Generating Stereochemically Acceptable Protein Pathways

by

Daniel W. Farrell

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

ARIZONA STATE UNIVERSITY

December 2010

Generating Stereochemically Acceptable Protein Pathways

by

Daniel W. Farrell

has been approved

September 2010

Graduate Supervisory Committee:

Michael F. Thorpe, Chair
Timothy J. Newman
Sefika Banu Ozkan
Marcia Levitus
Dmitry V. Matyushov

ACCEPTED BY THE GRADUATE COLLEGE

ABSTRACT

Understanding how proteins move from one conformation to another is critical for understanding how proteins perform their functions. A new computational method for rapid generation of all-atom pathways between two given conformations of a protein is presented, called geometric targeting (GT), available for use through an interactive web interface. The method is based on the philosophy that many essential features of motion in proteins can be determined solely by considering geometric relationships between atoms. A pathway is generated by pulling the system from initial coordinates towards a set of target coordinates, while enforcing geometric constraints to maintain covalent bond geometry, avoid overlap of atoms, avoid outlier Ramachandran regions and eclipsed torsion angles, and preserve hydrogen bonds and hydrophobic contacts. The pathways are not optimal in a minimum energy sense or a high flux sense; instead, they are stereochemically plausible all-atom pathways intended to give rapid insight into candidate motions for protein conformational changes. The method is applied to over twenty proteins and protein complexes, demonstrating the ability to handle large systems and highly non-linear motions. Pathways from GT for the protein “nitrogen regulatory protein C” are compared to pathways from the more traditional targeted molecular dynamics method, demonstrating that GT finds essentially the same motions as targeted molecular dynamics for this system in a factor of about 1000 less computational time. Current applications of the methodology include the following: input to umbrella sampling free energy calculations, cryo-electron microscopy structure fitting, and protein folding.

To Mom, Dad, Rebecca, Hannah, Angela, Joseph, and Steven, who make life wonderful, and to God in gratitude for my family and for all the happiness in my life

ACKNOWLEDGMENTS

I would first like to express deep thanks to my adviser Mike Thorpe for all the time and energy he has patiently invested to help me reach this point, for all the debates and discussions that have molded my thinking, for allowing me to pursue my own ideas, for giving me the pushes I needed to finish, and for helping and supporting me in so many ways. I also thank my adviser for the many years of financial support and for all the wonderful travel opportunities and he has provided over the years, to Europe, Barbados, and several conferences around the nation.

I express sincere thanks to Stephen A. Wells for sharing ideas and for patiently helping me to learn FRODA when he was a post-doc at ASU. I am indebted to Stephen's truly creative FRODA model, which sparked my own ideas that eventually led to this work. I also thank and acknowledge Brandon Hespenheide for many useful discussions about FIRST. I thank Kirill Speranskiy for the many hours of work he put into the webservers and script development, and for being flexible with changing demands.

Many collaborators made this work possible. I am very grateful to Ming Lei for noticing similarity between pathways of geometric targeting and targeted molecular dynamics and for all the subsequent insights and assistance he provided in the pathway comparison project. I thank Maria Kurnikova for seeing the potential of connecting geometric targeting with molecular dynamics and for all the hard work and computer resources she and Tatyana Mamonova have put into running molecular dynamics simulations for the umbrella sampling project. Two applications of the new geometric simulation method that are mentioned in this dissertation are the work of Tyler Glembo and S. Banu Ozkan (Zipping and Assembly Method of protein folding), and Craig C. Jolley (Cryo-EM structure fitting). I also thank Ileana Streinu for many interesting discussions about rigidity theory and for organizing the Barbados Workshop, and hosting me in Massachusetts. I owe much of my understanding of rigidity theory to Audrey Lee

and thank her for helpful conversations and her thorough dissertation. I thank Holger Gohlke for being a wonderful host in Germany and for collaborating on a project that eventually led me to the work of this dissertation. Thanks to Arjan van der Vaart for helpful discussions about this research.

I am deeply honored by the extremely generous fellowship provided by Diane and Tom Might of the ARCS Foundation, and I will remember to pass on the gift to someone else someday. I am likewise very grateful to the generous scholarships provided by Wally Stoelzel and the Molecular Imaging Corp. I thank the Graduate College for the Dissertation Fellowship Award which has financially supported me through much of this final year. I thank the Graduate Professional Student Association, the Graduate College, and Boehringer Ingelheim Fonds for travel funding. Special thanks to NSF/DOE for sending me to Lindau, an experience I will never forget. Research funding has been provided through NSF Grant DMS-0714953.

I thank my committee members for devoting the time to read through this dissertation and for their very helpful suggestions that have greatly improved this document. I especially thank Timothy Newman for countless entertaining and helpful discussions about biology, physics, careers, and academia. Thanks also to Banu Ozkan for her wisdom and advice over the years.

I thank the biophysics grad students at ASU for always being willing to share knowledge. Big thanks to Justin Spiriti for help in understanding NMR, and to Adam de Graff for his teamwork with many projects and for helpful discussions about new work in protein unfolding pathways. Thanks to Tyler Glembo for coming up with a great software name—FRODAN. Thanks to Jill Kolp for being a wonderful and helpful administrator.

Finally, a big thanks to my sister Angela Rose for help with formatting in Adobe InDesign, and to my dad for alerting me to relevant research articles.

TABLE OF CONTENTS

APPENDIX Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

CHAPTER 1 INTRODUCTION

Proteins are functional molecules built by cells to carry out the myriad of tasks vital for life. Many cellular activities are either performed or orchestrated by proteins, such as sensing the external environment, endocytosis, exocytosis, intra-cellular transport, metabolism, signal transduction and amplification, DNA replication and transcription, cell division, and cell motility (1-3). Proteins also underlie the operations of organs and tissues in higher organisms, acting as photoreceptor molecules in the eye (e.g., rhodopsin), contractile fibers in muscles (e.g., actin, myosin, and titin), digestive enzymes in the stomach and intestine (e.g., pepsin), and controlling action potentials along neuronal axons to communicate sensory input or activate muscles (e.g., voltage gated K^+,Na^+ channels) (1-3). Protein molecules come in a wide range of sizes (typically from hundreds of atoms to hundreds of thousands of atoms) and a wide range of shapes, each protein having structural and chemical features uniquely suited for it to carry out its particular function (1-3).

A key property of many proteins and protein assemblies is the ability of the structure to switch between various conformational states. A dramatic example is ATP synthase (2, 4), an assembly of 21-24 individual protein subunits that function together as a motor, with a rotating “rotor” unit and a stationary “stator” unit. A proton gradient drives rotation of the rotor, which is coupled to synthesis of ATP from ADP and inorganic phosphate. Another example is cowpea chlorotic mottle virus capsid (5, 6), an icosahedrally-symmetric assembly of 180 identical protein subunits that undergoes a large-scale swelling transition in which all the proteins move radially outward and rotate, expanding the capsid radius by 7-12%. Adenylate kinase is a monomeric protein with hinged domains that open and close like a clamshell as part of the enzyme’s catalytic cycle (7-9). Some membrane proteins have diaphragm-like gating mechanisms that can dilate or constrict dynamically to regulate the flow of ions or other small molecules

across a membrane [for example, bacterial K^+ channel (3)]. Some proteins possess flexible binding sites that adjust to optimally bind distinct ligands [e.g., antibody SPE7 (10)]. Conformational change is often an enabling feature that endows proteins and protein assemblies with functional capability.

An essential starting point for understanding the inner-workings of a protein is a determination of its 3D structure. Structure determination comes primarily from Xray crystallography (11), Nuclear Magnetic Resonance (NMR) spectroscopy (12), and Cryo-electron Microscopy (Cryo-EM) (13). Generally, the end result of one of these experiments is a 3D structure of a single conformational state of the protein (or a tightly clustered ensemble in the case of NMR), often with atomic-level resolution (Cryo-EM experiments are typically more coarse in resolution). Under different experimental conditions, such as by introducing a known binding partner (ligand), it is often possible to capture a protein in a different conformational state. By examining two or more static structural snapshots of a protein it is possible to identify regions of the protein that have moved, just as one might examine two still frames taken out of a movie scene and notice things that have moved, but this an incomplete picture of the dynamic system.

The ideal would be to capture a time-resolved 3D trajectory of a protein to observe the ensemble of states that it samples, the timings of transitions, and to see the pathways by which it switches between various conformational states, but this is not currently possible. One technique on the horizon is 4D electron microscopy (14, 15), a stroboscopic technique for imaging a system in 3 spatial dimensions plus time, although its application to protein folding and dynamics has so far not been demonstrated (16). However, even without actually tracking the 3D coordinates of atoms, modern experimental techniques have other ways of shining a light on conformational dynamics in proteins and biomolecular systems. For example, recent developments in NMR relaxation experiments, which monitor the response of nuclear spins to perturbations in

magnetic fields, have now enabled detection of equilibrium conformational exchange between a (high-populated) ground state conformation of a protein and a (less-populated) excited-state conformation, and the determination of kinetic and thermodynamic variables such as rate constants and changes in enthalpy and entropy (17, 18). Förster Resonance Energy Transfer (FRET) experiments can detect in real time when two fluorescent-dye-labeled sites in a molecular system are close together or far apart, which has revealed, for example, "shuttling" motion of HIV reverse transcriptase along DNA (19), and transient partial unwinding of DNA from spindle-like DNA-packaging proteins known as histones (20).

Computational modeling and simulation can be used to gain insight into protein dynamics at the atomic level, complementing experiment, but several factors make modeling of proteins very challenging. Proteins are heterogeneous systems, neither solid nor liquid, with little or no internal symmetry. Atoms are densely and irregularly packed, with highly correlated movements between distant parts of the system. Thermodynamically they are only marginally stable, with a mixture of interactions from a wide range of energy scales being critical to stability. The number of atoms in a protein is too large to be modeled quantum mechanically, but small enough that the finite size and surface of the protein are non-negligible features. Arguably the most vexing difficulty is the disparity between biologically-relevant timescales (many conformational changes require milliseconds to seconds) and the timescales of the thermal fluctuations (tens of femtoseconds). The timescale disparity is precisely the problem with straightforward molecular dynamics simulations (MD), in which atoms are modeled as classical particles in a classical Hamiltonian, and Newtonian equations of motion are integrated to simulate the dynamics. The time step, which must be lower than the fastest atomic motions in the system (the thermal fluctuations), can be no larger than 1-2 fs for proteins to keep the simulation stable. This is so far removed from the biological timescale that in

most situations there is essentially zero chance of a significant conformational change spontaneously occurring in the limited biological time accessible to a regular MD simulation.

The timescale problem makes predicting all-atom conformational changes very difficult. There is a class of computational methods that aim to answer what should be a more tractable problem: given two known protein conformations, generate a 3D pathway between them. These range from sophisticated but computationally intensive methods that explore pathway space to find minimum energy pathways or optimal pathways at some finite temperature, to the simple linear interpolation-based approaches in which atoms can unphysically move through each other as their Cartesian coordinates are interpolated from beginning positions to target positions. Background on computational protein pathway methods will be presented in this chapter.

In this dissertation, we present a new method for rapid generation of all-atom pathways in proteins. The method, called geometric targeting (GT), is based on the simplifying assumption that pathways in proteins are largely determined by geometric relationships between atoms. This assumption reflects the reality that proteins are highly constrained systems, in which effects such as excluded volume of atoms, covalent bond geometry, and other geometric considerations, place significant restrictions on the conformations that proteins can access. GT models the protein as a geometrically constrained system, building on prior work in constraint-based protein models that will be reviewed in this chapter. The geometric constraints are distance and angle constraints between atoms, some of which are inequality constraints, e.g., a maximum or minimum distance, and some of which are equality constraints. The constraints partition conformational space into an “allowed” region that meets the constraints, and a “disallowed” region that violates the constraints. All-atom pathways are generated by incremental movements of the system from an initial state towards a target state,

while enforcing the geometric constraints to keep the atoms in plausible geometric arrangements. Key advantages GT over other rapid and approximate pathway methods are all-atom modeling and dynamic collision avoidance. GT is a directed sampling approach, rather than a dynamical approach, since there are no velocities, accelerations, kinetic or potential energies, and no time steps. The pathways are not optimized because GT does not sample thermodynamically, but the pathways are "stereochemically-acceptable" because of the geometric constraints that must be satisfied by each pathway snapshot. The niche for GT is "back-of-the-envelope" calculation of plausible all-atom pathways, using several orders of magnitude less computational time than more intensive approaches, particularly attractive for large systems where other methods require prohibitive amounts of computational resources.

The new methodology (geometric targeting) presented in this dissertation is a successor to the method known as FRODA (Framework Rigidity Optimized Dynamics Algorithm). The new methodology follows the same philosophy as FRODA, but with a fundamentally different geometric model and "engine" for sampling conformational space and enforcing constraints. Although in this dissertation we emphasize the "geometric targeting" method, the new engine of the method can be run without targeting, and in this mode the method is called "geometric simulation" following Wells et al. (21). At the risk of confusing the reader, the name of the new software that implements both the new geometric targeting method and the new geometric simulation method is FRODAN (with an N), in recognition of the heritage of the method. The N in FRODAN stands for New, emphasizing that the software itself is completely new, rather than a modification of the existing FRODA software. The first 5 letters in FRODAN do not stand for anything, as their original meaning no longer seems applicable in the current methodology.

The present chapter contains background information that will be relevant to the rest of the dissertation. We begin with an overview of geometric aspects of protein structure, which underlies the philosophy of the geometric protein model in this and prior work. We describe one experimental source for insight into protein dynamics, specifically Nuclear Magnetic Resonance, which will serve as background for the free energy calculation of Chapter 4. We give an overview of the spectrum of computational pathway generation of methods. We describe prior work in rigidity theory and constraint based modeling of proteins, including the FRODA method. We finish Chapter 1 with a description of some of the limitations in FRODA that led to the development of the current methodology (FRODAN) presented in this dissertation, and a summary of FRODAN.

Several of the chapters in this dissertation contain work that is collaborative. In the text of each chapter we shall clarify in more detail the roles of the various participants in the work and in the writing. Here we give an overview of the organization of the dissertation. Chapter 2 is a published paper (22) by Daniel W. Farrell, Kirill Speranskiy, and M. F. Thorpe. The paper describes the methodology of geometric targeting, demonstrates its successful application to a wide variety of protein conformational changes [including some examples known to produce unphysical pathways at the Yale Morph Server (23) which uses linear interpolation with energy minimization], and introduces a public webserver for easily generating pathways through a web interface (webserver and web-interface to FRODAN software developed by Kirill Speranskiy). Chapter 3 is a submitted paper currently under peer-review, by Daniel W. Farrell, Ming Lei, and M. F. Thorpe, in which we compare GT pathways to previously published TMD pathways (24) for a complicated transition in nitrogen regulatory protein C. We demonstrate that GT captures essentially the same motions and some similar relative timing of events as TMD in this system, but requires a factor $\sim 10^3$ less computational

time. We also calculate various structure quality metrics to show that the pathway snapshots produced by GT have good stereochemical quality. Chapter 4 presents some applications of GT pathways and FRODAN software. We show that GT pathways can be used as input pathways into umbrella sampling free energy calculations, facilitated by calibration of parameters in GT against molecular dynamics (MD), which is the work of Daniel W. Farrell, Tatyana Mamonova, Maria Kurnikova, and M. F. Thorpe (unpublished). Chapter 5 contains concluding remarks.

GEOMETRIC ASPECTS OF PROTEIN STRUCTURE

The geometric targeting method presented in this dissertation makes the assumption that geometric aspects of protein structures greatly restrict allowed conformations of the protein, and that by considering these geometric aspects plausible candidate pathways between conformations can be generated. Here we briefly summarize some of the geometric aspects of proteins structure that can be understood in terms of distance or angle restrictions. These observations form the basis of many of the constraint-based protein models described later in this chapter and in geometric targeting. Notably absent from this discussion is electrostatic attraction/repulsion, which certainly affects protein structure, but is not easily represented by distance or angle restrictions.

Proteins are composed of amino acid residues chained end-to-end by covalent bonds (Fig. 1.1). The main chain atoms of each residue are bonded to the main chain atoms of the next, forming a continuous chain of covalent bonds. Side chain atoms of each residue are covalently attached at the Cα atom. Covalent bonds between typical protein atoms (C, H, N, O, S) are very stable. For example, the sp^3 C-C bond energy is about 80 kcal/mol (2) (at 300 K, RT = 0.596 kcal/mol, so the C-C bond energy is about $130RT$). Covalent bonds impose strict distance and bending angle geometries between bonded neighbors. To get a feel for how rigid the covalent bonding geometry is, consider that in a classical molecular mechanical approximation, the bending spring constant

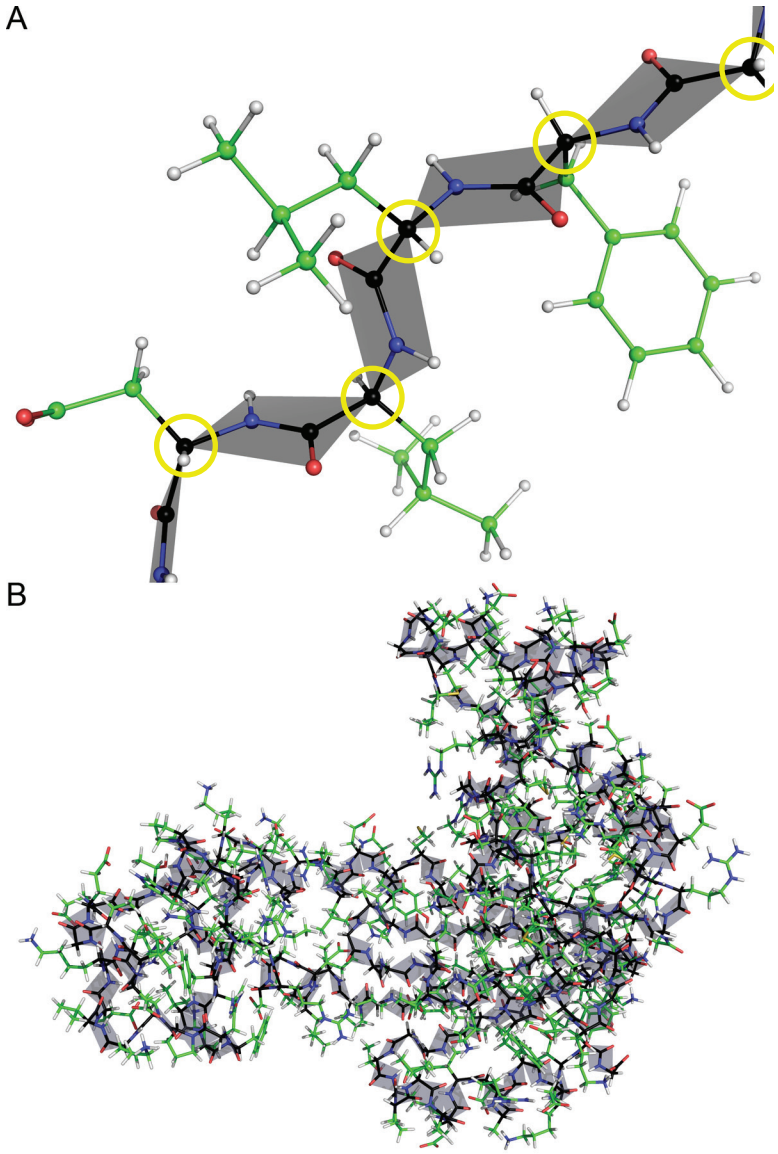

FIGURE 1.1 Geometric properties of protein chains. *(A)* shows a zoomed in view of the protein chain, with the "main chain" shown with black carbon atoms, and the "side chains" shown with green carbon atoms. All other colors are standard (white=hydrogen, blue=nitrogen, red=oxygen). Yellow rings mark the Cα atom of each amino acid residue. All bonds emanating from a Cα atom are rotatable, like hinges. Each Cα atom is connected via rotatable bonds to two rigid planar groups (amide planes, shown in gray) and to one side group shown in green. All bond distances and three-body angles are approximately rigid. *(B)* A view of an entire protein (adenylate kinase, Protein Data Bank ID 4AKE), colored the same as in *A* with the amide planes shown in gray, illustrating that the folded protein is made from a long chain.

between a pair of bonded *sp*³ carbons is $K_{\text{stretch}} = 310$ kcal/mol•Å^2, and the bending angle between H-C-H about an *sp*³ carbon has spring constant K_{bend}=35 kcal/mol•rad^2 (25). With these spring constants, the Boltzmann factor for stretching, $\exp\left(-K_{\text{stretch}}\Delta x^2 / RT\right)$, and the Boltzmann factor for bending, $\exp\left(-K_{\text{bend}}\Delta\theta^2 / RT\right)$, drop to 1% at a stretching of 0.09 Å or bending of 16° (at 300 K). Therefore, to a first approximation covalent bonds and angles can be considered to be rigid.

With fairly rigid covalent distance and angle geometry, the freedom for motion in proteins largely resides in the rotatability of dihedral angles about covalent bonds. Not all covalent bonds are rotatable, however. Single covalent bonds allow 360° dihedral rotations (although some dihedral angle values are more favored than others, as will be discussed below). Rotations about double bonds, or bonds with partial double bond character, are much more restricted than single bonds. One important case with partial double bond character is the peptide C-N bond, which joins consecutive amino acids in the protein. There are in principle two planar rotational states about a peptide bond, separated by 180°, known as *trans* and *cis* (in Fig 1.1 *A,* the planar groups shaded gray depict the *trans* configuration, with neighboring Cα atoms oriented diagonally from each other). The configuration known as *cis* (rotated by 180° relative to the *trans* configuration) is extremely rare (except in proline) (26), so to an approximation the peptide bond can be thought of as non-rotatable, being permanently in the *trans* configuration, yielding approximately rigid "amide planes" (2). Double (or partial double) bonds are also found in the side chains of amino acids, in planar ring configurations (for example, the hexagonal ring in phenylalanine, shown in Fig. 1.1 *A,* upper right) and planar non-ring configurations (such as the amide groups in glutamine). In planar rings, the individual partial double bonds do not permit 180° rotations because these would require breaking covalent bonds in the ring, which is energetically unfavorable. In planar non-ring groups, there are in principle two stable planar states about each (partial)

double bond, separated by 180° rotation; however, all (partial) double bonds in planar non-ring groups in the 20 standard amino acids have two-fold rotational symmetry, so the two states are indistinguishable [this observation is apparent from looking at the bond structures of the 20 standard amino acids, for example in (2)]. Therefore, in the 20 standard amino acids, double and partial double bonds only have one unique stable rotational state and as an approximation can be considered to be non-rotatable (except for the peptide bonds between adjacent residues).

In Fig 1.1 *A,* note that each Cα atom (circled in yellow) has two bonds extending along the main chain. These are single bonds and are therefore rotatable. The two rotational degrees of freedom along the backbone per Cα is what allows the linear protein chain to fold into a 3D conformation (Fig. 1.1 *B*). An important geometric aspect of the pair of rotatable backbone dihedral angles at each Cα is that certain combinations of angles are strongly disfavored. Fig. 1.2 shows the distribution of backbone dihedral angles from a recent survey of protein structures (27), known as a Ramachandran plot (28). The disallowed regions of Fig. 1.2 have a simple geometric explanation: they arise from pairs of backbone atoms in neighboring residues that clash (overlap) for certain combinations of dihedral angles (29).

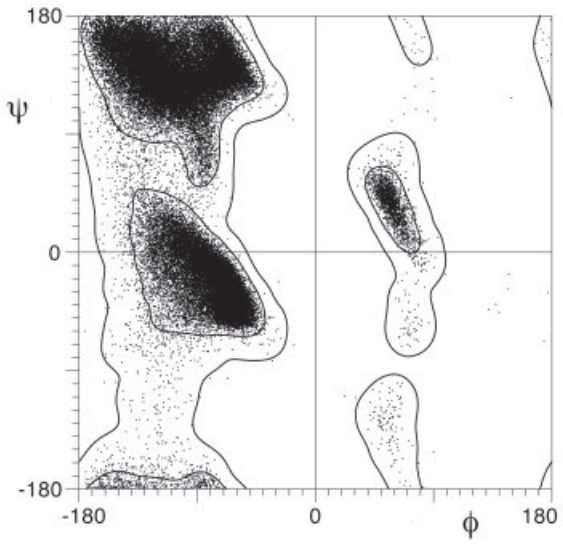

FIGURE 1.2 General Ramachandran plot. This plot shows the distribution of backbone dihedral angles (ϕ,ψ) from the 500-structure high-resolution database by Lovell et al. (30), containing 97,368 residues. All standard amino acid residues are included except glycine, proline, and residues that are followed by proline. Notice that the distribution is not uniform, and that there are certain favored regions and certain outlier regions that are sparsely populated. *Image reproduced from Lovell et al. (30) with permission from John Wiley and Sons.*

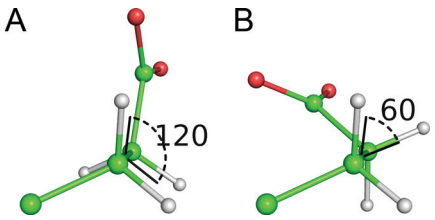

FIGURE 1.3 Eclipsed vs. staggered. (A) eclipsed conformation (energetically unfavorable) (B) staggered conformation (energetically favorable).

Like backbone dihedral angles, the dihedral angles about side chain single bonds also do not uniformly sample the full 360° of rotation. For a bonded sequence of atoms 1-2-3-4 connected by single bonds, where atoms 2 and 3 are each sp^3-hybridized and have four neighbors, the dihedral angle between atoms 2 and 3 tends to be found in "staggered" conformations (dihedral angles of +60, -60, and 180) avoiding energetically less-favorable "eclipsed" conformations (dihedral angles of 0, +120, -120), as shown in Fig. 1.3. The rotational energy barrier for hopping from one staggered basin to another (through a higher-energy eclipsed state) is ~2.80 kcal/mol (25), which is low enough to permit frequent thermal activation over the barriers at 300 K. An additional restriction on side chain dihedral angles is that combinations of dihedral angles that result in the overlapping of non-bonded atoms (fourth neighbors and above) are disallowed (30). Because of these and other effects, the dihedral angles of a side chain tend to cluster in preferred combinations called "rotamers" (30). While backbone dihedral distributions can be conveniently represented in 2D Ramachandran plots (e.g. Fig. 1.2), combinations of side chain dihedral angles are not so easily plotted because a side chain can have more than 2 rotatable dihedral angles. Instead, there are published "rotamer libraries" (30) that list the various preferred combinations of dihedral angles for each side chain and the relative likelihood of each combination.

Next, we show a space-filling view of a folded protein (Fig. 1.4), illustrating the tight and irregular packing of atoms. The spheres for the various atoms denote approximately the extent of the electron clouds. A strong repulsive force due to Pauli's exclusion principle keeps (non-bonded) atoms from overlapping. When two atoms with

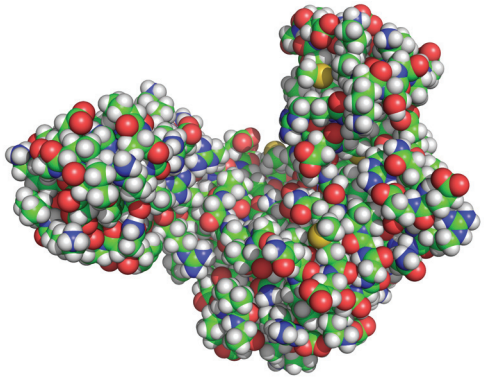

FIGURE 1.4 Space-filling view of a protein. The atoms in a protein are packed tightly and irregularly. Atoms are depicted as spheres, where the sphere radius represents the extent of the electron cloud. Covalently-bonded neighbors, which share electrons, are shown as interpenetrating spheres. Overlap between two non-bonded atoms is prevented due to the Pauli exclusion principle.

filled electron orbitals begin to come too close to each other, the outer orbitals overlap, which forces some electrons into higher energy orbitals since sharing of a spin-orbital is not allowed (Pauli exclusion principle). The rise in energy for atoms in close contact is rather sharp, rising approximately exponentially with decreasing pair separation, but is often modeled as $\sim 1/r^{12}$ for computational ease (31). To a first approximation, the non-overlap requirement can be thought of as a hard geometric restriction on distance, where a pair of non-bonded atoms cannot come closer than some minimum distance.

Two key factors that stabilize water-soluble folded protein are hydrogen bonds and hydrophobic interactions (2). Hydrogen bonds between backbone NH donors and backbone O acceptors stabilize two types of structural elements particularly common in 3D protein structures: α-helices and β-sheets (2), shown in Fig. 1.5 *A*-*B*. Hydrogen bonds involving side chain atoms are also important in protein structures, acting as topological constraints that hold two distinct regions of the protein together, or forming favorable hydrogen bonds with the surrounding water. Certain amino acid residues have hydrophobic side chain atoms (rich in CH_n groups), which do not hydrogen bond and therefore do not form favorable interactions with water. Hydrophobic side chains have a strong tendency to cluster together on the interior of the protein, minimizing their exposure to water, with polar (hydrogen-bonding) side chains positioned on the surface of the protein (2) (Fig. 1.5 *C*). Hydrophobic contacts within a protein therefore tend to be stable and help the protein keep its folded structure. However, hydrogen

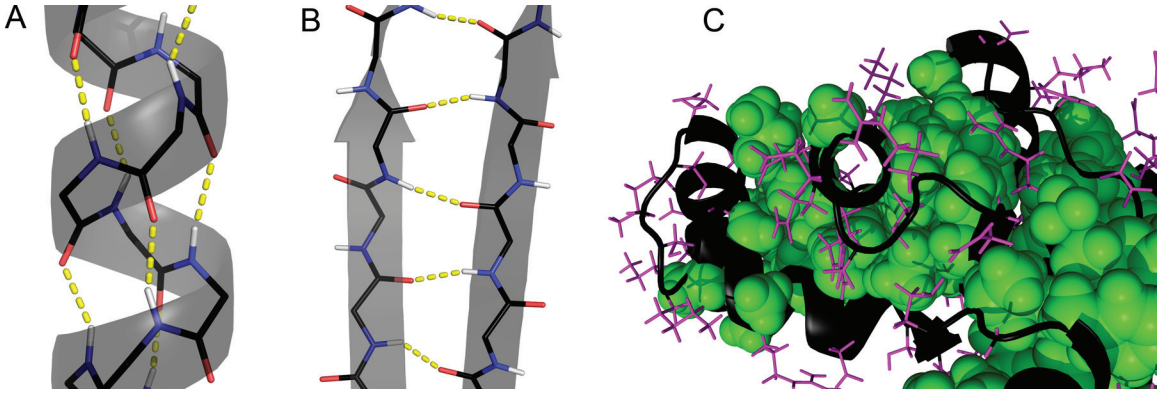

FIGURE 1.5 Protein structure stabilized by hydrogen bonds and hydrophobic contacts. *(A)* An α-helix, stabilized by backbone hydrogen bonds (yellow dashed lines). *(B)* A β-sheet, stabilized by backbone hydrogen bonds. *(C)* A view of a portion of the enzyme adenylate kinase, illustrating clustering of hydrophobic side chains on the interior of the protein, which here "glue" together distinct helices. All side chain atoms of hydrophobic residues (chosen here as Leu, Ile, Val, Phe, Trp, Met, Ala, Tyr) are depicted as green spheres. Polar (hydrogen bonding) residues are located on the surface to interact with the solvent (*magenta sticks*). Main chain atoms are not shown individually but are represented by the cartoon (*black*).

bonds and hydrophobic contacts are individually weaker than covalent bonds [their interaction strengths vary case-by-case but are typically less than 10 kcal/mol (2)], so they can spontaneously break and reform or make new contacts. When proteins change conformations, the change often involves the breaking of particular hydrogen bonds and hydrophobic contacts, which permits a collective motion of the atoms that would be impossible without the breakage of these interactions. Each stable conformation may have its own set of hydrogen bonds and hydrophobic contacts.

NUCLEAR MAGNETIC RESONANCE

This section serves as background for the comparison of computationally calculated free energy to experimentally measured free energy differences presented in Chapter 4. Structural, kinetic and thermodynamic data relating to proteins and protein dynamics can be determined by Nuclear Magnetic Resonance (NMR) spectroscopy and relaxation experiments [for a thorough treatment see (12)]. The protein in solution is placed in a strong, uniform magnetic field, causing nuclear spins to preferentially

align along the field. A radio frequency pulse then applied, perturbing the orientation of the spins. After the pulse is over, the spins gradually realign along the uniform field, precessing coherently about the field axis. The precession frequency depends on the local value of the magnetic field, which can be affected by the surrounding electrons and the local chemical environment. The precessing spins create a changing magnetic field, which can be detected with induction coils. The signal from each spin carries a characteristic oscillatory frequency (called a chemical shift) and decays as the spin relaxes and loses coherence with the other equivalent spins in other proteins. In small molecules the chemical shifts can each be identified with a particular chemical group in the system, but in proteins there are so many nuclei that chemical shifts tend to crowd the spectrum, making it difficult to distinguish specific nuclei. However, by applying series of pulses and varying the pulse separation, multi-dimensional spectra can be obtained that separate out the peaks and reveal interacting neighbors. With enough neighbor information, 3D protein structures can be determined.

If the protein undergoes conformational exchange between two well-defined states, the nuclear spins alternate between two chemical environments (with different characteristic precessional frequencies i.e. chemical shifts). Normally the chemical shifts of the less-populated conformation do not show up because rapid stochastic alternating between different precession frequencies causes dephasing of the spins, which broadens an already weak signal (18). However, in new Carr-Purcell-Meiboom-Gill (CPMG) relaxation dispersion experiments [summarized in (18)] rapid refocusing pulses are applied that sustain coherence of the precessing spins despite the stochastic switching between frequencies, so that the chemical shifts of both conformations are clearly visible in the spectrum. The precise amount of broadening depends on the rate constants of conformational exchange and on the time between refocusing pulses, so to determine rate constants, the time between refocusing pulses is varied and the corresponding amount

of broadening is measured. Thus rate constants, equilibrium constants, and equivalently, free energy differences, can be determined. Enthalpic and entropic differences can also be determined by measuring temperature dependence of the equilibrium constant (32). A few notable examples are the equilibrium exchange between catalytically relevant conformations in the enzymes dihydrofolate reductase (17) and cyclosporin A (33).

COMPUTATIONAL APPROACHES TO PROTEIN PATHWAYS

In this section, we present a sampling of some of the computational methods in existence for generating pathways between two given conformations of a protein. The list is not comprehensive, but it does illustrate the wide spectrum of methods in current use. We begin with perhaps the most widely used technique for predicting conformational change pathways in proteins, though not the most sophisticated, which is targeted molecular dynamics. Rigorous approaches highlighted in this section that find optimal pathways are nudged elastic band and transition path sampling, which are beginning to be applied to proteins but are not yet widely used. We also summarize linear interpolation with energy minimization, and elastic network-based models for pathway generation, which use approximations to make rapid pathways. We finish this section with a summary of umbrella sampling, a method for calculating free energy along a pathway. The targeted molecular dynamics and umbrella sampling methods descriptions also serve as background for Chapters 3 and 4.

Targeted molecular dynamics

Targeted molecular dynamics (34, 35) (TMD) is a biased variant of molecular dynamics (MD). MD-based approaches are classical dynamics simulations, in which the atoms are point particles (carrying mass and charge) whose motion is governed by a classical Hamiltonian called a molecular mechanical force field (31, 36). Typically, the potential energy includes terms for covalent bond stretching and bond bending, electrostatic energy, van der Waals attraction and repulsion, and periodic potentials for

dihedral bond rotations. Newtonian equations of motion must be integrated in small time steps or the integration will become unstable. As mentioned earlier, for proteins the time step can be no bigger than 1-2 fs (37), which makes biological timescales completely inaccessible to regular MD simulations in most cases.

