_i(t) - \mathbf{X}_{i,F}\right]\right\}}{\sum_{i=1}^{N} m_i} + \frac{\sum_{i=1}^{N} m_i\left[\mathbf{x}_i(t) - \mathbf{X}_{i,F}\right]^2}{\sum_{i=1}^{N} m_i} - \rho^2(t) \qquad [17]$$

$$= \frac{\sum_{i=1}^{N} m_i\left\{\mathbf{p}_i^2(t) + 2\mathbf{p}_i(t)\left[\mathbf{x}_i(t) - \mathbf{X}_{i,F}\right]\right\}}{\sum_{i=1}^{N} m_i} + \Phi\left[\mathbf{x}(t)\right] = 0 \quad .$$

Now, substituting Equation 15 into Equation 17 gives γ (and hence λ) as the solution of

$$a\gamma^2 + b\gamma + \Phi\left[\mathbf{x}(t)\right] = 0 \, , \qquad [18]$$

where

$$a = \frac{\sum_{i=1}^{N} m_i\left[\mathbf{X}_i(t-\Delta t) - \mathbf{X}_{i,F}\right]^2}{\sum_{i=1}^{N} m_i} , \qquad [19]$$

$$b = \frac{2\sum_{i=1}^{N} m_i\left[\mathbf{X}_i(t-\Delta t)-\mathbf{X}_{i,F}\right]\left[\mathbf{x}_i(t)-\mathbf{X}_{i,F}\right]}{\sum_{i=1}^{N} m_i} \qquad [20]$$

and

$$\Phi\left[\mathbf{x}(t)\right] = \frac{\sum_{i=1}^{N} m_i\left[\mathbf{x}_i(t)-\mathbf{X}_{i,F}\right]^2}{\sum_{i=1}^{N} m_i} - \rho^2(t) \qquad [21]$$

.

The transition pathway generated by TMD may contain artifacts and abnormalities, which need to be addressed. First, the mass weighting factors in the derivations above ensures a stationary center of mass only if all atoms are perturbed. Second, the angular momentum of the protein is not theoretically conserved, even when all atoms are perturbed. The perturbation calculations described above suppose the simulation and target structures are both aligned. So, every so often, the target structure is aligned (using the constrained atoms for the fitting) with the current structure.

A larger problem arises when considering the direction and magnitude of the TMD perturbation. Conventionally, the ρ(t) parameter, which specifies the desired RMSD with respect to the target structure, is set to decrease evenly and monotonically with time. In other words, at each time-step, the protein is forced to occupy the adjacent RMSD hypersphere, which is closer to the target value. As shown in Figure 5, even if the unconstrained molecule overshoots the next hypersphere, and gets closer to the target, the perturbation brings it back. And since these perturbations are only based on RMSD, they may force the crossing of large energy barriers.

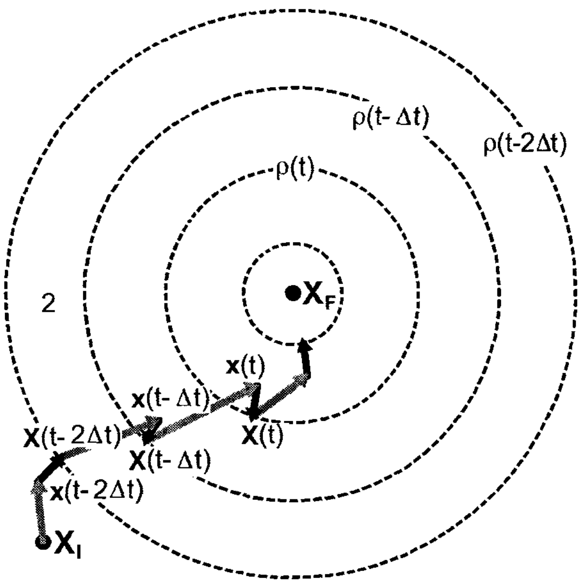


Figure 5. TMD movements along successive ρ-hyperspheres. The initial configuration (X_I) approaches the final configuration (X_F) by moving along adjacent RMSD hyperspheres (ρ). Newtonian and perturbing movements are labeled with grey and black arrows, respectively. And Newtonian and final configurations (i.e. after applying the perturbation) are labeled with, respectively, lower case and upper case letters.

The Restricted-Perturbation TMD method (43) addresses these problems by both restricting the total magnitude of perturbation and minimizing the RMSD with the target structure at each step. It is designed to prevent large barrier crossings. The total perturbation is obtained by summing Equation 15 over all perturbed atoms:

$$\sum_{i=1}^{N} \left|\mathbf{p}_i\right| = \gamma \sum_{i=1}^{N} \left|\mathbf{x}_i(t-\Delta t) - \mathbf{x}_{i,F}\right| . \qquad [22]$$

Restricting the total perturbation to a certain amount, P_F, means the scaling factor, γ, takes the value $\pm\gamma_F$, where

$$\gamma_F = \frac{P_F}{\sum_{i=1}^{N}\left|\mathbf{X}_i(t-\Delta t)-\mathbf{X}_{i,F}\right|} \quad [23]$$

Solving for ρ(t) (using Equation 10) gives the RMSD at each step,

$$\rho(t) = \sqrt{\frac{\sum_{i=1}^{3N} m_i\left|\mathbf{X}_i(t)-\mathbf{X}_{i,F}\right|^2}{\sum_{i=1}^{N} m_i} - \Phi[\mathbf{X}(t)]} \quad [24]$$

Since the first term under the square root sign in Equation 24 is positive, the RMSD, ρ(t), is minimized by maximizing the second term, $\Phi[\mathbf{X}(t)]=-a\gamma^2-b\gamma$ (Equation18). $a>0$ by Equation 19. So $\Phi[\mathbf{X}(t)]=-a\gamma^2-b\gamma$ is maximized by letting $\gamma=+\gamma_F$ when $b<0$ and $\gamma=-\gamma_F$ when $b>0$. Of course, setting $P_F=0$ means $\gamma_F=0$ (Equation 23), and hence $\Phi[\mathbf{X}(t)]=0$ (Equation 18). Thus setting $P_F=0$ recovers the unperturbed dynamics.

To further prevent large energy barrier crossings, the already restricted value of the total perturbation, P_F, is reduced when the perturbation vector, **p**, is poorly aligned with the unperturbed force vector, **F**, since the latter points downhill in energy ($\mathbf{F}=-\nabla E$). The sign of the following parameter checks the alignment:

$$C = \sum_{i=1}^{N}\left|\mathbf{p}_i\right|\cos(\mathbf{p}_i,\mathbf{F}_i) \quad [25]$$

This perturbation reduction results in a relative increase in both the current RMSD relative to the target structure (the RMSD using the full value of P_F would be lower) and hence the simulation length. Although the RMSD is minimized at each step, there is no reason why it cannot increase temporarily, in order for the protein to skirt the edge of an unstable conformational zone.

FREE ENERGY ANALYSIS

The relative stabilities of two protein conformers are measured by their respective free energies. Free energies, however, are rather difficult to obtain, and their accurate calculation, is an ongoing research problem (44-47). The origin of this difficulty lies in the close relationship between free energy and the thermodynamic partition function the latter being most difficult to calculate under simulation, since it requires a sampling of the entire configuration space. The difference in free energy between two structures is easier to calculate. The most rigorous methods for this are free energy perturbation (31) and thermodynamic integration (48). The former requires calculations on an entire path of structures between the two conformers. The latter requires the free energy to be a function of some parameter, which defines the thermodynamic state (i.e. conformation), thus allowing the free energy difference to be calculated by integrating it as a function of the change in the parameter. Since these are very time-consuming, attention has been focused on end-point calculations, where only the two structures being compared need to be analyzed (49-51). Most of these methods decompose the free energy into separate components. The Molecular-Mechanics Poisson-Boltzmann Surface-Area (MMPBSAA) method is a popular end-point method, which splits the free energy into four parts: gas phase molecular-mechanics energy, free energy of solvation – composing both electrostatic (Poisson-Boltzmann) and nonpolar (Surface-Area) components, and entropy (51-55). This dissertation uses the MMPBSA method to compare the relative stabilities of the different M20 loop conformers of DHFR under different ligand environments. The method is briefly described here.

Relative free energies of protein conformers may be computed using structures obtained from a protein/solvent MD simulation, but without referring to the explicit solvent coordinates. The basis for this reduction of coordinates lies in statistical thermodynamics and is briefly outlined here under the framework reviewed by Gilson et al (46). The free energy of protein conformation C (C represents the ensemble of protein configurations corresponding to that particular conformation), relative to the unfolded state, is

$$\Delta G^C = -kT\ln\frac{\int_C \exp\left(-\frac{U}{kT}\right)d\mathbf{x}_1,\ldots,d\mathbf{x}_N,d\mathbf{y}_1,\ldots,d\mathbf{y}_M}{\int \exp\left(-\frac{U}{kT}\right)d\mathbf{x}_1,\ldots,d\mathbf{x}_N,d\mathbf{y}_1,\ldots,d\mathbf{y}_M}, \qquad [26]$$

where k is Botzmann's constant, T is the temperature, U is the potential energy, $\mathbf{x}_i$ (i=1,...,N) are the protein coordinates and y_i (i=1,...,M) are the solvent coordinates. The lower integral is the classical configuration integral of the protein/solvent system, while the upper integral function is the sum over the configurations of conformation C. The ratio represents the probability of occurrence of conformation C. The protein potential of mean force, $W(\mathbf{x}_1,\ldots,\mathbf{x}_N)$, introduced by Kirkwood in 1935 (56), gives the average force over all solvent configurations, and depends on the volume, V:

$$\exp\left(-\frac{W(\mathbf{x}_1,\ldots,\mathbf{x}_N)}{kT}\right) = -kT\ln\frac{V^N\int \exp\left(-\frac{U}{kT}\right)d\mathbf{y}_1,\ldots,d\mathbf{y}_M}{\int \exp\left(-\frac{U}{kT}\right)d\mathbf{x}_1,\ldots,d\mathbf{x}_N,d\mathbf{y}_1,\ldots,d\mathbf{y}_M}. \qquad [27]$$

Rewriting the free energy in terms of W allows the solvent terms to be integrated out into a constant, which then cancels, and results in

$$\Delta G^C = -kT\ln\frac{\int_C \exp\left(-\frac{W}{kT}\right)d\mathbf{x}_1,\ldots,d\mathbf{x}_N}{\int \exp\left(-\frac{W}{kT}\right)d\mathbf{x}_1,\ldots,d\mathbf{x}_N}. \qquad [28]$$

The potential of mean force (PMF) represents an effective free energy potential (i.e. $W \approx G$). Using it allows implicit inclusion of solvent effects without explicit reference to solvent coordinates.

In the MMPBSA method (51-55), the free energy (or PMF) decomposes as

$$\Delta G = E_{MM} + \Delta G_{PB} + \Delta G_{SA} - TS \; . \qquad [29]$$

This calculation requires a set of protein coordinates (e.g. those obtained from an MD trajectory). Comparing Equation 29 to the familiar G=H-TS, suggests that the first three components and the last component are the enthalpy and entropy, respectively. The first term is simply the protein's gas phase energy calculated with a standard molecular-mechanics force field (discussed in the previous section). The second term is the electrostatic part of the free energy of solvation of the protein and is calculated using

$$G_{PB} = \sum_{i=1}^{N} \frac{q_i \phi_i}{2} \; , \qquad [30]$$

where q_i is the partial charge on each atom and ϕ_i is the potential at each atom (obtained by solving a linearized version of the Poisson-Boltzmann equation). Solving the Poisson-Boltzmann involves treating each charged atom as being encompassed by other charged atoms within a protein dielectric ($\varepsilon = 1$) as well as being surrounded by solvent dielectric ($\varepsilon = 80$). The Poisson-Boltzmann free energy term (Equation 30) represents the electrostatic interaction energy between explicit charges in the protein and the implicit charge of the solvent. So the actual term used in Equation 29 is difference between G_{PB} solved for $\varepsilon=80$ and G_{PB} for $\varepsilon=1$. The nonpolar part of the free energy of solvation represents the energy necessary to create a cavity in the solvent. It is taken to be proportional to the solvent accessible surface area (SASA)

$$\Delta G_{SA} = \gamma A + \beta \; , \qquad [31]$$

where γ is the constant of proportionality and β is the y-intercept. The fourth component of the free energy is the gas phase entropy of the protein and is obtained via quasiharmonic normal mode procedures (explained in the next section).

NORMAL MODE ANALYSIS

The complete vibrational pattern of any molecule decomposes into a set of independent non-interacting vibrations all occurring simultaneously. These are the normal modes. This research utilizes them in two different ways: first, to measure the degree of similarity between the vibrational distortions of DHFR complexes and the various conformational changes that can occur, and second, to calculate the entropy component for the relative free energy analysis of the DHFR MD simulations. Normal mode theory, as it applies to the two problems, is discussed below; but first, the basic formalism of the theory is developed (57-58).

Consider first the one-dimensional problem of a point mass moving in a harmonic potential

$$V(x) = \tfrac{1}{2}kx^2, \qquad [32]$$

where x specifies the displacement from equilibrium, which is defined as the location where the potential is at its minimum, and k is the force constant describing the stiffness of the potential:

$$k = \left(\frac{\partial^2 V}{\partial x^2}\right). \qquad [33]$$

The force acting on the particle is:

$$F = -\frac{dV}{dx} = -kx. \qquad [34]$$

Newton's equation of motion for this particle is

$$m\frac{d^2x}{dt^2} = -kx, \qquad [35]$$

or

$$\frac{d^2x}{dt^2} = -\omega^2 x \qquad \text{with} \qquad \omega = \sqrt{\frac{k}{m}}. \tag{36}$$

Solving this differential equation for x gives

$$x(t) = A \cdot \cos(\omega t + \delta). \tag{37}$$

Thus x moves about its equilibrium position (x=0) with amplitude A, frequency ω and phase δ. The constants of integration, A and δ, depend upon the initial conditions. This is called a harmonic oscillation.

Vibrations of a polyatomic molecule are described similarly. Consider a molecule with N atoms, described by their 3N Cartesian coordinates, x_i (i=1, 2, ..., 3N), and under the influence of a potential V. The Taylor expansion of V is

$$V(x_1, x_2, \ldots, x_{3N}) = V_0 + \sum_{i=1}^{3N} \left(\frac{\partial V}{\partial x_i} \right)_{x=0} x_i + \tfrac{1}{2} \sum_{i=1}^{3N} \sum_{j=1}^{3N} \left(\frac{\partial^2 V}{\partial x_i \partial x_j} \right)_{x=0} x_i x_j + \ldots \tag{38}$$

This form of the potential is simplified by assuming the following three criteria. First, the constant potential lies at the zero point of the energy scale (i.e. V_0=0). Second, motion occurs about the equilibrium configuration of the molecule, so that the partial derivative of the potential energy for each coordinate equals zero. Third, the atoms do not move very far from their equilibrium positions, so the third order terms and above are negligible and can be ignored. This third assumption is the only approximation made so far. It is called the harmonic approximation. So, V becomes

$$V(x_1, x_2, \ldots, x_{3N}) = \tfrac{1}{2} \sum_{i=1}^{3N} \sum_{j=1}^{3N} F_{ij} x_i x_j, \tag{39}$$

or

$$V = \tfrac{1}{2}\mathbf{x}^T\mathbf{F}\mathbf{x}, \tag{40}$$

a multidimensional harmonic potential, with harmonic force constants given by the symmetric matrix, **F**:

$$F_{ij} = \left(\frac{\partial^2 V}{\partial x_i \partial x_j}\right)_{\mathbf{x}=0}. \tag{41}$$

Newton's equations of motion for each coordinate are then

$$m_i \frac{d^2 x_i}{dt^2} = -\frac{\partial V}{\partial x_i} = -\sum_{j=1}^{3N} F_{ij} x_j, \tag{42}$$

a set of 3N coupled differential equations. They are coupled because the various points of the potential energy depend on the coordinates of more than one atom. There exists another related set of coordinates, y_j ($j = 1, 2, \ldots, 3N$), called the normal coordinates, in which, Newton's equations are uncoupled, thus giving the potential energy a diagonal form:

$$V(y_1, y_2, \ldots, y_{3N}) = \tfrac{1}{2}\sum_{i=1}^{3N} \lambda_i y_i^2, \tag{43}$$

with new constants, λ_i, or in matrix form,

$$V = \tfrac{1}{2}\mathbf{y}^T \Lambda \mathbf{y}, \tag{44}$$

where Λ is the diagonal matrix with the new constants, λ_i, occupying the diagonal positions and zeros elsewhere. The following shows the identity of these new constants and the relation between **x** and **y**. The original form of the potential can be transformed to mass-weighted Cartesian coordinates by

$$V = \tfrac{1}{2}\mathbf{x}^{T}\mathbf{M}^{1/2}\mathbf{M}^{-1/2}\mathbf{F}\mathbf{M}^{1/2}\mathbf{M}^{-1/2}\mathbf{x}, \qquad [45]$$

where $\mathbf{M}^{1/2}$ is the mass matrix. The diagonal elements of $M^{1/2}$ are the square roots of the atomic masses and the off diagonal elements are 0. $\mathbf{M}^{-1/2}$ is its inverse. Because **F** is symmetric and both $\mathbf{M}^{1/2}$ and $\mathbf{M}^{-1/2}$ are diagonal, the triple matrix product, $\mathbf{M}^{-1/2}\mathbf{F}\mathbf{M}^{1/2}$ is also symmetric and hence diagonalized by an orthogonal transformation matrix, **L**:

$$\mathbf{M}^{-1/2}\mathbf{F}\mathbf{M}^{1/2}\mathbf{L} = \mathbf{L}\Lambda, \qquad [46]$$

or

$$\mathbf{M}^{-1/2}\mathbf{F}\mathbf{M}^{1/2} = \mathbf{L}\Lambda\mathbf{L}^{-1}, \qquad [47]$$

which is an eigenvector/eigenvalue equation, with **L** being the eigenvector matrix and Λ the corresponding eigenvalue matrix (containing eigenvalues, λ_i , along its diagonal and zeros elsewhere). The potential then becomes

$$V = \tfrac{1}{2}\mathbf{x}^{T}\mathbf{M}^{1/2}\mathbf{L}\Lambda\mathbf{L}^{-1}\mathbf{M}^{-1/2}\mathbf{x}. \qquad [48]$$

The normal coordinates can be defined as

$$\mathbf{y} = \mathbf{L}^{-1}\mathbf{M}^{-1/2}\mathbf{x}. \qquad [49]$$

Then, since the transpose of the orthogonal matrix, **L**, is the same as its inverse, $\mathbf{y}^{T}$ is given by

$$\mathbf{y}^{T} = \mathbf{x}^{T}\mathbf{M}^{\frac{1}{2}}\mathbf{L}, \qquad [50]$$

and the potential simplifies to Equation 44. The eigenvector matrix, **L** are the normal modes, and the matrix, Λ, are the corresponding eigenvalues. In this formalism, the motion of the molecule is a superposition of independent harmonic vibrations in each of the coordinates y_i ($i = 1, 2, \ldots, 3N$). These 3N types of motion are called the normal modes of the molecule. Treating each normal mode as a one-dimensional problem, the potential and motion of the kth normal mode are given by

$$V_k = \tfrac{1}{2}\lambda_k y_k^2 \qquad [51]$$

and

$$y_k(t) = A_k\cos(\omega_k t + \delta_k), \qquad [52]$$

where A_k and δ_k are given by the initial conditions, and

$$\omega_k = \sqrt{\lambda_k}. \qquad [53]$$

Using the eigenvector matrix, **L**, to transform from normal coordinates back to Cartesian coordinates, the motion in Cartesian space of the kth normal mode becomes

$$x_i(t) = L_{ik}A_k\cos(\omega_k t + \delta_k) \qquad i = (1,2,\ldots,3N). \qquad [54]$$

So in any one normal mode, all 3N atoms oscillate with the same frequency ω_k (and same phase δ_k) and relative amplitudes given by the kth column in the eigenvector matrix **L**.

Since normal modes describe atomic vibrations, they can be used to calculate the mean square fluctuations of the particle positions, $\langle x_i^2 \rangle$. The equipartition theorem of statistical mechanics shows that the average potential energy of a harmonic oscillator is

$$\langle V \rangle = \langle \tfrac{1}{2}kx^2 \rangle = \tfrac{1}{2}k_BT, \qquad [55]$$

where k is the force constant, k_B is Boltzmann's constant and T is the temperature. This yields

$$\tfrac{1}{2}k\langle x^2 \rangle = \tfrac{1}{2}k_BT, \qquad [56]$$

or

$$\langle x^2 \rangle = \frac{k_BT}{k} = \frac{k_BT}{m\omega^2}. \qquad [57]$$

Similarly, in normal mode k,

$$\langle V_k \rangle = \langle \tfrac{1}{2}\lambda_k y_k^2 \rangle = \tfrac{1}{2}k_BT \qquad [58]$$

and

$$\langle y_k^2 \rangle = \frac{k_BT}{\lambda_k} = \frac{k_BT}{\omega_k^2}. \qquad [59]$$

Using Equation 52, the amplitude is calculated by

$$\langle y_k^2 \rangle = \langle A_k^2 \cos^2(\omega_k t + \delta_k) \rangle = \tfrac{1}{2}A_k^2. \qquad [60]$$

Thus

$$A_k^2 = \frac{2k_B T}{\omega_k^2} \quad [61]$$

In Cartesian space, the motion of normal mode k is given by Equation 54, so

$$\langle x_k^2 \rangle = \langle L_{ik}^2 A_k^2 \cos^2(\omega_k t + \delta_k) \rangle = \tfrac{1}{2} L_{ik}^2 A_k^2 = L_{ik}^2 \frac{k_B T}{\omega_k^2} \quad [62]$$

Then it is easy to show that the total mean square fluctuation of coordinate i (across all normal modes) is

$$\langle x_i^2 \rangle = k_B T \sum_{k=1}^{3N} \frac{L_{ik}^2}{\omega_k^2} \quad [63]$$

and similarly that the total mean square fluctuation for atom j is

$$\langle \Delta r_j^2 \rangle = \langle x_{3(j-1)+1}^2 + x_{3(j-1)+2}^2 + x_{3(j-1)+3}^2 \rangle, \quad [64]$$

or

$$\langle \Delta r_j^2 \rangle = k_B T \sum_{k=1}^{3N} \frac{L_{3(j-1)+1,k}^2 + L_{3(j-1)+2,k}^2 + L_{3(j-1)+3,k}^2}{\omega_k^2} \quad [65]$$

Here, the subscripts are changed from i to j to emphasize the use of positional coordinates, **r**, as opposed to one-dimensional coordinates, x, where $x_{3(j-1)+1}$, $x_{3(j-1)+2}$ and $x_{3(j-1)+3}$ correspond to x, y and z of atom j. This says that the mean square position fluctuations of atoms, in a molecule under the influence of a harmonic potential, are proportional to temperature and that the lowest frequency vibrations (i.e. the normal modes with the largest eigenvalues) contribute the most to atomic motion.

Finally, consider the covariance matrix, in which fluctuations in the distances between different atomic coordinates are calculated. The element in the ith row and jth column of this matrix (i.e. the covariance between coordinates i and j) is given by

$$\sigma_{ij} = \left\langle x_i x_j \right\rangle = k_B T \sum_{k=1}^{3N} \frac{L_{ik} L_{jk}}{\omega_k^2} \quad . \qquad [66]$$

In matrix notation this equation is

$$\Sigma = k_B T \cdot \mathbf{L} \Lambda^{-1} \mathbf{L}^T \quad . \qquad [67]$$

where Λ^{-1} is the inverse of the diagonal matrix of eigenvalues $\lambda_k = \omega_k^2$, and σ_{ij} are elements of Σ. Combining this equation with Equation 47 gives

$$\Sigma = k_B T \cdot \mathbf{M}^{-\frac{1}{2}} \mathbf{F} \mathbf{M}^{\frac{1}{2}} \quad \text{or} \quad \mathbf{F} = \frac{\mathbf{M}^{\frac{1}{2}} \Sigma \mathbf{M}^{-\frac{1}{2}}}{k_B T} \quad . \qquad [68]$$

Thus, normal modes are obtained from the covariance matrix also. Thus normal modes can be extracted from any source that yields a covariance matrix. The quasiharmonic method (see below) uses the covariance matrix calculated from an MD trajectory.

Normal mode analysis is applied in two different ways in this dissertation. The first method measures the degree of similarity between vibrational distortions of DHFR complexes and various M20 loop conformational changes. Normal mode theory shows how the vibrational distortions decompose into independent normal modes, each described by its own directional eigenvector with 3N components in Cartesian space. A conformational change is represented by a directional difference vector, i.e. the coordinates of the target conformation minus the coordinates of the initial conformation, also with 3N components in Cartesian space. Therefore, the degree of similarity between the two is measured simply by the magnitude of their dot product. Modes with large dot products indicate oscillations occurring along the path of conformational change.

Before using this measure to compare the vibrational overlap with two or more conformational difference vectors, a few considerations need to be addressed. First, all conformational difference vectors used in the comparison are normalized prior to calculating the dot products, since it is only the direction that is of interest and not the magnitude. The normalization is necessary because it eliminates any bias (e.g. if one difference vector is larger in magnitude than the other vectors). Second, a calculated dot product may be negative, which means that the normal mode eigenvector is pointing in the opposite direction to the conformational difference vector. However, normal modes are oscillatory motions, and therefore vibrate back and forth with the same amplitude in both directions. Therefore the sign of the dot product is meaningless. Hence only the absolute values of the dot products are used in the comparisons. Third, as previously shown, the largest atomic displacements occur in the lowest frequency modes. In fact, studies show that the few lowest frequency normal modes of proteins account for virtually all of the larger slow motions (e.g. conformational changes), while the vast majority of higher frequency modes involve only quick movements (e.g. bond stretches and angle bends) (59-61). Therefore, only the dot products involving the first few lowest frequency modes need to be examined (59-62). To examine a superposition of normal modes, the corresponding dot products are simply summed together. Because normal modes occur simultaneously, a dot product with one mode may be negative, while that with another is positive. So in this case, the absolute value is taken after the dot products are summed together. These considerations allow the dot products (and their sums) to be a good measure for determining which conformational change vector best overlaps with the protein vibrations.

The second use of normal mode analysis (NMA) is for calculating the entropy of an MD simulation structure. This data is used in the calculation of the relative free energy of the DHFR MD simulations. Two things are required: a set of normal modes (specifically their frequencies) that represent the motion of DHFR during molecular dynamics simulations and a way to extract the entropy given the frequencies. The normal modes are extracted in three steps. First, the positional covariance matrix, Σ, along the simulation trajectory is calculated. Next (using Equation 68), this matrix is converted into a force constant matrix. Finally, in what is called the Quasiharmonic Approximation, this 'effective force constant matrix' is used in the same diagonalization procedure (described above) to obtain the normal modes.

The subsequent entropy calculation from the normal mode frequencies requires some elementary statistical mechanics. In particular, the quantum formalism of the harmonic oscillator is used. The energy levels of the one-dimensional harmonic oscillator are given by

$$E_n = \hbar\omega\left(n + \tfrac{1}{2}\right) \qquad n = 1,2,\ldots \tag{69}$$

The partition function for this system is

$$q = \sum_n e^{-\frac{E_n}{k_B T}} \tag{70}$$

This sum is calculated by converting it to a geometric power series,

$$q = e^{-\alpha/2}\sum_n e^{-\alpha n} = \frac{e^{-\alpha/2}}{1-e^{-\alpha}} = \frac{1}{e^{\alpha/2} - e^{-\alpha/2}}, \tag{71}$$

where the dimensionless quantity is

$$\alpha = \frac{\hbar\omega}{k_B T} \tag{72}$$

The Helmholtz free energy and average energy of the oscillator are

$$A = -k_B T \ln\left(q\right) = \frac{\hbar\omega}{2} + k_B T \ln\left(1 - e^{-k_B T \hbar\omega}\right) \tag{73}$$

and

$$E = -\left[\frac{\partial(\ln q)}{\partial(k_B T)}\right] = \frac{\hbar\omega}{2} + \frac{\hbar\omega}{e^{k_B T \hbar\omega} - 1}, \qquad [74]$$

respectively. Calculating the entropy using the familiar thermodynamic equation,

$$S = \frac{E - A}{T}, \qquad [75]$$

yields

$$S = k_B\left[\frac{\alpha}{e^{\alpha} - 1} - \ln\left(1 - e^{-\alpha}\right)\right]. \qquad [76]$$

The partition function in systems with many degrees of freedom is

$$Q = \prod_k q_k \quad . \qquad [77]$$

Finally, plugging this result into Equations 73 and 74 and solving Equation 75 shows that the entropy of a harmonic system with many degrees of freedom is

$$S = k_B \sum_k \left[\frac{\alpha_k}{e^{\alpha_k} - 1} - \ln\left(1 - e^{-\alpha_k}\right)\right], \qquad [78]$$

with

$$\alpha_k = \frac{\hbar\omega_k}{k_B T} \qquad [79]$$

The entropy contribution to the free energy is then just TS, where T is the temperature.

CHAPTER III

M20 LOOP CONFORMERS OF DIHYDROFOLATE REDUCTASE

BACKGROUND

Escherichia coli dihydrofolate reductase (DHFR) catalyzes the reduction of 7,8-dihydrofolate (DHF) to 5,6,7,8-tetrahydrofolate (THF) by cofactor nicotinamide adenine dinucleotide phosphate (NADPH). The enzyme is comprised of 159 amino-acid residues and folds into a globular-shaped structure consisting of an 8-stranded β-sheet surrounded by 4 α-helices (Figure 6). DHFR regulates the amount of THF available for cellular processes. Notably, the biosyntheses of purine and thymidylate require THF, thus making it necessary for DNA production, and in turn, cell growth. Although DHFR is found in all cell types, the most rapidly proliferating ones (e.g. skin, hair, growing fetus, and cancer) are more dependent on it. The enzyme's ubiquitous nature and vital cellular role make it a very popular target in the development of anticancer and antimicrobial drugs (63). In fact, circa 1950, the first anticancer drug targeted this enzyme (64-66). These applications rely on a predetermined knowledge of the structure, especially its active site region.

The M20 loop (residues 14 to 24) is a flexible portion of dihydrofolate reductase, connecting β-strand A to α-helix B (Figure 6) that covers the active site. During catalysis this loop changes shape, allowing ligands access to the active site. More simply, it opens and closes as the reactant and product molecules bind to or detach from the protein. However, an in depth crystallography study by Sawaya and Kraut (67), on a large number of DHFR complexes, over a wide variety of ligands, shows the loop preferring one of three well-defined conformations: open, closed or occluded, with the shape determined by the identity of the ligand(s) present (Figure 7). In the closed conformation, the M20 loop is characterized by a short antiparallel β-sheet with hydrogen bonds in residues 16-19 forming a type III' hairpin turn [Met16(CO)-Ala19(NH) and Met16(NH)-Ala19(CO)] and a pair of hydrogen bonds to the FG loop [Gly15(CO)-Asp122(NH) and Glu17(NH)-Asp122(Oε2)] that maintain the shape of the loop's N-terminal portion. In this conformation, the loop fits snug against the cofactor's nicotinamide binding pocket sealing it from the solvent. The occluded conformation contains a short 3_{10}-helix [Glu17(CO)-Met20(NH)] and a pair of hydrogen bonds on its C-terminal end to the GH

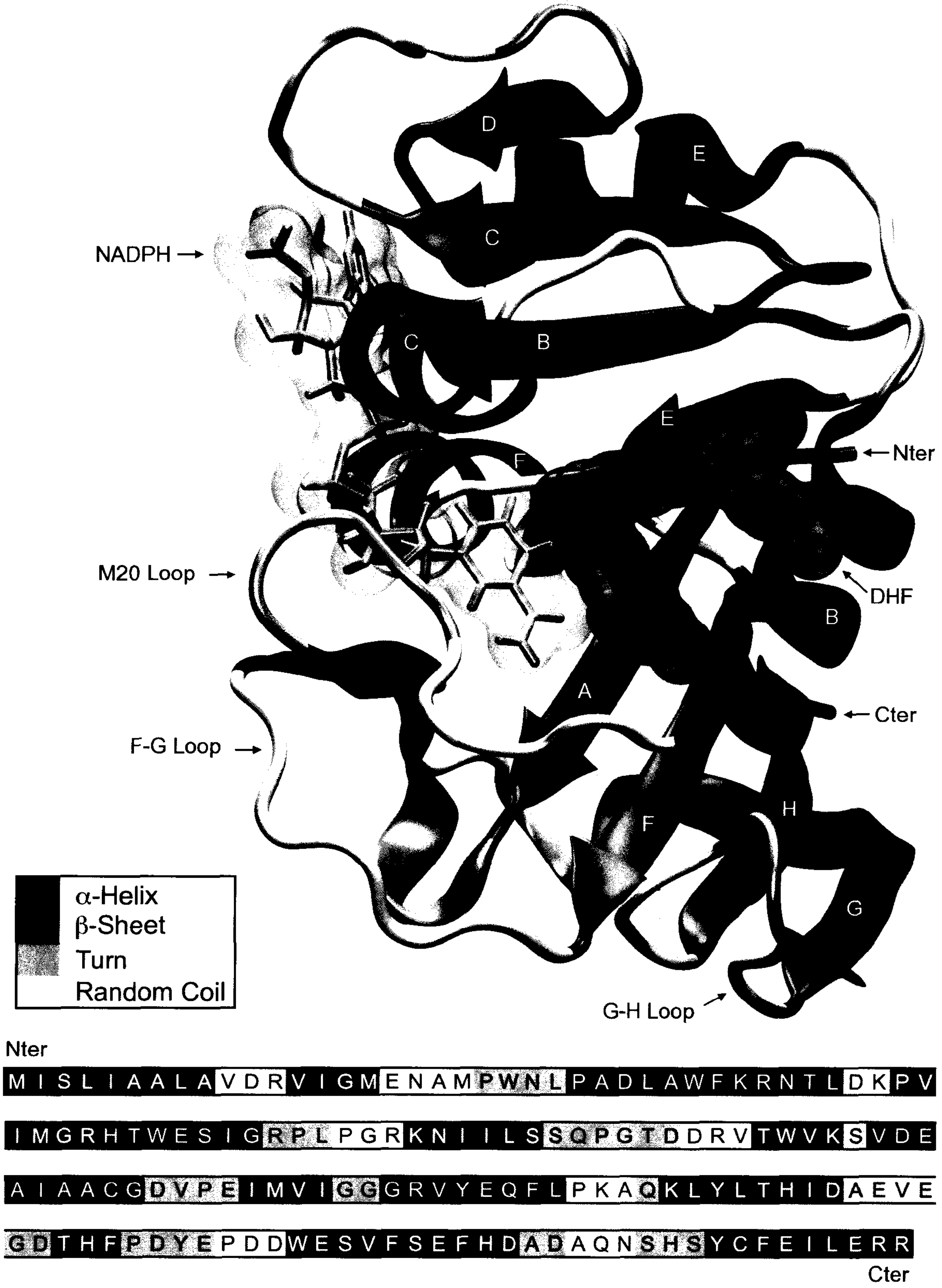


Figure 6. DHFR secondary structure. Shown at the top is the 3-D structure of DHFR with bound DHF and NADPH (PDB code 1RX2), and at the bottom, the corresponding 1-D amino acid residue sequence.

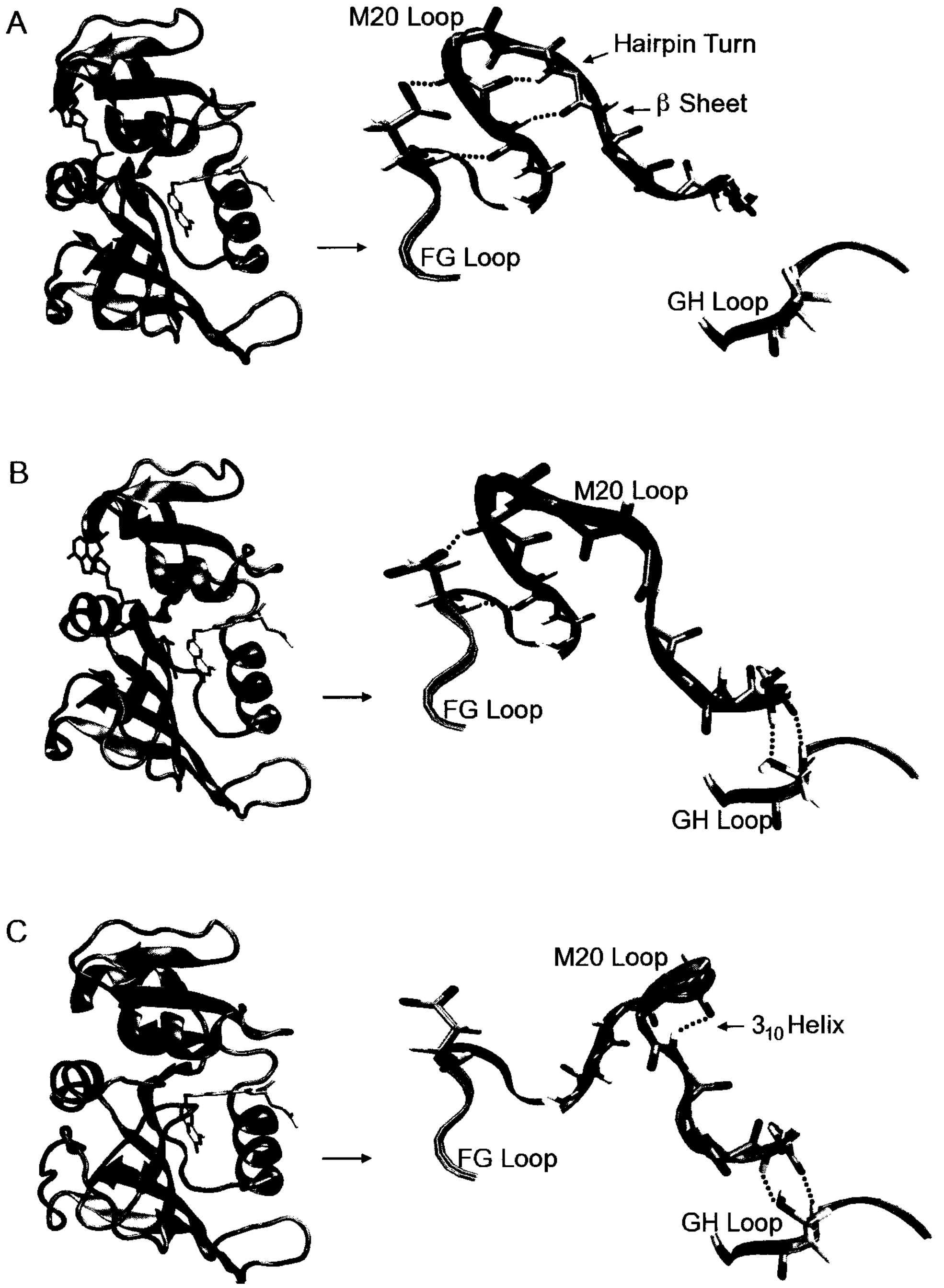


Figure 7. M20 loop conformers of DHFR. Shown are the closed (A), open (B) and occluded (C) conformers. The full protein images on the left show the highlighted M20 loop substrate and/or cofactor. Shown on the right are ribbon representations of the three major loops (M20, FG and GH), stick representations of the M20 loop backbone atoms and stick representations of all atoms in residues 122 and 148 of the FG and GH loops, respectively. This figure was constructed using PDB coordinates (1RX2 for closed, 1RD7 for open and 1RX7 for occluded).

loop [Asn23(CO)-Ser148(NH) and Asn23(NH)- Ser148(Oδ)]. Its N-terminal portion, being relieved of any interloop hydrogen bonds, protrudes into the active site, occluding the cofactor's nicotinamide binding pocket. The open conformer (also the most frequently occurring conformation crystallographically) contains one of the M20-FG interloop bonds [Gly15(CO)-Asp122(NH)] and both of the M20-GH interloop bonds. However, the loop's central portion does not resemble either of the other two conformers, but has a unique irregular shape, characterized by a large opening from the solvent to the interior of the active site. Although the open conformer is the most frequently occurring crystal conformer observed, it appears to be stabilized by crystal packing contacts. Every structure exhibiting the open conformer also reveals extensive contacts between the M20 loop of one enzyme and the GH loop of its two-fold symmetry related neighbor within the crystal lattice (Figure 8). It is possible that removing these contacts (as would exist *in vivo* conditions) would cause the loop to shift in order to restabilize itself.

Sawaya and Kraut were able to obtain isomorphous crystal structures (of space group $p2_12_12_1{}^B$) representing the five steady state catalytic intermediates of DHFR: E|NADPH, E|NADPH|DHF, E|NADP+|THF, E|THF and E|THF|NADPH (67). By using modified (but structurally similar) substrates and cofactors to mimic the reactive species, they found that each intermediate takes on either the closed or occluded conformation. From these results, they proposed an extended model of the catalytic cycle, which defines the loop conformation for each intermediate (Figure 9). The model shows the M20 loop closed in both the holoenzyme (E|NADPH) and the subsequent Michaelis complex (E|NADPH|DHF) and occluded the rest of the time. These results do not, however, provide any information on conformational changes for any complex or transitions between any two. Nuclear magnetic resonance (NMR) studies also lend support for Sawaya and Kraut's model. By using previously determined chemical shift markers (68-70), which distinguish between closed and occluded conformers, each catalytic intermediate (or representative complex mimicking a reactive intermediate) was observed to have the same M20 loop conformation seen by X-ray (71).

The interloop hydrogen bond contacts mentioned above, in addition to marking the different M20 loop conformers, are also vital for proper catalytic function. Mutagenesis studies targeting Asp122 of the FG loop (needed to stabilize the closed M20

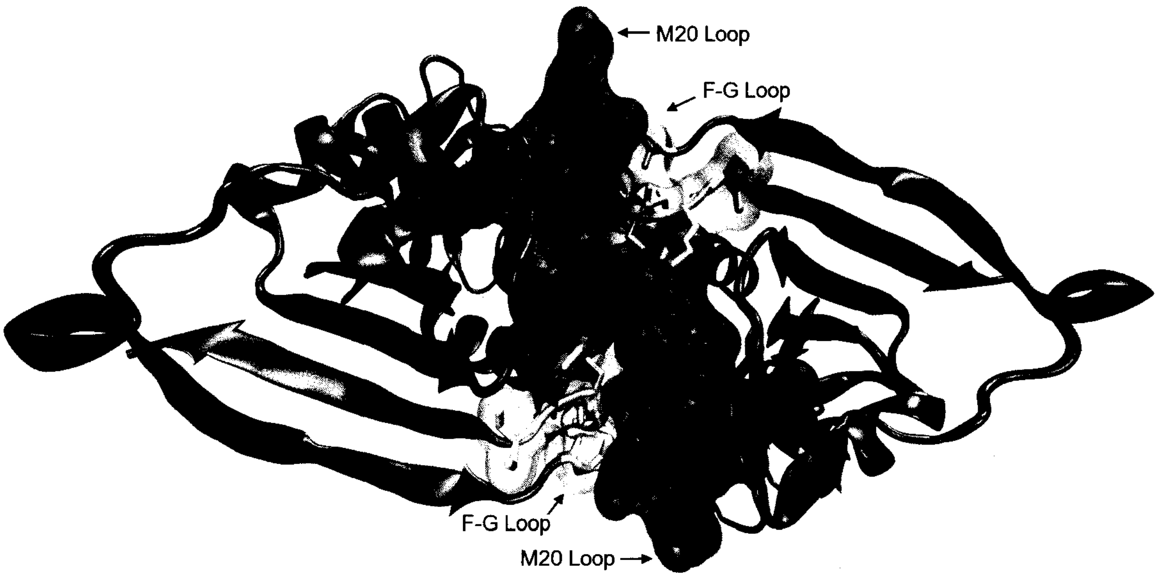


Figure 8. Crystal packing contacts of the open M20 loop conformer. Contacts exist between the FG loop and M20 loop of two symmetry-related proteins in the crystal lattice. Structure is drawn using the open M20 loop conformer of E. coli DHFR with folate bound (PDB code 1RD7).

