This is to certify that the
dissertation entitled

Modeling Non-Crystalline Networks

presented by

Ming Lei

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Physics

M. F. Thorpe
Major Professor's Signature

08-05-2003
Date

MODELING NON-CRYSTALLINE NETWORKS

By

MING LEI

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Physics and Astronomy

2003

ABSTRACT

MODELING NON-CRYSTALLINE NETWORKS

By

MING LEI

In this thesis, the author reports the modeling of both the static and the dynamical aspects of non-crystalline networks. Porous silicon and silica have attracted attention recently due to their unusual photoelectronic properties. Porosity is central to these striking properties which are not present in non-porous silicon and silica. We propose an algorithm that is effective in building fully-coordinated amorphous networks that are discontinuous in certain regions – that is, they contain large voids of mesoscopic or macroscopic dimensions. Such networks can be both porous and amorphous, and can also be finite in certain dimensions.

Voids of arbitrary shapes and sizes are first superimposed on a crystalline silicon network. The atoms in the pore regions are removed. Local "defects" are created, then eliminated, as pairs of them are brought together by a defect migration process. The network is fully coordinated after the defect migration process. The Wooten Winer Weaire (WWW) algorithm, is then applied to make the network amorphous. Oxygen is inserted on every silicon-silicon bond to create a porous silica network. Silica networks in the form of an amorphous fiber and an amorphous film are created by this procedure. Distortions due to surface effects are investigated. The local atomic arrangement in these discontinuous networks is similar to that in bulk amorphous silica.

Covalent bond lengths and angles in amorphous networks do not vary much because of the high energies associated with bond length and angle distortions. Therefore, they can be viewed as constraints that do not change with time in any significant way. Proteins, viewed as another type of non-crystalline network, are glued together

by covalent bonds, hydrogen bonds, hydrophobic interactions, and other interactions. The concentration of constraints in some regions of the proteins are so high that these regions are rigid. The other regions are flexible. The flexible regions of protein can exhibit large conformational changes. Protein functions and bio-activities are often coupled with these conformational changes.

We have built an algorithm that samples protein conformations randomly. It is called Rigidity Optimized Conformational Kinetics (ROCK). It is efficient, as it avoids sampling conformations for the rigid regions of the proteins. The constraints in the flexible regions of proteins inter-lock with each other to form complicated networks of rings. ROCK closes all the rings simultaneously at every step of sampling the protein conformations. It is the first algorithm that samples the protein conformations by following the closure of all the rings in a complicated network. All the bond length and angle constraints are exactly preserved in the conformations sampled by ROCK. Main chain dihedral angles are restricted in the preferred regions of the Ramachandran plot. The generated conformations have good stereo-chemistry properties.

ROCK samples a large scale conformational changes. Its capability is first demonstrated on a model molecule with two degrees of freedom. The conformations sampled by ROCK observes the same two symmetries which are present in the topology of the molecule. A large scale motion is shown in the conformations of HIV-1 protease sampled by ROCK. ROCK also samples conformational pathways between distinct conformations of proteins. Multiple conformational trajectories are explored by ROCK between the closed, the occluded, and the open conformations of DHFR. Since ROCK explores both the main chain flexibility and the side chain flexibility, it is a good tool in the studies of protein-ligand interactions, ligand design, protein motions etc.

To,

My wife Jing,

My parents,

My sister,

And

My mother in law

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude toward my adviser, Prof. Michael F. Thorpe, for his insightful guidance, invaluable encouragement and countless support. I did not have much research experience when I started my Ph.D. studies. It was under the direction of Mike that I finished the first research project in my life. Challenging and smart questions raised by Mike taught me how to check the results of research from a variety of ways, and how to ask the right questions to obtain the most interesting results. Mike is an exceptional expert on explaining complicated phenomenon by minimalist models. His teaching is invaluable to my future work. Mike gave me many suggestions on this thesis.

I am much obligated to Prof. Leslie A. Kuhn for her endless help in my graduate study. Leslie taught me how to do research systematically through her patient and friendly teaching and communications. I learned how to appreciate the charming characteristics of individual proteins from Leslie.

I am much obliged to Dr. Maria Zavodszky for her remarkable help in my research. It has always been a great pleasure to discuss a wide range of scientific topics with her. Candid feedback and innovative suggestions from Leslie and Maria are the key factors behind all the improvements in the program ROCK. Maria patiently and industriously scrutinized the manual of ROCK, and the Appendix and Acknowledgments of this thesis.

I am indebted to Prof. A. Roy Day of Marquette University for close collaboration and numerous discussions on the flexibility of small molecules.

I appreciate the generosity of Dr. Maria Kurnikova of University of Pittsburgh and Prof. Robert Cukier very much in sharing their molecular dynamic simulation data with me. Although these simulations are not discussed or cited in this dissertation, they greatly helped my understanding of the protein conformations. I thank very

much Prof. Petkov of Central Michigan University and Prof. Simon J. L. Billinge for their generous sharing of the synchrotron diffraction experimental data with me. I also thank very much Dr. Michael Wachhold, Dr. Kashturi Rangan and Prof. Kanatzidis for their sharing of the knowledge on adamantane with me.

I thank very much Prof. Walter Whiteley of York University, Prof. Normand Mousseau of University of Toronto, Prof. Chien-Peng Yuan, Prof. S. D. Mahanti and Dr. Claire Vieille for sharing their thoughts with me on the mathematics of rigidity analysis, on the protein dynamical trajectories, on the validity of the geometry aspects in proteins, and on the protein conformations in dihedral angle space. I also thank them a lot for their encouragement and help in various matters throughout my graduate career.

I benefited a lot from the discussions with former graduate students in Prof. Thorpe's and in Prof. Kuhn's groups: Dr. Brandon Hespenheide, Dr. Andrew John Rader, Dr. Mykyta Chubynsky and Mr. Valentin Levashov. We spent a lot of time discussing the geometry and biochemical aspects of constraints and rigidity in proteins. Dr. Brandon Hespenheide and Dr. Andrew John Rader helped me a lot on the usage of the program FIRST.

At last I would like to extend my thanks to my wife, my parents, my sister and my mother in law. I would not have been able to finish this challenging journey of the Ph.D. study without their support and understanding.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

IMAGES IN THIS THESIS ARE PRESENTED IN COLOR

LIST OF ABBREVIATIONS

2D ..... Two Dimension

3D ..... Three Dimension

CRN ..... Continuous Random Network

DARMSD ..... Dihedral Angle Root Mean Square Deviation

DCRN ..... Dis-Continuous Random Network

DDF ..... Density Distribution Function

DHF 7,8-dihydrofolate

DHFR ..... Dihydrofolate Reductase

DOF ..... Degrees of Freedom

ecDHFR ..... *Escherichia coli* Dihydrofolate Reductase

HIV ..... Human Immunodeficiency Virus

MC ..... Monte Carlo

MD ..... Molecular Dynamics

MM ..... Molecular Mechanics

NADPH ..... nicotinamide adenine dinucleotide phosphate

PDB ..... Protein Data Bank

PDF ..... Pair Distribution Function

QM ..... Quantum Mechanics

RCE ..... Ring Closure Equations

RDDF ..... Reduced Density Distribution Function

RDF ..... Radial Distribution Function

RMSD ..... Root Mean Square Deviation

RPDF ..... Reduced Pair Distribution Function

THF 5,6,7,8-tetrahydrofolate

WWW ..... Wooten Winer Weaire

Chapter 1: Introduction

Physics and geometry are "a marriage made in heaven", as said by Sir Michael Atiyah in his talk [1]. Geometry, which is a branch of mathematics, has been extensively utilized in the study of the physical world. Copernicus, for example, put forward the heliocentric theory to replace the geocentric theory after pondering on the geometrical properties of the observed planet orbits. Navigations from the middle age till today all rely on calculations using geometry. Fractal geometry, which is a newly invented branch in geometry, is the mathematical language of chaotic systems. Geometry and mathematics are of the most important tools in theoretical physics studies.

Complicated systems can of be too modeled using a few simple geometrical principles. Graner [2] nicely reviewed comprehensive aspects of building fluid foam models from geometrical considerations. Arns et al. [3] discuss how to build disordered material models whose micro-structures obey particular geometrical requirements. Free energy of the Langmuir mono-layer is examined by Lösche and his co-worker [4] based on an empirical Hamiltonian that accounts for the geometrical arrangement of micro-structures of the layers. Many more examples can be listed. All of them address certain properties of complicated materials from simple yet adequate geometrical thinking.

This thesis shows two more examples of the application of geometry in condensed matter physics. The first example is on the building of non-crystalline network models with geometrical restrictions on bond length and angles. The second example is about sampling conformations of non-crystalline networks. Conformations are spatial arrangements of atoms whose bond lengths and angles are correct. Both of these two examples show how to build complicated non-crystalline network models from geometrical rules.

This chapter is an introduction of the thesis. It explains the motivations of our research work. Chapter 2 discusses the algorithm to build non-crystalline networks, the results of which are examined in Chapter 3. The similarities and differences between glassy networks and proteins are investigated in Chapter 4. Though they share common characteristics, amorphous networks and proteins differ in that the former do not have multiple dynamical conformations yet the latter do. A new approach to sample protein conformations is elucidated in Chapter 5. Two applications of the algorithm are shown in Chapter 6. Chapter 7 summarizes this thesis and re-iterate how simple geometrical considerations can lead to complex non-crystalline network models.

1.1 Static Models of Non-Crystalline Networks

Short range structures of amorphous networks are very similar to those in the corresponding crystalline networks. For example, the bond lengths and angles of silicon atoms in the amorphous silicon networks are almost the same as those in the crystalline silicon networks. Those properties that rely on the short range structures, such as the vacuum UV absorption spectra and the electronic density of states curves, differ only in detail [5] between the amorphous and the crystalline networks. On the other hand, amorphous networks are remarkably different from crystalline networks in the characteristics related to the medium to long range order. For example the X-ray or neutron diffraction of amorphous materials detect broad peaks while the diffractions from crystalline materials show many fine sharp peaks. The modeling of amorphous networks has long been an interesting topic in condensed matter physics.

1.1.1 Continuous Random Network Models

In the early 1930s, it was hypothesized that amorphous material are composed of numerous micro-crystals. The orientations of the micro-crystals are not aligned so that the long range order of the crystals is destroyed. The X-ray powder diffraction pattern of the models built upon this hypothesis does not fit with the experimental data. Zachariasen [6] proposed that the amorphous material is not made up of micro-crystals. The atomic arrangement in an amorphous network does not have symmetry and periodicity. However Zachariasen did not propose an algorithm to build a continuous three-dimensional network which lacks symmetry and periodicity.

An amorphous network model that satisfies Zachariasen's criteria is built by hand by Bell and Dean [7, 8]. It is an amorphous silica model with 614 atoms. A hand built amorphous silicon model containing 440 atoms was reported by Polk [9] in 1971. Atomic coordinates in that model were detected by laser beams and saved into a computer for analysis [10]. These hand built models have free surfaces. The sizes of these models are limited in the order of tens of angstroms. Moreover, since a large portion of the atoms are at or close to the surfaces, the number of atoms that can be used to analyze the bulk properties of amorphous material is not large. Henderson [11] built a periodic amorphous network model by hand. There are only 61 atoms in the supercell of the periodic model.

Computational modeling is the area that thrived since the early attempt [12] when a periodic amorphous model with 54 atoms in the supercell was generated by computer. The scalability problem hindered further development of the algorithm behind the model. Guttman [13, 14] built amorphous network models in computers by linking atoms randomly. A subsequent relaxation procedure reduces the distortions of bond lengths and angles. Bonds are allowed to be switched in the relaxation procedure. The most successful algorithm today in building amorphous network models is the WWW technique proposed by Wooten, Winer, and Weaire [15]. The details

of the algorithm are discussed in Chapter 2. Several other algorithms are suggested thereafter, all of which are modifications of the original WWW technique. For example the improvement by Barkema and Mousseau [16] makes the computational modeling of device sized amorphous silicon models [17] with more than 10,000 atoms possible.

The amorphous network models built by the WWW and other similar techniques are called Continuous Random Network (CRN) models. By continuous it is meant that the networks are infinite, without disruption in the distribution of atoms. By random it is meant that the topology of an amorphous network model is different to that in the corresponding crystalline network. The CRN models built by the WWW algorithm match well with real amorphous silicon in terms of properties of electronic states, X-ray diffraction and the pair distribution function (PDF). The success of the WWW algorithm comes from the two geometrical principles in building CRN networks that are the essence of amorphous networks. Atoms are maintained as fully-coordinated. That is, all silicon and germanium atoms have four and exactly four nearest neighbors, while all oxygen, sulfur and selenium atoms have two and exactly two nearest neighbors. This principle originates from the preferred valences of these elements in semi-conductors. The topology of a CRN model is different from that of a crystalline network model. Randomly positioning atoms in a supercell will not create a CRN model. The WWW technique is a trustworthy method in creating a network whose topology is totally different from that of a crystalline network. The high quality models built by the WWW algorithm prove how geometrical approaches benefit the modeling of non-crystalline networks. As an example, we show in Section 3.4 how to model the amorphous metal-adamantane network starting from an amorphous gallium arsenide model. The powder diffraction pattern calculated from the model matches well with that measured in experiment.

Molecular dynamics (MD) simulations coupled with empirical or semi-empirical

potentials have been used to build amorphous silicon network models since 1985. Several empirical potentials have been invented and parameterized for the simulation of amorphous silicon. The SW potential by Stillinger and Weber [18], the potential invented by Biswas and Hamann [19], and the potential by Chelikowsky [20] are all composed of a two-body interaction part and a three-body interaction part, the first of which depicts the bond lengths vibrations while the latter of which describes the bond bending vibrations. Though the potential of Tersoff [21] has only pair wise interactions, the parameters in the interaction depend on the bond angles as well. Therefore the three-body interaction is implicitly calculated in the Tersoff's potential. The semi-empirical potential by Baskes et al. [22] is more complicated in form. It is supposed to be in good agreement with first principle calculations. The SW potential is most widely used in MD simulations. Despite much effort, the empirical and the semi-empirical potentials are not accurate on all phases of silicon. Furthermore, the connectivity is not guaranteed. There are dangling bonds in the amorphous network models built by MD simulations. Therefore MD is not the best way to generate amorphous models as of today. Car and Parrinello [23] generated a small amorphous silicon model of 54 atoms by first principle quantum mechanical calculations. The calculation cost however forbids further application of such algorithms at present, and the small size means that these models have serious strains due to the periodic boundary conditions.

1.1.2 Discontinuous Networks

Discontinuities in atom distributions lead to intriguing and unexpected properties that are not present in materials without discontinuities.

Porous silicon has application potential in harvesting solar energy. Canham et al. [24] report that porous silicon is photo-luminescent. The average diameter of the pores is about 13nm. The thickness of the silicon layers between the pores is on the

order of μm, measured by X-ray diffraction experiments [25]. Quantum confinement effects as well as the altered gaps between electronic states [26] are the causes of the photoluminescence phenomenon. The silicon layers between the pores are largely crystalline [27] rather than amorphous.

Porous silicon is bio-compatible and bio-degradable [28]. The body does not reject organs made of porous silicon. Porous silicon has the potential to be the platform of future biomedical implants and artificial organs.

Porous silica films are easy to fabricate. They have been used as chemical sensors and sources of photoluminescence. Zhao et al. [29] produced silica films whose porosities are between 51% and 75%. McDonagh et al. [30] applied sol-gel porous silica films to sensor the oxygen. Both Cohen and his co-workers [31] and Dag et al. [32] observe bright photoluminescence from the porous silica films containing nanoclusters of silicon. Amorphous silica films, though not porous, are photo-luminescent as well. Yoshida et al. [33] have found blue photoluminescence from slightly silicon doped amorphous silica films. The origin of the photoluminescence is believed to be in the silicon nanocrystals.

The structures and surface properties of amorphous silica films have been studied by a variety of techniques including electron diffraction [34], infrared spectroscopy [35], scanning reflection electron microscopy [36], Raman spectroscopy [37], and NMR [38] et al. *Ab initio* simulations [39, 40], MD simulations [41, 42], and Monte Carlo (MC) simulations [43] have all been used to study either amorphous silica films or the surface properties of amorphous silica.

All of these materials mentioned above are not continuous in the traditional sense in that they are not microscopically homogeneous. Porous silicon and silica contain virtually periodic voids that break the uniform distribution of the atoms over space. Amorphous silica films are not continuous because of the discontinuity of atom distributions over the surfaces.

Such discontinuities result in exciting and new material properties. Bulk silicon and silica, either crystalline or amorphous, are not photo-luminescent. On the other hand, porous silicon, porous silica and thin silica films all show photoluminescence effects. Though the origin of photoluminescence in these materials is currently being debated, it is almost certain that photo-luminescent characteristics in these materials involve the discontinuity in the spatial arrangement of atoms.

Though algorithms to build CRN models such as the WWW technique have been available for a long time, there is not yet a generic algorithm for building discontinuous random network (DCRN) models. Though the endeavor of building DCRN models may seem to be unnecessary at the first glance because such materials as *amorphous porous silicon* and *amorphous porous silica* are not much discussed in literatures yet, the author argues that these materials are not far fetched from being manufactured, considering the facts that 1) porous silicon and silica are easy to fabricate and 2) amorphous silicon and silica are stable. Since the porosity in crystalline silicon and silica leads to new properties, the porosity in amorphous silicon and silica is likely to bring exciting properties as well. The amorphous porous silicon models provide the first glimpse of the likely structural properties of such materials. Though not have been manufactured, the amorphous silicon film has been computationally modeled by Monte Carlo simulations using empirical potential [44] and by *ab initio* simulations [45]. The properties of amorphous silica films have also been examined, as described above.

Local bond geometries in the DCRN models should be similar to those in the crystalline networks. This requirement is the natural result of the strongly covalent characteristics of the glass forming elements such as silicon, germanium, oxygen, sulfur and selenium. The geometrical concepts which are the roots of the WWW algorithm also serve as the foundations of our algorithm to build the DCRN models. Step by step, Chapter 2 reveals the methodology to build the DCRN models.

1.2 Constraints and Flexibility Analysis

The empirical Keating potential is used in building DCRN models. The Keating potential reaches its minimum values of zero when bond lengths and angles are of their optimal values. The potential energy can be huge when distortions in bond lengths and angles are large. When it costs an infinite amount of energy to distort bond lengths and angles, every bond length and angle requirement is called a *constraint.*

At finite temperatures the atoms in non-crystalline networks are in constant motions, due to the thermal fluctuation energy of k_BT, in which k_B and T are the Boltzmann constant and the temperature respectively. The thermal fluctuation energy pushes the atoms so that they oscillate around the local potential minima. These oscillatory motions do not change the averaged relative orientation between atoms, not to mention the overall shapes of the networks. On the other hand, some non-crystalline networks have predominantly internal motions. Proteins are such examples. The scale of the protein internal motions is large, for example the relative distance between atoms in different conformations of HIV-1 protease can vary between 2.7Å and 8.0Å, shown in Section 6.2 in this thesis. These large scale motions do not result from the thermal fluctuations of bond lengths and angles. Rather they are caused by large thermal and ligand-induced fluctuations of the internal degrees of freedom (DOF) in the network.

The concept of constraints simplifies the analysis of the internal large scale motions of the non-crystalline networks. When bond lengths and angles are treated as constraints, the DOF of a network is simply the difference between the total number of degrees of freedom and the total number of independent constraints, as explained by Maxwell in 1864 [46]. The question of whether a non-crystalline network has large scale internal motions is answered by the counting of the DOF. A positive DOF in a network is correlated with the likelihood of large scale motions. Moreover, a network shows large scale motions without breaking any constraints when it undergoes motion

by sampling the DOF.

However the Maxwell counting is not exact. A procedure called *rigidity analysis* was first used by Jacobs et al. to count constraints in proteins, based on the pebble game algorithm [47, 48]. Since what we care about here is the flexibility properties of networks, the author uses the phrase *flexibility analysis* instead of rigidity analysis throughout this thesis. When only bond length constraints are counted, or when both bond length and bond angle constraints are counted, the DOF calculated by this procedure is exact for generic networks in 2D. Flexibility analysis is not exact for generic networks in 3D when only bond length constraints are included. However under usual conditions, though this has not been proven rigorously, its application to 3D networks is exact when both bond length and bond angle constraints are counted [49, 50, 51].

Flexibility analysis identifies the regions in generic networks that have positive numbers of DOF. These regions can have large scale internal motions that allow big changes in the relative orientations between atoms and the overall shapes of the networks. They are called the flexible regions. The other regions do not have large scale motions, hence they are called the rigid regions. Negative DOF are associated with over-constrained, or stressed regions of the network. Chapter 4 accounts for the details of the flexibility analysis.

1.3 Sampling Conformations

It is one matter to distinguish the flexible regions from the rigid regions in networks, yet another topic to search the spatial arrangements of atoms under which the bond lengths and angles obey all constraints. A spatial arrangement of atoms is called a *conformation* if bond lengths and angles in such an atomic arrangement observe the predefined constraints. In usual cases, one conformation is already obtained from modeling, from X-ray or neutron diffraction experiments, or from NMR

experiments. The problem is to find other conformations that satisfy the same set of constraints as the original one does.

1.3.1 Conformations of Proteins

Proteins, being non-crystalline and finite networks from a topological point of view, have one or more rigid regions and several flexible regions. The rigid regions of proteins serve as the stabilizing cores. At least one flexible region is typically close to the catalytic site in every protein. The catalytic functions of proteins are always coupled with conformational changes, either involving the side chain atoms or involving both the side chain and the main chain atoms. The main chain and the side chain of proteins are discussed in detail in Section 4.4.1. The flexible regions of proteins either uncover the functional sites, or make specific interactions with the substrates, or escort the substrates to and from the function sites, or directly assist the catalytic functions, through large scale conformational changes.

Adenylate kinase (ADK), for example, has large scale domain motions related to its catalytic cycle. Berry et al. [52] resolved the X-ray crystal structure of the protein in its ligand-binding conformation, which is called the "closed" conformation. The large flexible lid domain covers one of the bound substrates in one such conformation. The conformation of the lid domain is completely changed when there is no ligand bound to the protein [53, 54]. The conformational transitions of ADK are believed to be driven by the rotation of several dihedral angles at a few hinges [55]. Calmodulin, which regulates a variety of cellular processes, is another example of a protein that goes through large scale conformational changes in its catalytic pathway. Its binding of calcium initiates conformational changes, which in turn forces the other proteins that are bound to calmodulin to change their conformations also, resulting in an activation of certain biological functions of the cells [56, 57]. Zhang et al. [58] report the coupling of conformational changes with the electron transfer function of the

protein cytochrome bc_1.

1.3.2 Algorithms to Sample Protein Conformations

Proteins have to change their conformations to achieve certain biochemical functions. It is the ability to sample different conformations with functional significance and to carry out chemistry on their molecular particles that distinguishes proteins from other non-crystalline networks. Unfortunately, there is no way to detect all protein conformations from experiments. X-ray crystallography experiments identify one or several conformations of the same protein. NMR techniques reveal the most populated conformations, but they cannot detect the less populated ones. Computational modeling is indispensable in studying protein conformational transitions. Quite naturally, there have already been numerous such studies on sampling protein conformations.

Some algorithms on sampling protein conformations use databases of existing peptide conformations. The database used by Deane et al. [59] in their PETRA program stores low energy conformations of short peptide segments up to twelve residues long. The conformations are calculated by *ab initio* methods. Short peptide segments in the loop region of the proteins are then replaced by the conformations of the same or different segments in the database. To reduce the total calculation cost, filtering on several easily calculated criteria is first done to rule out the most impossible conformations. The empirical potential of the replacement peptide segment is then calculated and minimized. The application of algorithm is limited to the flexible loops on the surfaces of proteins.

Another category of algorithms samples protein conformations on the grid of dihedral angles [60, 61] systematically. Such algorithms are more appropriate for small molecules rather than on proteins, because the number of grid points to check grows exponentially with size.

Protein conformations can be sampled by MD simulations, in which protein motion trajectories are calculated by integrating Newton's equations. Every snapshot of the motion trajectory is a viable protein conformation. Recent developments in MD simulations are reviewed by Wang et al. [62]. The time step within MD simulations is typically one or two femto-seconds. Integration of Newton's equations is carried out at every time step. The MD simulations currently can reach one or two nanoseconds of real motion trajectories. Some fast protein conformational changes which are in the nanosecond time range can hence be simulated by MD algorithms. The slow protein conformational changes which are in the milliseconds to seconds time range are still out of the reach of the MD simulations.

Clever techniques have been presented to make the MD calculations run faster. In the multicanonical MD simulations [63, 64], the possibility distribution of sampling conformations at different energies is artificially flattened [65], so that the probability of jumping over energy barriers is enhanced. Multiple MD trajectories are sampled in parallel at different temperatures in the replica exchange MD algorithm [66, 67]. Trajectories at different temperatures are periodically exchanged so that they all have chances to overcome high energy barriers at high temperatures. Where there is a MD technique, there is a corresponding MC method. The multicanonical MC method [68, 69, 70, 71] and the replica-exchange MC algorithm [72] have all been applied on the studies of proteins.

1.3.3 Algorithms to Close Rings

Sampling conformations by MD or by MC algorithms is equivalent to exploring local minima on the rough and complicated energy surfaces which are constructed by the intricate potential functions. These potentials have numerous local minima and energy barriers. One way to sample the protein conformations is to follow the saddle points between the local minima [167]. In some studies, when to obtain an ensemble

of conformations that are as diverse as possible is sufficient, it is not necessary to know all the trivial peaks and troughs in the energy landscape. As discussed in Section 1.2, flexible regions of proteins have multiple conformations even in the most simplified potential, which is infinitely high or absolutely zero, depending on whether bond lengths and angles are distorted. Therefore it is very attractive to sample protein conformations obeying all bond lengths and angle constraints. The difficulty in sampling conformations in this way is to close all the rings in the proteins at every step otherwise the bond lengths and angles are distorted. A ring is a closed loop of bonds which connect atoms. There are two paths connecting any two atoms in a ring. Geometry solves the biochemical problem by providing algorithms to close the rings.

The ring closure equations (RCE) proposed by Gō and Scheraga [73] state the conditions under which a ring is closed when the six unknown dihedral angles in the ring are consecutively positioned, or when they are only separated by locked peptide bonds. A peptide bond is the bond between two amino acid residues in proteins. The ideal dihedral angle of a peptide bond is either 0° or 180°. Gō and Scheraga elucidate a procedure to convert the RCE, which are six independent equations, to a single variable equation with a single unknown dihedral angle. A subsequent work [74] exhausts the conformational space of a short cyclic peptide segment by following the solutions to the RCE. Gō and Scheraga [75] later developed a method to close the big rings when the rings have C_n, I, or S_{2n} symmetry. The conformations of gramicidin S, which is a cyclic molecule with 18 rotatable bonds, are sampled by this procedure [76], assuming the molecule has an exact C_2 symmetry. The conformations of the molecule *cyclo*-hexaglycyl under symmetries were also generated [77] and checked against conformations generated by MC algorithms [78].

Several algorithms have been invented to close the rings since the pioneering work of Gō and Scheraga. Wedemeyer and Scheraga [79] discovered that a ring is closed when its dihedral angles are roots of polynomial equations. The form and the

solutions to the polynomial functions of seven-fold and eight-fold rings are developed in that article. Wu and Deem [80] use three distance and angle constraints at the break point of a ring. These constraints are then transformed to a polynomial function of a single variable which can be solved numerically. Unlike all the other algorithms that optimize all of the involved dihedral angles simultaneously, the cyclic coordinate descent algorithm by Canutescu and Dunbrack [81] optimizes them one at a time. Bruccoleri and Karplus [82] allow bond angles to be relaxed when a ring cannot be closed exactly under the condition of fixed bond angles.

All the algorithms listed above close only a single ring. In some studies, the whole protein main chain whose two ends are fixed in space is treated as a big ring. In some other studies, a short protein main chain segment is the ring to be closed. The main chain dihedral angles are varied systematically or randomly so that protein conformations are built or sampled [83, 84, 85, 86].

A protein has multiple interlocking rings which have to be closed exactly. Gibson and Scheraga [87] write the bond length and angle constraints at the disulfide bond as the pseudo-potential of a ring. The ring closes if the pseudo-potential is zero. They show several examples where three or four rings containing disulfide bonds are closed simultaneously. Chapter 4, Chapter 5 and Chapter 6 of this thesis present our approach to sample protein conformations by closing all the rings in a complicated network simultaneously. Our algorithm differs from the algorithm by Gibson and Scheraga in several aspects. First, the hydrogen bonds are treated as components of rings in our algorithm, but not in Gibson and Scheraga's algorithm. Second, the fictitious potential of a ring utilized in our algorithm is different to that in Gibson and Scheraga's algorithm. As discussed in detail in Section 5.2, the fictitious potential of a ring is defined to be the sum of the squares of the RCE in our algorithm. Gibson and Scheraga define the sum of the bond length and angle constraints at the disulfide bonds as the pseudo-potential of a ring. Third, our algorithm is able to handle a large

number of inter-locked rings.

Though hydrogen bonds and hydrophobic interactions are important in proteins, previous studies do not regard them as forming rings. Therefore in the research works mentioned above few rings are identified in proteins. The flexibility analysis [88, 89, 90] has demonstrated that the predicted protein structural properties match better with what is measured in experiments when strong hydrogen bonds and hydrophobic interactions are counted as real constraints. Therefore proteins should be viewed as networks composed of covalent bonds, strong hydrogen bonds and hydrophobic interactions. The definitions of strong hydrogen bonds and hydrophobic interactions are discussed in Chapter 4.

A large number of rings appear when hydrogen bonds are treated the same way as the covalent bonds. These rings inter-lock with each other in such complicated ways that the rotation of any single bond can break several rings instead of one. The algorithms designed to close a single ring are not capable of sampling conformations for these complicated networks, because they are not created to close all the rings simultaneously. Chapter 5 presents a new algorithm that is efficient in closing all the rings in a network. It is a powerful tool for sampling large scale protein conformations.

Chapter 2: Modeling Discontinuous Random Networks

As discussed in Chapter 1, porous silicon and silica networks have unique properties and application potentials. This chapter demonstrates an algorithm to build fully coordinated amorphous silicon and silica networks with any desired characteristics of discontinuity in atomic distributions.

Section 2.1 is a brief review of the standard WWW technique in building CRN models. Section 2.2 explains how to build DCRN models in 3D with voids of arbitrary shapes and sizes. Defects, such as dangling bonds and hydrogenated silicon atoms are inevitable in real amorphous silicon and silica material. Section 2.4 discusses how to include defects into the DCRN models. Examples of amorphous film and fiber silica models are shown in Chapter 3.

2.1 WWW Algorithms on CRN Models

Amorphous silicon networks are made up of five-fold, six-fold, seven-fold, eight-fold rings and a small trace of four-fold rings and bigger rings. Crystalline networks are composed exclusively of six-fold rings. All rings mentioned in this thesis are irreducible rings. An irreducible ring is such a ring that the shorter path between two atoms in the ring is one of the shortest paths among all possible paths in the whole network connecting these two atoms. The WWW algorithm introduces five-fold and seven-fold rings to a silicon network by a bond switching process. In each bond switch process, the WWW algorithm exchanges a pair of nearest neighbors of two randomly chosen bonded atoms. Figure 2.1(a) shows a crystalline honeycomb network in 2D. The atom A in the figure has three nearest neighbors of atom B, C and D. The three

nearest neighbors of atom B are atom A, E and F. By breaking two bonds of AB and BE and creating two new bonds of AE and BD, the WWW algorithm alters the topology of the network, as shown in Figure 2.1(b). The strain in the new network is relaxed so the lengths of new bonds become acceptable, as shown in Figure 2.1(c). Comparing the networks in Figure 2.1(a) and (c) shows that the WWW algorithm destroys four six-fold rings in the original network, followed by the creation of two five-fold rings and two seven-fold rings. The bond switch process works identically in 3D, though the number of six-fold rings eliminated is different than in 2D. After tens of thousands of such bond switch processes, the topology of the network does not have any residues of the topology of the original crystalline network. A subsequent simulated annealing process with even more bond switches reduces the total energy of bond length and angle variations. The resulting network has a totally different topology from that of the initial crystalline network. Bond length distortions and bond angle distortions in the networks are also low. For the sake of simplicity, the phrase *bond distortions* is used from here on to denote both bond length distortions and bond angle distortions.

The Keating potential [91] used in the relaxation process in the WWW technique is composed of two parts: a bond length variation part and a bond angle variation part

$$E = \frac{3}{16}\frac{\alpha}{R_0^2}\sum_i\sum_j \left(\mathbf{R}_{ij}^2 - R_0^2\right)^2 + \frac{3}{8}\frac{\beta}{R_0^2}\sum_i\sum_{j<k}\left(\mathbf{R}_{ij}\cdot\mathbf{R}_{ik} - \cos\theta_{jik}R_0^2\right)^2 \qquad (2.1)$$

in which α and β are parameters that control the relative ratio of the bond length variation and bond angle variation. The parameter R_0 is the equilibrium bond length between a pair of bonded atoms. This value is 1.62Å for a bond between silicon and oxygen. The parameter θ_{jik} is the optimal angle between the two vectors of $\mathbf{R}_{ij}$ and $\mathbf{R}_{ik}$. This angle is 109.5° at silicon atoms and 147° at oxygen atoms. The sum over

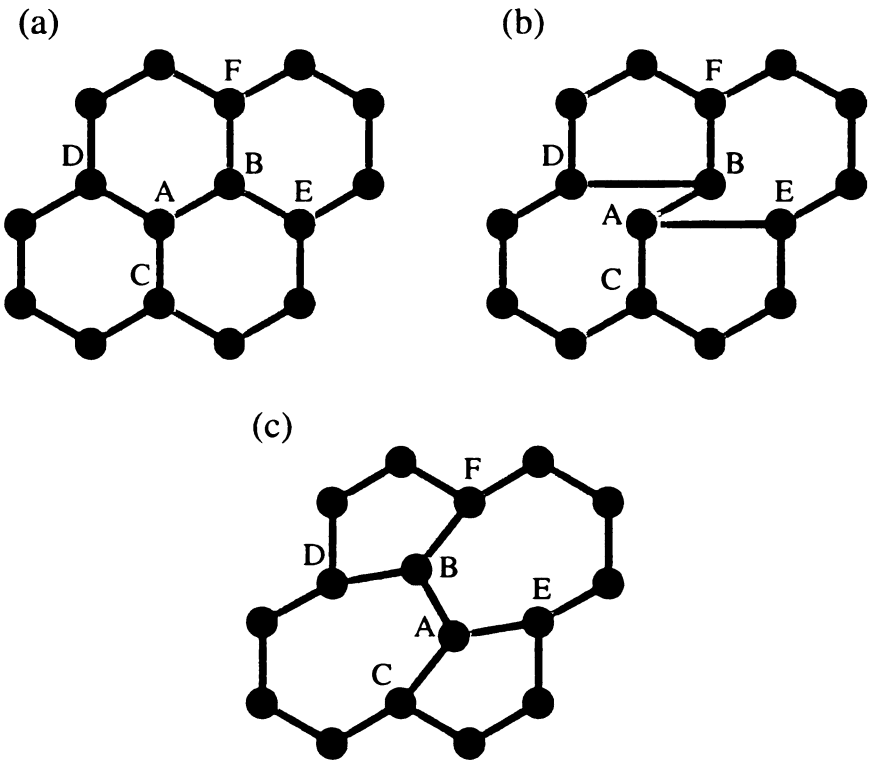


Figure 2.1: One bond switch step in the WWW algorithm is illustrated in 2D. Figure (a) shows the initial crystalline honeycomb network. Bond AD and bond BE are cut, followed by the creation of two new bonds of AE and BD. The resulting topology of the network is different from that in the original network, as shown in Figure (b). Atom coordinates are relaxed to reduce the bond distortions. Two five-fold rings and two seven-fold rings are produced while four six-fold rings are destroyed in this process, as shown in Figure (c).

i is over all atoms, and the sum over j and k are only over the nearest neighbors of atom i.

The Keating potential is simple in form. CRN models relaxed by Keating potential are low in bond distortions. In this sense the Keating potential is accurate enough in generating a static network model, though its first and second order derivatives are not as accurate as some other potentials.

High stress in bond lengths and angles are avoided by forbidding building bonds between two atoms if the distance between them is greater than 1.70 times the equilibrium distance R_0. Since the distance between a pair of second nearest neighbor atoms is $1.63 \times R_0$, this rule disallows direct connections of atoms that are further apart than second nearest neighbors.

Four-fold rings are more stressed in bond lengths and angles than the larger rings. The existence of four-fold ring in the network however helps the relaxation process, as discovered by Djordjević et al. [92]. So the overall effect of four-fold rings in amorphous structures is positive, at least in simulations of this kind.

Figure 2.2 shows an amorphous silicon CRN model built according to the WWW technique, starting from a crystalline silicon network. There are 1000 atoms in the supercell. The size of the supercell is 34.865Å in each side. Periodic boundary conditions are imposed on all three directions so an atom on one surface may be bonded to an atom on the opposite surface. All atoms have exactly four nearest neighbors. The mean average bond length deviation is 5.33% of the equilibrium distance and the mean bond angle deviation is 17.7°.

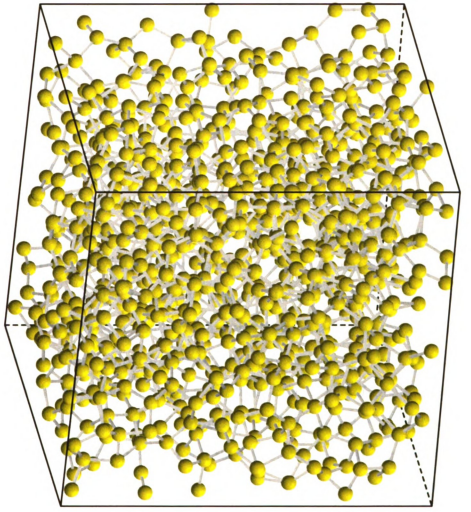


Figure 2.2: An amorphous silicon network model. There are 1000 atoms in the supercell. The size of the supercell is 34.865Å in each edge. The model is built according to the WWW algorithm from a crystalline silicon network.

2.2 Procedure to Build DCRN Silicon Network Models

The models built by the WWW algorithm do not have any internal and unnatural voids. As discussed in the beginning of Chapter 1, it is desirable to have an algorithm that is capable of building DCRN models such as porous amorphous silicon network models. Because the WWW technique is efficient and reliable in building CRN models, our algorithm of creating DCRN models is built upon the WWW method.

The WWW algorithm maintains the full coordination of atoms in every step of switching bonds. The method itself does not discriminate against networks with or without unnatural voids. As long as the atoms in the starting network are fully coordinated, the WWW technique can be applied to change the topology of the network. The effect of a bond switch step is localized. Only local topology and geometry are affected by a bond switch step. If the starting network has an internal void, but every atom is fully coordinated, the application of the WWW technique will not alter the overall shapes and sizes of the voids much. Therefore our task of building DCRN models is transformed to a simpler one of building a fully coordinated network with voids. The WWW algorithm will take care of the remaining job of making the network amorphous. In order for the bond distortions of the created DCRN models to be low, the distortions at every bond length and angle should also be as small as possible. This requirement rules out the simplest starting network of random distributed atoms within which voids are buried.

To build a fully coordinated network with voids for the usage as the input structure of the WWW technique, we first cut voids from a crystalline network. This procedure inevitably leaves partially coordinated atoms at the surface of the voids. Figure 2.3 illustrates the remaining atoms in a crystalline silicon network after a cylin-

der shaped pore is cut in the middle. Most silicon atoms are fully coordinated, shown as the yellow spheres in the figure. These atoms also have no bond distortions. The other atoms are partially coordinated. The atoms that have two and three nearest neighbors, which are named as *two-coordinated* and *three-coordinated* atoms in this thesis, are shown as the blue and green spheres in Figure 2.3. The network in the figure does not have any atoms that have only one nearest neighbor, which are called *one-coordinated* atoms. Because of existence of the partially coordinated atoms the WWW technique cannot be applied to this network. The bonds connecting these partially coordinated atoms have to be adjusted to make all atoms fully coordinated. It is worth noting that network shown in Figure 2.3 is periodic in all three directions. What shows in the figure is one supercell of an infinitely large network. The cylinder shaped pore in the middle of the supercell is infinitely long in the z direction which is virtually perpendicular to the paper.