In the TMD variant of MD (34, 35), a biasing force pulls atoms gradually towards a given target state, inducing the conformational change to take place on a much shorter time scale than the biological time scale. The bias is usually implemented as follows: A constraint is established on the root-mean-square distance (RMSD) of the current structure relative to the target structure, calculated over say all non-hydrogen atoms. The RMSD constraint distance is initialized as the RMSD between the initial structure and the target structure, and at each time step the RMSD constraint distance is decreased by some small amount. The constraint is integrated into the dynamics with a Lagrange multiplier. The force of constraint at any instant is along the vector from the current to the target positions, so is strongest for atoms that are furthest away. The magnitude of the force is whatever magnitude will cause the RMSD to equal the required value at that time step.

Although the use of bias makes the problem of finding a pathway computationally tractable, the biasing force alters the energy barriers and the natural dynamics (see Chapter 3). For this reason, TMD pathways are more properly regarded as plausible pathways than "optimal" pathways. Still, TMD is popular because it can produce a reasonable all-atom pathway that can reveal key aspects of a transition. Notable examples include the study by Ma et al. (38) of a large conformational change in the chaperone protein GroEL, revealing a two-stage transition, and an extremely large-scale simulation (2×10^6 atoms) of a tRNA entering the ribosome (39), revealing a loop that impedes tRNA entry and revealing how flexibility in the tRNA and the ribosome facilitates entry.

Nudged elastic band

Nudged elastic band (NEB) (40-43) seeks a minimum energy path (MEP) between two states. Given an initial set of pathway snapshots called "replicas" between the two states, a stretched spring is placed between the coordinates of each pair of replicas, forming an elastic band that stretches between the two end states. An objective function to be minimized is established, which is essentially a path integral that includes the "true" energy of the molecular mechanical potential and the spring energy between neighboring replicas. As the band relaxes through a minimization procedure, artifacts of the elastic band that would prevent the band from relaxing to the MEP, namely "corner cutting" and "sliding down," are avoided by projecting out certain components of the forces. In proteins, NEB has been applied to the opening-closing transition in adenylate kinase (44), and a loop transition in dihydrofolate reductase (45).

Transition path sampling

Transition path sampling (TPS) (46) is a method for finding an ensemble of room-temperature pathways connecting two basins A and B separated by a single transition state region [in proteins, where there can be several metastable states connecting the end points, the method must be performed piecewise (47)]. TPS requires an initial pathway connecting the two (meta)-stable states to begin, which need not be perfect. A random walk is performed in trajectory space, where each "trial move" is a new trial trajectory that connects the two basins, and a Metropolis criteria accepts or rejects each trial move. Each trial trajectory is created by randomly perturbing the momenta in a random snapshot of the trajectory, and then performing unbiased molecular dynamics forwards in time to make a new half-trajectory to state B, and backwards in time to make a new half-trajectory to state A (the move is rejected if the half-trajectories do not reach the right basin), then the move is unsuccessful and rejected). The procedure gradually converges to a stable ensemble of trajectories that connect the two basins. In proteins, TPS has been

successfully applied to an opening-closing transition in DNA polymerase β (48), a partial unfolding transition in photoactive yellow protein (47), and unfolding of the miniprotein Trp-cage (49).

Linear interpolation

Other approaches employ large approximations to facilitate rapid pathway generation. One is linear interpolation with local energy minimization, publicly available at the Yale Morph Server (23). This technique performs linear interpolation on the Cartesian coordinates of the atoms between two structures (all atoms except hydrogens are modeled) and performs a brief local energy minimization at each snapshot in an attempt to keep the structure sensible. While the approach is rapid and can in some cases produce visually reasonable pathways, it is also common for the method to produce severely unreasonable pathways with groups of atoms passing through each other on their way to the target (23).

Elastic network models applied to pathways

Another class of methods for rapid pathway generation use an elastic network representation, in which the protein is modeled as an interconnected spring network. These models do not consider all atoms, but instead only model one atom from each amino acid (the Cα atom). Harmonic springs are placed between pairs of Cα atoms that are near each other, within some maximum distance. A small-amplitude approximation is invoked to obtain normal modes. Pathways can be obtained from this model as follows. In Elastic Network Interpolation (50, 51), inter-Cα pair distances are interpolated from initial values to target values in a series of steps. A reaction coordinate is defined that represents the amount of interpolation, and at each value of the reaction coordinate the spring energy (in the small-amplitude approximation) is minimized. In the Plastic Network Model (90), normal modes and eigenvalues are calculated for the initial state and for the target state, defining two harmonic energy basins. The energy of an arbitrary

configuration of the Cα atoms is defined as the lower of the two basin energies. A minimum energy pathway (in the small amplitude approximation) is constructed between the two basins by conjugate peak refinement.

Although rapid, elastic network-based models of protein pathways neglect important geometric factors that heavily influence motion in proteins. One neglected aspect is the excluded volume of atoms. In reality, protein motions occur in packed environments, and atoms can hardly move without bumping into each other, which makes motion in proteins very complicated. In the elastic network interpolation method and the plastic network model, the packed environment is ignored due to the modeling of only one atom per amino acid. Furthermore, with no mechanism for detecting and avoiding collisions in these models, if two regions of the protein (that do not have springs between them) are transiently brought into each other's space during the pathway there is nothing to prevent them from overlapping. Another neglected aspect is the covalent bonding geometry. The approximately fixed covalent bond angles and distances place significant restrictions on motion between neighboring amino acids, which cannot be accounted for by placing a simple spring between Cα pairs. An updated version of the method pastes the full set of atoms onto the Cα scaffold after the pathway is completed, and minimizes the snapshots with a molecular mechanical potential to clear up any atom overlaps (52). However, by adding the atoms on at this late stage the atoms do not have an opportunity to influence that pathway. Some amino acids take up a large amount of space, and if this space is ignored during the generation of the pathway it seems plausible that pasting atoms onto the scaffold could lead to jammed situations in which clashes cannot be resolved with a downhill minimization that only takes the system to the nearest local minimum.

Umbrella sampling free energy calculations

Since chapter 4 addresses the use of geometric targeting pathways as input to umbrella sampling simulations (53), here we present a brief overview of umbrella sampling methodology that will serve as background. In its application to proteins, umbrella sampling is a method for enhancing sampling in molecular dynamics (MD) along a predefined coordinate, using a biasing potential to obtain thorough sampling in regions of conformational space that normally would be undersampled. The gathered statistics can be used to calculate free energy profiles, also called potentials of mean force, using the Weighted Histogram Analysis Method (WHAM) (54) to properly account for the effect of the bias.

To begin, consider a system of N classical particles in 3D whose positions can be represented with a $3N$-dimensional vector $\mathbf{r}^N$. In equilibrium at fixed temperature T, the probability density of finding the system at a point $\mathbf{r}^N$ is given by

$$p\left(\mathbf{r}^N\right)=\frac{\exp\left[-\beta E\left(\mathbf{r}^N\right)\right]}{Z}, \tag{1.1}$$

where $\beta=1/k_BT$, $\exp\left[-\beta E\left(\boldsymbol{r}^N\right)\right]$ is the Boltzmann factor, and Z is the partition function

$$Z=\int d\mathbf{r}^N \exp\left[-\beta E\left(\mathbf{r}^N\right)\right] \tag{1.2}$$

independent of q. Consider now some generalized coordinate $q=q\left(\mathbf{r}^N\right)$, which could be as simple as the distance between two particular atoms, or some more complicated function of the coordinates. Since multiple values of $\mathbf{r}^N$ can in principle map to the same value q, the probability density of finding the system at q must be obtained from a thermodynamic average

$$p(q) = \frac{\int d\mathbf{r}^N \exp\left[-\beta E\left(\mathbf{r}^N\right)\right] \delta\left(q - q\left(\mathbf{r}^N\right)\right)}{Z}. \tag{1.3}$$

We can define a free energy $F(q)$, also called a potential of mean force, through the following expression in analogy with (1.1),

$$p(q) = \frac{\exp\left[-\beta F(q)\right]}{Z} \tag{1.4}$$

where $F(q)$ is the numerator of (1.3). Taking the logarithm of (1.4), we obtain

$$F(q) = -kT \ln p(q) - kT \ln Z. \tag{1.5}$$

Inasmuch as molecular dynamics (MD) simulates a true thermodynamic ensemble, we can in principle (but not in practice) use MD to determine the free energy profile $F(q)$. To do so, we would calculate the value of q in every MD snapshot and construct a histogram representing the distribution $p(q)$. With this distribution, we obtain $F(q)$ from (1.5), neglecting the constant $-kT \ln Z$. The problem with this, as has been already discussed in this chapter, is the limited sampling ability of MD due to the small time step.

Umbrella sampling (53, 55) is a common method to overcome this problem, enabling the determination of free energy profiles in one or more dimensions. In "windowed" umbrella sampling, several independent MD simulations are run, each restrained to sample a particular window of the coordinate q by means of a restraining potential, usually

$$W_i(q) = \frac{1}{2} k_i (q - q_i')^2 , \tag{1.6}$$

where q_i' is the designated center of the window i, and k_i is the spring constant for window i. The use of windows enables comprehensive coverage of the parameter q, which would not otherwise be possible. The samples collected in each window can be used to calculate $F(q)$ after they are reweighted to account for the biasing potential. Reweighting is non-trivial, and the usual method is WHAM (Weighted Histogram Analysis Method) (54). The biased probability distribution of window i is, in principle,

$$p_i(q) = \frac{\exp\left[-\beta\left(F(q) + W_i(q)\right)\right]}{Z_i} \tag{1.7}$$

where Z_i is the partition function in the presence of the bias,

$$Z_i = \int d\mathbf{r}^N \exp\left[-\beta\left(E\left(\boldsymbol{r}^N\right) + W_i\left(q\left(\boldsymbol{r}^N\right)\right)\right)\right]. \tag{1.8}$$

If this were strictly true, unbiasing would be trivial, because we could combine (1.7) and (1.4) to solve for the unbiased distribution,

$$p(q) = \left(\frac{Z_i}{Z} \exp\left[+\beta W_i(q)\right]\right) p_i(q) . \tag{1.9}$$

The problem with this is that the biased partition function Z_i and biased $p_i(q)$ are incomplete, only containing information for a narrow range of sampled q. In WHAM (54), the estimate for the unbiased distribution $p(q)$ (and equivalently, the free energy profile $F(q)$) is obtained by combining the information from all histograms $p_i(q)$ in

a way that minimizes statistical errors and returns error estimates. We do not present or derive the WHAM equations here, referring the reader to these references (54, 55).

CONSTRAINT-BASED PROTEIN MODELING

The geometric targeting method (GT) for pathway generation, which is the focus of this dissertation, is related to prior work in the field of constraint-based protein modeling and simulation. Here, we describe the related models and clarify the relationship of the present work to these models. We begin with an introduction to rigidity theory and constraint counting. Rigidity theory is the basis for understanding FIRST (Floppy Inclusion and Rigid Substructure Topography) (56), a method for performing a static rigid cluster analysis on a given protein structure based on an assumed set of geometric constraints. We then briefly summarize ROCK (Rigidity Optimized Conformational Kinetics) (57), a method for exploring conformations of the FIRST rigid clusters, then move on to describe FRODA (Framework Rigidity Optimized Dynamics Algorithm) (21), a successor to ROCK. We shall also discuss some of the issues with the FRODA method, providing motivation for the development of the new geometric simulation/targeting methodology described in this dissertation and implemented in FRODAN. We then summarize FRODAN.

Rigidity theory

Here we introduce the concept of constraint counting and show how it can be used to characterize the rigidity in constraint networks. For thorough treatments of the subject, see the dissertations of Lee (58) and Sljoka (59). Although proteins are 3D objects that are modeled with 3D constraint networks in FIRST, it is easiest to begin with 2D examples. Fig 1.6 shows “bar-and-joint” frameworks in 2D—the vertices in the diagram are freely rotatable pin joints, and the edges in the diagram are “bars” with fixed distances. The question we shall concern ourselves with is whether a framework is rigid. A rigid framework in 2D has no internal degrees of freedom—it can move as a

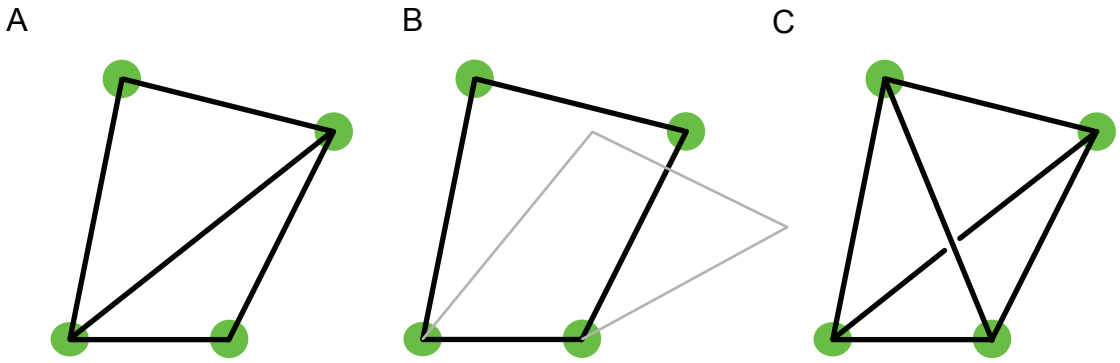

FIGURE 1.6 Bar-and-joint frameworks in 2D. Circles represent "joints", with 2 degrees of freedom in the plane. Lines represent "bars", which are rigid distance constraints between joints. *(A)* A minimally rigid framework. *(B)* An underconstrained framework with one internal degree of freedom (gray lines denote a possible deformation of the framework). *(C)* An overconstrained framework.

rigid object horizontally, vertically, and rotationally (3 rigid body degrees of freedom in the plane), but these are trivial global degrees of freedom, not internal degrees of freedom. If a framework in 2D has a total number of degrees of freedom f, the number of *internal* degrees of freedom is $f-3$. Each pin joint carries two degrees of freedom, so if there are n pin joints the total number of degrees of freedom of the pins only is $2n$, and the number of *internal* degrees of freedom is $2n-3$ (neglecting the constraining bars). Now we consider the effect of the bars. The bars are distance constraints, which remove degrees of freedom. If the number of bars is m, and if $2n-3=m$, then the number of bars exactly cancels out the number of internal degrees of freedom of the pin joints, and we have a minimally rigid framework (Fig. 1.6 *A*). Minimally rigid means just barely rigid, because if we remove any bar the framework will not be rigid anymore. If $2n-3>m$, then the number of bars do not quite cancel out all the internal degrees of freedom, and we have a flexible (underconstrained) framework (Fig. 1.6 *B*). If $2n-3<m$, then the number of bars is more than enough to rigidify the framework, and we have an overconstrained framework (Fig. 1.6 *C*) (extra bars are not independent distance constraints, but are redundant, adding reinforcement to the already rigid structure).

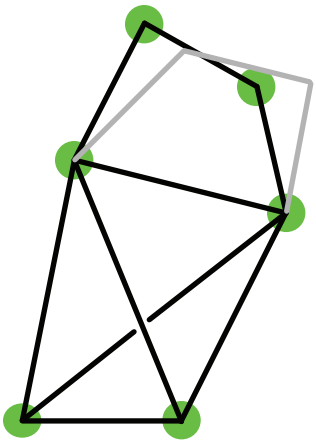

FIGURE 1.7 Bar-and-joint framework that violates the naïve counting condition. According to the not-quite correct counting condition for a minimally rigid framework $2n-3=m$, where n is the number of joints, and m is the number of bars, this framework should be minimally rigid. However, clearly there is one internal degree of freedom, located in the upper portion of the framework. The lower portion is overconstrained by 1 bar, and the upper portion is underconstrained by 1 bar, which on balance satisfies the count. This indicates that the counting condition is not correct, and needs to take into account the rigidity of subgraphs (subsets of the joints and the interconnecting bars between them).

The above rules are almost correct, but not quite. Consider Fig. 1.7. Here the constraint count gives $2n-3=m$, which according to the above rules should indicate a minimally rigid framework, however we see that the framework has an overconstrained region (bottom) and an underconstrained region (top). In other words, the count over all pin joints and bars is not a sufficient condition for determining minimal rigidity. If we were to consider the constraint count of the top subgraph only (subgraph here simply means a subset of joints and the bars the interconnect them), we would see that this subgraph is underconstrained. Similarly, if we apply the count to the bottom subgraph, we would find it is overconstrained. With this motivation, we state Laman's theorem (60) from 1970 to determine minimally rigid 2D bar-and-joint frameworks from constraint counting (paraphrased):

Theorem 1 (Laman) *A generic 2D bar-and-joint framework composed of n joints and m bars is minimally rigid if and only if (A)* $2n-3=m$*, and (B)* $2n'-3\geq m'$ *for every subset of joints* n' *and its interconnecting bars* m'.

"Generic" here means that there are no special symmetries in the framework, such as parallel lines or bars that share the same exact distance. At first thought it would seem very time consuming to perform this constraint count on all possible subgraphs, given a large number of joints and bars. However, an elegant algorithm called the 2D pebble game (61, 62) [or more precisely, the (2,3)-pebble game in the modern formalism of Lee

and Streinu (63)] performs the Laman constraint counting condition very efficiently and identifies overconstrained (rigid), minimally rigid, and underconstrained regions of the framework.

Proteins, however, are 3D objects, so we need a way to characterize rigidity in 3D. Unfortunately Laman's theorem does not generalize to the 3D bar-joint framework. However, for a different class of generic 3D frameworks, called body-bar-hinge frameworks, there is a constraint count that characterizes rigidity (64). Body-bar-hinge is the framework used in the current implementation of FIRST to characterize the rigidity of proteins, as described in these references (65, 66). Body-bar-hinge frameworks consist of "bodies," which are extended objects with 6 degrees of freedom, that are interconnected by "hinges" and "bars," as represented in Fig. 1.8. The bars are distance constraints between two points in two different bodies, removing 1 of the 6 degrees of freedom between two otherwise independent bodies. The hinges are joints that restrict the motion between two bodies, only allowing a dihedral rotation about the hinge axis. Thus hinges remove 5 of the 6 degrees of freedom between two otherwise independent bodies (equivalent to the effect of 5 bars) (67), leaving only one rotational degree of freedom. In

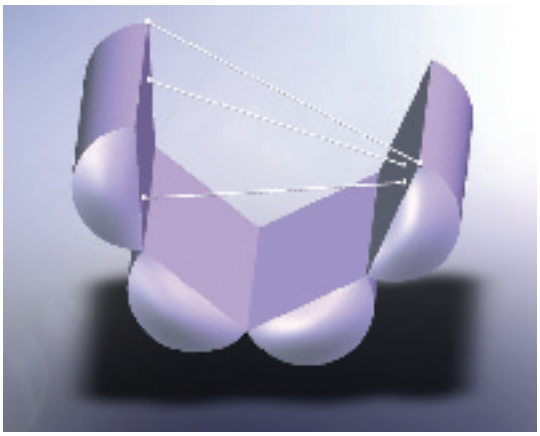

FIGURE 1.8 A generic body-bar-hinge framework in 3D. This is a 3D framework for which rigidity can be determined by a Laman-like counting condition. The framework consists of "bodies" (purple objects) with 6 degrees of freedom (3 translational and 3 rotational), "bars" (the thin white sticks that extend from the upper left body to the upper right body), and "hinges" (intersection lines between adjacent purple bodies). Bars remove 1 degree of freedom between a pair of bodies. Hinges remove 5 degrees of freedom between a pair of bodies, leaving only one dihedral rotation degree of freedom. For this reason, a hinge is equivalent to 5 bars for the purposes of constraint counting. *Image reproduced from Lee (58) with permission from Audrey Lee.*

analogy to Laman's theorem for 2D bar-and-joint frameworks, Tay's theorem (64) for 3D body-bar frameworks (hinges being 5 bars) is the following (paraphrased):

Theorem 2 (Tay) *A generic 3D body-bar framework composed of n bodies and m bars is minimally rigid if and only if (A)* $6n-6=m$*, and (B)* $6n'-6 \geq m'$ *for every subset* n' *and its interconnecting bars* m'*.*

As before, generic means that there are not special configurations such as hinges coincident at a point or parallel bars between bodies. Also, just as the 2D bar-and-joint constraint count was efficiently carried out by a 2D pebble game algorithm, Lee and Streinu have recently proven that there is a pebble game algorithm that performs the Tay $6n-6$ counting condition, called a (6,6)-pebble-game (63). In light of Tay's theorem (67), the constraint count performed by the (6,6)-pebble-game characterizes rigidity of body-bar-hinge frameworks and can determine 3D rigid clusters.

There is a final detail that must be mentioned in regards to the application of constraint counting to molecular systems. As we will see next with proteins, when molecules are modeled in the body-bar-hinge framework, rotatable covalent bonds translate to "hinges" in the framework. Atoms commonly have multiple covalent neighbors, and if these bonds are rotatable, the hinges are coincident at the center of the atom. This could potentially be a non-generic arrangement of hinges, making Tay's theorem inapplicable. According to the unproven "Molecular Conjecture" of Tay and Whiteley, Tay's theorem holds for body-bar-hinge frameworks representing molecules (68). In 2009, 25 years after the Molecular Conjecture was proposed, the Molecular Conjecture was finally proven for arbitrary dimension by Katoh and Tanigawa (69), a tremendous achievement in rigidity theory.

FIRST

FIRST (Floppy Inclusion and Rigid Substructure Topography) (56) is a method and software program that analyzes an input protein structure and predicts groups of atoms that are expected to behave as rigid bodies. It does not actually explore motion

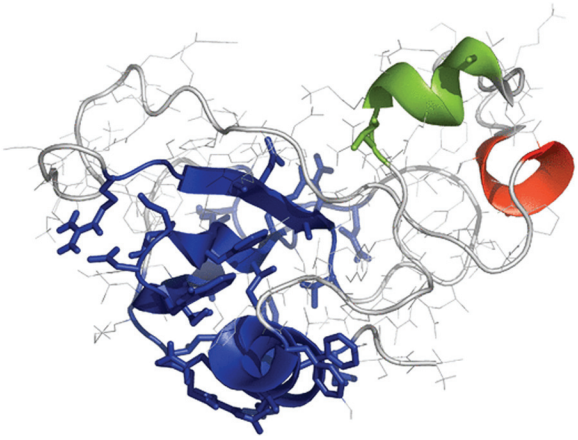

FIGURE 1.9 Rigid clusters from FIRST for the protein barnase. From a single input structure, a body-bar-hinge framework is created and a set of constraints is assumed. FIRST runs the pebble game to determine the rigid clusters in the framework. Only the three largest rigid clusters are shown here. In FIRST, all atoms are assigned to a rigid cluster, but these can be trivially small (as small as one atom plus its bonded neighbors). *Image reproduced from Wells et al. (21) with permission from IOP Publishing.*

of the rigid bodies. Fig. 1.9 shows a sample rigid cluster decomposition for the protein barnase. Only the largest three rigid clusters are shown—FIRST assigns every atom in the protein to a rigid cluster, but many of these are trivially small (consisting of an atom and its bonded neighbors). The rigid clusters are determined by modeling the protein as a constrained system. Constraints remove degrees of freedom, and the number of constraints may be high enough in certain regions that groups of atoms are left with no internal degrees of freedom—the rigid clusters. A noteworthy application is the use of rigid clusters from FIRST to coarse-grain an elastic network model (70, 71), improving the accuracy and predicted large-amplitude soft motions and the efficiency the calculation. Radestock and Gohlke (72) used rigidity predictions from FIRST to study stability differences between the protein structures of thermophilic organisms (thriving in high temperature environments like deep sea vents) and mesophilic organisms (thriving at moderate temperatures). The effective temperature of the rigidity analysis was modulated by varying the set of hydrogen bonds used as constraints, revealing folding/unfolding transition temperatures that tended to be higher for thermophilic proteins than for their mesophilic counterparts.

In the current implementation of FIRST (65, 66) the protein is represented as a body-bar-hinge framework (described earlier under the heading Rigidity Theory) as

shown in Fig. 1.10. Each atom is a “body” in the framework, and “hinges” (5 bars) interconnect bodies across rotatable covalent bonds. This construction has the effect of locking all covalent bond distances and three-body angles, leaving only 1 dihedral rotation degree of freedom between two atoms joined by a hinge. Six bars, rather than 5, are placed across bonds that are not rotatable, which effectively locks the two atoms together as one rigid body (e.g., across peptide bonds or double bonds, which have high energy barriers to rotation, or between a carbon and a terminal bonded hydrogen, for which bond rotation is meaningless).

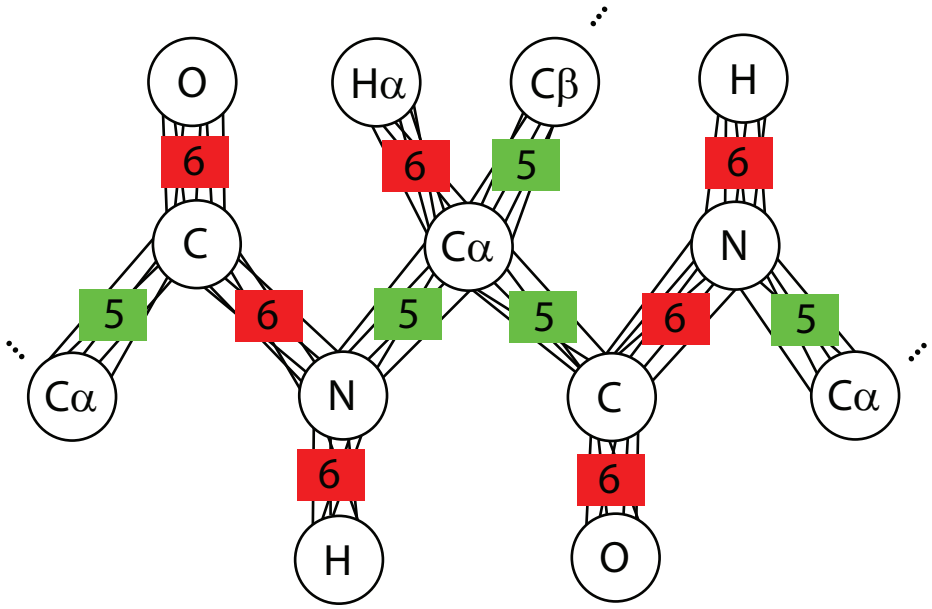

FIGURE 1.10 The body-bar-hinge framework of a protein. Each atom in the protein is a “body” in the framework with 6 degrees of freedom (if you like, think of the atom as an extended object like a tetrahedron with vertices extending towards bonded neighbors). Hinges (5 bars, denoted by a green square to indicate rotational freedom), allow only a single rotational dihedral degree of freedom between a pair of bodies (distances and bending angles fixed). Six bars (denoted by red square to denote a rotationally locked bond) are placed to remove all degrees of freedom between a pair, essentially locking the two bodies into the same rigid cluster. Notice that 6 bars are placed along the peptide N-C bond and along the C=O double bond preventing dihedral rotation. Six bars are also placed along bonds to terminal hydrogen atoms because rotations about these bonds are meaningless due to rotational symmetry.

In addition to the covalent bonds, hydrogen bonds identified in the input structure are also modeled as hinges (5 bars), allowing only a dihedral rotation about the H…A (hydrogen-to-acceptor) axis (hydrogen bond distances and 3-body angles are locked). Identification of hydrogen bonds is described in Jacobs et al (56), in which the structure is scanned for donor-hydrogen-acceptor atoms whose hydrogen bond energy score (56) is better than some user-defined value.

Hydrophobic contacts identified in the input structure, which are pairs of hydrophobic carbons or sulfurs (clarified below), are constrained with 2 bars. The 2-bar constraint removes 2 degrees of freedom between a pair of bodies in the body-bar-hinge framework, which is a looser constraint than a hinge (5 bars), but is a stronger constraint than a simple distance constraint between the pair (1 bar). The current criteria for identification of hydrophobic contacts or "tethers" [according to the FIRST User Guide (73), which describes an updated procedure compared to the published description (56)] is as follows. The input structure is scanned for pairs of carbons (or sulfurs) that are within some maximum distance of each other (3.9 Å for carbon-carbon pairs). The tether is only kept if both atoms are only bonded to carbons, sulfurs, or hydrogens (as an indicator of a hydrophobic environment). Furthermore, if a given atom has more than one tether extending from it to the atoms of a particular residue, only the tether with shortest distance is kept. Note that an atom is allowed to have more than one tether, as long as the tethers each go to a unique residue.

After the body-bar-framework is created and the bars assigned, rigid clusters are determined by performing a constraint count with the (6,6)-pebble-game of Lee and Streinu (63). The result of the rigidity analysis is a set of rigid clusters, interconnected by rotatable bonds (the hinges are the leftover rotatable covalent bonds and hydrogen bonds that did not become rigidified), and bars (the leftover hydrophobic tethers that did not become rigidified). In unconstrained regions of the framework, although there are no

macroscopic rigid clusters, trivially small rigid clusters are present. The smallest rigid clusters consist of a single atom and its covalently bonded neighbors.

ROCK

The next logical step after FIRST, a method known as ROCK (Rigidity Optimized Conformational Kinetics) (57) was developed to explore 3D conformations of the constrained protein system, taking as input the rigid clusters from FIRST. Because of constrained bond distances and angles, the only degrees of freedom in ROCK are the dihedral angles between adjoining rigid clusters. Because rigid clusters from FIRST can be joined together in ring topologies (e.g. rigid clusters A-F could be cyclically joined as ABCDEFA with dihedral angle rotations between each pair), and there can be rings within rings, sampling of conformational space in ROCK requires solving ring closure systems of equations to find dihedral angle solutions that do not break the rings.

FRODA

Like ROCK, the concept behind the FRODA method (Framework Rigidity Optimized Dynamics Algorithm) by Wells et al. (21) is to explore the allowed conformational space of the rigid clusters determined by FIRST. The difficulties of satisfying ring closure in ROCK were circumvented in FRODA. FRODA is a significant improvement compared to ROCK, enabling faster sampling and including dynamic collision avoidance between atoms. One noteworthy application of FRODA is the flexible fitting of all-atom protein structures into low-resolution structure data from cryo-electron microscopy (74). Portions of FRODA combined with the new methodology (FRODAN) presented in this dissertation have also been used as a component in computational protein folding methods (75). We will here describe FRODA methodology, and in the next section we discuss some limitations and problems that provide motivation for the development of the new methodology FRODAN.

The result of the FIRST rigidity analysis, which is passed into FRODA as input, is a set of rigid clusters that are interconnected by rotatable hinges (these are any leftover hinges, i.e. any rotatable covalent bonds and hydrogen bonds that did not become rigidified) and bars (leftover hydrophobic tethers that did not become rigidified). As FRODA moves the rigid clusters, it must maintain the hinge constraints between adjoining rigid clusters and the leftover hydrophobic tethers (2-bar constraints). Since it is not geometrically well-defined which two degrees of freedom are removed in a 2-bar constraint, FRODA converts these to inequality distance constraints (maximum distance constraints), which formally do not remove any degrees of freedom, but which do maintain a loose connection between the two atoms. In addition, FRODA must ensure that pairs of non-bonded atoms (in different rigid clusters) do not overlap as the clusters move, and it accomplishes this by enforcing minimum distance constraints.

In FRODA, the dihedral angles are not the explicit degrees of freedom, unlike ROCK. Instead, the model in FRODA consists of two types of mobile entities: atoms with three degrees of freedom, and rigid "ghost templates" with six (rigid body) degrees of freedom. Fig. 1.11 shows the atoms and ghost templates for an ethane molecule rather than a protein, to make it easier to understand the model. The ghost templates contain rigidly embedded "ghost atoms," whose positions serve as a guide for the (physical) atoms to help them maintain geometric relationships with other (physical) atoms within a rigid cluster, a point that will become clear shortly. Observe in Fig. 1.11 that the ghost templates extend across the rotatable bond, and that a (physical) atom can correspond to more than one ghost atom, each in a different ghost template.

To explore conformational space, a series of steps is taken, each step consisting of the following actions: (a) a random perturbation of the positions of the atoms, and (b) enforcement of the constraints. In the perturbation phase, each atom position is moved by a small random amount in a randomly chosen direction. The magnitude of

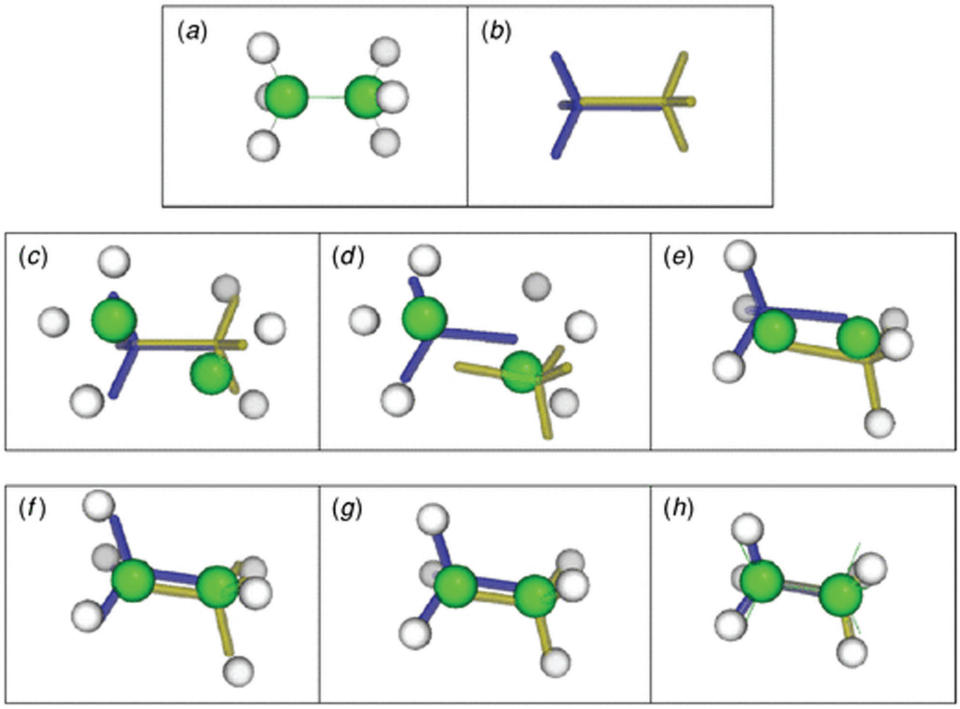

FIGURE 1.11 Original FRODA methodology. The example molecule in these panels is ethane, which is simpler to depict than a protein. In the assumption that bending angles and distances are rigid, ethane has one rotational degree of freedom about its central axis. In FRODA, the molecule is modeled with two types of mobile entities: atoms with 3 degrees of freedom and rigid ghost templates with 6 degrees of freedom. *(a)* The FRODA atoms, represented by spheres. *(b)* The rigid ghost templates of the FRODA model, depicted as sticks. The ghost templates correspond to FIRST rigid clusters. Observe that the two ghost templates overlap across the rotatable bond. *(c)* Atoms are perturbed by random small amounts in random directions. (*d-e)* First iteration of enforcement of constraints (simple case, ignoring non-overlap minimum-distance constraints and hydrophobic maximum distance constraints). *(d)* Ghost templates are moved to positions that best fit their corresponding atoms. *(e)* Atoms are moved to the mean position of their corresponding ghost atoms. *(f-g)* Second iteration of enforcement of constraints. *(f)* Ghost templates are fit to the atoms. *(g)* Atoms are moved to the mean position of their corresponding ghost atoms. *(h)* After several iterations of the enforcement of constraints procedure, the constraints are met to within tolerance. The system is now found in a different conformation from where it started. *Image reproduced from Wells et al. (21) with permission from IOP Publishing.*

the displacement is a random number between 0 and some maximum amount, usually set to 0.1 Å. After the perturbation, the atoms are clearly in violation of the constraints, since they no longer maintain the required rigid body relationships with each other. To enforce the constraints, the following iterative procedure is performed (for simplicity,

we first describe the enforcement procedure neglecting the steric overlap constraints and hydrophobic contact constraints): (a) Each ghost template is placed at the "best fit" location that minimizes the sum of square distances between the ghost atoms and the physical atoms. (b) The position of each physical atom is updated to the mean position of its ghost atoms. Continuing with the next iteration, ghost templates are re-fit to the new positions of the atoms, and the atoms are again updated to the new positions of the ghost atoms. The iterative process continues until each atom and its corresponding ghost atoms coincide within some threshold. The typical requirement is that the distance between any atom and any of its ghost atoms must be less than 0.125 Å. Once this point is reached, FRODA has successfully produced a new conformation that meets the original constraints.