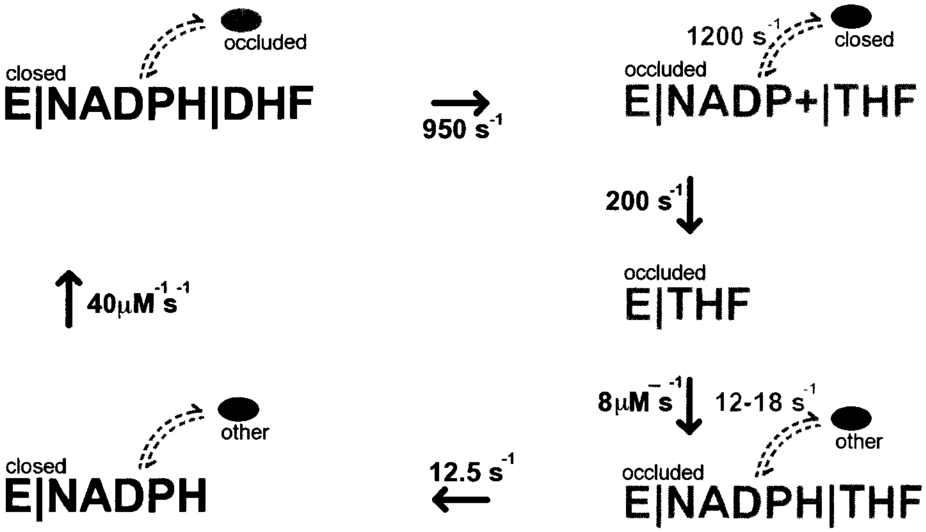


Figure 9. DHFR catalytic cycle. Steady-state rate constants in the forward direction are shown next to the straight black arrows connecting the catalytic intermediates. Also shown are conformational exchanges occurring between ground and excited states of the M20 loop. Exchange rate constants are shown next to the curved grey dashed arrows connecting the ground states and excited states. The M20 loop conformations (closed, occluded and other) in the ground states and excited states are specified.

loop conformer) (72) and Ser148 of the GH loop (needed for the occluded conformer) show the importance of these residues (73). First, substituting Asp122 with Asn, Ser or Ala (respectively decreasing in hydrogen bond forming ability) resulted in a decreased steady-state turnover rate. Not surprisingly, there was an increase in the cofactor dissociation rate from the Michaelis complex. This result was expected since without the hydrogen bonds between Asp122 and Gly15-Glu17, the M20 loop is better able to occlude the cofactor's binding pocket and hence push it back out of the active site. Unexpectedly, this mutation also increased the dissociation rate of the product NADP+ from the product complex E|NADP+|THF. This is strange since the crystal structure of this complex shows its M20 loop already occluded. Second, substituting Ser148 with Ala also decreased the steady-state turnover rate, and the NADP+ off rate from E|NADP+|THF was decreased. Without the hydrogen bonds connecting it to Ser148, the M20 loop should be freer to convert back to closed and trap the cofactor in the active site. Again, the mutation unexpectedly decreased the off rate of NADPH from E|NADPH|folate (also catalytically active, but with a slower reaction rate). The take home message is that destabilization of one M20 loop conformer also affects complexes which do not have that particular M20 loop conformation. The authors explained these strange results by suggesting that conformational exchange could be occurring in those complexes, which appear to only have one conformer via X-ray or NMR (72).

Evidence for closed-occluded conformational exchange within a DHFR catalytic intermediate was established by McElheny et al (74) by using NMR spin relaxation dispersion techniques (75). Nuclear spin relaxation results from time-dependent stochastic modulation of spin Hamiltonians, and when conformational dynamics occur on time scales comparable to or faster than the overall rotational correlation times, the spin relaxation rate constants are influenced. Measuring the dependence of transverse ^{15}N relaxation rate on the temporal separation between NMR pulses revealed exchange between a closed ground state and an occluded excited state occurring within DHFR|folate|NADP+ (the same ternary complex that the previously mentioned X-ray studies used for modeling the Michaelis complex of DHFR) (74). A more comprehensive study that followed revealed four of the five catalytic intermediates to be populating additional M20 loop conformers (76). Figure 9, which summarizes these results, shows that the Michaelis complex (closed ground state) populates an excited state conformer that resembles occluded. Likewise, the product complex (occluded ground state) populates an excited state conformer that resembles closed. E|THF|NADPH (occluded ground state) and E|NADPH (closed ground state) both populate a similar third conformation of unknown structure.

The observations described above suggest a mechanism for loop movement during the reactive step. Crystal structures show the occluded M20 loop occupying the same binding pocket as the cofactor's nicotinamide ring in the closed conformer. Therefore, as the Michaelis complex (E|NADPH|DHF) transitions from the closed ground-state to the occluded excited-state and back again (as observed by NMR), the nicotinamide ring must leave and reenter the binding pocket, respectively. Since hydride transfer can only occur in the closed conformation, the speed at which the reaction occurs may be limited by the speed at which conformational exchange occurs. Support for this theory comes from the similarity between the rate of conformational exchange within the product complex (1200 s^{-1}) and the steady-state hydride-transfer rate (950 s^{-1}) (77).

The combination of X-ray and NMR studies also sheds light on THF dissociation. The steady-state dissociation rate (12.5 s^{-1}) (77) from E|NADPH|THF, which is the rate-limiting step, is similar to the conformational exchange rate (12-18 s^{-1}) occurring in that complex (76). Thus product dissociation may occur from the excited state. In that case, the M20 loop in E|NADPH|THF would exchange conformation before the cycle proceeds to the next step. As noted in Figure 9, the excited state loop conformer of E|NADPH|THF resembles a structure other than closed or occluded. However, its structure does resemble the unknown exited-state conformer of the next catalytic intermediate, E|NADPH. Dissociation occurs more slowly from the product complex, E|NADP+|THF (2.4 s^{-1}) than from the ternary complex E|NADPH|THF (12.5 s^{-1}) (77). If the product complex were able to sample the unknown third conformer, like the ternary complex does, then, perhaps THF dissociation from E|NADP+|THF would occur just as quickly.

The third unknown M20 loop conformer observed by NMR (76) may be similar to the frequently seen open crystal structure (67). A separate NMR experiment by Kitahara et al (78) shows that under high-pressure the DHFR|folate binary complex adopts a second conformation apart from its ground state occluded form, which they propose may be the often-seen open crystal conformer (67). At normal pressure the second conformation is estimated to have a 10% population. In the two-dimensional $^{15}N/^{1}H$ HSQC spectrum, they observe splittings of Arg12, Val13 and Trp22 in the M20 loop and of Gly51 and Gly95 in α-helices C and F, respectively, all of which surround the cofactor binding site. The affected residues are similar to those areas with differences in the open and occluded crystal structures of DHFR|folate (PDB codes 1RD7 and 1RX7, respectively) (67). Additional calculations show a decrease in volume (ΔV = -25 mL/mol), enthalpy (ΔH = -43 kJ/mol) and entropy (ΔS = -0.15 kJ/K mol), as the complex transitions to the second higher energy conformer (78). These suggest the transition is accompanied by an increase in hydration, and a decrease in cavity volume. This agrees with Kitihara et al's (78) calculated decrease in cavity volume in the transition from the occluded conformer to open conformer in the crystal structures. The NMR evidence of the same open conformer as seen in crystallography is not well supported, especially after taking into account the crystal packing forces in all open conformer crystal structures (67) and the possible shift involved with removing those forces.

The considerations above demonstrate the close connection between change and stability for the M20 loop of E. coli DHFR. Crystallography reveals the stabilizing contact forces, within the M20 loop and between the M20 loop and the FG and GH loops, which are responsible for holding the separate M20 loop conformations in place (67). Mutagenesis shows that scaling down these contact forces (or removing them entirely) stabilizes one loop conformer over another and hence greatly reduces the enzyme's catalytic productivity (72-73). Furthermore, NMR experiments show how the M20 loop conformational exchange between the ground state and less stable excited state conformers allows catalysis to proceed from one intermediate to the next (78). The existence of the M20 loop conformational flexibility within E. coli DHFR is hypothesized to be the result of an evolutionary response to the increased NADP+ levels in E. coli over other cell types (67). That is, elevating the cellular concentration of NADP+ would result in product inhibition of DHFR. In response, E. coli developed an ability for the M20 loop of DHFR to occlude the cofactor binding pocket in order to prevent hindered catalysis by NADP+ binding. In other words, the cell forced the enzyme to change its active-site loop conformation to avoid a stagnant (or stable) environment.

This dissertation studies the M20 loop conformations of E. coli DHFR by addressing the following questions. How do different ligands affect the structure and dynamics of the M20 loop? More specifically, does changing the ligand shift the enzyme preference towards another M20 loop conformer? Does the 'open' M20 loop conformer observed in crystal structures play a part in the closed-occluded conformational transition? Does removal of the crystal packing forces destabilize the open crystal conformation? How do conformational changes occur? More specifically, what is the pathway for conformational change? The atomistic detail and timescale of these events make experimental inquiry difficult and therefore lend themselves exceedingly well to simulation investigation.

Numerous theoretical investigations have already been performed on E. coli DHFR. Certain classical MD studies have probed how conformation and dynamics contribute to catalytic function (79-81). One such study demonstrates that certain correlated motions appearing in the Michaelis complex disappear in the product complex (79). Another study reveals that the N5 atom of DHF (i.e. the protonation site) is solvent-exposed, whereas static X-ray images show it to be concealed (80). Another study shows how mutations distal to the active site affect the M20 loop conformation and the hydrogen-bonding network of the enzyme, both key factors in modulating the enzyme kinetics (81). One high-temperature classical MD study identified regions possibly serving as nucleation sites during folding (82). Another compares the unfolding of wild type with that of the G121V mutant and shows the two unfolding paths to be similar, but that the latter involves a lesser number of contacts connecting the M20 and FG loops (83). Several QM/MM approaches have looked into the hydride transfer step (84-94). Among these, Thorpe and Brooks have calculated the activation barriers along the reaction coordinate and have determined that these barriers fluctuate in time and depend on the M20 loop conformation (88,91,93). A classical look into the hydride-transfer step (via a combination of MD simulations on the Michaelis and product complexes along with free energy perturbation calculations) established that the pK_a on the N5 of the substrate DHF is modulated by the M20 loop conformation (95). A recent classical MD study found (using umbrella sampling methods) the minimum-energy transition pathway between the closed and occluded M20 loop conformers of the Michaelis complex (96).

Interestingly, during transition, the enzyme samples a relatively stable transition structure resembling the open conformer seen in X-ray studies (67).

A proper molecular modeling study on DHFR and its M20 loop aims to resolve the issues discussed in the preceding paragraphs. The conformational transitions will be characterized by three studies. First, the relative stabilities of the initial and final states of a transition are determined by running molecular dynamics simulations followed by free energy analysis on single enzyme complexes in multiple starting conformations. The following four complexes are studied: DHFR, DHFR/folate, DHFR/NADPH and DHFR/dihydrofolate. Apo and folate-bound complexes are chosen since NMR has characterized the distribution of M20 loop conformers for these structures. The dihydrofolate-bound form was chosen to study how a small change in the ligand affects loop behavior. The NADPH-bound form is chosen since it is an actual catalytic intermediate. Second, the inherent motional tendencies of these initial and final states are examined for any predisposition for conformational change. Normal mode analysis techniques are perfect for quantifying these motions. To understand how ligands affect this disposition, both the Apo enzyme and ligand bound forms are studied. Third, conformational transition pathways are mapped using targeted molecular dynamics simulations. Like the umbrella sampling method used in a previous study that generated the closed-occluded transition path of the Michaelis complex (96), targeted MD also generates a low energy transition path. But, since targeted MD only requires one simulation to generate the path, it is much quicker, and therefore much more desirable when many different paths need to be generated (as is required here). The pathways generated are inspected to see if the open (or shifted open) conformer is sampled during the transition. Again, both ligand-bound form and the free enzyme are studied to assess ligand effects. Similar protocols are used to impose THF dissociation and determine whether that forces sampling of the open (or shifted open) conformer. Together these results should characterize both the general behavior of the individual stable M20 loop conformers and the changes between them.

METHODS

Two sets of MD simulations are performed on DHFR, one under equilibrium conditions and one with external perturbation forces applied. The CHARMM software (97-99) with the CHARMM22 protein (34-35), substrate (100) and cofactor (101) parameters is used. In the first set, four complexes are simulated: E|NADPH, E|folate, E|DHF and E (Apo). Except for E|NADPH, each is run three times, corresponding to its M20 loop starting in the open, closed or occluded conformation (Table 1). The methods are the same in all eleven simulations. First, internal coordinate building procedures add the missing hydrogens to necessary positions. After construction, the complex is minimized by steepest descent for 500 steps, then by conjugant gradient for 1500 steps (or until achieving an energy tolerance of 0.001 kcal/mol), and lastly, by Adopted-Basis Newton-Rhapson method for 2000 steps. A diminishing harmonic restraining force is used to hold the backbone atoms in place during the steepest descent and conjugant gradient stages. van der Waals switching and electrostatic shifting functions handle nonbonded potential truncation from 8-Å to 10-Å. Next the complex is solvated by

Table 1. Equilibrium MD simulations of DHFR

Complex	M20 Loop*	PDB†	Time (ns)‡
NADPH	open	1RA1	10
	closed	1RX1	10
Folate	open	1RD7	10
	closed	1RX2	9
	occluded	1RX7	20
DHF	open	1RD7	7
	closed	1RX2	7
	occluded	1RX7	7
Apo	open	1RD7	7
	closed	1RX1	20
	occluded	1RX7	7

* Starting M20 loop conformation

† RCSB PDB code of structure used for building initial simulation coordinates.

‡ Total simulation time

placing it in the middle of a preconstructed truncated octahedral box containing 6861 TIP3P water molecules and then removing waters overlapping the protein. Prior to dynamics, the surrounding water is minimized for 200 steps of steepest descent while keeping the protein and ligand atoms fixed. This solvated structure serves as the initial coordinates for the simulation. Dynamics takes place in the microcanonical (NVE) ensemble under periodic boundary conditions, using a 2-fs time step. The SHAKE algorithm is used to constrain the lengths of all bonds involving hydrogen. A switching function is used to smoothly shut off the van der Waals potential from 8-Å to 11-Å. The particle mesh Ewald method handles electrostatic potential across the periodic boundaries. Production dynamics is preceded by 100 ps of equilibration, during which the protein is allowed to relax while the system is heated. During the initial 20 picoseconds of the equilibration, a harmonic restraining force holds the backbone in place while heating occurs. Starting at 60K the temperature is increased by 30K every 200 steps by reassigning the velocities until the desired temperature (298 K) is reached. Production dynamics extends for enough time, such that the subsequent calculated free energy is stabilized (further described in the next section). Coordinates are saved every 100 timesteps (or every 2 picoseconds).

The free energy at every 2000th timestep is calculated using the MMPBSA method (51-55). The molecular mechanics component uses the CHARMM22 force field (34-35). Infinite cutoffs (with a 1.0 constant dielectric) are employed for the nonbonded interactions. Thus, all nonbonded interactions are accounted for. The PB component also does not use a cutoff. Solvation energy calculations are performed using the PBEQ module of CHARMM, which uses a numerical grid-based approach to solve the Poisson-Boltzmann equation around the surface of the protein. A 0.4-Å sized grid unit cell is used along with a uniform number of grid points in the x,y and z directions to cover the protein and/or ligand surfaces. Dielectric constants of 80 and 1 are used for solvent and protein, respectively. The molecular (contact+reentrant) surface is created with a 1.4-Å radius and then used as the dielectric boundary between protein/ligand and solvent. The nonpolar surface area energy is estimated from the solvent-accessible-surface-area using a 1.4 Å solvent probe. The constant of proportionality, γ, and the y-intercept, β, (Equation 31) are set to 0.00542 kcal/(mol*Å^2) and 0.92 kcal/mol, respectively (55). The entropy component is estimated using quasiharmonic analysis of the MD trajectory. The calculations are performed on 2-ns portions of the trajectories, where the MMPBSA energy has reached a well-defined average. Further details on how the trajectory window is selected are specified in the Results and Discussion section of this chapter. Prior to calculating the covariance matrix, trajectory frames are reoriented by minimizing their mass-weighted root-mean-square difference (RMSD) with respect to the average structure along the selected 2-ns window. Entropy calculations are performed two ways; the first uses only heavy protein atoms (justified since the lengths of all bonds involving hydrogen are kept fixed during the simulation), and the second includes all protein atoms.

Two types of external perturbations are used for the second set of DHFR simulations: forced conformational changes and forced ligand dissociation. Table 2 outlines each complex and the corresponding forced action(s). Each initial structure is equilibrated using the same methodology as described above. Conformational changes utilized the Restricted-Perturbation TMD method (43). Target coordinates were taken from 1RX2 or 1RX7, depending on whether the target conformer is closed or occluded, respectively. Forces were applied to the M20 loop (residues 14 to 24) backbone atoms. Forced conformational change simulations are run for at least 400 ps. For the first 300 ps in the forced conformational change simulations, forward and backward perturbations are restricted to maximum RMSD moves of 0.001 Å and 0.0008 Å, respectively, per timestep. To ensure the transition occurs, the maximum RMSD moves are increased to 0.002 Å and 0.001 Å, respectively, in the final 100 ps. If required, simulations are extended another 100 ps (500 ps total) with forward and backward perturbations restricted to maximum RMSD moves of 0.004 Å and 0.002 Å, respectively.

The ligands are dissociated by applying an external force on a ligand atom to push it out of the active site. These runs last for 100 picoseconds, during which the pulling forces linearly increases from 0 to 1500 piconewtons. Two runs are performed using different pulling vectors for each dissociation simulation listed in Table 2. Pulling vectors are defined as a ligand atom coordinate minus a selected protein interior atom coordinate. The protein atoms are selected so that the defined pulling vectors point along low steric routes towards the solution. In the first cofactor dissociation simulation, nitrogen NN7 (at the reactive end of the cofactor molecule) is pulled away from the alpha carbon of

Table 2. External perturbation MD simulations of DHFR

Ligand	M20 Loop[*]	PDB[†]	Perturbation Type
DHFR	closed	1RX2	occluded
	occluded	1RX7	closed
DHFR | Folate	closed	1RX2	occluded
	occluded	1RX7	closed
DHFR | DHF	closed	1RX2	occluded
	occluded	1RX7	closed
DHFR | NADPH	closed	1RX1	occluded
	closed	1RX1	NADPH out
DHFR | DHF | NADPH	closed	1RX2	occluded
DHFR | THF | NADP+	closed	1RX2	occluded
	closed	1RX2	NADP+ out
	closed	1RX2	THF out
	occluded	1RX7,1RA1	THF out
DHFR | THF	closed	1RX2	occluded
	closed	1RX2	THF out
	occluded	1RX7	closed
	occluded	1RX7	THF out
DHFR | THF | NADPH	closed	1RX2	occluded
	closed	1RX2	THF out
	closed	1RX2	NADPH out
	occluded[‡]	1RX7,1RA1	closed
	occluded[‡]	1RX7,1RA1	THF out

[*] Starting M20 loop conformation

[†] RCSB PDB code of structure used for building initial simulation coordinates. Structures requiring 2 PDB codes use first for enzyme coordinates and second for cofactor coordinates.

[‡] The coordinates for bound cofactors in occluded conformers are obtained from 1RA1. In this crystal structure, the nicotinamide ring moeity is invisible. Since the CHARMM internal coordinate building procedures place the moeity outside the active site, the presence of the cofactor does not interfere with the occluded M20 loop.

ALA7 (Figure 10). In the second simulation, carbon NC7 (adjacent to the cofactor's nicotinamide ring) is pulled away from the carbonyl carbon of ALA7. In the first substrate dissociation simulation, carbon C (Figure 10) is pulled away from the alpha carbon of ALA7. And in the second simulation, carbon CB is pulled away from the carbonyl carbon of LEU112.

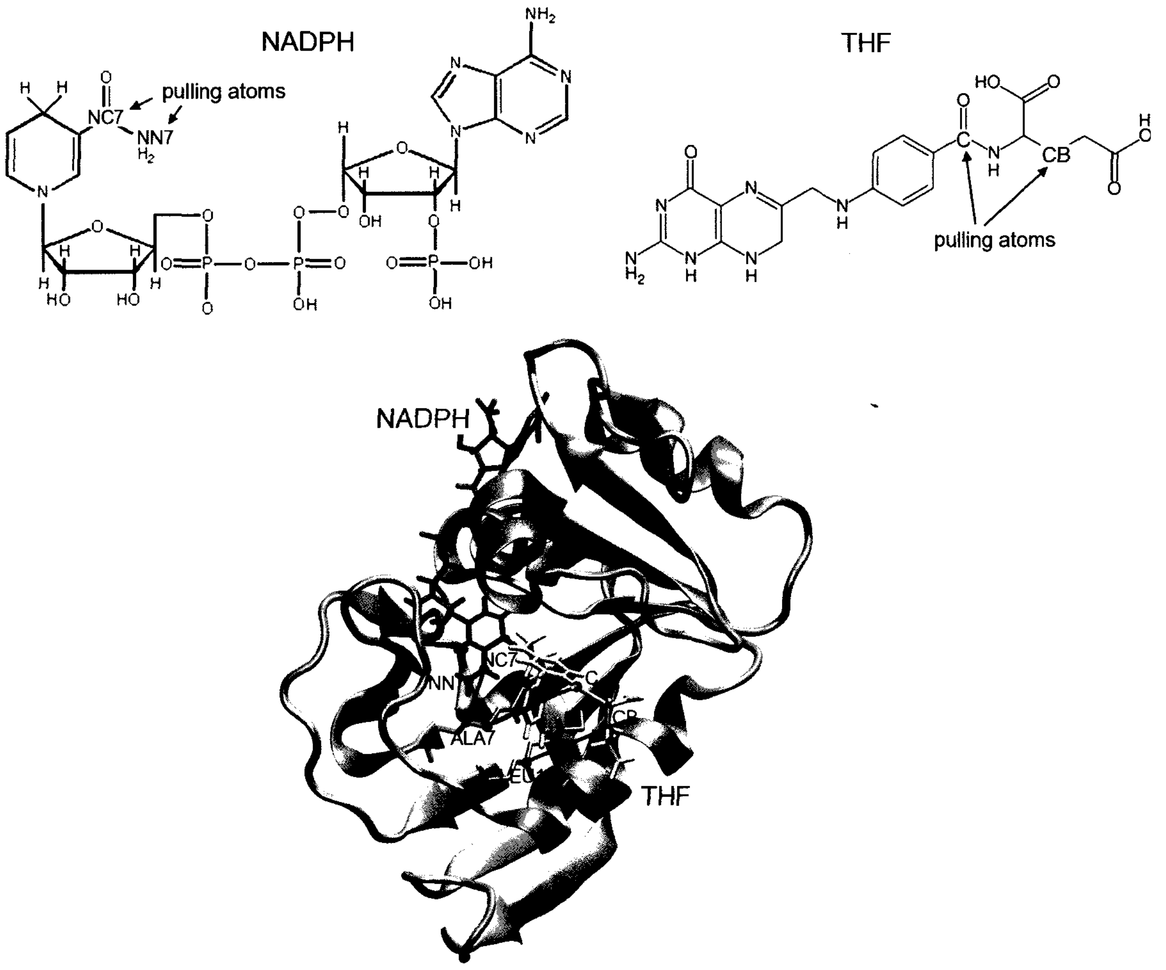


Figure 10. NADPH and DHF pulling atoms. Their positions relative to the protein atoms they are being pulled away from are displayed in the lower image.

An examination of other external perturbation simulations in the literature shows a lack of convention as to how fast these structural changes should be forced to occur, with simulation times ranging anywhere from 90 ps on up (102-105). The forced conformational change simulation times were set, based upon trial and error, by starting with a very small perturbation (0.0001 Å) in 1-2 ns runs, and slowly increasing the perturbation until the transition occurred. At least a 0.001 Å forward perturbation was needed to get any transition to occur. Certain structures required more than the 0.001 Å forward perturbation, so a final 100 ps was run with the perturbation doubled to 0.002 Å. The ligand dissociations were tested similarly (also using implicit solvent) by slowly increasing the force constant until dissociation occurred.

Normal modes are calculated for each starting structure used in the forced conformational change simulations listed in Table 2. The 13 structures are as follows: E|closed, E|occluded, E|folate|closed, E|folate|occluded, E|DHF|closed, E|DHF|occluded, E|NADPH|closed, E|DHF|NADPH|closed, E|THF|NADP+|closed, E|THF|closed, E|THF|occluded, E|THF|NADPH|closed and E|THF|NADPH|occluded. Since these are just the starting structures, they have not been altered by MD simulation. The first 7 structures, however, were simulated under equilibrium conditions. So normal mode calculations will also be performed on the minimized-average structures obtained from those equilibrium simulations. Normal modes are also calculated for the minimized-average structure obtained from the E|NADPH|open equilibrium simulation, since (as will be shown in the Results and Discussion section of this chapter) this structure shifts to a new conformation. Prior to obtaining the normal modes, each structure is first minimized by steepest descent for 2000 steps or until an energy change tolerance of 0.00001 kcal/mol is reached, followed by conjugant gradient for 20,000 steps or until a tolerance of 0.0000001 kcal/mol is reached. During minimization, van der Waals switching and electrostatic shifting functions handle nonbonded potential truncation from 8-Å to 10-Å. Studies show that only the lowest frequency modes are needed to describe conformational changes (59-61). Studies vary on the number of low-frequency modes used, with some using only one low-frequency normal mode (60,61), another using the lowest 20 (62) and another using the lowest 100 (106). The 20 lowest frequency normal modes are used in the dot product calculations.

After obtaining the normal modes, the dot products between the normal modes (discarding the first 6 translational and rotational eigenvectors) and the conformational change vectors are calculated. The conformational change vectors are defined as the coordinates of a second structure (**x2**) minus the coordinates of the structure for which the normal modes are calculated (**x1**). The second structure is superimposed onto the normal mode structure prior to calculating this vector. Only the M20 loop (residues 14-24) alpha-carbons are considered, so the dot product with the kth normal mode is calculated using

$$\mathrm{Dot}_k = \left|\mathbf{n}_k \cdot (\mathbf{x2} - \mathbf{x1})\right|, \qquad [80]$$

where $\mathbf{n}_k$ are the coordinates of the alpha-carbons in the M20 loop of the kth normal mode, **x2** are the coordinates of the alpha-carbons in the M20 loop of the second structure and **x1** are the coordinates of the alpha-carbons in the M20 loop of the normal mode structure. The caret indicates that the difference vector is normalized. The second structure coordinates used are the open (1RD7), closed (1RX2) and occluded (1RX7) X-ray conformers, and the new conformer (using the NADPH open MD minimized-average simulation structure). The dot products for each of the 5 lowest frequency modes (i.e. modes 7, 8, 9, 10 and 11) are individually examined to see which of the dot products is greatest. To assess the effects of multiple normal modes at once, the sum of the dot products for the 5 (modes 7 to 11), 10 (modes 7 to 16) and 20 (modes 7 to 26) lowest are calculated. For example, the dot product sum of modes 7 to 11 is calculated using

$$\mathrm{DotSum}_k = \sum_{k=7}^{11} \left| \mathbf{n}_k \cdot (\mathbf{x2} - \mathbf{x1}) \right| . \qquad [81]$$

RESULTS AND DISCUSSION

The data and analyses from the equilibrium MD simulations on the binary/Apo complexes are discussed first. Conformational stability is identified using extensive structural and energetic means. Then, the external perturbation MD simulations and their corresponding normal mode calculations are analyzed to see if the closed-occluded transition involves a sampling of the third (open) M20 loop conformer.

A total of 11 standard equilibrium molecular dynamics simulations (Table 3), covering 4 complexes, were performed. They can be divided into 4 classes according to ligand identity: NADPH, folate, DHF, and *apo*, or into 3 classes according to starting M20 loop conformation: open, closed and occluded. The general system coherence is assessed by examining the time series of the following quantities from the simulation trajectories: CHARMM total energies, backbone root mean square deviations (RMSD) with respect to the X-ray starting structure, and radius of gyration (RGY) (Figures 11-13, respectively). Both the CHARMM total system energies (Figure 11) and the protein radii of gyration (Figure 12) for all simulations remain constant, meaning, respectively, that throughout the simulation the systems remain energetically stable (or constant) in the

Table 3. Analysis of equilibrium MD simulations of DHFR.

Complex	Starting M20 loop	Time (ns)*	Stable (ns)†	M20 (stable)‡
apo	open	7	6-7	new
	closed	20	12.6-14.5	slightly shifted
	occluded	7	6-7	very conserved
NADPH	open	10	6-7	new
	closed	10	6-7	conserved
Folate	open	10	9-10	new
	closed	9	8-9	slightly shifted
	occluded	20	19-20	more shifted
DHF	open	7	6-7	new
	closed	7	6-7	conserved
	occluded	7	4-5	more shifted

* Total simulation time

† Stable 2-ns window of simulation

‡ Conformation of M20 loop backbone within the 2-ns window relative to where it started at. Shifts are either slight (a few residues move 1-2 Å) or more (several residues move up to 4 Å, but most sidechains still point the same general direction). New conformations have no resemblence to any of the three X-ray conformers).

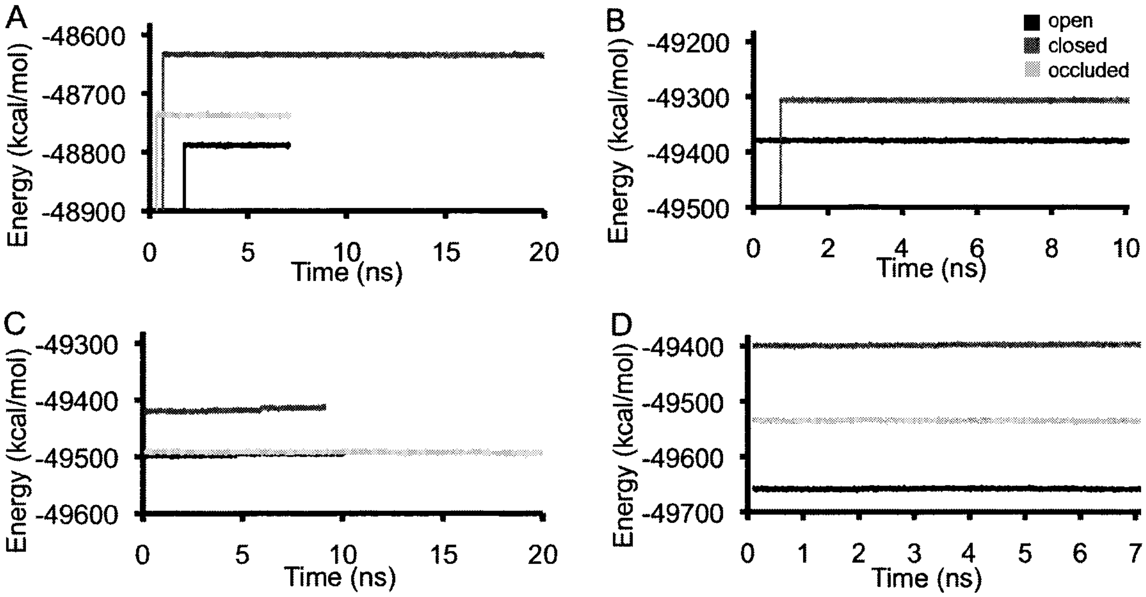


Figure 11. DHFR CHARMM total system energy vs. time. Shown are DHFR (A), DHFR|NADPH (B), DHFR|folate (C) and DHFR|DHF (D).

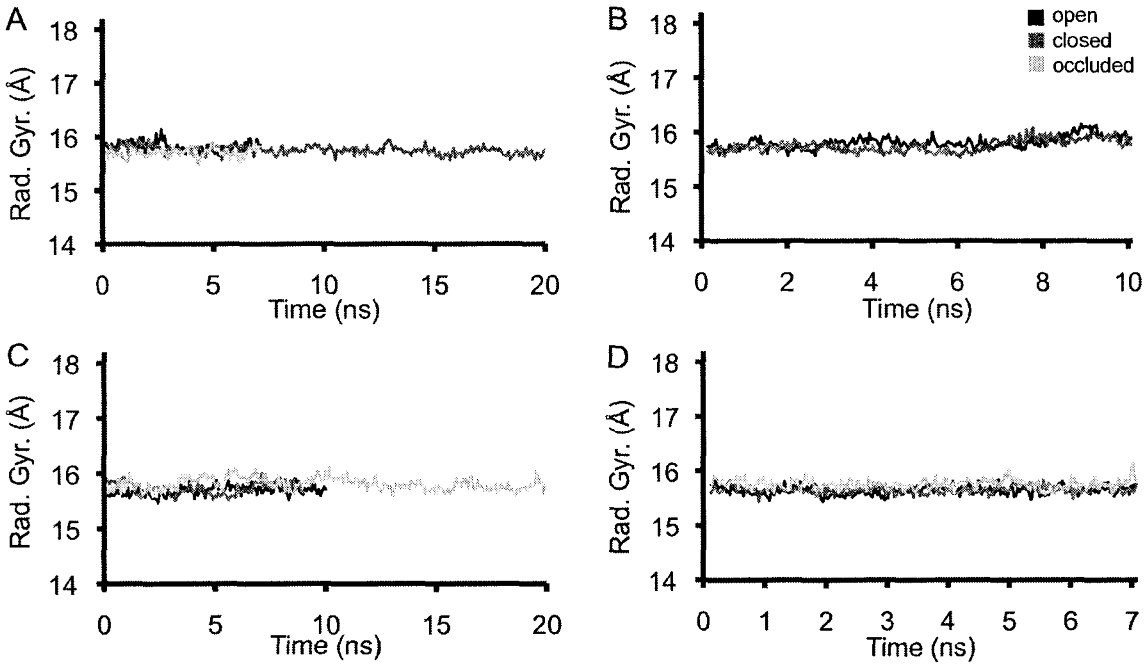


Figure 12. DHFR protein radii of gyration vs. time. Shown are DHFR (A), DHFR|NADPH (B), DHFR|folate (C) and DHFR|DHF (D).

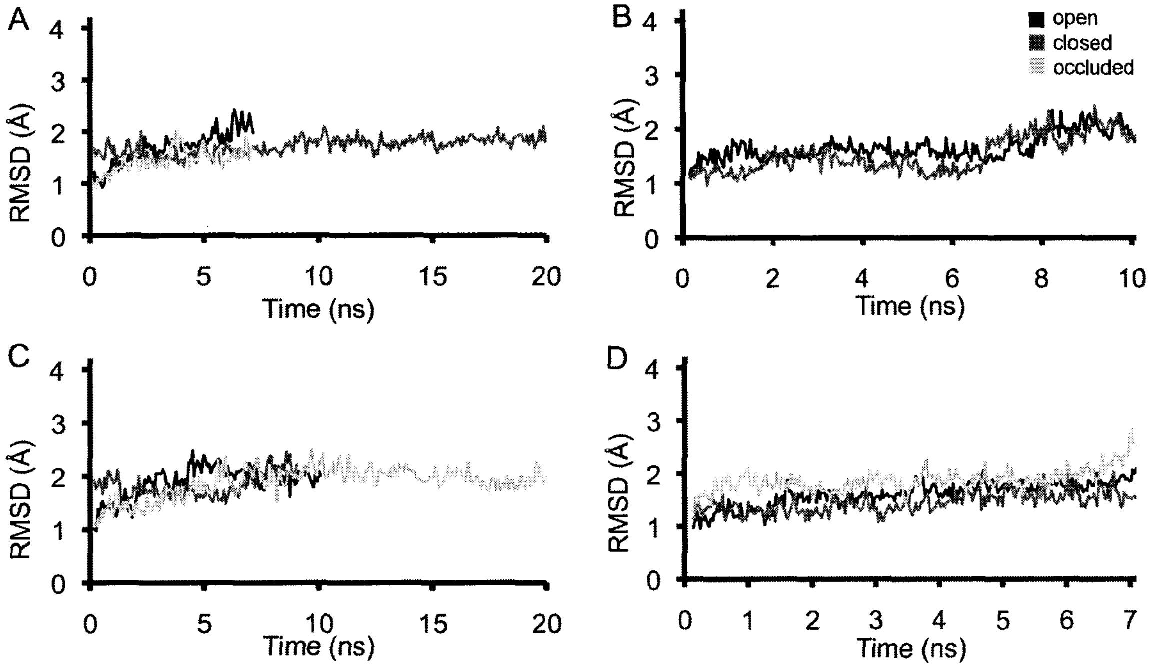


Figure 13. DHFR protein backbone RMSD vs. time. Shown are DHFR (A), DHFR|NADPH (B), DHFR|folate (C) and DHFR|DHF (D).

throughout the simulation the systems remain energetically stable (or constant) in the CHARMM force field and that the proteins do not abnormally shrink or grow. This result is not unexpected since previous DHFR simulations in the same force field also exhibited stability (44-50,69). The jumps in the total system energy plots (Figure 11) are normal, and occur when the atomic velocities are reassigned to maintain the temperature. Protein backbone RMSD values remain less than 2.5 Å and are relatively flat (Figure 13), indicating, respectively, that the protein conformation stays close to the native starting structure and that the structure fluctuates around an average.

Equilibrium Molecular Dynamics: Structural Analysis

With the overall systems regarded as stable, the specific conformational motion of the M20 loop represents how it probably behaves *in vivo*. Stable two-nanosecond simulation windows for all systems were determined (see Table 3) and used for various subsequent analyses, both structural and energetic. The determination of these windows is described later in the free energy analysis subsection; however, they represent portions of the simulation, during which the protein is both structurally and energetically equilibrated. Structural analysis of the protein during this interval proceeds by examining the following: alpha carbon root-mean square fluctuations (RMSF), visual inspection of the M20 loop region in the minimized-average structures, M20 loop backbone RMSD, existence of hydrogen bonds that characterize the three M20 loop conformers, residue centroid distance matrices of the M20 loop versus the entire protein and corresponding difference distance matrices of the M20 loop versus the entire protein. Table 3 describes the M20 loop conformation in the stable window. Most interestingly, all structures starting in the open conformer move to a new M20 loop conformer and remain there. It will be shown later, through visual inspection that these new conformers are all identical. Hence, the simulated open M20 loop conformer will be called 'new'.

Alpha carbon RMSF values were calculated during the stable 2-ns windows to show the ligand effect on protein dynamics. The plots show most residues fluctuating within 1 Å of the average position. Values greater than this, as indicated by the peaks on the RMSF plots (Figure 14) are the more flexible exterior loop regions (Figure 15). The M20 loop region (residues 14-24) of the new conformers (indicated by the blue lines in

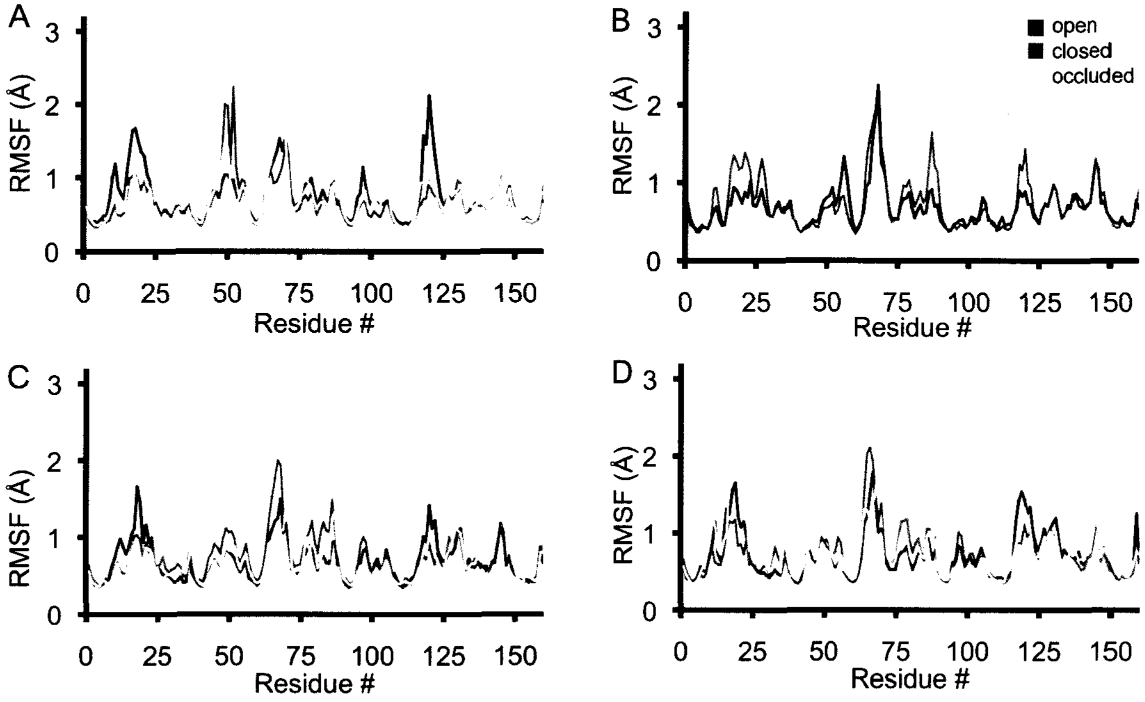


Figure 14. DHFR alpha carbon RMSF during the stable 2-ns window. Shown are DHFR (A), DHFR|NADPH (B), DHFR|folate (C) and DHFR|DHF (D).

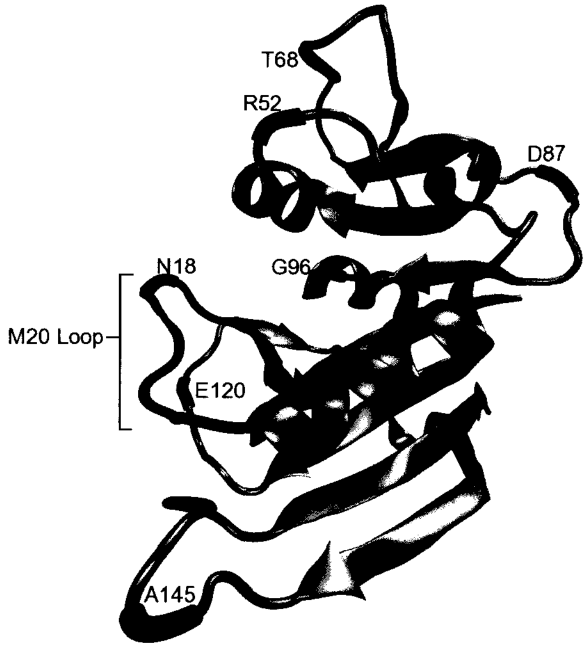


Figure 15. Fluctuating regions of DHFR. Largest fluctuations center around residues N18, R52, T68, D87, G96, E120 and A145.

Figure 14), fluctuate around 1.75 Å (Figures 14A, 14C and 14D), except when NADPH is bound (Figure 14B), where it fluctuates about 1 Å. This rigidity in the NADPH complex probably results from contacts between the loop and the proximal cofactor. The closed loop (red lines) behaves exactly opposite; it fluctuates 1 Å everywhere except in the NADPH complex, where it approaches ~1.5 Å (Figure 14B). The occluded loop (green lines) remains stable (~1 Å) in all but the DHF-bound complex (Figure 14D) where it has large 2-Å fluctuations. Visual inspections (Figure 16) of the minimized-average structures of the eleven simulations show a corresponding shift in the M20 loop of the occluded DHF simulation (green ribbon in Figure 16C).