Bonds linking to the partially coordinated atoms have to be rearranged for the network to be fully coordinated. To minimize distortions in bond lengths and angles as possible is the only principle in re-arranging the bonds in a porous network. The deletion of bonds and atoms is more favorable than the creation of new bonds, because the creation of new bonds inevitably brings bond distortions in the network by connecting two spatially separated atoms together. A simple rule is set that a bond should not be created to between two atoms if the distance between them is larger than 1.7 times the equilibrium distance. This rule is identical to the bond switching rule in the WWW algorithm, as already discussed in Section 2.1.

One-coordinated atoms are the easiest to handle: they are eliminated. The coordination numbers of their nearest neighbors are reduced by one. It is possible that some two-coordinated atoms are converted to be one-coordinated after this step. This step is repeated until the coordination numbers of all atoms in the remaining networks are at least two.

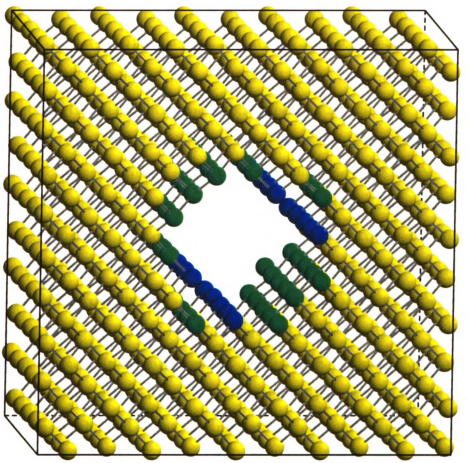


Figure 2.3: A crystalline network in which a cylinder is cut in the middle. The atoms at the surface of the cylinder shaped core are partially coordinated. Two-coordinated and three-coordinated atoms are rendered as blue and green spheres respectively. All the other atoms are fully coordinated, which are the yellow spheres. The atoms shown in this figure belong to one supercell of an infinite network. The network is periodic in all three directions.

When two three-coordinated atoms are bonded to each other, the bond can be cut so that a pair of three-coordinated atoms is transformed into a couple of two-coordinated atoms. In this way the number of three-coordinated atoms in the network is reduced by two, with the result that the number of two-coordinated atoms is increased by two. This result is favorable because the two-coordinated atoms are not as hard to deal with as three-coordinated atoms. Therefore it is preferred that the network has an even number of three-coordinated atoms, and any three-coordinated atom has another three-coordinated atom in its nearest neighbors.

Suppose a network has N_2 two-coordinated atoms, N_3 three-coordinated atoms and N_4 four-coordinated atoms. The total number of bonds will be $(N_2 \times 2 + N_3 \times 3 + N_4 \times 4)/2$. Because the first and the third terms in the denominator are all even and because the total number of bonds should be an integer, the second term in the denominator must also be even. Therefore the number of three-coordinated atoms must be even. This conclusion is valid for any network that is composed of two-, three- and four-coordinated atoms. Therefore the three-coordinated atoms can be grouped pair by pair. It is quite possible that all pairs of three-coordinated atoms are not directly bonded however. A three-coordinated atom may be far away from any other three-coordinated atoms. To translate a three-coordinated atom to be close to another three-coordinated atom, we have invented a defect migration process. It is thus named because a three-coordinated atom is a defect with a dangling bond. The defect of a three-coordinated atom is migrated toward another three-coordinated atom by cutting atoms and bonds, leaving more two-coordinated atoms along the migration pathway.

All three-coordinated atoms are first listed. The distances between the first and any other atoms on the list are calculated. The first one on the list is called the starting defect atom and the three-coordinated atom with the shortest distance to this atom is called the ending defect atom. The general strategy is to cut one bond

between the starting defect atom and one of its nearest neighbors. Suppose all three nearest neighbors of the starting defect are four-coordinated. In this simplest scenario the bond between the starting defect atom and one of its nearest neighbors whose distance to the ending defect atom is the shortest among all its nearest neighbors is cut. The starting defect atom has a coordination number of two after the bond is cut. The atom at the other end of the bond now has a coordination number of three, reduced from four before the bond is cut. The new three-coordinated atom is then the new starting defect atom. The new starting defect atom is closer to the ending defect atom than the old starting defect atom. This step can be repeated until the newly created defect atom is directly bonded to the ending defect atom. Then the bond between this pair of defect atoms can be cut, resulting in a reduced number of three-coordinated atoms in the network.

The three nearest neighbors of the three-coordinated atoms are not always four-coordinated. Defect migration rules are made case by case depending on the coordination numbers of the three nearest neighbors of the starting defect atom. The general principle is to remove as many partially coordinated atoms as possible in any single step. The details of the rules are:

- When the starting defect atom has one two-coordinated atom and two four-coordinated atoms as its nearest neighbors, the bond between the two- and the three-coordinated atoms is cut. The coordination number of the three-coordinated atom is reduced to be two. Since the two-coordinated atom has only one bond after the bond cut, it is removed according to our process of removing one-coordinated atoms.

- When the starting defect atom has two two-coordinated atoms and one four-coordinated atom as its nearest neighbors, both two-coordinated atoms and the three-coordinated atom are cut from the network. The four-coordinated atom is now the new starting defect atom with a coordination number of three.

- When the starting defect atom has three two-coordinated atoms as its nearest neighbors, its three neighbors and it are all removed.

- When the starting defect atom has one three-coordinated atom and two four-coordinated atoms as its nearest neighbors, the bond between the two three-coordinated atom is cut, as already stated before. This is the end of the defect migration process, which converts two three-coordinated atoms to be two-coordinated. The total number of three-coordinated atoms is reduced by two.

- When the starting defect atom has one three-coordinated atom, one two-coordinated atom and one four-coordinated atom as its three nearest neighbors, the three-coordinated atom and the two-coordinated atom are removed. The four-coordinated neighbor now has a coordination number of three. The coordination number of the three-coordinated nearest neighbor of the starting defect atom is reduced to be two.

- When the starting defect atom has one three-coordinated atom and two two-coordinated atoms as its nearest neighbors, the starting defect atom and the two-coordinated atoms are removed.

- When the starting defect atom has two three-coordinated atoms and one four-coordinated atom as its nearest neighbors, the starting defect atom is cut from the network. The coordination numbers of its two three-coordinated atoms are reduced to be two. The coordination number of the four-coordinated neighbor is reduced to be three. So the total number of three-coordinated atom in the network is reduced by two.

- When the starting defect atom has two three-coordinated atoms and one two-coordinated atom as its nearest neighbors, the starting defect atom and the two-coordinated atom are removed.

- When the starting defect atom has three three-coordinated atoms as its nearest neighbors, the starting defect atom is removed. The coordination numbers of the three three-coordinated atoms are all reduced to be two. The net result of this step is that the total number of three-coordinated atoms decreases by four.

Each defect migration step either removes some partially coordinated atoms and the starting defect atom, or making the defect one more step closer to the ending defect atom. When the starting and the ending defect atoms are directly bonded, the bond between them is cut so that they are converted to a pair of two-coordinated atoms. Since the number of three-coordinated atoms in the network is even, this procedure always works. There are only two types of atoms in the network after the defect migration procedure: the two-coordinated and the four-coordinated. Since we have only been cutting atoms and bonds so far, we have not created any new bonds whose bond lengths and angles are distorted. The distances between the two nearest neighbors of any two-coordinated atoms are all less than 1.70 times the equilibrium distance. Therefore by creating new bonds between the neighbors of the two-coordinated atoms and removing the two-coordinated atoms, all the remaining atoms in the network are four-coordinated. All the atoms are fully coordinated after this process.

Figure 2.4 shows a porous network in which all atoms are fully coordinated. The only type of distortions in the network is the bond length distortion at those bonds where two second nearest neighbor atoms are connected. It is built from the porous crystalline network which has partially coordinated atoms shown in Figure 2.3. Since most process involved in our bond rearrangement procedure cut bonds and atoms, the pore size of the fully coordinated network is larger than the original pore size when partially coordinated atoms are present.

Once a fully coordinated and porous network is built, the standard WWW algorithm can be applied to make the network amorphous. The resulting amorphous

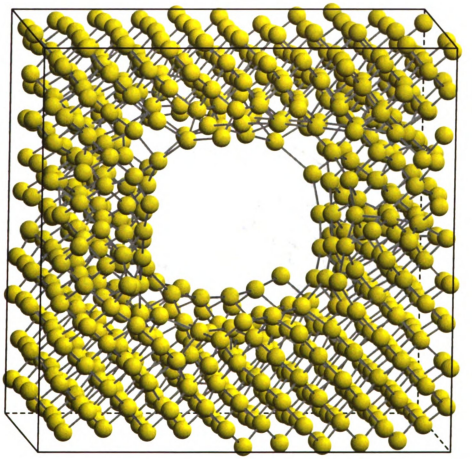


Figure 2.4: A porous network in which all atoms are fully coordinated. The porous network is built from the crystalline porous network shown in Figure 2.3 in which partially coordinated atoms exist, according to the bond rearrangement scheme examined in the text.

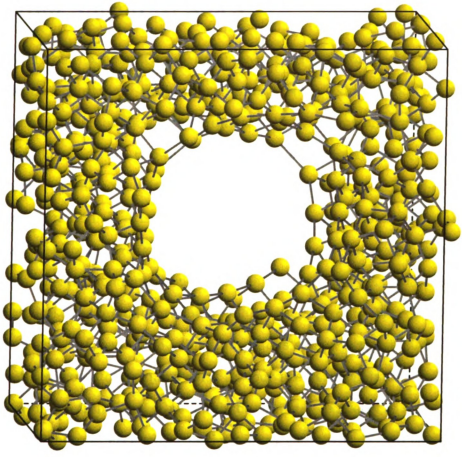


Figure 2.5: An amorphous porous network in which all atoms are fully coordinated. The bond length distortion is 10.1% of the equilibrium bond length and the bond angle distortion is about 15.2°.

porous network is shown in Figure 2.5. The variation of bond length is 10.1% of the equilibrium bond length, which is higher than the typical bond length variations of 5% in CRN models. The variation of bond angles is about 15.2° which is about the same as or somewhat higher than that in CRN models. The relatively big bond length variations are caused by the bond distortions at the surface of the pore.

2.3 From Silicon to Silica

By adding one oxygen atom between each silicon-silicon bond, the amorphous porous silicon networks are transformed to amorphous porous silica networks. All four nearest neighbors of the four-coordinated silicon atoms are oxygen atoms, and both two neighbors of the two-coordinated oxygen atoms are silicon atoms. Such amorphous porous silica models are ideal because small traces of silicon-silicon bonds and oxygen-oxygen bonds exist in the real world amorphous silica materials.

Because the bond angle variation potential energy of the oxygen atoms are soft compared to the stiff bond angle energy of the silicon atoms, the insertion of oxygen atoms help release the bond length and angle stress of the silicon atoms. Some silicon atoms are at the surfaces of the voids in a porous silicon network. The bond angle geometry of a silicon atom requires that its four nearest neighbors should be on the four corners of a tetrahedron. But if this bond angle geometry is satisfied, at least one nearest neighbor of each surface silicon atom will protrude into the voids. Since there should be no atoms in the voids, bond angles of silicon atoms at the surface have to be bent significantly. The insertion of oxygen between silicon-silicon bonds reduces the bond angle distortions at the surface. When an oxygen atom is placed on the surface of a void, both of its two nearest neighbors can be placed away from the surface. The bond angle geometry of the surface oxygen atom is correct, without the necessity of placing its nearest neighbors in the voids. Therefore the addition of oxygen atoms reduces the bond angle stress at the surfaces of the voids.

The amorphous fiber and amorphous film silica models shown in Chapter 3 are built by this method from the amorphous fiber and film silicon models. The bond angle distortions at the surface will be discussed in detail in Chapter 3.

2.4 Defects and Hydrogen

Hydrogenated amorphous silicon a-Si:H is a potential material for high efficiency solar cells [93, 94], thin film transistors [95, 96] and other applications. Dangling bond defects are often detected in the a-Si:H material. The dangling bonds affect the application performance of a-Si:H film. The defect migration process, which is designed to eliminate defects, can be used to introduce and distribute defects in the silicon network models. Therefore our algorithm can be used to build network models with arbitrary distribution of dangling bond defects or hydrogenated defects.

A pair of dangling bond defects can be created in a fully coordinated network by cutting a bond. The pair of defects then can be migrated to different directions. Slightly different from the defect migration rules in creating a fully coordinated network, the defects migrate according to a rule similar to the bond switch procedure in the WWW technique. Suppose atom B is a nearest neighbor atom of the defect atom A. Suppose atom C is a nearest neighbor atom of atom B. A bond between atom A and C is created, making atom A fully coordinated, followed by the cutting of the bond BC. Atom B has only three nearest neighbors after the procedure so it is the new defect atom. Atomic coordinates are relaxed to reduce the bond distortions. The net result of this step is that the defect migrates from atom A to atom B, as shown in Figure 2.6. The two defects are not spatially close to each other after several steps of this defect migration process. Defects are created and migrate repetitively until a predefined defect concentration and spatial distribution are reached. Our algorithm, which is initially designed to build fully coordinated networks, is an efficient and powerful procedure in planting defects in fully coordinated networks.

Most of the hydrogen atoms in the a-Si:H films are on the surfaces serving as terminators of dangling bonds. Few hydrogen atoms are buried deeply inside of the a-Si:H film. They are connected with silicon atoms. The Si-H bond in the interior of a-Si:H networks can be viewed as a special kind of defect involving dangling bonds

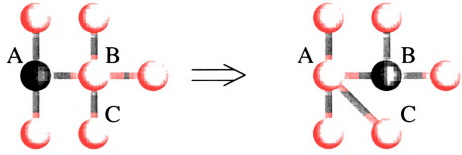


Figure 2.6: The defect migration procedure in planting defects in fully coordinated networks. Two defects are created when a bond in a fully coordinated network is cut. Defects are then migrated to different directions according to the procedure shown in this figure. A bond is created between the defect atom A and one of its second nearest neighbor atoms C. Atom A is then fully coordinated. The bond between atom C and atom B which is a nearest neighbor atom to both atom A and atom C is cut, making the atom B the new defect atom. Defect atom is rendered as green spheres while fully coordinated atoms are yellow spheres.

which are terminated by hydrogen atoms.

Hydrogen atoms can be implanted into our fully coordinated amorphous silicon models. Two dangling bonds are created when a bond between two silicon atoms is cut. As shown in Figure 2.7(a) a pair of hydrogen atoms can be attached to the dangling bonds to terminate them. Each hydrogen atom can go through the defect migration process so that these two newly inserted atoms are spatially separated. As shown in Figure 2.7(b), in each defect migration process the bond between the silicon atom A and the hydrogen atom is cut, followed by a creation of a new bond between atom A and one of its second nearest neighbors atom C. The hydrogen atom is migrated to atom B which is a nearest neighbor both of atom A and of atom C. The bond between atom B and atom C is cut. All atoms, including hydrogen and silicon atoms, are fully coordinated before and after this process. The net result is that the hydrogen atom is transferred from silicon atom A to silicon atom B. A subsequent relaxation process reduces the bond distortions. The two inserted hydrogen atoms are spatially separated after several steps of hydrogen atom migration process. By adding and migrating hydrogen atoms pair by pair, we can build amorphous silicon

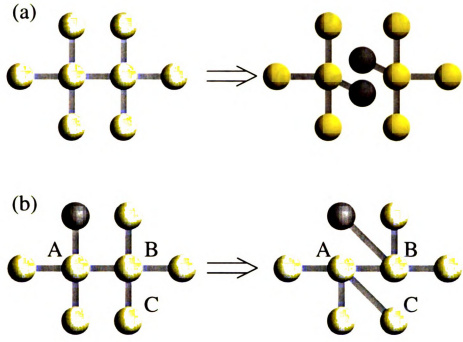


Figure 2.7: The procedure to add and migrate hydrogen atoms in the fully coordinated networks. Two dangling bonds are created when a silicon-silicon bond is cut. Two hydrogen bonds are then inserted to fulfill the valency of silicon atoms, as shown in Figure (a). Each hydrogen atom can be migrated to any random direction. The basic migration process involves the breaking of the bonds between atom A and H and between atom B and C, and the creating of two new bonds between atom A and C and between atom B and H. Hydrogen and silicon atoms are shown as gray and yellow spheres respectively.

models with any concentration of buried hydrogen atoms.

Chapter 3: Discontinuous Random Network Models

The algorithm described in Chapter 2 is able to build any DCRN models. The DCRN models can be an infinite amorphous network with voids built in such as an amorphous porous silicon model, or can be a network that is finite in certain dimensions. This chapter discusses two amorphous network models that are finite in certain dimensions. They are the amorphous fiber silica model in Section 3.1 which is infinitely long in the z direction but finite in the x and y directions, and the amorphous film silica model in Section 3.2 which is continuous in the x and y directions but finite in the z direction. The Pair Distribution Functions (PDF) of these two models are discussed in Section 3.3.

3.1 Amorphous Fiber Silica Model

3.1.1 Procedure to Build Amorphous Fiber Network Model

To build an amorphous fiber silicon model, a long fiber-like supercell containing 3 by 3 by 16 supercells of crystalline silicon is first set up, as shown in Figure 3.1. Because each supercell of crystalline silicon has 8 atoms, the fiber supercell is made up of 1152 atoms. The fiber supercell is periodic along the z axis, but finite in the x and y directions. The silicon atoms at the surface of the supercell are partially coordinated. Some atoms are one-coordinated and some are two-coordinated, shown as the red and blue spheres respectively in the figure. The other atoms are fully coordinated.

Silicon atoms in two adjacent side surfaces of the supercell, indicated by the two rectangular boxes in Figure 3.1, are removed. Only the outer-most atoms are removed

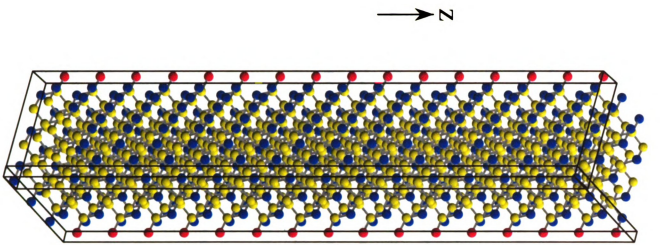


Figure 3.1: A big fiber supercell containing 3 by 3 by 16 supercells of crystalline silicon. There are 1152 atoms in the supercell. The supercell is periodic along the z axis but finite in the other two directions. The atoms at the surface of the supercell are partially coordinated. One- and two-coordinated atoms are colored as red and blue respectively.

in one surface. The outer-most atoms and their nearest neighbors are also removed in the other surface. The resulting network still has partially coordinated atoms, but all partially coordinated atoms are two-coordinated. By connecting the two nearest neighbors of the two-coordinated atoms and removing the two-coordinated atoms themselves, we obtain a fully coordinated network which is finite in the x and y directions and periodic in the z direction, as shown in Figure 3.2. There are 64 cases of a two-coordinated atom connected directly to another two-coordinated atom. New bonds are created between the two nearest neighbors of the two two-coordinated atoms while the two-coordinated atoms are removed. In this case 64 four-fold rings are created. We allowed the creation of four-fold rings in this case because the procedure in building fiber model is simpler in this way. The distortions at the 64 four-fold rings will be relaxed in the following WWW procedure. The boundary of the original supercell is also shown in the figure to clarify how much the supercell shrinks in x and y directions after this procedure. This bond rearrangement procedure removes almost half the atoms from the original supercell. There are only 640 atoms left in the supercell after this procedure. The remaining network is uniform in radius from top to bottom.

The WWW algorithm is then applied to the network to make it amorphous. Oxygen atoms are inserted afterward between every silicon-silicon bond to create an amorphous silica network. The final model which is shown in Figure 3.3 has 640 silicon atoms and 1280 oxygen atoms in one supercell. The length of the supercell is 114.8Å. Because the oxygen angles are soft, the fiber model is not stressed compared to the amorphous CRN silicon models. The root mean square deviation (RMSD) of Si-O bond is 0.107Å, which is about 6.6% of the optimal value. The bond angle RMSD at silicon and oxygen atoms are 13.2° and 13.8° respectively.

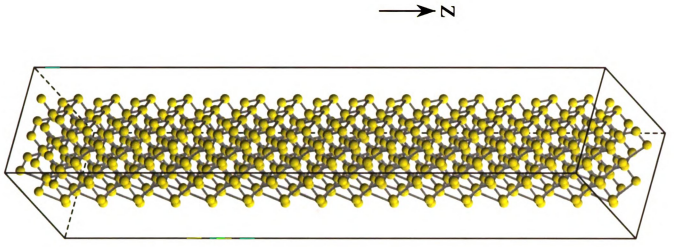


Figure 3.2: A fully coordinated network which is finite in the x and y directions and periodic along the z axis. There are 640 atoms in the supercell.

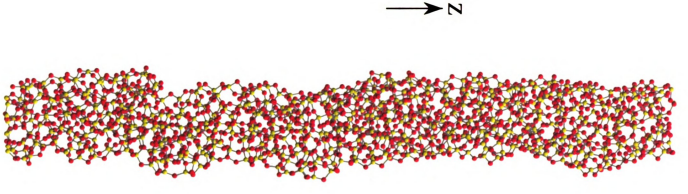


Figure 3.3: A fully coordinated amorphous fiber silica network. There are 640 silicon atoms (yellow) and 1280 oxygen (red) atoms in the supercell. The length of the supercell is 114.8Å. The average diameter of the fiber model is about 7.2Å.

3.1.2 Properties of Amorphous Fiber Network Model

The radius of the fiber varies a little bit from top to bottom due to local relaxations. To determine an average radius of this model, we cut the model into hundreds of slices perpendicular to the z axis. Each slice has its mean average center axis. The atomic number density of a slice is counted at all distances away from the center axis. The step size of the histogram is 0.12Å. Finally the number density is averaged over all slices and plotted in Figure 3.4. The total number density drops smoothly after about 6.1Å away from the center axis. It is almost zero at about 9.0Å. The average radius of the fiber model is roughly 7.2Å. As revealed in the figure, the number density of oxygen is roughly twice as that of the silicon at any distance away from the center axis as would be expected. The number density of oxygen is high at the center axis because there happens to be several atoms at the center line.

It seems there is a trend in Figure 3.4 that the total number density increases with the distance away from the center axis, but this is not true. Suppose there is an atom at the center axis of the fiber. The nearest neighbor of this atom will be at a certain distance away from the center axis, making the number density exactly zero at the space close to the center axis. The seemingly increasing number density from the center to the surface proves that the atoms at the center are more orderly packed than the atoms at the surface, so that the difference in number density at certain distances where atoms are placed and the number density at other distances where atoms are not likely to be placed is more obvious.

Because the bond angles of the oxygen atoms are softer than those of the silicon atoms, oxygen atoms are suitable to be positioned at the surface where the stress concentrates. Each oxygen atom is bonded to two nearest neighbors. The oxygen atom at the surface can point outward while its two nearest neighbors are placed inward. A silicon atom is at the center of a tetrahedron formed by its four nearest neighbors. Therefore it is not possible to place a silicon atom at the out most surface

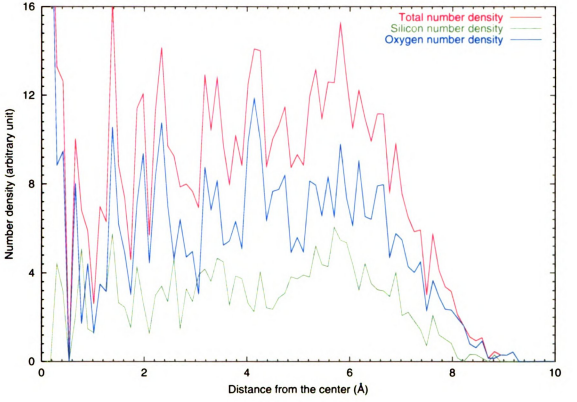


Figure 3.4: Atomic number density at all distances from the center axis of the fiber model. The box size of the histogram is 0.12Å.

without causing big bond distortions. Based on these arguments, we say the surface atoms should all be oxygen atoms. To count all the surface atoms, we first place a right circular cone on every oxygen atom in such a way that the tip of the cone is at the atom and the axis of the cone is perpendicular to the axis of the fiber. The axis of the cone coincides with the shortest line segment from the atom to the fiber axis. The half-cone angle is set to be 30°. If the cone encloses any other atoms the atom on which the cone is placed is not a surface atom. The blue atom shown in Figure 3.5 is such an example. Otherwise the atom is at the surface, which is exemplified by the green atom in the figure.

The surface atoms are also called the first layer atoms. Their nearest neighbors are called the second layer atoms. The atoms that are not first layer atoms but are bonded directly to the second layer atoms are called the third layer atoms, so on and so forth. There are seven layers in the fiber model. Because first layer atoms are exclusively made of oxygen atoms and because no like atoms are bonded in the model, all odd layers are made of oxygen atoms and all even layers are made of silicon atoms. There are 572, 503, 195 and 10 oxygen atoms in the first, the third, the fifth and the seventh layers, and 453, 163 and 24 silicon atoms in the second, the fourth and the sixth layers respectively. It is worth noting that the higher the layer indexes the closer the layer to the center of the fiber.

Figure 3.6 shows the bond length and angle distortions at different layers. The average bond length deviations from the optimal value are positive in the first two layers, but negative in the other layers. Since the first two layers are all on the fiber surface, we can reach the conclusion that the bond lengths are stretched on the surface, while compressed in the interior of the fiber. Therefore it is reasonable to say that atoms are more densely packed in the interior of the fiber model than they are on the surface. The bond angle deviations at the odd layers are all negative, while almost zero at even layers. Silicon atoms are at the even layer. This fact suggests that

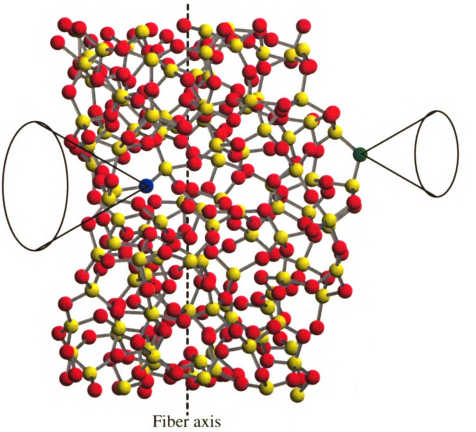


Figure 3.5: Surface atoms are the atoms further from the axis of the fiber (dotted line) than any spatially closed atoms. If a right circular cone whose axis is perpendicular to the fiber axis at an atom does not enclose any other atom, the atom upon which the cone is placed is a surface atom, such as the green atom. Otherwise the atom is not a surface atom, such as the blue one. Red and yellow atoms are oxygen and silicon atoms respectively.

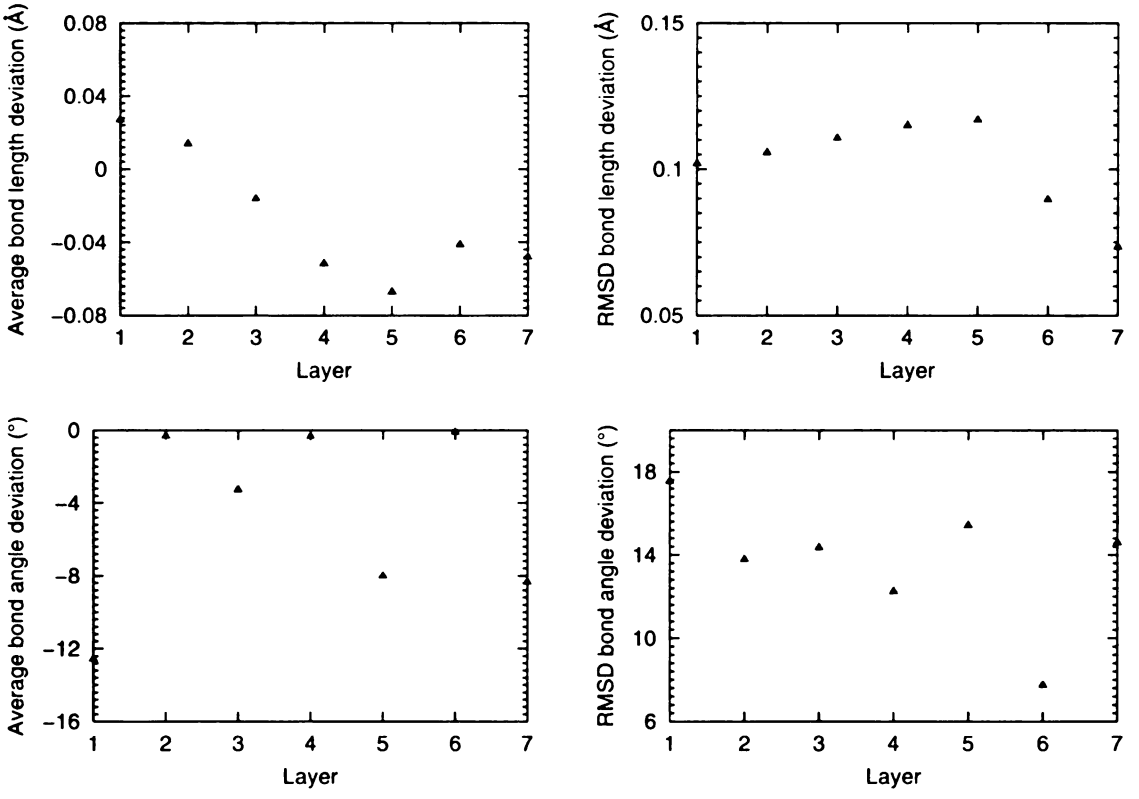


Figure 3.6: The average bond length deviations from the optimal value (top left panel), the root mean square deviation (RMSD) of bond lengths from the optimal value (top right panel), the average bond angle deviation from the optimal value (bottom left panel) and the RMSD of bond angle deviation from the optimal value (bottom right panel) in all seven layers of the fiber model. The first layer is the fiber surface, and the highest layers are in the center of the fiber.

the average bond angles at silicon atoms are very close to the optimal value of 109°. Average bond angles at oxygen atoms are all about 10° less than the optimal value of 147°. This is caused by the presence of non-negligible number of four-fold rings in the network. The RMSD of both bond length distortions and bond angle distortions are small in all layers, proving distortions in our fiber model are acceptable.

3.2 Amorphous Film Silica Model

3.2.1 Procedure to Build Amorphous Film Network Model

Fully coordinated amorphous film silica model can be easily built from crystalline silicon networks that are cut along the $(0, 0, 1)$ direction. All the atoms at the $(0, 0, 1)$ surface have two nearest neighbors, while all the non-surface atoms are fully coordinated, as shown in Figure 3.7. The example in the figure is made up of 4 by 4 by 8 supercells of crystalline silicon. It is periodic in x and y directions but finite in the z axis. Since each supercell of the crystalline silicon network has 8 atoms, the total number of atoms in the supercell of the film model is 1024.

Following the standard procedure described in Chapter 2, the two-coordinated atoms at the $(0, 0, 1)$ surfaces are removed after their nearest neighbors are connected by new bonds. The resulting network is fully coordinated. The number of remaining atoms in the supercell is reduced from 1024 to 960 after this procedure. The WWW algorithm is then applied on this network to make it amorphous. Oxygen atoms are added between every silicon-silicon bond to transfer the amorphous film silicon network to an amorphous film silica network. The supercell of the final result is shown in Figure 3.8. There are 960 silicon atoms and 1920 oxygen atoms in the supercell. The size of the supercell is 28.70Å along the x and y axes, and is 57.39Å along the z axis. The model is finite in z directions. The top and bottom surfaces of the supercell are the surfaces of the film model.

3.2.2 Properties of Amorphous Film Network Model

Similar to our studies in the amorphous fiber silica model, we can identify the layers in the amorphous film silica model. The atoms in the first layer are the surface atoms. The atoms in the second layer are the nearest neighbors to the first layer atoms, so on and so forth. All odd numbered layers in the model are made up of

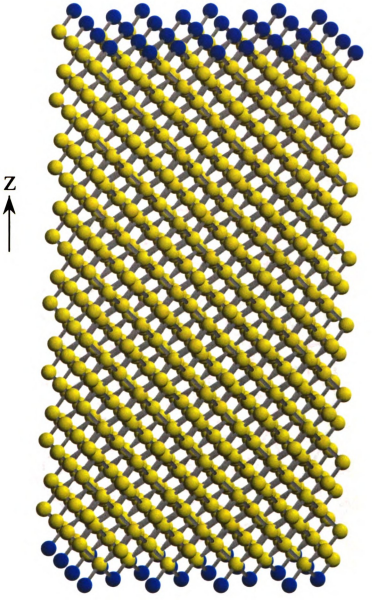


Figure 3.7: A network that is made up of $4 \times 4 \times 8$ supercells of crystalline silicon. Since the network is finite in z axis, atoms are not fully coordinated at the top and bottom surfaces of the network. The network is periodic in x and y directions. Fully coordinated silicon atoms are colored yellow, while two-coordinated atoms are colored blue.

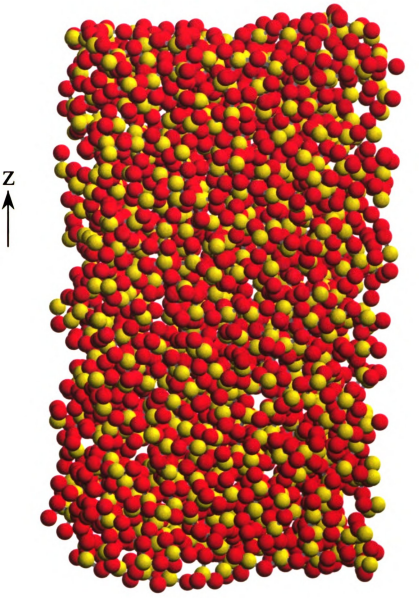


Figure 3.8: The supercell of an amorphous film silica network model. The sizes of the supercell are 28.70Å, 28.70Å and 57.39Å along the x, y and z directions respectively. The top and bottom surfaces of the supercell are the surfaces of the film silica network model. There are 960 silicon atoms (yellow) and 1920 oxygen atoms (red) in the supercell.

oxygen atoms while all even numbered layers are of silicon atoms. Figure 3.9 shows the bond distortions at all layers in the amorphous film model. The average deviations of bond lengths from the optimal value is highest at the surface, then drops almost linearly to be about zero at the fifth layer. The distortions caused at the surface is then limited to the outer-most five layers of the film. The RMSD of bond lengths from the optimal value is also the highest at the surface, then decreases to be normal within five layers. Since the amorphous fiber silica model has only seven layers, the whole fiber silica model are influenced by the surface effect. The average deviations of bond angles in the amorphous film silica model oscillate around zero at the silicon atoms, but have negative values at the oxygen atoms. The smaller average bond angle at the oxygen atoms is caused by the four-fold rings in the network. The RMSD of all silicon-oxygen bonds is 0.09Å, which is slightly better than that in the amorphous fiber silica model. The RMSD of bond angles at silicon atoms and at oxygen atoms are 9.52° and 12.64°, which are all improved when compared to the corresponding values in the amorphous silica model.

3.3 Distribution Functions

Due to the lack of long range order in amorphous networks, there is no way to predict positions and orientations of atoms far from the observation point based on the spatial arrangement of atoms nearby. The only information that is available in an amorphous network is the correlations between atomic distributions. Several distribution functions have been used to depict the atomic spatial distribution correlations. The nomenclatures of the distribution functions vary in literatures. The definitions used in this dissertation follow what are used by Thorpe et al. [97].

The Density Distribution Function (DDF) $\rho(r)$ is the probability to find an atom in a unit volume that is at a particular distance r away observed from an averaged

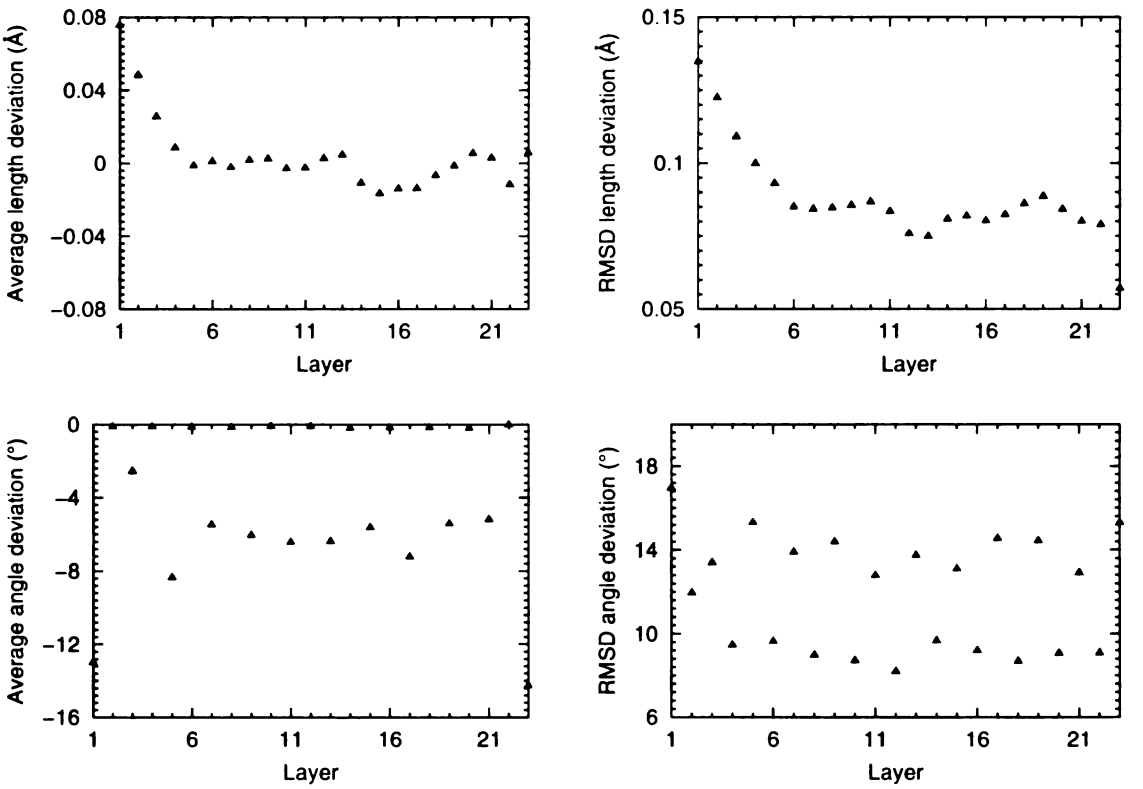


Figure 3.9: The average bond length deviations from the optimal value (top left panel), the RMSD of bond lengths from the optimal value (top right panel), the average bond angle deviation from the optimal value (bottom left panel) and the RMSD of bond angle deviation from the optimal value (bottom right panel) in all layers of the amorphous film silica network model. The first layer is the film surface, and the highest layers are in the center of the film.

atom. It is mathematically defined as

$$\rho(r) = \frac{1}{4\pi r^2}\frac{1}{N}\sum_{i\neq j}\delta(r - r_{ij}) \tag{3.1}$$

in which N is the total number of atoms in the network, r_{ij} is the distance between atom i and atom j and the sum is over all pairs of atoms excluding such terms as i and j are equal. $\delta(r - r_{ij})$ is a Dirac delta function.