The use of ghost templates is a clever way to keep track of the rigid geometry of the clusters as well as the geometric interconnections between clusters. Note that because the ghost templates extend across rotatable bonds, it is not necessary to explicitly constrain the distances and angles between two rigid clusters. As long as the ghost atoms from multiple templates coincide, the angle and distance constraints are automatically satisfied, while dihedral rotations about the adjoining bond are left unconstrained.

To handle the minimum distance constraints for preventing non-bonded overlap and the maximum distance constraints to preserve the pre-defined set of hydrophobic contacts, the iterative procedure for enforcement of constraints is modified as follows. Before atoms are moved to the mean position of their ghost atoms, a search is made for any pairs of non-bonded atoms that are closer than some contact distance value (determined by summing radii values for the atoms). Also, the pre-defined list of hydrophobic contact pairs is searched for any pairs that are farther apart than their allowed maximum distance. Each atom is then displaced by the sum of the following vectors: (a) the displacement vector that would move the atom to the mean position of

its ghost atoms; (b) the displacement vector that would move the atom directly away from an overlapping neighbor by half the overlap distance (if the atom overlaps with multiple neighbors, there is one such displacement vector added for each overlap); (c) the displacement vector that would move the atom directly towards a pre-defined hydrophobic partner by half the excess distance (if the atom violates more than one hydrophobic contact constraint, there is one such displacement vector added for each). The idea is that these additional movements towards/away from other atoms that are too far/too close will help the system converge to a state that satisfies the constraints. The iterative enforcement of constraints procedure continues until all constraints are met within some tolerance: the distance between any atom and any of its ghost atoms must be less than 0.125 Å, the distance between any two non-bonded atoms must be greater than 85% of the sum of their van der Waals radii, and the distance between pairs that

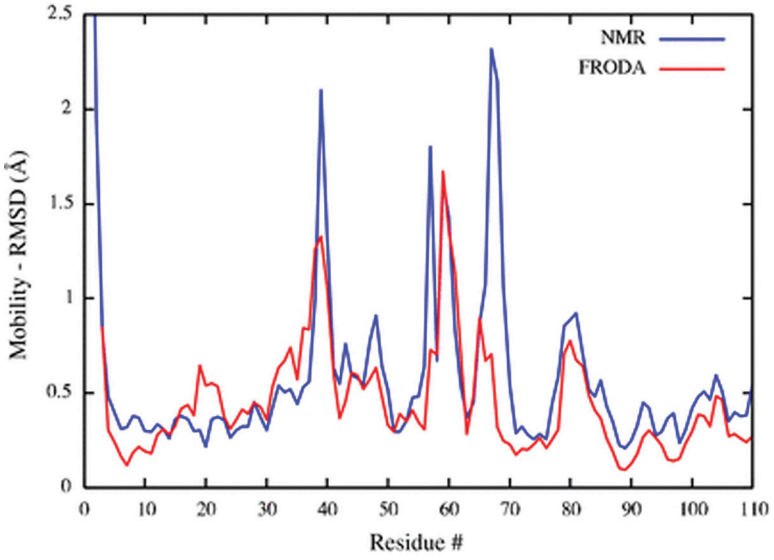

FIGURE 1.12 Mobility comparison between FRODA and NMR. Wells et al. generated an ensemble of conformations with FRODA for the protein barnase, and compared the mobility of each residue in the FRODA ensemble (*red*) to experiment (NMR) (*blue*). The mobility is measured as the root-mean-square fluctuation of each Cα atom. The FRODA and NMR curves have very similar features, showing that the geometric model in FRODA captures the correct mobility for barnase. *Image reproduced from Wells et al. (21) with permission from IOP Publishing.*

have a hydrophobic constraint must not exceed 0.125 Å beyond the maximum distance constraint (73).

A key result connecting FRODA with experiment, demonstrated in Wells et al. (21), was the prediction of Cα mobility in the protein barnase. Fig. 1.12, taken from (21), shows the mobility of each residue in barnase predicted by geometric simulation in FRODA, compared to NMR results. The FRODA and NMR curves shown in Fig. 1.12 have very similar features, showing that the geometric model in FRODA captures the correct mobility for barnase.

Targeting is one of the applications of FRODA described in Wells et al. (21), similar in principle to the geometric targeting method (Chapter 2). The goal is to produce a pathway from an initial structure to a target structure. In FRODA, targeting is achieved by modifying the perturbation step as follows. In addition to the random perturbation of the atoms, atoms are also perturbed by a small amount towards the target. Because certain hydrogen bonds and hydrophobic contacts of the initial structure were not compatible with the target structure, FRODA introduced the concept of "common constraints," which refers to the hydrogen bonds and hydrophobic contacts that are identified in both the initial and target structures. In FRODA targeting, only the common constraints are included in the FIRST rigidity analysis.

Problems and limitations in FRODA

Despite being a significant leap forward compared to ROCK in terms of speed of exploration and active prevention of collisions, there are several aspects in FRODA that are problematic, many of which are solved by the new methodology presented in this dissertation. The first issue relates to a fairly common occurrence when using FRODA, which was that the simulation would suddenly abort after repeated failed attempts to satisfy constraints. While it is difficult to diagnose exactly what was going wrong, the enforcement of constraints procedure itself seems a likely culprit. It is not clear that the

iterative enforcement procedure, consisting of fitting ghost templates to atoms, updating atom positions to the mean positions of the ghost atoms, and nudging atoms toward/away from each other by half the distance of their constraint violations, has any guarantee of reducing the total amount of constraint violations at each iteration. In particular, consider the movement of atoms by a pre-decided distance away from their overlapping neighbors (half the overlap distance). In a packed protein environment where there are multiple overlaps for each atom that need to be repaired, the arbitrary choice of the magnitude of the separation movement (half the overlap distance) could in principle create new overlaps which could be worse than the previous overlaps, which could lead to oscillatory behavior as atoms are bounced back and forth from one overlap to another. In the new methodology (Chapter 2), an explicit objective function is introduced that measures the total amount of constraint violation, and enforcement of constraints is achieved with the well-known conjugate gradient minimization algorithm (76). Also in the new methodology, there is only one type of mobile entity (a rigid unit) in the model, rather than two types (the atoms and the rigid ghost templates of FRODA that each have their own degrees of freedom). This is conceptually simpler and reduces the number of degrees of freedom.

Another problematic aspect with FRODA is that the rate of exploration of conformational space is limited by the small perturbation size. Just as MD is limited by a small time step, FRODA perturbations are limited to about 0.1 Å on each atom. With larger perturbations, the enforcement of constraints would sometimes get stuck and cause the simulation to abort. We can imagine the conformation of the protein as a point in 3N-dimensional space, and we can think of the FRODA steps (consisting of a perturbation of the system followed by enforcement of constraints) as a random walk of this point in high dimensional space. For random walks in any number of dimensions, the net distance traveled after m steps relative to the initial position grows as $\sqrt{m}$.

Put another way, the number of steps required to reach a net distance l from the initial position is proportional to l^2. Thus, the net distance traveled increases rapidly at first but gradually levels off, not because the exploration has reached any physical boundary, but because diffusive processes take very long times to travel long distances. In FRODA, the diffusion problem was manifest when a simulation of a protein would simply produce local jiggling of atoms, leaving large movements unexplored even though these motions were allowed in principle by the geometric constraints. In the new methodology of Chapter 2, the more robust enforcement of constraints procedure permits much larger random perturbations, effectively increasing the diffusion constant of random exploration. In addition, a new "momentum run-on" perturbation is introduced in Chapter 2 that uses each step's net motion in the subsequent step's perturbation, which can very quickly take the system far from its initial point.

Another issue in FRODA has to do with the underlying geometric model that it receives from FIRST, which consists of large rigid clusters obtained by treating hydrogen bonds and hydrophobic contacts as rigid. Let us consider these rigid clusters in the context of the pathway problem, where the system must move from state A to state B. While covalent bond distances and angles will be about the same in both A and B, hydrogen bond distances and angles can vary more significantly between A and B. Even if the hydrogen bond is "common" to both the initial and target structures, its geometry may be different. Because hydrogen bond constraints are included in the FIRST rigidity analysis with 5 bars, their geometries become rigid, so in FRODA they are unchangeable. Keeping hydrogen bond geometries rigid can prevent the system from reaching the target state, if these hydrogen bond distance and angle need to adjust to reach the target geometry but are not allowed because they are kept rigid. The same is true for "common" hydrophobic contacts that become rigid in the first analysis, but need to change their geometric relationships with their neighbors. The effect here could be

more than just a local one, as large scale motions may be inhibited if hydrogen bond and hydrophobic contact geometries are not allowed to change. In Chapter 2, a more flexible model is introduced in which hydrogen bond H…A (hydrogen-acceptor) distances have a maximum and a minimum distance, rather than a fixed distance, and a minimum D-H…A angle (D=donor), rather than a fixed angle. These "inequality constraints" do not formally remove any degrees of freedom, however, so they are not included in a rigidity analysis (just like minimum distance constraints for preventing overlap do not remove degrees of freedom and are not included in a rigidity analysis). Without hydrogen bonds and hydrophobic contacts in the FIRST rigidity analysis, the rigid clusters are reduced to trivially small rigid units that only take into account covalent bond geometry. Inequality constraints on hydrogen bonds and hydrophobic contacts still restrict motion significantly, but allow a small amount of flexibility in the geometries of these interactions.

Another area for improvement in FRODA is the manner in which targeting is accomplished. In FRODA, it is accomplished by perturbing atoms towards the target by a small amount at each step. This approach works in many cases, however in some scenarios the perturbation directly towards the target is minimally productive, for example, if a large group of atoms needs to rotate by 180°. The perturbation directly towards the target is orthogonal to the actual direction that atoms need to move to reach the target, and progress will be slow. If we measure progress by the RMSD (root-mean-square distance) to the target, improving the RMSD may require moving with a large component tangent to the RMSD surface rather than downhill. In the new methodology of Chapter 2, a constraint is imposed on the RMSD (root-mean-square distance) to the target (similar to TMD). At each step the RMSD is decremented towards zero, and the conjugate gradient minimization procedure iterative combines directions along the gradient and orthogonal to the gradient, gradually reducing the constraint violations of all

constraints including the RMSD constraint. Thus at each step forward progress is made towards the target, while the structural constraints are simultaneously enforced.

FRODAN

FRODAN (FRODA New) is a new method and software package intended as a successor to FRODA. FRODAN can be run in one of two modes: geometric simulation (a non-targeted exploration of conformational space, starting from a given input structure), and geometric targeting (targeted exploration from a given initial structure towards a given target structure). Chapter 2 of this dissertation describes the methodology in the context of geometric targeting, but it is important to note that both the targeted and non-targeted modes of operation use the same mathematical model and much of the same code. FRODAN resolves the issues with FRODA that were described earlier in this chapter: difficulty enforcing constraints, slow diffusive exploration of conformational space, overly-rigid hydrogen bonds and hydrophobic contacts, and unproductive targeting steps.

Some key differences and improvements in the underlying model and methodology are listed here. First, in FRODAN, rigid units are the only mobile entities in the system (atoms are rigidly embedded in the rigid units), making a conceptually simpler model than FRODA in which rigid ghost bodies and atoms each have their own degrees of freedom. Second, enforcement of constraints in FRODAN is accomplished by means of conjugate gradient minimization of an explicit objective function that measures the amount of constraint violation. The iterative conjugate gradient minimization is guaranteed to lower the amount of constraint violations at each iteration, in contrast to the iterative fitting procedure in FRODA. Third, the rate of diffusive exploration of conformational space is improved with much larger random perturbation steps. Fourth, an optional “momentum run-on” exploration scheme follows large amplitude motions, generating motion much more quickly than the random perturbation approach (in non-

targeted mode, the user can choose whether to use the random perturbation approach or the momentum run-on approach for exploring conformational space). Fifth, the hydrogen bonds and hydrophobic contacts in FRODAN are modeled with inequality constraints (e.g., maximum distance), rather than rigid constraints, allowing greater flexibility of hydrogen bond geometry and hydrophobic contact geometry. This is necessary in targeting where the distance and angle geometry of hydrogen bonds and hydrophobic contacts may be different between the initial and target structures. Sixth, new Ramachandran constraints and side chain torsion constraints have been added to keep dihedral angles in energetically good conformations. Seventh, the targeting method in FRODAN uses a new gradually changing RMSD constraint to efficiently pull the system towards the target. This is an improvement over the FRODA targeting approach, where the targeted perturbations of atoms can in some cases not be productive.

CHAPTER 2 GEOMETRIC TARGETING METHODOLOGY AND PATHWAY RESULTS

This chapter is contains the published paper "Generating Stereochemically Acceptable Protein Pathways" by Daniel W. Farrell, Kirill Speranskiy, and M. F. Thorpe (22). The paper describes geometric targeting methodology, a webserver to provide an easy interface to the method, and pathway results for over 20 proteins. Additional details regarding the methodology can be found in Appendix A of this dissertation. The paper was written by DWF, with some revisions by MFT. The method and software was developed by DWF with advising from MFT. All pathways were generated and analyzed by DWF. Figures and tables were made by DWF. The website and web-interface to the software were developed by KS, in consultation with DWF and MFT. The text and figures in this chapter are as published, with the following exceptions: Two of the supplementary tables originally published for this paper are superseded by tables in Chapter 4, so this chapter will reference the tables of chapter 4. One supplementary table is included in Appendix B. Supplementary movies cannot be included in the dissertation, but are available online at the publisher's website (http://doi.wiley.com/10.1002/prot.22810). Some of the mathematical notation has been updated to correspond with other notation used in this dissertation. For more details on the methodology beyond what is presented in this chapter, see Appendix A.

We describe a new method for rapidly generating stereochemically-acceptable pathways in proteins. The method, called geometric targeting, is publicly available at the webserver http://pathways.asu.edu, and includes tools for visualization of the pathway and creating movie files for use in presentations. The user submits an initial structure and a target structure, and a pathway between the two input states is generated automatically. Besides visualization, the structural quality of the pathways makes them useful as input pathways into pathway refinement techniques and further computations. The approach

in geometric targeting is to gradually change the system's RMSD relative to the target structure while enforcing a set of geometric constraints. The generated pathways are not minimum free energy pathways, but they are geometrically plausible pathways that maintain good covalent bond distances and angles, keep backbone dihedral angles in allowed Ramachandran regions, avoid eclipsed side-chain torsion angles, avoid non-bonded overlap, and maintain a set of hydrogen bonds and hydrophobic contacts. Resulting pathways for over 20 proteins featuring a wide variety of conformational changes are reported here, including the very large GroEL complex.

INTRODUCTION

The ability to determine pathways between different conformational states in proteins is key to understanding how structure influences function. Computational techniques of varying levels of sophistication have been introduced to find pathways in proteins, many of which are computationally intensive. In this work, we present a rapid and computationally-inexpensive method to produce pathways between two states of a protein called geometric targeting, publicly available on a webserver at http://pathways.asu.edu. The webserver provides a simple interface to the targeting method and includes features for visualization of the pathway and generating movie files.

The approach can be summarized as a gradual changing of the system's RMSD (root-mean-square distance) relative to the target structure while enforcing a set of geometric constraints. Underlying geometric targeting is the philosophy that the essence of conformational changes in proteins can be modeled purely from geometric considerations. Geometric targeting generates complicated, highly non-linear, all-atom pathways, and is broadly applicable to many classes of conformational changes and works even for very large systems. The generated pathways are not optimal pathways, but they are stereochemically-acceptable pathways in the sense that they prevent atom overlap, preserve bond distances and angles, preserve hydrogen bonds and hydrophobic

contacts, and keep backbone and side chain dihedral angles away from unfavorable configurations.

Two primary uses of the method and webserver are the visualization of conformational changes and the generation of input pathways for further computation or refinement. For visualization, there are obvious advantages to looking at a movie compared to looking at two superimposed static protein structures. It can be difficult to visually compare static structures, identify the regions that differ, and mentally figure out what motions could be involved in the transition. A movie showing a pathway between two states is a more natural way to learn what has changed and how the change takes place. The other primary use of geometric targeting is to create input pathways for use in more sophisticated techniques. Several techniques that explore the energy landscape to search for optimal pathways such as transition path sampling (46), string method (77-79), and nudged elastic band (41, 42), require an initial pathway to get started that is typically produced by simple interpolation between end states. In systems where interpolation produces a poor initial guess, pathways produced from geometric targeting may make better candidate input pathways. In an article that is currently in preparation with collaborators Tatyana Mamonova and Maria Kurnikova, we will show that a pathway generated from geometric targeting can be used as input into an umbrella sampling (53) free energy calculation (unpublished) (in this dissertation, the umbrella sampling work is in Chapter 4).

Besides the geometric targeting method introduced in this paper, various other approaches exist for finding pathways in proteins. Sophisticated techniques that perform rigorous searching of energy landscapes to determine optimal pathways include the aforementioned transition path sampling (46), string method (77-79), and nudged elastic band (41, 42), as well as the probabilistic roadmap method (80) and the finite temperature non-local exploration method of Branduardi et al. (81), and others (82, 83). See also these

review articles (84, 85). Steered molecular dynamics (86), targeted molecular dynamics (34, 35), and restricted perturbation-targeted molecular dynamics (87, 88) are intensive approaches that use biased dynamics to create pathways, recently extended to determine minimum free energy pathways (88). In contrast to these approaches, geometric targeting lacks a molecular mechanical force field and does not sample according to a Boltzmann distribution. While less rigorous than these approaches, geometric targeting can be thought of as a "back of the envelope" pathway calculation that considers only geometry and rapidly produces a plausible result at atomic-level detail. For very large systems, the sophisticated techniques may be intractable, making geometric targeting an alternative that can produce a stereochemically-correct all-atom pathway.

Other methods for creating pathways include elastic network models, which invoke a small-amplitude approximation on a system of interconnected springs to produce a simplified, smooth harmonic energy landscape. Examples include Elastic Network Interpolation models (50-52, 89) and the Plastic Network Model (90), which all use coarse-graining at the level of Cα atoms and in some cases rigid clusters of Cα atoms. Iterative cluster-normal mode analysis (91) (ic-NMA) includes all atoms, grouped into rigid clusters no smaller than a residue, with springs that attach pairs of atoms in distinct clusters. While successful at generating approximate transition pathways, some of the limitations of these models are the lack of atomic-level detail (excepting ic-NMA), the neglect of atomic-overlaps in the elastic network energy function, and an overly flexible protein backbone because of the neglect of covalent bond geometry. Compared to elastic network models, the advantages of geometric targeting are the all-atom geometric detail of the snapshots produced, and the dynamic prevention of atomic overlaps which allows more complicated motions in which atoms bump and move around each other. However, in some cases, elastic network based models do better at capturing the relative timing of events along the pathway (see Discussion).

Linear interpolation with energy minimization is a rapid technique for making pathways used by the Yale Morph Server (23, 92, 93), however these pathways are often not physically plausible because atoms and chains can pass through each other. In the Results section, we will show examples for which the Yale Morph Server's technique results in unphysical pathways, but for which geometric targeting produces plausible, non-linear, complex motions without chains passing through each other.

We wish to point out the relationship of the present geometric targeting method to prior work in constraint-based exploration of protein structures. Wells et al. described a method for exploring freedom in protein structures based on geometric constraints, called FRODA (21). One of the applications of FRODA described by the authors is targeting (21, 92). FRODA was the original idea that sparked the ideas for the geometric targeting method presented here, and the two methods share a similar philosophy but employ different underlying mathematical techniques. Some differences will be pinpointed in the Discussion section. Another technique, called tCONCOORD, also uses geometric constraints for sampling of protein structures (94, 95). Targeting and pathway generation are not possible uses of tCONCOORD, because each generated structure is completely uncorrelated with the previously-generated structure.

In this paper, we present results for over twenty proteins of various sizes and exhibiting a wide variety of conformational changes, including hinge motions, shear motions, loop rearrangements, side chain rearrangements, domain swapping, and other complex changes that are not easily classified.

WEBSERVER USAGE

The webserver is located at http://pathways.asu.edu. The webserver prompts the user to submit the initial and target protein structures in PDB format. The two proteins need not have identical atoms. Mutational differences and incomplete target structures are acceptable. The files also do not need to contain hydrogen atoms, as these will be added

automatically. Missing residues or atoms in the initial structure will not be modeled, however. If multiple chains exist, the webserver will prompt the user to decide how chains from one structure map to the chains of the other structure. After submitting the two structures, some automatic preprocessing takes place, and then the targeting begins. The targeting often completes in a few minutes. Depending on the size of the protein and the amount of structural difference between the two states, some runs can require an hour or longer. During the targeting, the web page is continuously updated, showing the current RMSD-to-target and current number of generated snapshots.

When the pathway is complete, the user views an animation of the pathway in an interactive Jmol window (96). The user can adjust the zoom level, rotate, and translate the protein while watching the pathway. A movie file can optionally be generated and downloaded. The atomic trajectory can also be saved to disk in PDB format for further analysis.

Various "Advanced Options" are also available when the user submits structures for targeting. As described later in the text, the user can opt to include random motion, enable backtracking, and choose whether to make certain hydrogen bond and hydrophobic contact constraints fixed or breakable. Additionally, the user may modify the hydrogen bond cutoff energy. Geometric constraints will be placed in the protein for hydrogen bonds that are stronger than the cutoff energy. Therefore, lowering this cutoff energy will result in fewer hydrogen bonds and hence a more flexible protein.

All targeting results are stored in the "File Cabinet," allowing a user to return to a previous targeting run, visualize the pathway, and download movies or trajectories as needed. Targeting runs continue even if a user navigates away from the page or disconnects from the internet, and the file cabinet can be used to access these runs.

METHODS

Preprocessing

When the user submits the initial and target PDB structures to the webserver, the webserver automatically carries out some preparatory steps behind the scenes before running the targeting. First, waters, hydrogens and ligands are removed. The program Reduce from the Richardson Lab is run on the initial and target structures to add hydrogens and optimally position them (97). Next, in order to determine which atoms from the initial structure correspond to which atoms in the target structure, a sequence alignment between the initial and target structures is performed by running ClustalW (98). This means that mutational differences or incomplete target structures are acceptable. Some atoms in the initial structure may have no matching counterpart in the target structure, and vice versa.

Geometric model

The targeting method begins by constructing a geometric model of the protein, using the initial structure as reference. The geometric model is an all-atom model, including hydrogens. In order to capture the geometric characteristics of covalent bonds, we subdivide each amino acid of the protein into rigid subcomponents based on the assumption of rigid bond distances, rigid 3-body angles. Dihedral angles for single covalent bonds are not constrained, but dihedral angles for higher bond orders (including peptide bonds) are treated as rigid. The grouping of atoms into rigid units is performed by the software package FIRST (56), run in a modified fashion so as to only include covalent bonding geometry in making rigid unit assignments. Fig. 2.1 shows an example of how the amino acid phenylalanine is subdivided into rigid units. In phenylalanine, the Cα plus its four covalent neighbors constitute one rigid unit, the Cβ plus its four neighbors are a second rigid unit, the phenyl ring's 6 carbon atoms plus their covalent neighbors are a third, and the peptide planes on both sides of the Cα are rigid. Within the 20 standard

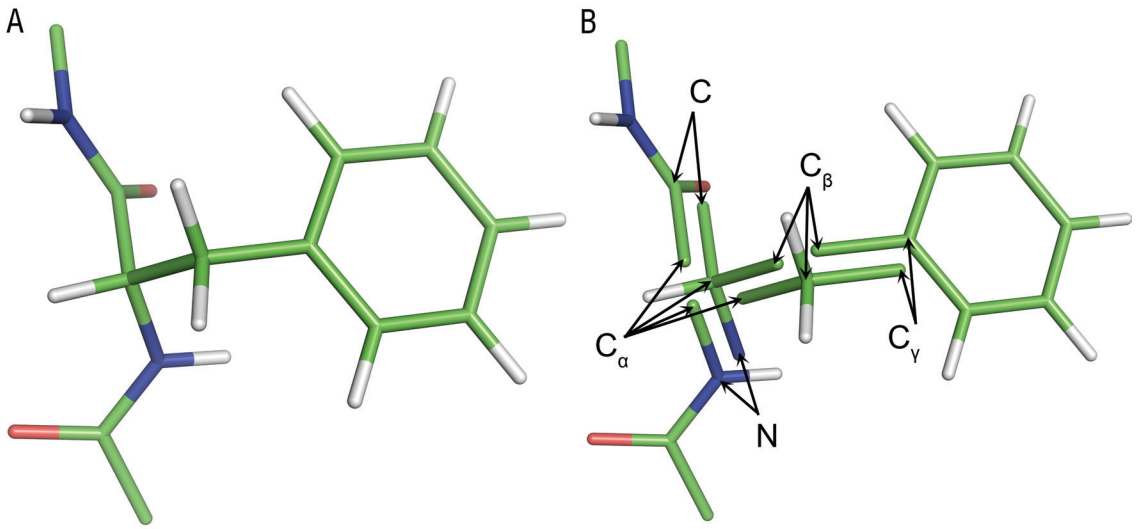

FIGURE 2.1 Decomposition of phenylalanine into rigid units. *(A)* Stick model of phenylalanine with main-chain atoms of neighboring residues. *(B)* Atoms are embedded within rigid units, which lock in place covalent bond distances and angles. Note that a single atom may have multiple copies, each belonging to a different rigid unit, as pointed out with arrows. Graphics were produced with PyMol (127).

amino acids, the largest rigid unit is the planar indole group found in tryptophan. Since there are no free atoms in the amino acids, the smallest rigid units have 3 atoms, such as the C-OH in the side chain of tyrosine or serine.

The rigid units are the mobile units of the system, each having 6 degrees of freedom (3 translational, 3 rotational). The rigid units contain "embedded atoms," whose positions depend entirely on the degrees of freedom of their corresponding rigid unit. Observe that several "embedded atoms" may correspond to the same (physical) atom, for example, the (split) Cα and Cβ atoms in Fig. 2.1 *B*. We define the position of the atom as being located at the mean of its embedded atoms. Therefore, the positions of the "atoms" depend on the positions of the "embedded atoms," which in turn depend on the rigid unit degrees of freedom.

Having divided the amino acids into rigid units, we then establish various geometric constraints that define allowed and disallowed configurations of the rigid units. The constraints are meant to maintain various aspects of stereochemical quality:

Covalent bond geometry between adjacent rigid units

To make the multiple copies of an shared atom shared by adjacent rigid units coincide, we constrain the distance between the multiple copies of a shared atom to be zero.

Non-bonded overlap

Between all non-bonded atoms (fourth neighbors and higher), we place a minimum distance constraint that keeps their separation greater than some cutoff value, the cutoffs depending on the types of atoms involved. The cutoff distances have been calibrated from MD simulations. This will be described in more detail in an article currently in preparation with collaborators Tatyana Mamonova and Maria Kurnikova (unpublished) (In this dissertation, the development of the cutoff distances are presented in Chapter 4. See Tables 4.1 and 4.4).

Backbone dihedral angles

In order to keep backbone dihedral angles out of the disfavored and disallowed regions of the Ramachandran plot (28), we establish additional minimum distance constraints between certain pairs of main-chain atoms. Note that we do not explicitly constrain the dihedral angles, but instead use distance constraints between atoms in order to block off the disallowed and disfavored backbone dihedral angles. Here we make use of the work of Ho et al. (29), in which they show that the disallowed and disfavored regions of the Ramachandran map can be understood simply from non-overlap constraints between certain pairs of main-chain atoms. We apply minimum distance constraints to the same main-chain pairs identified in their paper (see Table 2.1). These constraints apply to all 20 standard amino acids, including proline and glycine.

Main-chain pair	Minimum distance cutoff (Å)
Cβ…O	2.80
Cβ…N_{i+1}	3.00
O_{i-1}...O	3.10
O_{i-1}...N_{i+1}	3.00
O_{i-1}... Cβ	3.05
H... H_{i+1}	1.85

TABLE 2.1 Distance constraints for main-chain pairs. These minimum distance constraints are imposed on all of the standard 20 amino acids to keep main-chain dihedral angles out of the disallowed and disfavored regions of the Ramachandran plot. Ho et al. (29) found that the disallowed and disfavored regions arise because of steric clashes between these pairs of atoms. The distance cutoffs used here are slightly modified from the distances published by Ho et al. (29)

Side-chain torsion angles

To keep side-chain torsion angles away from unfavorable eclipsed configurations, we define constraints between 1-4 bonded atom pairs only in cases where atoms 2 and 3 are single-bonded and are each tetrahedrally-coordinated. To ensure that these dihedral angles remain staggered, it is not necessary to directly constrain the dihedral angles. Instead, we place a minimum distance constraint between each 1-4 pair, setting the minimum distance such that the dihedral angle between them comes no closer than 55°. These constraints partially account for rotamer configurations, but not completely, because bonds between non-tetrahedrally-coordinated atoms are left freely rotatable.

Hydrogen bonds and hydrophobic contacts

Hydrogen bonds and hydrophobic contacts are preserved by placing maximum-distance constraints between pairs of interacting atoms. For hydrogen bonds, we only place a constraint for those that have an energy score better than some cutoff value, typically -1.0 kcal/mol, as measured by an energy function (56, 99). The maximum distance constraint is placed between the hydrogen and the acceptor atoms and is set to the distance that is in the initial structure, but not less than 2.0 Å. For hydrophobic contacts, we place maximum distance constraints between pairs of hydrophobic side-chain carbon or sulfur atoms that are closer than 3.9 Å in the initial structure. We consider

only the hydrophobic residues Leu, Ile, Val, Phe, Trp, Met, Ala, Tyr. The maximum distance constraint for hydrophobic pairs is set to the distance in the initial structure plus an extra 0.5 Å.

Geometry in input structures takes precedence

In establishing the minimum distance constraints described above for non-overlap, Ramachandran, and side chain torsion, we make sure that the constraints do not conflict with the geometry in the input structures. If a pair of atoms in either the initial or target structures is found closer than would be normally allowed by a minimum-distance constraint, the minimum distance cutoff for the pair is altered and set to the actual distance in the input structure.

Targeting procedure

The strategy we use to bring the system from the initial to the target state is to impose a constraint on the RMSD-to-target, gradually decreasing this constraint towards 0 Å RMSD. While bringing the RMSD to zero, we also enforce the structural constraints to keep the snapshots stereochemically acceptable. By enforcing the structural constraints, atoms will be forced to move along curved trajectories, as they must maintain distance and angle relationships while moving towards the target.

We will first describe the targeting procedure in its most basic form, and then describe some optional modifications to the procedure. In the most basic form, the targeting procedure involves no random motion, producing a smooth pathway. Furthermore, only hydrogen bond and hydrophobic constraints that are common to both the initial and target structures are included. Otherwise, incompatible hydrogen bonds or hydrophobic contact constraints could prevent reaching the target state.

The targeting begins with the atoms in their initial positions from the submitted initial structure. The RMSD of the initial structure relative to the target structure,

calculated over all targeted atoms, is some number C_0. An RMSD step size δ is chosen, typically 0.1 Å or less. Each targeting step consists of the following actions:

1. Advance the RMSD constraint. Set the RMSD constraint to $RMSD < C_i$, where $C_i = C_{i-1} - \delta$, where subscript i denotes the step number.

2. Enforce constraints. The RMSD constraint and structural constraints are enforced simultaneously, causing the rigid units of the system must move and rotate, often taking atoms in curved paths. The process is described in the section Enforcement of Constraints.

3. Global fitting. Finish the step by globally rotating and translating the entire system to optimize the RMSD to the target.

4. If structure is acceptable, move on to next step. The criteria for judging whether the structure is acceptable are that the non-overlap constraints not be violated by more than 0.2 Å, and that the shared atoms between adjoining rigid units not be more than 0.2 Å apart. In the most basic form of targeting, the targeting steps are terminated here if the structure is not acceptable. This can happen when the targeting has run up against a particularly difficult obstacle that it cannot find a way to get around without violating structural constraints.

Random motion

The basic targeting procedure described above contains no random motion. The resulting pathway is deterministic, and atoms appear to move smoothly. To produce a random pathway, random motion can be optionally added to each targeting step as follows. At the beginning of each step, each rigid unit is randomly displaced and rotated, without regard for any constraints. The rest of the targeting step continues as usual. The constraint violations created by the random perturbation are restored during the "Enforce Constraints" portion of each step. The size of the perturbation is rather large, on the scale of 1 Å for translational displacement and 120° for rotational motion, so that rigid units

can hop over disallowed dihedral angle regions. This can cause some rigid units to get stuck during the enforcement of constraints, in which case the problem rigid units are restored to their original positions and orientations.

Options for handling of hydrogen bond and hydrophobic contact constraints

"Common" hydrogen bond and hydrophobic constraints are those that are found in the initial and target structures. In the basic targeting procedure, the common constraints are kept fixed throughout the targeting under the assumption that the interactions are present during the entire pathway. As an option, the common constraints can be made breakable, or can be not included, instead of kept fixed. When a breakable constraint becomes stretched beyond a certain amount, it "breaks" and is removed. This can be helpful if some hydrogen bond or hydrophobic contact that is found in both structures needs to transiently break during the pathway. "Non-common" constraints are those that are in the initial structure but not in the final structure. The basic setting is to simply not include the non-common constraints since they are incompatible with the final structure. Optionally, the user can choose to include the non-common constraints as breakable constraints. Having the non-common constraints included may improve the quality of the pathways, since they preserve favorable interactions until the moment they break.

Recovery methods

In the basic targeting procedure, if the shared-atom constraints and non-overlap constraints cannot be satisfied to within tolerance, the structure is deemed unacceptable and the targeting is terminated. Usually this does not happen until the very end, when the RMSD to target is quite low (<0.5 Å), and all the atoms are very close to their targets. It can sometimes happen earlier, when a particularly difficult obstacle in the pathway can cause the targeting to fail to produce an acceptable structure. A few recovery methods are available to try to help the protein move around the obstacle. The first is called "random

retry," which is to retry the last step using a random perturbation of the rigid units as described above in hopes that the random motion will help move past the obstacle. Typically up to 5 consecutive random retries are attempted.