In each of the 4 DHFR simulated complexes, at least one conformer remained mostly conserved (perhaps with slight shifts) and at least one changed to a new conformer (Table 3). Figure 16 shows the overlapped M20 loop regions of both simulation (colored) and X-ray (black) structures after superimposing the entire protein backbones of all structures. As mentioned previously, Figure 16A shows that all simulations starting in the open conformer change to a common new conformation. By comparing the middle orientations of Figure 16 to that of Figure 6 (which is similar), it is seen that the N-terminus end of the M20 loop (residues 14-19) of the new conformer folds in towards the protein. The slight shift of the DHFR|NADPH new conformer (red loop in Figure 16A) away from the other new conformers is due to the presence of the bound cofactor, since it sits between the loop and the enzyme. This conformer is actually similar to the closed conformer, though none of the simulations starting in the closed conformer approached this conformer. In the closed simulations (Figure 16B), the C-terminus portions of the M20 loop (residues 20-24) remain very well conserved compared to the X-ray conformer. In the DHF-bound complex (green), the N-terminus end shifts slightly away from the X-ray conformer. The M20 loop of the NADPH-bound complex (red) stays very conserved, which probably results from stabilizing interactions between it and the bound cofactor. The N-terminus ends of the folate (blue) and *apo* (yellow) complexes both fold in towards the protein (Figure 16B), probably to fill the void where the cofactor would bind. Perhaps if the DHF simulation were extended, or if different random starting velocities were used, it would have also shifted. In the occluded simulations (Figure 16C), only DHFR|*apo* (yellow) remains well conserved compared to

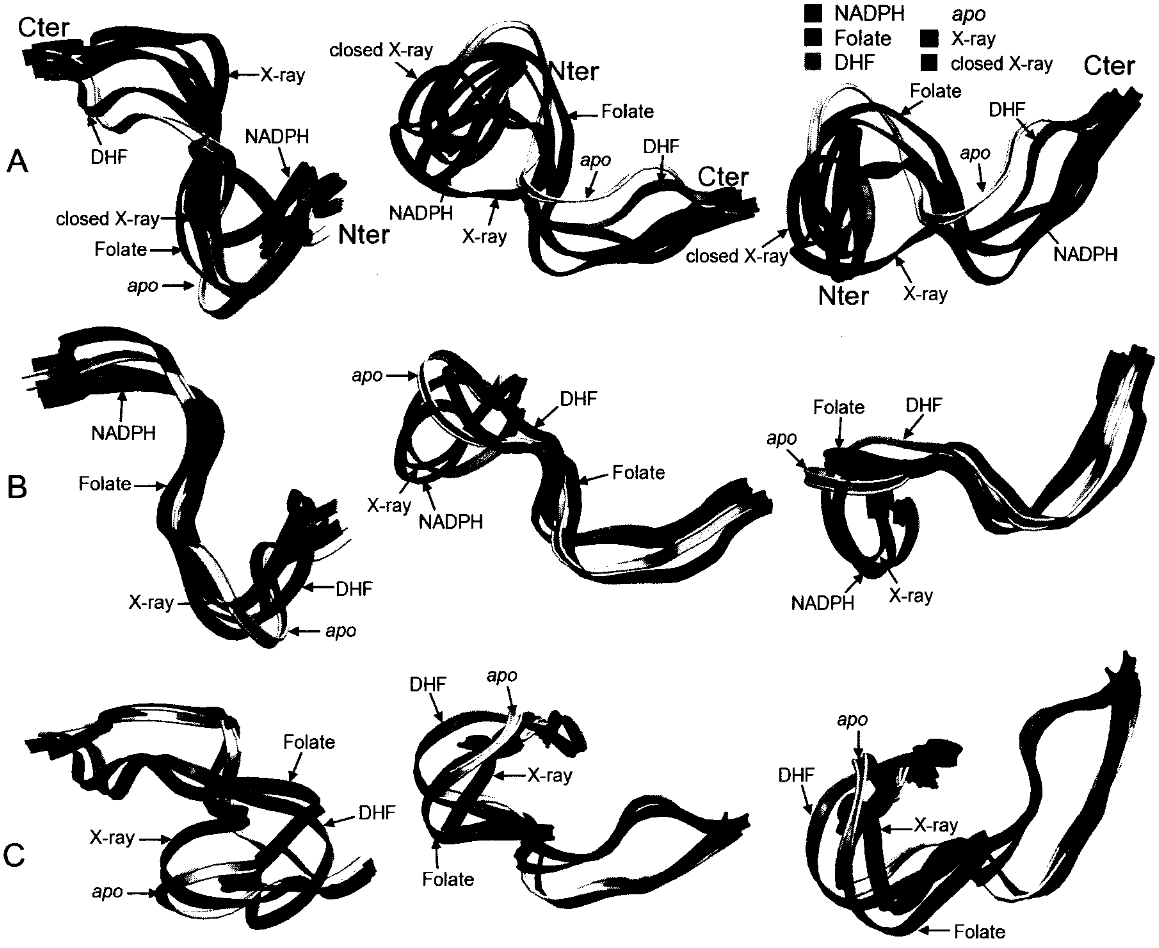


Figure 16. M20 loop portions of the superimposed simulation structures. Three different views are shown. The images are separated according to starting M20 loop conformation: open (A), closed (B) and occluded (C). Each figure (A, B, and C) shows the NADPH, folate, DHF and *apo* simulation structures, as well as the corresponding X-ray structure. C has no NADPH structure. Figure A also includes the closed X-ray structure (1RX1).

X-ray (black). The other two (DHF-green and folate-blue) both shift a good amount, probably a result of interactions with the pteridine ring. However, they do retain a loose resemblance to the occluded X-ray conformer.

The visual inspections above (Figure 16) show the relative mobility of the M20 loop's N-terminus portion over its C-terminus portion. This difference is also observed in the M20 loop backbone RMSD plots. After aligning the trajectories with respect to their X-ray starting points using the entire protein backbone, the RMSD of the M20 loop backbone in both its N-terminus (residues 14-19) and C-terminus (residues (20-24) portions were calculated and plotted (Figure 17). The N-terminus (thin dashed lines) moves farther away from the X-ray than the C-terminus (thick solid lines) in all cases, except the NADPH complex with a closed loop (red lines in Figure 17B), where both ends remain extremely well conserved. The formation of the new M20 loop conformer in the open simulations results in large RMSDs (Figure 17 - dotted blue lines).

The M20 loop conformations of the 11 simulations were further and more quantitatively characterized by three different methods: M20 loop backbone RMSDs, the presence (or absence) of X-ray conformation markers (17) and residue centroid difference distance matrices. In the first method, the simulation conformers are compared with the three X-ray conformers (open, closed and occluded), the new conformer (using the coordinates from the folate-bound minimized average structure from the open simulation), the folate-bound closed minimized average structure and the DHF-bound occluded minimized average structure (Table 4). The last two are included, since their M20 loops involved slight shifts away from X-ray. Before calculating the RMSDs for a given simulation conformer, the comparison structures are aligned using their entire protein backbones in a mass-weighted fitting. The RMSD calculations use only the M20 loop backbone atoms. Obvious matches occur when the RMSD is much less (at least by 1.0 Å) than all others involved in the comparison and are highlighted bold.

The *apo*-open and DHF-open are, by far, closer to the new conformer than any of the three X-ray structures. In addition, the folate-open does not match any others, confirming the presence of a new conformation. The NADPH-open structure has several comparisons to the open X-ray (2.8 Å), closed X-ray (2.2 Å) and new conformer (2.9 Å), which is in agreement with the observed shift away from the other open conformers. The fact that this new conformer is not seen in X-ray could mean one of two things, the conformer is too unstable to be crystallized or it is an artifact of the force field. The former circumstance (as discussed in the Background section of this chapter) is possible since the M20 loop of all open crystal structures contains interactions with one of the symmetry related crystal proteins (Figure 8). Removing these interactions (as exists *in vivo)* would likely cause a corresponding shift in the M20 loop conformation. The new conformer from the simulation may be the relaxed form of the open X-ray conformer.

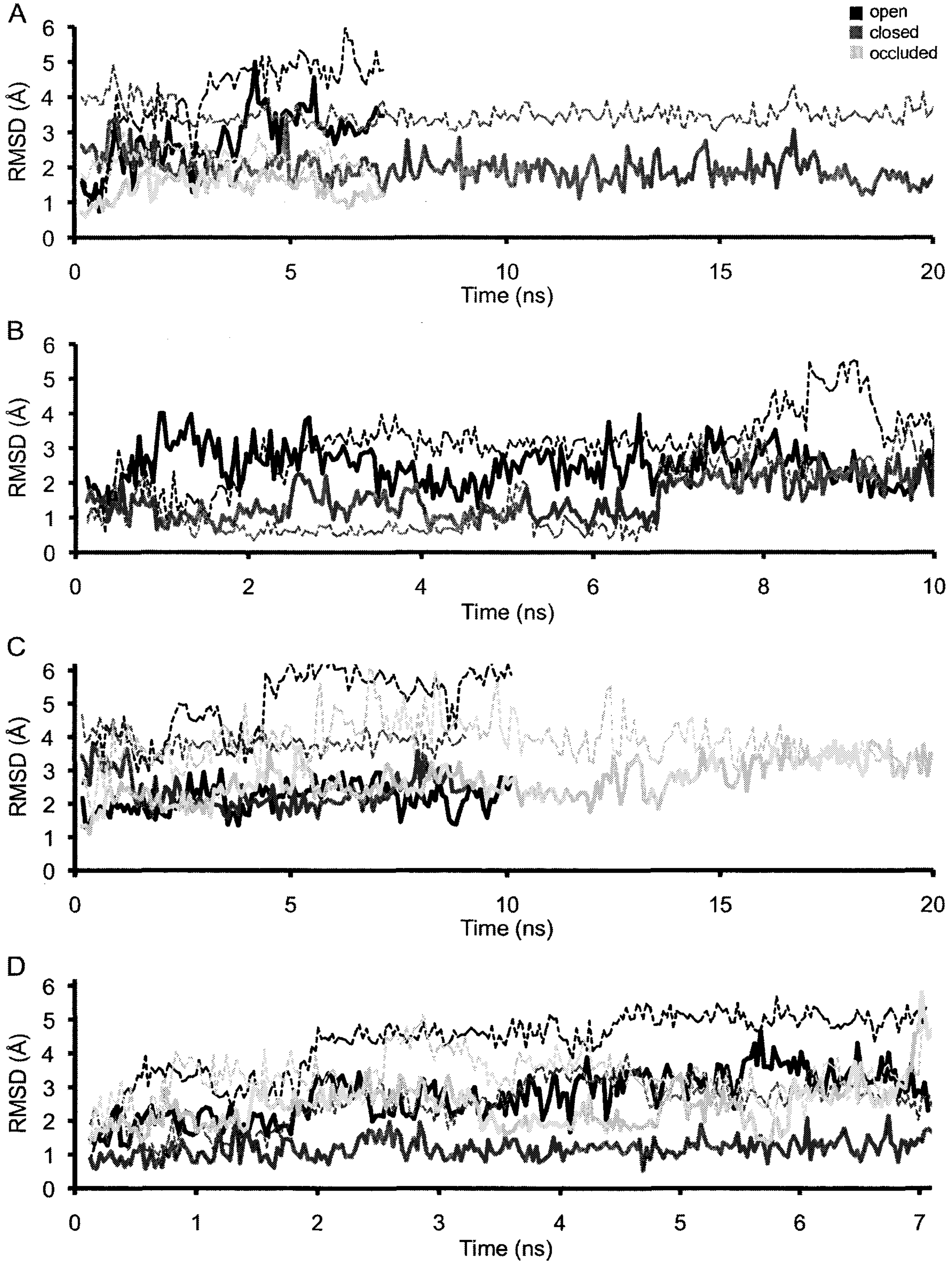


Figure 17. M20 loop N-terminus and C-terminus backbone RMSD vs. time. Shown are DHFR (A), DHFR|NADPH (B), DHFR|folate (C) and DHFR|DHF (D). N-terminus lines are drawn as thin-dotted lines and C-terminus as thick-solid.

Table 4. M20 loop backbone RMSD (Å)

Complex	Starting M20 Loop	Xray Open*	Xray Closed*	Xray Occluded*	New†	Folate closed‡	DHF occluded§
Apo	open	3.6	2.4	4.4	**1.8**	2.9	3.5
	closed	3.1	2.7	3.5	3.2	**1.1**	3.3
	occluded	3.9	3.7	**1.6**	4.7	2.9	3.4
NADPH	open	2.8	2.2	4.1	2.9	3.6	3.5
	closed	2.6	**0.7**	4.0	3.2	3.4	3.8
Folate	open	4.2	3.3	4.8	0.0	3.4	4.2
	closed	3.4	3.2	3.0	3.4	0.0	3.6
	occluded	4.0	4.0	3.3	4.6	3.8	**2.0**
DHF	open	4.4	3.2	5.2	**1.9**	4.1	3.7
	closed	3.6	2.1	4.6	1.8	3.3	4.0
	occluded	4.2	3.8	3.9	4.2	3.6	0.0

* open=1RA1, closed=1RX1, occluded=1RX7

† Folate-open minimized average simulation structure

‡ Folate-closed minimized average simulation structure

§ DHF-occluded minimized average simulation structure

The latter circumstance was tested by simulating the new conformer using the AMBER force field and seeing if the M20 loop remained stable. The new conformer (using the minimized-average *apo*-open structure – from the CHARMM simulation) was used as the starting point for a 2-ns AMBER run. The minimized-average structure of the AMBER run was calculated and compared with the minimized-average structure from the CHARMM run. Figure 18A shows that the M20 loop of both structures is very similar. The M20 loop RMSD calculations (Table 5) show that the structure resembles the new conformation more than the other X-ray conformers. As a control, the minimized-average structure from the *apo*-occluded simulation was also run for 2 ns in the AMBER force field. The *apo*-occluded was chosen since it remained close to its X-ray starting point when simulated under CHARMM (Figure 16C and Table 4). Figure 18B shows it also remained stable under AMBER. The corresponding RMSD calculations (Table 5) verify this. The control indicates that DHFR is well behaved when simulated under the AMBER force field. These results provide compelling evidence that the new M20 loop conformer is valid, not just an artifact of the CHARMM force field.

Table 5. M20 loop backbone RMSD of AMBER simulation structures (Å)

Complex	Starting M20 Loop	Xray Open†	Xray Closed†	Xray Occluded†	new‡
apo	new	4.2	2.6	5.0	**1.5**
	occluded	5.0	4.3	**2.2**	4.7

* open=1RA1, closed=1RX1, occluded=1RX7

† open=1RA1, closed=1RX1, occluded=1RX7

‡ *apo*-open minimized average simulation structure

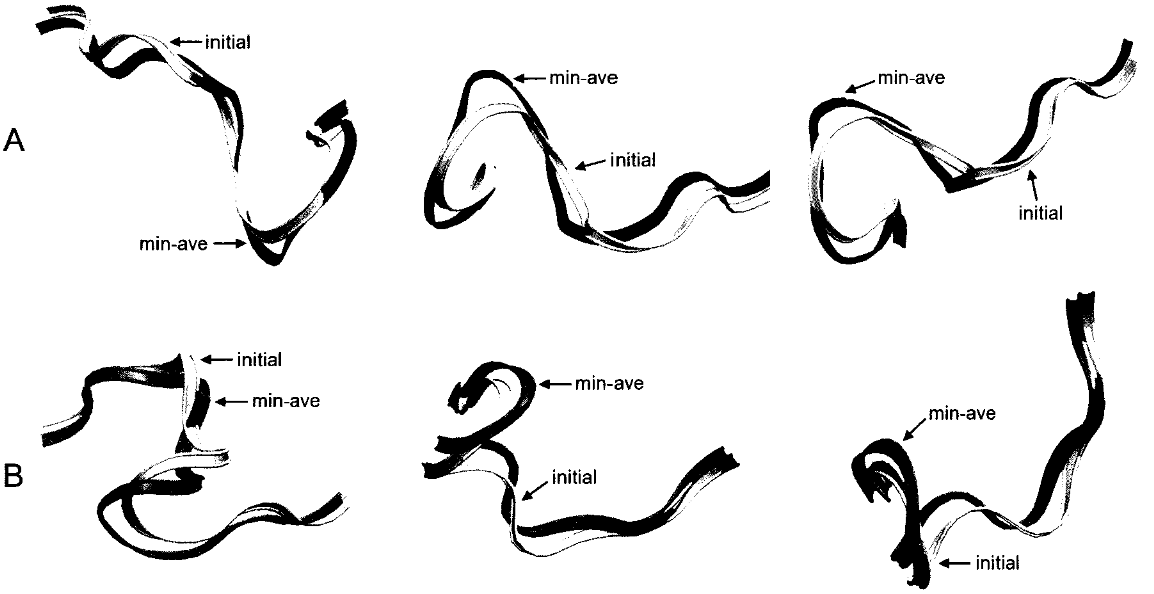


Figure 18. M20 loop portions of the superimposed AMBER simulation structures. Shown are the *apo*-new (A) and the *apo*-occluded (B) structures. The minimized-average structures and the initial starting structures are both drawn.

The folate and *apo* closed simulation structures both shifted away from their X-ray starting points (3.2 and 2.7 Å RMSDs, respectively). The small RMSD between them (1.1 Å) agrees with the visual drawing (Figure 16B) that they shifted to a similar conformation, where the N-terminus portion of the M20 loop folds in to occupy the region where the cofactor (missing) binds. The DHF closed simulation shifted slightly away from X-ray (2.1 Å RMSD) to a structure resembling the new conformer (1.8 Å RMSD). Visually, it appears close to X-ray, except for a distortion in the central loop residues (Figure 16B). The folate and DHF occluded simulations both shifted away from their starting points (3.3 and 3.9 Å RMSDs, respectively), but, the 2.0-Å RMSD between them shows the shifts were similar. The RMSD calculations show the three M20 loop conformers behaving very consistently. The open simulation structures all shift to a similarly-shaped new conformation. The closed is stable when either NADPH or DHF is bound, but shifts slightly (in a similar manner) to a more closed type conformer when folate is bound or in the *apo* form. Occluded is stable in the *apo* form, but slightly shifts when folate or DHF are bound.

The second characterization method is to look for the hydrogen bond markers. The closed X-ray conformer markers are the type III' hairpin turn and the two M20-FG interloop interactions, the occluded markers are a short 3_{10}-helix within the M20 loop and the two M20-GH interloop interactions and the open markers are aspects of both. Table 6 shows the corresponding distances in the X-ray and simulation structures. None of the structures (except NADPH-closed) contain all of the supposed markers. The *apo* and folate closed simulations lose most of the associated markers. Their visual renderings display a shift towards the protein in the N-terminus end of the M20 loop. Then once the hydrogen bonds were broken, the N-terminus portion of the M20 loop shifted in towards the protein. The open conformers each retain one M20-FG interloop interaction, but lose the M20-GH interloop interactions. In addition, one of the interactions in the intraloop type III' hairpin turn is formed. The visual renderings (Figure 16A) support this conclusion, where the simulation structures have formed a U-shaped bend, similar to that observed in closed, but in a different location than closed with respect to the rest of the protein. The *apo* occluded simulation remains close to X-ray, maintaining the M20-GH interloop hydrogen bonds, Asn23(CO)-Ser148(NH) and AsN23(NH)-Ser145(Oδ), at 3.1 Å and 3.5 Å, respectively, and the 3_{10}-helix hydrogen bond [Glu17(CO-Met20(NH)] at 2.7 Å. The 3_{10}-helix is lost in the occluded folate (5.2 Å) and DHF (4.5 Å) simulations. However, both possess larger spiral-type backbone folds. The DHF-bound structure actually shifts its 3_{10}-helix from Glu17(CO)-Met20(NH) over to Met16(CO)-Ala19(HN) with a 2.35 Å connection (Figure 19A), while the folate-bound structure shifts one more residue to form a Gly15(CO)-Met20(NH) π-helix connection (1.92 Å) (Figure 19B). Thus, losing the stabilizing 3_{10}-helix interaction causes a conformational shift in the 3_{10}-helical M20 loop to a larger backbone spiral (π-helical).

Table 6. Hydrogen bond markers in X-ray and simulation structures (Å)

Complex	M20	M20-FG[†]		Type III' hairpin[‡]		M20-GH[§]		3_{10}-helix[¶]
		A	B	A	B	A	B	
X-ray	open	2.1[‖]	4.4	6.4	6.8	2.0	1.9	7.2
	closed	**2.2**	**2.2**	**2.3**	**1.7**	3.4	3.8	6.7
	occluded	8.6	13.1	*2.3*	5.7	*2.1*	*2.0*	*2.2*
Apo	open	3.1	9.7	7.3	3.5	5.5	6.5	8.6
	closed	5.6	**2.4**	6.9	6.4	3.7	4.1	8.1
	occluded	5.8	10.4	*2.1*	4.4	*3.1*	*3.5*	*2.7*
NADPH	open	4.0	8.3	7.0	1.9	4.7	5.3	8.3
	closed	**2.3**	**3.1**	**2.2**	**1.9**	4.3	4.7	6.7
Folate	open	3.4	8.7	6.8	4.3	4.4	5.0	7.9
	closed	5.4	4.5	6.9	6.0	3.9	4.6	8.0
	occluded	8.8	4.8	5.4	7.7	8.1	11.4	5.2
DHF	open	2.3	8.3	6.9	2.4	5.7	7.6	7.8
	closed	**2.0**	8.7	6.5	3.3	4.4	4.8	7.6
	occluded	8.7	3.3	*2.3*	5.7	7.0	8.0	4.5

[*] X-ray coordinates are open (1RA1), closed (1RX1) and occluded (1RX7)

[†] Interloop H-bond closed markers: A=Gly15(CO)-Asp122(NH), B=Glu17(NH)-Asp122(Oε2)

[‡] Type III' Hairpin closed marker: A=Met16(CO)-Ala19(NH), B=Met16(NH)-Ala19(CO)

[§] Interloop H-bond occluded markers: A=Asn23(CO)-Ser148(NH), B=Asn23(NH)-Ser148(Od)

[¶] 3_{10}-helix occluded marker: Glu17(CO)-Met20(NH)

[‖] Open, closed and occluded X-ray markers are displayed blue (underlined), red (bold) and green (italic) The markers of simulation structures that are retained are indicated by the same scheme.

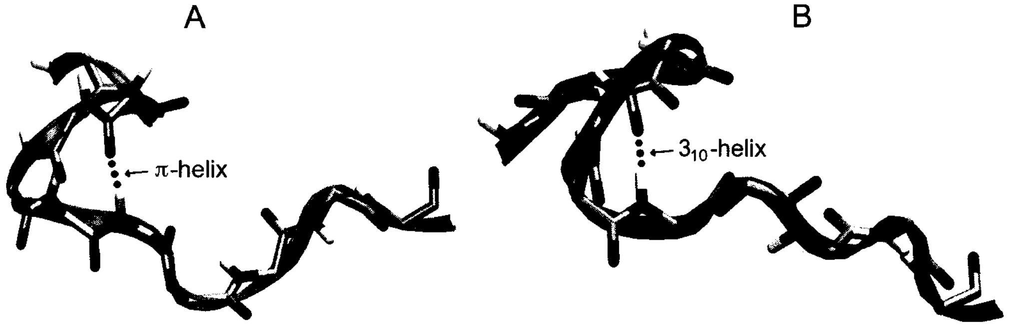


Figure 19. M20 loop helix shifts in folate and DHF occluded simulations. Shown are the minimized-average structures for DHF (A) and folate (B) simulations.

The third method for characterizing the simulation conformers is to compare the residue centroid difference distance matrices between the simulation structures and the X-ray conformers. First, the residue centroid distance matrices are calculated for each X-ray structure and each of the 11 simulation structures. To do this calculation, the centroid (i.e. the geometric center position of all atoms) in each of the 159 protein residues is calculated. Then, the distance between every pair of centroids is calculated and subsequently used to generate a 159x159 distance matrix. The matrices are symmetric, since, for instance, the distance between residues 14 and 24 is the same as the distance between residues 24 and 14. These matrices indicate the areas of close contacts within the protein. Figure 20 displays the M20 loop regions (residues 14 to 24) of these matrices for the three X-ray conformers as well as the new conformer (i.e. the one obtained from the open simulations). They show residues that are in contact with the residues of the M20 loop. However, using them to visually distinguish between conformers is difficult.

The residue centroid difference distance matrices, on the other hand, make it easy to match conformers. If two conformers are similar in structure, then they should have similar distance matrices, and subtracting their two distance matrices would generate a difference distance matrix with values close to zero. Therefore, matching the conformation of one structure to a set of other structures, involves calculating the corresponding difference distance matrices and looking for the one with the lowest values. The absolute values of the differences are used since the order of subtraction is unimportant. Figure 21 shows the difference distance plots used to characterize the simulation minimized average structures. Each of the eleven simulation structures is compared to the four following structures: X-ray open (1RA1), closed (1RX1) and occluded (1RX7) and the new conformer (the folate-bound open simulation minimized average structure). Some of the plots show clear matches, such as the *apo*-occluded simulation matching the occluded X-ray conformer (Figure 21C). Others, such as the *apo*-closed simulation (Figure 21B), show no obvious matches. The *apo*-open (Figure 21A), folate-open (Figure 21F) and DHF-open (Figure 21I) all show major deviations with respect to the open X-ray conformer in the central portion of the M20 loop, but have very close matches to the new conformer. The NADPH-open (Figure 21D) on the other

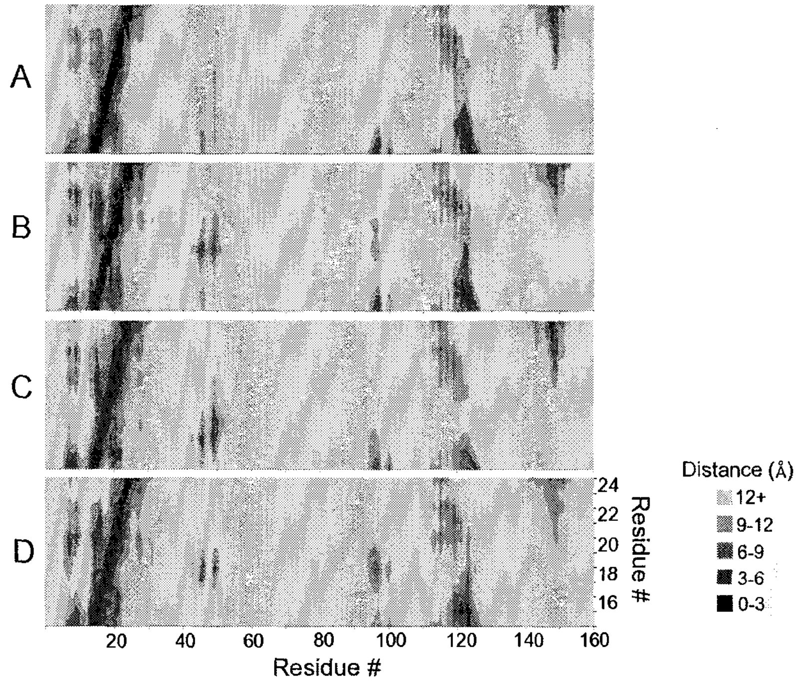


Figure 20. DHFR residue centroid distance matrix. Shown are M20 loop-region distance matrices for open-1RA1 (A), closed-1RX1 (B) and occluded-1RX7 (C) X-ray structures and the folate-bound minimized-average simulation structure (D).

hand matches the new conformer best for residues 20 to 24 and either the open or closed X-ray conformer best for residues 16 to 19, which agrees with the RMSD analysis.

To quantify the results of the plots shown in Figure 21, the average distance values on the matrices used to generate the plots are calculated. The calculations exclude the diagonal elements (i.e. res14|res14, res15|res15, …, res24|res24), since those values are all zero. The averages are calculated for the difference distance matrices of the 11 simulation structures with respect to the three X-ray M20 loop conformers and both the folate-closed and DHF-occluded simulation structures. These structures are the same structures as used in the M20 loop backbone RMSD comparisons (see Table 4), therefore it makes sense to compare the results of both methods. However, there are several noteworthy differences between using the M20 loop backbone RMSD to characterize the M20 loop conformation versus using the average value of the M20 loop region of the residue centroid difference distance matrix to characterize it. First, the RMSD calculation requires prior alignment of the entire protein backbones of the two structures. On the other hand, the distance matrices of the two structures can be calculated and subtracted

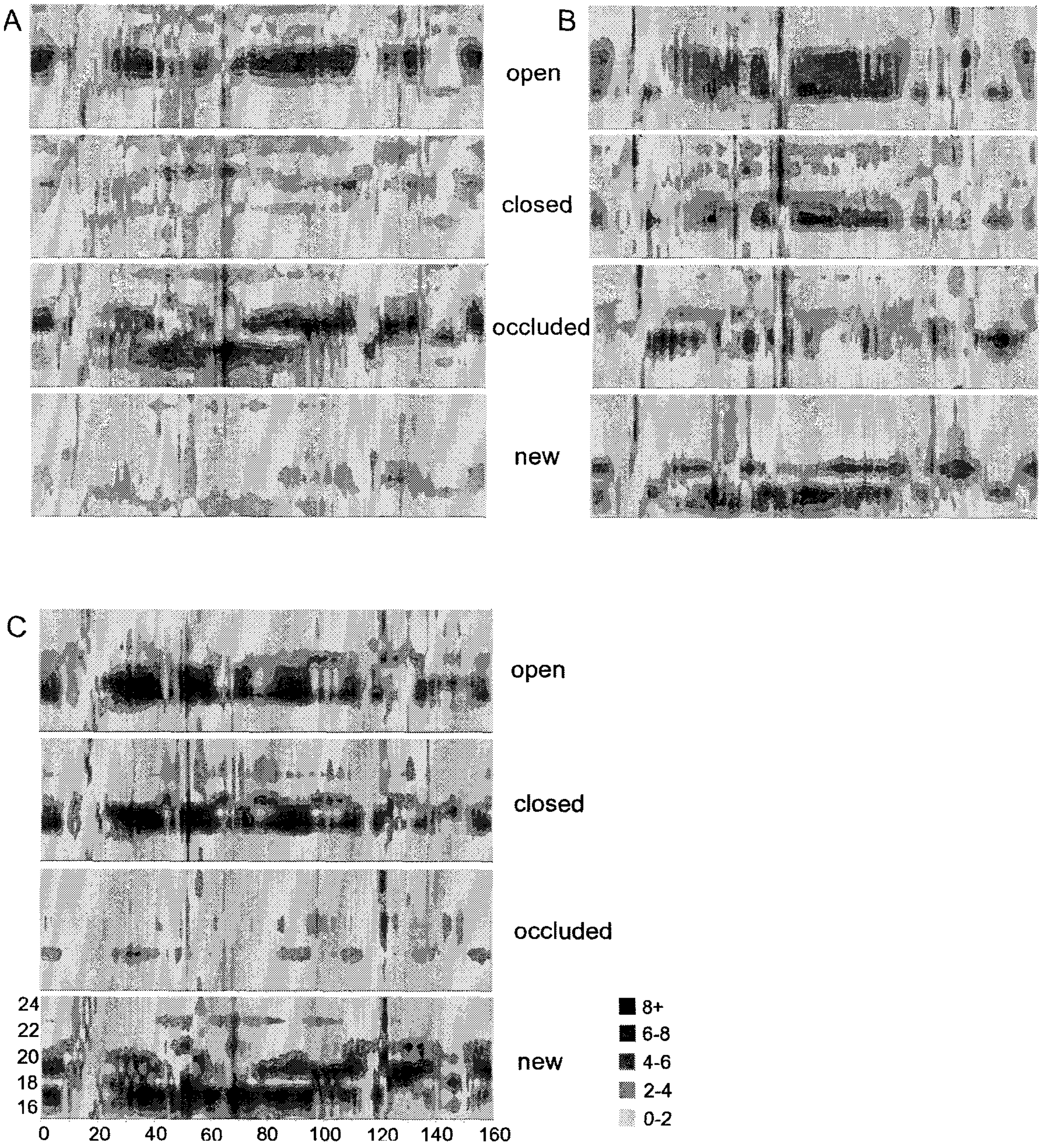


Figure 21. DHFR residue centroid difference distance matrices. Shown are the M20 loop regions for all 11 simulation structures: *apo*-open (A), *apo*-closed (B), *apo*-occluded (C), NADPH-open (D), NADPH-closed (E), folate-open (F), folate-closed (G), folate-occluded (H), DHF-open (I), DHF-closed (J) and DHF-occluded (K). Small distances (0-2 Å) indicate similarity in structure. Large distances (8+ Å) indicate deviation in structure.

Figure 21 Continued

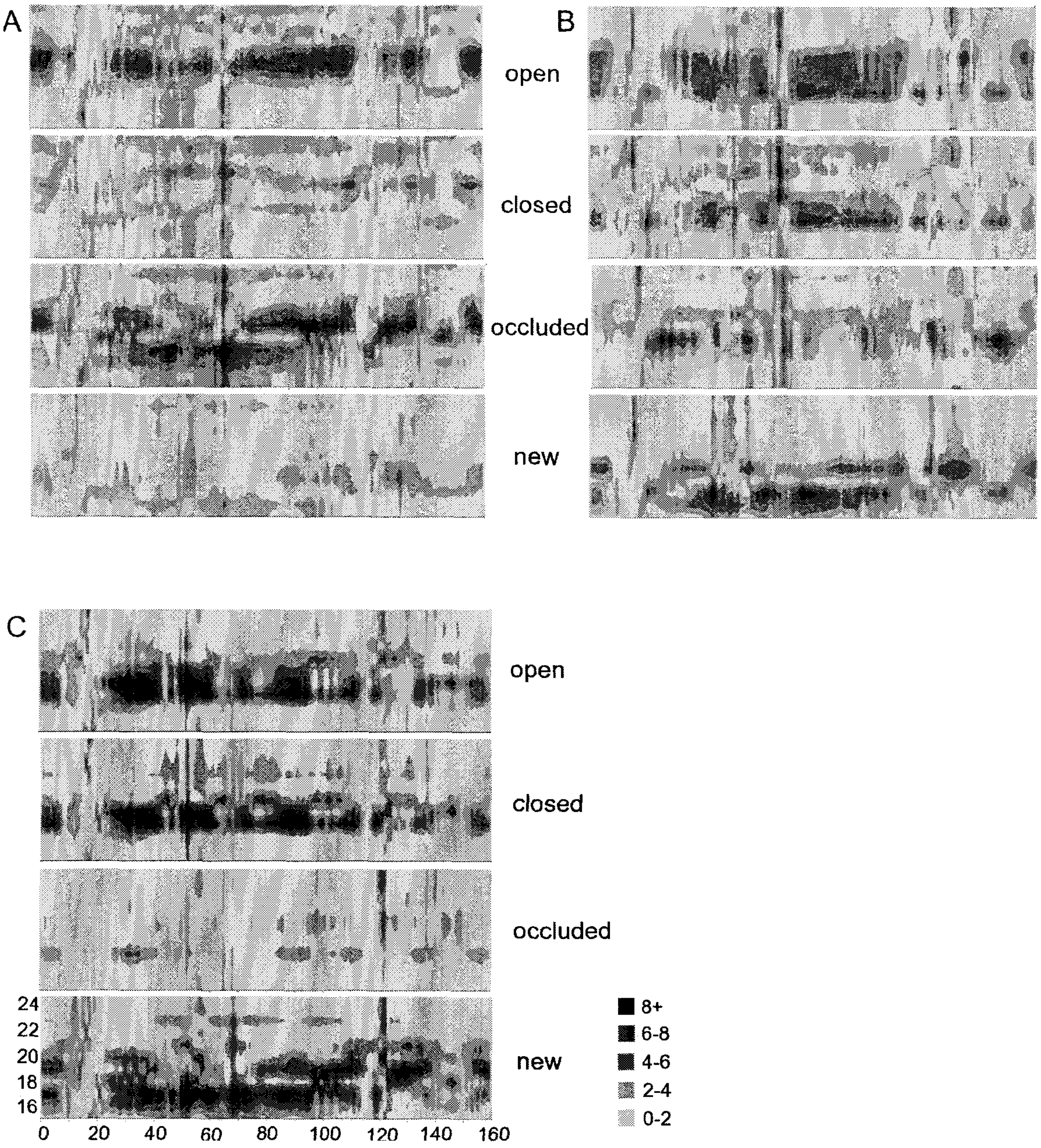
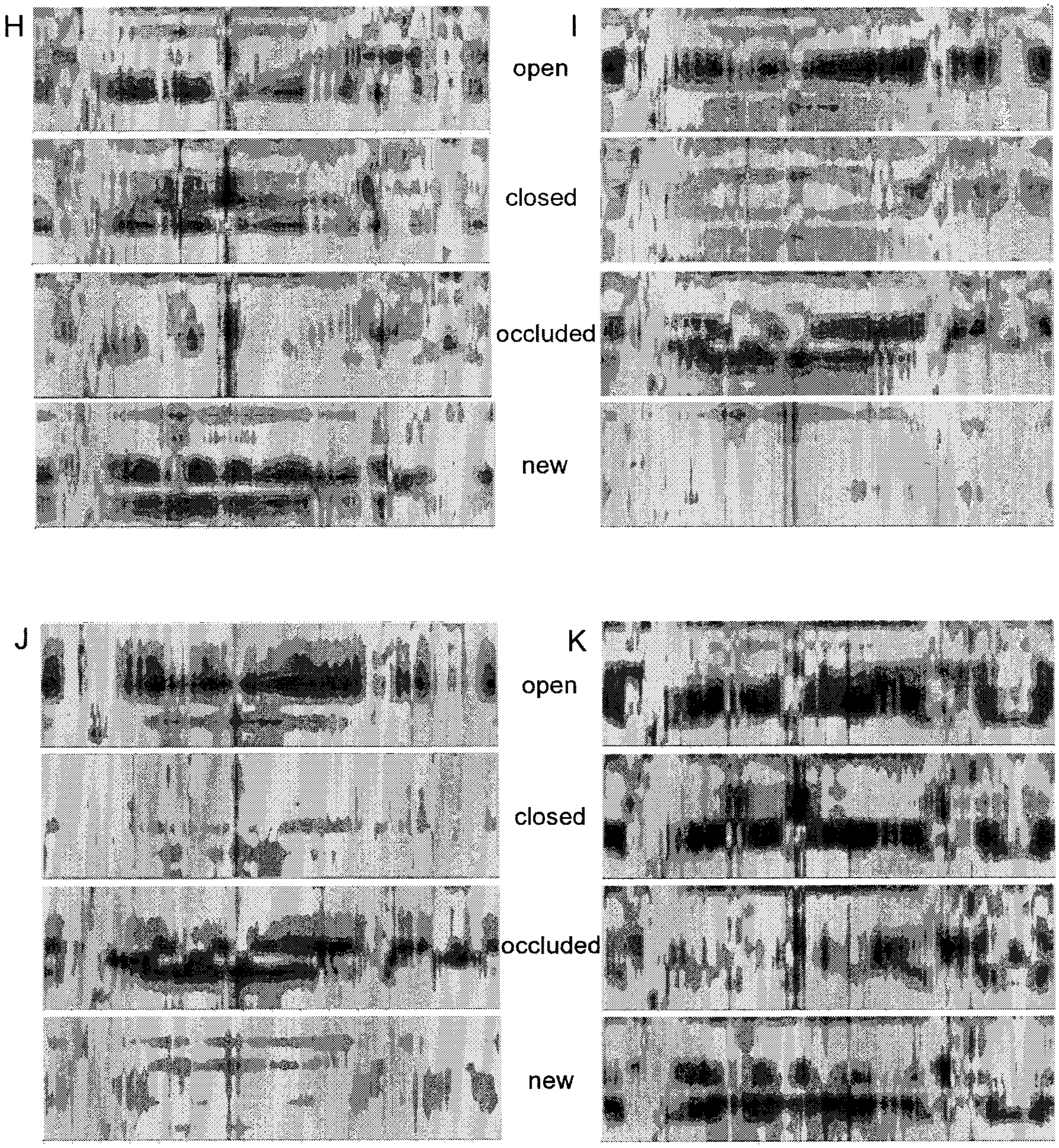


Figure 21 Continued



without alignment, because distances between two parts of a rigid structure always remain constant, even after rotating it (e.g. during an alignment procedure). Second, the RMSD calculation only involves the positions of the M20 loop backbone atoms of both proteins, in contrast to the average difference distance matrix calculation, which involves the positions of all protein atoms. So, the difference matrix average characterizes the conformation of the entire M20 loop, in contrast to the M20 loop backbone RMSD, which characterizes only the M20 loop backbone conformation. Also, the difference distance matrix averages are lower in value than the corresponding RMSDs of Table 4, which is due to an averaging effect from the very low numbers in large portions of the plots (as seen by the relatively large region light grey cells in the difference distance plots of Figure 21).

Table 7 shows the calculated averages and indicates (in bold) a match in the M20 loop conformation of a structure when one value is much lower than all others in the row. The obvious matches have values well below 1.50 Å. There is a very strong correlation between the clear matches in Table 7 and those from the RMSD calculations in Table 4. Except for the NADPH-bound complex, the simulations starting with the M20 loop open match the new conformer (i.e. the folate-open minimized average structure). The

Table 7. Residue centroid difference distance matrix averages (Å)

Complex	Starting M20 Loop	X-ray Open*	X-ray Closed*	X-ray Occluded*	New†	Folate closed	DHF occluded
Apo	open	2.17	1.60	2.36	**1.18**	1.59	2.15
	closed	1.98	1.81	1.72	1.81	**0.70**	1.82
	occluded	2.08	1.89	**1.13**	2.34	1.54	1.86
NADPH	open	1.60	1.48	2.07	1.67	1.76	2.32
	closed	1.62	**0.80**	2.06	1.93	2.05	2.65
Folate	open	2.45	1.93	2.44	0.00	1.73	2.43
	closed	2.11	2.05	1.78	1.73	0.00	1.97
	occluded	2.15	2.26	1.73	2.30	1.70	**1.32**
DHF	open	2.51	1.91	2.54	**1.16**	2.13	2.27
	closed	2.18	1.15	2.26	1.28	1.94	2.60
	occluded	2.87	2.64	2.29	2.43	1.97	0.00

* X-ray coordinates are open (1RA1), closed (1RX1) and occluded (1RX7)

† Folate-open minimized-average simulation structure

NADPH-bound complex does not clearly match any of the 6 structures. Rather it compares similarly to several structures: the X-ray open (1.60 Å), the X-ray closed (1.48 Å), the new (1.67 Å) and the folate-closed simulation structure (1.76 Å). The RMSD calculations also show a nonmatching trend of the NADPH-bound open simulation structure. The difference between the NADPH-bound structure and the other open simulation structures in these calculations most likely arises from the presence of steric distortions on the M20 loop by the bound cofactor. The NADPH-closed simulation stays well conserved with the X-ray starting conformation (0.80 Å). While the folate-closed and *apo*-closed both shift from their X-ray starting points (2.05 Å and 1.81 Å, respectively), the end up with very similar conformations (0.70 Å), which the RMSD calculations also show. The DHF-closed structure resembles the X-ray closed conformer (1.15 Å) and the new conformer (1.28 Å), a result also seen in the RMSD calculations, except that the RMSD calculations showed a slightly closer match to the new conformer. The *apo*-occluded simulation structure remains well conserved with the X-ray conformer (1.13 Å). While the folate-occluded and DHF-occluded shift away from the X-ray closed starting conformer (1.73 Å and 2.29 Å, respectively), they end up with a similar conformation (1.32 Å), a result also shown by the RMSD calculations.

So the structural behavior of the M20 loop is summarized as follows: the M20 loop either stays well conserved (e.g. in the NADPH closed and *apo*-occluded structures), shifts slightly as the folate and DHF occluded and the *apo* and folate closed conformers did or shifts to a new conformer as observed for the open conformer simulations. The folate and DHF occluded M20 loops shifted similarly, as did the *apo* and folate closed conformers. The conformational changes correlate with the disruption of hydrogen bond markers that were used to characterize the open, closed and occluded X-ray structures. Loss of M20-GH interloop contacts occurs with the open to new conformational change. Loss of both the M20-FG interloop contact and the type III' hairpin turn contacts occur with the *apo* and folate shifts in the closed conformers, whereas the NADPH and DHF closed simulations retain those contacts and their conformation. Loss of both of the M20-GH interloop contacts and the 3_{10}-helix contact correlates with the shift in the folate and DHF occluded conformations, unlike the *apo*-occluded conformer, which retains those contacts and its conformation. These results strongly suggest that intra and inter-loop hydrogen bond contacts help stabilize certain M20 loop conformations. In addition, ligands can have a subtle effect on the conformation of the M20 loop, as the loop shifts, depending on whether or not the nicotinamide site is empty.

Equilibrium Molecular Dynamics: Free energy Analysis

MMPBSA free energy analysis was performed on the 11 simulation structures. Before estimating the free energy, a 2-ns simulation window was chosen, in which the structure remains stable. Then, the MMPBSA energy is calculated at regular intervals during the window, and the average of those calculations is then taken. A minimum of two nanoseconds is required to obtain reasonable quasiharmonic estimates of the simulation structures' entropies. The protein needs enough time to sample a representative portion of phase space before the atomic positional covariance matrix is calculated. There are several criteria for selecting the 2-ns window from the entire trajectory. First, a visual inspection of the simulation interval must show the M20 loop either in a conserved state or fluctuating about an average conformation. Second, the M20 loop backbone RMSD must be relatively flat during this interval. Third, the MMPBSA energy (entropy not included), which is calculated throughout the entire trajectory, must have reached a well-defined average by the time the interval is sampled. This energy term represents the protein's stability within the force field and surrounding implicit solvent field. Performing these three tests is a cumbersome process, and it is difficult to satisfy all three stringent constraints perfectly. The most importance is placed on the visual inspection, since the goal of the calculation is to obtain the relative stabilities of 'separate' conformers.