The Radial Distribution Function (RDF) $T(r)$ is the total number of atoms within a thin spherical shell whose inner and outer radii are r and $r + dr$ respectively observed from an averaged atom. It is related to the DDF by:

$$T(r) = \rho(r)4\pi r^2 = \frac{1}{N}\sum_{i\neq j}\delta(r - r_{ij}) \tag{3.2}$$

An integration of $T(r)$ over all distance range produces $N - 1$ which is the total number of atoms in the network excluding the atom observed from.

The atomic distribution in a random network is virtually uniform when the distance r is large. Therefore RDF should be roughly $4\pi r^2\rho_0 dr$ in the limit of $r \to \infty$, in which ρ_0 is the average number density of the network. Hence $\rho(r)$ approaches the average number density ρ_0 in the long distance range. It is for this reason that the $\rho(r)$ defined in Equation 3.1 is called the density distribution function. The density distribution $\rho(r)$ is not the same as the average number density ρ_0 in the short and intermediate distance range because the atomic distribution is far from uniform when r is small. The DDF and its sister functions are great tools in analyzing the atomic arrangement in the short and intermediate distance range in amorphous networks.

The Reduced pair distribution function (RPDF) $G(r)$ is related to the DDF by the simple transformation

$$G(r) = 4\pi r\left[\rho(r) - \rho_0\right] \tag{3.3}$$

Obviously the RPDF approaches zero in the long distance range where $\rho(r)$ is virtually the same as ρ_0. RPDF is indirectly measured in X-ray or neutron powder diffraction experiments through a conversion from the diffraction pattern $s(q)$ by

$$G(r) = \frac{2}{\pi} \int_0^\infty [s(q) - 1] \sin(qr) q dq \tag{3.4}$$

Because of the easy conversion between RPDF and the powder diffraction pattern, RPDF and its sister functions have been extensively used as the standard tool to study the short and intermediate atomic spatial distributions. The DDF, RDF, RPDF and other variants are generally called the pair distribution function (PDF). There are two typical approaches in comparing the PDF obtained from modeling and from experiments. The first approach is to convert the diffraction pattern $s(q)$ to PDF, then compare the experimentally obtained PDF with the PDF from modeling in the real space. The second approach is to convert the PDF to a diffraction pattern, then compare the diffraction pattern from the modeling with that from the experiments in reciprocal space. The first approach is preferred because a peak in the real space has obvious meanings while a peak in the reciprocal space may not be so easy to explain.

One question we want to address is how much our amorphous fiber and film silica models differ from CRN models in term of PDF in the short to intermediate distance range. The PDF of our amorphous fiber and film models have to be close to that of CRN models, otherwise our models are likely to be unrealistic. The comparisons of the PDF of the amorphous fiber and film models with the PDF of the CRN silica model are thus not only a test of the validity of the built models, but also a test on the algorithm upon which the DCRN models are built.

But all variants of PDF are defined under the assumption that atoms are isotropically distributed in all directions. However the atomic distributions in both the fiber and the film models are anisotropic. Atoms in the fiber model are distributed along a

long cylinder, and atoms in the film model are distributed within two surfaces. The long range behaviors of the PDF of the fiber and film models are thus quite different from the PDF of CRN models. Figure 3.10 shows the RDF of three amorphous models: the CRN model, the amorphous fiber model and the amorphous film model. The RDF of the CRN model rises much faster than the RDF of the amorphous film model, which is roughly linear in long distance range. The RDF of the amorphous fiber has a maximum at about 10Å. It is almost flat in long distance range. Therefore it is not reasonable to compare the PDF of the amorphous fiber, amorphous film and CRN models directly. To reveal the differences in PDF of the three models in the short and intermediate distance range, we define a new distribution function called the Reduced Density Distribution Function (RDDF) $P(r)$ which is the radial distribution function of a network model divided by the radial distribution function of a uniform media that has the same overall shape as the model.

As discussed in Appendix A, the RDF of infinitely large uniform media, of uniform media in the shape of a film of thickness d, and of uniform media in the shape of a fiber of radius d are

$$T^u_{inf.}(r) = 4\pi r^2 \rho_0 \tag{3.5}$$

$$T^u_{film}(r) = \begin{cases} 4\pi r^2 (1 - \frac{r}{2d})\rho_0 & 0 \le r < d \\ 2\pi d r \rho_0 & r > d \end{cases} \tag{3.6}$$

$$T^u_{fiber}(r) = \begin{cases} 4\pi r^2 \rho_0 \left[1 - \frac{8d}{3\pi r}(1 + \frac{r^2}{4d^2})E(\frac{r}{2d}) + \frac{8d}{3\pi r}(1 - \frac{r^2}{4d^2})K(\frac{r}{2d})\right] & r < 2d \\ 4\pi r^2 \rho_0 \left[1 - \frac{8d}{3\pi r}(1 + \frac{r^2}{4d^2})E(\sin^{-1}(\frac{2d}{r}), \frac{r}{2d}) + \frac{16d^2}{3\pi r^2}(1 - \frac{r^2}{4d^2})K(\frac{2d}{r})\right] & r > 2d \end{cases} \tag{3.7}$$

in which ρ_0 is the average number density and functions E and K are the elliptic

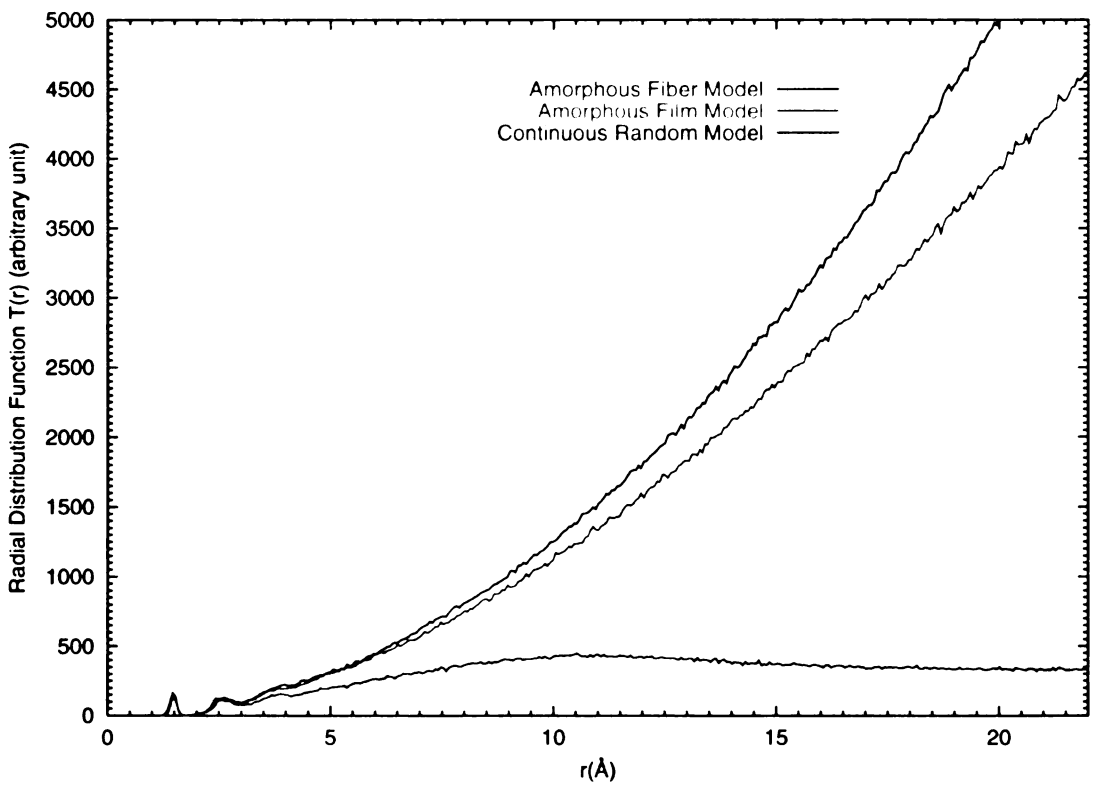


Figure 3.10: Radial Distribution Function of a CRN model (blue curve), of an amorphous film model (green curve) and of an amorphous fiber model (red curve).

integrals defined as

$$E(\phi, k) = \int_0^{\phi} \sqrt{1 - k^2 \sin^2 \theta} d\theta$$
$$K(k) = \int_0^{\frac{\pi}{2}} \frac{d\theta}{\sqrt{1 - k^2 \sin^2 \theta}} \tag{3.8}$$

The merit of the RDDF is that by dividing the RDF of uniform media of the same shapes, the RDF of amorphous models are transformed into new functions in which the peaks in the short and intermediate distance range are manifested. Figure 3.11(a) shows the RDF of the amorphous fiber model superimposed with the RDF of uniform media in the shape of a fiber. The peaks in the short and intermediate distance range are buried in the noise. Figure 3.11(b) shows the RDDF of the same fiber network model. Compared to the RDF, RDDF clearly exposes the first several peaks of the correlations in local atomic arrangement. The effects of the overall shapes on usual PDF are totally eliminated.

The RDF of infinitely large and uniform media is $4\pi r^2 \rho_0$, as given in Equation 3.5. Therefore the RDDF of CRN models which are infinite and isotopic is the RDF of the model divided by $4\pi r^2 \rho_0$, which is identical to the definition of the DDF of the CRN models except a constant factor ρ_0. The RDDF of the amorphous fiber and film models whose atomic distributions are anisotropic are different from the corresponding DDF.

Atomic correlations in the local and intermediate distance range are better disclosed in RDDF than in any other distribution functions. Figure 3.12 shows the RDDF of a CRN model, of the amorphous fiber model and of the amorphous film model. All three RDDF remarkably resemble each other. The first sharp peak comes from the nearest neighbor silicon-oxygen distance. The second peak is composed of the second nearest neighbor oxygen-oxygen correlations and the silicon-silicon correlations. A third peak is broad and barely visible in all three RDDF. The RDDF of

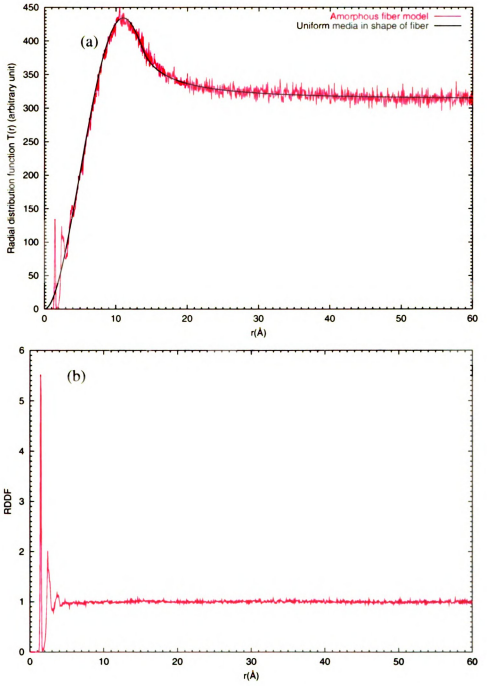


Figure 3.11: Figure (a) shows the RDF of the amorphous fiber silica model superimposed with the RDF of uniform media in the overall shape of a fiber. The RDF of amorphous fiber model divided by that of the uniform media produces the RDDF which is shown in Figure (b). The effects of shapes on the atomic distribution functions is absent in RDDF.

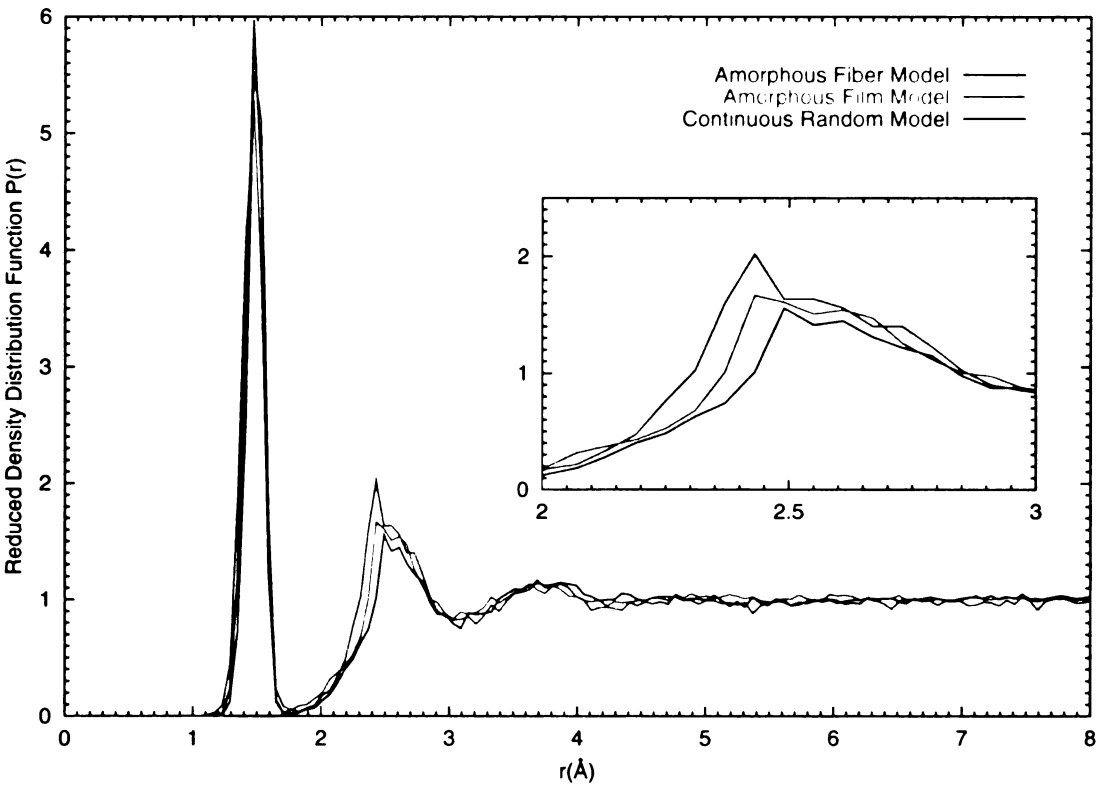


Figure 3.12: The RDDF of an amorphous fiber model (red curve), of an amorphous film model (green curve) and of a CRN model (blue curve). The inset is the close-ups of the second peaks in the RDDF of the three models.

all three models are featureless beyond the third peak.

The first peak in the RDDF of the CRN model is narrower and sharper than that in the RDDF of the amorphous fiber and film models. This is due to the smaller distortions in nearest neighbor silicon-oxygen bond lengths in the CRN model. The atoms at and close to the surfaces in the amorphous fiber and film models are unavoidably more distorted in bond lengths and angles than the atoms far away from the surface. Since the CRN model does not have any surface, the average bond lengths and angle distortions in CRN models are expected to be lower than those in the fiber and film models in which surfaces are present. This is proved by the fact that the RDDF of the first peak of the CRN model is sharper and narrower than that of the other two models.

The amorphous fiber model has the largest surface area, both in absolute and in relative terms. Therefore the distortions in the amorphous fiber model should be the largest among the three. Since the bond length distortions incurs higher potential penalties than the bond angle distortion does, most of the distortions are bond angle distortions, particularly oxygen bond angle distortions because the oxygen bond angles are softer than the silicon bond angles. The second peak in RDDF is the distribution of second nearest neighbor distances, which is an indirect measurement of the bond angle distortions. The second peak of the RDDF of the fiber model is the widest of the three RDDF, indicating the bond angle distortions in amorphous fiber model are larger than those in the other two models. The amorphous film model is in the middle between the amorphous fiber model and the CRN model in term of the width of the second peak.

The general forms of the RDDF of the three models are the same. The shapes and the positions of the peaks are virtually identical. In this sense the DCRN models generated by our algorithm – the amorphous fiber and amorphous film models shown in this chapter – do not differ much from the CRN models in terms of local and intermediate atomic distributions. Though the distortions of bond lengths and angles at the surfaces play roles in affecting local atomic distributions, they are not beyond reasonable levels so that the RDDF are not altered significantly.

3.4 Metal-Adamantane Network Model

Pivan et al. [98] synthesized a chemical material containing triogermanate adamantane $[Ge_4S_{10}]^{4-}$ units in 1994. X-ray diffraction experiment proves this material is crystalline. Each adamantane unit is a tetrahedron, as shown in Figure 3.13. Each adamantane unit is composed of ten sulfur atoms and four germanium atoms. Six out of the ten sulfur atoms – the gold atoms in the figure – are bonded to the four

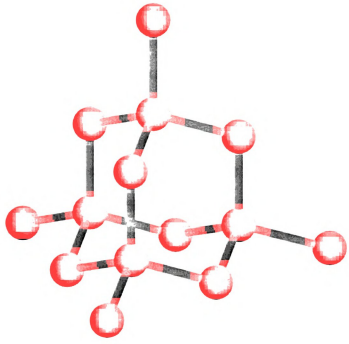


Figure 3.13: A $[Ge_4S_{10}]^{4-}$ adamantane unit. Each adamantane unit is composed of ten sulfur atoms and four germanium atoms. Four out of the ten sulfur atoms are connected to the sulfur atoms in the other adamantane units, or are bonded to metal atoms. These four sulfur atoms are colored in red. The other six sulfur atoms are bonded to the germanium atoms in the same adamantane unit. They are colored gold in the figure. Germanium atoms are colored as yellow. The whole adamantane unit forms a tetrahedron.

germanium atoms. The other four sulfur atoms are called the *terminal sulfur atoms.* One end of each terminal sulfur atom is bonded to a germanium atom in the same unit, and the other end may be connected to a metal atom or a terminal sulfur atom in another adamantane unit. Later studies [99, 100, 101] report metal-adamantane networks. All these networks are crystalline and micro-porous.

The material synthesized by Bonhomme and Kanatzidis [102] has mesoscopic structures. Thin layers of crystalline adamantane units are separated by long organic surfactant molecules. A novel metal-adamantane material is first reported by MacLachlan et al. [103]. Long surfactant molecules cluster to form large cylinder-shaped tunnels. These tunnels are well organized on a hexagonal network. The metal

atoms and the adamantane units fill between these tunnels. Therefore from a topological point of view, the metal atoms and the adamantane units form a porous network. A subsequent work [104] suggests that the metal-adamantane network is well ordered. On the contrary, Rangan and his co-workers [105] propose that the metal-adamantane network should be disordered. Later X-ray powder diffraction analysis by Wachhold et al. [106] confirms the lack of long range order in the metal-adamantane network. Hence the metal-adamantane network is not only porous, but also amorphous. A schematic diagram is plotted in Figure 3.14. The organic surfactant molecules form cylindrical channels. According to Wachhold et al. [106], the diameter of the channels ranges from 22Å to 32Å, depending on the lengths of the surfactant molecules. The separation between the channels are between 30Å to 44Å. Therefore the thickness of the metal-adamantane network can be as thin as about 10Å in some regions of the material. Since the distance between a terminal sulfur atom and the center of the adamantane unit is only 4.47Å, at most two layers of adamantane units are allowed in the thinest regions between the surfactant channels.

Metal atoms are tetrahedrally coordinated to the adamantane units. The adamantane units are also tetrahedrally bonded to other atoms. In this sense the amorphous metal-adamantane network resembles the amorphous gallium arsenide network, in which all atoms are tetrahedral. It is energetically favorable for the metal atoms to be bonded to adamantane units, and for the adamantane units to be connected to the metal atoms. All the rings in the network should be even-numbered. However a large amount of five-fold rings and seven-fold rings exist in the networks built by our algorithm. A large number of unlikely bonds will be present if the metal-adamantane network is modeled by our algorithm. For this reason we turn to other models to model the metal-adamantane network.

As the starting point we use an amorphous gallium arsenide model constructed by Barkema and Mousseau [107]. This model contains 1000 atoms. There are not any

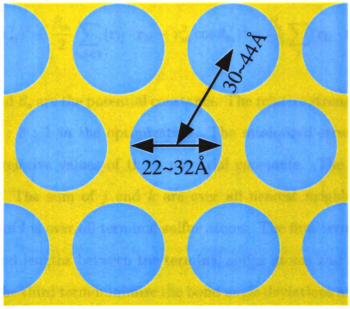


Figure 3.14: A schematic diagram of the proposed mesoscopically porous metal-adamantane material. Surfactant molecules cluster to form long cylindrical channels. The diameters of the channels range from 22Å to 32Å, depending on the lengths of the surfactant molecules. These surfactant channels can be approximately viewed as being on a hexagonal lattice. Average distance between channels are about 30Å to 44Å. These channels are sketched as the light blue circles in the figure. Metal atoms and adamantane units form cross-linked network between the channels.

odd-numbered rings in the model. The gallium and arsenic in the model are redefined to become metal atoms and adamantane units respectively. Since all the rings in the model are even-numbered, each metal atom is bonded to four adamantane units, and each adamantane unit is connected to four metal atoms. Since this is a bulk model, only the bulk properties of the metal-adamantane network can be studied.

The network is then relaxed by a Keating-like potential

$$E = \frac{\alpha}{2}\sum_{i,j}(r_{ij}^2 - r_m^2)^2 + \frac{\beta_m}{2}\sum_{i,j<k}(\mathbf{r}_{ij}\cdot\mathbf{r}_{ik} - r_m^2\cos\theta_m)^2 + \frac{\beta_s}{2}\sum_{l}(\mathbf{r}_{l1}\cdot\mathbf{r}_{l2} - r_s r_m\cos\theta_s)^2 \tag{3.9}$$

in which α, β_m and β_s are the potential constants. The relative strengths of $\alpha : \beta_m : \beta_s$ are set to be $10 : 3 : 1$ in the optimization. The minimized structure is not very sensitive to the relative values of these potential constants. The sum of i is over all metal atoms. The sum of j and k are over all nearest neighbors of the metal atoms. The sum of l is over all terminal sulfur atoms. The first term in the potential optimizes the bond lengths between the terminal sulfur atoms and the metal atoms. The second and the third term minimize the bond angle deviations at the metal atoms and at the terminal sulfur atoms respectively. The optimal bond length r_m between a metal atom and a terminal sulfur atom is 2.43Å. The optimal bond angle at the metal atoms is tetrahedral. The optimal bond angle at the terminal sulfur atoms is set to be 120°. Variations of the optimal terminal sulfur angle do not affect the structural properties of the relaxed network qualitatively. Adamantane units are held as rigid bodies in the optimization. They are rotated or translated, but their inner structures are not altered.

Figure 3.15 shows the comparison of the synchrotron powder diffraction pattern [108] of the real material with the calculated X-ray diffraction pattern of the modeled network. The calculated and the experimental diffraction patterns match well. There is an one to one correspondence between the peaks calculated from the

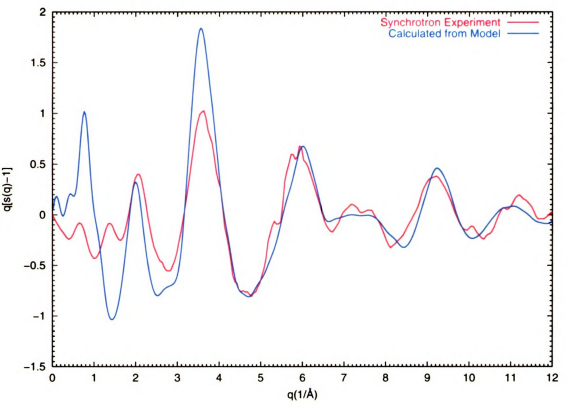


Figure 3.15: The comparison of the X-ray diffraction patterns measured from the real material (red curve) and calculated from the amorphous model (blue curve). The experimental data is in courtesy of Petkov et al. [108].

model and those measured from the real material, except the peak at $q = 1.5(1/\text{Å})$. The good match between the experimental data and the calculated data proves that the metal-adamantane network in the real material is amorphous. Though the real material is porous while the model is bulk, the local short distance order of the network model is identical to that in the real material.

We tried a variety of combinations of the potential parameters in Equation 3.9 to relax the network model. The width and the height of the peak vary with the change of parameters. The central positions of a few peaks shift a little bit under different parameters. But no matter under what combination of parameters, the peak at $q = 1.5(1/\text{Å})$ observed in experimental data is always missing in the diffraction pattern calculated from the amorphous metal-adamantane network models. Since a

peak at small q value implies a medium to long distance range order, the real material is seemingly more ordered in medium to long distance range than our bulk amorphous metal-adamantane network model. There are two possible causes of the observed peak at $q = 1.5(1/\text{Å})$. One is that the peak may be caused by the surfactant molecules. These surfactant molecules are packed in a more ordered way than the metal-adamantane network. Another possibility is that the metal-adamantane network in the real material is more ordered than that in our amorphous model. Since the thinest regions between the hexagonal cylindrical surfactant channels allow up to two layers of adamantane units, the metal-adamantane network may be more ordered than purely amorphous to fill the space more efficiently. However we cannot do further studies until we can design an algorithm to generate amorphous porous networks with only even-numbered rings, and more experimental information about the geometry of the surfactant is acquired, which is not available now.

Chapter 4: Constraints, Conformations and Flexibility

4.1 Constraints and Conformations

The feasibility to build amorphous networks has been analyzed in the previous two chapters. Amorphous silicon and silica networks have two characteristics: a topology that is different from that of the crystalline networks, and the spatial placement of atoms that makes the variations of bond lengths and angles small. A bond between two atoms is built when one or two electrons are distributed on the molecular orbitals formed by the two atoms. Each molecular orbital has its characteristic energy and electron density distribution. The bond is stable only when the distance between the two atoms is in a narrow range so that the shared electrons are distributed in the low energy orbitals. The bond breaks if the distance between the two atoms is elongated from its optimal value, when the shared electron(s) are driven into high energy molecular orbitals. This explains why the variations of the bond lengths, especially the bond lengths of covalent bonds are small. Bond angle variations are small because the electron interactions between three insulator or semi-conductor atoms favor certain low energy geometries.

Because bond length and angle variations are small, and because it costs energy to distort bond lengths and angles, it is reasonable to assume that the bond lengths and angles do not vary in time at all in moderate temperatures. We call the fixed bond lengths and angles, the bond length and angle *constraints* respectively. For the sake of simplicity, from here on the phrase *bond constraints* is used to designate both bond length constraints and bond angle constraints, unless otherwise stated.

Previous chapters presented how to build amorphous silicon and silica networks

that obey certain topology requirements and have good bond geometries. Once a network is built, the topology of the network as well as the bond constraints are determined. We call the spatial arrangement of atoms that satisfies a certain set of topology requirements and bond constraints a *conformation*. Once an amorphous network model is built, one conformation that satisfies the topology and bond constraints specified by the spatial arrangement of atoms in the network is found.

It is logical to ask the following two questions once one conformation is built: 1) Is the network model we built the only conformation that obeys the topology and constraints requirements? 2) How can we search all the other possible conformations that observe the same topology and bond constraints if we know, by some means, there should exist other conformations? This and the following two chapters address these two questions. This chapter discusses what types of networks can have more than one conformation. Chapter 5 depicts an algorithm in searching conformations for the flexible regions in non-crystalline networks. Chapter 6 shows a couple of examples.

4.2 Flexibility and Degrees of Freedom

A molecule is *flexible* when it has a positive number of internal degrees of freedom (DOF). Otherwise the molecule is *rigid*. The DOF can be easily counted by the simple equation:

$$DOF = 3 \times N - N_c - 6 \tag{4.1}$$

where N is the total number of atoms in the molecule and N_c is the total number of independent constraints. The number 6 comes from the six degrees of translational and rotational rigid body motions. The DOF represents the actual number of independently rotatable bonds in the molecules. The chain molecule in Figure 4.1(a) has four atoms, three independent bond length constraints and two bond angle con-

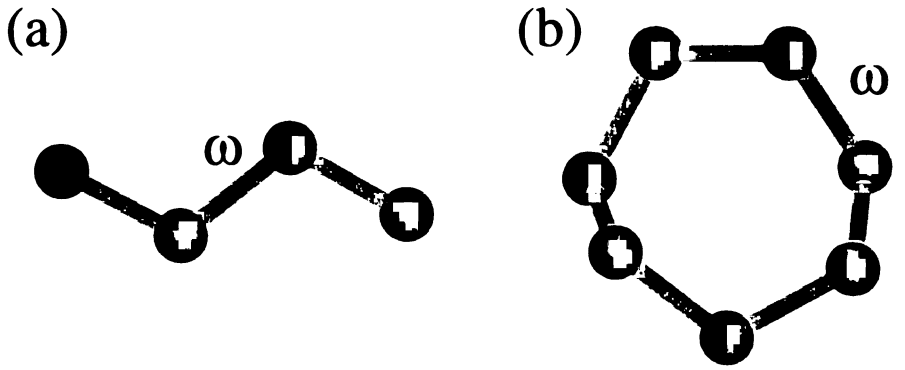


Figure 4.1: Both the chain molecule shown in (a) and the ring molecule shown in (b) have one DOF. The independently rotatable bonds ω in both molecules are marked as red. Figure (b) is one out of seven possibilities in selecting the independently rotatable bonds in the seven-fold ring.

straints. Therefore the molecule has $3 \times 4 - 3 - 2 - 6 = 1$ DOF. This single DOF is the rotation of the middle bond ω indicated in the figure. Rotations of the bond ω produce an infinite number of conformations, all of which comply with the same topology and bond constraints. This molecule is flexible with one DOF.

The flexibility of molecules made up of rings is also represented by the DOF, which is counted by Equation (4.1). The ring molecule in Figure 4.1(b) has seven atoms, seven independent bond length constraints and seven bond angle constraints. Hence the molecule has $3 \times 7 - 7 - 7 - 6 = 1$ DOF. Although all seven bonds in the molecule are rotatable, only one bond is independently rotatable. The other six bonds have to be rotated corresponding to the rotation of the single independently rotatable bond to close the seven-fold ring, otherwise one or more bond constraints will be violated. Any of the seven bonds can be taken as the independently rotatable bond, with the consequence that the remaining six bonds are not independently rotatable. The bond ω shown in the figure shows one out of seven choices of the independently rotatable bond. A rotation of this bond will produce multiple conformations for the seven-fold ring, as long as the other six bonds are rotated appropriately to close the ring.

There are N bond length constraints and N bond angle constraints in an N-fold ring. Thus $3 \times N - 2N - 6 = N - 6$ is the DOF of a N-fold ring. Therefore a ring whose size is less than or equal to 6 is rigid, whereas a ring whose size is greater than 6 is flexible. Flexibility not only means multiple conformations, but also means these conformations are in a continuous domain. The conformations of the chain molecule shown in Figure 4.1(a) can be labeled by the dihedral angle ω of the middle bond as shown in the figure. A conformation of the chain molecule whose ω value is ω_0 has two infinitely closed neighboring conformations at $\omega_0 + \delta$ and $\omega_0 - \delta$ in which the deviation δ is infinitely small. Each conformation of the seven-fold ring, shown in Figure 4.1(b), is a point in the seven dimensional space spanned by the seven dihedral angles of the seven bonds. For any point in the seven dimensional space there are neighboring points that are infinitely close to it. By repetitively transforming from one conformation to another close by, the seven-fold ring is able to transform continuously from an initial conformation to a very different one without breaking any bond constraints and topology requirements. It is worth noting that there may be multiple clusters of conformations that fit bond constraints and topology requirements. The conformations in each cluster are continuous, but bond constraints have to be broken for the molecule to transform between clusters of conformations. For example a generic seven-fold ring has two clusters of conformations, as will be discussed in Section 5.2.1. It does break bond constraints for the seven-fold ring to transform from any conformation in one cluster to an arbitrary conformation in another cluster, but not when the seven-fold ring transform between conformations in the same cluster.

A rigid molecule may have multiple conformations too, though its DOF is less than or equal to zero. A generic six-fold ring has two conformations, the boat and the chair, as shown in Figure 4.2. A six-fold ring is rigid because its DOF is zero though it has two instead of one conformation. The difference between a six-fold ring from

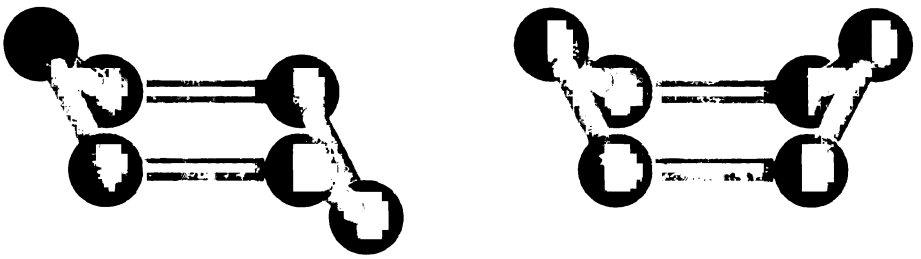


Figure 4.2: The chair (left) and the boat (right) conformations of a generic six-fold ring. A six-fold ring is rigid due to its zero DOF, though it has two distinct conformations. Bond constraints have to be violated for the six-fold ring to transform from the chair to the boat conformation, and vice versa.

a flexible ring is that the two conformations are not continuous. They are isolated single points in the six dimensional conformational space spanned by the six dihedral angles of the six bonds. There are not any other conformations that are infinitely close to either of the conformations, nor are these two conformations close to each other.

In summary, a network is flexible when it has a positive number of DOF. It is guaranteed that the network has multiple and continuous conformations. A rigid network may have multiple conformations, but we are not interested in sampling these conformations because bond constraints have to be violated for the network to transform from one isolated conformation to another. The continuous conformations of the flexible molecules are the intrinsically allowed motion of the networks, which are what we are specifically interested in.

4.3 Flexibility Analysis

Flexibility is a local property. DOF are usually confined to local regions in a network. For example when only bond lengths constraints are counted, the network shown in Figure 4.3(a) has one DOF. But only the right half of the network which are connected by gray bonds is flexible, while the left half is rigid.

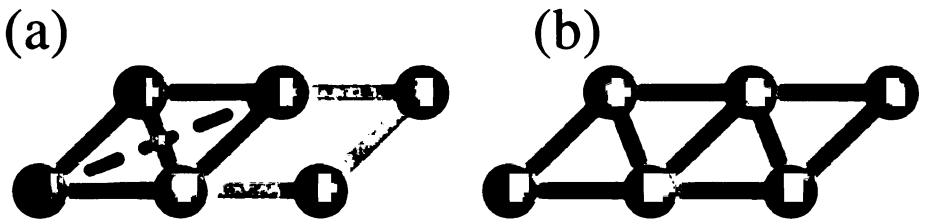


Figure 4.3: Two networks in 2D space with equal number of bond lengths constraints and atoms. The right half of the network shown in Figure (a) is flexible, whereas the left half of the network has an excessive constraint shown as the dashed bond. The network shown in Figure (b) is uniformly rigid. The blue and gray bonds are rigid and flexible bonds respectively.

Constraints and DOF should be counted in all local regions of a network in order to count and predict the distributions of DOF exactly. It is worth noting that only independent constraints should be counted. For example there are equal number of atoms and bond length constraints in the two networks in a 2D plane shown in Figure 4.3. The network shown in Figure 4.3(a) is flexible on its right half yet rigid on its left half. The left half of the network is rigid with or without the presence of the bond length constraint shown as the dashed bond. Therefore the bond length constraint shown as a dashed line is not an independent constraint. It should not be counted when calculating the DOF of the network. The network shown in Figure 4.3(b) does not have any excessive bond length constraint. It is uniformly rigid.

Flexibility properties of generic networks which include both bond length and bond angle constraints can be accurately analyzed by the software Floppy Inclusion and Rigidity Substructure Topography (FIRST) developed by Jacobs et al. [47, 51]. The software has been utilized to study phase transitions in random glass networks [51, 109].

Flexibility analysis on the amorphous fiber silica and amorphous film silica networks proves that all these two networks are rigid, without any flexibility in any local regions. This is indeed expected. The topology of our amorphous SiO_2 models is

that 1) every oxygen atom is connected to two silicon atoms; 2) every silicon atom is connected to four oxygen atoms. Therefore there are $4N$ bond length and $5N$ bond angle constraints in an amorphous silica network with N SiO_2 units. The DOF of such networks are $9N - 9N - 6 = -6$ which is negative regardless of the size of the network. Another type of non-crystalline networks, the proteins, on the other hand, is rich in flexibility and conformations. Therefore we shift the subject of study from amorphous silica networks to proteins in the following discussions.

4.4 Proteins

4.4.1 Amino Acids

Proteins are composed of twenty standard amino acid residues. The amino acid residues differ from each other by their side chain groups. The side chain groups are the R groups shown in Figure 4.4. The twenty standard amino acid residues differ from each other in their side chain groups. Main chains of amino acid residues are made up of a nitrogen atom, a carbon atom named as C_α to which the side chain atom is covalently bonded, and a carboxyl group. The dihedral angle of the bond between a main chain nitrogen atom and the C_α atom is called the angle ϕ, and the dihedral angle of the bond between the C_α atom and the carboxyl carbon atom is called the angle ψ. Statistical studies reveal that the angles ϕ and ψ are distributed in certain regions in the whole ϕ and ψ plane which is called the Ramachandran plot [110]. Amino acid residues are connected to form a chain through the covalent bonds between the carboxyl carbon atom of a residue and the nitrogen atom of the next residue, as shown in Figure 4.4.

Typical proteins are composed of one or several chains of amino acid residues. Each chain can contain hundreds of amino acid residues.

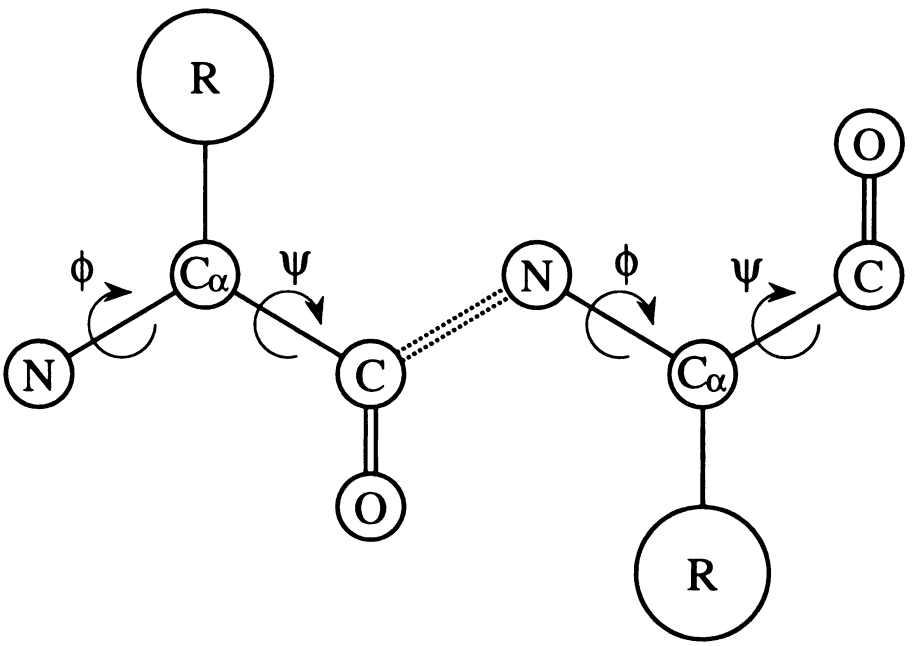


Figure 4.4: A short chain of two amino acid residues. Amino acid residues differ from each other by their side chain R groups. The main chain of residues is made of a nitrogen atom N, a carbon atom C_α and a carboxyl group. The dihedral angles of the bond between the nitrogen and the C_α atom and between the C_α and the carbon atom are called angle ϕ and ψ respectively. Amino acid residues are connected to form chains by the peptide bonds which are shown as the double dashed bonds in the figure.