Another available recovery method is "Backtracking." In backtracking, the targeting steps switch into reverse , taking the RMSD away from the target instead of closer to the target. The sign of the RMSD step δ is reversed so that the RMSD constraint C_i increases instead of decreases at each step. The inequality in the RMSD constraint is also switched to a greater-than sign, $RMSD > C_i$, to carry the system away from the target. The idea is to go back in RMSD, find a new starting point at the higher RMSD level, then return to forward steps, in hopes that this enables the system to get around an obstacle. The method used to find a new starting point at the elevated RMSD before returning to forward steps is called "momentum steps," described later. The first time that a targeting step fails to produce an acceptable structure, the system is backtracked by 1 Å, a new starting point is found, and then the procedure returns to regular forward steps. If the targeting again gets stuck, the backtracking method tries going back by 2 Å, then 4 Å, then 8 Å, etc., doubling the amount of backward motion each time. The backtracking can even take the protein back in RMSD farther than the initial state. All non-common constraints are removed during backtracking so they do not hinder the system from going back in RMSD.

Momentum steps

During backtracking, when the RMSD has been taken back to some higher value, we use "momentum steps" to find a new starting point before recommencing steps toward the target. Momentum steps are so named because the motion tends to persist in the same direction over many steps. Note that momentum is not actually conserved, since we are not integrating equations of motion, and there are no time steps or velocities. Here, the net translational and rotational change of each rigid unit is recorded for each step and

used as a perturbation in the next step. Throughout the momentum steps, the upper-bound RMSD constraint is kept active, ensuring that the RMSD does not go back further. A momentum step involves the following actions:

1. Store current configuration. The six degrees of freedom of each rigid unit are stored in a $6M$ -dimensional vector $\mathbf{q}_1$, where M is the number of rigid units.

2. Perturb rigid units by the last $\Delta\mathbf{q}$. The rigid units are translated and rotated by the $\Delta\mathbf{q}$ of the previous momentum step, or 0 if this is the first momentum step. The system is now at a new configuration $\mathbf{q}_2$.

3. Small Random Perturbation. Randomly perturb the rigid units (translationally and rotationally), but do so with a very small amplitude (atoms move by only about 0.05 Å). The system is now at $\mathbf{q}_3$.

4. Enforce constraints. Both the RMSD constraint and the structural constraints are enforced, bringing the system to state $\mathbf{q}_4$.

5. Global Fit to Target Structure. Remove any global translations and rotations by globally fitting to the target structure, bringing the system to state $\mathbf{q}_5$.

6. Calculate net change. Determine the net change of the degrees of freedom in this momentum step, $\Delta\mathbf{q} = \mathbf{q}_5 - \mathbf{q}_1$, for use in the next step. Then move on to the next step.

Due to the small random component being added in each step, components of the motion along soft directions gradually grow in size. Because constraints are enforced in each step, components of the motion that encounter constraints cannot persist more than one step and cannot grow. After several steps, large-amplitude motions develop, which enables fast movement to a new position.

Note that the RMSD constraint, which is kept active during the momentum steps, is only an upper-bound, so the RMSD of the system is free to decrease. Entropy, however,

usually keeps the RMSD as high as the constraint allows, since there tend to exist more states at high RMSD than low RMSD.

Enforcement of constraints

To explain how constraints are enforced, we must clarify the mathematical relationship between the rigid unit degrees of freedom and the positions of the atoms. Recall that the position of an atom is located at the mean position of its corresponding "embedded atoms," which in turn depend on the rigid unit degrees of freedom of their respective rigid units. Let $\bar{\mathbf{r}}$ be a $3N$-dimensional vector containing the (mean) positions of the N atoms, $\mathbf{r}$ be the $3N'$-dimensional vector containing the positions of the N' embedded atoms, and $\mathbf{q}$ be the $6M$-dimensional vector containing the rigid unit degrees of freedom of the M rigid units. For the translational degrees of freedom of a rigid unit, we use the Cartesian coordinates of the centroid of the rigid unit. For the rotational degrees of freedom of a rigid unit, we use three independent rotor variables from geometric algebra, B_x, B_y, B_z, as described in Wells (100). These three rotor variables can be interpreted as a 3-dimensional vector $\mathbf{B}$ that points along the axis of rotation and has a magnitude $|\mathbf{B}| = 2\sin\frac{\phi}{2}$, where ϕ is the angle of rotation. See Wells (100), for how these variables can be used to describe rigid body rotations (In this dissertation, the mathematics are presented in Appendix A).

To help the rigid units find their way to an acceptable state, we define an "energy function" that measures the total amount of constraint violation in the system. We then perform conjugate gradient minimization to find the configuration of rigid units that minimizes the constraint energy (101). In the constraint energy function, each constraint is represented by an energy term that is zero if the constraint is met and increases quadratically with the amount of constraint violation. It is important to recognize that the snapshots produced by geometric targeting lie within the flat portion of the energy landscape at energy zero (or near zero when some constraint violations cannot be fully

resolved). The non-zero region of the energy landscape only serves to guide the system back to the flat, zero energy region.

$$V = V_{\text{shared atoms}} + V_{\text{min dist}} + V_{\text{max dist}} + V_{RMSD} \tag{2.1}$$

$$V_{\text{shared atoms}} = \sum_{p<q}{}' \tfrac{1}{2} k \left| \mathbf{r}_p - \mathbf{r}_q \right|^2 \tag{2.2}$$

$$V_{\text{min dist}} = \sum_{i<j}{}' \begin{cases} \tfrac{1}{2} k \left(\left| \overline{\mathbf{r}}_i - \overline{\mathbf{r}}_j \right| - d_{ij}^{min} \right)^2, & \left| \overline{\mathbf{r}}_i - \overline{\mathbf{r}}_j \right| < d_{ij}^{min} \\ 0, & \left| \overline{\mathbf{r}}_i - \overline{\mathbf{r}}_j \right| \geq d_{ij}^{min} \end{cases} \tag{2.3}$$

$$V_{\text{max dist}} = \sum_{i<j}{}' \begin{cases} \tfrac{1}{2} k \left(\left| \overline{\mathbf{r}}_i - \overline{\mathbf{r}}_j \right| - d_{ij}^{max} \right)^2, & \left| \overline{\mathbf{r}}_i - \overline{\mathbf{r}}_j \right| > d_{ij}^{max} \\ 0, & \left| \overline{\mathbf{r}}_i - \overline{\mathbf{r}}_j \right| \leq d_{ij}^{max} \end{cases} \tag{2.4}$$

$$V_{\text{RMSD}} = \begin{cases} \tfrac{1}{2} kN \left(\text{RMSD} - C \right)^2, & \text{RMSD} > C \\ 0, & \text{RMSD} \leq C \end{cases} \tag{2.5}$$

In the above equations, $\mathbf{r}_p$ is the position of "embedded atom" p, $\overline{\mathbf{r}}_i$ is the position of atom i (mean position of its multiple embedded atoms), d_{ij}^{min} and d_{ij}^{max} are the minimum and maximum distance constraints for atoms i and j. The prime symbols in the summations denote that sums are only over pairs i, j or m, n for which a constraint exists.

Conjugate gradient minimization of V takes the system to a local minimum, using the gradient of the energy function to guide the search for the minimum (76, 101). The gradient must be taken with respect to the system's degrees of freedom $\mathbf{q}$ (the rigid unit degrees of freedom). Since the various terms of V are explicitly expressed as functions of the positions of the atoms $\overline{\mathbf{r}}$ or the positions of the embedded atom copies $\mathbf{r}$, rather than as functions of $\mathbf{q}$, chain rules must be used to obtain the derivatives $\partial V / \partial q_i$ for each degree of freedom q_i (see Appendix A of the dissertation). To make the system better-conditioned for conjugate gradient, the unitless rotor degrees of freedom are each

scaled by a characteristic length-scale so that they are comparable in magnitude with the translational degrees of freedom. In addition, diagonal elements of the second derivative matrix, $\partial^2 V / \partial q_i^2$, are calculated and used as a preconditioner (101).

To determine when to stop the conjugate gradient minimization, we make an estimate of how close each degree of freedom is from the local minimum. In the approximation that each degree of freedom lies in an independent parabolic well $\frac{1}{2} k \left(q_i - q_{i0} \right)^2$ for some unknown minimum-energy position q_{i0}, taking the ratio of $\partial V / \partial q_i$ to $\partial^2 V / \partial q_i^2$ gives $q_i - q_{i0}$, which is the estimate of the error. We stop the conjugate gradient minimization when this error estimate is below some tolerance value, typically 0.005 Å.

Ideally, conjugate gradient minimization would bring the energy to zero, meaning that all constraints are satisfied. In practice, local minima in the energy function can arise from mutually incompatible constraints (an RMSD constraint pulling a side chain through an eclipsed configuration, for example), preventing certain constraints from being fully satisfied.

RESULTS

We applied the geometric targeting method to over twenty proteins of various sizes and exhibiting a wide variety of conformational changes, including hinge motions, shear motions, loop rearrangements, side chain rearrangements, domain swapping, and other complex changes that are not easily classified. Some of these examples were selected from the Database of Macromolecular Movements (93). Results are summarized in Table 2.2. For each system, an initial targeting attempt was made using the following settings: no random motion, no backtracking, RMSD step size of 0.1 Å, common hydrogen bonds and hydrophobic contacts treated as fixed constraints, non-common hydrogen bonds and hydrophobic contacts left unconstrained. Random retry steps were used as a recovery method. With these settings, the targeting was successful for most

Protein Name	# Sub-units	# Atoms	Initial RMSD (Å)	**Final RMSD (Å)**	**CPU Time (min)**	# Steps	CPU Time per step per # atoms (ms)	Fig.	Movie
Basic settings[1]									
Toy Model 1	1	129	7.6	**0.0**	**0.0**	76		2 *A*	1
Collagenase	1	1770	8.1	**0.2**	**0.4**	85	0.16		
Calmodulin	1	2262	5.5	**0.0**	**0.2**	55	0.08		
Dihydrofolate Reductase	1	2489	1.9	**0.1**	**0.3**	26	0.31		
Pyrophosphokinase	1	2535	3.0	**0.1**	**0.3**	36	0.22		
Spindle Assembly Checkpoint Protein	1	3033	10.2	**0.2**	**1.9**	106	0.36	2 *B*	2
CD2	2	3096	23.2	**0.2**	**3.4**	237	0.28	2 *D*	3
Adenylate Kinase	1	3341	7.2	**0.1**	**0.6**	77	0.13		
Alcohol Dehydrogenase	1	5639	2.1	**0.2**	**1.2**	25	0.49		
Heparin Cofactor II	1	6931	6.1	**0.0**	**1.9**	61	0.27		
Diphtheria Toxin	1	7972	16.0	**0.2**	**4.7**	164	0.21	2 *C*	4
5'-Nucleotidase	1	8120	10.1	**0.2**	**2.8**	105	0.20		
Citrate Synthase	2	13182	3.3	**0.5**	**6.4**	34	0.85		
Pyruvate Phosphate Dikinase	1	13420	11.7	**0.1**	**4.2**	118	0.16		
DNA Polymerase	1	14563	6.6	**0.2**	**4.9**	70	0.29		
HIV-1 Reverse Transcriptase	2	15299	4.1	**0.3**	**8.6**	45	0.75		
Phosphofructokinase	4	19140	2.1	**0.2**	**8.2**	28	0.92		
Replication Factor C	6	29966	14.0	**0.0**	**16.9**	145	0.23		
Rho Transcription Termination Factor	6	37136	2.1	**0.2**	**15.9**	25	1.03		
GroEL	14	109718	11.2	**0.2**	**115.0**	115	0.55	2 *E*	5
Basic settings[1] **+ backtracking**									
Toy Model 2	1	1611	10.8	**0.3**	**9.3**	-	-	3 *A*	6
HIV Protease	2	3144	2.0	**0.2**	**1.6**	-	-	3 *B*	7
Dengue 2 Virus Envelope Glycoprotein	1	6129	12.1	**0.2**	**5.7**	-	-	3 *C*	8
Basic settings[1] **+ breakable hydrogen bond and hydrophobic contact constraints**									
Immunoglobulin E SPE7	2	3467	2.7	**0.5**	**1.3**	29	0.77		

[1]Basic settings: no backtracking, no random motion (except during retry steps), 5 random retry steps enabled, 0.1 Å RMSD step size, hydrogen bonds and hydrophobic contacts common to both structures treated as fixed constraints, non-common hydrogen bonds and hydrophobic contacts not included as constraints.

TABLE 2.2 Webserver pathway results. Pathways were generated for the listed examples. As indicated, many completed successfully even without random motion or backtracking. Other pathways were only successful when backtracking was enabled or when common constraints were made breakable, as indicated. The number of atoms in the initial state includes hydrogens. Initial and final RMSD are computed with respect to the target structure, using all targeted atoms including hydrogens. The reported CPU times correspond to a single processor. The number of steps reported includes random retry steps. The number of steps and time per step is not reported for backtracking runs, because these runs include a mixture of forward steps, backward steps, and momentum steps, each of which have different characteristic times per step. The "Fig." column lists figure numbers for

TABLE 2.2, continued
those pathways represented in the figures, and the "Movie" column lists the Supplementary Movie number (available at the publisher's website http://doi.wiley.com/10.1002/prot.22810).

systems. Proteins that could not reach their targets under these targeting settings were re-run with backtracking enabled in order to get around significant obstacles in the pathway. One system was unsuccessful even with backtracking enabled, but was successful when the common set of hydrogen bonds and hydrophobic contacts were allowed to break, instead of keeping them fixed (Table 2.2). All examples successfully reached their targets within very low all-atom RMSD (< 0.5 Å). Runs typically completed within minutes, with the largest case GroEL requiring almost 2 hours. Run times scaled roughly in proportion to the number of atoms and the RMSD difference between the two states. Rather than describing in detail the results for each protein, we highlight below a few examples that suffice to demonstrate the versatility and robustness of the method (movies for these examples can be found in the online Supplementary Material at the publisher's website, http://doi.wiley.com/10.1002/prot.22810). Protein Data Bank (102) (PDB) IDs and chain information for all examples are listed in Appendix B, Table B.1.

A few of the successful examples discussed below are known to yield unphysical pathways under the linear interpolation method of the Yale Morph Server (23), namely diphtheria toxin and GroEL, with groups of atoms passing through each other as discussed in their paper and available for viewing at their website (93). A third unphysical example not discussed in their paper but available on their website (93) is spindle assembly checkpoint protein.

Fig. 2.2 *A* shows results for a toy model system designed to illustrate how the targeting procedure can produce highly non-linear pathways, without the use of backtracking or random motion. Toy model 1 is a poly-alanine fragment of length 12 residues starting in an extended beta-like configuration that was targeted to an alpha-helix

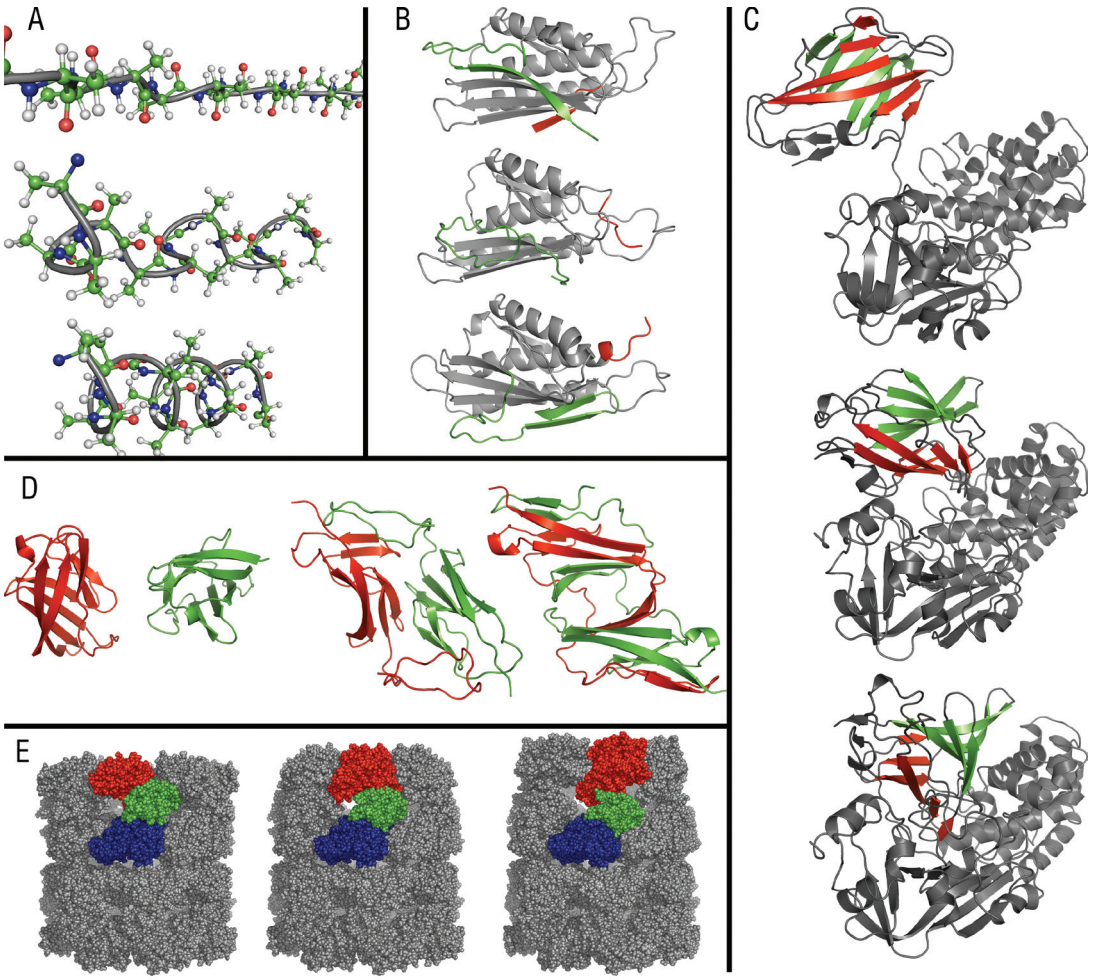

FIGURE 2.2 Example pathways that completed successfully without the use of backtracking or random motion. Each panel shows three pathway snapshots: the initial structure, an intermediate snapshot, and the final snapshot. Colored regions in green, red, and blue *(B-E)* highlight particular motions, described in the text. *(A)* Toy Model 1. A poly-alanine strand of beta sheet gradually folds to alpha helix while maintaining geometric constraints. *(B)* Spindle assembly checkpoint protein. A strand of beta sheet (red) passes under a loop and joins with an alpha helix, while a second strand of beta sheet (green) moves from the left to right side of the beta sheet. *(C)* Diphtheria Toxin. A very large domain rotation of nearly 180° is shown. Observe the relative position of the red and green colored beta sheets over the course of the rotation, in which the green beta sheet begins behind the red beta sheet, then rotates to be on top, then rotates further to end up on the right of the red. *(D)* CD2. Two monomers (red and green) dimerize, forming a domain-swapped dimer. *(E)* GroEL. Each subunit in the upper ring of the large 14-subunit complex undergoes a transition involving a large clockwise rotation and upward tilt of the apical domain (red) and a downward tilt of the intermediate domain (green), while the equatorial domain (blue) remains relatively unchanged. For clarity, only one subunit is colored.

configuration. The initial structure had no hydrogen bonds or hydrophobic contacts, so the only constraints active in the system were the minimum distance constraints between atoms (preventing overlap, disfavored Ramachandran regions, and eclipsed side-chain configurations), and the shared atom constraints between connected rigid units. Random motion was not used, so the motion is driven solely by the gradually changing RMSD constraint which pulls the system closer and closer to the target. As atoms are pulled towards their targets, they follow curved trajectories due to the enforcement of structural constraints. The Ramachandran constraints and non-eclipsing side chain constraints pose particularly difficult obstacles, creating disallowed regions in dihedral angle space that must somehow be crossed in order to reach the target. Sometimes the rigid units are observed to make sudden jumps as the RMSD constraint pulls them from one allowed region to another, moving over a disallowed region. The RMSD constraint energy term in the constraint energy function lifts the system over a barrier created by a minimum-distance energy term, and minimization carries the system downhill to the other side.

The conformational change in spindle assembly checkpoint protein is complicated (Fig. 2.2 *B*), involving a strand of beta sheet (red) that passes under a loop and joins with an alpha helix, while a second strand of beta sheet (green) moves from the top to bottom side of the beta sheet. On the Yale Morph Server (23) which uses linear interpolation with energy minimization, the polypeptide chains can be seen to pass through each other in an unphysical manner. With geometric targeting, the chains are observed to bump into each other and move around each other to avoid atomic overlap in reaching the target.

The pathway generated for diphtheria toxin shows a very large domain rotation of nearly 180° (Fig. 2.2 *C*), created without any backtracking or random motion. Observe the relative position of the red and green colored beta sheets over the course of the rotation, in which the green beta sheet begins behind the red beta sheet, then rotates to be on top, then rotates further to end up on the right of the red. The initial state was

taken from a domain-swapped state (only one monomer shown), and the final state was the native, non-swapped state. The pathway generated here by geometric targeting is in contrast to the result obtained from the linear-interpolation-based method at the Yale Morph Server (23), which produces an unphysical pathway with atoms passing through each other. Krebs et al. (23) declared the conformational change in diphtheria toxin to be "impossible" to compute, speculating that only a complete unfolding and refolding of the domain could explain the conformational change. The insight gained from the geometric targeting approach is that a plausible pathway does exist that does not involve unfolding/refolding, although the method makes no prediction as to the actual or optimal pathway. This is in harmony with the results from elastic network interpolation on this protein (89).

A misfolding pathway is shown in Fig. 2.2 *D*, as two monomers of the protein CD2 dimerize to form a domain-swapped dimer. The motion is complicated and non-linear, involving the domains opening up and inter-digitating, which is notable considering that the pathway was generated without random motion and without any backtracking. In the figure, the two monomers were given separate colors to highlight how their beta-strands intermingle.

The conformational change in the large, 14-subunit GroEL complex is shown in Fig. 2.2 *E*. Two 7-subunit rings are stacked on top of each other, viewed from the side as the top ring undergoes an opening and twisting transition. Although the initial and final states have 7-fold rotational symmetry, symmetry was not enforced in the pathway and all atoms were explicitly simulated. In the figure, the apical, intermediate, and equatorial domains of one subunit are given distinct colors to show how they change in the pathway. GroEL is another case that results in an unphysical pathway when run on the Yale Morph Server (23). It is important to note that a targeted molecular dynamics (38) study of a single GroEL monomer indicates that the intermediate domain motion occurs first, followed by the apical domain motion. In the pathway generated from geometric

targeting, all motions occur simultaneously because it is stereochemically plausible to do so, and because the energetics of the system are not considered. Still, the types of movements involved in the transition can be seen in the pathway, even if the relative timing of events is not accurate.

Next we present some results for pathways that required backtracking in order to successfully reach the target. Toy model 2 was created to help illustrate how difficult obstacles can cause the targeting to get stuck, and how backtracking can sometimes help in these cases. The model is a 4-helix bundle in the initial state. The target state was created by moving one of the helices to the opposite side of the bundle, shown in Fig. 2.3 *A* (fourth snapshot). The other three helices form a wall that the mobile helix must somehow pass to reach its target state on the other side. The helices making up the wall cannot separate because there are hydrogen bond and hydrophobic constraints between helices that are common to both the initial and target structures, which are kept fixed. Without backtracking, the gradually decreasing RMSD simply pulls the mobile helix into the wall formed by the other three (not shown). With backtracking enabled, instead of quitting when stuck, the RMSD is backed, momentum steps are performed to find a new starting state, and the forward steps commence again. However, as shown in the second snapshot of Fig. 2.3 *A*, the system is still stuck, with the mobile helix trying to go around the side of the wall, but unable to because of its loop attachment to the wall. After more backtracking attempts, the helix is observed to flip over the top of the wall to reach the other side (Fig. 2.3 *A*, third snapshot).

In HIV protease, initial and target structures were chosen that would require the two flexible arm regions to move past each other (Fig. 2.3 *C*). The green-colored region is behind the red region in the initial state, but is in front of the red region in the target state. Without backtracking and without the use of random motion, the RMSD constraint pulls these two arm regions into each other, causing them to collide and get stuck (Fig. 2.3 *C*,

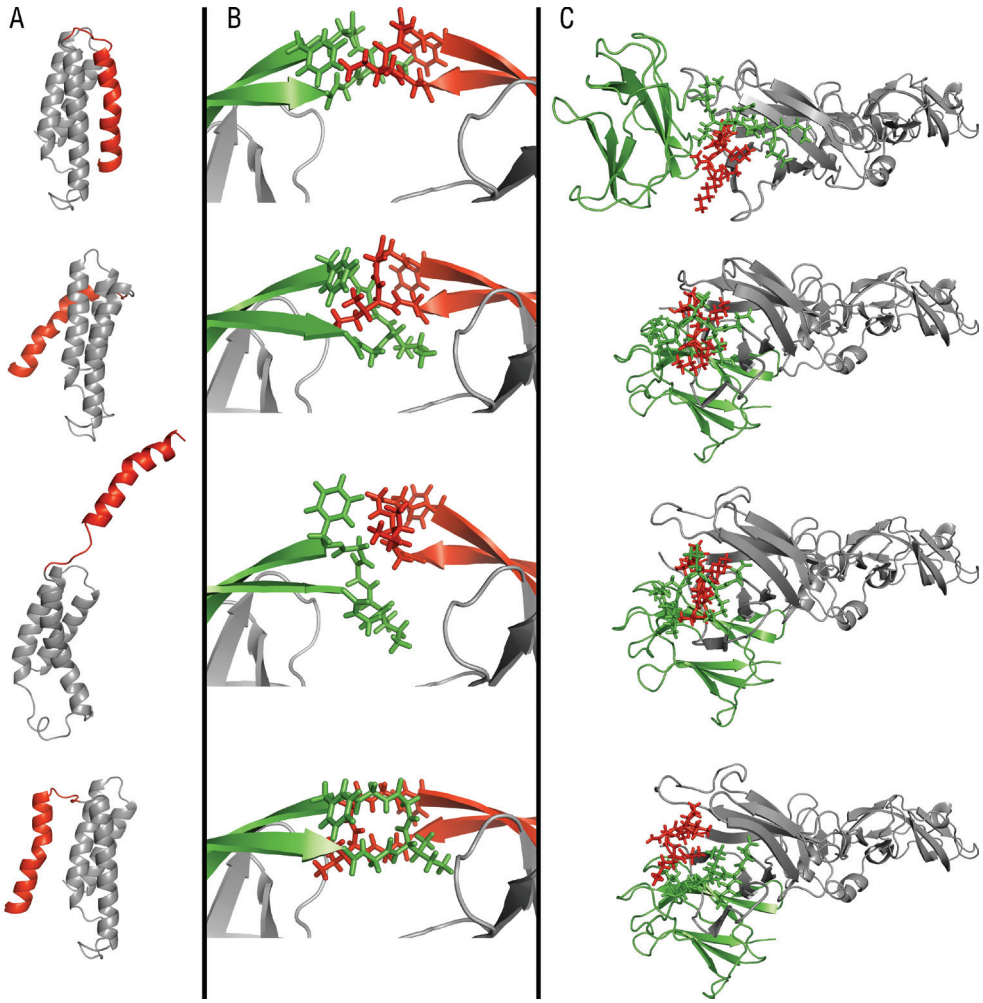

FIGURE 2.3 Example pathways that required backtracking. Each panel shows 4 pathway snapshots: the initial structure, a snapshot when the system encounters an obstacle, a snapshot when the protein moves around the obstacle, and the final snapshot. *(A)* Toy Model 2. One helix (red) of a 4-helix bundle must move from the right side to the left, the other three helices forming a wall. The helix tries to go around the side of the wall (second snapshot), but is unable to because of its loop attachment to the wall. After more backtracking attempts, the helix is observed to flip over the top of the wall (third snapshot) to reach the other side. *(B)* HIV Protease. Two arm regions (green and red) must somehow switch places, with the red-colored arm moving from front to back. Targeting initially gets stuck (second snapshot), but after backtracking finds a way around the obstacle (third snapshot). *(C)* Dengue 2 Virus Envelope Glycoprotein. A hinged domain (green) closes up against the stable portion of the protein (gray), but a small loop region (red) that needs to move gets pinned by the closing domain (second snapshot). With backtracking enabled, the system finds a new starting configuration that is not obviously very different (third snapshot), however when forward steps recommence the red loop region is able to slip out and move to its target position (fourth snapshot).

second snapshot). With backtracking enabled, the arms back away after colliding. During the momentum steps that follow the backward steps, the red region moves over the top of the green region (Fig. 2.3 *C*, third snapshot). As forward steps begin again, the system has moved around the obstacle, enabling the RMSD constraint to pull the system to the target state.

The main part of the motion in dengue 2 virus envelope glycoprotein, shown in Fig. 2.3 *C*, is the closing of a hinged domain (green) against the stable portion of the protein (gray). This motion is accomplished easily, however a small loop region (red) that needs to move gets pinned by the closing domain (Fig. 2.3 *C*, second snapshot). With backtracking enabled, the system finds a new starting configuration that is not obviously very different (Fig. 2.3 *C*, third snapshot), however when forward steps recommence the red loop region is able to slip out and move to its target position (Fig. 2.3 *C*, fourth snapshot).

In the antibody Immunoglobulin E SPE7, we performed targeting between two structures that exhibited some loop and side chain conformational differences in the heavy chain. We found that we had to make the common hydrogen bond and hydrophobic contact constraints breakable in order to successfully reach the target. Typically, common constraints are kept fixed, under the assumption that if the interactions are present in the initial state and in the final state, they are also present at intermediate states. In this protein, all targeting attempts with common constraints kept fixed were unsuccessful, even with backtracking and random motion activated. When common constraints are made breakable, however, targeting is successful without backtracking and without random motion, indicating that the pathway requires some hydrogen bond or hydrophobic contact to transiently break and reform.

All examples presented so far did not use random motion. To demonstrate the use of random motion, we performed additional targeting runs on the HIV protease system

with random motion activated. Recall that without random motion, the arms of the protein could only move past each other if backtracking was used. With random motion added to each RMSD step, and using an RMSD step size of 0.01 Å, we find that targeting is successful without backtracking. The random motion enables the two colliding arms to find a way to slip past each other without getting stuck. Fig. 2.4 shows the variability in the pathway introduced by the random motion. The superimposed snapshots shown were taken from three independent targeting runs, at 1.24 Å RMSD from the target. Interestingly, the random motion was not successful when used in conjunction with a larger RMSD step size of 0.1 Å. With a large RMSD step, the RMSD changes so rapidly that the random motion does not have enough opportunity to prevent the arms from getting stuck.

DISCUSSION

Several successful examples of the application of geometric targeting have been presented, illustrating that the technique is generally applicable to a wide variety of conformational changes. Especially promising is the application to very large systems, demonstrated in the 14-subunit GroEL complex, for which an all-atom pathway was produced in under two hours on a single CPU. A significant improvement compared to linear-interpolation techniques (23) has been demonstrated in that the pathways produced by geometric targeting do not have chains passing through each other. The geometric constraints between atoms serve to keep the system in plausible geometric configurations,

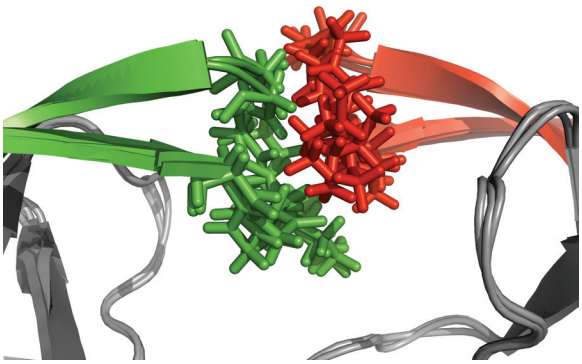

FIGURE 2.4 Pathways with random motion. Three snapshots are superimposed, each taken from the same intermediate point of three different random runs in HIV Protease. The random motion enables the arm regions to find a way to pass each other.

redirecting atoms along curved paths as the target is approached. Furthermore, backtracking, random motion, and breakable constraints have been shown to enable the system to get past particularly difficult obstacles.

Beyond the results presented, we have also found that creating a geometric pathway between an unfolded state (linear polypeptide) of a protein and its folded state is successful in some cases (data not presented). It is not successful for some folds, for example in knotted proteins. Though broadly applicable to diverse kinds of conformational changes, there are sure to be other examples of pathological transitions where geometric targeting is not successful.

By including random motion in the pathways, it is possible to create random pathways. Because the random perturbations are performed at the level of individual rigid units, however, the diffusive motion of the rigid units has little effect on domains. The domain motions are more heavily influenced by the changing RMSD constraint than they are by the random motion. In principle, using a very small RMSD step size would give the random motion more opportunity to have effect at the level of domains.

A limitation to be aware of in geometric targeting is the tendency for events to happen simultaneously along the pathway, due to the flat zero-energy landscape within which all configurations reside that meet the geometric constraints. In the GroEL case, for example, the intermediate domain and apical domain motions occur simultaneously in the geometric pathway, whereas targeted molecular dynamics (38) shows the domain movements occurring sequentially. Similarly, in adenylate kinase, the geometric pathway has the ATP-binding domain and the lid domain closing together simultaneously, rather than sequentially, as was found using elastic network models (52, 90) and in the work of Arora et al. (44) which the authors used nudged elastic band (41, 42) and umbrella sampling (53). As an attempt to induce sequential domain movement in GroEL and adenylate kinase, we tried rerunning them with non-common constraints included as

breakable constraints (instead of the usual setting, which is to not include the non-common constraints), but the correct timing was still not obtained. What is learned from geometric targeting in these two cases is that there is no steric reason why the timing of events must be a certain way, and the change can be performed while preserving all the geometric constraints. The timing must be due to energetics.

In many cases, the motions determined by geometric targeting may accurately capture the essential geometric features of a transition. On the other hand, it is important to recognize the fundamental limitations of geometric targeting that arise from its neglect of thermodynamics and lack of Boltzmann weighting. Some geometric features may be successfully captured, but features that depend on energetic considerations will be missed, such as transient hydrogen bonds and the timing of events, which could have significant effects on the pathway.

It is interesting to consider what would happen if an all-atom molecular mechanical force field was used instead of the constraint energy function, and if the rigid units were reduced to single atoms. The procedure would then be to perform energy minimizations at each RMSD level in an attempt to produce a low-energy pathway. We have not tried this, and we do not know whether such a technique would be an improvement or if new problems would arise, such as getting trapped in local minima.

It is also interesting to examine the set of all pathways to the target, under different conditions—keeping common constraints only, adding randomness and adding the option of backtracking. It is also interesting to reverse the direction of the targeting between the two protein conformations. This leads to a plethora of data that will be the subject of a future study in which “bottlenecks” along the pathway are identified—these are narrow regions of phase space through which the structure passes.

The geometric targeting method introduced in this paper has some similarities and differences with the FRODA-targeting method published earlier (21). The similarities

are in the overall idea, which in both cases is to move the system towards the target state while enforcing a set of geometric constraints to keep the structure stereochemically acceptable. Though similar in overall idea, the underlying geometric model and manner in which constraints are enforced are quite different, leading to improvements in speed, ability to successfully enforce constraints, and ability to more closely reach the target. Differences in the model and method that facilitate these improvements include the following: the use of a constraint energy function with conjugate gradient minimization to enforce constraints; the use of an explicit RMSD constraint in the targeting; small rigid units instead of large rigid clusters; maximum-distance constraints for hydrogen bonds instead of rigid distance and angle constraints; new minimum-distance constraints calibrated from MD simulations; minimum-distance constraints for maintaining good Ramachandran quality; and minimum-distance constraints for favorable side-chain torsional configurations.