The 3-test procedure is demonstrated here for the NADPH-bound open simulation. Visual inspection shows the M20 loop moves to the new conformation after 2 ns of dynamics (Figure 22), and remains there. Corroboration is provided by the M20 loop backbone RMSD with respect to the superimposed X-ray starting point (Figure 23). The loop moves fairly quickly up to a 3 Å RMSD and is relatively flat for the remainder of the simulation. Towards the end of the simulation, the RMSD does increase to around 4 Å, but does return to 3 Å. The MMPBSA energy is plotted in Figure 24. The running average (dotted line) slowly decreases until about 8 ns of dynamics have

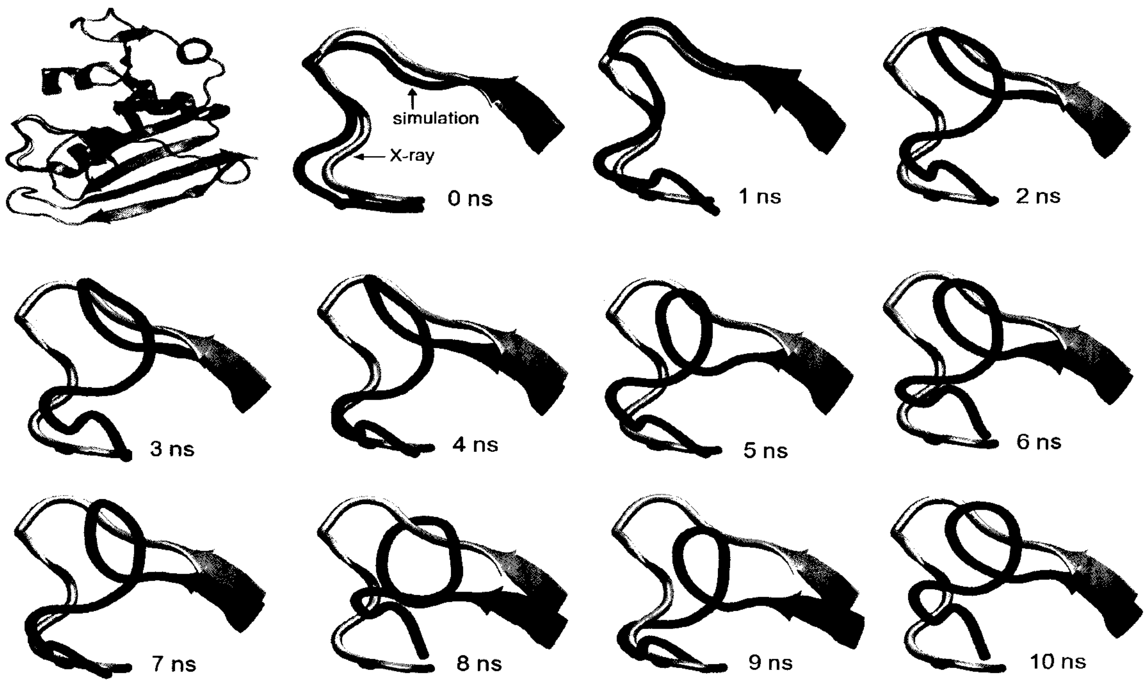


Figure 22. M20 conformations along the NADPH-bound open simulation trajectory. Snapshots of the M20 loop portion are shown after every 1 ns of simulation time. The superimposed open X-ray structure (1RA1) is also drawn.

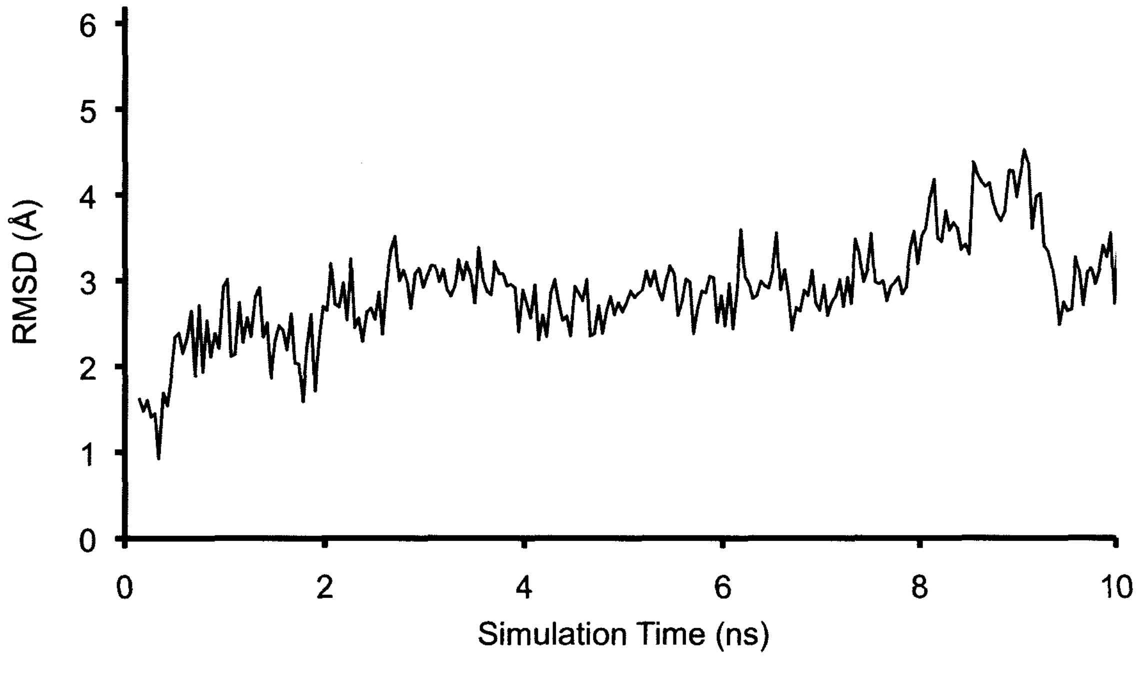


Figure 23. M20 loop backbone RMSD of NADPH-bound open simulation. The X-ray starting point (1RA1) is superimposed using protein backbone atoms in a mass-weighted best fit. The time-series RMSD between the simulation structure and X-ray is calculated using only coordinates of the backbone atoms of the M20 loop.

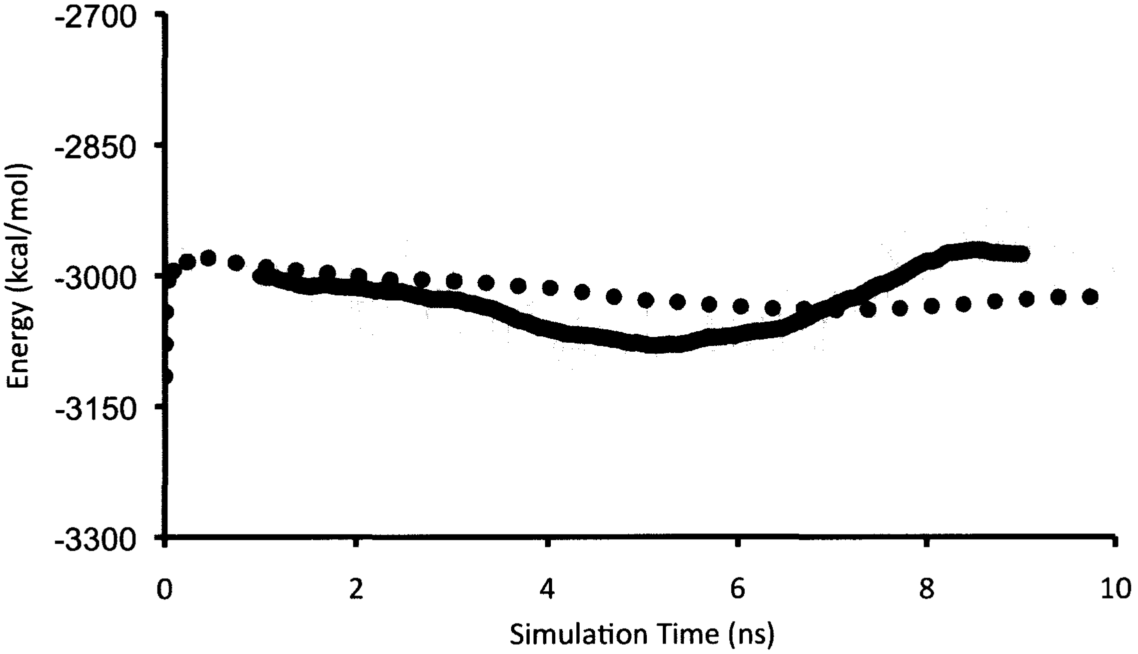


Figure 24. MMPBSA energy of NADPH-bound open simulation. The MMPBSA energy is plotted as a thin line, the running average as a thick dotted line and the average of the encompassing 2 ns (i.e. the value at 1 ns is the 0-2 ns average) as a thick solid line.

completed, where it starts to rise again. So the shift in structure towards the end of the simulation (as indicated by the RMSD bump to 4 Å) is manifested here as a resulting increase in the running average of the MMPBSA energy. Since 2 ns of the trajectory are necessary for calculating the average, it is helpful to see how this average changes as the simulation progresses. Therefore, the encompassing 2-ns average is also plotted (solid blue). Each plotted value is the average of the surrounding 2 ns. That is, the value plotted at 1 ns is the average from 0 to 2 ns, and the value plotted at 7 ns is the average from 6 to 8 ns. The dip in the thick blue line from 4 to 6 ns indicates that 3 to 7 ns is a good section of the simulation, from which to choose the 2-ns window. Therefore, after considering the answers to these three tests, the window is chosen to be from 5 to 7 ns, and the resulting MMPBSA average is -2423.5 kcal/mol. The minimized average structure across this window is taken to represent the simulated complex's structure, and it is the same structure used in the above structural analysis portion of this section.

The MMPBSA average energies are combined with their corresponding entropy estimates. The same 2-ns windows are used to calculate the quasiharmonic entropy (described in Chapter II). The M20 loop conformer relative free energies for each complex are displayed in Table 8 below. The relative energies are calculated for each protein complex by setting the lowest energy conformer to zero. Two observations are immediate: the separate conformers differ significantly in energy and the 'new' conformers are the most stable in the *apo* and folate-bound complexes. These results disagree with the experimental NMR data discussed in Background section of this chapter (78), since, in the folate-bound complexes, the free energy of the new conformer is estimated to be about ~90 kcal/mol lower than the other two. Using the familiar relation, $\Delta G=-RT\ln K$, this corresponds to a population of the new, much greater than 99.9999%, whereas the NMR observations place the population of the closed at 90% and the other (probably the open conformer) at about 10%. Also, X-ray only observes the folate-bound complex in both the open and occluded M20 loop conformations (67), whereas the free energy calculations predict that occluded is high in energy. In addition, experimental NMR studies observe the apo complex in 2 conformers (107), but two of comparable energies are not calculated with this method. Interestingly, the NADPH-closed simulation is lower in energy than NADPH-open, which agrees with

Table 8. Relative free energies of M20 loop conformers (kcal/mol)

Ligand	Starting M20 Loop	E_{MM}	ΔG_{PB}	ΔG_{SA}	TS (heavy)	ΔG (heavy)	TS (all)	ΔG (all)
Apo	open	-87.3	-2386.4	50.3	1903.0	0.0	2246.0	0.0
	closed	-46.7	-2390.8	50.0	1893.2	45.8	2235.0	47.0
	occluded	-95.1	-2350.3	49.3	1904.5	25.9	2248.1	25.3
NADPH	open	195.4	-3222.6	51.4	1931.9	25.6	2273.4	31.2
	closed	158.4	-3172.2	49.6	1968.9	0.0	2316.1	0.0
Folate	open	-747.0	-2566.7	48.2	1917.6	0.0	2259.1	0.0
	closed	-600.5	-2611.4	49.6	1935.2	85.6	2277.3	85.0
	occluded	-474.2	-2722.6	49.7	1946.5	89.5	2291.6	85.9
DHF	open	-716.6	-2631.8	49.2	1953.4	70.0	2297.9	67.9
	closed	-837.3	-2593.7	48.3	1940.0	0.0	2282.4	0.0
	occluded	-689.2	-2631.7	49.8	1941.2	110.4	2285.1	108.9

experiment (76). The all-atom entropies were also calculated (Table 8). Not surprisingly, the trend is exactly the same. Thus, these numbers cannot be trusted for comparing the relative free energies of DHFR conformers. Therefore, we applied other methods to try and investigate the link between ligand structures and M20 loop conformation.

External Perturbation Molecular Dynamics and Normal Mode Analysis

Conformational changes and ligand dissociations were forced using the methods discussed in Chapter II. Table 2 (in the Methods section of this chapter) describes the 31 simulations: the 9 forced ligand dissociation simulations (each one run twice) and the 13 forced conformational change simulations. Both types will be similarly analyzed to determine whether their trajectories involve the protein sampling either the open X-ray or new conformer. In addition, whether or not the ligand has an effect on the pathway followed will be investigated. The main tool for making this determination is the M20 loop backbone RMSD. To analyze a simulation, each frame of the trajectory is first aligned with a reference structure by superimposing the protein backbone atoms in a mass-weighted fit. Then, the RMSD of just the M20 loop backbone atoms is calculated for each frame. The following analysis is divided into two parts: forced ligand dissociations and forced conformational changes.

Each forced ligand dissociation simulation finished with the ligand successfully dissociated. Some crashed before the 100 ps because their ligands traveled too far from the protein and reached into neighboring periodic boxes causing problems. In most of the THF dissociation simulations, RMSD plots show that the M20 loop remained nearer to its starting conformation than any of the others (Figure 25). Only the E|THF closed → THF out simulation involves a possible path through the new M20 loop conformation (Figures 25E and 25F). Although the graphs clearly indicate that the simulation structures in both cases maintain their closed starting conformation, the loops do get close (~ 2 Å) to the new conformation as dissociation continues. But, the close pass may be more of a random fluctuation in RMSD than a large relative drop. All second dissociation runs starting from the occluded conformer involve the M20 loop getting far from its starting conformation (Figures 25D, 25H and 25L). Visual inspection shows that in all three cases the M20 loop protrudes further into the active site to occlude part of the THF binding site. The

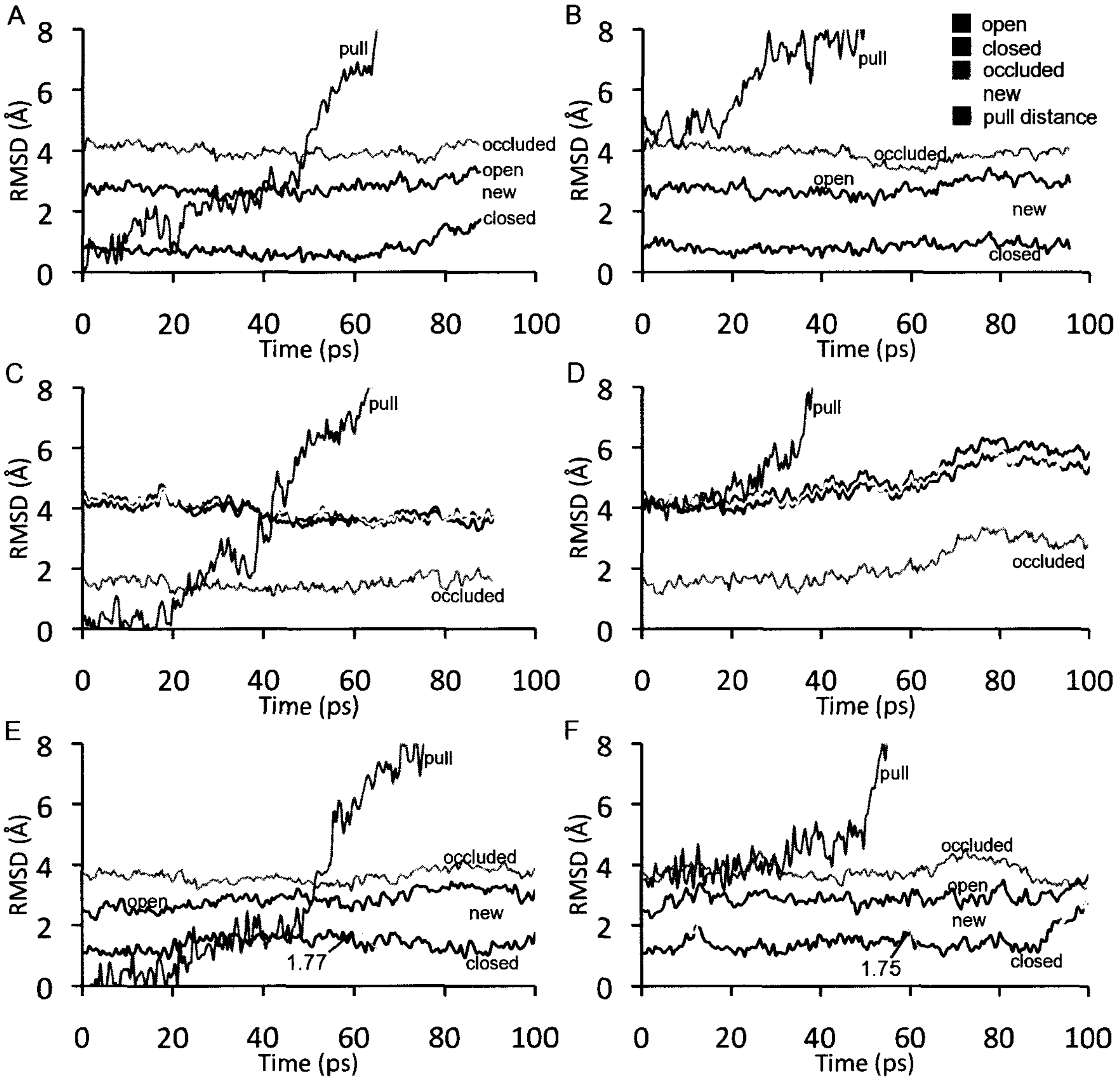
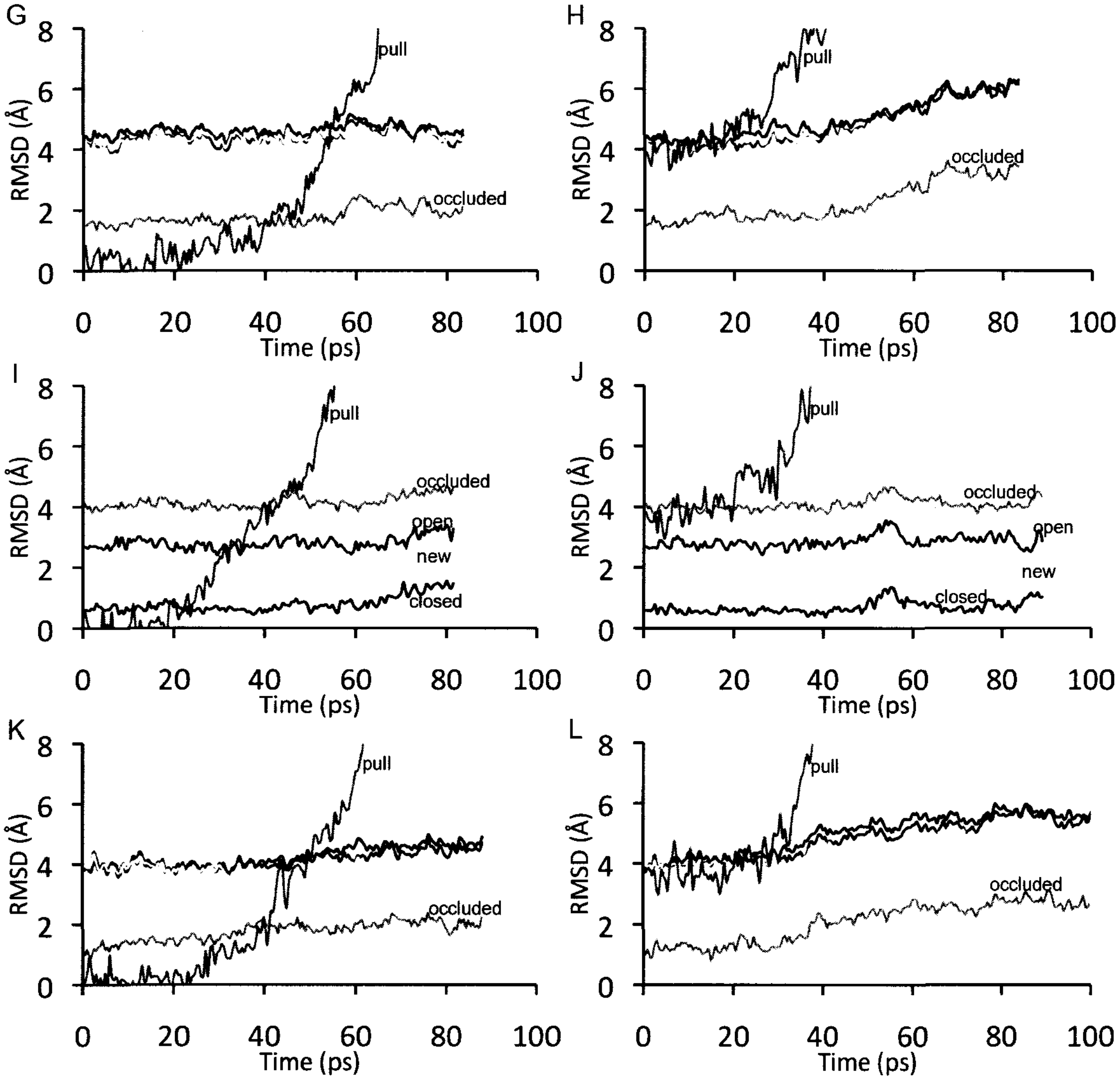


Figure 25. M20 loop backbone RMSD during forced THF dissociation simulations. Shown are E|THF|NADP+ closed → THF out (A-B), E|THF|NADP+ occluded → THF out (C-D), E|THF closed → THF out (E-F), E|THF occluded → THF out (G-H), E|THF|NADPH closed → THF out (I-J), E|THF|NADPH occluded → THF out (K-L). In each simulation pair, the first run involves pulling the C atom of folate and the second run involves pulling the CB atom (see Figure 10). Simulation M20 loop backbone RMSD is plotted relative to the following structures: open-1RA1, closed-1RX1, occluded-1RX7 and new-NADPH open simulation. Because it is large, the distance between pulling atoms is plotted 12 Å less than actual. Low RMSD points are noted in Å.

Figure 25 Continued



corresponding first simulation runs do not exhibit this behavior. In all dissociations from the closed conformer, the M20 loop remains conserved.

As THF dissociates, it leaves along a direct route to the solvent (Figure 26), with no protein loop blocking the way as happens when the cofactor leaves (the M20 loop blocks the way). Several of the simulations show that as THF leaves, it interacts with α-helix B and the loop connecting α-helix C to β-strand C, as if to pull them out into solution with itself (Figure 26 shows this effect for the first run of E|THF|closed → THF out). The

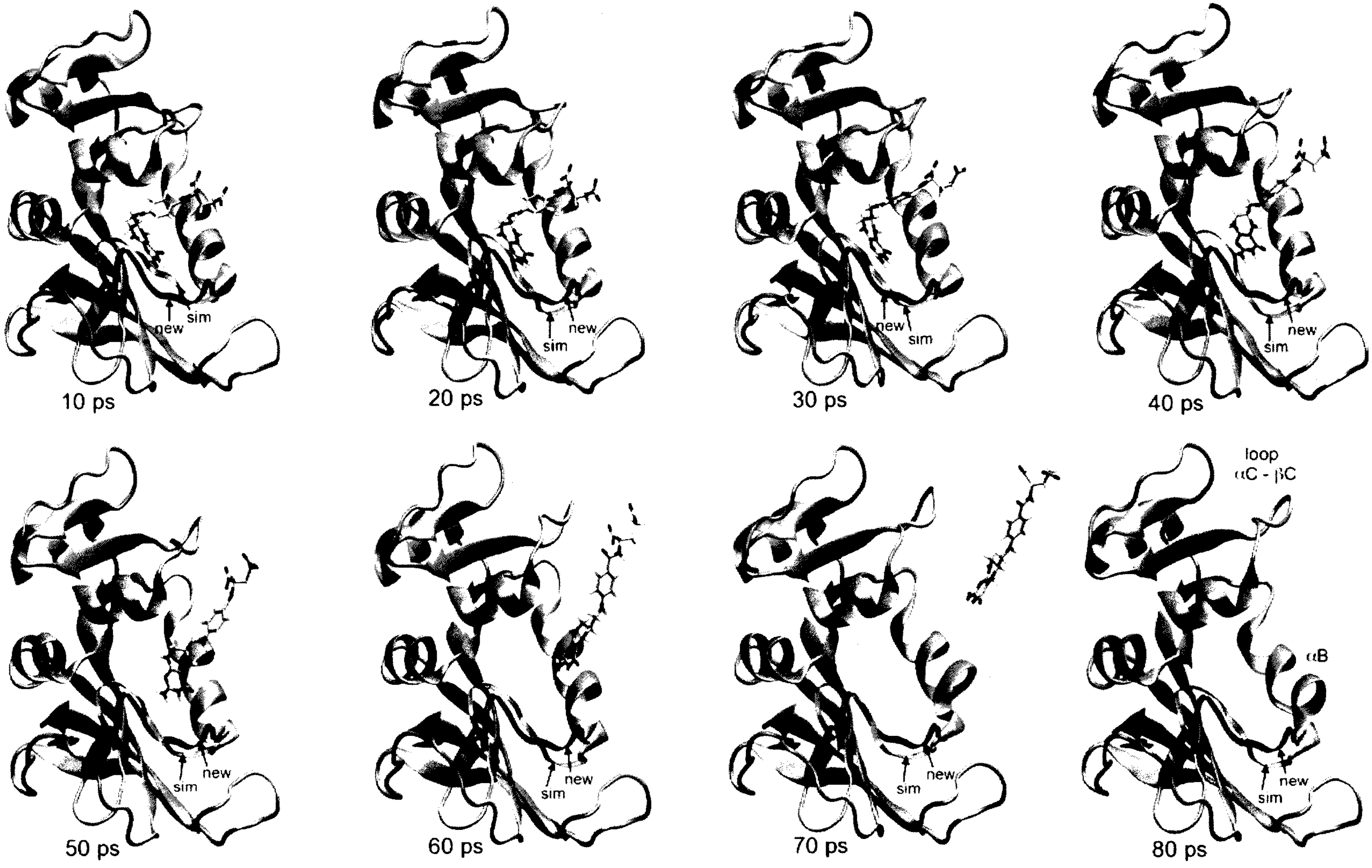


Figure 26. THF dissociation from E|THF|closed (1st run). The superimposed new conformer (NADPH open simulation) is also shown.

effect is stronger in some simulations than in others, but its occurrence does not seem to depend on either the starting M20 loop conformation or the identity and location of any cofactor present. It is interesting though, that α-helix B (residues 25-36) is directly connected to the M20 loop (residues 14-24).

It is interesting to compare the THF dissociation to experimental kinetic data. Figure 25 shows THF dissociating more quickly from E|THF|NADPH (occluded or closed) than from E|THF|closed. This might be interpreted as agreeing with the kinetic experiments (77), which shows a 1.4 s^{-1} rate from E|THF and a 12.5 s^{-1} rate from E|THF|NADPH. However, crystallography (17) and NMR (76) studies indicate that E|THF has an occluded M20 loop. Figure 25 shows dissociation occurring just as quickly from E|THF|occluded than from E|THF|NADPH|occluded, thus disagreeing with the kinetic data. Also, there seems to be no difference between THF dissociation from E|THF|NADPH and E|THF|NADP+. This disagrees with the experimental data, which shows the latter to be about 5 times quicker. Therefore these forced dissociations cannot reproduce experimental data.

Most of the cofactor dissociations involve the M20 loop sampling conformations close to the new conformation. The M20 loop RMSD graphs show a common characteristic among these simulations to be that once dissociation begins, the RMSD relative to the starting closed conformation increases, while that relative to the new conformer decreases (Figure 27). This trend continues until the two RMSDs meet each other, where at this time, the simulation structure is sampling a conformation equally near to both the closed and new conformers. In the first E|NADPH|closed → NADPH out dissociation, the M20 loop reaches 1.87 Å and 1.98 Å relative to the closed and new conformers, respectively (Figure 27A), and in the second run, it reaches 1.65 Å and 1.81 Å, respectively (Figure 27B). Figures 28 and 29 show visual renderings of the first and second runs, respectively. The first E|THF|NADP+|closed → NADP+ out run does not involve a decrease in M20 loop RMSD relative to the new conformer, but that relative to the closed conformer suddenly increases after dissociation begins at ~ 50 ps and continues to steadily increase to above 4 Å until the simulation finishes (Figure 27C). Visual analysis of the trajectory reveals that this major distortion in the loop conformation results from the C-terminus portion entering in to occlude the missing cofactors binding pocket, a job normally reserved for the N-terminus portion of the loop (shown visually in Figure 30). But this does not happen in the second E|THF|NADP+|closed → NADP+ out dissociation run (shown visually in Figure 31). The

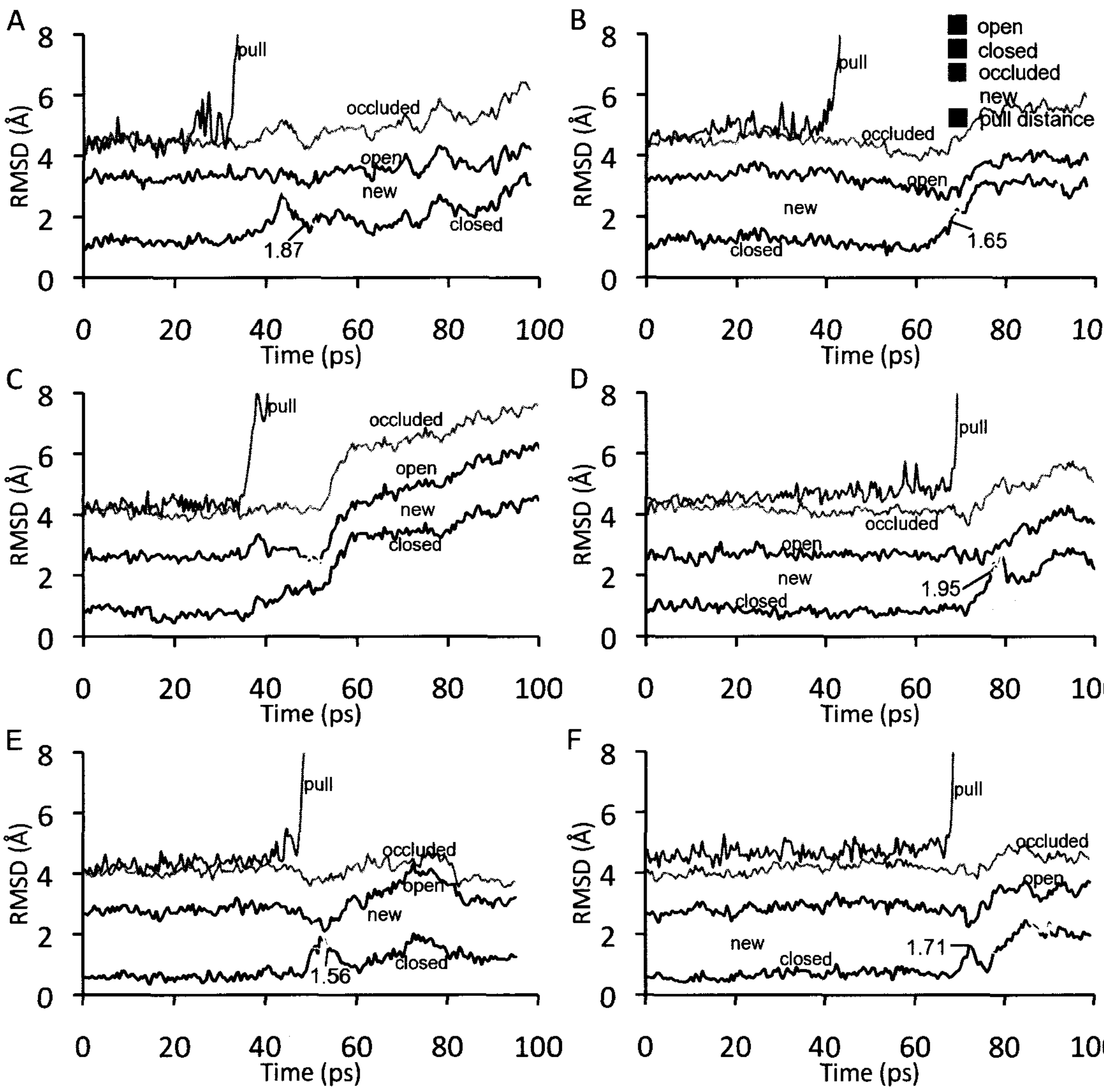


Figure 27. M20 loop backbone RMSD during forced cofactor dissociation simulations. Shown are E|NADPH closed → NADPH out (A-B), E|THF|NADP+ closed → NADP+ out (C-D), E|THF|NADPH closed → NADPH out (E-F). In each simulation pair, the first run involves pulling the NN7 atom of the cofactor and the second run involves pulling the NC7 atom (see Figure 10). Simulation M20 loop backbone RMSD is plotted relative to the following structures: open-1RA1, closed-1RX1, occluded-1RX7 and new-NADPH open simulation. Distance between pulling atoms is also plotted. Low RMSD points are noted in Å.

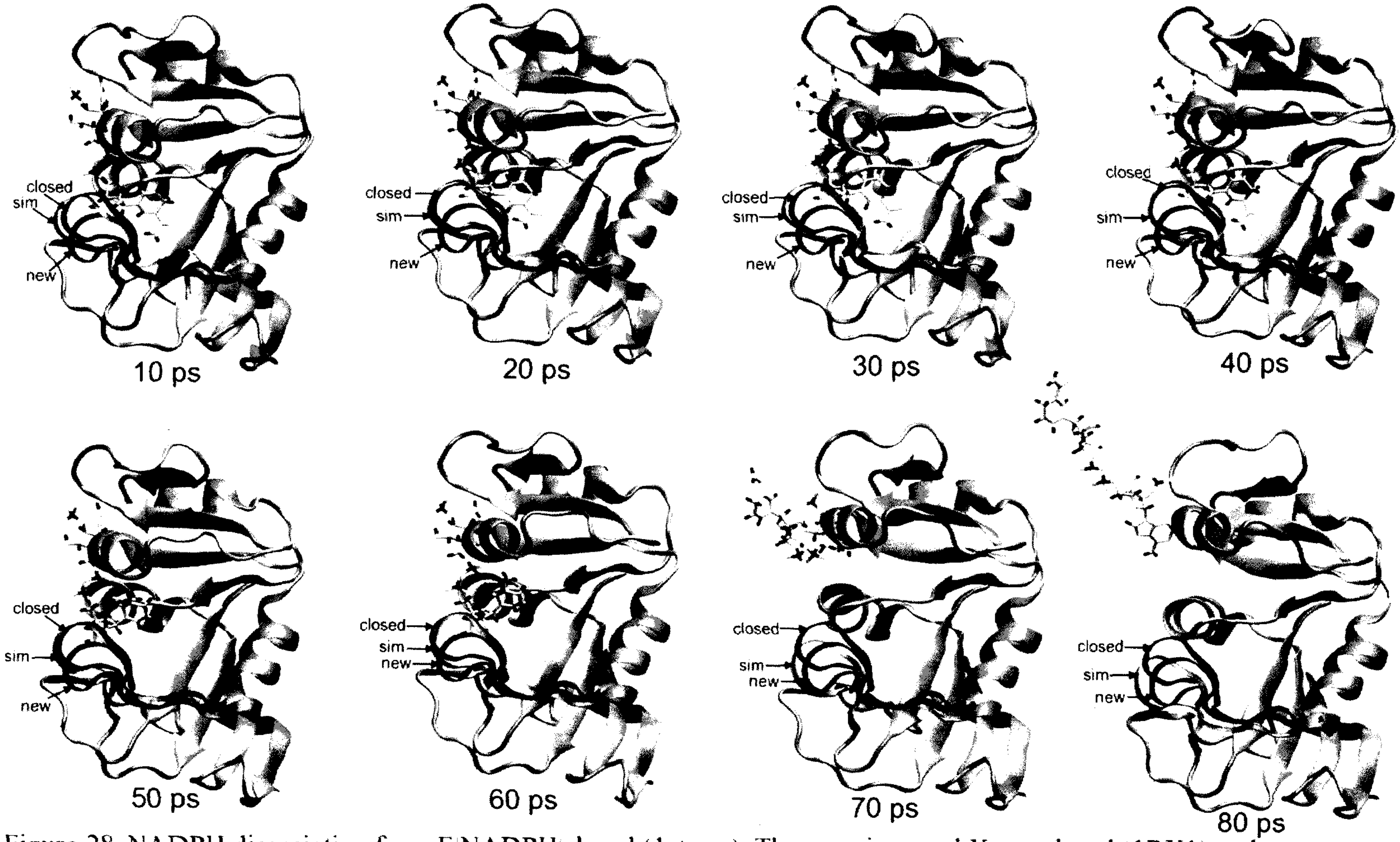


Figure 28. NADPH dissociation from E·NADPH·closed (1st run). The superimposed X-ray closed (1RX1) and new (NADPH open simulation) conformers are also drawn.

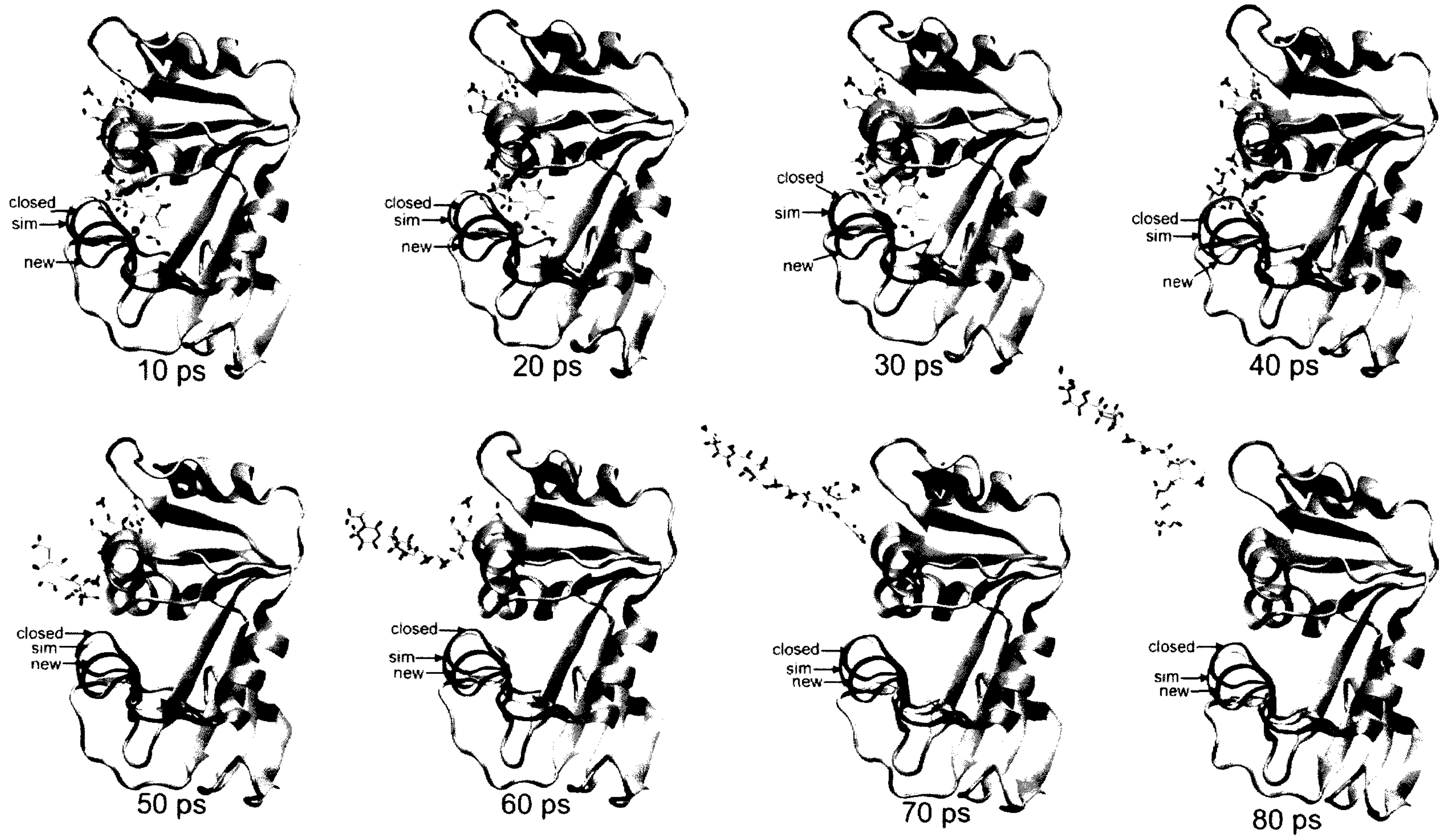


Figure 29. NADPH dissociation from E|NADPH|closed(2ndrun). The superimposed X-ray closed (1RX1) and new (NADPH open simulation) conformers are also drawn.

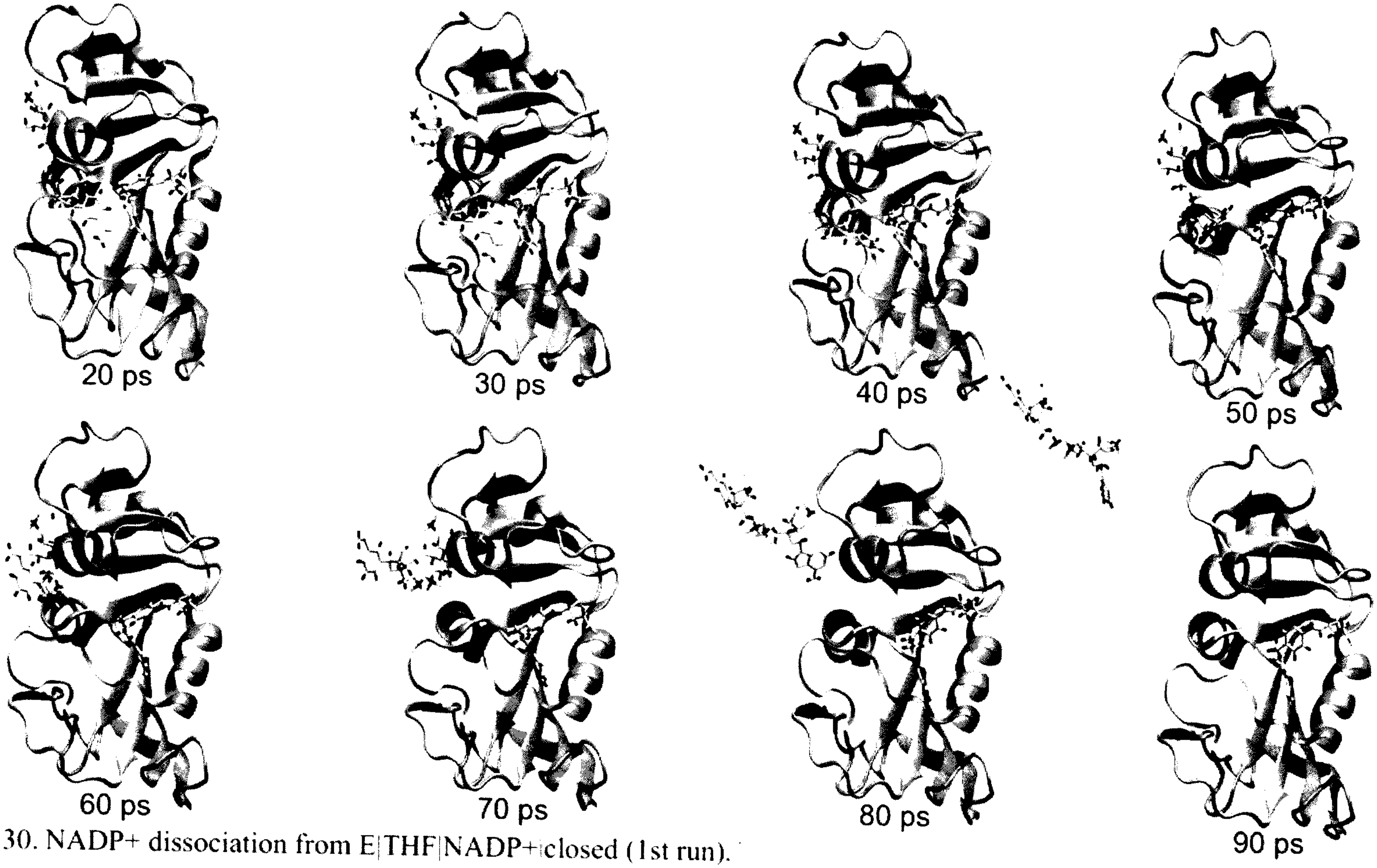


Figure 30. NADP+ dissociation from E|THF|NADP+|closed (1st run).