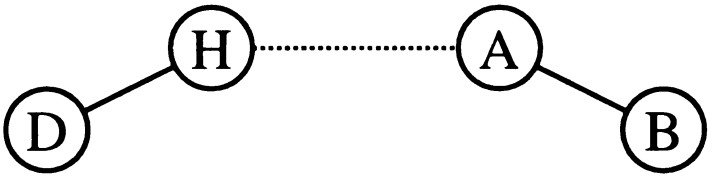


Figure 4.5: A hydrogen bond is an interaction between a hydrogen atoms and a polar atom that is not covalently bonded to the hydrogen atom. The polar atom that is covalently bonded to the hydrogen atom is the hydrogen bond donor atom, designated by the circled letter *D* in the figure. The hydrogen bond acceptor atom which is the polar atom that is not covalently bonded to the hydrogen atom is donated by the circled letter *A* in the figure. A non-hydrogen atoms covalently bonded to the hydrogen bond acceptor atom is usually called the base atom, indicated by the circled letter *B* in the figure. The hydrogen bond is shown as the dashed line.

4.4.2 Hydrogen Bonds and Hydrophobic Interactions

Hydrogen bonds and hydrophobic interactions link one or multiple chains of amino acids into complicated networks. Hydrogen bonds are formed between a hydrogen atom that is attached to a polar atom and another nearby polar atom. The polar atom that is covalently bonded to the hydrogen atom is called the hydrogen bond donor, shown as the atom *D* atom in Figure 4.5. The polar atom that is close but not covalently bonded to the hydrogen atom is called the hydrogen bond acceptor atom, shown as the atom *A* in Figure 4.5. Both the donor and the acceptor atoms can be nitrogen or oxygen atoms. A non-hydrogen atom that is covalently bonded to the acceptor atom is called the base atom which is shown as the atom *B* in Figure 4.5.

Hydrogen bonds favor certain geometries [111]. The distance between the donor and the acceptor atom is typically in a range between 2.5Å and 3.5Å. The distance between the donor and the hydrogen atom is in a narrow range around 1.0Å. The distance between the hydrogen and the acceptor atom is between 1.5Å to 2.5Å. The donor-hydrogen-acceptor angle, which is usually called the hydrogen bond angle, is commonly in a large range from 120° to 180°. Strong hydrogen bonds can be distinguished from weak hydrogen bonds based on geometry considerations. For example

the bond angles of strong hydrogen bonds are usually larger than 150°. There have been several empirical hydrogen bond potentials [112, 113, 114]. All of them evaluate energies of hydrogen bonds based on the geometries of bond lengths and angles.

Hydrophobic interactions are very important in biological systems. They arise from the weak polarizability of water molecules around biochemical molecules. The attraction between the water molecules is stronger than that between the water molecules and the non-polar side chains or main chains of proteins. Therefore the water molecules tend to weave a net of hydrogen bonds around the proteins and force the proteins to collapse tightly so that all the non-polar atoms are buried inside of the proteins. This effect makes the non-polar atoms in proteins appear to attract each other. This phenomenon is called the hydrophobic effect.

Hydrogen bonds and hydrophobic interactions lock a chain of amino acid residues into a network. Showing only covalent bonds in its one dimensional chain configuration, the protein has the maximum number of DOF. Any bond, such as the main chain ϕ and ψ dihedral angles shown in Figure 4.4 can rotate freely in such configurations. The protein explores the maximum allowed conformational space. When a hydrogen bond forms between a hydrogen atom and a hydrogen bond acceptor atom, or when a hydrophobic interaction stably forms when two non-polar atoms are close to each other, the addition of the constraints restricts the available motions of the protein. When more and more hydrophobic interactions or hydrogen bonds form within the protein, motions in some regions of the protein can be restricted so much that these regions are effectively rigid. We call these regions the *rigid* cores of the proteins. The concentration of constraints may not reach the critical value in some other regions of the proteins so that these regions of protein can still have internal motions. We call these regions the *flexible* regions. The hydrogen bonds and hydrophobic interactions collectively link the amino acid residues in proteins into networks of interactions.

It is common for a protein to have one or multiple rigid cores and one or several

major flexible regions. The rigid core functions as the stabilization cores of the protein. The flexibility in other regions in the proteins allows the proteins to access a variety of conformations. The constant change of conformations in proteins is the driving force of protein motions and functions.

4.4.3 Interactions in Proteins

Potential energy of proteins in typical molecular dynamics (MD) simulations such as the ff94 [115] potential used in Amber [116] and the CHARMM [117] potential used in CHARMM [118] have forms like:

$$E = \sum_{bonds} K_r(r - r^{eq})^2 + \sum_{angles} K_\theta(\theta - \theta^{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}^1 2} - \frac{B_{ij}}{r_{ij}^6} + \frac{q_i q_j}{r_{ij}}\right] \quad (4.2)$$

where the first, the second, the third and the last terms are the bond length vibration energy, the bond angle vibration energy, the dihedral angle rotation energy and the sum of the electrostatic and van der Waals interactions respectively. Bond length vibration energy increases rapidly with small variation in bond length. For example the energy increases by about 3 kcal/mol when the bond length of a typical covalent bond deviates about 0.08Å from its equilibrium value. The bond angle vibration energy is weaker than the bond length vibration energy, though it incurs noticeable energy changes as well. The energy increases by about 3 kcal/mol when a typical bond angle changes by about 0.3°. The exact value depends on the types of atoms of the bonds. The dihedral angle rotation energy is usually small. A deviation of about 13° in the dihedral angle increases the energy by about 3 kcal/mol. Though energies of both bond length vibration and bond angle vibration can increases to infinity, the dihedral angle rotation energy is capped at V_n, which ranges from 1 kcal/mol to 20 kcal/mol, depending upon the types of bonds. The van der Waals energy does

not vary much in usual environments where non-bonded atoms do not closely contact with each other.

Our constraint concept can be regarded as a result of two steps of simplification from the usual potentials used in MD. First, since the bond length and angle vibration energies are high compared to the other terms, only these two terms are taken into account while the other energies are ignored. Second, since bond lengths and angles do not vary much under common conditions, the high frequency vibration of bond lengths and angles are neglected. Shown in Figure 4.6 is a schematic diagram of the energy in the conformational space. The dotted line represents the energy when all terms in Equation 4.2 are included. The dihedral angle rotation energies and the van der Waals interactions produce small ripples on the smoother surface which is shaped by the bond length and angle vibrations. When these smaller interactions are ignored, the energy landscape is much smoother, shown as the dashed line. The solid lines in the figure show the simplifications introduced by our constraint concept. The energy is either zero or infinite, depending on whether bond constraints are obeyed or violated.

Our constraint concept ignores the details of the energy surface. It has two obvious advantages. The first one is that flexibility analysis which is based on the constraint concept can be applied to predict those flexible regions in proteins upon which computation resources can then be exclusively applied. The second one is that all the conformations obeying the bond constraints comprise the intrinsic conformations of the proteins which are not affected by any other effects.

4.4.4 Flexibility Analysis on Proteins

The flexibility analysis software FIRST that has been applied on random glass networks has been successful in predicting flexibility properties [88] and the folding cores [89] of proteins. As stated in Section 4.4.2, hydrogen bonds and hydropho-

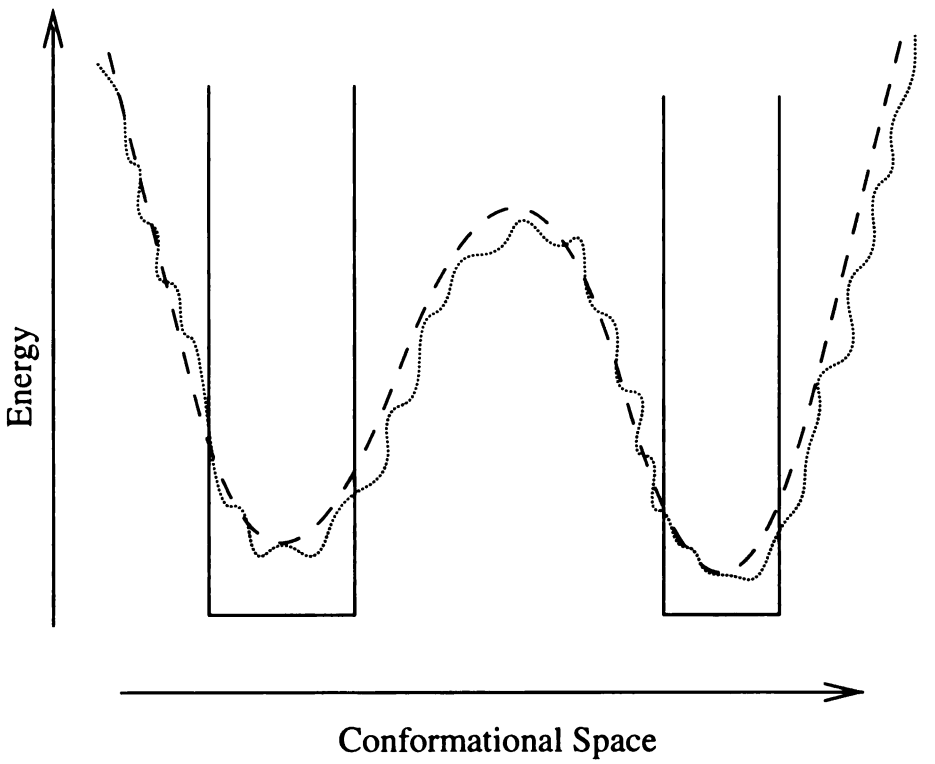


Figure 4.6: Schematic figure of the simplifications of the common empirical potentials of proteins. The x-axis is a general coordinate in the conformational space. The y-axis is the potential energy. The dotted line represents the potential energy of proteins when all terms in Equation 4.2 are included into the calculation. When the dihedral angle rotation energy and the van der Waals interaction are neglected, the potential surface becomes smoother, shown as the dashed line. The solid line represents the energy when bond length and angle vibrations are disallowed.

bic interactions are as important as covalent bonds in proteins. FIRST includes the strong hydrogen bonds and the hydrophobic interactions when counting constraints. The methodology of treating hydrogen bonds and hydrophobic interactions are comprehensively discussed by Jacobs et al. [88] and by Rader et al. [89]. The author summarizes the main points here for the sake of completeness and clarity.

Hydrogen bonds have a great range of energies. Some of the bonds can be almost as strong as covalent bonds, while some others are as weak as van der Waals interactions. The Mayo's hydrogen bond potential [112] is used in FIRST to analyze the strength of the hydrogen bonds. The stronger the hydrogen bonds are the lower their energies are. FIRST utilizes a step function in selecting hydrogen bonds. Any hydrogen bonds whose energies are less than a cut off value are taken by FIRST to be the strong hydrogen bonds, whereas all the other hydrogen bonds are ignored. FIRST then treats the strong hydrogen bonds the same way as it handles the covalent bonds by counting the bond length and bond angle as constraints. Every strong hydrogen bond brings three additional constraints to the proteins: 1) the bond length constraint between the hydrogen atom and the donor atom; 2) the bond angle constraint of the angle formed by the donor, the hydrogen and the acceptor atom; and 3) the bond angle constraint of the angle formed by the hydrogen, the acceptor and the base atom. The bond angles are counted as constraints because the strengths of hydrogen bonds are dependent on the angles.

Contrary to the covalent and the hydrogen bonds which are highly directional dependent, hydrophobic interactions do not have angular preferences. Thus the bond angles of hydrophobic interactions are not counted by FIRST as constraints. Moreover, the distance between two non-polar atoms that are believed to have hydrophobic interactions can be in a relatively large range. Intuitively only one inequality distance constraint should be counted for each hydrophobic interaction. After many tests Rader et al. [89] find that protein unfolding process is better described when

two instead of one constraint per hydrophobic interaction are used.

4.4.5 Validity of Flexibility Analysis on Proteins

Since hydrogen bonds and hydrophobic interactions are not as clearly defined as covalent bonds, exactly which hydrogen bonds and hydrophobic interactions to be included in flexibility analysis as sources of constraints is prone to errors.

There are two sources of possible errors in selecting hydrogen bonds as constraints. The first one comes from the empirical hydrogen bond potential used in FIRST. Mayo's hydrogen bond potential assigns low energies to the hydrogen bonds whose bond lengths and angles resemble those of typical strong hydrogen bonds. But it may give a very high energy to a hydrogen bond whose geometry is not perfect yet acceptable [119]. Therefore the number of acceptable hydrogen bonds is a little underestimated in FIRST. The second source of errors is the on and off hydrogen bond selecting switch used by FIRST. Suppose there are two hydrogen bonds whose energies are -0.99 kcal/mol and -1.01 kcal/mol respectively. A hydrogen bond cut-off value at -1.00 kcal/mol will render the first bond as a weak hydrogen bond and the second one as a strong hydrogen bond respectively, though the two bonds are very close in energies. FIRST then counts three constraints at the location of the second hydrogen bond, but zero at the place of the first one. Since flexibility is a local property, this step like function in selecting hydrogen bonds tends to lead to either an underestimate or overestimate of local constraints at various locations.

However all the errors related to the hydrogen bonds are acceptable from the statistical point of view. The average number of hydrogen bonds per residue is 1.1 [120]. Hence there are more than 110 hydrogen bonds in a moderately sized protein having more than 100 residues. This number is large enough that small amount of errors, for example five mistakenly handled hydrogen bonds, do not affect the overall distributions of flexibility in proteins. This explains why the flexibility analysis has been

proved to be a good tool in studying the relationship between protein structures and functions [90].

Hydrophobic interactions used in FIRST have undergone significant fine tunings. The improved hydrophobic interaction definition [121] are said to increase the accuracy of FIRST noticeably [119]. The flexibility data used in this thesis are analyzed by a beta version of FIRST that utilizes this refined hydrophobic interaction definition. This version of FIRST is not available to the public yet.

4.4.6 Advantage of Flexibility Analysis

As stated in Section 4.1, we are interested in finding the conformations of non-crystalline networks, i.e. proteins in this specific study. Interactions such as the covalent, the strong hydrogen bonds and the hydrophobic interactions inter-lock with each other so that some regions of the proteins are practically rigid while some other regions of the proteins are still flexible. By counting the concentration of constraints in all possible local areas in proteins, we are able to distinguish the flexible regions from the rigid cores of the proteins prior to any computationally expensive trials involving searching conformations. In the next chapter, we show that by concentrating our calculations on the flexible regions only, we are able to avoid wasting valuable computational resources on the rigid cores of proteins, which should not have multiple conformations at all due to the high density of constraints in these regions. It is for this reason that our algorithm in sampling protein conformations is called Rigidity Optimized Conformational Kinetics (ROCK), because the flexibility analysis helps our algorithm avoid sampling conformations for the rigid cores of the proteins. Details of the algorithm are discussed in the next chapter.

Chapter 5: Rigidity Optimized Conformational Kinetics (ROCK)

As stated in Chapter 4, we want to address two questions in the second half of this thesis: 1) How to determine whether a network has one or multiple conformations; 2) How to search the conformations of a network if it is known from flexibility analysis that this network should have multiple conformations?

The flexibility analysis algorithm described in Chapter 4 answers the first question. The algorithm pinpoints the regions that have positive numbers of DOF by counting the local bond constraints. This chapter describes an algorithm that searches the possible conformations that obey the same topology and the bond constraints as the input molecule.

The concept of constraints is in fact that of a simplified potential. The system has exactly zero energy when all bond constraints are satisfied. A violation of any bond constraint, even as tiny as one degree in a bond angle constraint or a tenth Angstrom in bond length constraint, will increase the energy of the network to infinity.

5.1 Ring Clusters and Side Branches

Contrary to amorphous silicon or silica networks in which every atom is present in several rings, protein networks have many dangling ends. Figure 5.1 shows the topology graph of a small portion of HIV-1 protease. There are two types of atoms in this network from a topological point of view: the ring cluster atoms and the side branch atoms. Ring cluster atoms are those atoms participating in large rings linked by covalent and strong hydrogen bonds. The rings equal to or smaller than six-fold are not counted as parts of the ring clusters unless they are connected directly or

indirectly by other rings. For example the four fold ring shown in Figure 5.1 is not counted as a component of the ring cluster. There are at least two sets of completely different bonds that connect any pair of ring cluster atoms. All the other atoms are the side branch atoms.

Traditionally atoms in proteins are classified as main chain atoms and side chain atoms based on their biochemical properties, as shown in Figure 4.4. In most cases, the main chain atoms are ring cluster atoms though exceptions exist. In any given protein some of the side chain atoms are connected to form ring clusters by strong hydrogen bonds. Side branch atoms are usually exclusively made up of side chain atoms.

There are no ring clusters when a protein is in its random coil state, except the single five-fold rings of proline. All atoms are side branch atoms in this stage. Ring clusters form as hydrogen bonds appear when some residues are close to each other. Both side chain and main chain atoms can participate in hydrogen bonding. As a consequence the ring clusters are composed of both atom types. When protein folds, more and more hydrogen bonds form, therefore the average size of a ring cluster grows while the number of side branch atoms is reduced. In its native state, the protein usually has one or several large ring clusters with many small side branches dangling around. Whether an atom is classified as a ring cluster atom or a side branch atom is affected by the definition of hydrogen bonds.

It is very easy to generate conformations for side branch atoms. A rotation of any dihedral angles in a side branch produces a valid new side branch conformation, because the rotation does not change the topology of the networks, nor does it change the bond lengths or bond angles.

However it is not easy to search conformations for rings. A disturbance of a dihedral angle in a flexible ring introduces large bond distortions at the point where the ring breaks. Though such distortions can be minimized by a subsequent optimiza-

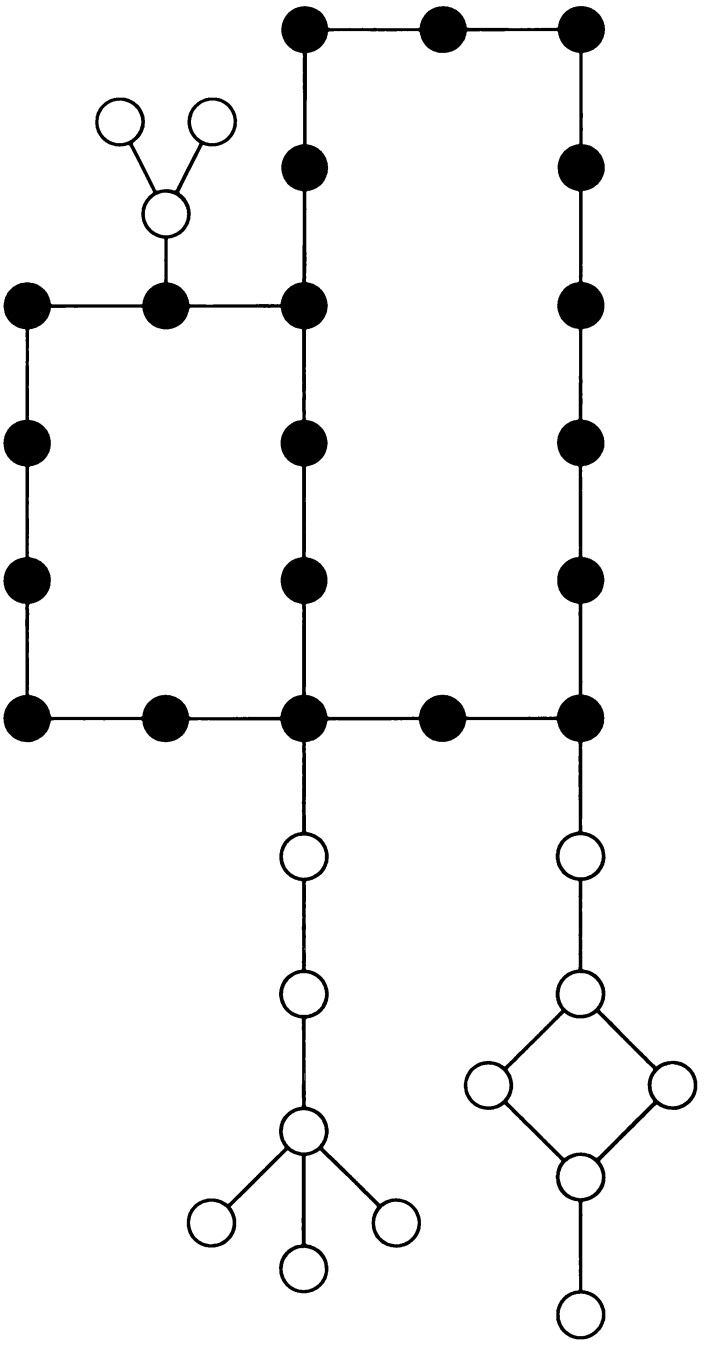


Figure 5.1: Topology graph of a small portion of the protein HIV-1 protease. The solid circles are atoms in the ring clusters, because there are two different sets of bonds between any pair of atoms in the ring cluster. The open circles are atoms in the side branches.

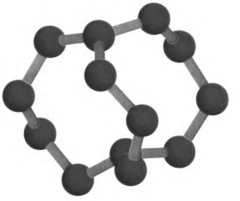


Figure 5.2: A simple molecule made up of two rings. The dihedral angles in the rings are all correlated. A disturbance in one dihedral angle, no matter in the left ring or in the right ring, requires the changes of all dihedral angles in both rings for the two rings to close.

tion process, the generated structures are likely to have high energies because such an initial distortion can produce molecular conformations far away from energetically favorable regions. The major difference between a molecule with and a molecule without flexible rings is that the dihedral angles in the rings are correlated. A rotation of one dihedral angle in a ring demands rotations of other dihedral angles for the ring to close. The dihedral angles in a molecule without any rings can vary independently without causing distortions in bond lengths and angles.

It is even more difficult to close all the rings in a ring cluster. Figure 5.2 shows a simple flexible molecule with only two rings. All the dihedral angles in both the left and the right rings are correlated. A rotation of one dihedral angle in the left ring not only requires all the other dihedral angles in the left ring to rotate correspondingly for the left ring to close, but also requires the rotation of dihedral angles in the right ring, otherwise the right ring will break due to the change of dihedral angles of the bonds shared by both the left and the right ring. All of the rings in the molecule have to be closed simultaneously in such circumstances to avoid bond distortions.

The correlation in dihedral angles in rings requires a new algorithm in finding conformations for ring clusters. Section 5.2 introduces the ring closure equations

and their solutions. Section 5.3 explains how the method to solve the ring closure equations can be applied to sample conformations of networks with multiple rings. Section 5.4 briefly mentions the method to anchor side branch atoms back to the ring clusters. The whole procedure of our algorithm ROCK is listed in Section 5.5.

5.2 Sampling Conformation of a Single Ring

When the bond lengths and angles of a ring are known, as is usual, the unknown dihedral angles of the ring should be one set of solution to the following ring closure equations [73] for the ring to close:

$$\begin{aligned} \mathbf{p}_0 + \mathbf{T}_0\mathbf{R}_1\mathbf{p}_1 + \mathbf{T}_0\mathbf{R}_1\mathbf{T}_1\mathbf{R}_2\mathbf{p}_2 + \cdots + \mathbf{T}_0\mathbf{R}_1\cdots\mathbf{T}_{N-2}\mathbf{R}_{N-1}\mathbf{p}_{N-1} &= \mathbf{0} \\ \mathbf{T}_0\mathbf{R}_1\mathbf{T}_1\mathbf{R}_2\cdots\mathbf{T}_{N-2}\mathbf{R}_{N-1}\mathbf{T}_{N-1}\mathbf{R}_N &= \mathbf{I} \end{aligned} \quad (5.1)$$

where

$$\mathbf{T}_i = \begin{pmatrix} \cos\theta_i & -\sin\theta_i & 0 \\ \sin\theta_i & \cos\theta_i & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \mathbf{R}_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\omega_i & -\sin\omega_i \\ 0 & \sin\omega_i & \cos\omega_i \end{pmatrix} \quad (5.2)$$

are rotation matrices. The quantities $\mathbf{0}$, $\mathbf{p}_i$ and $\mathbf{I}$ are the zero vector $(0,0,0)^T$, distance vector $(d_i,0,0)^T$ and unit matrix respectively. The quantities d_i, θ_i and ω_i are the bond length, angle and dihedral angle of the ith bond in the ring. The definition of bond lengths, angles and dihedral angles are shown in Figure 5.3.

The bond lengths and angles are fixed parameters and the dihedral angles are unknown variables. The fact that a ring is flexible if and only if the size of the ring is bigger than six determines that only six out of twelve equations in Equation 5.1 are independent. If $N-6$ dihedral angles of a single N-fold ring are known, the values of

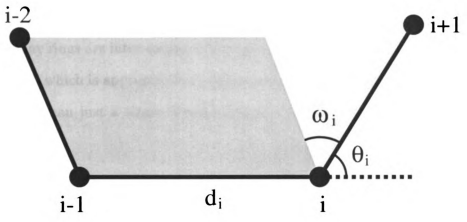


Figure 5.3: Definition of bond lengths, angles and dihedral angles in ring. Bond length d_i is the distance between the ith and $(i-1)$th atoms. Bond angle θ_i is the supplementary angle of the angle formed by the $(i-1)$th, the ith and the $(i+1)$th atoms. The dihedral angle ω_i is the angle between the two planes: one plane passes through the $(i-2)$th, the $(i-1)$th and the ith atoms and the other plane goes through the $(i-1)$th, the ith and the $(i+1)$th atoms.

the six unknown dihedral angles can be solved from the six independent ring closure equations. Gō et al. [73] prove that the ring closure equations can be reduced to four nonlinear equations each of which has only one variable. Each of these single variable equations can have zero or multiple solutions. Each solution to the single variable equation corresponds to a set of solution of the six unknown dihedral angles to the ring closure equations. Therefore, if $N-6$ dihedral angles are set, there can be multiple or no solutions of the six remaining dihedral angles to the ring closure equations. These solutions are discrete in generic cases, where there is no special symmetry. A case of no solution implies that the combination of $N-6$ dihedral angles is so inappropriate that the ring does not close no matter how to adjust the other 6 dihedral angles.

Ring closure equations cannot be reduced to single variable equations when the six unknown dihedral angles are not consecutive in the ring, unless they are separated by locked bonds such as peptide bonds. Only numerical methods can solve the ring closure equations in generic cases. Gō and Scheraga's method is also limited to single

rings. Their method cannot be used to sample conformations of complicated networks in which many rings are inter-locked with each other. For these reasons we introduce a new approach which is appropriate to sample conformations of macromolecules which contain more than just a single flexible ring. The fictitious ring closure potential is given by ƒ

$$\begin{aligned} \mathfrak{f} = & \left[\mathbf{p}_0 + \mathbf{T}_0\mathbf{R}_1\mathbf{p}_1 + \cdots + \mathbf{T}_0\mathbf{R}_1 \cdots \mathbf{T}_{N-2}\mathbf{R}_{N-1}\mathbf{p}_{N-1}\right]^2 \\ & + \sum_{i,j=1}^{3} \left[\mathbf{T}_0\mathbf{R}_1\mathbf{T}_1\mathbf{R}_2 \cdots \mathbf{T}_{N-2}\mathbf{R}_{N-1}\mathbf{T}_{N-1}\mathbf{R}_N - \mathbf{I}\right]_{ij}^2 \end{aligned} \tag{5.3}$$

which is the sum of the squares of the differences between the left and the right sides of the original ring closure equations 5.1.

We solve the ring closure equations by minimizing the fictitious ring closure potential ƒ which is zero only at those points that are solutions to the original ring closure equations. To sample the conformations of a single N-fold ring with $N > 6$, one can systematically or randomly try all the possible combinations of $N-6$ dihedral angles of the ring, and minimize the fictitious ring closure potential ƒ with respect to the six unknown dihedral angles at every step. Each zero fictitious potential ƒ corresponds to a set of solutions to the ring closure equations, which in turn suggests a new conformation of the ring. We have utilized the limited-memory BFGS source code [122], which is a quasi-Newton unconstrained nonlinear optimization algorithm, to minimize the function ƒ. The optimized fictitious potential may not have a zero value at some combinations of the $N - 6$ dihedral angles. It may be due to the nonexistence of a solution of the remaining six dihedral angles, or it may be due to the inefficiency of the L-BFGS method to find a solution. The L-BFGS algorithm is a local potential optimization algorithm so it is not capable of finding the global minimum point under any conditions. Only when the function value of ƒ is zero do we consider a set of solutions to all the ring closure equations has been found.

5.2.1 Conformations of a Seven-Fold Ring

At every step while working on a single N-fold ring, ROCK first randomly selects and rotates $N - 6$ dihedral angles from their values in the previously accepted conformation. Then it minimizes the fictitious potential $\mathfrak{f}$ with respect to the six remaining dihedral angles. The new conformation, if the fictitious potential is zero, is subjected to checks on van der Waals overlaps before accepted. A van der Waals overlap occurs when the distance between two non-bonded atoms is smaller than the sum of their van der Waals radii times a coefficient. A small coefficient such as 0.6 or 0.7 represents a soft van der Waals repulsion between non-bonded atoms. A large coefficient such as 0.9 or 1.0 represents a stiff van der Waals repulsion. A generated conformation is rejected if the number of van der Waals overlaps is not zero. The whole process is a random walk process. It is not exactly a Monte Carlo approach because it does not accept or reject conformations based on Metropolis [123] criterion on the energy of the generated conformations. Since ROCK searches conformations in the vicinity of the last one it accepted, it is capable of exploring the conformations space in a quasi-continuous manner. It is not able to jump from one cluster of conformations to another cluster of conformations, if there are multiple conformational clusters. This point is manifested in the conformational space of a seven-fold ring. The bond lengths and angles are exactly 1.54Å and 67° for all bonds (it is worth noting that bond angle values are given according to the definition shown in Figure 5.3). Due to the perfect seven-fold symmetry, three pairs of dihedral angle correlations are enough to describe the conformational space of the seven-fold ring. They are: the correlations of the nearest neighbor dihedral angles ϕ_1 vs. ϕ_2, the correlations of the second nearest neighbor dihedral angles ϕ_1 vs. ϕ_3, and the correlations of the third nearest neighbor dihedral angles ϕ_1 vs. ϕ_4, as shown in figure 5.4.

Gō and Scheraga's algorithm [73] of reducing ring closure equations to four nonlinear single variable equations is able to find all possible conformations of the seven

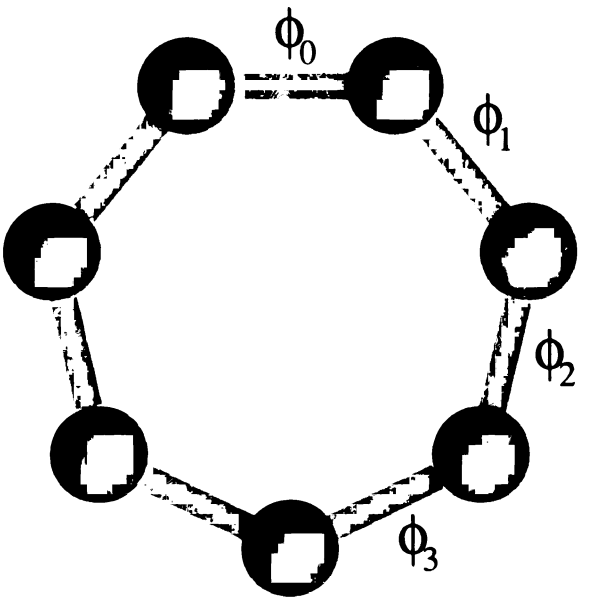


Figure 5.4: A seven fold ring whose bond lengths and angles are exactly the same. Correlations of ϕ_1 vs. ϕ_2, ϕ_1 vs. ϕ_3 and ϕ_1 vs. ϕ_4 reveals the whole conformational space of the seven fold ring due to its seven-fold symmetry.

fold ring in a single run. The left column in Figure 5.5 shows the conformations sampled by Gō and Scheraga's method. The seven-fold ring has two well separated clusters of conformations. One cluster of conformations forms a twisted loop in the seven dimensional conformational space which is spanned by the seven dihedral angles. The other cluster of conformations forms an ellipse in the conformational space. Gō and Scheraga's algorithm is able to find all possible conformations in both clusters in one run.

Our algorithm ROCK solves the ring closure equations by minimizing the fictitious ring closure potential. It does not jump between clusters. Starting from a conformation in the twisted loop cluster, it is able to find all the other conformations in the same cluster, as shown in the middle column of Figure 5.5. But it is not able to find any conformation in the ellipse cluster. Or, if the starting conformation is in the ellipse cluster, ROCK will find all conformations in the ellipse cluster, but not any single conformation in the twisted loop cluster, as shown in the right column in

Figure 5.5. The fact that ROCK is not capable of jumping between clusters of conformations can be viewed as a disadvantage in sampling conformations. Or it in fact can be regarded as a safeguard against jumping over high energy barriers. The bond distortions, which are necessities for the ring to transform from a conformation in one cluster to a conformation in another cluster, are equivalent to high energy barriers. Since our goal is to explore low energy protein conformations at room temperature, it is an advantage of ROCK that it does not overcome obvious potential barriers easily.

5.3 The Complexity Associated with a Network of Rings

As stated in Section 5.1, it is difficult to generate conformations for a flexible network with lots of inter-connected rings due to the correlations between all of the dihedral angles in the network. Gō and Scheraga's method, although it is able to generate conformations for single rings when the six unknown dihedral angles are consecutively arranged, is not efficient in searching conformations for networks. ROCK can be applied on any ring system because of its simplicity. We define a fictitious total ring closure potential of all the rings in a network as

$$\mathcal{F} = \sum_{\text{all rings}} \mathfrak{f} \tag{5.4}$$

which is the sum of the fictitious ring closure potential of every ring in the network. The total fictitious potential of the whole network can then be minimized with respect to all the rotatable and unknown dihedral angles, in the hope to find a zero potential point that is the new conformation of the ring network.

However the computation cost of solving multiple nonlinear equations simultaneously makes it infeasible to solve ring closure equations of all rings concurrently.

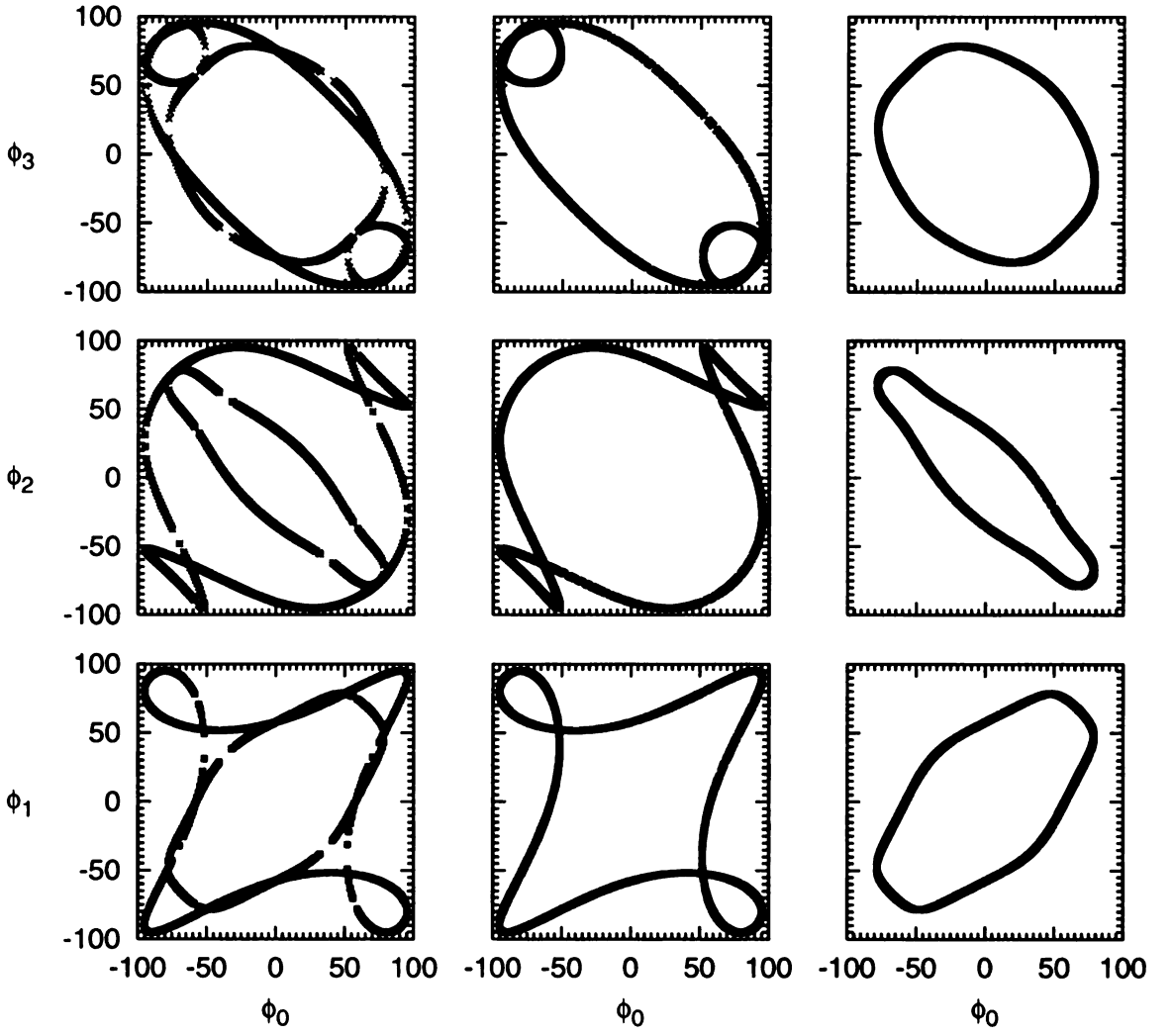


Figure 5.5: Complete conformations of a seven-fold ring calculated by Gō and Scheraga's algorithm (left column) and by minimizing the fictitious ring closure potential (middle and right columns). The top, middle and bottom rows show correlations between nearest neighbor, second nearest neighbor and third nearest neighbor dihedral angles respectively. The seven-fold ring has two clusters of conformations. One cluster of conformation form a twisted loop in the conformational space. The other cluster looks like an ellipse. Gō and Scheraga's algorithm is able to find both clusters of conformations in one run (shown in left column). Our method of minimizing the fictitious ring closure potential however does not jump between clusters. The conformations sampled by ROCK are either confined to the twisted loop cluster (shown in middle column), or to the ellipse cluster (shown in right column), depending on the initial conformation of the seven-fold ring.

Suppose a generic network of rings has N bonds and $\mathcal{M}$ DOF. In principle one can randomly select and disturb $\mathcal{M}$ dihedral angles in the network, and minimize the total fictitious ring closure energy of all the rings in the network, with respect to the total number of variables $\mathcal{N} = N - \mathcal{M}$. Since the computational cost of minimizing such nonlinear potential increases in the order of $\mathcal{N}^3$, this method is not capable of handling networks with hundreds of bonds, which are common in the flexible regions in small to moderate sized proteins.

It is more preferable to handle a network of rings in a ring by ring fashion. Suppose the number of rings in the network is n, the average number of variables per ring is thus $\mathcal{N}/n$. The computational cost of solving ring closure equations for every ring one by one for one time is thus on the order of $(\mathcal{N}/n)^3 \times n = \mathcal{N}^3/n^2$. Since both total number of rings n and total number of variables $\mathcal{N} - \mathcal{M}$ scale linearly with the total number of bonds N in common protein structures, the computation cost in theory scales linearly, if ring closure equations of all rings can be solved one by one. In practice, however, any other rings in the network break when dihedral angles in one ring are rotated to close the ring, as have been explained in Section 5.1.