CONCLUSION

We have created a new method for pathway generation in proteins, with an easy-to-use webserver, that is quick and produces stereochemically-correct pathways. When compared to more sophisticated and computationally expensive methods like targeted molecular dynamics, this method can be thought of as a "back-of-the-envelope" calculation. It is a quick and easy method to gain preliminary insights into a pathway. The geometric constraints used here model the physical reality that motion in proteins is highly constrained. While the neglect of energetic considerations certainly affects the details of the outcome, in many cases, geometric considerations alone may be sufficient to capture the essential translational and rotational motions that make up the actual pathway. At a minimum, these pathways are useful for visualization purposes, to easily see what is changing and what motions might be involved in the change. But beyond

visualization, the stereochemical quality of these pathways makes them candidates for input to more intensive quantitative approaches.

Future planned developments on the pathways website include extensions of the technique to handle RNA, DNA and ligands. This site is a companion to http://flexweb.asu.edu which uses similar techniques to explore undirected motion.

CHAPTER 3 PATHWAY COMPARISON BETWEEN GEOMETRIC TARGETING AND TARGETED MOLECULAR DYNAMICS

This chapter contains a submitted paper "Comparison of Pathways between Geometric Targeting Method and Targeted Molecular Dynamics in Nitrogen Regulatory Protein C" by Daniel W. Farrell, Ming Lei, and M. F. Thorpe. In this paper, previously published pathways from targeted molecular dynamics (TMD) are compared with new pathways from geometric targeting (GT) for the protein "nitrogen regulatory protein C." The paper also extends the methodology of the Chapter 2, describing a procedure for dynamic making/breaking of constraints. The paper was written by DWF, with some revisions by ML and MFT. DWF designed and implemented the dynamic making/breaking of constraints, generated the GT pathways, carried out the comparison of pathways between the two methods, and generated the tables and figures (except Fig. 3.1). ML supplied the TMD pathways from previously published work (24), supplied analysis scripts used to compare trajectories, and supplied Figure 3.1. Many discussions and exchanges of ideas between DWF, ML, and MFT went into the work. The paper below is as submitted, with the following exceptions: Supplementary Movies referred to in the text are not included in the dissertation.

Geometric targeting (GT) is a recently-introduced method for rapidly generating all-atom pathways from one protein state to another, based on geometric rather than energetic considerations. To generate pathways, a bias is applied that gradually moves atoms towards a target structure, while a set of geometric constraints between atoms is enforced to keep the structure stereochemically acceptable. In this work, we compare conformational pathways generated from GT to pathways from the much more computationally intensive and commonly-used targeted molecular dynamics technique (TMD), for a complicated conformational change in the signaling protein Nitrogen Regulatory Protein C. We show that the all-atom pathways from GT are similar to

previously reported TMD pathways for this protein, by comparing motion along six progress variables that describe the various structural changes. The results suggest that for Nitrogen Regulatory Protein C, finding an all-atom pathway is primarily a problem of geometry, and that a detailed force-field in this case constitutes an extra layer of detail. We also show that the pathway snapshots from geometric targeting have good structure quality, by measuring various structure quality metrics. Transient hydrogen bonds detected by the two methods show some similarities but also some differences. The results justify the usage of GT as a rapid, approximate alternative to TMD for generating stereochemically-acceptable all-atom pathways in highly constrained protein systems.

INTRODUCTION

The geometric targeting method (GT) (22) has recently been introduced as a rapid way to generate all-atom pathways from one protein structure to some known target structure, usually in a matter of minutes, available at the Geometric Pathways Webserver http://pathways.asu.edu. GT is based on the philosophy that essential features of protein conformational changes can be captured by solely considering geometric relationships between atoms. The present geometric targeting method (22) is related to prior work by Wells et al. (21), who described a method for exploring freedom in protein structures based on geometric constraints called FRODA [Framework Rigidity Optimized Dynamics Algorithm]. In GT, the protein is modeled as a constrained geometric system, with constraints established to enforce various aspects of structure quality: preserve covalent bond geometry, prevent overlap of atoms, avoid forbidden Ramachandran regions (28) for backbone dihedral angles, avoid eclipsed side-chain torsion angles, and maintain hydrogen bonds and hydrophobic contacts. To generate a pathway, the GT method takes steps from an initial structure towards a target structure, decreasing the RMSD to the target by increments of some RMSD step size δ, while enforcing the constraints so that the structure remains stereochemically acceptable. Atoms can follow

complicated paths as they maintain proper bonding geometry and bump and move around each other. GT is fast because no energetics are considered, and there is no molecular mechanical force field. The generated pathways are not minimum free energy paths or optimal high-flux pathways, but they are stereochemically acceptable pathways that can give insight into the motions involved in conformational changes. Because the method is computationally lightweight and often yields results within a few minutes, the generated pathways can be thought of as "back-of-the-envelope" or "first guess" pathways.

In the paper that introduced GT (22), stereochemically-plausible all-atom pathways were reported for proteins exhibiting large domain motions, loop rearrangements, side-chain rearrangements, domain swapping, and other complicated motions, in systems as large as the 14-subunit GroEL complex. It was also demonstrated that in cases where linear interpolation with energy minimization (23) yields severely unphysical pathways with atoms passing through each other, GT finds pathways in which the atoms do not interpenetrate. What has not yet been studied, however, is how well GT pathways compare with pathways from more computationally intensive and detailed techniques, which is the topic of this paper.

In this work, we use GT to generate all-atom pathways in nitrogen regulatory protein C (NtrC), and we compare these to recently-reported pathways (24) generated from the more commonly-used but more computationally-intensive targeted molecular dynamics approach (34) (TMD). The motion that NtrC undergoes as presented in the recent TMD work is rather complicated, sequentially consisting of tilting of the α4 helix by ~30°, axial rotation of the same helix by ~120°, the removal of half of a helical twist at one end of the α4 helix and the addition of half of a helical twist at the other end, with adjustments in the loop regions that lie sequentially before and after the α4 helix. Our aim is to show that even though the geometric model employed by GT is a large approximation and requires a factor ~10^3 less computational time than TMD, the motion

in the pathways is quite similar between the two methods. We demonstrate the similarity by comparing the pathways along six progress variables that describe the multiple structural changes. Transient hydrogen bonds predicted by GT show some overlap with those predicted by TMD. We also show with various structure quality metrics that the all-atom snapshots in the GT pathways have good structure quality. The results suggest that for NtrC, finding an all-atom pathway is primarily a problem of geometry—finding how the atoms can move from point A to point B while keeping the covalent bonding geometry intact, avoiding overlaps of atoms, etc.—and that a detailed molecular mechanical force field in this case constitutes an extra layer of detail.

Other approximate methods for creating pathways such as Elastic Network Interpolation (50-52, 89) and the Plastic Network Model (90) model the protein as a system of interconnected springs and invoke a small-amplitude approximation to derive normal modes. We have not applied these types of models to NtrC; however we point out that these methods use coarse-graining at the level of Cα atoms, which is lower in spatial resolution than the all-atom GT and TMD pathways considered here. In addition, these methods do not have a mechanism to dynamically avoid overlap of atoms, which is critical in the case of NtrC because the complicated helix reorientation and loop rearrangements involve many atoms bumping and sliding around each other. Furthermore, Lei et al. (24) demonstrated with quasi-harmonic analysis (103) that the transition in NtrC does not overlap well with the low-frequency modes, so it is not expected that an elastic network model would describe the transition well.

Although this is the first detailed comparison of GT pathways with more computationally intensive approaches, the earlier paper on GT (22) did touch on this subject. It was noted that a GT pathway for adenylate kinase showed the two mobile domains moving from the open state to the closed state simultaneously. Arora and Brooks (44) using umbrella sampling (53) in combination with nudged elastic band (41), and

Kubitzki and de Groot (104) using a technique called TEE-REX, have found that during the final part of the closure of the two domains, one domain finishes closing before the other, consistent with a prediction from the Plastic Network Model (90). Because energetics are neglected in GT, the domains in a GT pathway moved simultaneously (22) as there was no geometric reason for a sequential movement. However, the TEE-REX results do show simultaneous domain movement during the portion of the transition between open and partially-closed (104), so the GT results do agree with this portion. The earlier GT paper (22) also reported a pathway for a single subunit of GroEL in which the apical and intermediate domains move simultaneously. This differs from a TMD result (38), in which electrostatic attraction between the intermediate domain and the bound ligand drives the closure of this domain first, followed by the apical domain rotation. These examples indicate that predicting relative timing of events is not a strength of GT; however, ordering of events in TMD should also be viewed skeptically, as Apostolakis et al. (105) have shown that TMD pathways are not reversible. The present results in NtrC indicate that the geometric model in GT can capture some of the same relative timing of events seen in TMD when events are geometrically coupled.

TMD is itself not a perfect standard, since it employs biasing forces to pull the system towards the target state. The biasing force in TMD, which gradually changes over the course of the simulation and pulls strongest on the atoms furthest away, destroys proper thermodynamics. The resultant transition pathways cannot rigorously be interpreted as being representative of the optimal high-flux pathway, and have been shown to cross high free-energy barriers (87). A more appropriate interpretation is to regard TMD pathways as non-optimized stereochemically-plausible pathways that may share some features with the optimal transition pathway, but may get minor or major details wrong due to the pulling. Other computational methods can rigorously generate minimum-energy pathways or high-flux pathways in proteins with atomic-level resolution

(41, 46), but these tend to be even more computationally demanding than TMD and face difficulties of sampling the space of possible pathways in seeking out an optimized pathway. Despite the limitations of TMD, it is valuable for gaining insight into the nature of a transition, identifying residues that may play roles in stabilizing intermediate states, and providing pathways that can be input into free energy calculations. We use TMD for comparison here.

New in this work is a dynamic treatment of hydrogen bonds and hydrophobic contacts in the GT method. In the original GT paper (22), hydrogen bonds were determined based solely on the two input structures (initial structure and target structure). Hydrogen bonds that were found in only one of the two end structures were left unconstrained, while those common to both end structures were given a maximum hydrogen-to-acceptor distance constraint, requiring the hydrogen bonds to remain intact during the entire pathway. Similar logic was applied to pairs of hydrophobic atoms in contact. In the present work, while we do still impose permanent constraints on hydrogen bonds and hydrophobic contacts that are common to both end structures, other hydrogen bonds and hydrophobic contacts form transient constraints that keep the pair of atoms together temporarily. In addition, new minimum-angle constraints are used in addition to the maximum-distance constraints to better model hydrogen bond geometry.

RESULTS AND DISCUSSION

Five random pathways in NtrC were generated with geometric targeting (GT), using an RMSD step size 0.01 Å. As described in the geometric targeting paper (22), each step includes a fairly large random perturbation to the system combined with the step towards the target. Supplementary Movie 1 shows an all-atom animation of GT pathway run 1. Each GT pathway took approximately 7 CPU minutes (~0.1 CPU hours) to complete on a single processor. Each targeted molecular dynamics (TMD) pathway

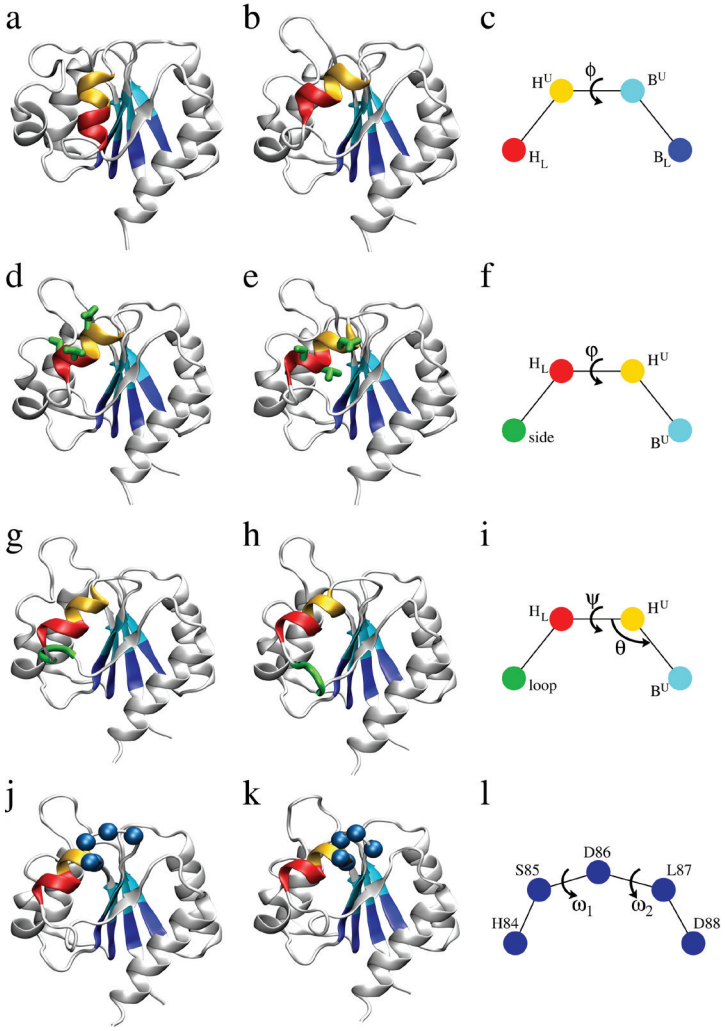

FIGURE 3.1 Previously published TMD results. Lei et al. (24) obtained a segmented transition pathway between the active state of NtrC *(a)* and the inactive state *(k)* using TMD. The pathways has 4 stages, each row of the figure depicting one stage. The beginning and ending conformations of each stage are shown in the left-hand and middle columns respectively. The right-hand column depicts the pseudo angles and pseudo dihedral angles that serve as progress variables for each stage of the transition. Colored circles in the right-hand column denote pseudo atoms, representing the centers of mass of a region of the protein, with the corresponding regions colored in the left-hand and middle panels. For example, in *(c)*, the pseudo-dihedral angle ϕ is defined in terms of the pseudo atoms "HL" (the lower half of the α4 helix), "HU" (the upper half of the α4 helix), "BU" (the upper half of the beta sheet), "BL" (the lower half of the beta sheet), which are shown in panels *(a)* and *(b)* with colors that match the circles in *(c)*. In the segmented TMD pathway, the α4 helix tilts (first row), rotates about its axis (second row), loses a half-helical turn at the C-terminus (third row), gains a half-helical turn at the N-terminus (fourth row). *Adapted from Lei et al. (24), reproduced with permission from Elsevier Ltd.*

reported in Lei et al. (24) took about 200 CPU hours to complete (about 24 hours on 8 processors); a factor 2000 more computational time.

Comparison of pathway motion

We compare the motion in the five random pathways from GT against the 4 previously published TMD pathways (24). Figure 3.1 is adapted from Figure 4 of Lei et al. (24), showing a segmented transition pathway obtained with TMD. The two left columns of Fig. 3.1 show snapshots taken from the TMD at the beginning and ending of each segment of the transition. Four consecutive stages were found (Fig. 3.1, first two columns), described by six progress variables (Fig. 3.1, right-hand column) (24). Each progress variable captures a unique geometric aspect of the motion in the transition. The progress variables are pseudo angles and pseudo dihedral angles, expressed in terms of the positions of pseudo-atoms that are placed at the centers of mass of particular regions of the protein. The pseudo atoms are "HU" (the upper half of the α4 helix), "HL" (the lower half of the α4 helix), "BU" (the upper half of the beta sheet), "BL" (the lower half of the beta sheet), "side" (backbone of D88, V91 and S92) and "loop" (the loop region at the C-terminal end of the α4 helix). The first three progress variables defined by Lei et al. (24) describe the orientation of the α4 helix, relative to the stable beta sheet core: its tilt (dihedral angle BL-BU-HU-HL), its axial rotation (dihedral angle BU-HU-HL-side), and opening angle away from the beta sheet (angle BU-HU-HL). The fourth progress variable describes the loss of one helical turn at the C-terminal end of the α4 helix (dihedral angle BU-HU-HL-loop). The fifth and sixth progress variables (dihedral angle between four consecutive C-alphas His84 to Leu87, and Ser85 to Asp88) describe the addition of one helical turn at the N-terminal end of the α4 helix.

The first observation to make about the GT pathways (Supplementary Movie 1) is that the core of the α4 helix remains folded as it tilts and rotates on its axis, and as helical turns are added/subtracted from its ends, as also occurs in the TMD pathways. This is not

surprising, because we have imposed permanent hydrogen bond constraints for hydrogen bonds that are present in the initial and final structures. There are six consecutive backbone hydrogen bonds within the α4 helix that are found in both end structures, connecting oxygens from residues 85-90 to hydrogens of residues 89-94. The permanent hydrogen bond constraints keep the helix folded, but do not rigidify the helix, because they are inequality distance and angle constraints (see Materials and Methods)

Figure 3.2 shows the transition motion along the six progress variables for the 5 GT pathways (gray), and the 4 TMD pathways (black). In each panel of Fig. 3.2, the pathway starts at the left and finishes at the right, with all progress variables plotted against the RMSD-to-target (decreasing RMSD is used rather than increasing time, because there is no notion of time in the GT pathways).

First progress variable

Fig. 3.2 *a* shows the first progress variable, measuring the ~30° change in the tilt dihedral angle of the α4 helix. In the GT pathways, the helix begins to tilt immediately at RMSD 3.8 Å and reaches the target orientation around RMSD 2.6 Å. In TMD, the helix also begins tilting immediately, but in most of the TMD pathways the tilt reaches the target earlier than in GT, around 2.9 Å. One TMD pathway reaches the target later than GT, at 2.3 Å.

Second progress variable

The second progress variable is depicted in Fig. 3.2 *b*, showing the ~120° helix rotation about its axis. In the GT runs, the rotation does not start immediately, but commences somewhere between 3.2 and 2.7 Å and finishes at 1.6 Å. The axial helix rotation in TMD begins at almost the same point as in GT (between 3.5 and 2.9 Å), but in two of the TMD pathways the rotation finishes at 2.7 Å, much earlier than the GT pathways, while the other two TMD pathways do finish around the same time as the GT pathways.

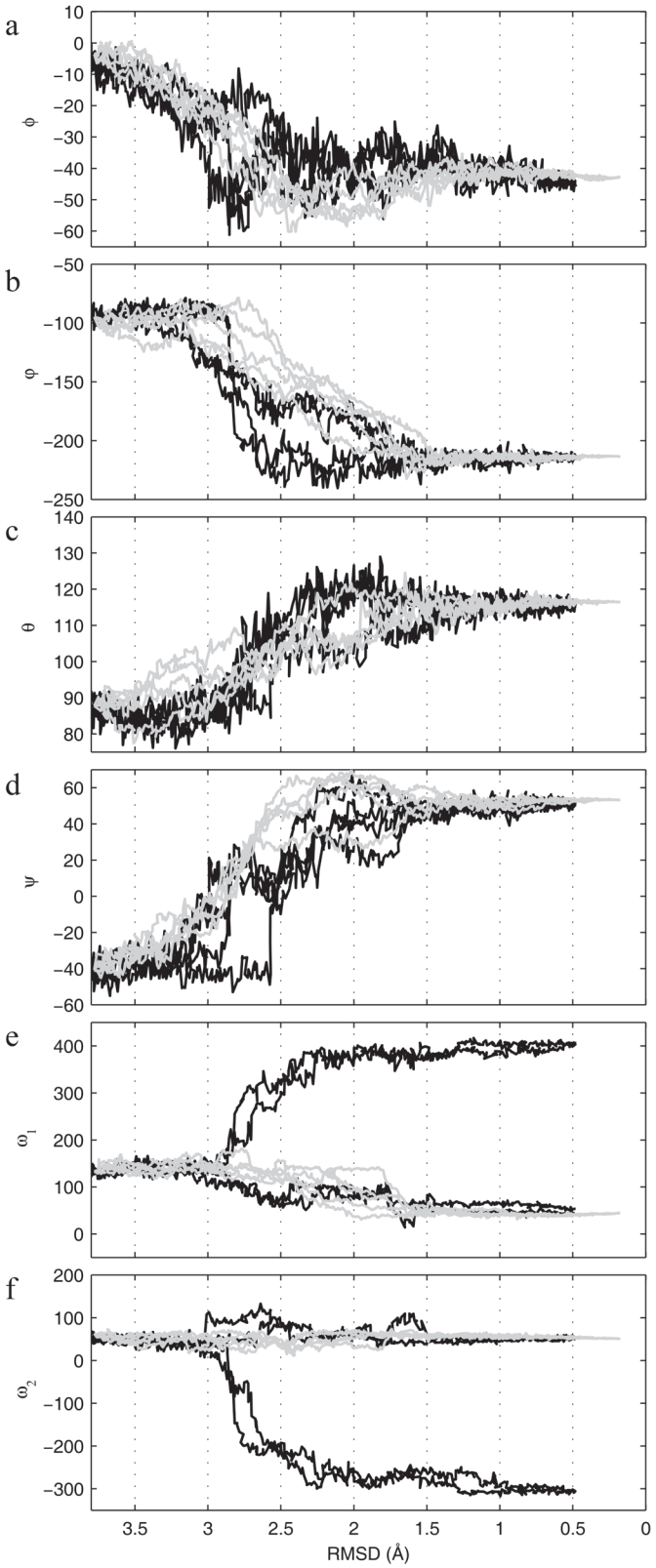

FIGURE 3.2 Comparison of motion between TMD and GT pathways. Each panel *(a-f)* shows motion along one of six progress variables that track various aspects of the transitional motions. Four TMD pathways (black) and five geometric targeting pathways (gray) are shown for each progress variable. Since there is no notion of time in GT, the steadily decreasing RMSD to the target is used as the horizontal axis. The TMD and GT pathways show similar transitions and similar start and end times for each transition. These GT pathways were generated using RMSD step size 0.01 Å.

Third and fourth progress variables

In the third and fourth progress variables (Fig. 3.2 *c*-*d*), the GT and TMD transitions both start around 3.0 Å (some GT pathways start earlier) and end around 1.7 Å. In the third progress variable (Fig. 3.2 *c*), one GT pathway shows similar plateau behavior to TMD around 2.0 Å. In the fourth progress variable (Fig. 3.2 *d*), both TMD and GT show a rapid change in angle taking place between 3.0 Å and 2.5 Å, with similar plateau behavior at 30°.

Fifth and sixth progress variables

In the fifth and sixth progress variables (Fig. 3.2 *e-f*), the TMD pathways bifurcate, showing two different routes to the target angle (despite the bifurcation, the finishing angles are the same, differing by 360°). In two of the TMD pathways, the dihedral angles rotate about a full turn (~300° for the fifth progress variable and ~360° for the sixth), while in the other two TMD pathways the dihedral angles change much less. The GT pathways follow the more direct of the two TMD routes in each case (Fig. 3.2 *e-f*). One possible reason that the GT pathways do not bifurcate here is the permanent hydrogen bonds in the α4 helix. In TMD the hydrogen bonds at the ends of the helix transiently break, even though they are present in the initial and target states, which endows the helix ends with more flexibility than in GT. In TMD the transition begins around 2.8 Å and finally settles around 1.3 Å (Fig. 3.2 *e-f*). In GT, only the fifth progress variable shows any significant change, with beginning points that range from 3.0 to 1.8 Å, and finishing around 1.7 Å

Lei et al. observed that the transition in TMD is segmented in four consecutive stages with some variability in the start and end times (24). In terms of RMSD to target, the start and end points of the four stages, considering all 4 TMD pathways, can be summarized from the results above as follows: first stage (first progress variable) 3.8 to 2.9 Å; second stage (second progress variable) 3.2 to 2.7 Å; third stage (third and fourth

progress variables) 3.0 to 1.7 Å; fourth stage (fifth and sixth progress variables) 2.8 to 1.3 Å. In GT, it does appear that stage 1 precedes stage 2 (start/end times 3.8-2.6 Å for stage 1, 3.0-1.6 Å for stage 2). Stages 2 and 3 in GT coincide (3.0-1.7 Å) and cannot be distinguished as separate stages, but this is also the case with 2 of the 4 TMD pathways in stage 2 (Fig. 3.2 *b*). Stage 4 in GT does take place last, as it does in TMD.

Supplementary Movie 2 shows a TMD pathway and a GT pathway superimposed. The movie shows that in both pathways the helix tilts, rotates, and turns are added/removed at the two ends of the α4 helix. The movie also illustrates the more gradual axial rotation of the α4 helix in GT as compared to TMD, and there are some minor differences between GT and TMD in the loop motions at the ends of the α4 helix.

	Targeted Molecular Dynamics	Geometric Targeting
Steric Clashes		
Number of pairs of atoms in a severe steric clash, per 100 atoms, averaged over all pathway snapshots	2.5	6.4
Bad Rotamers		
Number of residues in rotameric configurations that carry <1% weight, averaged over all pathway snapshots	3.2 / 101	5.9 / 101
Ramachandran		
Number of residues in the outlier region, averaged over all pathway snapshots	3.2 / 121	4.0 / 121
Number of residues in the allowed region, averaged over all pathway snapshots	8.5 / 121	14.5 / 121
Number of residues in the favored region, averaged over all pathway snapshots	109.3 / 121	102.5 / 121
Covalent Bond Geometry		
Fraction of bonds with bad distances, averaged over all pathway snapshots	2.7%	0.1%
Fraction of bonds with bad angles, averaged over all pathway snapshots	7.7%	0.2%

TABLE 3.1 Structure quality metrics comparison. Metrics are compared between the TMD pathway presented in Lei et al. (24) and one GT pathway (RMSD step size 0.01 Å, run 1). The simple geometric model used in GT produces pathway snapshots that are almost as high in quality as the TMD pathway. Structure quality metric were calculated from MolProbity (106). See (106) for more precise definitions of these metrics. Metrics have been averaged over all snapshots in the pathway.

Structure quality comparison

Table 3.1 shows various structure quality metrics calculated for the main TMD pathway presented in Lei et al. (24) and one GT pathway (RMSD step size 0.01 Å, run 1). Metrics were calculated with MolProbity (106). The precise definitions of these metrics can be found in the MolProbity paper (106). Each metric is averaged over all snapshots in the pathway. The number of severe steric clashes and the number of bad rotamers is low for both TMD and GT, although the TMD scores better in both of these metrics. The number of residues in outlier regions of the Ramachandran plot are about the same in both TMD and GT, but TMD has a higher fraction of residues in the favored regions than does GT. Interestingly, GT scored better than TMD in covalent bond geometry. This is probably because the GT covalent bond geometry is locked in rigid units, but the TMD covalent bond geometry may bend or stretch due to the biasing force or due to thermal fluctuations.

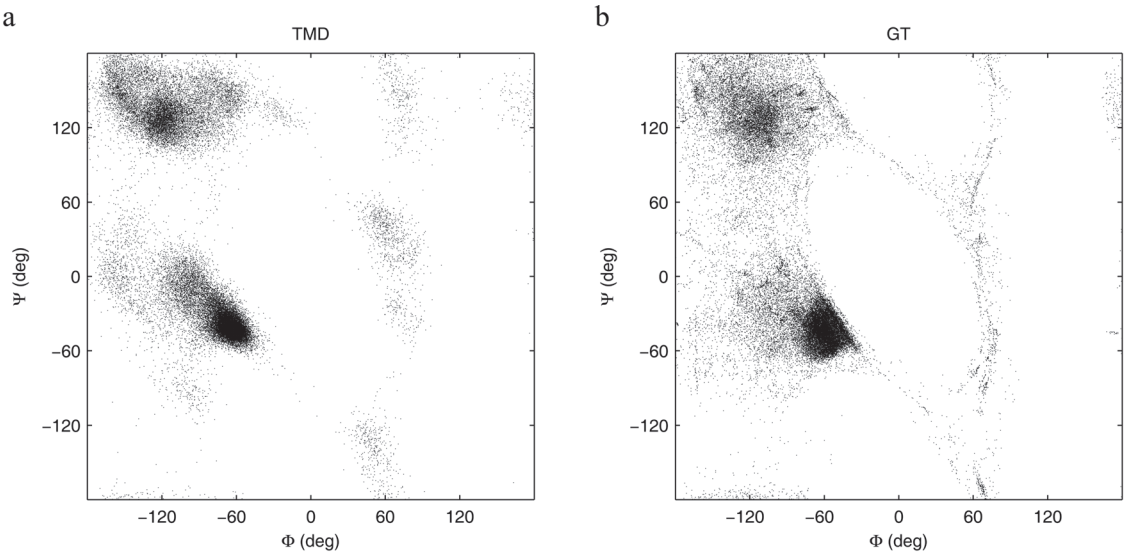

FIGURE 3.3 Ramachandran plots. Backbone dihedral angles are compared between one TMD pathway and one geometric targeting pathway. The simple geometric model in GT captures the essence of the Ramachandran map, with outlier regions successfully produced from the geometric constraints. *(a)* Phi/psi pairs from 351 TMD snapshots are plotted from the TMD pathway presented in Lei et al. *(b)* Phi/psi pairs from 364 GT snapshots are plotted, from 1 GT pathway (RMSD step size 0.01 Å).

Fig. 3.3 compares the Ramachandran plot (28) for one TMD pathway [the main pathway presented in Lei et al. (24)] and one GT pathway (RMSD step size 0.01 Å, run 1). All phi-psi pairs from ~350 evenly-spaced snapshots in each pathway are plotted. Fig. 3.3 *B* shows that the simple geometric model in GT captures the essence of the Ramachandran map, with outlier regions successfully produced from the geometric constraints. The points are more diffuse in GT than TMD in the favored and allowed regions, because in GT there is no molecular mechanical energy function to provide attractive forces.

Transient hydrogen bonds

In Table 3.2, we compare transient hydrogen bonds identified in the 5 GT simulations with those previously reported in Table 3 of Lei et al. (24) from the 4 TMD simulations. Lei et al. (24) reported transient hydrogen bonds that met the following criteria: (a) donor-hydrogen-acceptor angle is greater than 100°; (b) hydrogen-acceptor distance is less than 2.5 Å, (c) the hydrogen bond is maintained for 50 ps; (d) the hydrogen bond is not transiently stable in equilibrium MD simulations of the end states; (e) at least one of the atoms is a side-chain atom; (f) at least one of the atoms in the hydrogen bond is in residues 82-104; In addition, transient bonds were filtered from the list if (g) the hydrogen bonded atoms were from the same residue, or (h) involved a backbone atom hydrogen-bonded to a side chain atom ±1 residues away. Twenty hydrogen bonds were identified according to these criteria (some of which were water-mediated) and reported previously (24), repeated here in Table 3.2 for comparison. Six of these were present in at least two of the four TMD pathways. These 6 appear in the top rows of Table 3.2 as “top-ranked” transient hydrogen bonds, while those appearing in only 1 TMD pathway are listed as “lower-ranked” in Table 3.2.

For the GT pathways, the criteria for reporting a transient hydrogen bond in Table 3.2 is slightly different from the TMD criteria. Instead of the distance and angle criteria

	TMD		GT	
	Donor	**Acceptor**	**Donor**	**Acceptor**
Top-Ranked Transient Hydrogen Bonds	S85:Oγ	D86:Oδ	H84:Nδ	D86:Oδ
	Q96:Nε	A90:O	S85:Oγ	D86:Oδ
	Q96:Nε	V91:O	Q96:Nε	S92:Oγ
	Q96:Nε	A93:O	K104:Nζ	H84:Nε
	A98:N	Y101:Oη		
	Y101:Oη	Q96:Oε		
Lower-Ranked Transient Hydrogen Bonds	T82:Oγ	S85:Oγ	L66:N	Q96:Oε
	H84:N	D86:Oδ	K70:Nζ	Q96:O
	H84:N	Y101:Oη	K70:Nζ	Q96:Oε
	H84:Nδ	D86:Oδ	H84:N	T82:Oγ
	H84:Nδ	D88:Oδ	Q95:Nε	V91:O
	S85:N	D88:Oδ	Q95:Nε	Q96:Oε
	S85:Oγ	T82:Oγ	Q96:Nε	S92:O
	Q96:N	Y101:Oη	Q96:Nε	Y94:O
	Y101:Oη	S85:Oγ	Q96:Nε	Q95:Oε
	Y101:Oη	A98:O	A98:N	Q96:Oε
	K104:N	H84:Nε	F99:N	Q96:Oε
	K104:N	Y101:Oη	Y101:Oη	H84:O
	K104:Nζ	H84:Nε	Y101:Oη	Q96:Oε
	K104:Nζ	D88:Oδ	K104:N	D86:Oδ
			K104:Nζ	D10:Oδ
			K104:Nζ	D86:Oδ

TABLE 3.2 Transient hydrogen bonds. Transient hydrogen bonds are compared between the 4 TMD pathways presented in Lei et al. (24) and the 5 GT pathways of this work (RMSD step size 0.01 Å). The top-ranked hydrogen bonds (top rows) for TMD are those appearing in at least 2 of the 4 pathways, and the lower-ranked hydrogen bonds only appear in 1 of the 4 pathways. For GT, the top-ranked bonds are those appearing in at least 3 of the 5 pathways, while the lower-ranked hydrogen bonds appear in 1 or 2 of the 5 pathways. Four hydrogen bonds are found in both TMD and GT (shaded entries). Three of the 4 top-ranked GT bonds (shaded) also appear in TMD. Two of the 6 top-ranked TMD bonds (shaded) also appear in GT. One bond is found in both the TMD and GT top-ranked sets (S85:OG…D86:Oδ).

1 and 2 above, we report in Table 3.2 the hydrogen bonds that were added as transient geometric constraints during the geometric targeting (see Materials and Methods). We only report hydrogen bond constraints that were maintained for at least 5 steps, lasting about the same fraction of the pathway as the 50 ps for TMD pathways (criteria 3 above).

We also remove initial and target state hydrogen bonds from the list analogous to criteria 4 above, by not listing bonds if hydrogen and acceptor distance was within 5 Å in the initial or target structures. The other criteria 5 through 8 were applied to GT without modification. Twenty-one transient hydrogen bonds met these criteria and are listed in Table 3.2. As the top-ranked transient hydrogen bonds from GT, we have chosen those that were found in at least 3 of the 5 GT pathways (there are 4 of these, top rows of Table 3.2). The lower-ranked hydrogen bonds (bottom rows of Table 3.2) appear in 1 or 2 of the 5 pathways.

Bonds that are found in both TMD and GT are shown as shaded entries in Table 3.2. The lower-ranked hydrogen bonds (bottom rows, Table 3.2) show little overlap between the two methods. Three of the 4 top-ranked GT bonds (shaded, GT column) appear in TMD, while two of the 6 top-ranked TMD bonds (shaded, TMD column) appear in GT. One bond is found in both the TMD and GT top-ranked sets (S85:OG… D86:Oδ). Another two of the 6 top-ranked TMD bonds involve side chains that form transient hydrogen bonds with backbone atoms of the α4 helix, requiring the helix to transiently interrupting the helix hydrogen bonds. In GT, the helix hydrogen bonds in the α4 helix are permanent, so it is not likely for side chain atoms to form stable bonds with the helix backbone.