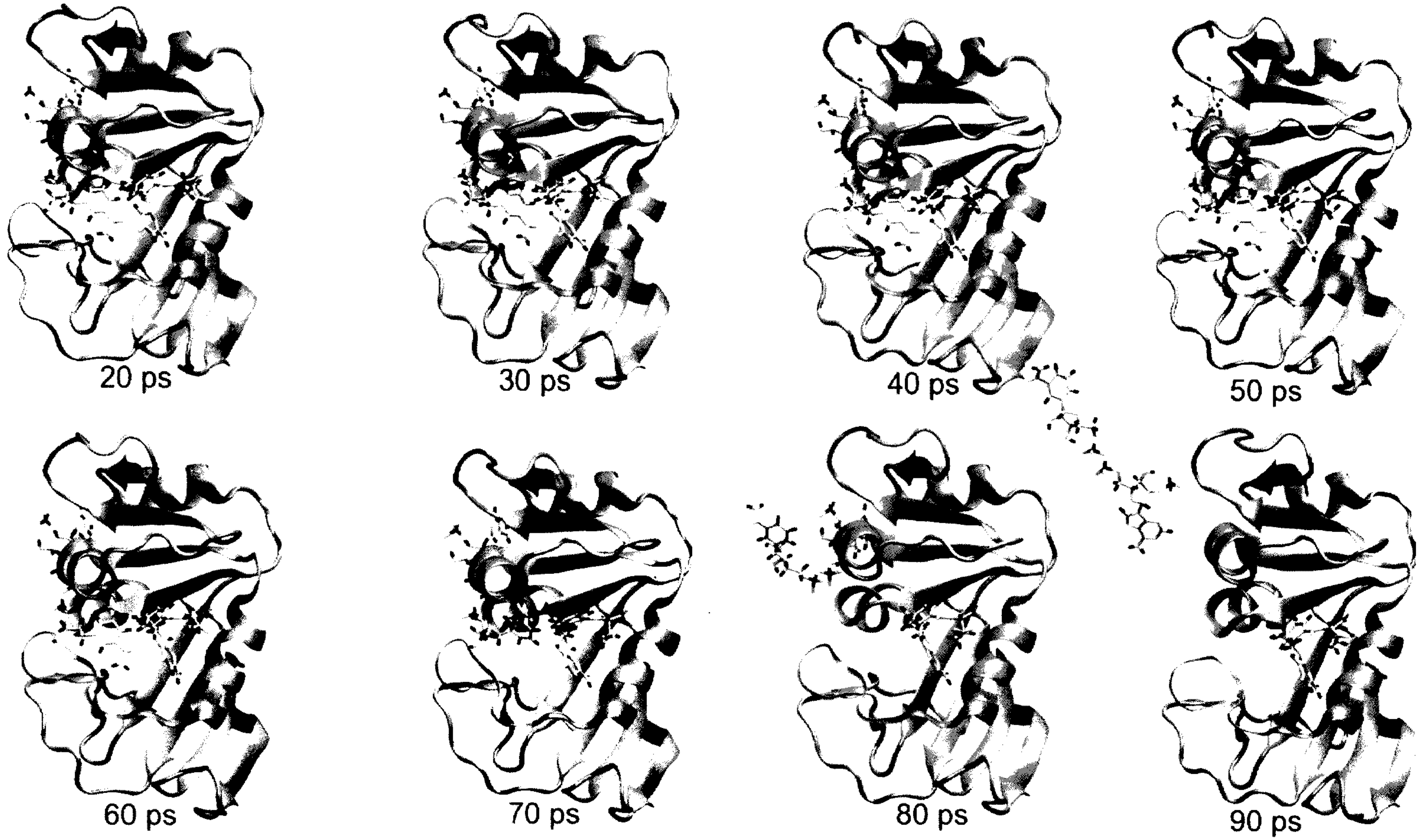


Figure 31. NADP+ dissociation from E|THF|NADP+|closed (2nd run).

second run dissociation begins at ~ 70 ps, and right afterwards, the RMSD relative to the closed starting conformer increases sharply while that relative to the new conformer decreases (Figure 27D). The nearest passes to the new conformer occur in E|THF|NADPH closed → NADPH out dissociations, where the RMSD reaches a minima of 1.50 Å and 1.74 Å in the first and second runs, respectively (Figures 27E and 27F).

At each near pass, the RMSD indicates that the simulation structure resembles both the closed and new M20 loop conformers. To gain a better comparison, residue centroid difference distance graphs are calculated on the five near pass structures (Figure 32). In all five cases, the plots indicate a clear resemblance to both the closed and new conformers. Each deviates similarly from the closed conformer, with major differences from residues 40 to 80 and minor differences elsewhere. Residues 40 to 80 contain beta-strands B, C and D, and alpha-helices C and E, all of which make up what is known as the adenosine-binding subdomain, since it is responsible for binding the adenosine portion of the cofactor. So, the M20 loop has similar proximity relative to the rest of the enzyme as the closed X-ray conformer does, except in the adenosine-binding subdomain. Since the effect is observed in all five structures, it is probably not an artifact of improper methodology, suggesting that it could take place *in vivo*. These results agree with the crystallography studies (67), which show that subdomain rotations do occur during catalysis. The difference plots relative to the new conformer show smaller magnitude deviations in adenosine-binding subdomain portion of the closed conformer-reference plots. But, more deviations with the closed conformer-reference plots, much similarity with respect to the locations of deviation is found among the five new conformer reference plots. This suggests that the M20 loop conformation is similar in these five dissociation snapshots. Visual renderings confirm this (Figure 33). These observations suggest that cofactor dissociation causes the M20 loop to a sample conformation resembling both the new and closed conformers.

Another interesting observation is that cofactor dissociation happens earlier in the first run than in the second run. Figure 27 shows clearly where the pulling distance sharply increases, marking the beginning of dissociation, and in each case it happens about 25 ps earlier in the first run. It might simply be that the first runs involve a more direct route for the ligand out to the solvent. Another possible explanation is that in the first runs, the NN7 atom, which lies on the very end of the cofactor, is being pulled (Figure 10). But in the second runs, the NC7, which lies one bond away from the end, is being pulled. As it is pulled, atoms attached to both sides of it drag alongside and interact

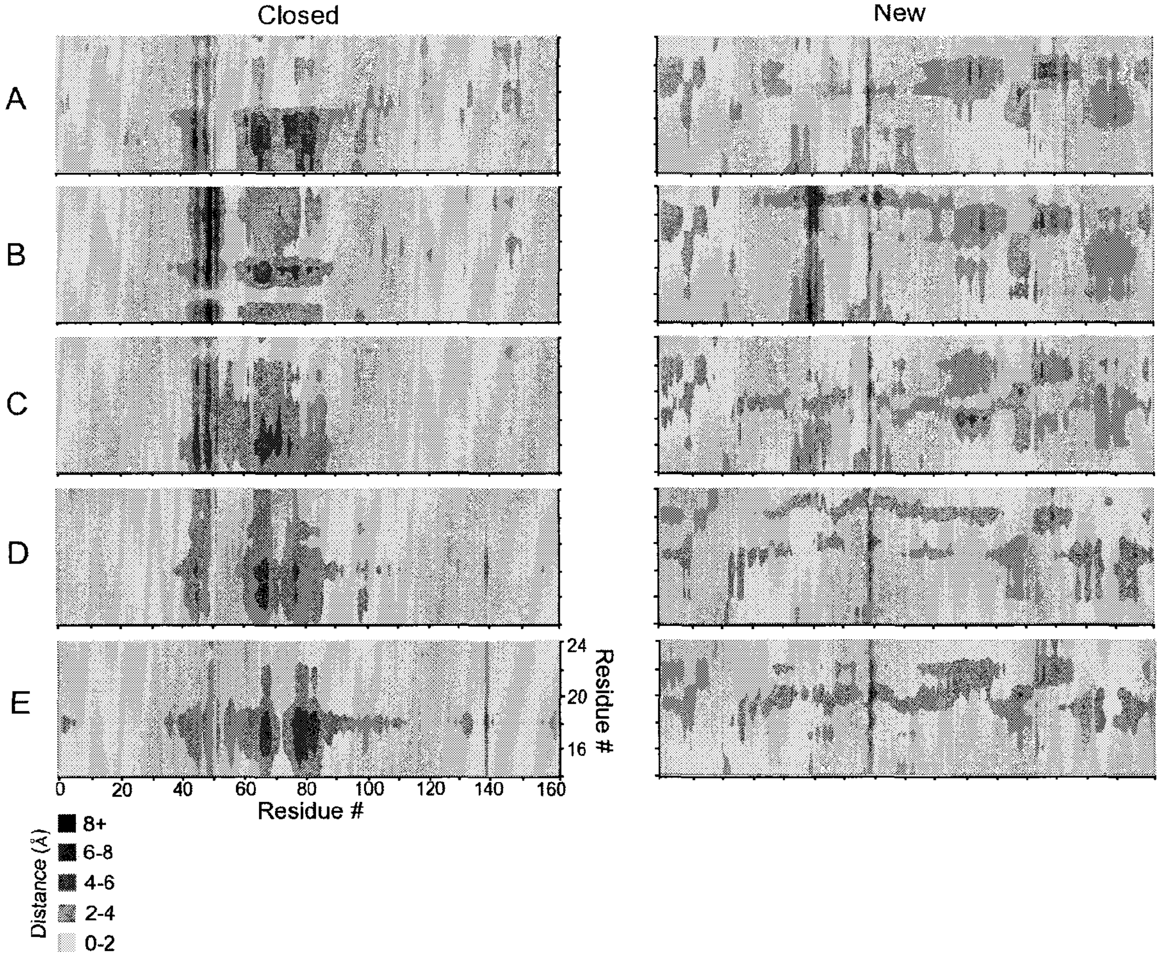


Figure 32. Residue centroid difference distance plots for cofactor dissociation structures. The plots correspond to the following simulation snapshot structures: 50.3 ps in the 1st E|NADPH closed → NADPH out run (A), 66.2 ps in the 2nd E|NADPH closed → NADPH out run (B), 76.6 ps in the 2nd E|THF|NADP+ closed → NADP+ out run (C), 53.4 ps in the 1st E|THF|NADPH closed → NADPH out run (D) and 72.0 ps in the 2nd E|THF|NADPH closed → NADPH out run (E). Difference distance matrices are plotted relative to both the X-ray closed conformer (1RX1) and new conformer (using the NADPH open simulation).

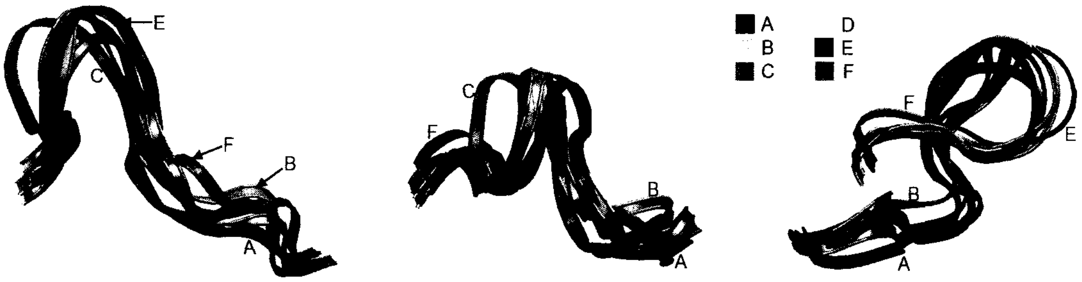


Figure 33. M20 loops of superimposed near-pass cofactor dissociation structures. Shown are snapshots of the first E|NADPH closed → NADPH out run at 50.3 ps (A), the second E|NADPH closed → NADPH out run at 66.2 ps (B), the second E|THF|NADP+ closed → NADP+ out run at 76.6 ps (C), the first E|THF|NADPH closed → NADPH out run at 53.4 ps (D), and the second E|THF|NADPH closed → NADPH out run at 72.0 ps (E). The new conformer using the NADPH open simulation is also shown (F).

with the protein. If the pulling atom were an end atom, there may be less of a drag force, as there are only atoms attached to one side. Hence pulling on an end atom may allow dissociation to be faster.

Closed-occluded conformational changes were forced using the restricted-perturbation targeted molecular dynamics method (43). The amount of simulation time and the magnitude of the restricted-perturbation required to reach a low M20 loop backbone RMSD with respect to the target varied with each complex (Figure 34). As stated in the Methods section of this chapter, maximum-allowed perturbations were increased after 300 ps, and again after 400 ps. Although most complexes achieved conformational change within 300 ps, each was run for 400 ps. E occluded → closed (Figure 34B) clearly needed to be extended to 500 ps with increased perturbations.

No passes near either the X-ray open or new M20 loop conformers were observed in these simulations. In all cases, the M20 loop RMSD clearly shows the following concerning the simulation structure: it gets further away from the starting conformation, it approaches the target conformation and it does not get near the other conformers. Certain simulations involve decreases in the RMSD relative to the other conformers (Figures 34D, 34E, 34F, 34K and 34M), but in each case, the RMSD relative to the target conformer is clearly much lower. An example visual rendering (Figure 35) shows E|THF transitioning from the occluded starting conformer to the closed target conformer.

All forced conformational-change simulations involving the cofactor demonstrate behavior inconsistent with the experimental data. First, X-ray studies show that whenever the M20 loop is occluded, the cofactors nicotinamide ring cannot bind in the active site.

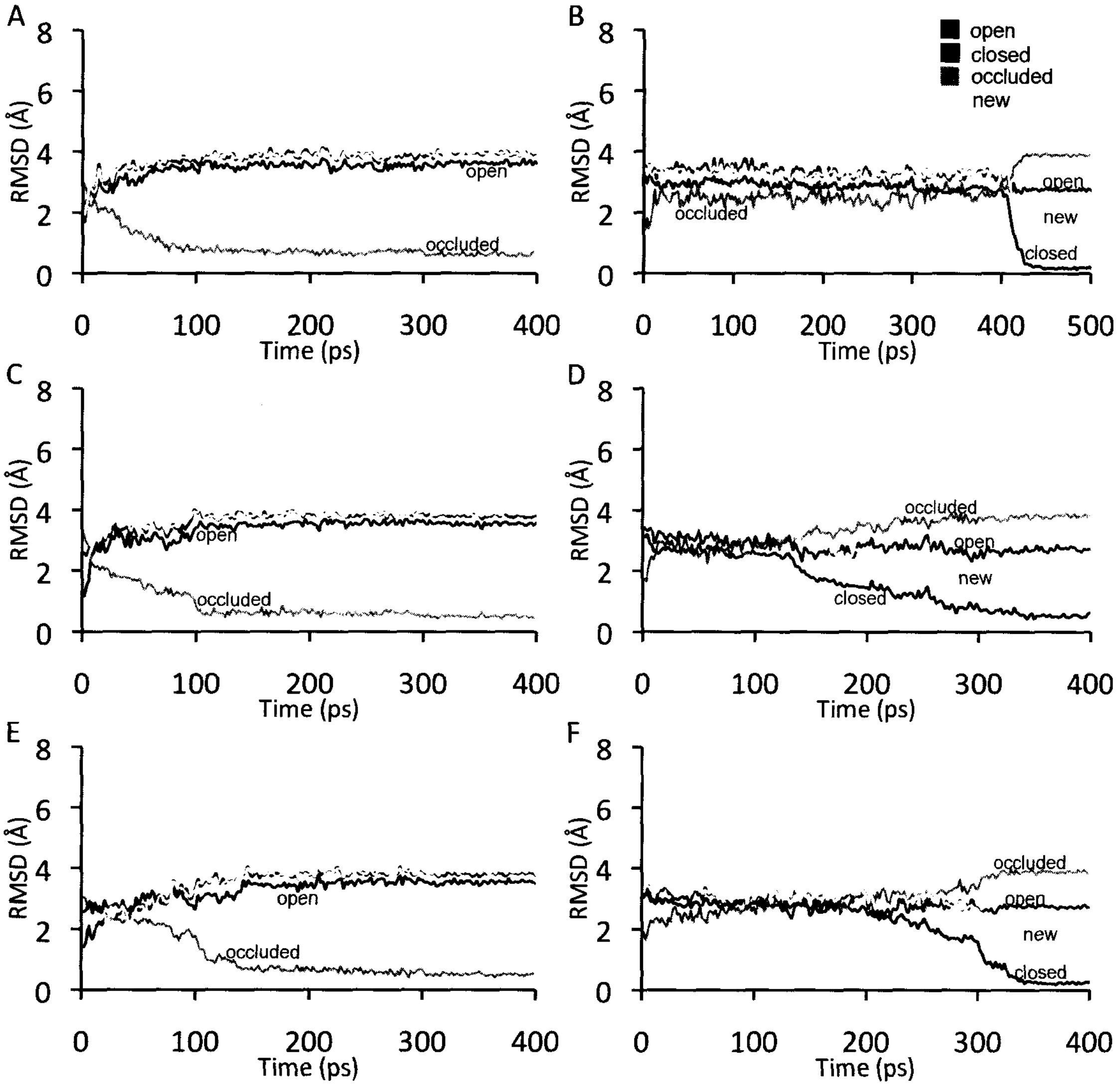
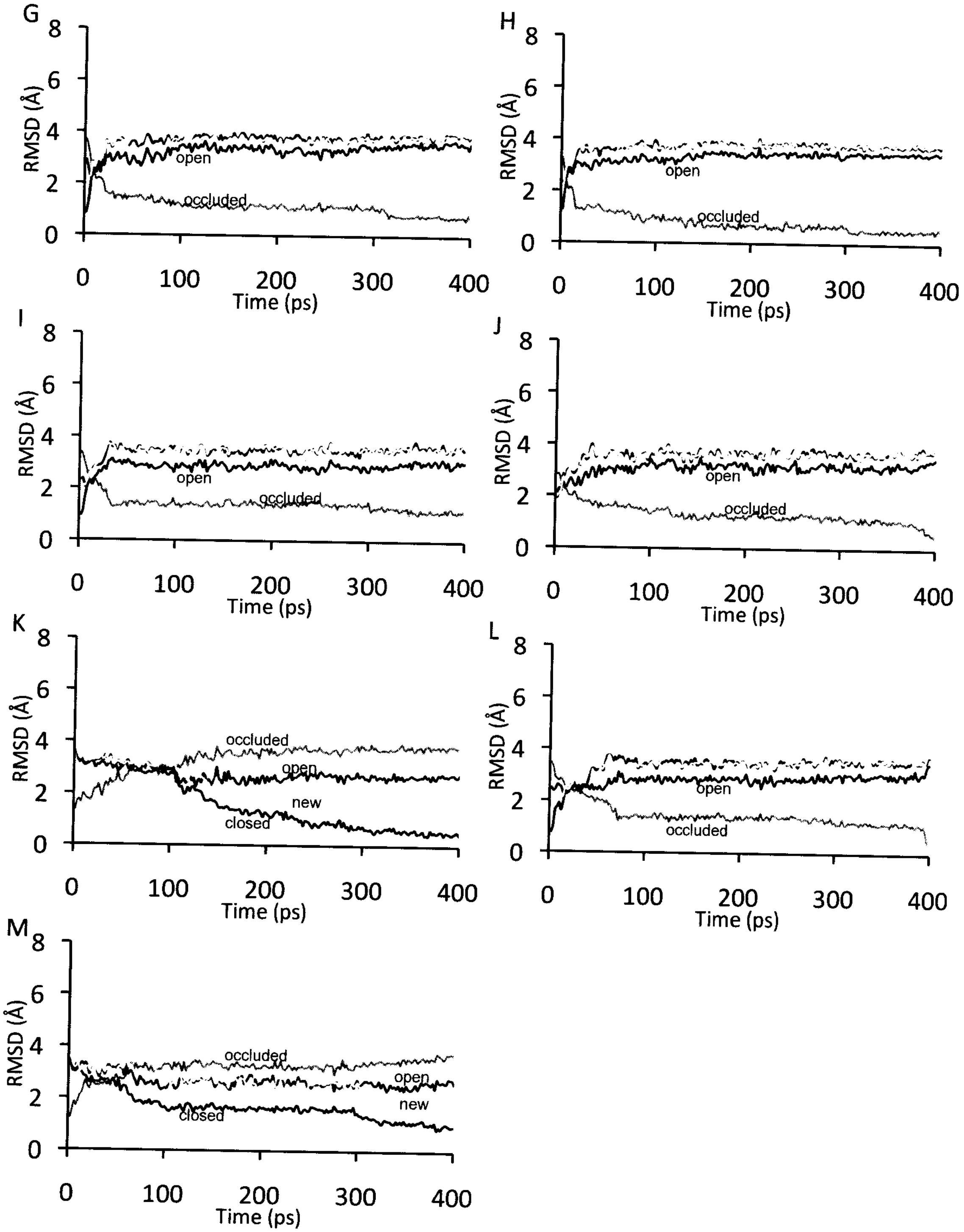


Figure 34. M20 loop backbone RMSD during forced conformational change simulations. Shown are E closed → occluded (A), E occluded → closed (B), E|folate closed → occluded (C), E|folate occluded → closed (D), E|DHF closed → occluded (E), E|DHF occluded → closed (F), E|NADPH closed → occluded (G), E|NADPH occluded → closed (H), E|DHF|NADPH closed → occluded (I), E|THF|NADP+ closed → occluded (J), E|THF|NADP+ occluded → closed (K), E|THF occluded → closed (L), E|THF|NADPH closed → occluded (M) and E|THF|NADPH occluded → closed (N). Simulation M20 loop backbone RMSD is plotted relative to the following structures: X-ray open (1RA1), X-ray closed (1RX1), X-ray occluded (1RX7) and new (NADPH open simulation).

Figure 34 Continued



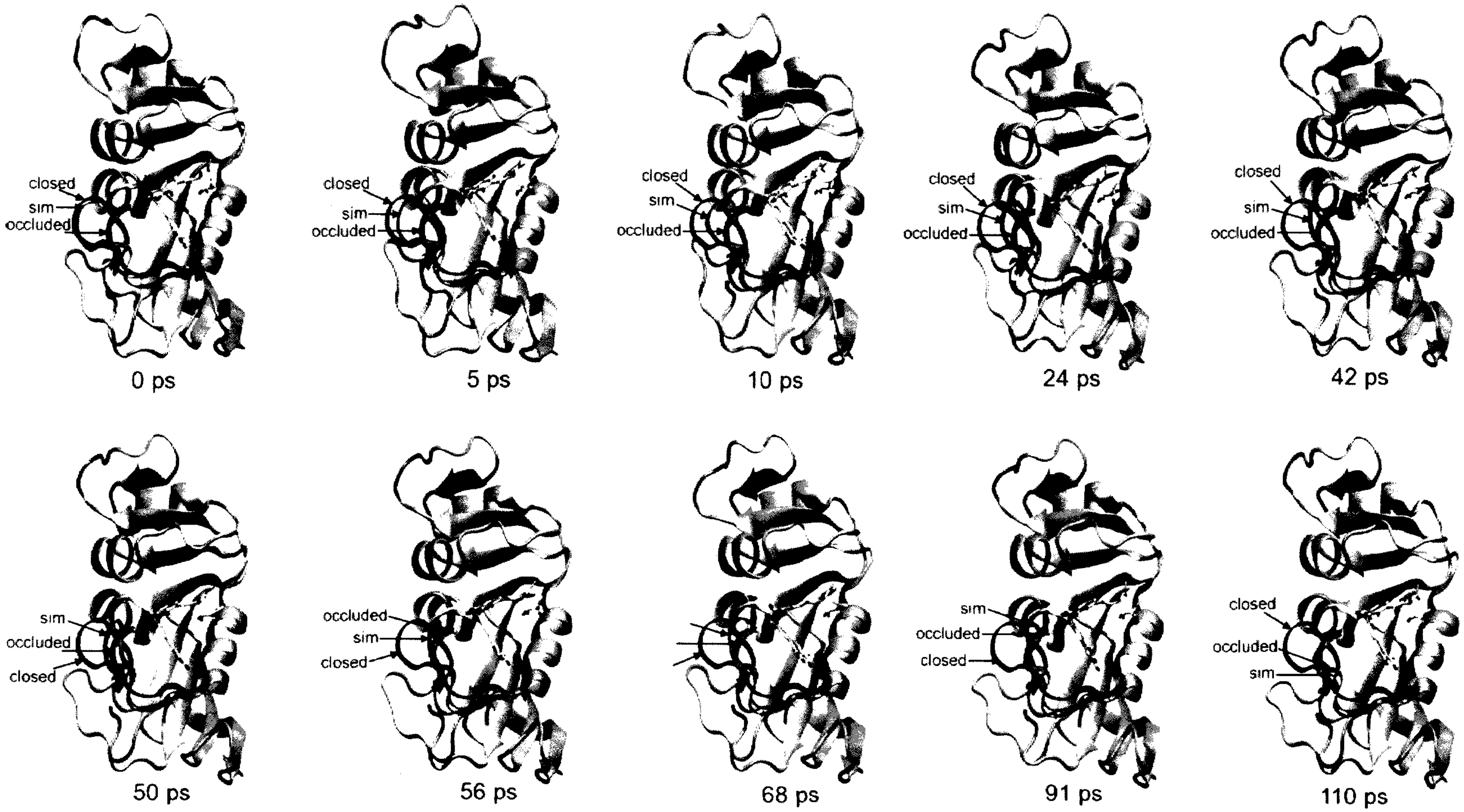


Figure 35. Closed-to-occluded conformational change in E|THF. M20 loops of superimposed X-ray closed (1RX1) and X-ray occluded (1RX7) structures are also drawn.

However, in each closed-to-occluded simulation, the cofactor is not expelled from the binding pocket to make way for the M20 loop. Instead as the M20 loop changes to the occluded conformation, it pushes the nicotinamide ring further into the binding pocket. This may result from the unnatural conditions produced from forcing the conformational change too quickly. *In vivo*, the enzyme allows time for random fluctuations in cofactor and loop motion to occur. Without allowing enough time for them in the simulation before introducing the perturbation, the cofactor is not given the chance to leave the active site before the occluding M20 loop enters the active site and blocks the exit. The equilibrium simulations (explained earlier) showed no such random fluctuations in cofactor movement occurring after 10 ns, indicating that the barrier for the nicotinamide leaving the active site is probably high. Also, X-ray studies show that if the cofactor is bound and the M20 loop is in the closed conformation, then the nicotinamide ring occupies the binding pocket. The occluded-to-closed simulations involving the cofactor start with the nicotinamide portion outside of the active site. But as the conformational change occurs, the nicotinamide ring remains outside. Again, there is probably not enough time *in silico* to allow the nicotinamide ring to enter the active site before inducing the change in the M20 loop conformation. As such, the restricted-perturbation method is unreliable when the cofactor is bound.

The normal modes were calculated for all starting structures studied in the forced conformational exchange simulations (a.k.a. the non-simulated structures) as well as several structures obtained from the equilibrium MD studies (a.k.a. the simulated structures). These modes describe the vibrational character of the DHFR structures. Here the modes are probed to find whether the structures exhibit tendencies to vibrate in the direction of an M20 loop conformational change. According to the NMR experiments (74,76) discussed in the Background section of this chapter, several catalytic intermediates sample excited state M20 loop conformers (Figure 9). So the ground state structures should contain vibrations pointing towards their respective excited states. Thus dot products of the mode vectors with the corresponding conformational difference vector (i.e. the excited state structure minus the ground state structure) should be larger than dot products with other conformational difference vectors. Using the normal modes for each structure, the dot product calculations (described in the Methods section of this chapter) were performed. Tables 9-13 list the dot product results for the 5 lowest frequency normal modes. The lowest frequency normal mode (which contains the largest atom position movements) is mode 7; modes 1-6 are only translations and rotations. Superpositions of normal modes are examined by dot product sums. Tables 14, 15 and 16 show the dot product sums of modes 7-11, 7-16 and 7-25, respectively.

Table 9. Normal mode dot products (mode 7)

Complex	M20 Loop[*]	open	closed	occluded	new[†]
DHFR	closed	0.025	-	0.015	0.012
	closed[‡]	0.027	-	0.043	0.012
	occluded	0.031	0.036	-	0.031
	occluded[‡]	0.005	0.038	-	0.008
DHFR | Folate	closed	0.013	-	0.025	0.001
	closed[‡]	0.002	-	0.001	0.003
	occluded	0.009	0.013	-	0.013
	occluded[‡]	0.010	0.013	-	0.030
DHFR | DHF	closed	0.019	-	0.018	0.008
	closed[‡]	0.047	-	0.044	0.048
	occluded	0.034	0.052	-	0.035
	occluded[‡]	0.008	0.002	-	0.008
DHFR | NADPH	closed	0.016	-	0.018	0.012
	closed[‡]	0.042	-	0.008	0.032
	new[†‡]	0.010	0.005	0.005	-
DHFR | DHF | NADPH	closed	0.002	-	0.009	0.000
DHFR | THF | NADP+	closed	0.019	-	0.014	0.002
DHFR | THF	closed	0.015	-	0.024	0.005
	occluded	0.010	0.005	-	0.013
DHFR | THF | NADPH	closed	0.028	-	0.032	0.039
	occluded	0.013	0.047	-	0.023

[*] M20 loop conformation of the complex for which the normal modes are calculated for

[†] NADPH-open minimized-average simulation structure.

[‡] Minimized-average structure from the equilibrium simulation.

Table 10. Normal mode dot products (mode 8)

Complex	M20 Loop[*]	open	closed	occluded	new[†]
DHFR	closed	0.008	-	0.025	0.001
	closed[‡]	0.009	-	0.003	0.003
	occluded	0.065	0.045	-	0.059
	occluded[‡]	0.040	0.051	-	0.037
DHFR | Folate	closed	0.006	-	0.010	0.001
	closed[‡]	0.027	-	0.030	0.015
	occluded	0.003	0.023	-	0.017
	occluded[‡]	0.031	0.016	-	0.024
DHFR | DHF	closed	0.000	-	0.017	0.010
	closed[‡]	0.015	-	0.026	0.030
	occluded	0.005	0.021	-	0.003
	occluded[‡]	0.033	0.016	-	0.031
DHFR | NADPH	closed	0.024	-	0.005	0.031
	closed[‡]	0.019	-	0.004	0.027
	new[†‡]	0.020	0.017	0.001	-
DHFR | DHF | NADPH	closed	0.016	-	0.020	0.004
DHFR | THF | NADP+	closed	0.018	-	0.006	0.025
DHFR | THF	closed	0.003	-	0.015	0.000
	occluded	0.047	0.020	-	0.031
DHFR | THF | NADPH	closed	0.010	-	0.010	0.008
	occluded	0.020	0.014	-	0.014

[*] M20 loop conformation of the complex for which the normal modes are calculated for

[†] NADPH-open minimized-average simulation structure.

[‡] Minimized-average structure from the equilibrium simulation.

Table 11. Normal mode dot products (mode 9)

Complex	M20 Loop[*]	open	closed	occluded	new[†]
DHFR	closed	0.003	-	0.019	0.021
	closed[‡]	0.047	-	0.020	0.004
	occluded	0.006	0.032	-	0.005
	occluded[‡]	0.002	0.023	-	0.012
DHFR | Folate	closed	0.023	-	0.014	0.042
	closed[‡]	0.007	-	0.015	0.011
	occluded	0.055	0.032	-	0.020
	occluded[‡]	0.008	0.012	-	0.008
DHFR | DHF	closed	0.024	-	0.002	0.035
	closed[‡]	0.019	-	0.019	0.005
	occluded	0.025	0.005	-	0.029
	occluded[‡]	0.007	0.005	-	0.008
DHFR | NADPH	closed	0.016	-	0.009	0.030
	closed[‡]	0.040	-	0.014	0.041
	new[†‡]	0.019	0.007	0.016	-
DHFR | DHF | NADPH	closed	0.008	-	0.003	0.031
DHFR | THF | NADP+	closed	0.030	-	0.003	0.027
DHFR | THF	closed	0.027	-	0.003	0.041
	occluded	0.073	0.028	-	0.038
DHFR | THF | NADPH	closed	0.011	-	0.037	0.016
	occluded	0.037	0.049	-	0.045

[*] M20 loop conformation of the complex for which the normal modes are calculated for

[†] NADPH-open minimized-average simulation structure.

[‡] Minimized-average structure from the equilibrium simulation.

Table 12. Normal mode dot products (mode 10)

Complex	M20 Loop[*]	open	closed	occluded	new[†]
DHFR	closed	0.026	-	0.004	0.021
	closed[‡]	0.016	-	0.041	0.013
	occluded	0.008	0.003	-	0.004
	occluded[‡]	0.016	0.015	-	0.017
DHFR | Folate	closed	0.019	-	0.013	0.006
	closed[‡]	0.013	-	0.013	0.012
	occluded	0.024	0.013	-	0.006
	occluded[‡]	0.025	0.010	-	0.016
DHFR | DHF	closed	0.012	-	0.005	0.028
	closed[‡]	0.041	-	0.046	0.020
	occluded	0.014	0.008	-	0.017
	occluded[‡]	0.014	0.008	-	0.007
DHFR | NADPH	closed	0.021	-	0.007	0.026
	closed[‡]	0.000	-	0.013	0.008
	new[†‡]	0.017	0.032	0.027	-
DHFR | DHF | NADPH	closed	0.002	-	0.009	0.035
DHFR | THF | NADP+	closed	0.010	-	0.000	0.022
DHFR | THF	closed	0.020	-	0.003	0.005
	occluded	0.022	0.023	-	0.024
DHFR | THF | NADPH	closed	0.014	-	0.026	0.007
	occluded	0.021	0.016	-	0.025

* M20 loop conformation of the complex for which the normal modes are calculated for

† NADPH-open minimized-average simulation structure.

‡ Minimized-average structure from the equilibrium simulation.

Table 13. Normal mode dot products (mode 11)

Complex	M20 Loop*	open	closed	occluded	new†
DHFR	closed	0.019	-	0.003	0.004
	closed‡	0.007	-	0.025	0.023
	occluded	0.038	0.016	-	0.031
	occluded‡	0.022	0.013	-	0.010
DHFR | Folate	closed	0.012	-	0.003	0.008
	closed‡	0.028	-	0.021	0.020
	occluded	0.028	0.011	-	0.020
	occluded‡	0.002	0.013	-	0.012
DHFR | DHF	closed	0.037	-	0.006	0.006
	closed‡	0.026	-	0.038	0.023
	occluded	0.002	0.002	-	0.002
	occluded‡	0.069	0.017	-	0.071
DHFR | NADPH	closed	0.017	-	0.004	0.005
	closed‡	0.000	-	0.002	0.002
	new†‡	0.004	0.006	0.008	-
DHFR | DHF | NADPH	closed	0.026	-	0.007	0.009
DHFR | THF | NADP+	closed	0.000	-	0.005	0.002
DHFR | THF	closed	0.002	-	0.006	0.010
	occluded	0.047	0.000	-	0.008
DHFR | THF | NADPH	closed	0.015	-	0.002	0.035
	occluded	0.016	0.021	-	0.020

* M20 loop conformation of the complex for which the normal modes are calculated for

† NADPH-open minimized-average simulation structure.

‡ Minimized-average structure from the equilibrium simulation.

Table 14. Normal mode dot product sums (modes 7 to 11)

Complex	M20 Loop*	open	closed	occluded	new†
DHFR	closed	0.031	-	0.060	0.007
	closed‡	0.035	-	0.040	0.018
	occluded	0.071	0.061	-	0.060
	occluded‡	0.050	0.111	-	0.051
DHFR | Folate	closed	0.009	-	0.045	0.044
	closed‡	0.033	-	0.053	0.010
	occluded	0.101	0.028	-	0.024
	occluded‡	0.009	0.018	-	0.002
DHFR | DHF	closed	0.006	-	0.014	0.004
	closed‡	0.118	-	0.121	0.056
	occluded	0.041	0.063	-	0.051
	occluded‡	0.073	0.028	-	0.125
DHFR | NADPH	closed	0.028	-	0.021	0.093
	closed‡	0.063	-	0.033	0.037
	new†‡	0.037	0.010	0.007	-
DHFR | DHF | NADPH	closed	0.055	-	0.030	0.018
DHFR | THF | NADP+	closed	0.019	-	0.022	0.077
DHFR | THF	closed	0.009	-	0.003	0.051
	occluded	0.033	0.021	-	0.039
DHFR | THF | NADPH	closed	0.036	-	0.103	0.057
	occluded	0.049	0.018	-	0.041

* M20 loop conformation of the complex for which the normal modes are calculated for

† NADPH-open minimized-average simulation structure.

‡ Minimized-average structure from the equilibrium simulation.

Table 15. Normal mode dot product sums (modes 7 to 16)

Complex	M20 Loop*	open	closed	occluded	new†
DHFR	closed	0.117	-	0.107	0.043
	closed‡	0.051	-	0.057	0.032
	occluded	0.068	0.055	-	0.041
	occluded‡	0.025	0.064	-	0.026
DHFR | Folate	closed	0.044	-	0.003	0.041
	closed‡	0.027	-	0.005	0.042
	occluded	0.150	0.036	-	0.012
	occluded‡	0.033	0.018	-	0.010
DHFR | DHF	closed	0.062	-	0.002	0.000
	closed‡	0.111	-	0.144	0.071
	occluded	0.012	0.043	-	0.031
	occluded‡	0.152	0.072	-	0.191
DHFR | NADPH	closed	0.044	-	0.071	0.118
	closed‡	0.084	-	0.008	0.081
	new†‡	0.057	0.038	0.013	-
DHFR | DHF | NADPH	closed	0.059	-	0.022	0.008
DHFR | THF | NADP+	closed	0.061	-	0.007	0.080
DHFR | THF	closed	0.005	-	0.003	0.045
	occluded	0.043	0.016	-	0.056
DHFR | THF | NADPH	closed	0.045	-	0.044	0.004
	occluded	0.059	0.054	-	0.060

* M20 loop conformation of the complex for which the normal modes are calculated for

† NADPH-open minimized-average simulation structure.

‡ Minimized-average structure from the equilibrium simulation.

Table 16. Normal mode dot product sums (modes 7 to 26)

Complex	M20 Loop*	open	closed	occluded	new†
DHFR	closed	0.072	-	0.094	0.031
	closed‡	0.121	-	0.026	0.148
	occluded	0.120	0.067	-	0.077
	occluded‡	0.069	0.018	-	0.033
DHFR | Folate	closed	0.192	-	0.031	0.122
	closed‡	0.016	-	0.032	0.004
	occluded	0.067	0.223	-	0.064
	occluded‡	0.134	0.066	-	0.022
DHFR | DHF	closed	0.028	-	0.019	0.054
	closed‡	0.044	-	0.017	0.072
	occluded	0.095	0.174	-	0.159
	occluded‡	0.118	0.001	-	0.139
DHFR | NADPH	closed	0.022	-	0.051	0.026
	closed‡	0.072	-	0.041	0.020
	new†‡	0.032	0.034	0.038	-
DHFR | DHF | NADPH	closed	0.018	-	0.019	0.024
DHFR | THF | NADP+	closed	0.005	-	0.008	0.031
DHFR | THF	closed	0.076	-	0.020	0.104
	occluded	0.010	0.047	-	0.038
DHFR | THF | NADPH	closed	0.054	-	0.084	0.001
	occluded	0.089	0.014	-	0.056

* M20 loop conformation of the complex for which the normal modes are calculated for

† NADPH-open minimized-average simulation structure.

‡ Minimized-average structure from the equilibrium simulation.

The validity of the normal mode dot product method described above is determined by comparing its calculation results with experimental data. The NMR studies discussed in the Background section of this chapter provide three different validation tests. First, E|DHF|NADPH|closed is observed by NMR to populate an excited state M20 loop conformer that closely resembles the occluded X-ray conformer (76). So the dot products of its low-frequency mode vectors with the closed-to-occluded conformational difference vector should be larger than the dot products with the closed-to-open or closed-to-new vectors. Second, E|THF|NADPH|occluded populates an excited state conformer not resembling closed (76). So its low-frequency modes should not have the largest dot products with the occluded-to-closed vector. Third, E|NADPH|closed populates an excited state conformer not resembling occluded (76). So its low frequency modes should not have the largest dot products with the closed-to-occluded vector.

More consideration is given to the first validation test because it involves an event that is less unlikely to occur randomly than the events of the second or third tests. The first test has three possible results: the closed-to-occluded dot product is the largest, the closed-to-open dot product is the largest or the closed-to-new dot product is the largest. So there is only a 33% chance of passing the first test. Similar reasoning shows that there is a 66% chance of passing the second test and a 66% chance of passing the third test.

To pass the first validation test, the dot products of the low-frequency mode vectors of E|DHF|NADPH|closed with the closed-to-occluded conformational difference vector must be larger than those with the closed-to-open or closed-to-new vectors. Out of the five lowest frequency modes, only modes 7 and 8 pass the test. Mode 7 has a 0.009 value with occluded, but only 0.002 and 0.000 values with open and new, respectively (Table 9). Mode 8 has a 0.020 value with occluded, but only 0.016 and 0.004 values with open and new, respectively (Table 10). Also, the larger 0.020 value of mode 8 indicates that mode to vibrate more along the closed-to-occluded direction than mode 7. Although modes 7 and 8 favor the occluded direction, modes 9, 10 and 11 do not. In fact, the dot product sums, which describe modal superposition, show that modes 7 to 11, together, vibrate more towards the open direction (0.55) than towards the occluded (0.030) or new (0.018) directions (Table 14). The superposition of modes 7 to 16 (Table 15) exhibits a similar trend, but the superposition of modes 7 to 26 (Table 16) shows the closed-to-new sum (0.024) is larger than the closed-to-open (0.018) and closed-to-occluded (0.019) sums. Thus, the calculations do not strongly support the first validation test.

To pass the second validation test, the dot products of the low-frequency mode vectors of E|THF|NADPH|occluded with the occluded-to-closed conformational difference vector must not be largest. Calculations on the individual normal modes 7 to 11 show that three out of the five modes (7, 9 and 11) do have the largest dot products with the occluded-to-closed vector (Tables 9-13). However, the dot product sums of modes 7 to 11 are 0.049 for occluded-to-open, 0.041 for occluded-to-new and only 0.018 for occluded-to-closed (Table 14). Thus the superposition of modes 7 to 11 vibrates more along the occluded-to-open and occluded-to-new directions than along the occluded-to-closed direction. The dot product sums of modes 7 to 16 show much less variation: 0.059 for occluded-to-open, 0.060 for occluded-to-new and 0.054 for occluded-to-closed (Table 15). But the dot product sums of modes 7 to 26 (like the sum of modes 7 to 11) show the occluded-to-closed direction (0.014) to be less favored than the occluded-to-open (0.089) or the occluded-to-new (0.056) directions (Table 16). Thus calculations do mostly support the second validation test.

To pass the third validation test, the dot products of the low-frequency mode vectors of E|NADPH|closed with the closed-to-occluded conformational difference vector must not be largest. E|NADPH|closed was one of the structures studied by equilibrium molecular dynamics simulations, so it makes sense to examine the normal modes of both native and simulated structures. Calculations on both structures do, in fact, show smaller dot products of mode vectors 7 to 11 of E|NADPH|closed (shown in Tables 9 to 13, respectively) with the closed-to-occluded vector than with the closed-to-open or closed-to-new vectors. For example, mode 8 of the simulated structure has dot product values of 0.019 and 0.027 with the open X-ray and new conformers, respectively, but only 0.004 with the occluded X-ray conformer. The same dot products of the modes from the non-simulated structure are 0.024 and 0.031 with the open X-ray and new conformers, respectively, and 0.005 with the occluded X-ray conformer. Mode 9 follows a similar trend, that is, the dot products in both the simulated and non-simulated E|NADPH|closed structures are lower with the closed-to-occluded vector than with the closed-to-open or closed-to-new vectors. Mode 7 of the simulated structure exhibits the trend, and mode 10 of the non-simulated structure does as well. Mode 11 does not exhibit the trend for either. Thus, some of the individual low frequency modes of E|NADPH|closed do vibrate towards M20 loop conformations other than occluded.