In order to solve the ring closure equations for all rings in a network as efficient as possible, we design a procedure that minimizes the fictitious ring closure potentials of an expanding network. The algorithm first tries to close the ring which has the smallest number of unknown dihedral angles in the whole network. This ring is called the seed. After succeeding at closing this ring, the algorithm then minimizes the sum of fictitious ring closure potentials of the seed and of up to five more rings that share bonds with the seed. The newly added rings and the old seed is now the new seed of the network. If all rings in the seed can be closed simultaneously, ROCK then adds up to five more rings that share bonds with the seed to be the newly expanded seed. Step by step, ROCK adds rings to the expanding seed, and then minimizes the sum of the fictitious ring closure potential of all the rings in the seed. Because all the

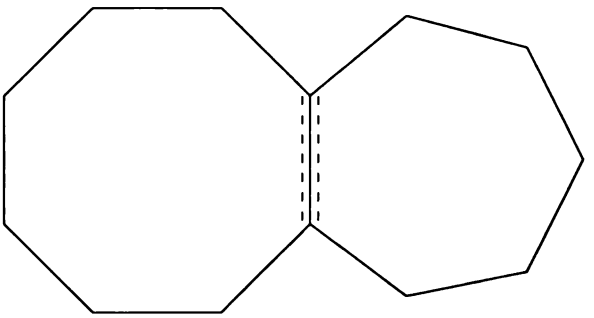


Figure 5.6: A simple network with two DOF. The left side eight-fold ring and the right side seven-fold ring share one bond which is shown as the dashed line.

rings in the seed are already closed when new rings are added, only small adjustments in dihedral angles are necessary to close all the rings in the seed concurrently. The whole process stops when all the rings in the network are added to the seed. The total calculation cost of this procedure is lower than minimizing the total fictitious ring closure potential of all the rings directly.

One crucial step in generating a new conformation for a network of rings is to select a set of bonds to be randomly rotated about their original values. Inappropriate selections of bonds results in unsuccessful trials and wasted computation time. For example there are two DOF in the network in Figure 5.6. One can select and rotate two bonds in order to find a new conformation for the network. But the selections of the two bonds cannot be arbitrary. If two selected bonds are all in the seven-fold ring, there are only five variables left unknown in the ring closure equations of the seven-fold ring. But each set of ring closure equations has six independent equations. In most cases, there is not a solution for the five variables in six independent equations. This trial is most likely to be unsuccessful due to the wrong combination of selected bonds to be randomly rotated. The possibility to close the rings after two bonds are rotated is greatly enhanced if one selected bond is in the seven-fold ring and the other is in the eight-fold ring, or if both selected bonds are in the eight fold ring.

ROCK utilizes the following procedure to select a set of bonds to rotate. It is

designed to avoid choosing the wrong combination of bonds to rotate from constraint theory point of view.

1. Randomly select a freely rotatable bond.

2. Count the DOF in the ring to which the selected bond belongs. Go back to step 1 if the DOF is negative which means the ring is over-constrained. One more constraint is counted because a randomly selected bond is equivalent to one more constraint on the dihedral angles. The ring is considered to be the seed of the network.

3. Expand the seed by one more ring by adding one ring that shares bonds with the rings in the seed to the seed. The DOF of the seed is calculated. Every randomly selected bond is counted as one more constraint. Go back to step 1 if the DOF is negative.

4. Repeat step 3 until all rings in the network are included in the seed. The bond selected in step 1 is then officially selected to be randomly rotated. Any local area the network is not over constrained by the rotation of this bond.

5. Repeat step 1 to step 4 until a desired number of bonds has been selected.

The procedure listed above ensures that the randomly selected bonds do not over constrain any local area of the network. Though it does not guarantee that there are six unknown variables for each set of ring closure equations, it does help reduce the rate of unsuccessful trials. Randomly selecting and rotating fewer bonds than the DOF further improves the rate of successful trials.

The discussion carried so far assumes every bond in a network is freely rotatable. There are bonds which, however, should be considered to be locked. The peptide bonds, for example, favor either *trans* or *cis* conformation. There are energy barriers between these two conformations. The dihedral angles of these bonds should be kept

unchanged from their values in the initial conformation. Each fixed bond adds one constraint to the network. A seven-fold ring with one fixed bond has zero DOF which is the same as that of a six fold ring.

5.4 Conformations of Side Chains

Once new conformations of the flexible ring clusters are generated, side branch atoms are anchored back to the ring clusters with correct bond lengths and angles. Side branch atoms are first randomly disturbed to sample the conformations of the side branches. The coordinates of the side branch atoms are then relaxed in the Cartesian coordinates so that 1) bond lengths and angles of side branch atoms are undistorted from the original values; 2) there are no van der Waals overlaps between side branch atoms themselves and between side branch atoms and ring cluster atoms; and 3) chiralities at side branch atoms are not changed. An atom is called a chiral center if the atom has equal to or more than three nearest neighbors. The bond lengths and angles between the chiral atom and its nearest neighbors are the same as those in the mirror image. The atomic arrangement of a chiral center and its three nearest neighbors is not identical to that in the mirror image. A chiral center atom and its three nearest neighbors are illustrated in Figure 5.7.

Since the bond lengths and angles are equality constraints, while the checks against van der Waals overlaps and chirality flips are inequality constraints, ROCK calls the program DONLP2 [124] to minimize the function

$$f(x) = \sum_{\text{bonds}} (r - r^0)^2 + \sum_{\text{angles}} (\theta - \theta^0)^2 \tag{5.5}$$

subject to a collection of inequality constraints of van der Waals repulsions

$$g_1(x) = r_{ij}^2 - r_v^2 \geq 0 \tag{5.6}$$

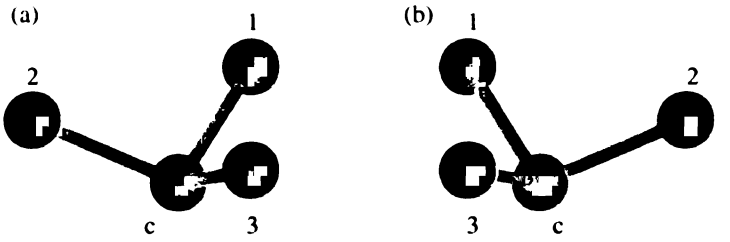


Figure 5.7: A chiral center atom c and its three nearest neighbors of atom 1, atom 2 and atom 3. Figure (b) shows the mirror image of Figure (a). The bond lengths and angles between the chiral center atom and its nearest neighbors are the same as those in the mirror image. The atomic arrangement of the chiral center and its nearest neighbors is not identical to that in the mirror image. It goes counter-clockwise from atom 1 to atom 2 to atom 3 in the original image, but it goes clockwise from atom 1 to atom 2 to atom 3 in the mirror image.

and a set of inequality constraints that chirality at atoms should not be flipped

$$g_2(x) = [\mathbf{r}_{ij} \cdot (\mathbf{r}_{ik} \times \mathbf{r}_{il})] \left[\mathbf{r}^0_{ij} \cdot (\mathbf{r}^0_{ik} \times \mathbf{r}^0_{il})\right] \geq 0 \tag{5.7}$$

in which r and θ are bond lengths and angles in the trial conformation, r^0 and θ^0 are the bond lengths and angles in the initial conformation, r_{ij} is the distance between two non-bonded atoms of atom i and atom j, r_v is the sum of the van der Waals radii of the atom i and atom j, $\mathbf{r}_{ij}$, $\mathbf{r}_{ik}$ and $\mathbf{r}_{il}$ are the vectors between atom i and its bonded neighbors in the trial conformation, and $\mathbf{r}^0_{ij}$, $\mathbf{r}^0_{ik}$ and $\mathbf{r}^0_{il}$ are the vectors between atom i and its bonded neighbors in the initial conformation. The function $f(x)$ is minimized to be zero when the bond lengths and angles of the side branch atoms are identical to the corresponding values in the initial conformation. The sign of the dot and cross product of the three vectors $\mathbf{r}_{ij}$, $\mathbf{r}_{ik}$ and $\mathbf{r}_{il}$ specifies the chirality of atom i. The chirality of atom i in the generated conformation is identical to the chirality of the same atom in the initial conformation if and only if the sign of the dot and cross product of the three vectors at the atom in the generated conformation is the same as the sign of the same product in the initial conformation. DONLP2 is a non-linear optimization program that can optimize a function subject to both

equality and inequality constraints. The inequality constraints of van der Waals overlaps forces the distances between any pair of non-bonded atoms to be larger than a critical value r_v which is the sum of van der Waals radii of these two atoms. When all the inequality constraints are satisfied there are no van der Waals overlaps between any pair of atoms in the protein, nor chirality at any atoms are flipped.

After randomly disturbing the side branch atoms from their original Cartesian coordinates, ROCK checks the distances between every pair of non-bonded atoms to build the van der Waals overlap list. It constructs the inequality constraints $g_l(x)$ according to the list. Then it minimizes the function $f(x)$ subject to the inequality constraints. Once the function is minimized to practically zero, ROCK builds a new van der Waals overlap list to begin a new round of minimization. A new conformation of the side branch is found when there are no van der Waals overlaps, when the function $f(x)$ is practically zero, and when the chirality at every atom is not flipped.

5.5 Workflow

According to the algorithms outlined above, we wrote a FORTRAN program package ROCK to sample conformations of the flexible regions in proteins. Since it relies on flexibility analysis to sort the flexible regions from the rigid cores of the proteins, it is preferable to have the flexibility analysis result from FIRST ready before the program is run. The program works in the following procedure:

1. Read in the initial protein conformation. Calculate the bond lengths and angles.

2. Read in the flexibility properties of the protein analyzed by FIRST. If the rigidity analysis result is not available, ROCK counts the distribution of DOF by itself by counting constraints in all local regions in the protein.

3. Find the rigid cores in the protein. Starting from one non-rotatable bond as the seed, ROCK adds nearest neighbor non-rotatable bonds to the seed until

there are no non-rotatable bonds can be included. ROCK finds all rigid cores in the proteins by this method. It then fixes the orientation and position of the largest rigid core in space. All the other smaller rigid cores and flexible ring clusters move relative to the largest rigid core.

4. Find all flexible ring clusters and side branches.

5. Randomly select and rotate several bonds in each flexible ring cluster according to the procedure described in Section 5.3.

6. Minimize the total fictitious ring closure potential $\mathcal{F}$ of the ring cluster defined in Equation 5.4 by the L-BFGS algorithm [122]. Go back to step 5 if the fictitious potential cannot be minimized to be zero.

7. Check for van der Waals overlaps between the ring cluster atoms themselves and between the ring clusters atoms and the rigid core atoms. Go back to step 5 if van der Waals overlaps exist.

8. Randomly disturb the side branch atoms away from their positions in the previous conformation.

9. Utilize the DONLP2 [124] algorithm to find a new conformation of each side branch with zero bond distortions, with correct chiralities at every atom and without any van der Waals overlaps. Go back step 8 for nine extra trials on each branch if new conformations cannot be found, or go back to step 5 to start from the beginning if ten consecutive searches of side branch conformations fail.

10. Accept the new conformation. Go back to step 5 to search for another conformation by using this new conformation as the starting point. Stop the whole process when a predefined number of new conformations are generated.

ROCK also checks the quality of main chain ϕ and ψ angles against the Ramachandran plot [110] to ensure the stereo-chemical quality of the generated conformations. Eighteen out of the twenty standard residues (alanine, arginine, asparagine, aspartic acid, cysteine, glutamine, glutamic acid, histidine, isoleucine, leucine, lysine, methionine, phenylanaline, serine, threonine, tryptophan, tyrosine and valine) are checked against the Ramachandran plot generated by Morris et al. [125]. The other two residues, glycine and proline, are checked against two Ramachandran plots specially designed for glycine and proline. The Ramachandran plots of glycine and proline are discussed in Appendix B. The main chain ϕ and ψ angles of the twenty standard residues are restricted to the *core* and the *allowed* regions of the Ramachandran plot by default. ROCK rejects any conformations which have one or more residues whose ϕ and ψ angles are not in the allowed regions of the Ramachandran plot.

Chapter 6: Results and Discussions

ROCK has been tested on several macromolecules. This chapter shows three examples. The first example is a model molecule made up of four ten-fold rings. The second example is the conformations of the human immunodeficiency virus type 1 (HIV-1) protease, which is one of the most important proteins controlling the life cycle of the virus. The third example shows multiple randomly generated pathways between the occluded, the closed and the open conformations of the protein dihydrofolate reductase (DHFR).

6.1 Model Molecule $H_8C_8S_{20}$

Figure 6.1 shows the model molecule $H_8C_8S_{20}$. Figure 6.1(a) shows the topology of the molecule, while Figure 6.1(b) shows a low energy conformation. From the topological point of view, eight carbon atoms are positioned on the corners of a rectangular box, connected either by double or by single sulfur atoms. Eight hydrogen atoms complete the valency of the carbon atoms. The bond lengths between carbon and sulfur, sulfur and sulfur, and carbon and hydrogen atoms are 1.805Å, 2.019Å and 1.120Å respectively. These values are the optimal bond lengths used in the MM3 force fields [126]. All the bond angles at carbon atoms are exactly tetrahedral (109.5°). To avoid van der Waals overlaps, the bond angles of sulfur atoms are increased to be 135°, rather than a more realistic value such as 95°.

The model molecule has 36 atoms, 60 independent bond angle and 40 bond length constraints. Note that there are only 5 not 6 independent bond angle constraints at the carbon atoms. The total number of DOF is thus 2. The molecule was chosen by us to have both many interlocking rings as well as two internal DOF, to make the conformational space both non-trivial and easy to display. The conformational space

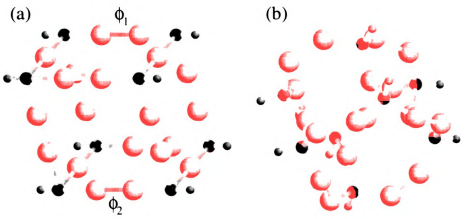


Figure 6.1: The model molecule $H_8C_8S_{20}$. In the topological graph shown in Figure (a), eight carbon atoms (green spheres) are at the corners of a rectangular box. Di-sulfur linkages are at eight out of twelve edges of the box. Single sulfur linkages occupy the other four edges. Sulfur atoms are represented by large yellow spheres. Hydrogen atoms (small Gray spheres) are connected to carbon atoms. A low energy conformation is shown in Figure (b). This molecule has two DOF.

of the molecule can be plotted on a 2D graph, in which the two variables representing the two DOF are the two axes. Any two dihedral angles can serve as the two axes on which the conformational space is projected. A convenient choice is the two dihedral angles ϕ_1 and ϕ_2 as shown in Fig. 6.1(a). As illustrated in the figure, these two dihedral angles are equivalent in the sense that they are interchangeable. The topology of the model molecule also shows mirror symmetry over the plane cutting through the centers of ϕ_1 and ϕ_2. Because a dihedral angle changes its sign under a mirror symmetry operation, and because the mirror image of a ring has the same bond lengths and angles as the ring does, a ring is closed when all its dihedral angles change sign. Therefore if there is a conformation at a certain combination of ϕ_1 and ϕ_2, there is also a conformation when the signs of ϕ_1 and ϕ_2 are changed. Because of these two symmetries, one quarter of the whole $2\pi \times 2\pi$ plane expanded by the two dihedral angles ϕ_1 and ϕ_2 is enough to depict the conformational space accessible to the molecule. These symmetries are $\{\phi_1, \phi_2\} \rightarrow \{\phi_2, \phi_1\}$ and $\{\phi_1, \phi_2\} \rightarrow \{-\phi_1, -\phi_2\}$.

Symmetries in conformational space hold true in the model molecule because it

does not have any bond length and angle variations. The lengths of all bonds between carbon and sulfur atoms are the same, the lengths of all bonds between sulfur and sulfur atoms are the same, the angles of all carbon atoms are the same, and the angles of all sulfur atoms are the same. The symmetries discussed above are not apparent in any single conformation of the model molecule, but are obvious in the ensemble of all conformations.

6.1.1 Conformations of Model Molecule $H_8C_8S_{20}$

Our program successfully generates 10,000 conformations in 53 hours on an Athlon AMD MP 1900+ processor. The search was carried out with a random walk procedure in the ϕ_1, ϕ_2 space. Van der Waals overlap is allowed in the first 5,000 conformations but not in the later 5,000. As shown in Figure 6.2(a), the conformations cover almost the whole two dimensional space when van der Waals overlaps are allowed. The conformations are not as densely packed in the vicinity of $\phi_1 \sim 0$ or $\phi_2 \sim 0$ as in the other regions. This is because the distribution of number of solutions to the ring closure equations is not uniform. As discussed in Section 5.2, 6 dihedral angles in a single N-fold ring can have zero or several solutions to the RCE, if the other $N-6$ dihedral angles are given. Similarly, because our model molecule has two degrees of freedom, for a given pair of ϕ_1 and ϕ_2 values, there can be none to multiple sets of solutions of the remaining dihedral angles to the RCE. There are fewer solutions to the RCE at points in $\phi_1 \sim 0$ and $\phi_2 \sim 0$ than at points in other regions. The distribution of the number of solutions in the (ϕ_1, ϕ_2) plane is directly reflected by the frequency of our program finding solutions in any regions. A low resolution grid search and a random search in conformational space confirm this point.

Our program implements a hard sphere model in checking for van der Waals overlaps. The distances between any two non-bonded atoms have to be greater than the sum of their van der Waals radii otherwise the conformation is abandoned. Van

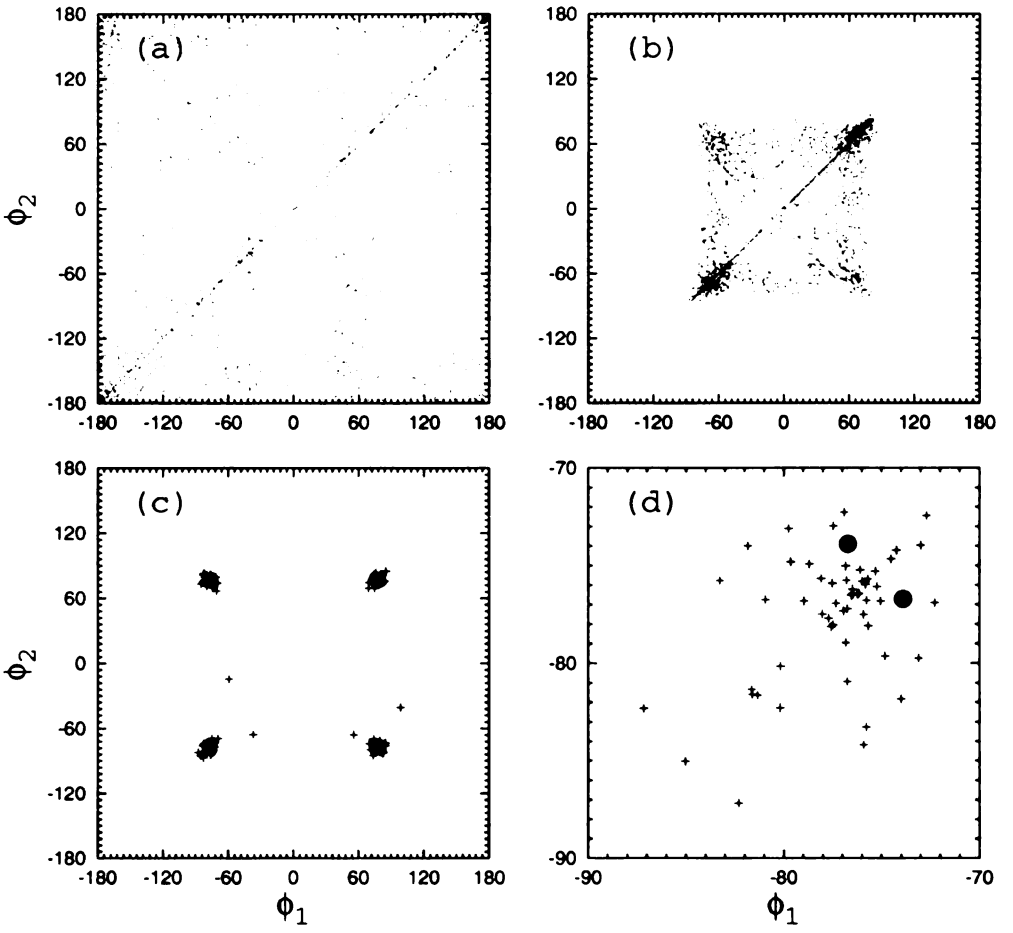


Figure 6.2: The distribution of conformations projected on the two axes ϕ_1 and ϕ_2. The top left Figure (a) shows the generated conformations when van der Waals overlaps are allowed. The top right Figure (b) shows the generated conformations when van der Waals overlaps are prohibited. The bottom left Figure (c) shows the optimized conformations in the MM3 force field. The bottom right Figure (d) shows a close up of one of the four clusters of optimized conformations in Figure (c). Two degenerate global minimum states are shown as solid circles in Figure (d).

der Waals radii used in our program are 1.632Å, 1.72Å and 1.296Å for carbon, sulfur and hydrogen respectively. These values are all 0.8 times the optimum radii used in the van der Waals interactions in the MM3 force field. The van der Waals potential in MM3 force field is soft at the optimum distance. It does not rise sharply until the distance between two atoms is considerably shorter than the sum of their optimum radii. The potential increases by roughly 4 kcal/mol when the distance between two atoms decreases to 0.8 times the optimum distance. By allowing a maximum 4 kcal/mol potential penalty in the van der Waals interactions, our program samples as large a conformational space as possible, while avoiding generating conformations whose van der Waals potentials would be unreasonably high. Therefore our hard sphere radii of atoms are set to be 0.8 times the optimum radii used in the MM3 force field. As shown in Figure 6.2(b), the conformational space sampled when the hard sphere interaction is turned on is considerably reduced from the conformational space when the van der Waals overlap effect is not included. This is indeed expected. The figure shows clearly that some regions are sampled more frequently than other regions. Van der Waals constraints, in addition to the uneven distribution in the number of solutions to the RCE, cause the nonuniform distribution of conformations.

The 5000 conformations without van der Waals overlap are optimized further by an external software package TINKER [127] with the MM3 force field. To be consistent with the model molecule, the optimal angles of sulfur atoms in MM3 force field are adjusted to be 135°. The conformational space of the optimized structures is plotted in Figure 6.2(c). The energies of most optimized conformations are all below -58 kcal/mol. The global minimum energy is -61.998 kcal/mol. All but a few optimized conformations lie in one of the four symmetric clusters. The 5000 conformations generated when van der Waals overlap is included are also optimized by the same force field. The energy of the global minimum structure is also -61.998 kcal/mol, and has the same conformation as above.

It is worth noting that the ensemble of conformations sampled by our program shows the two expected symmetries of $\{\phi_1, \phi_2\} \rightarrow \{\phi_2, \phi_1\}$ and $\{\phi_1, \phi_2\} \rightarrow \{-\phi_1, -\phi_2\}$. As shown in Figure 6.2, these two symmetries can be vaguely identified among the 5000 conformations generated when the van der Waals overlaps are allowed, and is obvious among the 5000 conformations generated when the van der Waals overlaps are disallowed. The fact that the two symmetries required by the topology of the molecule are manifested in the ensemble of conformations sampled by our program is a necessary yet insufficient proof that our algorithm samples the whole conformational space of this model molecule.

It is clear that our algorithm samples conformational space efficiently by using this hierarchical approach. Maximum conformational space is sampled when van der Waals overlaps are allowed. As expected, the conformational space is considerably reduced when van der Waals overlap is prohibited. The conformational space is further reduced when the molecular structure is optimized in a full force field such as MM3. By exploring the conformational space with only bond length and angle constraints while forbidding van der Waals overlap, our algorithm is able to explore the conformational space where local minima of full force field are most likely to reside. Without bond length distortion, bond angle distortion and van der Waals overlaps, every generated conformation is accessible by the model molecule at moderate temperatures.

6.2 Conformations of HIV-1 Protease

6.2.1 Structures and Functions of HIV-1 Protease

HIV-1 protease is vital for the reproduction of the HIV virus. The HIV virus replicates several proteins that are essential for the viral maturation process on a long peptide chain which is called the "polyprotein". The proteins in the long polypro-

tein chain are not active. There is a time window for the HIV-1 protease to cut the polyprotein into several pieces to activate these proteins before the polyprotein degrades. The binding of prohibitory molecules to the HIV-1 protease therefore hinders the activation of the proteins in the HIV polyprotein, with the consequence that the HIV virus cannot reach its matured stage. The importance of the HIV-1 protease in the life circle of the HIV virus has made itself the primary pharmaceutical target for curbing the acquired immune deficiency syndrome (AIDS) which is caused by the HIV virus. Several drugs designed to bind to the HIV-1 protease with great affinity have shown positive effects on AIDS patients.

Since the first 3D structure of the HIV-1 protease was published [128], more than 200 structures of the protease have been reported on various resources [129]. Most of the structures are bound with inhibitors. The inhibitor-free structure of HIV-1 protease contains two identical amino acid chains. Each chain is made up of 99 residues. The two chains are glued together by hydrogen bonds and hydrophobic interactions. The molecule has an exact C_2 symmetry. Shown in Figure 6.3 are a ligand-free HIV-1 protease structure (1HHP) obtained from X-ray crystallography experiment [130]. The two chains are shown as red and blue ribbons respectively. The big free volume in the middle of the protein is the binding site. Polyproteins are locked and cut in this region. The catalytic sites ASP25-THR26-GLY27 in both chains are rendered as spheres.

It is worth noting that all conformations of the HIV-1 protease observed in experiments are either bound with ligands or are in the closed conformation, similar to the one shown in Figure 6.3. In such a conformation the catalytic site is not covered by the two flaps at the top of the protease due to the large void that is immediately above the site, but the short distance between the two flaps forbids the polyprotein to approach to the catalytic site from the top. There should be many open conformations where the two flaps are widely open so that the polyprotein can pass through.

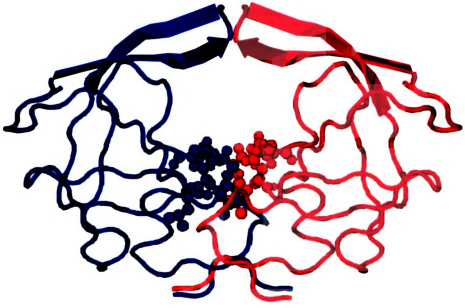


Figure 6.3: The ligand-free structure of the HIV-1 protease (1HHP). The two identical chains are rendered as red and blue ribbons respectively. The flexible flaps at the top of the protein are rendered as widened strands. The catalytic site is indicated by spheres.

These open conformations are not observed in X-ray crystallography because other molecules in adjacent unit cells in the protein crystal force the protease to take a unique and closed conformation.

The two flaps at the top of the protease are flexible. NMR experiments show that the two flaps, residue 48 to 55 which are shown as the widened ribbons in Figure 6.3, have two types of motions in different time range. One motion is the slow motion of the flexible flaps in the time range of μs-ms [131]. The other motion is the fast curl in and curl out motion of the tips of the flexible flaps (residue 49 to 53) which is in the sub-ns time range [132]. The fast motion confirmed what is seen in the MD simulation [133], though Freedberg et al. [132] do not agree on the scale of the flexibility shown in the simulation. Another MD simulation by Carlson's group [134] shows that the fast curl motion of the tips of the flaps is absent when water molecules are first equilibrated before the MD simulation. Carlson argues that the

curl motion observed in MD simulations is actually caused by the voids between water molecules and the flaps, if the water molecules are not in their optimal positions at the beginning of the simulation. Whether the curl motion is real is yet to be cleared. The MD simulation shows that the characteristic of the curl motion is that the ϕ and ψ angles of the GLY51 residue are in different regions in the Ramachandran plot as other glycine residues in the protease are.

The motion of the flexible flaps is important in the function of HIV-1 protease. The distance between the two tips of the flaps in crystal structures is only 2.7Å. Such distance is not large enough for the polyprotein to pass through, bearing in mind that the diameter of a water molecule is 2.8Å. The two flaps have to undergo a large range of conformational transformations to catch a small part of the polyprotein. The fast sub-ns range motion enables the protease to adapt to a suitable conformation to better interact with the polyprotein that is passing nearby. The flaps then drag a part of the polyprotein chain to the reaction center via the slow motion.

We would like to answer two questions pertaining to the conformations of the flexible flaps of the HIV-1 protease: 1) How large the distance between the tips of the flexible flaps can be in all possible conformations? This distance have to be large enough so that the amino acid chain of the polyprotein can pass through the voids between the flaps. 2) Are there conformations in which the tips are curved inward? Are the ϕ and ψ angles of the GLY51 residue different than those of other glycine residues in the protease? We utilize the combination of FIRST and ROCK to address these questions.

Our algorithm samples the possible conformations but not the dynamical trajectories. Therefore our algorithm can study the statistical behavior of possible trajectories such as the maximum distance between two atoms and the distribution of dihedral angles et al. Our algorithm cannot answer the questions related to the absolute dynamical trajectories. However we do get hints of what a real trajectory looks

like by studying the statistical behavior of an ensemble of possible conformations.

6.2.2 Flexibility Analysis on HIV-1 Protease

The software package FIRST is used to analyze the flexibility properties of the HIV-1 protease. FIRST assumes a hydrophobic interaction whenever two non-bonded carbon atoms are within a certain distance. This criterion works well in the interior of the proteins where hydrophobic interactions are stable. On the other hand, the hydrophobic interactions on the surface of the proteins are fragile. The only interaction between two non-bonded carbon atoms is the van der Waals interaction, which is weak and easy to break. The hydrophobic interactions in the interior of the protein are stable not because of the interactions between the non-bonded atoms themselves are strong, but because of the overall confinement and compression effects of the water molecules around the proteins. The presence of water molecules around the proteins strengthens the hydrophobic interactions in the interior of the proteins. However water weakens or even destroys the hydrophobic interactions on the surface of the protein. Two non-bonded carbon atoms on the protein surface may accidentally be close to each other in fractions of the whole protein motion trajectory, but the contacts are easily destroyed by the collision of water molecules which are always in thermal motion. Another argument of why the hydrophobic interactions are stable in the interior but not on the surface is that the concentration of hydrophobic interactions in the interior of proteins is higher than that on the surface. The high concentration hydrophobic interactions interlock with each other so all of the interactions are strengthened. Therefore the analysis of hydrophobic interactions should consider the environments in the surrounding of the potential hydrophobic interactions. FIRST however does not take the environmental effects on hydrophobic interactions into consideration. Its simple definition of hydrophobic interactions may produce false positives in some cases.

FIRST identifies six hydrophobic interactions between the two flexible flaps. From the fact that these two flexible flaps are in constant motion between the open and the closed states, though the open states are not observed in X-ray crystallography experiments, we know that these six hydrophobic interactions must be short lived and weak, otherwise the two flexible flaps will be always in the closed conformation. The fact that these six pairs of non-bonded carbon atoms are in close contacts in the crystals does not mean these atoms always stay together. For this reason we manually removed these six hydrophobic interactions. A flexibility analysis without these six interactions shows that majority of the HIV-1 protease is rigid with small flexible regions, as shown in Figure 6.4. The two flaps on the top are flexible, as required by the biochemical function of these two flaps. The four additional flexible regions in the protein are not of interest to us. Not shown in the figure are many other smaller flexible regions which are exclusively made up of flexible side chain atoms.

6.2.3 Conformations of HIV-1 Protease

ROCK generates 600 conformations of the protein HIV-1 protease obeying all the constraints specified in the flexibility analysis. Because a large portion of the protease is rigid, the calculational power of ROCK is concentrated on the two top flexible flaps and on the other smaller flexible regions shown in Figure 6.4. ROCK also generates conformations for the small flexible regions which involve only side chain atoms. The total CPU time is 6 hours and 40 minutes on an AMD Athlon 1900+ (real frequency is 1.6GHZ) processor. All of the generated conformations do not have van der Waals overlaps. All main chain ϕ and ψ angles are restricted to the core and the allowed regions in the Ramachandran plot. The Ramachandran plot is discussed in detail in Appendix B.

Figure 6.5 shows the superimposition of the 600 conformations in the ribbon diagram. The top figure is viewed from the side and the bottom figure is viewed from

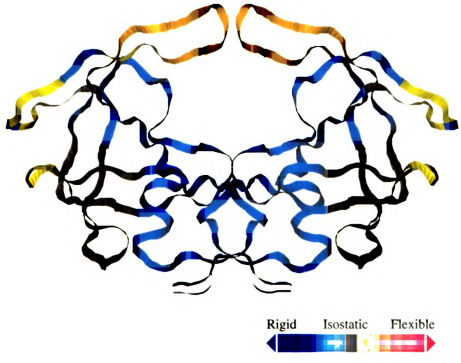


Figure 6.4: The flexibility properties of the HIV-1 protease. The protease has one major rigid core as shown in blue and gray. The blue regions are over constrained. Lots of bond constraints have to be cut to make the blue regions flexible. The gray regions are almost flexible but still rigid. The two flaps on the top of the protein are flexible, indicated by the color gold. Four additional regions in the protease are flexible, shown as the yellow regions in the figure. The regions colored by gold have greater density of DOF than the regions colored by yellow.

the top. Only the protein main chain motion is shown in the figure. The flexible and the rigid regions of the proteins are colored by yellow or gold and blue or gray respectively. Figure 6.6 shows the superimposition of the same 600 conformations in wire diagram. All bonds are shown in the figure. The motion of flexible side chains is revealed in this figure but not in Figure 6.5.

The distance between the tips of the two flexible flaps in all conformations indicates how big the conformational space the protein can sample. The distance between the flaps is defined as the smallest distance between any atom in one flap and any atom in the other flap. The distance between the two flaps is 2.7Å in the crystal structure. This distance is too small for any peptide chain or small molecules to pass through. The distance can be as large as more than 8.0Å however, as shown in Figure 6.7. As shown in the figure, the distance between the two flaps is smaller than 3.0Å only occasionally. The distance oscillates around 5.0Å in most of the conformations. The distance of 5.0Å is large enough for water molecules to pass through, but not large enough for a thick peptide or a big inhibitor molecule to pass. So the protein does not open up enough in most of the time. There are occasions however that the distance is larger than 8.0Å, which is large enough for a peptide chain to go through. Once a short segment of polyprotein can go through the gap between the two flexible flaps, the flaps falls back to the closed conformations to hold the polyprotein tightly. The catalytic residues in the HIV-1 protease then cut the polyprotein. Our calculation shows that the first step of the whole catalytic function of the protease, which is the opening of the two flexible flaps, is indeed possible without any external driving force.

The RMSD of C_α atoms indicate the average deviation of the main chain atoms in generated conformations from those in the crystal structure. Its mathematical form u is calculated by

$$u^{(j)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i^{(j)} - \mathbf{r}_c^{(j)})^2} \tag{6.1}$$

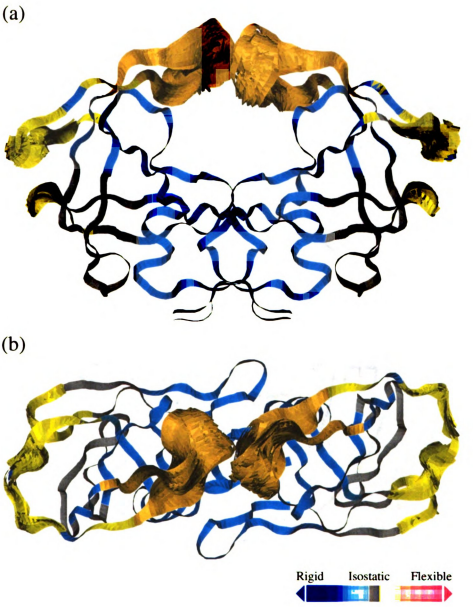


Figure 6.5: Superimposition of 600 conformations of HIV-1 protease generated by ROCK shown in ribbon diagram. The residues in the rigid regions of the protein are plotted in blue and gray. The flexible residues are shown in yellow, gold and red. Figure (a) shows the side view while Figure (b) shows the view from the top.

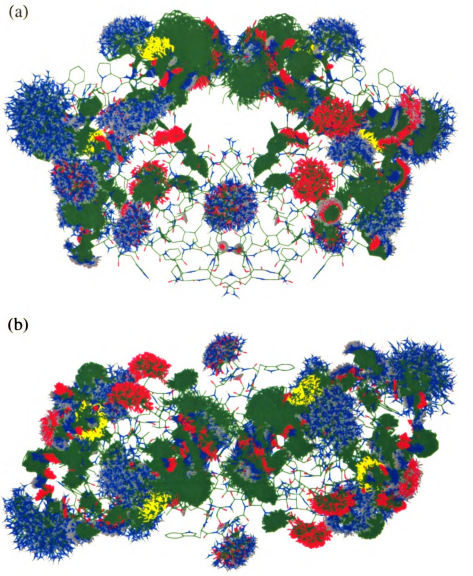


Figure 6.6: Superimposition of 600 conformations of HIV-1 protease generated by ROCK shown in wire diagram. Atoms are not shown in the figure but the bonds are colored by the atom types at the two ends of the bonds. Carbon, nitrogen, oxygen and hydrogen atoms are colored in green, blue, red and gray respectively. The conformations shown in this figure are the same as those in Figure 6.5. Figure (a) shows the side view while Figure (b) shows the view from the top.

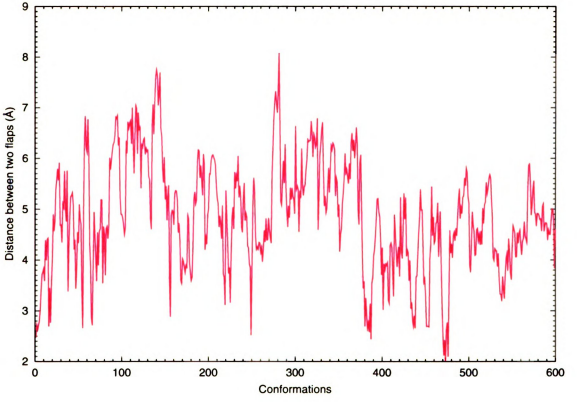


Figure 6.7: Distance between the two flexible flaps in all generated 600 conformations. The distance is only 2.7Å in the crystal structure. The distances in most of the generated conformations are around 5.0Å. The largest distance is more than 8.0Å.

for the RMSD of the C_α atom of the jth residue. $\mathbf{r}_i^{(j)}$ is the coordinate of the C_α atom of the jth residue in the ith conformation. $\mathbf{r}_c^{(j)}$ is the coordinate of the C_α atom of the jth residue in the crystal structure. The sum is over all N conformations.

Atoms are in constant motion even in crystal structures. The averaged atomic motion in X-ray crystallography data is given by the Debye-Waller B factor which is related to the RMSD by

$$B^{(j)} = 8\pi^2 (u^{(j)})^2 \tag{6.2}$$

Figure 6.8 shows the calculated and the measured RMSD of the C_α atoms. The calculated data is based on the coordinates of the C_α atoms in the 600 generated conformations. The measured data is converted from the measured Debye-Waller factor provided in the crystal structure of the HIV-1 protease [130]. Because the motion of the protease is restricted in the crystal, the RMSD of the C_α atoms in the crystal structure does not have much interesting features. It oscillates around 0.7Å without any major peaks. The protease in the crystal structure does not sample many conformations because of crystal contacts and volume constraints. Our calculation shows that the protein should have large conformational changes in three regions: from residue 15 to residue 18, from residue 35 to residue 42 and from residue 45 to residue 56. Prominent motion is shown in residue 45 to residue 56, which are the flexible flaps at the top of the HIV-1 protease.

The calculated RMSD is only non-zero in the flexible regions. This is expected because our algorithm fixes the positions of the atoms in the rigid cores of the proteins. Our algorithm artificially eliminates the small scale fluctuations of all atoms around their equilibrium positions. It illustrates only the large scale conformational changes that are beyond the averaged fluctuations.

The calculated RMSD of the two chains, shown as the red and the blue curves in Figure 6.8 do not overlap with each other. This is because our calculation is not able to sample all possible conformations. These two curves will be identical when

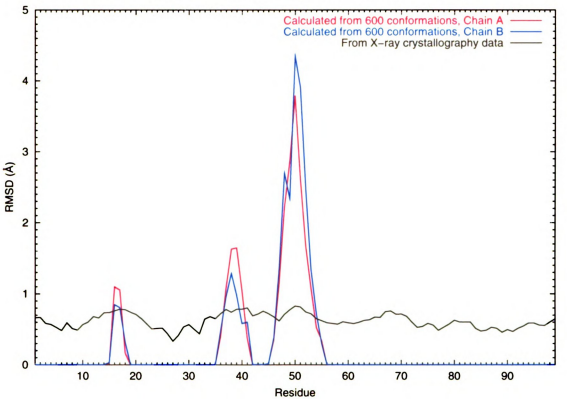


Figure 6.8: The comparison of the calculated and the measured RMSD of the main chain C_α atoms in each residue. The red and the blue curves are calculated based on the 600 generated conformations. The black curve is converted from the Debye-Waller factor in the X-ray crystallography data of the crystal structure of the HIV-1 protease.

the generated conformations of one chain are the same as those of the other chain. Since the conformations of the two chains are independently generated, the generated conformations of the two chains will be symmetric only when they are the complete conformations of the two chains. Given the number of DOF in the HIV-1 protease is large, the calculation time required to sample the complete conformations would be astronomical.