Sensitivity to pulling rate

To test how the GT pathways are affected by the pulling rate, we ran other GT simulations using different RMSD step sizes. The five GT pathways that we have described in this paper were generated with RMSD step size 0.01 Å. For larger RMSD step sizes (0.02 Å, 0.05 Å, 0.10 Å), we found that the pathways as measured by the six progress variables were very similar to the pathways generated with 0.01 Å step size (results not shown), although the progress variable curves have less fluctuations in these

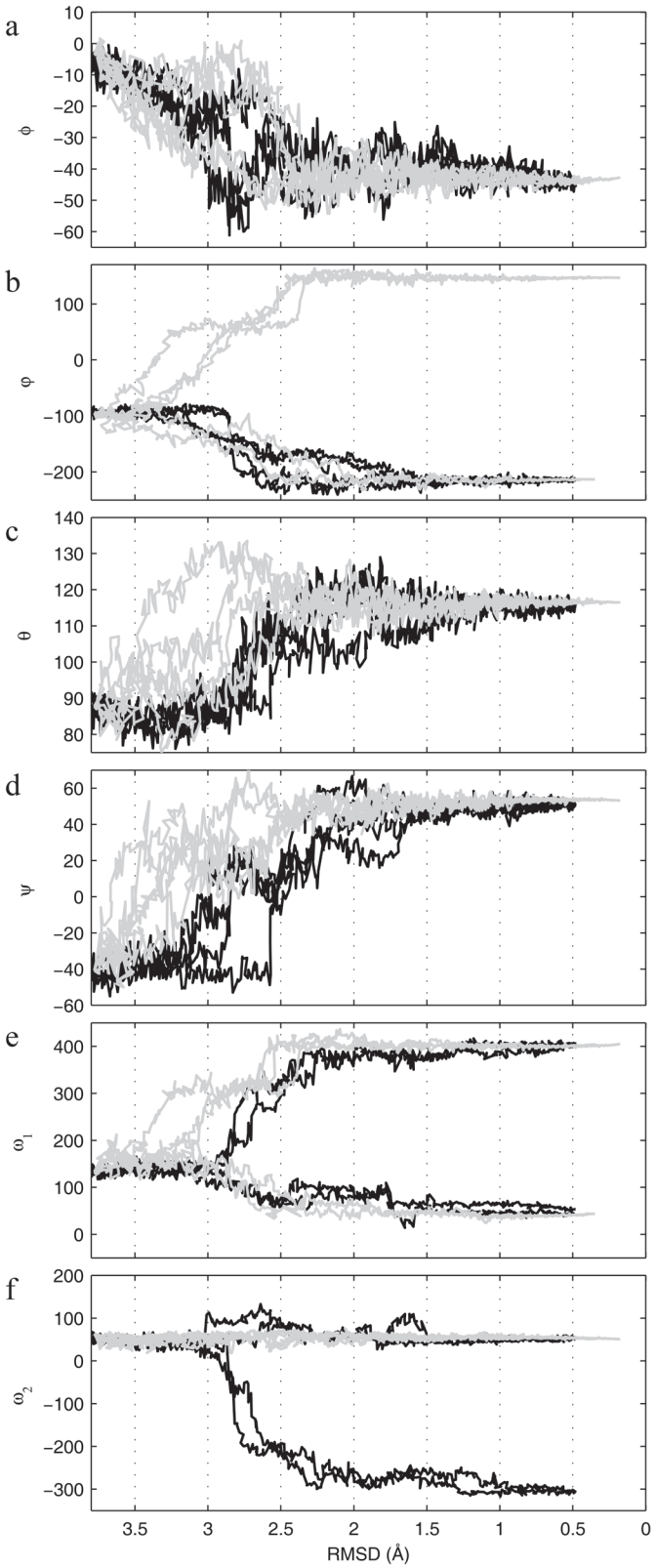

FIGURE 3.4 Slower pulling rate in GT leads to novel pathways. GT pathways were generated with RMSD step size 0.001 Å, a factor 10 smaller than was used for the pathways of Fig. 2. Each panel *a-f* shows motion along one of six progress variables that track various aspects of the transitional motions. Four TMD pathways (black) and five geometric targeting pathways (gray) are shown for each progress variable. The RMSD to the target is used as the horizontal axis. Because of the smaller step size, the random motion in GT has more opportunity to sample pathways away from the more direct route. Notably, in several pathways in *b*, the α4 helix rotates about its axis in the reverse direction compared to TMD (the tracks in *b* bifurcate in two different directions, but end up at the same angle, separated by 360°). This reverse rotation of the helix is a novel, stereochemically acceptable pathway route that is not observed in TMD.

cases because the random motion has less opportunity to impact the trajectory at high pulling rates.

At smaller RMSD step sizes, however, the GT pathways begin to explore other ways of reaching the target. With a smaller RMSD step size, it takes more steps to reach the target, and the bias in each step is smaller, giving the random motion more opportunity to take the pathway away from the most direct route. Five GT pathways were generated with RMSD step size 0.001 Å, a factor 10 smaller than before, and requiring a factor 10 more computational time. Figure 3.4 shows these five GT pathways (gray) compared with the TMD pathways (black). In all six progress variables, the GT pathways with 0.001 Å step size (Fig. 3.4 *A*-*F*) fluctuate more and deviate more from TMD than the GT pathways with 0.01 Å (Fig. 3.2). This means that GT is exploring other geometrically plausible pathways that were not accessible with the 0.01 Å step size. One notable difference can be seen in Fig. 3.4 *b*, which shows several GT pathways in which the α4 helix rotates about its axis in the reverse direction compared to TMD (the tracks in Fig. 3.4 *b* bifurcate in two different directions, but end up at the same angle, separated by 360°). This "reverse" rotation of the helix is about 240°, rather than 120° in the more direct route. It is evident from watching an animation (Supplementary Movie 3) that the reverse rotation is the natural unwinding direction at the C-terminal end and the natural winding direction at the N-terminal end because of the right-handed helicity of the helix, which possibly facilitates the addition/removal of a half helical turn at the two ends. On the other hand, in the reverse rotation, charged residues N88 and S92 must pass by a hydrophobic surface, which is likely to be unfavorable. GT gives no indication as to which alternative is preferred energetically, but what it does reveal is that the 240° reverse rotation is geometrically plausible in addition to the 120° rotation. Other techniques such as umbrella sampling (53) would be necessary to calculate free energy barriers to quantitatively determine which direction of helix rotation is more favorable.

The structure quality and Ramachandran maps of the 0.001 Å step-size pathways are not shown, because they are very similar to the results for the 0.01 Å step-size pathways.

CONCLUSION

We have shown that the motion in the geometric targeting (GT) pathways of nitrogen regulatory protein C (NtrC) is very similar to the motion in targeted molecular dynamics pathways (TMD), as measured by six progress variables. In a factor ~10^3 less computational time compared to TMD, we have produced stereochemically-acceptable all-atom pathways that capture the essence of the transition in NtrC. A difference is that the GT pathways lack some of the sudden movements observed in the TMD pathways, and the GT pathways are less-clearly segmented than in TMD. Overall, however, the start and end times for the various motions are similar between the two methods. All-atom structure quality is good in the GT pathways, especially considering the approximate nature of the model, although it is not as high as TMD. Some of the same top-ranked transient hydrogen bonds were identified in both methods.

Lei et al. (24) discuss in their paper that in the TMD pathways, NtrC does not require local unfolding or "cracking motions" as has been suggested by a coarse-grained study (107). The α4 helix tilts and rotates without unfolding in the TMD pathways, as also in the GT pathways reported here. It should be pointed out that in GT, the stability of the helix was assumed from the beginning, because of the helix hydrogen bonds common to the initial and final states that were treated as permanent geometric constraints. GT shows it is geometrically plausible for the transition to take place without local unfolding, although GT by itself is neutral on the question of whether local unfolding is energetically preferable.

The pathway similarity between GT and TMD was observed for RMSD step sizes of 0.01 Å or greater. We have also shown that when the RMSD step size is reduce by a factor 10, GT has freedom to sample pathways that deviate more from the TMD results,

producing in this case a novel stereochemically-acceptable reverse rotation of the α4 helix. In terms of free energy, GT gives no indication as to whether this pathway is more or less favorable than the more direct rotation observed in the TMD and in the other GT pathways. We do not know whether TMD would also produce the reverse rotation of the helix with smaller step sizes, but it would require a factor 10 more computational resources to test this.

The similarity of the pathways between GT (at RMSD step sizes of 0.01 Å or greater) and TMD may partially be due to the fact that GT and TMD both apply the same type of bias to the same two conformational end states. The similarity of the pathways may also arise from the highly-constrained nature of proteins in general. Geometric considerations such as covalent bonds, non-overlapping atoms, maintenance of hydrogen bonds, etc., do in reality severely limit accessible conformational space and restrict the possible ways that a protein can move from state A to state B. The neglect of energetics in GT appears not to have a significant effect on the overall pathway motions in this case. Further studies of other proteins would be necessary to test how generalizable the level of similarity is between GT and TMD. In situations where a rapid and approximate pathway is desired, the present results justify the use of GT as a first-look into possible all-atom transition pathways.

MATERIALS AND METHODS

Geometric targeting

The geometric targeting method (GT) is described in detail in this reference (22). Here we describe two additions to the previous work (22): (a) how hydrogen bonds are modeled with new angular inequality constraints, in combination with inequality distance constraints described previously (22), and (b) new dynamic constraints for transient hydrogen bonds and hydrophobic contacts.

Hydrogen bonds that are found common to both the initial and target structures are given permanent inequality constraints that preserve the hydrogen bonds throughout the pathway. Identification of hydrogen bonds uses a maximum distance constraint between the hydrogen and acceptor atoms, requiring the hydrogen-acceptor pair to stay together during the entire pathway. Constraints are only added for hydrogen bonds that score better than a cutoff energy value (-1.0 kcal/mol) according to a modified Mayo energy function (56, 99). The maximum distance is set as the greater of the two hydrogen-acceptor distances from the two structures, and not less than 2.0 Å. In addition, the donor-hydrogen-acceptor angle is constrained (new in this work) to be greater than 100°. In order to preserve good quality secondary structure, backbone-backbone hydrogen bonds with NH…O angle greater than 140° are constrained to keep the angle above 140°, otherwise the angle is constrained to be greater than 100°. In addition, backbone-backbone hydrogen bonds with H…OC angles greater than 130° are constrained to keep the angle above 130°. Pairs of hydrophobic atoms closer than 3.9 Å are given a maximum distance constraint to keep the atoms together, with constraint distance equal to the greater of the two distances from the two structures, plus an extra 0.5 Å. Hydrogen bonds and hydrophobic contacts that are not found in both structures, or that come into contact temporarily during the pathway, are treated dynamically as described below.

As described in the geometric targeting paper (22), constraints are enforced by minimizing a pseudo energy function that measures the amount of constraint violation. Minimization brings the system to the flat region of the landscape at energy zero where the constraints are met. The energy term for the minimum-angle constraints $\theta_i \geq \theta_i^{min}$ is

$$V_{\text{min angle}} = \sum_{\text{i}} \begin{cases} \frac{1}{2} k \left(\theta_i - \theta_i^{min} \right)^2, & \theta_i < \theta_i^{min} \\ 0, & \theta_i \geq \theta_i^{min} \end{cases} \tag{3.1}$$

which is flat in the region where the constraint is satisfied. Angles are measured in radians. All constraints use the same spring constant k, although in principle these are adjustable.

During the course of the targeting, hydrogens and acceptors may transiently move into proximity, and pairs of hydrophobic carbons may move into proximity. At the beginning of each step before any movements of the system are performed, a search is made for new hydrogen bonds and new hydrophobic contacts, and constraint distances and angles are established. The constraint distances are updated to tighter values in subsequent steps if the pair of atoms move closer together. These constraints are transient and are only maintained temporarily (permanent hydrogen bond and hydrophobic constraints are established for those that exist in both the initial and target structures, and these are never removed). Also at each step, after adding new constraints, but before any movements of the system are performed, any transient constraints that currently are in force are considered for removal as follows. Transient hydrogen bonds are removed with probability 0.1, and transient hydrophobic contact constraints are removed with probability 0.2. Transient constraints are also removed if the previous step's minimization procedure had difficulty enforcing the constraint, for example when a hydrogen bond must break in order to move the system closer to the target. Specifically, if a transient hydrogen bond distance constraint is violated by more than 0.02 Å, it is removed, and if a transient hydrophobic contact constraint is violated by more than 0.01 Å it is removed.

Note that the Geometric Pathways Webserver (pathways.asu.edu) and the standalone version of the software available through the website have been updated with the changes noted here.

Targeted molecular dynamics

We used four TMD pathways previously reported in Lei et al. (24), The details of the preparation of the initial and target structures and the TMD simulations are described in their paper, but here we summarize the protocol used. NMR structures for the active state (PDB ID 1DC8) and the inactive state (PDB ID 1DC7) are refined in molecular dynamics simulations that include NOE (Nuclear Overhauser Effect) distance restraints and subsequently minimized. These refined active and inactive structures are used as the initial and target structures of the TMD simulation. The active state structure of NtrC is first submerged in a water sphere. The protein and solvent system is then simulated at 300K with stochastic boundary condition in CHARMM. During the 3.5 ns simulation, the RMSD to the inactive state structure is forcefully and linearly reduced from 3.8 Å to 0.3 Å by the TMD algorithm. Hydrogen atoms are constrained to heavy atoms by the SHAKE algorithm. The time step is 2 fs.

CHAPTER 4 APPLICATIONS

In this chapter we demonstrate ways that pathways from geometric targeting (GT) and the FRODAN software have been used to make contact with experiment. The main application we present is the use of GT pathways as input into umbrella sampling free energy calculations, used here to calculate the relative free energy between two conformations in the enzyme dihydrofolate reductase. The work described is a collaborative effort involving Daniel W. Farrell, Tatyana Mamonova, Maria Kurnikova and M. F. Thorpe (unpublished). The roles of the contributors are clarified in the text. The text was written by DWF, except some portions of the methodology written by TM and MK where noted.

There are two other applications that use FRODAN software in a non-targeted fashion (i.e. geometric simulation rather than geometric targeting), which we mention here but do not describe in detail. One is the flexible fitting of high resolution protein structural data from one conformation into low resolution density maps from cryo-electron microscopy that show the protein or protein complex in a different conformation. This application was first implemented by Craig C. Jolley in the original FRODA method, described in (74). CCJ has implemented a much faster cryo-EM fitting algorithm in the new FRODAN software (unpublished), in consultation with DWF. Fitting atomistic structures into low-resolution cryo-EM data is very useful because in the fitted structure one can identify residues that have switched from one set of neighbors to another in the conformational change, which could be key residues that stabilize the transition. The other application we mention here briefly is the use of geometric simulation in the zipping and assembly method of protein folding by (75), where geometric simulation is used during a portion of the method to combine two folded fragments into a single larger folded fragment. Geometric simulation speeds up this step significantly. This application of geometric simulation has been developed Tyler J. Glembo and S. Banu Ozkan, using

functionality developed by Stephen A. Wells in the original FRODA software combined with portions of the new FRODAN software. Specifically, the work of Glembo and Ozkan uses the new FRODAN "engine" for the enforcement of constraints and the new FRODAN "momentum run-on" perturbation for enhanced exploration efficiency. Integration of FRODAN software components into the original FRODA software to permit their combined usage was implemented by DWF.

GEOMETRIC TARGETING AS INPUT TO UMBRELLA SAMPLING

This section describes the use of geometric targeting pathways as input to umbrella sampling free energy calculations in the protein dihydrofolate reductase. The work is collaborative involving Daniel W. Farrell, Tatyana Mamonova, Maria Kurnikova and M. F. Thorpe (unpublished). There are two main aspects to this study: (a) calibrating the constraints of GT to produce structures that are compatible with molecular dynamics (calibration designed and carried out by DWF), and (b) the umbrella sampling (performed by TM and MK). The text in this section is written by DWF, with some methodology written by TM and MK where noted. Research in this project is ongoing, but the work so far demonstrates a proof-of-concept that snapshots from GT can be used as initial snapshots for umbrella sampling, for the purpose of calculating the relative free energy between two conformations.

Recent advances in NMR relaxation experiments (see Chapter 1) have revealed in at least two cases that an enzyme in a particular catalytic state spontaneously samples conformations that resemble the conformations of its other catalytic states. The picture that has emerged is that a protein's conformation under given conditions is characterized by a native ensemble, exchanging between a ground-state conformation (the native state) and other higher-energy less-populated conformations. Furthermore, conformational exchange has been shown to be a rate-limiting step in catalysis. These new insights from NMR provide new opportunities for computational methods to make testable predictions

about conformational exchange and its relationship to catalytic function. In particular, umbrella sampling (53) can be used to calculate free energy differences between conformations to assess their relative stability. In umbrella sampling, independent molecular dynamics simulations are run, each in a different "window" along some reaction coordinate (Chapter 1). These windowed molecular dynamics simulations each need an initial conformations to get started. In this section, we test whether snapshots from GT pathways can serve as the initial conformations in the umbrella sampling MD.

Our test system is the enzyme dihydrofolate reductase (DHFR), shown in Fig. 4.1. DHFR produces THF (tetrahydrofolate), a key compound for cell growth (108), from DHF (dihydrofolate), with the help of cofactor NADPH. With bound substrate DHF, the enzyme adopts a "closed" conformation, with the Met-20 loop closing over the binding site of DHF and the nicotinamide ring of NADPH (109) facilitating catalysis. In the product states, the Met-20 loop is found in an "occluded" conformation, blocking the nicotinamide ring from the binding site. Crystallographically, structures are known for the closed (PDB ID 1RX2) and occluded (PDB ID 1RX6) states. NMR relaxation experiments have uncovered much about the dynamics of this loop motion (17, 109, 110), finding that the loop exchanges between both occluded and closed conformations even

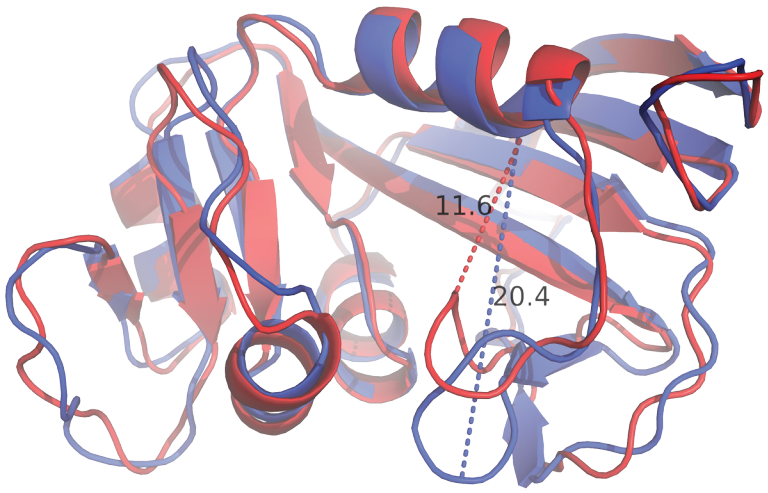

FIGURE 4.1 Dihydrofolate reductase. The closed (blue) and occluded (red) states are superimposed. Structures shown are the two MD-equilibrated end conformations. The Glu17 Cα to Asp27 Cα pair distance, which is the chosen reaction coordinate in the umbrella sampling, is shown for both structures in Å.

though only one of these states is the stable ground state of the enzyme at a given point in the catalytic cycle (17). Here use umbrella sampling to measure the free energy difference between closed and occluded states in the apo-enzyme, comparing this with the experimental free energy measurement between closed (ground state) and occluded state (excited state) in a complex that models the Michaelis complex DHFR:NADPH:DHF (17). The free energy difference obtained by McElheny et al. is 2.1 kcal/mol (17). We have not included the cofactor or substrate in this calculation, but it is suspected that the conformational exchange is a property of the enzyme itself, and we investigate this by testing whether the calculated free energy difference in the apo-state is similar to the experimentally measured free energy difference in the model Michaelis complex.

Several other approaches have been used to generate a set of input pathway frames for umbrella sampling. One method is to choose a reaction coordinate in terms of some angle or pair distance in the protein, and then to step the system through intermediate values of the reaction coordinate using restraining forces in MD until the end value is reached (111, 112). Equilibration between steps allows the other degrees of freedom to adjust to each move of the chosen reaction coordinate. This approach only works when a suitable reaction coordinate can be pre-determined from the two conformational end states, such that forcing the system along the coordinate produces the desired conformational change, and yields intermediate states are not too high in energy. Another approach is to use a spontaneous pathway from molecular dynamics, if such a pathway can be observed in reasonable time (111, 113). Targeted molecular dynamics simulations (34, 35) have also been used to generate input pathway frames, using a constraint on the RMSD relative to the target structure, or a constraint on the difference in RMSD relative to both the initial and final structures (114-116), to bias the dynamics from one conformation to the other. Notably, minimum energy pathways from nudged

elastic band (41, 42) have also been used to provide input pathway frames to umbrella sampling (44, 45).

Note that most of the above methods for generating an initial pathway do not involve finding a minimum energy pathway or a high-flux pathway at finite temperature. To take a conservative view, a free energy profile obtained from umbrella sampling that begins with non-optimal pathway snapshots cannot be trusted to yield the "true" free energy barriers, since the pathway likely takes the system over barriers that are higher than in the optimal case. If we consider the calculated free energy difference between end states, rather than the free energy profile, this quantity should in principle be path independent. As a practical issue, the precision in the estimate of end-state free energy difference is highly sensitive to the rises and falls in free energy along the chosen pathway, so a suitable input pathway does not take the system over high free energy barriers or low dips. Furthermore, if the pathway has steep free energy gradients, very tight restraints are needed in order to keep the system in its intended region of the reaction coordinate. Tight restraints result in very narrow sampling regions, which means that many closely-spaced umbrella simulations must be performed to adequately cover the steep region. A successful calculation of relative free energy depends therefore on the ability to create a pathway that is reasonably low in free energy connecting the two conformational states of the protein, though not necessarily an optimal pathway.

GT and MD employ very different models and different parameters, so it is not a trivial matter to simply transfer a structure from one method into the other. GT does not use a molecular mechanical force field like MD, instead modeling the protein as a geometrically constrained system. Each method has its own parameters, and a "good" structure in one model is not necessarily a "good" structure in the other. In a prior attempt to transfer geometrically-generated snapshots into umbrella sampling a few years ago (unpublished) using the FRODA method, the restrained MD simulations seeded with

FRODA snapshots would pull strongly to one side of the window and not sample the intended region. It was suspected that non-bonded atoms in FRODA were coming too close, resulting in high van der Waals energies (unpublished).

To diminish the structural differences between GT snapshots and MD, we tailor the non-overlap constraints (minimum distance constraints) of GT so that atoms do not come closer than they typically would in an MD simulation. At first thought, it might seem that appropriate cutoff distances could be obtained by summing the van der Waals radii that are built into the molecular mechanical force field of MD. However, the separations between non-bonded atoms in MD are affected by more than just the van der Waals energy term. For example, in OH…O hydrogen bonding, the bound hydrogen allows the two oxygens to come closer than two oxygens without a bound hydrogen (O…O), even though the van der Waals radius of the oxygen is the same in both cases. Arguments against using molecular mechanical van der Waals parameters alone to characterize the effective repulsive interactions between atoms in MD are more thoroughly discussed in the work of Hornak and Simmerling (117).

Here, our approach is to calibrate the minimum distance constraints of the GT model against MD trajectories. We do not simply use the minimum distances observed in MD, because the minimum distance for a pair is a rare event, highly unfavorable energetically. If such distances were used as minimum distance constraints in GT, it would be perfectly acceptable for pairs of atoms to be found in these highly unfavorable close contacts (in GT there is no repulsion unless pairs of atoms are closer than the minimum-allowed distance). Instead of basing the constraints on the absolute minimum distances observed in MD, we base the constraints on the minimum *stable* contact distances observed in MD. We define a “stable contact distance” in MD as a pair distance averaged over some length of time, here chosen as 200 ps. The idea here is that if a pair of atoms in MD maintains some average separation distance over 200 ps, that distance is

a stable contact distance, and the minimum distance constraint in GT should be no larger than this value. We do not want the GT constraints to be so large that they prevent pairs of atoms from coming as close as they do in MD in an averaged sense.

Note that we do not choose constraint distances from pair distance histograms or a potential of mean force between pairs of atoms. As we will illustrate in Results and Discussion, pair distance histograms in folded proteins can give a false impression of what distances are stable and what distances are repulsive. Problems associated with using pair distance histograms for determining contact distances in folded proteins have been noted earlier by Li and Nussinov (118) in their determination of effective atomic radii from protein crystal structures.

Type Name	Type Description
CTa	Amber CT type (*sp3* carbon) with only carbon and hydrogen neighbors
CTb	Amber CT type (*sp3* carbon) with at least one neighbor that is not carbon or hydrogen
C	Amber C type (carbonyl *sp2* carbon)
CA	Amber CA type (aromatic *sp2* carbon in 6-membered rings and CE of Arg)
C_	All other carbon types
N	Amber N type (*sp2* nitrogen in amides)
N3	Amber N3 type (*sp3* nitrogen)
N_	All other nitrogen types
O	Amber O type (*sp2* oxygen in amides)
O2	Amber O2 type (*sp2* oxygen in anionic acids, COO-)
OH	Amber OH type (*sp3* oxygen with bonded hydrogen)
O_	All other Amber oxygen types
S	All sulfur types
HN	Amber H type (hydrogen attached to nitrogen)
HS	Amber HS type (hydrogen attached to sulfur)
HO	Amber HO and HW type (hydrogen attached to oxygen/water)
HA	Amber HA type (hydrogen attached to aromatic carbon)
HC	Amber HC type (hydrogen attached to aliphatic carbon with no electron-withdrawing substituents)
H_	All other hydrogen types

TABLE 4.1 Atom types used for defining minimum distance constraints. These atom types are based on the Amber atom types (25).

Methodology

Calibration of minimum distance constraints

As a reference for calibration, we used trajectories from three equilibrium molecular dynamics simulations (MD trajectories supplied by Tatyana Mamonova and Maria Kurnikova, see below). To establish minimum distance constraints for geometric targeting, a set of atom types was defined (Table 4.1) based on the Amber atom types (25), including hydrogens. Rather than assign additive hard-sphere radii to the atom types, we allowed for each unique pair of atom types to have a unique constraint distance.

The following steps were performed for each unique pair of atom types. The pair distances of all non-bonded pairs of atoms that ever came closer than 4.5 Å during any of the three MD simulations were measured. Covalent first, second, and third neighbor

	CTa	CTb	C	CA	C_	N	N3	N_	O	O2	OH	O_	S	HN	HS	HO	HA	HC	H_
CTa	3.45																		
CTb	3.40	3.40																	
C	3.15	3.25	3.10																
CA	3.40	3.35	3.10	3.40															
C_	3.40	3.50	3.25	3.30	3.70														
N	3.10	3.05	3.00	3.05	3.15	3.10													
N3	3.65	3.80	3.25	3.55	3.80	3.80	-												
N_	3.40	3.20	3.15	3.25	3.30	3.05	3.65	3.35											
O	3.05	3.00	2.80	3.15	3.15	2.75	2.75	2.75	2.95										
O2	3.35	3.15	2.95	3.20	3.45	2.80	2.70	2.70	3.05	4.15									
OH	3.25	3.10	3.15	3.30	3.55	2.85	2.90	2.90	2.65	2.55	2.95								
O_	-	-	-	-	-	-	-	-	-	-	-	-							
S	3.35	3.55	3.30	3.65	3.75	3.20	3.70	3.85	3.20	3.60	3.40	-	-						
HN	2.45	2.60	2.25	2.40	2.40	2.20	3.30	2.10	1.80	1.75	1.95	-	2.90	1.85					
HS	3.35	3.30	3.00	2.95	4.55	2.60	-	4.25	1.95	4.30	4.10	-	-	2.25	-				
HO	2.80	2.90	2.35	3.05	3.65	2.35	3.10	3.05	1.70	1.65	2.10	-	2.55	1.95	3.55	2.45			
HA	2.85	2.70	2.75	2.65	2.80	2.60	3.30	2.75	2.45	2.60	2.60	-	2.80	2.10	3.20	2.20	2.35		
HC	2.70	2.75	2.60	2.65	2.70	2.50	2.65	2.65	2.45	2.50	2.55	-	2.80	1.85	2.85	2.25	2.25	2.20	
H_	2.65	2.45	2.55	2.60	2.70	2.50	3.05	2.55	2.25	2.50	2.50	-	2.70	2.00	2.60	2.20	2.10	2.10	2.05

TABLE 4.2 Minimum smoothed pair distances observed in reference MD trajectories. From three reference MD trajectories, pair distance trajectories for individual atom pairs were smoothed with a 200 ps windowed average. Entries here indicate the minimum observed smoothed pair distance for each possible pair of atom types. Because of symmetry, only the bottom half of the table is shown. Gray entries denote those that had low numbers of samples (see Table 4.3), and a dash (-) denotes no data. Distances are in Å.

pairs were discarded, because our focus here is on non-bonded pair distances. The pair distance vs. time of each atom pair was smoothed using a 200-ps windowed average. The individual atom pairs were grouped according to the pair of atom types. A preliminary constraint distance for the pair of atom types was assigned as the minimum smoothed pair distance of any individual pair (Table 4.2). Certain pairs of atom types had very little data (Table 4.3), so for these atom type pairs, the constraint distance was lowered (subjectively) to that of similar types (Table 4.4).

Protocol for MD simulations used in calibration of GT constraints

Tatyana Mamonova and Maria Kurnikova provided equilibrium molecular dynamics simulations, used by Daniel W. Farrell for calibrating the pair-specific non-overlap cutoff distances of geometric targeting. They provided the following text

	CTa	CTb	C	CA	C_	N	N3	N_	O	O2	OH	O_	S	HN	HS	HO	HA	HC	H_
CTa	431																		
CTb	312	77																	
C	296	174	38																
CA	427	223	144	77															
C_	83	61	47	32	5														
N	261	169	547	127	48	115													
N3	11	3	19	6	0	1	0												
N_	103	39	61	31	10	24	1	10											
O	900	612	733	209	62	690	11	34	351										
O2	81	78	72	26	5	50	15	33	59	1									
OH	95	65	101	28	5	59	3	7	49	13	4								
O_	0	0	0	0	0	0	0	0	0	0	0	0							
S	28	20	19	19	2	16	0	1	15	1	4	0	0						
HN	577	232	585	89	38	334	5	21	472	62	73	0	9	204					
HS	7	6	6	5	0	3	0	0	2	0	1	0	0	2	0				
HO	44	53	43	53	4	54	4	9	18	11	4	0	1	48	1	3			
HA	222	167	125	72	19	92	3	18	98	9	14	0	14	51	7	30	33		
HC	839	468	679	434	83	809	23	89	1169	148	88	0	35	945	10	96	322	1221	
H_	354	121	219	207	62	139	13	77	781	76	48	0	23	439	7	65	119	900	164

TABLE 4.3 Sample size. For each pair of atom types, the average number of "close contact" atom pairs per MD snapshot was calculated. "Close contact" here means having a smoothed pair distance within 1 Å of the minimum observed smoothed distance. The average was calculated separately for each of the three separate MD trajectories. The entries in the table are the sum of these three averages. The value 40 was arbitrarily chosen as an indicator of low statistics. Values lower than 40 are shown in gray.

	CTa	CTb	C	CA	C_	N	N3	N_	O	O2	OH	O_	S	HN	HS	HO	HA	HC	H_
CTa	3.45																		
CTb	3.40	3.40																	
C	3.15	3.25	3.10																
CA	3.40	3.35	3.10	3.40															
C_	3.40	3.50	3.25	3.30	3.30														
N	3.10	3.05	3.00	3.05	3.15	3.05													
N3	3.10	3.05	3.00	3.05	3.15	3.05	3.05												
N_	3.40	3.20	3.15	3.25	3.30	3.05	3.05	3.05											
O	3.05	3.00	2.80	3.15	3.15	2.75	2.75	2.75	2.95										
O2	3.35	3.15	2.95	3.20	3.15	2.80	2.70	2.70	3.05	4.15									
OH	3.25	3.10	3.15	3.30	3.30	2.85	2.85	2.85	2.65	2.55	2.95								
O_	3.25	3.10	3.15	3.30	3.30	2.85	2.85	2.85	2.65	2.55	2.95	2.95							
S	3.30	3.30	3.30	3.30	3.30	3.20	3.20	3.20	3.20	3.20	3.20	3.20	3.60						
HN	2.45	2.60	2.25	2.40	2.40	2.20	2.20	2.10	1.80	1.75	1.95	1.95	2.55	1.85					
HS	2.80	2.90	2.35	2.95	3.05	2.60	2.60	2.60	1.95	1.95	1.95	1.95	2.55	1.95	2.45				
HO	2.80	2.90	2.35	3.05	3.05	2.35	2.35	2.35	1.70	1.65	2.10	2.10	2.55	1.95	2.45	2.45			
HA	2.85	2.70	2.75	2.65	2.65	2.60	2.60	2.60	2.45	2.60	2.60	2.60	2.55	2.10	2.20	2.20	2.35		
HC	2.70	2.75	2.60	2.65	2.70	2.50	2.50	2.65	2.45	2.50	2.55	2.55	2.55	1.85	2.25	2.25	2.25	2.20	
H_	2.65	2.45	2.55	2.60	2.70	2.50	2.50	2.55	2.25	2.50	2.50	2.50	2.55	2.00	2.20	2.20	2.10	2.10	2.05

TABLE 4.4 Minimum distance constraints, adjusted after considering low statistics. Some of the original constraint distances of Table 4.2 that had low statistics were (subjectively) lowered to that of similar types. The modified list of cutoff distances is given below. Grayed values mark those had low statistics (Table 4.4). All cutoffs involving type O_ were set to the cutoffs of type OH, since no data was available for type O_. All cutoffs involving type N3 were lowered to the cutoffs of type N, except for one (N…O2) in which the original cutoff was already lower. As there was no data for S…S interactions, the value of 3.60 was used from Word et al. (97). Distances are in Å.

describing their protocol, edited by DWF. Three different MD simulations were used as reference in the calibration: barnase (PDB access code 1A2P), an AMPA type glutamate receptor ligand binding domain (GluR2) (PDB access code 1FTO), and dihydrofolate reductase (PDB access code is 1RX2). The MD simulations followed a standard protocol as described in (119), using the AMBER 7 package (120) with the Cornell et al. force field (25). Equilibrium simulations were performed at a constant temperature of 300 K. Proteins were solvated in TIP3P water. Periodic boundary conditions were applied with the particle mesh Ewald (PME) method for long-range electrostatics as implemented in AMBER. The temperature was controlled using a Berendsen thermostat (121). Equilibration simulations were performed at constant pressure and equilibrium

simulations were performed at a constant volume to speed up the simulation. Total equilibrium simulation time for the GluR2 protein was 10 ns. The first 3 ns of the simulations were discarded and the last 7 ns were collected at every 0.5 ps. Total simulation time of barnase was slightly less than 8 ns, the last 4 ns were collected every 1 ps (only the last 4 ns were used in the calibration). The DHFR protein was simulated by about 3 ns, 2.5 ns were used in the calibration, collected every 1 ps.

Preparation of conformational end states

The closed (PDB 1RX2) and occluded (PDB 1RX6) states of DHFR were prepared for all atom simulations as follows. All missing atoms (including hydrogens) were added and heteroatoms removed from the structures. The structures were then slightly minimized using Sander (AMBER 7) using the standard protocol of steepest descent and conjugated gradient to remove the clashes, solvated in TIP3P water, and equilibrated for 500 ps in MD simulation at 300K.

Geometric targeting pathway generation

The final equilibrated closed state and occluded state were input into the geometric targeting method, and a pathway was generated from the occluded state to the closed state. The RMSD step size was 0.01 Å, and no random motion was added to the simulation. Hydrogen bonds and hydrophobic contacts common to both end structures were treated with fixed geometric constraints, and all other hydrogen bonds and hydrophobic contacts were left unconstrained. The minimum distance constraints used were those obtained in this work, scaled by a factor 0.95 to allow extra room for movement. As described in Chapter 2, for any pairs of atoms in the initial and target states that had distances less than the minimum distance constraints, the constraint distance was overridden and set to the distance from the input structure.