The vibrational trend of E|NADPH|closed is, perhaps, better seen with an examination of the dot product sums. The sum of modes 7 to 11 (Table 14) in the simulated structures shows a larger tendency for them to vibrate more along the closed-to-open vector (0.063) than along the closed-to-occluded (0.033) or closed-to-new (0.037) vectors. In contrast, the non-simulated structure has the largest value with the closed-to-new vector (0.093). Both structures show the smallest values with the closed-to-occluded vector, which indicates that the vibrational tendency of these modes is not towards occluded. Modes 7 to 16 (Table 15) in the simulated structure strongly favor vibration along the closed-to-open (0.084) and closed-to-new (0.081) directions as opposed to the closed-to-occluded direction (0.008). These same modes in the non-simulated structure vibrate more along the closed-to-new vector (0.118) than the closed-to-occluded (0.071) or the closed-to-open (0.044) vectors. Modes 7 to 26 (Table 16) in the simulated structure vibrate more along the closed-to-open vector (0.072) than the closed-to-occluded (0.041) or closed-to-new (0.020) vectors. Modes 7 to 26 in the non-simulated structure vibrate more along the closed-to-occluded direction (0.051) than the closed-to-open (0.022) or closed-to-new (0.026) directions. Thus, incorporating modes 17 to 26 shifts the vibrational tendency of the non-simulated structure to be more along the closed-to-occluded vector than the other two conformational difference vectors. Evidence supporting the third validation test is not very strong.

Although the normal mode analysis method passed certain parts of the validation tests, the evidence is not strong enough to conclude it as very good (i.e. at predicting vibrational tendencies of the M20 loop towards certain defined conformational changes). This is not surprising since in order for conformational change to occur, the structure may need to move further away from the target conformation before getting closer to it, whereas the normal modes describe only vibrational tendencies in the immediate vicinity of a structure. Molecular motion proceeds according to various constraints (e.g. dihedral angle rotations), which normal modes do not fully consider.

CONCLUSIONS

The open, closed and occluded X-ray M20 loop conformations of Escherichia coli dihydrofolate reductase were examined. The closed and occluded conformers stayed close to their native conformations in several complexes through the help of hydrogen bonds within the M20 loop and between the M20 loop and neighboring loops. Loss of these hydrogen bonds was accompanied by corresponding shifts in structure. This was most apparent when simulating the open X-ray conformer; each case involved a loss of a pair of interactions connecting the M20 loop to the GH loop resulting in a similar shift in conformation. Structural analysis showed this shift to be different enough from the three X-ray structures to name it a new M20 loop conformation.

Ligand dissociations, which were simulated by pulling on certain atoms, revealed that in five of the six occurrences, as the cofactor leaves the binding pocket from the closed conformer, it pushes its way past the M20 loop, causing the loop to change structure. In all five cases, this change resembles the new conformation observed in the equilibrium simulations. The M20 loops in these five near-pass structures were similar in conformation to each other as well as to the new conformation. Substrate dissociation simulations revealed no such sampling by the M20 loop.

Certain methods were seen to be unreliable for examining the hypotheses. The restricted-perturbation targeted molecular dynamics method used to produce the conformational change failed to expel the cofactors nicotinamide ring from the active site as the M20 loop was forced into the occluded conformation. This directly contradicts crystallography studies, which show that the nicotinamide ring lies tethered outside the active site whenever the loop is occluded. Likewise in contradiction, occluded-to-closed transitions failed to bring the tethering nicotinamide ring into the binding site. The MMPBSA energetic analysis failed to calculate the relative free energies of the different M20 loop conformers of various equilibrium DHFR simulation structures. This was immediately apparent, since the estimated population of the folate-bound open M20 loop conformer was less than 0.0000001%, whereas experimental NMR studies predict it to be roughly 10%. Finally, the dot products between the low-frequency normal mode eigenvectors of a complex and the four conformational change difference vectors failed to identify tendencies of the complex for vibration towards any one specific conformational change.

Future research should be done to see if different free energy analysis methods (e.g. umbrella sampling or free energy perturbation) are better at calculating the relative free energies of the various M20 loop conformers of dihydrofolate reductase than the MMPBSA method. Also, while the restricted-perturbation targeted molecular dynamics method provided interesting results, other path-determining methods (108) might furnish better mappings of the closed-occluded transition. Also, recently developed vibrational methods (109-110) that are able to calculate the pathway of conformational change between two given protein structures should be tried.

CHAPTER IV

CONSERVED CONTACT NETWORKS OF GREEK-KEY PROTEINS

BACKGROUND

The topological fold of a protein is the overall global placement of the polypeptide backbone with respect to itself, while disregarding sidechains and secondary structure. Although the general subject of topology is vast, a few examples should illustrate the essential ideas pertaining to protein folds. Figure 36 shows three sets of equivalent different topological folds or motifs: the trefoil knot (Figure 36A), the V-shaped bend (Figure 36B) and the straight line (Figure 36C). The two different proteins containing trefoil knots, E. coli YBEA (Figure 36A left) and RNA 2'-O-Ribose Methyltransferase (Figure 36A center), have different α-helix lengths, but equivalent topologies. Independence of secondary structure is illustrated by topologically equivalent V-shaped motifs in both the all α-helix human death domain (Figure 36B left) and the all β-sheet human titin (Figure 36B center).

This structural ambiguity reduces the vast 3-dimensional fold space into a smaller topological fold space, more manageable in size (111). Classifying proteins as such departs from other popular protein classification schemes, such as CATH (112,113), SCOP (114), FSSP (115) and DALI (116), in that it completely disregards both secondary structure and evolutionary sequence heredity. Instead it emphasizes commonality between proteins with equivalent topology but different secondary structure (111,117). One such commonality is the conserved residue long-range interaction network. Network theory (or graph theory) is an important tool for describing any system of interconnected items (or nodes), such as the world-wide web, electrical power grids, the national highway system and biological systems (118-120). In a protein network, residues are the nodes and long-range interactions the connections (121). Greene and Higman (117) have proposed that evolutionarily conserved long-range residue interaction networks both govern and stabilize the native topology of a protein. The proving ground used for testing this hypothesis is the Greek-Key topology, a common fold found among many proteins that vary in both secondary structure and function. The name 'Greek-Key' derives from its resemblance to the Greek-Key meander design in art and architecture.

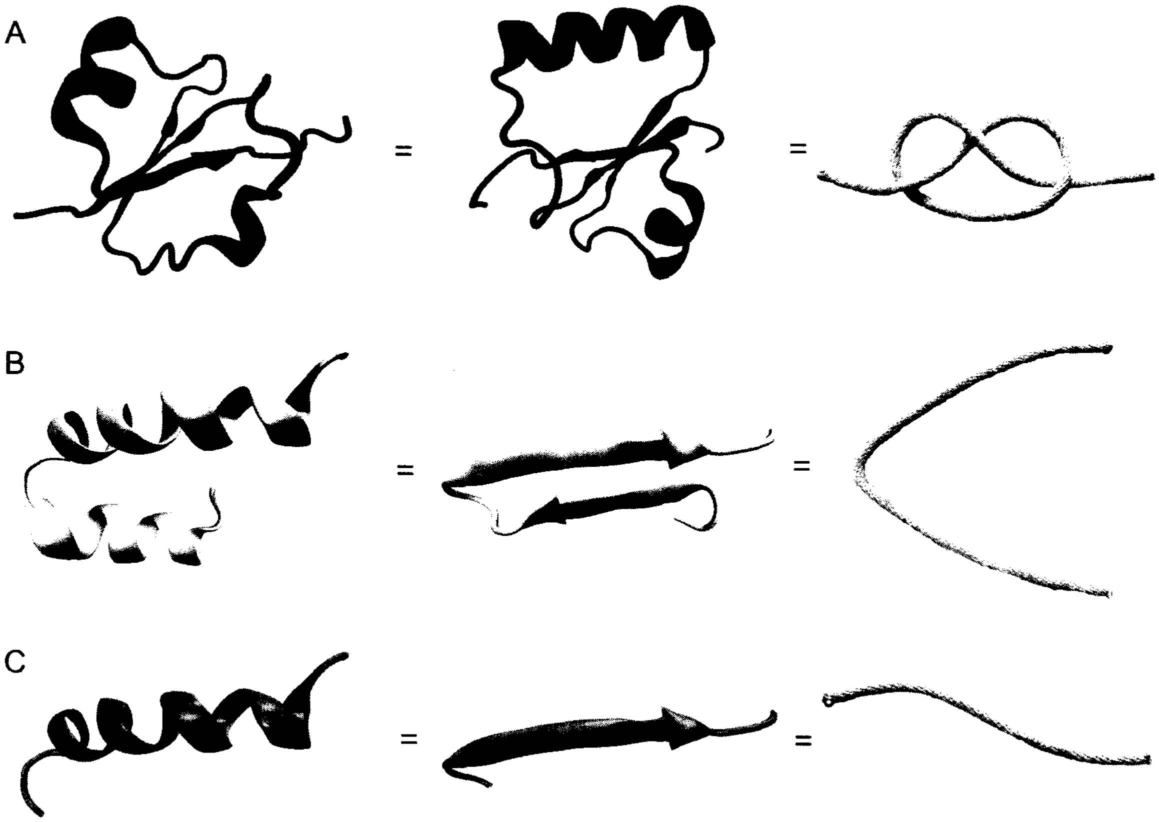


Figure 36. Equivalent protein folding motifs. Shown above are the trefoil knot (A), the V-shaped bend (B) and the straight line (C). The left trefoil knot is from residues 67 to 123 of E. coli YBEA (PDB code 1NS5). The center trefoil knot is from residues 188 to 239 (PDB code 1IPA). The left V-shaped bend is from residues 122 to 154 of human death domain (PDB code 1E3Y). The right V-shaped bend is from residues 45 to 62 of human titin (PDB code 1TIT). The left straight line is from residues 135 to 154 of human Fas-associated death domain (PDB code 1E3Y). The center straight line is from residues 54 to 62 of human titin (PDB code 1TIT). The rope images on the right are photographs taken by the author.

The general procedure designed for testing the hypothesis follows five steps, in which part of the work in this dissertation composes the final step. First, three sets of protein domains are selected to represent three different superfamilies, the all α–helix death domains, the α/β-plaits and the all β-sheet immunoglobulins (Table 17). Although these differ in secondary structure, they share a common Greek-Key fold (Figure 37). Second, find the native contacts in each selected protein. A contact occurs when heavy atoms from two residues, separated by at least eight other residues, are within a certain cutoff distance of each other. This cutoff distance is 7 Å for both the death domains and the α/β-plaits, but only 6 Å for the immunoglobulins since they are composed solely from the more tightly-packing β-sheet secondary structure units. The PERL programs for calculating the native contacts were written by Ahmed and Greene (unpublished data). An additional BASH program, which (exactly) reproduces the results of the PERL programs, was written (for purposes of smooth operability between it and other analysis programs) by the author, for use in the fifth step below.

Third, the proteins within each set are structurally aligned by Greene using the Combinatorial Extension program (122) and refined with Monte Carlo simulations (unpublished data). Then, a sequence alignment list is generated by grouping together residues from different proteins that occupy equivalent structural positions. Sequence gaps occur when a protein contains no residue at the corresponding equivalent position. Table 18 shows an example sequence alignment list for the death domain superfamily. Structurally equivalent positions correspond with the table columns. Sequence gaps correspond with dashes. Apart from truncation of N-terminus and/or C-terminus residues, the entire connected sequences from each protein are contained in these alignment lists. In other words, removing the gaps the original sequence.

In the fourth step, the conserved contact networks are generated by Pothen and Greene via one of two methods. In the direct method, the conserved network is generated

Table 17. PDB codes of selected superfamily protein sets.*

Death Domains	α/β-Plaits	Immunoglobulins
1E3Y	1Q5Y:A	1TIT
1DDF	1RIS	1WIT
1D2Z:A	1UOS:A	1TLK
1N3K:1	2ACY	2VAA:B
1C15:1	1URN:A	3CD4
1UCP:1	1B7F:A	1CQK:A
1DGN:1	1GH8:1	1TEN
1A1W	1RKJ:A:1	1G84:1
1CRD:1	2AW0:1	1HE7
		1HNG:A

* *Letters/numbers after colon represent chain ID/frame number*

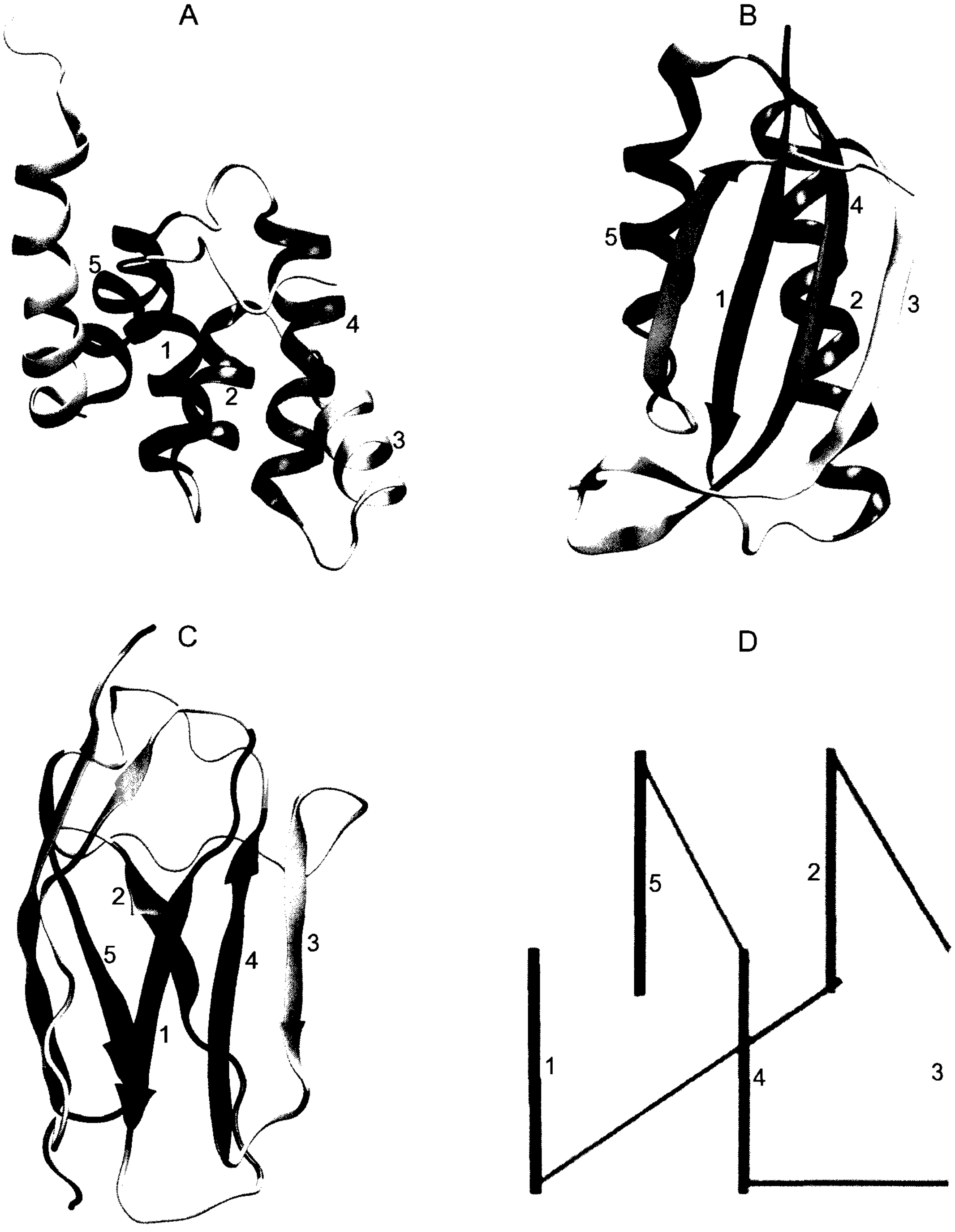


Figure 37. Proteins sharing a Greek-Key topology. Proteins shown are human Fas-associated death domain - 1E3Y (A), ribosomal S6 - 1RIS (B) and human titin - 1TIT (C). A generic stick representation of Greek-Key topology is shown in D (117). The five equivalent topological segments are numbered 1 to 5, starting from the N-terminus.

Table 18. Sequence alignment list for the death domains.*

1E3Y	SHMGEEDLCAAFNVICDNV-G-KDWRRLARQL------
1DDF	------LSKYITTIAGVMT-L-SQVKGFVRKN------
1D2Z	MAIRLLPLPVRAQLCAHLD-ALDVWQQLATAV------
1N3K	----AEYGTLLQDLTNNIT-L-EDLEQLKSACKE----
1C15	---DAKARNCLLQHREALEKD-IKTSYIMDHM------
1UCP	----GRARDAILDALENLT-A-EELKKFKLKLLSVPLR
1DGN	--------KKRRIFIHSVG-A-GTINALLDCL------
1A1W	---MDPFLVLLHSVSSSLS-S-SELTELKYLCLG----
1CRD	----KQVLRSLRLELGAEV-L-VE-GLVLQYL------

1E3Y	----K----VS------DTKIDSIEDRYPRNLTERVRE
1DDF	----G----VN------EAKIDEIKNDNVQDTAEQKVQ
1D2Z	----K----LY------PDQVEQISSQKQRGRS-ASNE
1N3K	----D----IP------SEKSEEI----------TTGSA
1C15	----ISDGFLT------ISEEEKVRN-EPTQQQRAAML
1UCP	EGYGR----IP----------RGALLS-------MDALD
1DGN	----LEDEVIS------QEDMNKVRDEN-DTVMDKARV
1A1W	----R----VG------KRKLERV----------QSGLD
1CRD	----Y----QEGILTENHIQEINAQT----TGLRKTML

1E3Y	SLRIWK-NTEKENATVAHL---VGALRSCQMNLVADLV
1DDF	LLRNWH-QLHGKKEAYDTL---IKDLKKANLCTLAEKI
1D2Z	FLNIWG-GQYN--HTVQTL---FALFKKLKLHNAMRLI
1N3K	WFSFLESHNKLDKDNLSYI---EHIFEISRRPDLLTMV
1C15	IKMILK-K--DNDSY-VSF---YNALLHEGYKDLAALL
1UCP	LTDKLV-SFYLETYGAELT---ANVLRDMGLQEMAGQL
1DGN	LIDLVT-GKGPKSCC--KF---IKHLCE-EDPQLASKM
1A1W	LFSMLL-EQN--DLEPGHTELLRELLASLRRHDLLRRV
1CRD	LLDILP-SRGPKAF--DTF---LDSL--QEFPWVREKL

1E3Y	QEVQQARDLQNRS
1DDF	QTIILKDITS---
1D2Z	KDYVSEDLH----
1N3K	VDYRTRVLKISEE
1C15	HDGI---------
1UCP	QAATHQ-------
1DGN	GL-----------
1A1W	DDFE---------
1CRD	KKAREEAMTD---

* Vertical columns indicate corresponding sequence positions between the aligned proteins. Dashes indicate gaps in the chain (i.e. removing them produces the actual sequence).

by examining each native contact of the first protein in the sequence alignment list. If both residues forming a contact are in the alignment, then the corresponding slots (i.e. structurally equivalent positions) of the second protein are checked. If it also contains residues in those slots, and if they form a contact, then the programs moves on to the third protein. Similarly, all proteins in the alignment list are checked to see if they contain the contact. If so, the contact is considered conserved. Doing this for each native contact in the first protein generates the conserved network. The toggle method (Figure 38), introduced by Higman and Greene (117,121), is applied in this work. Toggling takes into account potential secondary structural stabilization. For example, suppose that all but one protein contain a contact and closer inspection of the corresponding two residues in the rogue protein shows that although the first residue does not contact the second, it does

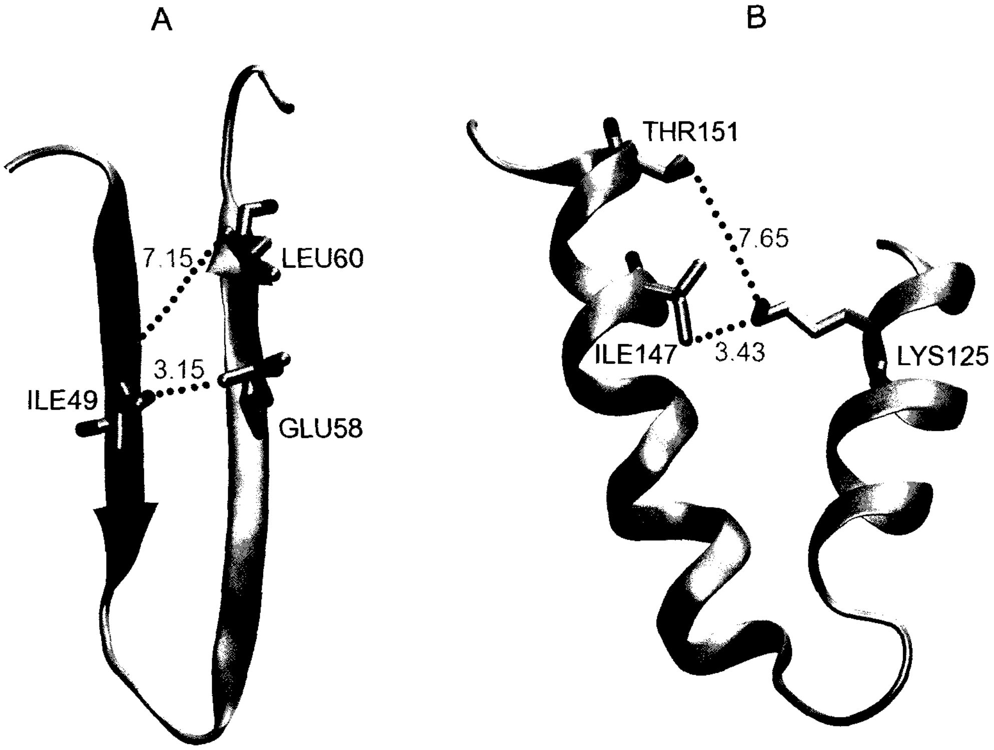


Figure 38. Direct versus Toggle contacts. In the all β-sheet human immunoglobulin - 1TIT (A), there is no direct contact between ILE49 and LEU60, but toggling down 2 residues in the β-strand reveals one between ILE49 and GLU58. In the all α-helix human Fas-associated death domain - 1E3Y (B), there is no direct contact between LYS125 and THR151, but toggling up 4 residues in the helix reveals one between LYS125 and ILE147. Contact distances are in angstroms.

contact a residue near the second. This may occur if the second residue lies on a β-sheet, since the side chains of every other residue in the β-strand occupy similar positions (Figure 38A). Similarly, the side chains of every forth residue in an α-helix point in the same direction (Figure 38B). Toggling allows for these small sequence shifts within secondary structure units, and hence more natural variability within biological structures. Contacts related in such a way are still considered equivalent. So in the case of the rogue protein described above, the contact would be conserved. A third network is made from this by deleting any contacts containing acidic or basic residues (i.e. arginine, aspartic acid, glutamic acid or lysine). This allows the final network to contain only hydrophobic interactions, removing potential hydrophilic interactions with the solvent.

The three conserved networks each differ in size; Direct (D), Toggle (T) and Toggle with no acidic/basic residues (TN). The T network is a superset of both of the other two networks. The D network would be a subset of the TN network if none of the proteins in the alignment contained any acidic/basic residues (R, D, E or K) among their direct conserved interactions, however, this situation does not apply to the three superfamilies being studied. Despite this, the D networks are still much smaller than the TN networks. For example, the D network of the death domains contains just 4 interactions, whereas the corresponding TN network contains 14 (Figure 39). The entire

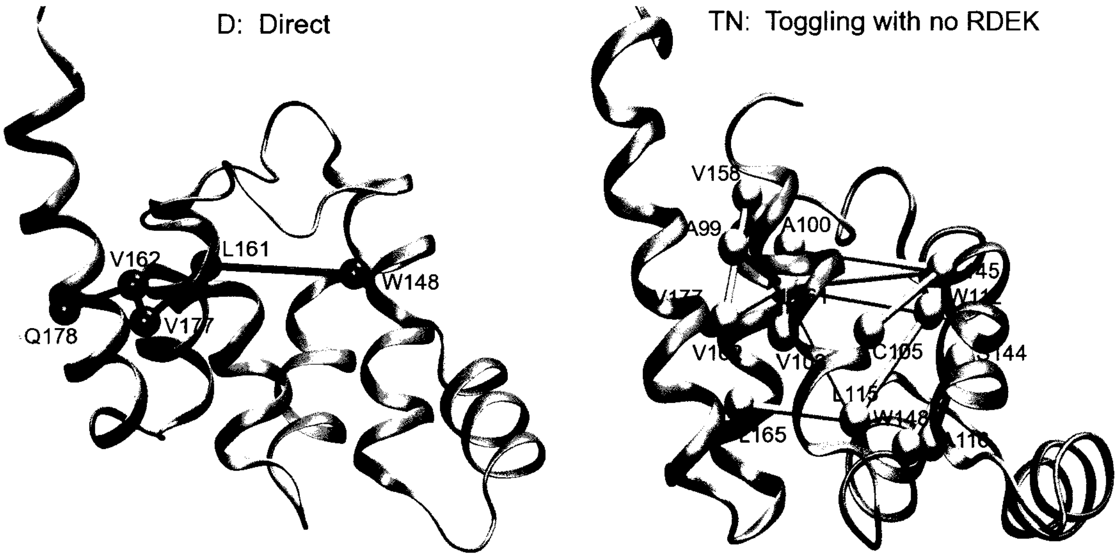


Figure 39. D and TN conserved networks in human Fas-associated death domain (1E3Y).

conserved networks are shown in Tables 19-21 for the all α-helix human Fas-associated death domain, the α/β-plait ribosomal S6 and the all β-sheet human titin, respectively. Pothen and Greene generated this data (unpublished data). Counting shows that the smallest network is the D network of the death domains, which contains 4 interactions. The largest is the T network of the α/β-plaits, which contains 109 interactions.

Recent computational studies have identified (using a different protocol to the one used in the present study) a list of sixteen conserved residues in the death domains (123,124). Experimental studies have examined the folding of human Fas-associated death domain in the death domain superfamily containing mutations at conserved and nonconserved sites (123). The conservation of tryptophan made Trp112 and Trp148 interesting sites for mutation. In the present study, Trp112 is not found in the D network, but the T network contains Trp112-Ser144 and Trp112-Val141 interactions (the TN network only has Trp112-Ser144). Trp148 is found in all 3 networks of the present study (Trp148-Leu161 in the D network, and Leu115-Trp148 and Trp148-Leu161 in both the T and TN networks). The reduction in sidechain size in the Trp148Phe mutant resulted in large destabilizing effects on the protein structure. Trp112Phe had lesser effects, but the rate of folding was significantly decreased. The non-conserved His160Gly mutant showed hardly any effect on the structure or folding kinetics. Combining the conserved Trp112Phe with the non-conserved His160Gly in a double mutant resulted in little additive effect, even though the conserved Trp148 has a long-range interaction with His160. This experimental study shows significant progress towards understanding how conserved residue interaction networks stabilize the structure and folding of proteins.

The fifth step in testing the hypothesis, is performing MD simulations on an example protein from each superfamily. These three proteins (human Fas-associated death domain, ribosomal S6 and human titin) are interrogated under high-temperature conditions to see if their conserved interactions do, in fact, stabilize their Greek-key topologies. Higher temperature means more kinetic energy, hence more residue shaking. If the conserved interactions do contain some inherent stability over other non-conserved native contacts, then that should be apparent from the simulations. Molecular dynamics is used to make this determination. The chosen proteins from each superfamily are simulated at several different temperatures, the highest being enough to denature them.

Table 19. Conserved networks for human Fas-associated death domain (1E3Y)*

D	148	161	162	162													
	161	177	177	178													
T	99	99	100	101	101	101	102	102	103	103	103	103	104	104	105	105	105
	161	176	145	141	142	145	142	176	141	161	173	176	141	173	141	142	145
	106	107	109	111	111	112	112	113	113	115	115	115	115	115	116	125	145
	141	173	141	141	165	141	144	140	141	144	148	161	164	165	144	147	161
	148	158	158	158	161	162	162	162	163	163							
	161	177	178	180	177	177	178	180	177	178							
TN	99	100	103	105	112	115	115	115	116	145	148	158	161	162			
	161	145	161	145	144	148	161	165	144	161	161	177	177	177			

* Contacts are indicated between by pairs of vertically-separated sequence numbers.

Table 20. Conserved networks for ribosomal S6 (1RIS)

D	4	4	5	5	6	6	6	6	7	7	8	8	8	9	9	9	37
	63	64	62	63	61	62	63	89	61	88	59	60	87	59	86	87	63
	37	37	38	39	39	40											
	64	65	64	62	63	62											
T	2	3	3	4	4	4	4	5	5	5	5	5	6	6	6	6	6
	65	64	65	62	63	64	65	62	63	64	65	89	61	62	63	64	65
	6	6	7	7	7	7	7	7	7	8	8	8	8	8	8	8	8
	88	89	60	61	62	63	87	88	89	26	58	59	60	61	62	63	86
	8	8	8	9	9	9	9	9	9	9	9	10	10	10	10	10	10
	87	88	89	58	59	60	61	86	87	88	89	26	58	59	60	61	86
	10	11	11	11	11	12	12	12	25	25	26	26	26	28	28	29	29
	87	58	59	86	87	58	59	86	78	79	63	75	79	75	78	75	78
	29	30	30	30	30	30	31	32	33	33	33	35	35	36	36	36	37
	79	63	65	75	78	79	75	75	75	78	79	65	66	64	65	66	63
	37	37	37	38	38	38	38	38	39	39	39	39	40	40	40	40	41
	64	65	66	62	63	64	65	66	62	63	64	65	61	62	63	64	62
	41	41	42	42	43	43	44										
	63	64	61	62	62	63	61										
TN	4	4	6	6	6	7	7	8	8	8	9	10	26	26	29	30	30
	63	65	63	65	88	63	88	26	63	88	88	26	63	75	75	63	65
	30	32	33	35	37	37	40										
	75	75	75	65	63	65	63										

* Contacts are indicated between by pairs of vertically-separated sequence numbers.

Table 21. Conserved networks for human titin (1TIT)*

D	17	19	20	20	21	22	22	23	23	30	30	31	31	32	33	33	33
	62	60	59	60	58	57	58	56	73	75	76	74	75	74	72	73	74
	34	34	34	35	35	36	36	70	71	71							
	49	58	72	70	71	69	70	82	81	82							
T	17	18	18	18	18	19	19	19	19	20	20	20	20	21	21	21	21
	62	59	60	61	62	58	59	60	61	57	58	59	60	56	57	58	59
	21	22	22	22	23	23	23	23	24	25	25	30	30	31	31	31	32
	60	56	57	58	55	56	57	73	56	56	75	75	76	73	74	75	73
	32	32	32	33	33	33	33	34	34	34	34	34	34	35	35	35	35
	74	75	76	72	73	74	75	49	58	71	72	73	74	70	71	72	73
	35	36	36	36	36	69	70	71	71	72							
	74	69	70	71	72	82	82	81	82	82							
TN	18	19	19	20	21	21	21	23	23	25	25	30	31	32	32	34	34
	60	58	60	60	56	58	60	56	73	56	75	75	75	73	75	58	73

* Contacts are indicated between by pairs of vertically-separated sequence numbers.

The conserved networks are compared with randomly chosen native contact networks to show that inherent stability does exist.

This work supports a previous study on proteins from these same three superfamilies (121). In the previous study, it was revealed how conserved contact networks guide the formation of the Greek-Key topology. And a network common to all three superfamilies was employed. The current study extends the idea of stability to an already folded protein and involves separate networks for each superfamily. The network in the previous study was constructed using only one protein from each superfamily. But in this study, several proteins from each superfamily are used to determine an individual network for each superfamily.

METHODS

The molecular dynamics simulations use CHARMM (97-99) on a Dell 1950 cluster with dual-core 2.99-GHz Intel Xeon processors. Initial coordinates for the all α-helix death domain (from the two-domain Fas-associated death domain), the α/β-plait ribosomal protein and the all β-sheet immunoglobulin-like domain from titin are obtained from PDB structures 1E3Y, 1RIS and 1TIT, respectively. Using the CHARMM22 all-atom force field (34-35), each protein is simulated at 7 different temperatures: 300, 350, 400, 450, 500, 550 and 600 K. The same protocols used for simulating DHFR (see the Methods section of Chapter III) are used here, except for a few minor differences, which are listed here. Dynamics employs a 0.001-fs timestep. More heating is required during the initial 20-ps phase of equilibration in order to reach the final temperature (300, 350, 400, 450, 500, 550 or 600 K). All simulations extend for 10 nanoseconds of production dynamics, using a 1-fs time-step. Coordinates are saved every 100 steps, resulting in 100,000 frames per trajectory for analysis.

Determining whether the set of conserved contacts behaves any differently in simulation than a set (equal in number) of random native contacts requires a random selection protocol. First, the native contacts are numbered from 1 to NT, where NT is the total number of native contacts. So if there are NC conserved contacts, then NC unique random integers ranging from 1 to NT must be generated. Obtaining a random integer from 1 to NT is accomplished as follows: first, a random real number from 0 up to, but not including 1, is generated; second, it is multiplied by NT; third, the fraction part is removed, and fourth, 1 is added. The generation of the random real number ($0 \leq x < 1$) is done with the GNU FORTRAN 90 random_number() intrinsic function. The random networks for 1E3Y, 1RIS and 1TIT are shown in Tables 22, 23 and 24, respectively.

Table 22. Random contacts of human Fas-associated death domain (1E3Y)*

D	116	118	118	129													
	125	155	168	140													
T	90	90	91	94	95	96	96	97	98	99	99	100	101	107	111	111	112
	101	142	142	180	158	159	187	177	177	158	173	145	177	172	137	172	126
	112	112	113	113	114	114	114	115	115	116	116	119	119	119	119	119	125
	130	141	126	141	126	169	170	164	170	126	144	147	148	151	152	164	148
	126	129	157	159	159	159	161	163	163	165							
	140	139	184	180	181	184	177	174	177	178							
TN	90	96	97	101	101	104	107	111	112	119	121	125	157	159			
	101	158	156	149	156	119	177	170	128	148	147	148	184	180			

* Contacts are indicated between by pairs of vertically-separated sequence numbers.

Table 23. Random contacts of ribosomal S6 (1RIS)*

D	2	3	4	6	6	9	9	9	10	10	11	12	14	27	30	31	32
	92	94	93	61	75	58	84	86	19	23	85	46	84	63	40	75	75
	33	34	37	37	49	75											
	73	68	63	65	87	90											
T	1	1	2	2	3	3	3	4	4	4	4	4	4	5	5	5	6
	36	68	67	68	36	91	96	64	67	68	69	90	92	63	94	96	26
	6	6	6	6	6	7	7	7	7	8	8	8	8	8	8	9	9
	61	72	76	89	91	43	61	63	89	58	59	61	84	85	89	46	52
	9	9	9	9	10	10	11	11	11	12	12	12	13	13	14	14	18
	85	86	88	89	23	86	58	84	86	55	56	84	45	57	59	85	84
	19	19	19	21	22	22	22	23	25	25	26	27	27	28	29	30	30
	42	59	61	82	63	82	83	42	79	82	75	37	40	78	79	40	63
	30	30	30	32	32	33	33	33	34	34	34	35	36	37	38	38	39
	64	65	78	75	78	68	75	77	66	67	71	67	67	65	63	65	63
	39	40	40	40	41	43	43	45	46	47	47	47	48	48	49	53	53
	65	61	62	63	61	59	63	59	59	56	60	87	57	59	60	83	86
	59	60	65	75	77	78	80										
	86	87	90	88	88	88	89										
TN	2	4	4	4	5	5	6	9	10	11	12	14	29	31	35	36	36
	65	64	67	91	63	92	79	89	82	83	45	85	67	75	71	65	68
	38	39	41	42	47	60	80										
	66	64	61	61	56	87	90										

* Contacts are indicated between by pairs of vertically-separated sequence numbers.

Table 24. Random contacts of human titin (1TIT)*

D	2	3	3	3	5	8	8	13	14	15	17	19	19	22	25	25	32
	28	24	78	79	26	22	23	84	63	86	63	60	84	56	55	73	51
	32	34	36	37	67	68	69	69	69	72							
	73	51	70	68	86	84	82	83	85	82							
T	1	2	2	2	2	3	4	4	6	7	8	8	8	9	10	11	11
	77	26	27	28	78	24	25	78	24	24	34	73	81	82	82	83	85
	12	12	13	13	13	13	14	15	16	16	17	17	17	18	20	22	23
	83	87	61	63	87	88	88	64	61	63	60	61	62	62	58	58	54
	23	23	24	27	29	29	30	31	31	32	34	34	36	36	36	36	36
	55	80	54	53	52	54	54	56	73	51	49	74	60	66	68	72	84
	37	37	46	47	48	50	65	69	70	72							
	71	72	60	58	61	59	86	85	83	81							
TN	3	3	5	6	7	14	19	21	21	21	24	35	36	36	64	65	70
	25	78	24	25	24	63	60	34	58	84	54	70	68	72	86	88	82

* Contacts are indicated between by pairs of vertically-separated sequence numbers.

RESULTS AND DISCUSSION

The protein conformational stability during simulation was assessed by calculating the backbone RMSD versus native structure time series (Figure 40) for all seven simulations on each of the three proteins. A general trend common to all three proteins is that the RMSD increases with temperature. Increasing the temperature means increasing the kinetic energy and hence, the protein fluctuations. Comparing the three proteins shows that the thermophile ribosomal S6 stays closest to its native conformation. It maintains flat RMSDs of about 3 Å or less for temperatures 300-550 K (Figure 40B). At 550 K, the RMSD strays up to around 5 Å during the second quarter of the 10-ns simulation and then comes back down to about 3 Å. At 600 K, the thermophile unfolds; its RMSD steadily rises to more than 20 Å. Figure 41B shows evenly spaced snapshots of this unfolding. The Greek-Key topology is held until about 7 ns into the simulation. The death domain protein stays relatively close to its native structure (< 5 Å) for temperatures up to 450 K (Figure 40A). The 500 K simulation remains stable (with an RMSD ~5 Å for the first 4 ns), then jumps to around 10 Å and slowly increases after that. Figure 41A shows snapshots of the 500 K

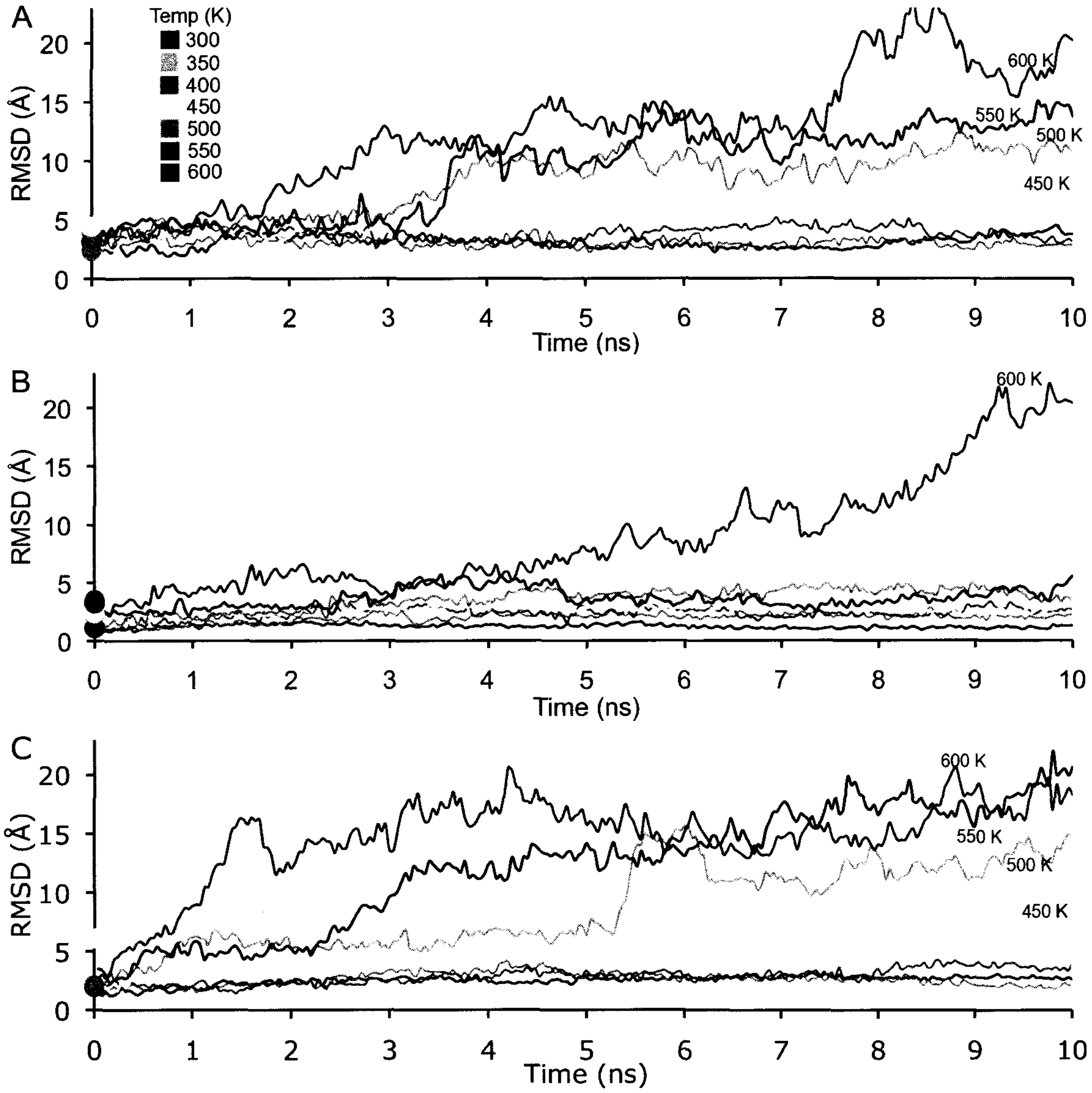


Figure 40. Greek-Key proteins backbone RMSD vs. time. Shown are 1E3Y (A), 1RIS (B) and 1TIT (C). Plots combine data from all high-temperature simulations: 300, 350, 400, 450, 500, 550 and 600. Filled circles on the y-axis indicate the RMSD of the minimized-average simulation structure.

simulation. Although much secondary structure is held for the duration, inspection reveals that the native Greek-Key topology is lost after 5 ns. Not surprisingly, the 550 and 600 K simulations end up even further from the native structure (Figure 40A). Titin shows an RMSD trend similar to that of the death domain. From 300 to 400 K, it stays within 3 Å of the native structure at 450 K within 8 Å, but like the death domain ends up unfolding at 500 K, as indicated by its mid-simulation jump in RMSD to ~15 Å (Figure

40C). At 550K, titin gradually strays up to 20 Å away from its native structure. And at 600 K, it quickly increases to an RMSD ~15 Å, and after that it increases far above 20 Å. Figure 41C shows titin's gradual unfolding at 500K. These results reveal that the three

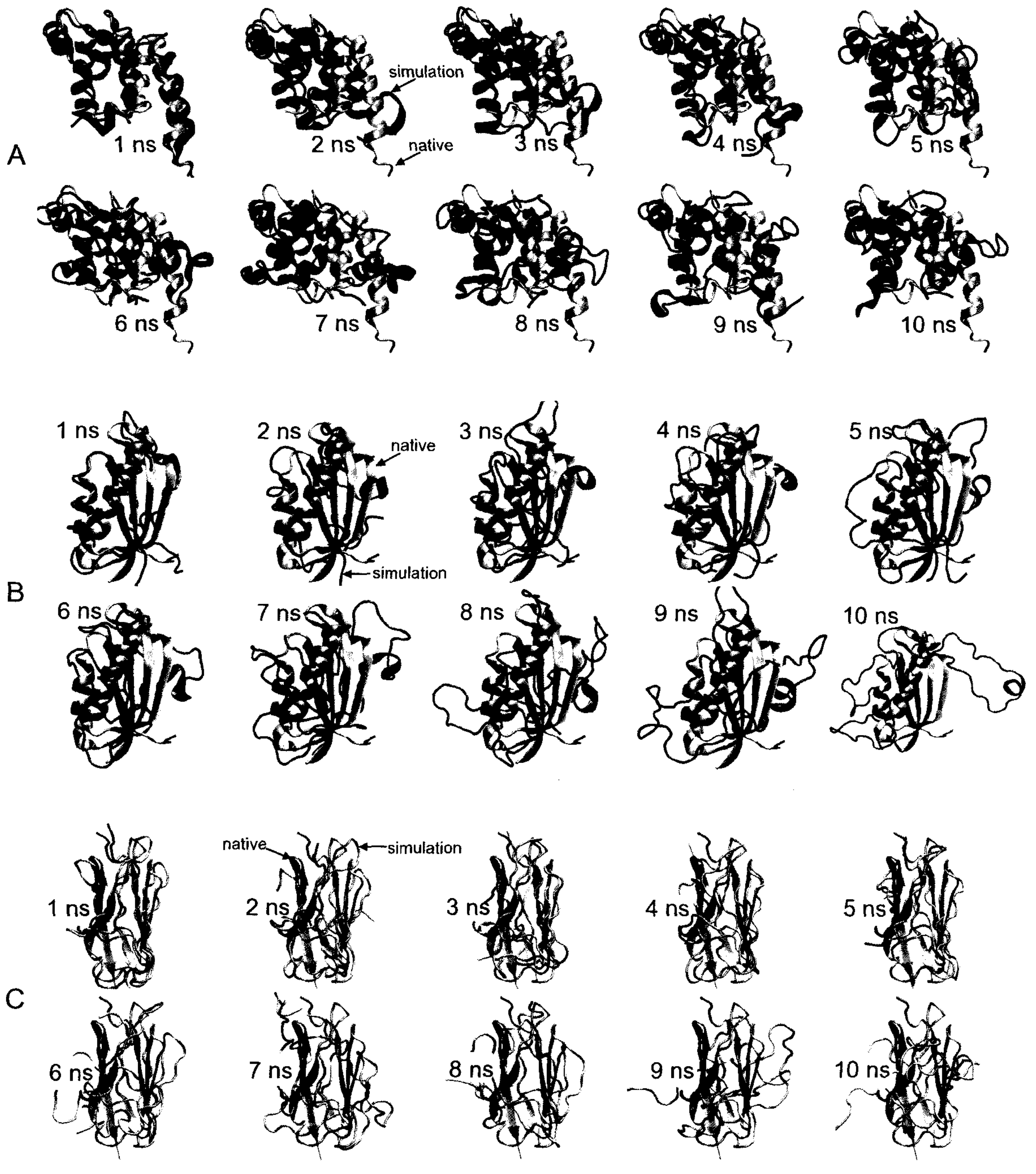


Figure 41. Protein snapshots after each nanosecond during unfolding simulations. Shown are 1E3Y at 500K (A), 1RIS at 600K (B) and 1TIT at 500K (C). Simulation structures are superimposed onto the respective native structures.

proteins vary in their ability to remain in a folded state as temperature increases. And as expected, the thermophile has a greater ability than the other two.