Scott et al. [133] publish a similar figure in the discussion of their MD simulation on the same protease. Because MD simulation catches both the slow and the fast motions in proteins, the RMSD calculated from the MD simulation has a oscillating background of 1.2Å. The highest RMSD value of about 4.8Å is observed at the 50th residue in one chain. This data is comparable to the RMSD of the 50th residue in our generated conformations. The RMSD of the C_α atoms in the three flexible regions identified by FIRST are all distinguishable from the background in the MD simulation. MD simulation shows that the RMSD of the C_α atoms in residue 75 to residue 83 are higher than the background noise, which implies residues from 75 to 83 are flexible. However flexibility analysis predicts these residues to be rigid. The flexibility analysis may have falsely identified a few hydrogen bonds or hydrophobic interactions in this region.

The distribution of the main chain ϕ and ψ angles of the glycine residues in the conformations generated by our algorithm are qualitatively different from the those of the conformations sampled by the MD simulation reported by Scott et al. [133]. As shown in the four panels in Figure 6.9, the distributions of ϕ and ψ angles of four glycine residues that are on the tips of the flexible flaps are very narrow in the whole Ramachandran plot. The main chain ϕ and ψ angles of the GLY51 residue in both chains are distributed in such a narrow range that they can be considered as not having changed at all in all of the 600 conformations. A structural analysis reveals that these two dihedral angles are all included in a ten-fold ring which is formed by

the main chain atoms of GLY49, ILE50 and GLY51. The hydrogen bond between the GLY52 main chain hydrogen atom and the GLY49 main chain oxygen atom closes the ten-fold ring, as indicated by the dashed bond in Figure 6.10. The peptide bond of the ILE50, GLY51 and GLY52 residues are also included in this ten-fold ring, as indicated by the red bonds in the figure. Because peptide bonds are not rotatable, the ten-fold ring has in fact only one DOF. The dihedral angles of all rotatable bonds, including the main chain ϕ and ψ angles of the residue GLY51, are hence much limited.

On the contrary, the distribution of the main chain ϕ and ψ angles of the GLY51 residue in the conformations created by the MD simulation [133] covers a significant portion of the whole Ramachandran plot. Assume the bond lengths and angles of the covalent bonds do not vary much in the MD simulation, the bond length and angle of the hydrogen bond that closes the ten-fold ring must undergo great distortions to allow the ϕ and ψ angles of GLY51 to vary much. This implies that the hydrogen bond of the ten-fold ring is not stable in the MD simulation. Because the energy of the hydrogen bond is -3.14 kcal/mol in the crystal structure, which is a typical value of a strong hydrogen bond, flexibility analysis by FIRST lists this bond as stable constraints. The discrepancy between the hydrogen bond stability predicted by the flexibility analysis and that by the MD simulation has to be eliminated in future studies.

Since the distribution of the main chain ϕ and ψ angles of the GLY51 residue is limited to a small region in the Ramachandran plot, the flexible flaps in the conformations generated by our algorithm do not have the curl in motion discovered in the MD simulation. Even then, the two flexible flaps are widely open in some conformations, for example the distance between the two flaps can be as large as 8.0Å. Therefore we conclude that the curl motion is not related to the open and close motion of the flexible flaps at the top of the HIV-1 protease.

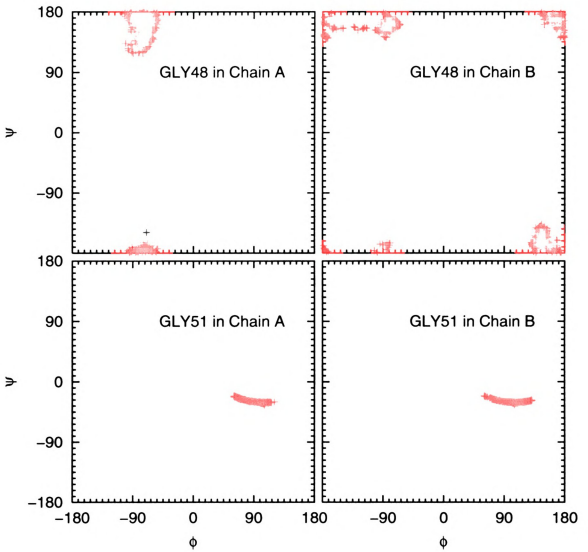


Figure 6.9: The distribution of main chain ϕ and ψ angles of the GLY48 residue in chain A (top left panel), of the GLY48 residue in chain B (top right panel), of the GLY51 residue in chain A (bottom left panel) and of the GLY51 residue in chain B (bottom right panel). The distribution of main chain ϕ and ψ angles of GLY48 in both chains covers much larger space in the Ramachandran plot than the distribution of ϕ and ψ angles of the GLY51 residues does.

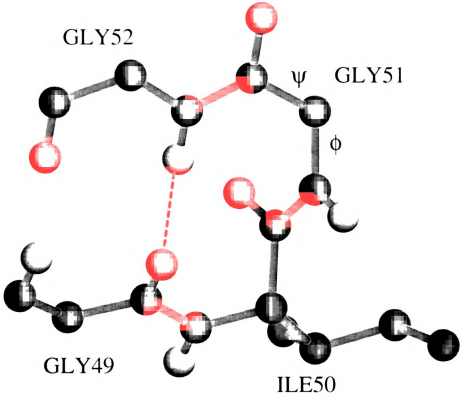


Figure 6.10: The tip of the flexible flaps at the top of the HIV-1 protease is made up of GLY49, ILE50, GLY51 and GLY52. The main chain ϕ and ψ angles are included in the ten-fold ring which is closed by the hydrogen bond between the main chain hydrogen atom of GLY52 and the main chain oxygen atom of GLY49. The hydrogen bond is indicated as the dashed bond. Three non-rotatable peptide bonds, which are the red bonds in the figure, are also in the ten-fold ring. Green, blue, red and gray spheres are carbon, nitrogen, oxygen and hydrogen respectively.

6.3 Conformational Pathways of DHFR

6.3.1 Structures and Functions of DHFR

The enzymatic protein DHFR catalyzes the reduction of 7,8-dihydrofolate (DHF) or folate to 5,6,7,8-tetrahydrofolate (THF) with the help of the coenzyme nicotinamide adenine dinucleotide phosphate (NADPH). THF is the one carbon carrier in the synthesis of many amino acids. It also plays a key role in the bio-synthesis of purine and thymidylate, which are essential components of DNA. Therefore the activity of DHFR, which transforms folate or DHF to THF, indirectly controls the bio-synthesis of DNA. The inhibition of DHFR inevitably results in the blocked DNA synthesis, which ultimately leads to the death of the cells. The key role it plays in the DNA metabolism has made DHFR the target of anti-cancer drugs [135]. These drugs hinder the growth of cancer cells by blocking the activity of DHFR. Because cancer cells grow faster than human cells, anti-cancer drugs that bind to DHFR do not affect the human cells as much as they do cancer cells.

Because of the importance it bears, the protein DHFR is present in all living organisms, including archae, prokaryotes and eukaryotes. It is also extensively studied. Structures of DHFR complexed with various ligands have been determined by X-ray crystallography and by NMR techniques. Up to date there are already 105 DHFR structures in the Protein Data Bank (PDB) [136]. Statistical analysis [137] shows that there are three *Escherichia coli* DHFR (ecDHFR) conformations: the open, the closed and the occluded conformations. Crystal structures of vertebrate DHFR proteins are always found in one conformation which resembles the closed conformation of ecDHFR [138, 139].

Figure 6.11(a) shows the superimposition of the three conformations of ecDHFR. The closed, the occluded and the open conformations are represented by the protein 1RX1, 1RX6 and 1RA9 respectively [140]. The three conformations are almost iden-

tical except in the loop region of residue 14 to 24, which are conventionally called the M-20 loop. Figure 6.11(b) shows the close up of the loop region.

The M-20 loop covers the binding site of the ecDHFR. Its movement is believed to be coupled with the catalytic reaction of the protein. The M-20 loop catches the ligands when it is in a particular conformation. The loop then escorts the ligand to the binding site through proper conformational changes. Once the ecDHFR finishes the catalyzing process, the M-20 loop then opens, guides the reduced ligands out of the binding site, and then releases it. The NMR experiments by Falzone et al. [141] prove the frequency of the M-20 loop conformational change is the same as the disassociation rate of the THF.

Beyond the function of steering the ligands in and out of the binding site, the motion of the M-20 loop may participate in the catalytic reaction directly by coupling with the reaction coordinates, according to the theory that protein motions may activate catalytic reactions [142]. Through their hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) simulation, Agarwal et al. [143] identify the coupling between the side chain motions and the reaction coordinate of the ecDHFR. Simulation of Shrimpton et al. [138] reports similar results on chicken DHFR.

Sawaya et al. [137] conclude from systematic studies that the closed conformation of the ecDHFR is in the first half cycle of the catalytic reaction of the protein, while the occluded conformation is in the second half cycle. There must be conformational pathways between these two distinct conformations. The role of the open conformation is not clear though. The open conformation is not seen in either the first or the second half of the catalytic reactions of ecDHFR. The open structure, shown as the blue bend in Figure 6.11, is not in the middle of the closed (the yellow bend) and the occluded (the red bend) conformations in real space. However, Sawaya et al. find that the open conformation is in the middle of the closed and the occluded conformations in terms of structural and biochemical characteristics, such

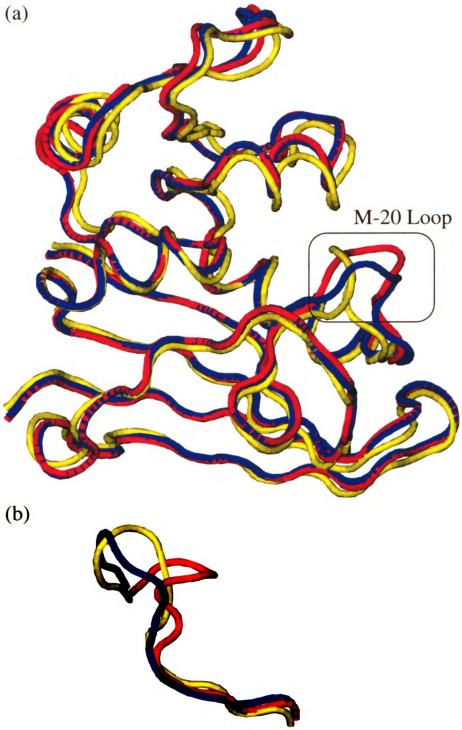


Figure 6.11: The superimposition of the open (blue), the closed (yellow) and the occluded conformations (red) of ecDHFR. The closed, the occluded and the open conformations are represented by the protein 1RX1, 1RX6 and 1RA9 respectively. Figure (a) shows the whole proteins while Figure (b) shows the close up of the M-20 loops.

as the pattern of hydrogen bonds, the packing with ligands, the secondary structures and the distribution of main chain dihedral angle.

The importance of the motion of the M-20 loop requires studies of its detailed motion patterns. There are two questions to be answered in understanding the conformational changes of the M-20 loop. The first question relates to the conformational pathways. Since the occluded and the closed conformations of the M-20 loop are detected and proved to be intermediates of the catalytic reactions, there must be conformational pathways between these two distinct conformations. It is not clear whether there is only one conformational pathway, or there are many. The second question concerns the existence of the open conformations. The open conformation is in the middle of the occluded and the closed conformations when examined from the point of the main chain dihedral angles, but is outside of the range of these two conformations in real space, as shown in Figure 6.11. It is not clear whether the conformational pathways between the occluded and the closed conformations involved the open conformation.

The NMR technique, which measures an ensemble of conformations, is the right tool to gain the structural data of the most populated conformations, but not the tool to extract the structures of a conformation that is less populated. For this reason the intermediate conformations between the occluded, the closed and the open conformations are never observed in experiments. Simulation techniques have to be utilized to gain insight into the conformational pathways of the M-20 loop.

MD simulation, being the standard method in exploring protein conformational changes, is in practice not be able to sample the whole conformational pathways between the open, the closed and the occluded conformations of ecDHFR, because the time range of the ecDHFR conformational changes its conformation is beyond the calculation capacity of current MD simulation techniques. The ecDHFR changes its conformation at a rate of about $20s^{-1}$ [144]. The best MD simulations reach a

couple of microseconds while the majority of MD simulations in literatures are in the nanosecond time range. Yet several MD simulations have been tried on ecDHFR [145, 146]. The MD simulation by Rod et al. [147] shows very promising results. It shows multiple transitions between the three distinct conformations, though the simulation is only 10ns long which is far less than the $20s^{-1}$ conformational change rate reported in experiments. It also predicts several other distinct conformations that are not detected by experiments.

In practice, the MD simulation is limited in its capacity to sample long time scale protein conformational changes partly because it wastes time on high frequency motions, partly because it includes the whole protein in calculation. The protein ecDHFR is very stable. Except in the M-20 loop region, the protein does not undergo much structural transformation under usual conditions. The structural stability of the protein support our interpretation of proteins as flexible regions anchored on motionless rigid cores. The rigid core of DHFR does not change its conformation in a full catalytic reaction cycle. MD simulations waste time on calculating the dynamic trajectory of the rigid cores of proteins which is not of interest in our study of the motion of the M-20 loop.

ROCK eliminates calculation endeavor on any high frequency motions by fixing the bond lengths and angles. The flexibility analysis enables our algorithm to sample the conformations of the flexible regions of the proteins only. These two advantages make our algorithm a powerful tool in sampling the conformational changes that is beyond the scope of present MD simulations. Our algorithm is capable of sampling the conformational pathways between the occluded and the closed conformations of the protein ecDHFR. We would like to answer the questions of how similar these conformational pathways are and whether the open conformation appears in the pathways. Section 6.3.2 examines the flexibility characteristic of the protein. It is worth noting that our algorithm samples the protein conformations by a random walk procedure.

Section 6.3.3 shows how a small modification enables our algorithm to search the random yet directional pathways. Section 6.3.4 discusses the calculation results. Unlike the MD simulation which determines whether a snapshot of a trajectory is closer to which of three conformations (open, closed or occluded) according to the characteristics of the main chain dihedral angles, our analysis are mainly done in real space.

6.3.2 Flexibility Analysis on DHFR

Experiments extract more than one distinct conformation for some proteins, such as HIV-1 and DHFR. The existence of multiple conformations makes it easier to predict which hydrogen bonds or hydrophobic interactions are stable in the whole path of the protein conformational changes. If a hydrogen bond or a hydrophobic interaction is present in one conformation but not in the others, it is safe to say that this interaction is not stable. Therefore by comparing the distinct protein conformations we can identify those interactions that are truly stable. The quality of the flexibility analysis is improved by including only the interactions that are truly stable in the whole protein motion trajectory.

In principle the more distinct conformations of a protein there are the more reliable the prediction of the stable hydrogen bonds and hydrophobic interactions is in flexibility analysis. We use only one conformation in the flexibility analysis of HIV-1 protease however because our algorithm is powerful enough to provide much insight into the intrinsically allowed motions of proteins on the basis of only one structure. After all, many proteins have only one distinct structure which is determined from experiments.

Since the occluded and the closed conformations are two observed intermediate states in the catalytic pathways of the enzyme ecDHFR, these two structures are good indicators of which hydrogen bonds and hydrophobic interactions are stable in the

whole enzyme reaction pathways. We exclude the unstable hydrophobic interactions and hydrogen bonds from our flexibility analysis. Of those hydrogen bonds that are present in both conformations, the energies are taken to be the higher values (weaker bonds) of the same bonds in the two conformations, because the higher energy of a bond in two conformations tells how unstable the hydrogen bond can be.

Information from the open conformation is not used in the analysis of stable hydrogen bonds and hydrophobic interactions. One question we would like to address is whether the open conformation is accessible on the pathways for the protein to transform between the closed and the occluded conformations. The inclusion of any information from the open conformation in the flexibility analysis would bring bias in favor of showing open conformation in the pathways.

The protein structures 1RX1 and 1RX6 [140] are used in this study to represent the closed and the occluded conformations of ecDHFR. The two conformations are identical in amino acid residues. Therefore the covalent bonds are the same in the two conformations. Both conformations are bound with ligands. Ligands and surrounding water molecules are removed since they are irrelevant to the sampling of intrinsically allowed conformations of the protein. Both protein structures are from X-ray crystallography with a resolution of 2.0Å. Polar hydrogen atoms are added to both conformations by the Unix version of the software WhatIF99 [148].

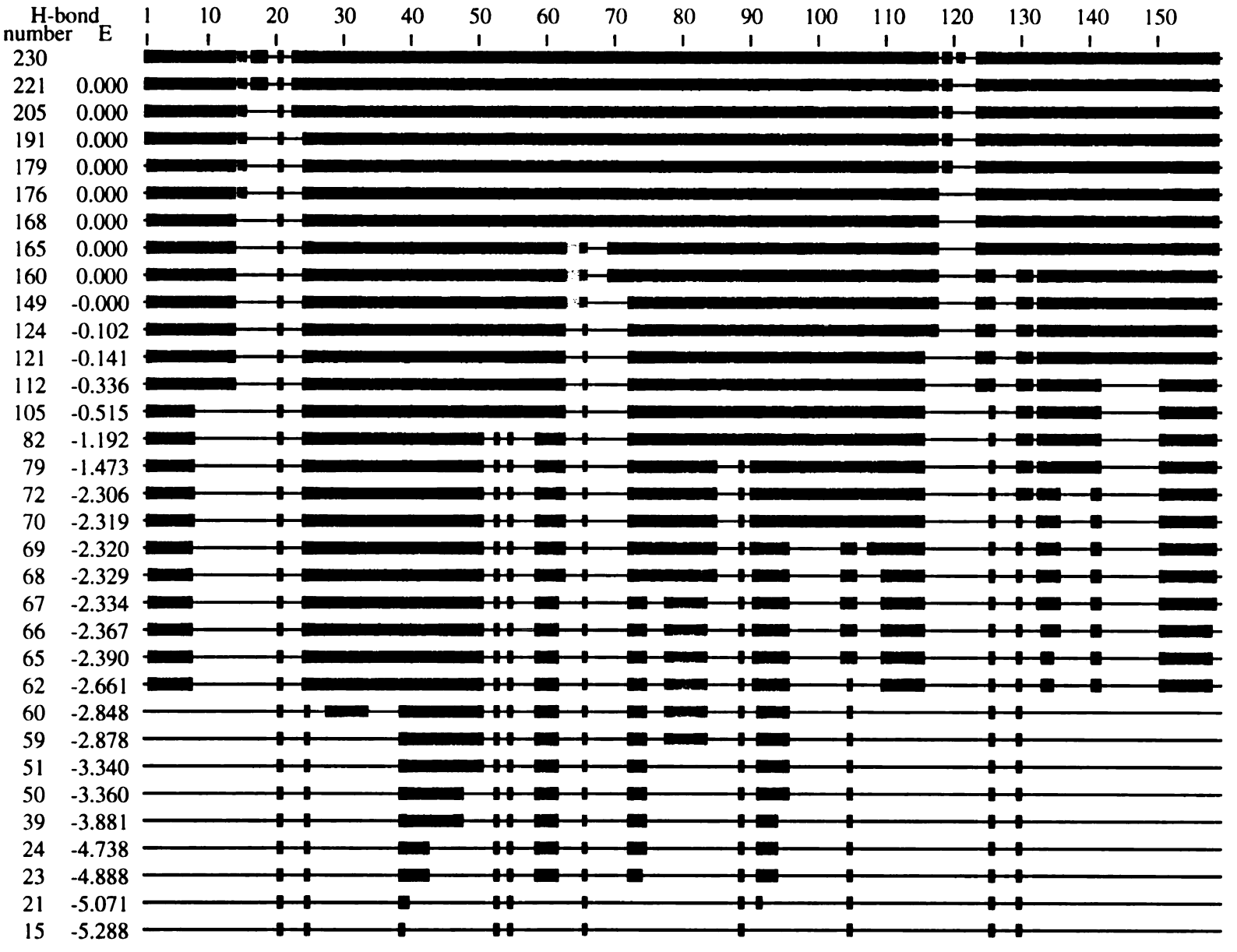
The software FIRST [88] is applied to both conformations. It first calculates the energies of the potential hydrogen bonds based on geometrical considerations. Hydrogen bonds in the two conformations are then re-organized so that the hydrogen bonds are identical in both conformations. A potential hydrophobic interaction is listed when a pair of hydrophobic centers is closer than 0.7Å plus the sum of the van der Waals radii of the two centers. According to Dr. Zavodszky's experience [119] on the usage of FIRST on several proteins, biochemical properties match better with the predictions from FIRST when the critical distance of hydrophobic interactions is

0.7Å plus the sum of van der Waals radii of a pair of hydrophobic centers. A possible hydrophobic interaction is taken as a stable hydrophobic interaction when it is present in both conformations. This procedure creates a list of interactions and constraints that are common to both conformations. The two conformations, when obeying this single set of constraints and interactions, are identical in flexibility properties.

The selection of hydrogen bonds depends on the hydrogen bond cut off energy. Figure 6.12 shows the structural properties of the protein under different hydrogen bond cut off energies. Each horizontal line represents the protein under one particular hydrogen bond cut off energy. The thick bar shows rigid cores while the thin lines are flexible loops. The red and thick bars in the figure denote the biggest rigid cluster in the protein. From top to bottom, hydrogen bonds are accumulatively cut according to their energies. Majority part of the protein is in the single and large rigid core until hydrogen bonds whose energies are higher than -2.661 kcal/mol, which is equivalent to more than 2300 K in temperature, are all cut. The structure of ecDHFR is very stable in this sense.

In ambient temperatures those hydrogen bonds whose energies are less than -1.0 kcal/mol should be considered as stable interactions. The flexibility properties of ecDHFR, when all hydrogen bonds whose energies are less than -1.0 kcal/mol are counted as constraints while all the other hydrogen bonds are eliminated, is shown in Figure 6.13. The core of the protein is rigid as expected. The M-20 loop is flexible. The PRO21 residue is shown as a rigid residue because its main chain ϕ angle is locked by a five fold ring. Appendix B.3 discusses the distribution of main chain dihedral angles of proline residues in detail. In addition to the M-20 loop, the residues from 118 to 129 and from 142 to 150 are also flexible. Because these two ranges of residues are close to the M-20 loop in the coordinate space, the conformational changes of these two ranges of residues are coupled to the conformational fluctuation of the M-20 loops through van der Waals interactions, hydrogen bonds and hydrophobic interactions.

Figure 6.12: The flexibility of the ecDHFR at different hydrogen bond cut off energies. Each horizontal line represents the protein ecDHFR at a particular hydrogen bond cut off energy. Thin line segments signify flexible loop while thick and colored line segments show rigid regions. Thick line segments colored by the same color belong to a same rigid region. The first column of data on the left are the number of hydrogen bonds remaining in the protein structure. The second column of data on the left are the hydrogen bond cut off energy. The data on the top are the residue numbers.


H-bond number E
1 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
230
221 0.000
205 0.000
191 0.000
179 0.000
176 0.000
168 0.000
165 0.000
160 0.000
149 -0.000
124 -0.102
121 -0.141
112 -0.336
105 -0.515
82 -1.192
79 -1.473
72 -2.306
70 -2.319
69 -2.320
68 -2.329
67 -2.334
66 -2.367
65 -2.390
62 -2.661
60 -2.848
59 -2.878
51 -3.340
50 -3.360
39 -3.881
24 -4.738
23 -4.888
21 -5.071
15 -5.288

Residues from 63 to 72, which are shown in color red at the top of the protein in the figure, are also flexible. The conformational changes in these residues do not have obvious biochemical functions.

6.3.3 Sampling Directed Pathways

ROCK samples conformations randomly. Starting from one conformation, it searches a nearby conformation in random directions. In order to search pathways directed from the occluded or the closed conformation of ecDHFR to the other, a simulated annealing procedure [149, 150] is incorporated in our algorithm. The RMSD d between the main chain atoms in a generated conformation and the corresponding atoms in the target conformation is the pseudo-energy of the generated conformation. It is calculated as

$$d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - \mathbf{r}_i^t)^2} \tag{6.3}$$

in which $\mathbf{r}_i$ is the coordinate of the ith atom in the generated conformation and the $\mathbf{r}_i^t$ is that of the same atom in the target conformation. The sum over the index i is over all main chain atoms of interested residues. In this case the sum is over all main chain atoms in the M-20 loop region. Those conformations whose RMSD to the target conformation are smaller than those of their immediately proceeding accepted conformations are always accepted. The other conformations are accepted with a probability of $\exp[-\Delta d/T^*]$, in which Δd is the change of the RMSD to the target conformation since the last accepted conformation and T^* is the pseudo-temperature. The pseudo-temperature is always proportional to the current RMSD to the target conformation by

$$T^* = d\tau \tag{6.4}$$

In this way the pseudo-temperature is high when the RMSD to the target conformation is big so that it is possible for the protein to overcome some pseudo-energy

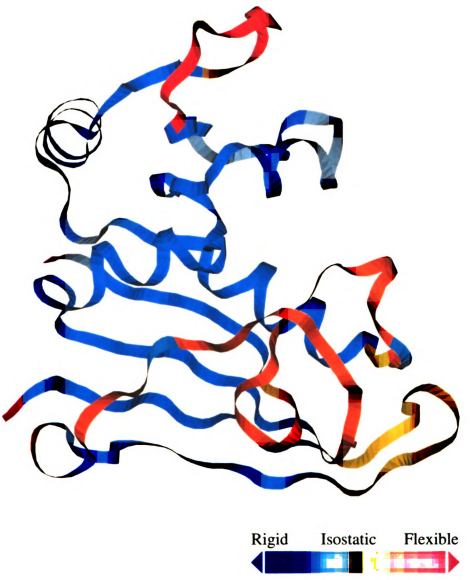


Figure 6.13: Flexibility properties of ecDHFR. The protein shown in this figure is 1RX6 which is in the occluded conformation. Only those hydrogen bonds and hydrophobic interactions that are present in both the occluded and the closed conformations are included in the flexibility analysis. Residues shown in blue and gray are rigid. Residues shown in yellow, gold and red are flexible. The flexibility property is analyzed by the software FIRST.

barriers. The pseudo-temperature is low when the RMSD to the target conformation is small so that only those conformations which are closer to the target conformation than their previously accepted conformations are accepted. The setting of the parameter τ depends on the properties of the conformational space of the proteins being studied. It is set to be 0.05 in this calculation.

The generated conformations are dragged toward the target conformation when the RMSD decreases. Because the trajectories generated by our algorithm are not driven by any empirical potential, they are not the real dynamic pathways of the proteins. However since all of the conformations generated by our algorithm are feasible conformations, the connections of all these conformations link to form possible conformational pathways. These conformational pathways are statistically correct. The properties of an ensemble of possible conformational pathways generated by our algorithm will be the same as those of an ensemble of conformational pathways generated by MD or any other algorithms, when the size of the ensemble is large.

6.3.4 Conformational Pathways of DHFR

Pathways from the Occluded to the Closed Conformation

Six trajectories starting from the occluded conformation targeting at the closed conformation are generated by our algorithm. All of these calculations generate thousands of conformations within two to three days of CPU time on a single AMD Athlon 1900+ processor. The parameter settings for all of the six pathways are the same except the initial random seeds. The random numbers used in the program ROCK are generated by the program authored by L'Ecuyer et al. [151].

The region of our interest is the M-20 loop. There are 11 residues in the M-20 loop. Each residue has four main chain atoms. Therefore the conformations are best defined in a $3 \times 4 \times 11 = 132$ dimensional space. To simplify the analysis, we define three reference points in the high dimensional space, which are the occluded, the

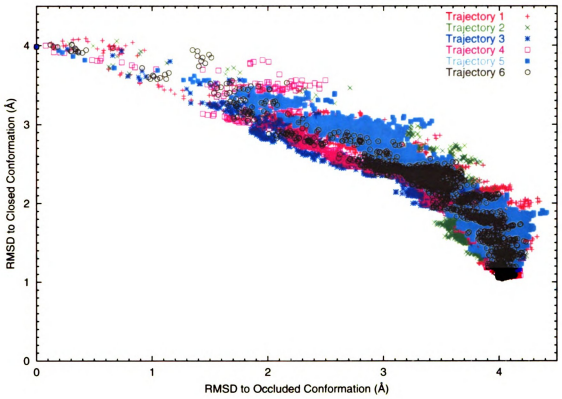


Figure 6.14: The correlation of RMSD of the six trajectories to the occluded and to the closed conformations. The RMSD of all six trajectories are exactly 0.0Å to the occluded and roughly 4.0Å to the closed conformations in the beginnings. The calculations are terminated when the RMSD to the closed conformations are below 1.0Å.

closed and the open conformations. All the conformations in the high dimensional space then can be projected to a simpler three dimensional space, in which the RMSD of a conformation to the three reference points are its coordinates. Trajectories of conformations are easily tracked and examined in the three dimensional space, at the cost that some information is lost when trajectories in high dimensional space is expressed in the newly built three dimensional space.

Figure 6.14 illustrates the correlations between the RMSD of generated conformations to the occluded and the closed conformations. Since the calculation begins from the occluded conformation, the RMSD of conformational trajectories to the occluded conformation are exactly zero at the beginnings. The RMSD of trajectories

to the closed conformation are more than 4.0Å at the first several snapshots. Our calculations are terminated when the RMSD to the closed conformation are below 1.0Å. Because the bond lengths and angles in the starting conformation, the occluded conformation in this case, are not exactly the same as those in the ending conformation which is the closed conformation, our calculation is not capable of driving the RMSD to be exactly zero because the bond lengths and angles are not changed in our algorithm. In a simple test, we build a conformation in which the bond lengths and angles are identical to those of the occluded conformation, and the dihedral angles are the same as those of the closed conformation. The RMSD of the best fit of this manually built conformation to the closed conformation is about 0.6Å. Therefore driving the RMSD down to the vicinity of 0.6Å is the limit of our algorithm when the bond lengths and angles are not disturbed. In reality the lowest RMSD observed in calculations is between 0.8Å and 1.0Å.

All six trajectories do not pass the open conformation, as shown in Figure 6.15. All six trajectories are distinctly away from the open conformation at any point. The smallest RMSD of the trajectories to the open conformation is about 2.0Å. This fact is repeated in Figure 6.16 in which the correlations of RMSD to the closed and to the open conformations are shown.

All the correlations between the RMSD of the trajectories to the three conformations prove that these six trajectories do not differ from each other much. All trajectories fall into same regions in all correlations shown in Figure 6.14, Figure 6.15 and Figure 6.16. This is reasonable considering that there are not obvious barriers among the occluded, the closed and the open conformations. As shown in Figure 6.11, it is not as crowded in the M-20 loop region as it is in other regions in the protein, even when the van der Waals repulsion of the side chains are taken into account. Therefore the M-20 loop can transform from the occluded conformation to the closed conformation without as much difficulty as conformational changes in other regions

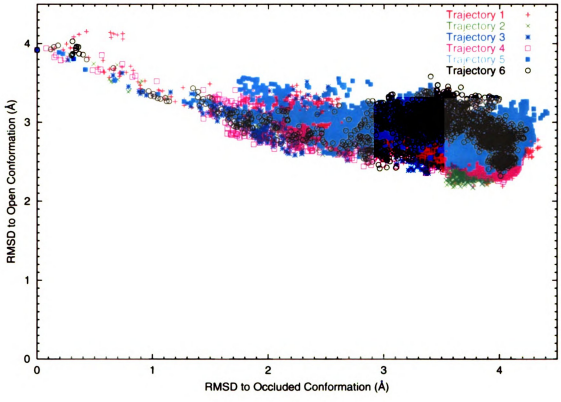


Figure 6.15: Correlations of the RMSD of trajectories to the occluded and to the open conformations. All six trajectories are not close to the open conformation at any point.

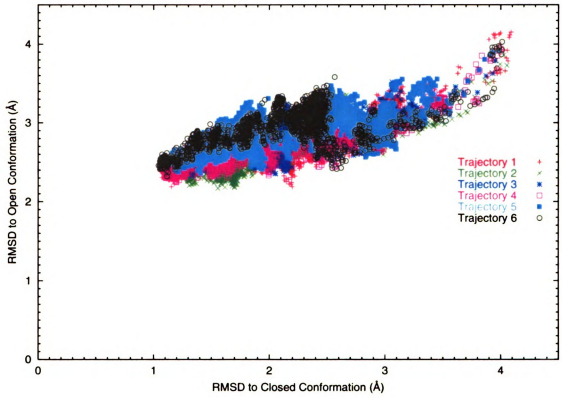


Figure 6.16: Correlations of the RMSD of trajectories to the closed and to the open conformations. This figure, together with the two previous figures, proves that the six trajectories do not differ from each other much.

of the protein encounter. Since the overall shapes of the correlations are almost linear in all of the three figures, we hypothesize that the conformational pathways between the occluded and the closed conformations are roughly linear connections between the two.

Real vs. Main Chain ϕ and ψ Space

Similar to the analysis in real space, we define the dihedral angle RMSD θ (DARMSD) of the trajectories to the three conformations as

$$\theta = \sqrt{\frac{1}{2N}\sum_{i=1}^{N}[(\phi_i - \phi_i^c)^2 + (\psi_i - \psi_i^c)^2]} \tag{6.5}$$

in which ϕ_i and ψ_i are main chain dihedral angles of a generated conformation and ϕ_i^c and ψ_i^c are the corresponding dihedral angles in the closed conformation. The sum is over all residues of interest, which are residues in the M-20 loop in this case.

Since there are 11 residues in the M-20 loop, the conformations are best described in a 22 dimensional space which is spanned by the 22 main chain dihedral angles. Similar to the analysis in real space, we define three reference points which are the closed, the occluded and the open conformations respectively. Any conformation is then expressed in a smaller three dimensional space in which the RMSD to the three reference points are its coordinates.

Figure 6.17 shows the correlations of DARMSD of the six trajectories to the occluded and to the closed conformations. Surprisingly, the DARMSD of the six trajectories to the closed conformation are always bigger than 50°. Some trajectories do not show any inclination to smaller DARMSD values at all. The DARMSD of the six trajectories to the closed conformation in real space are all below 1.0Å when calculations are terminated, but the DARMSD of the same trajectories to the closed conformation in dihedral angle space oscillate around big values.

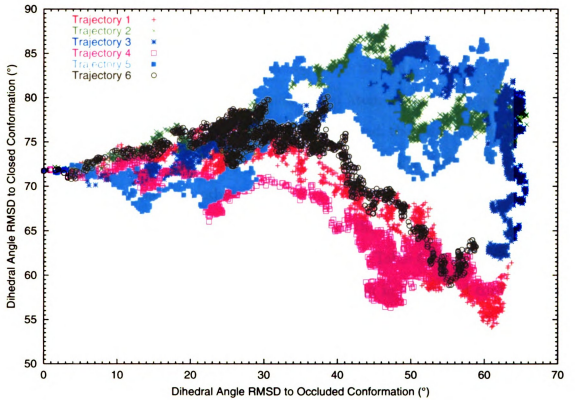


Figure 6.17: Correlation of the RMSD of the six trajectories to the occluded and to the closed conformations. The DARMSD is defined in Equation 6.5.

To further investigate the difference between the RMSD in real space and the DARMSD in dihedral angle space, we analyzed the similarities between a generated conformation and the closed conformation. This generated conformation is one of the conformations which are very near to the closed conformation in term of RMSD in real space. Table 6.1 lists the difference in coordinates of the main chain atoms of the M-20 loop between the two conformations. The RMSD between these two conformations is mere 1.057Å in real space. The biggest distance between corresponding atoms in the two conformations is only 1.216Å. All these data prove that the generated conformation is almost identical to the closed conformation.

Figure 6.18 shows the M-20 loop main chain atoms of both the generated and the closed conformations. Though not perfectly matched, the two conformations wind around each other like the two chains of the double helix. It is obvious that these two conformations resemble each other.

Though these two conformations are virtually the same in real space, they are quite different in the main chain ϕ and ψ angle space. Table 6.2 lists the main chain ϕ and ψ angles of the M-20 loop in the generated and in the closed conformations. The main chain dihedral angles are calculated by the software ViewerLite 5.0 [152] to eliminate the possibility that our program may be erroneous in calculating the main chain ϕ and ψ angles. The DARMSD between the two conformations is 56.9°. Differences in ϕ and ψ angles are also listed in the table. The difference between some corresponding dihedral angles can be more than 100°.

When the RMSD between two conformations is exactly zero in real space, the DARMSD between them must also be zero in dihedral angle space. Therefore, it is intuitive to assume that when the RMSD is small in real space the DARMSD must also be small in dihedral angle space. In contrast our calculations show that the RMSD in dihedral angle space can be very large even when it is small in real space.

It is easy to prove that the case of small DARMSD in dihedral angle space

		Δx	Δy	Δz	R
ILE14	N	-0.305	1.112	-0.073	1.155
	C_α	-0.550	1.077	0.132	1.216
	C	-0.163	0.298	0.123	0.361
GLY15	N	0.016	0.344	-0.248	0.424
	C_α	0.406	-0.391	-0.202	0.599
	C	0.259	0.023	-0.600	0.654
MET16	N	0.542	-0.683	0.308	0.925
	C_α	0.288	-0.304	0.111	0.433
	C	-0.402	-0.164	0.169	0.466
GLU17	N	0.457	-0.281	0.196	0.571
	C_α	-0.001	-0.152	0.190	0.250
	C	-0.852	0.139	0.762	1.151
ASN18	N	0.302	-0.736	0.025	0.796
	C_α	-0.136	-0.632	0.371	0.745
	C	-0.450	-0.377	0.056	0.590
ALA19	N	-0.030	-0.244	-0.312	0.397
	C_α	-0.169	-0.111	-0.544	0.580
	C	0.584	-0.081	-0.498	0.772
MET20	N	-0.195	0.045	0.246	0.317
	C_α	0.390	0.190	0.425	0.607
	C	0.277	0.104	0.076	0.305
PRO21	N	0.458	0.357	-0.045	0.582
	C_α	0.375	0.277	-0.130	0.484
	C	-0.147	0.102	0.492	0.524
TRP22	N	-0.015	0.643	-0.783	1.013
	C_α	-0.403	0.590	-0.427	0.832
	C	0.241	0.223	-0.141	0.357
ASN23	N	-0.708	0.421	-0.152	0.838
	C_α	-0.333	-0.092	0.132	0.370
	C	-0.414	-0.053	0.349	0.544
LEU24	N	0.052	-0.372	-0.545	0.662
	C_α	0.054	-0.352	-0.477	0.595
	C	-0.237	0.000	-0.701	0.740

Table 6.1: Difference in coordinates of main chain atoms in the M-20 loop between the generated and the closed conformations. The difference in coordinates in x, y and z are listed, together with the distance R between corresponding atoms in the generated and in the closed conformations.

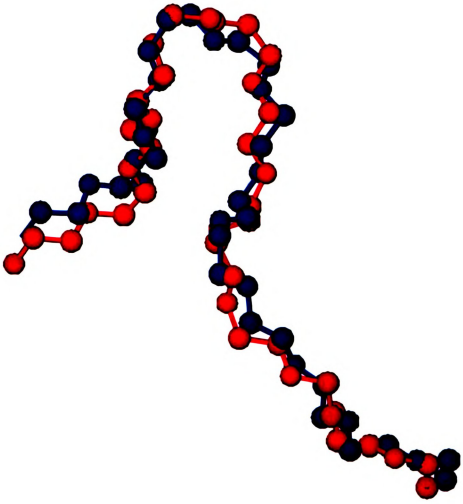


Figure 6.18: Superimposition of the M-20 loop of the generated and of the closed conformations. One conformation is colored red while the other is colored blue. The two conformations are almost identical.