Umbrella sampling using pathway from geometric targeting as input

The initial coordinates for the umbrella sampling (US) simulations were taken from 10 GT snapshots spaced roughly equally along the reaction coordinate. The pathway snapshots were solvated in TIP3P water and Na+ were added to neutralize the system using the Leap module of Amber7. In order to relax water around the solute a 500 ps MD simulation in the *NPT* ensemble was performed with the backbone atoms constrained by a harmonic constraint of 4 kcal/mol-Å^2 at 300K. Then equilibration was performed at constant volume followed for up to 2.5 ns. The temperature was controlled using a Berendsen thermostat (121). Periodic boundary conditions were applied with Particle Mesh Ewald (PME) method using a 12 Å cutoff for nonbonded interactions. The SHAKE algorithm was applied (122). A 2 fs timestep for all simulations was employed.

For each window we performed 1-3 ns MD simulation in the *NVT* ensemble with a harmonic distance restraint of 6, 4 or 2 kcal/mol-Å^2 placed between the Cα atoms of Glu17 and Asp27. In regions of the reaction coordinate where more statistics were needed, additional windows were created from nearby windows, taking a snapshot from the nearby MD simulation as input to the new window. The coordinates for the final window were taken from the equilibrium MD simulation described above in "Preparation of conformational end states" (DHFR PDB ID 1RX2, Cα-Cα distance is 20.28 Å). For each window collect histograms of the Cα-Cα distance between Glu17 and Asp27 (the chosen reaction coordinate). The weighted histogram analysis method (WHAM) was applied to produce the potential of mean force (PMF) curve (54).

Results and discussion

We first show data to justify our use of the minimum 200 ps averaged pair distance from MD as the distance constraint in GT. Fig. 4.2 shows pair distance trajectories for five selected CTa-CTa pairs (CTa atom type defined in Table 4.1) from the DHFR reference MD simulation. Instantaneous pair distances vs. time are shown

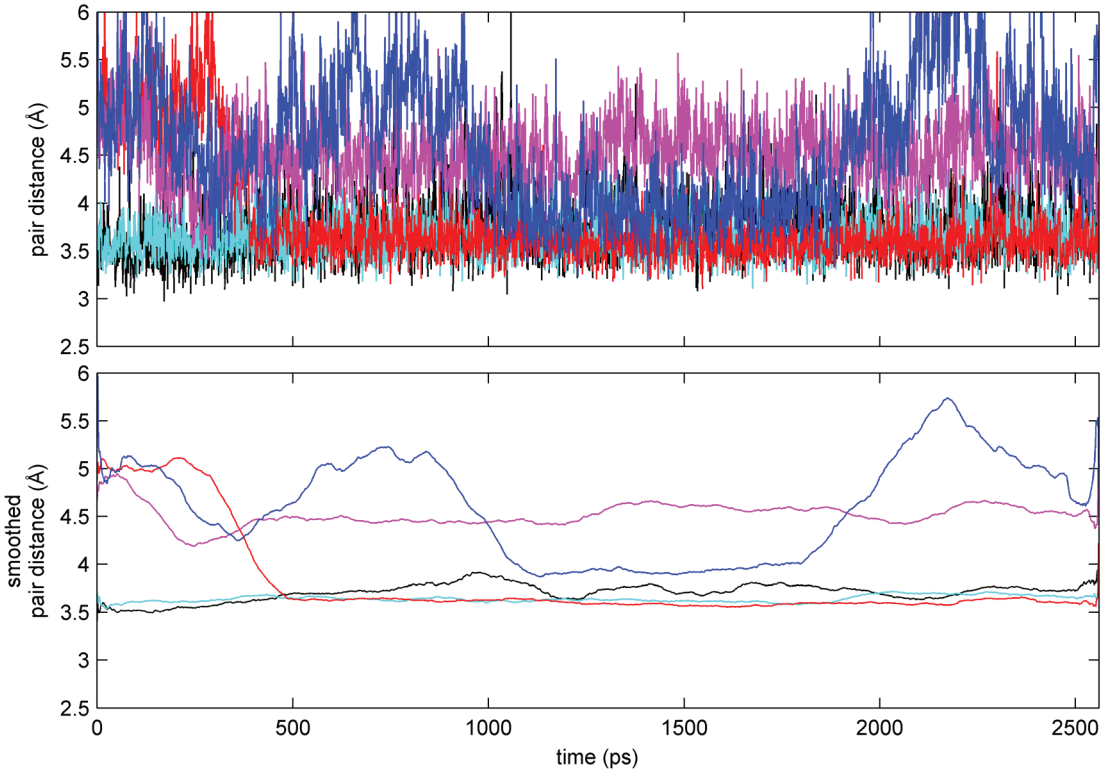

FIGURE 4.2 Pair distance trajectories for five selected carbon-carbon pairs from the DHFR reference MD simulation. The selected pairs were of pair type CTa-CTa (CTa is the atom type for tetrahedrally-coordinated carbon with only hydrogen and carbon neighbors, defined in Table 4.1). *(A)* the instantaneous pair distances vs. time of the selected pairs. *(B)* the 200-ps smoothed pair distances vs. time. In both panels *(A)* and *(B)*, the following atom pairs are represented: black=(Ala9 CB, Val119 CG2), cyan=(Ile41 CD1, Ile91 CG2), red=(Ile61 CG1, Val72 CG2), blue=(Val 78 CG1, Lys106 CD) and magenta=(Ile61 CD1, Val72 CG1). Three of the selected pairs come as close as about 3.5 Å, with fluctuations on the order of 0.5 Å (black, red, cyan). The other selected pairs (blue, magenta) have their closest average distances around 4.0 Å.

in Fig. 4.2 *A*, and smoothed pair distances are shown in Fig. 4.2 *B* (smoothed with 200 ps windowed average). We have selected three CTa-CTa pairs that have stable contact distances centered around ~3.5 Å with fluctuations on the order of 0.5 Å [Fig. 4.2 *A*-*B*, black=(ALA 9 CB, VAL 119 CG2), cyan=(ILE 41 CD1, ILE 91 CG2), red=(ILE 61 CG1, VAL 72 CG2)], and two CTa-CTa pairs that do not come as close [Fig. 4.2 *A*-*B*, blue=(VAL 78 CG1, LYS 106 CD) and magenta=(ILE 61 CD1, VAL 72 CG1)]. We see that CTa-CTa pairs can come as close as about 3.5 Å stably, and therefore the minimum

distance constraints that we assign to CTa-CTa pairs in GT should be no larger than 3.5 Å. Otherwise, we would prevent these stable contacts from forming. Furthermore, it does not seem necessary to use a distance much smaller than 3.5 Å as a constraint, because these distances only appear in the MD as transient fluctuations and therefore appear to be in a repulsive region (less energetically favorable).

There are a few reasons why the pairs represented by the blue and magenta curves in Fig. 4.2 do not form stable contacts around ~3.5 Å like the other pairs in the figure. Part of the reason is that atoms cannot freely diffuse in a protein. In a stable fold, there is some amount of wiggle room, but atoms are slaved to their Cα atoms via non-stretchable covalent bonds and in some cases are prevented from approaching each other closely unless the Cα atoms themselves were to move closer. Another reason is that the surfaces of CTa carbon atoms are "decorated" with hydrogen atoms, and depending on the orientation of the contact, the bonded hydrogens might be located on the CTa-CTa axis (pushing the carbons farther apart) or off axis (allowing a closer carbon-carbon distance). In the pair represented by the magenta curve, for example, the hydrogens are on-axis, and there is not an easy way within the stable fold of the protein for this orientation to change. In choosing the ~3.5 Å distance as the constraint, we are therefore decoupling the CTa-CTa repulsion from the hydrogen-hydrogen repulsion, which shall have its own constraint value, and from the effects of non-stretchable covalent bonds, which GT accounts for through other means.

Fig. 4.3 shows a histogram of pair distances between pairs of type CTa-CTa from the equilibrium MD trajectory of DHFR. First, second, and third covalent-neighbor pairs were excluded to examine the non-bonded distances. The histogram has been scaled by $1/4\pi r^2$ to give a distribution per unit surface area [however this neglects the fact that atoms at the protein surface are not surrounded by a full sphere of protein atoms (117)]. The histogram begins to rise around 3.0 Å and plateaus between 4.0-4.5 Å. We

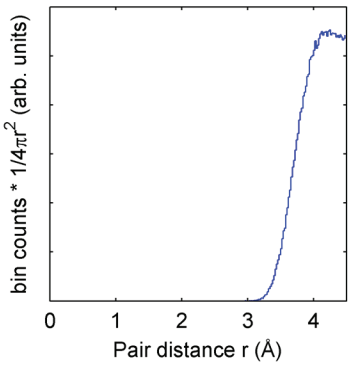

FIGURE 4.3 Pair histogram for carbon-carbon pairs of type CTa-CTa from the DHFR reference MD simulation. CTa is the atom type for tetrahedrally-coordinated carbon with only hydrogen and carbon neighbors, defined in Table 1. Bin spacing is 0.05 Å. Histogram was multiplied by $1/4\pi r^2$. The histogram begins to rise around 3.0 Å and plateaus between 4.0-4.5 Å. Observe that 3.5 Å is deep in the tail region, even though we have seen (Fig 4.2) that CTa-CTa pairs can stably persist at 3.5 Å. It would therefore be a misinterpretation to conclude from the histogram that pairs at distances less than 4.0 Å experience an effective repulsive force pulling them towards the 4.0 Å distance.

observed above that CTa-CTa pairs can stably maintain long-lived fluctuations about 3.5 Å, but on the pair histogram 3.5 Å is deep in the tail region. It would therefore be a misinterpretation to conclude from the histogram that CTa-CTa pairs at distances less than 4.0 Å experience an effective repulsive force pulling them towards the 4.0 Å distance. For this reason we have chosen to base the minimum distance constraints of GT in the smoothed pair trajectories rather than the pair histograms of the reference MD simulations.

In Fig. 4.4 we present the results of the fully atomistic umbrella sampling in explicit water (see Methods), performed by Tatyana Mamonova and Maria Kurnikova, using the pathway from GT as input snapshots for the restrained MD simulations. The free energy difference between the end states is ~2 kcal/mol, with the closed state favored over the occluded state, which matches well with the experimental value of 2.1 kcal/mol (17) and with published free energy calculations in this system by Arora and Brooks (45). The profile itself contains a prominent, sharp bump around 17 Å. As discussed in the introduction, we should not assume this profile is indicative of the optimal transition pathway, since the input pathway could have taken the system through a non-optimal route. The peak may be an artifact of the pathway.

FIGURE 4.4 Potential of mean force between closed and occluded states of DHFR. The free energy difference between the end states is ~2 kcal/mol, with the closed state favored over the occluded state, which matches well with the experimental value of 2.1 kcal/mol (4). The profile itself contains a prominent, sharp bump around 17 Å. We should not assume this profile is indicative of the optimal transition pathway, since the input pathway could have taken the system through a non-optimal route. The peak may be an artifact of the pathway.

As a proof-of-concept, it is demonstrated that the pathway snapshots from GT can be used as input to umbrella sampling. The work does not as yet address the question of what gains in efficiency there might be from using GT as opposed to other methods. On the one hand, GT pathways are easy to generate and take little time. On the other hand, upon transfer into MD, the GT snapshots must be solvated in explicit water and equilibrated before statistics can be collected. If we compare the total time required to obtain the initial equilibrated snapshots to the more common approach of using the restraining potential itself to gradually pull the system through the transition, it is not clear whether the use of GT has saved any time. A related question is whether the extra time invested up front in obtaining an optimal initial pathway, e.g. generating a minimum

energy pathway from nudged elastic band as was done by Arora and Brooks (45), ends up saving time in the subsequent umbrella sampling because of lower or less steep free energy barriers. Perhaps a clear advantage of using an initial pathway from GT could be demonstrated in a more complicated conformational change where other pathway methods cannot easily be applied. Another question concerns the assumption that the calculated free energy difference between end states is path independent. As a check, it would also be worthwhile to repeat the umbrella sampling with different initial pathways to test how the results are affected by the initial pathway snapshots.

With the calibrated minimum distance constraints, the GT snapshots stayed well-centered within their windows (not shown) rather than pulling to the side as happened in prior attempts (see Introduction). The use of GT snapshots as input to umbrella sampling, facilitated by the calibrated constraints, is a new accomplishment that had not been achieved previously. However, it is reasonable to ask if a simpler set of constraint distances than the ones used here, using fewer atom types for example, would work just as well. It is also possible that a different approach to calibrating the distances could improve the distances. For example, in the work of Hornak and Simmerling (117), pair distance histograms were obtained from an MD simulation of a “soup” of free amino acids. These pair histograms were uninfluenced by folded protein structure and without surface effects, unlike the pair histograms obtained in this work from MD simulations of folded proteins. Further work would be needed to check whether the pair histograms of (117) have the same difficulties as the pair histograms calculated here, masking the stable contact distances. If not, they could be used to determine an improved set of distance constraints.

CHAPTER 5 CONCLUSION

Building off of prior work in constraint based modeling of proteins, we have created a conceptually simple all-atom model of protein pathways that casts complex interactions into simple geometric criteria. We have demonstrated that the method can rapidly produce stereochemically acceptable pathways in proteins as large as the 14-subunit GroEL complex, and for conformational changes with complicated motions like the reorganization of beta sheet topology of spindle assembly checkpoint protein or the dimerization and domain swapping of the protein CD2. These complicated motions are beyond the capacity of linear interpolation with energy minimization, and probably also the elastic network based models given their lack of collision avoidance although this has not been tested. Structure quality metrics calculated for the nitrogen regulatory protein C (NtrC) pathways show quantitatively that GT produces structures with surprisingly good all-atom local stereochemistry considering the level of approximations of the model.

There is significant value in having a method that can rapidly produce reasonable all-atom candidate pathways. Throughout science, simplifications and approximations yield physical intuition into phenomena that are often much more complicated in reality. The GT method is designed to give intuition into motions involved in conformational change. The results in nitrogen regulatory protein C demonstrate that the model predicts the same fundamental pathway features as much more detailed MD simulations in this case, despite being in comparison a very coarse method. It would be interesting to compare with more rigorous approaches such as nudged elastic band. Certainly as with any approximation it is important to keep in mind the limitations and potential room for error in GT. In GT, simultaneity of events should be viewed with skepticism. Furthermore, pathway features that depend on energetics may be wrong, while features that arise out of geometric necessity, such as translational and rotational motion of

domains or structural elements, obstacle avoidance, squeezing through narrow passages, sequential motions whose simultaneous movement is geometrically forbidden, etc., should be more reliable. It may be that geometry plays a more significant role in very large machine-like proteins and complexes that have multiple moving parts, such as ATP synthase, SWI/SNF, myosin, and the ribosome, than it does in smaller protein systems. The usual pathway methods are nearly impossible to apply to very large systems, and GT may have a unique advantage in this domain. It will be interesting to produce pathways in these larger systems, perhaps targeting only a portion of the system to test the mechanical response of the rest of the system. GT pathways may elucidate how various moving parts in a large protein complex are mechanically coupled.

Making contact with experiment, we have used GT pathways as input to an MD-based calculation of the free energy difference between two conformations of dihydrofolate reductase matching a result from NMR (Chapter 4), and we have mentioned the application of FRODAN to cryo-EM fitting and computational protein folding predictions (Chapter 4). We have not, however, compared conformational change pathways from GT with experiment. The difficulty here is that experimental techniques cannot usually determine pathways. One way to make a comparison against experiment would be to generate pathways between structures for which there is a known, non-trivial intermediate structure from crystallography, as in (123). Another avenue for comparing GT pathways with experiment is currently being pursued in new work by Adam M. R. de Graff, Gareth Shannon, Daniel W. Farrell, Philip M. Williams, and M. F. Thorpe (paper in preparation), in which protein unfolding pathways under an applied pulling force are generated with GT and compared to MD and experiment. In pulling experiments, the two ends of a folded protein chain are pulled in opposite directions using atomic force microscopy (124) or optical tweezers (125), obtaining force profiles vs. end-to-end distance. From experimental force profiles it is possible to identify end-to-end distance

values of some stable unfolding intermediates, as well as distances at which sudden unfolding events occur. An experimental technique called Φ-value analysis (126) can identify the portions of a protein that have changed environments between the folded state and the unfolding transition state, revealing in some cases which structural elements unfold first under a pulling force. By setting the spring constants in the GT model to physically reasonable values and breaking springs as they become over-stressed by pulling, de Graff et al. (paper in preparation) show that the model can produce unfolding pathways that in many cases match MD at lower computational cost, which in turn agree with many but not all aspects of the experimental results.

Another way to connect GT pathways with experiment could be to investigate why a certain mutation affects the ability of a protein to undergo conformational change, if the effect of the mutation cannot be understood from looking at the conformational end points. A pathway from GT might reveal that the residue of interest participates in a transient interaction along the pathway, explaining the effect of the mutation. Related to this point, further studies involving the fitting of high-resolution structures into low-resolution cryo-EM density maps (Chapter 4) could reveal key residues that change environments upon conformational change, which may be candidates for mutation.

Generating very large quantities of pathways for a particular protein system using different random seeds may be a useful application, and we have already seen that GT can in some cases find two significantly different routes that are geometrically plausible (Chapter 3). The optimal pathway between given endpoints should in principle be a member of the ensemble of generated pathways, however in practice it may be infinitesimally unlikely for a pathway close to the optimal pathway to be generated. The effect of the RMSD bias, which pulls hardest on the atoms farthest from the target, may cause an entire ensemble of generated pathways to take the wrong route. One way to overcome some of the effects of the RMSD bias would be to implement a constraint

on the *difference* between the RMSD-to-initial structure and the RMSD-to-target, as in (114-116). The constraint on the RMSD difference, rather than the RMSD itself, allows the system to venture far in RMSD from both the initial and target structures, rather than requiring a monotonic decrease in RMSD. Another way to remove bias could be to implement a simplified "elastic band" model of geometric targeting, like nudged elastic band (NEB) (see Chapter 1). In the flat region of the pseudo energy landscape of GT, an elastic band approach may allow replicas to freely stretch between the end conformations, avoiding the side walls of the pseudo energy function. If the elastic connections between replicas are given a non-zero equilibrium length, many different string-like pathways within the flat allowed region could possibly be generated without RMSD bias. A further difficulty with generating an ensemble of pathways in GT is that simultaneity of large geometrically-independent motions may be inevitable. An optimal pathway may involve sequential motions, but it would be extremely unlikely in GT for random perturbations to cause one large collective motion to happen first while another independent large motion by chance remained unperturbed. The flat energy landscape in GT cannot encode sequential motions except where geometrically coupled.

The use of GT pathways as input to other methods has potential. The proof-of-concept that GT pathways can start off umbrella sampling free energy calculations is significant, and merits further work with more complicated transitions than the one presented in Chapter 4, where other pathway methods choke or are not easily applicable. The key here is that the input pathway to umbrella sampling need not be optimal in order to obtain an accurate estimate of the relative free energy between the endpoints, but the input pathway does need to avoid sharp changes in free energy. Another technique that could benefit from an input pathway from GT is nudged elastic band (NEB) (Chapter 1), a technique that refines an initial pathway. NEB begins with the replicas located at the two end conformations, with none in between, the band tightly stretched across the

expanse from one end to the other (43). The relaxation of the band allows the replicas to spread out along a pathway that is initially close to linear. There may be situations where it would be advantageous to seed the NEB with GT pathway snapshots, giving the NEB a head start, skipping past the initial relaxation of the band where the pathway is nearly linear.

The two webservers http://pathways.asu.edu (for geometric targeting) and http://flexweb.asu.edu (for geometric simulation) are open for public use.

REFERENCES

1. Phillips, R., J. Kondev, and J. Theriot. 2009. Physical biology of the cell. Garland Science, New York.

2. Garrett, R. H., and C. M. Grisham. 2005. Biochemistry. Thomson Brooks/Cole, Belmont, CA.

3. Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. 2002. Molecular biology of the cell. Garland Science, New York.

4. Nakanishi-Matsui, M., M. Sekiya, R. K. Nakamoto, and M. Futai. The mechanism of rotating proton pumping ATPases. Biochim. Biophys. Acta-Bioenerg. 1797:1343-1352.

5. Speir, J. A., S. Munshi, G. Wang, T. S. Baker, and J. E. Johnson. 1995. Structures of the native and swollen forms of cowpea chlorotic mottle virus determined by X-ray crystallography and cryo-electron microscopy. Structure 3:63-78.

6. Miao, Y., J. E. Johnson, and P. J. Ortoleva. All-Atom Multiscale Simulation of Cowpea Chlorotic Mottle Virus Capsid Swelling. The Journal of Physical Chemistry B 114:11181-11195.

7. Muller, C. W., G. J. Schlauderer, J. Reinstein, and G. E. Schulz. 1996. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure 4:147-156.

8. Henzler-Wildman, K. A., M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. Kern. 2007. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature 450:913-U927.

9. Henzler-Wildman, K. A., V. Thai, M. Lei, M. Ott, M. Wolf-Watz, T. Fenn, E. Pozharski, M. A. Wilson, G. A. Petsko, M. Karplus, C. G. Hubner, and D. Kern. 2007. Intrinsic motions along an enzymatic reaction trajectory. Nature 450:838-U813.

10. James, L. C., P. Roversi, and D. S. Tawfik. 2003. Antibody Multispecificity Mediated by Conformational Diversity. Science 299:1362-1367.

11. Drenth, J. M. J. 2007. Principles of protein x-ray crystallography. Springer.

12. Cavanagh, J. 2007. Protein NMR spectroscopy principles and practice. Amsterdam ; Boston : Academic Press. xxv, 885 p. : ill. ; 824 cm.

13. Frank, J. 2006. Three-dimensional electron microscopy of macromolecular assemblies : visualization of biological molecules in their native state. New York.

14. Zewail, A. H. 2010. Four-Dimensional Electron Microscopy. Science 328:187-193.

15. Flannigan, D. J., B. Barwick, and A. H. Zewail. 2010. Biological imaging with 4D ultrafast electron microscopy. Proceedings of the National Academy of Sciences of the United States of America 107:9933-9937.

16. Zewail, A. H. 2010. FILMING THE INVISIBLE IN 4-D. Scientific American 303:74-81.

17. McElheny, D., J. R. Schnell, J. C. Lansing, H. J. Dyson, and P. E. Wright. 2005. Defining the role of active-site loop fluctuations in dihydrofolate reductase catalysis. Proceedings of the National Academy of Sciences of the United States of America 102:5032-5037.

18. Mittermaier, A., and L. E. Kay. 2006. New Tools Provide New Insights in NMR Studies of Protein Dynamics. Science 312:224-228.

19. Liu, S., E. A. Abbondanzieri, J. W. Rausch, S. F. J. L. Grice, and X. Zhuang. 2008. Slide into Action: Dynamic Shuttling of HIV Reverse Transcriptase on Nucleic Acid Substrates. Science 322:1092-1097.

20. Gurunathan, K., and M. Levitus. 2009. Single-Molecule Fluorescence Studies of Nucleosome Dynamics. Current Pharmaceutical Biotechnology 10:559-568.

21. Wells, S., S. Menor, B. Hespenheide, and M. F. Thorpe. 2005. Constrained geometric simulation of diffusive motion in proteins. Physical Biology 2:S127-S136.

22. Farrell, D. W., K. Speranskiy, and M. F. Thorpe. 2010. Generating stereochemically acceptable protein pathways. Proteins: Structure, Function, and Bioinformatics 78:2908-2921.

23. Krebs, W. G., and M. Gerstein. 2000. The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. Nucleic Acids Research 28:1665-1675.

24. Lei, M., J. Velos, A. Gardino, A. Kivenson, M. Karplus, and D. Kern. 2009. Segmented Transition Pathway of the Signaling Protein Nitrogen Regulatory Protein C. Journal of Molecular Biology 392:823-836.

25. Cornell, W. D., P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. 1995. A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. J. Am. Chem. Soc. 117:5179-5197.

26. Fersht, A. 1999. Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding. W.H. Freeman, (1998 printing).

27. Lovell, S. C., I. W. Davis, W. B. Adrendall, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. 2003. Structure validation by C alpha geometry: phi,psi and C beta deviation. Proteins-Structure Function and Genetics 50:437-450.

28. Ramachandran, G. N., C. Ramakrishnan, and V. Sasisekharan. 1963. Stereochemistry of Polypeptide Chain Configurations. Journal of Molecular Biology 7:95-&.

29. Ho, B. K., A. Thomas, and R. Brasseur. 2003. Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. Protein Science 12:2508-2522.

30. Lovell, S. C., J. M. Word, J. S. Richardson, and D. C. Richardson. 2000. The penultimate rotamer library. Proteins-Structure Function and Genetics 40:389-408.

31. Leach, A. R. 2001. Molecular modelling : principles and applications. New York, Harlow, England.

32. Mulder, F. A. A., A. Mittermaier, B. Hon, F. W. Dahlquist, and L. E. Kay. 2001. Studying excited states of proteins by NMR spectroscopy. Nature Structural Biology 8:932-935.

33. Eisenmesser, E. Z., O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern. 2005. Intrinsic dynamics of an enzyme underlies catalysis. Nature 438:117-121.

34. Schlitter, J., M. Engels, and P. Kruger. 1994. Targeted Molecular-Dynamics - a New Approach for Searching Pathways of Conformational Transitions. Journal of Molecular Graphics 12:84-89.

35. Schlitter, J., M. Engels, P. Kruger, E. Jacoby, and A. Wollmer. 1993. Targeted Molecular-Dynamics Simulation of Conformational Change - Application to the T[--]R Transition in Insulin. Molecular Simulation 10:291-&.

36. Frenkel, D., and B. Smit. 2002. Understanding molecular simulation : from algorithms to applications. Academic Press, San Diego.

37. Case, D. A., T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. 2005. The Amber biomolecular simulation programs. Journal of Computational Chemistry 26:1668-1688.

38. Ma, J. P., P. B. Sigler, Z. H. Xu, and M. Karplus. 2000. A dynamic model for the allosteric mechanism of GroEL. Journal of Molecular Biology 302:303-313.

39. Sanbonmatsu, K. Y., S. Joseph, and C. S. Tung. 2005. Simulating movement of tRNA into the ribosome during decoding. Proceedings of the National Academy of Sciences of the United States of America 102:15854-15859.

40. Mills, G., H. Jónsson, and G. K. Schenter. 1995. Reversible work transition state theory: application to dissociative adsorption of hydrogen. Surface Science 324:305-337.

41. Jónsson, H., G. Mills, and K. W. Jacobsen. 1998. Nudged elastic band method for finding minimum energy paths of transitions. In Classical and Quantum Dynamics in Condensed Phase Simulations. B. J. Berne, G. Ciccoti, and D. F. Coker, editors. World Scientific, Singapore. 385–404.

42. Chu, J. W., B. L. Trout, and B. R. Brooks. 2003. A super-linear minimization scheme for the nudged elastic band method. Journal of Chemical Physics 119:12708-12717.

43. Bergonzo, C., A. J. Campbell, R. C. Walker, and C. Simmerling. 2009. A Partial Nudged Elastic Band Implementation for Use With Large or Explicitly Solvated Systems. Int. J. Quantum Chem. 109:3781-3790.

44. Arora, K., and C. L. Brooks, 3rd. 2007. Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. Proc Natl Acad Sci U S A 104:18496-18501.

45. Arora, K., and C. L. Brooks. 2009. Functionally Important Conformations of the Met20 Loop in Dihydrofolate Reductase are Populated by Rapid Thermal Fluctuations. J. Am. Chem. Soc. 131:5642-5647.

46. Bolhuis, P. G., D. Chandler, C. Dellago, and P. L. Geissler. 2002. TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark. Annual Review of Physical Chemistry 53:291.

47. Vreede, J., J. Juraszek, and P. G. Bolhuis. 2010. Predicting the reaction coordinates of millisecond light-induced conformational changes in photoactive yellow protein. Proceedings of the National Academy of Sciences 107:2397-2402.

48. Radhakrishnan, R., and T. Schlick. 2004. Orchestration of cooperative events in DNA synthesis and repair mechanism unraveled by transition path sampling of DNA polymerase Î²'s closing. Proceedings of the National Academy of Sciences of the United States of America 101:5970-5975.

49. Juraszek, J., and P. G. Bolhuis. 2006. Sampling the multiple folding mechanisms of Trp-cage in explicit solvent. Proceedings of the National Academy of Sciences 103:15859-15864.

50. Kim, M. K., G. S. Chirikjian, and R. L. Jernigan. 2002. Elastic models of conformational transitions in macromolecules. Journal of Molecular Graphics & Modelling 21:151-160.

51. Kim, M. K., R. L. Jernigan, and G. S. Chirikjian. 2005. Rigid-cluster models of conformational transitions in macromolecular machines and assemblies. Biophysical Journal 89:43-55.

52. Feng, Y., L. Yang, A. Kloczkowski, and R. L. Jernigan. 2009. The energy profiles of atomic conformational transition intermediates of adenylate kinase. Proteins 77:551-558.

53. Torrie, G. M., and J. P. Valleau. 1977. NON-PHYSICAL SAMPLING DISTRIBUTIONS IN MONTE-CARLO FREE-ENERGY ESTIMATION - UMBRELLA SAMPLING. Journal of Computational Physics 23:187-199.

54. Kumar, S., D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. 1992. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. 1. The Method. Journal of Computational Chemistry 13:1011-1021.

55. Mills, M., and I. Andricioaei. 2008. An experimentally guided umbrella sampling protocol for biomolecules. J Chem Phys 129:114101.

56. Jacobs, D. J., A. J. Rader, L. A. Kuhn, and M. F. Thorpe. 2001. Protein flexibility predictions using graph theory. Proteins-Structure Function and Genetics 44:150-165.

57. Lei, M., M. I. Zavodszky, L. A. Kuhn, and M. F. Thorpe. 2004. Sampling protein conformations and pathways. Journal of Computational Chemistry 25:1133-1148.

58. Lee, A. 2008. Geometric constraint systems with applications in CAD and biology. University of Massachusetts Amherst. xvi, 156 p.

59. Sljoka, A. 2006. Counting for rigidity, flexibility and extensions via the pebble game algorithm. York University. xxii, 173 leaves.

60. Laman, G. 1970. GRAPHS AND RIGIDITY OF PLANE SKELETAL STRUCTURES. Journal of Engineering Mathematics 4:331-&.

61. Jacobs, D. J., and B. Hendrickson. 1997. An algorithm for two-dimensional rigidity percolation: The pebble game. Journal of Computational Physics 137:346-365.

62. Jacobs, D. J., and M. F. Thorpe. 1995. GENERIC RIGIDITY PERCOLATION - THE PEBBLE GAME. Physical Review Letters 75:4051-4054.

63. Lee, A., and I. Streinu. 2008. Pebble game algorithms and sparse graphs. Discrete Mathematics 308:1425-1437.

64. Tay, T. S. 1984. RIGIDITY OF MULTI-GRAPHS .1. LINKING RIGID BODIES IN N-SPACE. Journal of Combinatorial Theory Series B 36:95-112.

65. Hespenheide, B. M., D. J. Jacobs, and M. F. Thorpe. 2004. Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. J. Phys.-Condes. Matter 16:S5055-S5064.

66. Thorpe, M., M. Chubynsky, B. Hespenheide, S. Menor, D. J. Jacobs, L. A. Kuhn, M. I. Zavodsky, M. Lei, A. J. Rader, and W. Whitley. 2005. Flexibility in Biomolecules. In Current topics in physics : in honor of Sir Roger J. Elliott. R. A. Barrio, K. Kaski, and R. J. Elliott, editors. Imperial College Press, London. 97-112.

67. Tay, T. S. 1989. LINKING (N-2)-DIMENSIONAL PANELS IN N-SPACE-II - (N-2,2)-FRAMEWORKS AND BODY AND HINGE STRUCTURES. Graphs and Combinatorics 5:245-273.

68. Tay, T. S., and W. Whiteley. 1984. Recent advances in the generic rigidity of structures. Structural Topology 9:31-38.

69. Katoh, N., and S.-i. Tanigawa. 2009. A proof of the molecular conjecture. In Proceedings of the 25th annual symposium on Computational geometry. ACM, Aarhus, Denmark.

70. Ahmed, A., and H. Gohlke. 2006. Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. Proteins: Structure, Function, and Bioinformatics 63:1038-1051.

71. Gohlke, H., and M. F. Thorpey. 2006. A natural coarse graining for simulating large biomolecular motion. Biophysical Journal 91:2115-2120.

72. Radestock, S., and H. Gohlke. 2008. Exploiting the Link between Protein Rigidity and Thermostability for Data-Driven Protein Engineering. Engineering in Life Sciences 8:507-522.

73. FIRST version 6.2.1 User Guide. http://flexweb.asu.edu/software/first/user_guides/FIRST6.2.1_user_guide.pdf.

74. Jolley, C. C., S. A. Wells, P. Fromme, and M. F. Thorpe. 2008. Fitting low-resolution cryo-EM maps of proteins using constrained geometric simulations. Biophysical Journal 94:1613-1621.

75. Glembo, T. J., and S. B. Ozkan. 2010. Union of Geometric Constraint-Based Simulations with Molecular Dynamics for Protein Structure Prediction. Biophysical Journal 98:1046-1054.

76. Fletcher, R., and C. M. Reeves. 1964. Function Minimization by Conjugate Gradients. Comput. J. 7:149-&.

77. Weinan, E., W. Q. Ren, and E. Vanden-Eijnden. 2005. Finite temperature string method for the study of rare events. Journal of Physical Chemistry B 109:6688-6693.

78. Maragliano, L., A. Fischer, E. Vanden-Eijnden, and G. Ciccotti. 2006. String method in collective variables: minimum free energy paths and isocommittor surfaces. J Chem Phys 125:24106.

79. Pan, A. C., D. Sezer, and B. Roux. 2008. Finding transition pathways using the string method with swarms of trajectories. Journal of Physical Chemistry B 112:3432-3440.

80. Yang, H. J., H. Wu, D. W. Li, L. Han, and S. H. Huo. 2007. Temperature-dependent probabilistic roadmap algorithm for calculating variationally optimized conformational transition pathways. Journal of Chemical Theory and Computation 3:17-25.

81. Branduardi, D., F. L. Gervasio, and M. Parrinello. 2007. From A to B in free energy space. Journal of Chemical Physics 126.

82. Elber, R., and M. Karplus. 1987. A METHOD FOR DETERMINING REACTION PATHS IN LARGE MOLECULES - APPLICATION TO MYOGLOBIN. Chemical Physics Letters 139:375-380.

83. Fischer, S., and M. Karplus. 1992. CONJUGATE PEAK REFINEMENT - AN ALGORITHM FOR FINDING REACTION PATHS AND ACCURATE TRANSITION-STATES IN SYSTEMS WITH MANY DEGREES OF FREEDOM. Chemical Physics Letters 194:252-261.

84. Elber, R. 2005. Long-timescale simulation methods. Curr Opin Struct Biol 15:151-156.

85. Christen, M., and W. F. Van Gunsteren. 2008. On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. Journal of Computational Chemistry 29:157-166.

86. Isralewitz, B., M. Gao, and K. Schulten. 2001. Steered molecular dynamics and mechanical functions of proteins. Current Opinion in Structural Biology 11:224-230.

87. van der Vaart, A., and M. Karplus. 2005. Simulation of conformational transitions by the restricted perturbation-targeted molecular dynamics method. Journal of Chemical Physics 122:-.

88. van der Vaart, A., and M. Karplus. 2007. Minimum free energy pathways and free energy profiles for conformational transitions based on atomistic molecular dynamics simulations. Journal of Chemical Physics 126.

89. Song, G., and R. L. Jernigan. 2006. An enhanced elastic network model to represent the motions of domain-swapped proteins. Proteins 63:197-209.