The minimized-average structures are calculated along the entire 10 ns for all simulation trajectories remaining within 10 Å RMSD of the native structures and used in subsequent contact analyses. This corresponds to 300-450K for both the death domain and titin, and 300-550K for ribosomal S6. RMSDs going higher than 10 Å indicate an unfolding gradient occurring, so finding those average (or minimized-average) structures makes little sense. Backbone RMSDs of the minimized-average simulation structures relative to the native structures are reported in Figure 40 by solid circles on the y-axis. These values are all lower than 5 Å (except for the 450 K minimized-average structure of titin, which has a 6.095 Å RMSD), which indicates similarity to the native backbone conformations. Figure 42 shows these structures superimposed onto their respective native structures. Death domain and titin (Figures 42A and 42C, respectively) both show their 450 K structure (yellow) a little shifted from the other lower temperature structures and native state. Likewise, the 550 K structure (red) of ribosomal S6 (Figure 42B) displays a couple of loop regions that do not overlap well with the other lower temperature structures. In the other high temperature simulations with RMSDs > 10 Å, the proteins do not fluctuate about an average structure, but unfold in many (if not all) sections, hence it is not meaningful to present them as minimized-average structures.

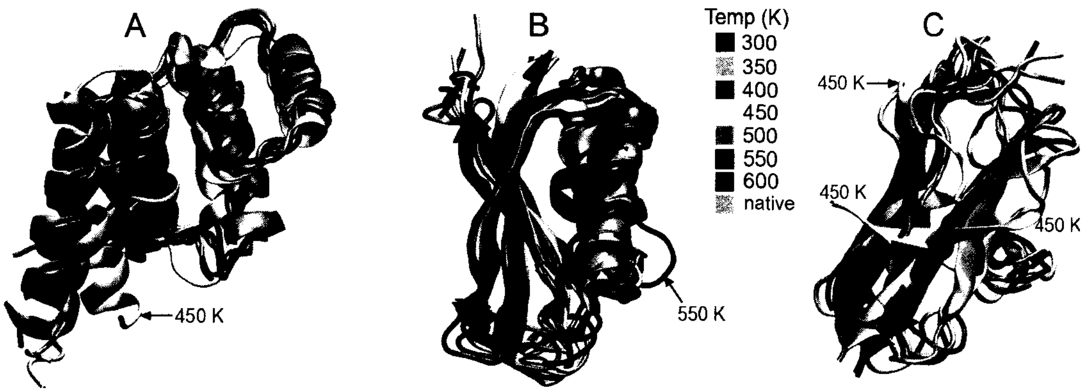


Figure 42. Superimposed minimized-average simulation structures. Shown are 1E3Y (A), 1RIS (B) and 1TIT (C). Simulation structures are superimposed onto the native structure.

Contact Analysis

The three conserved contact networks (Direct - TN, Toggling - T and Toggling without acidic/basic residues - TN) determined in the first part of this study are assessed to see if they behave differently (under simulation) than randomly selected contact networks. For each protein, the three networks contain different numbers of contacts, so three sets of random contact networks are chosen for each of the three proteins, or nine random sets in all. The conserved networks for 1E3Y, 1RIS and 1TIT are shown in Tables 19-21, and the random networks are shown in Tables 22-24. Three tests are performed to demonstrate that the conserved networks do, in fact, behave differently than randomly chosen networks. The same random contact networks were used for all three tests. The first checks the average contact distances in each of the minimized average simulation structures to see if it changes with temperature. A steeper increase in the random contact distances will indicate that they fall apart easier, and hence play a lesser role in protein stability than the conserved contacts do. The second test compares the average simulation root-mean-square fluctuations (RMSF) of the contact distances. Larger fluctuations in the random contact distances indicate less inherent stability within those contacts and more of a tendency to break apart than the conserved contacts. The third test examines a high-temperature unfolding simulation in each protein. The number of conserved contacts kept is compared with the number of random contacts kept as the protein unfolds. Here, smaller numbers of random contacts maintained indicate that, during unfolding, the random contacts break before the conserved contacts do.

In the first test, the conserved contact distances are calculated for a minimized-average structure and then averaged. This calculation is repeated for each minimized-average structure and for the random contacts. Figure 43 displays these averages as distance vs. temperature graphs. To determine the rate of increase of the contact distance average with temperature, linear least squares regression lines are fitted to the data. Clearly the slopes of the conserved lines (solid) are less than those of the random lines (dashed). To quantify this comparison, we define the network score to be the slope of the random regression line divided by the slope of the conserved regression line. Network scores greater than 1 indicate an inherent stability of the conserved network. The regression line slopes and network scores are shown in Table 25. As indicated by all the

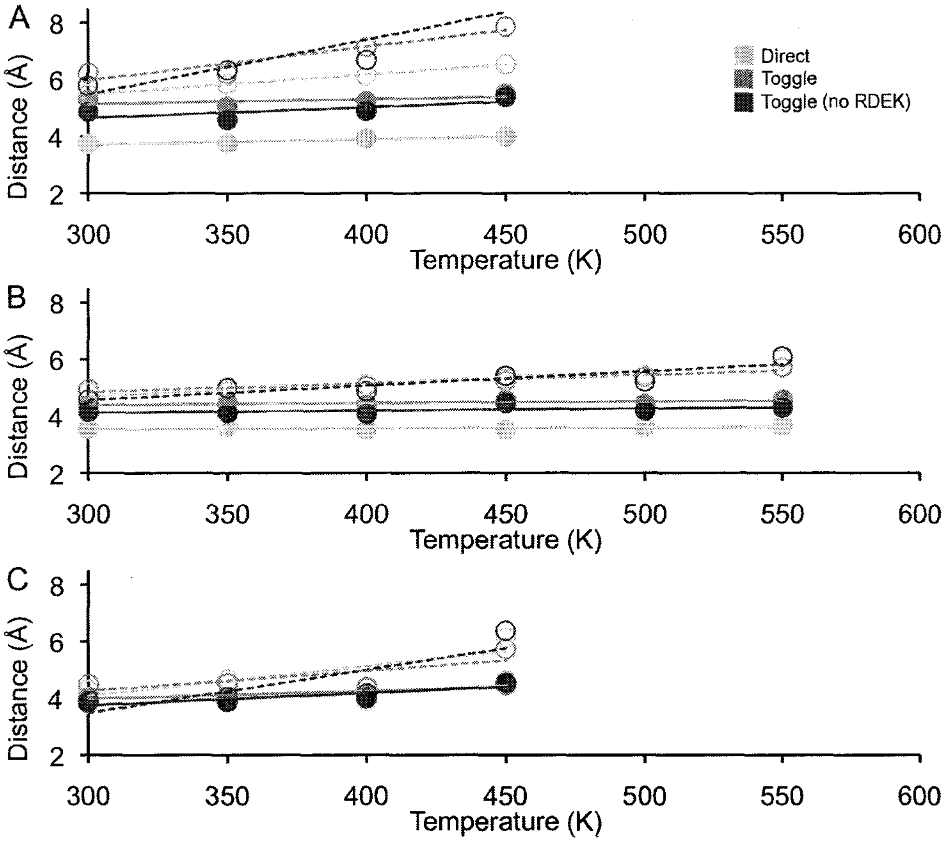


Figure 43. Average contact distance vs. temperature (random). Graphs show average conserved/random contact distances in the simulation minimized-average structures and their associated linear least-squares-fitted trendlines. Plots show 1E3Y (A), 1RIS (B) and 1TIT (C) and include the data from each algorithm used to determine the conserved network: Direct, Toggling and Toggling without acidic/basic residues. Conserved contact data are plotted as solid lines and random contact data as dashed lines.

Table 25. Regression line slopes and network scores (random)

	Direct			Toggle			Toggle no RDEK		
PDB	Cons.	Rand.	|R/C|*	Cons.	Rand.	|R/C|	Cons.	Rand.	|R/C|
1E3Y	0.0018	0.0070	3.8	0.0016	0.0117	7.4	0.0038	0.0192	5.1
1RIS	0.0003	0.0045	14.6	0.0006	0.0029	4.8	0.0007	0.0050	6.9
1TIT	0.0046	0.0104	2.3	0.0027	0.0071	2.6	0.0043	0.0152	3.5

* Network score = Random/Conserved

network scores being greater than 1, the conserved networks in each protein break later than other contacts as the temperature is increased. In other words, on average, the contacts within the conserved networks are more robust, and can withstand more thermal shaking than other native contacts. So they contribute a larger portion of the stabilizing energy that holds the protein together. Without them, the contact distances, on average, will increase and cause the protein to fall apart.

Comparing the conserved networks with sets of random contacts has two problems. First, different random network selections would, of course, yield different network scores, perhaps even resulting in scores less than 1. Repeating our procedures several times and averaging the network scores from each trial would overcome this obstacle. However, it is better to compare the conserved networks with the entire set of native contacts. The latter method was employed. The average distance of all native contacts was calculated for each minimized-average structure. Figure 44 and Table 26 provide these results. In all cases, the network scores are above 1, which shows that the conserved contacts are held together more tightly than randomly chosen contacts.

The second problem is that the random contact distances in the native structures of the three proteins, are, on average, already greater then the conserved contact distances. Figure 43 shows the difference clearly for the 300-K minimized-average structures, which are similar enough to the native structures. The random network average distances (hollow spheres) are greater than the conserved network average distances (solid spheres). However, the random native contacts were chosen using the same criteria used to identify the conserved native contacts (i.e. the two residues of the contact are sequentially separated by at least 9 other residues and the contact is less than 7 Å long for the death domain and ribosomal S6 and 6 Å long for titin). So, a question is raised of whether the conserved contacts behave differently only because of their shorter contact distances. The experiments were repeated using a different set of random contacts, having contact distances less than or equal to the greatest distance in the entire conserved networks. This process may create three different cutoff distances for each protein, since each of the three conserved networks may have a different greatest conserved contact distance. The results obtained from these parallel studies are shown in Figures 45 and 46 with corresponding Tables 27 and 28 (below). In

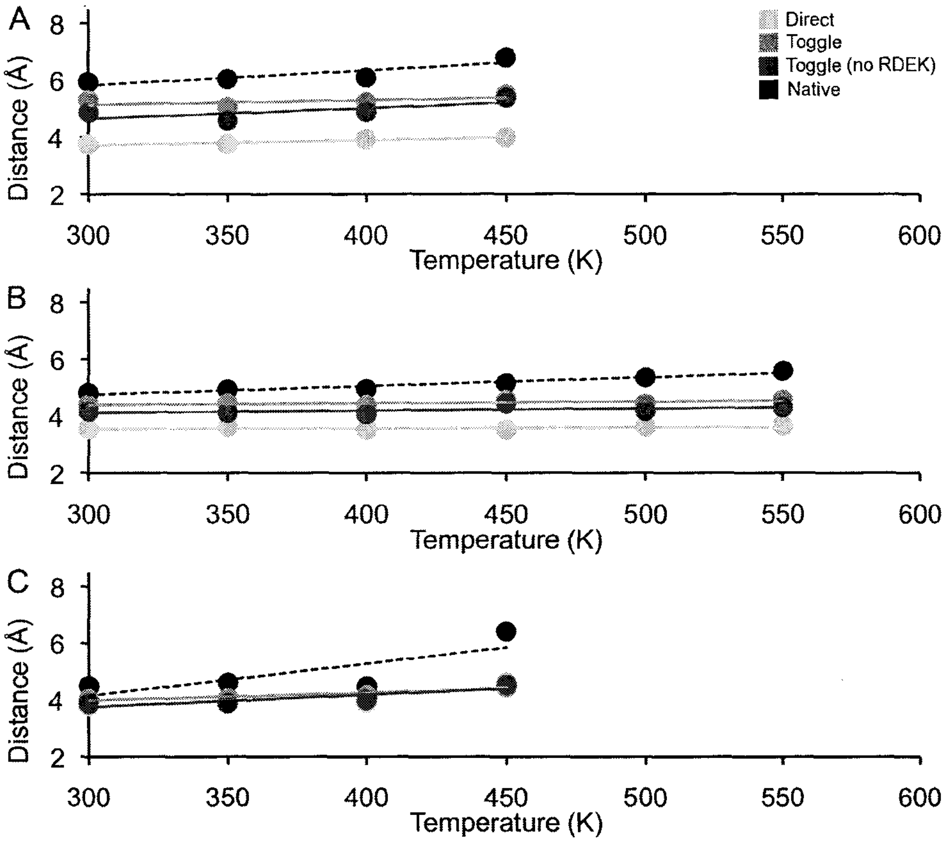


Figure 44. Average contact distance vs. temperature (native). Graphs show average contact distances in the simulation minimized-average structures and their associated linear least-squares-fitted trendlines. Plots show 1E3Y (A), 1RIS (B) and 1TIT (C) and include the data from each algorithm used to determine the conserved network: Direct, Toggling and Toggling without acidic/basic residues. Conserved contact data are plotted as solid lines and native contact data as dashed lines.

Table 26. Regression line slopes and network scores (native)

		Direct		Toggle		Toggle no RDEK	
PDB	Native	Cons.	|N/C|*	Cons.	|N/C|	Cons.	|N/C|
1E3Y	0.0053	0.0018	2.9	0.0016	3.3	0.0038	1.4
1RIS	0.0030	0.0003	9.8	0.0006	5.0	0.0007	4.2
1TIT	0.0113	0.0046	2.5	0.0027	4.1	0.0043	2.6

* Network score = Native/Conserved

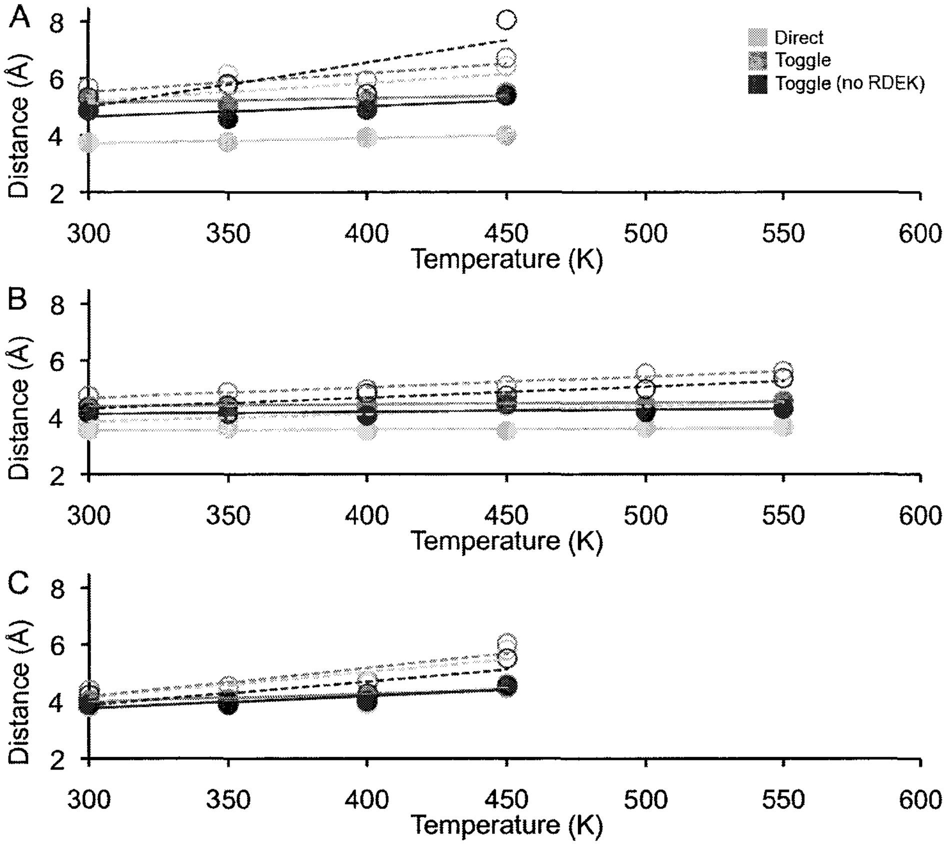


Figure 45. Average contact distance vs. temperature (lower-distance random). Graphs show average conserved/random contact distances in the simulation minimized-average structures and their associated linear least-squares-fitted trendlines. Plots show 1E3Y (A), 1RIS (B) and 1TIT (C) and include the data from each algorithm used to determine the conserved network: Direct, Toggling and Toggling without acidic/basic residues. Conserved contact data are plotted as solid lines and random contact data as dashed lines.

Table 27. Regression line slopes and network scores (lower-distance random)

PDB	Direct			Toggle			Toggle no RDEK		
	Cons.	Rand.	|R/C|*	Cons.	Rand.	|R/C|	Cons.	Rand.	|R/C|
1E3Y	0.0018	0.0064	3.5	0.0016	0.0066	4.2	0.0038	0.0155	4.1
1RIS	0.0003	0.0027	8.7	0.0006	0.0037	6.1	0.0007	0.0038	5.3
1TIT	0.0046	0.0088	1.9	0.0027	0.0101	3.7	0.0043	0.0083	1.9

* Network score = Random/Conserved

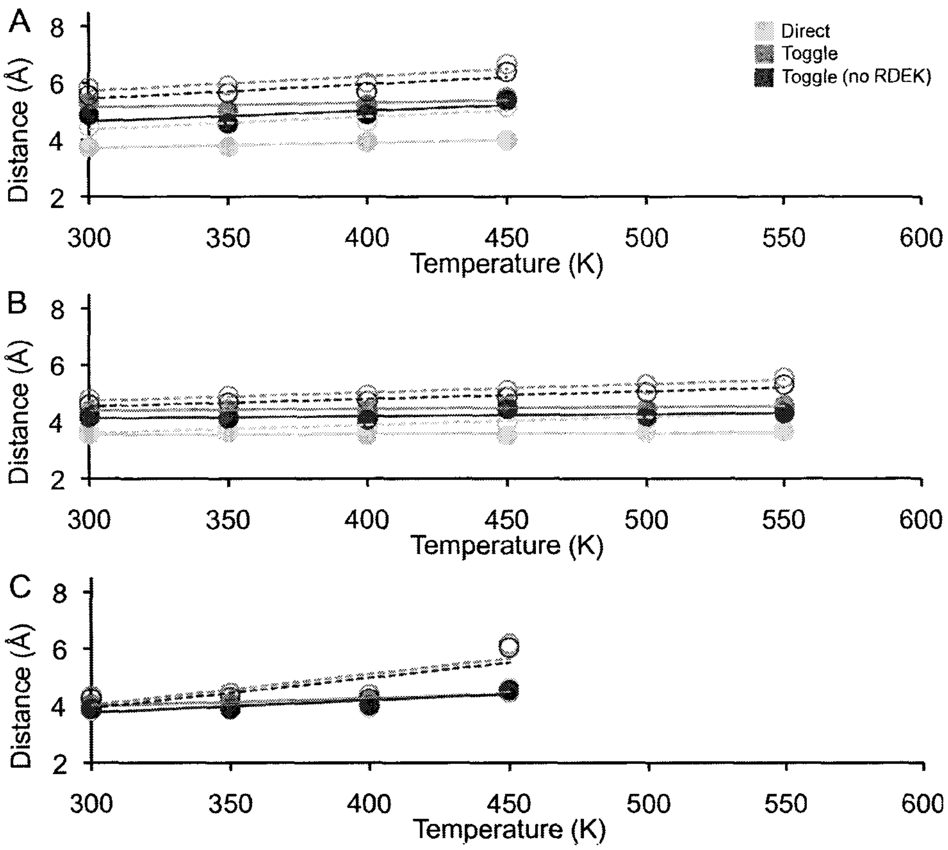


Figure 46. Average contact distance vs. temperature (lower-distance native). Graphs show average conserved/native contact distances in the simulation minimized-average structures and their associated linear least-squares-fitted trendlines. Plots show 1E3Y (A), 1RIS (B) and 1TIT (C) and include the data from each algorithm used to determine the conserved network: Direct, Toggling and Toggling without acidic/basic residues. Conserved contact data are plotted as solid lines and native contact data as dashed lines.

Table 28. Regression line slopes and network scores (lower-distance native)

PDB	Direct			Toggle			Toggle no RDEK		
	Cons.	Native	|N/C|*	Cons.	Native	|N/C|	Cons.	Native	|N/C|
1E3Y	0.0018	0.0045	2.5	0.0016	0.0051	3.3	0.0038	0.0051	1.3
1RIS	0.0003	0.0028	9.0	0.0006	0.0029	4.8	0.0007	0.0027	3.7
1TIT	0.0046	0.0108	2.4	0.0027	0.0110	4.0	0.0043	0.0107	2.5

* Network score = Native/Conserved

all cases, the network scores are above 1. These results confirm that the inherent stability of the conserved contacts over other contacts is not due to their shorter contact distances.

The results from the second test of the contact distance RMSF averages versus temperature graphs are given in Figure 47. The graphs are scaled so that only the data from simulations whose RMSD is less than or equal to 10 Å (300-450 K for death domain and titin, and 300-550 K for ribosomal S6) are shown. Beyond that, the proteins unfold and distance fluctuations are meaningless. Clearly, in all cases, the fluctuations of the distances of the random contacts (dashed lines) are greater than those of the conserved contacts (solid lines). The difference between random and conserved contact distance fluctuations is magnified at higher simulation temperatures. 1RIS shows this clearly; at 300 K, it is difficult to tell the difference between the plotted points, but at 550 K, the points are clearly separated. As before, the conserved networks' RMSF averages were compared with those of the entire native contacts (Figure 48), those of the lower distance set of random contacts (Figure 49) and those of the lower distance set of native contacts (Figure 50). Again, in all cases, the results remain unchanged. Results from the second test confirm that for folded Greek-Key proteins, conserved contacts do not fluctuate in distance as much as other contacts do. However, if conserved interactions do maintain and stabilize the native topology, then they are expected to be more rigid than other interactions.

The third test, perhaps, most clearly shows the difference between the conserved contacts and the random contacts. Here, the fraction of conserved contacts remaining is compared with the fraction of random contacts remaining as the protein gradually unfolds. At each frame in the simulation trajectory, the conserved interactions are assessed to determine how many still satisfy the criteria of being a contact. The fraction of contacts remaining is then the number still in contact divided by the original number of conserved contacts. Contacts are lost as the protein unfolds. So the fraction of native contacts remaining should decrease with simulation time. The same calculations are done on the set of random contacts, which requires an unfolding or partial unfolding simulation for each protein. As seen in the RMSD graphs (Figure 40), ribosomal S6 only unfolds partially at 600 K simulation so this is the only trajectory that can be used. Death domain and titin both unfold at 500, 550 and 600 K. The 500 K simulations are chosen for both

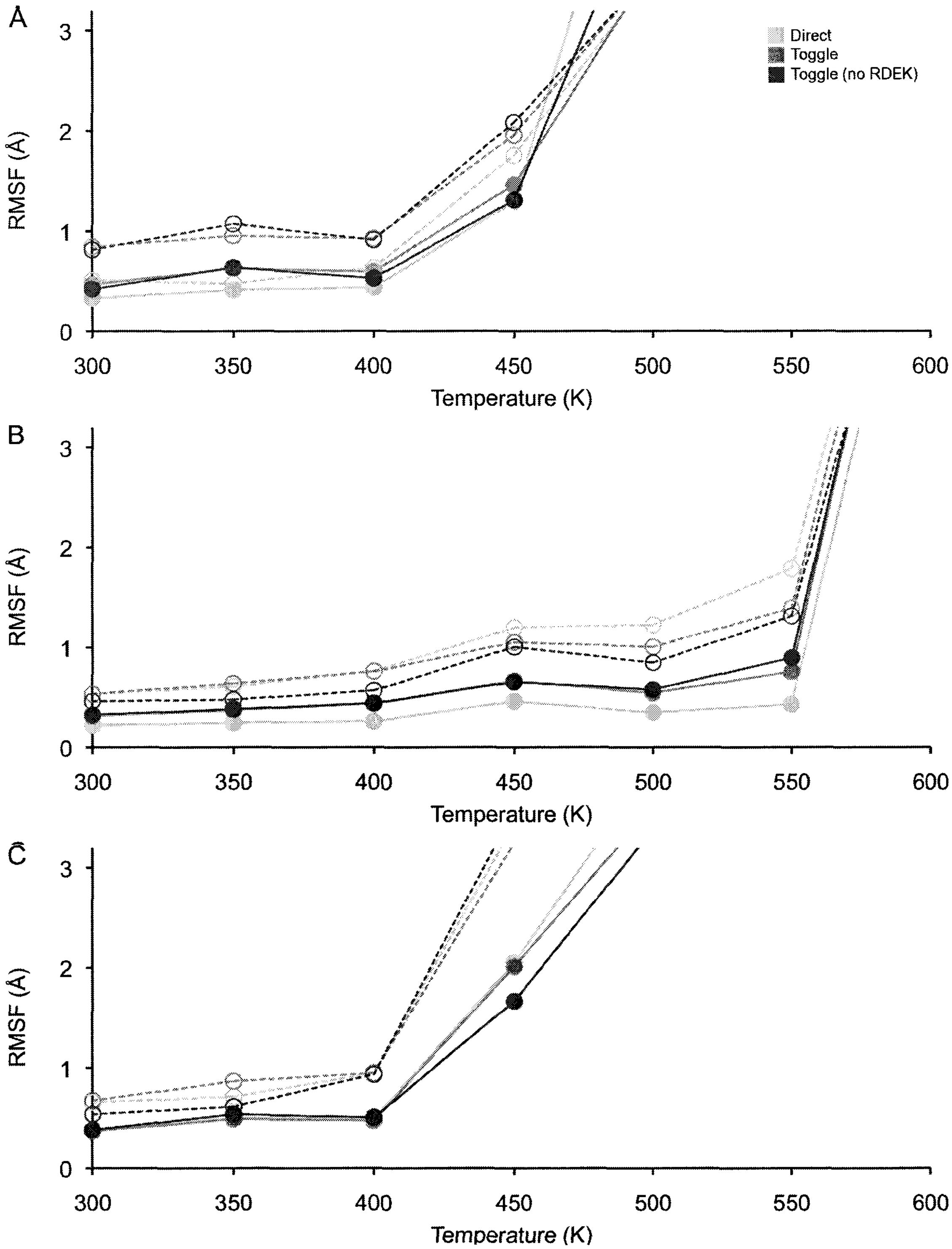


Figure 47. Average contact distance RMSF vs. temperature (random). Plots show 1E3Y (A), 1RIS (B) and 1TIT (C) and include the data from each algorithm used to determine the conserved network: Direct, Toggling and Toggling without acidic/basic residues. Conserved contact data are plotted as solid lines and random contact data as dashed lines.

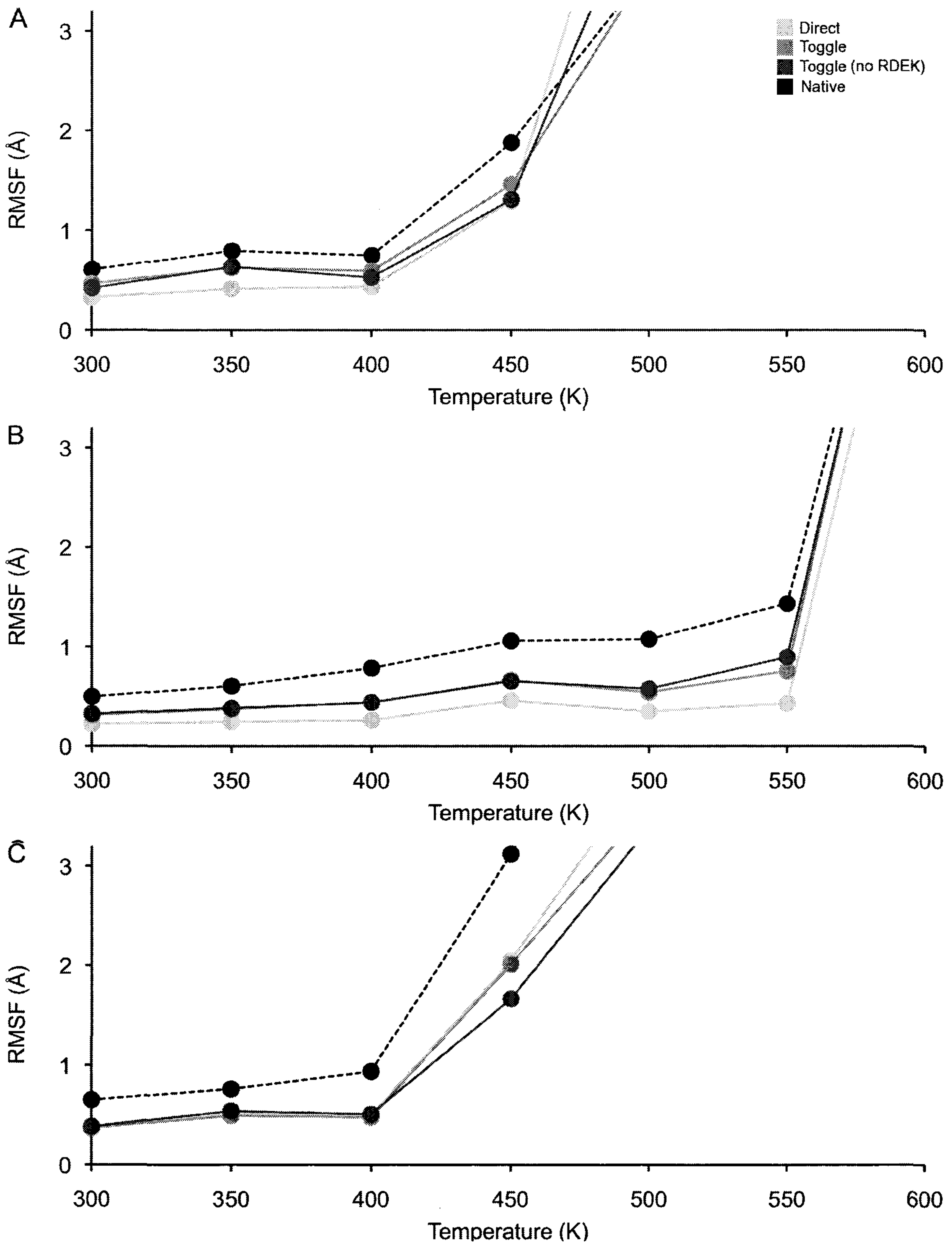


Figure 48. Average contact distance RMSF vs. temperature (native). Plots show 1E3Y (A), 1RIS (B) and 1TIT (C) and include the data from each algorithm used to determine the conserved network: Direct, Toggling and Toggling without acidic/basic residues. Conserved contact data are plotted as solid lines and native contact data as dashed lines.

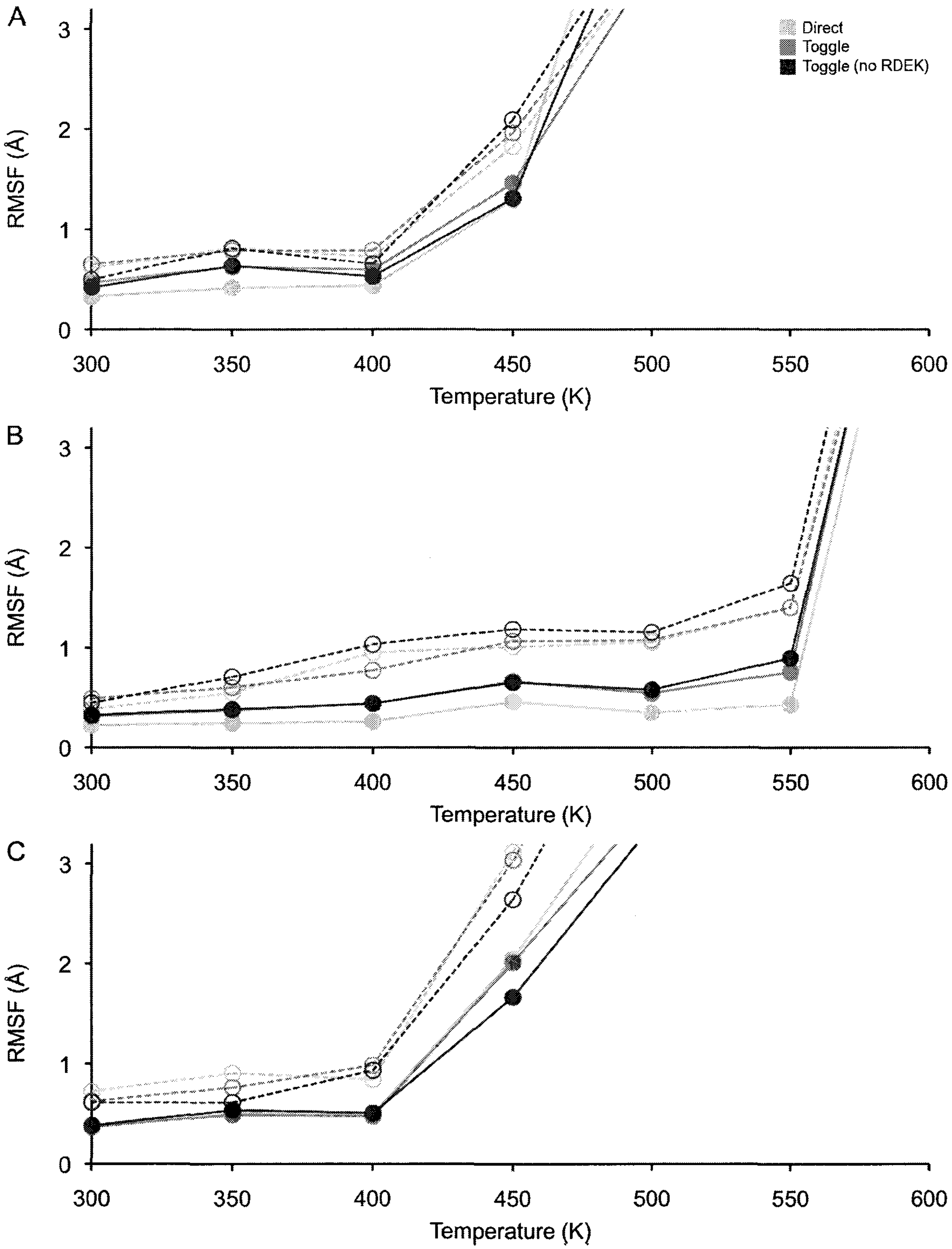


Figure 49. Average contact distance RMSF vs. temperature (lower-distance random). Plots show 1E3Y (A), 1RIS (B) and 1TIT (C) and include the data from each algorithm used to determine the conserved network: Direct, Toggling and Toggling without acidic/basic residues. Conserved contact data are plotted as solid lines and random contact data as dashed lines.

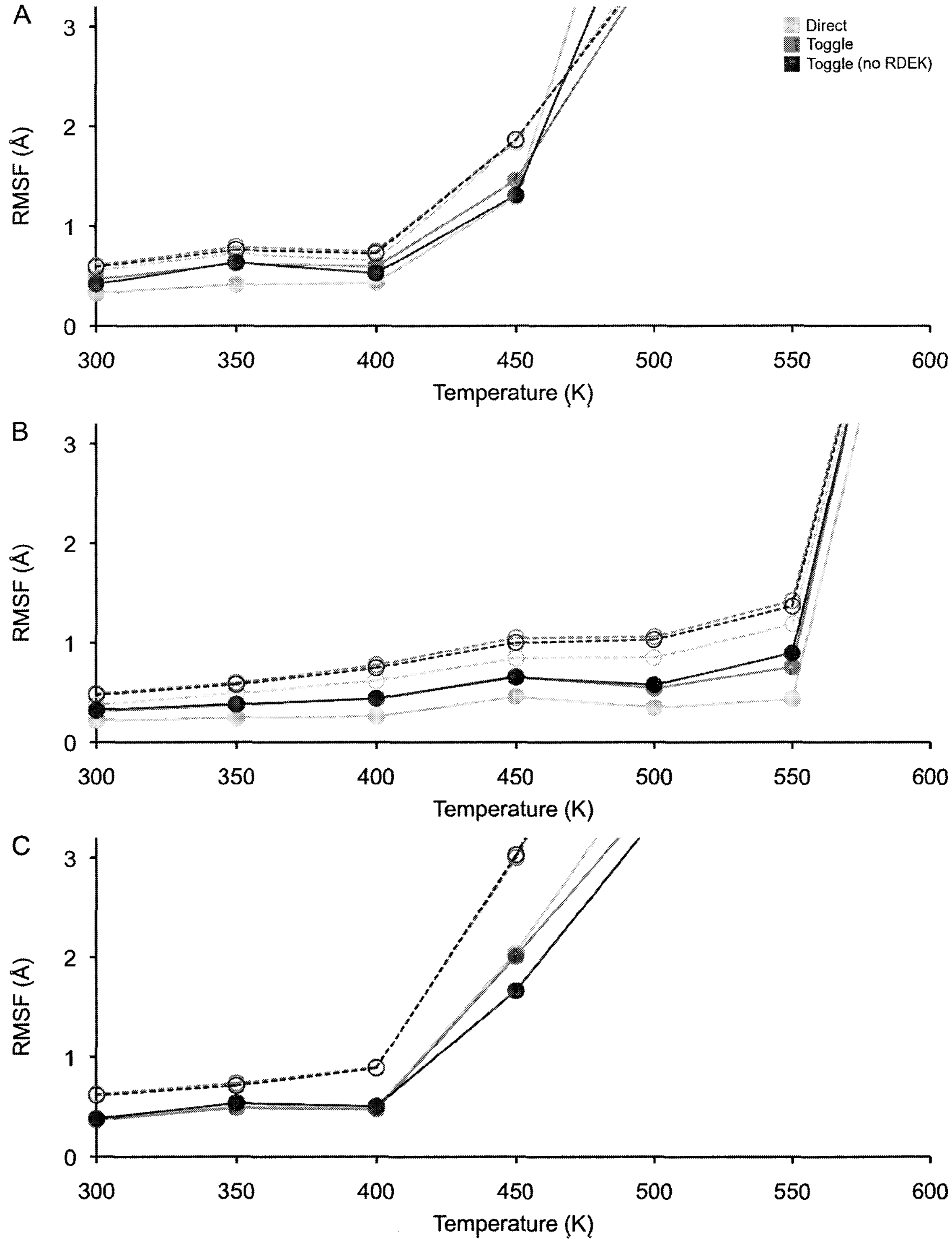


Figure 50. Average contact distance RMSF vs. temperature (lower-distance native). Plots show 1E3Y (A), 1RIS (B) and 1TIT (C) and include the data from each algorithm used to determine the conserved network: Direct, Toggling and Toggling without acidic/basic residues. Conserved contact data are plotted as solid lines and native contact data as dashed lines.

both proteins since, in them, the unfolding process is more gradual. The results are shown in Figure 51. Complete breakage of the tertiary structure is seen when the fraction decreases closed to 0 or drops to about 0.2. This occurs at ~3 ns for death domain, ~7 ns for ribosomal S6 and ~7 ns for titin. Clearly, in each protein, the conserved networks (solid lines) retain more contacts longer than the random networks (dashed lines) do. Interestingly, but perhaps not surprisingly, the thermophile ribosomal S6 maintains most conserved contacts for ~7 ns at 600 K, while titin only maintains them for ~7 ns and death domain for ~3 ns, both at the lower temperature of 500 K. Yet again, additional calculations comparing the fractions of conserved contacts remaining with both the fraction of native contacts, the fraction of lower-distance random contacts and the fraction of lower-distance native contacts are performed. Graphs of these results are shown in Figures 52, 53 and 54, respectively. Again, in all cases, the results are the same, as was observed for the completely random contact set. These results show that as Greek-Key proteins unfold, conserved contacts break later than other contacts do suggesting that when a Greek-Key protein folds, conserved contacts are the first to form.

The three tests showed that all three conserved networks (D, T and TN) exhibited more stabilizing effects on Greek-Key topology than random contacts do suggesting that any of these three algorithms would have sufficed in producing a conserved interaction network. In fact, the collected data make no clear distinction on which algorithm serves as the best for all three proteins. However, when considering only ribosomal S6, the Direct method appears the best. From Table 25, the slope of the average contact distance in the first test is less for the D network contacts (0.0003) than for the network contacts (0.0006 and 0.0007 for T and TN, respectively). The second test also shows that the contact distances in the D network, on average, fluctuate less than those in the T or TN networks. In the third test, more of the D network contacts are held for longer as the protein unfolds than contacts from the other two networks are. Thus, the Direct network is the best choice for ribosomal S6. It is not as clear in the other two proteins. The slopes in the first test indicate that the T network is the best for titin (Table 25), but there is no clear distinction in the contact distance fluctuations (Figure 47) or the fraction of contacts remaining (Figure 51). The first test slopes put the D and T networks about equal for death domain, but there is no distinction in the second and third tests.