	Generated		Closed			
	$\phi(^\circ)$	$\psi(^\circ)$	$\phi(^\circ)$	$\psi(^\circ)$	$\Delta\phi(^\circ)$	$\Delta\psi(^\circ)$
ILE14	-79.8	-46.1	-111.7	-13.4	-31.9	32.7
GLY15	179.1	133.7	-162.0	161.4	18.9	27.7
MET16	-54.0	167.6	-147.4	109.8	**-93.4**	-57.8
GLU17	29.3	-58.9	50.6	59.6	21.3	**118.5**
ASN18	163.0	24.0	53.9	31.8	**-109.9**	7.8
ALA19	-153.7	105.9	-148.3	154.7	5.4	48.8
MET20	-54.1	127.3	-99.9	123.5	-45.8	-3.8
PRO21	-79.7	57.6	-66.8	-46.5	12.9	**-104.1**
TRP22	-120.1	-145.9	-65.6	178.0	54.5	-37.1
ASN23	165.5	118.1	-134.8	89.2	59.7	-28.9
LEU24	176.2	92.2	-117.4	78.2	66.4	-14.0

Table 6.2: Main chain ϕ and ψ angles of residue 14 to 24 in the generated and in the closed conformations. Angles larger than or close to 100.0° are marked by bold font.

and large RMSD in real space is also possible. Suppose one dihedral angle in a chain molecule is rotated by 180°. The RMSD between the resulting and the initial conformations is large because the flip of one dihedral angle can drag part of the molecule far away from its original coordinates. The DARMSD between the two conformations is small because only one dihedral angle is altered.

The lack of correlations between the RMSD in real space and the DARMSD in dihedral angle space has been seen before. In their study of the variations of main chain ϕ and ψ angles in different conformations of proteins, Korn and Rose [153] discover that the main chain conformation of a protein is not deformed when the main chain dihedral angles are rotated compensatorily. The effect of a big change in the main chain ψ angle in one residue may be offset either by the rotation of the main chain ϕ angle in the next residue by the same amount in the opposite direction, or by small rotations of several main chain dihedral angles in nearby residues. Vieille [154] observes the same phenomenon in her analysis of the conformational fluctuations in a MD simulation trajectory.

The counter-intuitive fact that the RMSD in real space is not correlated with the

DARMSD in dihedral angle space disqualifies the usage of the main chain dihedral angles in the analysis of the similarities between conformations. Whether two conformations are alike should be investigated in real space. On the other hand, the analysis of the variations of the main chain ϕ and ψ angles can be used to pinpoint the exact positions of the conformational changes. Whether to inspect the deviations in real space or in dihedral angle space therefore depends on the questions to be answered.

Pathways from the Occluded and the Closed Conformations to the Open Conformation

The six conformational trajectories between the occluded and the closed conformations do not pass the open conformation. It is not clear whether it is because the trajectories do not need to pass by the open conformation, or because the constraints in the occluded and the closed conformations inhibit the sampling of the open conformation. To clarify this point, two more conformational trajectories are sampled by the ROCK. One trajectory starts from the occluded conformation targeting at the open conformation. The other one starts from the closed conformation targeting at the open conformation.

A trajectory between the closed and the open conformations is successfully explored. The RMSD between generated conformations and the open conformation drops to around 1.0Å after ROCK samples about 2000 conformations. A trajectory between the occluded and the open conformations is also found. The best RMSD between generated conformations and the open conformation is around 1.2Å. The closed conformation is not on the trajectory between the occluded and the open conformations. The occluded conformation is not on the trajectory between the closed and the open conformations too.

To summarize, ROCK successfully generates direct pathways between any two of the three conformations – the closed, the occluded and the open conformations.

The pathways between any two conformations do not pass the third one. Therefore it is possible that the three conformations form a triangle in the high dimensional conformational space. Direct pathways between any two conformations can be built, without the necessity of passing through the third.

However our calculations are not conclusive. The trajectories are sensitive to the constraints. The addition or removal of one or more constraints could change the characteristics of the trajectories. The removal of constraints enables a protein to sample larger conformational space. The addition of constraints compresses the conformational space of a protein. For example an addition of one or few crucial constraints may disqualify the six conformational trajectories between the occluded and the closed conformations. The ecDHFR may have to go through the open conformation along its pathways between the occluded and the closed conformations in this case. The influences of the constraints on the properties of protein motion trajectories deserve further analysis.

Chapter 7: Summary and Perspectives

7.1 Summary

This thesis covers two sub-topics of modeling non-crystalline networks: the modeling of discontinuous networks and the sampling of conformations of non-crystalline networks. We propose and test two algorithms on these two topics. The first algorithm is dedicated on the re-arrangement of dangling bonds in network models so that 1) all atoms in the resulting networks are all fully coordinated and 2) the bond length and angle distortions are within acceptable limits. The second algorithm samples conformations of non-crystalline networks which are composed of complicated ring clusters. Proteins are such networks.

The building of the DCRN models starts from crystalline networks. Voids are cut from the networks, resulting in dangling bonds at the surface of the voids. The defects involving dangling bonds at three-coordinated atoms are transferred to their nearest neighboring atoms at each step of the defect migration process. The defects are then removed when they come together. The defect migration process creates fully coordinated networks.

The networks are randomized to be amorphous by the WWW technique. Since the bond switch process is local in the WWW technique, the voids in the initial networks are roughly unchanged before and after the WWW process. The resulting network is random, but discontinuous in the sense that it can have built in voids of any shapes and sizes. The insertion of oxygen between all silicon-silicon bonds makes amorphous silica models. This algorithm is demonstrated to build the amorphous fiber silica and the amorphous film silica models. The distortions of bond lengths

and angles are reasonable in both models. The variation of distortions in atom layers suggests that the surface effects diminish within five layers to the surfaces.

Since the overall shapes of the amorphous fiber silica and film silica models are confined in one or two dimensions, the PDF of these two models are not directly comparable with that of the CRN silica models. The effects of the overall shape on the PDF can be eliminated by dividing the RDF of the models over that of the continuous and uniform media of the same shape. The obtained RDDF of both the amorphous fiber silica and film silica models are quantitatively in good agreement with that of CRN silica models, except a small shoulder at the second peak which is caused by distortions at the surfaces.

The procedure to build the DCRN models can be used to build any glassy network models. Moreover, defects and hydrogen can be introduced into a fully coordinated random network by reversing the defect migration process used to build the fully coordinated networks. Hydrogenated networks can be built with any concentration and distribution of hydrogen.

If desired, models produced by this algorithm can be further optimized by other techniques, for example by MD simulations with empirical potentials or by *ab initio* algorithms. The refined models are then suitable for the studies of the electron states of defects, of the photo luminescent characteristics of voids, and of the electron distributions around hydrogen atoms et al.

The amorphous metal-adamantane network is modeled starting from a CRN gallium arsenide model. Gallium and arsenic atoms are replaced by the metal atoms and the adamantane units respectively. Since the amorphous gallium arsenide model does not have odd-numbered rings, each metal atom is bonded to four adamantane units and each adamantane unit is bonded to four metal atoms. The calculated X-ray powder diffraction pattern fits that measured in the synchrotron experiment.

There are only covalent bonds in the DCRN models. The concentration of the

covalent bonds in DCRN networks is so high that DCRN models we examined are all rigid. The concentration of bonds in another type of non-crystalline network, namely the protein, is not high enough to restrict the relative positions of atoms, so some regions of proteins have multiple conformations. Proteins have evolved in such a way that some regions of the proteins are rigid while the other regions are flexible. Constraints in the rigid regions stabilize the proteins and define a template for interacting with other molecules. The flexible regions of proteins can carry out biological functions. Since distortions of bond lengths and angles are energetically unfavorable, it is preferred to sample the conformations without disturbing the bond lengths and angles. This task is not easy for proteins, because there are lots of rings in proteins which are composed of covalent and hydrogen bonds. The concentration of the bonds is high, so virtually all these rings are inter-locked with each other. The bonds linking the rings bring extra constraints on how rings are relatively positioned against each other. It is therefore vital to build an algorithm that closes all the rings simultaneously.

Inspired by the RCE proposed by Gō and Scheraga [73], we define a fictitious ring closure potential which is the sum of the squares of the RCE. A set of dihedral angles is the solution to RCE if and only if it makes the fictitious ring closure potential to be zero. Since the fictitious potential is non-negative everywhere, the potential is at one of its minima whenever its value is zero. Therefore by minimizing the fictitious potential, our algorithm is able to numerically solve the RCE. This method can close any rings, regardless of how the non-rotatable and rotatable dihedral angles are mixed. The main advantage of this method is that by minimizing the total fictitious potential of all the rings in a complicated ring cluster, it is able to close all the rings simultaneously. All the bond constraints are concurrently satisfied. The ability to close all rings and to generate stereo-chemically correct structures makes our algorithm efficient in sampling conformations for proteins, which are composed

of numerous inter-locked rings.

Our method of solving the RCE is tested on a model molecule, which is made up of four inter-connected rings. The molecule has only two DOF, so its conformations can be easily projected on a 2D graph. The distributions of the conformations generated by our algorithm show two symmetries which are identical to the two symmetries of the topology of the molecule. This fact supports the claim that, given enough calculation time, our algorithm can sample all conformations of a small macromolecule, if these conformations are continuous in the conformational space. While all the dihedral angles in a large macromolecule cannot be sampled exhaustively, the combination of ROCK and FIRST makes ROCK sample the important DOF in the system.

The application of flexibility analysis [88] to proteins reduces the calculation cost by differentiating the flexible regions from the rigid regions in proteins. The rigid regions of proteins have negative or zero DOF. They may have multiple conformations, but these conformations are well separated in the conformational space, so that it is reasonable to assume these rigid regions do not sample more than one conformation under usual conditions. The program ROCK samples conformations only for the flexible ring clusters which are identified from the flexibility analysis. It saves calculation time by avoiding sampling conformations for the rigid regions of proteins.

ROCK first perturbs the rotatable dihedral angles in the flexible regions of proteins by modest rotations which are typically within 5° to 10°. It then closes all the rings by minimizing the total fictitious ring closure potential. It randomly samples side branch conformations as well. It has three minimal requirements on side branch conformations: 1) the bond lengths and angles should be undistorted from their original values; 2) there should be no van der Waals overlaps between non-bonded atoms; and 3) the chirality at all chiral centers should not be inverted. In the case when a bond in the side branch is not rotatable, a distance constraint between the two nearest neighboring atoms of the two ends of the bond is imposed, so that the bond

is effectively locked.

ROCK also checks the quality of the generated protein conformations on the main chain Ramachandran plot. It rejects those conformations whose main chain ϕ and ψ angles are not in the preferred regions in the plot (see Appendix B). The distributions of the dihedral angles in the side chains have certain patterns [155]. An additional check on the quality of side chain conformations against the dihedral angle distributions in the rotamer library [156, 157] could be added to the program. However the additional check would bring an additional calculational cost to our algorithm. The side chains often adopt to non-rotameric dihedral angles. Moreover, at least in the examples of HIV-1 protease and DHFR, the main interest of sampling the main chain conformations does not require elegant algorithms on the side chain conformation sampling, if there is a feasible set of side chain conformations for the given main chain conformation. By keeping only the bond length and angle constraints, ROCK samples the conformations that are consistent with the constraints.

ROCK has been applied on two proteins, the HIV-1 protease and DHFR. Our algorithm did a very good job on sampling the conformations of the HIV-1 protease. The distances between the tips of the flaps can be as large as 8.0Å in some conformations. This is indeed a big distance considering the fact that such distance is only 2.7Å in the crystal structure of the protease which is used as the initial conformation. Our calculation shows that a curling motion is not necessary for the two flaps to open. A detailed analysis of the bond constraints in the tips of the flaps suggests that such curling motion may not exist at all. Because the hydrogen bond between the main chain hydrogen atom of the residue GLY52 and the main chain oxygen atom of residue GLY49 encloses the main chain ϕ and ψ bonds of residue GLY51 into a ten-fold ring which has only one DOF, the distribution of the main chain ϕ and ψ angles of GLY51 is limited in a very narrow region in the Ramachandran plot. One explanation is that this hydrogen bond is not dynamically stable, though it has a

favorable energy in the crystal structure. However, it is safe to say that any curling motion is independent of the slow flap motions.

ROCK has been used to generate six conformational trajectories between the occluded and the closed conformations of the protein ecDHFR, one trajectory between the closed and the open conformations, and one trajectory between the occluded and the open conformations. Based on the properties of these conformational trajectories, we hypothesize that the open, closed and occluded conformations of ecDHFR form a triangle in the conformational space. The direct conformational pathways between any two conformations do not necessarily pass the third one. The six conformational trajectories between the occluded and the closed conformations are very similar to each other.

We also analyzed the similarities between the generated conformations and the closed conformation. The RMSD in real space between the generated conformations and the target conformation can be as low as 1.0Å. The DARMSD in dihedral angle space between the same conformations and the closed conformation are always large. This analysis yields the surprising fact that there are virtually no correlations between the RMSD in real space and the DARMSD in dihedral angle space. Two conformations that are close in terms of RMSD in real space may have little similarity in their dihedral angles. Real space RMSD values of two conformations that are similar in their dihedral angles may be large. Therefore, caution should be taken in choosing how to analyze protein trajectories. Conformations should be compared in real space to answer the question of whether they are similar. The analysis in dihedral angle space is good at locating the dihedral angles that have big variations and result in conformational changes.

7.2 Limitations

Both the algorithms to build the DCRN models and to sample protein conformations are somewhat limited in their applications. The former algorithm is suitable for modeling amorphous networks made of strong covalent bonds. The atoms in the network models are limited to the group IV semiconductor elements of silicon and germanium, and to the group VI elements of oxygen, sulfur and selenium. The four strong covalent bonds of group IV elements usually form a tetrahedron. Covalent bonds of group VI elements favor certain geometries too. The bonds of all elements mentioned above can be described simply by geometries of bond lengths and angles. It is for this reason that our algorithm to build the DCRN models and the WWW technique to build the CRN models are able to predict the characteristics of these networks based on simple geometry considerations. Both our algorithm and the WWW technique do not apply when the covalency of elements is weak. Two other group IV elements, tin and lead, bond with their nearest neighbors more like metal than covalent elements. Our procedure is not appropriate in building glassy networks for these two elements. Valency electrons in the metallic atoms flow freely around, causing the bonds around metallic atoms to lack fixed bond angle geometries. Therefore, our algorithm is not suitable for the building metallic glassy networks or semi-conductor glassy networks with metal atoms.

Proteins are mainly built of carbon, nitrogen, oxygen and hydrogen. All these atoms are bonded together by covalent bonds and by hydrogen bonds with certain optimal geometries of bond lengths and angles. Our algorithm to sample protein conformations is also based on these simple geometry considerations. Care should be taken when sampling conformations of proteins with buried metallic atoms to model their bonds correctly.

All of the limitations in both algorithms are caused by the lack of consideration of quantum effects and electronic states. In general, our algorithms cannot be used to

study reactions related to the re-distribution of electrons. The initial and final states of the hopping of a hydrogen atom in the amorphous silica network can be simulated in our algorithm, but the detailed bond breaking and reforming mechanism cannot be. Though our program ROCK can sample conformations of ecDHFR, it does not study such questions as how a hydrogen atom is transferred from NADPH to the DHF in ecDHFR. These types of questions can be addressed by QM/MM hybrid simulations.

7.3 Applications and Perspectives

Non-crystalline networks have rich physics due to the lack of long range order. Local atomic arrangements in non-crystalline networks are roughly identical to those in crystalline networks. For example, the bond angles at silicon atoms in non-crystalline networks are all roughly 109°, which is the optimal silicon bond angle in crystalline networks. But non-crystalline networks do not have any long range order as crystalline networks do. The absence of long range order brings two difficulties in simulations: the complicated energy landscape and the large number of atoms included in the simulation.

The number of atoms involved in *ab initio* calculations is usually between 10 and 100. Due to the long range translational and rotational symmetries, electronic properties of roughly 10 or more atoms in a small supercell in the crystalline network are identical to those of all of the other atoms in the whole network. Therefore, the simulations in crystalline networks can be very accurate through the use of *ab initio* algorithms. Simulations in non-crystalline networks, however, do not have such an advantage. Because of the dearth of long range order, an infinite number of atoms, in principle, should be included in the simulation. The supercell of an amorphous network typically has thousands or even millions of atoms. The supercell of MD simulations on proteins with explicit water contains one large protein surrounded

by tens of thousands of water molecules. In all of these cases, the precise quantum calculations have to be confined to a small region of the network, for example a dozen or fewer atoms on the surface of amorphous silica, or ten atoms or so in the catalytic sites of proteins. The large scale structures and the long range motions of non-crystalline networks have to be modeled by empirical algorithms.

The local distortions in bond lengths and angles lead to complicated energy landscapes. Since atoms in one supercell are equivalent to the atoms in all the other supercells in a crystalline network, the energy levels of the one supercell are degenerate with those of the other supercells. The energy landscapes of crystalline networks are clean and easy to interpret. The degeneracy is lost when bonds are distorted. The consequence of the distortions is a very complicated energy landscape with numerous local energy minima. The energy barriers between these local energy minima may vary a lot. The energy barriers of the amorphous networks, for example, are high enough so that the networks are trapped in local minima even under high temperatures. An amorphous silicon network is in one local minimum in the whole energy landscape of silicon networks, because its potential is higher than that of a crystalline silicon network. The energy barriers of proteins, on the other hand, are low enough so that proteins can transform from one conformation to another fairly freely under room temperature.

Our model molecule described in Section 6.1 exemplifies how complicated the energy landscape can be even for small molecules. It has more than 230 local minima, some of which are shown in Figure 6.2. The energy landscape of this simple molecule containing only 36 atoms is already complicated, not to mention a large protein or a large amorphous fiber silica network.

The protein energy landscapes have been studied by various techniques [158, 159, 160, 161]. The funnel picture [162, 163, 164, 165] has been accepted by many scientists. It states that a protein has one global minimum conformation, which is the folded

state, and many other local minimum conformations, which are the intermediate or the unfolded states. The whole landscape is very rough, with numerous local minima.

The sampling of all local energy minima is not possible for large proteins. The sampling of all minima for small molecules and for large proteins in the vicinity around a given conformation are computationally feasible. Our program ROCK cannot be utilized directly to sample the local minima in protein energy landscapes, because the effective force field used in ROCK is not the same as the usual ones used in MD simulations. The force field used in ROCK is a sum of a series of infinitely high step functions. Any violations of the bond length and angle constraints or any van der Waals overlaps push the effective potential to be infinitely high. The dihedral angles are freely rotatable, as long as the rotations do not result in van der Waals overlaps and the break of rings. The force fields used in standard MD simulations are much more complicated. Potentials of bond length and angle variations are high but finite. Many weak interactions, such as the dihedral angle rotation interaction and the electrostatic interaction, are included in the force fields used in standard MD simulations. The advantage of ROCK is that by ignoring the details of energy variations caused by the weak interactions, it is capable of easily jumping over small energy barriers that require calculation cost in other algorithms utilizing more complex force fields. The coupling of ROCK and commonly used force fields enables a fast sampling of conformations without losing details of energy landscape. Such an algorithm can be a good tool in scanning the maps of local minima of the protein energy landscape. The sampling of local minima of the model molecule in Section 6.1 is such an example. There are also methods of predicting saddle points [166, 167, 168] around local minima in complicated energy landscapes. Once all local minima and saddle points are known, statistical quantities such as transition rate, relative populations and conformational pathways are easy to obtain.

The goal of protein docking is to filter tens of thousands of potential ligands

(drugs) for their ability to match the binding sites of proteins. In principle, both the flexibility of ligands and that of the proteins should be taken into account in docking studies. In many studies, however, only the flexibility of ligands is considered, for example, the methods used in DOCK [169] and in a modified version of DOCK [170], and the more complicated approach by Given et al. [171]. Or, in some cases, only the side chain flexibility of the proteins is included in the calculations [172, 173]. It is only recently that the flexibility of main chains of proteins catches the attention in the protein docking studies [175]. Since ROCK is efficient at sampling conformations for ring clusters, it is a perfect tool to build conformational libraries in which protein main chain and side chain conformations are stored. Zavodszky et al. [176, 177] report improved docking when flexible drugs are docked to the ensemble of protein conformations created by ROCK. Flexibility of the ligands is handled by the tool, SLIDE [178, 179].

It is interesting that the length scale of both the pores in the mesoscopic porous network models and proteins are in the order of nanometers. Physics on the length scale of angstroms requires quantum mechanics to interpret. Physics on larger length scales such as micrometers and millimeters can be safely described by classical mechanics of continuous media. It is in the length scales of nanometers to micrometers that classical mechanics based on single atoms is the suitable tool, on the condition that the interactions involved are not subject to large fluctuations. Covalent bonds are the stable interactions in amorphous material. We include covalent bonds, strong hydrogen bonds and hydrophobic interactions as stable interactions in our sampling of protein conformations. The algorithm to build DCRN models, the algorithm of ROCK, and MD simulations, are dedicated trials to build a set of geometrical language that is precise and fast to describe the physics in the nanometer length scale and up. All these efforts will facilitate future studies and simulations of larger and more complicated systems, such as protein-protein interactions, viruses, semiconductors in

the nanometer range and the micro-machines.

APPENDIX

Appendix A: Radial Distribution Function of Uniform Media

RDF of infinite and uniform media is $T(r) = 4\pi r^2 \rho_0$ in which ρ_0 is the average density. The RDF is proportional to r^2 in any distance range in the infinite media. The RDF of uniform media distributed in the shape of a film and in the shape of a fiber do not follow this square law due to the anisotropic mass distribution.

Suppose uniform mass is distributed in an infinitely wide and flat film of thickness d in space. Let $z = 0$ plane is the center of the film and let $z = d/2$ and $z = -d/2$ be the top and bottom surfaces of the film. Suppose a point $(0, 0, z_0)^T$ which is on the z axis is the observation point. Due to the mirror symmetry of mass distribution about the $z = 0$ plane, the RDF observed from a point whose z coordinate is positive is the same as that observed from its mirror point about the $z = 0$ plane. Therefore we limit our considerations to points above the $z = 0$ plane. Since the observation point should be inside of the mass distribution we have $z_0 \leq d/2$.

RDF at distance r observed from this point is proportional to the surface area of a sphere of radius r that is buried in the film. When the distance r is small, as shown in Figure A.1(a), complete surface of a sphere that is centered at the observation point is within the top and the bottom surfaces of the film. In this case, the RDF observed from the particular point at $(0, 0, z_0)^T$ is

$$T(r; z_0) = 4\pi r^2 \rho_0 \tag{A.1}$$

which is valid in the range of $r \in [0, d/2 - z_0)$.

When the radius r of the sphere increases beyond the point of $r = d/2 - z_0$, the sphere crosses with the top surface of the film, as shown in Figure A.1(b). The

surface area of the sphere that is still buried inside of the film is less than $4\pi r^2$. Simple calculation shows that the RDF observed from the point of $(0, 0, z_0)^T$ is

$$T(r; z_0) = (2\pi r^2 + \pi dr - 2\pi z_0 r)\rho_0 \tag{A.2}$$

which is valid when r is in the range of $r \in [d/2 - z_0, d/2 + z_0)$.

When the radius r of the sphere is larger than $d/2 + z_0$, the sphere crosses with both the top and the bottom surfaces of the film, as shown in Figure A.1(c). The RDF in the large distance range is

$$T(r; z_0) = 2\pi dr\rho_0 \tag{A.3}$$

which is valid when $r \in [d/2 + z_0, +\infty)$.

Therefore the RDF observed from one particular point of $(0, 0, z_0)^T$ is:

$$T(r; z_0) = \begin{cases} 4\pi r^2\rho_0 & 0 \le r < \frac{d}{2} - z_0 \\ (2\pi r^2 + \pi dr - 2\pi z_0 r)\rho_0 & \frac{d}{2} - z_0 \le r < \frac{d}{2} + z_0 \\ 2\pi dr\rho_0 & \frac{d}{2} + z_0 \le r \end{cases} \tag{A.4}$$

RDF of the whole mass distribution is the averaged RDF observed at every point in the media. Because of the translational symmetry in the x and y directions, any point whose z component is z_0 observes the same RDF as shown in Equation A.4. Therefore the average is only necessary along the z axis where z_0 ranges from 0 to $d/2$.

When $0 \le r < d/2$, the first case in Equation A.4 is valid when $0 \le z_0 < d/2 - r$, while the second case in the equation is valid when $d/2 - r \le z_0 < d/2$. There is no z_0 in the full range of $0 \le z_0 \le d/2$ satisfies the third condition in Equation A.4. The

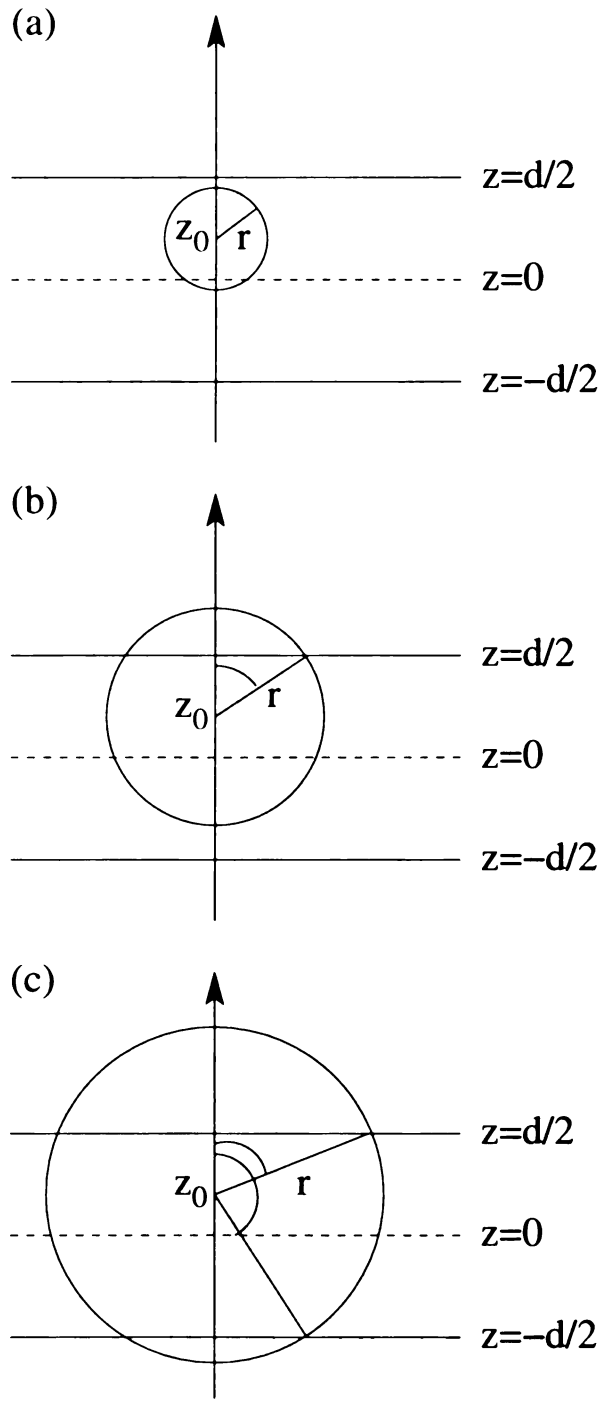


Figure A.1: A sphere of radius r. It can be totally inside of a continuous media of the shape of infinitely wide film (Figure (a)), or it may cross with one surface of the film (Figure (b)), or it may cross with both surfaces of the film (Figure (c)).

RDF of the film media is hence

$$\begin{aligned} T(r) &= \frac{1}{\frac{d}{2}}\left[\int_0^{\frac{d}{2}-r} T(r;z_0)dz_0 + \int_{\frac{d}{2}-r}^{\frac{d}{2}} T(r;z_0)dz_0\right] \\ &= \frac{1}{\frac{d}{2}}\left[\int_0^{\frac{d}{2}-r} 4\pi r^2\rho_0 dz_0 + \int_{\frac{d}{2}-r}^{\frac{d}{2}} (2\pi r^2 + \pi dr - 2\pi z_0 r)\rho_0 dz_0\right] \\ &= 4\pi r^2(1-\frac{r}{2d})\rho_0 \end{aligned} \tag{A.5}$$

which is valid when $r \in [0, d/2)$.

When $d/2 \leq r < d$, the first case in Equation A.4 is invalid for all possible values of z_0 in the range of $[0, d/2]$. The second case is valid in the range of $r-d/2 \leq z_0 \leq d/2$ while the third case is valid in the range of $0 \leq z_0 < r - d/2$. An average of $T(r; z_0)$ over all possible z_0 values produce

$$T(r) = 4\pi r^2(1-\frac{r}{2d})\rho_0 \tag{A.6}$$

which is valid when $r \in [d/2, d)$. The RDF of the whole media in this distance range happens to be in the same form as that in the range of $r \in [0, d/2)$. Similar analysis produces

$$T(r) = 2\pi dr\rho_0 \tag{A.7}$$

when $r > d$.

Therefore the RDF of an uniform mass distribution in a film of thickness d is

$$T(r) = \begin{cases} 4\pi r^2(1-\frac{r}{2d})\rho_0 & 0 \leq r < d \\ 2\pi dr\rho_0 & r > d \end{cases} \tag{A.8}$$

The RDF of the film mass distribution is proportional to r instead of r^2 in the large distance. Similar to the RDF of infinite uniform distribution, the RDF of mass distribution in the shape of a film is proportional to $4\pi r^2$ in the small distance range

where $r \to 0$.

The RDF of the mass distribution in a form of a fiber can be calculated by the same technique, though the derivation is much more complicated due to the complicated integration of the spherical surface area that is buried inside of a cylinder. The final result is

$$T(r) = \begin{cases} 4\pi r^2 \rho_0 \left[1 - \frac{8d}{3\pi r}(1 + \frac{r^2}{4d^2})E(\frac{r}{2d}) + \frac{8d}{3\pi r}(1 - \frac{r^2}{4d^2})K(\frac{r}{2d})\right] & r < 2d \\ 4\pi r^2 \rho_0 \left[1 - \frac{8d}{3\pi r}(1 + \frac{r^2}{4d^2})E(\sin^{-1}(\frac{2d}{r}), \frac{r}{2d}) + \frac{16d^2}{3\pi r^2}(1 - \frac{r^2}{4d^2})K(\frac{2d}{r})\right] & r > 2d \end{cases} \tag{A.9}$$

in which d is the radius of the fiber and functions E and K are elliptic integrations defined as

$$E(\phi, k) = \int_0^{\phi} \sqrt{1 - k^2 \sin^2\theta}\, d\theta$$
$$K(k) = \int_0^{\frac{\pi}{2}} \frac{d\theta}{\sqrt{1 - k^2 \sin^2\theta}} \tag{A.10}$$

The RDF of mass distributions of an infinitely long fiber is proportional to $4\pi r^2$ in the short distance range, and approaches to a constant in the long distance range.

Figure A.2 shows the RDF of mass distribution in infinite media, in a film of thickness d and in a fiber of radius d. The difference in the three RDF in the long distance range is obvious.

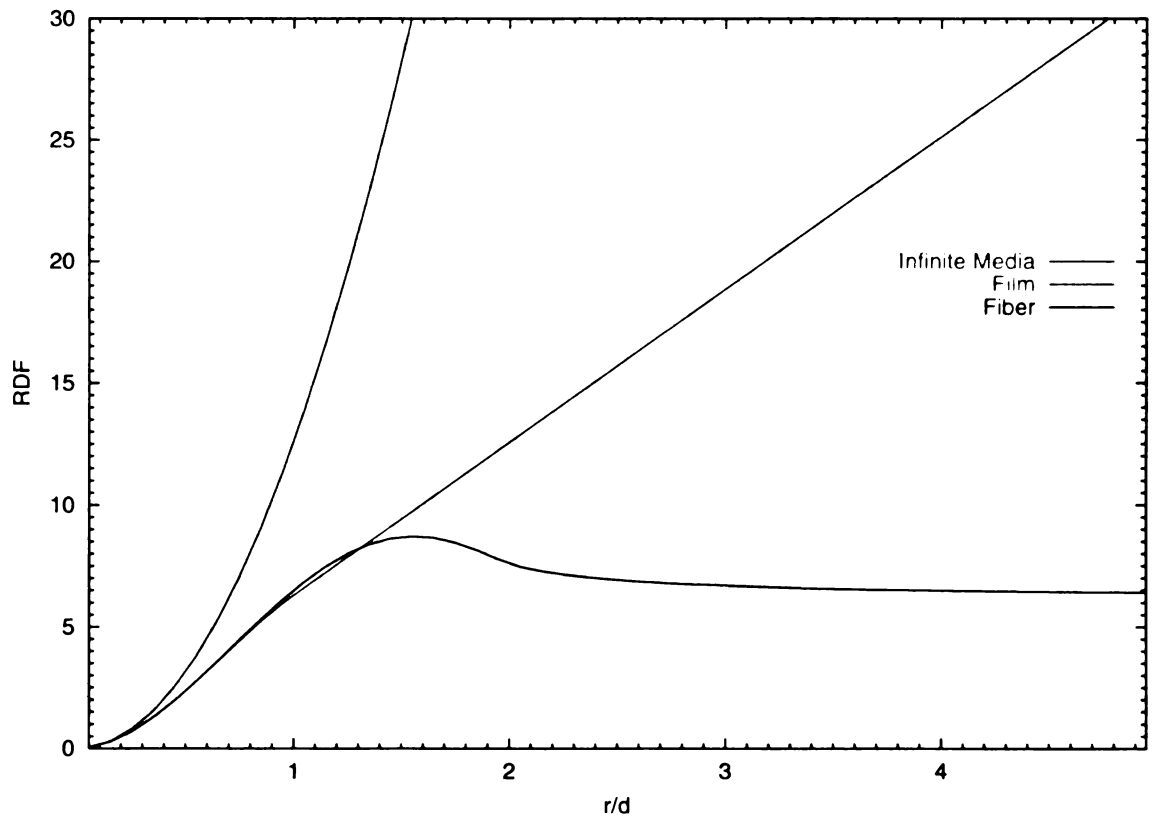


Figure A.2: RDF of uniform mass distribution in infinite (red curve), in a film of thickness d (blue curve) and in a fiber of radius d (green curve).

Appendix B: Ramachandran Plot

B.1 Ramachandran Plot of Residues Other Than Glycine and Proline

As discussed in Chapter 4, proteins are made up of 20 amino acid residues. As shown in Figure 4.4, the main chain ϕ and the ψ dihedral angles are freely rotatable. Certain values of these two dihedral angles would bring the side chain groups in close contact with the main chain atoms, resulting in high van der Waals potential energies, which are unfavorable. Ramachandran et al. [110] systematically examined the intra- and inter-residue van der Waals contacts at all possible combinations of the main chain ϕ and ψ dihedral angles. Their study led to the construction of the Ramachandran plot, which shows the favorable combinations of the ϕ and ψ angles that do not produce van der Waals collisions.

The Ramachandran plot was later refined to reflect the statistical preferences for ϕ and ψ dihedral angle values observed in high resolution crystal structures of proteins [125, 180, 181]. These distributions are used to assess the quality of experimentally determined or predicted protein structures.

Except for glycine and proline, the other 18 standard amino acid residues share a similar distribution of favored ϕ and ψ angles in the Ramachandran plot, shown in Figure B.1. The values of the ϕ and ψ angles are most likely to be in one of the three large regions of the plot. The ϕ and ψ angles of right handed α helices, of left handed α helix's, and of β sheets are in the regions whose centers are at ($\phi = -100°$, $\psi = -30°$), ($\phi = 60°$, $\psi = 50°$) and ($\phi = -120°$, $\psi = 150°$) respectively.

The plot by Morris et al. [125] is widely used in protein structure validation. Based on the experimentally observed statistical preferences, the boundaries of these regions can be set at various cut off values, depending on what percentage of the angles

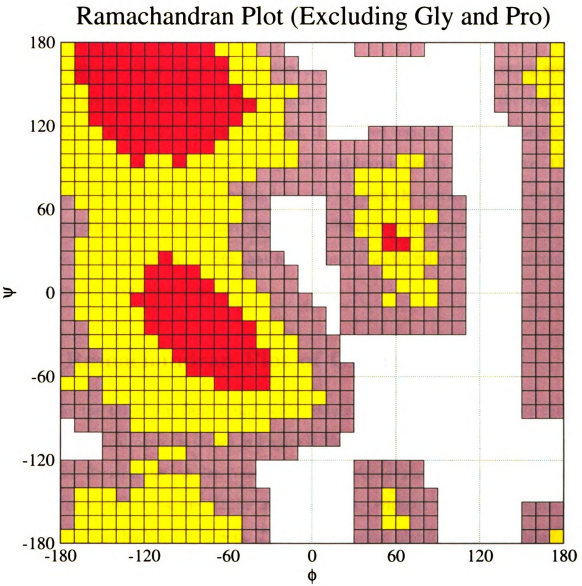


Figure B.1: Reproduction of the Ramachandran plot by Morris et al. [125]. Red, yellow and brown blocks represent the core, allowed and generously allowed regions, respectively. White regions represent the disallowed regions. A residue has van der Waals collisions with its adjacent residues if its ϕ and ψ angles are in the white regions in the plot.

fall within the given boundaries. Over 90% of the residues in the protein structures fall in the *core* regions. Most of the remaining 10% of the residues are in the *allowed* regions. Morris et al. designate the regions that are within 20° of the allowed regions as *generously allowed* regions. The percentage of the residues in the generously allowed regions is at most 1% or 2% in well-resolved protein structures. The remaining areas of the plot are the *disallowed* regions. Figure B.1 is a reproduction of the original data by Morris et al. [125]. The core regions, the allowed regions, and the generously allowed regions are represented by red, yellow and brown blocks in the figure. Our program ROCK utilizes this Ramachandran plot to check the stereo-chemical quality of residues other than glycine and proline in generated conformations, and can be set to discard those conformations that have ϕ and ψ angle values in the generously allowed and/or disallowed regions.

B.2 Ramachandran Plot of Glycine

Unlike the other standard amino acid residues, glycine has only one side chain atom, which is hydrogen, as shown in Figure B.2(a). All other residues contain larger side chains. Because the van der Waals radius of a hydrogen atom is small, the ϕ and ψ angles of a glycine residue can be in a large range without introducing intra- or inter-residue van der Waals contacts. The Ramachandran plot for glycine is therefore qualitatively different from the plot of the majority of residues which is shown in Figure B.1.

The Ramachandran plot containing the distribution of ϕ and ψ angles of glycine residues in high resolution crystal structures [182] was provided to Dr. Zavodszky by Dr. Laskowski (University College London) . The data show the numbers of occurrences of glycine main chain ϕ and ψ angles in protein structures in $8° \times 8°$ blocks. The number of occurrences in every block is first converted into the frequency of

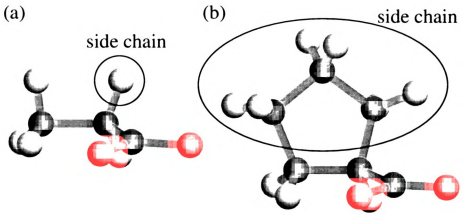


Figure B.2: The glycine residue shown in Figure (a) has only one hydrogen atom in its side chain. The side chain of proline shown in Figure (b) links with the main chain to form a five-fold ring. Carbon, nitrogen, oxygen and hydrogen atoms are represented by green, blue, red and white spheres respectively. The side chain atoms in both residues are enclosed in circles.

occurrence by dividing the number of occurrence by the total number of occurrences of ϕ and ψ angles in the whole $360° \times 360°$ range. Starting from the block with the highest frequency of occurrence, we mark the blocks in order of their decreasing occurrence until 90% of occurrences are accounted for. These blocks form the core regions in the glycine Ramachandran plot. All the other blocks with a non-zero frequency of occurrence are labeled as allowed regions of the glycine Ramachandran plot. The generously allowed regions were defined following the same procedure as in the case of the non-glycine and non-proline residues. The final glycine Ramachandran plot is shown in Figure B.3. Our glycine Ramachandran plot is in qualitative agreement with the plot created by Lovell et al. [181].