90. Maragakis, P., and M. Karplus. 2005. Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. J Mol Biol 352:807-822.

91. Schuyler, A. D., R. L. Jernigan, P. K. Qasba, B. Ramakrishnan, and G. S. Chirikjian. 2009. Iterative cluster-NMA: A tool for generating conformational transitions in proteins. Proteins 74:760-776.

92. Flores, S., N. Echols, D. Milburn, B. Hespenheide, K. Keating, J. Lu, S. Wells, E. Z. Yu, M. Thorpe, and M. Gerstein. 2006. The Database of Macromolecular Motions: new features added at the decade mark. Nucleic Acids Res 34:D296-301.

93. Echols, N., D. Milburn, and M. Gerstein. 2003. MolMovDB: analysis and visualization of conformational change and structural flexibility. Nucleic Acids Res 31:478-482.

94. Seeliger, D., J. Haas, and B. L. de Groot. 2007. Geometry-based sampling of conformational transitions in proteins. Structure 15:1482-1492.

95. Seeliger, D., and B. L. De Groot. 2009. tCONCOORD-GUI: Visually Supported Conformational Sampling of Bioactive Molecules. Journal of Computational Chemistry 30:1160-1166.

96. Angel, H. 2006. Biomolecules in the computer: Jmol to the rescue. Biochemistry and Molecular Biology Education 34:255-261.

97. Word, J. M., S. C. Lovell, J. S. Richardson, and D. C. Richardson. 1999. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. Journal of Molecular Biology 285:1735-1747.

98. Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. 2007. Clustal W and clustal X version 2.0. Bioinformatics 23:2947-2948.

99. Dahiyat, B. I., D. B. Gordon, and S. L. Mayo. 1997. Automated design of the surface positions of protein helices. Protein Science 6:1333-1337.

100. Wells, S. A., M. T. Dove, and M. G. Tucker. 2002. Finding best-fit polyhedral rotations with geometric algebra. J. Phys.-Condes. Matter 14:4567-4584.

101. Shewchuk, J. R. 1994. An Introduction to the Conjugate Gradient Method Without the Agonizing Pain. http://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf.

102. Berman, H., K. Henrick, and H. Nakamura. 2003. Announcing the worldwide Protein Data Bank. Nature Structural Biology 10:980-980.

103. Karplus, M., and J. N. Kushick. 1981. Method for estimating the configurational entropy of macromolecules. Macromolecules 14:325-332.

104. Kubitzki, M. B., and B. L. de Groot. 2008. The Atomistic Mechanism of Conformational Transition in Adenylate Kinase: A TEE-REX Molecular Dynamics Study. Structure 16:1175-1182.

105. Apostolakis, J., P. Ferrara, and A. Caflisch. 1999. Calculation of conformational transitions and barriers in solvated systems: Application to the alanine dipeptide in water. The Journal of Chemical Physics 110:2099-2108.

106. Davis, I. W., A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall, J. Snoeyink, J. S. Richardson, and D. C. Richardson. 2007. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Research 35:W375-W383.

107. Latzer, J., T. Shen, and P. G. Wolynes. 2008. Conformational switching upon phosphorylation: A predictive framework based on energy landscape principles. Biochemistry 47:2110-2122.

108. Schnell, J. R., H. J. Dyson, and P. E. Wright. 2004. STRUCTURE,DYNAMICS, AND CATALYTIC FUNCTION OF DIHYDROFOLATE REDUCTASE. Annual Review of Biophysics & Biomolecular Structure 33:119-C-113.

109. Boehr, D. D., D. McElheny, H. J. Dyson, and P. E. Wright. 2006. The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis. Science 313:1638-1642.

110. Boehr, D. D., D. McElheny, H. J. Dyson, and P. E. Wright. 2010. Millisecond timescale fluctuations in dihydrofolate reductase are exquisitely sensitive to the bound ligands. Proceedings of the National Academy of Sciences of the United States of America 107:1373-1378.

111. Mamonova, T., M. J. Yonkunas, and M. G. Kurnikova. 2008. Energetics of the Cleft Closing Transition and the Role of Electrostatic Interactions in Conformational Rearrangements of the Glutamate Receptor Ligand Binding Domainâ€ Biochemistry 47:11077-11085.

112. Novak, B. R., D. Moldovan, G. L. Waldrop, and M. S. d. Queiroz. 2009. Umbrella Sampling Simulations of Biotin Carboxylase: Is a Structure with an Open ATP Grasp Domain Stable in Solution? The Journal of Physical Chemistry B 113:10097-10103.

113. Mamonova, T., and M. Kurnikova. 2006. Structure and Energetics of Channel-Forming Proteinâ˜Polysaccharide Complexes Inferred via Computational Statistical Thermodynamics. The Journal of Physical Chemistry B 110:25091-25100.

114. Banavali, N. K., and B. Roux. 2005. The N-Terminal End of the Catalytic Domain of Src Kinase Hck Is a Conformational Switch Implicated in Long-Range Allosteric Regulation. Structure 13:1715-1723.

115. Banavali, N. K., and B. t. Roux. 2005. Free Energy Landscape of A-DNA to B-DNA Conversion in Aqueous Solution. J. Am. Chem. Soc. 127:6866-6876.

116. Yu, H. B., L. Ma, Y. Yang, and Q. Cui. 2007. Mechanochemical coupling in the myosin motor domain. I. Insights from equilibrium active-site simulations. PLoS Computational Biology 3:199-213.

117. Hornak, V., and C. Simmerling. 2004. Development of softcore potential functions for overcoming steric barriers in molecular dynamics simulations. Journal of Molecular Graphics and Modelling 22:405-413.

118. Li, A. J., and R. Nussinov. 1998. A set of van der Waals and Coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. Proteins-Structure Function and Bioinformatics 32:111-127.

119. Mamonova, T., B. Hespenheide, R. Straub, M. F. Thorpe, and M. Kurnikova. 2005. Protein flexibility using constraints from molecular dynamics simulations. Physical Biology 2:S137-S147.

120. Case, D. A., D. A. Pearlman, J. W. Caldwell, I. T.E. Cheatham, J. Wang, W. S. Ross, C. L. Simmerling, T. A. Darden, K. M. Merz, R. V. Stanton, A. L. Cheng, J. J. Vincent, M. Crowley, V. Tsui, H. Gohlke, R. J. Radmer, Y. Duan, J. Pitera, I. Massova, G. L. Seibel, U. C. Singh, P. K. Weiner, and P. A. Kollman. 2002. AMBER 7. University of California, San Francisco.

121. Berendsen, H. J. C., J. P. M. Postma, W. F. Vangunsteren, A. Dinola, and J. R. Haak. 1984. MOLECULAR-DYNAMICS WITH COUPLING TO AN EXTERNAL BATH. Journal of Chemical Physics 81:3684-3690.

122. Ryckaert, J. P., G. Ciccotti, and H. J. C. Berendsen. 1977. NUMERICAL-INTEGRATION OF CARTESIAN EQUATIONS OF MOTION OF A SYSTEM WITH CONSTRAINTS - MOLECULAR-DYNAMICS OF N-ALKANES. Journal of Computational Physics 23:327-341.

123. Weiss, D. R., and M. Levitt. 2009. Can Morphing Methods Predict Intermediate Structures? Journal of Molecular Biology 385:665-674.

124. Borgia, A., P. M. Williams, and J. Clarke. 2008. Single-molecule studies of protein folding. Annual Review of Biochemistry 77:101-125.

125. Cecconi, C., E. A. Shank, C. Bustamante, and S. Marqusee. 2005. Direct observation of the three-state folding of a single protein molecule. Science 309:2057-2060.

126. Ng, S. P., R. W. S. Rounsevell, A. Steward, C. D. Geierhaas, P. M. Williams, E. Paci, and J. Clarke. 2005. Mechanical unfolding of TNfn3: The unfolding pathway of a fnIII domain probed by protein engineering, AFM and MD simulation. Journal of Molecular Biology 350:776-789.

127. The PyMOL Molecular Graphics System, version 1.2r1. Schrödinger, LLC.

APPENDIX A

RIGID BODY DEGREES OF FREEDOM

During enforcement of constraints (Chapter 2), the conjugate gradient minimization of the constraint-enforcing energy function [Eqs. (2.1)-(2.5)] requires derivatives of the energy function with respect to the translational and rotational degrees of freedom of the rigid units. In this section we derive the first derivatives of the energy function. We begin by present the mathematics used by both FRODAN and the original FRODA (21) for parametrizing rigid body rotations with three independent rotor variables (100). In what follows, we do not use the geometric algebra language (bivectors, etc.) used by Wells et al. (100), as the simpler language of vectors is sufficient for our purposes.

We begin with a single rigid unit from FRODAN with six degrees of freedom (three translational and three rotational), in which are embedded a set of N "embedded atoms." We designate the center of mass of the rigid unit as the center of rotation (all atoms are considered to have unit mass). The center of rotation is located at (Cartesian) position $\mathbf{R} = (X, Y, Z)$ with respect to some external coordinate system. These are the three translational degrees of freedom of the rigid unit. For simplicity, we consider just one of the embedded points in the rigid unit, whose (unrotated) position measured *with respect to the center of rotation of the rigid unit* is $\mathbf{r}' = (x', y', z')$. Prior to rotation, the absolute position $\mathbf{r}$ of this atom with respect to the external coordinate system is $\mathbf{r} = \mathbf{R} + \mathbf{r}'$. Representing the rotation by an orthogonal 3x3 matrix $\mathbf{M}$, the absolute position $\mathbf{r}$ is given by

$$\mathbf{r} = \mathbf{R} + \mathbf{M}\mathbf{r}' \tag{A.1}$$

where $\mathbf{M}\mathbf{r}'$ indicates matrix multiplication, and the vectors $\mathbf{R}$, $\mathbf{r}$, and $\mathbf{r}'$ are column vectors.

There is a general form for the rotation matrix that allows easy encoding of arbitrary rotations about an arbitrary axis (100). Let $\hat{\mathbf{B}}$ be a unit vector designating the axis of rotation (passing through the designated center of rotation), and let φ be a *positive* rotation angle about this axis, ranging from $0 \le \varphi \le \pi$ (the rotation angle follows a right-hand convention, measured counter-clockwise if the axial vector points out of the page). Let us also define a vector $\mathbf{B} = (B_x, B_y, B_z)$ as

$$\mathbf{B} \equiv \left(2 \sin \frac{\varphi}{2} \right) \hat{\mathbf{B}} \tag{A.2}$$

which points along the rotation axis $\hat{\mathbf{B}}$ and has magnitude $B = |\mathbf{B}| = 2 \sin(\varphi / 2)$. Observe that for small angles, $B \approx \varphi$, and that $B \le 2$ because of the sine function. The three components of vector $\mathbf{B}$ are the three rotational degrees of freedom of the rigid unit that uniquely specify the axis and angle of the rotation. We also define a scalar W as

$$W \equiv \cos \frac{\varphi}{2} \tag{A.3}$$

which in light of Eq. (A.2) and using a little trigonometry is equivalent to

$$W = \left(1 - \frac{B^2}{4} \right)^{\frac{1}{2}} \tag{A.4}$$

Note that W is not an independent parameter since it depends completely on B. The general form for the rotation matrix $\mathbf{M}$ parametrized by $\mathbf{B}$, which we state here without derivation [see Wells et al. (100)] is

$$\mathbf{M} = \begin{bmatrix} 1-\frac{B^2-B_x^2}{2} & -WB_z+\frac{B_xB_y}{2} & WB_y+\frac{B_zB_x}{2} \\ WB_z+\frac{B_xB_y}{2} & 1-\frac{B^2-B_y^2}{2} & -WB_x+\frac{B_yB_z}{2} \\ -WB_y+\frac{B_zB_x}{2} & WB_x+\frac{B_yB_z}{2} & 1-\frac{B^2-B_z^2}{2} \end{bmatrix}. \tag{A.5}$$

To summarize, the six degrees of freedom of a rigid unit in FRODAN are $\left(X,Y,Z,B_x,B_y,B_z\right)$. The instantaneous position of an embedded atom $\mathbf{r}$ is a function of these degrees of freedom and a fixed reference position $\mathbf{r}'$

$$\mathbf{r} = f\left(X,Y,Z,B_x,B_y,B_z|\mathbf{r}'\right) \tag{A.6}$$

where the transformation f is given by Eqs. (A.1) and (A.5).

In FRODAN, the instantaneous positions of the embedded atoms are tracked, being updated continuously as the degrees of freedom change. One thing that must be cared for, however, is that the magnitude of the vector B be kept sufficiently below its maximum allowed value of 2 (corresponding to 180°). At $B > 2$, the W parameter becomes imaginary [see Eq. (A.4)], and the rotation matrix is no longer valid. The way FRODAN handles this is at the beginning of every step, the reference coordinates $\mathbf{r}'$ of all rigid units (which denote unrotated positions relative to the center of rotation) are updated to $\mathbf{r}-\mathbf{R}$, and all rotors $\mathbf{B}$ are reset to the zero vector. This essentially re-initializes the unrotated state of each rigid unit to the current orientation, allowing each step to begin with $\mathbf{B}=0$. This resetting of the rotors is also performed if at any time during conjugate gradient minimization a rotor magnitude B rises above some threshold (currently 1.0).

Recall from Chapter 2 that the physical "atoms" of the system may have multiple copies, each in a different rigid unit. We define the position of an atom, $\bar{\mathbf{r}}$, as being

located at the mean of its multiple "split" copies that are embedded in various rigid units. The bar in $\overline{\mathbf{r}}$ indicates averaging over the M embedded atoms that correspond to the particular physical atom,

$$\overline{\mathbf{r}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{r}_m \tag{A.7}$$

In the constraint enforcing potential V from Chapter 2 [Eqs. (2.1)-(2.5)], most energy terms are explicit functions of the physical atom positions $\overline{\mathbf{r}}$. However, one term (the shared-atoms term) is an explicit function of the embedded atom positions $\mathbf{r}$ and cannot be expressed in terms of the physical atom positions. It is a simple task however, to convert the other energy terms from functions of atom positions $\overline{\mathbf{r}}$ to functions of the embedded atom positions $\mathbf{r}$, through Eq. (A.7), enabling V to be expressed as a function of embedded atom positions $\mathbf{r}$

$$V = V\left(\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_P\right) \tag{A.8}$$

where P is the total number of embedded atoms in the system, including copies of the same physical atom.

Conjugate gradient minimization is used to enforce constraints (Chapter 2). The procedure requires that we take the gradient of V with respect to the rigid unit degrees of freedom Q. However, there is a problem with the degrees of freedom as they are currently defined. The translational degrees of freedom $\left(X, Y, Z\right)$ have units of length (Å), whereas the rotational degrees of freedom $\left(B_x, B_y, B_z\right)$ are unitless. The conjugate gradient procedure uses the magnitude of the previous iteration's gradient in determining the next direction to search (76, 101), but this magnitude is meaningless if the variables do not share the same units. Recall that for small values of $B = \left|\mathbf{B}\right|$, we have $B \approx \varphi$.

Since B is approximately the rotation angle, we can scale each rotational degree of freedom by radius value unique to each rigid body. We choose the maximum radial distance of any embedded atom relative to the center of rotation as this scaling distance. Thus, by applying a scale factor to the rotational parameters, all degrees of freedom X, Y, Z, B_x, B_y, B_z can be expressed in units of Å. In what follows, to keep the math simpler we work with the *unscaled* rotational parameters.

In calculating derivatives, we will necessarily use the chain rule to connect changes in the Cartesian positions of the embedded atoms to changes in the rigid unit degrees of freedom. For a rigid unit containing M embedded atoms, the derivative operator for a rigid unit degree of freedom Q is

$$\frac{\partial}{\partial Q} = \sum_{i=m}^{M} \frac{\partial x_m}{\partial Q}\frac{\partial}{\partial x_m} + \frac{\partial y_m}{\partial Q}\frac{\partial}{\partial y_m} + \frac{\partial z_m}{\partial Q}\frac{\partial}{\partial z_m} \tag{A.9}$$

where the subscript m labels the embedded atoms in the rigid unit. Note that the sum does not include the embedded atoms of other rigid units, since the degree of freedom Q only affects the embedded atoms in its own rigid unit. The derivatives $\frac{\partial x}{\partial Q}, \frac{\partial y}{\partial Q}, \frac{\partial z}{\partial Q}$ in Eq. (A.9) are determined from Eq. (A.1), restated below in an expanded form that facilitates taking derivatives (and dropping the subscript m for simplicity)

$$\begin{aligned} x &= X + M_{xx}x' + M_{xy}y' + M_{xz}z' \\ y &= Y + M_{yx}x' + M_{yy}y' + M_{yz}z' \\ z &= Z + M_{zx}x' + M_{zy}y' + M_{zz}z' \end{aligned} \tag{A.10}$$

where $M_{ij} = M_{ij}\left(B_x, B_y, B_z\right)$. From Eq. (A.10), the derivatives $\frac{\partial x}{\partial Q}, \frac{\partial y}{\partial Q}, \frac{\partial z}{\partial Q}$ with $Q = X, Y,$ or Z are

$$\begin{aligned}
&\frac{\partial x}{\partial X}=1,\ \frac{\partial x}{\partial Y}=0,\ \frac{\partial x}{\partial Z}=0\\
&\frac{\partial y}{\partial X}=0,\ \frac{\partial y}{\partial Y}=1,\ \frac{\partial y}{\partial Z}=0\\
&\frac{\partial z}{\partial X}=0,\ \frac{\partial z}{\partial Y}=0,\ \frac{\partial z}{\partial Z}=1
\end{aligned} \tag{A.11}$$

The derivatives $\frac{\partial x}{\partial Q}, \frac{\partial y}{\partial Q}, \frac{\partial z}{\partial Q}$ with $Q = B_x, B_y,$ or B_z are more complicated. From Eq. (A.10), these derivatives are

$$\begin{aligned}
\frac{\partial x}{\partial B_x} &= 0 + \left(\frac{B_z B_x}{4W}+\frac{B_y}{2}\right)y' + \left(\frac{-B_x B_y}{4W}+\frac{B_z}{2}\right)z'\\
\frac{\partial x}{\partial B_y} &= -B_y x' + \left(\frac{B_y B_z}{4W}+\frac{B_x}{2}\right)y' + \left(W-\frac{B_y^2}{4W}\right)z'\\
\frac{\partial x}{\partial B_z} &= -B_z x' + \left(-W+\frac{B_z^2}{4W}\right)y' + \left(\frac{-B_y B_z}{4W}+\frac{B_x}{2}\right)z'\\
\frac{\partial y}{\partial B_x} &= \left(\frac{-B_z B_x}{4W}+\frac{B_y}{2}\right)x' - B_x y' + \left(-W+\frac{B_x^2}{4W}\right)z'\\
\frac{\partial y}{\partial B_y} &= \left(\frac{-B_y B_z}{4W}+\frac{B_x}{2}\right)x' + 0 + \left(\frac{B_x B_y}{4W}+\frac{B_z}{2}\right)z'\\
\frac{\partial y}{\partial B_z} &= \left(W-\frac{B_z^2}{4W}\right)x' - B_z y' + \left(\frac{B_z B_x}{4W}+\frac{B_y}{2}\right)z'\\
\frac{\partial z}{\partial B_x} &= \left(\frac{B_x B_y}{4W}+\frac{B_z}{2}\right)x' + \left(W-\frac{B_x^2}{4W}\right)y' - B_x z'\\
\frac{\partial z}{\partial B_y} &= \left(-W+\frac{B_y^2}{4W}\right)x' + \left(\frac{-B_x B_y}{4W}+\frac{B_z}{2}\right)y' - B_y z'\\
\frac{\partial z}{\partial B_z} &= \left(\frac{B_y B_z}{4W}+\frac{B_x}{2}\right)x' + \left(\frac{-B_z B_x}{4W}+\frac{B_y}{2}\right)y' + 0
\end{aligned} \tag{A.12}$$

Applying the operator $\frac{\partial}{\partial Q}$ from Eq. (A.9) to the energy function V, the derivative $\frac{\partial V}{\partial Q}$ with respect to any rigid unit degree of freedom Q is

$$\frac{\partial V}{\partial Q} = \sum_{i=m}^{M} \frac{\partial x_m}{\partial Q}\frac{\partial V}{\partial x_m} + \frac{\partial y_m}{\partial Q}\frac{\partial V}{\partial y_m} + \frac{\partial z_m}{\partial Q}\frac{\partial V}{\partial z_m}. \tag{A.13}$$

It is straightforward to apply the chain rule in Eq. (A.13) for each energy term in Eqs. (2.1)-(2.5), taking Cartesian derivatives of each term, and substituting the appropriate partial derivatives from Eqs. (A.11) and (A.12).

APPENDIX B

SUPPLEMENTARY MATERIAL

Protein Name	# Sub-units	Initial PDB	Final PDB	Chain Information
5'-Nucleotidase	1	1HP1	1HPU	Chain A from initial structure targeted to chain C of final.
Adenylate Kinase	1	4AKE	1AKE	Chain A from initial structure targeted to chain A of final.
Alcohol Dehydrogenase	1	8ADH	6ADH	Chain A from initial structure targeted to chain A of final.
Calmodulin	1	1CFD	1CFC	Chain A from initial structure targeted to chain A of final.
CD2	2	1HNG	1CDC	Chains AB from initial structure targeted to chains AB of final.
Citrate Synthase	2	5CSC	6CSC	Chains AB from initial structure targeted to chains AB of final.
Collagenase	1	1NQD	1NQJ	Chain A from initial structure targeted to chain B of final structure
Dengue 2 Virus Envelope Glycoprotein	1	1OAN	1OK8	Chain A from initial structure targeted to chain A of final.
Dihydrofolate Reductase	1	1RX2	1RX6	Chain A from initial structure targeted to chain A of final.
Diphtheria Toxin	1	1DDT	1MDT	Chain A from initial structure targeted to chain A of final.
DNA Polymerase	1	1IH7	1IG9	Chain A from initial structure targeted to chain A of final.
GroEL	14	1KP8	1AON	The chains in these two PDB files are not labeled consistently. The correct mapping of chains that we used in targeting was ABCDEFG to FEDCBAG (top ring) and HIJKLMN to HIJKLMN (bottom ring).
Heparin Cofactor II	1	1JMO	1JMJ	Chain A from initial structure targeted to chain A of final.
HIV Protease	2	2HB4	2AZ8	Since both PDB files only contain one subunit (chain A) in the asymmetric unit, the full biological unit (2 subunits) was downloaded from PDB and relabled as chains AB. Chains AB from initial structure were targeted to chains AB of final structure.
HIV-1 Reverse Transcriptase	2	1DLO	2HMI	Chains A+B from initial structure targeted to chains A+B of final.
Immunoglobulin E SPE7	2	1OAQ	1OCW	Chains HL from initial structure targeted to chains HL of final.
Phosphofructokinase	4	4PFK	6PFK	Since initial PDB file 4PFK only contains one subunit (chain A) in the asymetric unit, the full biological unit (4 subunits) was downloaded from PDB and relabeled as chains ABCD. These were targeted to chains ABCD of final structure.
Pyrophosphokinase	1	1HKA	1Q0N	Chain A from initial structure targeted to chain A of final.
Pyruvate Phosphate Dikinase	1	1KBL	2R82	Chain A from initial structure targeted to chain A of final.
Replication Factor C	6	2CHV	2CHQ	Since final PDB file 2CHQ only contains 3 subunits (chains ABC) in the asymmetric unit, the full biological unit (6 subunits was downloaded from the PDB and relabeled as chains ABCDEF. These were targeted to chains ABCDEF of the final structure.
Rho Transcription Termination Factor	6	3ICE	3ICE	Same PDB file was used for both initial and final structures. Chains ABCDEF of initial state were targeted to chains BCDEFA, to simulate one step in the cyclic transition of this protein.
Spindle Assembly Checkpoint Protein	1	1DUJ	1KLQ	Chain A from initial structure targeted to chain A of final.
Toy Model 1	1	-	-	
Toy Model 2	1	-	-	

TABLE B.1 PDB IDs and chain information (referenced in Chapter 2).

APPENDIX C

COPYRIGHTED MATERIAL

The content of Chapter 2, and the supplementary table in Appendix B, are published in Farrell, D. W., K. Speranskiy, and M. F. Thorpe. 2010. Generating stereochemically acceptable protein pathways. Proteins: Structure, Function, and Bioinformatics 78:2908-2921. Copyright 2010 Wiley-Liss, Inc. Reprinted in this dissertation with rights licensed to the authors in the Copyright Transfer Agreement. The article and all supplementary material can be obtained from the publisher at http://doi.wiley.com/10.1002/prot.22810

The content of Chapter 3 is a submitted manuscript, Farrell, D. W., M. Lei, and M. F. Thorpe. Comparison of Pathways between Geometric Targeting Method and Targeted Molecular Dynamics in Nitrogen Regulatory Protein C. Submitted to Journal of Molecular Biology, July 22, 2010.

Several figures in this dissertation are reproduced with permission from other sources, listed below. The written permissions obtained from the publisher or author are included in the pages that follow.

Figure 1.2 was reprinted from Lovell, S. C., I. W. Davis, W. B. Adrendall, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. 2003. Structure validation by C alpha geometry: phi,psi and C beta deviation. Proteins-Structure Function and Genetics 50:437-450, with permission from John Wiley and Sons.

Figure 1.8 was reprinted from Lee, A. 2008. Geometric constraint systems with applications in CAD and biology. University of Massachusetts Amherst. xvi, 156 p.

Figures 1.9, 1.11, and 1.12 were reprinted from Wells, S., S. Menor, B. Hespenheide, and M. F. Thorpe. 2005. Constrained geometric simulation of diffusive motion in proteins. Physical Biology 2:S127-S136, with permission from IOP Publishing and M. F. Thorpe.

Figure 3.1 was reprinted from Lei, M., J. Velos, A. Gardino, A. Kivenson, M. Karplus, and D. Kern. 2009. Segmented Transition Pathway of the Signaling Protein Nitrogen Regulatory Protein C. Journal of Molecular Biology 392:823-836, with permission from Elsevier.

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Aug 04, 2010

This is a License Agreement between Daniel W Farrell ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2482191459161
License date	Aug 04, 2010
Licensed content publisher	John Wiley and Sons
Licensed content publication	Proteins: Structure, Function and Bioinformatics
Licensed content title	Structure validation by Cα geometry: &phis;,ψ and Cβ deviation
Licensed content author	Lovell Simon C., Davis Ian W., III W. Bryan Arendall, et al
Licensed content date	Jan 8, 2003
Start page	437
End page	450
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Original Wiley figure/table number(s)	Figure 4A
Will you be translating?	No
Order reference number	
Total	0.00 USD
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one if its group companies (each a "Wiley Company") or a society for whom a Wiley Company has exclusive publishing rights in relation to a particular journal (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your Rightslink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

1. The materials you have requested permission to reproduce (the "Materials") are protected by copyright.

2. You are hereby granted a personal, non-exclusive, non-sublicensable, non-transferable, worldwide, limited license to reproduce the Materials for the purpose specified in the licensing process. This license is for a one-time use only with a maximum distribution equal to the number that you identified in the licensing process. Any form of republication granted by this licence must be completed within two years of the date of the grant of this licence (although copies prepared before may be distributed thereafter). Any electronic posting of the Materials is limited to one year from the date permission is granted and is on the condition that a link is placed to the journal homepage on Wiley's online journals publication platform at www.interscience.wiley.com. The Materials shall not be used in any other manner or for any other purpose. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher and on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Material. Any third party material is expressly excluded from this permission.

3. With respect to the Materials, all rights are reserved. No part of the Materials may be copied, modified, adapted, translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Materials without the prior permission of the respective copyright owner. You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Materials, or any of the rights granted to you hereunder to any other person.

4. The Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc or one of its related companies (WILEY) or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto.

5. WILEY DOES NOT MAKE ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND WAIVED BY YOU.

6. WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

7. You shall indemnify, defend and hold harmless WILEY, its directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

8. IN NO EVENT SHALL WILEY BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

9. Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

10. The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

11. This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

12. These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in a writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

13. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

14. WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

15. This Agreement shall be governed by and construed in accordance with the laws of England and you agree to submit to the exclusive jurisdiction of the English courts.

16. Other Terms and Conditions:

BY CLICKING ON THE "I ACCEPT" BUTTON, YOU ACKNOWLEDGE THAT YOU HAVE READ AND FULLY UNDERSTAND EACH OF THE SECTIONS OF AND PROVISIONS SET FORTH IN THIS AGREEMENT AND THAT YOU ARE IN AGREEMENT WITH AND ARE WILLING TO ACCEPT ALL OF YOUR OBLIGATIONS AS SET FORTH IN THIS AGREEMENT.

V1.2

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK10826740.
Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

If you find copyrighted material related to this license will not be used and wish to cancel, please contact us referencing this license number 2482191459161 and noting the reason for cancellation.

Questions? customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

(The following is an excerpt from an email sent by Audrey Lee)

Audrey Lee-St. John <astjohn@mtholyoke.edu> Wed, Aug 4, 2010 at 1:31 PM
To: Daniel Farrell <dwf@asu.edu>

Hi Dan,

...
You are welcome to use the figure from my dissertation.
...

-Audrey

Audrey Lee St. John
Visiting Assistant Professor
Computer Science Department
Mount Holyoke College
South Hadley, MA 01075
USA

http://minerva.cs.mtholyoke.edu

Office: Clapp 200
Email: astjohn@mtholyoke.edu

To: permissions@iop.org,
Cc:
Bcc:
Subject: reproducing figures in a dissertation
From: Daniel Farrell <dwf@asu.edu> - Saturday 07/08/2010 22:17

I am writing a PhD dissertation at Arizona State University entitled Generating Stereochemically Acceptable Conformational Change Pathways in Proteins. I would like permission to reproduce and include in my dissertation the following figures from an IOP journal:

✓ Figs. 1 and 3 from
Wells, S., S. Menor, B. Hespenheide, and M. F. Thorpe. 2005. Constrained geometric simulation of diffusive motion in proteins. Physical Biology 2:S127-S136.

Thank you,
Dan Farrell

PERMISSION TO REPRODUCE AS REQUESTED
IS GIVEN PROVIDED THAT:

(a) the consent of the author(s) is obtained
(b) the source of the material including author/editor, title, date and publisher is acknowledged.

IOP Publishing Ltd
Dirac House
Temple Back
BRISTOL
BS1 6BE

9/8/2010 .....
Date Rights & Permissions

To: Permissions <permissions@iop.org>,
Cc:
Bcc:
Subject: requesting permission to reproduce figure
From: Daniel Farrell <dwf@asu.edu> - Friday 10/09/2010 05:38

I am writing a PhD dissertation at Arizona State University entitled Generating Stereochemically Acceptable Conformational Change Pathways in Proteins. I would like permission to reproduce and include in my dissertation the following figures from an IOP journal:

✓ Fig. 8 from
Wells, S., S. Menor, B. Hespenheide, and M. F. Thorpe. 2005. Constrained geometric simulation of diffusive motion in proteins. Physical Biology 2:S127-S136.

This is in addition to a previous request by me (already granted) to reproduce Figs. 1 and 3 from the same work.

Thank you,
Dan Farrell

PERMISSION TO REPRODUCE AS REQUESTED
IS GIVEN PROVIDED THAT:

(a) the consent of the author(s) is obtained
(b) the source of the material including author/editor, title, date and publisher is acknowledged.

IOP Publishing Ltd
Dirac House
Temple Back
BRISTOL
.....
BS1 6BE

10/9/2010 .....
Date Rights & Permissions

Michael Thorpe <Michael.Thorpe@asu.edu> Fri, Sep 10, 2010 at 10:12 AM
To: Daniel Farrell <dwf@asu.edu>
Dan:

It is my pleasure to give you permission to reproduce the figures you list below.

Mike

-- Mike Thorpe

Leverhulme Visiting Professor (2009-2011)
From May 15 until Nov 15, 2010
Rudolf Peierls Centre for Theoretical Physics, room 4.1
1 Keble Road, Oxford OX1 3N
Office phone: +44 (0) 1865 273975
Home phone: +44 (0) 1865 425946
Cell phone: +44 (0) 7910 788621

Foundation Professor of Physics, Chemistry & Biochemistry
Bateman Physical Sciences PSF 359
Arizona State University, Tempe, AZ 85287−1504
Office phone: +1 480.965.3085
Home phone: +1 480.491.2549
Cell phone: +1 602.463.1042

> -----Original Message-----
> From: Daniel Farrell [mailto:dwf@asu.edu]
> Sent: Friday, September 10, 2010 5:42 AM
> To: Michael F Thorpe
> Subject: permission to reproduce figures
>
> Mike,
>
> Can you grant me permission to reproduce Figs. 1, 3, and 8 from
>
> Wells, S., S. Menor, B. Hespenheide, and M. F. Thorpe. 2005.
> Constrained geometric simulation of diffusive motion in proteins.
> Physical Biology 2:S127-S136.
>
> for use in my dissertation? I have obtained permission from IOP, but
> they also require that I obtain permission from one of the authors. I
> will include your response in an appendix in my dissertation.
>
> Thank you,
> Dan Farrell

ELSEVIER LICENSE
TERMS AND CONDITIONS

Sep 15, 2010

This is a License Agreement between Daniel W Farrell ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Daniel W Farrell
Customer address	PSF 470 Tempe, AZ 85287-1504
License number	2510020634725
License date	Sep 15, 2010
Licensed content publisher	Elsevier
Licensed content publication	Journal of Molecular Biology
Licensed content title	Segmented Transition Pathway of the Signaling Protein Nitrogen Regulatory Protein C
Licensed content author	Ming Lei, Janice Velos, Alexandra Gardino, Alexsandr Kivenson, Martin Karplus, Dorothee Kern
Licensed content date	25 September 2009
Licensed content volume number	392
Licensed content issue number	3
Number of pages	14
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Generating Stereochemically Acceptable Protein Pathways
Expected completion date	Sep 2010

Estimated size (number of pages)	130
Elsevier VAT number	GB 494 6272 12
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and

conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article. If this license is to re-use 1 or 2 figures then permission is granted for non-exclusive world rights in all languages.

16. **Website**: The following terms and conditions apply to electronic reserve and author websites:
Electronic reserve: If licensed material is to be posted to website, the web site is to be password-protected and made available only to bona fide students registered on a relevant course if:
This license was made in connection with a course,
This permission is granted for 1 year only. You may obtain a license for future website posting,
All content posted to the web site must maintain the copyright information line on the bottom of each image,
A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com , and
Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

17. **Author website** for journals with the following additional clauses:

All content posted to the web site must maintain the copyright information line on the bottom of each image, and
he permission granted is limited to the personal version of your paper. You are not allowed to download and post the published electronic version of your article (whether PDF or HTML, proof or final version), nor may you scan the printed edition to create an electronic version,
A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx , As part of our normal production process, you will receive an e-mail notice when your article appears on Elsevier's online service ScienceDirect (www.sciencedirect.com). That e-mail will include the article's Digital Object Identifier (DOI). This number provides the electronic link to the published article and should be included in the posting of your personal version. We ask that you wait until you receive this e-mail and have the DOI to do any posting.
Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

18. **Author website** for books with the following additional clauses:
Authors are permitted to place a brief summary of their work online only.
A hyper-text must be included to the Elsevier homepage at http://www.elsevier.com

All content posted to the web site must maintain the copyright information line on the bottom of each image
You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version.
Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

19. **Website** (regular and for author): A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx. or for books to the Elsevier homepage at http://www.elsevier.com

20. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

21. **Other Conditions**:

v1.6

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK10848756.
Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

If you find copyrighted material related to this license will not be used and wish to cancel, please contact us referencing this license number 2510020634725 and noting the reason for cancellation.

Questions? customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.