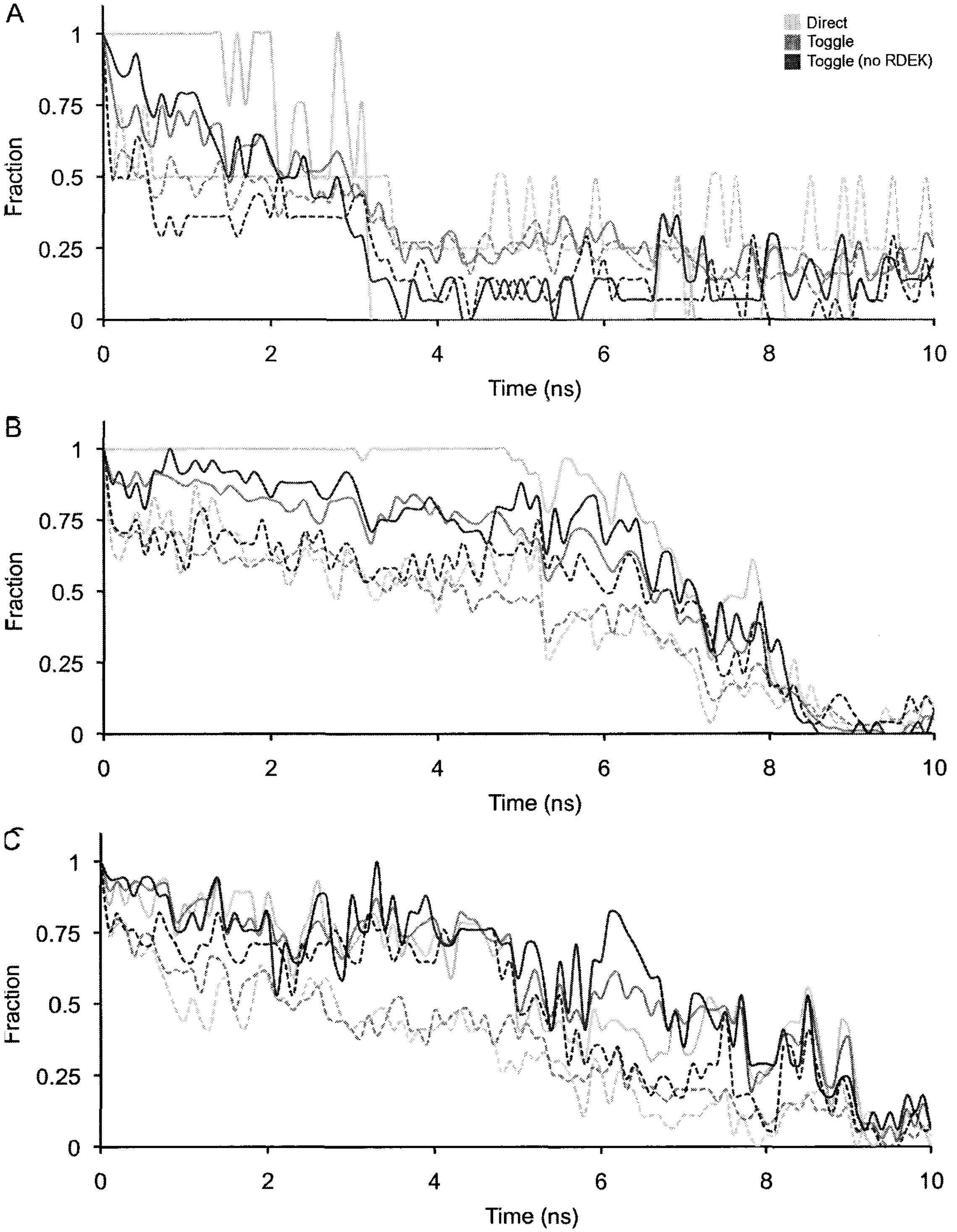


Figure 51. Fraction of contacts remaining vs. time (random). Plots show 1E3Y (A), 1RIS (B) and 1TIT (C) and include the data from each algorithm used to determine the conserved network: Direct, Toggling and Toggling without acidic/basic residues. Conserved contact data are plotted as solid lines and random contact data as dashed lines.

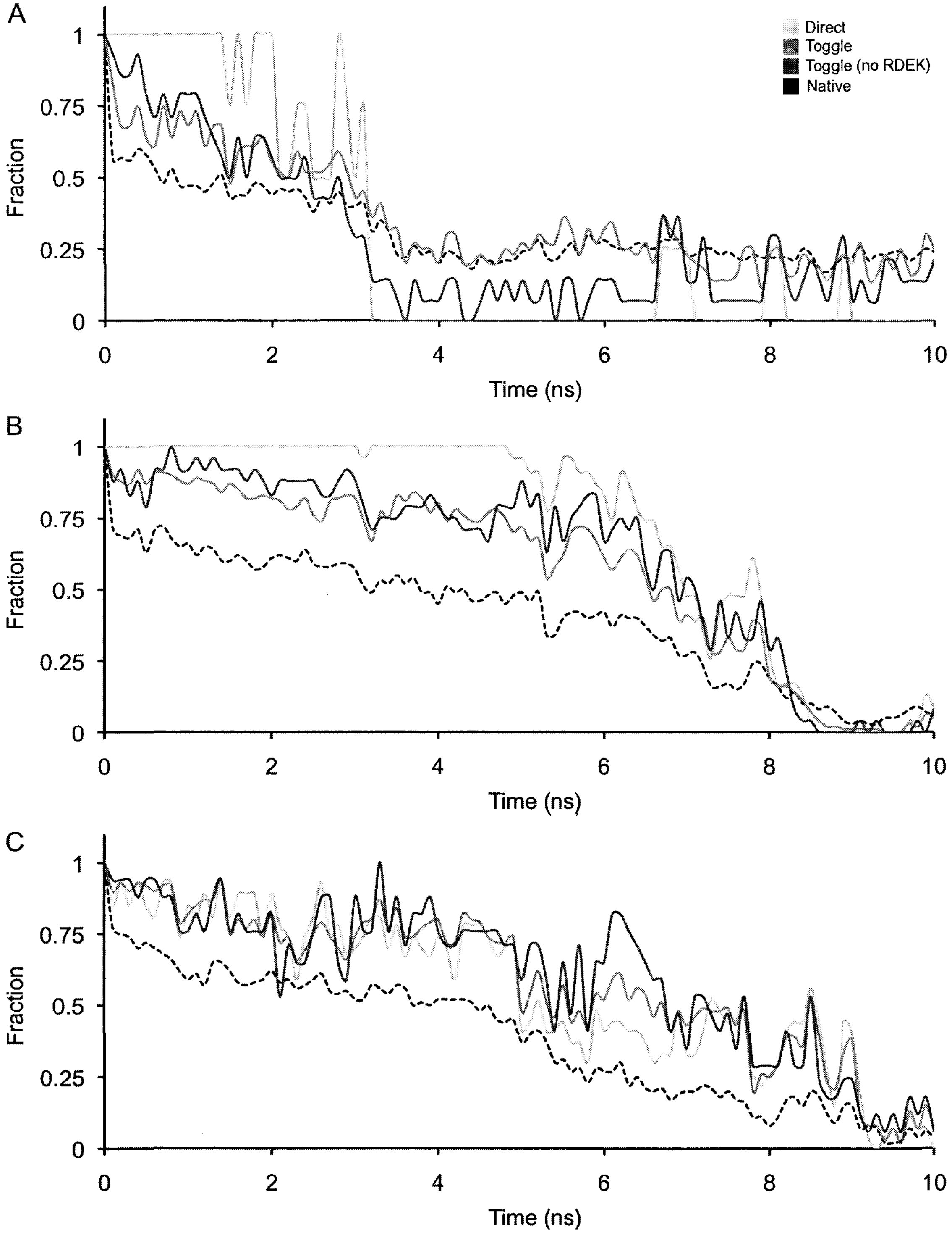


Figure 52. Fraction of contacts remaining vs. time (native). Plots show 1E3Y (A), 1RIS (B) and 1TIT (C) and include the data from each algorithm used to determine the conserved network: Direct, Toggling and Toggling without acidic/basic residues. Conserved contact data are plotted as solid lines and native contact data as dashed lines.

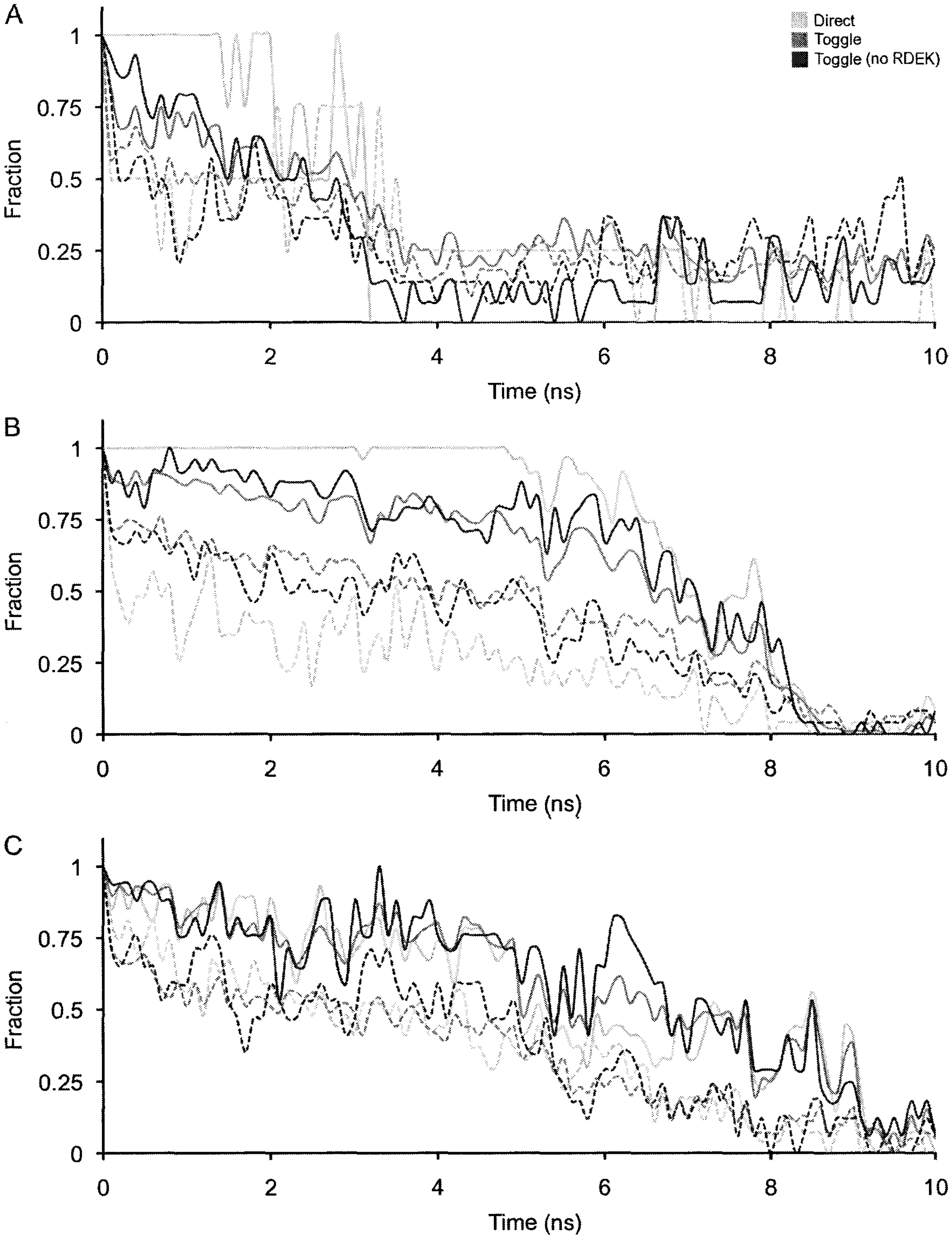


Figure 53. Fraction of contacts remaining vs. time (lower-distance random). Plots show 1E3Y (A), 1RIS (B) and 1TIT (C) and include the data from each algorithm used to determine the conserved network: Direct, Toggling and Toggling without acidic/basic residues. Conserved contact data are plotted as solid lines and random contact data as dashed lines.

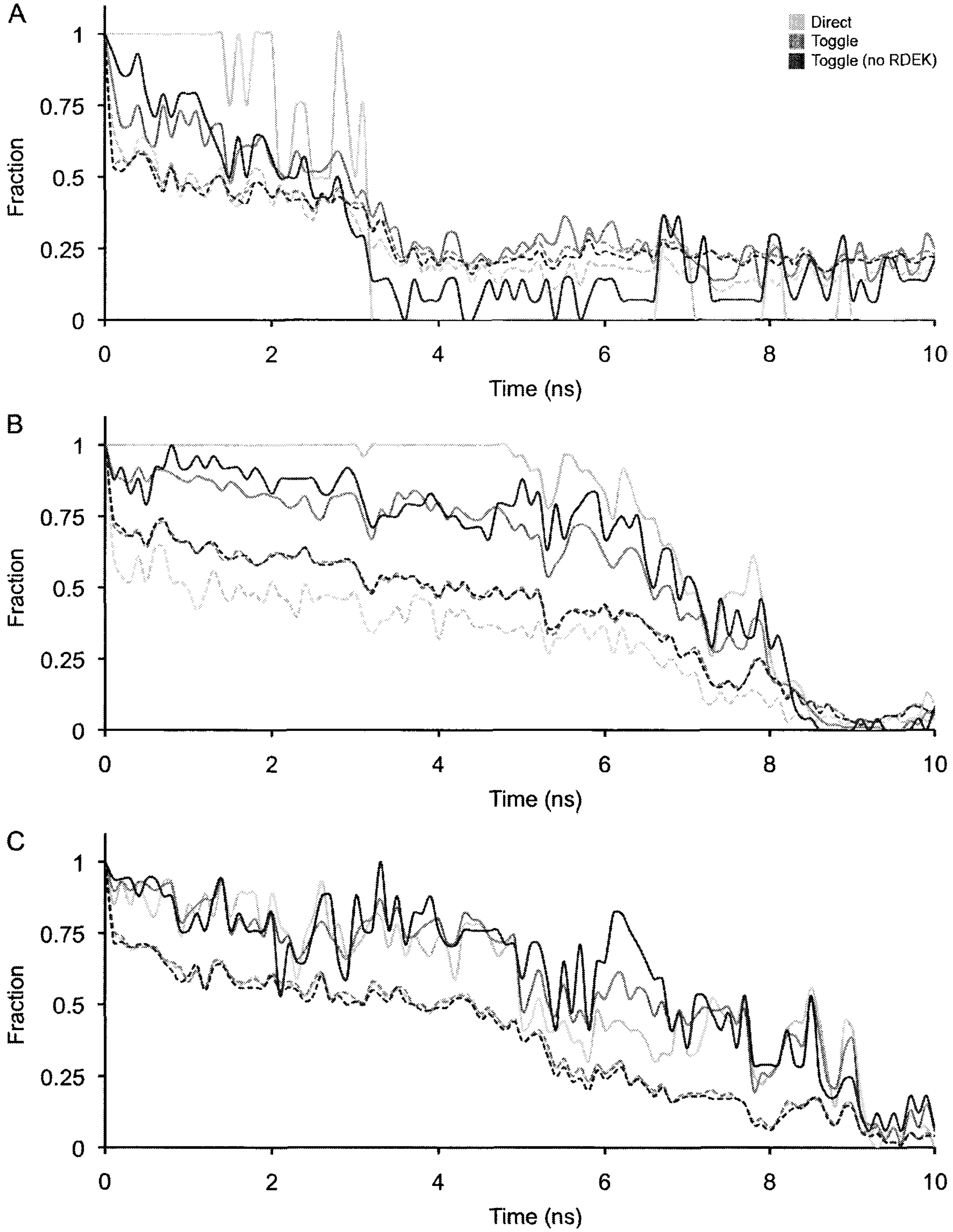


Figure 54. Fraction of contacts remaining vs. time (lower-distance native). Plots show 1E3Y (A), 1RIS (B) and 1TIT (C) and include the data from each algorithm used to determine the conserved network: Direct, Toggling and Toggling without acidic/basic residues. Conserved contact data are plotted as solid lines and native contact data as dashed lines.

CONCLUSIONS

The three Greek-Key proteins, (human Fas-associated death domain, ribosomal S6 and human titin) were subjected to high temperature simulations for the purposes of examining the conserved long-range residue interaction networks found among their respective protein superfamilies. Under folded conditions, the conserved contacts were found to be more resistant to breaking and were less prone to fluctuating in length than other random native contacts. During unfolding simulations, the conserved contacts were the last ones to break before the protein unfolded. These effects were subjected to much scrutiny to make sure they were not just a random occurrence. Thus, the conserved networks were found to be a major stabilizing force that prevents the Greek-Key topology from changing. The trends remained the same for all three conserved networks (Direct, Toggling and Toggling with no acidic/basic residues).

The analysis in the Results and Discussion section of this chapter raises the question of what the nature of the conserved and random contact networks is. The conserved contacts exhibit different behavior than the random contacts. But what is the definition of a contact? And are the conserved contacts similar, in nature, to the random contacts? Both sets of contacts are chosen from among the set of all native contacts. So the only defining difference between the two is that the conserved contacts of a particular protein are found among all of the other proteins in the entire superfamily. In other words, the conserved contacts are hereditary and the random contacts are not. Therefore the question of whether or not the conserved contacts are of the same type as the random contacts are depends on whether (or not) all native contacts are of the same type.

The Methods section of this chapter describes the algorithm used to define a native contact. It states that two residues form a direct contact if, first, they are separated by 9 or more other residues and, second, the distance between any heavy atom from one residue and any heavy atom from the other residue is lower than a certain cutoff distance. Toggling the residue position to account for secondary structure increases the number of contacts, but the general idea remains the same. This algorithm acknowledges a large variation in the types of residue-residue contacts that make up the set of native contacts. For instance, a certain native contact may only have one heavy atom-heavy atom interaction below the cutoff while another native contact may have 20 interactions. But they both satisfy the requirements of being a native contact. Another source of variation is the absence of the chemical properties of the native contacts. Some contacts may result from polar interactions, others from nonpolar van der Waals type forces, and still others from water-mediated hydrogen bonding. But, again, each type satisfies the requirements of being a native contact.

So, in theory, the conserved contacts could be of a much different nature than the random contacts. This would only happen if the algorithm that is used to classify a native contact as conserved also places weight upon one type of contact over another. In other words, do the conserved contacts have proportionately more hydrogen-bonding contacts than its superset native contacts do? It is, of course, possible for such differences in the chemical type to occur. And given enough time, a more stringent protocol for generating the sets of conserved and random could be developed. But it is of the opinion of the author and of the principal investigator that such a line of questioning leads to an unnecessary complication of protocol, at least at this stage. Both the conserved network generation and the native contact definition algorithms are kept simple.

The next step in this research is to test whether these conserved networks guide the formation of the Greek-Key topology as proteins fold. Progress in this area has already been made. Greene has conducted preliminary folding simulations (unpublished data) using conventional simulated-annealing structure-determination protocols with the Crystallography & NMR System software package (125). Normally, these programs fold a structure using an entire set native contact distances as restraints. However, correct Greek-Key topologies are being generated when using only conserved contact distances as restraints. These preliminary results should be checked with other software packages (e.g. CHARMM).

Also, work is underway on finding the common interaction network among the three superfamilies studied here (i.e. the all α-helix death domains, the α/β-plaits and the all β-sheet immunoglobulins). The differences in tertiary structure among these superfamilies make aligning the proteins very difficult. With the structures aligned, Monte Carlo methods are being used to find the common conserved contacts.

CHAPTER V

SUMMARY

The molecular dynamics simulations conducted in this dissertation sheds light on how DHFR M20 loop conformational changes and Greek-Key protein foldings occur. The absence of time-resolved X-ray structures for these processes required the simulation accuracy to be rigorously tested via other indirect comparisons to experimental data (e.g. by measuring the backbone RMSD relative to the native structure to make sure the simulated structure maintains a native-like shape). The conformational changes and foldings were both characterized by the stabilizing properties that prevented them from occurring. In the DHFR simulations, the stabilizing hydrogen bonds between the M20 loop and other nearby loops kept the M20 loop from changing its conformation. Their loss resulted in corresponding shifts within the M20 loop. In the Greek-Key protein simulations, conserved contacts were found to be more important than other native contacts in maintaining the overall native-like topology. Loss of these contacts resulted in the protein unfolding. These two cases demonstrate the symbiotic relationship of change and stability.

As discussed in Chapter I, direct observation of phenomena is invaluable in gaining a good understanding of it. Correct understanding leads to the construction of accurate models. But also the process of producing accurate models leads to better understanding. This dissertation work required the production of accurate models of protein motion. Once used, these models generated very detailed descriptions where the positions of all atoms were solved with angstrom-level precision at different points, and on a picosecond-level timescale. That level detail and finely-spaced timescale allowed the protein to be 'observed' moving in real time. The generated simulation trajectories allowed the dissertation hypotheses to be attacked in a brute-force manner. In other words, the simulations generated enough (actually, much more than enough) information to answer the questions posed in the Introduction sections of Chapters III and IV, the only problem was how to go about retrieving that information.

The DHFR project required knowledge of the motion of the M20 loop residues, and the simulation trajectories provided the motion of every atom. Given that information, the M20 loop conformation was able to be characterized by several different methods: visual renderings, backbone RMSDs, M20 loop N-terminus and C-terminus backbone RMSDs, hydrogen-bond distances, alpha-carbon RMSFs, residue centroid difference distance matrices and free energy analyses. The list far from exhausts all possible analyses; nevertheless it does allow a thorough description of the M20 loop behavior in the different complexes.

The Greek-Key project required knowledge of the overall backbone topology and many inter-residue distances as a function of temperature; again the simulation trajectories provided more than enough information. Visual renderings and backbone RMSDs helped identify which proteins remained folded in their native-like topology. Various structures along the trajectories and calculated minimized-average structures were analyzed by intensive contact analyses methods. Most of the analyses involved using knowledge of the atomic positions to calculate the contact distances in the random and conserved interaction lists. These calculations were taken a step further to yield the average conserved contact distance and the average random contact distance – that is only two values were retrieved using knowledge of all atomic positions. But the brute force method worked. Again, there were many other methods that could have showed that the conserved contacts behaved differently than the random contacts do. One method that was performed, but not described in this dissertation, was the use of calculated residue-residue correlation matrices to show that conserved contacts move in correlation more than random contacts do.

The idea is that these models provided a complete enough picture of the proteins' behavior, and therefore understanding the models allowed the proteins to be understood. The hypotheses of this dissertation were examined to a satisfactory level, as discussed in the Conclusion sections of chapters III and IV. But many aspects of protein motion are still not well understood, and no models will ever be completely accurate. However, the simulations described in this dissertation did allow direct observation of the modeled phenomena, with mechanistic detail unlike any known laboratory experiment today. The future development of new simulation tools will result in further and better understanding of these phenomena. Because theoretical models are never proven, only disproven, a perfect understanding of the phenomena they describe will always remain elusive.

REFERENCES

1. Bolton W, Cox JM, Perutz MF (1968) Structure and function of haemoglobin. IV. A three-dimensional Fourier synthesis of horse deoxyhaemoglobin at 5.5 A resolution. *J Mol Biol* 33:283-297.

2. Perutz MF, Miurhead H, Cox JM, Goaman LC, Mathews FS, McGandy EL, Webb LE (1968) Three-dimensional Fourier synthesis of horse oxyhaemoglobin at 2.8 A resolution: (I) X-ray analysis. *Nature* 219:29-32.

3. Humphrey W, Dalke A, Schulten K (1996) VMD – Visual Molecular Dynamics. *J Molec Graphics* 14:33-38.

4. Vitaglioano L, Vergara A, Bonomi G, Merlino A, Verde C, di Prisco G, Howes BD, Smulevich G, Mazzarella L (2008) Spectroscopic and crystallographic characterization of a tetrameric hemoglobin oxidation reveals structural features of the functional intermediate relaxed/tense state. *J Am Chem Soc* 130:10527-10535.

5. Guallar V, Jarzecki AA, Friesner RA, Spiro TG (2006) Modeling of ligation-induced helix/loop displacements in myoglobin: toward an understanding of hemoglobin allostery. *J Am Chem Soc* 128:5427-5435.

6. Wider G (2000) Structure determination of biological macromolecules in solution using NMR spectroscopy. *Biotechniques* 29:1278-1294.

7. Smyth MS, Martin JHJ (2000) x Ray crystallography. *J Clin Pathol – Mol Pathol* 53:8-14.

8. Liu HL, Hsu JP (2005) Recent developments in structural proteomics for protein structure determination. *Proteomics* 5:2056-2068.

9. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D (2007) A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* 450: 913-916.

10. Moffat K (2001) Time-resolved biochemical crystallography: a mechanistic perspective. *Chem Rev* 101:1569-1581.

11. Anfinrud P, Schotte F (2005) X-ray fingerprinting of chemical intermediates in solution. *Science* 309:1192-1193.

12. Bourgeois D, Schotte F, Brunori M, Vallone B (2007) Time-resolved methods in biophysics. 6. Time-resolved Laue crystallography as a tool to investigate photo-activated protein dynamics. *Photochem Photobiol Sci* 6:1047-1056.

13. Bourgeois D, Weik M (2009) Kinetic protein crystallography: a tool to watch proteins in action. *Crystallogr Rev* 15:87-118.

14. Schotte F, Lim M, Jackson TA, Smirnov AV, Soman J, Olson JS, Phillips GN, Wulff M, Anfinrud PA (2003) Watching a protein as it functions with 150-ps time-resolved X-ray crystallography. *Science* 300:1944-1947.

15. Friedrich S, Jayashree S, Olson JS, Wulff M, Anfinrud PA (2004) Picosecond time-resolved X-ray crystallography: probing protein function in real time. *J Struc Biol* 147:235-246.

16. Gerhard H, Friedrich S, Anfinrud PA (2004) Unveiling functional protein motions with picosecond X-ray crystallography and molecular dynamics simulations. *Proc Natl Acad Sci USA* 101:15330-15334.

17. Aranda R IV, Levin EJ, Schotte F, Anfinrud PA, Phillips GN Jr (2006) Time-dependent atomic coordinates for the dissociation of carbon monoxide from myoglobin. *Acta Crystallogr D* 62:776-783.

18. Bourgeois D, Vallone B, Arcovito A, Sciara G, Schotte F, Anfinrud PA, Brunori M (2006) Extended subnanosecond structural dynamics of myoglobin revealed by Laue crystallography. *Proc Natl Acad Sci USA* 103:4924-4929.

19. Cammarata M, Levantino M, Schotte F, Anfinrud PA, Ewald F, Choi J, Cupane A, Wulff M, Ihee H (2008) Tracking the structural dynamics of proteins in solution using time-resolved wide-angle X-ray scattering. *Nat Methods* 5:881-886.

20. Carver TE, Brantley RE Jr, Singleton EW, Arduini RM, Quillin ML, Phillips GN Jr, Olson JS (1992) A novel site-directed mutant of myoglobin with an unusually high O_2 affinity and low autooxidation rate. *J Biol Chem* 267:14443-14450.

21. Lim MH, Jackson TA, Anfinrud PA (1995) Midinfrared vibrational-spectrum of CO after photodissociation from heme: Evidence for a ligand docking site in the heme pocket of hemoglobin and myoglobin. *J Chem Phys* 102:4355-4366.

22. Schlichting I, Berendzen J, Phillips GN Jr, Sweet RM (1994) Crystal structure of photolysed carbonmonoxy-myoglobin. *Nature* 371:808-812.

23. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 9:646-652.

24. Karplus M, Kuriyan J (2005) Molecular dynamics and protein function. *Proc Natl Acad Sci USA* 102: 6679-6685.

25. Scheraga HA, Khalili M, Liwo A (2007) Protein-folding dynamics: overview of molecular dynamics simulation techniques. *Annu Rev Phys Chem* 58:57-83.

26. Kincaid D, Cheney EW (2009) *Numerical Analysis: Mathematics of Scientific Computing,* (AMS Bookstore, Providence), *3rd Ed.*

27. Press WH, Teukolsky SA, Vetterling WT, Glannery BP (2007) *Numerical Recipes - The Art of Scientific Computing,* (Cambridge University Press, New York), *3rd Ed.*

28. Trefethen LN (1997) *Numerical Linear Algebra,* (SIAM, Philadelphia).

29. Alder BJ, Wainwright TE (1957) Phase transition for a hard sphere system. *J Chem Phys* 27:1208.

30. Alder BJ, Wainwright TE (1959) Studies in molecular dynamics. I. General Method. *J Chem Phys* 31:459.

31. Allen MP, Tildesley DJ (1989) *Computer Simulation of Liquids*, (Oxford University Press, Oxford).

32. Frenkel D, Smit B (2002) *Understanding Molecular Simulation – From Algorithms to Applications,* (Academic Press, San Diego), *2nd Ed.*

33. Leach AR (2001) *Molecular Modelling – Principles and Applications,* (Pearson Education Limited, Harlow), *2nd Ed.*

34. MacKerell AD Jr, Bashford D, Bellott M, Dunbrack RL Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE III, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102:3586-3616.

35. MacKerell AD Jr, Feig M, Brooks CL III (2004) Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* 25:1400-1415.

36. Wang J, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem* 21:1049-1074.

37. Schuler LD, Daura X, van Gunsteren WF (2001) An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J Comput Chem* 22:1205-1218.

38. Jorgensen WL, Tirado-Rives J (1988) The OPLS force field for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc* 110:1657-1666.

39. Jorgensen WL, Maxwell S, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* 117:11225-11236.

40. Rizzo C, Jorgensen W (1999) OPLS all-atom model for amines: resolution of the amine hydration problem. *J Am Chem Soc* 121:4827-4836.

41. Hammes-Schiffer S, Benkovic SJ (2006) Relating protein motion to catalysis. *Annu Rev Biochem* 75:519-541.

42. Schlitter J, Engels M, Kruger P, Jacoby E, Wollmer A (1993) Targeted molecular dynamics simulation of conformational change – application to the T-R transition in insulin. *Mol Simulation* 10:291-308.

43. van der Vaart A, Karplus M (2005) Simulation of conformational transitions by the restricted perturbation-targeted molecular dynamics method. *J Chem Phys* 122:114903.

44. Beveridge DL, DiCapua FM (1989) Free-energy via molecular simulation—applications to chemical and biomolecular systems. *Annu Rev Biophys Biomol Struc* 18:431-492.

45. Straatsma TP, McCammon JA (1992) Computational alchemy. *Annu Rev Phys Chem* 43:407-435.

46. Gilson MK, Given JA, Bush BL, McCammon JA (1997) The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys J* 72:1047-1069.

47. Gilson MK, Zhou HX (2007) Calculation of protein-ligand binding affinities. *Annu Rev Biophys Biomol Struc* 36:21-42.

48. Straatsma TB, McCammon JA (1991) Multiconfiguration thermodynamic integration. *J Chem Phys* 95:1175-1188.

49. Vorobjev YN, Hermans J (1999) ES/IS: estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model. *Biophys Chem* 78:195-205.

50. Aqvist J, Luzhkov VB, Brandsdal BO (2002) Ligand binding affinities from MD simulations. *Acct Chem Res* 35:358-365.

51. Srinivasan J, Cheatham TE, Cieplak P, Kollman PA, Case DA (1998) Continuum solvent studies of the stability of DNA, RNA and phosphoramidate-DNA helices. *J Am Chem Soc* 120:9401-9409.

52. Massova I, Kollman PA (1999) Computational alanine scanning to probe protein-protein interactions: a novel approach to evaluate binding free energies. *J Am Chem Soc* 121:8133-8143.

53. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE III (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res* 33:889-897.

54. Wang W, Kollman PA (2001) Computational study of protein specificity: the molecular basis of HIV-1 protease drug resistance. *Proc Natl Acad Sci USA* 98:14937-14942.

55. Swanson JMJ, Henchman RH, McCammon JA (2004) Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys J* 86:67-74.

56. Kirkwood JG (1935) Statistical mechanics of fluid mixtures. *J Chem Phys* 3:300-313.

57. Wilson EB, Decius JC, Cross PC (1955) *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra,* (General Publishing Company, Ltd, Toronto).

58. Woodward LA (1972) *Introduction to the Theory of Molecular Vibrations and Vibrational Spectroscopy*, (Clarendon Press, Oxford).

59. Brooks B, Karplus M (1985) Normal modes for specific motions of macromolecules: Application to the hinge-bending motion of lysozyme. *Proc Natl Acad Sci USA* 82:4995-4999.

60. Tama F, Sanejouand YH (2001) Conformational change of proteins arising from normal mode calculations. *Protein Eng* 14:1-6.

61. Dobbins SE, Lesk VI, Sternberg MJE (2008) Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc Natl Acad Sci USA* 105:10390-10395.

62. Petrone P, Pande VS (2006) Can conformational change be described by only a few normal modes. *Biophys J* 60:1583-1593.

63. McGuire JJ (2003) Anticancer antifolates: current status and future directions. *Curr Pharm Des* 9: 2593-2613.

64. Hitchings GH (1989) Selective inhibitors of dihydrofolate reductase. *Angew Chem Int Ed* 28: 879-885.

65. Hitchings GH, Elion GB, VanderWerff H, Falco EA (1948) Pyrimidine derivatives as antagonists of pteroyglutamic acid. *J Biol Chem* 174: 765-766.

66. Burchenal JH, Karnofsky DA, Kingsley-Pillers EM, Southam CM, Myers WPL, Escher GC, Craver LF, Dargeon HW, Rhoads CP (1951) The effects of the folic acid antagonists and 2,6-diaminopurine on neoplastic disease, with special reference to acute leukemia. *Cancer* 4:549.

67. Sawaya MR, Kraut J (1997) Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: crystallographic evidence. *Biochemistry* 36:586-603.

68. Zaborowski E, Chung J, Kroon GJ, Dyson HJ, Wright PE (2000) Backbone H^N, C^α, C' and C^β assignments of the 19 kDa DHFR/NADPH complex at 9 °C and pH 7.6. *J Biomol NMR* 16:349-350.

69. Osborne MJ, Venkitakrishnan RP, Dyson HJ, Wright PE (2003) Diagnostic chemical shift markers for loop conformation and cofactor binding in dihydrofolate reductase complexes. *Protein Sci* 12:2230-2238.

70. Osborne MJ, Schnell J, Benkovic SJ, Dyson HJ, Wright PE (2004) Backbone dynamics in dihydrofolate reductase complexes: role of loop flexibility in the catalytic mechanism. *Biochemistry* 40:9846-9859.

71. Venkitakrishnan RP, Zaborowski E, McElheny D, Benkovic SJ, Dyson HJ, Wright PE (2004) Conformational changes in the active site loop of dihydrofolate reductase during the catalytic cycle. *Biochemistry* 43:16046-16055.

72. Miller GP, Benkovic SJ (1998) Strength of an interloop hydrogen bond determines the kinetic pathway in catalysis by escherichia coli dihydrofolate reductase. *Biochemistry* 37:6336-6342.

73. Miller GP, Wahnon DC, Benkovic SJ (2001) Interloop contacts modulate ligand cycling during catalysis by escherichia coli dihydrofolate reductase. *Biochemistry* 40:867-875.

74. McElheny D, Schnell JR, Lansing JC, Dyson HJ, Wright PE (2005) Defining the role of active-site loop fluctuations in dihydrofolate reductase catalysis. *Proc Natl Acad Sci USA* 102:5032-5037.

75. Palmer AG (2004) NMR characterization of the dynamics of biomolecules. *Chem Rev* 104:3623-3640.

76. Boehr DD, McElheny D, Dyson HJ, Wright PE (2006) The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* 313:1638-1642.

77. Fierke CA, Johnson KA, Benkovic SJ (1987) Construction and evaluation of the kinetic scheme associated with dihydrofolate reductase. *Biochemistry* 26:4085-4092.

78. Kitahara R, Sareth S, Yamada H, Ohmae E, Gekko K, Akasaka K (2000) High pressure NMR reveals active-site hinge motion of folate-bound escherichia coli dihydrofolate reductase. *Biochemistry* 39:12789-12795.

79. Radkiewicz JL, Brooks CL III (2000) Protein dynamics in enzymatic catalysis: exploration of dihydrofolate reductase. *J Am Chem Soc* 122:225-231.

80. Shrimpton P, Allemann RK (2002) Role of water in the catalytic cycle of E. coli dihydrofolate reductase. *Prot Sci* 11:1442-1451.

81. Rod TH, Radkiewicz JL, Brooks CL III (2003) Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proc Natl Acad Sci USA* 100:6980-6985.

82. Khavrutskii IV, Price DJ, Lee J, Brooks CL III (2007) Conformational change of the methionine 20 loop of escherichia coli dihydrofolate reductase modulates pKa of the bound dihydrofolate. *Protein Sci* 16:1087-1100.

83. Sham YY, Ma B, Tsai CJ, Nussinov R (2002) Thermal unfolding molecular dynamics simulation of escherichia coli dihydrofolate reductase: thermal stability of protein domains and unfolding pathway. *Proteins Struct Funct Bioinf* 46:308-320.

84. Swanwick RS, Shrimpton PJ, Allemann RK (2004) Pivotal role of Gly 121 in dihydrofolate reductase from escherichia coli: the altered structure of a mutant enzyme may form the basis of its diminished catalytic performance. *Biochemistry* 43:4119-4127.

85. Castillo R, Andres J, Moliner V (1999) Catalytic mechanism of dihydrofolate reductase enzyme. A combined quantum-mechanical/molecular-mechanical characterization of transition state structure for the hydride transfer step. *J Am Chem Soc* 121:12140-12147.

86. Agarwal PK, Billeter SR, Rajagopalan R, Benkovic SJ, Hammes-Schiffer S (2002) Network of coupled promoting motions in enzyme catalysis. *Proc Nat Acad Sci* 99:2794-2799.

87. Cummins PL, Greatbanks SP, Rendell AP, Gready JE (2002) Computational methods for the study of enzymic reaction mechanisms. 1. Application to the hydride transfer step in the catalysis of dihydrofolate reductase. *J Phys Chem B* 106:9934-9944.

88. Agarwal PK, Billeter SR, Hammes-Schiffer S (2002) Nuclear quantum effects and enzyme dynamics in dihydrofolate reductase catalysis. *J Phys Chem B* 106:3283-3293.

89. Thorpe IF, Brooks CL III (2003) Barriers to hydride transfer in wild type and mutant dihydrofolate reductase from E coli. *J Phys Chem B* 107:14042-14051.

90. Watney JB, Agarwal PK, Hammes-Schiffer S (2003) Effect of mutation on enzyme motion in dihydrofolate reductase. *J Am Chem Soc* 125:3745-3750.

91. Garcia-Viloca M, Truhlar DG, Gao J (2003) Reaction-path energetics and kinetics of the hydride transfer reaction catalyzed by dihydrofolate reductase. *Biochemistry* 42:13558-13575.

92. Thorpe IF, Brooks CL III (2004) The coupling of structural fluctuations to hydride transfer in dihydrofolate reductase. *Proteins Struct Funct Bioinf* 57:444-457.

93. Wong KF, Watney JB, Hammes-Schiffer S (2004) Analysis of electrostatics and correlated motions for hydride transfer in dihydrofolate reductase. *J Phys Chem B* 108:12231-12241.

94. Thorpe IF, Brooks CL III (2005) Conformational substates modulate hydride transfer in dihydrofolate reductase. *J Am Chem Soc* 127:12997-13006.

95. Sergi A, Watney JB, Wong KF, Hammes-Schiffer S (2006) Freezing a single distal motion in dihydrofolate reductase. *J Phys Chem B* 110:2435-2441.

96. Arora K, Brooks CL III (2009) Functionally important conformations of the Met20 loop in dihydrofolate reductase are populated by rapid thermal fluctuations. *J Am Chem Soc* 131:5642-5647.

97. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) CHARMM: A program for macromolecular energy, minimization and dynamics calculations. *J Comput Chem* 4:187-217.

98. MacKerell, AD Jr, Brooks BR, Brooks CL III, Nilsson L, Roux B, Won Y, Karplus M (1998) CHARMM: The energy function and its parameterization with an overview of the program. In: The Encyclopedia of Computational Chemistry (John Wiley & Sons, Chichester), pp 271-277.

99. Brooks BR, Brooks CL III, Mackerell AD, Nilsson L, Petrella J, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: the biomolecular simulation program. *J Comput Chem* 30:1545-1615.

100. Radkiewicz JL, Brooks CL III (2000) Protein dynamics in enzymatic catalysis: exploration of dihydrofolate reductase. *J Am Chem Soc* 122:225-231.

101. Pavelites JJ, Bash PA, Gao J, MacKerell Jr AD (1997) A molecular-mechanics force field for NAD+, NADH and the pyrophosphate groups of nucleotides. *J Comput Chem* 18:221-239.

102. Apostolakis J, Ferrara P, Caflisch A (1999) Calculation of conformational transitions and barriers in solvated systems: Application to the alanine dipeptide in water. *J Chem Phys* 110:2099-2108.

103. Flynn TC, Swint-Kruse L, Kong Y, Booth C, Matthews KS, Ma J (2003) Allosteric transition pathways in lactose repressor protein core domains: Asymmetric motions in a homodimer. *Protein Sci* 12:2523-2541.

104. Lempiainen H, Molnar F, Gonzalez MM, Perakyla M, Carlberg C (2005) Antagonist- and inverse agonist-driven interactions of the vitamin D receptor and the constitutive androstane receptor with corepressor protein. *Mol Endocrinol* 19:2258-2272.

105. Hu J, Ma A, Dinner AR (2008) A two-step nucleotide-flipping mechanism enables kinetic discrimination of DNA lesions by AGT. *Proc Natl Acad Sci USA* 105:4615-4620.

106. Valadie H, Lacapcre JJ, Sanejouand YH, Etchebest C (2003) Dynamic properties of the MscL of Escherichia coli: a normal mode analysis. *J Mol Biol* 332:657-674.

107. Cayley, Dunn, King (1981) Kinetics of substrate, coenzyme, and inhibitor binding to Escherichia coli dihydrofolate reductase. *Biochemistry* 20:874-879.

108. Elber R (2005) Long-timescale simulation methods. *Curr Opin Struct Biol* 15:151-156.

109. Zheng W, Brooks BR (2005) Normal-modes-based prediction of protein conformational changes guided by distance constraints. *Biophys J* 88:3109-3117.

110. Zheng W, Brooks BR (2006) Modeling protein conformational changes by

iterative fitting of distance constraints using reoriented normal modes. *Biophys J* 90:4327-4336.

111. Cuff A, Redfern OC, Greene L, Sillitoe I, Lewis T, Dibley M, Reid A, Pearl F, Dallman T, Todd A, Garratt R, Thornton J, Orengo C (2009) The CATH hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. *Structure* 17:1051-1062.

112. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH—a hierarchic classification of protein domain structures. *Structure* 5:1093-1108.

113. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* 35:D291-D297.

114. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536-540.

115. Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G (1992) A database of protein structure families with common folding motifs. *Prot Sci* 1:1691-1698.

116. Holm L, Sander C (1998) Dictionary of recurrent domains in protein structures. *Proteins* 33:88-96.

117. Higman VA, Greene LG (2006) Elucidation of conserved long-range interaction networks in proteins and their significance in determining protein topology. *Physica A* 368:595-606.

118. Strogatz SH (2001) Exploring complex networks. *Nature* 410:268-276.

119. Albert R, Jeong H, Barabasi A (1999) Diameter of the world-wide web. *Nature* 401:130-131.

120. Wolf YI, Karev G, Koonin EV (2002) Scale-free networks in biology: new insights into the fundamentals of evolution? *BioEssays* 24:105-109.

121. Greene LH, Higman VA (2003) Uncovering network systems within protein structures. *J Mol Biol* 334:781-791.

122. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11:739-747.

123. Li H, Wojtaszek JL, Greene LH (2009) Analysis of conservation in the Fas-associated death domain protein and the importance of conserved tryptophans in structure, stability and folding. *Biochim Biophys Acta, Proteins Proteomics* 1794:583-593.

124. Greene LH, Hamada D, Eyles SJ, Brew K (2003) Conserved signature proposed for folding in the lipocalin superfamily. *FEBS Letters* 553:39-44.

125. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998) Crystallography & NMR System: a new software suite for macromolecular structure determination. *Acta Cryst* D54:905-921.

VITA

Jeffrey Andrew Tibbitt
Department of Chemistry and Biochemistry
Old Dominion University
4541 Hampton Boulevard
Norfolk, VA 23529-0126

Education:

(2010) Ph.D. Chemistry	Old Dominion University
(2001) B.S. Chemistry	Virginia Polytechnic & State University
(1998) A.S.	Tidewater Community College
(1997) A.A.S. Culinary Arts	Johnson & Wales University

Presentations:

"Calculating relative free energies of M20 loop conformers of E. coli dihydrofolate reductase", Tibbitt, J.; Radkiewicz, J. L. Abstracts, 36th Middle Atlantic Regional Meeting of the American Chemical Society, Princeton, NJ, June 8-11, 2003 (poster presentation).

"Conserved networks underlying protein structure, folding and dynamics", Greene, L. H.; Pothen, J.; Tibbitt, J. A.; Li, H. Division of Computers in Chemistry, 238th ACS National Meeting, Washington D.C., August 16-20, 2009 (oral presentation).

"Investigating M20 loop conformers of dihydrofolate reductase", Tibbitt, J. A.; Poutsma, J. Division of Computers in Chemistry, 233th ACS National Meeting, Chicago, IL, March 25-29, 2007 (poster presentation).

Publications:

Tibbitt JA, Pothen JJ, Morris JD, Will AD, Morris C, Ahmed T, Poutsma J, Munyanyi A, Li H, Osgood C, Luo LS, Greene LH (2010) Conserved networks in proteins govern topology by encoding a fold-determining core. (in preparation)

Tibbitt JA, Shi C, Poutsma J (2010) Ligand effect on M20 loop conformation of E. coli dihydrofolate reductase. (in preparation)

Shi C, Tibbitt JA, Poutsma J (2010) Molecular dynamics on R67 dihydrofolate reductase. (in preparation)