B.3 Ramachandran Plot of Proline

Proline is a unique amino acid. Unlike all the other residues whose side chains are bonded to their main chains through one bond, proline has two covalent bonds between its main chain and side chain forming a five-fold ring, as shown in Figure B.2.

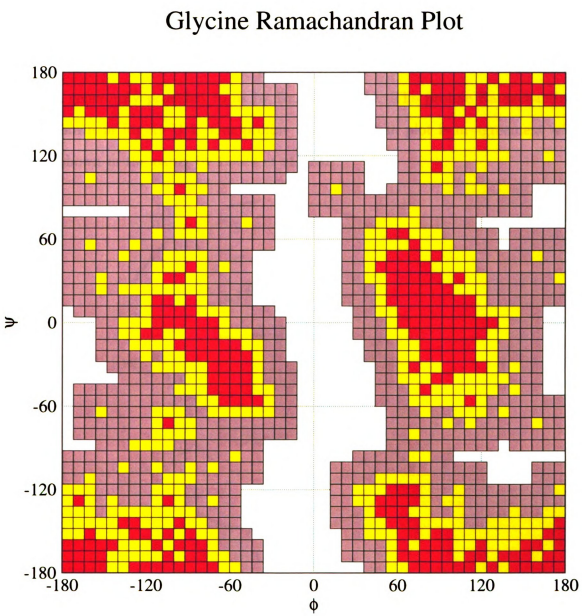


Figure B.3: The glycine Ramachandran plot shown in $8° \times 8°$ resolution. The red, yellow and brown blocks are the core, allowed and generously allowed regions. The white regions are disallowed regions.

As discussed in Chapter 4, a five-fold ring is rigid, so the main chain ϕ angle which is in the five-fold ring of proline is not rotatable. The ϕ angles of proline residues in all high resolution protein structures are therefore restricted within a narrow range.

Following the same procedure by which we created the glycine Ramachandran plot, as explained in Section B.2, we generated the proline Ramachandran plot from the data generously provided to us by Dr. Laskowski. The proline Ramachandran plot is shown in Figure B.4. Our proline Ramachandran plot is in agreement with the one shown in the paper by Lovell et al. [181].

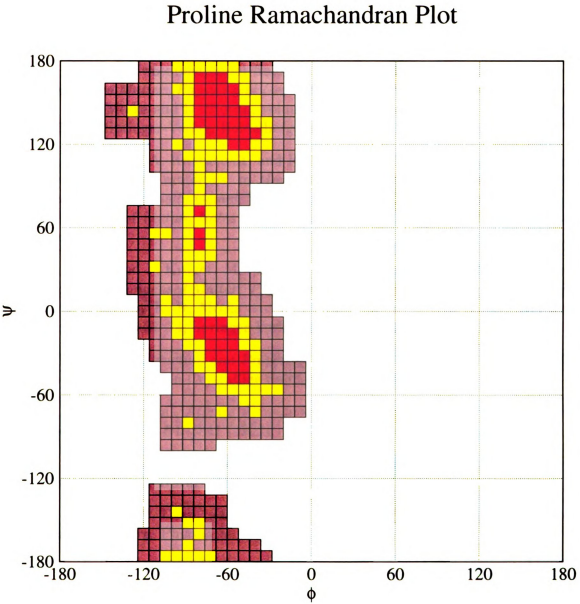


Figure B.4: The proline Ramachandran plot shown in 8° × 8° resolution. The red, yellow and brown blocks are the core, allowed and generously allowed regions. The white regions are disallowed regions.

BIBLIOGRAPHY

[1] M. F. Atiyah, Geometry and physics: A marriage made in heaven, lecture addressed to Department of Physics at University of Michigan., 2001, video and slides of this lecture are collected by the Web Lecture Archive Project. They are available for free download at http://www.wlap.org.

[2] F. Graner, Two-dimensional fluid foams at equilibrium, in *Morphology of Condensed Matter: Physics and Geometry of Spatially Complex systems*, edited by K. R. Mecke and D. Stoyan, volume 600 of *Lecture Notes in Physics*, pages 187–211, Springer-Verlag, Heidelberg, 2002.

[3] C. H. Arns, M. A. Knackstedt, and K. R. Mecke, Characterising the morphology of disordered materials, in *Morphology of Condensed Matter: Physics and Geometry of Spatially Complex systems*, edited by K. R. Mecke and D. Stoyan, volume 600 of *Lecture Notes in Physics*, pages 37–74, Springer-Verlag, Heidelberg, 2002.

[4] M. Lösche and P. Krüger, Morphology of langmuir monolayer phases, in *Morphology of Condensed Matter: Physics and Geometry of Spatially Complex systems*, edited by K. R. Mecke and D. Stoyan, volume 600 of *Lecture Notes in Physics*, pages 37–74, Springer-Verlag, Heidelberg, 2002.

[5] P. H. Gaskell, *Structure and Properties of Glasses - How Far Do We Need To Go?*, Journal of Non-Crystalline Solids **222**, 1 (1997).

[6] W. H. Zachariasen, *Atomic Arrangement in Glass*, Journal of the American Chemical Society **54**, 3841 (1932).

[7] R. J. Bell and P. Dean, *Properties of Vitreous Silica: Analysis of Random Network Models*, Nature **212**, 1354 (1966).

[8] R. J. Bell and P. Dean, *The Structure of Vitreous Silica: Analysis of Random Network Theory*, Philosophical Magzine **25**, 1381 (1972).

[9] D. E. Polk, *Structural Model for Amorphous Silicon and Germanium*, Journal of Non-Crystalline Solids **5**, 365 (1971).

[10] D. E. Polk and D. S. Boudreaux, *Tetrahedrally Coordinated Random-Network Structure*, Physical Review Letters **31**, 92 (1973).

[11] D. Henderson, *Random Tetrahedral Network with Periodic Boundary Conditions*, Journal of Non-Crystalline Solids **16**, 317 (1974).

[12] W. Y. Ching, C. C. Lin, and L. Guttman, *Structural Disorder and Electronic Properties of Amorphous Silicon*, Physical Review B **16**, 5488 (1977).

[13] L. Guttman, Simulation of continuous random network models with periodic boundary conditions, in *American Institute of Physics Conference Proceedings*, pages 224–228, 1974.

[14] L. Guttman, Vibrational spectra of four-coordinated random networks with periodic boundary conditions, in *American Institute of Physics Conference Proceedings*, volume 31, pages 268–272, 1975.

[15] F. Wooten, K. Winer, and D. Weaire, *Computer-Generation of Structural Models of Amorphous Si and Ge*, Physical Review Letters **54**, 1392 (1985).

[16] G. T. Barkema and N. Mousseau, *High-quality Continuous Random Networks*, Physical Review B **62**, 4985 (1985).

[17] R. L. C. Vink, G. T. Barkema, M. A. Stijnman, and R. H. Bisseling, *Device-size Atomistic Models of Amorphous Silicon*, Physical Review B **64**, 5214 (2001).

[18] F. H. Stillinger and T. A. Weber, *Computer Similation of Local Order in Condensed Matter Phases of Silicon*, Physical Review B **31**, 5262 (1985).

[19] R. Biswas and D. R. Hamann, *Interatomic Potentials for Silicon Structure Energies*, Physical Review Letters **55**, 2001 (1985).

[20] J. R. Chelikowsky, J. C. Phillips, M. Kamal, and M. Strauss, *Surface and Thermodynamic Interatomic Force Fields for Silicon Clusters and Bulk Phases*, Physical Review Letters **62**, 292 (1989).

[21] J. Tersoff, *New Empirical Approach for the Structure and Energy of Covalent Systems*, Physical Review B **37**, 6991 (1988).

[22] M. I. Baskes, J. S. Nelson, and A. F. Wright, *Semiempirical Modified Embedded-atom Potential for Silicon and Germanium*, Physical Review B **40**, 6085 (1989).

[23] R. Car and M. Parrinello, *Structural, Dynamical, and Electronic Properties of Amorphous Silicon: An ab initio Molecular Dynamics Study*, Physical Review Letters **60**, 204 (1988).

[24] L. T. Canham, *Silicon Quantum Wire Array Fabrication by Electrochemical and Chemical Dissolution of Wafers*, Applied Physics Letters **57**, 1046 (1990).

[25] D. Buttard, D. Bellet, and G. Dolino, *X-ray-diffraction Investigation of the Anodic Oxidation of Porous Silicon*, Journal of Applied Physics **79**, 8060 (1996).

[26] V. Lehmann and U. Gösele, *Porous Silicon Formation: A Quantum Wire Effect*, Applied Physics Letters **58**, 856 (1991).

[27] A. G. Cullis, L. T. Canham, and P. D. J. Calcott, *The Structural and Luminescence Properties of Porous silicon*, Journal of Applied Physics **82**, 909 (1997).

[28] V. Chin, B. E. Collins, M. J. Sailor, and S. Bhatia, *Compatibility of Primary Hepatocytes with Oxidized Nanoporous Silicon*, Advanced Materials **13**, 1877 (2001).

[29] D. Zhao et al., *Continuous Mesoporous Silica Filmns with Highly Ordered Large Pore Structures*, Advanced Materials **10**, 1380 (1998).

[30] C. McDonagh, B. D. MacCraith, and A. K. McEvoy, *Tailoring of Sol-gel Films for Optical Sensoring of Oxygen in Gas and Aqueous Phase*, Analytical Chemistry **70**, 45 (1998).

[31] Y. Cohen et al., *Spin-on Nanostructured Silicon-Silica Film Displaying Room-Temperature Nanosecond Lifetime Photoluminescence*, Advanced Materials **15**, 572 (2003).

[32] Ö. Dag et al., *Photoluminescent Silicon Clusters in Oriented Hexagonal Mesoporous Silica Film*, Advanced Materials **11**, 474 (1999).

[33] S. Yoshida, T. Hanada, S. Tanabe, and N. Soga, *Blue Photoluminescence from Si-doped Amorphous Silica Films by RF Sputtering*, Japanese Journal of Applied Physics Part I - Regular Papers, Short Notes and Review Papers **35**, 2694 (1996).

[34] T. Morioka et al., *Study of the Structure of Silica Film by Infrared Spectroscopy and Electron Diffraction Analyses*, Monthly Notices of the Royal Astronomical Society **299**, 78 (1998).

[35] V. M. Izgorodin et al., *Composition and Surface Properties of a Silicon Dioxide Film Deposited by a Plasma-Chemical Technique*, High Energy Chemistry **36**, 426 (2002).

[36] K. Z. Baba Kishi, *Scanning Reflection Electron Microscopy of Surface Topography by Diffusely Scattered Electrons in the Scanning Electron Microscope*, Scanning **18**, 315 (1996).

[37] M. Yoshikawa et al., *Characterization of Fluorine-Doped Silicon Dioxide Film by Raman Spectroscopy*, Thin Solid Filmns **310**, 167 (1997).

[38] S. Yoshida, T. Hanada, S. Tanabe, and N. Soga, *Annealing Characteristics of Si Doped Amorphous Silica Films by RF Sputtering*, Journal of Materials Science **34**, 267 (1999).

[39] B. Civalleri et al., *Quantum Mechanical ab initio Characterization of a Simple Periodic Model of the Silica Surface*, Journal of Physical Chemistry B **103**, 2165 (1999).

[40] C. Mischler, W. Kob, and K. Binder, *Classical and Ab-initio Molecular Dynamic Simulation of an Amorphous Silica Surface*, Computer Physics Communications **147**, 222 (2002).

[41] A. Roder, W. Kob, and W. Binder, *Structure and dynamics of amorphous silica surfaces*, Journal of Chemical Physics **114**, 7602 (2001).

[42] C. Wang, N. Kuzuu, and Y. Tamai, *Molecular Dynamics Study on Surface Structure of a-SiO2 by Charge Equilibration Method*, Journal of Non-Crystalline Solids **318**, 131 (2003).

[43] V. I. Bogillo, L. S. Pirnach, and A. Dabrowski, *Monte Carlo Simulation of Silica Surface Dehydroxylation Under Nonisothermal Conditions*, Langmuir **13**, 928 (1997).

[44] G. Hadjisavvas, G. Kopidakis, and P. C. Kelires, *Structural Models of Amorphous Silicon Surfaces*, Physical Review B **64**, 5413 (2001).

[45] K. A. Kilian, D. A. Drabold, and J. B. Adams, *First-Principles Simulations of a-Si and a-Si:H Surfaces*, Physical Review B **48**, 17393 (1993).

[46] J. C. Maxwell, *On the Calculation of the Equilibrium and Stiffness of Frames*, Philosophical Magzine **27**, 294 (1864).

[47] D. J. Jacobs and M. F. Thorpe, *Generic Rigidity Percolation: The Pebble Game*, Physical Review Letters **75**, 4051 (1995).

[48] D. J. Jacobs and M. F. Thorpe, *Generic Rigidity Percolation in Two Dimensions*, Physical Review E **53**, 3682 (1996).

[49] T. S. Tay and W. Whiteley, *Recent Advances in the Generic Rigidity of Structures*, Structural Topology **9**, 31 (1984).

[50] W. Whiteley, Rigidity of molecular structures, in *Rigidity Theory and Applications*, edited by M. F. Thorpe and P. M. Duxbury, pages 21-46, New York, 1999, Kluwer Academic/Plenum Publishers.

[51] D. J. Jacobs, *Generic Rigidity in Three-Dimensional Bond Bending Networks*, Journal of Physics A: Mathematical and General **31**, 6653 (1998).

[52] M. B. Berry et al., *The Closed Conformation of a Highly Flexible Protein - the Structure of Escherichia-Coli Adenylate Kinase with Bound AMP and AMPPNP*, Proteins: Structure, Function, and Genetics **19**, 183 (1994).

[53] C. W. Müller, G. J. Schlauderer, J. Reinstein, and G. E. Schulz, *Adenylate Kinase Motions During Catalysis: an Energetic Counterweight Balancing Substrate Binding*, Structure **4**, 147 (1996).

[54] G. J. Schlauderer and G. E. Schulz, *The Structure of Bovine Mitochondrial Adenylate Kinase: Comparison with Isoenzymes in Other Compartments*, Protein Science **5**, 434 (1996).

[55] M. Gerstein, G. Schulz, and C. Chothia, *Domain Closure in Adenylate Kinase: Joints on Either Side of Two Helices Close Like Neighboring Fingers*, Journal of Molecular Biology **229**, 494 (1993).

[56] A. Crivici and M. Ikura, *Molecular and Structural Basis of Target Recognition by Calmodulin*, Annual Review of Biophysics and Biomolecular Structure **24**, 85 (1995).

[57] A. Houdusse, M. Silver, and C. Cohen, *A Model of Ca^{2+}-free Calmodulin Binding to Unconventional Myosins Reveals How Calmodulin Acts as a Regulatory Switch*, Structure **4**, 1475 (1996).

[58] Z. Zhang et al., *Electron Transfer by Domain Movement in Cytochrome bc_1*, Nature **392**, 677 (1998).

[59] C. M. Deane and T. L. Blundell, *A Novel Exhaustive Search Algorithm for Predicting the Conformation of Polypeptide Segments in Proteins*, Proteins: Structure, Function, and Genetics **40**, 135 (2000).

[60] P. Güntert et al., *Conformational Analysis of Protein and Nucleic Acid Fragments ith the New Grid Search Algorithm FOUND*, Journal of Biomolecular NMR **12**, 543 (1998).

[61] R. A. Dammkoehler, S. F. Karasek, E. F. B. Shands, and G. R. Marshall, *Sampling Conformational Hyperspace: Techniques for Improving Completeness*, Journal of Computer-Aided Molecular Design **9**, 491 (1995).

[62] W. Wang, O. Donini, C. M. Reyes, and P. A. Kollman, *Biomolecular Simulations: Recent Developments in Force Fields, Simulations of Enzyme Catalysis, Protein-Ligand, Protein-Protein, and Protein-Nucleic Acid Noncovalent Interactions*, Annual Review of Biophysics and Biomolecular Structure **30**, 211 (2001).

[63] N. Nakajima, H. Nakamura, and A. Kidera, *Multicanonical Ensemble Generated by Molecular Dynamics Simulation for Enhanced Conformational Sampling of Peptides*, Journal of Physical Chemistry B **101**, 817 (1997).

[64] U. H. E. Hansmann, Y. Okamoto, and F. Eisenmenger, *Molecular Dynamics, Langevin and Hybrid Monte Carlo Simulations in a Multicanonical Ensemble*, Chemical Physics Letters **259**, 321 (1996).

[65] K. B. Kamal and J. P. Sethna, *Multicanonical Methods, Molecular Dynamics, and Monte Carlo Methods: Comparison for Lennard-Jones Glasses*, Physical Review E **57**, 2553 (1998).

[66] Y. Sugita and Y. Okamoto, *Replica-exchange Molecular Dynamics Method for Protein Folding*, Chemical Physics Letters **314**, 141 (1999).

[67] Y. Sugita and Y. Okamoto, *Replica-exchange Multicanonical Algorithm and Multicanonical Replica-exchange Method for Simulating Systems with Rough Energy Landscape*, Chemical Physics Letters **329**, 261 (2000).

[68] B. A. Berg and T. Neuhaus, *Multicanonical Algorithms for First Order Phase Transitions*, Physics Letters B **267**, 249 (1991).

[69] J. Lee, *New Monte Carlo Algorithm: Entropic Sampling*, Physical Review Letters **71**, 211 (1993).

[70] J. Higo et al., *Two-Component Multicanonical Monte Carlo Method for Effective Conformation Sampling*, Journal of Computational Chemistry **18**, 2086 (1997).

[71] U. H. E. Hansmann and Y. Okamoto, *Prediction of Peptide Conformation by Multicanonical Algorithm - New Approach to the Multiple-minima Problem*, Journal of Computational Chemistry **14**, 1333 (1993).

[72] R. H. Swendsen and J.-S. Wang, *Replica Monte Carlo Simulation of Spin-Glasses*, Physical Review Letters **57**, 2607 (1986).

[73] N. Gō and H. A. Scheraga, *Ring Closure and Local Conformatinal Deformations of Chain Molecules*, Macromolecules **3**, 178 (1970).

[74] N. Gō and H. A. Scheraga, *Calculation of the Conformation of the Pentapeptide cyclo-(Glycylglycylglycylprolylprolyl). I. A Complete Energy Map*, Macromolecules **3**, 188 (1970).

[75] N. Gō and H. A. Scheraga, *Ring Closure in Chain Molecules with C_n, I, or S_{2n} Symmetry*, Macromolecules **6**, 273 (1973).

[76] M. Dygert, N. Gō, and H. A. Scheraga, *Use of a Symmetry Condition to Compute teh Conformations of Gramicidin S*, Macromolecules **8**, 750 (1975).

[77] N. Gō and H. A. Scheraga, *Calculation of the Conformation of cyclo-Hexaglycyl*, Macromolecules **6**, 525 (1973).

[78] N. Gō and H. A. Scheraga, *Calculation of the Conformation of cyclo-Hexaglycyl 2: Application of a Monte Carlo Method*, Macromolecules **11**, 552 (1978).

[79] W. J. Wedemeyer and H. A. Scheraga, *Exact Analytical Loop Closure in Proteins Using Polynomial Equations*, Journal of Computational Chemistry **20**, 819 (1999).

[80] M. G. Wu and M. W. Deem, *Efficient Monte Carlo methods for cyclic peptides*, Molecular Physics **97**, 559 (1999).

[81] A. A. Canutescu and R. L. Dunbrack, Jr, *Cyclic Coordinate Descent: A Robotics Algorithm for Protein Loop Closure*, Protein Science **12**, 963 (2003).

[82] R. E. Bruccoleri and M. Karplus, *Chain Closure with Bond Angle Variations*, Macromolecules **18**, 2767 (1985).

[83] K. A. Palmer and H. A. Scheraga, *Standard-Geometry Chains Fitted to X-ray Derived Structures: Validation of the Rigid-Geometry Approximation I: Chain Closure through a Limited Search of "Loop" Conformations*, Journal of Computational Chemistry **12**, 505 (1991).

[84] K. A. Palmer and H. A. Scheraga, *Standard-Geometry Chains Fitted to X-ray Derived Structures: Validation of the Rigid-Geometry Approximation II: Systematic Searches for Short Loops in Proteins: Applications to Bovine Pancreatic Ribonuclease A and Human Lysozyme*, Journal of Computational Chemistry **13**, 329 (1992).

[85] J. Moult and M. N. G. James, *An Algorithm for Determining the Conformations of Polypeptide Segments in Proteins by Systematic Search*, Proteins: Structure, Function, and Genetics **1**, 146 (1986).

[86] M. J. Dudek and H. A. Scheraga, *Protein Structure Prediction Using a Combination of Sequence Homology and Global Energy Minimization I. Global Energy Minimization of Surface Loops*, Journal of Computational Chemistry **11**, 121 (1990).

[87] K. D. Gibson and H. A. Scheraga, *Energy Minimization of Rigid-Geometry Polypeptides with Exactly Closed Disufide Loops*, Journal of Computational Chemistry **18**, 403 (1997).

[88] D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe, *Protein Flexibility Prediction Using Graph Theory*, Proteins: Structure, Function, and Genetics **44**, 150 (2002).

[89] A. J. Rader, B. M. Hespenheide, L. A. Kuhn, and M. F. Thorpe, *Protein Unfolding: Rigidity Lost*, Proceedings of the National Academy of Sciences of the United States of America **99**, 3540 (2002).

[90] B. M. Hespenheide, A. J. Rader, M. F. Thorpe, and L. A. Kuhn, *Identifying Protein Folding Cores from the Evolution of Flexible Regions During Unfolding*, Journal of Molecular Graphics and Modelling **21**, 195 (2002).

[91] P. N. Keating, *Effect of Invariance Requirements on the Elastic Strain Energy of Crystals with Application to the Diamond Structure*, Physical Review **145**, 637 (1966).

[92] B. R. Djordjević, M. F. Thorpe, and F. Wooten, *Computer Model of Tetrahedral Amorphous Diamond*, Physical Review B **52**, 5685 (1995).

[93] Y. Tawada, H. Okamoto, and Y. Hamakawa, *a-SiC:H/a-Si:H Heteriojunction Solar Cell Having Moer Than 7.1% Conversion Efficiency*, Applied Physics Letters **39**, 237 (1981).

[94] H. Iida et al., *Efficiency of the a-Si:H Solar Cell and Grain Size of SNO_2 Transparant Conductive Film*, IEEE Electron Device Letters **4**, 157 (1983).

[95] S. C. Deane, F. J. Clough, W. I. Milne, and M. J. Powell, *The Role of the Gate Insulator in the Defect Pool Model for Hydrogenated Amorphous Silicon Thin Film Transistor Characteristics*, Journal of Applied Physics **73**, 2895 (1993).

[96] S. K. Kim, Y. J. Choi, K. S. Cho, and J. Jang, *Coplanar Amorphous Silicon Thin Film Transistor Fabricated by Inductively Coupled Plasma Chemical Vapor Deposition*, Journal of Applied Physics **84**, 4006 (1998).

[97] M. F. Thorpe, V. V. Levashov, M. Lei, and S. J. L. Billinge, Notes on the analysis of data for pair distribution functions, in *From Semiconductors to Proteins: Beyond the Average Structure*, edited by S. J. L. Billinge and M. F. Thorpe, pages 105–128, Kluwer Academic/Plenum Publishers, New York, 2002.

[98] J. Y. Pivan, O. Achak, L. Michéle, and L. Daniel, *The Novel Thiogermanate $[(CH_3)_4N]_4\,Ge_4S_{10}$ with a High Cubic Cell Volumn. Ab Initio Structure Determination from Conventional X-ray Powder Diffraction*, Chemistry of Materials **6**, 827 (1994).

[99] C. L. Bowes et al., *Dimetal Linked Open Frameworks: $[(CH_3)_4N]_2\,(Ag_2, Cu_2)Ge_4S_{10}$*, Chemistry of Materials **8**, 2147 (1996).

[100] O. M. Yaghi, Z. Sun, D. A. Richardson, and T. L. Groy, *Directed Transformation of Molecules to Solids: Synthesis of a Microporous Sulfide from Molecular Germanium Sulfide Cages*, Journal of the American Chemical Society **116**, 807 (1994).

[101] C. Cahill and J. B. Parise, *Synthesis and Structure of $MnGe_4S_{10} \cdot (C_6H_{14}N_2) \cdot 3H_2O$: A Novel Sulfide Framework Analogous to Zeolite Li-A (BW)*, Chemistry of Materials **9**, 807 (1997).

[102] F. Bonhomme and M. G. Kanatzidis, *Structurally Characterized Mesostructured Hybrid Surfactant-Inorganic Lamellar Phases Containing the Adamantane $[Ge_4S_{10}]^{4-}$ Anion: Syntesis and Properties*, Chemistry of Materials **10**, 1153 (1998).

[103] M. J. MacLachlan, N. Coombs, and G. Ozin, *Non-aqueous Supramolecular Assembly of Mesostructured Metal Germanium Sulphides from $(Ge_4S_{10})^{4-}$ Clusters*, Nature **397**, 681 (1999).

[104] M. J. MacLachlan et al., *Mesostructured Metal Germanium Sulfides*, Journal of the American Chemical Society **121**, 12005 (1999).

[105] K. K. Rangan et al., *Aqueous Mediated Synthesis of Mesostructured Manganese Germanium Sulfide with Hexagonal Order*, Chemistry of Materials **11**, 2629 (1999).

[106] M. Wachhold et al., *Mesostructured Metal Germanium Sulfide and Selenide Materials Based on the Tetrahedral* $[Ge_4S_{10}]^{4-}$ *and* $[Ge_4Se_{10}]^{4-}$ *Units: Surfactant Templated Three-Dimensional Disordered Frameworks Perforated with Worm Holes*, Journal of Solid State Chemistry **152**, 21 (2000).

[107] The amorphous GaAs model is generously provided by G. T. Barkema and N. Mousseau.

[108] V. Petkov, K. K. Rangan, M. G. Kanatzidis, and S. J. L. Billinge, Structure of crystallographically challenged materials by profile analysis of atomic pair distribution functions: Study of $limos_2$ and mesostructured $mnge_4s_{10}$, in *Materials Research Society Proceedings*, edited by P. G. Allen, S. M. Mini, D. L. Perry, and S. R. Stock, volume VI of *Application of Synchrotron Radiation Techniques to Materials Science*, page EE1.5, Materials Research Society, 2001.

[109] M. F. Thorpe, M. V. Chubynsky, D. J. Jacobs, and J. C. Phillips, *Non-Randomness in Network Glasses and Rigidity*, Glass Physics and Chemistry **21**, 160 (2001).

[110] G. N. Ramachandran and V. Sasiskharan, *Conformation Of Polypeptides And Proteins*, Advances in Protein Chemistry **23**, 283 (1968).

[111] E. N. Baker and R. E. Hubbard, *Hydrogen Bonding in Globular Proteins*, Progress in Biophysics and Molecular Biology **44**, 97 (1984).

[112] B. I. Dahiyat, D. B. Gordon, and S. L. Mayo, *Automated Design of the Surface Positions of Protein Helices*, Protein Science **6**, 1333 (1997).

[113] F. Fabiola, R. Bertram, A. Korostelev, and M. S. Chapman, *An Improved Hydrogen Bond Potential: Impact on Medium Resolution Protein Structures*, Protein Science **11**, 1415 (2002).

[114] R. Balasubramanian, R. Chidambaram, and G. N. Ramachandran, *Potential Functions for Hydrogen Bond Interactions II: Formulation of an Empirical Potential Function*, Biochimica et Biophysica Acta, International Journal of Biochemistry and Biophysics **221**, 196 (1970).

[115] W. D. Cornell et al., *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules*, Journal of the American Chemical Society **117**, 5179 (1995).

[116] D. A. Pearlman et al., *AMBER, A Package of Computer Programs for Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules*, Computational Physics Communication **91**, 1 (1995).

[117] A. D. J. MacKerell et al., *All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins*, Journal of Physical Chemistry B **102**, 3586 (1998).

[118] B. R. Brooks et al., *CHARMM - A Program For Macromolecular Energy, Minimization, and Dynamics Calculations*, Journal of Computational Chemistry **4**, 187 (1983).

[119] M. I. Zavodszky, personal communication.

[120] D. F. Stickle, L. G. Presta, K. A. Dill, and G. D. Rose, *Hydrogen-Bonding in Globular-Proteins*, Journal of Molecular Biology **226**, 1143 (1992).

[121] B. H. Hespenheide et al., Improved definition of hydrophobic interactions in rigidity analysis, to be published.

[122] D. C. Liu and J. Nocedal, *On the Limited Memory Method for Large Scale Optimization*, Mathematical Programming B **45**, 503 (1989).

[123] N. Metropolis et al., *Equation of State Calculations by Fast Computing Machines*, Journal of Chemical Physics **21**, 1087 (1953).

[124] P. Spellucci, *A SQP Method for General Non-linear Programs Using Only Equality Constrained Subproblems*, Mathematical Programming **82**, 413 (1998).

[125] A. L. Morris, M. W. MacArthur, E. G. Hutchioson, and J. M. Thornton, *Stereochemical Quality of Protein Structure Coordinates*, Proteins: Structure, Function, and Genetics **12**, 345 (1992).

[126] N. L. Allinger, Y. H. Yuh, and J. H. Lii, *Molecular Mechanics. The MM3 Force Field for Hydrocarbons: I*, Journal of the American Chemical Society **111**, 8551 (1989).

[127] http://dasher.wustl.edu/tinker.

[128] M. Miller et al., *Structure Of Complex Of Synthetic HIV-1 Protease With a Substrate-Based Inhibitor At 2.3Å Resolution*, Science , 1149 (1989).

[129] M. Kumar and M. V. Hosur, *Adaptability And Flexibility Of HIV-1 Protease*, European Journal of Biochemistry **270**, 1231 (2003).

[130] S. Spinelli et al., *The 3-Dimensional Structures of the Aspartyl Protease from the HIV-1 Isolate Bru*, Biochimie **73**, 1391 (1990).

[131] R. Ishima et al., *Flap Opening and Dimer-Interface Flexibility in the Free and Inhibitor Bound HIV Protease and Their Implications for Function*, Structure **7**, 1047 (1999).

[132] D. I. Freedberg et al., *Rapid Structural Fluctuations Of The Free HIV Protease Flaps In Solution: Relationship To Crystal Structures And Comparison With Predictions Of Dynamics Calculations*, Protein Science **11**, 221 (2002).

[133] W. R. P. Scott and C. A. Schiffer, *Curling of Flap Tips in HIV-1 Protease as a Mechanism for Substrate Entry and Tolerance of Drug Resistance*, Structure **8**, 1259 (2000).

[134] H. A. Carlson, personal communications.

[135] C. Shih et al., *LY231514, A Pyrrolo[2,3-d]pyrimidine-based Antifolate That Inhibits Multiple Folate-requiring Enzymes*, Cancer Research **57**, 1116 (1997).

[136] http://www.pdb.org.

[137] M. R. Sawaya and J. Kraut, *Loop and Subdomain Movements in Mechanism of Escherichia Coli Dihydrofolate Reductase: Crystallographic Evidence*, Biochemistry **36**, 586 (1997).

[138] P. Shrimpton, A. Mullaney, and R. K. Allemann, *Functional Role for Tyr31 in the Catalytic Cycle of Chicken Dihydrofolate Reductase*, Proteins: Structure, Function, and Genetics **51**, 216 (2003).

[139] V. Cody et al., *Ligand-induced Conformational Changes in the Crystal Structure of Pneumocystis Carinii Dihydrofolate Reductase Complexes with Folate and NADP+*, Biochemistry **38**, 4303 (1999).

[140] V. M. Reyes, M. R. Sawaya, K. A. Brown, and J. Kraut, *Isomophous Crystal Structures of Escherichia Coli Dihydrofolate Reducatase Complexed with Folate, 5-Deazafolate, and 5,10-dideazatetrahydrofolate: Mechanistic Implications*, Biochemistry **34**, 2710 (1995).

[141] C. J. Falzone, P. E. Wright, and S. J. Benkovic, *Dynamics of a Flexible Loop in Dihydrofolate Reductase from Escherichia Coli and Its Implication for Catalysis*, Biochemistry **33**, 439 (1994).

[142] D. Antoniou and S. D. Schwartz, *Internal Enzyme Motions as a Source of Catalytic Activity: Rate-Promoting Vibrations and Hydrogen Tunneling*, Journal of Physical Chemistry B **105**, 5553 (2001).

[143] P. K. Agarwal et al., *Network of Coupled Promoting Motions in Enzyme Catalysis*, Proceedings of the National Academy of Sciences of the United States of America **99**, 2794 (2002).

[144] G. G. Hammes, *Multiple Conformational Changes in Enzyme Catalysis*, Biochemistry **41**, 8221 (2002).

[145] E. Y. Lau and J. T. Gerig, *Effects of Fluorine Substtution on the Structure and Dynamics of Complexes of Dihydrofolate Reductase (Escherichia Coli)*, Biophysical Journal **73**, 1579 (1997).

[146] J. L. Radkiewicz and C. L. Brooks III, *Protein Dynamics in Enzymatic Catalysis: Exploration of Dihydrofolate Reductase*, Journal of the American Chemical Society **122**, 225 (2000).

[147] T. H. Rod, J. L. Radkiewicz, and C. L. Brooks III, *Correlated Motion and the Effect of Dismal Mutations in Dihydrofolate Reductase*, Proceedings of the National Academy of Sciences of the United States of America **100**, 6980 (2003).

[148] G. Vriend, *WHAT IF: a Molecular Modelling and Drug Design Program*, Journal of Molecular Modelling and Graphics **8**, 52 (1990).

[149] H. Gould and J. Tobochnik, *An Introduction to Computer Simulation Methods*, Addison-Wesley Publishing Company, 1998.

[150] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi, *Optimization by Simulated Annealing*, Science **220**, 671 (1983).

[151] P. L'Ecuyer and S. Cote, *Implementing a Random Number Package with Splitting Facilities*, ACM Transactions on Mathematical Software **17**, 98 (1991).

[152] http://www.accelrys.com.

[153] A. P. Korn and D. R. Rose, *Torsion Angle Differences as a Means of Pinpointing Local Polypeptide Chain Trajectory Changes for Identical Proteins in Different Conformational States*, Protein Engeering **7**, 961 (1994).

[154] C. Vieille, personal communication.

[155] J. Janin, S. Wodak, M. Levitt, and B. Maigret, *Conformation of Amino-Acid Side-Chains in Proteins*, Journal of Molecular Biology **125**, 357 (1978).

[156] R. L. Dunbrack Jr. and K. Martin, *Backbone-Dependent Rotamer Library for Protein Application to Side-Chain Prediction*, Journal of Molecular Biology **230**, 543 (1993).

[157] S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson, *The Penultimate Rotamer Library*, Proteins: Structure, Function, and Genetics **40**, 389 (2000).

[158] H. Frauenfelder et al., *The Role of Structure, Energy Landscape, Dynamics and Allostery in the Enzymatic Function of Myoglobin*, Proceedings of the National Academy of Sciences of the United States of America **98**, 2370 (2001).

[159] K. W. Plaxco, K. T. Simons, I. Ruczinski, and D. Baker, *Topology, Stability, Sequence and Length: Defining the Determinants of Two-state Protein Folding Kinetics*, Biochemistry **39**, 11177 (2000).

[160] T. Kortemme et al., *Similarities Between the Spectrin SH3 Domain Denatured State and its Folding Transition State*, Journal of Molecular Biology **297**, 1217 (2000).

[161] D. T. Leeson et al., *Protein Folding and Unfolding on a Complex Energy Landscape*, Proceedings of the National Academy of Sciences of the United States of America **97**, 2527 (2000).

[162] W. A. Eaton et al., *Fast Kinetics and Mechanisms in Protein Folding*, Annual Review of Biophysics and Biomolecular Structures **29**, 327 (2000).

[163] M. Karplus, *Aspects of Protein Reaction Dynamics: Deviation from Simple Behavior*, Journal of Physical Chemistry B **104**, 11 (2000).

[164] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, *theory of Protein Folding: The Energy Landscape Perspective*, Annual Reviews of Physical Chemistry **48**, 545 (1997).

[165] D. J. Brockwell, D. A. Smith, and S. E. Radford, *Protein Folding Mechanisms: New Methods and Emerging Ideas*, Current Opinion in Structural Biology **10**, 16 (2000).

[166] L. Angelani et al., *Saddles in the Energy Landscape Probed by Supercooled Liquids*, Physical Review Letters **85**, 5356 (2000).

[167] N. Mousseau and G. T. Berkema, *Traveling Through Potential Energy Landscapes of Disordered Materials: The Activation-Relaxation Technique*, Physical Review E **57**, 2419 (1998).

[168] N. Mousseau, P. Derreumaux, G. T. Barkema, and R. Malek, *Sampling Activated Mechanisms in Proteins with the Activation-Relaxation Technique*, Journal of Molecular Graphics and Modelling **19**, 78 (2001).

[169] T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, *DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases*, Journal of Computer-Aided Molecular Design **15**, 411 (2001).

[170] D. M. Lorber and B. K. Shoichet, *Flexible Ligand Docking Using Conformational Ensembles*, Protein Science **7**, 938 (1998).

[171] J. A. Given and M. K. Gilson, *A Hierarchical Method for Generating Low-Energy Conformers of a Protein-Ligand Complex*, Proteins: Structure, Function, and Genetics **33**, 475 (1998).

[172] A. R. Leach, *Ligand Docking to Proteins with Discrete Side-chain Flexibility*, Journal of Molecular Biology **235**, 345 (1994).

[173] R. Najmanovich, J. Kuttner, V. Sobolev, and M. Edelman, *Side-chain Flexibility in Proteins upon Ligand Binding*, Proteins: Structure, Function, and Genetics **39**, 261 (2000).

[174] H. Clauben, C. Buning, M. Rarey, and T. Lengauer, *FLEXE: Efficient Molecular Docking Considering Protein Structure Variations*, Journal of Molecular Biology **308**, 377 (2001).

[175] H. A. Carlson and J. A. McCammon, *Accommodating Protein Flexibility in Computational Drug Design*, Molecular Pharmacology **57**, 213 (2000).

[176] M. I. Zavodszky, M. Lei, M. F. Thorpe, and L. A. Kuhn, Modeling protein main chain flexibility for docking, to be published.

[177] M. I. Zavodszky, *Modeling Flexibility in Protein-Ligand Recognition*, PhD thesis, Michigan State University, 2003.

[178] V. Schnecke et al., *Screening a Peptidyl Database for Potential Ligands to Proteins Including Side-Chain Flexibility*, Proteins: Structure, Function, and Genetics **33** (1998).

[179] V. Schnecke and L. A. Kuhn, *Virtual Screening with Solvation and Ligand-Induced Complementarity*, Perspectives in Drug Design and Discovery **20**, 171 (2000).

[180] G. J. Kleywegt and T. A. Jones, *Phi/Psi-chology: Ramachandran Revisited*, Structure **4**, 1395 (1996).

[181] S. C. Lovell et al., *Structure Validation by C_α Geometry: ϕ, ψ and C_β Deviation*, Proteins: Structure, Function, and Genetics **50**, 437 (2003).

[182] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, *PROCHECK: a program to check the stereochemical quality of protein structures*, Journal of Applied Crystallography **26**, 283 (1993).